Application of Next Generation Sequencing In the Characterisation of Variants Causing Haemoglobinopathies

Shooter, Claire Miranda

Awarding institution:
King's College London

King's College London

PhD Thesis

Application of Next Generation Sequencing In the Characterisation of Variants Causing Haemoglobinopathies

Author:

Claire Shooter

Supervisors:

Prof. Swee Lay Thein

Dr. Barnaby Clark

A thesis submitted for the degree of Doctor of Philosophy

In the Department of Molecular Haematology

Faculty of Life Sciences and Medicine

Acknowledgements

I have been hugely lucky over the course of my PhD to receive help from so many sources. First and foremost, I would like to extend my thanks to my supervisors, Professor Swee Lay Thein and Dr Barnaby Clark for their constant support over the last four years, and the knowledge, energy and enthusiasm they dedicated to this project.

Every person in the Department of Molecular Pathology (now ViaPath) at King's College Hospital has contributed to my work in one way or another, so I would like to thank them all for their knowledge and help in the lab. More importantly however, I would like to thank them for the constant source of excellent conversation, with nothing deemed off-topic for a working environment. I particularly enjoyed the time we were winched off the building for the greater good, the laughs at our Christmas parties and the abundance of cake in the office.

The Department of Molecular Haematology at King's College London were also instrumental to my completion of this work. I received a huge amount of support from the whole department, but particularly from Helen for her help in the lab, and Stephan for always being on hand to give me advice, and to indulge whatever half-baked ideas I had in any given week.

I also invaded the Genomics Facility at Guy's Campus and the Institute of Psychiatry (both King's College London) for several weeks during my time here. I would like to extend gratitude to Efterpi Papouli, Athina Gkazi and Bradley Smith for allowing me to run amok in their laboratories and for their commiserations when everything went wrong.

"*Carefully directed ignorance is the key to all knowledge*" – Terry Pratchett

Abstract

Haemoglobinopathies are a group of diseases caused by abnormal structure, function or quantity of the globin chain subunits that form haemoglobin. They are a global health problem with 7% of the population being asymptomatic carriers. The UK has implemented new-born screening and antenatal screening to identify affected individuals. A wide variety of variants causing haemoglobinopathies have been documented, from single base substitutions to large insertions, deletions and complex rearrangements. Next Generation Sequencing (NGS) could streamline laboratory diagnosis as it has the potential to identify both single nucleotide changes and large rearrangements in a single assay. With this objective in mind, we evaluated NGS and its ability to detect all varieties of causative variants in thalassaemia. DNA samples from persons with known thalassaemia-causing mutations and unknown novel variants were selected for NGS. DNA samples were fragmented and prepared for sequencing. The two genomic regions that are affected by variants causing thalassaemia, were enriched for using in-solution bait capture and sequenced on an Illumina platform. Sequence 'reads' of the DNA fragments were aligned to a reference sequence of the human genome using NextGene software (SoftGenetics) and sequence indels and small variants were reported. Dosage changing events were identified by changes in coverage between test and negative control samples. The breakpoints of structural variants could be identified by the presence of sequences that spanned the breakpoint region. The method was honed by analysis of samples with known, previously characterised variants and tested in samples with unknown novel variants. The findings were confirmed by Gap-PCR and sanger sequencing. We conclude that NGS is a technique that represents an improvement on the current diagnostic standard. An examination of the features of the structural rearrangements identified in this study revealed that multiple mutagenic mechanisms contribute to the range of variants that affect the alpha and beta globin gene loci.

Table of Contents

Supplementary 1: Shooter, C., H. Rooks, S. L. Thein and B. Clark (2015). "*Next Generation Sequencing Identifies a Novel Rearrangement in the HBB Cluster Permitting to-the-Base Characterisation.*" Human Mutation 36(1): 142-150.

Supplementary 2: Shooter, Claire, et al. "*First Reported Duplication of the Entire Beta Globin Gene Cluster Causing an Unusual Sickle Cell Trait Phenotype*." BritishJournal of Haematology (2014)

List of Figures

List of Tables

Glossary

All terms are defined in the text upon first use

p. Falciparum	*Plasmodium Falciparum*
BAC	Bacterial Artificial Chromosome
CAE gel test	Cellulose Acetate Gel Test
CGH Array	Comparative Genomic Hybridization
CNC	Computer Numerical Control
CMT1A	Charcott-Marie Tooth Syndrome
DNA	Deoxyribonucleic Acid
DSB	Double-stranded break
E.Coli	*Escherichia coli*
FASTA	file format
FASTQ	file format
FBC	Full Blood Count
FoSTeS	Fork Stalling and Template Switching
gDNA	genomic DNA
HbA	Haemoglobin A
HbA_2	Delta Haemoglobin
HBA1	Alpha Globin 1 Gene
HBA2	Alpha Globin 2 Gene
HBB	Beta Globin Gene
HBD	Delta Globin Gene
HbF	Foetal Haemoglobin

HBE	Epsilon Globin Gene
HBG1	Gamma Globin 1 Gene
HBG2	Gamma Globin 2 Gene
HBM	Mu Globin Gene
HBS	Sickle Globin Gene
HbS	Sickle Haemoglobin
HR	Homologous Recombination
HPLC	High Performance Liquid Chromatography
HPFH	Hereditary Persistence of Foetal Haemoglobin
IEF Gel Test	Isoelectric Focussing Gel Test
KCL	King's College London
KCH	King’s College Hospital
LINE repeat	Long INterspersed Element repeat
LTR	Long Terminal Repeat
MCH	Mean Corpuscular Haemoglobin
MCV	Mean Corpuscular Volume
MLPA	Multiplex Dependent Ligation Probe Amplification
MMEJ	Microhomology-mediated end joining
NAHR	Non-Allelic Homologous Recombination
NGS	Next Generation Sequencing
NHEJ	Non-Homologous End Joining
NIPND	Non-Invasive Prenatal Diagnosis
NTC	No Template Control

NTDT	Non-Transfusion Dependent Beta Thalassaemia
ORI	Origin of Replication
PCR	Polymerase Chain Reaction
pHBB	Pseudo Beta Globin Gene
PND	Prenatal Diagnosis
qPCR	Quantitative Polymerase Chain Reaction
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic Acid
ROA	Relative of Affected
ROI	Region of Interest
RPKM	Repeats Per Kilobase Exon Model Per Million Mapped Reads
SSA	Single Strand Annealing
SINE repeat	Short INterspersed Element repeat
STR	Short Tandem Repeat
TDBT	Transfusion Dependent Beta Thalassaemia
ΦX174	Phi X 174 bacteriophage

Abbreviations

Chr	Chromosome
StDev	Standard Deviation
NegC StDev	Standard Deviation from Negative Control Average
bp	base pair(s)
Kb	Kilobase(s)
Mb	Megabase(s)
cM	CentiMorgan

Greek Letters

α	Alpha
β	Beta
γ	Gamma
δ	Delta
ε	Epsilon
ζ	Zeta
μ	Mu
θ	Theta

Introduction

Haemoglobinopathies are a group of disorders caused by genetic variants that prevent the adequate production or function of haemoglobin, the molecule responsible for oxygen transport around the body. These diseases can be extremely heterogeneous in both their underlying genetic cause and clinical severity. New DNA sequencing technologies have the potential to streamline the process of identifying genetic variants that cause haemoglobinopathies in a diagnostic setting. These new technologies must be evaluated for their utility in diagnosing these disorders, as their implementation could improve patient care.

Haemoglobin

The Haemoglobin Molecule

Red blood cells are responsible for transporting gasses around the body: they supply oxygen from the lungs to other tissues and also from a pregnant woman to her foetus. They also remove the waste gasses cells produce, ferrying them back to the lungs for expulsion. The molecule within red blood cells that permits gas transport is haemoglobin. Haemoglobin is a metalloprotein made up for four subunits: two alpha (or alpha-like) subunits and two beta (or beta-like) subunits (see Figure 1). Each of the four units contains a heme group with an iron core to which oxygen can bind. One molecule of haemoglobin can therefore transport four oxygen molecules simultaneously. A single red blood cell may contain enough haemoglobin to bind 1 billion molecules of oxygen.

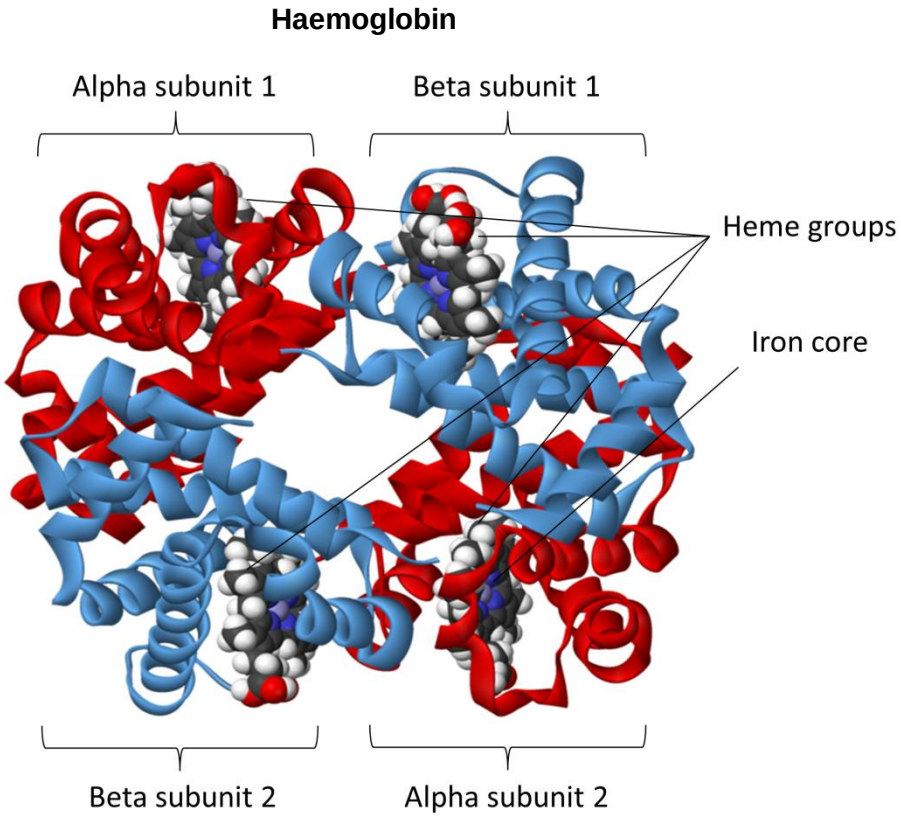


Figure 1: Haemoglobin. A molecule of haemoglobin is made up of four subunits; two alpha like subunits and two beta like subunits. Each subunit contains a heme group with a ferrous core to which an oxygen molecule can reversibly bind (Image adapted from (Wheeler 2007)).

Haemoglobin takes advantage of 'cooperative binding' to deliver oxygen from areas where its partial pressure is very high (the lungs) to areas where it is low. 'Cooperative Binding' means that as oxygen binds to haemoglobin, the shape of the molecule changes in a way that increases its affinity for additional oxygen atoms. Therefore, in an oxygenous environment a lot of oxygen binds to haemoglobin, while in an oxygen poor environment it is released readily. Haemoglobin's oxygen affinity is reduced by hydrogen ions and carbon dioxide, both of which are found in respiring tissues. The higher the respiration rate of a tissue, the greater the release of oxygen it triggers in nearby red blood cells. This is known as the Bohr Effect (Benesch and Benesch 1961, Berg JM, Berg et al. 2002).

Every cell in the body requires oxygen in order to respire and creates carbon dioxide as a waste product of respiration. It is therefore critical to the operation of all tissues and organs in the body that haemoglobin performs its function correctly. The structure and production of haemoglobin are both determined by DNA. Iron deficiency can have similar effects but for the individuals in this study iron deficiency had been excluded. Heme synthesis disorders also occur, but are rarer and individuals that have problems

with heme synthesis suffer from porphyria and present with a different phenotype. This is beyond the remit of this investigation and will not be discussed further.

The Globin Genes

The structure of the globin subunit proteins (or globin chains) are encoded by genes. The sequence of nucleotide bases within the gene dictates the sequence of amino acids that must be joined together to construct these proteins. Humans have eight genes which code for haemoglobin subunits, as well as five globin-like pseudogenes which are organised into two clusters on chromosomes 11 and 16 (See Figure 2). The globin genes are highly similar in their structure, sequence and function. The subtle differences in their sequence and expression confer different physical properties and functions (Czelusniak J 1982, Steinberg and Adams 1991)

The Globin Gene Loci

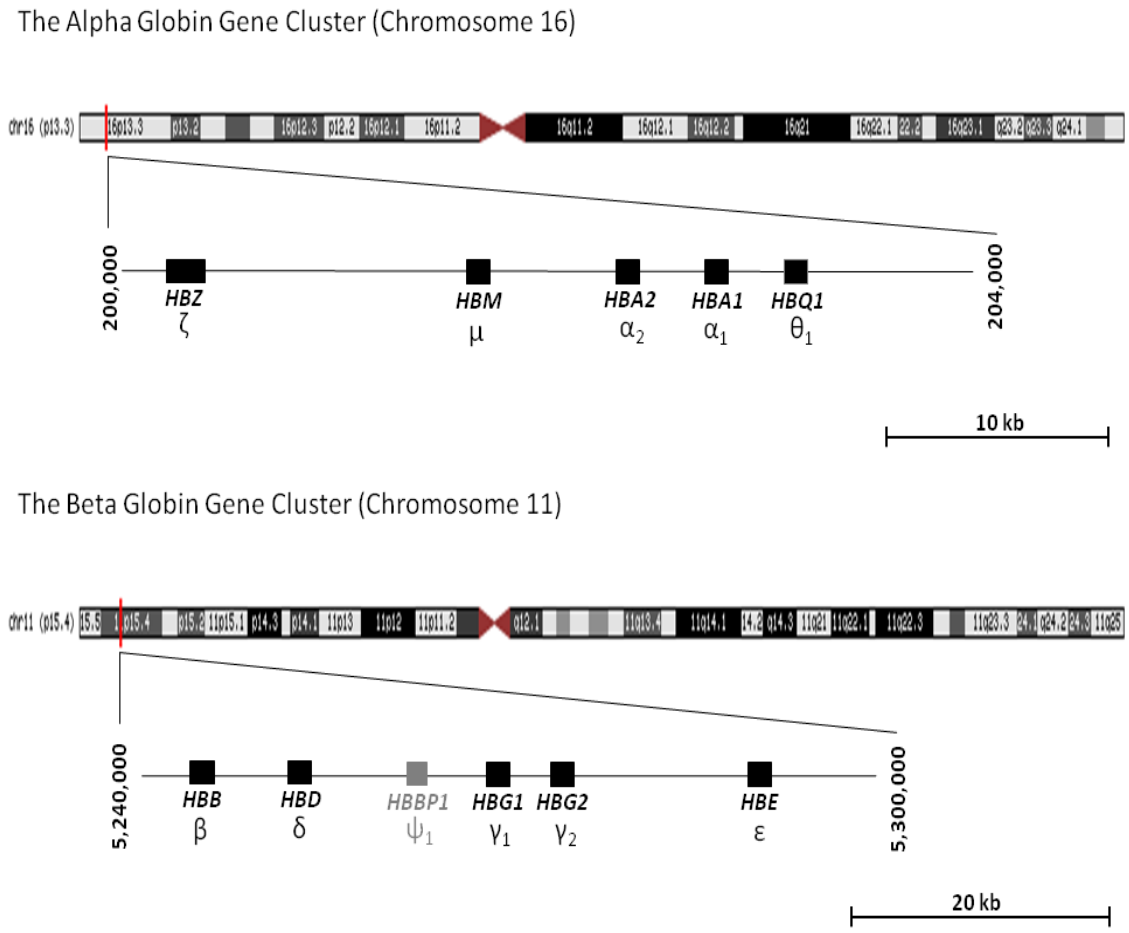


Figure 2: The Globin Gene Loci. Upper panel: The alpha globin gene cluster on chromosome 16, Lower panel: The beta globin gene cluster on chromosome 11. Chromosome schematic pictures obtained from the UCSC Genome Browser

The alpha globin gene cluster is situated on the short arm of chromosome 16 p13.1p (Waye and Chui 2001). The cluster consists of four alpha pseudogenes, the zeta globin gene (*HBZ*) the mu globin gene (*HBM*) and two alpha globin genes – alpha 1 (*HBA1*) and alpha 2 (*HBA2*) and theta globin (*HBQ1*) (Steinberg and Adams 1991). Forty kilobases upstream of the gene cluster is an erythroid-specific DNAse1 hypersensitive region which is essential for the expression of the genes (Sharpe, Summerhill et al. 1993). The *HBA1* and *HBA2* genes are identical except for their promoter regions. Both genes are used to synthesise alpha globin subunits throughout life, but the alpha globin 2 (HBA2) subunit is more highly expressed than alpha globin 1 (HBA1). This is believed to be due to the subtle differences between their promoter regions (Liebhaber 1986).

Although low levels of mRNA from the theta globin pseudogene have been detected *in vivo*, no transcribed product of the gene has been identified and mutations removing the gene do not appear to have any clinical consequences (Albitar, Peschle et al. 1989). The other pseudogenes are thought to have been inactivated over time by mutations and deletions in their promoter regions (Fischel-Ghodsian, Nicholls et al. 1987).

The beta globin gene cluster is located on chromosome 11 p15.5q. The cluster contains the beta and beta-like globin genes; beta globin (*HBB*), pseudobeta globin (*HBBP1*), delta globin (*HBD*), gamma globin 1 (*HBG1*), gamma globin 2 (*HBG2*) and epsilon globin (*HBE*) which bind to alpha or alpha-like subunits to form a haemoglobin molecule. The expression of these genes is controlled by the locus control region (LCR) centromeric of the globin gene cluster (Hardies, Edgell et al. 1984). *HBB* is the primary beta-like globin gene expressed during adulthood. Expression of the other globin genes change over the course of development. *HBD* is postulated to have arisen from gene conversion of *HBB*, given their extremely high homology. Despite this, delta haemoglobin protein (HBD) only makes up around 2-3% of total adult haemoglobin compared to the beta globin gene due to transcriptional deficiencies and low half-life mRNA. A change in the promoter region sequence is believed to confer the decreased transcriptional output (Humphries RK 1982, Hardies, Edgell et al. 1984, Kosche K 1984). Like *HBA2* and *HBA1*, the *HBG1* and *HBG2* genes are highly homologous. They differ by one amino acid residue, where codon 136 denotes glycine in *HBG2* and alanine in *HBG1*. Like the alpha globin genes, they are not produced at equal rates: gamma globin 2 (HBG2) is the dominant beta-like globin subunit produced up until approximately 6 weeks after birth.

Haemoglobin Switching

The reason that multiple globin genes exist is that their different chemical properties are necessary at different stages during development (see Figure 3) (Weatherall 2001). The alpha and beta globin gene loci are activated approximately three weeks after conception. In early embryos, four of the globin genes are active, producing four different forms of haemoglobin. Three of these are considered embryonic haemoglobins: Hb Portland (combining Zeta and Gamma globin chains), Hb Gower I (zeta and epsilon globin chains), and Hb Gower II (alpha 1 and epsilon globin chain). Foetal haemoglobin (HbF: alpha and gamma globin chains) is also produced early in pregnancy and becomes the most abundantly produced form of haemoglobin from approximately eight weeks post-conception until six weeks after birth. At this stage it is gradually replaced by adult haemoglobin (HbA: alpha and beta globin chains) (Figure 4). HbF continues to be produced at low levels (<5%) throughout adult life and a small amount of HbA_2 (HbA_2: alpha and delta globin) is also produced (Choi and Engel 1988, Weatherall 2001).

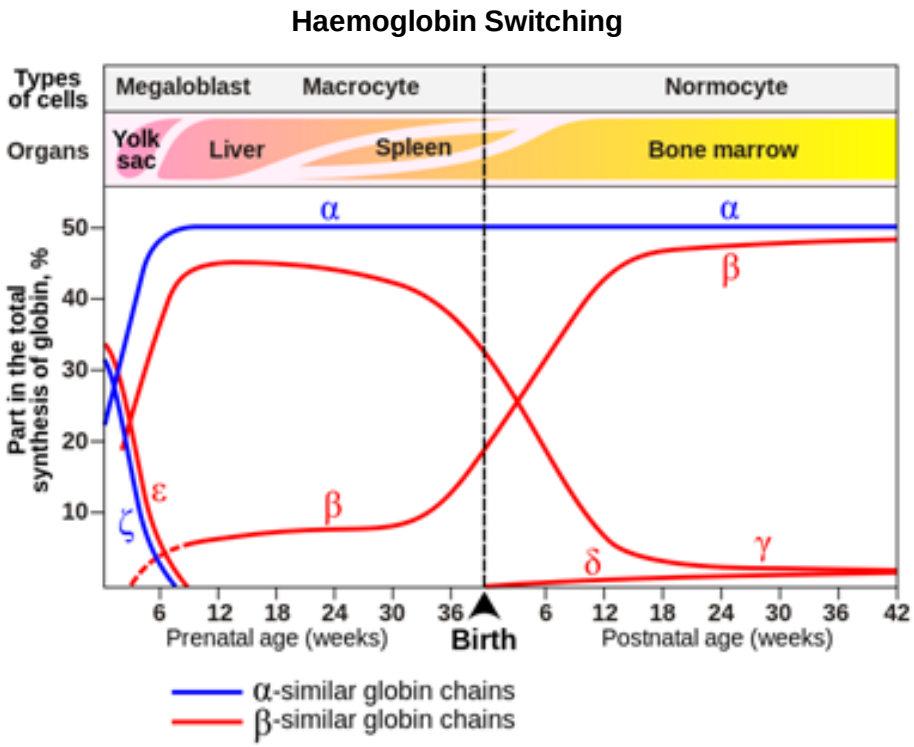


Figure 3:Haemoglobin Switching. Graph shows changes in globin gene expression during prenatal and post-natal development (Wetherall 2001)

Haemoglobin Proteins Produced By The Combination Of Different Globin Subunits

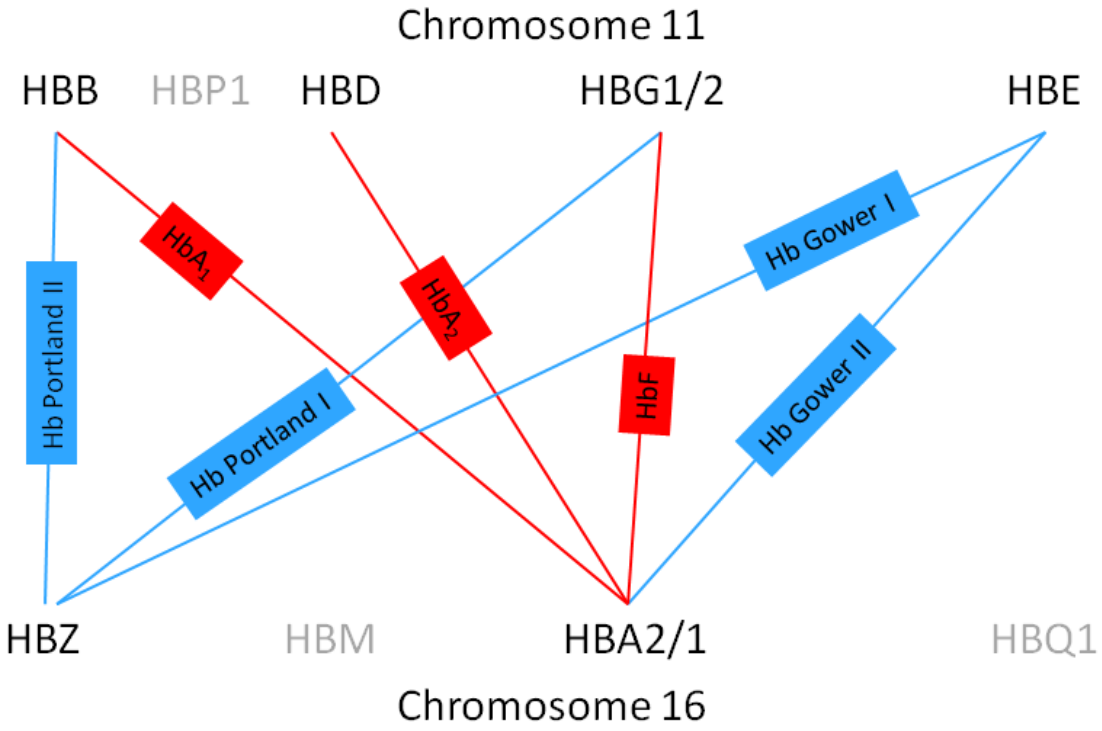


Figure 4: Haemoglobin Proteins Produced By The Combination Of Different Globin Subunits. Haemoglobin types expressed during adulthood are coloured in red.

Haemoglobinopathies

Haemoglobinopathies are a group of diseases which affect either the structure or production of one or more of the haemoglobin subunits. These diseases are caused by genetic mutations that affect the globin genes or the regions that control them (Trent 2006). Haemoglobinopathies are the most common group of genetic disorders worldwide, and affect approximately 7% of the global population (Weatherall 2008). They are more common in individuals of Middle Eastern, African, Mediterranean and South East Asian ancestry, and the incidence of different forms of haemoglobinopathy varies regionally (Huehns and Shooter 1965). Genetic variants that alter the amino acid sequence of the globin chains results in haemoglobin variants that can have altered properties and function. The most common and best studied is the sickle cell variant which causes sickle cell disease. Genetic variants that result in a reduced quantity of globin chain being produced give rise to the thalassaemia's and these are subdivided by the globin chain affected, the most common forms are alpha thalassaemia and beta thalassaemia (Weatherall and Clegg 2001).

Sickle Cell Disease

Sickle cell disease is the most common form of haemoglobinopathy. It is estimated 2.8% of the global population are carriers of the sickle cell variant, which has clinical

consequences in a homozygous or compound heterozygous state (Modell 2008). The causative mutation is a single base substitution (A>T) within the coding region of the beta globin gene. Clinical symptoms of the disease begin to appear around six weeks after birth when adult haemoglobin (HbA) production (incorporating the affected gene) overtakes foetal haemoglobin (HbF) production. The mutation (also designated rs334) changes the amino acid sequence of the gene, replacing glutamic acid with valine at codon 6, after the N terminal methionine is cleaved from the processed chain (Ingram 1957). This produces a variant form of the beta globin molecule which can combine with alpha globin to form sickle haemoglobin (HbS). The valine residue in HbS is hydrophobic, and reduces the solubility of the molecule in its deoxygenated state and allows it to bind to other HbS molecules. This results in the formation of irreversible intracellular polymers of deoxygenated haemoglobin molecules which alter the structure of the red blood cell and affect membrane function. The accumulation of polymers contorts the red blood cells, and they eventually become irreversibly sickled (Bunn 1997).

The structure of HbS is similar enough to HbA that it does not have significantly different oxygen affinity. The HbS globin subunit interacts slightly less efficiently with alpha globin subunits than the beta globin subunit. For this reason, most heterozygous carriers of the sickle cell variant will produce HbS globin at a rate of 40:60 HbA globin.

Normal haemoglobin is extremely hydrophilic but HbS is hydrophobic and under low oxygen conditions, the deoxygenated HbS molecules interact and polymerise. In this conformation the cells are ‘sticky’ and less flexible than normal red blood cells. Their inflexibility and their stickiness trigger vaso-occlusion of capillaries and small blood vessels, which reduces blood flow and extends the time spent in a hypoxic environment. If the vaso-occlusions are not cleared the number of red cells undergoing polymerisation increases and the blood flow is reduced further, creating a cycle which is difficult for the body to resolve. These acute episodes are thought to induce what used to be known as a sickle cell crisis and in some cases requires hospitalisation with opiate medication for pain relief. The acute attacks overlay a chronic condition of anaemia and a hypercoagulable state. The balance between acute and chronic is different between individuals and influenced by both genetic and environmental factors. Sickled cells are extremely fragile with a lifespan of 16-20 days (compared to a normal red cell span of 120 days), resulting in haemolytic anaemia.

The clinical consequences of sickle cell disease can be wide ranging: The most common symptoms arise from the anaemic element of the condition and include

fatigue, dizziness, headaches, pallor and jaundice. The onset of sickle cell crisis can be accompanied by episodes of serious pain. Downstream consequences of sickle cell crisis include hand and foot syndrome in children (pain and swelling in the hands and feet), splenic crisis (in which the spleen attempts to hold more red blood cells than it should, exacerbating anaemia), increased vulnerability to infection (as a consequence of spleen damage), acute chest syndrome (due to infected or sickled cells becoming trapped in the lungs), ulcers, pulmonary hypertension, stroke, eye problems, priapism and other issues linked to the occlusion of small blood vessels. In severe cases, if sickle cell crisis affects several organs simultaneously, multiple organ failure may develop (Stuart and Nagel).

For a disease that stems from one single, precise genetic change, sickle cell disease has an incredibly heterogeneous phenotype. Many other genetic or environmental factors can influence disease severity and progression. A major factor that influences the severity of sickle cell disease is the amount of HbF produced by the individual (Ali 1970). All people produce small amounts of HbF throughout adulthood, but in some individuals the level is slightly elevated. People who have this trait and also have sickle cell disease experience a less severe disease state. Knowledge of these modifying factors is important for accurate diagnosis, counselling and patient care.

Alpha Thalassaemia

Normal individuals have four copies of the alpha globin genes; two on each copy of chromosome16 (αα/αα). Alpha thalassaemia is caused by variants that inactivate these genes. Inactivation of a single alpha globin gene (αα/α-) has no clinical consequences. Inactivation or loss of two alpha globin genes (αα/-- or α-/α-) has no clinical consequence to the affected individual but they have a microcytic hypochromic red cell morphology. Inactivation or loss of three alpha globin genes (α-/--) causes HbH disease. This is often a mild condition and the reduced availability of alpha globin chains forces excess beta chains to combine to form β^4 tetramers, otherwise known as HbH. Individuals with a α-/-- genotype (three alpha globin genes deleted) have excess gamma chains early in gestation and this permits $_\gamma{}^4$ tetramers to form, known as Hb Bart's. Hb Bart's disappears 3 to 6 months after birth as gamma globin production is replaced by beta globin production (Harteveld and Higgs 2010). Both HbH and Hb Bart's are unstable globin variants, and have a higher affinity for oxygen than normal haemoglobin. This results in poor oxygen delivery to tissues and microcytic hypochromic anaemia (Clegg and Weatherall 1967). Inactivation or loss of all four alpha globin genes (--/--) means that no normal haemoglobin can be produced after the

first embryonic switch. This condition is called Bart's Hydrops Fetalis and is lethal, usually in utero, or occasionally shortly after birth. Death in utero often results in late-term miscarriage of the affected foetus with potentially serious physical and psychological consequences for the mother (Lehmann 1970). Hb Bart's Hydrops Fetalis is common where alpha thalassaemia deletions which remove both the alpha globin genes on the same chromosome are common; most notably Southeast Asia and the Mediterranean. In Africa, alpha plus thalassaemia is very common with the single alpha globin gene deletion –α being carried by 0.33 of the population. As the alpha (--/αα) deletion of both genes on the same chromosome is rare in Africa and the Middle East, HbH disease and severe alpha thalassaemia is not a clinical concern (Chui and Waye 1998).

Alpha globin gene loss is most commonly caused by deletions. Duplications of the alpha globin genes possible, with some individuals documented as having eight copies of the alpha gene, when a duplication is inherited on both chromosomes. These duplications have not been shown to have any thalassaemic consequences in isolation even though there is evidence that they are expressed (Harteveld, Refaldi et al. 2008). This may reflect that alpha globin chains are chaperoned by alpha globin chaperone protein prior to tetramer formation with a beta globin like chain (Kihm, Kong et al. 2002). Common deletions that cause alpha plus thalassaemia (single gene deletions) include the 3.7 Kb deletion and the 4.2 Kb deletion. Common deletions that cause alpha zero thalassaemia (two genes deleted from the same chromosome) are the SEA (Southeast Asian) deletion, MED (Mediterranean), alpha 20.5 deletion and FIL (Filipino) deletion. Single base changes causing alpha thalassaemia include Hb Constant Spring (Alpha 2 codon 142 TAA > CAA) which causes a premature stop codon and Hb Quong Sze (Alpha 2 codon 125 CTG > CCG) which creates an unstable alpha-like globin chain variant (Chui and Waye 1998). These non-deletional mutations can have more severe consequences compared to single gene deletions as cellular energy is spent creating a non-productive transcript, while in the presence of a deletion all transcriptional energy is focused on the remaining transcriptionally active genes (Chui, Fucharoen et al. 2003). In the UK, antenatal screening tries to identify instances where two alpha globin genes deleted from the same chromosome, as these mothers are at increased risk of having a child without any alpha globin genes resulting in Bart's Hydrops Fetalis. As alpha zero thalassaemia is rare in African populations but the -α/-α two gene deletion is relatively common, antenatal cases with African ethnicity are not considered at risk for alpha thalassaemia. Their greatest risk is HbH disease (-α/--) and this is usually a mild condition.

Large deletions removing upstream genes in addition to the alpha globin genes can cause a form of alpha thalassaemia that is associated with mental retardation. This is described as ATR syndrome (Wilkie, Zeitlin et al. 1990). A related condition with a distinct phenotype has also been reported, which is associated with a deletion on the X chromosome (ATR-X) posited to contain a trans-acting factor which effects the expression of the alpha globin genes (Gibbons, Suthers et al. 1992).

Beta Thalassaemia

Beta Thalassaemia is caused by mutations that remove or inactivate the beta globin gene, reducing availability of beta globin chains. The majority of variants causing beta thalassaemia are point mutations and different populations have different and overlapping subsets of pathogenic variants. Deletions are rare, except for the 619 bp deletion found within some Indian ethnic groups and the Middle East. Heterozygous carriers of beta thalassaemia are normally symptomless with microcytic hypochromic indices and a raised HbA_2 and follow a recessive mode of inheritance (Cao A 2000).

The disease state occurs when a person is homozygous for beta thalassaemia or compound heterozygous for beta thalassaemia and another variant. As there is such a wide number of different variants, with varying clinical effects, the severity of the condition is a spectrum. At the genetic variant level beta thalassaemia causing variants are classified as either β^0 or β^+ which reduce expression from the allele entirely or partially, respectively. Individuals that inherit two β^0 variants would be expected to have a severe beta thalassaemia condition and all other combinations would be less severe. The clinical diagnosis uses the term thalassaemia major or transfusion dependent beta thalassaemia (TDBT) and requires a complete clinical picture before classification. It is possible for an individual to be homozygous for two β^0 variants and not present with beta thalassaemia major as the lack of beta globin is substituted by HbF, leaving them clinically well. Milder forms of beta thalassaemia – termed thalassaemia intermedia or non-transfusion dependent thalassaemia (NTDT) – arise when there is some beta globin gene expression. Currently, blood transfusion is used to manage beta thalassaemia patients and the additional red cells contain iron which causes iron overload. Therefore these individuals also require iron chelation therapy to manage their excess iron, which if left unchecked can cause organ damage. In rare cases, beta thalassaemia can be an autosomal dominant condition, where individuals with only one affected allele also experience a disease state. Hyper unstable haemoglobin variants causing autosomal dominant thalassaemia are limited to the third exon of the beta globin gene as these escape clearance by nonsense mediated decay (Thein 1992).

The HbVar database currently lists 279 reported variants causing beta (or variants on, e.g. delta-beta) thalassaemia (Patrinos, Giardine et al. 2004).

The symptoms of beta thalassaemia begin to appear around six weeks after birth where adult haemoglobin production surpasses foetal haemoglobin production. The primary clinical feature of beta thalassaemia is chronic hypochromic microcytic anaemia due to insufficient globin chain synthesis and ineffective erythropoesis. In addition to this, aggregates of unused alpha globin subunits collect inside red blood cells. The presence of these aggregates (termed 'inclusion bodies') reduces the lifespan of mature red blood cells, worsening the person's anaemia (Shinar E 1990). Individuals who are homozygous for beta thalassaemia produce little or no functional adult haemoglobin and their survival is usually dependent on receiving frequent blood transfusions (Weatherall 2001). Beta thalassaemia intermedia has lower clinical severity than beta thalassaemia major. This group produce a greater amount of functional haemoglobin than those with thalassaemia major, which may be due to the inheritance of ameliorating functions or less severe thalassaemia variants (Thein 1993).

The chronic anaemia beta thalassaemia causes can have many downstream clinical consequences. In most countries thalassaemia major is treated with frequent blood transfusions, which results in a different disease natural history. The clinical features of thalassaemia in countries providing transfusions are as follows: Common clinical symptoms of beta thalassaemia are pallor, fatigue, irritability, feeding problems, diarrhoea and splenomegaly. Consequences of iron overload resulting from repeated treatment by blood transfusion in early life include growth retardation; failure of sexual maturation ; issues with the heart, liver (such as hepatitis, fibrosis cirrhosis), and endocrine glands (resulting in the abnormal hormone production which may create various other knock-on effects); hypersplenism (and subsequently reduced immune response); venous thrombosis; osteoporosis and hypogonadism (Cao and Galanello 2010). In countries where frequent blood transfusions are not available, thalassaemia major is characterised by growth retardation, pallor, jaundice, poor musculature, hepatosplenomegaly, leg ulcers and skeletal changes (Galanello and Origa 2010). Without regular transfusions the life expectancy for individuals with beta thalassaemia major is less than a year. With regular transfusions and iron chelation therapy, life expectancy is normal. In 71% of patients who receive regular blood transfusions the eventual cause of death is cardiac complication as a result of iron overload (Borgna-Pignatti, Cappellini et al. 2005). Bone marrow transplant can cure beta thalassaemia but is currently associated with an unacceptable risk of mortality for the

majority of sufferers (Trent 2006). Current research into a cure for thalassaemia focuses on gene therapy approaches (namely in boosting foetal haemoglobin production or somehow reversing the switch from foetal to adult haemoglobin production) (Donnall Thomas, Sanders et al. 1982).

Beta thalassaemia carriers are symptom free, but the presence of the variant can still be detected from microcytic, hypochromic red cell indices accompanied by an elevated HbA_2 level (>3.4%) . Both sickle cell disease and beta thalassaemia show a great amount of diversity in the clinical severity of the disease experienced by patients. A major factor in this is the inheritance of modifying genetic factors. These include concurrent inheritance of other variants affecting the same globin gene loci and variants affecting the globin subunit's counterpart (concurrent alpha and beta thalassaemia result in a milder form of both diseases because the imbalance between the production of the two subunits is lessened) and many other genetic factors, yet to be identified, plus environmental factors. The most important modifier of disease severity identified to date is the level of HbF (foetal haemoglobin) produced in adulthood. In most adults, trace amounts of HbF persist in the peripheral blood, <1% of total haemoglobin. A minority, however, show elevated levels of HbF with the genetic cause either linked to the beta globin loci or due to variants in trans acting factors. Although a complete understanding of the molecular mechanism of HbF regulation is not yet known the complex trait is strongly genetic and is termed 'hereditary persistence of foetal haemoglobin' (HPFH) (Thein 2009). Genomic regions, or quantitative trait loci (QTL) associated with HPFH have been identified on chromosomes 11p and 6q, but their precise mechanism has yet to be determined (Craig 1996) (Menzel, Garner et al. 2007). When associated with haemoglobinopathy, HPFH has been shown to have a positive impact on disease outcome. For this reason, understanding what causes HPFH is of extreme interest to understanding haemoglobinopathies and improving patient care.

Other Thalassaemias

Alpha and beta are the most commonly seen thalassaemias, but mutations can affect any or multiple globin genes producing conditions with varying clinical severity. Thalassaemias affecting only the delta globin subunit go unreported as they have no clinical impact although they can mask a raised HbA_2 level (which is the differentiator between alpha and beta thalassaemia) and can cause diagnostic delays as a result (Bouva, Harteveld et al. 2006). Thalassaemias affecting epsilon globin or gamma

globin have the greatest effect early in development and may either go unnoticed or require intra-uterine transfusion, but be tolerable postnatally.

Deletions in the beta globin gene loci are rare. Deletions have various clinical effects depending on the breakpoints and are therefore important to identify and record accurately. Deletions may remove all or part of the beta globin gene and are classified as beta thalassaemia deletions. Deletions that remove the delta and beta globin genes are termed δβ thalassaemia deletions and are often associated with moderate increases in HbF (7-15%). HPFH deletions are larger and remove one gamma globin gene plus the delta and beta globin genes. These deletions result in elevated levels of HbF (15-30%), which is significantly higher than that in beta or δβ thalassaemia. Whole beta globin loci deletions have also been documented, associated with a mild increase in HbF production to (2-5%).Individuals present with a haematological picture similar to alpha thalassaemia; microcytic hypochromic with a normal to low level of HbA_2. These $\varepsilon_\gamma\delta\beta$ thalassaemia deletions are sub classified into two groups. Group I are deletions that remove the globin genes and group II remove the locus control region which silences the globin genes that are left intact. Duplications at this locus had not been documented prior to this study

The range of variants that can cause haemoglobinopathies is unusually high; many diseases with known genetic origins are associated with specific variants. Identifying the underlying variants in haemoglobinopathies can be clinically useful, but also demanding given their variety. How do mutations that cause haemoglobinopathies come about, what causes such a range of variants at these loci, and why are they so prevalent in African, South East Asian and Mediterranean populations?

Mutation

Mutations are changes to the genetic code that can occur anywhere in the genome and have an associated connotation that they will have a negative impact on the host. Mutation is a process that mostly occurs during meiosis, where mistakes can be made during DNA replication. Many mutations result in variants that are considered to be ‘silent’ where they fall outside gene coding regions, or occur within the coding region but do not cause any changes to the gene’s function or product. Since the advent of next generation sequencing the amount of genetic variation under investigation has increased and the use of the term mutation has changed, with the journal *Human Mutation* tightly limiting its use to explain a process. All genetic differences resulting from mutation should be described as genetic variation such as sequence variants,

pathogenic variants or structural variants. NB: some of the analysis tools used in this study have yet to adopt this change, and still refer to variants as “mutations”.

Single Nucleotide Changes

Single nucleotide changes (often called ‘single nucleotide polymorphisms’ or ‘SNPs’) can be as phenotypically severe as a large deletion that removes a megabase of sequence. When nucleotide changes occur in the coding region of a gene and change its amino acid sequence they are described as non-synonymous gene variants. These can have detrimental consequences for the structure and function of the protein the gene produces. Single base changes in the beta globin gene are responsible for structural haemoglobinopathies including sickle cell disease (Steinberg M.H. 2001), HbC disease and HbE disease. Single base changes can also reduce or prevent the production of any functional protein by the affected gene if they are within splice sites or promoters. Changes such as these that affect the dosage of globin genes cause thalassaemia. Single nucleotide changes causing beta thalassaemia such as the IVS-II-1 (G>A) variant and Sardinian codon 39 nonsense variant are common (Najmabadi, Karimi-Nejad et al. 2001). Single nucleotide changes causing alpha thalassaemia also occur but are rarer. Small genetic changes such as single base substitutions and small deletions lend themselves well to the HGVS naming nomenclature used to document thalassaemia-causing variants. In this system all genetic variants are called against a standard transcript or a genomic reference sequence. This standardises how the majority of genetic variation described and allows instant understanding of the position of the variant within the gene. For example, a mutation described as “HBB:c.315+1G>A”, affects the gene HBB, with the +1 indicating the variant is 1bp within the intron (after codon 315) changing a base from guanine to adenine. From this information, the variant can be surmised to affect mRNA splicing.

Structural Rearrangements

Recombination is an exchange of genetic material between two chromosomal regions, which is used to repair double stranded breaks (DSBs) in DNA which occur during cell division. Homologous recombination (HR) is an exchange of similar or identical nucleotides which generates sequence variation in gametes and is an important mechanism for evolution (Schleif 1993). In non-allelic homologous recombination (NAHR), a double stranded break is repaired via an unequal exchange of genetic material which can result in a structural rearrangement (Stankiewicz and Lupski 2002). Structural rearrangements include duplications, deletions, translocations and inversions of DNA sequence. These are currently defined as “*sequence variants of at least 50bp*

in size" (Mills, Walter et al. 2011). Smaller sequence variants are known as 'SNPs' (single base changes) or 'indels' (insertions, inversions or deletions of up to 50bp of sequence). Structural rearrangements can affect multiple megabases of DNA. Deletion variants are a common cause of alpha thalassaemia (Trent, 2006) and occasional cause of beta thalassaemia. While structural rearrangements can happen anywhere in the genome, they are reported more frequently in some regions than others. The number of rearrangements associated with alpha thalassaemia is higher than many other genetic diseases. It has been suggested that the large number of globin genes we possess may be due to historic duplication and gene conversion events, mediated by elements at this locus that encourage recombination events (Michelson and Orkin 1983, Steinberg and Adams 1991). Elements within the DNA sequence that caused these events could be responsible for the structural rearrangements we see in thalassaemia.

Certain features of DNA sequence can affect its chance of being involved in a recombination event by influencing the 3D structure of the chromosome on which it is situated. Features such as mirror repeats, palindromic sequences and z-DNA motifs can result in the formation of Non-B (or non-canonical) DNA structures which are associated with non-homologous recombination (Bacolla 2009). As non-homologous recombination is often associated with detrimental phenotypic consequences, evidence of regions with a high rate of non-homologous recombination may not be reflected in the general 'recombination rate' as calculated based on the inheritance of broadly functional chromosomes in the normal population. In fact, only 60% of identified recombination events are linked to known recombination hotspots, as our understanding of where and how DNA recombines is not yet complete (Mills 2012). Non-homologous recombination is probably responsible for most alpha zero deletion events – the break-points for these tend to involve small partially homologous sequences which are known to be important for some mechanisms of non-homologous repair of double stranded DNA breaks which occur during meiosis (Nicholls, Fischel-Ghodsian et al. 1987). Alu repeats are associated with deletion breakpoints and are frequent around the alpha globin gene cluster. They may have structural significance or just provide the regions of partial homology. Where these occur around the beta globin gene loci, their association with structural rearrangements is substantially lower than the alpha globin gene locus (Henthorn, Smithies et al. 1990).

The alpha globin gene cluster may be situated in an area particularly prone to non-homologous recombination (alternatively, the location may just be particularly tolerant of rearrangements, as many resulting phenotypes such as duplications have no clinical

affect). The cluster is located near the telomere on the short arm of chromosome 16. In males, chromatin forms tighter loops towards the telomeres. These smaller loops are associated with an increased rate of recombination, where DNA from different loops which are brought into close physical proximity to one another can recombine (Paigen, Szatkiewicz et al. 2008) (Paigen and Petkov 2010). Non-homologous recombination events in these regions could create gene deletions. The entire alpha globin gene cluster (both the genes and surrounding DNA) also has a high (60%) GC content (Fischel-Ghodsian 1987) which could influence the recombination rate of the region.

Recombination events can be difficult to characterise: their breakpoints are often situated in repetitive regions that are hard to sequence, they may affect very large regions of DNA, and in the case of balanced rearrangements (inversions, insertions and translocations), the standard techniques used in molecular diagnostic laboratories may not detect them at all. For this reason, it can be difficult for diagnostic laboratories to identify characterise novel rearrangements causing thalassaemia or develop tests to detect further instances of the variant in the future. In addition, diagnostic labs may only need to detect the presence of a variant and do not need to completely characterise the it for patient management. Where these rearrangements are phenotypically silent (as duplications and inversions often are) they may go undetected in the normal population, so estimating the rate at which these events occur can be challenging (Nobrega, Zhu et al. 2004) (Weischenfeldt, Symmons et al. 2013) (Firth, Richards et al. 2009). Rearrangements recorded on the alpha and beta globin gene loci to date are illustrated in Figure 5 and Figure 6.

Variants in HbVar Affecting the Beta Globin Gene Cluster

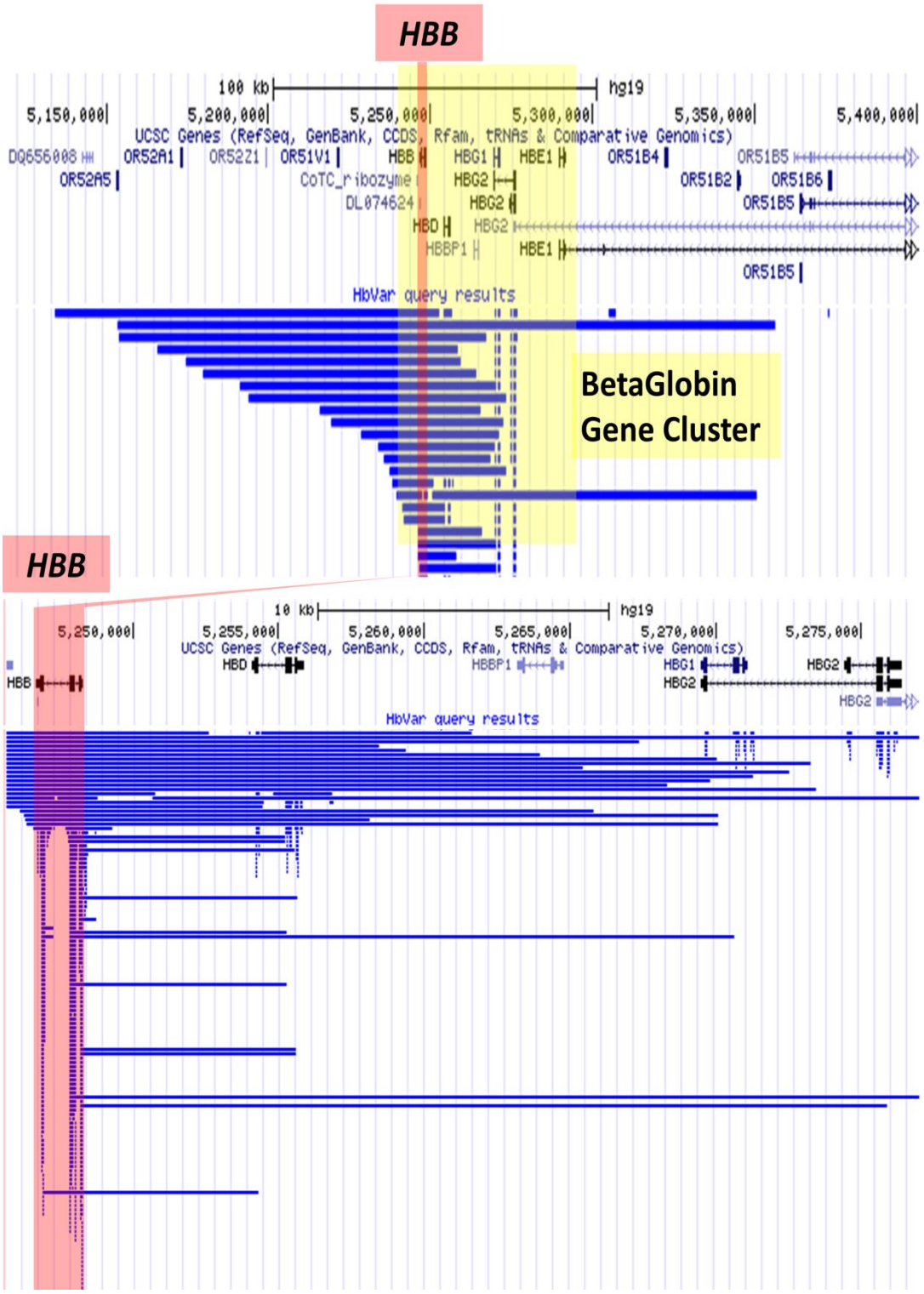


Figure 5: Variants in HbVar Affecting the Beta Globin Gene Cluster. Variants (blue lines) are displayed in the UCSC genome browser, with gene positions shown at top of image. The Globin gene cluster is highlighted in yellow and position of the beta globin gene is also highlighted in red. Two images on different scales are provided to illustrate the range of variants.

Variants in HbVar Affecting the Alpha Globin Gene Cluster

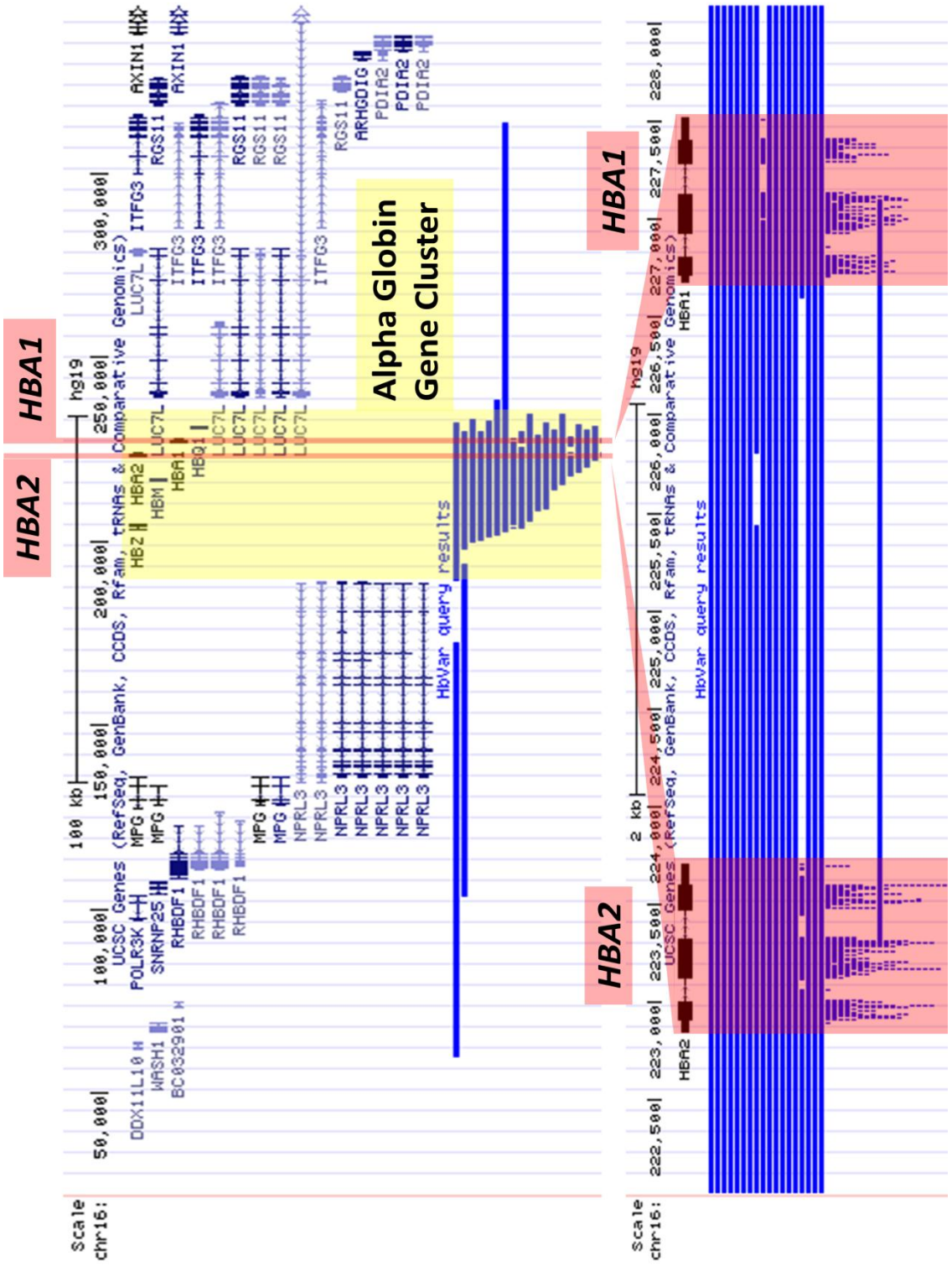


Figure 6 Variants in HbVar Affecting the Alpha Globin Gene Cluster. Variants (blue lines) are displayed in the UCSC genome browser, with gene positions shown at top of image. The Globin gene cluster is highlighted in yellow and position of the alpha globin genes are also highlighted in red. Two images on different scales are provided to illustrate the range of variants.

Recombination Hotspots

Some parts of the genome are subject to a higher rate of recombination than others. These regions are called recombination hotspots. The rate of recombination can be calculated by studying the haplotypes that exist in different regions. A haplotype is the precise sequence of nucleotides on a single chromosome. A chromosome can have many different possible haplotypes even in healthy individuals, as new variants with no functional consequences accumulate over generations. As some regions of the genome have a high rate of recombination, chromosomal haplotypes tend to split into smaller 'blocks' (Paigen and Petkov 2010). The pattern of SNPs within a block is highly conserved and they are almost always inherited together. The pattern of blocks making up a chromosome, however, can change significantly between populations. By comparing the haplotypes of different chromosomes across a region of interest, the rate of recombination can be calculated based on the number of points at which different haplotypes with a common ancestor diverge from one another (Lupski 2004). From this, clear 'hotspots' where recombination has occurred frequently in the past can be seen. These regions are typically surrounded by large areas of 'linkage disequilibrium' where the recombination rate is significantly lower than average. A recombination hotspot has been previously reported in the beta globin gene cluster, between the HBB and HBD genes. The alpha globin gene cluster on the other hand, is >30Kb from the closest calculated recombination hotspot (Chakravarti, Buetow et al. 1984, Ardlie, Kruglyak et al. 2002, Gabriel, Schaffner et al. 2002).

Spread and Distribution of Haemoglobinopathies

The areas where haemoglobinopathies are most common are all regions where the virulent malaria strain *p.falciparum* was once endemic (Weatherall 2008). Evidence suggests that carriers of haemoglobinopathies have partial resistance to malaria infection, which is why the variants became common in these areas as they conferred a selective advantage (Taylor, Parobek et al. 2012). The mechanism by which haemoglobinopathies might provide malaria resistance is not fully understood. Evidence suggests that because haemoglobinopathies reduce the lifespan of red cells, the malaria parasite is unable to complete the stage of its life cycle that occurs within red blood cells, which prevents systemic infection of the individual. Haemoglobinopathy carriers may also experience a partial benefit from this (Ayi, Turrini et al. 2004). Sickle Cell disease and HbC disease are also hypothesised to protect against malaria through their disruption to intracellular trafficking networks that the parasite needs to modify to propagate (Cyrklaff, Sanchez et al. 2011).

Migration of people between these areas was once rare. Most populations which have high rates of haemoglobinopathies due to a selective pressure have a distinct set of variants that arose, and then remained within that area. The population of Sardinia is a good example, where 95.7% of persons who carry a beta thalassaemia variant have the same codon 39 nonsense variant (Rosatelli, Dozy et al. 1992). Migration between all countries has increased enormously over the last century. London has a particularly diverse population in which 50 nationalities are represented by at least 10,000 individual residents, and over 301 languages were reported to be spoken in the 2005 census (Statistics 2011). This diversity means that hospitals must be prepared to test for pathogenic variants that would not be common in the native population. In addition, ethnicity is difficult to collect and in a metropolitan city it is an unreliable guide to the variants that may be detected (Clark and Thein 2004). Therefore tests must detect all types of thalassaemia causing variants which cannot be predicted by ethnicity.

Routine Diagnosis of Haemoglobinopathies in the UK

Diagnosis of haemoglobinopathies in the UK can occur at any time in an individual's life depending on how and when they present with symptoms. Most diagnoses are made as a new-born or antenatally due to two major screening programmes. The new-born screening programme is universal and aims to detect neonates affected by sickle cell disease. As a consequence of the methods used this screening program can also detect the absence of HbA and therefore severe forms of beta thalassaemia. The first line screening method is either High performance liquid chromatography (HPLC) or mass spectrometry, both of which abnormal protein chains. Antenatal screening aims to detect mothers affected by a haemoglobinopathy and the first line screening methods are HPLC and a full blood count (FBC). Screening is divided into high and low prevalence regions depending on the affected birth rate. Mothers that are detected as carriers or affected by a haemoglobinopathy have their partners called for testing and if they are positive, the couple have counselling so they are aware of the implications and their options. The majority of individuals tested are unaffected and are reported as normal. Genetic testing is only required when the screening results from the analysis are not clear, variants need to be identified to allow informed counselling, and in the assessment of alpha thalassaemia. Genetic testing is also required for prenatal diagnosis which is offered on the NHS, to couples at risk of an affected child. It is best practise to identify the pathogenic variants in the parents prior to foetal sampling, however this does not happen for cases of sickle cell disease as the penetrance of the variant is close to 100%, and the protein screening test represents the genotype.

Genetic investigations follow a diagnostic algorithm looking for common variants first and then moving on to rarer causes if the first test is negative. The investigation is guided by the phenotype or the primary screening results using it as a guide to look for alpha or beta thalassaemia or an alpha or beta globin gene variant. The diagnosis of thalassaemia can be time consuming, particularly if the variant is rare. For example, if an individual has $\varepsilon_{\gamma}\delta\beta$ thalassaemia the screening results will appear like alpha thalassaemia. A screen for common alpha deletions is started and then the *HBA1* and *HbA2* genes are sequenced. Upon review an MLPA assay will be carried out to detect rare deletions and only once this is negative will an MLPA assay be carried out on the beta globin loci to detect the causative structural variant. This diagnosis could take up to one month to complete at a time when mothers may be anxious about their pregnancy and haemoglobinopathy risk.

A description of these routine diagnostic techniques is provided in Table 1 These techniques still have limitations leaving a very small proportion of cases undiagnosed. These include balanced translocations and small point mutations in the LCR that may affect globin gene transcription. The main drawback is the approach to testing which favours individuals with common variants and penalises those with rare variants. In addition, some individuals will have multiple haemoglobinopathy variants requiring multiple tests. Managing this workload is complex and requires dedicated staff and processes within a molecular diagnostic facility aimed at providing services for a repertoire of conditions. Having a single technique that detects all the genetic variation in one assay would standardise the workflow, provide an equitable service to all and detect all genetic variation allowing it to be related to the referring phenotype, thereby giving a better diagnosis (Clark 2004).

Newborn Screening

The new-born screening program aims to identify babies born with sickle cell disease. In 94% of cases, children born with sickle cell disease will develop functional asplenia within 5 years of birth due to vaso-occlusion of small vessels within the spleen by sickled red blood cells (Milne 1991, Pai and Nahata 2000). Functional asplenia is a form of immunodeficiency that is associated with increased risk of infection, and increased severity of infections. Initiating penicillin prophylaxis soon after birth has been found to significantly reduce mortality associated with a number of infections in patients with sickle cell disease (Gaston, Verter et al. 1989). Fast and accurate new-born screening for sickle cell disease can therefore have a significant impact on patient health. Early diagnosis is also beneficial to children with beta thalassaemia major,

where regular blood transfusions maintaining a haemoglobin concentration of >95-105 g/L can prevent splenomegaly (Cao and Galanello 2010).

Antenatal Screening and Prenatal Diagnosis of Haemoglobinopathies

The NHS Sickle Cell and Thalassaemia Screening Programme aims to allow people to make informed choices before conception and during pregnancy.

Women are screened from 8 weeks of gestation and if they are affected or carrying a haemoglobinopathy, the partner is also screened. If they are both carrying a haemoglobinopathy, the couple are counselled and offered prenatal diagnosis (PND). The antenatal screening plan aims to offer diagnosis to all women in the UK to allow informed decision making, and to perform at least 50% of diagnoses before 12 weeks and 6 days of pregnancy.

PND requires foetal sampling, either by chorionic villus sampling (CVS) in early pregnancy (9.5-12.5 weeks of gestation) or by amniocentesis (14-20 weeks of gestation). Accurate testing is crucial at this stage, as the results may determine whether the mother decides to continue with the pregnancy or not.

The diagnostic workflow depends on the availability of the father for testing. In PND of sickle cell disease, if the father is available two tests are performed: A TaqMan real time PCR assay and a restriction fragment length polymorphism (RFLP) assay are used to directly detect the presence or absence of the sickle cell mutation. Microsatellites are analysed as a third assay to detect maternal cell contamination in the gDNA prep. If the father is unavailable for testing, beta globin gene sequencing is also performed to detect the presence of compound heterozygous states that could result in a sickle cell disease genotype. In prenatal diagnosis of thalassaemia, either alpha genotyping or beta sequencing are performed, depending on the variant identified in the parents. MLPA is also performed to support the diagnosis. This workflow is most accurate when both parental samples are available but in the absence of the paternal sample the accuracy will be less as the pathogenic variant is not known.

Table 1 Techniques used in the diagnosis of Haemoglobinopathies

Blood/Molecular/Protein laboratory techniques used in the screening and diagnosis of haemoglobinopathies	
Test	**Details**
FBC -MCV -MCH	A full blood count counts the number of red blood cells, white blood cells and platelets in a blood sample and also calculates red cell indices: mean corpuscular volume and mean corpuscular haemoglobin. An automated blood count analyses the contents of blood by the number, size and light absorption of the cells in a standard volume of blood. Low MCV and MCH are indicative of haemoglobinopathy.
HPLC	High performance liquid chromatography is used to separate out and identify the different forms of haemoglobin present in a blood sample. Red blood cells are lysed to release their haemoglobin, which is then electrophoresed through a gel. Differences in mass and charge alter the rate at which different molecules cross the gel.
Sickle solubility test	A blood sample is treated with a reducing agent such as sodium dithionite which reduces the available oxygen in the blood. Lysis is induced in the red blood cells, releasing their haemoglobin. HbS will form liquid crystals when released, giving the mixture a cloudy appearance. In the absence of HbS the mixture will remain clear. The test detects the presence of HbS but not whether the subject is a carrier, compound heterozygous or homozygous for the condition.
CAE /IEF gel test	A Cellulose acetate electrophoresis (CAE) test identifies HbS in a blood sample by its migration across a cellulose acetate strip when exposed to an alkaline pH. HbS migrates noticeably more slowly across the gel than other haemoglobins, due to an extra positive charge. Isoelectric focussing (IEF) separates out the different haemoglobin variants present in a sample by their migration across a gel with a pH gradient to their isoelectric point, which is determined by their molecular charge. This can detect differences of 0.02pH units between different haemoglobin variants. This is more costly than a CAE test, but allows differentiation of the isoelectrically similar Hb

	D-Punjab from HbS (Wajcman 2011).
Restriction length polymorphism (RFLP)	When a genetic variant causes a sequence change that affects a restriction enzyme recognition site it may be detected using a RFLP assay. The region is amplified by PCR and then digested with the restriction assay. The digested fragments are then resolved on an agarose gel and the presence of specific size fragments will indicate if the variant is present and its zygosity.
CGH Array	Comparative Genomic Hybridization is used to detect dosage changing variants. DNA from a test sample and a reference sample are labelled with different fluorophores. the comparative levels of hybridization (which releases the fluorescent tag) of DNA from these samples to DNA probes corresponding to a region of interest is measured. Regions were a deletion or duplication has occurred are identified by their decreased or increased, respectively, binding to probes in the affected region. (*NB: CGH array is performed in some laboratories, but not at the Dept. Molecular Pathology at KCL, where Next Generation Sequencing strategies are being perused instead*)
MLPA	Multiplex ligation dependent probe amplification uses a single primer set to amplify multiple oligonucleotide pairs from across a genomic region of interest. Each separate probe group creates a different sized product. The amplified products can be separated out by capillary electrophoresis. Any products which are absent due to deletions affecting the target sequence are quickly identified, and the approximate size of the deletion can be estimated by the number of probe groups affected.
Gap-PCR	A Gap-PCR uses primers to amplify the break-point product of a known deletion via PCR. Primer sets for multiple variants can be pooled together in a multiplex reaction, providing they amplify under the same thermal cycling conditions. Multiplex Gap-PCR is used to quickly test for multiple common deletions in alpha thalassaemia. It is also useful to confirm newly identified breakpoints.
Dye-Terminator Sequencing	Dye-terminator sequencing uses the sanger-sequencing method to establish the precise sequence of PCR products amplified from patient DNA. Fluorescent dyes are used to label free nucleotides

	which fluoresce upon incorporation into DNA fragment, revealing its sequence. This can identify point mutations and small indels which are commonly associated with beta thalassaemia and haemoglobin variants.
Real time PCR/ Taq Man assay	TaqMan qPCR is used to detect known genetic variants in a DNA sample, and to determine whether it is present in a homozygous or heterozygous form. Fluorescent probes bind specifically to the target alleles and are detected when the extended primer releases the fluorophore during amplification.
STR analysis for Maternal contamination	Short Tandem Repeats (STRs) are highly variable sequences with multiple alleles in the population, so it is likely the mother and father will have different alleles. Amplifying the repeats with a fluorescent primer and sizing the PCR product allows detection of the maternal and foetal genotypes. Maternal contamination of the foetal sample can be detected when the mother heterozygous and the allele not inherited by the foetus appears in the foetal sample.

New techniques for DNA sequencing and analysis are constantly emerging, and have become increasingly sophisticated over the last century. Techniques used today can produce a wealth of information, making interpretation and storage of this data challenging. Wet laboratory DNA analysis techniques are becoming increasingly supported by computational tools for data handling, and as laboratory techniques have evolved, so have the fields of computational biology and bioinformatics.

The Technological Evolution of DNA Sequencing and Analysis

Bioinformatics

The classical definition of bioinformatics is "*The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information*" - attributed to Fredj Tekaia - Institut Pasteur (Fenstermacher 2005)

Bioinformatics differs from the equally important field of computation biology, which is defined as "*The development and application of data-analytical and theoretical methods, mathematical modelling and computational simulation techniques to the*

study of biological, behavioural, and social systems" - NIH Biomedical Information Science and Technology Initiative Consortium.

The boundaries between these two disciplines are blurred and elements of both are usually used together in biological research.

The discipline of bioinformatics grew out of the increasing importance of protein modelling, and also the need to handle large amounts of data produced by sequencing technology. Bioinformatics governs how large datasets are acquired, stored, evaluated and analysed. A crucial element of the field is successfully combining the findings of multiple fields together – often taken from huge databases – to discover new relationships between one finding and another. These studies were quick to produce data: such as the relevance of the 3D structure of mRNA on its natural selection (Konings, Hogeweg et al. 1987).

DNA Sequencing Technology

The first genome to be completely sequenced was that of the ΦX174 bacteriophage, stored in notebooks and then hand-typed for publication. Using data in this format was unwieldy, and the sequence was transferred to a punch card computer by scientists at Cambridge. Scientist Rodger Staden then created an interactive program (The Staden Package) which allowed biologists with little knowledge of computer science to access the data. This is arguably the first use of bioinformatics in the field of genomics (Krawetz and Womble 2003) (Fenstermacher 2005).

The ΦX174 bacteriophage genome was sequenced using the 'plus and minus' method developed by Frederik Sanger (Sanger, Coulson et al. 1982). In this method, two polymerase chain reactions were performed: the first amplified fragments into random lengths, and then terminated synthesis with the incorporation of a ^{32}P labelled nucleotide. The product was then divided into aliquots, to which a single type of nucleotide base was added for a second reaction. The products were all electrophoresed on an acrylamide gel. If the sequence was a single base longer in one of the reactions, this was concluded to be the nucleotide present at that position in the fragment. Problems arose where homopolymers occurred and the number of bases incorporated had to be estimated based on the position of the band on the gel. This slow process was eventually sufficient to sequence the entire bacterial genome. Two years later, Sanger published an improved, faster sequencing method which has the same underlying chemistry as methods used today (Hutchison 2007).

DNA sequencing became more rapid after the development of the dideoxy sequencing method by Frederik Sanger in 1977. Dideoxy or ‘chain termination method’ sequencing initially sequenced only single stranded DNA cut into small fragments (~100bp) by a restriction enzyme. The fragments were mixed with a DNA polymerase enzyme, a short specific primer and free nucleic acids. This mixture was then divided into four separate reactions. To each of these, a different radioactively labelled chain-terminating nucleotide was added. At some point, one of the fragments in the reaction tube would incorporate one of these bases, rather than an unlabelled base, into the DNA fragment. The lack of a 3’ group attached to this base prevented further replication of the fragment. The products of each of the four reactions were added to separate lanes on an acrylamide gel. Small fragments (which had been incorporated a terminating base early in synthesis) travelled further down the gel than slower moving large fragments. Comparing fragment sizes in different lanes showed the relative positions of different bases. From this, the entire sequence of the fragment could be deduced (shown in Figure 7). This method originally required single stranded DNA, but was later modified to work for double stranded molecules. Shortly after the publication of the Sanger method of sequencing, the Maxam-Gilbert Method was also released. This method used similar mechanism but different underlying chemistry, with the advantage that a cloning step was not required. The method became highly popular for this reason, but fell out of favour due to improvements to the sanger-sequencing method, coupled with its complexity and use of hazardous chemicals (França, Carrilho et al. 2002).

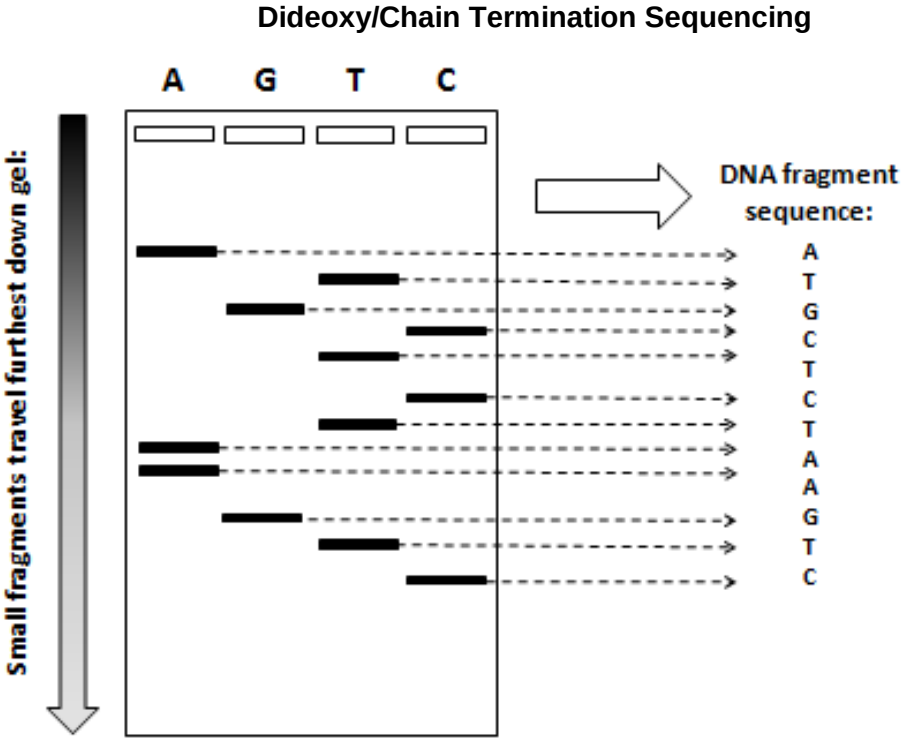


Figure 7: Dideoxy or 'chain termination' sequencing, developed by Fredrik Sanger. The different positions at which chain termination occur in the different reactions show positions where that base appears in the sequence. For a 10bp sequence for example, if the reaction containing 'A' chain terminating nucleotides created products that were 3, 4 and 5bp; the 'G' reaction created products that were 1, 2 and 6 bp; the 'C' reaction an 8bp product and the 'T' reaction 7, 9 and 10bp products, the sequence of the original fragment would be determined to be 'GGAAAGTCTT'

The completion and digital recording of the first bacteriophage genome was quickly followed by the completion of other small genomes. This eventually led to the creation of an international digital database to store the information, called GenBank in 1982. The availability of this data to other scientists led to the rapid development of tools to analyse and manipulate the data, such as the FASTA format for storing and searching for sequence information. Improved computing technology also allowed increased automation of the Sanger method, resulting in platforms such as the Life Technologies Genome Analyser, which is capable of sequencing hundreds of base pairs of multiple samples in a few hours.

Milestones in Genome Sequencing

As the number of small genomes that had been sequenced successfully grew, so did the attraction of the challenge of sequencing the human genome. Plans to sequence the human genome were officially made in 1985, with the project starting in 1990. The project was completed in 2003, covering 99% of the genome, 92% of that with an accuracy of 99.99% (Schmutz, Wheeler et al. 2004). The project started out using

sanger-sequencing methods, but adopted more sophisticated approaches as new technology became available. A significant portion of the project was carried out via shotgun-sequencing, developed in principle in 1977.

Shotgun sequencing combined sanger-sequencing with clonal amplification and also introduced paired-end sequencing. The technique was cumbersome and computationally more taxing than its predecessors, but invaluable for novel genome assembly. DNA is sheared randomly into fragments, normally >2 and <150Kb in length. A particular size is amplified clonally in a vector (such as bacterial artificial chromosomes (BAC)) and then each fragment is sequenced by chain-termination from both ends. The sequencing of both ends of each fragment is a crucial aspect of this technique: it allows two distant sequences to be linked, as they are known to be (i) orientated in opposite directions from one another and (ii) a distance from one another that is approximately equivalent to the fragment length. Multiple clonal fragments from circular vectors can be cut into linear molecules with different start and end positions. By sequencing both ends of each of these fragments (sanger-sequencing can now sequence up to 1000bp DNA), the entire sequence of the fragment can eventually be resolved with greater speed and accuracy than by single read sequencing.

Shotgun Sequencing Methodology

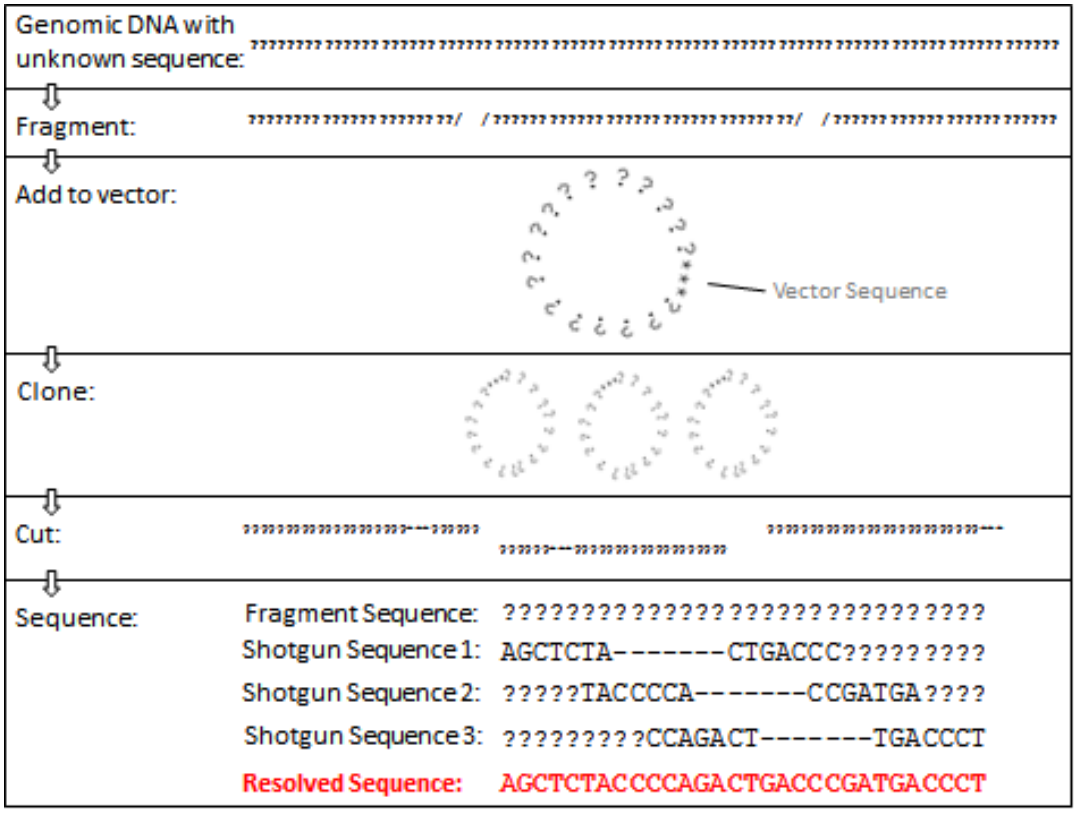


Figure 8: Shotgun Sequencing Methodology. Genomic DNA is sheared at random into small fragments (2-150Kb). Specific size selected fragments are added into a vector such as BACs and clonally amplified. The clonal copies are cut at random positions to create linear DNA strands. The fragments are sequenced from both ends. The original sequence can be resolved by compiling overlapping sequences into 'contigs'.

The human genome project was completed ahead of schedule and under budget, thanks to the boom in sequencing technology and bioinformatics that it created. Data from the project is freely available online or to download and several newer builds improving on the initial assembly have been released. The data, hosted by the UCSC Genome Bioinformatics Site and Ensembl Genome site, have since been cited by innumerable research papers. The completion of this project inspired even more ambitious sequencing challenges, many of which have already been completed: the 1000 genomes project; the compilation of the Neanderthal genome; the $10,000 genome project. The $1000 human genome challenge was to sequence a human genome for less than $1000 dollars and was completed by Illumina when they released their X10 sequencing instrument (Sheridan 2014). The 100,000 genomes project (sequencing 100,000 genomes from people with rare disease or cancer) is still underway. These sequencing milestones have been accompanied by the development of many different bioinformatics tools to analyse the data (Baird, Driscoll et al. 1981, Conesa, Götz et al. 2005, Li, Yu et al. 2009, McKenna, Hanna et al. 2010). The

projects themselves have been made possible by the development of increasingly sophisticated sequencing technologies: namely the development of Next Generation Sequencing.

Timeline of Significant Advances in Genome Sequencing

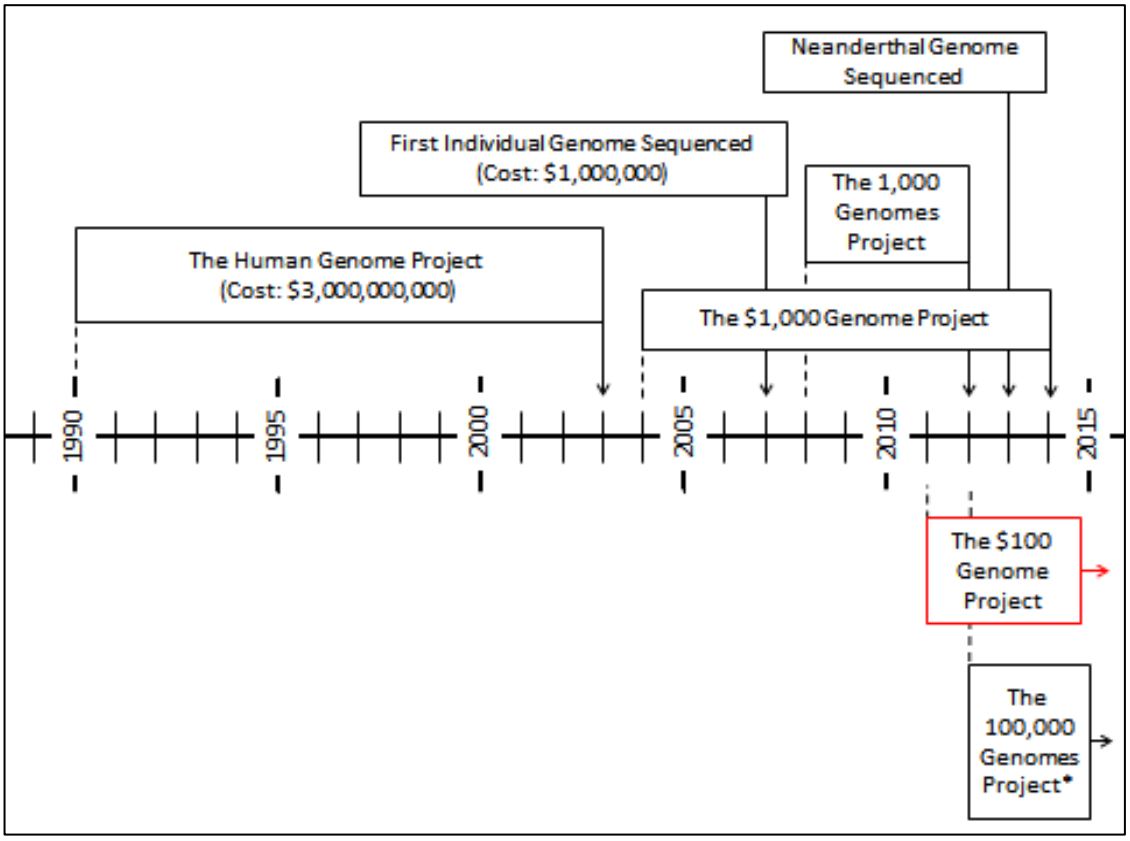


Figure 9 Timeline of Significant Advances in Genome Sequencing. The Human Genome Project. 1990-2003; The $1,000 Genome Project, 2004-2014; The first individual genome sequenced, 2007; The 1,000 Genomes Project, 2008-2012 (Consortium 2012); The completion of the Neanderthal genome in 2013 (Prufer, Racimo et al. 2014); The $100 Genome Project, started by Genia in 2011 and ongoing; The 100,000 Genomes Project, started in 2012 by Genomics England and ongoing

NGS sequencing

Next Generation sequencing platforms are united by their ability to sequence many thousands of DNA fragments, from multiple samples, simultaneously. Many different platforms have been developed, with different advantages and disadvantages for various sequencing applications. Three companies produce the majority of NGS platforms currently used:

Illumina

Illumina platforms operate through ‘sequencing by synthesis’. DNA is prepared for sequencing in advance by shearing it into small fragments, blunt ending and adenylating the fragments and then ligating adaptors and identifying index tags to the

ends of the fragments. The DNA fragments are denatured to single stranded and ligated to a glass slide called a ‘flow cell’. The flow cell has eight deep channels to maximise its surface area. It is coated with adaptors complimentary to those of the DNA fragments, which are able to ligate to them. The DNA fragments are then ‘bridge amplified’: they bend over forming a bridge shape between two of the adaptor molecules on the flow cell. The fragment is then amplified, beginning from the end of the ligated adaptor to which the DNA molecule has attached and extending across the bridge. Both molecules break away from one another and straighten out through another denaturing step. The process is repeated, and both clonal fragments undergo new bridge amplifications. Each amplification cycle doubles the number of clones produced. The result is many clusters of identical single stranded DNA fragments in very small, localized area. This allows detection of the florescent signal generated from the clone.

The actual sequencing of the fragments employs a ‘sequencing by synthesis’ technique. Fluorescently labelled nucleotides are washed over the flow cell. Different fluorescent tags are attached to the four different nucleotides. The nucleotides polymerase against the single stranded DNA fragments ligated to the flow cell to create double stranded molecules. Each time a nucleotide is incorporated into a new molecule, it release its fluorescent label. A high definition camera within the instrument records the combined signal emitted by each clonal cluster of DNA fragments, and interprets this as the fragment sequence. It also assigns a quality score to each nucleotide it calls.

Thermo Fisher Scientific (formerly Life Technologies)

Ion torrent platforms (Thermo Fisher Scientific) employ semiconductor sequencing and emulsion PCR. DNA is sheared into small single stranded fragments. Each fragment is ligated to a magnetic bead. An emulsion PCR is performed that creates an isolated PCR reaction for each bead. The PCR reaction continues until the bead is completely covered with clonal copies of the original DNA fragment. The beads are then added to an ion-chip covered with wells large enough for only a single bead. Different DNA bases are washed sequentially over the chip. If that base is incorporated into the DNA fragments on the bead, they release a hydrogen ion. The change in pH this causes in the solution surrounding the chip is detected by an ion sensitive layer at the base of the well. A different nucleotide is then washed over the chip and the process is repeated. Homopolymers cause release of a larger number of ions, and can be identified the pH change they create.

Oxford Nanopore

Single molecule sequencing is an emerging technology, with the first platform operating using this chemistry – the MinION (Oxford Nanopore, UK) released this year. Sequencing produces '2D' reads of DNA molecules: one read of the forward strand followed by a read of the reverse strand connected to it by a hairpin structure, to produce a consensus for each base position. The platform is extremely small and low cost; library construction is simple and doesn't involve amplification steps (Cherf, Lieberman et al. 2012). Sequencing single molecules is a potentially enormous advance for sequencing technology which could allow accurate sequencing of repetitive regions and immediate resolution of novel variants. Recent studies examining the potential of the MinION platform to sequence the phage lambda genome reported achieving an average read length of 5Kb, but with a high base call error rate (Quick, Quinlan et al. 2014). Work is currently underway to improve the accuracy of MinION sequencing, but it is not yet at a stage where it can be used as a reliable diagnostic tool (Madoui, Engelen et al. 2015).

Next generation sequencing has many advantages compared to previous sequencing technologies: its high sequencing capacity can be used to assemble an entire genome, to sequence a targeted region of interest from multiple samples, or to sequence target regions in individuals at a very high read depth. High read depth can allow detection of non-germline, minor allele and low copy sequences normally masked by the genomic background. NGS also has the advantage that while regions of interest can be targeted specifically, this can be done with a high tolerance for novel sequences that might exist within those areas. This is something that cannot be done in PCR-based technologies, where prior knowledge of the target sequences is necessary for primer design. The technology has huge diagnostic potential, and in the context of haemoglobinopathy diagnosis could potentially replace the battery of tests currently employed. In addition to matching the accuracy of the current diagnostic standard, NGS has the scope to improve upon it by allowing other clinically important variants outside the globin genes to be included in diagnosis, such as those associated with risk of preeclampsia and modifiers of haemoglobinopathy disease state, as well as sex determination.

Although NGS platforms represent a great advance in sequencing technology they still have some shortcomings: platforms are costly; PCR contaminants and excessive PCR duplicates can hinder results; semiconductor and pyrosequencing platforms have issues with accurate sequencing of homopolymer repeats; base calling accuracy tends to diminish with read length; false positive and negative rates can be high if data is not

treated with appropriate stringency; long repetitive regions cannot be targeted accurately. The sequencing capacity of the machine also means a trade-off is required between number of samples sequenced, size of region sequenced, and read depth produced. Many of these problems may cease to be with the development and refinement of new 'third generation' sequencing technologies such as nanopore sequencing.

Data analysis itself is also an issue: alignment programs struggle to correctly align reads situated in highly homologous sequences such as repeats or very similar genes, such as *HBG1* and *HBG2*. Analysis can also be hindered by the bulk of sequence that these platforms produce, which many non-specialised centres many not have the computing capacity to store or manipulate.

Summary, AIM

At the outset of this project NGS sequencers were just becoming widespread in diagnostic labs. Their application to haemoglobinopathies was questionable as not all globin genes need to be sequenced diagnostically and the clinically relevant genes and regions are small, lending themselves well to Sanger Sequencing. CGH array had been investigated by the diagnostic group at Kings College Hospital and others (Phylipsen, Chaibunruang et al. 2012) looking at mapping and characterising large structural haemoglobinopathy variants. CGH array technology was effective at identifying deletions and duplications but breakpoint mapping was not possible in every case and balanced translocations could not be detected. CGH array was applied to haemoglobinopathies as another methodology in the battery of techniques used for diagnosis, whereas NGS had potential to replace all tests, with a single methodology, which would give a definitive diagnosis in all cases. The benefits of streamlining the diagnosis could standardise the diagnostic approach and align with other genetic tests carried out in the lab for other diseases. This could therefore make the lab more efficient and ensure all variants are identified within the 10 working day turnaround time for antenatal assessment, irrespective of how common they are.

Before starting the project it was assumed that small sequence variants would be detected and that the large structural variants would be a challenge, although the lab had no experience of NGS data generation or analysis. Once this assessment of NGS is complete, the aim will be to implement it into routine diagnosis. This thesis describes the technological assessment of NGS as applied to haemoglobinopathies, a disease group that covers a complete range of mutation types and would therefore have relevance to other conditions.

Methods

Sample Collection

Ethics approval: Use of Residual Clinical Samples

Extracted DNA, residual from clinical testing, does not contain cells and therefore does not come under the Human Tissue Act (2004) but falls under common law and professional guidance, this is provided by the Royal College of Pathologists.

In accordance with the Royal College of Pathologists guidance:

- The retention and storage of pathological records and specimens (5th edition, April 2015)
- Guidance on the use of clinical samples for a range of purposes that are not within the remit of Research Ethics Committees (RECs) (3rd edition)
- Consent and confidentiality in clinical genetic practice: Guidance on genetic testing and sharing genetic information

Prior consent was obtained for genetic testing, long term storage of DNA samples and use of residual DNA samples for controls or test development.

The Sample Cohort

Residual clinical samples came from patients who had undergone genetic testing for haemoglobinopathies at King's College Hospital. These samples came from individuals of both genders, all ages and a variety of ethnic backgrounds. In most cases a deletion or duplication had previously been identified by MLPA or Sothern blotting. Therefore the genetic cause for the haemoglobinopathy had been identified but the structural variant had not been characterised.

Red Cell Indices

For some samples, haematological data was available which could be used to predict whether the variants in the sample affected the alpha or beta globin gene loci. This data and its normal interpretation are summarised in Table 2. Globin percentage data is obtained through an HPLC assay, and the other indices through a full blood count (FBC). For details of these, refer to Table 1.The 'normal' reference ranges vary between different regions. Values in the table reflect those used at KCH.

Table 2 Red cell indices

Metric	Meaning	Diagnostic use
HbA_2 %	Percentage of HbA_2 in blood (composed of alpha and delta globin)	Raised in beta thalassaemia trait (>3.4%) reference range is generally set at 2-3.4%
HbF %	Percentage of foetal haemoglobin in blood (composed of alpha and gamma globin)	May indicate diserythropoesis. Raised in some conditions, including some forms of beta thalassaemia Noted when above 5%.
RBCx1012/L	Red Blood Cell Count	Raised in thalassaemia (varies with age, gender and other factors). The normal ranges used for men and women are: 4.5-6.5 males 3.8-5.8 women
Hb	Concentration of haemoglobin in the blood	Haemoglobin less than 100 g/L indicates anaemia, which can be caused by haemoglobinopathies or iron deficiency
MCV	Mean corpuscular volume (volume of red blood cells)	Normal range is 78-100 fL Less than this indicative of thalassaemia
MCH	Mean corpuscular haemoglobin (amount of haemoglobin within red blood cells)	Normal range is 27-32 pg Less than this is indicative of thalassaemia

Automated DNA extraction

DNA processed by the Molecular Pathology laboratory at King's College Hospital was extracted using the QIAsymphony (Qiagen, Germany). Extraction was performed in accordance with the manufacturer's protocol, as part of standard laboratory practice. The primary blood collection tube was loaded onto the instrument with a pre-labelled 2ml DNA collection tube in a corresponding rack. 1ml of EDTA blood was extracted,

using the Qiagen Midi kit eluted into 200 µL of buffer yielding 30-150 ng/µL gDNA. If more DNA was required then the programme was run multiple times eluting into the same 2ml screwcap collection tube.

Phenol-Chloroform DNA Extraction (manual)

Solutions referred to throughout 'Methods' are listed in Appendix 2: List of Solutions

Phenol-Chloroform DNA extraction was used to purify DNA from blood samples collected from outside the routine diagnostic route at King's College Hospital. This procedure was performed by Mrs Helen Rooks, Department of Molecular Haematology, King's College London. The details of the procedure are given in Appendix 4.

DNA Quantitation

Before the DNA was used in any downstream reactions a number of quality metrics were assessed including concentration and purity. This checks that the extraction process has been successful and ensures the quality of the extracted DNA is of a high standard and reduces the downstream failure rate. Assessing the size range and distribution of DNA fragments in the sample was also important during NGS library preparation and was measured in a few ways depending on the availability of apparatus.

Nanodrop

The Nanodrop Spectrophotometer ND-8000 (Thermo Scientific, USA) was used to quantify DNA extracted from blood as part of routine laboratory practise. The concentration of a dsDNA sample (in ng/µL) is made based on its absorbance of light in a 260 nm wavelength over 1-2 ms with the assumption that a DNA concentration of 50 µg/ml absorbs 1 unit. The Nanodrop calculates the purity of a DNA sample based on the ratio of the sample's absorbance of 260 nm and 280 nm light wavelengths. A ratio of >1.8 was considered acceptable with minimal contamination.

The Nanodrop was initialized upon power-up by adding 2 µL water to each pedestal and selecting 'Initialize' in the software. The pedestal was wiped clean with a lens cleaning tissue and a 3 µL blank (molecular grade water or AE buffer for Qiagen Symphony extracted DNA) was added to the pedestals. After a blank measurement was taken the pedestal was wiped clean again. 3 µL of DNA sample was placed on the pedestal and its concentration was measured. Results were displayed in a table and stored electronically and automatically. The pedestal was cleaned between different samples with a tissue and distilled water.

QuBit

The QuBit 2.0 Fluorometer (Life Technologies, USA) Broad Range and High Sensitivity kits were used according to the manufacturer's protocol. The QuBit measures the DNA concentration of a sample by measuring the fluorescent signal emitted by the QuBit reagent, which fluoresces only when bound to DNA. The High Sensitivity kit was used when sample concentration was expected to be <5 ng/µL. For other applications the Broad Range kit was used. For each use of the instrument, a fresh QuBit Assay Mix of 1:100 QuBit Reagent : QuBit Sample Buffer was made up. ~200 µL per sample is required, plus 380 µL to make up the two standards necessary for quantitation. 5 µL of dsDNA was added to 0.5 ml QuBit assay tubes. 195-198 µL QuBit Assay Mix was added to make a final volume of 200 µL. The DNA and Assay Mix were mixed by pipetting and vortexing briefly. Two standards (Standard #1 and Standard #2, provided with the kit) were made up by mixing 10 µL of each standard with 190 µL of the Assay Mix in a QuBit Assay Tube. The QuBit was turned on and the desired assay function was selected. Both standards were placed in the QuBit and measured for calibration. These two standards are used to draw a standard curve from which unknown samples were calculated. Assay tubes containing sample DNA were then measured. The dilution factor was programmed into the machine in order to calculate the concentration of the original sample by reading from the standard curve. This technique is more accurate than UV absorbance measurement by the Nanodrop, as there is no interference from proteins and other contaminants and does not measure single base nucleotides. Qubit concentrations are particularly useful at low DNA concentrations where other techniques are less accurate.

Agarose gel electrophoresis

For a 1.5% gel, 1.5 g agarose powder (BioGene, UK) was added to 100 ml 1x TBE Buffer and heated in a microwave on full power until the agarose had dissolved and the solution was clear (approx. two minutes). The dissolved agarose was cooled by holding the flask under a running cold tap until it was comfortable to hold by hand. 5 µL of 10 mg/ml ethidium bromide was added to the agarose solution (final concentration 0.5 ng/ml). The solution was poured into a gel tray and left to set with a comb with appropriate number of slots for the number and volume of the samples. DNA samples (2-5 µL) were made up to 10 µL with nuclease free water and 1.5 µL 6x bromophenol blue gel loading dye (Sigma, USA). Once the gel was set it was submerged in a gel tank containing 1x TBE Buffer. The DNA samples were loaded into the slots, along with an appropriate sized DNA ladder, often 50 ng of 1 Kb DNA ladder (Invitrogen). The gel

was run at 100 V for 1-2 hours until the individual fragments had resolved, separation was viewed using a gel documentation system (Syngene, UK?). DNA fragment size was estimated by comparison to the DNA ladder electrophoresed in parallel with the samples. Gel images were saved as (.SGD and .JPEG file types) using GeneSnap software (Syngene).

Bioanalyser

The 2100 Bioanalyser (Agilent, USA) measures DNA fragment size, distribution and concentration by digital electrophoresis. In principle, the DNA is labelled and electrophoresed through a gel-dye matrix to separate the different sized fragments and the electrophoresis is visualised in real time. This method is very accurate at visualising small fragments of similar size <500 bp which can be difficult to resolve by agarose gel electrophoresis. The Bioanalyser also has the added advantage that it can measure the concentration of the various size fractions by comparing to known standards run in parallel. All reagents were equilibrated to room temperature for 30 minutes before use. The reagents are light sensitive and were kept out of bright light in a box for this period. If necessary, a new aliquot of gel-dye mix was prepared: The DNA Dye Concentrate was vortexed and 25 µL was added to a DNA Gel Matrix vial. The solution was vortexed and centrifuged at 2240 x g briefly in a micro centrifuge. The mixture was transferred to a supplied spin filter tube. The spin filter tube was centrifuged at 2240 x g for 15 minutes. The gel-dye mix was stored in aliquots for use for up to 2 weeks.

Setting up the Bioanalyser Chip Priming Station

Figure 10: Setting up the Chip Priming Station. Figure modified from images in the Agilent DNA 1000 Kit Quickstart Guide G2938-90015 (for URL, see appendix 2)

A new DNA chip was removed from its packaging and the glass was inspected for any cracks. The chip was placed in the chip priming station. The base of the chip priming station was set to position 'C' and the syringe holder was set to the lowest position (See Figure 10). Nine µL of the Gel-Dye Mix was added to the well, marked with a circled 'G' (See Figure 11).

A Bioanalyser Chip

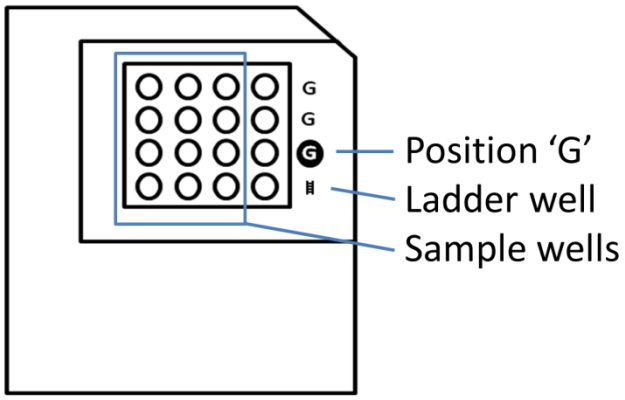


Figure 11: A Bioanalyser Chip. Positions of sample wells, ladder well and well 'G' labelled.

Any bubbles in the gel were removed by pipetting. Ensuring the syringe plunger was fully retracted, the gel priming station was locked into place on top of the chip. The plunger was slowly depressed until it could be secured under the catch. After exactly

one minute, the catch was released. After 5 seconds, the plunger was pulled back to its starting position and the chip was removed from the chip priming station. A further 9 µL of Gel-Dye Mix was added to the two additional wells marked with 'G'. Five µL of Marker was added to all the 13 wells of the chip (12 sample wells and ladder well) that were not marked with 'G'. One µL of DNA ladder was added to the well with the 'ladder' symbol. One µL of sample was added to each of the sample wells, or if less than 12 samples were being run, 1 µL of deionised water was added to any unused wells. The chip was placed horizontally in the provided chip vortex adapter and vortexed at 2400 x g for exactly one minute. The chip was placed in the Bioanalyser and 'start run' was selected. It was necessary to start the run within five minutes of removing the chip from the vortex for best results.

Closing the lid of the Bioanalyser lowers an electrode into each well. The electrodes are activated one well at a time. The electrodes produce a charge which causes the polar DNA molecules to migrate through the gel. A laser shone on the well causes the dye in the gel to fluoresce if it is bound to a DNA molecule. The size of the fragments in the well is measured by the speed at which they migrate through the gel, and their concentration is measured by the strength of fluorescence the sample emits. The Bioanalyser 2100 Expert Software produces an electropherogram of the sample. The size and concentration of the DNA in the sample can be determined by comparison to a DNA ladder run in a separate well and to lower and upper size markers included in the sample buffer which is added to each well. The results were exported as .xad run file and a .pdf report file.

Tapestation

DNA fragment length was calculated using the 2200 Tapestation (Agilent, USA). The D1000, D1000 High Sensitivity and Genomic DNA kits were used according to the manufacturer's protocol. The D1000 High Sensitivity kit was used when sample concentration was expected to be <5 ng/µL. The Genomic DNA kit was used when the mean DNA fragment size was expected to exceed 500 bp. For all other assays the DNA 1000 kit was used.

The reagents and screen tape were equilibrated to room temperature for 30 minutes before use. As the reagents are light sensitive, they were kept in a box away from bright light during this time. For the D1000 and gDNA assays, 3 µL of Sample Loading Buffer was added to an 8 well PCR strip tube (Starlab). One µL of D1000 ladder was added to the first tube. One µL of DNA sample was added to each other tube. For the High Sensitivity kit, 2 µL Sample Loading Buffer was added to the 2 µL of Ladder and 2

µL DNA samples. The tubes were centrifuged briefly in a minifuge to bring all liquid to the bottom, then placed in a 96 well plate adapter on a vortex mixer and vortexed at 2400 x g for one minute. The tubes were centrifuged again to return all liquid to the bottoms of the wells. The tubes were placed in the Tapestation in the strip tube adapter, with the ladder in the top right position. The tip rack was filled with Agilent Tapestation tips, and the used tip tray was emptied prior to the run. The tape was placed in the tape reader with the barcode at the front, on the left hand side where it could be scanned by the instrument. Lids were removed from the strip tubes. The instrument was closed and the run was initiated. The Tapestation measures the size and concentration of DNA fragments in the sample using a similar methodology to the Bioanalyser, but each sample migrates through a lane in the screentape, rather than across a well. The Tapestation produces an image of the electrophoresis through the screentape that is converted into an electropherogram trace whereas the Bioanalyser produces an electrophoresis trace as its primary output. As a result the Bioanalyser is more accurate at sizing fragments but is slower and prone to more issues. The results from the Tapestation were saved as instrument/tape specific file types (.D1K, .D1KHS, etc.) and exported as a .doc report file.

Sample Fragmentation

Covaris

The Covaris E220 (Covaris, USA) was used to shear genomic DNA into fragments of various size ranges by sonication. Three to five µg of high molecular weight genomic DNA for each sample was made up to a total volume of 130 µL in nuclease free water and then placed in AFA fibre crimp cap 6x16 mm microtubes (Covaris). The Covaris E220 tank was filled with Millipore water up to level 6, so that tubes would be immersed in the water during sonication. The chiller and pump were switched on and left for 45 minutes to degas and cool the water to 4° C prior to use. Samples were placed in a 96 position rack and sheared to the desired size according to the parameters provided by the manufacturer listed in Table 3.

Table 3 Shearing parameters for Covaris E220

Desired fragment size:		200 bp	400 bp	500 bp
Parameter	**Description**			
PiP	The instantaneous power in Watts transmitted to the acoustic transducer during the “on” time of each acoustic burst. *	175	140	105
DF	Percentage of time that the transducer is “on” and creating acoustic waves*	10%	5%	5%
Cycles Per Burst	The number of waves generated by the transducer in a burst.*	200	200	200
Treatment time	Time in seconds of treatment	180	55	80
Temperature	Temperature of waterbath in Celsius	7°C	7°C	7°C
Water bath level	Level to which waterbath is filled	6	6	6

*Parameters for shearing as defined by the Covaris User Guide. These parameters were used for creating 200 bp fragments.

Post shearing a small aliquot of samples was run on the Bioanalyser or Tapestation to check to size distribution of the fragments. It was important to remember the size limitation of the technique as some larger fragments may not be visible when run with specific consumables. Comparing the total estimate of DNA from the Tapestation to the Qubit was useful.

Bioruptor

This instrument is similar to the Covaris in that it shears DNA to a desired size range using sonication. This instrument provides less control over the shearing parameters and yields fragments in a broader range around the desired fragment size.

To shear DNA into ~500 bp fragments 3.5 µg DNA was made up to 100 µL with molecular grade water. The DNA sample was placed in a 1.5 ml microcentrifuge tube in a tube holder and inserted into the bioruptor’s chilled water bath. The program used to shear the DNA was 15 cycles of 30 seconds ‘ON’, 30 seconds ‘OFF’ high intensity sonication. No other parameters were available to modify the distribution of fragments and this was a limitation of the apparatus.

Size Selection with SPRI-Select Beads (Beckman Coulter, USA)

SPRI beads are 1.5 nm in diameter. They comprise of a polystyrene core coated with a layer of magnetite and an outer layer of carboxyl molecules. DNA is a polar molecule that can bind reversibly to the carboxyl molecules on the bead's surface in a manner dependent on the concentration of polyethylene glycol (PEG) and salt to which it is exposed. DNA-bound SPRI beads can then be extracted from a solution via magnetic separation. This process is used to purify DNA from solutions that may contain PCR inhibitors or other impurities. The amount of DNA that can bind to a SPRI bead is dependent on the concentration of PEG buffer. With a concentrated buffer solution small fragments of DNA (≥100 bp) can be bound and as the buffer is diluted the beads are only able to bind larger fragments, therefore fragments <100 bp are never captured (See Figure 12).

SPRI beads can therefore be used to purify DNA, concentrate it, or, by changing the buffer conditions, extract DNA fragments of a specific size from a wider population. Beckman Coulter produce a range of beads for different DNA purification applications: AMPure XP beads are a generic bead used to remove contaminants from a DNA sample and to concentrate DNA fragments or PCR products; SPRIselect beads are similar to AMPure beads except the buffer and bead concentrations are more accurately prepared, ensuring the highest consistency in size selection and purification. SPRI-select and pre-PCR AMPure bead steps were performed on different sides of the laboratory ensuring PCR product purification (using AMPure beads) never contaminated the NGS workflow (SPRI-Select beads). CleanSeq beads are specifically designed to purify small DNA fragments for Sanger sequencing, removing residual unincorporated labelled nucleotides. Using these beads to purify sequencing templates allowed the process to be easily automated on a BioMek NX (Beckman Coulter, USA).

'SPRIselect' beads were used to select a specific size range of DNA fragments after sonication. Figure 12 shows a titration of different SPRIselect concentrations and its effect on the ability to recover a DNA ladder. This was used as a guide for our DNA selection experiments. Using two independent bead captures at different concentrations permitted capture of a specific size range of DNA from the broader range produced by Covaris/Bioruptor sonication..

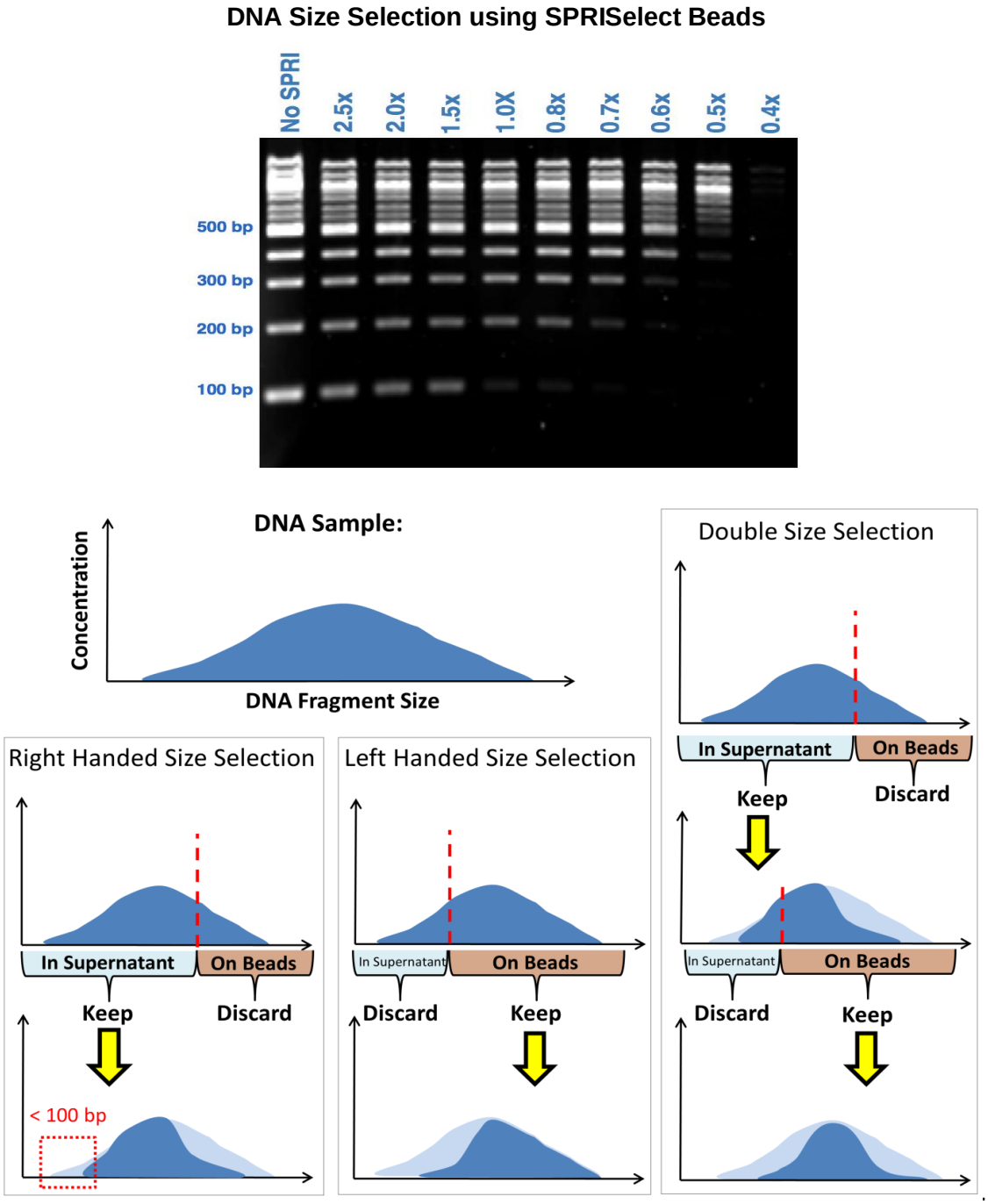


Figure 12: Size selection using SPRIselect beads. Upper image: DNA fragment ranges selected by different ratios of SPRI beads to sample (See Appendix 1 for URL).
Lower image: A DNA sample contains DNA fragments of various sizes at different concentrations. SPRI-select beads can be added to the DNA sample. The beads bind preferentially to larger fragments. The range of fragment sizes bound to the beads depends on the amount of beads relative to the concentration of the DNA sample. This attribute is used to select specific size ranges of DNA fragments: a right handed size selection binds unwanted large fragments to the beads, then discards the bead pellet leaving the desired smaller fragments in the supernatant. A left-handed size selection removes unwanted small fragments by discarding the supernatant and purifying the bead-bound DNA. A double size selection can be performed to acquire a specific range of fragment sizes by removing and discarding large fragments in a right handed size selection, followed by removing and discarding small fragments in a left handed size selection. The beads do not bind to DNA fragments smaller than 100 bp and these are removed in either selection protocol

The beads were shaken thoroughly to re-suspend them in the PEG buffer solution. A 0.4x to 2x ratio of beads to DNA sample volume was added to DNA in 1.5 ml tubes. For example, if a 0.6 x concentration was required for 100 µL DNA, 60 µL SPRIselect beads were added. The solution was mixed briefly by vortexing and then left at room temperature for five minutes. The tubes were then placed in a magnetic rack and left for 3-5 minutes until the solution was clear and the beads had formed a pellet on the side of the tube against the magnet.

For right-handed size selection the supernatant was aspirated and discarded (containing small unbound DNA fragments). Keeping the tubes with the bead pellet in the rack, 500 µL of 85% ethanol (EtOH) was added to the tube and left to stand for one minute without mixing. The ethanol was aspirated and discarded. This was repeated for a total of two washes. The pellet was left to dry at room temperature for 10 minutes. Once the bead pellet was dry, the tubes were removed from the magnetic rack and the pellet re-suspended in the desired volume of molecular grade water.

Left-handed size selection can also be performed. As with right-handed selection, the DNA and beads were mixed together and incubated at room temperature for five minutes, then placed in a magnetic rack. Once the beads have formed a pellet against the magnetic rack and the solution had cleared, the solution was aspirated and retained while the beads (which had bound the larger DNA fragments) were discarded. Additional beads are then added to the solution (taking into account the PEG it already contains) to bind to the remaining small DNA for purification.

Higher throughput bead size selection can be performed in a 96 well plate, using 200 µL rather than 500 µL of EtOH to wash the bead pellet. This process can be carried out by a liquid handling robot.

AMPure XP Bead Purification

AMPure XP beads (Beckman Coulter, USA) are identical to SPRIselect beads except that the bead and buffer concentrations are less accurately made so they are not suitable for fragment size selection.

AMPure beads were vigorously re-suspended and allowed to equilibrate to room temperature for 30 minutes before use. Fresh 80% ethanol was made up using pure ethanol and molecular laboratory grade water. The beads were shaken again before use to re-suspend. The beads (at a volume of 1.8 x beads to the volume of the sample) were added to tubes/plates containing the sample. The beads and sample were mixed by vortexing or pipetting and incubated at room temperature for five minutes. The

samples were transferred to a magnetic plate or rack, where they were incubated for a further 3-5 minutes, until a pellet of beads had collected on the side of the tube against the magnet and the supernatant had cleared. The supernatant was removed and discarded. 500 µL 80% ethanol (for 1.5 ml tubes, or 200 µL for 300 µL 96 well plates or strip tubes) was added to the tube, left for one minute and then aspirated and discarded. This step was repeated for a total of two washes. The sample tubes remained in the magnetic rack during this time and care was taken not to disturb and aspirate any of the pellet. The samples were removed from the magnetic rack and added to a 37°C heat block, or left at room temperature until any trace remnants of ethanol had evaporated. The pellet was then re-suspended in 15-50 µL of nuclease free water.

Vacuum Concentration

Prior to use, the Vacuum Concentrator Plus (Eppendorf, USA) was run for 30 minutes with no samples in it to heat it to 30°C. Samples in open microcentrifuge tubes or PCR plates covered with perforated seals were then placed in the concentrator. Once started, the evaporation status of the samples was inspected every 10 minutes. The samples were removed once all DNA was completely lyophilized. The settings for concentration were: Break ON, Temperature 30°C, Mode V-AQ. The rotor speed was 530 x g. Care was taken to ensure the concentrator was well balanced, with equal volumes of liquid to be evaporated from complimentary positions.

Next Generation Sequencing

For the experiments performed Agilent SureSelect XT library preparation was used in conjunction with Agilent “in-solution” bait capture target enrichment. These DNA fragment libraries were then sequenced on either the Illumina HiSeq 2000 or the Illumina MiSeq. The methodology was well suited to capturing a large contiguous genomic region in a single step without the need for multiplex PCR, therefore the relative proportions of DNA fragments would be preserved during the process allowing identification of copy number variants by relative coverage analysis.

Samples were prepared for sequencing as described in Figure 13. Samples from different individuals were tagged with a 6-8 bp nucleotide index and pooled at an equimolar concentration prior to sequencing. The instrument performs three ‘reads’ of the base sequence of each DNA fragment: one read of the index tag and two reads of the DNA fragment sequence (one forward and one reverse). The instrument outputs the base calls of each individual DNA fragment sequenced and assigns them a quality

score. The 6-8 bp index sequences are read and the various samples are segregated by the sequencing instrument. The output from the sequencer is two FASTQ format files for each indexed sample, one containing the forward fragment sequences and the other containing the reverse sequences. The analysis and interpretation of this data is described later.

Sample Preparation Process for Illumina Sequencing

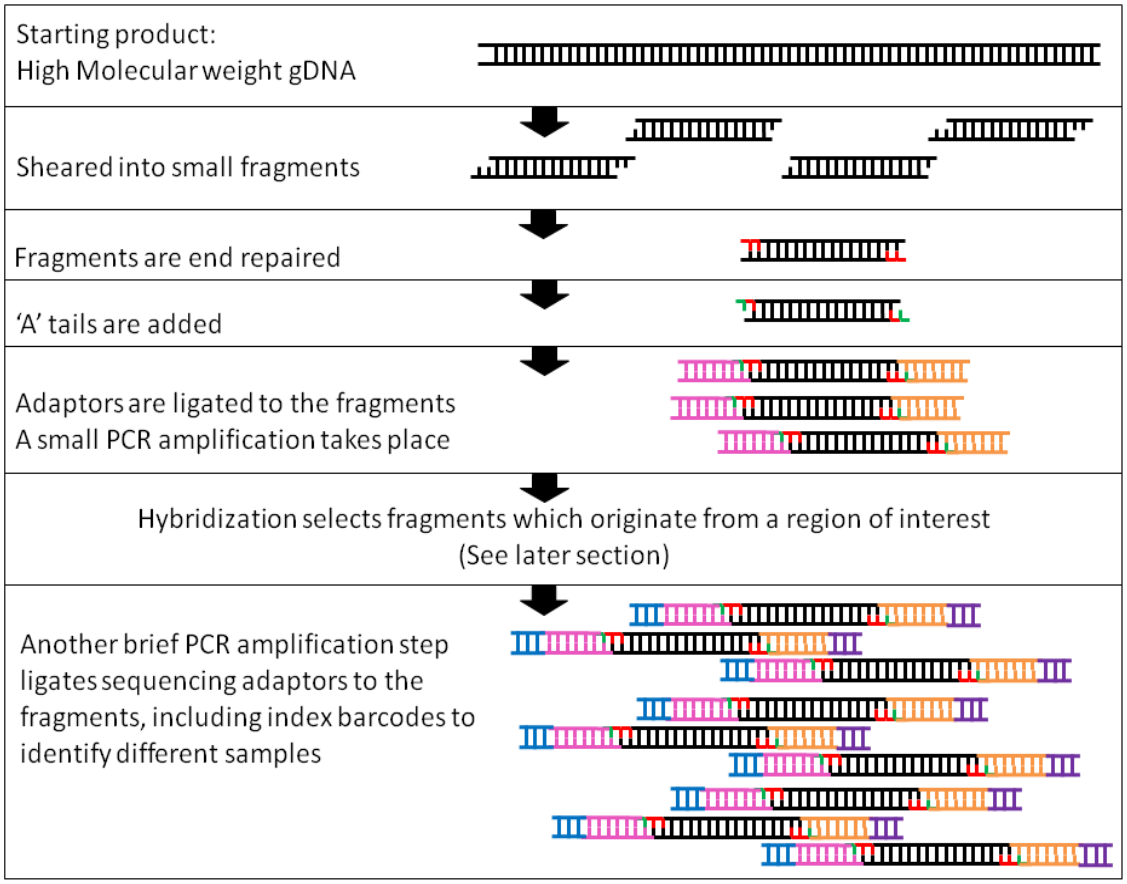


Figure 13: Sample Preparation Process for Illumina Sequencing. DNA is sheared into small fragments, blunt ended, adenylated and ligated to adaptors. Fragments that originated from the target region of the genome for this study were selected for via hybridization to a custom library of oligonucleotide probes, with sequences corresponding to sequences from the target region. These fragments were then isolated from the remainder of the sample through successive wash steps. Index tags were ligated to the fragments and then samples were pooled at an equimolar concentration for sequencing on the Illumina sequencing platforms.

Agilent SureSelect Bait Capture Library Design

RNA baits, 120bp in length, were designed to be complimentary to the region of the human genome that we wanted to sequence. The RNA bait length was automatically adjusted to ensure the Tm between baits did not vary significantly. Baits could be designed to cover the region of interest (ROI) at 1x tiling - so that every base is covered by one bait - or at greater depth up to 5x tiling (every base in the ROI is covered by 5 different baits). For an RNA bait to bind a DNA fragment at a Tm 65°C

they must share at least 40 bp of complimentary sequence. Therefore, baits can capture DNA fragments with sequences extending beyond the bait covered region. This is a helpful feature when trying to sequence through short repetitive regions. The RNA baits are biotinylated and therefore after the baits have hybridised to the DNA they are mixed with iron-containing streptavidin beads, which can be selected for with a magnet. Off-target DNA fragments not bound to the beads are washed away, allowing for enrichment of the ROI. The selected fragments are then sequenced (Figure 14).

SureSelect Target Enrichment Workflow

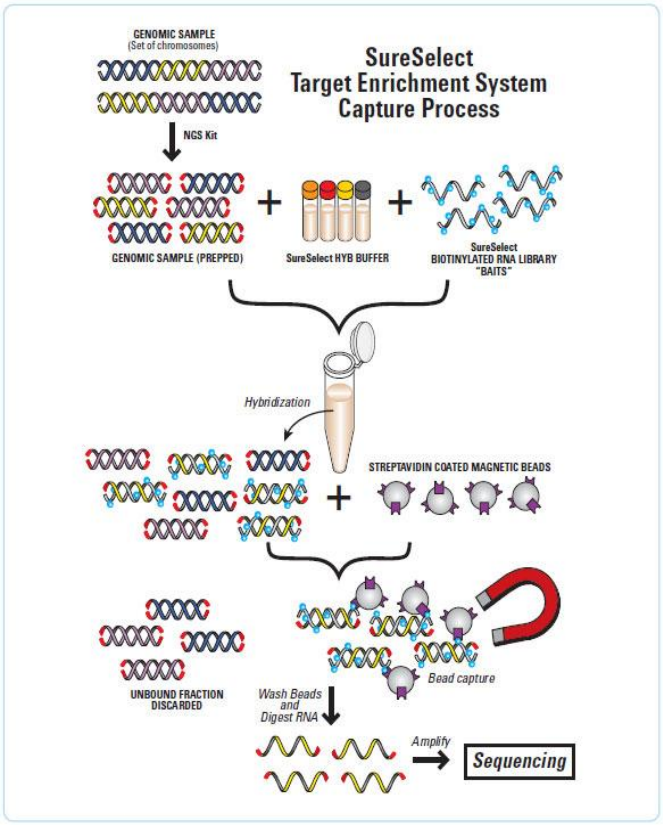


Figure 14 SureSelect Target Enrichment System Workflow. Taken from Agilent web page: 'SureSelect Process – how it works' (For URL see Appendix 1)

Initial Library Design

Bait Capture Library 1 was designed using the in-browser application eArray (Agilent). It targeted 2 regions – one on chromosome 11 and one on chromosome 16 – covering both globin gene loci in two contiguous regions. Non-exonic regions were also included in the design which covered a total of 4.7 Mb: 260 Kb of the alpha globin loci on chromosome 16 and 4.4 Mb on chromosome 11. The small region on the alpha globin gene locus was an error not detected until after submission. The library was designed with a 1x tiling frequency throughout, meaning the baits were head to tail and did not

overlap. Each design can have up to 57,750 baits and therefore this bait limit caused a design constraint. Baits were not designed against repetitive regions because it would not be possible to target sequences specifically from the region of the chromosome we were interested in, as it would capture sequence from every location in the genome where the repeat sequence occurred.

Baits were permitted to extend 20 bp into repetitive regions. Hybridization between RNA baits and DNA fragments required a minimum 40 bp match, so this was expected to be enough to specifically capture sequence from repetitive regions without also picking up large amounts of off target repetitive sequence. By including some of the repeat sequence we hoped to be able to sequence through some of the smaller repetitive regions. A summary of the library design parameters is shown in Table 4.

Table 4 eArray design parameters for Bait Capture Library 1

Parameter	**Selection**
Sequencing Technology	Illumina
Sequencing Protocol	Paired-end long read
Bait length (bp)	120
Bait tiling frequency	1x
Allowed overlap into avoided regions (bp)	20
Strand	Sense
Species	H. Sapiens
Genome Target Intervals	Chr11: 3,244,823 – 7,264,822 Chr16: 0-260,000
Avoid Standard Repeat Masked Regions	Window masker and repeat masker

The library was 'boosted' to compensate for high GC baits which were predicted to perform poorly. The degree of boosting was taken from SureSelect sequence data which showed lower coverage in GC rich regions (Guy's Regional Genetics Laboratory). The boosting parameters are detailed below in Table 5. The GC content of each bait was calculated using an online pipeline at http://galaxyproject.org/ and separated into different percentage groups. The baits in each group could then be amplified accordingly. 'Orphan baits' are baits which are >20 bp from their nearest adjacent bait. These were also boosted to ensure they captured their region efficiently. These orphan baits were further boosted according to their GC content. Low GC/ high AT content was considered less of an issue at the time as the dataset used was from exon targeted panel and genes are considered to be more GC rich. The final baits included in the library design are shown in Figure 15.

Table 5 Boosting parameters for baits in Bait Capture Library 1

GC content	Boost for Non-Orphan Baits	Boost for Orphan Baits
<50%	1x	4x
50-55%	1x	4x
55-60%	2x	6x
60-65%	5x	6x
65-70%	6x	8x
>70%	10x	10x

Bait placement in Bait Capture Library 1

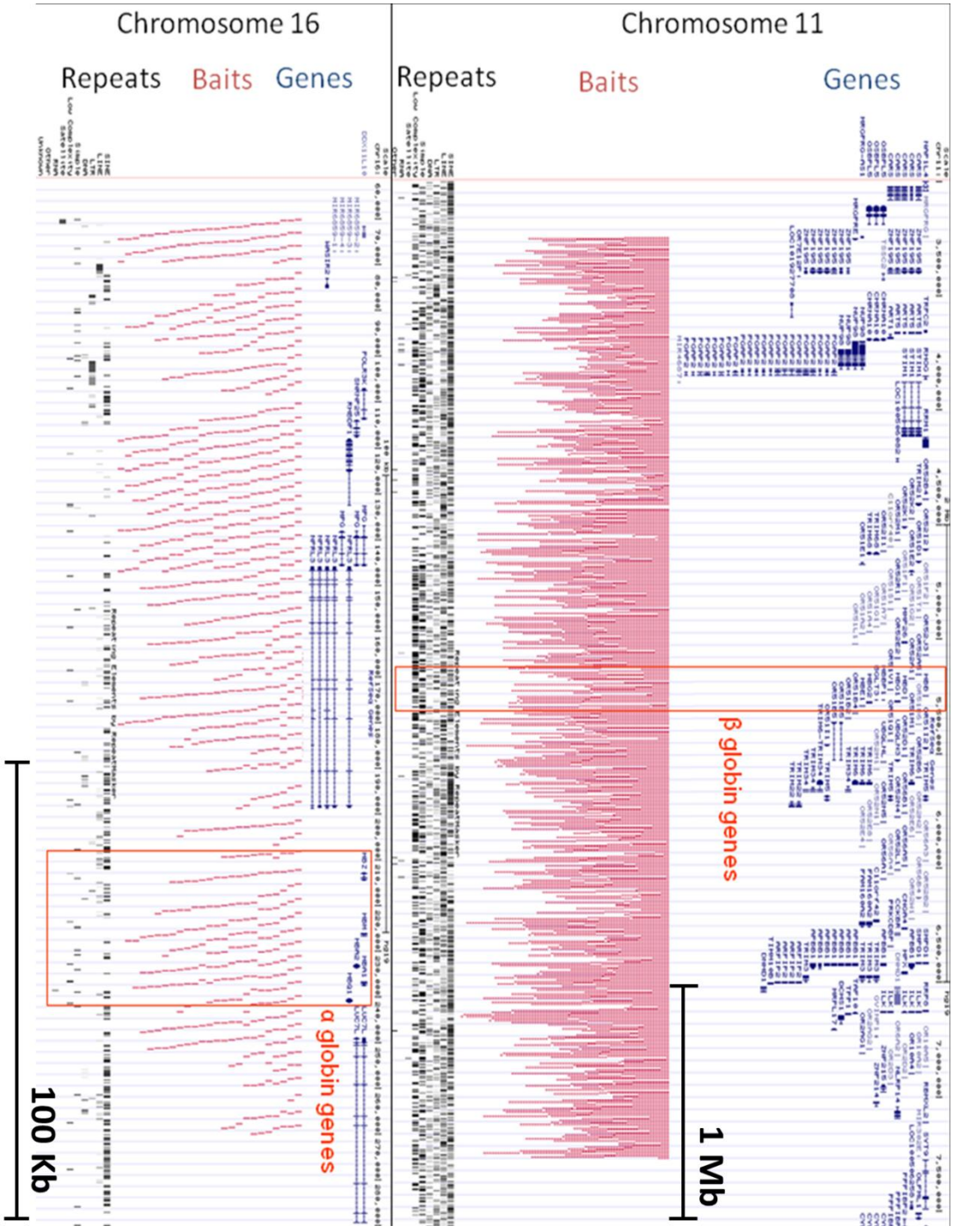


Figure 15 Baits Placement in Bait Capture Library 1. Baits displayed in the UCSC Genome Browser. Panels for each chromosomal region show baits in red, genes in blue and repeats in black. Orange bars highlight the positions of the globin genes.

Second Library Design

The second bait design (Bait Capture Library 2) targeted different regions and used different bait selection parameters to the first design. This was performed using SureDesign (Agilent), the next version on from eArray.

The design was extended to cover a larger region of chromosome 16. Regions on chromosomes 2 and 6 that influence the level of HbF in adults at steady state and can influence the severity of haemoglobinopathies were also included in the design. The X and Y chromosome were also targeted so that the assay could also be used for sex determination (which can be helpful in resolving or confirming sample mix-ups). The covered regions are listed in Table 6.

Table 6 Target Regions for Bait Capture Library 2

Region of Interest	Chromosome	Start of Region	End of Region	Size
Beta globin gene cluster	chr11	4000078	7000522	3 Mb
Alpha globin gene cluster	chr16	60046	2185789	2 Mb
Region containing disease-modifying covariants	chr2	60574969	60730041	156 Kb
Region containing disease-modifying covariants	chr6	135374939	135505003	130 Kb
Sex determination	chrX	11312846	11318819	6 Kb
Sex determination	chrY	2655007	2655666	600 bp

We relied on Agilent's newly developed boosting parameters to the library design. These are listed in Table 7.

Table 7 Max Performance boosting parameters for Bait Capture Library 2 as recommended by manufacturer

GC content	Boost for Non-Orphan Baits	Boost for Orphan Baits
<30%	2x	4x
30 - 60%	1x	2x
60 - 66%	3x	3x
66 -72%	8x	8x
> 72%	16x	16x

Tiling density and boosting parameters varied between different regions of the design, depending how crucial we expected that region to be for diagnosis: baits were placed most densely in the regions of the globin genes where thalassaemia-causing variants were most likely to occur. The regions covered by the design and the bait placement strategy assigned to the regions are outlined in Table 8. The new design covered approximately 2.8 Mb of sequence, falling into the <3 Mb library category (which requires a slight adjustment to the hybridization stage of sample preparation). The positions of baits on chromosomes 11 and 16 is shown in Figure 16 and the additional regions included on chromosomes 2, 6, X and Y are shown in Figure 17.

Table 8 Regions included in Bait Library Design 2, bait placement conditions and sequence covered

Region name	Bait placement: -Tiling* -Placement** -Boost***	Position	Bp Covered	Probes
Region 1 Core Region	- 2x -Moderate Stringency* -Max performance**	Chr11: 4750000-6250000 Chr16: 0-600000	1,290,000	34,016
Region 2 Flanking regions	- 1x -Moderate Stringency -Max performance	Chr11: 4,500,000-4,750,000 Chr11: 6,250,000-6,500,000 Chr16: 600,000-1,000,000	635,547	16,730
Region 3 Upstream ATX genes	- 1x -Moderate Stringency -Max performance	PKD1 chr16: 2,138,711-2,185,899 TSC2 chr16: 2,097,990-2,138,713 (CDS only)	26,014	1,323
Region 4 Orphans	- 1x -Moderate Stringency -Max performance	Orphans from Regions 1, 2, 3	-	1687
Region 5 Fringe regions	- 1x -Moderate Stringency -Balanced	500bp segments placed every 5Kb: Chr11: 4,00,000 – 4,750,000 Chr11: 6,500,000-7,000,000 Chr16:1,000,000-2,097,990	206,065	2191
Region 6 Sex determination	- 2x -Moderate Stringency -Max performance	ChrY: SRY Gene ChrX: AMELX Gene	1,933	41
Region 7 Research genes	- 1x -Moderate Stringency -Max performance	Chr2: 60,575,000-60,625,000 Chr2: 60,710,000-60,730,000 Chr6: 135,375,000-135,505,000	200,003	1,489
Total			**2,359,562**	**57,477**

*Tiling density is the number of baits tiled over each base position in the target region. 1x tiling covers each base once. 2x tiling overlaps bait positions by 50%, so each base is covered by 2 baits.

**The bait placement algorithms avoid tiling over repetitive regions. The stringency with which repetitive regions are avoided can be: ‘least’ (3 separate repeat-masking tools mask avoided regions); moderate (2 separate repeat-masking tools mask avoided regions) or ‘most’ (any sequences masked by the Repeat Masker repeat-masking tool are avoided).

***Boosting parameters determine the number of replicates of high-GC baits which are included in the library: ‘Max Performance’ boosts all probes in the design region according to the parameters in Table 7; ‘Balanced’ uses less probes and only applies a boost in coding regions; ‘no boosting’ doesn’t boost any probes.

Bait Capture Library 2 Bait Locations: Chromosomes 11 and 16

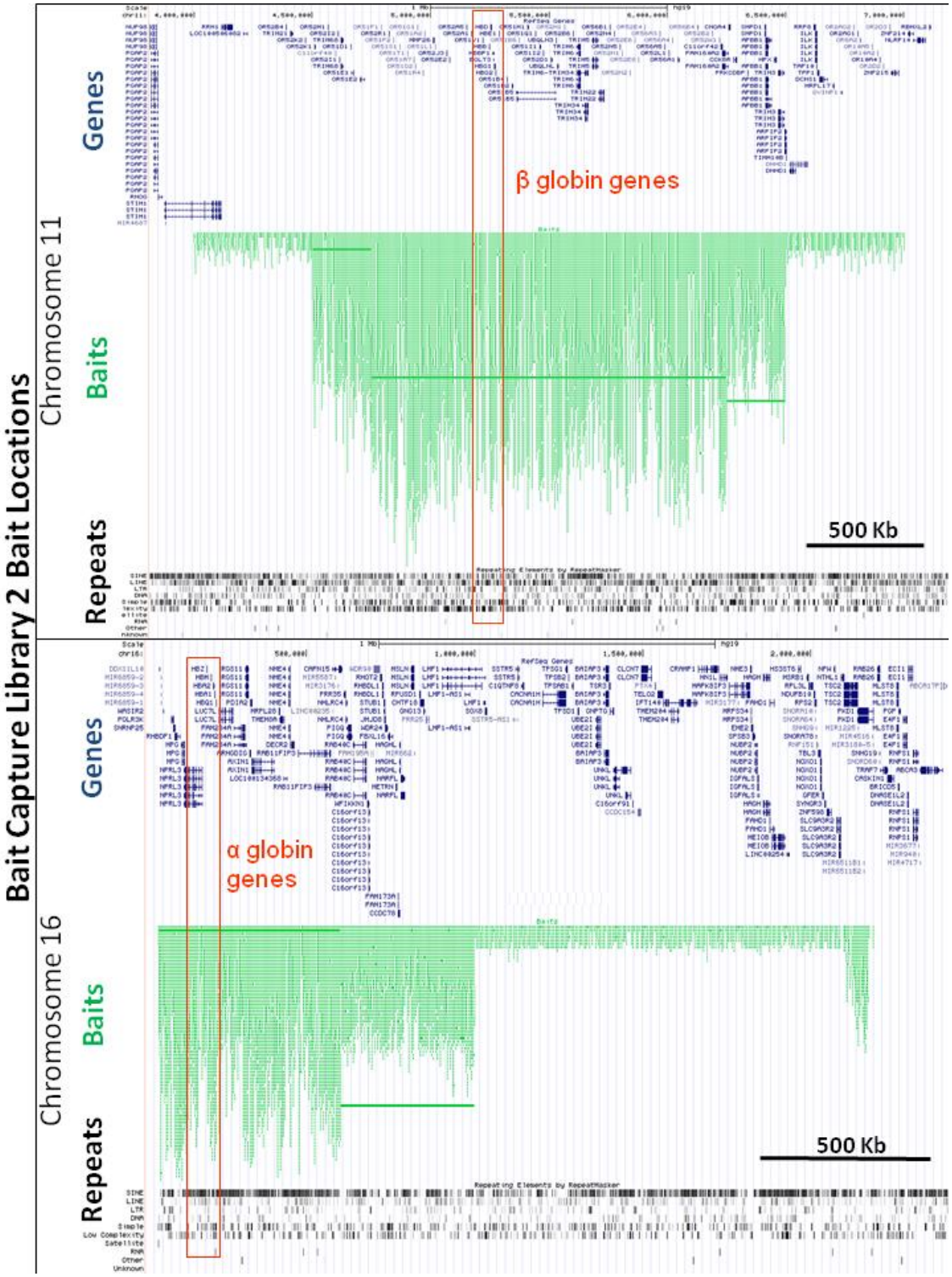


Figure 16: Bait Capture Library 2 Bait Locations: Chromosomes 11 and 16. Genes are shown in blue, baits are shown in green and repeats are shown in black. NB: Green lines appearing on the graph are due to incorrect rendering of all dense tiling in the genome browser.

Bait Capture Library 2 Bait Locations: Chromosomes X, Y, 6 and 2

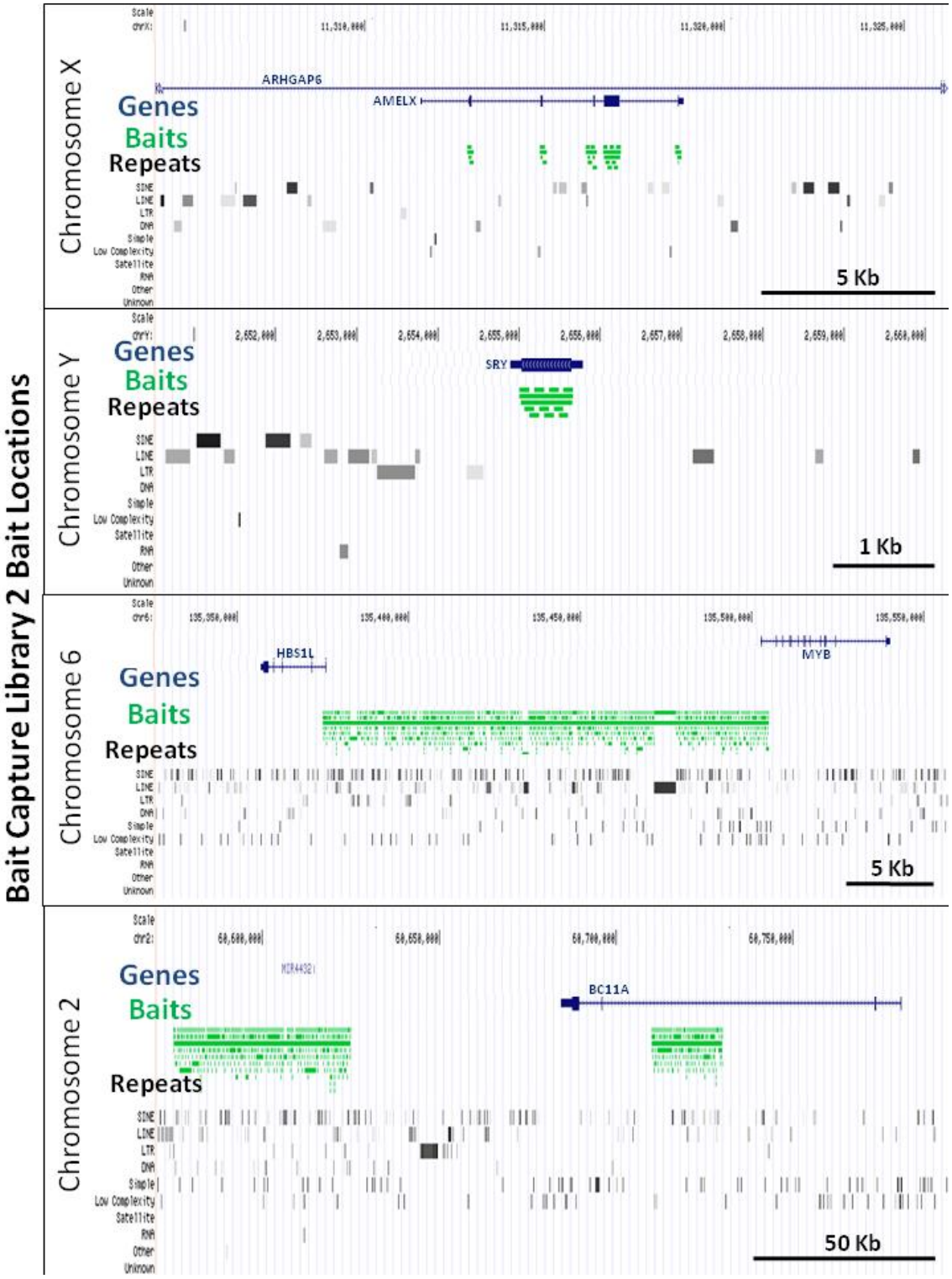


Figure 17: Bait Capture Library 2 Bait Locations: Chromosomes X, Y, 6 and 2. Bait positions are shown in the UCSC Genome Browser. Genes are shown in blue, baits are shown in green and repeats are shown in black. NB: Green lines appearing on the graph are due to incorrect rendering of dense tiling in the genome browser.

Manual Library Construction Process

Experiments were performed in accordance with SureSelect Library Preparation kit (Agilent, USA) protocol version G7530-90000 (2012). For an overview of this process, refer to Figure 13 and Figure 14. Reagents from the following kits were used: Agilent SureSelect Library Preparation Kit (Part #G9612A), Agilent SureSelect Custom Target Enrichment Kit (Bait Capture Library 1 Part # 5190-4827; Bait Capture Library 2 Part #5190-4817), Agilent Herculase II Fusion Kit (Part #600675). The contents of reagent mixes are listed in 'Appendix 1: List of Solutions'.

DNA samples

Non-degraded genomic DNA (3 – 5 µg) was sheared into fragments (variously 150 bp, 500 bp, 600 bp) using the Covaris or Bioruptor in accordance with the manufacturer's protocol. Shearing was checked using the Tapestation or Bioanalyser.

The sheared product was purified using AMPure XP beads, with an input volume of 180 µL beads, eluting in 50 µL nuclease-free water.

End repair

End repair reaction mix was made up on ice for each DNA sample (see appendix). For multiple samples, a master mix was made with 0.5 reactions excess. The end repair mix (52 µL) was added to 48 µL DNA sample and mixed by pipetting. The reaction mix was incubated in a thermal cycler for 30 minutes at 20°C without a heated lid.

This step is followed by an AMPure XP bead clean-up with using a bead volume equivalent to 1.8x the reaction volume, eluting in 32 µL nuclease-free water. This ensured recovery, clean-up and concentration of all DNA fragments from the reaction with no size selection.

Adenylation

Adenylation reaction mix was made up on ice for each DNA sample (see appendix). For multiple samples, a master mix was made with 0.5 reactions excess. 20 µL reaction mix was added to 30 µL DNA sample and mixed by pipetting. The mix was incubated on a thermal cycler for 30 minutes at 37°C without a heated lid.

This step is followed by an AMPure XP bead clean-up with using a bead volume equivalent to 1.8x the reaction volume, eluting in 15 µL nuclease-free water. This ensured recovery, clean-up and concentration of all DNA fragments from the reaction with no size selection.

Adapter Ligation

Adaptor ligation reaction mix was made up on ice for each DNA sample (see appendix). For multiple samples, a master mix was made with 0.5 reactions excess. 37 µL reaction mix was mixed with 13 µL DNA sample and mixed by pipetting. The mix was incubated for 15 minutes at 20°C on a thermal cycler without a heated lid.

This step is followed by an AMPure XP bead clean-up using a bead volume equivalent to 1.8x the reaction volume, eluting in 32 µL nuclease-free water. This ensured recovery, clean-up and concentration of all DNA fragments from the reaction with no size selection

Library amplification

The optimal quantity of DNA for this reaction is 250 ng which was quantified on the QuBit. Surplus DNA was stored at –20°C. Pre-Hybridization PCR mix was made up on ice for each DNA sample (See appendix). For multiple samples, a master mix of the reactants was made with 0.5 reactions excess. The thermal cycler program was started (Table 9).

Table 9 Thermal Cycler program for Library Amplification

Step:	Step 1	Step 2 (x6)			Step 3	Step 4
Temperature:	98°C	98°C	65°C	72°C	72°C	4°C
Time:	2 minutes	30 seconds	30 seconds	1 minute	10 minutes	Hold

NB: *Step 2 was repeated for a total of 5 cycles for QIAsymphony extracted samples and a total of 6 cycles for phenol-chloroform extracted samples due to their different amplification efficiencies*

This step is followed by an AMPure XP bead clean-up using a bead volume equivalent to 1.8x the reaction volume, eluting in 30 µL nuclease-free water. This ensured recovery, clean-up and concentration of all DNA fragments from the reaction with no size selection

During library preparation the size and concentration of the fragments was measured on a Bioanalyser or a Tapestation (D1000 kit). As adapters were ligated to the DNA fragments they increased in size. Unincorporated adapters or non-ligated fragments could be seen and measured in the analysis therefore assessing reaction efficiency. Small peaks (<100 bp) indicate the presence of primer dimer which must be removed by further AMPure purification. A ‘hump’ in the peak can indicate over amplification of a

subset of DNA fragments within the library (See MiSeq Sample 5 for an example of this).

The concentration of prepared libraries was calculated using the QuBit (Broad Range Kit). 500 ng of DNA was used for library hybridization.

Hybridization

To avoid library evaporation, the labware for hybridization was tested prior to use: The combination of PCR plate, thermal cycler and sealing caps to be used were incubated at 65°C for 24 hours with wells containing 30 µL of water. Labware combination was changed if evaporation exceeded 33%. NB: This step of the process is complex and for clarity key terms have been colour coded.

Five hundred ng of adapter ligated DNA was placed in 96 well plate and then completely lyophilized using a Vacuum Concentrator Plus (Eppendorf). A SureSelect Custom Target Enrichment Kit was used to hybridize the DNA to a custom bait library. The process is summarised below in Figure 18.

Hybridization Buffer was prepared at room temperature for each reaction Occasionally a precipitate formed in the mix, in which case it was incubated at 65°C for 5 minutes. SureSelect Capture Library Mix was prepared on ice in a separate tube. The solution was mixed by pipetting and kept on ice until use. SureSelect Block Mix was prepared on ice in a separate tube (for reagents used to make each mix, see Appendix 1: List of Solutions)..

The lyophilized DNA libraries were reconstituted in 3.4 µL of molecular grade water in separate wells of a PCR plate. 5.6 µL SureSelect Block Mix was added to each DNA sample. The DNA and Block Mix were mixed by pipetting up and down. The wells were sealed firmly with caps and placed in the thermal cycler. The thermal cycler program (Table 10) was started.

Table 10 Thermal Cycler Program for Hybridization

Step:	Step 1	Step 2
Temperature:	95°C	65°C
Time:	5 minutes	Hold
Lid Temp: 105°C		

Reagents were loaded into the plate as illustrated in Figure 18: Once the thermal cycler reached 65°C, 40 µL Hybridization Buffer for each sample was loaded into separate

wells of the PCR plate in the thermal cycler and firmly sealed with caps. After the Hybridization Buffer had been incubated in the PCR plate at 65°C for 5 minutes, 7 µL of the Capture Library for each sample was added to separate wells of the PCR plate in the thermal cycler. The wells were sealed firmly with caps and the program allowed to run for a further 2 minutes. The lids were then removed and using a multichannel pipette, 13 µL of Hybridization Buffer was transferred into each of the wells containing the Capture Library. With the plate remaining in the thermal cycler, the lids were then removed from the wells containing the prepared DNA Library and SureSelect Block Mix. The entire contents (7 µL) of the wells were transferred to the wells containing the Capture Library and Hybridization Buffer using a multichannel pipette and mixed by pipetting slowly up and down 10 times before being sealed. A thermal mat was placed on top of the plate to improve the seal and reduce evaporation. The thermal cycler was then closed, and the reactants left to incubate at 65°C for 24 hours.

PCR Plate Setup for SureSelect Hybridization

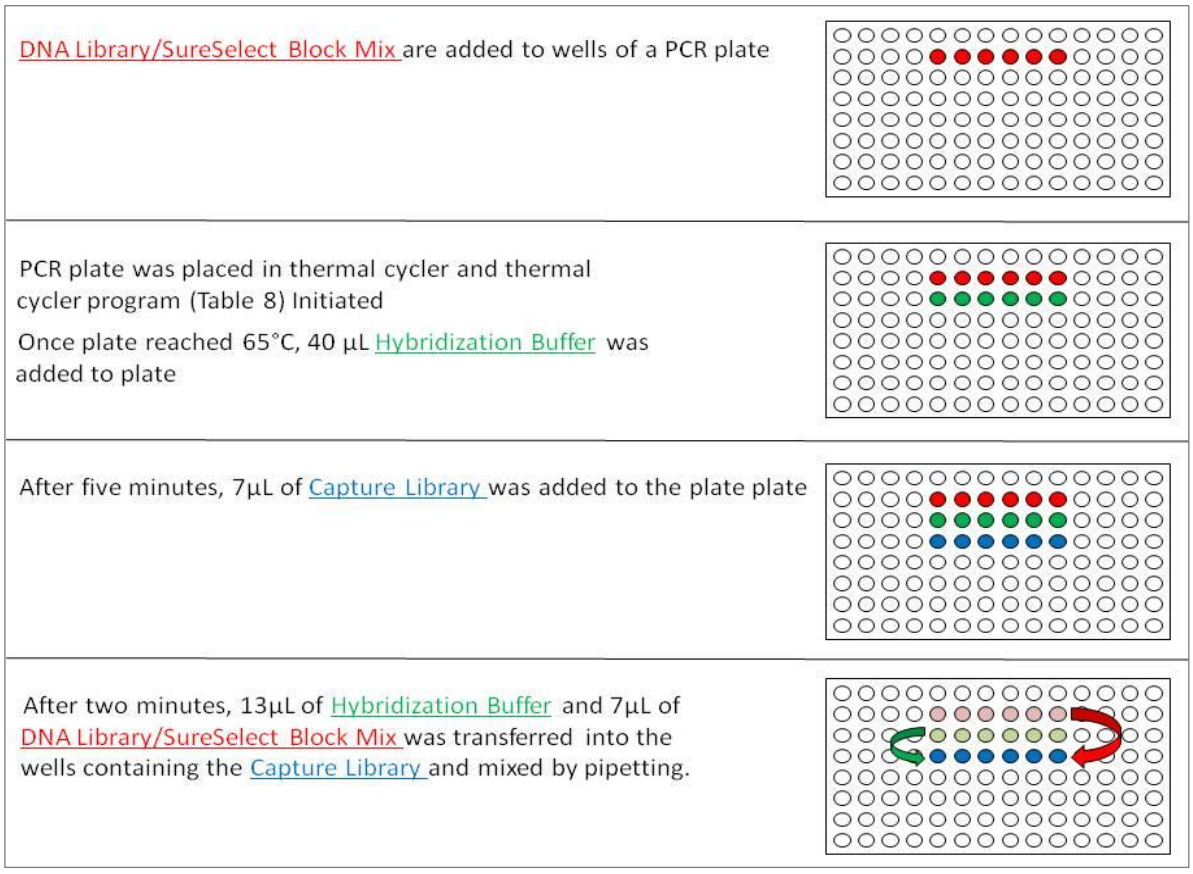


Figure 18 PCR Plate Setup for SureSelect Hybridization. Reagents and the wells to which they should be added in a 96 well plate are colour-coded.

Separation of bait-bound DNA using Streptavidin beads

Washing the beads

Streptavidin beads bind to the biotinylated baits which are hybridised to the DNA fragments of interest. The Streptavidin beads contain iron and can be held by a magnet, therefore while the beads bind the DNA fragments from the region of interest, the unwanted DNA fragments are washed away.

The Streptavidin beads were prepared 30 minutes prior to the completion of the 24 hour hybridization stage by washing them to remove preservatives: Dynalbeads, MyOne Streptavidin T1, were vigorously re-suspended on a vortex mixer. For each hybridization, 50 µL of beads were pooled in a single 1.5 ml microcentrifuge tube (max 250 µL or 5 reactions). The beads were washed by adding 200 µL (per 50 µL beads) of SureSelect Binding Buffer, vortexed for 5 seconds and then placed in a magnetic rack. Once the beads had collected against the magnet, the supernatant was removed and discarded. After this had been repeated for a total of three washes, the beads were re-suspended in 200 µL (per 50 µL beads) SureSelect Binding Buffer and divided into separate tubes for each hybrid capture.

Binding the Hybridized library to the beads

The hybridized libraries were added directly from the thermal cycler at 65°C into the microcentrifuge tubes containing 200 µL of room temperature beads and buffer. The tubes were inverted 3-5 times to mix and then attached to a Nutator and mixed gently at room temperature for 30 minutes. SureSelect Wash Buffer 2 was added to a 65°C heat block during this time for use later. After 30 minutes the microcentrifuge tubes were centrifuged briefly to remove any liquid from the lid. The tubes were then added to a magnetic separator rack and the supernatant was removed from the beads after standing for five minutes.

Wash stage one

The tube was removed from the magnetic rack and the bead pellet was re-suspended in 500 µL SureSelect Wash Buffer 1 (room temperature) by mixing on a vortex mixer for five seconds. The sample was then incubated for 15 minutes at room temperature and mixed briefly on a vortex mixer every three minutes. The microcentrifuge tubes were briefly spun in a centrifuge, and the supernatant was removed and discarded. This was repeated twice for a total of 3 washes in SureSelect Wash Buffer 1.

Wash stage two

The bead pellet was resuspended in 500 µL Wash Buffer 2 (pre warmed to 65°C) by vortexing for five seconds. The samples were incubated 65°C for 10 minutes, vortexing briefly every three minutes. The samples were placed in a magnetic rack. Once the supernatant cleared it was removed and discarded. This was repeated for a total of

three washes in pre-warmed Wash Buffer 2. The DNA was eluted with 30 µL nuclease free water.

Selecting optimal bar codes for sequencing runs

The barcodes used to identify different samples in a pool are supplied within the Agilent XT Library Preparation kit. When low numbers of barcodes are being used some combinations perform better than others. This is because the MiSeq camera uses the first cycles of sequencing by synthesis (which read the barcodes) to calibrate the camera and this requires it to see each different base fluoresce in each cycle. If neither the red nor green channel is activated for any of the samples at one base position, it will not be recorded. This causes the MiSeq to record incorrect barcode sequences which may inhibit de-multiplexing. Pooling recommendations distributed by Agilent were followed for low-plexity pools. These are outlined in the Table 11.

Table 11 Optimal index combinations for low sample pools

Pool of 2 samples:	
Index #6	GCCAAT
Index #12	CTTGTA
Pool of 3 samples:	
Index #4	TGACCA
Index #6	GCCAAT
Index #12	CTTGTA
Pool of 6 samples:	
Index #2	CGATGT
Index #4	TGACCA
Index #5	ACAGTG
Index #6	GCCAAT
Index #7	CAGATC
Index #12	CTTGTA

Add index tags to library

Indexing PCR mix was prepared using kit reagents (See appendix) on ice. For multiple samples, a master mix of the reagents with 0.5x excess was prepared. 35 µL was added to 14 µL captured on-bead DNA with 1 µL of the appropriate indexing tag for a total reaction volume of 50 µL. The thermal cycler program (Table 12) was started.

Table 12: PCR program for post-capture indexing PCR

Step:	Step 1	Step 2 (x12)			Step 3	Step 4
Temperature:	98°C	98°C	57°C	72°C	72°C	4°C
Time:	2 minutes	30 seconds	30 seconds	1 minute	10 minutes	Hold

This step is followed by an AMPure XP bead clean-up with using a bead volume equivalent to 1.8 x the reaction volume, eluting in 32 µL nuclease-free water. The clean-up was performed in a designated post-PCR area. Recovery, clean-up and concentration captured all DNA fragments with no size selection.

Quantifying and pooling the DNA

The optimal amount of sequencing data per sample depends on the capture library size. Bait Capture Library 1 is a custom library of 4.26 Mb and the optimal output per sample is 426 Mb. For Bait Capture Library 2 which covers 2.78 Mb the optimal output per sample is 278 Mb.

The Bioanalyser or Tapestation High Sensitivity kit was used to determine the quality and size range of DNA fragments in each sample, and to detect any primer dimer (if primer dimer was found additional clean-ups with the AMPure XP beads were performed until the dimer disappeared). The QuBit High Sensitivity kit was used to determine the concentration of the DNA samples and the pM concentration could be calculated using the molecular weight of an average nucleotide base pair and the mean fragment size taken from the Tapestation.

Samples were pooled together to make an equimolar 2 nM pool in molecular grade water. These prepared libraries were stored at -20°C if not being sequenced immediately.

Sequencing on HiSeq 2000 (in collaboration with Alexander Smith)

Cluster generation and sequencing on the HiSeq 2000 was performed off site at the Department of Haematological Medicine, Rayne Institute, King's College London by Dr Alex Smith. One 12 Sample Run was performed. Cluster generation was carried out using reagents from the Illumina HiSeq PE Cluster Kit (v2) (Illumina part # PE-401-4001). Sequencing was performed using the HiSeq PE Sequencing Kit (v2) (Illumina # FC401-4001).Tris-Cl and NaOH were obtained from generic laboratory suppliers. The

steps involved in cluster generation and HiSeq setup according to the current version of the protocol supplied by Illumina are outlined in Appendix 3.

Sequencing on MiSeq

The following steps are in accordance with the protocol provided by Illumina "*Preparing Libraries for MiSeq (protocol part number 15039740 D)*". The MiSeq could produce 7Gb data using the V2 2x 250 bp reagent kits and 13.2-15 Gb data using the V3 2x 300 bp reagent kits.

Denaturation and Dilution of DNA Sample Pool

DNA samples that had been prepared for NGS were pooled together at a 2 nM concentration. 4-6 samples were pooled per run.

The reagent cartridge was removed from -20°C storage and placed in a room temperature water bath up to the 'max' line on its' side (approximately one hour to thaw and stabilise). The tube of HT1 Buffer packaged with the cartridge was set aside at room temperature until thawed. Once thawed, both were stored at 2-8°C until used.

0.2 M NaOH was prepared freshly. Mixing 800 µL molecular grade water with 200 µL of 1.0 M NaOH stock. The tube was inverted to mix. The dilution can be used for up to 12 hours after preparation.

2 nM pooled DNA library (5 µL) and 0.2 M NaOH (5 µL) were added to a microcentrifuge tube and vortexed briefly to mix. The sample was then micro centrifuged for 1 minute at 280 x g and incubated for 5 minutes at room temperature to denature the double stranded DNA. HT1 Buffer was removed from 2-8°C storage and 990 µL of the buffer was added to the DNA/NaOH mix. This diluted the DNA to form a 10 pM denatured library in 1mM NaOH, which was placed on ice.

The library could be retained at a 10 pM concentration, or further diluted using pre-chilled HT1 to 6 pM (360 µL DNA : 240 µL HT1) or 8 pM (480 µL DNA : 120 µL HT1) depending on desired clustering density. Dilution was followed by pulse mixing and brief centrifugation, and the denatured diluted library was stored on ice until use.

Denaturation and Dilution of Phi-X Control

2 µL of the Phi-X control was combined with 3 µL 10 mM Tris-Cl PH 8.5 with 0.1% Tween 20 in a microcentrifuge tube. 5 µL 0.2 M NaOH (prepared within the last 12 hours) was added to the Phi-X control. The mixture was vortexed briefly and

centrifuged at 280 x g for 1 minute. The tube was then incubated for 5 minutes at room temperature.

Pre-chilled HT1 (990 µL) was added to the 10 µL Phi-X library. The denatured library can be stored at -20°C for further use.

For a v3 reagent kit, Phi-X control is used at this concentration (20 pM). For a v2 reagent kit, a final dilution was performed prior to running, by combining 375 µL Phi-X with 225 µL pre-chilled HT1 to give a final working concentration of 12.5 pM and inverted several times to mix.

Spiking the Denatured DNA Library with Phi-X

Diluted denatured Phi-X control (6 µL) was added to 594 µL diluted denatured DNA pool, resulting in a 1% Phi-X spike. The mixture was stored on ice until use.

Setting up a Run on the MiSeq

The following steps are in accordance with the protocol provided by Illumina "*MiSeq System User Guide (part #15027617)*"

Before removing reagents from cold storage, ensure that any pre-run washes required by the MiSeq have been carried out, with the wash cartridge filled with laboratory grade water.

The thawed reagent cartridge was removed from 4°C storage. A lint free lab tissue was used to clean the foil seal over the reservoir labelled 'load samples'. The foil was then carefully pierced using a 1 ml pipette tip. 600 µL prepared library with Phi-X spike was loaded into the opened reservoir, with care taken to avoid getting any sample on the foil seal.

On the MiSeq instrument screen:

Open 'Manage Instrument' and select 'Reboot' to restart the software. Reopen 'Manage Instrument' and select 'Sequence'. On the first screen – BaseSpace Options – select 'Use BaseSpace for Storage and Analysis' and click 'NEXT'. A prompt appears to load the flow cell.

Remove flow cell from cold storage at 4°C. Carefully extract flow cell from container using tweezers. Avoid touching glass with tweezers. The flow cell was carefully rinsed with laboratory grade water. The cell was then dried with lint-free lens cleaning tissues. The flow cell compartment was opened. A button releases the latch for the flow cell

stage. The flow cell was placed carefully on the stage, without touching the glass and the latch closed to hold the cell in place. The compartment was then closed.

Once the flow cell is in place, the next prompt on the MiSeq instrument screen instructs the user to load the reagents and reagent cartridge.

PR2 reagent bottle was removed from cold storage and inverted gently to mix. Reagent compartment was opened. Sipper handle was raised and locked into place. PR2 bottle was placed in right side of upright reagent chiller with lid removed. The waste bottle on the right of the upright reagent chiller was emptied and replaced. The lever was then lowered

Reagent Cartridge was inserted into cartridge space on left hand of reagent chiller and the door was closed. Screen must show cartridge RFID has been read.

A sample sheet was uploaded to the Experiment manager. Sample sheets can be created in advance with experiment manager or for experienced users as .csv files. Run parameters listed in sample sheet are shown in (Table 13).

Table 13 Experiment Manager Sample Sheet for MiSeq Sequencing

Parameters							
IEMFileVersion	4						
Investigator Name	[Initials]						
Experiment Name	[Experiment]						
Date	[Date]						
Workflow	Generate FASTQ						
Application	FASTQ Only						
Assay	TrueSeq LT						
Chemistry	Default						
Reads	[151/251/351 dependent on kit]						
Sample Datasheet							
Sample ID	Sample Name	Sample Plate	Sample Well	17_Index_ID	Index	Sample Project	Description
[ID]	[Name]	[blank]	[blank]	Index number	Index Sequence	[Project name]	[blank]

After checking all the run parameters are correct the user tells the instrument to perform a pre-run check. Once this check is completed, the run can be started.

Various MiSeq kits were used according to their availability and the practicalities of each kit for the experiment at hand. Variously, the V2 2x150, 2x500, V3 2x150, V3 2x250 and V3 2x300 kits were used.

All sequencing data produced from the MiSeq was stored on BaseSpace, a cloud computing environment provided by Illumina. Three reads are taken of each fragment cluster: an identifying read (that reads the indexing barcode), a forward read of the DNA fragments (length of which depends on sequencing kit used) and a reverse read of the fragments (Figure 19). The 2 main sequencing files generated for each sample are FASTQ forward and a FASTQ reverse read. FASTQ format uses one ASCII character to denote both the nucleotide and quality score of each position in a sequence read. Illumina machines output data in FASTQ format which also includes sequence identifiers, such as flow cell lane and X-Y co-ordinates of the cluster of origin within flow cell. Format conversion removes reads where the quality scores for the reads do not meet user defined criteria. The quality information is stripped from the remaining sequences, converting them into sequence-only FASTA format.

Illumina sequencing platforms sequence each DNA fragment on the flow cell in a forward and reverse direction (see Figure 19). This creates 2 ‘reads’ for each DNA fragment which are linked by their names. ‘Read one’ reads the fragment in a 5’ to 3’ direction and ‘Read 2’ reads in a 3’ to 5’ direction. If one read from a DNA fragment does not pass format conversion its partner is also removed at this stage.

Generation of Read 1, Read 2 and Read 3 on Illumina Sequencing Platforms

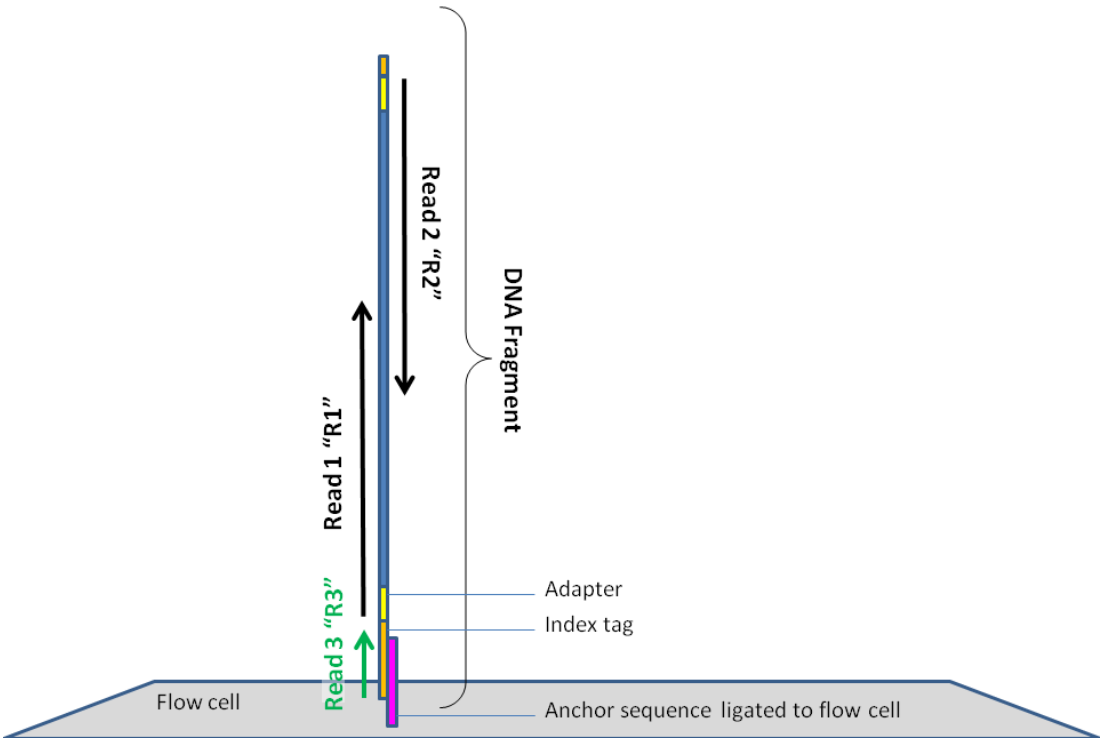


Figure 19: Generation of Read 1, Read 2 and Read 3 on Illumina Sequencing Platforms. Three reads are taken of each DNA fragment ligated to the flow cell: Read 1 and Read 2 read the sequence of the DNA fragment from opposite directions for (n) bases depending on the reagent kit used. Read 3 reads the index tag so that the fragment sequences can be attributed to the correct sample from the DNA pool. The Read 1 and Read 2 data is then output as a pair of linked files in FASTQ format grouped by the index identified in Read 3.

The PhiX control that is spiked into the pooled DNA samples is used by BaseSpace to produce quality metrics about the run according to the alignment of read data generated by the PhiX control to the PhiX reference genome (Table 14). This quality data can be used to make inferences about the quality of the run for the pooled DNA samples. These values are updated throughout the progress of the run, thus inspection of the values after the run has completed will indicate values for either cycle 300, or cycles 1-300. Quality scores are calculated for each base call generating logarithmic values of call certainty (Q10 = 10% chance of wrong base call; Q20 = 1%, Q30 = 0.1%, Q40 = 0.01%).

Table 14 MiSeq Run Quality Metric Details, from the BaseSpace User Guide (Illumina)

Metric	Explanation
Yield total	Number of bases sequenced in total
Yield perfect	Number of reads that contain no base calling errors in the PhiX control
Yield >=3 errors	Number of reads containing three base call errors or less in the PhiX control
Aligned	Percentage of total reads on flow cell that aligned to the PhiX genome
Perfectly Aligned (%)	Proportion or bases in PhiX reads that align perfectly to the reference sequence (at read cycle indicated in square brackets)
>=3 errors (%)	Proportion or reads in PhiX reads that align with three base call errors or less to the reference sequence (at read cycle indicated in square brackets)
Error rate	Overall error rate in all reads aligning to the PhiX genome
Q >=3- (%)	Percentage of bases with a quality score >30
Tiles	The flow cell is divided into strips which are subdivided into 'tiles'. Tiles are the data unit of processing corresponding to an imaging area of the flow cell.
Density	Number of clusters (thousands per mm^2) detected. Optimal cluster density for 865-965 k/mm^2 for v2 chemistry and 1200-1400 k/mm^2 for v3 chemistry
PF	Number of clusters that pass filtering (+/- one standard deviation)
Phas./Prephas	Number of DNA molecules within a cluster that fall out of phase (by jumping ahead or lagging behind) the rest, based on the first 25 cycles.

Reads	Number of clusters (millions)
Reads PF	Number of clusters passing filtering (millions). Filtering removes clusters that cannot be consistently identified in the first 25 cycles of sequencing.
Q>=30	Bases with a quality score greater or equal to 30. Scores expected per kit: MiSeq Reagent Kit V2 2x150bp >80% >Q30 2x250bp >75% >Q30 MiSeq Reagent Kit V3 2x300 >70% >Q30
Yield	Number of bases passing filtering
Cycles error rated	Cycles error rated by PhiX control
Aligned	Percentage of reads aligned to PhiX genome
Error rate	Error rated determined by PhiX alignment during cycles 1-35, 1-75 and 1-100
Intensity cycle 1	The intensity of the wavelength of light detected by the MiSeq for each base. This data can have diagnostic use if a run performs unexpectedly poorly.
Intensity cycle 20	

Data Analysis (NextGene)

Format conversion

Raw sequencing data was converted from FASTQ format to FASTA format for alignment to the reference sequence (Hg19).

Format Conversion was performed using the standard quality filtering settings recommended by NextGene which are listed in Table 15.

Table 15 Default format conversion settings in NextGene

Parameter	Setting
File Format Type	Illumina
Median Score Threshold	>= 20
Max number of uncalled bases	<=3
Called Base Number of Each Read	>=25
Trim or Reject Read when	>= 3 bases score <= 16
Paired Reads Data	TRUE

Alignment settings

Alignment settings for HiSeq data are listed in Table 16 Alignment Settings for HiSeq Data. Whole genome alignment settings, which were adopted later and also used for alignment of MiSeq data are listed in Table 17.

Table 16 Alignment Settings for HiSeq Data

Parameter	Setting
Reference Sequence	Human v37_2
Instrument Type	Illumina
Application Type	SNP/Indel Discovery
Steps	
Do Condensation	FALSE
Do Assembly	FALSE
Do Alignment	TRUE
[Sequence Alignment]	
Whole Gene Project	TRUE
Directory of Index	C
Do WGA With Paired Index	FALSE
Do Methylation Detection	FALSE
Reads	Allowable Mismatched Bases 2
Allowable Ambiguous Alignments 500	
Seed	35 Bases, Move Step
Allowable Alignments 500	
Overall	Matching Base Percentage >= 80.0
Detect Large Indels	FALSE
Sequence Range Checked	FALSE
Hide Unmatched Ends	FALSE

Mutation Percentage <= 15.00	
SNP Allele <= 3 Counts	
Total Coverage <= 30	
Except for Homozygous	TRUE
Mutation Filter Use Original	TRUE
Allow Software to Delete Mutations	FALSE
Forward and Reverse Balance	FALSE
Forward and Reverse Balance Value <= 0.050	
Load Assembled Result Files	FALSE
Load Sage Data	FALSE
Delete Small Homopolymer Indels if F/R <= 0.25	FALSE
Load Paired Reads	FALSE
Save Matched Reads	TRUE
Highlight Anchor Sequence	FALSE
Ambiguous Gain/Loss	FALSE
Detect Structure Variations	FALSE
Check Mismatch Base Number	TRUE 50 Bases
Check Mismatch Ratio	TRUE 0.300 Length

Table 17 Whole Genome Alignment Settings

Parameter	**Setting**		
Reference Sequence	Human v37_3_dbSNP_135		
Condensation	FALSE		
Assembly	FALSE		
Alignment	TRUE		
Reads: Allowable Mismatched Bases	2		
Reads: Allowable Ambiguous Alignments	500		
Seeds: Bases	80		
Seeds: Move step	5		
Seeds: Allowable Alignments	500		
Overall: Matching base percentage	80		
Detect Large Indels	TRUE		
Sample Trim	FALSE		
Mutation Filter: Except for homozygous	TRUE		
Mutation Filter: Mutation Percentage	**SNPs**	**Indels**	**Homopolymer Indels**
	20	20	20
Mutation Filter: SNP Allele Count	3	3	3
Mutation Filter: Total Coverage Count	5	5	5
Mutation Filter: Balance Ratios and Frequency	FALSE	BR = 0.1 F = 80	BR = 0.8 F = 80
File Type: Paired Reads	TRUE Size range 0 bases to 600 bases		
Saved Matched Reads	FALSE		
Highlight Anchor Sequence	FALSE		
Ambiguous Gain/Loss	FALSE		

Alignment Output

Visualisation of alignment in NextGene Viewer

The NextGene Viewer provides read-by-read visualisation of the data aligned to the reference sequence.

The NextGene Viewer

Figure 20: The NextGene Viewer. (A) Shows reference position. (B) Shows a rainfall chart of alignment across a scalable region, where grey bars indicate the level of coverage. Grey, purple and blue points on the chart show positions where the sample sequence differs from the reference (grey: dismissed as misalignment, purple: accepted known variant, blue: accepted novel variant). (C-G) show an enlarged region of the reference sequence above, corresponding to the position of the blue + in (B). (C) Shows position on the chromosome, (D) Indicates if the region is translated, (E) shows the same mutation calls as seen in (B) for the enlarged region, (F) shows the reference sequence, and below it the consensus sequence from the reads that align to the position (G) shows the pileup of reads aligning to the reference sequence. A known/accepted SNP (heterozygous G>A) is highlighted in purple in the window. Several bases of mismatch that have not been accepted as a genuine deletion in one read are highlighted in grey.

Expression Report

The expression report contains data related to the number of sequences aligning to the reference sequence. The report takes a .BED file as input. A .BED file containing the locations of all the baits in the bait capture library was used. This had the advantage of excluding repetitive regions near our design where varying amounts of off-target sequence in different samples could skew data interpretation. Data collected in the expression report is outlined below in Table 18.

Table 18 Data included in Expression Report

Data	Description
Chromosome	1-23, X/Y
Chromosome Position Start	Start of region covered by bait
Chromosome Position End	End of region covered by bait
CDS	Gene, if region contains sequence that is transcribed
Average Read Count	Average of coverage at each base position within the region
Read Count	Total number of reads where the central base of the read is between 'Chromosome Position Start' and 'Chromosome Position End'
RPKM	'Reads per kilobase exon model per million mapped reads' – a normalised measure of coverage calculated based on the size of the reference sequence, the average coverage of the position in question, and the coverage of the sequence surrounding it for 1 kilobase

Rejected reads

Reads are rejected from the alignment if they don't fit the alignment criteria in Table 17. Rejected reads are stored in separate FASTA files according to the criteria on which they were rejected from the alignment. Rejected read files and the criteria under which reads are assigned to them are listed below in Table 19.

Table 19 List of files produced to which rejected reads are assigned

File	Basis of rejection from alignment
Unmatched	Not enough bases in read match reference sequence
Opposite Direction Paired Reads	Both reads in pair match the reference in the correct orientation from one another (i.e. they are read from 'opposite directions' and match the sequence as expected) but do so outside the expected gap distance (e.g. greater than 600bp apart)
Same Direction Paired Reads	Same direction reads both match the reference sequence, but in the same orientation as one another. Normally, one read needs to be 'flipped' to match the reference because it aligns in the opposite direction
Single Reads	Only one read from the pair matches the reference sequence

Variant Report

Variants in the aligned sequences are filtered and scored by NextGene. Any variants that pass the filtering criteria defined in Alignment Settings (Table 17) are included in the variant report. The data included in the Variant Report is listed below in Table 20.

Table 20 Data included in Variant Report

Data	Description
Chromosome	1-23, X/Y
Chromosome Position Start	Start position of variant
Chromosome Position End	End position of variant
CDS	Gene, if variant occurs within its coding region
Coverage	Number of reads that cover position of variant
Score	Phred score assigned to mutation
Reference Nucleotide	Expected sequence at position of variant
SNP db_xref	dbSNP reference number, if variant has been reported previously
Genotype	Genotype of individual at position of variant
Mutation	Nature of mutation (insertion, substitution, deletion)
Mutant allele frequency	Percentage of reads covering position that show variant rather than reference sequence
A% Ratio	Percentage of reads at position showing ‘A’ at position of variant
G% Ratio	Percentage of reads at position showing ‘G’ at position of variant
C% Ratio	Percentage of reads at position showing ‘C’ at position of variant
T% Ratio	Percentage of reads at position showing ‘T’ at position of variant
Amino Acid Change	If variant occurs within CDS, expected amino acid change at position brought about by variant

In-Browser Bioinformatics tools

BLAST

The National Centre for Biotechnology Information (NCBI) Basic Local Alignment Search Tool (Nucleotide) was used to search for sequence motifs <20bp in the human genome.

http://blast.ncbi.nlm.nih.gov/Blast.cgi

University of California Santa Cruz Genome Bioinformatics Site
The UCSC Genome Browser

The UCSC Genome Browser was used obtain information about sequences of interest, such as the genomic co-ordinates of genes, the sequences of repeat elements and locations of SNPs. FASTA files for regions of interest were downloaded from this site. Custom information (such as the positions of baits in a capture library design) were compared against sequence features by uploading them as a custom track (BED file).

https://genome.ucsc.edu/cgi-bin/hgGateway?redirect=manual&source=genome.ucsc.edu

BLAT

The UCSC BLAT Tool was used to query sequences >20bp in length and map them to the genome in the UCSC Genome Browser.

https://genome.ucsc.edu/cgi-bin/hgBlat?hgsid=445223818_NkJ0R0XwzfRU4JRvb4aAe5EglBSQ&command=start

In-Silico PCR

The UCSC In-Silico PCR tool was used to check primers for correct location, product length, Tm and specificity/uniqueness.

https://genome.ucsc.edu/cgi-bin/hgPcr?hgsid=445223818_NkJ0R0XwzfRU4JRvb4aAe5EglBSQ

EMBL-EMI Pairwise Sequence Alignment

Pairwise sequence alignment was used to compare multiple FASTA sequence files and find regions of homology within them.

http://www.ebi.ac.uk/Tools/psa/

Primer3Plus

Gap PCR primers were used to confirm the breakpoints of variants found by NGS data analysis. A FASTA sequence of the expected break point region was uploaded to primer3plus as the target region. FASTA sequence was obtained from the UCSC Genome Bioinformatics Site. Repetitive elements were masked to lower case. The FASTA sequence typically included 500 bp of sequence upstream and downstream from the breakpoint. If the regions were highly repetitive a larger target sequence was used to ensure the primers produced a unique product. The breakpoint position was highlighted as the target region. The Tm range for the primers was set to 57-62°C. Primer pairs with similar melting temperatures, with at least 3 gua9 nucleotides in the last 5 5' bases were selected preferentially. Selected primers were evaluated in the UCSC in-silico PCR tool.

http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi/

Break Point Confirmation:

Gap-PCR

Gap-PCR was used to confirm the breakpoints of variants identified using NGS. Gap PCR primers were designed in Primer3Plus. If the primers targeted a rearrangement which meant they would not product a product in DNA samples from control subjects, additional 'control' primer pairs were designed. These would create 2 PCR products in a normal control, each using one of the 'test' primers for use with the test DNA sample in which the variant occurred. This allowed the primers to be tested and their reaction conditions optimized. Upon receipt 10 mM working dilutions were made of all primers using nuclease free water and the stocks and dilutions were stored at -20°C.

The optimal reaction conditions for the primers were determined. The initial optimization PCR amplification experiment reaction conditions are as follows: Optimal Tm as buffer is not optimised

A PCR mastermix was made for 4 PCR conditions, using two control samples plus no template controls, for each primer set (10 reactions). For each reaction: 10 µL Qiagen Multiplex PCR MasterMix (Qiagen, Germany), 1 µL 10mM forward/ reverse PCR primer, 1-2 µL template DNA (30-200ng/ µL) and nuclease free water to make a total volume of 20 µL.

The reagents and DNA were mixed briefly on a vortex mixer and centrifuged to bring all liquid to the bottom of the tubes or plate. The tubes or plate were then placed in a

Verity Veriflex thermal cycler (Thermo, USA). The thermal cycling program used is outlined in Table 21.

Table 21 Thermal cycling conditions for optimizing Gap PCR conditions. A different annealing temperature is used for each of the 4 reactions prepared for the samples. The temperatures used are the primer Tm -4°C, primer Tm +2°C, primer Tm, primer Tm +2°C, primer Tm +4°C.

Step	1	2 (35x)			3	4
Temperature	95°C	95°C	[Veriflex Step]	72°C	72°C	4°C
Time	10 minutes	30 seconds	30 seconds	1 minute (or 1 minute per Kb of PCR product)	10 minutes	Hold

The PCR products were electrophoresed in a 1.5% agarose gel. The optimal annealing temperature was selected and used for amplification of the novel PCR product in the test sample.

PCR products >1Kb did not amplify well under these conditions. These products were amplified using the following reagent mix (for one reaction): 5 µL LongAmp Taq Reaction Buffer (NEB, USA), 1 µL 10mM forward/ reverse primer, , 0.75 µL dNTPs, 1-2 µL 30-200ng/ µL template DNA, nuclease-free water to make a total volume of 25 µL. Reactions using this kit were amplified in a Verity Veriflex thermal cycler using the program outlined in Table 22. This process may also require optimisation to successfully amplify a product, through trial of multiple annealing temperatures and extension times

Table 22 PCR reaction conditions for LongAmp Taq Polymerase reaction kit

Step	1	2 (35x)			3	4
Temperature	95°C	95°C	64°C	65°C	65°C	4°C
Time	10 minutes	30 seconds	30 seconds	10 minutes	10 minutes	Hold

Dye Terminator Sequencing

Dye-terminator sequencing was used to confirm the precise breakpoints of variants found by NGS after successful amplification of a Gap-PCR product covering the break point region.

AMPure Purification

The Gap PCR product was purified using AMPure beads on the BioMek NX robot platform: 80 µL AMPure XP beads were added to the reaction and mixed by pipetting up and down. The mixture was left to incubate for 5 minutes at room temperature. The mixture was then transferred to a magnetic rack. After 4 minutes, the supernatant was aspirated and discarded. 150 µL 85% Ethanol was added to the tube, left for one minute and then aspirated and discarded. This step was repeated once, for a total of two ethanol washes. The pellet was left to dry on the magnet for 5 minutes at room temperature. The samples were then removed from the magnetic rack and eluted in 35 µL nuclease-free water. The mixture was left to incubate for 3 minutes at room temperature and then transferred to a magnetic rack. The samples were left to incubate on the rack for 5 minutes at room temperature. 30 µL of the supernatant was then removed and transferred to a fresh PCR plate.

Dye-terminator PCR reaction

Purified PCR products were prepared for dye-terminator sequencing in triplicate, with separate reactions for forward and reverse sequencing of the product. The following reagent mix was prepared per reaction: 0.5 µL Dye-Terminator Ready Mix, 0.5 µL Dye-Terminator Buffer, 4 µL DNA, 4 µL nuclease free water, 1 µL 5pM forward or reverse PCR primer, total volume 10 µL.

The reactants were mixed by vortexing and centrifuged briefly to bring all liquid to the bottoms of the tubes. The reactants were placed in a Verity thermal cycler and the program outline in Table 23 was initiated.

Table 23 Thermal cycling program for preparation for dye-terminator sequencing

Step	1 (40x)			2
Temperature	95°C	52°C	60 °C	10 °C
Time	30 seconds	15 seconds	2 minutes	10 minutes

CleanSeq Purification

The PCR product was returned to the BioMek NX for purification with CleanSeq beads: 30 µL CleanSeq beads were added to the reaction and mixed by pipetting up and down. The mixture was left to incubate for five minutes at room temperature. The mixture was then transferred to a magnetic rack. After 5 minutes, the supernatant was aspirated and discarded. To wash the bead pellet, 150 µL 85% EtOH was added to the tube, left for one minute and then aspirated and discarded. This step was repeated once, for a total of two ethanol washes. The pellet was left to dry on the magnet for five

minutes at room temperature. The samples were then removed from the magnetic rack and eluted in 35 µL nuclease-free water. The mixture was left to incubate for three minutes at room temperature and then transferred to a magnetic rack. The samples were left to incubate on the rack for five minutes at room temperature. Ten µL of the supernatant was then removed and transferred to a fresh MicroAmp Optical 3130 plate (Life Technologies).

Dye-Terminator Sequencing

The MicroAmp Optical 3130 plate was fitted with a rubber septa and placed in the ABI Genetic Analyser 3130. Dye-terminator sequencing was performed using standard laboratory parameters.

Dye Terminator sequence data was analysed using Mutation Surveyor software (Biogene USA) or Sequencher Software (Gene Codes, USA). In both cases dye terminator sequences were aligned to reference sequences. When using Sequencher, the reference sequence for the region that was amplified was downloaded from the UCSC Table Browser application (Hg.19). Mutation Surveyor was more aligned with diagnostic applications while Sequencher was useful for building contigs and identifying breakpoint sequences.

Automation of Sample Preparation Process on BioMek FX^{P}

During the course of the project the BioMek FX^{P} liquid handling robot was purchased with the aim of automating the library preparation and target enrichment steps of the sequencing process. This system was selected over the Agilent Bravo system as it was a more flexible platform and was not tied in to a single companies chemistry. It also had the potential to automate the entire process while the Agilent Bravo required some manual intervention steps. The cost of both instruments was similar but the Bravo would work almost out of the box while the methods needed to be developed on the BioMek. Developing an automated method sounded relatively straight forward but with limited experience of automation this proved challenging.

After the robot install and initialisation a programme, developed by Beckman, was installed on the robot. This was modified on site by a field application specialist to accommodate the deck layout and specific laboratory requirements. The process of establishing a method was estimated to take 3 months. In reality this took almost 9 months with three different application specialists trying three different programmes. Each programme was divided into library preparation, pre hybridisation and post hybridisation. Library preparation was relatively easy to automate and was working

within 2 months. The target enrichment process which included 24 hour incubation at 65 °C and a complex arrangement of washes was much more difficult to automate. After sequencing the libraries the off target sequence was very high indicating that little or no target enrichment had taken place. Examining this process, taking it apart and reprogramming the robot took time and various strategies were taken to best reproduce the manual steps, which used volumes exceeding those possible on the robot.

Due to the time lost on the automation of the library prep and the focus of my PhD studies routine diagnostic staff took on the responsibility of developing the automated solution with Beckman. Once it became available it was towards the end of my research and therefore limited samples were available to prove consistency of processing.

Results Chapter One: Preliminary Investigation

The initial sequencing experiments in this study were performed before on-site facilities for sample preparation and sequencing had been established. Sample preparation was performed at the Genomics Facility at Guy's and St Thomas's Hospital, London. The facility has previous experience in preparing samples for NGS and were able to provide advice and all necessary equipment for the procedure.

Samples were then sequenced on the HiSeq 2000 platform at the Rayne Institute, London, by Dr A. Smith from the Department of Haematological medicine who was the operator of this platform.

Data obtained from Dr Smith in FASTQ format was then analysed using the NextGene (SoftGenetics) package that was acquired by the KCL Department of Molecular Pathology for this purpose. Multiple approaches to analysis were investigated in order to determine the best way to use this software for the targets of this investigation.

This initial experiment investigated the ability of this combination of sample preparation, sequencing and analysis to characterise the variants and rearrangements present in a range of DNA samples, including known and novel variants. The chapter describes and evaluates the strengths and weaknesses associated with each step of this process. This information was used to modify these processes in subsequent chapters to optimize the process for diagnostic use. The novel variant cases came from a collection amassed by Prof Swee Lay Thein over a period of 30 years. Throughout this period, when new sequencing technologies became available Professor Thein and researchers in her laboratory attempted to use these techniques to resolve the breakpoints of these novel variants. The cases examined in this investigation were not resolvable using MLPA, CGH array, Southern blotting or qPCR.

As this investigation went on, more resources became available for use in preparing the DNA samples and analysing the data. When the project started, both sequencing and sample preparation had to occur off-site, due to the lack of suitable equipment on site, and lack of experience with the techniques used. Data analysis was performed using an off-the-shelf software package with which no one on site was familiar, and determining the best way to use this software to interpret our data was a significant undertaking. Figure 21 shows a timeline of the resources used over the course of the sequencing experiments performed as a part of this investigation.

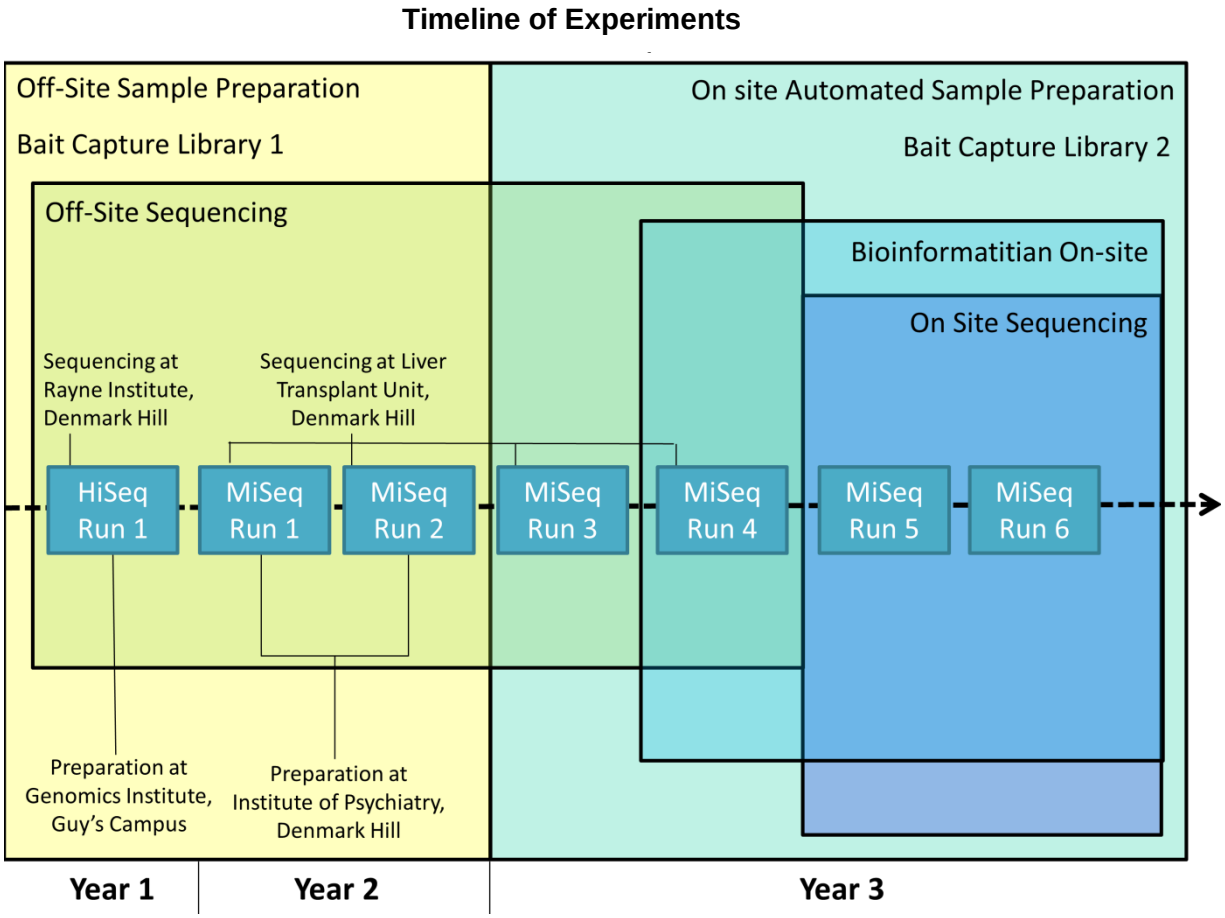


Figure 21: Timeline of Experiments. Different locations for sample preparation and sequencing, different platforms, different capture libraries and manual and automated sample preparation protocols were used in various combinations as they became available. The preliminary investigation involved HiSeq Run 1 and MiSeq Run 1 and 2 (Years 2 and 3).

HiSeq Sample Preparation: Run 1

Sample Details

HiSeq Run 1 sequenced 12 DNA samples obtained from the diagnostic laboratory. The batch included negative control samples, positive control samples, and test samples from patients with unknown, uncharacterised structural rearrangements (Table 24). Southern Blots or MLPA had been performed by the diagnostic laboratory and gave a rough idea of the size and type of the unknown rearrangements, and confirmed that no structural rearrangements existed in the negative controls. To increase negative control data available, samples known not to have variants affecting one of the globin gene clusters were used as negative controls for the region of their unaffected cluster. Therefore, alpha thalassaemia samples could be used as negative controls for beta globin gene analysis, and vice versa. HiSeq Sample 12 (a duplicate of HiSeq Sample 11) was used as an additional control for the alpha globin region, but not analysed further for the beta globin locus.

Table 24: Sample Details for HiSeq Run.

Sample	Beta Globin Locus	Alpha Globin Locus
HiSeq Sample 1	Negative Control	Test (Duplication)
HiSeq Sample 2	Negative Control	Negative Control
HiSeq Sample 3	Negative Control	Test (Duplication)
HiSeq Sample 4	Negative Control	Negative Control
HiSeq Sample 5	Test (Deletion)	Negative Control
HiSeq Sample 6	Positive Control ($-^{\text{Asian Indian Indel}}/$)	Negative Control
HiSeq Sample 7	Test (Duplication)	Positive Control ($\alpha-^{3.7}/$)
HiSeq Sample 8	Negative Control	Test (Duplication)
HiSeq Sample 9	Negative Control	Test (Deletion)
HiSeq Sample 10	Test (Deletion)	Negative Control
HiSeq Sample 11	Test (Deletion)	Negative Control
HiSeq Sample 12	N/A	Negative Control

Sample Preparation

The samples were prepared for sequencing in two batches of six for ease of handling. Sample preparation took place in the Genomics Centre at Guys Campus, under the guidance of the experienced staff at this facility. Sample preparation was in accordance with the protocol as outlined in Methods. The samples were run on the Tapestation prior to sequencing using the D1000 DNA kit, to check concentration and fragment size distribution (Figure 22). The size of fragments in each sample ranged from 200 bp – 500 bp, all with a mean peak at approximately 350 bp. This is the expected product size range with an input fragment size of ~150 bp, after the addition of adapters and index tags (200 bp). The size distribution of the products for all 12 samples was extremely uniform.

Tapestation Data Showing Results of Library Preparation for HiSeq Samples 1-12

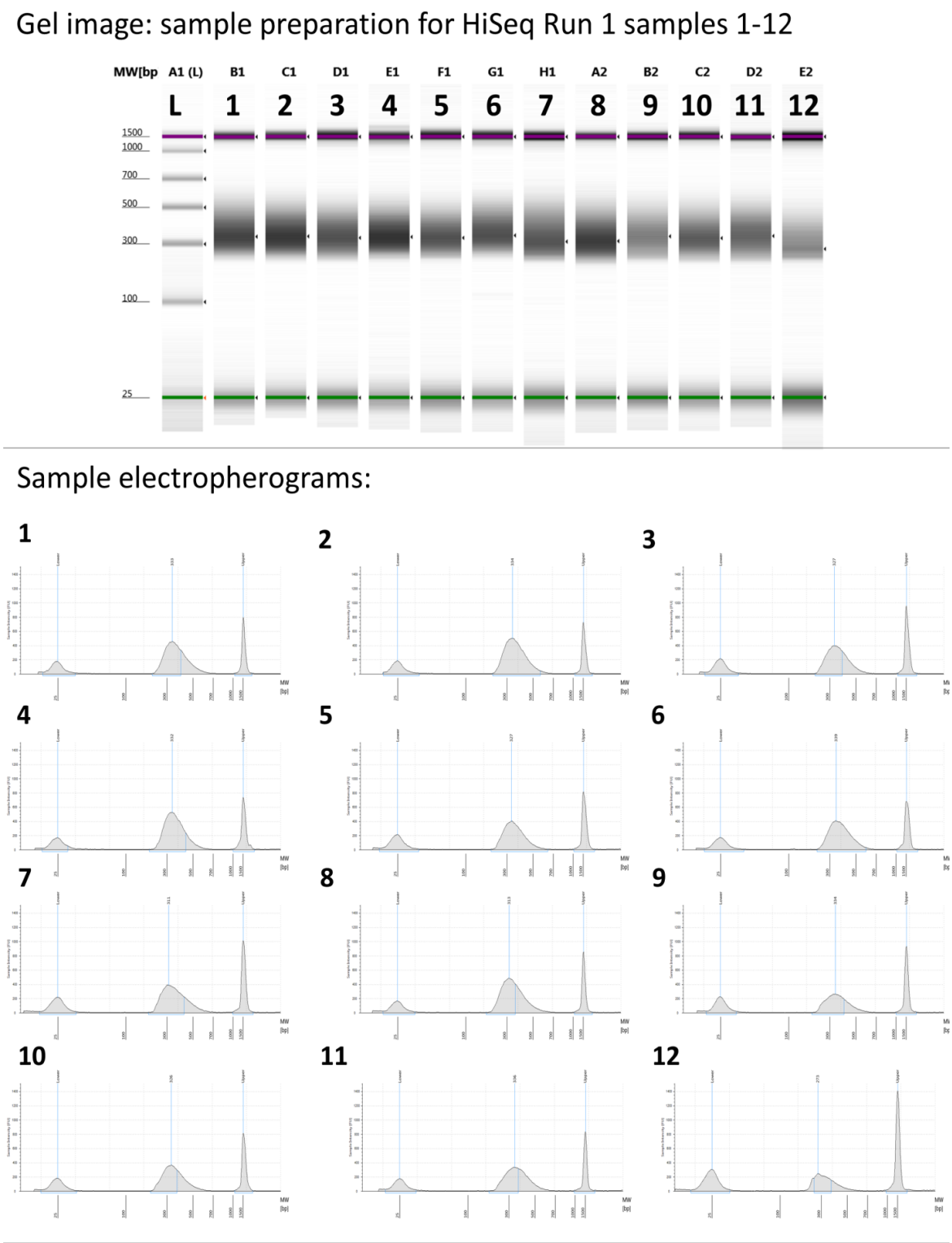


Figure 22: Tapestation Data Showing Results of Library Preparation for HiSeq Sample 1-12. (S1-12, lanes 1 to 12). Top panel shows each trace as a band, next to a DNA ladder (lane L). Lower panel shows individual traces for each sample. Sample yield is broadly similar (1000-2000 pg/µl) in all samples, as is size range (150-500 bp size range, mean size approximately 350 bp).

Sequencing Results

Samples were sequenced on the HiSeq 2000 in a batch of 12. The HiSeq platform had been recommended over other available platforms (MiSeq and Genome Analyser) for its high sequencing capacity. The samples were sequenced off-site at the Department of Haematological Medicine, Rayne Institute, King's College London and run quality data, including cluster density, was not provided with the raw data. The HiSeq 2000 platform generates two 97 bp reads of each DNA fragment ligated to the flow cell: 'Read 1' (R1) and 'Read 2' (R2) along with addition to a 6bp index read (Figure 19). The sequencing data is in FASTQ format. Multiple FASTQ files were generated for each sample. These were condensed into a pair of files for each sample, the first containing 'Read 1' (R1) of each DNA fragment that was sequenced, and the second file containing 'Read 2' (R2) of each fragment. Read 3 data is not employed for any further analysis after it has been used to separate the R1 and R2 data into individual sample files. The R1 and R2 sequences from each DNA fragment are given unique identifying names that relate them to one another using the format "*XXXX:XXX:XXX_1"/"XXXX:XXX:XXX_2*".

FASTQ (read names, base calls, base quality scores) format files were converted into FASTA (read names, base calls) format. The conversion program also removes any reads that do not meet a series of user-defined quality score thresholds based on the base quality and score information. Remaining bases called as 'N' in reads that pass quality filtering are later ignored during alignment, which is performed based on the sequences of the bases in the read that were called successfully. Standard parameters (see Methods) were used for format conversion (please refer to Methods for more detailed descriptions of these metrics).The results of format conversion for HiSeq Run 1 are shown in Table 25 as averages of the entire sample cohort. The majority of reads in all samples passed the standard quality filters. Of those that fail, the majority are rejected based on their 'median score', i.e. the median quality score of each base position in the read. A minority of reads are filtered out by "uncalled bases". Some reads in which bases have a low quality score towards the end of the read (a common issue with Illumina sequencing platforms) can be 'trimmed' so that the remainder of the read that does pass quality filtering can be included in analysis.

Table 25: Success of Format Conversion. Average for 12 samples.

Average values for all 12 samples	R1		R2	
	Average	Standard Deviation	Average	Standard Deviation
Total Reads in the Input File	15392225	6770274	19250149	11371742
Reads Converted Successfully	14223852(92.4%)	6222500(0.27%)	18087229(93.95%)	10707225(0.22%)
Reads Failed to Convert	1168373	548447	1162921	665870.9
Reads Filtered by "Median Score"	1146658 (98.14%)	538818.9(0.06%)	1122329(96.51%)	642301.9(0.12%)
Reads Filtered by "Uncalled Bases" (too many bases called as 'N')	3815.583 (0.32%)	1657.816 (0.01%)	1824.75 (0.15%)	1103.059 (0.01%)
Reads Filtered by "Called Base Number in Read" (not enough bases are *not* called as 'N')	0	0	94 (0.008%)	106.4522(0.004%)
Reads Filtered After Trimming	17899.67	7982.387	38673	22434.98
Reads Trimmed	2286045	1018293	3261744	1874167
Reads Trimmed by "Quality Score"	2286045	1018293	3261744	1874167
Trimmed Bases	53106560	23823152	61581995	35274987
Trimmed Bases by "Quality Score"	53106560	23823152	61581995	35274987

The average quality score for each base position in each read (from base 1-96) is calculated per sample: a score >30 is classed as 'good'; a score >20 and <30 is classed as 'acceptable'; a score <20 is classed as 'poor'. This data is shown in Figure

23. The majority of positions in both R1 and R2 reads obtained a ‘good’ score. In both reads, quality scores deteriorate for the last 10 called positions. This drop in quality towards the end of the read was seen in all runs performed on Illumina instruments.

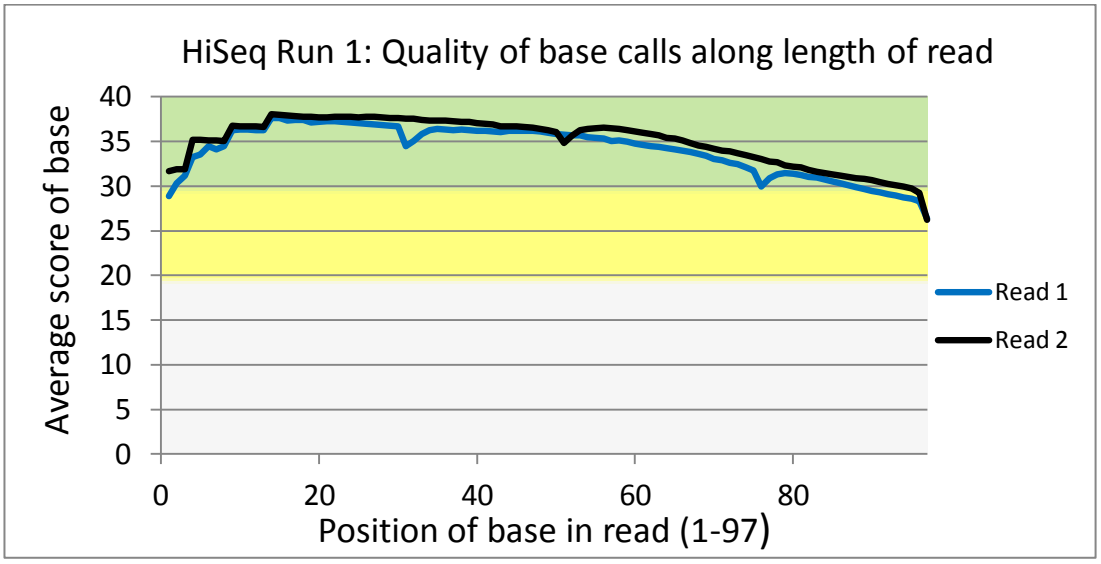


Figure 23: HiSeq Run 1: Quality of Base Calls Along Length of Read. A score >30 is deemed ‘good’; <30 and >20 is deemed ‘acceptable’ and <20 is deemed ‘poor’.

Alignment of FASTA Data to the Reference Sequence

Read data was aligned to a reference sequence of the human genome. Hg.19 v.37.3, dbSNP v135. NextGene attempts to align each FASTA format read to the position in the reference that has the closest match to its sequence. If the read can be aligned to multiple positions in the reference with equal accuracy, ‘ambiguous alignment’ is allowed which assigns the read to all the positions at which it is an equally good match. Reads that do not perfectly match the reference sequence are aligned to the position that they best fit. The mismatch in the read sequence compared to the reference is noted. If it fits certain parameters (such as the same mismatch being present in a large proportion of the reads. - see later section for more details), it will be recorded as a variant. If a mismatch does not fit these parameters, it is considered to be a spurious result and while the read may not be rejected from the alignment, the mismatch will not be recorded as a variant. Reads in which a certain percentage of bases do not match the reference sequence are rejected from the alignment entirely. The alignment parameters for all these conditions are user defined, so alignment stringency can be tailored to the needs of an individual project.

Opting for a ‘paired read alignment’ instructs the software to match R1 and R2 sequences from each fragment in pairs. The two reads in each pair must align in opposite orientation to one another on the reference sequence. Paired reads that do

not align in opposite orientations are rejected from the alignment (see later section: Same Direction Reads). The paired reads must also align at a set distance from one another, based on the size of the DNA fragments that were sequenced. For example, according to Figure 22 the fragment size range for this run is 100-500 bp, therefore the two reads of each fragment are expected to align within this distance of one another along the reference sequence). Reads that align to the reference in the correct (opposite) orientation, but exceed the acceptable distance are also rejected from the alignment (see later section: Opposite Direction Reads).

Optimizing Alignment Settings

The FASTA data from the samples in HiSeq Run 1 was subjected to multiple alignments with different stringencies to determine the best alignment parameters for this study.

Initially, the data obtained from the Rayne Institute was not formatted in the manner expected for paired read data (i.e. one file for all R1 data and another file for all R2 data) and so paired read alignment was not successful. Instead, each read was aligned individually to the reference sequence. This increased the number of ambiguous and spurious alignments for each sample, as only half of the data that was available to aid accurate alignment was used. This also meant reads could not be rejected from the alignment based on read or gap distance requirements. Once this issue was identified, the data was re-ordered using the ‘merge files’ and ‘arrange paired reads’ tools in the NextGene package. Analysis was then restarted using paired read alignment.

Table 26: Alignment Statistics for HiSeq Run 1.

Alignment Statistics	Samples HiSeq Run 1												
	Average	1	2	3	4	5	6	7	8	9	10	11	12
Read alignment statistics													
Perfectly Matched Reads Count	**17821683**	1.18E+07	7.14E+06	1.27E+07	1.21E+07	1.05E+07	1.14E+07	2.46E+07	2.76E+07	2.82E+07	2.28E+07	1.76E+07	2.74E+07
Matched Reads Count	**25545951**	1.66E+07	1.05E+07	1.79E+07	1.69E+07	1.46E+07	1.62E+07	3.48E+07	3.95E+07	3.93E+07	3.31E+07	2.57E+07	4.15E+07
Unmatched Reads Count	**294791**	1.29E+05	1.18E+05	1.23E+05	1.15E+05	3.83E+05	1.48E+05	2.62E+05	5.21E+05	2.93E+05	6.42E+05	2.07E+05	5.95E+05
Short Reads Count	**1052672**	6.67E+05	2.00E+00	7.14E+05	6.68E+05	6.12E+05	6.24E+05	1.32E+06	1.66E+06	1.82E+06	1.42E+06	1.11E+06	2.01E+06
Number of Matched Bases	**2500514169**	1.63E+09	1.01E+09	1.75E+09	1.67E+09	1.42E+09	1.58E+09	3.42E+09	3.90E+09	3.87E+09	3.25E+09	2.50E+09	4.02E+09
Number of reads aligning to target region	**9705129**	6.93E+06	2.71E+06	7.69E+06	7.15E+06	2.71E+06	6.65E+06	1.24E+07	1.65E+07	1.56E+07	1.26E+07	9.97E+06	1.56E+07
% of reads aligning to target region	**53**	59	38	60	59	26	58	50	60	55	55	57	57
Average Read Length	**93**	93	92	93	93	93	93	93	93	93	93	93	92
Average Coverage (Reference Sequence)	**9**	8	6	9	8	6	10	6	12	9	9	10	13
Average Coverage (Target Region)	**363**	376	184	358	371	184	275	492	757	598	515	402	662
Unmatched Bases Recorded as Mutations													
Mismatches	**3327366**	2.10E+06	1.34E+06	2.40E+06	2144764	1.75E+06	2.03E+06	4.63E+06	5.45E+06	4.76E+06	4.23E+06	3.67E+06	5.42E+06
Deletions	**371600**	2.57E+05	1.66E+05	2.71E+05	238018	1.91E+05	2.24E+05	4.65E+05	6.44E+05	5.36E+05	4.59E+05	3.94E+05	6.13E+05
Insertions	**184569**	1.23E+05	7.06E+04	1.34E+05	118994	9.33E+04	1.09E+05	2.42E+05	3.23E+05	2.60E+05	2.26E+05	2.06E+05	3.08E+05
Unmatched Bases NOT Recorded as Mutations													
Mismatches	**9998683**	5.88E+06	4.55E+06	6.23E+06	6016358	5.37E+06	6.08E+06	1.27E+07	1.53E+07	1.41E+07	1.43E+07	1.01E+07	1.94E+07
Deletions	**397897**	2.10E+05	1.78E+05	2.04E+05	208162	2.07E+05	2.36E+05	5.00E+05	6.56E+05	4.66E+05	6.79E+05	3.42E+05	8.88E+05
Insertions	**198774**	1.07E+05	8.81E+04	9.93E+04	104902	1.05E+05	1.18E+05	2.47E+05	3.40E+05	2.28E+05	3.40E+05	1.64E+05	4.44E+05

Several rounds of analysis attempted to align the read data to the specific region of interest targeted by the bait capture library (chr11:3,500,000-7,500,000, chr16:0-260,000). This proved to be unhelpful, as off-target reads that had evaded elimination during the hybridization stage of sample preparation were 'shoehorned' into partial alignments within the region of interest. This created false positive variant calls where a large number of reads from another part of the genome showed the same deviations from the part of the truncated reference sequence that they best aligned to, and were therefore counted as legitimate variants.

Preliminary alignments were performed using the standard alignment parameters described in Methods. Higher and lower stringency alignment options were explored, but were found to be less suitable for the requirements of this study (discussed later).

Statistics from alignment using standard parameters for all samples and the run average are shown in Table 26. An average of 96.54% of reads that passed format conversion were successfully aligned to the reference sequence. Hybridization was found to have achieved only 53% specificity for sequences originating from the target region. The high sequencing capacity of the HiSeq 2000 platform meant that despite this, the targeted region was still sequenced at a high read depth (average coverage 400x).

Alignment Results

Once alignment is complete, NextGene produces several reports that were useful for data analysis in this investigation. Additionally, the NextGene Viewer allows exploration of the exact base sequence of the reads aligning across the reference sequence.

Variant characterisation for diagnostic use follows two different pathways using different outputs from NextGene: one for indel/single nucleotide variant detection and another for structural rearrangement detection (Figure 24). Indel/single nucleotide variant detection is provided by the Mutation Report. Structural rearrangement detection begins with gross identification of dosage changes utilising the Expression Report, followed by breakpoint identification. The NextGene Viewer may provide sufficient data to determine the rearrangement breakpoints, but if this is not possible resolution variously requires the use of the Mutation Report, Opposite Direction Read Report or Same Direction Read Report. A 'Structural Variant' report provided by later releases of the software included an extremely high amount of spurious data from misalignment of off-target reads, and failed to include any of the genuine

rearrangements found (and confirmed via Gap PCR) in these sequencing experiments. This report was disregarded in favour of the strategy described here.

NextGene Variant Detection Workflow

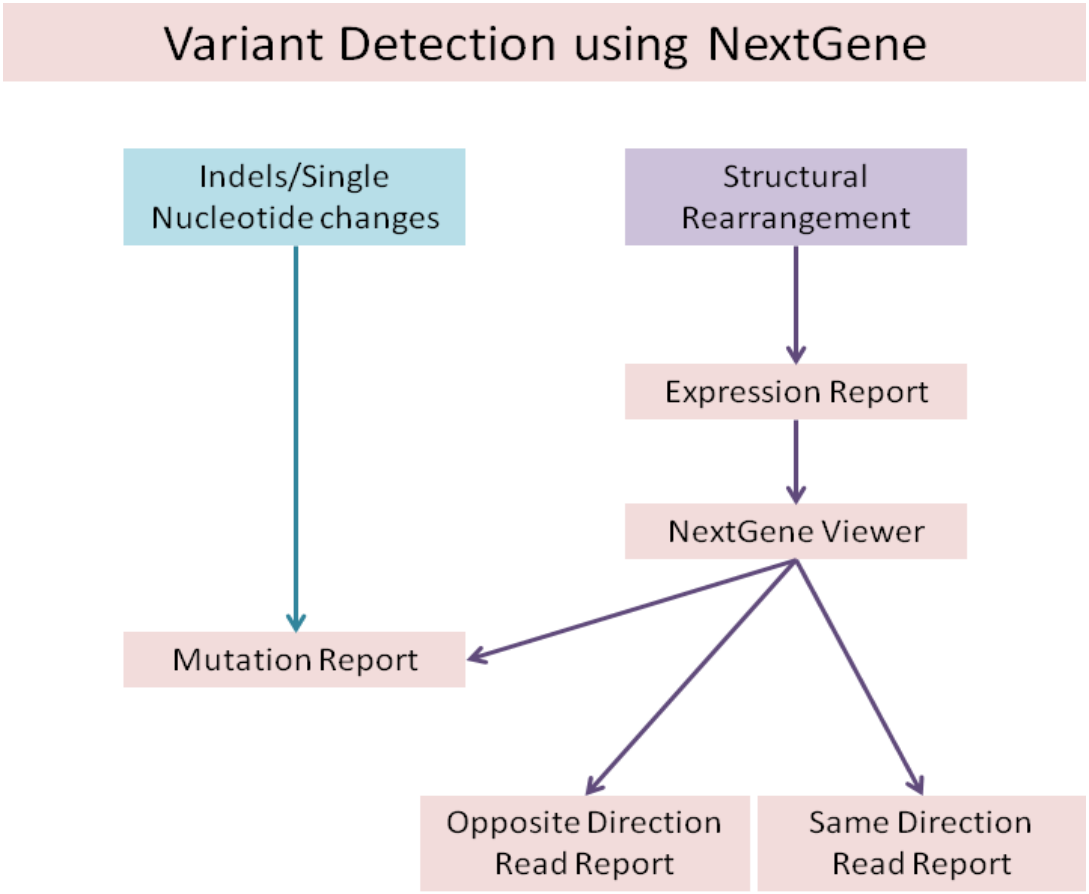


Figure 24: NextGene Variant Detection Workflow. The NextGene outputs required to detect and characterise Indels, single nucleotide changes and structural rearrangements.

Mutation Report (for characterising single nucleotide variants and indels)

A mutation report was generated for each sample. The report contained a list of all positions where a given proportion of reads contained sequences that differed from the reference sequence (to be included, a position must be covered by at least 15 reads, and at least 3 must exhibit the same deviation from the reference sequence). The mutation report includes the chromosome, position, base change, mutant allele frequency, CDS and score. If the variant is already published in the dbSNP database, its identifying number is also listed. The 'score' is a phred score determined by a number of factors including penalties for whether the variant is only recorded in the end of reads, only in forward or only reverse reads, or within homopolymer sequences. Mutation scores are considered to be reliable if they are >20. Scores of >15 should be treated with caution. Lower scores are likely to be due to sequencing or alignment errors.

Haemoglobinopathies such as sickle cell disease – and many cases of beta thalassaemia – are caused by single nucleotide changes, so for diagnostic use it is crucial that these variants can be reliably identified and mapped to the correct globin gene. Two samples in the patient cohort sequenced in HiSeq Run 1 had known single nucleotide variants: a codon 44 deletion in one sample, and rs334 (sickle cell variant) in another. Both these variants were successfully identified in the mutation report (see individual sample analysis).

The ability of NGS sequencing to detect single nucleotide variants has been widely investigated and many publications report that the technique is extremely sensitive to these variants. This will not be explored further in this analysis then confirming the presence of mutations previously confirmed in the samples.

Expression Report

The expression report is designed primarily for use in RNA sequencing, where the relative proportion of different sequences indicates the level to which those genes are expressed. In this study, the report was used to identify DNA dosage changes that increased or decreased the number of reads aligning to the reference sequence. Relative increases in the amount of sequence aligning to a position indicated a duplication had occurred in the sample, while a relative decrease indicated a deletion.

Bait performance varies widely across bait capture libraries, so rather than directly comparing the coverage achieved at different positions, a relative measure, RPKM, was used. An 'RPKM value' was calculated for each bait-covered position. This is a normalised measure of coverage across a region of interest (each 120 bp bait) expressed per million reads aligned to the reference sequence for a single sample. These values were used to compare coverage across the entire bait tiled region and between different samples. The RPKM values of each bait-covered position in test samples were plotted on a graph, where the X-axis showed their position on the chromosome and the Y axis showed the value by which the RPKM value for that bait differed between the sample and the average of the negative controls for that region (Figure 25). These values were plotted on a log2 scale as this provided direct comparison to CGH array data. This scale also proved to be an effective way of quickly identifying dosage changes: A negative value indicates lower coverage in the sample than in the controls. A positive value indicates higher coverage than in the controls. A value of 0 indicates no difference between the coverage of this position in the sample and the controls. To look for variations that could affect the globin genes, the coverage across chromosome 11p was viewed on an enlarged scale, showing positions

5,000,000-5,500,000. This was not necessary to inspect the smaller target region on chromosome 16 from positions 60,000-2,600,000.

Plotting Variation in RPKM Values between Samples and Negative Controls

Bait ID	Position on chromosome	RPKM Negative control sample 1	RPKM Negative control sample 2	RPKM Negative control sample 3	Control average	RPKM Test sample 1	Log2 ratio (Test Sample 1 / Control Average)
Bait 1	2,000,500 – 2,000,620	240	265	200	235.00	220	-0.095
Bait 2	2,000,620 – 2,000,740	400	430	440	423.33	512	0.274
Bait 3	2,000,740 – 2,000,860	230	300	220	250.00	165	-0.599

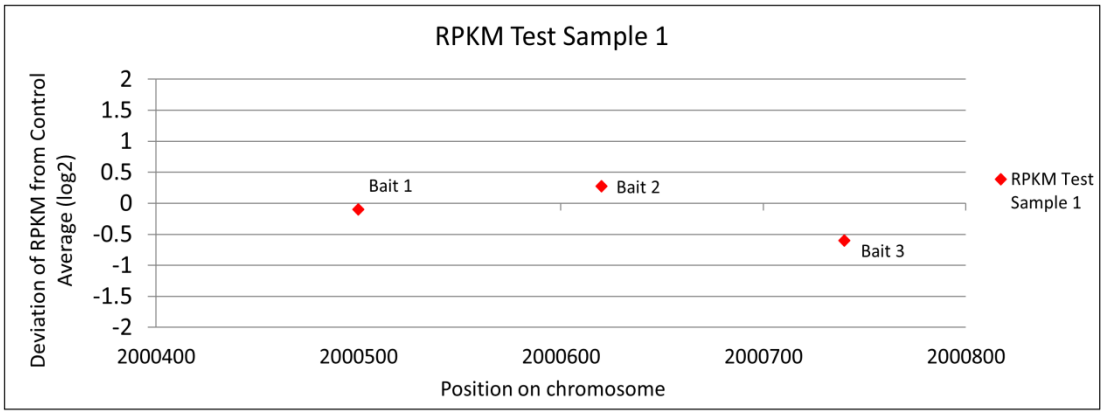


Figure 25: Plotting Variation in RPKM Values between Samples and Negative Controls. RPKM values according to chromosomal position (X axis), and deviation from negative control average (Y axis). RPKM values are used as a normalised measure of the number of reads aligning to each part of the genome covered by a bait in the capture library. The difference between the amounts of sequence obtained for test samples compared to normal controls are plotted on a log2 scale.

The coverage of some regions varied greatly in both the samples and the negative controls. Presence of this variation in the controls indicates that it is not clinically significant, and may be due to CNV events or differences in the hybridization efficiency or specificity of different baits. To identify sample specific coverage differences which are likely to be clinically significant, the coverage graph was overlaid with standard deviation data derived from the mean RPKM from the negative control samples (“NegC StDev”) (see Figure 26). Deletions in test samples were quickly apparent when their coverage data was graphed using this method (graphs shown below). The start and end positions of the deletions as indicated by the RPKM data were then investigated in the NextGene viewer, which shows all the aligned reads, their genomic location and their sequences.

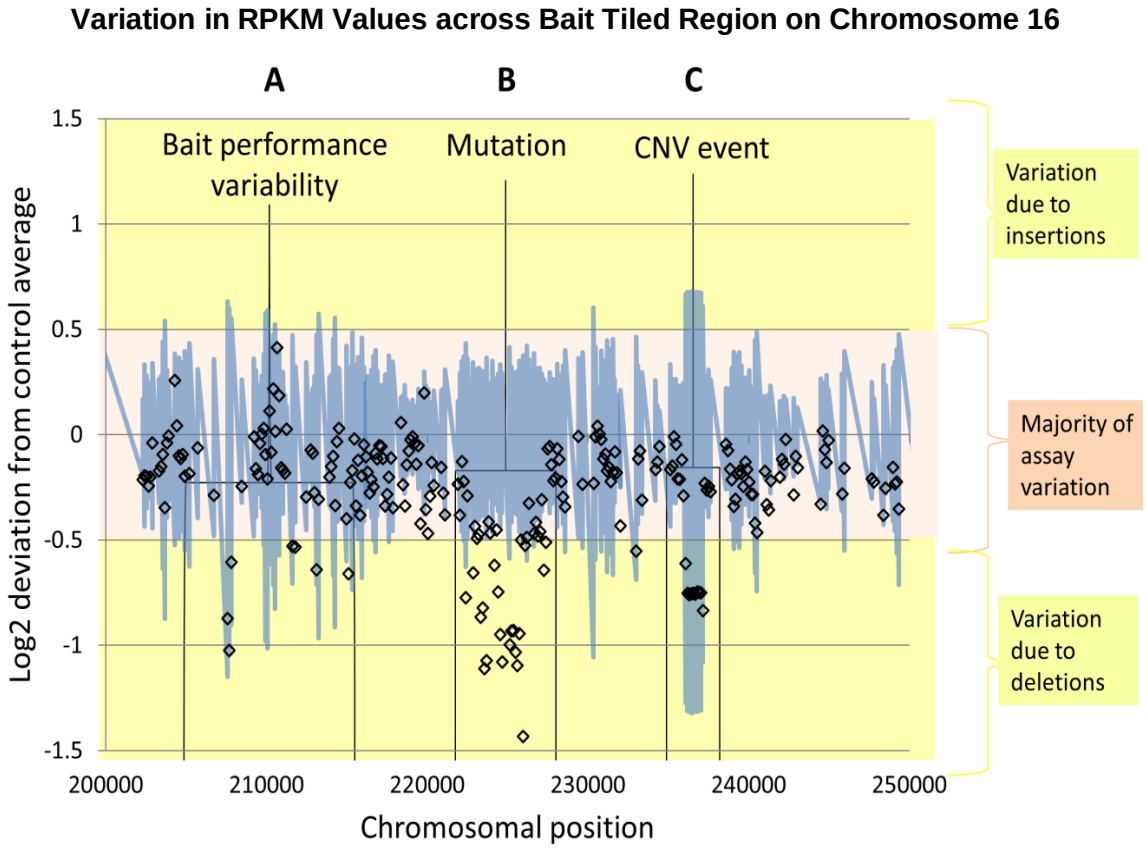


Figure 26: Variation in RPKM Values Across Bait Tiled Region on Chromosome 16. X axis shows position on chromosome 16 and Y axis shows deviation from the negative control RPKM average on a log 2 scale. Black diamonds indicate individual bait values, plotted according to their position and on the chromosome, and the deviation from the negative control average at that position. Increases in RPKM values relative to the negative control average can indicate duplications, while decreases can indicate deletions. The standard deviation in negative control samples from their average is shown by a grey line. The negative control standard deviation values can be used to distinguish random variation in common in the assay from genuine variants.

(A) shows a region where the sample RPKM values decrease relative to the negative control average. However, this variation is also seen in the standard deviation from the negative controls, therefore this can be dismissed as random variation in the assay.

(B) Shows a sustained region where RPKM values for the baits in the sample are lower than the negative control average. This indicates a deletion of this region the sample that is not present in the negative controls.

(C) Shows a sustained region where the RPKM values of the baits are lower than the negative control average. This implies a deletion of this region, but the same region also appears to be commonly deleted in the controls, which is therefore likely to be unrelated to the sample phenotype. Figure taken and adapted from(Shooter 2014).

NextGene Viewer

Once the broad region affected by a structural rearrangement has been identified using RPKM data, the NextGene Viewer is used to locate the precise breakpoints. Visually inspecting the read pileup at the approximate start and end positions of the rearrangement can reveal reads that have crossed the breakpoint position. These reads are identified by strings of misaligned bases, all originating from the same position (indicating the first breakpoint of the rearrangement) and showing the same

alternate sequence to the reference. In order to be accepted into the alignment, these reads may only contain a few bases of breakpoint sequence, as otherwise they are trimmed or rejected from the alignment. In this case, the entire read sequence can be retrieved from the original FASTA files by looking up the read name. Once the whole R1 and R2 sequences are obtained, a BLAT query of their sequences can be used to reveal the deletion breakpoints (Figure 27). BLAT (Blast-like alignment tool) is a pairwise sequence alignment algorithm hosted by the UCSC genome bioinformatics site which finds matches in the human genome to FASTA format sequences. It provides quality scores based on the quality of each match it returns. Misaligned reads at breakpoint positions are often not included in the mutation report. This is because many reads that cover the position may be rejected on the basis of (i) the amount of sequence within the read that misaligns; (ii) the position to which the second read aligns (see opposite/same direction reads). The alignment process will also attempt to 'shoehorn' the read sequence into the best possible alignment to the reference sequence, attempting to match as many bases in the misaligned sequence as possible to the reference. If any part of the misalignment is accepted into the mutation report, it does so as a series of disjointed small variants, rather than as one large variant.

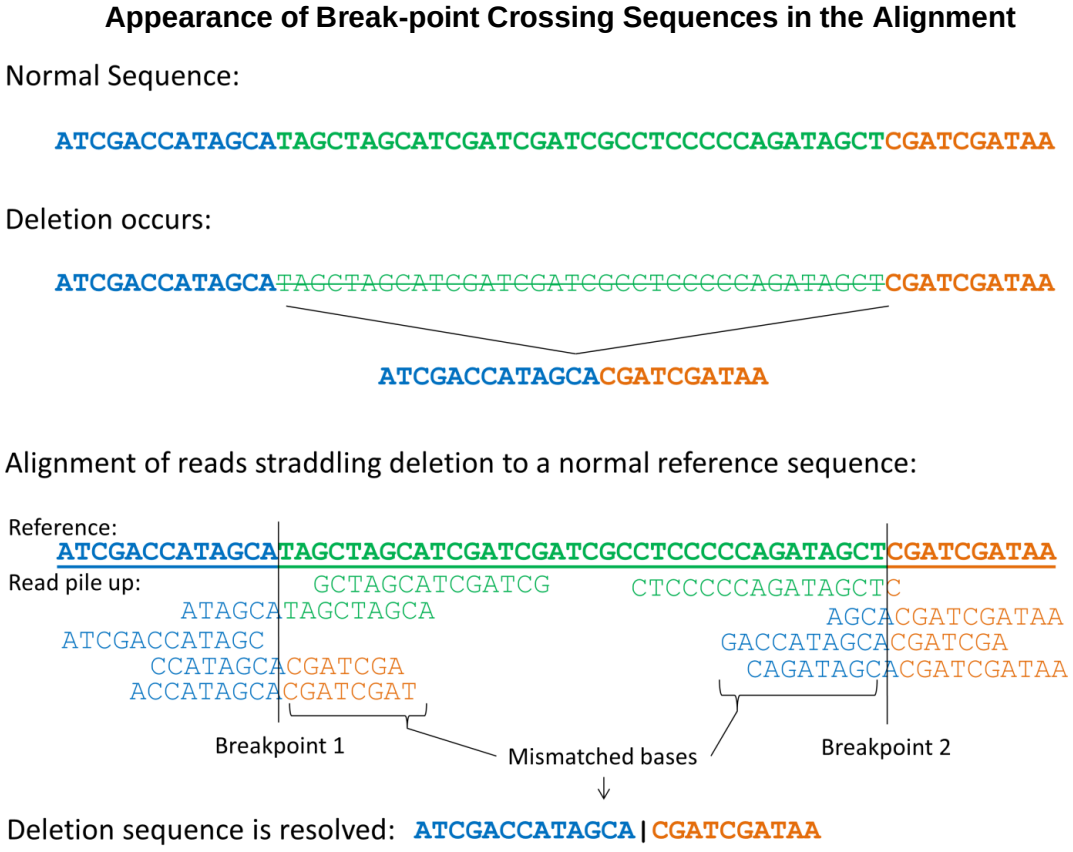


Figure 27: Appearance of Break-point Crossing Sequences in the Alignment. Diagram shows a hypothetical deletion (upper panel), and how reads covering the deletion break point are identified in the alignment (lower panel). Reads crossing the deletion breakpoint lack the green coloured

bases, so in the alignment strings of unmatched bases will appear either side of the region that is deleted in the sample. NB: If a read contains too many bases that do not correspond to the reference sequence, it may be rejected from the original alignment.

Mutation Report (for characterising structural variants)

The mutation report includes mutant allele frequencies for each listed variant. If there is doubt about the presence or start/end points of a rearrangement based on the RPKM data, this can provide additional evidence of whether a rearrangement has occurred, and what region it covers. In balanced sequences, mutant allele frequencies should either be 1:1 (heterozygous) or 0:1 (homozygous). A long stretch of 0:1 mutant allele frequencies strengthens the suggestion that a particular region is deleted. A stretch of mutant allele frequencies with ratios of 1:2 or 1:1:1 indicate duplicated regions. However, SNPs occur at a low frequency of approximately one per kilobase, and many are common enough in the general population that homozygosity can be expected. As such, this report is only occasionally informative in characterising structural variations.

Opposite Direction Read Report

Reads can be rejected from the alignment because they align to the reference sequence in the correct (opposite) orientation from one another, but at an unexpected gap distance (Figure 28). These reads are stored in an 'opposite direction reads' file, which lists the name of each rejected read pair, the positions in the reference sequence the reads align to, and the distance between them. Large numbers of reads are rejected on this basis from across the reference sequence, including in negative controls. These dispersed reads are dismissed as noise.

Structural rearrangements bring together two positions in the genome that are not normally adjacent to one another. If DNA fragments that cross a deletion breakpoint are captured, the two reads will align at unexpected distances from one another. As such, many breakpoint reads may be rejected from the alignment and stored in the opposite direction reads report. This report can be inspected to identify any 'pile up' of multiple rejected reads that align between the same two normally distant genomic positions. Overlaying RPKM data with the positions at which these reads align can help determine the precise breakpoints of a rearrangement which can then be identified in the NextGene Viewer, or by BLAT query of the reads listed in the opposite direction read report.

Same Direction Read Report

Same direction reads are rejected from the alignment because although they match the reference sequence, they are aligned in the wrong (i.e. the same) orientation to one another (Figure 28). Reads that cross the breakpoints of an inversion are rejected by

this condition and so this is a crucial report for detecting these rearrangements. Inversions are balanced rearrangements, so they cannot be detected using expression data. It is important to detect these rearrangements as they can have clinical significance. Complex inversion-deletion events have also been reported in thalassaemia, so knowing whether an inversion is involved in a rearrangement is critical for designing primers to fully characterise it.

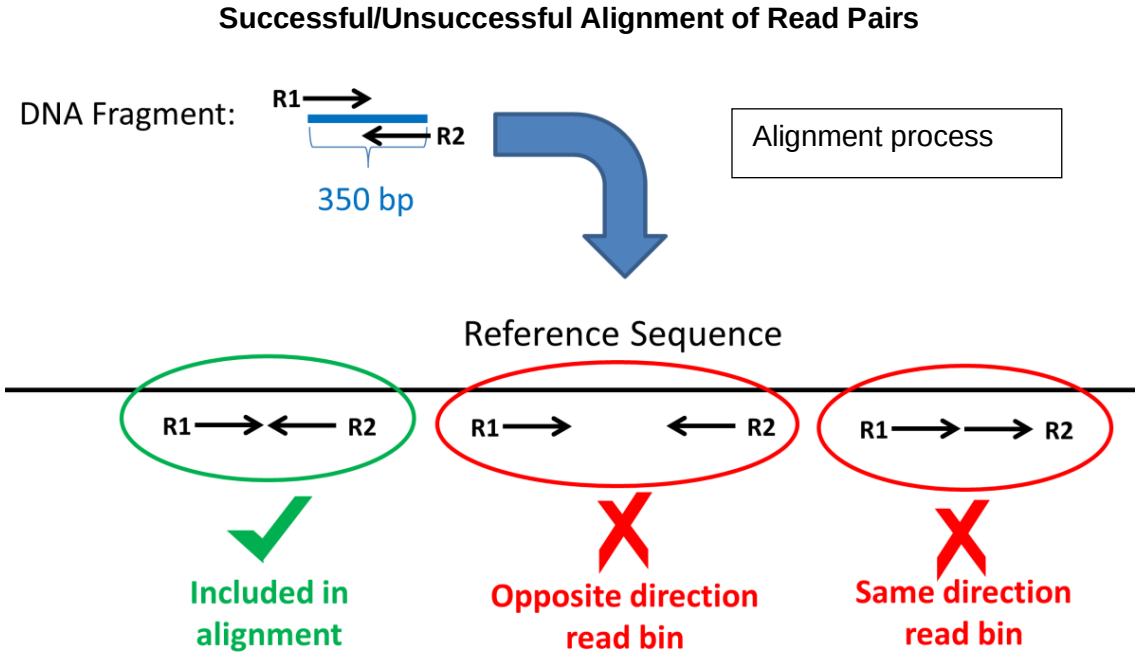


Figure 28: Successful/Unsuccessful Alignment of Read Pairs. A DNA fragment that cannot be aligned correctly to the reference sequence may be rejected as Opposite Direction (correct orientation of Read1 and Read 2 from one another but incorrect gap distance) or Same Direction (incorrect orientation of R1 and R2) reads.

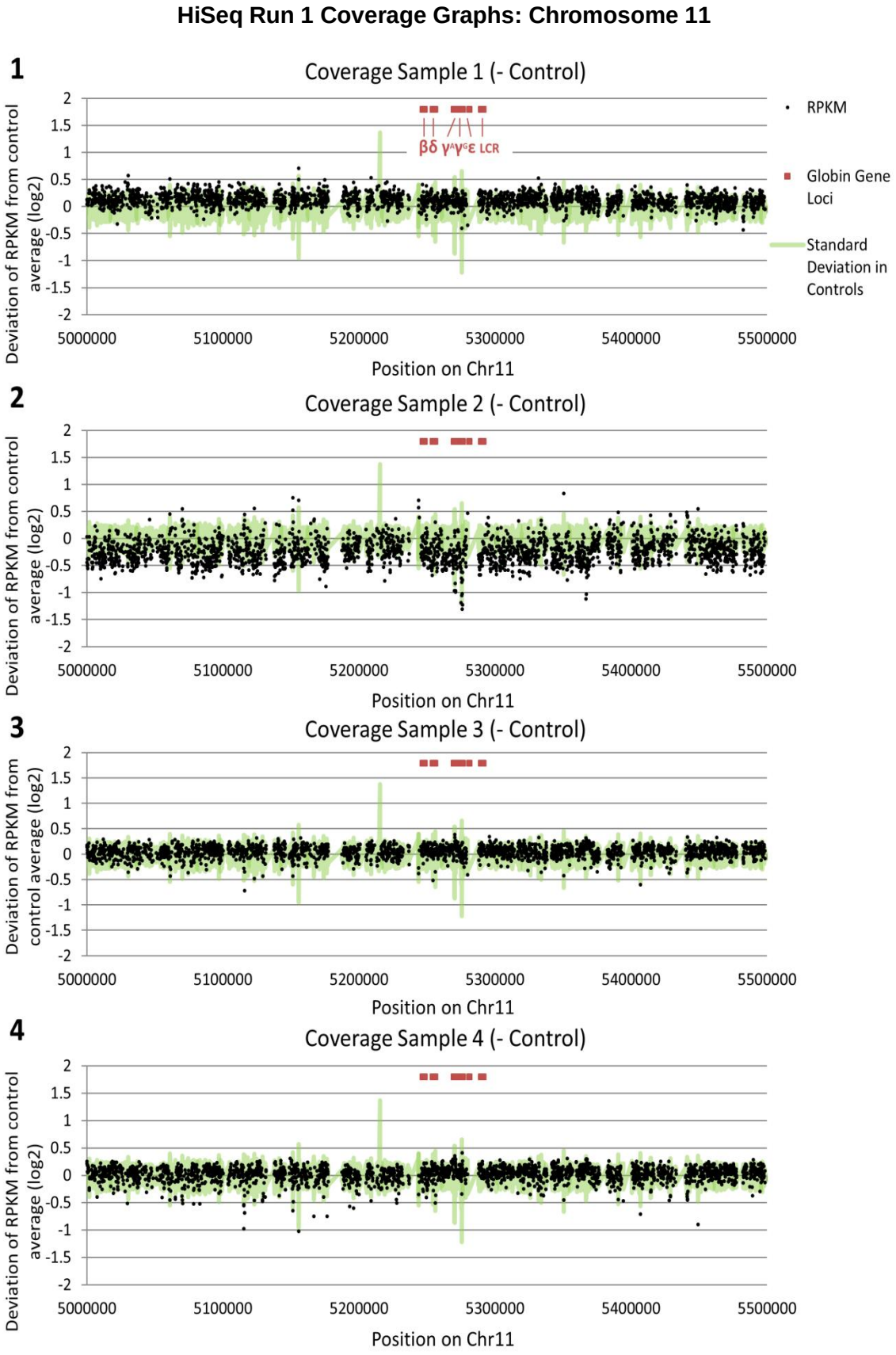


Figure 29: HiSeq Run Coverage Graphs Chromosome 11. Part 1- RPKM plots across Chr11 for HiSeq Samples 1-4.

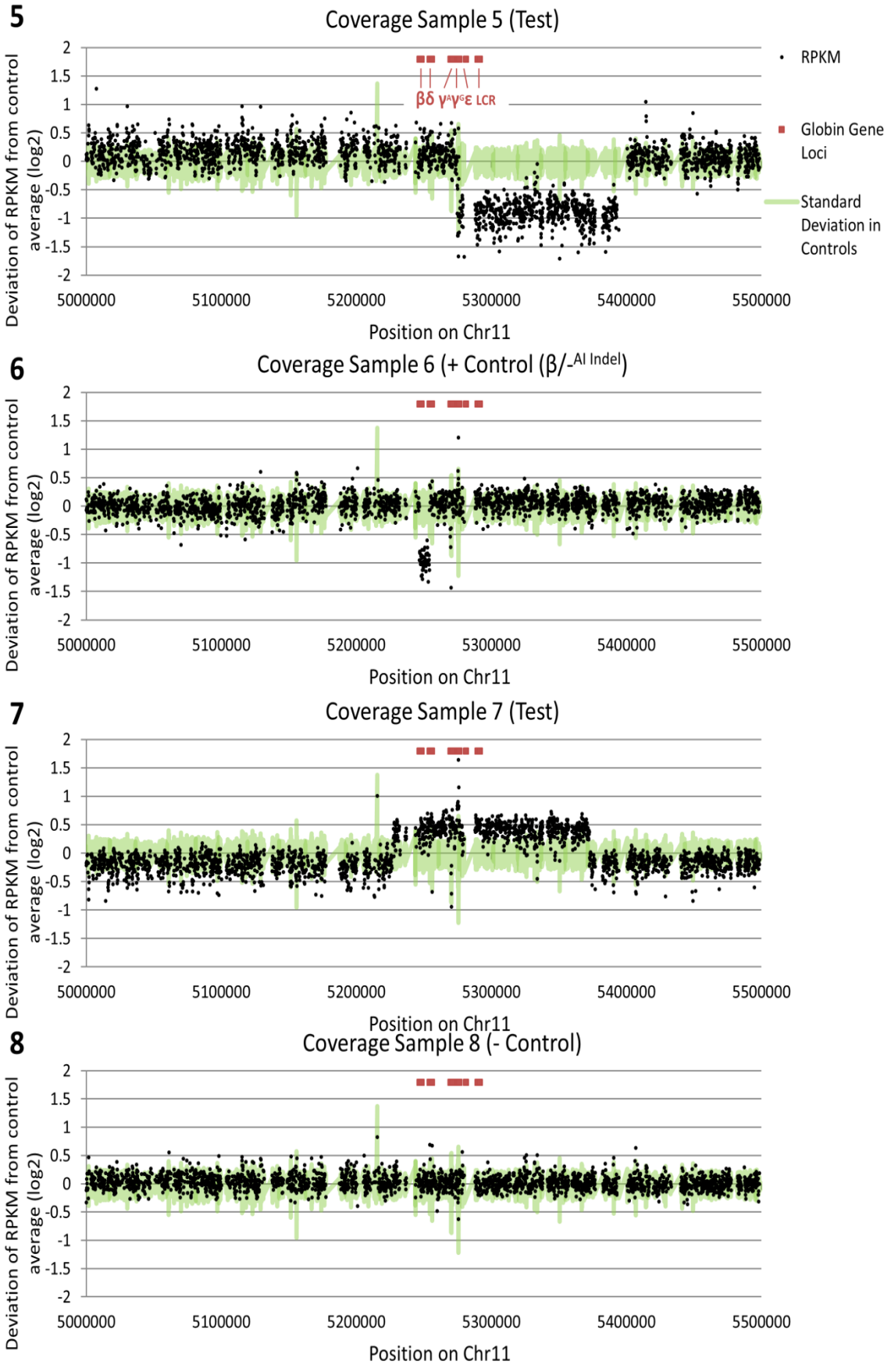


Figure 29 continued: RPKM plots across Chr11 for HiSeq Samples 5-8.

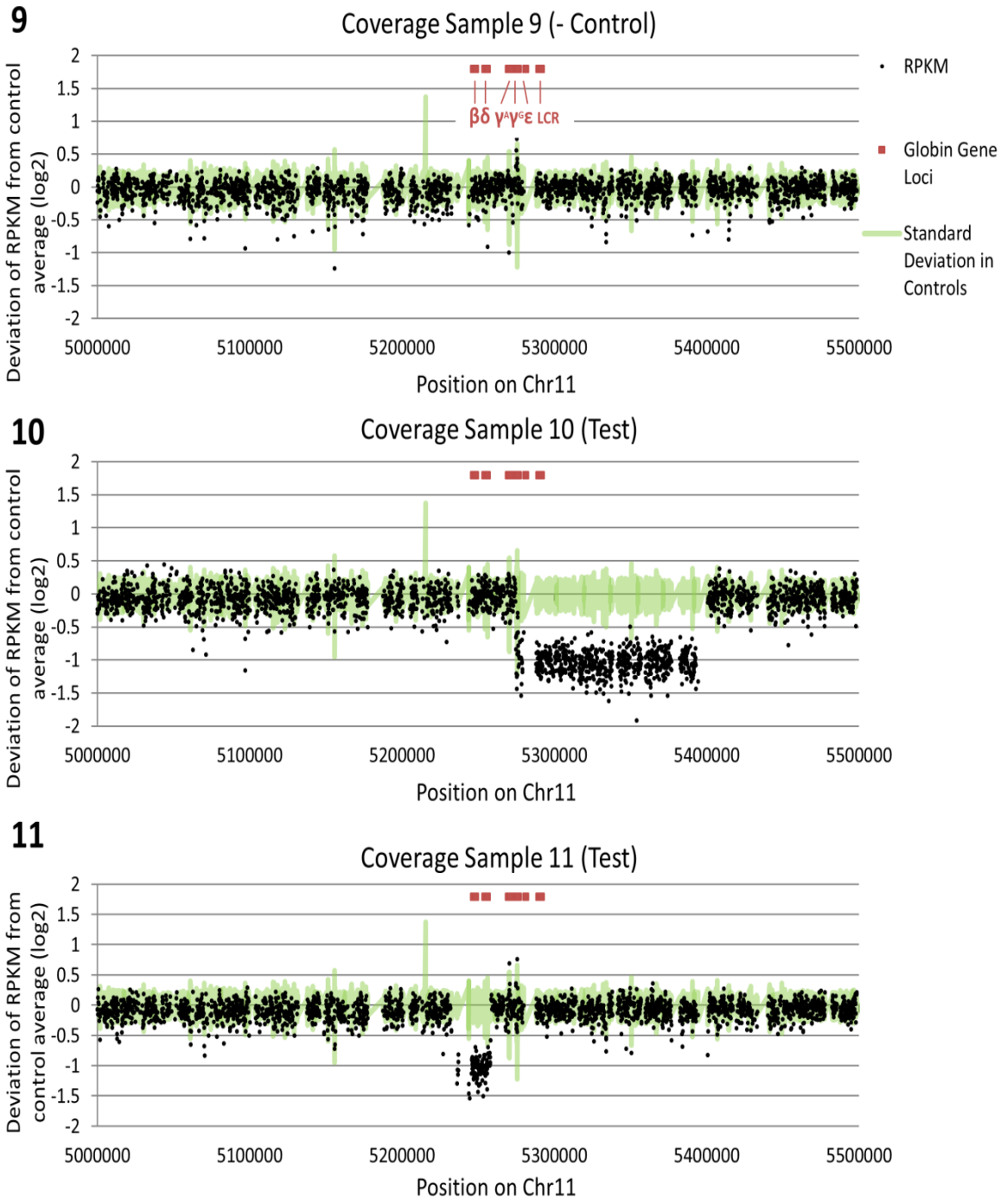


Figure 29 continued: RPKM plots across Chr11 for HiSeq Samples 9-11.

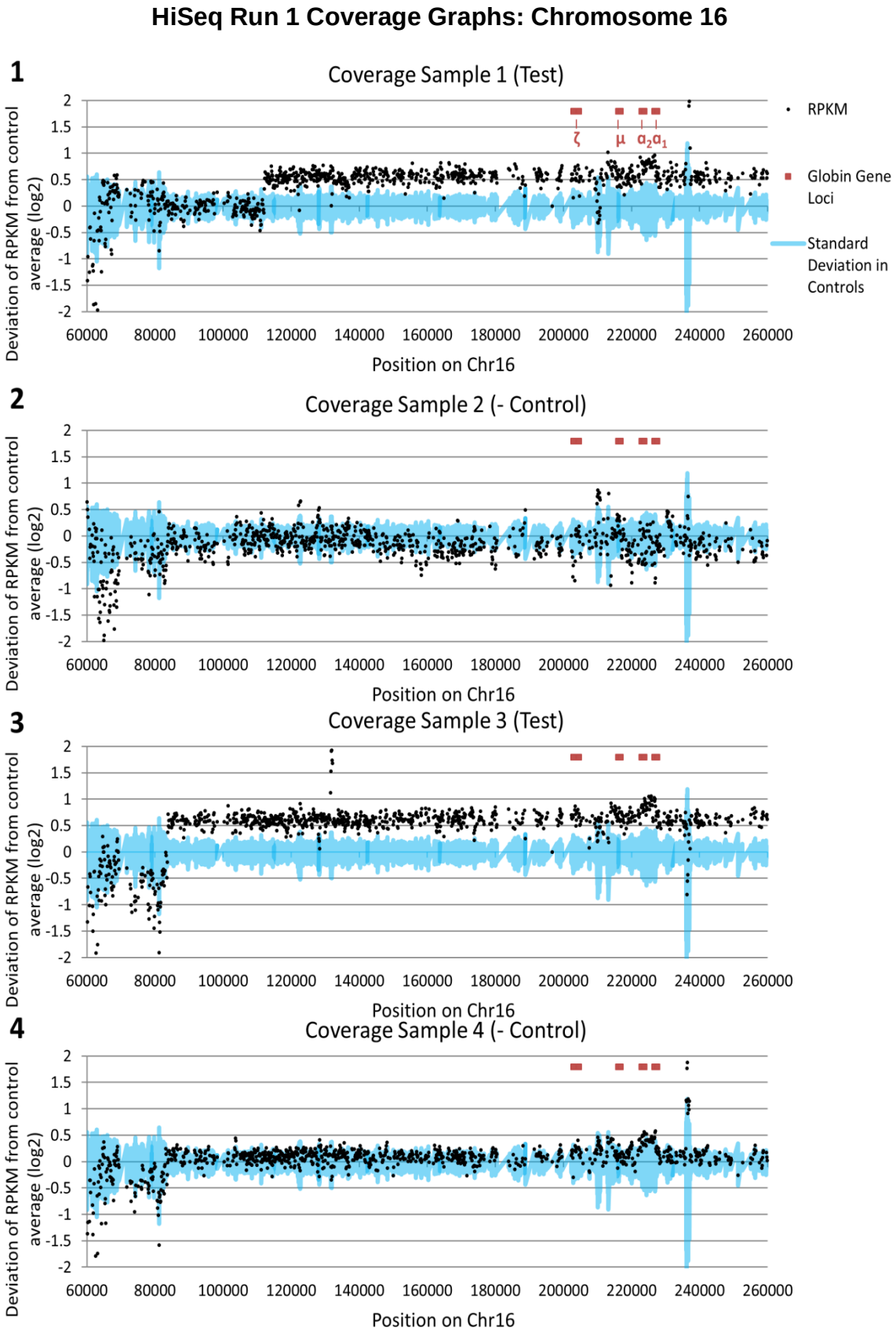


Figure 30: HiSeq Run Coverage Graphs: Chromosome 16. Part 1 - RPKM plots across Chr16 for HiSeq Samples 1-4.

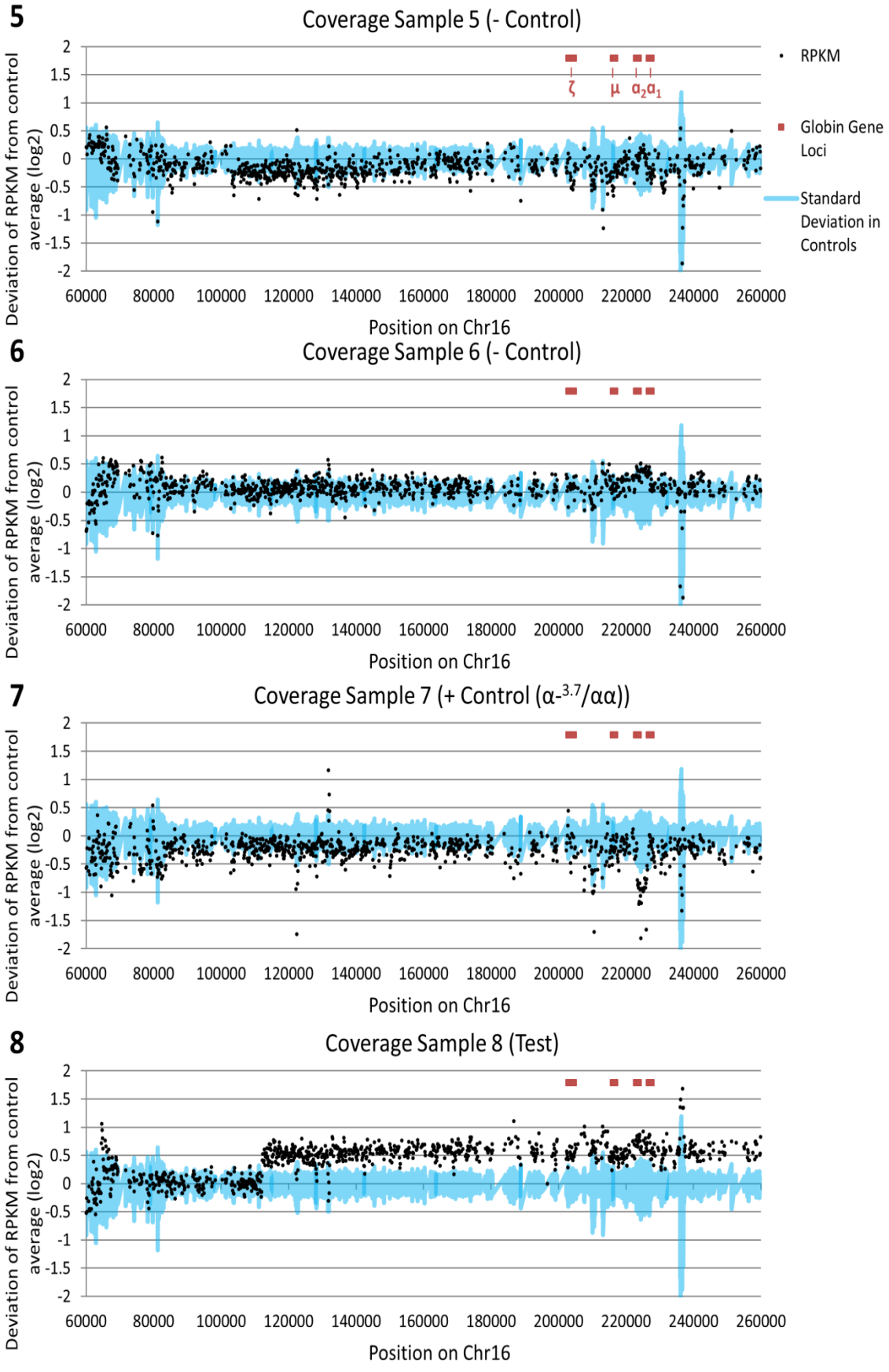


Figure 30 continued: RPKM plots across Chr16 for HiSeq Samples 5-8.

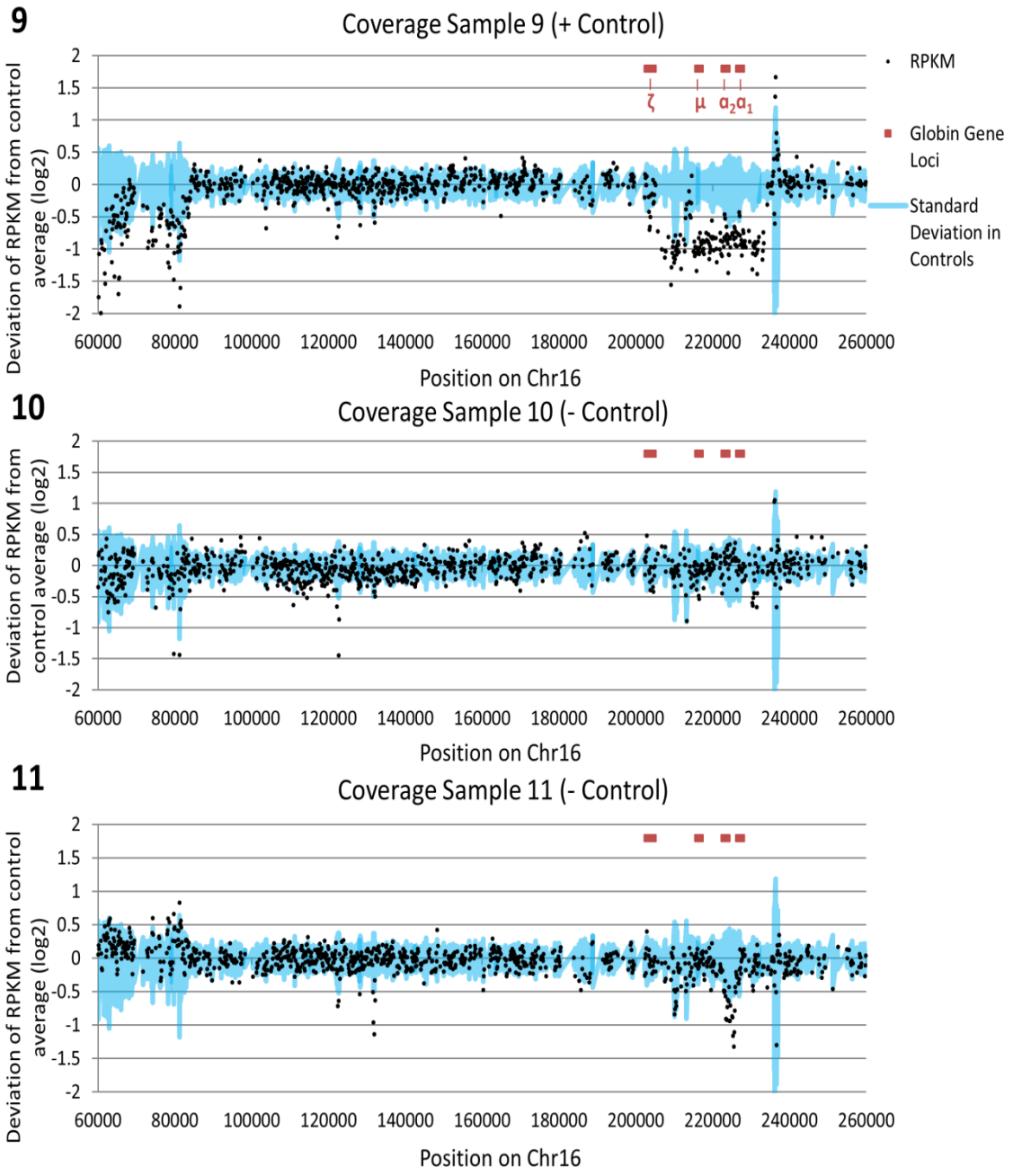


Figure 30 continued: RPKM plots across Chr16 for HiSeq Samples 9-11.

Data Analysis in Negative Control: HiSeq Sample 4

HiSeq Sample 4 in the sample cohort was a negative control. MLPA had confirmed that the individual did not carry any structural rearrangements at either the chr11 or chr16 globin gene locus. This individual has no family history of haemoglobinopathies and is not expected to carry any variants related to haemoglobinopathy.

Mutation Report for HiSeq Sample 4

The mutation report for HiSeq Sample 4 listed 4,091 variants within the region of interest. 1,049 of these were in the CDS, and 35 caused an amino acid change with a

score >15. In the globin genes, 37 variants were recorded, of which 2 occurred in the CDS, neither of which caused an amino acid change (Table 27).

Table 27: Contents of Mutation Report for Negative Control: HiSeq Sample 4.

	Within ROI (Chr11:3,500,000-7,500,000) (Chr16:0-260,000)	**Within Globin Genes**
Total Variants	4091	-
Within Genes	1049	37
Within CDS	228	2
Amino Acid Changing	46	0
Score >15	35	0
Homozygous	15	0

Detection of structural variants for HiSeq Sample 4

Chromosome 11

Coverage appeared uniform for the negative control sample across the region of interest on chromosome 11. This indicated that, as expected, no dosage changes had occurred in the sample at this position. A small number of bait-covered positions exceeded the negative control standard deviation. This is likely to be due to noise in the assay: there were no positions where this deviation was sustained across a continuous region, and there was no evidence of break point reads in the NextGene viewer at these regions.

Chromosome 16

Like chromosome 11, coverage across chromosome 16 in the negative control was uniform with no evidence of dosage changes. A significant increase in coverage was recorded at one position (~chr16:235,000). High variation in this region across the sample cohort was reflected in the negative control standard deviation recorded at this position. This position was evidently duplicated or deleted with a high frequency in the sample cohort, and was therefore likely to be a copy number variant. Query of the position in dbVAR revealed a copy number variant had been reported at this position which matched the data: esv2662940 frequently causes deletions at this position. The frequency of deletions in samples used to calculate the negative control average caused samples where this position was intact (such as in negative control sample 4)

to appear to show a duplication at the position, due to their relatively increased coverage.

Opposite and Same Direction Read Reports

Both reports contained a large number of entries from positions across the genome. Many entries had partial alignment to the regions of interest on chromosome 11 and 16, but with the other read aligning to other positions across the genome. It is likely that these fragments are from other regions of the genome, but were captured based on their partial alignment to the reference region covered in the bait library. Other fragments were the result of incomplete target enrichment during sample preparation, which resulted in 47% of sequence reads aligning outside the target regions (Table 26). These reads were dismissed as noise. There was some 'pile up' of opposite and same direction reads on chromosome 11 between *HBG1* and *HBG2*, and on chromosome 16 between *HBA1* and *HBA2*. This is due to the extremely high homology of these genes, which causes issues with correctly aligning reads from these locations to the correct position in the reference sequence. These were also considered to be noise and ignored. No other pile-ups of multiple reads showing alignment to the same two locations were identified in the control data.

Characterising Structural Variants in HiSeq Sample 6 Positive Control: Asian Indian Inversion-Deletion (Figure 29.6)

HiSeq Sample 6 is a positive control for a known variant causing thalassaemia with an elevated level of HbF: HbVar 1038; commonly known as the Asian-Indian Inversion Deletion ("AI Indel"). This variant consists of two deletions (0.6 Kb and 7.5 Kb) flanking an inversion (15.5 Kb). The two deleted regions are clearly visible in the RPKM plot (the smaller deletion is harder to see as it is only represented by four consecutive baits). Three of the four break points these two deletions create were clearly visible in the NextGene viewer (Figure 31).

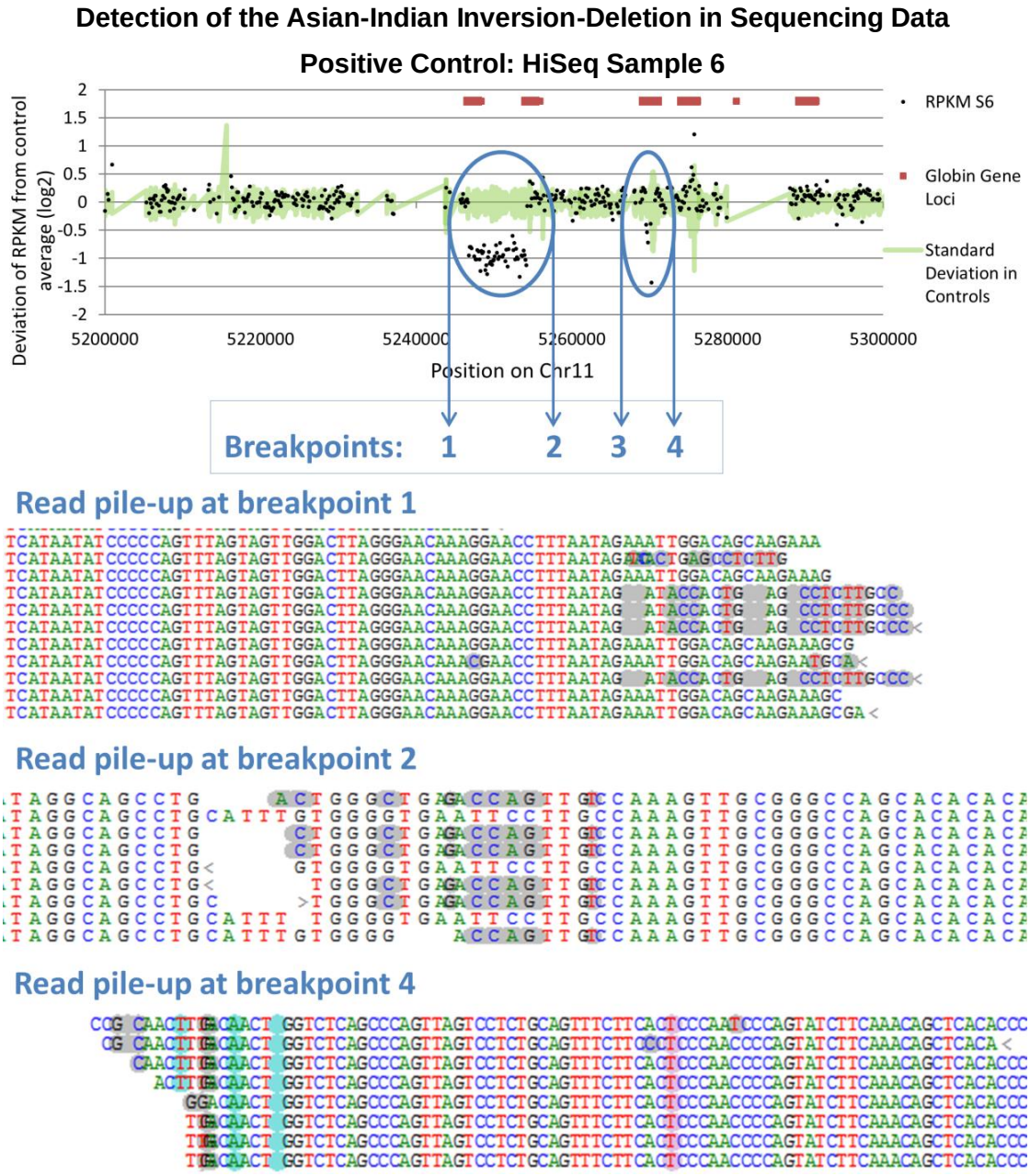
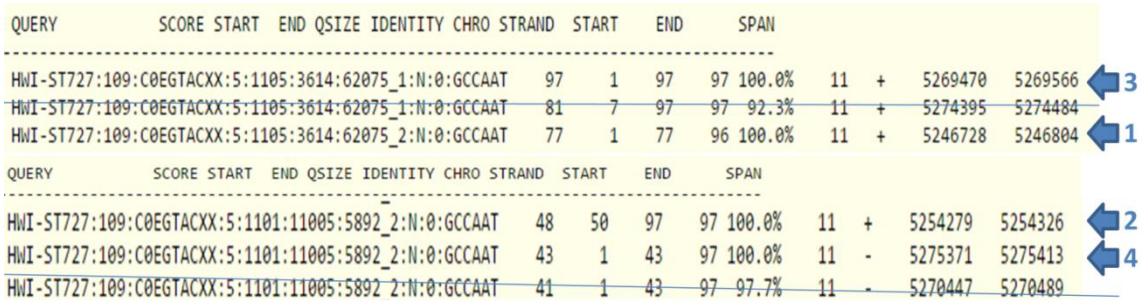


Figure 31: Detection of the Asian-Indian Inversion-Deletion in Sequencing Data. RPKM values across the bait-tiled region across chromosome 11 for HiSeq Sample 6 (Asian Indian Inversion-deletion). Graph: The two deletions identified in the upper panel (circled in blue) result in four breakpoint locations (1-4). Below: At three of the four locations, reads containing breakpoint sequence could be identified in the NextGene Viewer at locations shown in graph (highlighted in grey and blue). Breakpoint 3 occurs in a repetitive region and was not captured.

The two deletions involved in this variant are clearly visible in the RPKM plot. Inspecting the breakpoint regions as indicated by this plot in the NextGene Viewer revealed that three of the four breakpoints of this rearrangement had been included in the alignment. The remaining breakpoint could be determined by retrieving full read sequences from reads that aligned to the reference at the other side of this rearrangement, and querying them in BLAT. This revealed the co-ordinates of the two sequences brought together by the rearrangement (Figure 32).

BLAT Query of Sequences Aligning to AI-Indel Breakpoint Positions



QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN	
HWI-ST727:109:C0EGTACXX:5:1105:3614:62075_1:N:0:GCCAAT	97	1	97	97	100.0%	11	+	5269470	5269566		3
HWI-ST727:109:C0EGTACXX:5:1105:3614:62075_1:N:0:GCCAAT	81	7	97	97	92.3%	11	+	5274395	5274484		
HWI-ST727:109:C0EGTACXX:5:1105:3614:62075_2:N:0:GCCAAT	77	1	77	96	100.0%	11	+	5246728	5246804		1
QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN	
HWI-ST727:109:C0EGTACXX:5:1101:11005:5892_2:N:0:GCCAAT	48	50	97	97	100.0%	11	+	5254279	5254326		2
HWI-ST727:109:C0EGTACXX:5:1101:11005:5892_2:N:0:GCCAAT	43	1	43	97	100.0%	11	-	5275371	5275413		4
HWI-ST727:109:C0EGTACXX:5:1101:11005:5892_2:N:0:GCCAAT	41	1	43	97	97.7%	11	-	5270447	5270489		

Figure 32: BLAT Query of Sequences Aligning to AI-Indel Breakpoint Positions. Multiple matches are listed for each query. The BLAT query returns (left to right) query name, score, start and end of match, length of query, proportion of query that matches the reference sequence, chromosome, direction, start and end point of sequence match in genome.

Although in this instance the breakpoint sequences identified in NextGene were sufficient to characterise the entire inversion deletion (as BLAT indicates the orientation of the sequences conjoined at both breakpoints), whether or not it could be detected in the opposite/same direction read reports was also investigated.

Opposite Direction Reads

The RPKM data plot was overlaid with the locations of read pairs from the opposite direction read report (Figure 33). Reads were dismissed as noise where one half aligned to a position on another chromosome in this case, because we knew that these were spurious in this rearrangement. Each read pair was assigned an arbitrary Y-Axis value; this facilitated visual identification of clusters of reads pairs with similar start and end positions. The plot does not reveal any pile-up of pairs of opposite direction reads at the break-point locations of this rearrangement. This is due to the inversion of the intervening sequence, which means that reads covering the deleted region are in fact same direction, rather than opposite direction reads.

Locations of Opposite Direction Reads in the AI-Indel

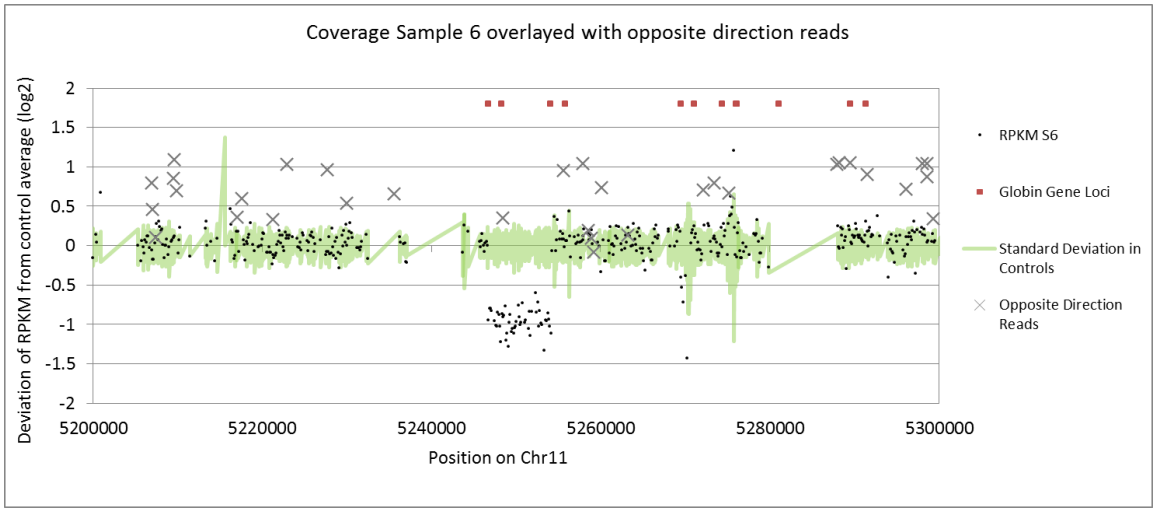


Figure 33: Locations of Opposite Direction Reads in the AI-Indel. Opposite direction reads (grey crosses) are overlaid on the RPKM plot for HiSeq Sample 6 (an enlarged version of the rearrangement region is shown here). The X-axis value shows the position on the reference sequence each read aligns to. Each read pair was assigned an arbitrary Y axis value. Thus, crosses with the same Y axis value are different halves of the same read.

Same Direction Reads

The sequence between the two deleted regions in the AI Indel is inverted. Therefore, a certain pattern of same direction reads covering the area should be expected, outlined in Figure 34.1. Each deletion is covered by one set of DNA fragments. The two reads taken from each fragment align either side of the deleted region. With two deletions, this means there are a total of four break point regions. The inversion covers the region between the two deletions, and therefore includes one breakpoint from each of the deletions. This means that each set of reads covering the deletion breakpoints are same direction reads, as one read has been inverted. The pattern this produces is shown below, where pink lines indicate fragments covering one deletion, and blue lines indicating fragments covering the other. The way NextGene attempts to align these fragments to the reference sequence, and the resulting same direction read pairs is also shown.

Figure 34.2 shows the actual clustering of same direction reads overlaid over the RPKM data plot. For clarity, the clustered reads that conform to the predicted read pile-up are also indicated in pink and blue. Other same direction reads which occur as background noise are in grey. An additional region of read pile-up is noticeable for reads aligning to the blue breakpoint pair. This due to ambiguous alignment between *HBG1* and *HBG2* that confounds the data at this position.

Predicted and Observed Positions of Same Direction Reads in the AI Indel

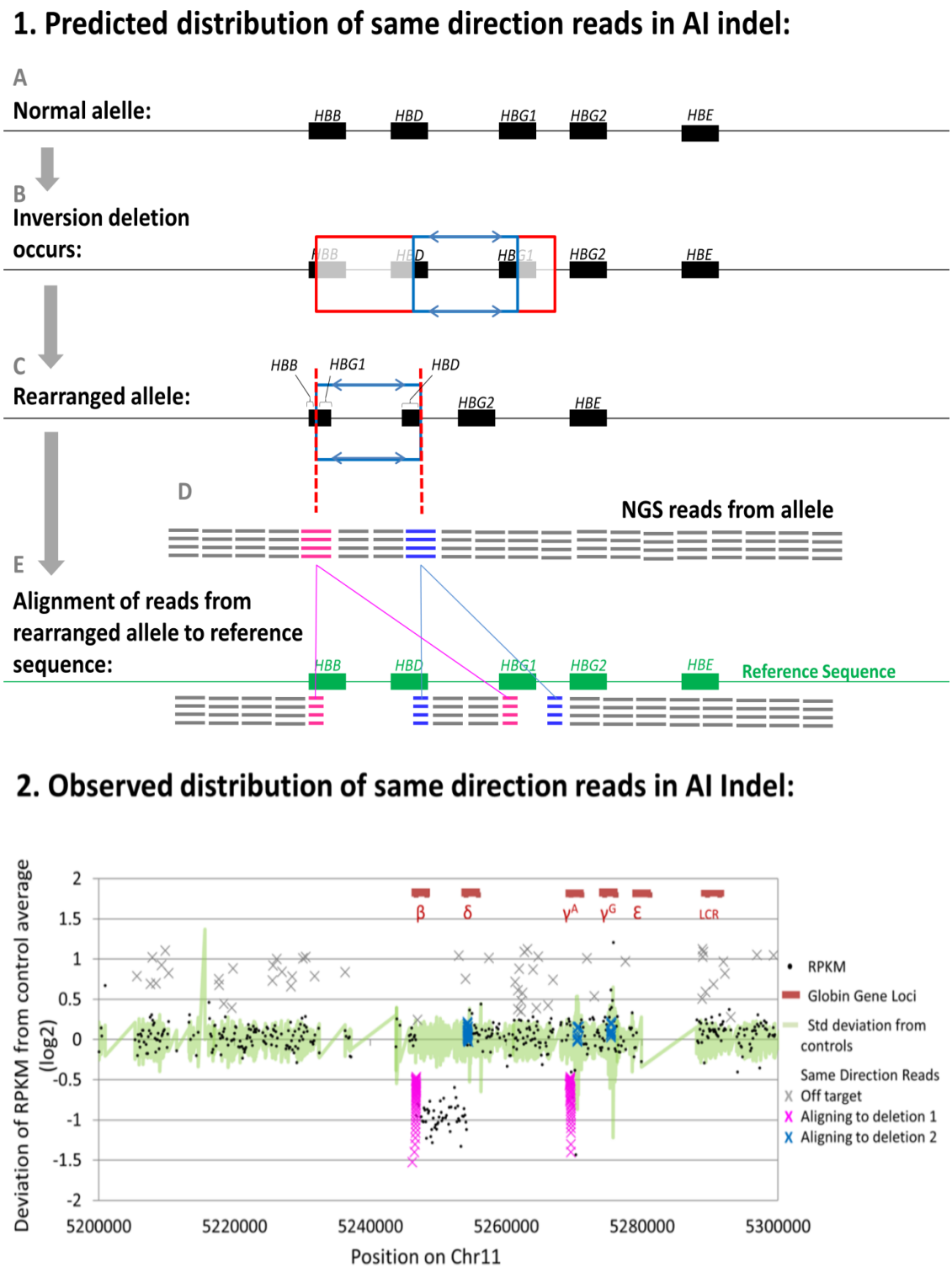


Figure 34: Predicted and Observed Positions of Same Direction Reads in the AI Indel.

1: Same direction read data for the Asian-Indian inversion deletion. The layout of the Asian-Indian inversion deletion is described in the upper panel (1). 1A: normal organisation of the globin gene loci. 1B: the rearrangement occurs, and two regions (outlined in red) are deleted, and the intervening region (outlined in blue) is inverted. 1C: This results in a rearranged allele that truncates *HBB*, *HBD* and *HBG1*, also inverting the remaining sequence of *HBD* and *HBG1*. 1D: When DNA from this region is sequenced, most of the captured fragments will align normally to the reference sequence (1E). Reads that originate from fragments that cross the inversion-deletion breakpoints (pink and blue) will be rejected from the alignment as same-direction reads.

2: Plotting the positions of these reads against the positions they align to on the reference sequence results in two clusters of linked reads (pink and blue) that straddle each break point of the inversion deletion. The observed distribution of same direction reads across the region of the globin gene loci each pair of same-direction reads are assigned the same arbitrary Y-axis value so that linked values can be identified. Grey crosses show same direction reads that are spurious noise. Pink crosses show a group of reads where the two halves cluster around the same regions: the first break point of the inversion deletion, as predicted in 1D. A second group shows same direction reads that cross the second inversion deletion break point. Some of these reads have been erroneously aligned to HBG2 instead of HBG1 due to the extremely high homology between these two genes.

To-the-base resolution of this variant was possible, as three of the four breakpoints were successfully identified and the two regions affected by the deletion were clearly visible. Characterisation of this variant was complicated by the presence of the inverted region, as same direction reads related to this variant were obscured by the noise occurring between the two gamma globin genes. In this patient an additional variant was identified by looking through the NextGene Viewer, a 17 bp deletion was clearly visible within *HBG1* (Figure 35). This deletion is a known polymorphism recorded in the dbSNP database, named Rs72324537 (Intron variant; Clinical significance N/A (Sherry, Ward et al. 2001)). Multiple positions are visible where NextGene has attempted to 'shoehorn' the deletion containing sequences into aligning. Correctly aligning data at this position may have been particularly challenging for the software due to the repetitive nature of its surroundings. This software, or perhaps the settings used in this alignment, may not be reliable in detecting small indels which it is unable to align correctly, and are too small to be identified by inspecting the coverage reports.

An Indel in the AI Indel Alignment not Called by NextGene

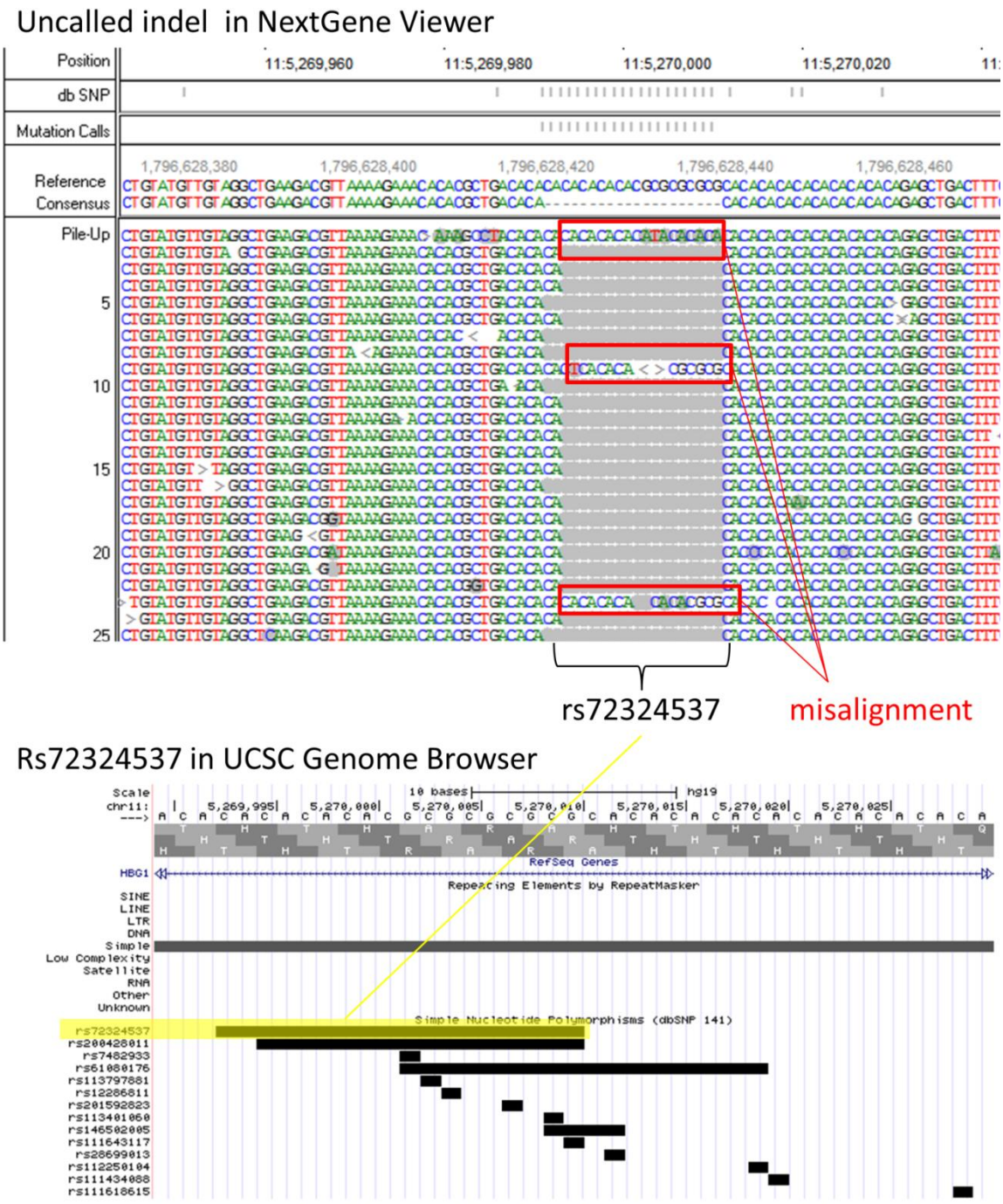


Figure 35: An Indel in the AI Indel Alignment not Called by NextGene. Upper: A 17 bp deletion is clearly visible in the NextGene Viewer, but is not called by the software. The variant occurs in a simple repeat. Lower: The deletion as recorded in dbSNP, displayed on the UCSC genome browser.

Characterising Structural Variation in the Test Samples

All test samples showed clear regions where their coverage was substantially higher or lower than the range seen in negative controls. It was possible to estimate the size and dosage of the rearrangements in these samples based on the number of baits affected by the rearrangement (Table 28).

Table 28: Rearrangements identified in sample cohort from RPKM plots. Where available, HbVar database ID numbers are included for positive control samples for previously reported rearrangements. "+ Cntrl" = Positive Control; "-C =" Negative Control.

	Sample	Rearrangement	Co-ordinates	Size (bp)
Alpha Globin Gene Cluster	Test Sample 1 Figure 30.1	Het Deletion **Alpha Globin Gene Cluster**	Chr16: 112,159 - ?	Cannot be estimated
	Test Sample 3 Figure 30.2	Het Duplication **Hb.Var.1076 (α-$^{3.7}$/)** ***HBA1***	Chr16: 83,214 - ?	Cannot be estimated
	Positive Control Figure 30.7	+Cntrl Het Deletion **Alpha Globin Gene Cluster**	Chr16: 223,197-226,313	2,759 - 3,236
	Test Sample 8 Figure 30.8	Het Duplication **Alpha Globin Gene Cluster**	Chr16: 111,919 - ?	Cannot be estimated
	Positive Control Figure 30.9	+Cntrl Het Deletion **"British Deletion"** **Alpha Globin Gene Cluster**	Chr16: 205,187- 234, 194	28,767 - 29,247
Beta Globin Gene Cluster	Test Sample 5 Figure 29.5	Het Deletion **Beta LCR**	Chr11: 5,274,634- 5,393,003/5,400,736	118,129 - 126,342
	Positive Control Sample 6 Figure 29.6	+ Cntrl Het Inversion-Deletion **HbVar.1038 (Al InDel)** ***HBB, HBD, HBG1, HBG2***	Chr11: 5,254,616-5,246,648 And 5,269,528-5,270,000	1,028 and 472
	Test Sample 7 Figure 29.10	Het Duplication **Beta LCR**	Chr11: 5,226,289 - 5,372,652	146,0363
	Test Sample 10 Figure 29.10	Het Deletion **Beta LCR**	Chr11: 5,274,634- 5,393,003/5,400,736	28,767 - 29,247
	Test Sample 11 Figure 29.11	Het Deletion ***HBB, HBD***	Chr11: 5,232,355 - 5,258,687	21,572 - 26,322

Variant Characterisation: HiSeq Samples 1, 8 (Figure 30.1, Figure 30.8)

The duplication on chromosome 16 in HiSeq Sample 1 extended beyond the region covered by the bait design (See Figure 30.1). The start position of the duplication (as indicated by the change in coverage relative to the negative control samples) is ~chr16:112,164. HiSeq Sample 8, a related individual, showed a similar RPKM plot (See Figure 30.8) as expected. The pile-up of sequences aligning to the reference at the duplication start point was examined in the NextGene Viewer. Numerous reads that

had aligned to the reference sequence at this position contained a string of bases that did not match the reference sequence (Figure 36 B). The ID numbers of these reads were recorded and identified in the original FASTA files for the sample to obtain the full Read 1 and Read 2 sequences for these DNA fragments (Figure 36 C). The Read 1 and Read 2 sequences of these reads were queried using the BLAT tool hosted by the UCSC Genome Browser to identify matches to these sequences within the human reference genome. A number of matches were obtained for the sequences (Figure 36 D), only two merited high confidence scores (based on the high percentage of bases from the query included in the match and the number of bases matching in continuous sequence).

Characterising a Novel Rearrangement in HiSeq Sample 1

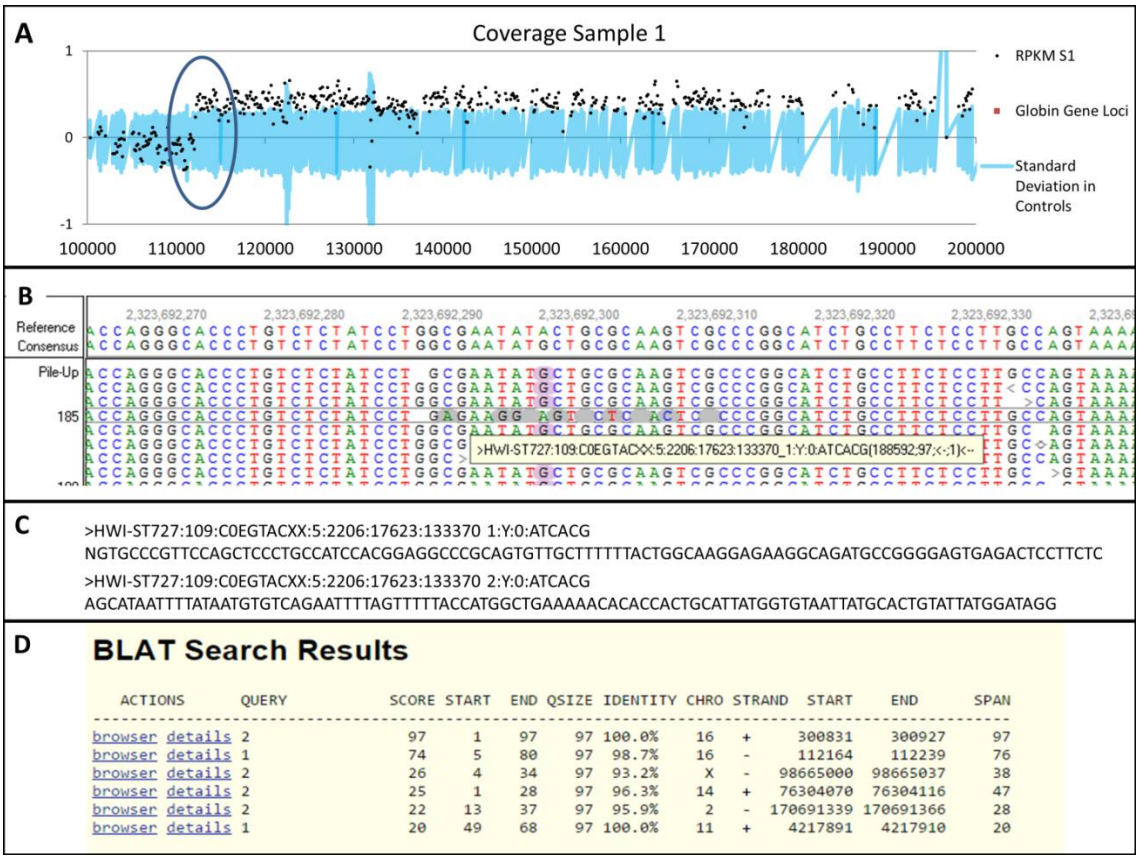


Figure 36: Characterising a Rearrangement in HiSeq Sample 1. (A) The start point for the duplication (circled in blue) (B) Read pile-up at the estimated duplication start position (C) The whole read sequence is obtained from the original FASTA files for the sample (D) The read sequence is queried in BLAT.

The confidence score UCSC assigns to query results is based on the absolute number of bases in the query that match the reference sequence at this location, the number of bases that match the reference sequence continuously, and the number of mismatches between the query sequence and the reference sequence. A blanket threshold for acceptable scores for all queries could not be established, as the length of sequences

queried varied greatly and this has a significant impact on the score a query is assigned. For this reason, as a general rule, the matches with the three highest scores for each read were considered as possible sources for the original sequence. If other matches had a similarly high score, these were also considered.

UCSC BLAT showed that these reads from the rearrangement breakpoint region in sample 1 contained sequences from two different points in the reference genome, which had been brought together by the rearrangement. Part of each read aligned to the reference sequence beginning from position chr16:112,164 (where these reads were identified in the NextGene Viewer); while the remaining read sequence was a positive match to a region 118Kb downstream, ending at position chr16:300,927. This meant that the sequence normally found at position 112,164 was repeated after position 300,927 in the genome, the intervening sequence having been duplicated. Both portions of the sequence in each read matched either the positive or negative strand of the UCSC reference sequence, meaning that the rearrangement was top-to-tail in orientation, rather than inverted (where one match would be to the positive strand, and the other to the negative strand). The duplication size is 118,763 bp. The layout of the rearrangement is shown in Figure 37A.

The novel duplication was confirmed by Gap-PCR. Primers were designed to capture the duplication breakpoint indicated by the NGS data: Pr1: 5'-GCTGCTTGGGTTTAATGAGG-3' 5'-AAGTCAGTGTGTCCCCGTTC-3'. These primers were expected to make a unique product of approximately 700 bp across the duplication break point. PCR amplification produced a single band of the expected size in the patient, and no band in a negative control (Figure 37B). The PCR product was sequenced (Figure 37C) and confirmed the same break points of the duplication (as suggested by the NGS data) in both HiSeq Sample 1 and HiSeq Sample 8, both showing identical break points for the rearrangement in both family members.

Schematic of a Novel Duplication in HiSeq Samples 1 and 8

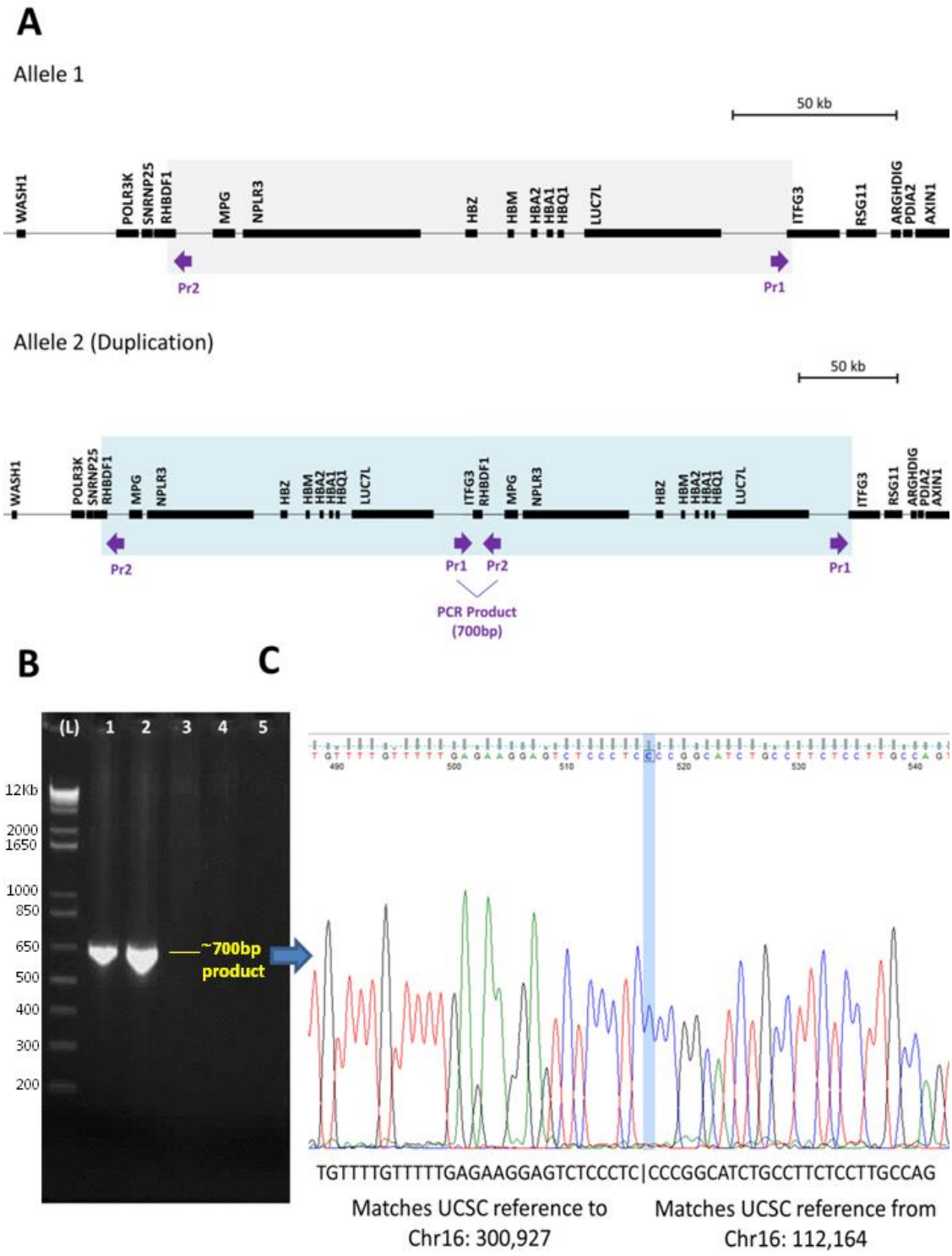


Figure 37: Schematic of a Novel Duplication in HiSeq Samples 1 and 8. (A) Schematic of rearrangement, showing primer locations on normal and duplicated allele (B) Gel image for Gap PCR. Lanes 1-6: (1) 1Kb+ DNA Ladder (2) HiSeq Sample 1 (3) HiSeq Sample 8 (4) HiSeq Sample 2 (unaffected relative) (5) Negative control (6) No template control (NTC). (C) Chromatogram showing sequence of PCR product from HiSeq Sample 1.

Variant Characterisation: HiSeq Sample 3 (Figure 30.3)

HiSeq Sample 3 shows a duplication that extends from approximately chr16: 83,214 to outside the bait-covered region. Inspection of the estimated start point of the duplication in the NextGene Viewer did not reveal any reads showing mismatched sequences at this position. One nearby region did show strings of mismatched bases, but the region was repetitive and BLAT query of these reads produced multiple high-scoring matches across the genome.

Opposite and same direction read analysis

To investigate whether any reads that contained the variant sequence had been rejected from the alignment outright, both the opposite and same direction read reports were examined. The opposite direction reads report data contained thousands of entries derived from off-target regions and alignment errors. As such, the report was filtered to only include entries where at least one part of the read pair was within 5Kb of the suspected duplication breakpoint covered by the bait design. No read 'pile-up' like that seen in the positive control sample was revealed that could represent reads crossing the duplication break points. The same direction reads report was filtered in the same way. This revealed nine individual read pairs where one half aligned to the approximate region of the duplication start point (66,635-68,951) and the other half aligned in the same orientation to the same region on chr15. This may indicate that a translocation had occurred, resulting in an additional copy of this region of chromosome 16 situated on chromosome 15. These reads appear in the same direction read report, which would also indicate that the translocation was inverted. Comparison to the negative control same direction report revealed similar reads, however, so this was dismissed as noise due to homology between the regions.

Single nucleotide variant analysis

The mutation report generated by the NextGene software for the region was inspected. No mismatches that could represent the duplication breakpoints had been recorded in the report. The mutation quality score cut-off point was reduced from >20 to >10. This leniency is used in order to account for penalties that the software may apply to genuine variants which are affected by the duplication. In reports from negative control samples that are known to be balanced, variants show a mutant allele frequency of 1:0 for homozygous variants or 1:1 for heterozygous variants. In the patient sample, from position chr16:60,842 onwards many heterozygous SNPs were recorded with allele frequencies of 1:2 indicative that three copies of the DNA at the positions of these

SNPs are present. In the patient, 80% of heterozygous variants recorded after position chr16:60,842 from the telomere show this 'duplicated' allele frequency (i.e. 1:2 or 2:1). No variants were present at a 1:1:1 ratio, which would imply three separate genotypes for this region (although SNPs where >2 possible genotypes exist are rare). This suggests that one of the two normal alleles present in this individual has been identically duplicated due to an intrachromosomal recombination event.

Variant Characterisation: HiSeq Sample 5/Sample 10 (Figure 29.5, Figure 29.10)

The RPKM plots for HiSeq Samples 5 and 10 indicate they have the same - or a very similar - deletion on chr11. This was expected, as these two individuals are related. The deletion removes exon 3 of *HBG2*, *HBE* and the beta globin 'Locus Control Region' (LCR), thus down regulating the expression of all the globin genes on the chromosome. The RPKM plot indicates the variant is 118,129 – 126,342 bp in size. A more accurate estimation could not be made as one breakpoint fell within a 6 Kb repetitive region that was not included in the bait capture. Investigation of the known breakpoint showed some reads with a small amount of mismatch. A BLAT query of the original reads revealed multiple (>50) matches for the non-aligning sequence. The highest scoring of these was for a region of *HBG1* that was highly homologous to this part of the *HBG2* gene. The other match on chr11 was located 58 Kb downstream of the breakpoint. That region was covered by an orphan bait which did not appear to have captured any sequence at this position.

Opposite direction and same direction read reports revealed a large number of misaligned reads that partially matched the suspected breakpoint position chr11: 5,274,684 in *HBG2*. The majority of these appeared to result from misalignments of read data from *HBG1*. The remainder showed matches across the genome, with no pile-up of multiple instances of the same mismatch that could indicate a genuine link between two distal sequence regions.

The mutation report confirmed the presence of a deletion between the positions indicated by the RPKM plot; all but two of the SNPs recorded between these two loci in bait-tiled regions showed allele frequencies of 90-100%, indicating there was no alternative allele throughout this 122Kb region. The two exceptional entries in the report were dismissed as spurious data: they were not located in a bait covered region, meaning that read pileup at these locations was more likely to include off-target sequences. Also, both variants were indels, for which allele frequency counts are less

accurate in NextGene, as some reads containing the variant are aligned to the reference incorrectly.

Gap-PCR primers were designed between the known break-point at chr11:5,274,684 and the first unaffected position after the 6 Kb repeat covered by the bait design (chr11:5,400,736) but these failed to create any product, even with a 10 minute extension time to allow for a product of up to 10 Kb (the maximum expected size of this product being 6.5Kb). Without any information about the breakpoints from the sequence, or from the opposite or same direction read reports, this rearrangement could not be fully characterised. The same rearrangement was later encountered in two other samples sequenced on the MiSeq (with longer insert sizes), and in this case to-the-base characterisation of the variant was achieved. The breakpoints were confirmed for both the MiSeq samples and the two samples discussed here. The novel rearrangement was reported in Human Mutation (Shooter, Rooks et al. 2015) and named the 'English V' inversion - deletion. Please see later section for detailed discussion of the successful characterisation of this rearrangement.

Variant Characterisation: HiSeq Sample 7 (Figure 29.7)

HiSeq Sample 7 was an antenatal referral from a woman who had an unusually low level of HbS (13%) with an HbF level of 0-2% and hypochromic microcytic red cell indices. Co-inheritance of alpha thalassaemia with the sickle variant was investigated as this is known to reduce the fraction of HbS in a sickle cell carrier. She was subsequently confirmed to be heterozygous for the alpha thalassaemia 3.7 deletion (HbVar ID: 1076), but this single alpha globin gene is not sufficient to explain the unusually low HbS percentage. Beta globin gene sequencing did not identify any thalassaemia variants which may be in *cis* with the sickle cell variant that could reduce the HbS level. A HbS Taqman assay that indicates the relative abundance of HbS and HbA was performed, and indicated that an additional copy of HbA could be present in the patient. The NextGene mutation report listed rs334 – the sickle cell disease variant - with a mutant allele frequency of 1:2, suggesting a duplication of the normal allele may have occurred (Table 29).

Table 29 Mutation report listing for rs334 in HiSeq Sample 7

Position	Reference Nucleotide	Coverage	Score	Mutation Call	Mutant Allele Frequency
Chr11: 5248232	T	422	18.8	T>AT	33.18

Inspection of the RPKM data for the covered region on chr11 indicated the presence of a large duplication (146,033-146,513 bp) extending from position chr11:5,226,876 to chr11:5,372,562, encompassing the entire beta globin gene cluster (Figure 39). The mutation report assigned allele frequencies of 1:2 or 2:1 for heterozygous SNPs in this region.

We would expect that a duplication would increase the coverage of the affected region by a third (+0.3 on a log2 scale) relative to the negative control average. We found that the general variability of the assay meant that bait positions from balanced regions frequently showed variation as great as + 0.3. For this reason, we elected to use a deviation of +0.5 from the negative control average as an approximate indicator of duplicated regions in this study rather than + 0.3. Figure 38 shows that this cut-off provides a clearer distinction between duplicated and balanced regions in three confirmed duplications from this study.

Bait Coverage in Balanced Versus Duplicated Regions for Three Samples

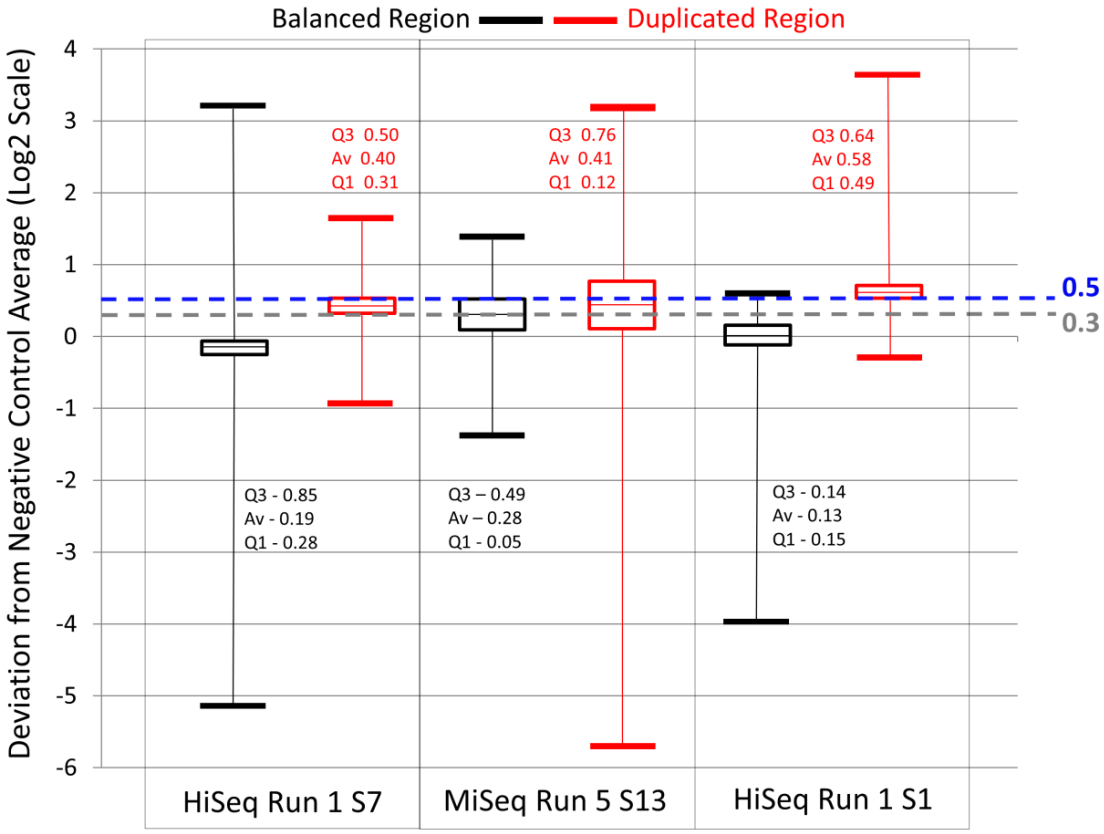


Figure 38: Bait Coverage in Balanced Versus Duplicated Regions for Three Samples. Box and whisker plots show range of Log2 values in duplicated versus balanced regions for three samples. Black plots show range of balanced values, while red plots show range of values in confirmed duplicated regions. Quartile values indicated by 'Q' and average values indicated by 'Av' are shown for each plot.

Inspection of the breakpoint regions in the NextGene Viewer revealed misaligned sequences in the read pile-up near both locations. BLAT query of reads containing these sequences in the original FASTA files confirmed that these misaligned sequences were the breakpoints of the duplication. The precise breakpoints were revealed to be positions chr11:5,226,885-5,372,677. Gap PCR and sequence analysis confirmed that the duplication was in a head-to-tail orientation (Figure 40). Eleven bp of novel sequence (CACCTCCACTT) was inserted at the break point. The 11 bp insertion, together with two last two bases of the duplicated region created a 13 bp mirror repeat. Previous studies have pointed out an association between mirror repeats and the incidence of duplication events. This scenario is unlike other incidents where mirror repeats are associated with duplications, as the duplication appears to create the mirror repeat, rather than the initial presence of a mirror repeat causing subsequent duplication. The variant was reported to HbVar as Hg19.Chr11 g.5372677_5372678insCACCTCCACTTdup5226885_5372677 and recorded with the ID 2961.

Characterisation of a Novel Duplication in HiSeq Sample 7

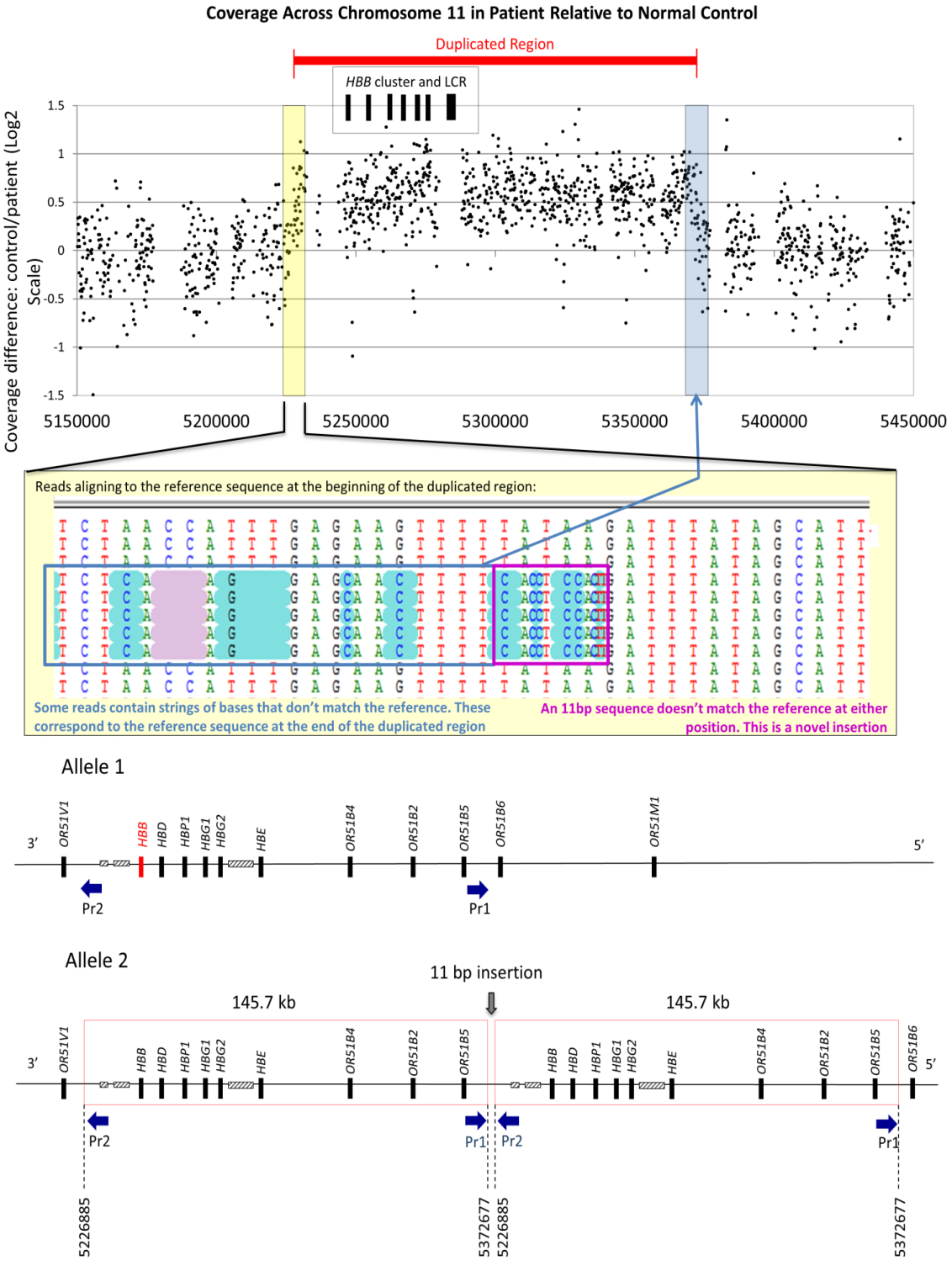


Figure 39: Characterisation of a Novel Duplication in HiSeq Sample 7. Upper panel shows duplicated region indicated by RPKM plot. Middle panel shows reads aligning to the reference sequence at the position of the first break point. Lower panel shows schematic of duplication according to BLAT query results.

Breakpoint Confirmation for Novel Duplication in HiSeq Sample 7

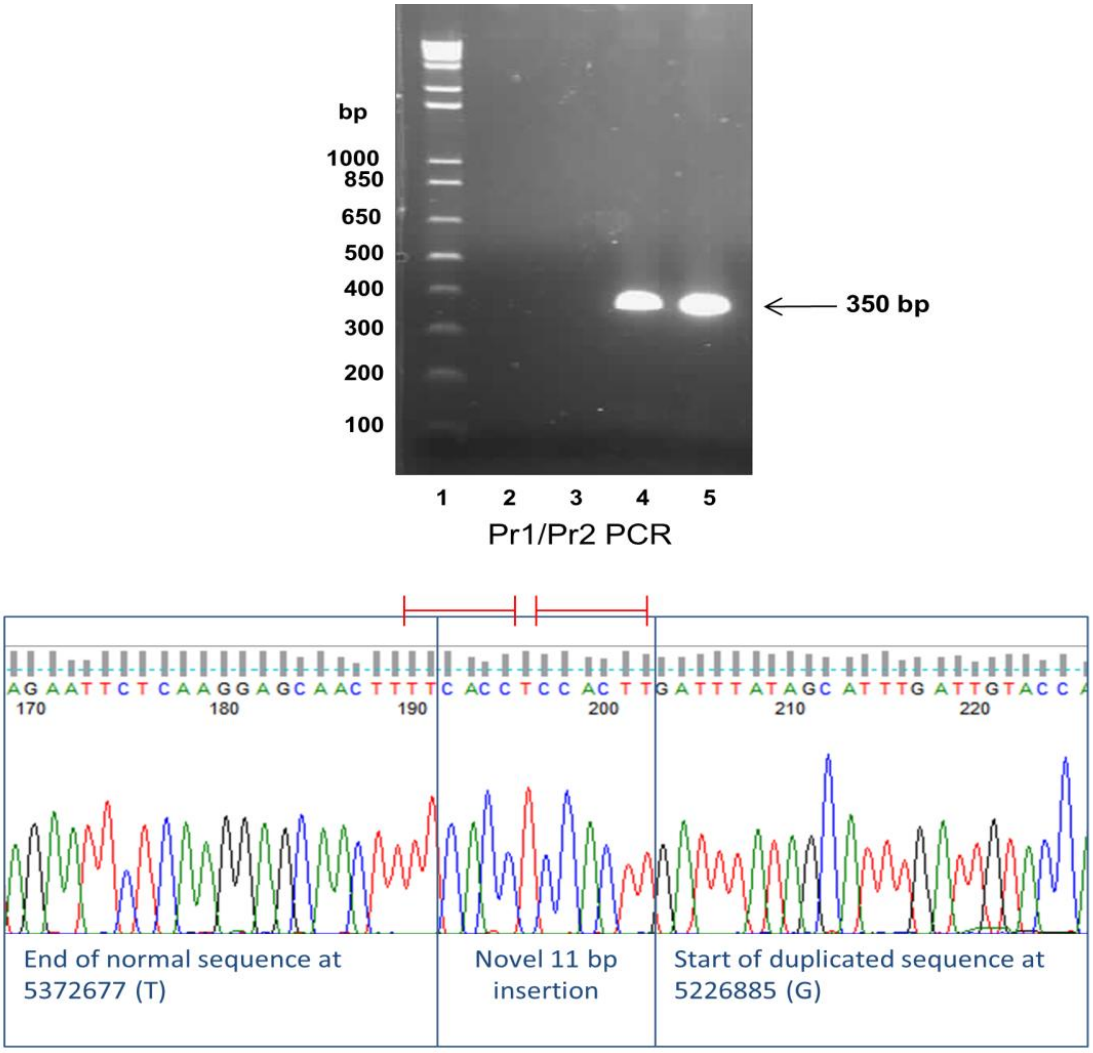


Figure 40: Breakpoint Confirmation for Novel Duplication in HiSeq Sample 7 by Gap PCR. Upper panel: Gel image showing (Lanes 1-5): (1) 1Kb plus DNA ladder (2) NTC (3) negative control (4) whole blood extracted DNA sample from sample 7 (5) saliva extracted DNA sample from sample 7. A 350 bp product is produced in lane 4 and lane 5. Lower panel: chromatogram of sequenced PCR product confirming the break point sequence identified by NGS analysis.

The RPKM plot for chr16 showed the presence of the 3.7 deletion. Inspection of the break point locations in the NextGene Viewer did not reveal any misaligned sequences. This is likely to be due to the extremely high homology between the sequences at the start and end points of the 3.7 deletion. There were no opposite or same direction reads identified in the location. This is expected, as the breakpoint is within exon/intron 1 between *HBA2* and *HBA1* and both have almost identical sequence. Therefore, any breakpoint-spanning sequences would align across the reference like normal paired end reads. The 3.7 Kb deletion was only detectable by the RPKM plot, supported by the lack of any heterozygous variants within the region that had acceptable quality scores.

Variant Characterisation: HiSeq Sample 9 (Positive Control for British Deletion HbVar.1091) (Figure 29.9)

HiSeq Sample 9 showed a deletion of approximately 30 Kb (co-ordinates Chr16:205,500-234,800) removing globin genes *HBA1*, *HBA2* and *HBM*. Both break-points appeared to be situated in repetitive regions that were not included in the bait design. Some variation outside the normal range made the position of the 5' break point particularly hard to estimate. Inspection of the region showed a large amount of misalignment within the region that was supposedly deleted across approximately 1 Kb of sequence. The misalignment in this region was not the same in all reads; several different sequences showed partial alignment to the reference. Although the region is not classed as repetitive by Repeat Masker, there was no pile-up of misaligned reads at either of the suspected breakpoints. The opposite direction reads report did not contain any reads which mapped across the breakpoint locations, with a gap distance equal to the estimated size of the deletion, or pile-up between either breakpoint position and any other region. Large numbers of same direction reads aligned to the regions of both break points, but no trends stood out that could indicate a legitimate rearrangement over the background of partial matches to similar repetitive elements. When the precise known co-ordinates of this deletion were investigated in the NextGene viewer, a mismatch of two bases could be seen in reads at the telomeric break point location. It was established that there was a high level of homology in the breakpoint region, meaning that only two bases in the breakpoint reads did not match what was expected in the reference sequence. These two bases had previously been dismissed as a SNP rather than a rearrangement breakpoint. Therefore, without prior knowledge of the co-ordinates of this variant, to-the-base characterisation of this rearrangement would not have been achieved. Having now identified this 2 bp change as the breakpoint for this deletion, we will be able to confirm future cases of the rearrangement based on the RPKM data and this signature mismatch.

Variant Characterisation: HiSeq Sample 11 (Figure 29.11)

The RPKM plot for Sample 11 showed a deletion removing the delta and beta globin gene loci. Both deletion breakpoints as estimated by the RPKM data lie in unmapped repeat regions. The first break point appears to be between positions chr11: 5,232,475-5,236,266 (LINE repeat, size: 3,791 bp) and the second between positions chr11: 5,258,078-5,258,567 (SINE repeat, size: 500 bp). The layout of this region of the genome is shown in Figure 41. The size of the deletion is therefore estimated as 21,572 bp – 26,332 bp. There were no strings of mismatched bases at either region that could represent the breakpoint sequence. No opposite direction reads or same

direction reads were found to cluster around either breakpoint, or cluster significantly anywhere in the wider surrounding region (excluding the expected mismatch between the two alpha globin gene loci). This variant could not be resolved with to-the-base accuracy using this methodology. This case could not be investigated further via Gap PCR due to insufficient DNA being available.

A Novel Deletion in HiSeq Sample 11 That Could Not Be Characterised

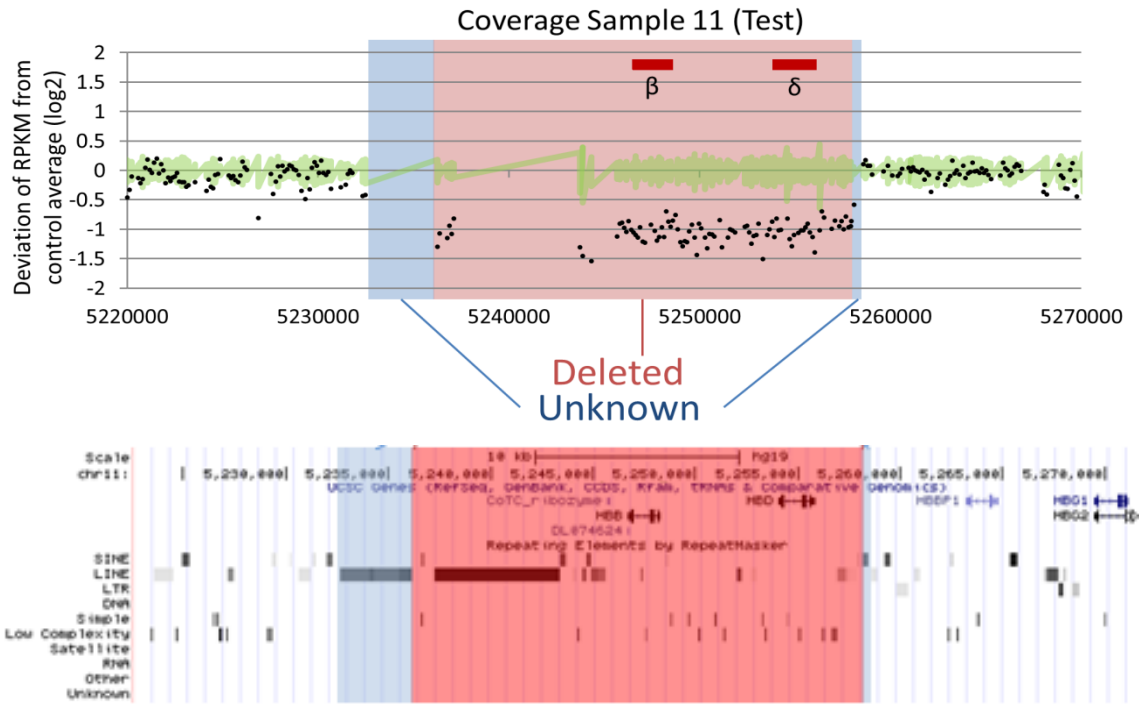


Figure 41: A Novel Deletion in HiSeq Sample 11 That Could Not Be Characterised. The region highlighted in red appears to be deleted. The flanking regions highlighted in blue are repeats (LINE and SINE) that are not covered by the bait design library. Lower panel shows the layout of this region, including repeats, in the UCSC genome browser.

Conclusions from HiSeq 2000 (Run 1)

The HiSeq allowed 12 samples to be sequenced simultaneously, achieving a depth of coverage for each sample that was sufficient to identify the structural variants they were expected to contain (mean coverage 400 x). Although all expected variants could be identified by the RPKM plots, not all of the rearrangements present could be resolved (Table 30). The region covered by the bait design on chromosome 16 was not large enough to encompass some of the structural variants identified; therefore in these cases only one end of the variant was mapped. Despite this, in one case, a duplication could be resolved with to-the-base accuracy using reads from the one captured break point. Other variants could not be resolved where the break-point within the target region occurred in a repeat sequence not included in the bait library, and thus the size of the variant could not be estimated. The Asian Indian inversion deletion positive

control sample was successfully characterised by this assay due to the fact that three of the rearrangement's four breakpoints were situated in non-repetitive sequence. In spite of this, there was still some confusion about the position of the fourth breakpoint, due to the high homology between the two gamma globin genes producing additional read pile-up. A novel duplication affecting the beta globin gene locus was successfully characterised, where neither breakpoint fell within sequence that was repetitive, or highly homologous to another region.

Multiple variants on both clusters could not be characterised due to their breakpoints falling in either repetitive and unmapped, or highly homologous sequences. This included the (α-$^{3.7}$/) deletion, which occurs between two highly homologous sequences. Although two known single nucleotide variants were successfully detected by NextGene and included in the mutation report, another variant was identified in the viewer which had not been detected, giving a false negative call. The uncalled variant was a 17 bp indel. Although this variant is a polymorphism with a silent phenotype, the fact that such a variant could be missed by software intended for diagnostic use is concerning. It is possible that increased read length will improve the ability of the software to confidently align reads containing this variant to the reference sequence and produce an accurate mutation call.

Table 30: Summary of Results from HiSeq Run 1.

	Sample	Rearrangement	Co-ordinates	Resolved
Alpha Globin Gene Cluster	HiSeq Sample 1 Figure 30.1	Het Deletion **Alpha Globin Gene Cluster**	Chr16: 112,164 - 300,927	**Yes**
	HiSeq Sample 3 Figure 30.2	Het Duplication **Hb.Var.1076, (α-$^{3.7}$/) *HBA1***	Chr16: 83,214 - ?	**No**
	HiSeq Sample 7 Figure 30.7	+ Cntrl Het Deletion **Alpha Globin Gene Cluster**	Chr16: 223,197-226,313	**No**
	HiSeq Sample 8 Figure 30.8	Het Duplication **Alpha Globin Gene Cluster**	Chr16: 112,164 - 300,927	**Yes**
	HiSeq Sample 9 Figure 30.9	+ Cntrl Het Deletion **"British Deletion"** **Alpha Globin Gene Cluster**	Chr16: 205,187- 234,194	**No**
Beta Globin Gene Cluster	HiSeq Sample 5 Figure 29.5	Het Deletion **Beta LCR**	Chr11: 5,274,634- 5,393,003/5,400,736	**No**
	HiSeq Sample 6 Figure 29.6	+ Cntrl Het Inversion-Deletion **HbVar.1038 (Asian-Indian Inversion Deletion)** ***HBB, HBD, HBG1, HBG2***	Chr11: 5,254,616-5,246,648 And 5,269,528-5,270,000	**Yes**
	HiSeq Sample 7 Figure 29.10	Het Duplication **Beta LCR**	Chr11: 5,226,885-5,372,677	**Yes**
	HiSeq Sample 10 Figure 29.10	Het Deletion **Beta LCR**	Chr11: 5,274,634- 5,393,003/5,400,736	**No**
	HiSeq Sample 11 Figure 29.11	Het Deletion ***HBB, HBD***	Chr11: 5,232,355 - 5,258,687	**No**

Inter-sample variation was high across both regions included in the design. This variation made it hard to identify small dosage changes reliably. In some cases the variation also prevented clear determination of where dosage changing variants began and ended in the alignment, due to noise in the surrounding region. Average deviation from the Log2 negative control average within the negative control cohort for chromosome16 was +0.199/-0.244, and for chromosome 11 was +0.16/-0.204. There was a particularly large degree of coverage variability in the region of the bait design nearest the telomere on chromosome 16:60,000-80,000Kb. In this region, the average standard deviation from the negative control average among the negative control samples was -0.47/+0.35. This variation does not appear to be the result of copy number variation, because the variation does not imply that clearly defined segments are variously present or absent, as is the case with other CNVs (Figure 26). Features

of the sequence at this region that may be responsible for this variability include elevated GC percentage (which correlates with increased coverage variation across the chromosome (Figure 42)), large segmental duplications that confer a high homology between this position and regions of other chromosomes, and human chained self-alignments. Multiple copy number variants are also recorded within this area (Figure 43). These features of the sequence may contribute to the coverage variability seen in this region.

Controlling for variability is necessary to reduce noise in the assay and improve its ability to detect rearrangements. A longer read length will aid this, as it allows alignment to be performed with a higher degree of specificity. Longer fragment length could also improve specificity, as alignment then relies on two non-overlapping sequences occurring at a given distance from one another. This would improve the ability of this assay to reliably sequence into repetitive regions, where breakpoints are frequently situated.

Although the HiSeq produced a wealth of sequencing data and good coverage, the read length (2x97 bp) was short and limited sequencing into repetitive regions. By shearing the DNA to a larger size and increasing the read length it was hoped that more of the adjoining repeat regions could be sequenced. The main concern of this approach would be the coverage read depth, as this type of sequencing can only be done on a MiSeq, not the HiSeq. By moving to the MiSeq platform the assay can be run in a diagnostic laboratory and removes the need for collaboration with large genomic sequencing facilities. The MiSeq platform has lower sequencing capacity than the HiSeq 2000, which means that fewer samples can be included in each run to produce the same read depth for the targeted region. High read depth is crucial to accurately detecting structural rearrangements, so it may be necessary to significantly reduce the number of samples per run for this assay to retain diagnostic use on the smaller platform.

Impact of GC Content on Bait Variability

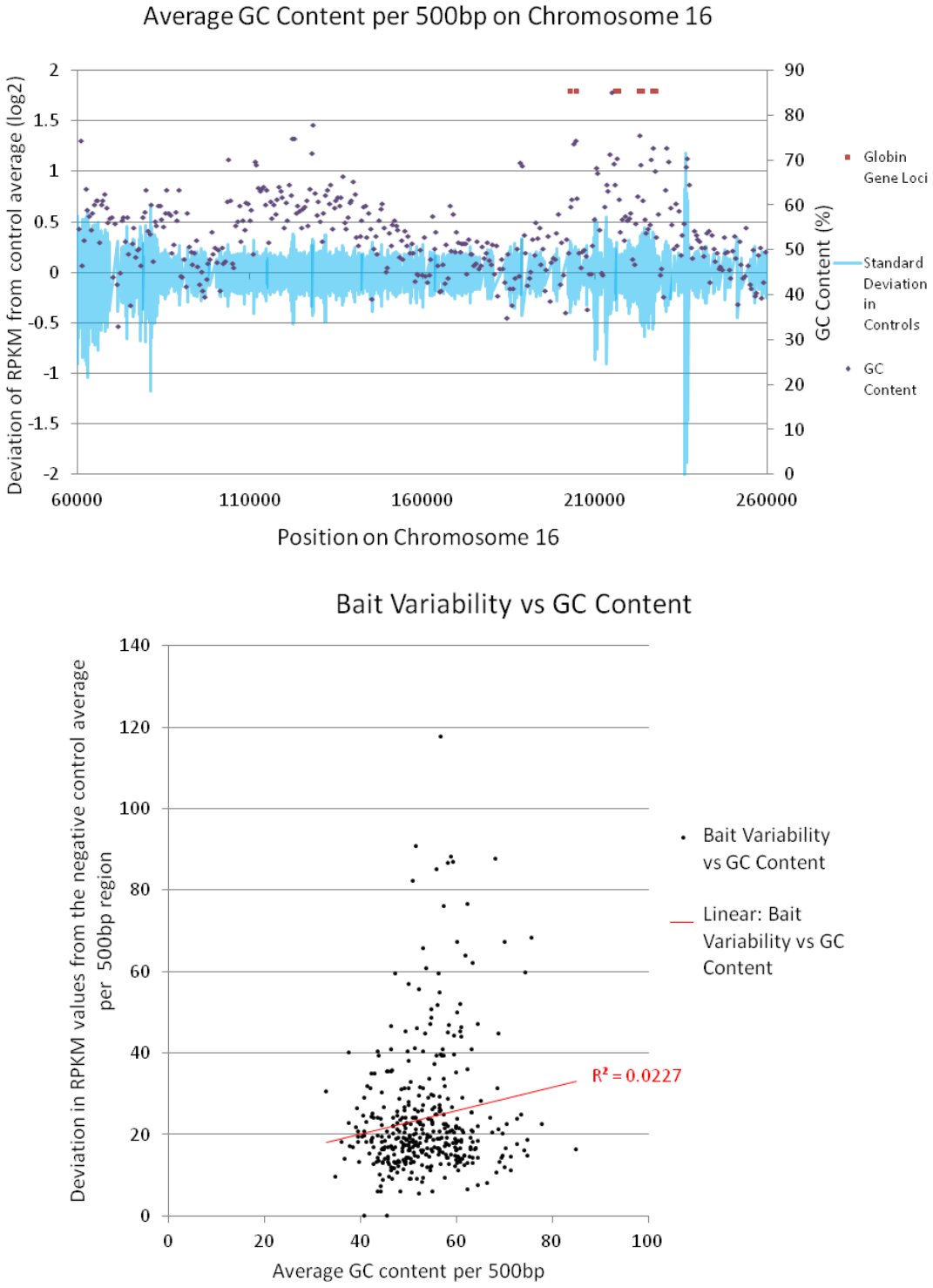


Figure 42: Impact of GC Content on Bait Variability. Top: The GC content calculated per 500 bp of the bait tiled region vs. standard deviation in bait position RPKM values between negative controls. Bottom: Correlation between amount of standard deviation from the average RPKM value per 500 bp segment of the covered region on chromosome 16, and the GC content of each segment.

Genomic Features of the Region of High Coverage Variability on Chromosome 16

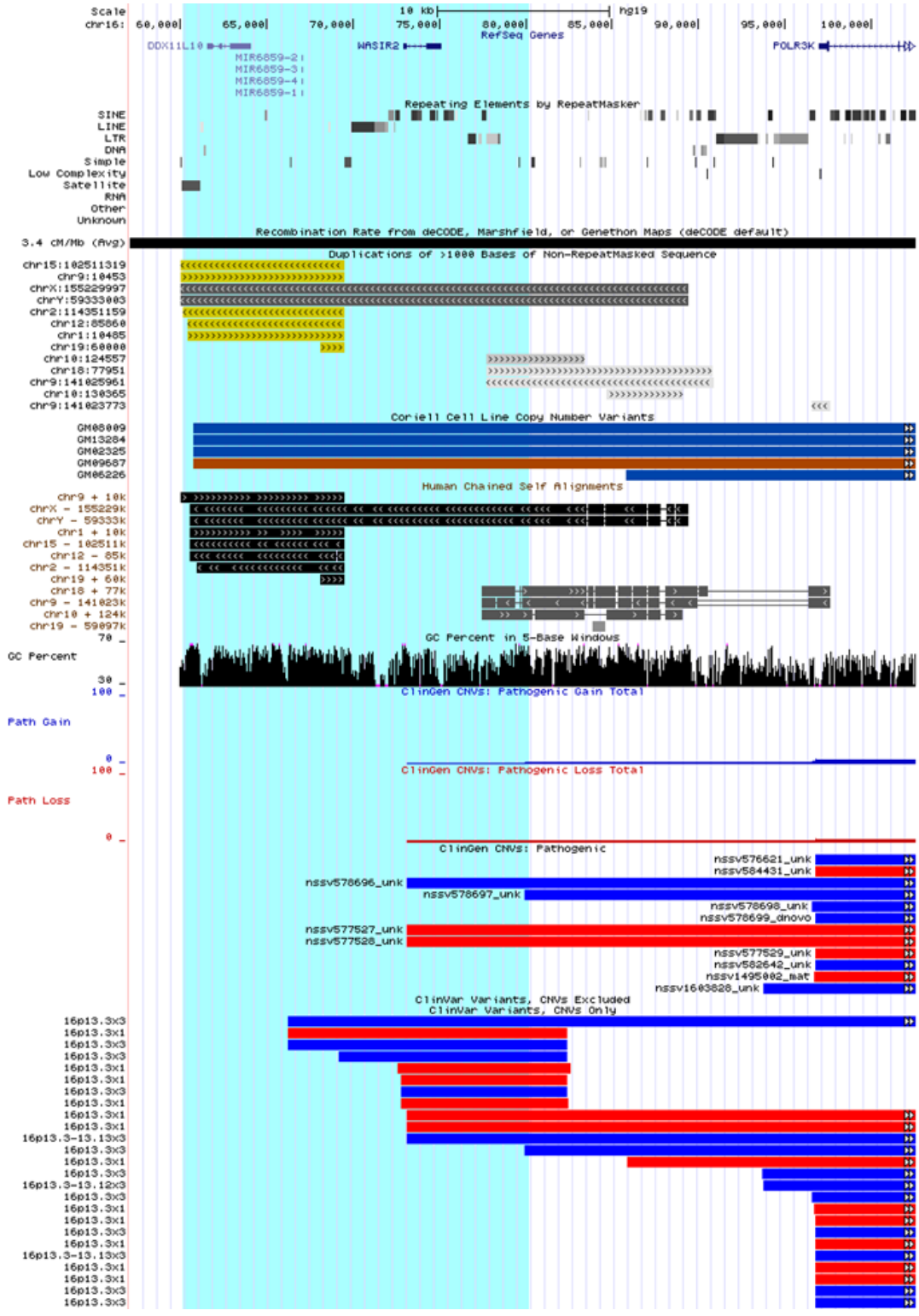


Figure 43: Genomic Features of the Region of High Coverage Variability on Chromosome 16 16:60,000-100,000. The region that produces highly variable coverage during sequencing is highlighted in blue.

A major advantage of this assay over CGH array or MLPA is that it is capable of detecting inverted sequence. In the case of the positive control Asian Indian inversion deletion, this was accomplished with to-the-base accuracy. This technique also has the

potential to detect insertions and translocations, however none of the samples in this cohort have either form of rearrangement. It should be considered in further evaluations of this assay that while insertions could be theoretically detected, it would not be possible to determine their size if the inserted sequence did not correspond to any sequences included in the bait capture library.

MiSeq Runs 1 and 2

To investigate whether moving to the MiSeq platform would improve the capability of this assay to detect structural variants, two MiSeq sequencing runs were performed (off-site). These runs contained a smaller number of samples to account for the lower sequencing capacity of this instrument than the HiSeq 2000.

Sample Details

DNA samples were sequenced on the MiSeq in batches of four (MiSeq Run 1) and five (MiSeq Run 2), using 2x 250 bp read kits. Each run contained a positive control; a negative control and at least one test sample (see Table 31). In MiSeq Run 2, DNA from the same patient was erroneously prepared twice, both as MiSeq Sample 7 and MiSeq Sample 9. The second preparation of this DNA (MiSeq Sample 9) was excluded from variant analysis. MiSeq runs 1 and 2 focussed on characterising variants affecting the beta globin gene locus on chromosome 11.

Table 31: Sample Details for MiSeq Run 1 and 2.

MiSeq Run 1		
Sample	Beta Globin Locus	Alpha Globin Locus
Sample 1	Positive control (-^{HPFH1}/)	Negative Control
Sample 2	Test (deletion)	Negative Control
Sample 3	Test (deletion)	Negative Control
Sample 4	Negative Control	Negative Control
MiSeq Run 2		
Sample	Beta Globin Loci	Alpha Globin Loci
Sample 5	Positive control (-$^{619bp\ deletion}$/)	Negative Control
Sample 6	Test (deletion)	Negative Control
Sample 7	Test (duplication)	Negative Control
Sample 8	Negative Control	Positive Control (α-$^{3.7}$/)
Sample 9	Duplicate of Sample 7	Duplicate of Sample 7

Sample Preparation

Samples were prepared in accordance with the protocol (See methods). An issue with the previous run was that the read size on the HiSeq 2000 was not long enough to accurately align reads containing long mismatches where they covered rearrangement break points. Another issue was that the small read size and small overall DNA fragment size impaired alignment to repetitive regions, which were frequently the sites of rearrangement breakpoints that were subsequently not captured. MiSeq sequencing kits allow a longer read length than available on the HiSeq (2x150, 2x250, 2x300). To accommodate this, a larger starting fragment size was used. This was achieved by shearing the DNA samples using a Bioruptor (as a Covaris was not available on site) to approximately 500 bp (see methods). MiSeq Run 1 was prepared successfully; producing libraries with a peak fragment size of approximately 650 bp (Figure 44). Five samples were prepared for sequencing, but only the highest quality four preparations were included in this run. Sample 4 from this run was later included in MiSeq Run 2 as Sample 5. Due to an error, this sample was then prepared again and was included in MiSeq Run 2 twice (as Sample 7 and Sample 9). Sample 7 showed better data and was used for analysis, while the duplicate, sample 9 was excluded from variant analysis but used for comparing inter-run variability (see later section). Sample preparation took place in Dr Chris Shaw's laboratory at the Institute of Psychiatry, Denmark Hill Campus, King's College London. Dr Athina Gkazi gave assistance and advice with the sample preparation process.

Bioanalyser Traces of Samples Prepared for Sequencing in MiSeq Run 1

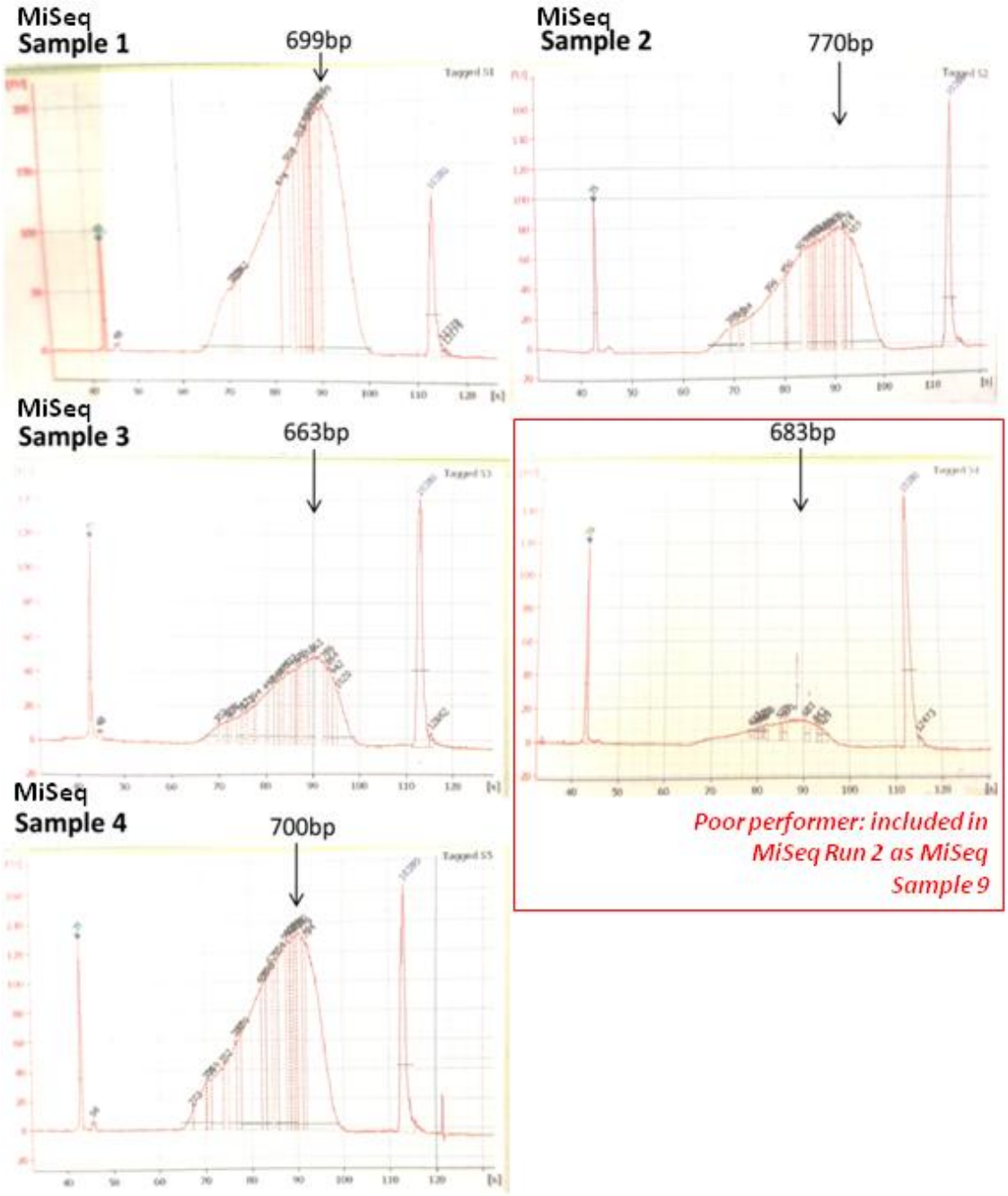


Figure 44: Bioanalyser Traces of Samples Prepared for Sequencing in MiSeq Run 1. Electropherograms show fragment size distribution in each sample following post-hybridization indexing PCR and clean-up.

Bioanalyser Traces of Samples Prepared for Sequencing in MiSeq Run 2

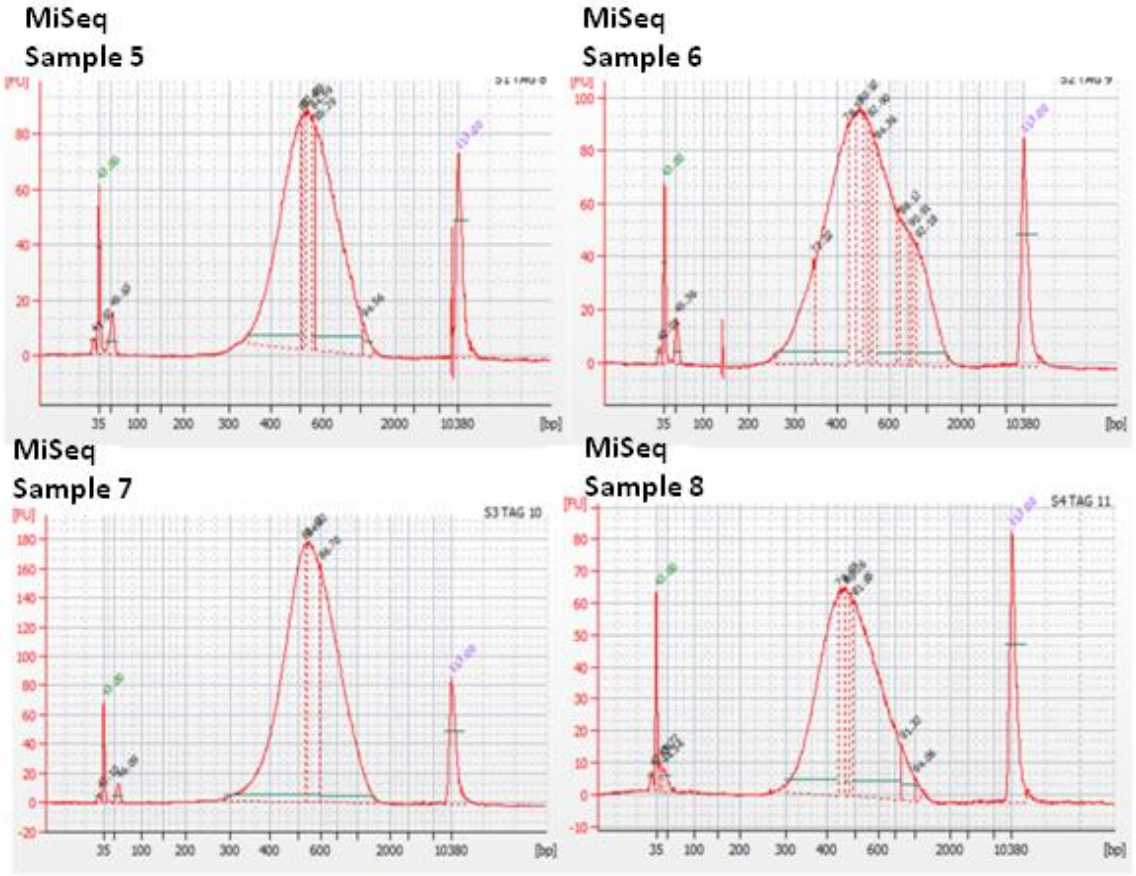


Figure 45: Bioanalyser Traces of Samples Prepared for Sequencing in MiSeq Run 2 Samples 5-8 (NB: Sample 9 was prepared in previous batch). Upper panel shows gel image and lower panel shows electropherogram trace. Each sample shows a size distribution of 300-1,000 bp, with a peak fragment size of around 500 bp. Sample 6 shows a slight hump on the peak which is indicative of over amplification during library prep.

Sequencing

Samples were sequenced on the MiSeq using the v2 2x250 reagent kit. MiSeq Run 1 had been slightly overloaded and the cluster density was higher than recommended for this sequencing kit (Table 32). MiSeq Run 2 achieved acceptable cluster density (Table 33). In both runs the quality scores of the bases in R2 tail off towards in the last few bases of the read (Figure 47). This is a known issue with the long read lengths of the v2 2x250 kits. The quality drop is more pronounced in MiSeq Run 1 read 2. This is likely to be due to the overloading of the flow cell in this run, meaning that the amount of reagents remaining was a limiting factor on sequencing quality in the second half of the . All indexed samplesamples were represented evenly in both runs (with no one sample over or underrepresented by >10% of total reads), indicating that pooling had been carried out correctly (Figure 46).

Reads Sequenced per Index in MiSeq Run 1 and MiSeq Run 2

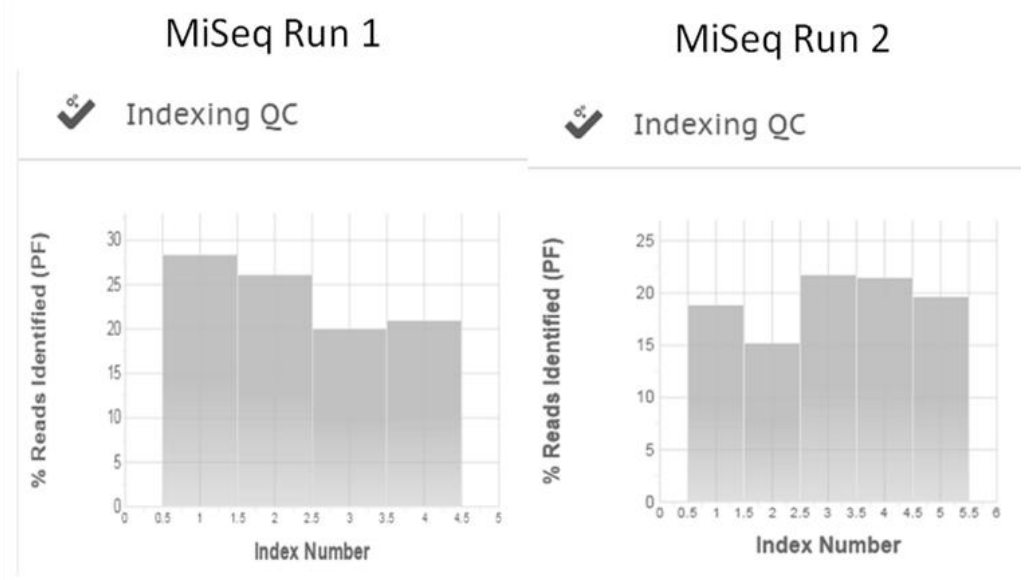


Figure 46: **Reads sequenced per index in MiSeq Run 1 and MiSeq Run 2**.

Table 32 Sequencing Statistics MiSeq Run 1.

MiSeq Run 1

Level	Yield Total (G)	Projected Total Yield (G)	Aligned (%)	Error Rate (%)	Intensity Cycle 1	% >= Q30
Read 1	6.24	6.24	0.32	0.83	185	88.07
Read 2	0.12	0.12	0	0	568	86.88
Read 3	6.24	6.24	0.3	1.49	151	76.6
Total	12.6	12.6	0.31	1.16	301	82.28

Read 1

Lane	Tiles	Density (K/mm2)	Clusters PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield(G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1
1	28	1773 +/- 19	79.23 +/- 1.38	0.065 / 0.128	31.48	24.94	88.07	6.24	250	0.32 +/- 0.01	0.83 +/- 0.07	0.14 +/- 0.01	0.20 +/- 0.02	0.24 +/- 0.02	185 +/- 18
Read 2															
1	28	1773 +/- 19	79.23 +/- 1.38	1.311 / 0.000	31.48	24.94	86.88	0.12	0	0	0	0	0	0	568 +/- 60
Read 3															
1	28	1773 +/- 19	79.23 +/- 1.38	0.088 / 0.130	31.48	24.94	76.6	6.24	250	0.30 +/- 0.01	1.49 +/- 0.14	0.43 +/- 0.24	0.50 +/- 0.11	0.55 +/- 0.09	151 +/- 12

Table 33 Sequencing Statistics, MiSeq Run 2.

MiSeq Run 2

Level	Yield Total (G)	Projected Total Yield (G)	Aligned (%)	Error Rate (%)	Intensity Cycle 1	% >= Q30
Read 1	4.63	4.63	0.74	0.85	199	94.52
Read 2	0.09	0.09	0.00	0.00	599	94.37
Read 3	4.63	4.63	0.72	1.01	168	87.76
Total	9.35	9.35	0.73	0.93	322	91.17

Read 1

Lane	Tiles	Density (K/mm2)	Clusters PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield(G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1
1	28	1035 +/- 17	93.15 +/- 0.17	0.062 +/- 0.075	19.88	18.52	94.52	4.63	250	0.74 +/- 0.01	0.84 +/- 0.07	0.09 +/- 0.02	0.12 +/- 0.02	0.16 +/- 0.03	199 +/- 12

Read 2

Lane	Tiles	Density (K/mm2)	Clusters PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield(G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1
1	28	1035 +/- 17	93.15 +/- 0.017	0.574 +/- 0.00	19.88	18.5	94.3	0.09	0	0	0	0	0	0	5.99 +/- 37

Read 3

Lane	Tiles	Density (K/mm2)	Clusters PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield(G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1
1	28	1035 +/- 17	93.15 +/- 0.17	0.08 +/- 0.072	19.88	18.52	87.76	4.63	250	0.72 +/- 0.01	1.01 +/- 0.07	0.14 +/- 0.01	0.2 +/- 0.02	0.24 +/- 0.03	168 +/- 9

Format Conversion

After sequencing, s were converted from FASTQ format to FASTA format using NextGene. The majority of reads passed format conversion successfully (

Table 34). The majority of reads that were not converted were rejected based on their median score. A minority of reads were removed when they exceed the limit for bases called as ‘N’, or when insufficient bases were called as A,G,C or T.

Table 34: Success of Format Conversion. Average for MiSeq Run 1 and MiSeq Run 2

	MiSeq Run 1				MiSeq Run 2			
	Read 1		Read 2		Read 1		Read 2	
	Average	Std Dev	Average	Std Dev	Average	Std Dev	Average	Std Dev
Total Reads in the Input File	5953839	870731	5953839	870731	3495808	436384	3495808	436384
Reads Converted Successfully	5731766	711602	5558507	968676	3472168	433353	3356935	416783
Reads Failed to Convert	222073	222710	395332	168740	23639	3158	138873	21068
Reads Filtered by "Median Score"	165511	196443	322591	155326	15261	2034	114519	16626
Reads Filtered by "Uncalled Bases" (read exceeds acceptable number of bases called as 'N')	5338	602	5203	827	1892	251	3085	437
Reads Filtered by "Called Base Number in Read" (read does not have enough bases not called as 'N')	14	16	25	16	15	6	15	4
Reads Filtered After Trimming	51210	26148	67513	14429	6472	896	21254	4082
Reads Trimmed	2524390	1274577	3287320	705121	644800	124514	1685649	178406
Reads Trimmed by "Quality Score"	2524390	1274577	3287320	705121	644800	124514	1685649	178406
Trimmed Bases	181504593	82907587	227412362	42218735	37149047	6751485	83141061	10750576

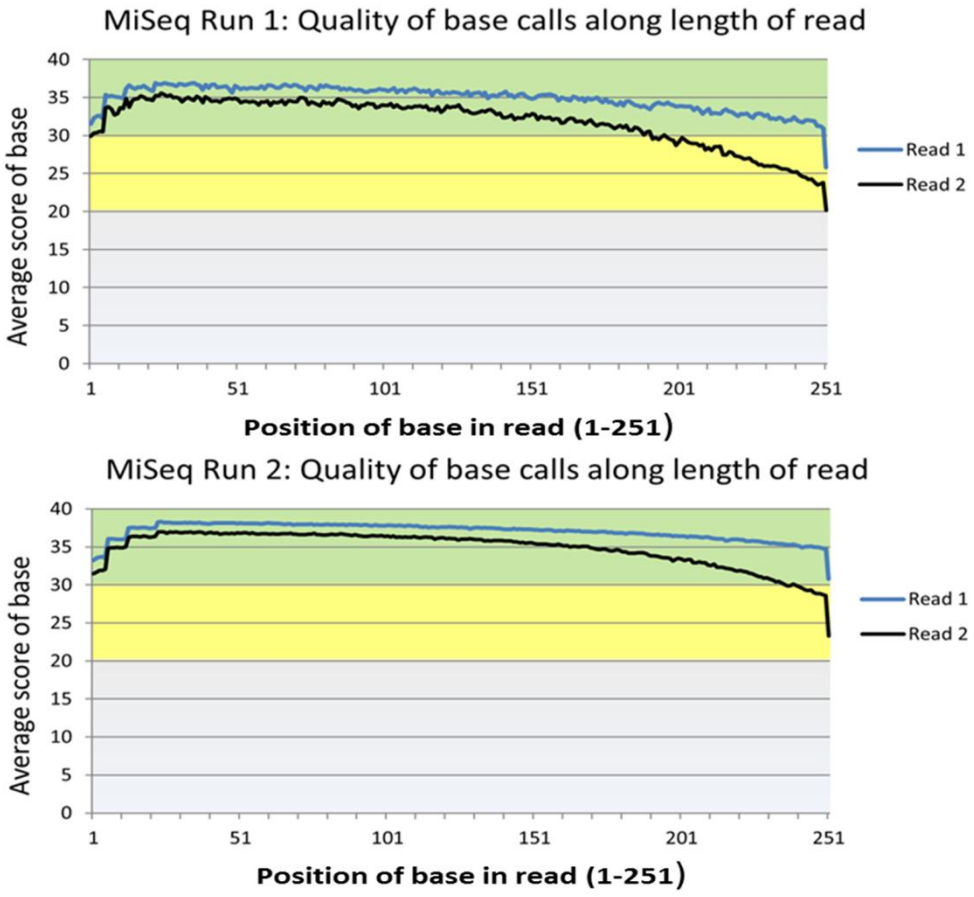


Figure 47: MiSeq Run 1 and 2: Quality of base calls along length of read. A score >30 is deemed 'good'; <30 and >20 is deemed 'acceptable' and <20 is deemed 'poor'.

Sequence Alignment

The majority of reads in both runs aligned successfully to the reference sequence (Table 35). In Run 1 37% of reads were on target, yielding average coverage of the ROI of 200x. In MiSeq Run 2 40% of reads were on target, and average coverage of the ROI was 180x. The reduced coverage per sample in Run 2 is contributed to by the inclusion of five samples in the batch, reducing the total reads produced for each sample.

Table 35: Alignment results for MiSeq Run 1 and MiSeq Run 2.

	MiSeq Run 1				MiSeq Run 2			
	Sample 1	Sample 2	Sample 3	Normal Control	Sample 1	Sample 2	Sample 3	Normal Control
Perfectly Matched Reads Count	4571292	3561360	3863200	5250276	4071571	3707711	2572340	3385457
Matched Reads Count	10674608	8163557	8737009	11871294	7521900	7247949	4893438	6392714
Unmatched Reads Count	1326794	951864	1079785	1522447	178168	269565	519708	194940
Short Reads Count	471414	378271	361358	502197	142370	128166	95146	130681
Number of Unmatched Bases That are Recorded as Mutations								
Mismatches	4459089	3191150	3238723	4459455	3271129	3215075	1886271	2790841
Deletions	790157	535502	538468	730572	635896	521192	323224	513688
Insertions	475713	335683	326272	451891	374969	321482	205237	288034
Number of Unmatched Bases That are NOT Recorded as Mutations								
Mismatches	28308228	22816832	24110993	33484832	9553211	12748729	10595859	9157927
Deletions	3367801	2726893	2975405	4172822	871909	1616346	1645132	851666
Insertions	2273159	1837438	2006035	2811485	584816	1074854	1066043	576355
Average Read Length	238	238	238	237	248	246	240	247
Average Coverage (Genome)	12	11	9	9	9	10	9	10
Average Coverage (ROI)	221	165	173	238	174	184	184	180
% of Reads on Target	38	37	36	37	26	39	51	46

Coverage data

With only a single negative control in each run, and all the samples containing variants from the same chromosome 11, it was not possible calculate a negative control average RPKM value. Instead, potential variants were identified by a sustained deviation from the RPKM values of the single negative control of >0.5 on a log2 scale (Figure 48, Figure 49).

MiSeq Run 1 Coverage Graphs: Chromosome 11

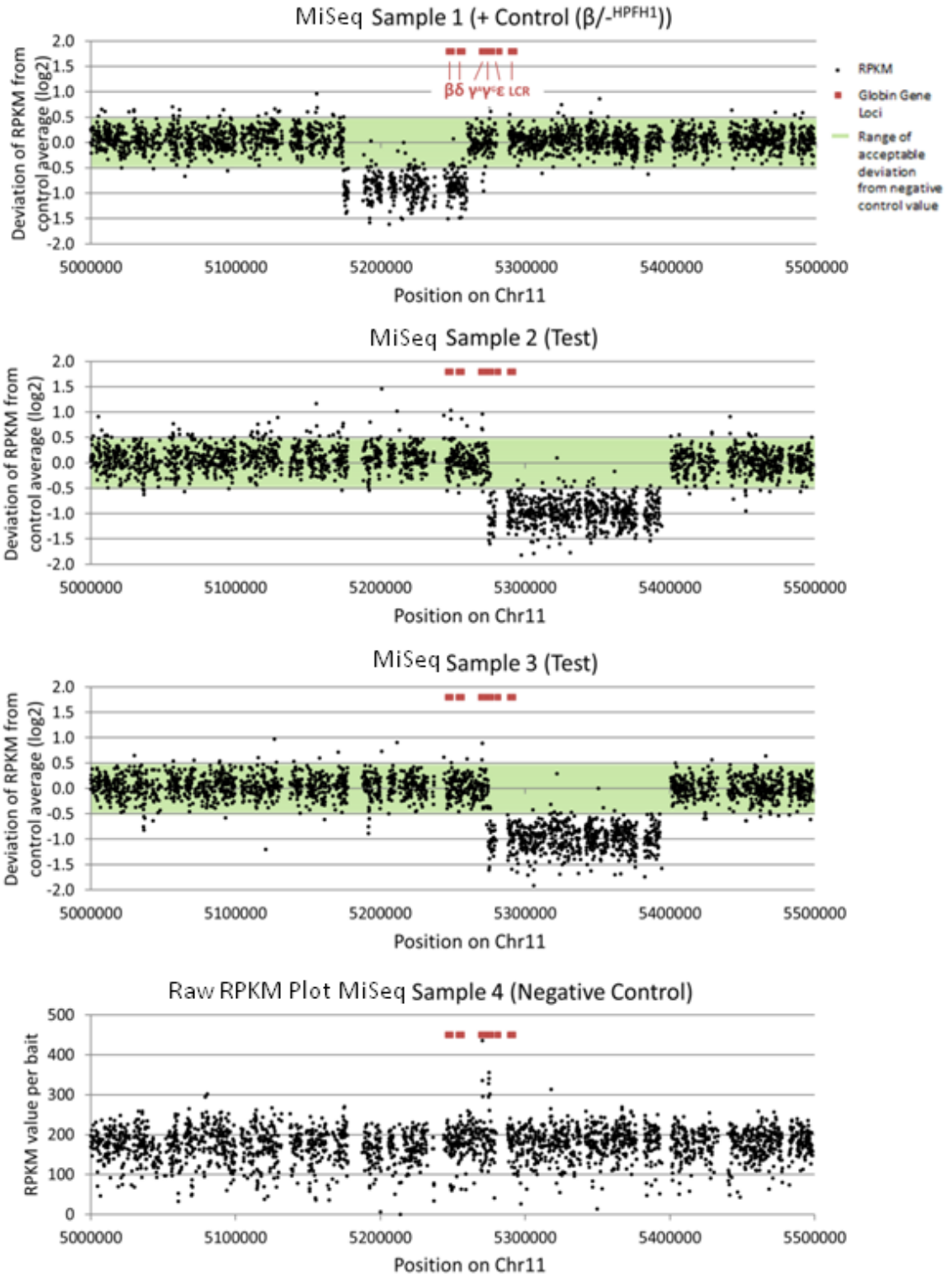


Figure 48: MiSeq Run 1 Coverage Graphs: Chromosome 11. RPKM plots for MiSeq Run 1 Samples 1-3, plus raw RPKM plot for Sample 4 (normal control). Green bar indicates region of acceptable (+/-0.5) variation from negative control value.

MiSeq Run 2 Coverage Graphs: Chromosome 11

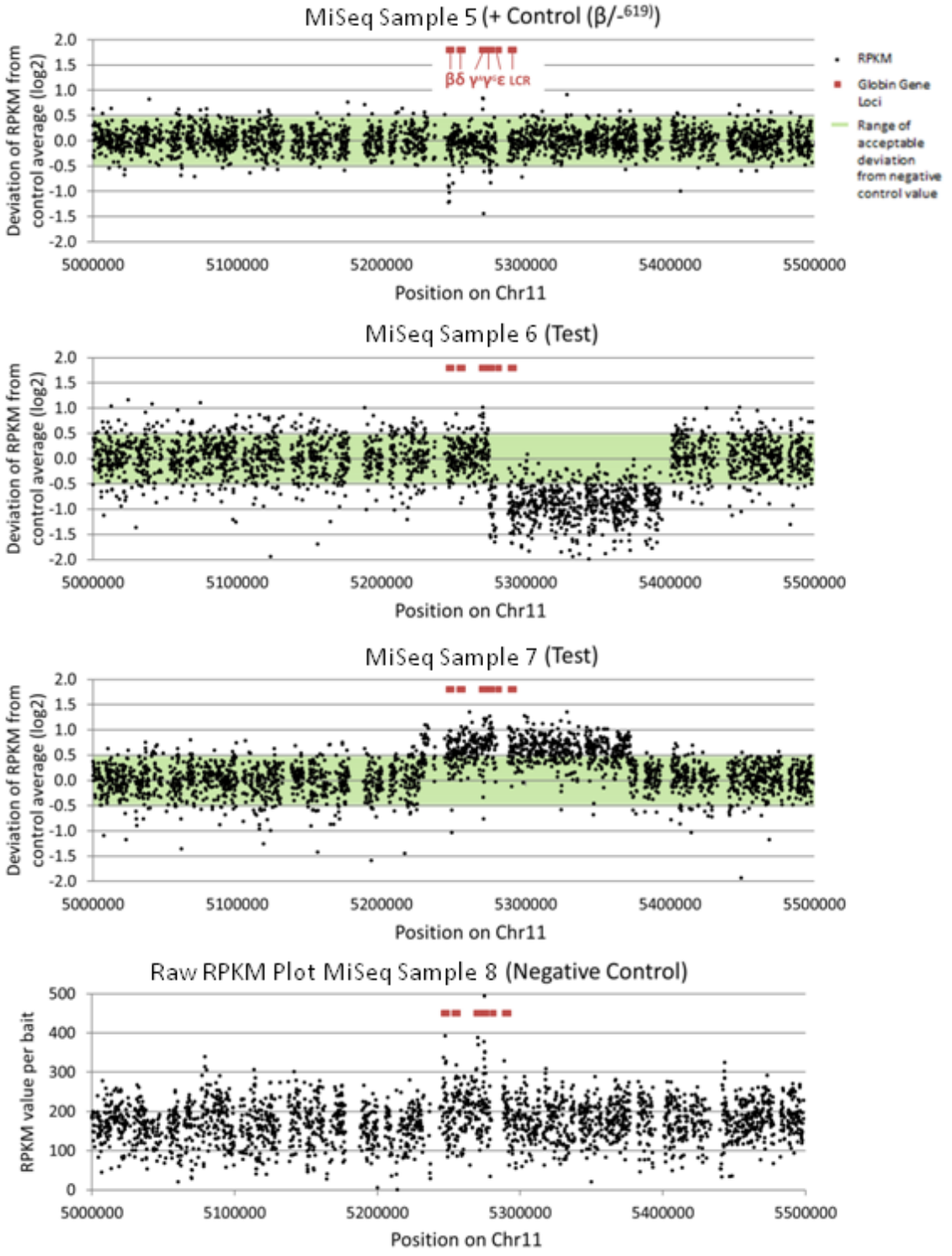


Figure 49: MiSeq Run 2 Coverage Graphs: Chromosome 11. Samples 5-7, plus raw RPKM plot for Sample 9 (normal control). Green bar indicates region of acceptable (+/-0.5) variation from negative control value.

Variant Characterisation: MiSeq Run 1, Sample 1 (Figure 48.1)

MiSeq Sample 1 (Figure 48.1) was a positive control for the HPFH-1 (HbVar.1021), a deletion of 85 Kb between co-ordinates chr11: 5,174,451 - 5,259,369 removing *HBB*

and *HBD*. The variant is clearly visible on the RPKM plot, estimated at between chr11: 5,174,346 (+/-240 bp) – and 5,259,047 – 5,259,874 (where a repeat of 830 bp is not covered by the bait design). Inspection of the 5' breakpoint region showed multiple reads aligning to the reference sequence at position ~5,259,360 that exhibited the same mismatched string of bases. Query of reads containing this string in the FASTA data files showed that they spanned the deletion breakpoint, aligning partially at both the 5' and 3' breakpoints of the deletion as listed in HbVar. Despite the gap in the bait design across this region, the long DNA fragment length and read length allow alignment of multiple sequences to this region (Figure 50), and to-the-base characterisation for this variant.

Detection of the HPFH1 Deletion in Sequencing Data

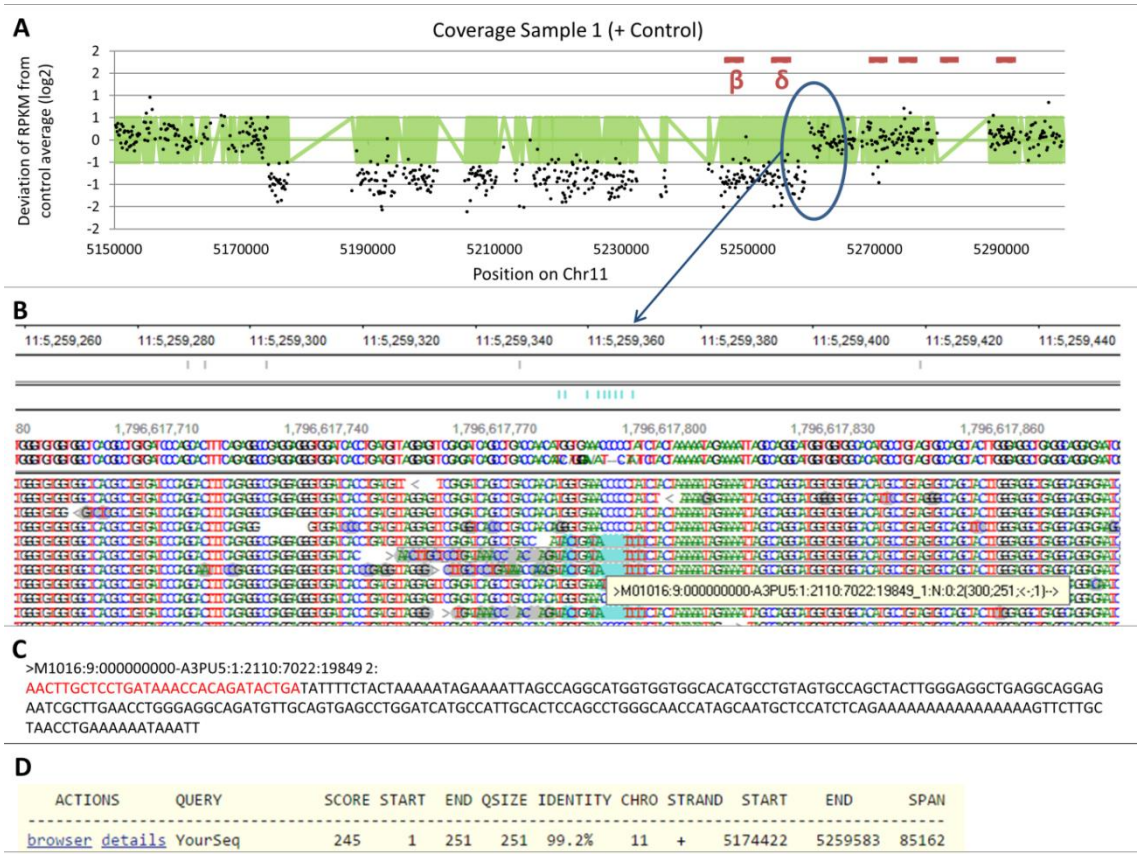


Figure 50: Detection of the HPFH1 Deletion in Sequencing Data. (A) RPKM plot showing deleted region relative to globin gene positions. (B) The read pile up at the 3' break point region (circled in blue) reveals breakpoint reads. (C&D) BLAT query of the original sequences of these reads reveal deletion breakpoints.

Variant Characterisation: MiSeq Samples 2 and 3 (Figure 48.2, Figure 48.3)

MiSeq Samples 2 and 3 (Figure 48) were obtained on separate occasions from two unrelated individuals. Neither variant could be resolved via Gap-PCR, although a large deletion removing the LCR was clearly visible by MLPA. NGS sequencing revealed that the same region was deleted in both samples, removing approximately 126 Kb of

sequence including the last exon of *HBG2*, *HBE* and the LCR (Figure 51A). The 5' breakpoint of the deletion was obscured by a 6 Kb LINE repeat which was not included in the bait design. Multiple reads aligning to the region of the 3' break point showed strings of mismatched sequences (Figure 51B). The deletion was similar to that of HiSeq Samples 5 and 10, in whom the break-points of the variant could not be resolved. BLAT query of these sequences revealed that the first portion of the read matched the reference sequence up to position chr11: 5,274,684 within *HBG2* (and also partially matched a region highly homologous to this within the *HBG1* gene) (Figure 51C/D). The second portion of the read matched the reference sequence 58 Kb downstream between positions 5,215,647 and 5,215,692 on the reverse strand, and also aligned with high homology to an overlapping sequence (5,215,689 - 5,215,734) on the positive strand. The same 3' breakpoint had also been identified in HiSeq data, but the shorter read length had meant that this highly homologous match downstream on chr11 was obscured by the many other high scoring matches to this common repetitive sequence.

Characterisation of a Novel Variant in MiSeq Samples 2 and 3

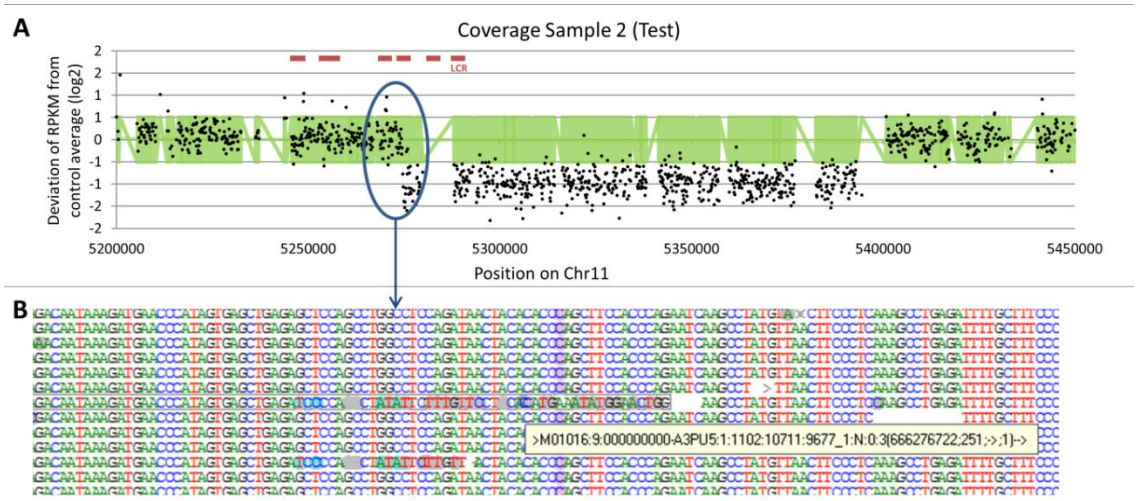


Figure 51: Characterisation of a Novel Variant in MiSeq Run 1 Samples 2 and 3. (A) RPKM plot showing deletion relative to positions of globin genes (B) Reads containing breakpoint sequence identified at 5' breakpoint (C&D) BLAT query of the original sequences of these reads reveal multiple matches to the reference sequence including inverted break point positions.

Overlaying the opposite direction reads over the RPKM data did not reveal any pile up that could indicate the deletion break points. A single same direction read aligned between the two positions identified by BLAT of the misaligned sequences described above which suggested that an inversion may have occurred between these two points. Close examination of the sequence at position 5,215,647 – 5,215,692 revealed that the region was a 159 bp reverse-compliment palindrome (meaning that the sequence on the positive strand read 5'-3' was identical to that of the negative strand when read in from its 5'-3') (Figure 52). It also appeared that the bait covering this region – which was an orphan situated between two repeats – was not functioning correctly: <5 reads aligned to this position in any of the sequenced samples from the MiSeq or HiSeq runs. This is likely to be related to the palindromic nature of the sequence which may cause the bait itself or DNA fragments containing this sequence to take on unusual physical conformations.

A Palindromic Sequence on Chromosome 11 Impeding Bait Performance

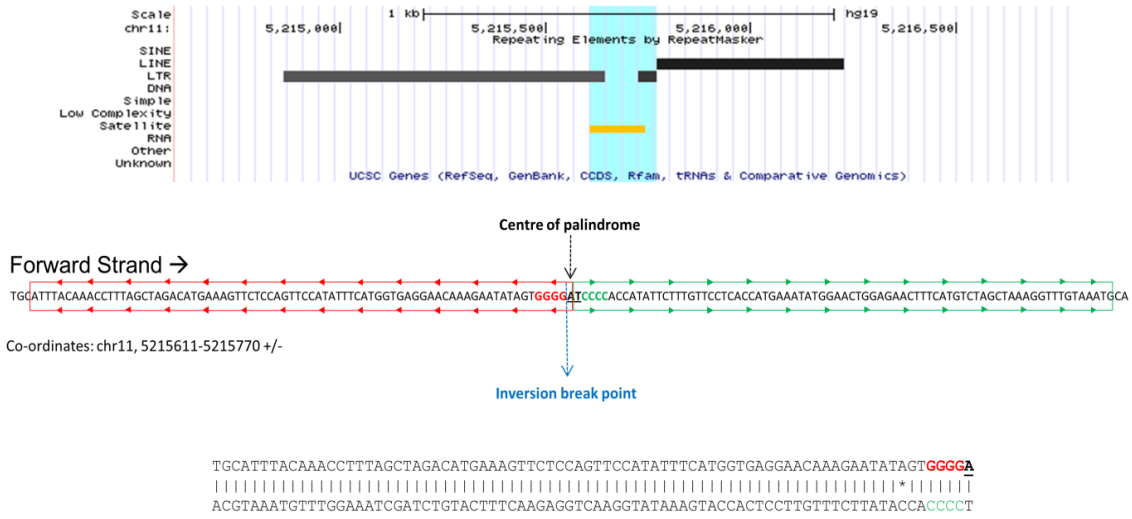


Figure 52: A Palindromic Sequence on Chromosome 11 That Impeded Bait Performance. Palindromic region at position 5,215,611-5,215,770. The nature of this sequence renders the bait that covers this position (marked in orange) non-functional

Confirmation of Novel Inversion-Deletion by Gap-PCR

Gap-PCR primers were designed to confirm the suspected inversion break points, and to identify the deletion breakpoints. A 1,031 bp product confirming the inversion sequence was successfully amplified using primers Pr1: 5'AGCTGGTTGGTCCGTTTTGG-3' and Pr3: 5'-CTCTGCATCATGGGCAGTGAG'3 (See methods). The product was dye-terminator sequenced on the ABI 3130 platform and the resulting chromatogram examined in Sequencher. The chromatogram showed that the variant breakpoints were 5,215,690 – 5,274,684 (Two bases at the break point position were expected in the normal sequence at both locations, so their origin was

ambiguous – the variant breakpoints could also be 5,215,692-5,274,682). Additional primers were designed between the inverted sequence (in the closest section of unique, non-repetitive sequence to the break point) and the unique, unaffected sequence region occurring immediately after the 6 Kb LINE repeat obscuring the 5' end of the deletion (Figure 53). The primer sequences were Pr2: 5'-AGTGCAAAGGATGCCAGGAC'3 and Pr4: 5'-GAGCAAGTGCATGCAAGGAGA-'3. A unique Gap-PCR product of 4,499 bp was successfully amplified using a long-range polymerase (See Methods). These revealed that the deletion breakpoints were 5,215,722 (inverted) – 5,397,195.

Breakpoint confirmation for novel variant in MiSeq Samples 2, 3 and 6 and HiSeq Sample 10

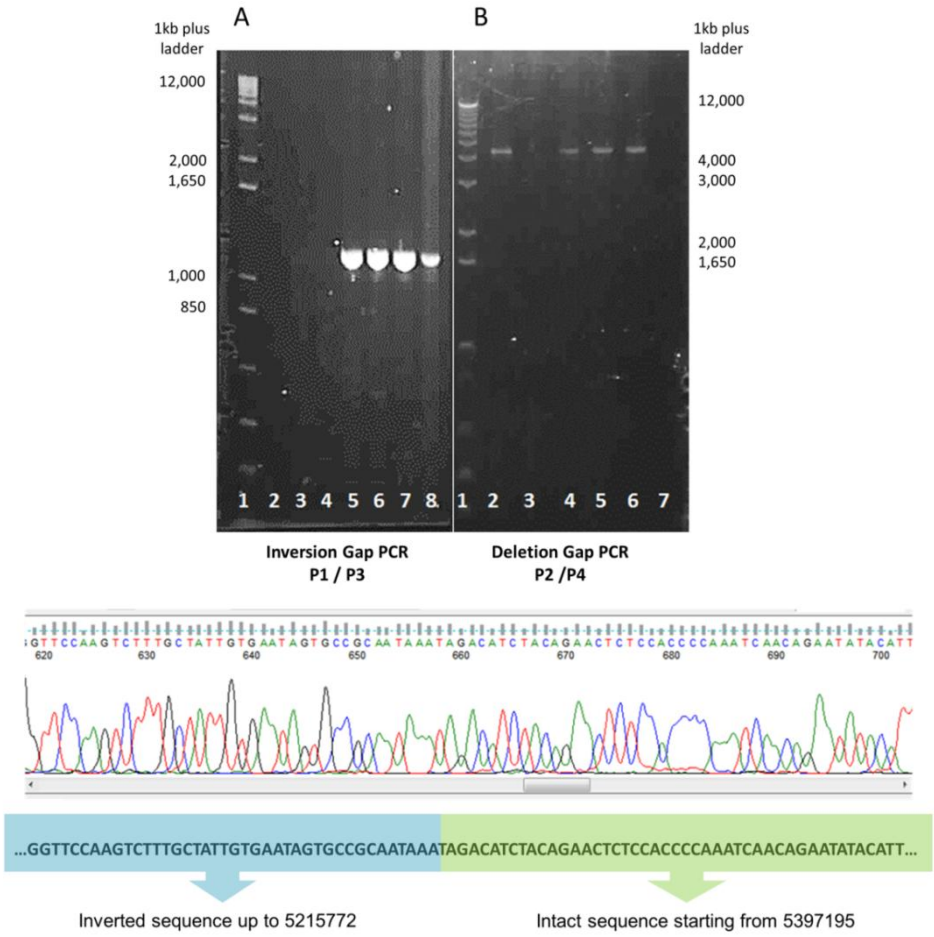


Figure 53: Breakpoint confirmation for novel variant in MiSeq Samples 2, 3 and 6 and HiSeq Sample 10. Gel images: (A) Inversion PCR product. Lanes 1-8: (1) 1Kb+ ladder (2) Blank (3) Negative Control (4) Negative Control (5) MiSeq Sample 2 (6) MiSeq Sample 3 (7) MiSeq Sample 6 (8) HiSeq Run 1 Sample 10. (B) Deletion PCR product Lanes 1-7 (1) 1Kb+ ladder (2) MiSeq Run 1 Sample 2 (3) negative control (4) MiSeq Run 1 Sample 3 (5) HiSeq Run 1 Sample 5 (6) HiSeq Run 1 Sample 10 (7) Blank. Chromatogram: Sequence analysis of the inversion Gap PCR product for MiSeq Run 1 Sample 2 showing inversion-deletion breakpoint.

An 82 bp section of the inverted sequence was also removed by the deletion, which was 122,593 bp in length. The inclusion of inverted sequence in the deletion indicates

that the inversion event happened before the deletion event. The rearrangement concatenated two normally distant, but highly homologous LINE repeats (Figure 54). This novel inversion-deletion was submitted to HbVar with the common name English V, ID 2935, and is described as: Chr11 Hg19 (build 37.3) g.5215690_5274684invdel5215690_5215772del5274684_5397195 (Shooter, Rooks et al. 2015).

Schematic of the Novel Rearrangement in MiSeq Samples 2, 3 and 6 and HiSeq Sample 10 (English V Inversion Deletion)

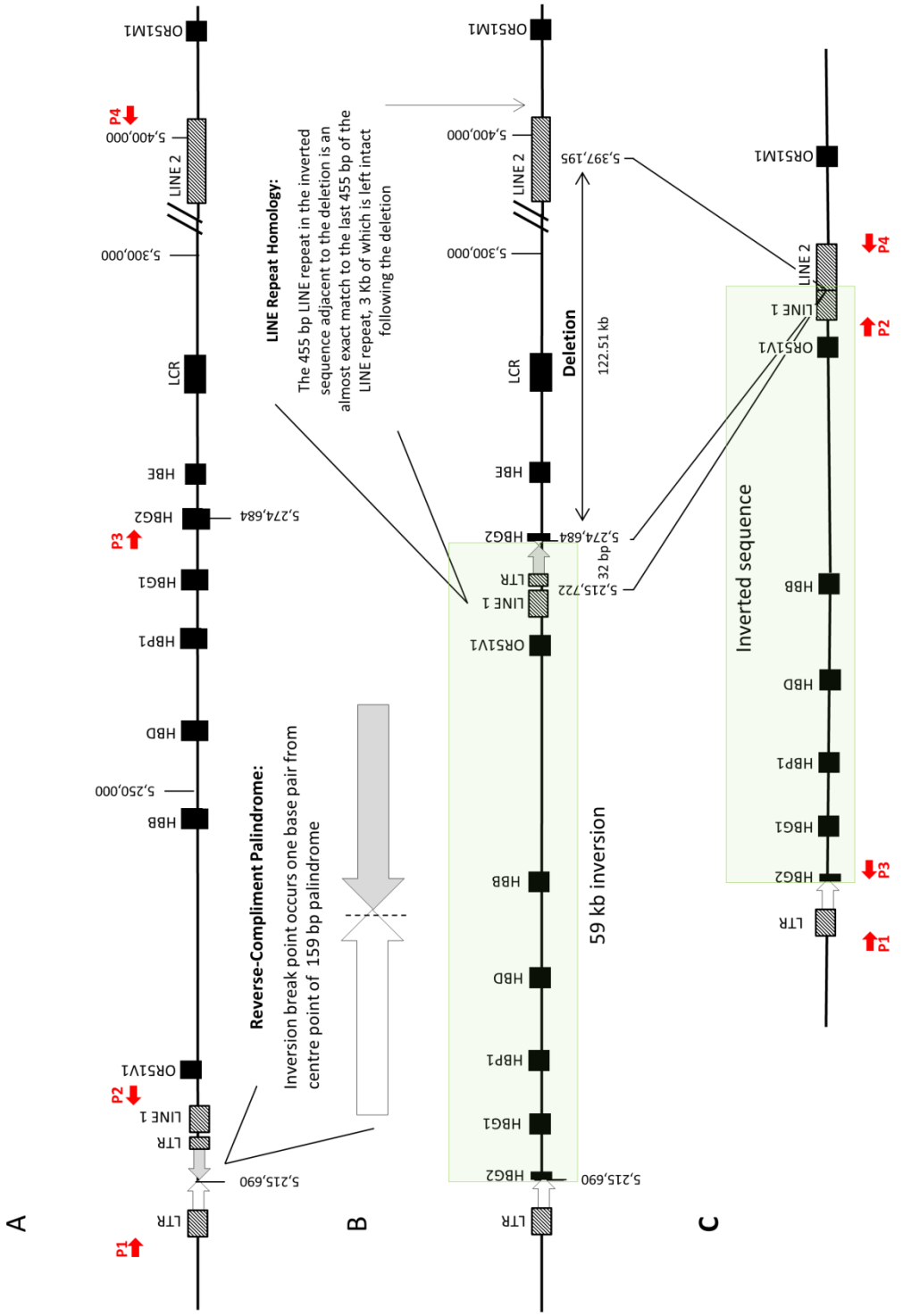


Figure 54: Schematic of the Novel Rearrangement in MiSeq Samples 2, 3 and 6 and HiSeq Sample 10 (English V Inversion Deletion). (A) normal layout of beta globin gene cluster and surrounding region of chromosome 11, plus positions of primers for Gap-PCR. (B) Inversion event (indicated by a light green box) occurs, followed by deletion removing 82 bp of the inverted sequence and 122,511 bp of upstream sequence (C) rearranged chromosome (figure from Shooter et al, 2014).

Variant Characterisation: MiSeq Sample 5 (Figure 49.1)

MiSeq Sample 5 is a positive control for the well-characterised 619 bp deletion (HbVar ID.979) which removes the sequence between chr11: 5,246,619 - 5,247,237, inserting 7 bp of novel sequence ('TCTACTT') at the deletion break point. The small size of the variant made it difficult to identify from the RPKM data, particularly as in this case the cut-off used was >+0.5/<-0.5 from the negative control, rather than the more accurate means of detection by comparison to the standard deviation of variation from the average of multiple negative controls. The deletion is more prominent when the RPKM data is viewed at a higher resolution (Figure 55). Investigation of the read pile-up at the expected breakpoint regions from the RPKM data showed large numbers of reads that had crossed the break point had aligned to the reference sequence. BLAT query of the sequences of reads containing these mismatched bases showed a split alignment between the expected break point positions, and also the 7 base pair insertion. The read pile-up also shows an additional pile up of multiple reads containing a different sequence of mismatched bases. All of the aligned sequences containing this novel sequence of bases were identical in length and ended with a similar string of novel bases. BLAT query of this sequence indicated that they were published PCR primer sequences, indicating that this sample had been contaminated by a PCR product.

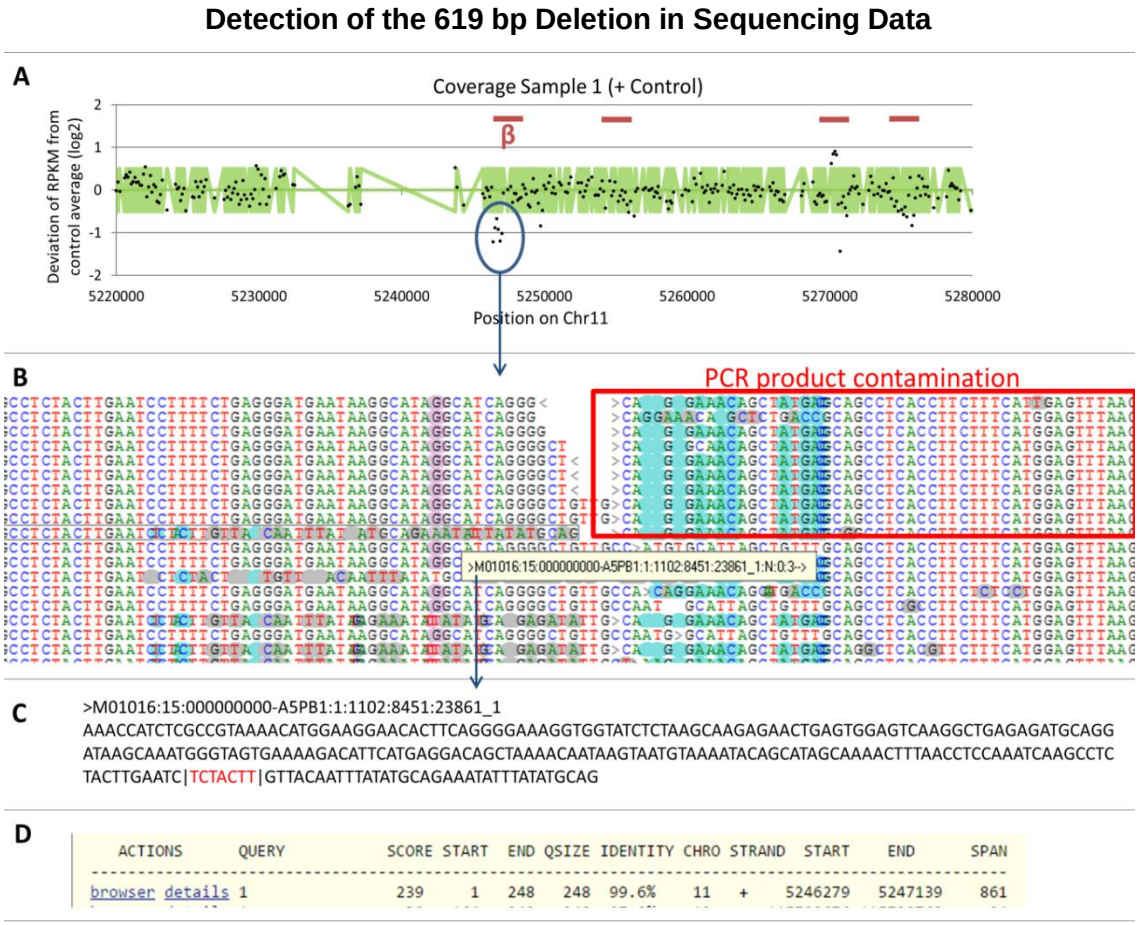


Figure 55: Detection of the 619 bp Deletion in Sequencing Data. (A) RPKM plot showing deletion relative to positions of globin genes (B) misaligning reads in the NextGene viewer at deletion breakpoint region (C&D) BLAT query of original FASTA sequences identifies breakpoint, including 7 bp insertion (indicated in red).

Identifying this PCR contaminant was important, as it shows the necessity for good laboratory practise and isolation of post-PCR products. As NGS is a high sensitivity technique, minor contaminations have the potential to influence how the results of sequencing are interpreted.

Variant Characterisation: MiSeq Sample 6 (Figure 49.2)

MiSeq Sample 6 was also identified to have the newly characterised English V deletion. This sample had previously been sequenced on the HiSeq 2000 (HiSeq Sample 5) in which the variant could not be resolved due to the shorter read length. Here, the variant could be resolved by BLAT query of reads aligning to the reference sequence at position chr11: 5,274,684 which crossed the inversion breakpoint. As with MiSeq Samples 2 and 3, no sequences aligned to the reference sequence at this position, as they had not been captured by the non-functional bait covering this region. The deletion breakpoints could once again only be resolved through Gap-PCR.

We investigated whether the rearrangement was on the same haplotype background in all, or whether this rearrangement had arisen independently in these people. Mutation reports for all three samples sequenced on the MiSeq were generated. Variant calls data from the HiSeq 2000 excluded from the analysis to increase the uniformity of the datasets being compared. The reports were filtered to the broad region of the rearrangement (chr11:4,500,000-6,000,000). Variants within the region deleted by this rearrangement (chr11:5,215,690-5,215,722 and 5,274,684-5,399,855) were excluded, as any variants here would belong to the unaffected chromosome.

For each sample, we selected variants that were (a) homozygous, and therefore definitely present on the rearranged allele, and (b) had a phred score of >15 (Table 36). We searched for these variants in the mutation reports of the other two samples to identify any variants that were unique to the rearranged allele in that sample.

The data showed that >98% of variants that could definitely be assigned to the rearranged allele were present in all three samples. This suggests that the haplotype background for the variant was the same in all three cases, and that it has a common point of origin.

Table 36 Comparison of variants *in cis* with the English V deletion in three samples

Variants in the mutation report	**MiSeq Sample 2**	**MiSeq Sample 3**	**MiSeq Sample 6**
Number of variants recorded in **region**	4818	4641	4749
Number of those variants with a **Phred score >15**	3497	3121	3315
Number of variants (Phred score >15) that were **homozygous** (and thus definitely present on rearranged allele)	1403	1408	1514
Number of variants (Phred score >15) homozygous in this sample **not listed in all three samples**	18 (1.2%)	19 (1.3%)	16 (1.1%)
Number of those variants (Phred score >15) homozygous **unique** to this sample	8 (0.6%)	6 (0.4%)	4 (0.3%)

Variant Characterisation: MiSeq Sample 3 (Figure 49.3)

MiSeq Sample 3 was known to be heterozygous for the (α-$^{3.7}$/) deletion on chromosome 16, for the sickle cell variant on chromosome 11, and heterozygous for a duplication on chromosome 11 which was resolved with to-the-base accuracy in HiSeq Run 1 (Sample 7). The novel duplication identified on the HiSeq 2000 was once again clearly visible in the RPKM data, and in the reads aligning to the break point regions in the NextGene viewer (Figure 56). The sickle cell mutation was listed in the mutation report with a mutant allele frequency of 31%.

Appearence of a duplication identified in HiSeq sample 7 when resequenced on the MiSeq platform (as MiSeq Sample 7)

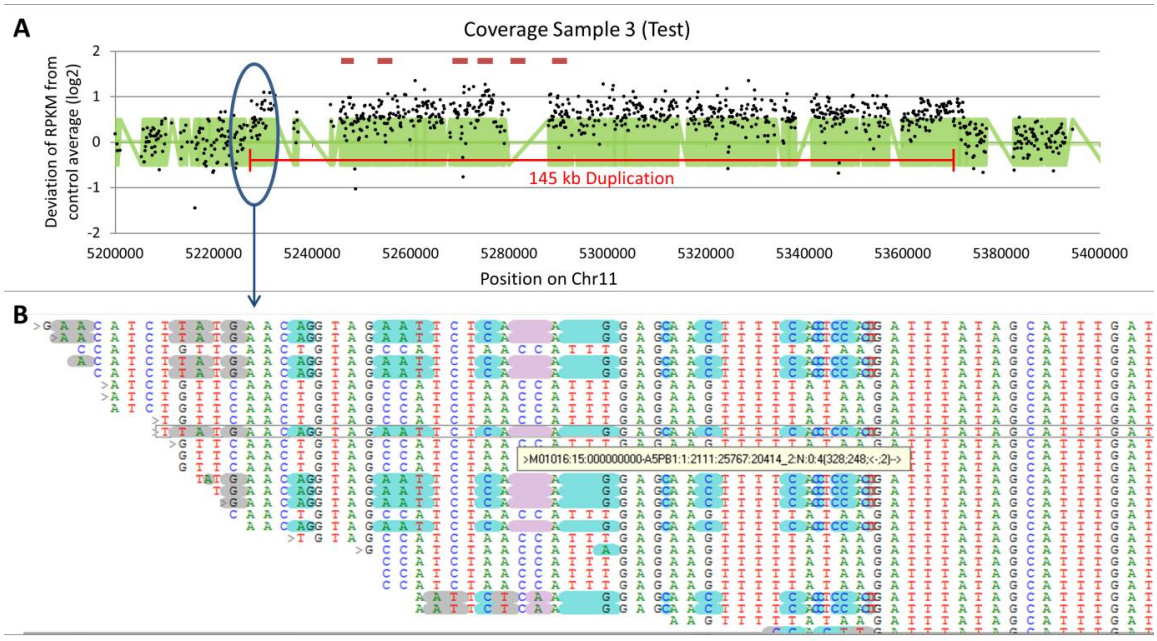


Figure 56: Appearence of a duplication identified in HiSeq sample 7 when resequenced on the MiSeq platform (as MiSeq Sample 7)**. (A) RPKM plot identifies novel duplication of approximately 145Kb encompassing entire globin gene cluster. (B) Read pile-up at approximate duplication start point includes multiple reads showing string of mismatched bases. (C&D) BLAT query of reads containing misaligned reads indicate they represent the break point of the duplication. Duplication includes insertion of 11 bp (red) creating 13 bp mirror repeat (underlined).**

The 3.7 Kb Deletion

Detecting the (α-$^{3.7}$/) deletion in this sample had been challenging using the HiSeq 2000 data, as the homology between the breakpoint regions prevented any opposite direction reads or break point sequences from being generated. We investigated whether the longer fragment and read lengths of the MiSeq improved the visibility of

this rearrangement during analysis. As this variant is located on chromosome 16, the two other samples in the Run (which both had variants on chromosome 11) could be used as additional negative controls for this locus.

Detection of the (α-$^{3.7}$/) Deletion in Sequencing Data

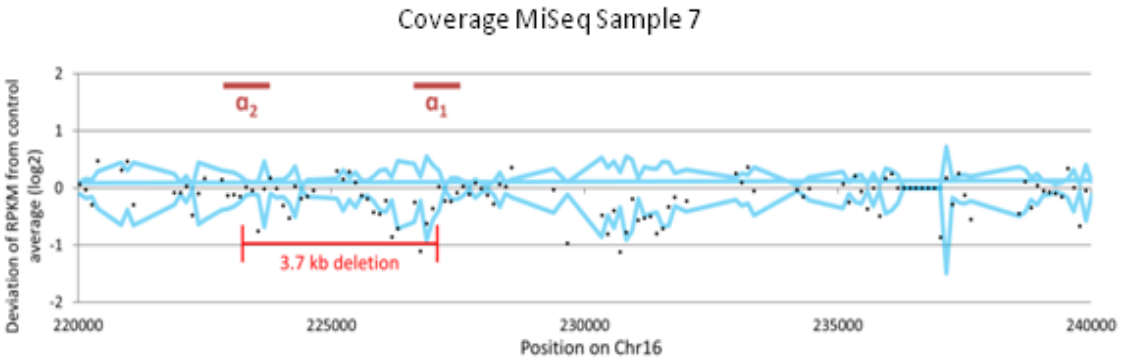


Figure 57: Detection of the (α-3.7/) Deletion in Sequencing Data. Blue line indicates standard deviation seen in samples known to be negative for structural variants at this locus. Expected position of 3.7 Kb deletion indicated by red bar.

The deletion is faintly visible in the RPKM data (Figure 57), but many of the bait positions in the affected region do not exceed the negative control standard deviation. This would make this variant difficult to detect in persons who were not previously known to carry the deletion. The variant is also not visible as misaligned reads in the sequencing data, or in the opposite direction reads report, despite both break points covered by the bait design. Alignment ambiguity affecting this region may also be responsible for the faint signal of this deletion in the RPKM data. This is a common variant, and detecting its presence can be diagnostically important. Issues with detecting this deletion will be discussed in detail in a later section.

Evaluation of MiSeq Vs HiSeq Sequencing Platforms

The HiSeq platform allowed sequencing of 12 samples simultaneously (in one lane) at an average read depth (across the bait-covered region) of 363x. Samples were sequenced on the MiSeq in smaller batches as the sequencing depth required to characterise variants had not been established. A run of four samples on the MiSeq achieved an average coverage of 387x across the bait tiled region. A run of five samples achieved an average coverage of 266x. The coverage provided by the HiSeq 2000 was sufficient to identify deletions and duplications in the RPKM data, however some variants could not be resolved with to-the-base accuracy. There were several reasons for this, as discussed earlier. One important factor was the read length (2x97 bp) which was not sufficient to align reads that contained a large amount of novel sequence to the reference. This problem was compounded by the fact that many of the

variant break points were within repetitive regions not included in the bait capture library. The small read length meant that it was challenging to align reads originating from these regions to the correct positions in the reference sequence. The bait capture design did not cover enough of chromosome 16 to capture both break points of several large rearrangements affecting the alpha globin gene loci. This meant that not only could the break point sequences not be resolved for these variants, their approximate size and number of genes affected could not be determined.

Sequencing on the MiSeq platform allowed longer read lengths of 2x250-2x300 bp. Larger DNA fragments could be sequenced for longer distances with this technique, and several variants affecting the beta globin gene cluster on chromosome 11 could be fully characterised, where this was not possible from HiSeq 2000 sequencing data. The 3.7 Kb deletions could not be reliably identified or resolved with to-the-base accuracy using either sequencing platform. The HiSeq 2000 showed the deletion with greater clarity in the RPKM plot, due to the greater number of controls for this region included in the run. Better alignment algorithms may be required to reliably detect this variant by assigning reads to *HBA1* and *HBA2* with greater accuracy. It was suggested that the changes to the bait library could improve the ability of this assay to detect rearrangements. As described earlier, this new design would have a higher tiling density across the regions most likely to be involved in structural rearrangements, cover a larger region of chromosome 16, and use newer bait tiling algorithms developed by Agilent to improve performance (See Methods).

On both platforms, bait performance varied between different samples and between different runs (See Figure 58). Some regions showed more inter-sample and inter-batch variation than others. This variation created noise that interfered with the detection of some rearrangements, including the 619 bp and 3.7Kb deletions. Figure 58 shows the variance between different samples, where any bait that show low variation in the RPKM values obtained from the two samples (<+/-0.5) is plotted in green, while baits showing significant variation (>+/-0,5) are plotted in red. Figure 58.**A** shows the same sample prepared twice on the HiSeq 2000 (Sample 11 and Sample 12). Only 2.2% of bait covered positions had significantly different RPKM values between the two identical samples. **B** Shows the variation between two sequencings of the same sample on different platforms: the HiSeq 2000 and the MiSeq. Significant variation is recorded between the two sequencing experiments at 20% of bait-covered positions. It should be noted that other factors, such as the different fragment and read sizes used, facility in which the samples were prepared, and level of experience of the researcher changed between these sequencing experiments in addition to the platform, so this is

not a particularly accurate measure of the difference between the platforms. **C** shows the difference in RPKM values for the same negative control sample prepared and sequenced separately for MiSeq Run 1 and 2, and the variation between these two runs was 9.8%. **D** Shows two identical iterations of a sample that were prepared separately and then included in the same sequencing run on the MiSeq (MiSeq Samples 7 and 9). Significantly different RPKM values between these two samples are recorded 4.7% of bait positions. This suggests that the sample preparation process accounts for approximately half the variability seen between different MiSeq runs. That the HiSeq platform showed least variability was unexpected. This may have been impacted by preparing these samples under close supervision from the experienced staff at Guy's Campus. MiSeq Run 1 and MiSeq Run 2 were prepared with considerably less guidance, and sample preparation during MiSeq Run 2 suffered numerous setbacks which were eventually determined to be due to the temperature of the laboratory in which preparation was carried out (the lab in which these samples were prepared rose to a temperature of 34°C in the summer which was eventually determined to cause severe evaporation of the ethanol used during wash stages).

These comparisons provide a rough idea of the variation between samples prepared in different batches and sequenced in different runs and on different platforms. This comparison was not planned in advance, so the comparisons made given the data available are not ideal. It appears that some variation occurs between different instances of sample preparation. The introduction of the automated sample preparation platform may reduce this variation, and increase the sensitivity of the assay.

Bait Variability on Chromosome 11 between Different Samples, Runs and Platforms

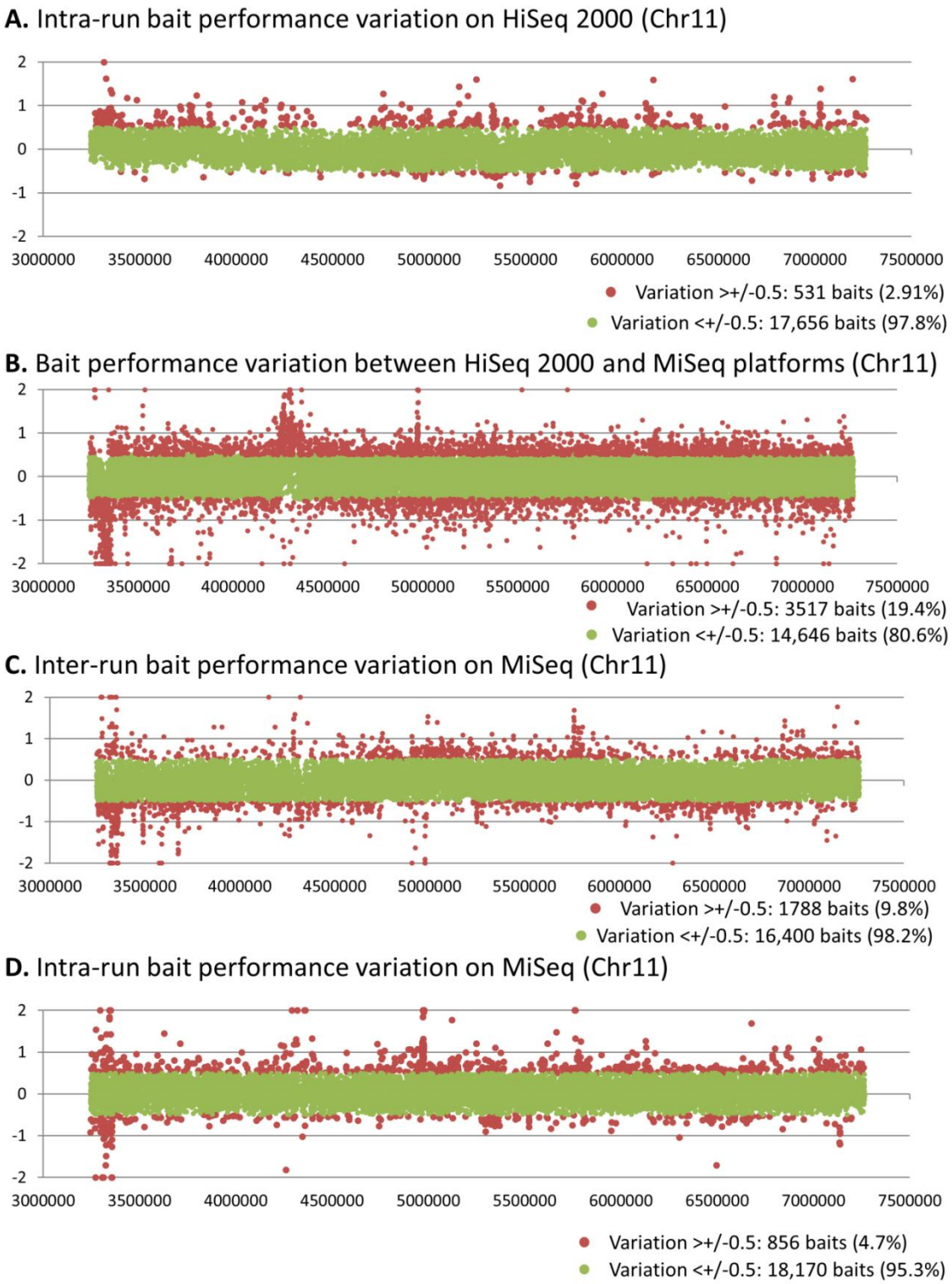


Figure 58: Bait Variability on Chromosome 11 between Different Samples, Runs and Platforms. A-C: Variation between different preparations of the same sample (this was not possible for comparison D). (A) Variation between instances of same sample run twice on HiSeq 2000 (HiSeq Samples 11 and 12). (B) Variation between the same sample sequenced on HiSeq and MiSeq platforms shows significant variation in 20% of baits. (C) Inter-run variation on the MiSeq platform between the same sample is 10% (D) intra-run variation on the MiSeq platform is 4.7%

The high inter-run variability on the MiSeq platform means than multiple negative controls should be included in each run to distinguish genuine rearrangements from random noise. The noise itself may be reduced by improving the uniformity of the sample preparation process, which can differ substantially between samples. This is demonstrated by the variability in success of amplification at the final stage of the process within MiSeq Run 1 (Figure 44). We expect this variability is the culmination of many minor differences in the treatment of samples at each stage of the library preparation process (i.e. concentration, quantity and purity of DNA sample; number of PCR cycles employed; exact hybridization conditions; quality of post hybridization washes). Inter-sample variability may additionally be improved by the introduction of the new bait capture design, with more intelligent bait placement and more baits covering regions expected to perform poorly.

Redesigning the Bait Capture Library

The process of redesigning the bait capture library is described in Methods. The performance of the new design was evaluated on the MiSeq. The key requirements of the new library were to:

- Reduce the amount of off-target sequence captured
- Extend the regions covered in the bait design to cover all variants affecting the globin gene loci up to the largest published rearrangement size
- Reduce the variability in bait performance

The new library, Bait Capture Library 2, was used for target enrichment in subsequent runs. The performance of the new library in comparison to Bait Capture Library 1 used for target enrichment in the sequencing runs performed in this chapter is evaluated in a later section. To reduce inter-batch and inter-sample variation that occurred due to differences in sample preparation conditions, a robotic platform was set up to automate the sample preparation procedure.

Automated Sample Preparation Using the BioMek FX^{p} Robotic Platform

Automating the library preparation procedure adds value to the diagnostic workflow: It means more samples can be processed in parallel, with a more uniform approach, which should result is less inter-sample variation. We believed that the hybridisation reaction during library prep contributed to a large portion of the batch to batch variation

due to the sensitivity of this step. Reducing this variability by automation would be helpful.

Issues with the performance of the robotic platform caused the failure of several s which would otherwise have been used to evaluate the robotic sample preparation method. The failing runs used the last of Bait Capture Library 1.

Initial issues with the robotic preparation platform included:

i) Errors with the program, in which tips were re-used at inappropriate times, causing cross-contamination of samples

ii) Errors with the hardware, (a) the robotic arms would fail to eject tips correctly and continue its operations with lower than expected clearance of the robot platform, leading to physical crashes and damage to the machinery (b) incorrect calculation of the height of the labware on the robot platform, causing damage to the gripper arms.

iii) Pipetting errors, leading to incorrect reagent mix assembly and sample loss.

iv) Hybridization errors, in which the extremely sensitive hybridization stage of sample preparation failed. Factors contributing to this included: (a) the behaviour of the thermal cycler (b) the makeup of reagent mixes (c) the manner in which reagent mixtures were combined with DNA libraries.

Errors i-iii were eventually resolved, but errors with hybridization on the robot platform persisted (See Results Section: MiSeq Run 3). Failure at this stage was costly, due to expenditure of the capture library, and particularly problematic because the failure of this stage could not be detected until sample preparation and sequencing had been completed and sequence alignment had been performed. For this reason, once the problems with robotic sample prep in pre and post-hybridization had been resolved, the hybridization step was still performed manually.

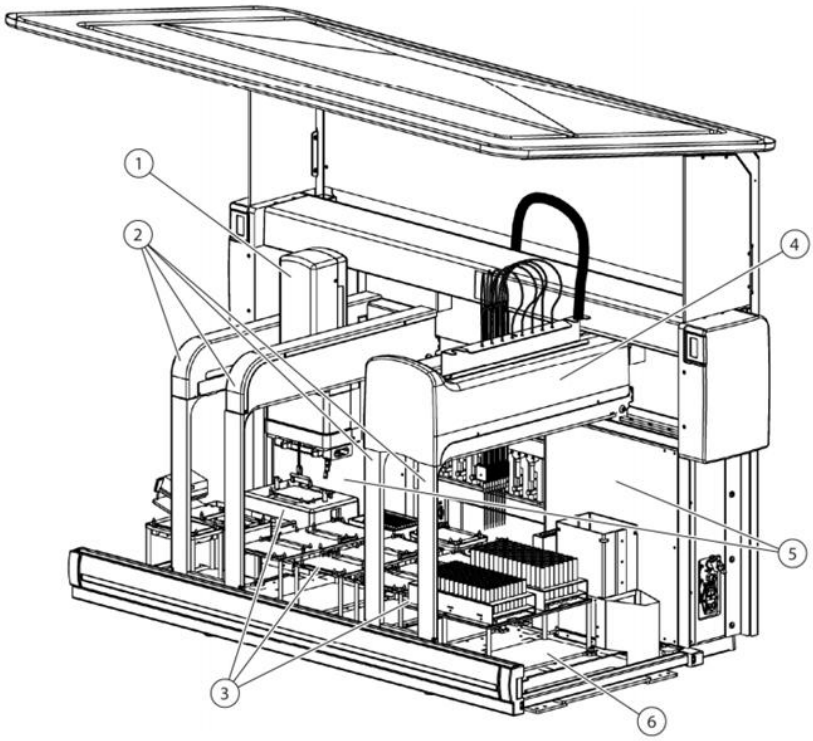
Validation of both the new library and the automated sample preparation platform were carried out over the course of the runs performed in Results Chapter 2.

Results Chapter 2: Development of an NGS Sequencing Methodology Suitable for Diagnostic Use

Introducing an Automated Sample Preparation Platform into the Diagnostic Laboratory

The aim of this research is to implement Next Generation Sequencing for routine diagnostic use in the Molecular Pathology laboratory at King's College Hospital. The manual sample preparation workflow for NGS is extremely time consuming and can result in high variation both within and between different runs. To increase the throughput of this process and with the intention of also reducing variability, a BioMek FX^P (Beckman Coulter) laboratory automation workstation was purchased. This is an automated liquid handling platform that can be used to mix and distribute reagents at each stage of the Agilent sample preparation procedure, as well as the size selection, washing and PCR steps. The workstation consists of two CNC (computer numerical control) arms which move over a modular deck. The arms move along the X and Y axes across the workstation on bridges. The arms can be fitted with tip adapters to transfer liquids between different positions on the workstation, or with grabbers to move items around the workstation. The workstation is protected by a cover and a light curtain, which will freeze all movement on the platform if it is activated. The basic layout of the workstation is shown in Figure 59.

The BioMek FX^P Laboratory Automation Workstation



1. Multichannel *Pods and Heads*
2. *Bridges*
3. *ALPs*
4. Span-8 *Pods and Heads*
5. *Towers*
6. *Deck*

Figure 59: The BioMek FX^P Laboratory Automation Workstation - Main Components. Image taken from Beckman Coulter BioMek FX^P User Manual (PN 987834AF), P31. 'ALPs' are Automated Labware Positions.

The layout of the BioMek FX^P workstation at the Department of Molecular Pathology is shown in Figure 60. This platform has two CNC arms: 'Pod 1', with a 96 well head (used for transferring reagents between plates and performing washes) and a grabber (for moving plates and tip racks around the platform) and 'Pod 2' with a span-8 head (with 8 individually manoeuvrable probes for mixing and dispensing reagents). The deck can take a variety of different plates, such as 96 well PCR plates and 1 ml capacity round-bottomed plates for washing reagents. Automated Labware Positioners (ALPs) included in the workstation are the T-Robot thermal cycler (for use in PCR and incubation stages), tip loading and disposal points, two static heating/cooling peltier modules, a shaking peltier module and a non-temperature controlled shaking module.

The BioMek FX^P Setup in the Dept. Molecular Pathology

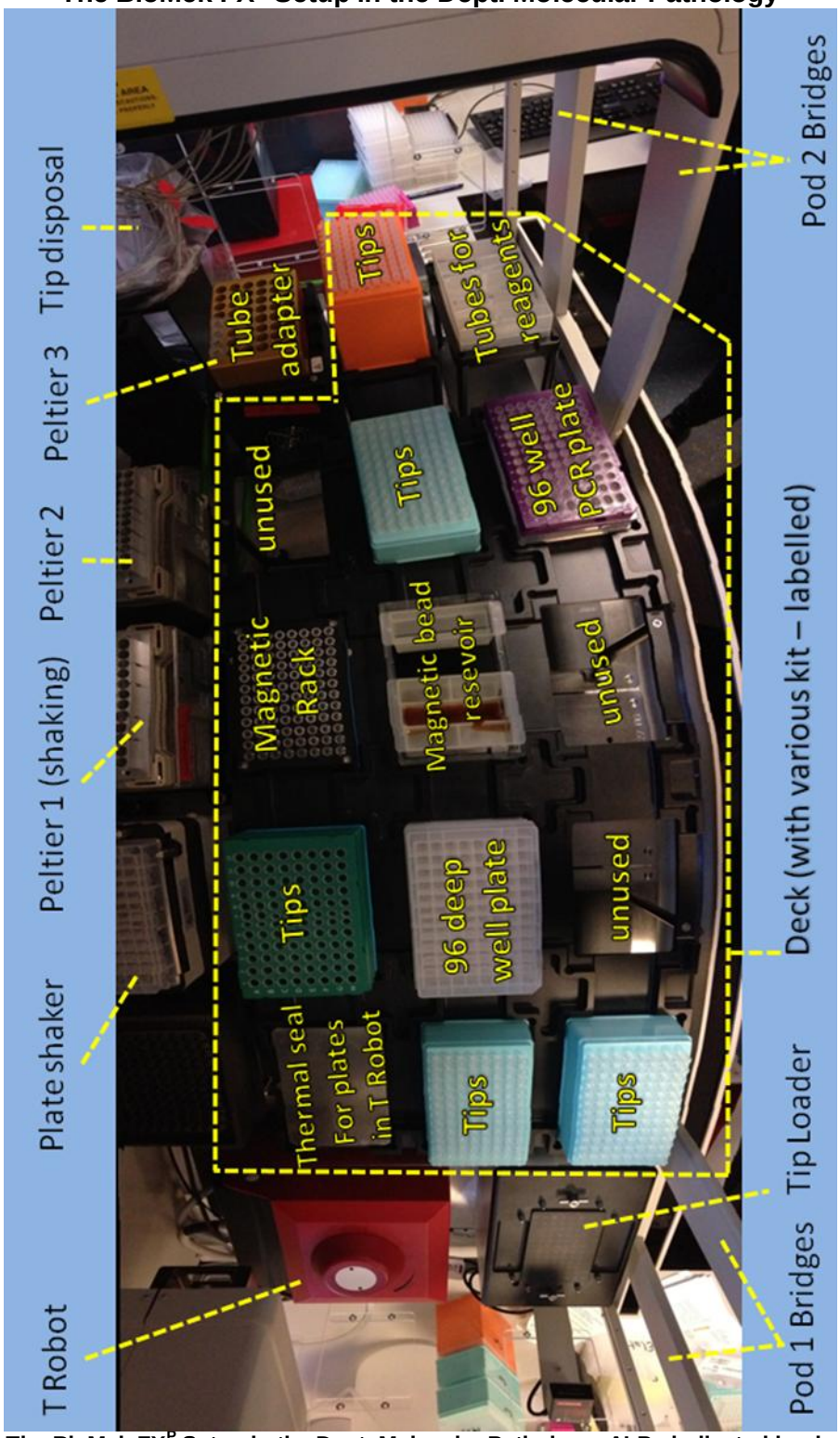


Figure 60: The BioMek FX^P Setup in the Dept. Molecular Pathology. ALPs indicated by dashed lines. Various pieces of kit arranged on the deck are highlighted in yellow. Picture taken by Frances Smith.

The instrument was set up by engineers from Beckman Coulter. A Beckman Coulter product specialist wrote a series of programs that would perform the Agilent SureSelect Automated Sample Preparation protocol. Four programs were used to manage different stages of sample preparation. The advantage of this was that each program

could then be optimised separately. This also allowed the workstation layout to be changed between stages, rather than all reagents for the entire process being thawed at the start of sample preparation. The division of the process between programs is outlined in (Figure 61). The last stage of sample preparation is an AMPure clean-up after the post-hybridization PCR. To reduce the risk of PCR products from this stage contaminating unamplified samples in future preparations, this stage was performed on a separate BioMek platform already operating in the laboratory – the NX^P. The NX^P program was already in routine use for purifying PCR products in the laboratory and no optimization of this step was required. The Pre-Hybridization PCR stage only consists of six cycles of amplification, so the contamination risk from this stage was considerably lower and continuing the sample preparation procedure on the same platform was considered to be acceptable.

Division of Automated SureSelect Sample Preparation between Programs and Equipment

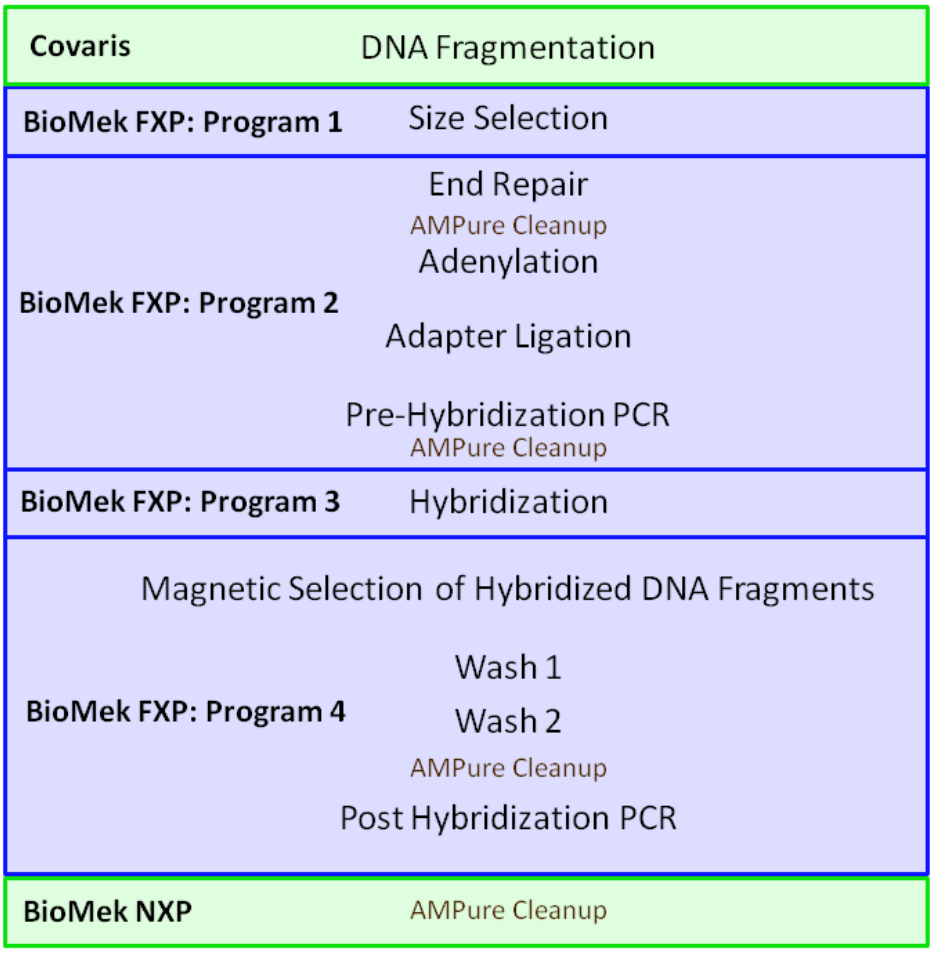


Figure 61 Division Of The SureSelect Sample Preparation Process Between Programs And Equipment.

A second program was also created that performed a SPRIselect double size-selection on DNA samples that had been fragmented using the Covaris which would be performed before the Agilent sample preparation program.

There was a plethora of problems with implementing automated sample preparation. Issues encountered and fixes made are summarised in Table 37.

The pre and post-hybridization stages of sample preparation were successfully implemented on the platform fairly quickly, but automating the hybridization stage of the process was challenging. At the end of the experiments outlined in Table 37, the remains of the first bait capture library had been exhausted. While a second library was designed and we awaited its delivery, the issues with the automated hybridization stage were tackled by the Molecular Pathology operations manager Frances Smith, using a different capture library panel for another study, in collaboration with Beckman Coulter engineers. Frances Smith found that factors affecting the success of hybridization included:

- Wash Buffer 2 was not properly heated by the peltier module during the post-hybridization clean-up. This buffer needs to be heated to 65°C for optimal performance
- The amount of DNA available for the hybridization stage was often lower than expected. Several changes were made to increase pre-hybridization DNA yield:
- DNA input concentration, which was increased to ensure the optimal amount of DNA for hybridization was available.
- Bead carryover during clean-up stages may have reduced sample yield
- Pipetting during the elution stage of the post-hybridization clean-up was causing bubbles to form, which decreased the amount of liquid aspirated at this stage, and therefore reduced the DNA yield.
- After the 24 hour hybridization stage is complete, the samples are aspirated from the T-robot and transferred to a fresh plate. It was noted that in some wells several microlitres of DNA remained in the old plate. The aspiration speed at this step was reduced in order to ensure all of each sample was transferred to the new plate.
- The ethanol used at the clean-up stages can sit on the workstation for up to five hours before the final time it is used during library prep over the course of a day. Changing the ethanol halfway through the day significantly increased pre and post hybridization yields. This suggested that the ethanol evaporating over the day had decreased its concentration resulting in lower DNA yields.

- Additional steps where tip changes should have been included in the program but were not, where contamination could occur, were identified. These steps were modified.

By the time these issues had been completely resolved, all the runs performed for this study had been completed: MiSeq Runs 3 and 4 as part of the process of developing the automated protocol, and MiSeq Runs 5 and 6 with automated sample preparation and manual hybridization.

Table 37 Issues Encountered Implementing Automated Sample Preparation on the BioMek FX^P Automated Sample Preparation Workstation

Run Details	Issues Encountered	Solutions Implemented
Water run with DNA samples, wash reagents and dummy sample preparation reagents to estimate sample loss during wash stages	50% of starting DNA was lost between end repair and pre-hybridization PCR steps	Increased length of elution stage of AMPure clean-up steps. Increased pipette mixing throughout wash steps.
Reagent run with DNA samples, wash reagents, sample preparation reagents	Large amounts of DNA were lost between end repair and pre-hybridization PCR. The level of loss varied between samples. We suspected the sample loss was due to insufficient mixing of the sample preparation reagents, leading to ineffective amplification during the pre-hybridization PCR.	Pipette mixing of reagents was increased at all stages.
Reagent run with DNA samples, wash reagents, and sample preparation reagents followed by manual hybridization	Hybridization was performed manually to check that the sample preparation steps were working correctly. No issues were encountered during sample preparation. The samples were sequenced on the MiSeq and showed approximately 55% of sequence was on target. There was good uniformity between the samples.	Proceeded to attempt automated hybridization on remaining prepared DNA
Samples prepared successfully up to the hybridization stage in the previous batch were used to test the automated hybridization protocol	Virtually no sequence DNA fragments were on target, meaning that the hybridization step had failed. The sequence that did align showed high inter-sample variability The workstation T-Robot had been programmed to 'reinitialize' once the thermal cycling program reached the 36 hour 65°C hold step. This was intended to be a checking step to ensure it had reached 65°C. Re-initialization required the T-robot to open its lid, and this caused the temperature drop briefly to 25°C. We suspected this had caused the hybridization step to fail.	Removed re-initialization step Reduced hybridization time from 36 hours recommended by Beckman Coulter to 24 hours as specified in SureSelect Automated Sample Preparation Protocol
Reagent run with DNA samples, wash reagents, and sample prep reagents. Automated hybridization stage following fixes to previous run.	Pod 1 failed to eject its tips correctly and crashed into the workstation deck. The arm was no longer in correct alignment with the deck which led to unequal pipetting across the plates and subsequently incorrect reagent mix distribution, and loss during clean-up stages.	An engineer rehomed the arm and replaced the faulty part that was not ejecting tips correctly

Table 37 Continued

Run Details	Issues Encountered	Solutions Implemented
Water run prior first solo run without Beckman Coulter engineers present	A bug in the program had been created when the robot arm was realigned which prevented the hybridization stage from working when less than eight samples were selected for sample preparation	Engineers were called to fix the program
Water run to confirm no more errors occurred in the program	Pod 2 was unable to confirm that it has shucked tips successfully. The error could be overridden by selecting 'ignore' in the error window that popped up, but this meant that the platform required constant supervision to clear these errors	The issue was fixed by engineers
Reagent run following fixes of previous errors	Pod 2 did not pipette reagent mix into one of the samples during the adenylation stage, despite there being excess reagent mix remaining. The AMPure XP beads placed on the platform in initial setup ran out during the final clean-up stage after post-hybridization PCR, leading to sample loss. Other stages in the program were also found to intermittently suffer from the reagent mix issue found at the adenylation stage. It was also noted that at one stage the Span-8 did not change tips at a point where there was a risk of contamination.	An engineer was called to rectify the pipetting error. The excess volume in each reagent mix and volume of beads allocated for sample prep were increased to ensure they wouldn't run out The step that could potentially cause contamination was modified so that the tips were changed
Reagent run following fixes of previous errors	Library construction was completed without issues until the hybridization stage. During hybridization, excess capture library was added to the samples, because the '<3Mb Library' setting had not been selected. A tip box was placed in the wrong position on the deck and Pod 1 crashed into it, causing damage to the its grabber and pushing it out of alignment	A pop-up was added into the program to check that the correct capture library size option had been selected. An engineer fixed the robot arm.
Reagent run following fixes of previous errors, using DNA prepared up to the pre-hybridization stage of the last run	Samples were sequenced on the MiSeq, and the amount of on-target sequence achieved for each sample was found to be extremely variable, from approximately 60% to <20%.	Engineers were requested to resolve the issues with the hybridization stage of the program.

MiSeq Run 3 for Validation of New Capture Library and the Automated Sample Preparation Platform

Sample Details

Four samples were prepared for sequencing on the MiSeq using the BioMek FX^P automated sample preparation platform. The samples were from four individuals who were homozygous controls for the HbC variant. The samples would be used primarily to evaluate sample preparation on the BioMek FX^P, but information about the haplotype background on which the HbC variant occurs was also of interest. The samples were sheared to 200 bp on the Covaris and post-shearing clean-up, sample preparation and hybridization were all performed on the BioMek FX^P. Target enrichment for this Run was performed using Bait Capture Library 2 (See appropriate methods and results sections for details of the new capture library), with the hope that this Run could be used to evaluate the performance of the new library.

Sample Preparation

The samples were sheared to 200 bp. As these samples had single nucleotide variants rather than large rearrangements, this smaller fragment size was not expected to have a negative impact on interpreting the sequencing results. During sample preparation it was noticed that the master mixes were not being dispensed into the sample tubes correctly. Furthermore, the beads that were added to the BioMek FX^P deck during setup of the protocol ran out before the last clean-up stage was completed, resulting in sample loss in three of the four samples.

After fixes were made to the protocol to address these problems (Table 37) sample preparation produced good results at the pre-hybridization stage. During the hybridization stage, excess capture library was added to the samples as the 'Library Size <3Mb Protocol' option (in which 2µl of capture library are used per sample, as opposed to 5µl in the <3Mb protocol) had not been checked. After the 24 hour hybridization period had been completed, Pod 1 crashed into a tip box that had been left on a deck position listed as blank in the sample preparation program. The crash caused damage to the grabbers and moved Pod 1 out of alignment. Subsequently, the post-hybridization sample preparation steps were performed manually for this run.

Sequencing

The samples were pooled to an equimolar concentration and sequenced on the MiSeq platform using the MiSeq V3 2x250 reagent kit. The run data showed that pooling had

been successful and the run had produced an acceptable cluster density for the V3 2x250 kit (Table 38, Figure 62).

Table 38 Sequencing Statistics for MiSeq Run 3

Run Summary MiSeq Run 3

Level	Yield Total (G)	Projected Total Yield (G)	Aligned (%)	Error Rate (%)	Intensity Cycle 1	% >= Q30
Read 1	4.74	4.74	1	0.7	185	89.67
Read 2	0.09	0.00	0.00	0.00	624	92.55
Read 3	4.74	4.74	0.95	1.43	122	78.73
Total	9.58	9.58	0.98	1.07	310	84.28

Read 1

Lane	Tiles	Density (K/mm2)	Clusters PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30	Yield(G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1
1	28	1149 +/- 17	88.86 +/- 0.8	0.08 / 0.069	21.36	18.97	89.67	4.74	250	1.00 +/- 0.02	0.07 +/- 0.07	0.12 +/- 0.01	0.14 +/- 0.01	0.16 +/- 0.02	185 +/- 12

Read 2

1	28	1149 +/- 17	88.86 +/- 0.80	0.00 / 0.00	21.36	18.97	92.55	0.09	0	0	0	0	0	0	624 +/- 49

Read 3

1	28	1149 +/- 17	88.96 +/- 0.8	0.109 / 0.066	21.36	18.97	78.73	4.74	250	0.95 +/- 0.02	1.43 +/- 0.15	0.34 +/- 0.03	0.38 +/- 0.04	0.43 +/- 0.05	122 +/- 12

Proportion of Reads on the Flow Cell from Each Sample Index MiSeq Run 3

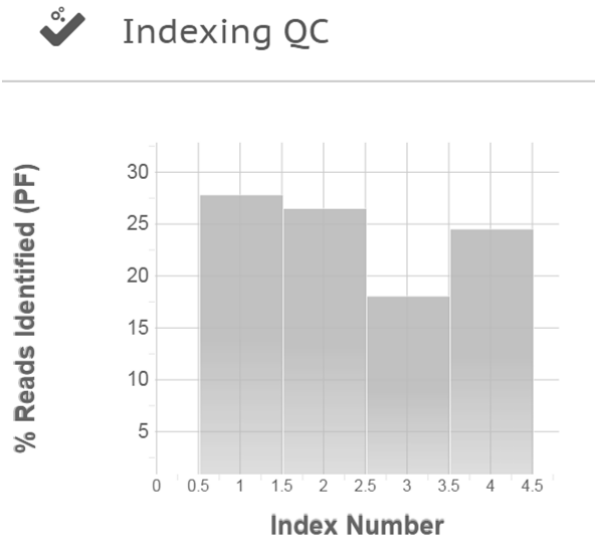


Figure 62: Proportion of Reads on the Flow Cell from Each Sample Index MiSeq Run 3.

Format Conversion

Reads in the FASTQ files were successfully converted to FASTA format in 97% of cases, all passing quality filters. Of those reads that did not pass quality filters, 79% failed due to not meeting the “Median Score” requirement (Table 39). The quality of base calls along Read 1 and Read 2 remained high along the entire length of both reads (Figure 63).

Table 39 Format Conversion statistics for MiSeq Run 3

Format Conversion Statistics MiSeq Run 3	Average	StDev
Total Reads in the Input File	2157240	143441.8
Reads Converted Successfully	2100138	136080.8
(%)	97.35	0.21
Reads Failed to Convert	57101.5	7799.49
Reads Filtered by "Median Score"	45411.75	6605.024
(%)	79.52	2.35
Reads Filtered by "Uncalled Bases"	0	0
Reads Filtered by "Called Base Number in Each Read"	0	0
Reads Filtered After Trimming	11689.75	1780.231
Reads Trimmed	283925.3	22358.05
Reads Trimmed by "Quality Score"	283925.3	22358.05
Reads Trimmed by "Homopolymer Trimming"	0	0
Reads Trimmed by "Sequence Trimming"	0	0
Reads Trimmed by "Specified Length"	0	0
Trimmed Bases	12856252	1502435
Trimmed Bases by "Quality Score"	12856252	1502435
Trimmed Bases by "Homopolymer Trimming"	0	0
Trimmed Bases by "Sequence Trimming"	0	0
Trimmed Bases by "Specified Length"	0	0

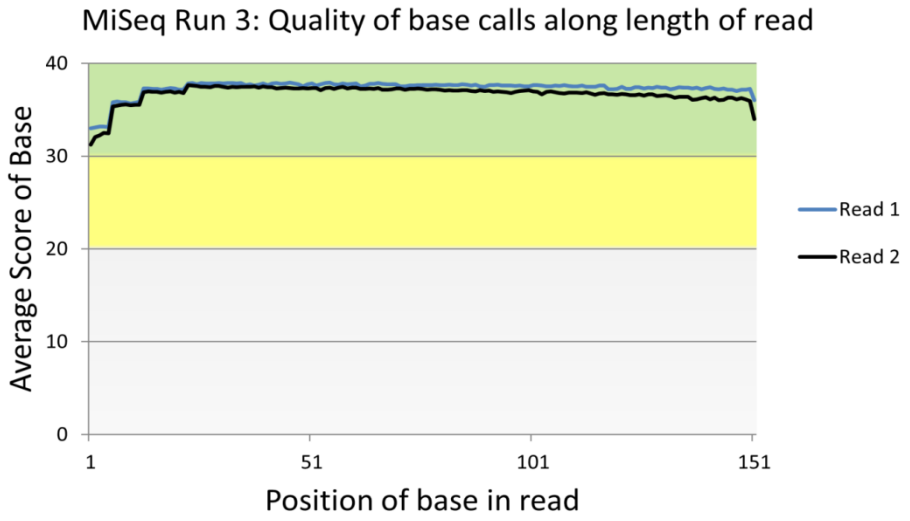


Figure 63: Quality of Base calls across Length of Read 1 and Read 2 for MiSeq Run 3.

Sequence Alignment

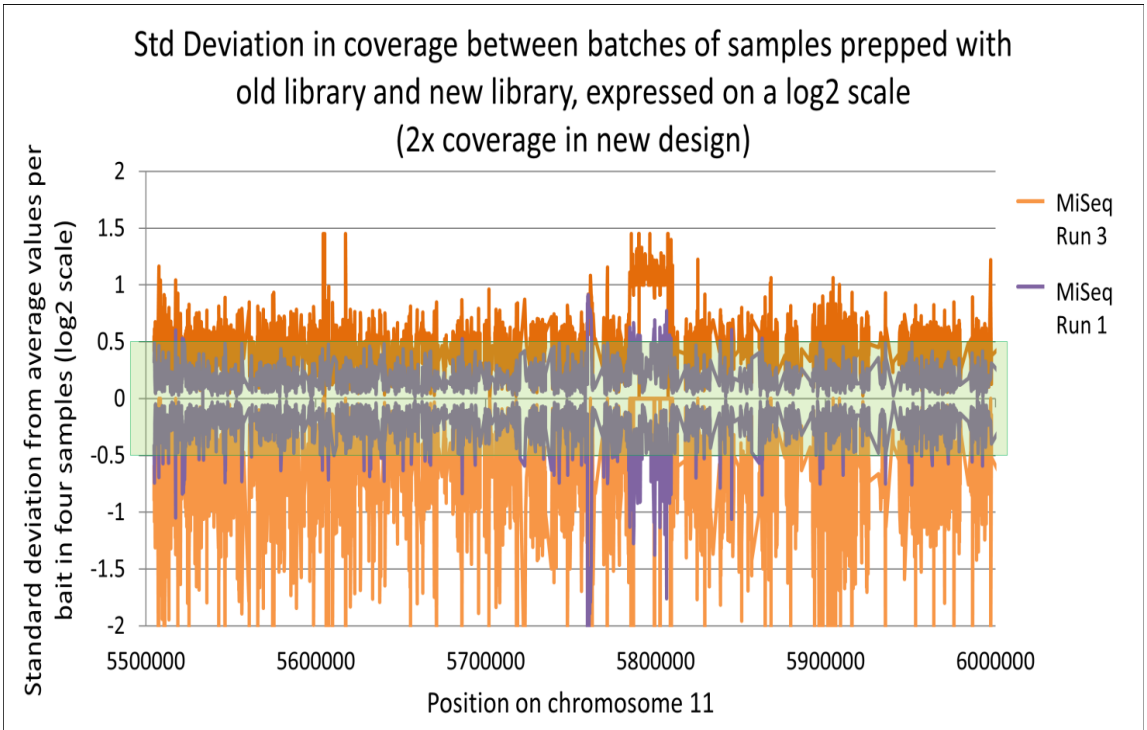
Reads were aligned to a reference sequence of the Human Genome in NextGene. The alignment statistics revealed that hybridization on the BioMek FX^{P} automated library prep workstation had not worked: only 10% of sequence was on target, covering the region of interest at an average read depth of 13x (Table 40). This was not sufficient to produce reliable variant calls for use in haplotyping these samples, as variant score calculation assigns penalties for low coverage (See Methods).

Table 40 Alignment Statistics for MiSeq Run 3

	Average	StDev	MiSeq Sample 10	MiSeq Sample 11	MiSeq Sample 12	MiSeq Sample 13
Perfectly Matched Reads	2337869	209283	2531802	2114793	2204792	2500090
Matched Reads Count	3660802	260877	3925557	3353591	3542753	3821306
Unmatched Reads Count	585355	26345	621325	561846	588012	570236
Short Reads Count	68323	7948	73382	58383	65605	75920
Number of unmatched bases recorded as mutations						
Mismatches	97653	5059	103595	97383	91259	98375
Deletions	8760	1207	10071	9428	7419	8123
Insertions	5486	274	5487	5866	5356	5233
Number of unmatched bases NOT recorded as mutations						
Mismatches	4418668	341141	4526273	4338614	4811565	3998220
Deletions	844645	92298	858349	839347	952927	727957
Insertions	475141	50493	482045	474725	533458	410335
Average Read Length	273	4.1	277	279	273	272
Average Coverage (ROI)	13	4	416427	452822	326266	321177
Reads on Target (%)	10	2	10.60	13.50	9.20	8.40

There was extremely high inter-sample variation across the region of interest in this run. This may be primarily the result of the extremely low level of coverage, but could also indicate that the success of hybridization and uniformity of sample preparation between the samples were different in the manually and automated sample preparation/hybridization procedures. A 0.5 Mb segment of the region of interest on chromosome 11 (chr11:5,500,000 - 6,000,000) was selected to compare inter-sample variation within this run to inter-sample variation within MiSeq Run 1. This region was selected as it was not involved in any of the structural variants included in MiSeq Run 1. The samples in MiSeq Run 1 had been prepared for sequencing manually, and had been enriched using the old bait capture library. The new capture library used for enrichment of MiSeq Run 3 increased bait tiling density across this region from 1x to 2x, with the specific aim of reducing inter-sample variability. In MiSeq Run 1, the amount of baits showing high inter-sample coverage variability (i.e. the standard deviation from the average of the four samples was >0.5 on a log2 scale) in this region was 1.5%. In MiSeq Run 3, 29.5% of baits within the same region showed high inter-sample coverage variability (Figure 64).

Comparing Inter-Sample Variation between MiSeq Run 1 and MiSeq Run 3 Shows Substantially More Variability in MiSeq Run 3



	Number of Baits Covering Region	Baits with highly variable coverage within the run (standard deviation from the average was >0.5 on a log 2 scale)	
		Number of baits	% of baits in region
MiSeq Run 1 (Old capture library)	3422	58	1.54%
MiSeq Run 3 (New capture library)	6907	2005	29.03%

Figure 64: Comparing Inter-Sample Variation between MiSeq Run 1 and MiSeq Run 3 Shows Substantially More Variability in MiSeq Run 3. Graph: Standard deviation within the sample cohort from the RPKM average for each bait position in the design. The graph depicts a 'snapshot' region of chromosome 11. Orange line shows MiSeq Run 3 (using the new bait capture library) and purple line shows MiSeq Run 1 (using the old bait capture library). Green box indicates acceptable limits of inter-sample variation (+/- 0.5). Table: number of baits in each design showing high inter-sample variability in coverage.

This data shows that hybridization is still ineffective on the robotic platform, and that the automated sample preparation may result in high-inter sample variation in coverage of the bait tiled region. These issues must be resolved before automated sample preparation can be introduced into the diagnostic laboratory for routine use. The low coverage and high variability mean that this data could not be used to accurately detect structural variants in prepared samples.

MiSeq Run 4

Sample Details

Eight samples were prepared for sequencing on the BioMek FX^P platform following fixes to the method implemented after the previous run (Table 37). A larger number of samples were included in the run to get a better idea of inter-sample variation within the run. We suspected that mixing issues that had been identified over the course of optimizing the automated library preparation protocol may result in lower quality sample preparation at some positions than others. By including eight samples we would be able to identify the positions on the plate where sample preparation was worst. The eight samples including a range of positive and negative controls for rearrangements affecting the alpha globin gene cluster on chromosome 16, to see whether the new bait capture library covered a large enough region of chromosome 16 to cover all these rearrangements.

Sample Preparation

The samples were fragmented to 400 bp on the Covaris. Sample preparation was then carried out on the BioMek FX^P automated library preparation workstation. Target enrichment was performed using Bait Capture Library 2. Sample preparation had to be halted due to a crash on the BioMek FX^P that required realignment of the robot arm, but this did not appear to affect sample preparation negatively. Tapestation traces at various stages during sample prep showed consistency between the samples, and that there was not unacceptable sample loss across the preparation process. After the final post-PCR clean-up some primer dimer was visible on the Tapestation 4, 5, 6, 7 and 8. Two additional clean-ups were necessary to remove this product. (NB: These products are more obvious after the second clean-up because the first clean-up trace was measured using the D1000 kit instead of the D1000 High sensitivity kit)

Sequencing Results

The samples were sequenced twice on the MiSeq, as the first run (MiSeq Run 4a) had an unworkably low cluster density (96 +/- 38 K/mm^2), resulting in under loading of the flow cell. A second run (MiSeq Run 4b) was performed loading 10pM of the pooled samples onto the flow cell. A MiSeq V3 2x150 reagent kit was used for this run, as no more V3 2x250 kits were available in the lab. This run achieved an acceptable cluster density (1215 K/mm^2), so the run data and FASTQ output were downloaded and analysed. Of the eight pooled samples, two samples accounted for 47% of the reads

that were produced (Figure 65). This was most likely to be the result of a quantification or pipetting error.

Proportion of Reads Achieved Per Index for MiSeq Run 4

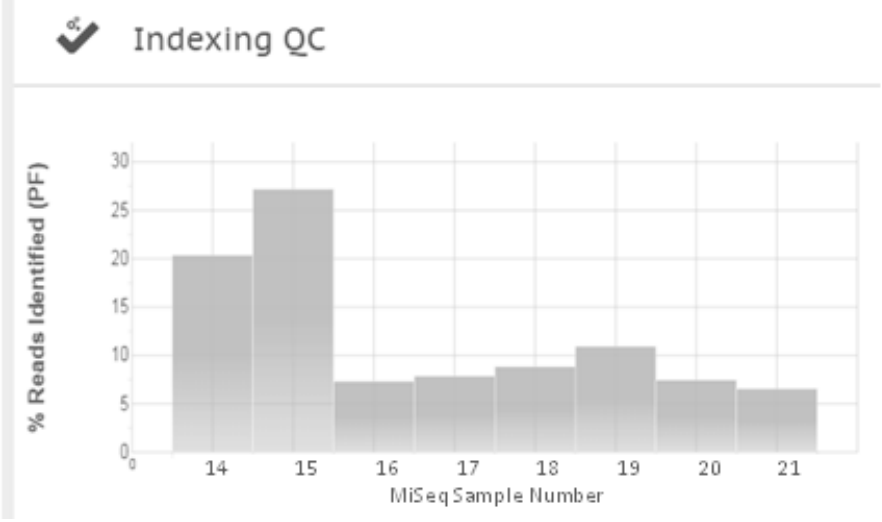


Figure 65: Proportion of Reads Identified on the Flow Cell Containing Each Index Tag, MiSeq Run 4b. MiSeq Samples 14 and 15 together account for 47.5% of total reads.

The read data was converted from FASTQ to FASTA. In all samples >99% of reads passed quality filtering.

Sequence Alignment

Alignment of the FASTA data showed that overall the alignment statistics corresponded to the amount of DNA sequenced on the flow cell, with the two overloaded samples showing higher values than the other 6 for most criteria. The only statistic where this isn't the case is in the amount of on-target sequence for Sample 14, which is significantly less than that achieved for Sample 15. Sample 14 had been in position A1 of the plates used during sample preparation on the BioMek FX^{P} platform. Issues had been identified with the accuracy of pipetting at this position, which may have affected hybridization efficiency for this sample. The same issue was seen in the data from both the first and second MiSeq runs. On-target capture ranged from 18% to 49% (Table 41). This was substantially lower than expected. The average coverage of the ROI was too low to effectively use this data, and an extremely high level of inter-sample variation was found across the bait-covered regions. It was concluded that there were still problems with the hybridization stage of the sample preparation process at this time. These issues were not resolved until after data collection for this project ended. For MiSeq Run 5 and MiSeq Run 6 hybridization was performed manually.

Table 41: Sequence alignment results for MiSeq Run 4.

Sample:	MiSeq Sample 14	MiSeq Sample 15	MiSeq Sample 16	MiSeq Sample 17	MiSeq Sample 18	MiSeq Sample 19	MiSeq Sample 20	MiSeq Sample 21
Perfectly Matched Reads Count	5361733	7111730	1961834	2060562	2413861	3057601	2003812	1786561
Matched Reads Count	7039055	9338180	2575922	2767601	3154975	3950873	2641565	2343100
Unmatched Reads Count	377582	374890	99633	133488	156047	153993	137902	93692
Short Reads Count	180348	281767	67139	105171	72524	84452	74931	69162
% of reads that are short	3.36	3.96	3.42	5.10	3.00	2.76	3.74	3.87
# Unmatched bases recorded as mutations	1613421	2029716	558279	551994	652153	692221	524806	431811
# Unmatched bases NOT recorded as mutations	3234048	3199479	975337	1025094	1379124	1453857	1115933	770770
Average Read Length	144	143	143	143	144	144	143	143
Average Coverage (in whole genome alignment)	1	2	2	1	1	1	1	2
Number of Covered Bases	6.22E+08	5.19E+08	1.3E+08	2.05E+08	3.04E+08	3.3E+08	2.28E+08	1.45E+08
Average coverage (in alignment to baited region)	79	231	78	58	40	70	43	59
% of reads that are duplicated	3	4	2	0	0	0	0	1
Total duplicate reads removed from file	516458	950506	168368	57694	51614	73540	44352	59496
Average gap distance	360	380	330	410	375	370	375	360
# reads on target	1306336	3811490	1266755	950720	644160	1144105	717407	982652
% reads on target	18.6	40.8	49.2	34.4	20.4	29.0	27.2	41.9

MiSeq Run 5

Following the redesign of the Agilent Bait Capture Library, two sequencing runs were performed on the MiSeq to test the ability of this technology to characterise a range of rearrangements affecting the alpha and beta globin gene loci. Some samples that had been previously run with the old library design were re-run with the new design. This was in order to determine whether the improved bait placement strategy and coverage associated with reduced noise had increased the assay's ability to detect variants (single nucleotide changes and structural variants).

Sample Details

Prior to this run, a sequencing experiment was performed as part of another study using a cohort of 16 samples. The samples were prepared in tandem, target enriched using Bait Capture Library 2, pooled and sequenced on the MiSeq in a single run. This experiment was only concerned with genotyping the samples, so high read depth was not a priority. Sequence alignment from this run showed that with a batch of 16 samples, the coverage achieved across the bait-tiled region was approximately 250x. We considered this to be a satisfactory read depth for variant detection and as such, increased the number of samples included in the following two runs in this study.

Fourteen samples were prepared in tandem for sequencing to evaluate Bait Capture Library 2 on the MiSeq and to provide NGS data for several known structural variants that commonly cause alpha thalassaemia. The sample group included eight positive controls for alpha thalassaemia variants commonly encountered in the diagnostic laboratory, four test samples predicted to have unknown variants affecting the alpha globin gene cluster from screening data, one negative control, and one test sample that had been previously sequenced on the MiSeq using the old bait capture library (Table 42)

At the time that MiSeq Run 5 was ready for sequencing, a MiSeq platform was brought into the Molecular Pathology laboratory. Prior to this, MiSeq sequencing had been carried out on an instrument in another department. In order to validate the new instrument, we sequenced MiSeq Run 5 twice, once on the machine used for previous runs, and once on the new instrument. Subsequently, the first sequencing run – **MiSeq Run 5a** – was not analysed in great detail. General run statistics will be described, followed by a more detailed account of the analysis of the second round of sequencing for these samples – **MiSeq Run 5b.**

Sample Preparation

Sample preparation was performed using the automated BioMek FX^P platform in accordance with the manufacturer's protocol (see Methods). Hybridization was performed manually, as this part of the process had not yet been optimized on the BioMek FX^P. Size selection was also performed on the robot.

Size Selection Using SPRISelect Beads

For MiSeq Run 1 and MiSeq Run 2, DNA had been sheared to a mean fragment size of 500 bp in order to take advantage of the longer read length of the MiSeq platform and to improve fragment mappability. We found that the sample preparation process favoured smaller fragments and these were preferentially amplified, leading to a smaller than expected average sample size at the sequencing stage.

To combat this, we introduced a double size selection step to remove small fragments. Several samples were sheared on the Covaris E220 using a range of settings specified by the manufacturer. Fragmented DNA was electrophoresed on the Tapestation using the gDNA kit (Figure 66). We found that the manufacturer's settings for shearing to 400bp and 500 bp produced the most DNA in our target size range of 400-600 bp (Figure 66 D&E). The 500 bp shearing settings produced a mean fragment size of 736

bp, but with the largest proportion of fragments in our target range relative to the smaller fragments we intended to exclude from sample preparation.

Results of DNA Fragmentation on the Covaris E220 According to Manufacturer-Recommended Settings

	A	B	C	D	E	F	G	H
Manufacturer's settings:	150	200	300	400	500	800	1000	1500
Duty Factor	10%	10%	10%	10%	5%	5%	5%	2%
Peak Incident Factor	175	175	140	140	105	105	105	140
Time	450	180	80	55	80	50	40	15
Cycles per burst Waterbath Temperature Water level	200 6-8°C 6							
Mean fragment size achieved	150	200	360	494	736	839	952	1405

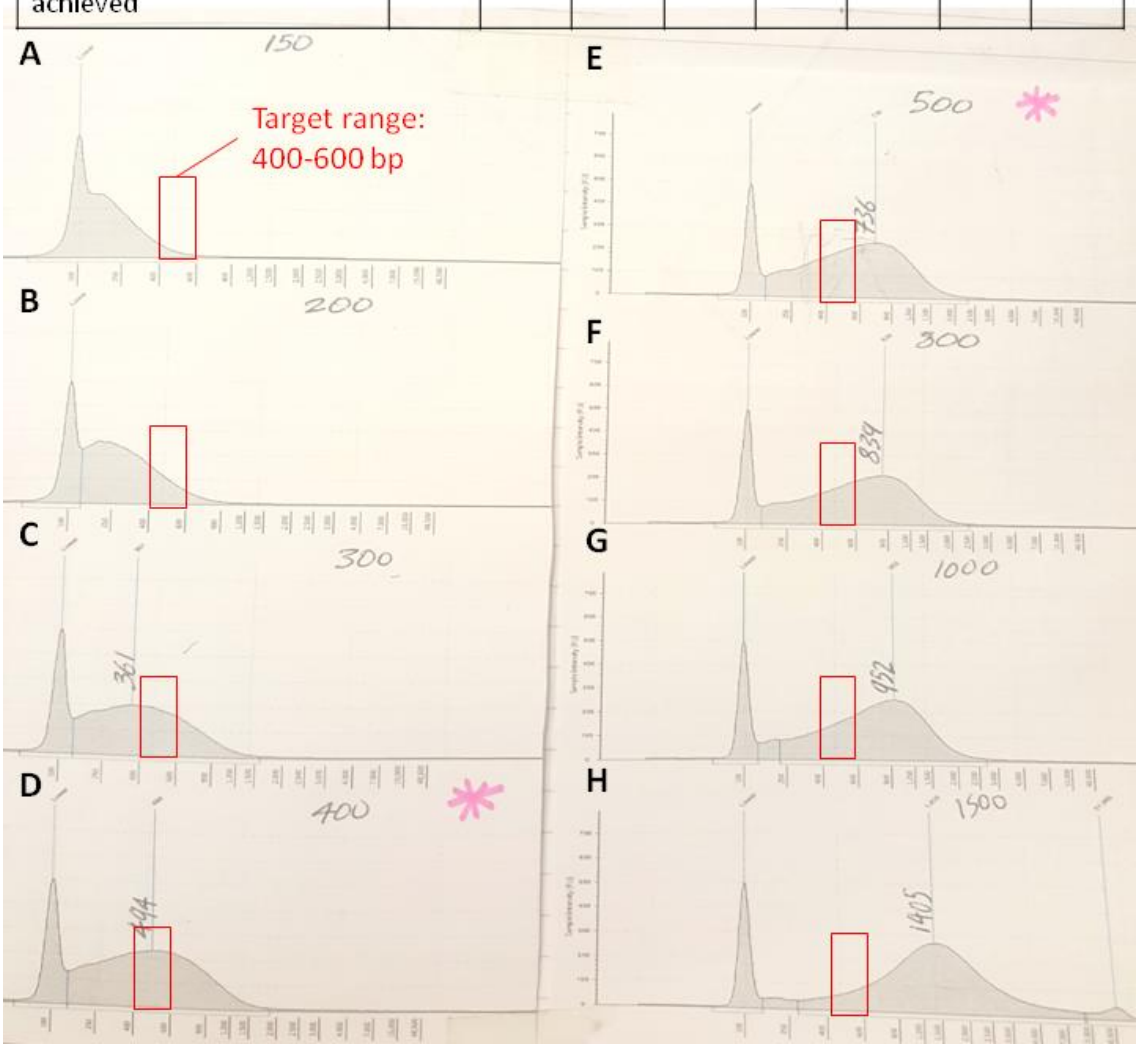


Figure 66 Results of DNA Fragmentation on the Covaris E220 According to Manufacturer-Recommended Settings. Settings and resulting mean fragment size are shown in the table, and charts below with corresponding numbers show the range of fragments produced. Target fragment range is indicated by a red bar.

To tighten the fragment size range prior to sample preparation, DNA sheared using the 500bp fragmentation settings was purified twice with SPRIselect beads. The first purification used a concentration of 0.4x SPRIselect beads per sample volume. This removed fragments <~250 bp from the sample. A second purification using a concentration of 0.6 x SPRIselect beads per sample volume removed fragments >~800 bp from the sample (for further details see Methods). This produced a mean fragment size of 694 bp, but with a substantial reduction in small fragments (Figure 67). The volume of input DNA was increased from 3 µg to 5 µg to accommodate the increased sample loss resulting from this purification process.

Bioanalyser Traces Before and After Double Size Selection to Concentrate DNA Fragments Of Desired Size

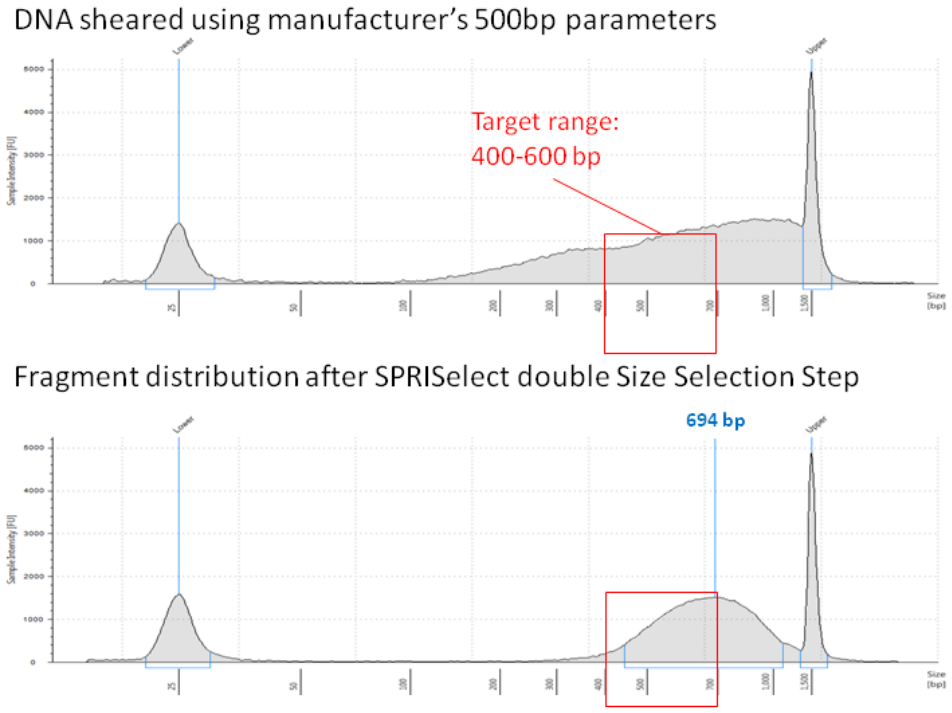


Figure 67: Bioanalyser Traces Before and After Double Size Selection to Concentrate DNA Fragments Of Desired Size. Electropherogram traces (Tapestation DK 1000 kit) show sample post fragmentation above, and post size selection below. Target fragment is size represented by red box.

Following fragmentation, samples were prepared according to the protocol as outlined in Methods. 2nM of each sample was added to an equimolar pool for sequencing. 9pM of pooled DNA was loaded onto the flow cell.

Table 42: Sample List, MiSeq Run 5

Sample	Status	Variant	Locus
MiSeq 22	Test sample	Duplication	Alpha
MiSeq 23	Positive control (---^{FIL}/)	Deletion (HbVar.1094)	Alpha
MiSeq 24	Positive control (--^{Thai}/)	Deletion (HbVar.1095)	Alpha
MiSeq 25	Previously characterised	Duplication	Beta
	Positive control (α-$^{3.7}$/)	Deletion (HbVar.1076)	Alpha
MiSeq 26	Positive control (α-$^{3.7}$/--^{SEA})	Compound heterozygous deletions (HbVar.1086, HbVar.1076)	Alpha
MiSeq 27	Positive control (--^{SEA}/)	Deletion (HbVar.1086)	Alpha
MiSeq 28	Positive control (--^{MED}/)	Deletion (HbVar.1078)	Alpha
MiSeq 29	Positive control (α-$^{4.2}$/)	Deletion (HbVar.1079)	Alpha
MiSeq 30	Positive control (--$^{20.5}$/)	Deletion (HbVar.1088)	Alpha
MiSeq 31	Positive control (ααα/)	Insertion	Alpha
MiSeq 32	Negative control	None	None
MiSeq 33	Test Sample	Deletion	Alpha
MiSeq 34	Test Sample	Duplication	Alpha
MiSeq 35	Test Sample	Duplication	Alpha

Sequencing Statistics

Samples were sequenced using the v3 2x300 MiSeq Reagent Kit (MS-102-3001), which was prepared in accordance with the manufacturer's protocol (see Methods). Sequencing showed that MiSeq Samples 32 and 35 had been under loaded when creating the equimolar sample pool (Figure 68). The cluster density achieved on the flow cell was within acceptable limits (600-1400 k/mm^2) for the v3 kit. The large fragment size used in this experiment meant that it was difficult to produce cluster densities towards the higher end of this range, due to the space required for long fragments to bend over during bridge amplification. The quality of base calls at the ends of both reads dropped below acceptable levels (Figure 69). This is a known feature of the longer read length of the v3 kit.

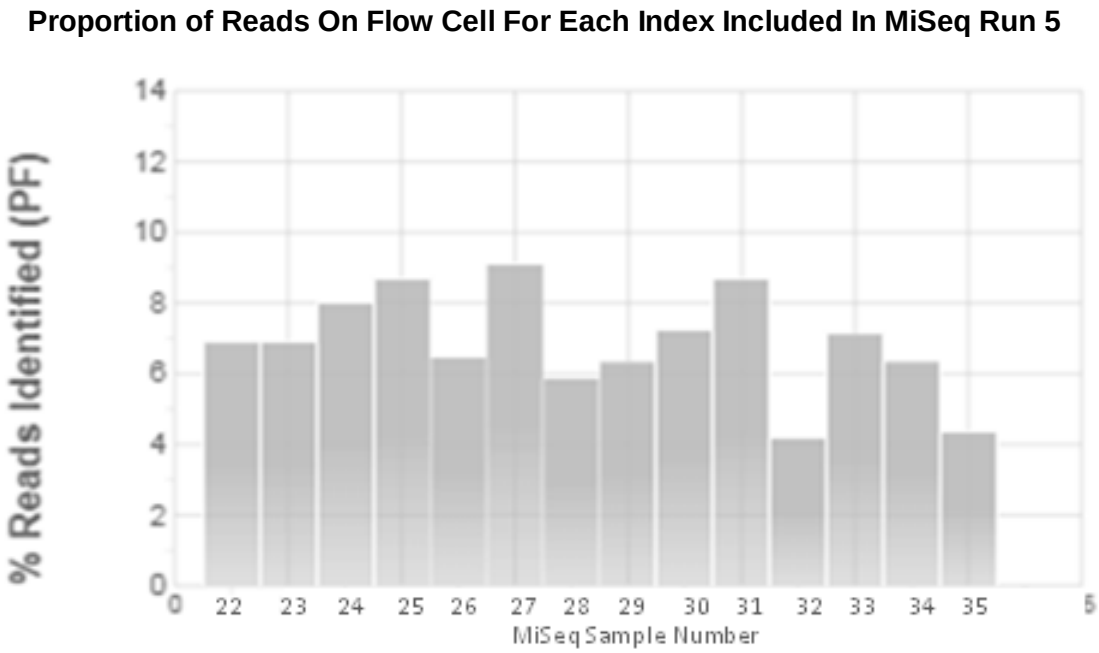


Figure 68: Proportion of Reads On Flow Cell For Each Index Included In MiSeq Run 5.

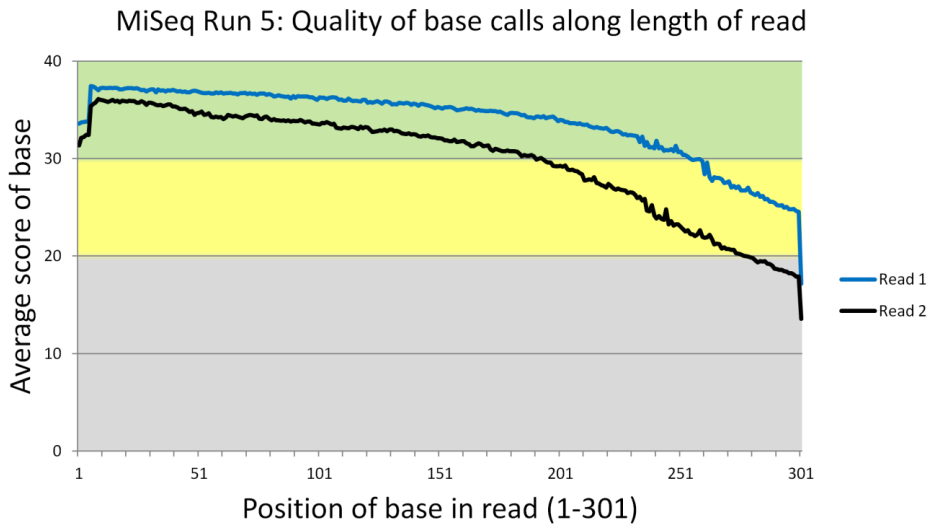


Figure 69: Quality of Base calls Across Read 1 and Read 2 During Sequencing MiSeq Run 5.

Format Conversion

An average of 98% of reads in the sample input files were successfully converted from FASTQ to FASTA format, passing all quality filters. 98.2% of reads that did not pass format conversion failed to meet the 'Median Score Threshold' criteria (Table 43).

Table 43: MiSeq Run 5 Format Conversion Statistics.

Format Conversion Statistics MiSeq Run 5	Average	StDev
Total Reads in the Input File	1238963	264830.3
Reads Converted Successfully	1216821	261649.7
%	98.21%	0.17
Reads Failed to Convert	22142.07	3423.95
Reads Filtered by "Median Score"	21748.93	3361.86
%	98.22%	21.18
Reads Filtered by "Uncalled Bases"	0.71	0.99
Reads Filtered by "Called Base Number in Each Read"	0	0
Reads Filtered After Trimming	392.42	71.91
Reads Trimmed	650491.6	125778.9
Reads Trimmed by "Quality Score"	650491.6	125778.9
Reads Trimmed by "Homopolymer Trimming"	0	0
Reads Trimmed by "Sequence Trimming"	0	0
Reads Trimmed by "Specified Length"	0	0
Trimmed Bases	25404692	4197521
Trimmed Bases by "Quality Score"	25404692	4197521
Trimmed Bases by "Homopolymer Trimming"	0	0
Trimmed Bases by "Sequence Trimming"	0	0
Trimmed Bases by "Specified Length"	0	0

Sequence Alignment

Sample data in FASTA format was aligned to the reference sequence in NextGene using standard parameters. The average coverage across the entire genome was 6x, and across the bait-tiled region was 230x, indicating that hybridization had been successful.

Variant Detection

Detection of large structural variants

Large structural variants are identified where the RPKM (a normalised measure of the number of reads covering a location) differs from the values for the bait-covered positions in a negative control. A comparative increase in coverage implies a duplication, and a decrease in coverage implies a deletion. As coverage is highly variable throughout the regions of interest, multiple controls are used to generate a control average to compare with the RPKM data from samples. If the RPKM data from the sample deviates by more than one standard deviation from the negative control average (the negative control standard deviation, or for short: **'NegC StDev'**), this suggests that that region has been affected by a structural rearrangement.

Accurate detection of large structural variants was not possible using this dataset. The low number of negative controls resulted in high levels of noise in the data which prevented reliable identification of structural variants some positive controls, as well as the test samples. Figure 70 shows the RPKM plot for MiSeq Sample 24 (--^{Thai}/), with the position of the known variant ('Thai 1 Deletion', HbVar.1095) indicated by an orange bar. While 76% of baits covering the deleted region did show a significant departure from the 'Neg C StDev', this was matched by 56% of baits exceeding the NegC StDev in the balanced region (Table 44). This level of variation in the balanced region prevents differentiation between genuine dosage changes and noise. More negative controls are necessary to reduce the level of noise in the data.

Detection of the (--^{THAI}/) Variant Using Coverage Data from Two Negative Controls

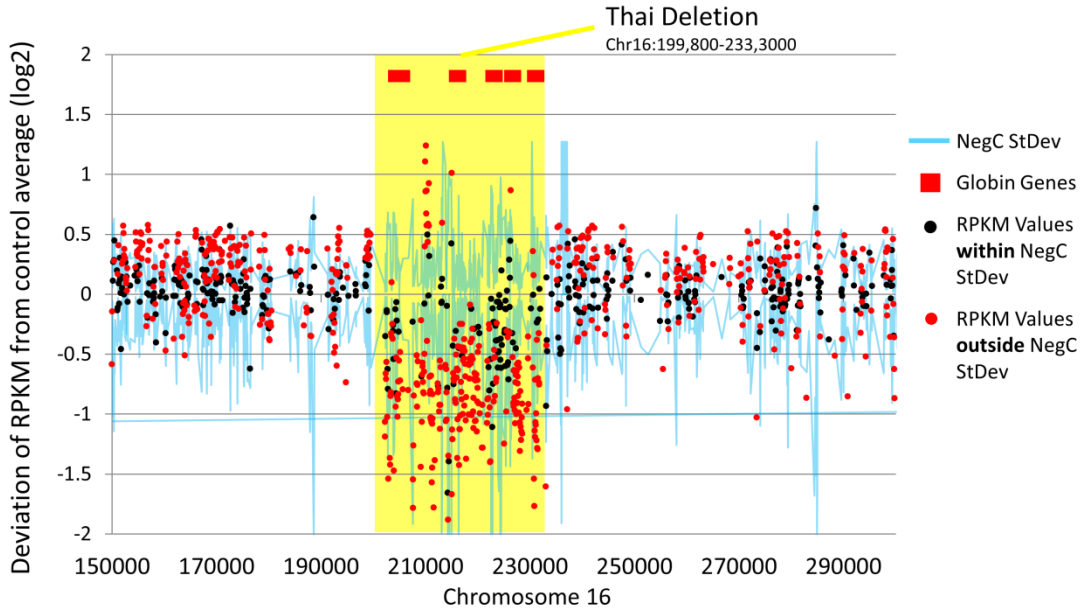


Figure 70: Detection of the (--THAI/) Variant Using Coverage Data from Two Negative Controls. Graph shows RPKM plot for MiSeq Sample 24. The position of the known deletion in highlighted in yellow. Baits where RPKM was within the limits of the NegC StDev are plotted in black and baits where RPKM exceeds the NegC StDev are plotted in red

Table 44: Difference in RPKM Values between a Positive Control Sample and the Average Values for Two Negative Controls.

Difference in RPKM values between a positive control sample and the average values for two negative controls.	Proportion of baits within NegCStDev	Proportion of baits exceeding NegC StDev
In deleted region (chr16)	24%	76%
In balanced region (chr16)	44%	56%

For this reason, variant detection analysis of **MiSeq Run 5a** was discontinued. The pooled samples were re-sequenced in order to compare the performance of the off-site MiSeq with the new instrument in the lab (discussed below). We attempted to integrate the data from both runs to increase the read depth across the region, but found that this was not possible, as the MiSeq platform recycles its naming system between runs, meaning that many reads from the two runs would have identical names, and thus the files could not be merged together. Subsequently, structural variant analysis for these samples was performed using the data from **MiSeq Run 5b**, with samples from **MiSeq Run 6** that were balanced at the alpha globin gene locus used as additional negative controls.

MiSeq Run 5b

Sample details

The pooled sample libraries from MiSeq Run 5 were re-run on the MiSeq using a new 2x300 v3 MiSeq Reagent Kit. The samples were prepared for sequencing in accordance with the manufacturer's protocol as before, and sequenced on the MiSeq. This produced a cluster density similar to the previous run (861 +/- 31 K/mm^2). Lower cluster densities are a consequence of the longer fragment read length (as each fragment requires more space during bridge amplification) and this density is considered to be acceptable by Illumina using the 300bp read kit. Base call quality dropped at the end of Read 1 and 2, as in the previous run.

Format Conversion

Format conversion was performed to convert sequence files from FASTQ to FASTA format. 98.4% (+/- 0.11) of reads passed format conversion (Table 45).The majority of those that failed were rejected based on the 'Median score threshold' filtering parameter. 50% of reads required trimming based on their quality score, which is a consequence of lower base calling confidence towards the end of the long reads.

Table 45: Format Conversion Statistics MiSeq Run 5b.

Format Conversion Statistics MiSeq Run 5b	Average	StDev
Total Reads in the Input File	1462224	257543.7
Reads Converted Successfully	1439009	254459.3
(%)	98.4	0.11
Reads Failed to Convert	23214.08	3458.64
Reads Filtered by "Median Score"	21972.92	3250.08
(%)	94.65	0.42
Reads Filtered by "Uncalled Bases"	1133.41	217.47
Reads Filtered by "Called Base Number in Each Read"	0	0
Reads Filtered After Trimming	107.75	18.18
Reads Trimmed	807144.6	140730.9
Reads Trimmed by "Quality Score"	808158.4	134227.6
Reads Trimmed by "Homopolymer Trimming"	0	0
Reads Trimmed by "Sequence Trimming"	0	0
Reads Trimmed by "Specified Length"	0	0
Trimmed Bases	23689727	3721451
Trimmed Bases by "Quality Score"	23689727	3721451
Trimmed Bases by "Homopolymer Trimming"	0	0
Trimmed Bases by "Sequence Trimming"	0	0
Trimmed Bases by "Specified Length"	0	0

Comparison between Sequencing Metrics from MiSeq Run 5a and MiSeq Run 5b

Comparison of the quality metrics from the off-site and new on-site instruments showed that they had a comparable performance (Table 46). A comparison between variant calling from each run was performed by the on-site bioinformatician, Dr David Brawand, using his bespoke alignment and analysis pipeline (in development). Variant call information was compared for all fourteen samples sequenced in this run, using four independent variant callers to create a consensus list of variants in each sample (variant callers were FreeBayes, Samtools, Platypus and Varscan). The consensus variant calls from the data produced by the two sequencing instruments showed 97% concordance between them (data not shown). We concluded that the new instrument produced data that was comparable to that produced at the other site where routine sequencing was already in operation.

Table 46: Comparison between quality metrics from off-site run 5a and new on-site instrument run 5b.

Metric	MiSeq Run 5a	MiSeq Run 5b
Library type	Globin Mapping	Globin Mapping
MiSeq	Liver	MHL
Read length	2x300bp	2x300bp
Loading conc. (pM)	9	9
Cluster density	850	861
%PF clusters	90.83	96.45
Phasing/prephasing (%)	0.121/0.084	0.174/0.071
Read number PF (million)	18.01	20.58
Intensity cycle 1 read 3	164	312
% >Q30	78.6	82.4
Total yield (G)	10.9	12.4
PhiX % aligned	2.21	2.16
PhiX error rate %	2.47	2.29

Sequence Alignment

FASTA data was aligned to the reference sequence using standard alignment parameters, with a paired gap distance of 0-800 bp to accommodate the mean fragment size of the run which was 600 bp. Across the 14 sample run, an average of 57% of reads aligned to the target region (of a total of approx. 2.6 million matched reads), giving the target region 225x coverage (Table 47). Four percent of reads could not be aligned to the reference sequence. The coverage obtained for each sample was broadly equal, indicating that they had been correctly pooled prior to sequencing.

Table 47: Alignment Statistics for MiSeq Run 5b.

	Average	StDev	22	23	24	25	26	27	28	29	30	31	32	33	34	35
Perfectly Matched Reads	1234203	267796	1649156	1458066	1128143	1267040	1556321	1283272	1244714	1114516	1200988	1531125	1161787	703452	788001	1244714
Matched Reads Count	2600690	524154	3295156	3270378	2290842	2752065	3231266	2626092	2640206	2551237	2526960	3113163	2387501	1588595	1740632	2640206
Unmatched Reads Count	119240	75204	91425	234300	68844	84345	198458	93928	106814	315287	84553	95284	77549	86071	63400	106814
Short Reads Count	70581	8948	69437	63409	61308	80778	79070	64798	72793	78032	77933	66883	62436	66816	87689	72793
Number of unmatched bases recorded as mutations																
Mismatches	896915	193557	1102961	1130329	720681	1086060	1114657	940447	964814	931208	858357	986609	815461	538172	556566	964814
Deletions	177109	44998	224433	226645	146989	261152	230829	164884	162028	179230	158433	183424	165388	107234	108549	162028
Insertions	131727	27092	168477	160626	112802	154466	162062	125417	136350	140541	132155	145556	113381	82711	81981	136350
Number of unmatched bases NOT recorded as mutations																
Mismatches	3157986	1513875	3396315	7640121	2154652	2553247	3919457	2826810	2878867	4395990	2559375	3714116	2365366	1816836	1698087	2878867
Deletions	433984	312370	457459	1376442	242174	282513	557404	362832	358580	717014	294333	520810	284810	197169	141237	358580
Insertions	266229	197398	276106	864876	141745	167496	352234	216563	217422	437820	182852	320335	173968	123725	86426	217422
Average Read Length	273	4.1	277	279	273	272	274	274	272	271	271	275	274	265	264	272
Average Coverage (genome)	7	0.8	8	8	7	8	7	6	6	7	6	7	7	7	6	6
Average Coverage (ROI)	225	7.5	232	237	224	229	223	212	219	222	217	234	235	228	218	219
Reads on Target (%)	57	1.6	59	59	57	58	57	54	55	57	55	58	58	56	54	55

Variant Detection

The alpha globin locus on chromosome 16 contains five globin genes of similar size, structure and sequence. The most abundantly expressed genes, *HBA1* and *HBA2*, share very high homology and other highly homologous regions are common across

the alpha globin gene cluster. Deletions are common causes of alpha thalassaemia and this is believed to be due to non-homologous recombination occurring between these positions (See Results Chapter 3). Carriers of alpha thalassaemia variants are protected from the severe effects of *p.falciparum* malaria and this has increased its prevalence to as high as 1:3 in some populations where *p.falciparum* malaria is endemic. Being able to detect these common alpha thalassaemia variants is important if we are to offer a single diagnostic test for all haemoglobinopathies.

The aim of this sequencing run was to evaluate the capability of the capture library (which now covered a larger part of chromosome 16 than the previous library) to resolve large structural variants affecting the alpha globin gene locus. The previous capture library only covered chr16: 60,000 to 260,000 bp, whereas the new capture library extends from chr16: 60,000 to 2,000,000 bp. Baits cannot be designed against the telomeric region of chromosome 16 (0 to 60,000 bp). Known alpha zero and alpha plus variants commonly encountered in the diagnostic laboratory were used as positive controls. All heterozygous variants (test samples and positive controls) could be identified from the RPKM plots (Table 48), with the use of the additional negative controls afforded by resequencing the samples and including negative controls from a subsequent run (MiSeq Run 6). The coverage depth achieved per sample in the 14 sample pool was adequate for variant detection with the introduction of these additional negative controls. Many variants in the run could be identified broadly, based on just the two duplicates of the negative control in MiSeq Run 5a and 5b. The (α-$^{3.7}$/) deletion and its counterpart, the triplicated alpha ($\alpha\alpha\alpha$/) variant were exceptions to this (see Table 48 and also later section devoted to these variants) until the additional negative controls from MiSeq Run 6 were added. It appeared that for these two variants, several controls would be necessary to reliably detect their presence.

With the additional control data, all variants could be identified easily in the RPKM data (with the exception of the (α-$^{3.7}$/) deletion). Despite this, not all of the variants could be characterised with to-the-base accuracy (Table 48). In these cases, the breakpoint sequences were either within repetitive regions and not captured, or situated in regions with such high homology to one another that the mutant sequences could not be distinguished from the expected sequence at either breakpoint position in the reference sequence.

Table 48: Variant Characterisation in MiSeq Run 5b. NB: Same direction read data was not analysed in positive control samples that were known not to involve inversions.

Sample		Mutations	Visible in RPKM plot	Visible in Opposite Directrion Read report	Visible in Same Direction Read Report	To-The-Base Characterization achieved
MiSeq Sample 22	Test sample	Duplication	Yes	No	No	No
MiSeq Sample 23	+ Cntrl ($\alpha\alpha/--^{FIL}$)	Deletion	Yes	No	-	No
MiSeq Sample 24	+ Cntrl ($\alpha\alpha/--^{Thai}$)	Deletion	Yes	No	-	No
MiSeq Sample 25	+ Cntrl ($\alpha^{-3.7}/\alpha\alpha$)	Deletion	Yes	No	-	No
MiSeq Sample 26	+ Cntrl ($\alpha^{-3.7}/--^{SEA}$)	Compound heterozygous Deletions	$\alpha^{-3.7}$ No $--^{SEA}$ Yes	No No	-	No No
MiSeq Sample 27	+ Cntrl ($\alpha\alpha/--^{SEA}$)	Deletion	Yes	Yes	-	Yes
MiSeq Sample 28	+ Cntrl ($\alpha\alpha/--^{MED}$)	Deletion	Yes	Yes	-	Yes
MiSeq Sample 29	+ Cntrl ($\alpha\alpha/\alpha^{-4.2}$)	Deletion	Yes	No	-	No
MiSeq Sample 30	+ Cntrl ($\alpha\alpha/--^{20.5}$)	Deletion	Yes	Yes	-	Yes
MiSeq Sample 31	+ Cntrl ($\alpha\alpha/\alpha\alpha\alpha$)	Deletion	Yes	No	-	No
MiSeq Sample 32	- Cntrl ($\alpha\alpha/\alpha\alpha$)	None	-	-	-	-
MiSeq Sample 33	Test Sample	Deletion	Yes	No	No	Yes
MiSeq Sample 34	Test Sample	Duplication	Yes	No	No	Yes
MiSeq Sample 35	Test Sample	Duplication	Yes	No	No	No

MiSeq Run 5b Coverage Graphs Chromosome 16

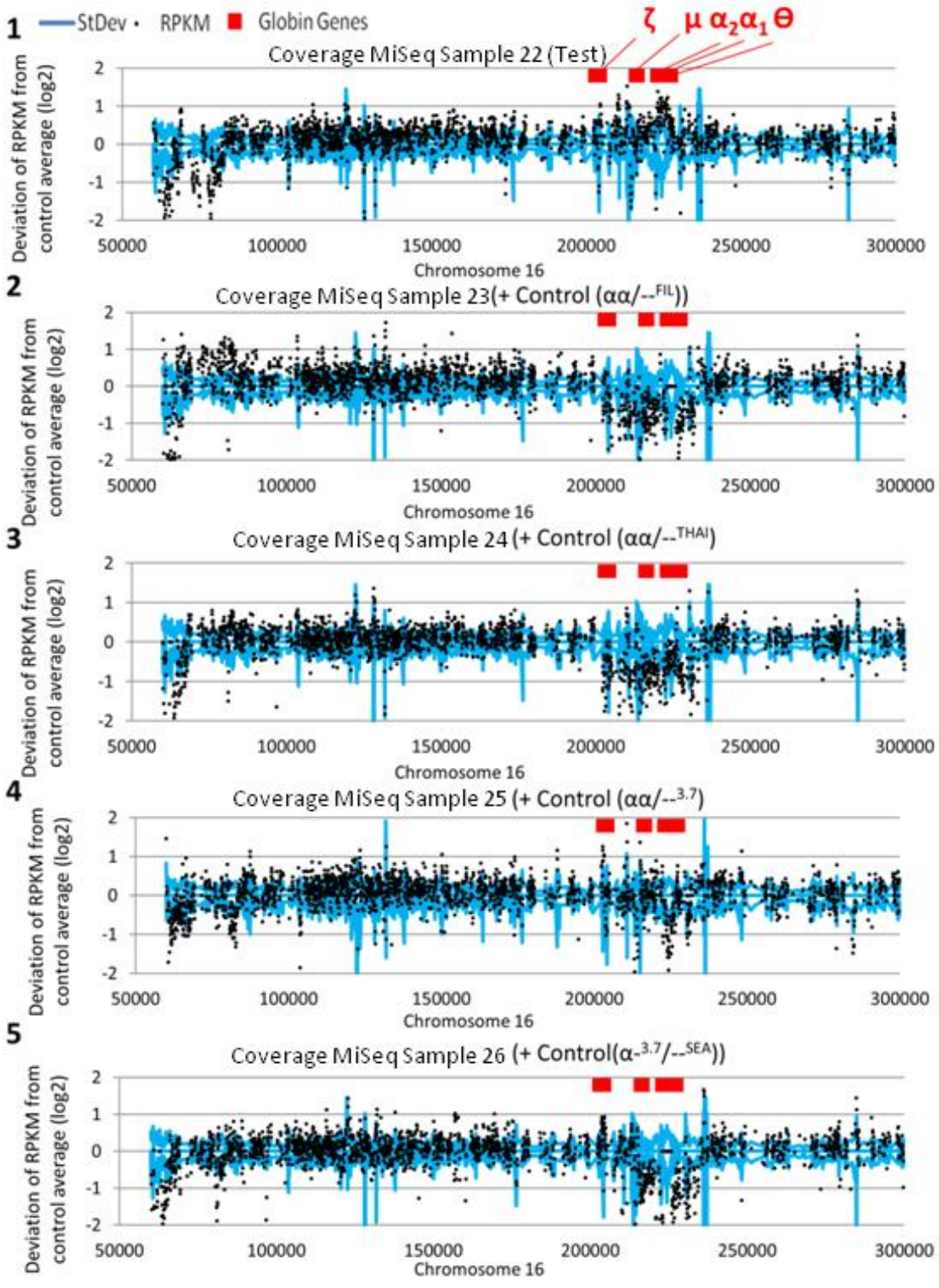


Figure 71: MiSeq Run 5b Coverage Graphs Chromosome 16. Plots show depth of coverage (Y axis) at each bait-covered position on chromosome 16 (X axis) in relation to the average coverage of that position in negative controls on a log2 scale. Combined NegC StDev from this run and the negative controls in MiSeq Run 6 is shown in blue. The positions of the alpha globin genes are shown in red horizontal blocks. Graphs: (1) Test sample22 (2) positive control sample 23 (3) positive control sample 24 (4) positive control sample 25 (5) positive control sample 26.

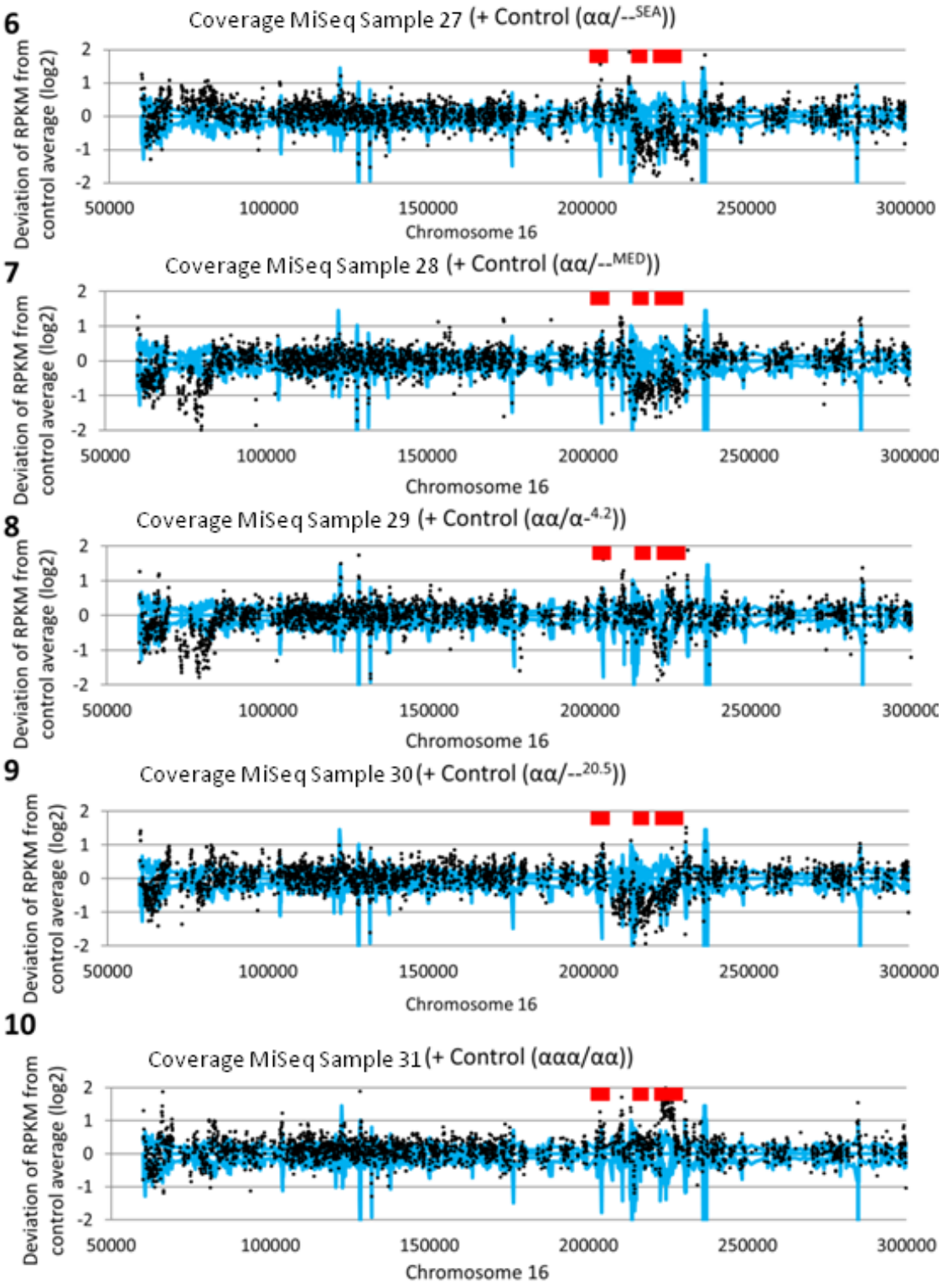


Figure 71 continued: (6) positive control sample 27 (7) positive control sample 28 (8) positive control sample 29 (9) positive control sample 30 (10) positive control sample 31.

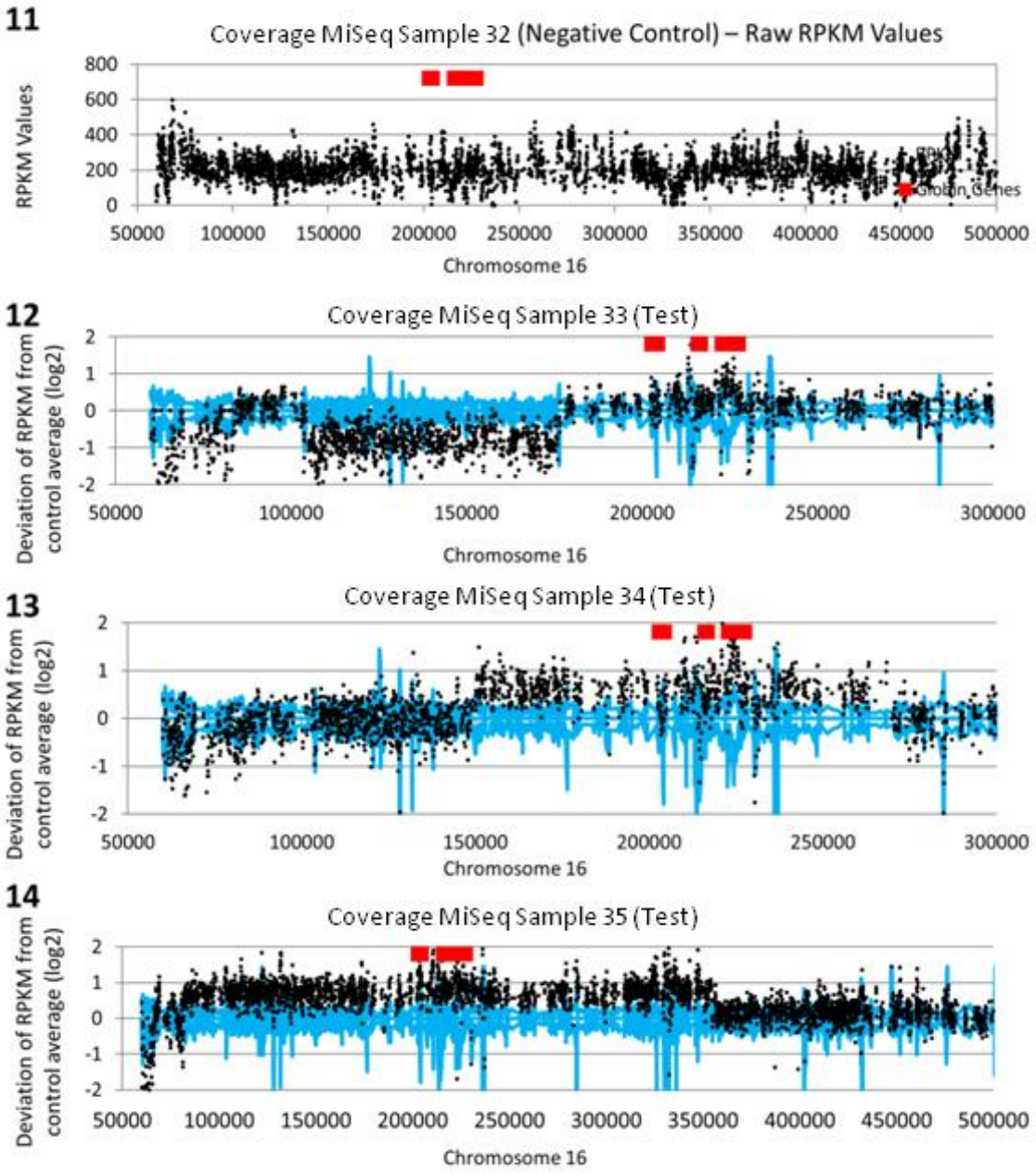


Figure 71 continued: (11) negative control sample 32: plot shows raw RPKM data rather than deviation from average (12) test sample 33 (13) test sample 34 (14) test sample 35: scale zoomed out to show entire rearrangement.

Variant Characterisation: Positive Controls

Ten positive controls for known rearrangements affecting the alpha thalassaemia locus were included in this run, comprising of eight heterozygous deletions, one insertion and one compound heterozygote with two known deletions affecting the same globin gene locus (Variant list: Table 42. Positions of known rearrangements in the run: Figure 72). All variants except the (α-$^{3.7}$/) deletion could be identified on RPKM plots. Three of the 10 alpha thalassaemia variants (--^{SEA}/, --^{MED}/, --$^{20.5}$/) could be identified by pile-up of reads at the break point locations in the opposite direction read report. These

rearrangements could also be identified with to-the-base accuracy by inspection of the read pile-up in the NextGene Viewer at the break point positions estimated by the RPKM data. Some break point reads were also included in the opposite direction read report. No inversions were included in the run (no inversion rearrangements that affect the alpha globin gene cluster have been reported), so same direction read data was not analysed for the positive control samples.

To-the-base characterisation was not permissible where the rearrangement break-points were situated in repetitive regions not covered by the bait design, or where the homology between the start and end points of the rearrangement was too high. The alpha globin gene cluster contains many highly homologous sequences, including the globin genes themselves and many Alu repeats. Even if the two homologous regions are included in the bait design, they may be so alike that breakpoint-crossing reads show very little deviation from the reference sequence. If multiple other regions include this homologous sequence, this can cause a higher level of noise and misalignment between these positions due to inadequate alignment by NextGene, further obscuring any misalignment genuinely related to a rearrangement breakpoint.

Six variants with known breakpoints were included in this run. The three that showed the highest degree of sequence homology around their breakpoints could not be characterised with to-the-base accuracy. The three variants with a lower degree of homology between their break point regions were characterised successfully. Figure 73 shows Pip plots for the homology of the breakpoint regions of these deletions: 1 Kb sections of the reference sequence around each pair of breakpoints are compared. A graph is created where the X-axis represents the first breakpoint region from base 1-1000, and the Y-axis represents breakpoint region 2 from base 1-1000. A dot in the graph area represents a position of homology. A line in the graph area shows an extended region of homology. The pip plot for the (--^{Thai}/) deletion shows that both this deletion's breakpoints occur in completely homologous sequences. The same is true for the (α-$^{3.7}$/) deletion (and also the (ααα/) insertion). The (--^{FIL}/) deletion has moderately homologous break point regions. The extreme homology between the approximate breakpoint regions of the (α-$^{3.7}$/) deletion and the triple alpha insertion made these variants challenging to identify in the RPKM data. The lack of identifiable break point sequences in conjunction with this was concerning. These variants are common, and it would be essential for a diagnostic assay to be capable of detecting them. The (α-$^{4.2}$/) deletion is also believed to have homologous breakpoints, but was not included here as the breakpoints have not yet been reported in the literature.

In MiSeq Run 5b, the (--^{SEA}/) deletion in a normal heterozygous state (Positive control Sample 6) could be resolved with to-the-base accuracy, as could the (--^{MED}/) (Positive control Sample 7) and (--$^{20.5}$/) (Positive control Sample 9) deletions. A previously identified duplication of the beta globin cluster on chromosome 11 (Positive Control Sample 4 - not shown) could also be identified successfully with to-the-base accuracy, conferring with previous sequencing data (See Results: MiSeq Run 2, Sample 3).

Positions of Variants Included as Positive Control Samples in MiSeq Run 5 in UCSC Genome Browser

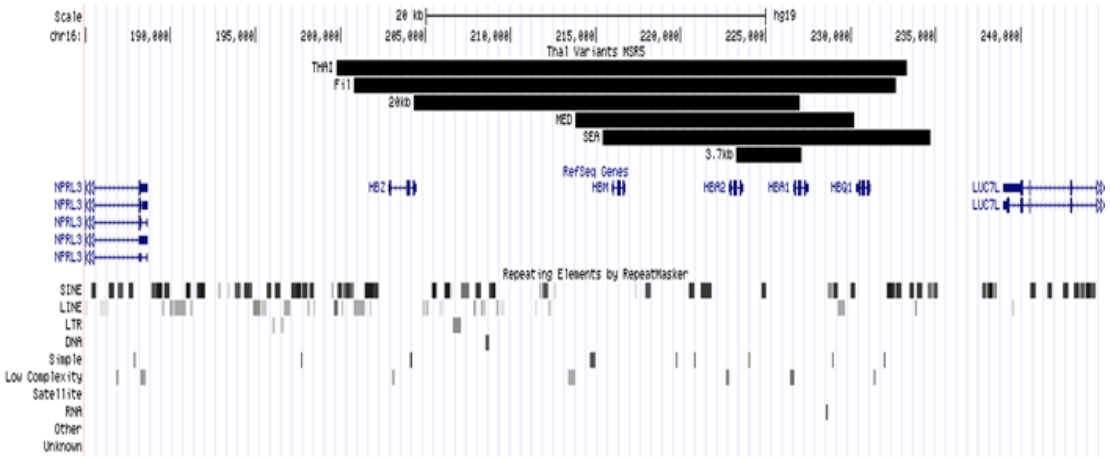


Figure 72: Positions of Variants Included as Positive Control Samples in MiSeq Run 5 in UCSC Genome Browser. The region removed by each variant is depicted by a black bar, with the variant name to the left. NB: The (α-$^{4.2}$/) deletion is not included as its breakpoints are not defined in HbVar. The (ααα/) variant affects the same region as the (α-$^{3.7}$/) deletion.

Homology of Breakpoint Sequences in MiSeq Run 5 Positive Control Variants

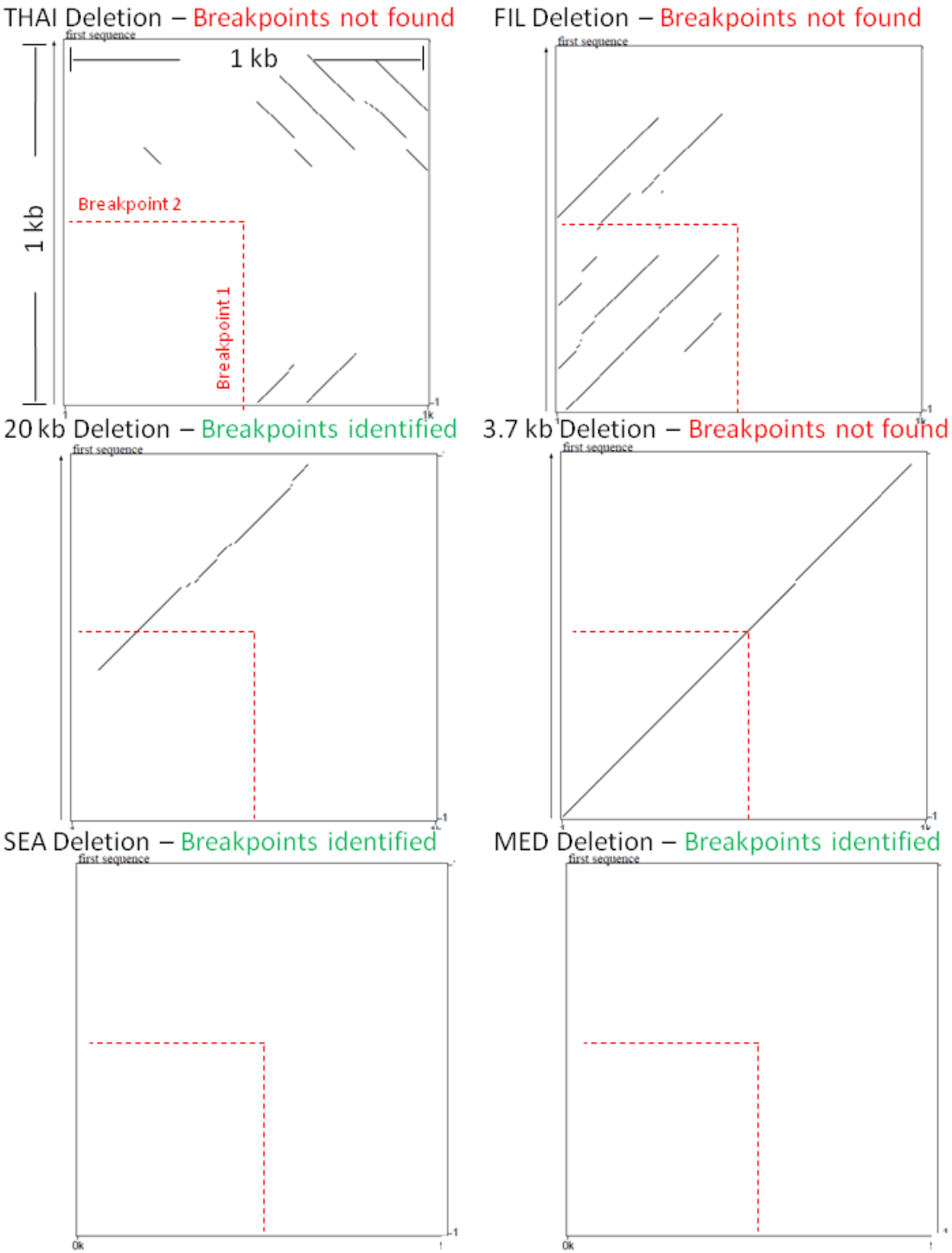


Figure 73: Homology of Breakpoint Sequences in MiSeq Run 5 Positive Control Variants. Each square plot concerns a different variant. The X axis represents 1 Kb of sequence centring on the 5' break point. The Y axis represents 1 Kb of sequence centring on the 3' break point. A red dotted line indicates the breakpoint position on each axis, and thus the position in the graph area at which they intersect. The graph area depicts homologous parts of these sequences as dots and lines. The (α-$^{3.7}$*I*) deletion breakpoints have almost absolute homology, a single diagonal line running across the entire length of the plot. The (--^{THAI}*I*), (--$^{20\ Kb}$*I*) and (--^{FIL}*I*) deletion breakpoints are adjacent to highly homologous regions. The (--^{SEA}*I*) and (--^{MED}*I*) deletion breakpoint regions have no homology. The variants where breakpoint sequences could not be identified show a higher level of homology to one another than the variants where to-the-base characterisation was achieved.

Detecting the (α-$^{3.7}$/) Deletion and (ααα/) Insertion

The most commonly occurring rearrangement of the alpha globin gene cluster is non-homologous recombination between the homologous 'z box' regions of *HBA1* and *HBA2*. The recombination event results in a 3,804 bp deletion on one allele (creating a single functional alpha globin gene by joining the remaining exons of *HBA1* and *HBA2* together) and a 3,804 bp insertion (creating an additional alpha globin gene) on the other allele. These are referred to as the (α-$^{3.7}$/) deletion (HbVar.1076) and triple alpha (ααα/) (Figure 74). The (α-$^{3.7}$/) deletion is the most common variant affecting the alpha globin gene locus. Co-inheritance of this variant with an alpha zero thalassaemia can result in HbH disease. The (ααα/) insertion can have clinical consequences when co-inherited with beta thalassaemia.

Formation of the (α-$^{3.7}$/) Deletion and (ααα/) Insertion via Misalignment and Reciprocal Crossover during Meiosis

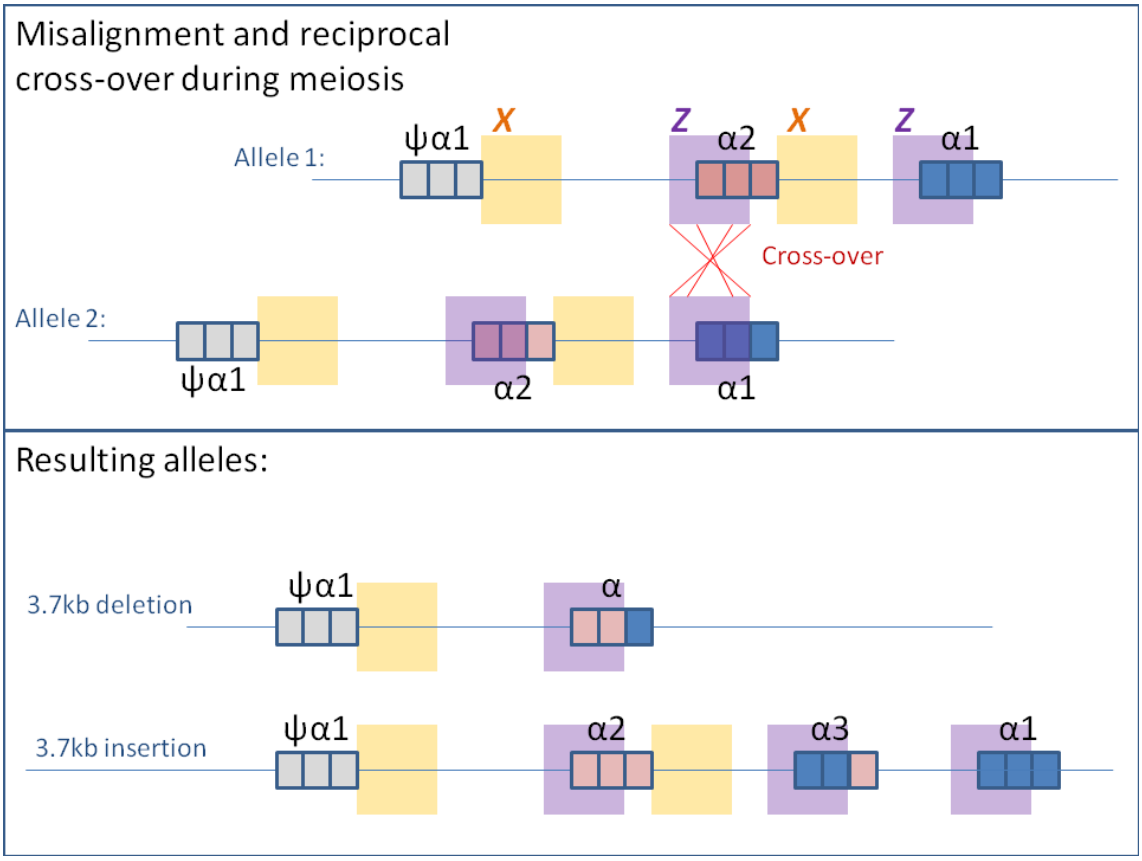


Figure 74: Formation of the (α-$^{3.7}$/) Deletion and (ααα/) Insertion via Misalignment and Reciprocal Crossover during Meiosis. The cross-over results in alleles with a reciprocal deletion and insertion of 3804bp, and the triplicated alpha globin gene (ααα/α-$^{3.7}$). Recombination between the X-box homologous regions results in the (α-$^{4.2}$/) alpha thalassaemia deletion, and its reciprocal (ααα$^{anti-4.2}$/).

Detection of these variants via the analysis techniques developed to this point in this study was not possible, due to the high degree of homology between the two alpha

globin genes (Figure 75) which cause multiple alignment issues in NextGene. Issues with the strategy used for other rearrangements characterised during this study are listed below.

DotPlot Showing Homology between *HBA1* and *HBA2* Gene Regions

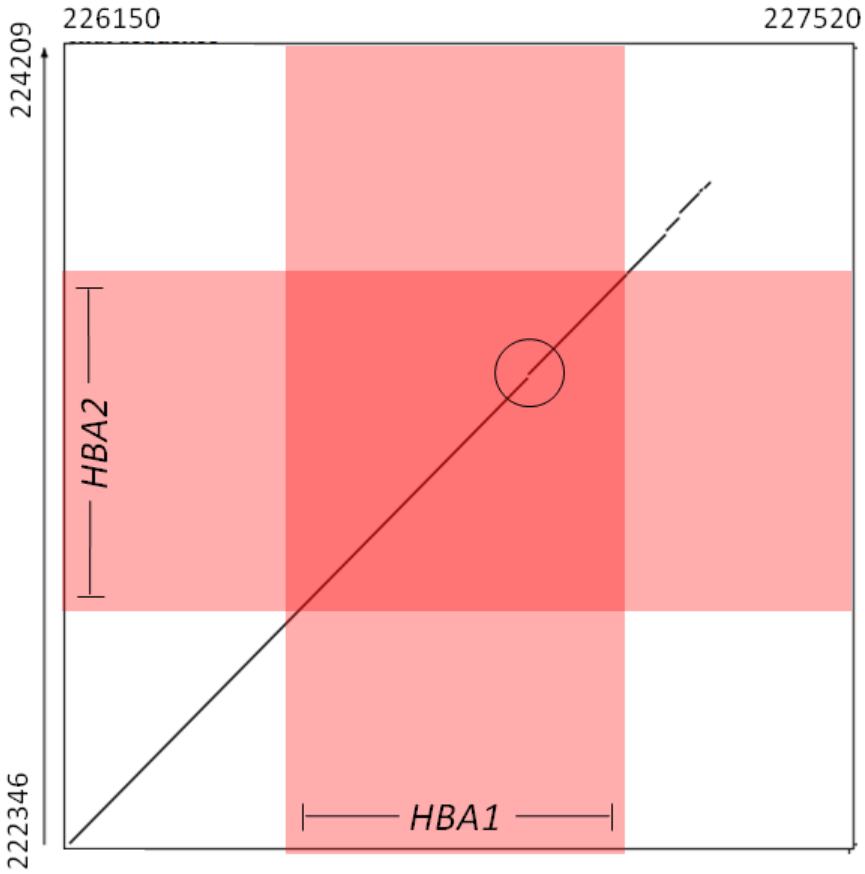


Figure 75: Dot Plot Showing Homology between *HBA2* and *HBA1*. The X axis represents the HBA2 gene and 1 Kb of surrounding sequence. The Y axis represents the HBA1 gene and 1 Kb of surrounding sequence. Homologous stretches of sequence are represented by dots or lines in the chart area. Between HBA2 and HBA1 only a single base change breaks the homology of the two sequences (circled).

NextGene Viewer

These variants bring together (or insert) homologous sequences, leaving no unique signature. As such, like the (--^{Thai}/) and (--^{FIL}/) deletions, these variants cannot be identified in the NextGene Viewer, as none of the reads that capture the rearrangement breakpoint produce any detectable misalignment (Figure 76).

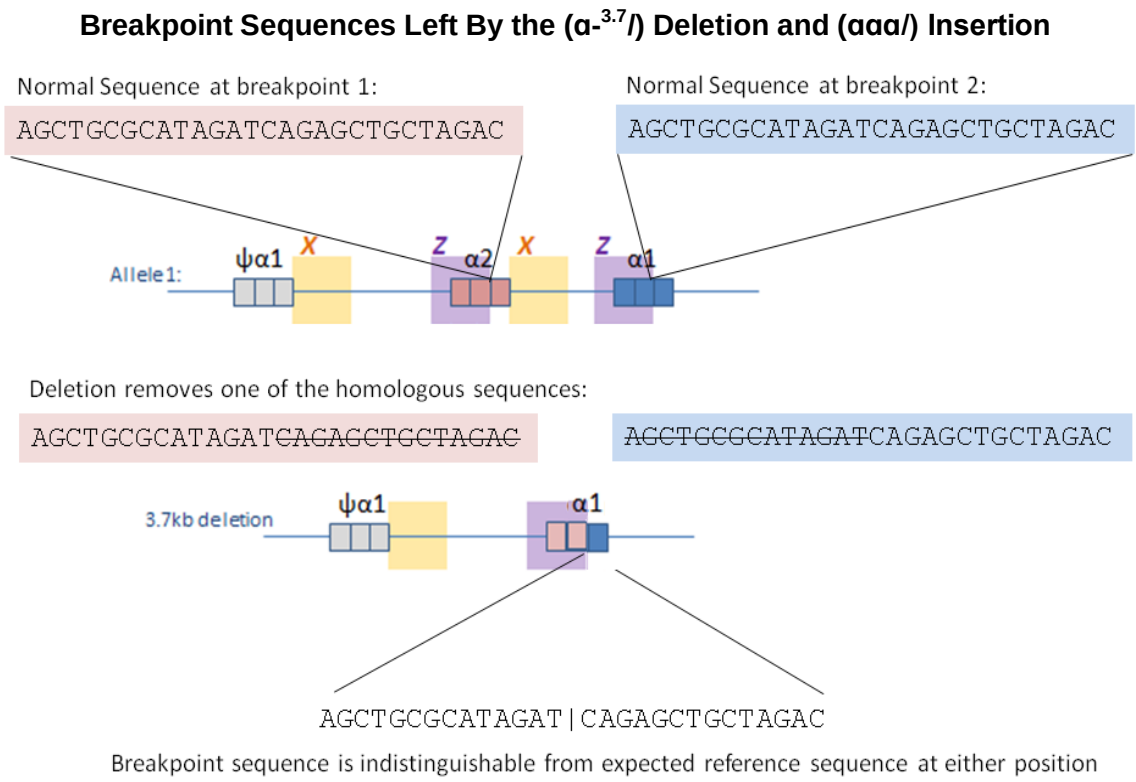


Figure 76: Breakpoint Sequences Left by the (α-$^{3.7}$/) Deletion and (ααα/) Insertion. The breakpoint sequences resulting from these rearrangements are indistinguishable from the expected normal sequence, and as such align perfectly to the reference sequence.

Opposite Direction Read Data:

The homology between *HBA1* and *HBA2* confounds the data in the opposite direction read report at this location by simultaneously creating multiple false positives in balanced alignments, and false negatives in alignments containing these rearrangements. False positives (reads which are rejected from normal alignment as opposite direction reads when they are in fact from normal sequence) occur when NextGene erroneously assigns one half of a read pair to *HBA1* and the other half to *HBA2*. Pile-up of opposite direction read pairs between the two alpha globin genes is high both in samples and in negative controls (Table 49). False negatives (reads which originate from fragments that crossed the rearrangement break point) are created when NextGene erroneously aligns both reads to the same alpha globin gene and accepts them into the aligned data.

Table 49: Opposite direction reads between *HBA1* and *HBA2* in samples with and without 3.7 variants. Data shown is (i) the number of opposite direction reads for each sample where positions are within 1 Kb of the *HBA1* and *HBA2* genes and (ii) the number of those reads that show the expected gap distance for the 3.7 Kb rearrangements.

Samples in MiSeq Run 5b	**(αα/αα)**	**(αα/αα)**	**(αα/αα)**	**(αα/αα)**	**(ααα/αα)**	**(α-$^{3.7}$/αα)**
Opposite direction reads aligning to *HBA1*/*HBA2* region	9	16	6	20	16	16
Opposite direction reads aligning to *HBA1*/*HBA2* region with a 3000-4500 bp gap distance	0	15	5	15	1	11

Thus, the data cannot be used to identify these rearrangements due to the high level of noise, compounded by the reduced level of genuine reads recorded between these two regions.

Same Direction Read data is not applicable in this case because the rearrangements do not involve sequence inversion.

Mutation report

SNP Detection using the mutation report may bolster evidence of a rearrangement from the RPKM data, but single nucleotide changes do not occur with a high enough frequency (average of one per Kb) to reliably identify these rearrangements. It is unlikely that enough SNPs would occur within this region on the balanced allele that it would produce a notably large string of homozygous SNPs to indicate a deletion. SNPs showing an allele frequency indicative of the insertion (1:2) would also be unlikely to occur with a high enough frequency to identify this region.

RPKM

RPKM plots are the only means available through NextGene capable of detecting these two variants. However, the sequence that these rearrangements remove/insert is identical to the sequence that is expected to follow the breakpoint in the normal genome (Figure 76). This means that this rearrangement can only be identified by the dosage change it creates. The high homology of the two sequences also means that the impact of the dosage change is diluted, as it is spread between the sequences of *HBA1* and *HBA2*. The fact that misalignment is frequent between these positions is

evidenced by the number of opposite direction reads aligning across the two regions - even in negative control samples. Reliably identifying this variant requires multiple negative controls to compare the amount of sequence captured in the affected region to other normal regions within the affected sample. In this run, only the negative control and Test Sample 33 have a normal dosage of sequence at this position.

MiSeq Run 5 includes samples with the following rearrangements: (α-$^{3.7}$/αα) deletion heterozygote, triplicated alpha (ααα/αα) heterozygote, (α-$^{3.7}$/--^{SEA}) compound heterozygote. The run only contained a single negative control for the alpha globin gene cluster, which was insufficient for detecting these rearrangements in RPKM plots.

MiSeq Run 6 contained a large number of samples which could be used as negative controls for alpha globin rearrangements. Crucially, both runs contained the same number of samples, meaning that the coverage per sample was comparable between the runs. Inter-run variability was still high, with the normal control from MiSeq Run 5 outside the standard deviation of the negative control average ('NegC StDev') from MiSeq Run 6 at 66% of bait positions (Figure 77).

MiSeq Run 5 Negative Control Deviation from MiSeq Run 6 Negative Control Average

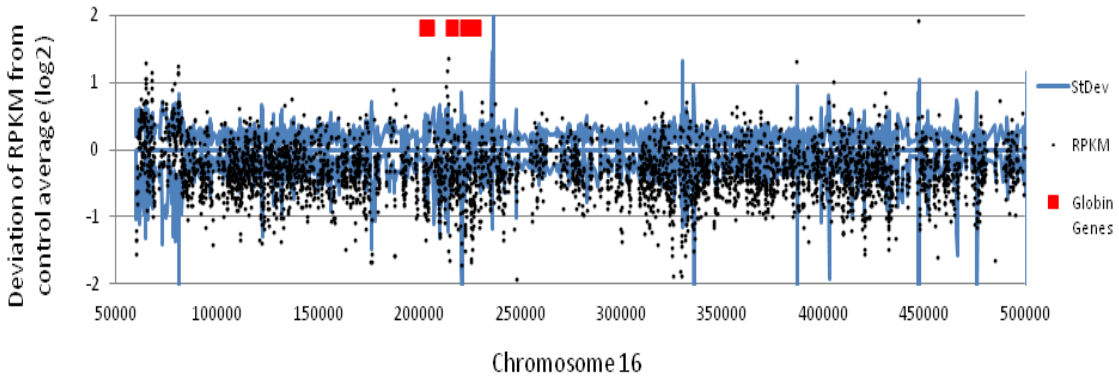


Figure 77: MiSeq Run 5 Negative Control Deviation from MiSeq Run 6 Negative Control Average. RPKM values for MiSeq Run 5 (black) and the negative control average of MiSeq Run 6 (blue).

Combining these controls with the negative control from MiSeq Run 5 improved the visibility of the (α-$^{3.7}$/) rearrangements. The deletion and insertion could be identified visually in heterozygous states. In compound heterozygous states, the region that was absent on both chromosomes had an RPKM value of zero. As calculating the difference in the sample from the negative control average for these positions resulted in a 'divide by zero' error, it was necessary to assign an arbitrary value to baits that returned this error so that they could be plotted correctly on the graph. The arbitrary value assigned was -5. This improved visibility of the (α-$^{3.7}$/) rearrangement when it

occurred in a compound heterozygous state. As compound heterozygotes showed complete absence of sequence in the smaller region affected by both rearrangements, they could be clearly distinguished from single copies of either rearrangement (Figure 78).

The number of RPKM values falling outside the NegC StDev values was compared between the region of the rearrangement and the same number of bait positions from a balanced region. The effect of the number of negative controls used to calculate NegC StDev on these regions was investigated in a negative control, (αα/) duplication (het) and (α-$^{3.7}$/) deletion (het) samples. As the number of negative controls used to calculate NegC StDev was decreased, the number of baits falling outside these values from the balanced region increased, while in the affected region they decreased (Table 50). Three negative controls provided sufficient data to discern deletions from balanced regions. The insertion remained harder to detect. This is, in part, because the relative dosage change caused by a heterozygous duplication (30%) is smaller than a heterozygous deletion (50%), meaning that fewer affected baits can be distinguished from the noise across the region. For this reason, it may be prudent to increase the number of negative controls either included in the same runs, or from other runs to more clearly define consistent noise across the region.

Table 50: Effect of negative control number on number of positions exceeding NegC StDev in test and control regions and samples. 'Del' indicates the (α-$^{3.7}$/) positive control, 'dup' indicates the (ααα/) insertion positive control, and 'Neg' indicates an additional negative control. Data shows how number of positions exceeding NegC StDev increases as *N*(negative controls) increases in the balanced region, and decreases in the affected region in positive controls.

Affect of negative control number on number of positions exceeding NegC StDev in test and control regions and samples															
Negative controls:	N=6			N=5			N=4			N=3			N=2		
Rearrangement:	Del	Dup	Neg	Del	Dup	Neg	Del	Dup	Neg	Del	Dup	Neg	Del	Dup	Neg
Affected Region (total baits = 61)															
# Baits with RPKM within NegC StDev	3	15	40	0	17	36	3	16	38	6	18	34	18	22	37
# Baits with RPKM outside NegC StDev	58	46	21	61	44	25	58	45	23	55	43	27	43	39	24
(%)	**95**	**75**	**34**	**100**	**72**	**41**	**95**	**74**	**38**	**90**	**70**	**44**	**70**	**64**	**39**
Balanced Region (total baits = 104)															
# Baits with RPKM within NegC StDev	67	68	77	53	69	71	58	65	67	64	69	64	58	60	55
# Baits with RPKM outside NegC StDev	37	36	27	51	35	33	46	39	37	40	35	40	46	44	49
(%)	**36**	**35**	**26**	**49**	**34**	**32**	**44**	**38**	**36**	**38**	**34**	**38**	**44**	**42**	**47**

The exact breakpoints of the (α-$^{3.7}$/) deletion are not known, as they do not produce any signature sequence which does not match the reference. They may not be the same in every case of the deletion, occurring at any pair of matching points between the two iterations of the *Z* box. For this reason, diagnosis must rely on RPKM values alone, when they indicate a rearrangement corresponding to the approximate size and position of the (α-$^{3.7}$/) rearrangements, and by the lack of break point sequences in the read pile up. For the purposes of diagnosis, lack of break point sequences indicates that regardless of where the break has occurred, the new gene created by the rearrangement will be functional. The position of the insertion was less clear than the position of the deletion, and a larger number of controls may be necessary for their reliable detection than for the deletion (Figure 79).

Detection of the (α-$^{3.7}$/) Deletion in Heterozygous and Compound Heterozygous States with Varying Numbers of Negative Controls (enlarged image overleaf)

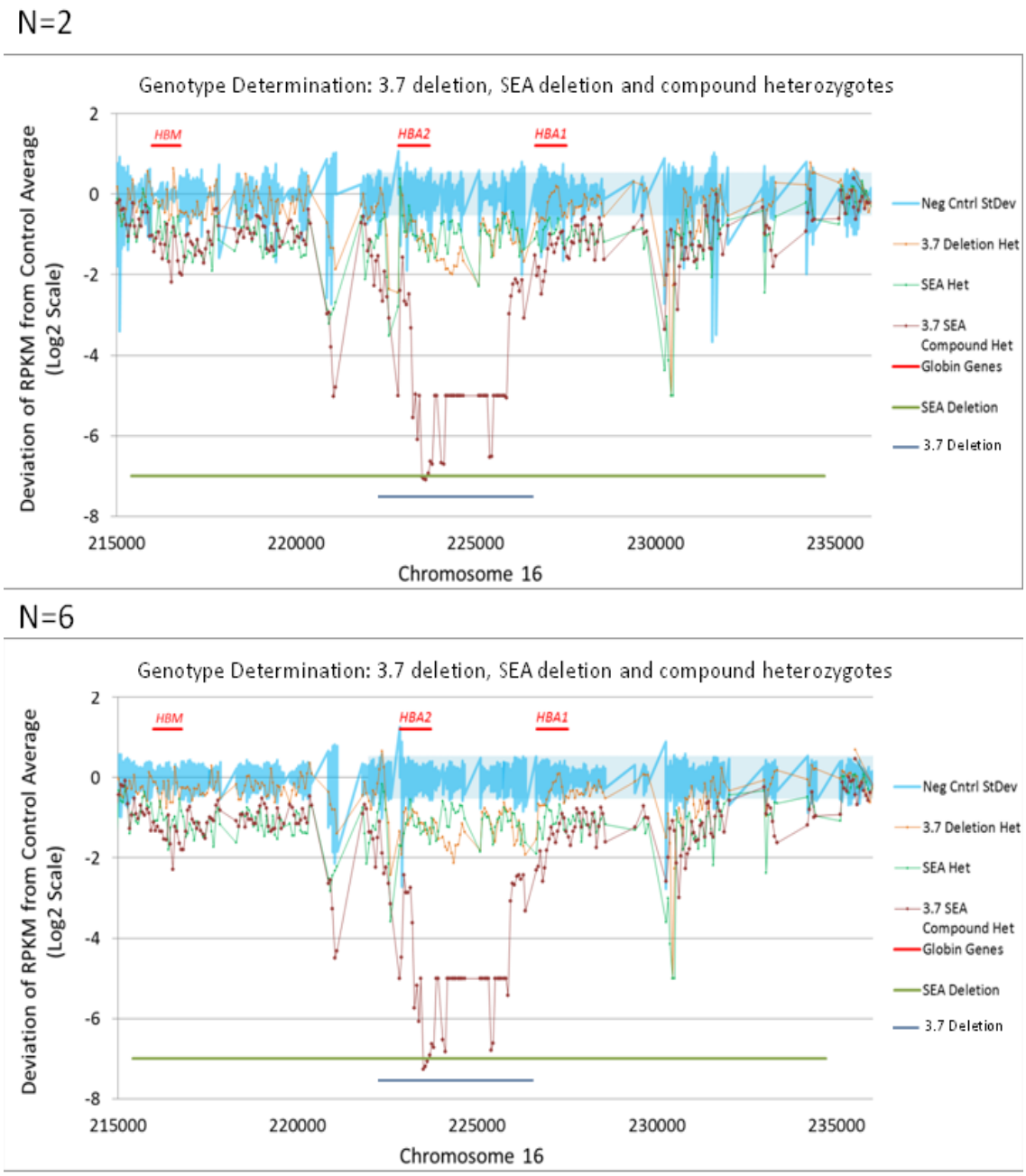


Figure 78: Detection of The (α-$^{3.7}$/) Deletion in Heterozygous and Compound Heterozygous States with Varying Numbers of Negative Controls. Both graphs: light blue line indicates standard deviation in the negative controls from the negative control average. Orange line indicates RPKM value per bait position in a sample heterozygous for the (α-$^{3.7}$/) deletion. Green line indicates RPKM value per bait position in a sample heterozygous for the (--^{SEA}/) deletion. Brown line indicates RPKM value per bait position in a compound heterozygous (α-$^{3.7}$/--^{SEA}) sample. Positions of the globin genes are indicated by red bars. Known region affected by the (α-$^{3.7}$/) rearrangement is indicated by the blue bar at the bottom of the graph, and the ($^{-SEA}$/) deletion by a green line. X axis shows position on chromosome 16 and Y axis shows deviation from negative control average on a Log2 scale. Upper panel shows data when two negative controls are used. Lower panel shows data when 6 negative controls are used.

Simplified Version of Figure 78

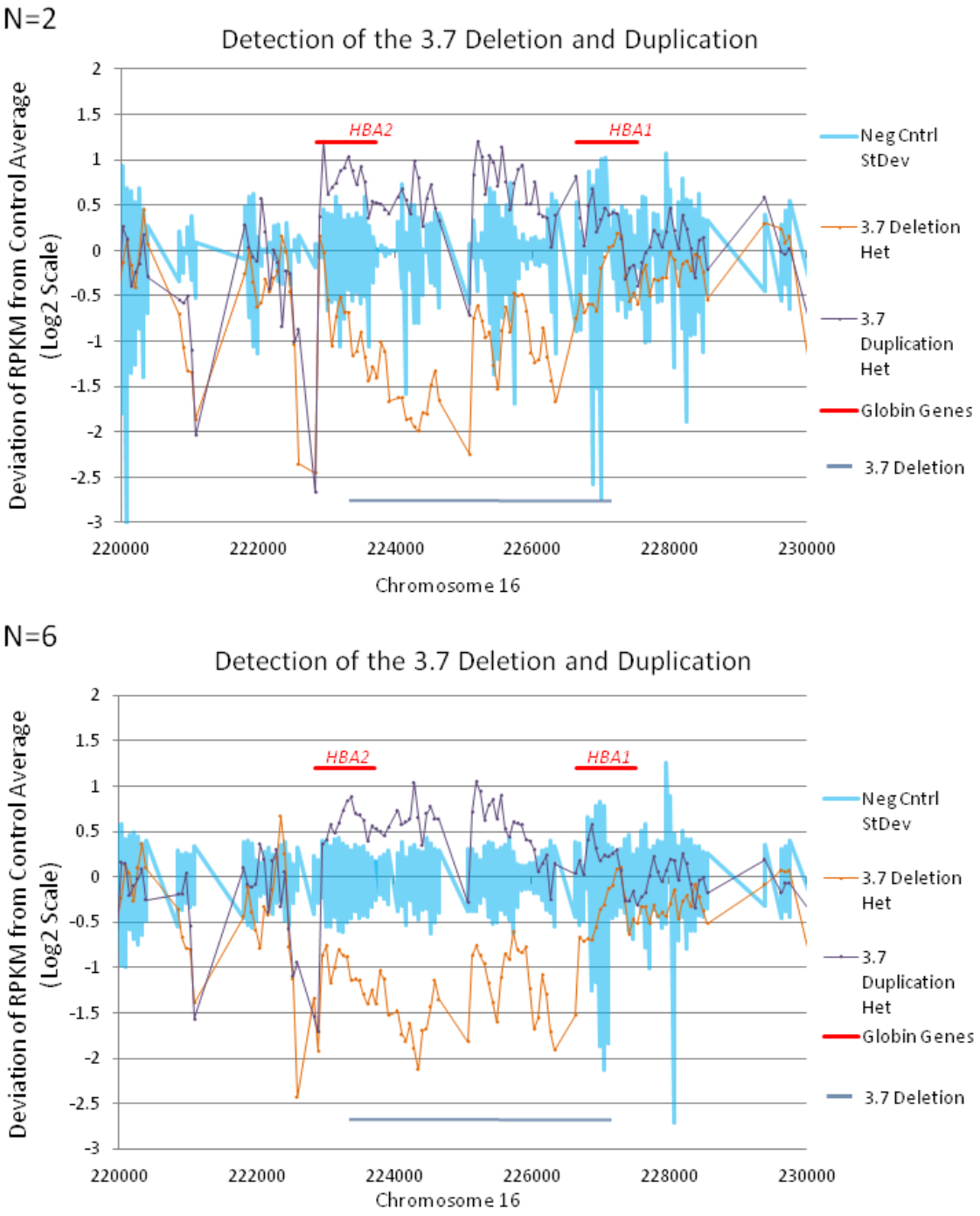


Figure 79: Simplified Version of Figure 78. Graphs focus on region affected by (α-$^{3.7}$/) Deletion and (ααα/) Insertion, showing distinction between heterozygous deletions and limits of NegC StDev.

Variant Characterisation: Test Samples

Run 5b included four test samples with novel rearrangements affecting the alpha globin gene locus. All four rearrangements could clearly be identified from the RPKM plots. Three samples showed a novel copy number variant at the same position, with a pile-up of reads with strings of misaligned sequence at this position. In all three cases this

copy number variant removed different numbers of an iteration of the same sequence which is repeated multiple times at this location (Figure 80). This region was highly homologous to the same chromosomal location on chr18 and also to regions on chrX and chrY. This CNV presented as misaligned sequences in the variant report, which were evaluated as potential breakpoints for the rearrangements investigated in two of these samples, increasing the time required to complete characterisation of those variants.

A Copy Number Variant in Three Test Samples from MiSeq Run 5b

Figure 80: A Copy Number Variant in Three Test Samples from MiSeq Run 5b. Green and red text is used to show iterations of the same sequence. The grey highlighted region shows the region deleted by the copy number variant in MiSeq Samples 22, 33, 35.

MiSeq Sample 22 (Test)

MiSeq Sample 22 showed a duplication that began between positions Chr16: 82,980 and 83,274. This was consistent with the results of sequencing this sample on the HiSeq 2000 (HiSeq Run 1 Sample 3). The duplication appeared to extend beyond the bait covered region (>Chr16:2,097,990) making it >2Mb in length (Figure 81).

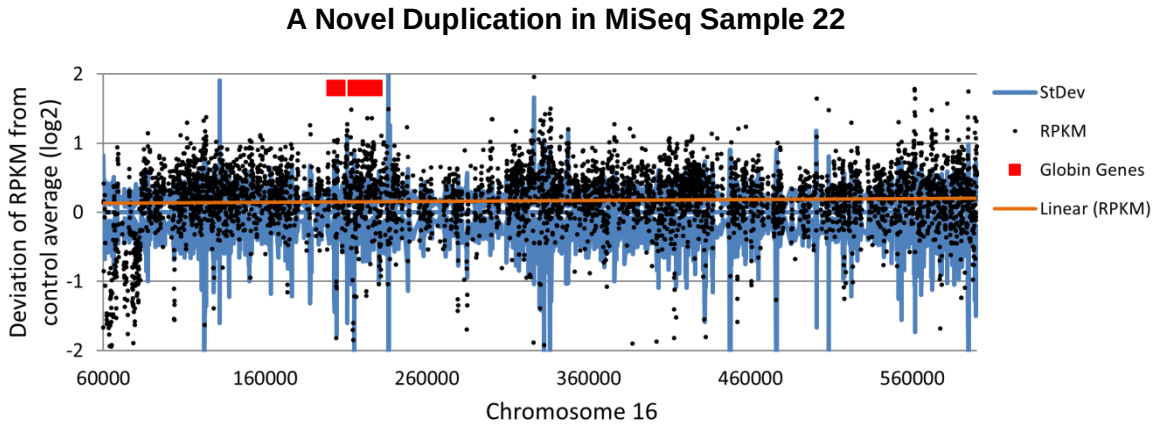


Figure 81: A Novel Duplication in MiSeq Sample 22. RPKM data indicates a duplication from approximately chr16:83,000 extending beyond the bait tiled region. A trend line is included to better illustrate the relative increase in coverage, due to the high level of coverage variability in this sample.

The mutation report showed a mix of 'duplicated' (1:2) and 'balanced' (1:1) allele frequencies for heterozygous SNPs occurring within the duplicated region which did not help inform variant characterisation. Opposite direction and same direction read reports and visual inspection of the alignment in NextGene failed to identify any misaligned read pile-up in the covered break-point region that could represent the start position of the duplication. Many small mismatches were found which BLAT query showed were spurious alignments from other regions of the genome with high homology to this region of chr16. The target region did show a small deletion of 43 bp. Without complete coverage of the region, or any candidate regions for the 5' breakpoint identified by the usual methodology, this variant could not be resolved.

MiSeq Sample 33 (Test)

MiSeq Sample 33 had been referred to the clinic for sequencing with the following red cell indices: HbA_2 2%, HbF 0.2%, RBC 6.1, Hb 133, MCV 71.4, MCH 21.7. This was consistent with an alpha thalassaemia phenotype. CGH array data had indicated that MiSeq Sample 33 had a deletion removing the tip of chromosome 16, up until approximately chr16:175,000. RPKM data for this sample showed a clear deletion of approximately 70 Kb ending at the same approximate position, but with a small intact region between the deletion and the telomere (Figure 82C). The RPKM data also showed a negative dosage change affecting the tip of the chromosome.

The telomeric region of the bait-tiled region also appears deleted (Figure 82B). However, as discussed previously, this region is highly variable and this degree of variation at this position has been seen in multiple samples. It is not possible to confirm the presence or absence of this region via Gap PCR, as rather than bringing together two sequences against which Gap PCR primers could be designed; in a telomere tip

deletion this sequence is simply absent. The start point of the possible telomeric deletion was also situated in an unmapped repetitive region. If the position had been covered by the bait design, it may have been identifiable by reads in the alignment that started at the exact same position, followed by an immediate drop in read depth, which would indicate that one copy of chromosome 16 ended abruptly at that position. In the absence of this data, the presence or absence of this region could not be determined by coverage or alignment data.

We also speculated that the RPKM data may indicate an inversion event had occurred between the small intact region and the balanced region, followed by a single telomeric deletion event removing both the deletion within the bait tiled region and the potentially deleted telomeric region of the bait-tiled region (Figure 82A).

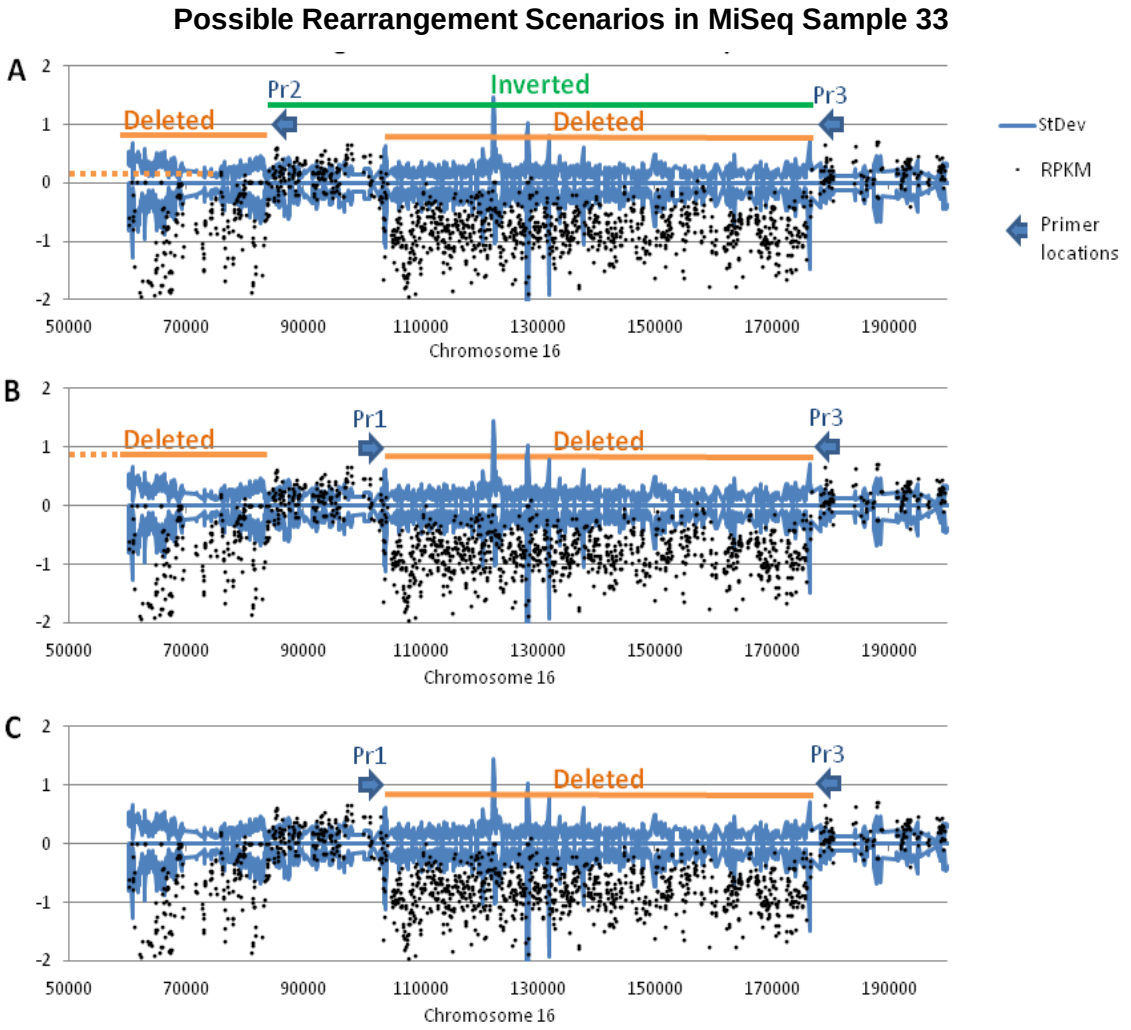


Figure 82 Possible Rearrangement Scenarios in MiSeq Sample 33, Based on RPKM Data Plot. (Scenario A) An inversion (green) occurs between the intact regions of the chromosome, followed by a single deletion (orange) removing the inverted sequence between chr16:110,000-175,000 and the telomeric region. (Scenario B) Two separate deletions have occurred, (Scenario C) a single deletion has occurred and the dosage change indicated by baits between positions chr16:60,000-80,000 is due to the bait variability seen in multiple samples at this region. Three primers were designed to determine whether the sequence between chr16:110,000-170,000 had been removed by a deletion (Pr1 and Pr3) or an inversion deletion (Pr2 and Pr3).

A telomeric deletion would not leave any misaligned sequence at the breakpoint or any reads in the opposite or same direction reads reports. The high homology between this region of chromosome 16 and other parts of the genome produced a high number of same direction reads partially aligning to these regions, but none gave a strong consensus towards a genuine rearrangement. Primers were designed across the break points of the deletion as estimated by the RPKM plot, and additionally between the intact sequence at position 84,174+ and the intact sequence after the deletion, to detect a potential inversion between these two points (Figure 82).

The primer pair that had been designed assuming a straightforward deletion (Pr1-Pr3) successfully amplified a PCR product. The primer pair that had been designed

assuming an inversion (Pr2-Pr3) did not make a product (Figure 83). Sanger sequencing analysis of the deletion product on the 3130 Sequencer revealed the deletion break point in several reactions. The deletion breakpoints were chr16: (104,934-104,943) - (177,954-177,966). The breakpoints shared 12 bp of homologous sequence, meaning the precise start and end points could not be determined beyond this accuracy (Figure 83). The deletion was 73 Kb in length.

Breakpoint Confirmation for a Novel Deletion in MiSeq Sample 33

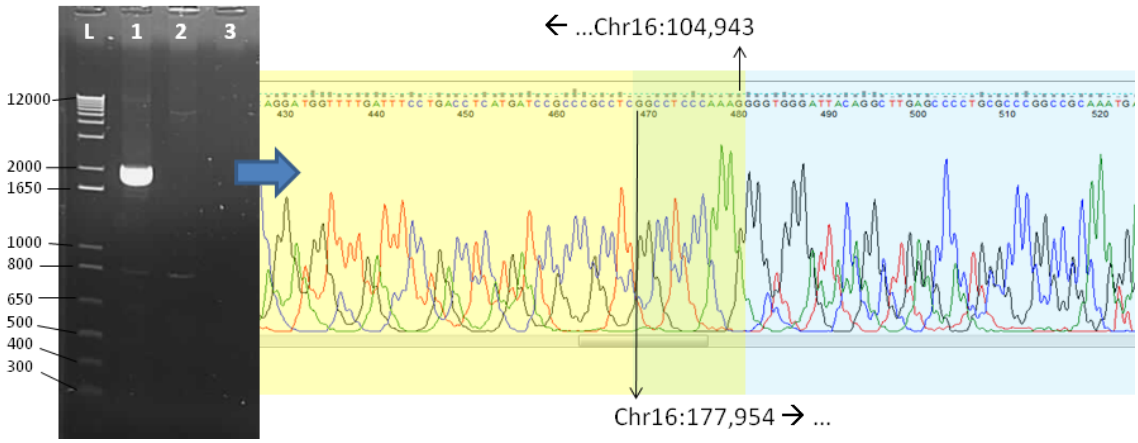


Figure 83: Breakpoint Confirmation for a Novel Deletion in MiSeq Sample 33. Gel Image and Chromatogram of Sequenced PCR Product from Pr1 and Pr3. Gel image: (L) 1 Kb+ Ladder (1) MiSeeq Sample 33 (2) Negative Control (3) No Template Control. Chromatogram: Base calls that match the reference sequence up to position 104,931 highlighted in yellow. Base calls that match the reference sequence from position 177,942 onwards highlighted in blue. The 12 bp ambiguous sequence is highlighted in green.

The deletion left the alpha globin gene cluster intact, but removed the main regulatory region, HS40, which is located 40 Kb upstream from *HBZ*. *MPG* and parts of *NPLR3* and *RHBDF1* were also removed (Figure 84A). The deletion breakpoints both occur in gene introns within Alu repeats (Figure 84B). The breakpoint itself is located in a highly homologous region. Multiple sequences in the surrounding region also show high homology to one another (Figure 84C).

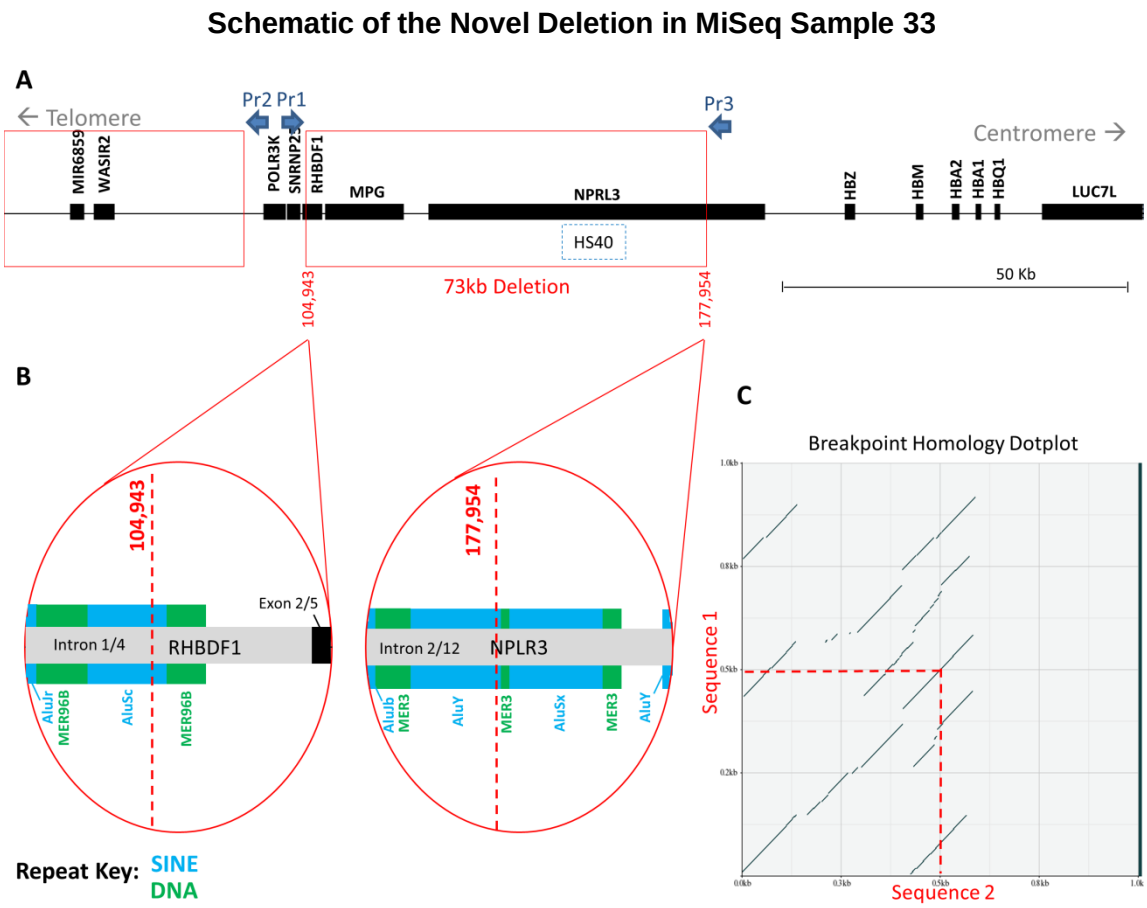


Figure 84: Schematic of the Novel Deletion in MiSeq Sample 33. (A) The deletion removes 73 Kb of sequence between positions 104,934-104,943 and 177,954-177,966. The deletion removes *MPG* and parts of *RHBDF1* and *NPLR3*. The HS40 alpha globin regulatory region is removed by the deletion, while the alpha globin gene cluster remains intact, (B) The deletion breakpoints occur in the introns of RHBDF1 and NPLR3 within two Alu Repeats. (C) Pip Plot showing homology between the 1 Kb of sequence surrounding each breakpoint.

MiSeq Sample 34 (Test)

MiSeq Sample 34 showed a duplication encompassing the alpha globin gene cluster. Both duplication breakpoints were situated within unmapped repetitive regions. The duplication was approximately 120 Kb within the region chr16: 148,000-270,000. At the centromeric break point of the duplication, two repeats occurred in close proximity. The covered region between the two repeats showed highly variable coverage, and it was not clear whether this region was part of the duplication or signified the start of the balanced region. Primers were designed against the unique sequence outside this region. The primers were designed with the assumption that this duplication was top-to-tail in orientation, as reported in the literature for all duplications of this cluster, have been. Several PCR conditions were tested, and one intermittently produced a 10 Kb band in the sample that did not appear in a negative control sample.

The PCR conditions were as follows:

Forward Primer sequence: TCTACTGCCCTCTCCCTTCA

Reverse Primer sequence: CAGGGTACGGCTTCTCTCAA

The reagent mix (per sample) was: 2.5 µl LongAmp taq buffer (NEB), 2.5 µl MgCl2 25mM (Qiagen), 0.1 µl LongAmp taq polymerase (NEB), 1 µl forward primer 10 pmol/µl. 1 µl reverse primer 10 pmol/µl, 9.9 µl molecular biology grade water, 4 µl DNTPs (Qiagen). 4 µl DNA template (providing approx 200ng total). The thermal cycling program for the reaction is listed in (Table 51).

Table 51: Thermal Cycling Conditions for Gap PCR Confirmation of Duplication in Test Sample 34.

Step	1	2 (30 cycles)			3	4
Temperature	94°C	94°C	58°C	68°C	68°C	10°C
Time	2 minutes	15 seconds	20 seconds	6 minutes	11 minutes	∞

This reaction produced a weak band with intermittent success. Final characterisation of the duplication breakpoints was performed by Dr Xunde Wang (Sickle Cell Branch, National Heart Lung and Blood Institute, NIH, USA) who cloned the PCR product into a pCR2.T/A cloning vector. Clones of with the correct size inserts were confirmed by RFLP analysis and sanger sequencing analysis using the primers M13F(-20), GTAAAACGACGGCCAGT; M13R-pUC(-40), CAGGAAACAGCTATGAC, followed by walking sequence primers, NPRL3-R1-seq, CAGAAATGACCAATCCCAGG; NPRL3-R2-seq, TCTTCTGGCACCACCTGTTC, NPRL3-R3-seq, GGCCCTTCTCTCCGCAGTTAA. This revealed that the breakpoints of the duplication were chr16:(148,451-148454)-(269,038-269,041) (Figure 85).

Breakpoint Confirmation for a Novel Deletion in MiSeq Sample 34

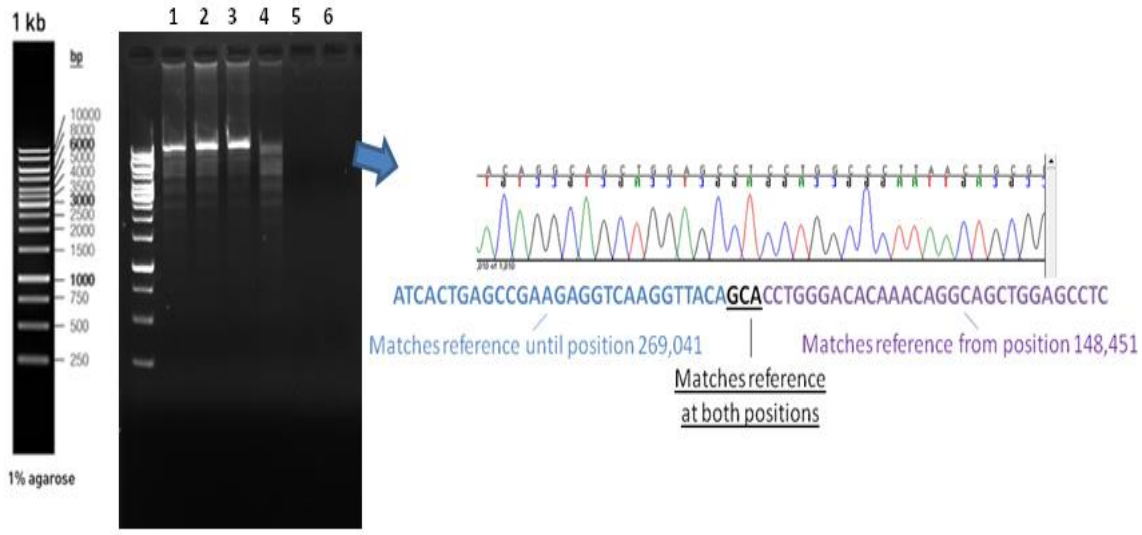


Figure 85: Breakpoint Confirmation for a Novel Deletion in MiSeq Sample 34. Gel Image and Chromatogram of Gap PCR for Sample 13. Gel image: (1) Sample 34 (proband); (2) Sibling of proband; (3) Parent of affected; (4) Negative Control; (5) No Template Control; (6) Blank. Chromatogram: Breakpoint sequence is identified, where bases in blue match the reference sequence until position 269,041, followed by three ambiguous bases that could be from either position, followed by bases in purple that match the reference sequence position from position 148,451. This picture is by courtesy of Dr Xunde Wang (NHLBI/NIH)

The duplication of 120,590 bp encompassed the entire alpha globin gene cluster as well as part of *NPLR3* upstream of the cluster and *LUC7L* downstream. The duplication was top-to-tail in orientation and brings together two Alu repeats with a high degree of homology to one another (Figure 86). The proband had inherited this duplication in conjunction with a copy of the beta thalassemia c.135delC variant on chromosome 11 Table 52, The excess alpha produced by the total number of alpha globin genes, combined with the reduced beta globin chain synthesis resulting from this variant led to a beta thalassaemia phenotype of intermediate severity, requiring intermittent blood transfusions.

Table 52 Mutation report listing for rs80356820 in MiSeq Sample 34.

Position	Reference Nucleotide	Coverage	Score	Mutation Call	Mutant Allele Frequency
Chr11: 5247987	G	188	15.6	delG	49.47

Schematic of the Novel Duplication in MiSeq Sample 34

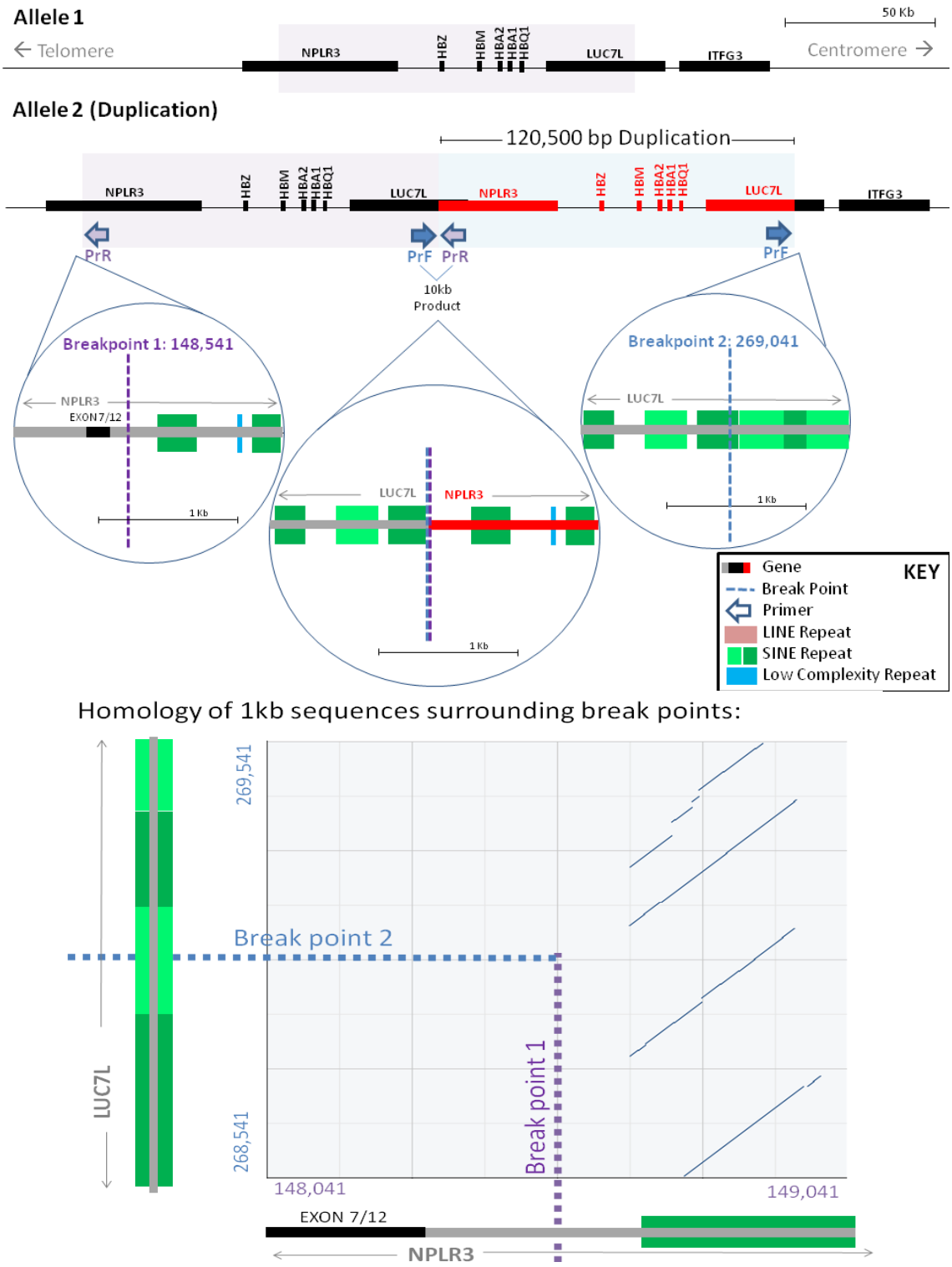


Figure 86: Schematic of the Novel Duplication in MiSeq Sample 34. Upper panel: 120,500 bp of sequence beginning within NPLR3 and ending in LUC7L, encompassing the entire alpha globin gene cluster, was duplicated. The duplicated sequence was reinserted immediately adjacent to the original sequence in top-to-tail orientation. Primers had been designed under the assumption that this was the layout of the duplication, and a 10 Kb PCR product was produced across the breakpoint sequence.

Lower panel: a dot plot of the sequence surrounding the breakpoints (500 bp upstream and downstream of the precise break point position) showed that two Alu repeats brought into close proximity by the duplication had a high level of homology to one another.

MiSeq Sample 35 (Test)

MiSeq Sample 35 shows a duplication of approximately 274,000 bp encompassing the alpha globin gene cluster (Figure 87A). Noise in the RPKM data plot at the 3' region of the duplication makes it difficult to estimate the position of the break point (Figure 87B).

Issues Estimating Rearrangement Size from RPKM Data in Sample 35

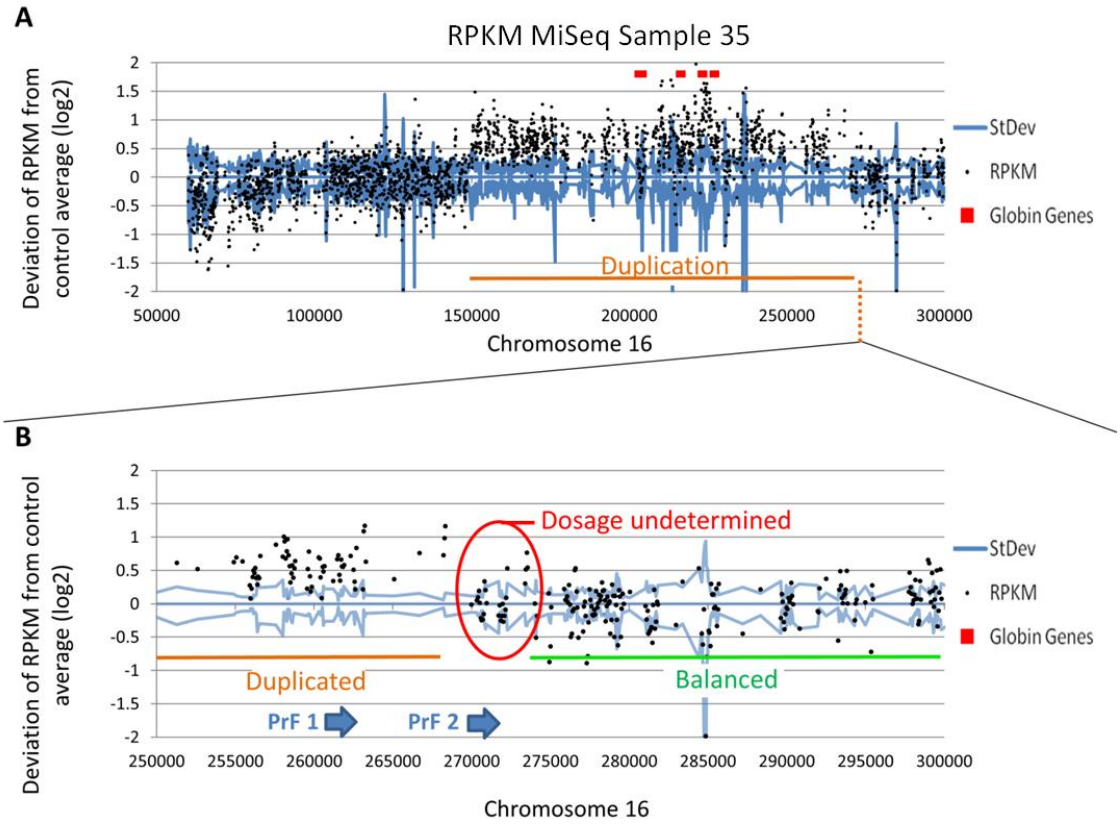


Figure 87 Issues Estimating Rearrangement Size from RPKM Data in MiSeq Sample 35 (A) Rearrangement in MiSeq Sample 35. (B) Expanded view of 3' breakpoint region - end point of duplicated region is unclear due to the presence of a repeat that is not covered in the bait design, plus noise in the RPKM data which makes the position at which normal sequence dosage resumes unclear.

Both the approximate start and end positions of the duplication indicated from the RPKM data include repeats not covered by the bait library. These repeats share a region of high homology (See dotplot, Figure 88) and were considered to be the most likely locations of the duplication breakpoints, as no opposite or same direction reads or misaligned sequence strings were obtained from the read data. Primers were designed against positions that appeared to be balanced. Two separate primers were designed against the 3' region of the duplication because noise in the data made the duplication end point unclear. We made two assumptions about the most likely layout of the duplication: that the duplication was in the adjacent and top-to-tail layout, and that the breakpoints would occur between regions of high homology. A dotplot of the Alu repeats situated at the start and end positions of the duplication identified a region

of high homology which was considered to be the most likely position for the duplication breakpoints. This allowed us to estimate the probable size of the PCR products our primers might produce.

DotPlot Showing Homology at Break Point Regions for MiSeq Sample 35

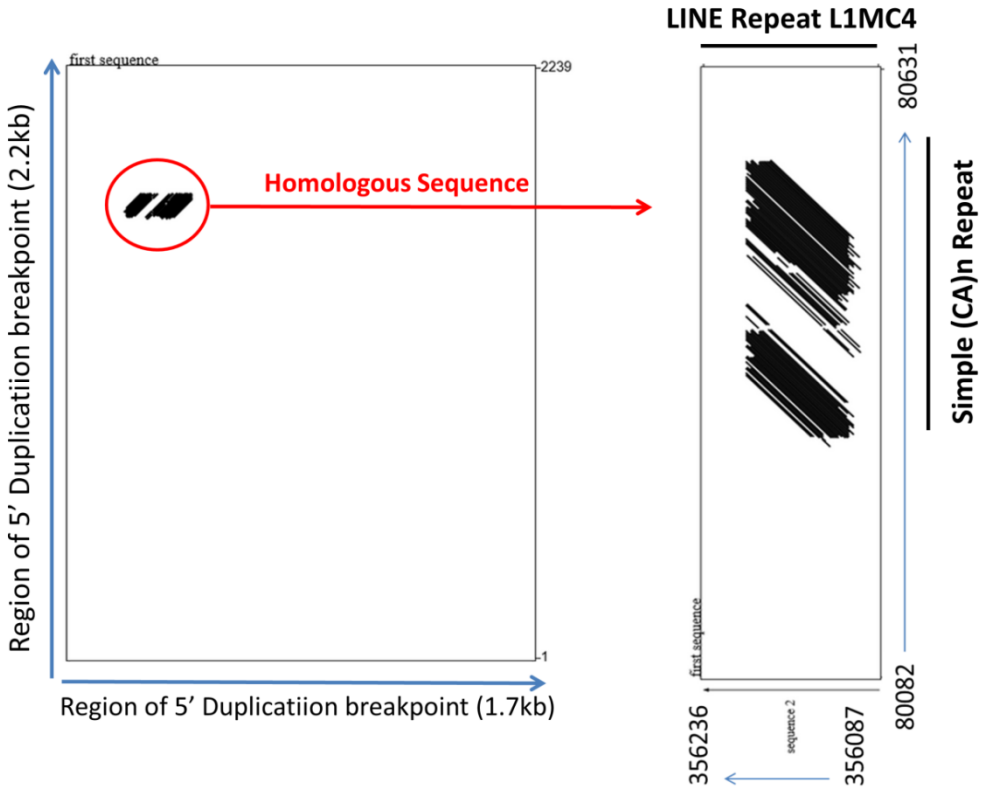


Figure 88 DotPlot Showing Homology at Break Point Regions for MiSeq Sample 35. The two breakpoint regions share a homologous region between a simple repeat (5') and a LINE repeat (3').

In the event that either assumption was false, we also used these primers to test for a head-to-head duplication (where Pr2 could effectively be used as both the forward and reverse primer for the inversion break point, assuming again that the duplicated region was adjacent to the original copy of the sequence), and used a range of thermal cycling conditions to accommodate products of up to 10 Kb. At the time of writing we were unable to isolate a unique PCR product from any of the primers that were designed. The presumed layout of the rearrangement is shown in Figure 89, but is not confirmed.

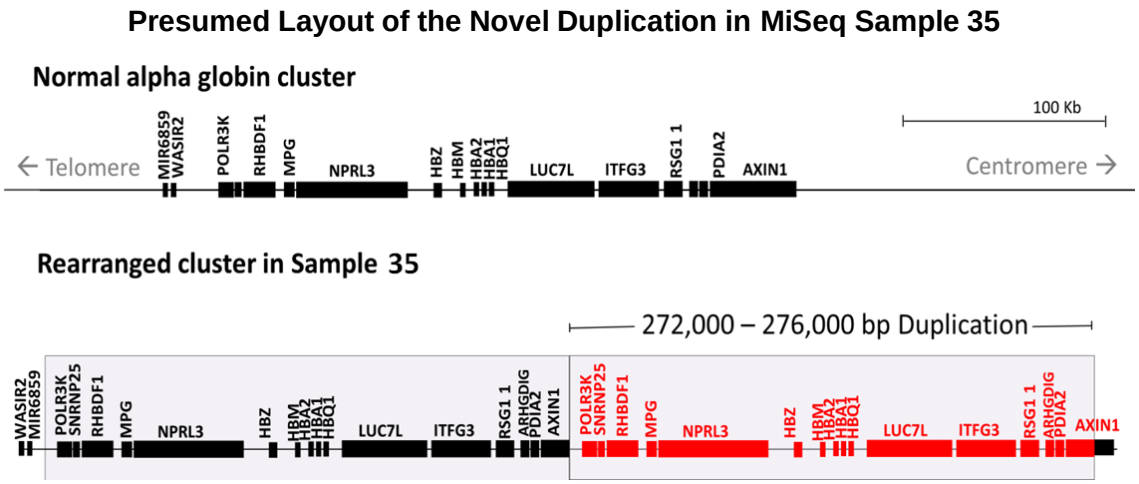


Figure 89 Presumed Layout of the Novel Duplication in MiSeq Sample 35.

MiSeq Run 6

Thirteen samples with a range of novel rearrangements affecting both the alpha and beta globin gene loci were prepared for sequencing on the MiSeq (Table 53). The run also included three controls that were negative for rearrangements at both loci. The run included samples with rearrangements affecting the beta globin gene cluster to evaluate the performance of the new bait capture library, along with further samples with rearrangements affecting the alpha globin gene loci, which appeared to be more challenging. Sequencing a mixture of alpha and beta rearrangements in a single run allowed samples with alpha rearrangement to be used as negative controls for evaluation of the beta locus and vice versa, increasing the number of negative controls available for each locus within the batch. Three of the samples (40, 41, 42) were a family trio. Sample 44 was included as part of a duo of siblings who had a discordant phenotype, however, the sibling was not successfully prepared (Figure 90). Sample 36 was a relative of affected individual (ROA) HiSeq Run 1 Sample 11 and Sample 46 was a relative of affected individual (ROA) MiSeq Run 6 Sample 48.

Samples were fragmented to an average of 600 bp using double size selection to tighten the fragment size range. The samples were then prepared for sequencing on the BioMek FX^{P} platform (asides hybridization which was performed manually). The double size selection step ensured that fragments of the target size (500-600bp) were highly represented in the finished libraries, as the smaller fragments which had been preferentially amplified in earlier runs had been removed. Sample preparation failed for one sample during either the hybridization or post-hybridization clean-up stage. The reason for this was not known (Figure 90). The remaining 12 samples were prepared

successfully and were taken forward for sequencing. The samples were pooled and diluted to a concentration of 2nM before being sequenced on the MiSeq using a V3 2x300 reagent kit. Sequencing was successful, generating 861k/mm^2 clusters on the flow cell. All sample indexes were equally represented on the flow cell. The quality of base calls dropped towards the end of Read 1 and Read 2 due to their long length using this sequencing kit (Figure 91).

Table 53: Sample Details MiSeq Run 6.

Sample	Status	Mutation	Affected Locus
MiSeq Sample 36	Relative of affected	Deletion	Beta
MiSeq Sample 37	Test	Deletion	Alpha
MiSeq Sample 38	Test	Deletion	Alpha
MiSeq Sample 39	Test	Deletion	Alpha
MiSeq Sample 40	Relative of affected	Duplication	Beta
MiSeq Sample 41	Relative of affected	Duplication	Beta
MiSeq Sample 42	Test	Duplication	Beta
MiSeq Sample 43	Test	Deletion	Alpha
MiSeq Sample 44	Test	B0-39 Indel plus Unknown	Beta
MiSeq Sample 45	Test	Duplication	Alpha
MiSeq Sample 46	Relative of affected	Unknown	Alpha
MiSeq Sample 47	Negative Control		
MiSeq Sample 48	Negative Control		
MiSeq Sample 49	Negative Control		

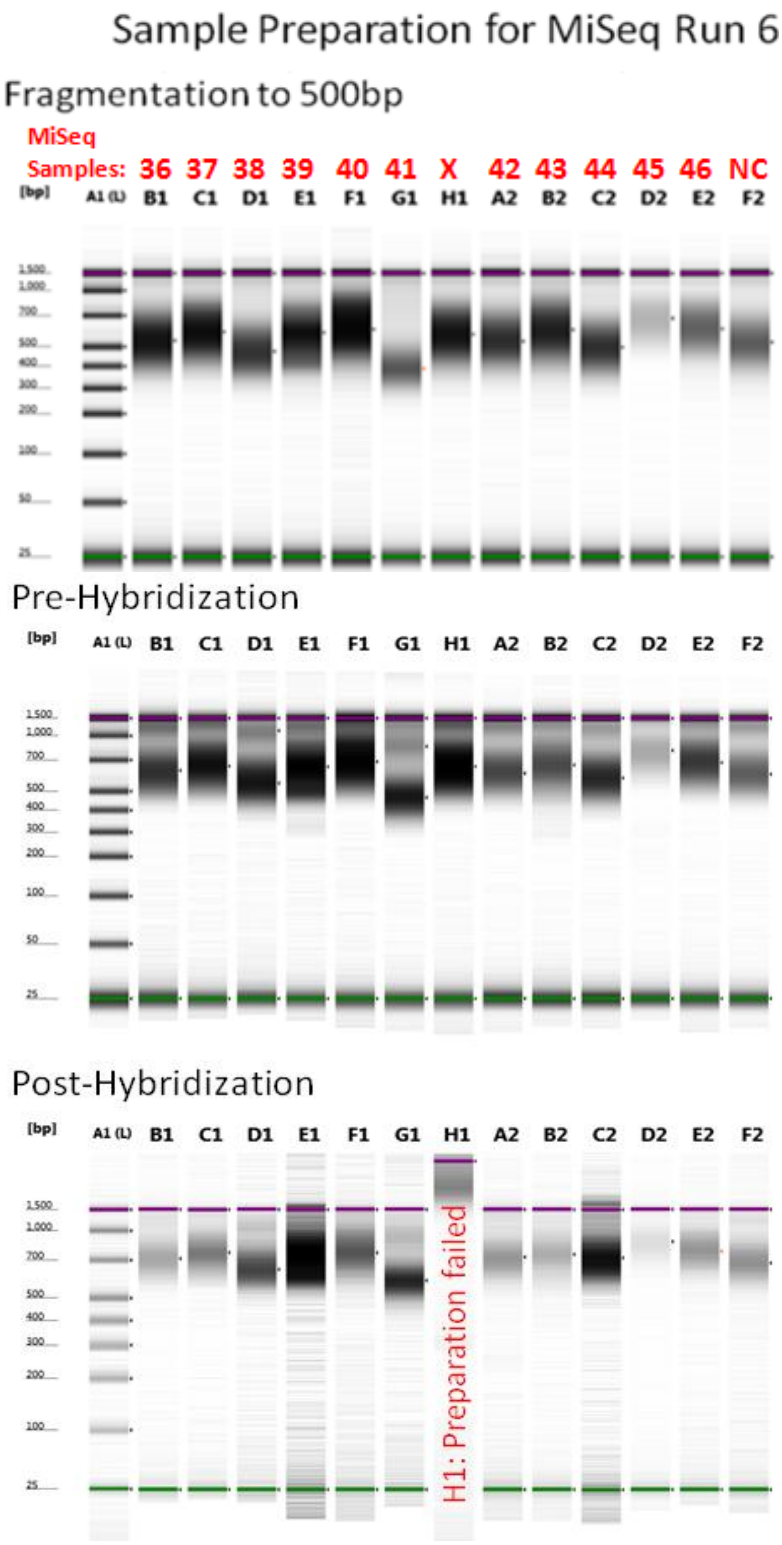


Figure 90 Electropherograms showing samples from MiSeq Run 6 at three stages of library preparation. Stages shown are fragmentation, pre-hybridization and post-hybridization. NB: the negative control Sample 47 was prepared once and then different indexes were added to the aliquots of the sample to create three negative control samples to be run on the MiSeq. Lane H1 shows a sample that was prepared for sequencing but was shown to have failed at the post-hybridization stage for unknown reasons.

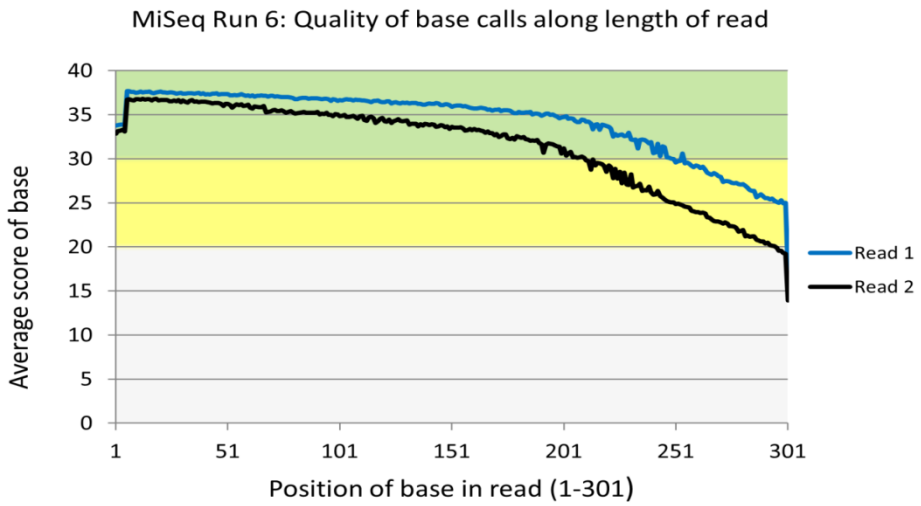


Figure 91: Base Call Qualities for Read 1 and Read 2, MiSeq Run 6.

Format Conversion

Sequencing data from each sample was downloaded from BaseSpace in FASTQ format, converted to FASTA format and filtered according to standard parameters (see Methods). The majority (93.5%) of reads passed the quality filtering metrics (Table 54). Of those that failed to pass, 96.7% failed the 'Median Score Threshold' parameter. The drop in base call quality score towards the ends of the long reads meant that 83% of reads required trimming. The average read length after trimming was 271 bp (Table 54).

Table 54: Format Conversion Statistics for MiSeq Run 6 (sample average and standard deviation)

Format Conversion MiSeq Run 6	Average	StDev
Total Reads in the Input File	1161829	358196.9
Reads Converted Successfully %	1088325 93.56	340555.4 1.29
Reads Failed to Convert	73504	23189.58
Reads Filtered by "Median Score“ %	71140.29 6.23	22387 1.26
Reads Filtered by "Uncalled Bases"	780.2857	256.8172
Reads Filtered by "Called Base Number in Each Read"	0	0
Reads Filtered After Trimming	1583.429	653.4567
Reads Trimmed	910029.3	287757.1
Reads Trimmed by "Quality Score"	910029.3	287757.1
Reads Trimmed by "Homopolymer Trimming"	0	0
Reads Trimmed by "Sequence Trimming"	0	0
Reads Trimmed by "Specified Length"	0	0
Trimmed Bases	47581785	15971162
Trimmed Bases by "Quality Score"	47581785	15971162
Trimmed Bases by "Homopolymer Trimming"	0	0
Trimmed Bases by "Sequence Trimming"	0	0
Trimmed Bases by "Specified Length"	0	0

Sequence Alignment

Reads that passed quality filtering were aligned to the reference sequence using standard parameters, with an expected gap distance of 0-800 bp. The region of interest was covered at an average depth of 250x, with 61% of reads on target (Table 54). Of reads that passed quality filtering, 95% were successfully aligned to the reference sequence.

Table 55: Sequence Alignment MiSeq Run 6

	Average	StDev	MiSeq Sample Number													
			36	37	38	39	40	41	42	43	44	45	46	47	48	49
Matched Reads	2057756	603965	2470607	2665922	2062646	2335184	805257	2006919	1840762	3169820	2531861	2387767	2276892	2386635	2905665	2346402
Perfect Reads	941824	283750	1177893	1223931	943900	1044539	362783	897897	818062	1563391	1201906	1140639	1095551	1151384	1400802	1132218
Unmatched Reads	86708	15131	99637	104707	79565	96785	60925	78628	83373	112643	140729	148472	97637	138776	96832	99992
Short Reads	1111383	33628	115261	126448	147531	121063	39977	118018	85265	161225	123760	121857	99286	104704	151901	120363
Substitutions Called	743680	225907	848337	993631	675177	896714	291473	756747	719903	962568	811757	733690	798785	788693	920392	738061
Substitutions Uncalled	1544708	341133	1794632	1900700	1555301	1698492	852477	1466645	1466612	2264328	1994753	1908705	1716517	1912575	2063358	1722145
Deletions Called	205210	69459	208272	283167	169461	282927	80552	206880	226851	217927	183655	172085	190826	210071	229395	173021
Deletions Uncalled	125881	24925	154406	157711	91892	131394	120114	99768	134658	177637	186276	177721	172709	202475	150002	138478
Insertions Called	129057	43576	149306	184737	133808	143262	41884	121347	110944	175249	138965	124121	147316	141637	156725	124748
Insertions Uncalled	85772	17359	105436	106009	59250	93704	78151	72083	97765	113739	129731	140337	110047	131395	100937	90450
Average Read Length	271	3	273	272	266	271	275	269	272	271	272	271	274	273	270	271
Average Coverage (Genome)	8	1	9	7	8	8	9	9	10	8	8	8	11	9	8	8
Average Coverage (ROI)	250	10	246	240	243	247	270	257	274	244	242	247	254	249	241	241
% On-target	61	3	60.85	58.15	59.50	60.11	66.21	61.56	66.26	60.21	60.12	60.64	62.04	61.65	60.16	60.48

Variant Characterisation MiSeq Run 6

Seven test samples and four samples from potentially affected relatives of patients had been sequenced (Relative of Affected: 'ROA'). RPKM plots were created for all samples, showing the coverage across the ROI compared to the average coverage in negative controls, and also in test samples with variants at the alternate loci (Figure 92, Figure 93). To-the-base characterisation of the test samples and determination of presence or absence of the variant in related samples was possible in 8 of 11 cases (Table 56). One relative of a test sample was found to also carry the test sample variant.

Table 56: Variant Characterisation in Sample Cohort, MiSeq Run 6

Sample	Status	Mutation	Affected Locus	Result
MiSeq Sample 36	ROA	Deletion	Beta	Negative
MiSeq Sample 37	Test	Deletion	Alpha	Unresolved
MiSeq Sample 38	Test	Deletion	Alpha	Resolved
MiSeq Sample 39	Test	Deletion	Alpha	Resolved
MiSeq Sample 40	ROA	Duplication	Beta	Negative
MiSeq Sample 41	ROA	Duplication	Beta	Resolved
MiSeq Sample 42	Test	Duplication	Beta	Resolved
MiSeq Sample 43	Test	Deletion	Alpha	Resolved
MiSeq Sample 44	Test	B0-39 Indel plus Unknown	Beta	B-039 Indel confirmed, negative for additional variants
MiSeq Sample 45	Test	Duplication	Alpha	Unresolved
MiSeq Sample 46	ROA	Unknown	Alpha	Negative
MiSeq Sample 47	Negative Control			
MiSeq Sample 48	Negative Control			
MiSeq Sample 49	Negative Control			

MiSeq Run 6 Coverage Graphs Chromosome 11

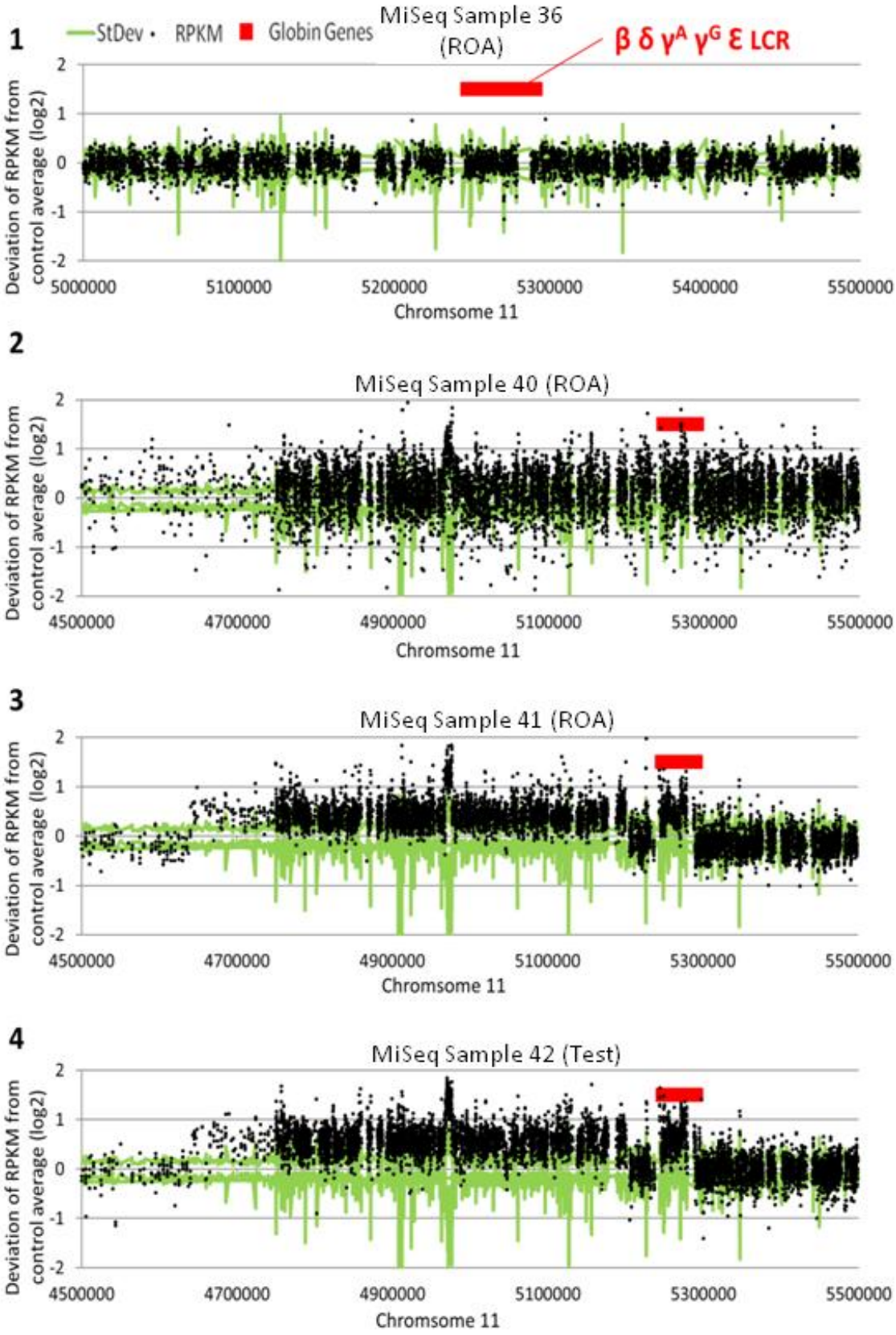


Figure 92: MiSeq Run 6 Coverage Graphs Chromosome 11. Plots show depth of coverage (Y axis) at each bait-covered position (X axis) in relation to the average coverage of that position in negative controls on a Log2 scale. Standard deviation from the average in the negative controls is shown in green, the positions of the beta globin gene locus are shown in red. (1) Sample 36 Relative of affected (ROA) (2) Sample 40 ROA (3) Sample 41 ROA (4) Sample 42 Test

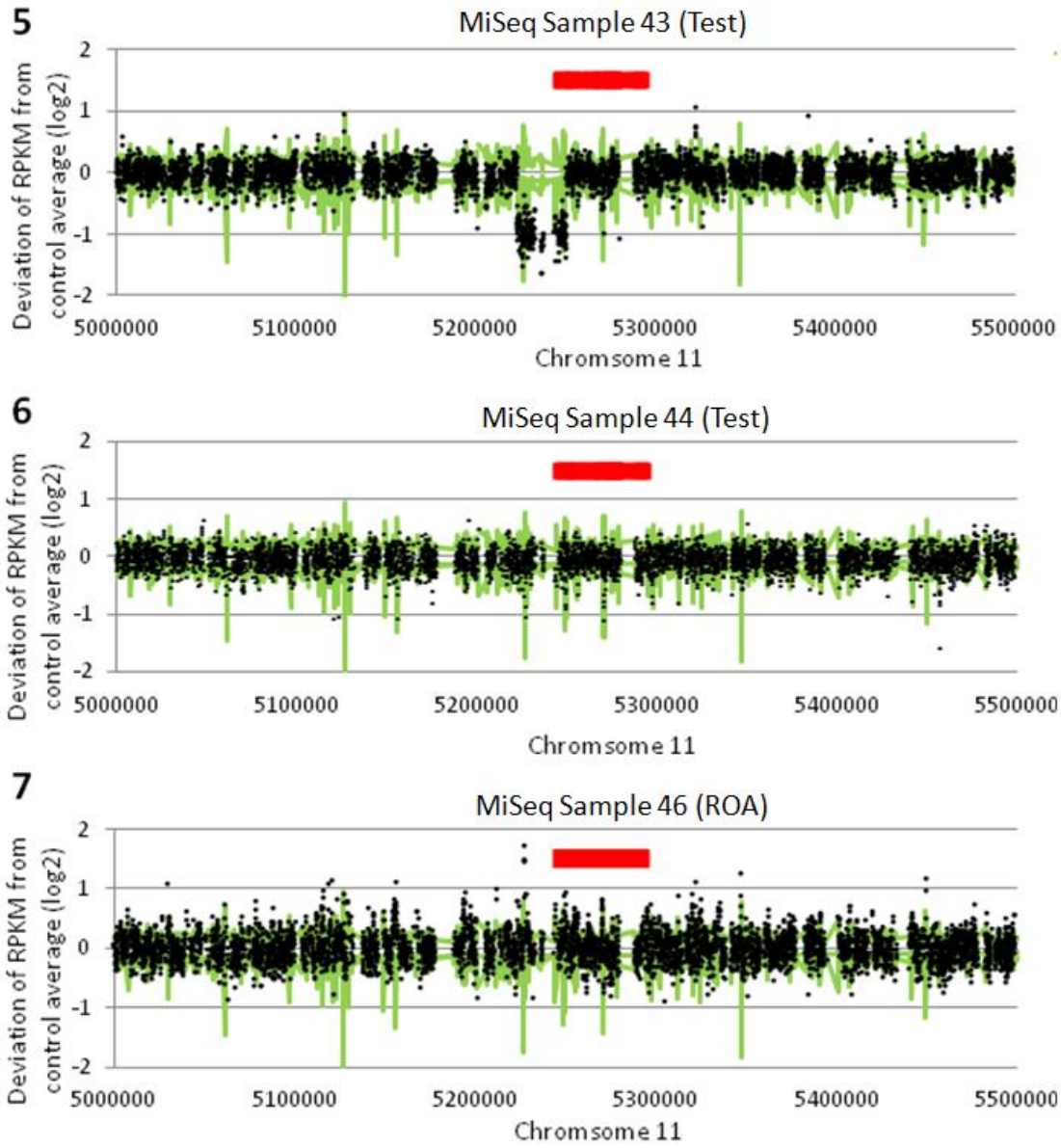


Figure 92 Continued: (5) Sample 43 (Test) (6) Sample 44 (Test) (7) Sample 46 (Test)

MiSeq Run 6 Coverage Graphs Chromosome 16

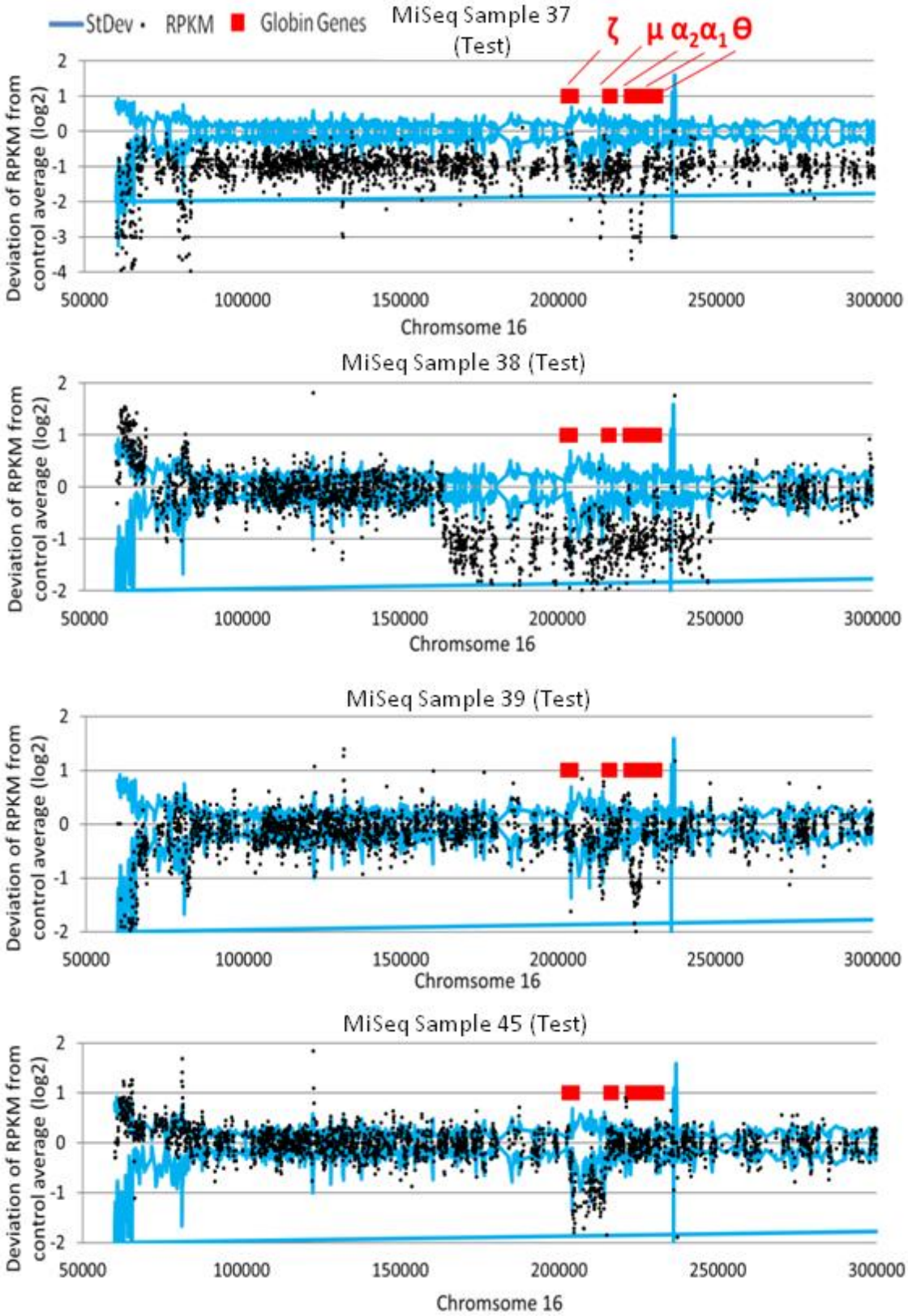


Figure 93: MiSeq Run 6 Coverage Graphs Chromosome 16. (1) Sample 37 (Test) (2) Sample 38 (Test) (3) Sample 39 (Test) (4) Sample 45 (Test)

MiSeq Sample 36, Figure 92.1 (ROA)

Sample 36 was a relative of an affected sample from a previous run (MiSeq Sample 34). This individual appeared to be negative for the structural variant affecting that sample.

MiSeq Sample 37, Figure 93.1 (Test)

This sample was compound heterozygous carrier of the 3.7 Kb deletion and a large novel deletion removing the entire tip of chromosome 16, including the alpha globin gene cluster. The resulting genotype was (--/α-$^{3.7}$) with a clinical phenotype of HbH disease. The patient had red cell indices of 1.9% HbA_2, 0.2% HbF, 6.01 RBC, Hb 9.2, MCV 49.1, MCH 15.3. The RPKM data indicated that the large deletion removes 791-793 Kb of sequence (Figure 94). The deletion breakpoint was in a repetitive region not included in the bait design comprising a LINE repeat and two SINE repeats (Figure 95).

Two Deletions Affecting the Alpha Globin Gene Cluster in MiSeq Sample 37

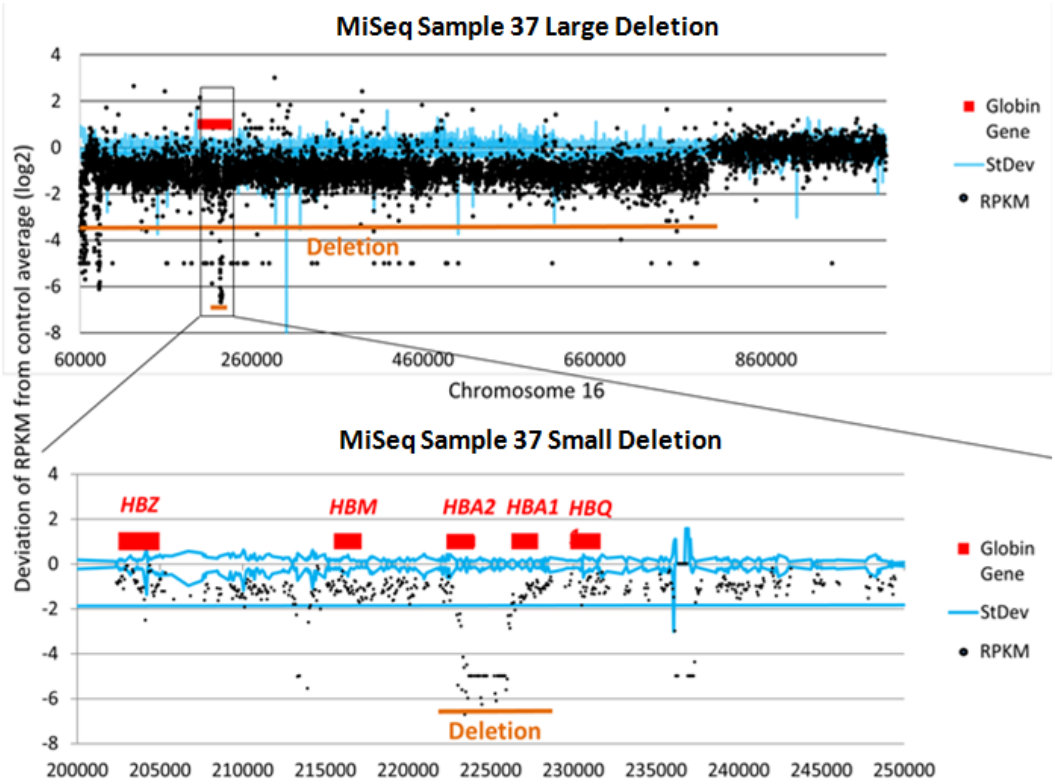


Figure 94: Two Deletions Affecting the Alpha Globin Gene Cluster in MiSeq Sample 37. Upper: A large novel heterozygous deletion removing all bait-covered sequence - including the entire alpha globin gene cluster - up until position ~chr16:789,357. Lower: The co-inherited (α-$^{3.7}$/) deletion reduced the coverage of the region affected by both rearrangements to zero. This returns an error when calculating the deviation in coverage from the negative control average, so the error calls have been replaced by the value -5 so that they can be correctly plotted on the chart.

No breakpoint data was captured for this deletion: The rearrangement may have brought together part of the unmapped telomere region of the chromosome and the unmapped repetitive region, in which case neither part of the breakpoint sequence

would be captured by the bait library. This would also be the case if the chromosome was bluntly severed at this location. Even if the deletion breakpoint was in a captured region, in this situation it may not leave any abnormal sequences behind that could be identified in the alignment. It is possible that the breakpoint in this situation could be identified by a sudden drop in coverage with reads from multiple DNA fragments showed the exact same start position on the chromosome. This deletion could not resolved with to-the-base accuracy, as the bait capture library did not cover the breakpoint position, and it is not possible to design Gap-PCR primers against this region. Theoretically, it may be possible to capture this breakpoint via southern blotting, but due to time constraints this was not investigated.

Approximate Region of Chromosome 16 Removed By Large Deletion in MiSeq Sample 37

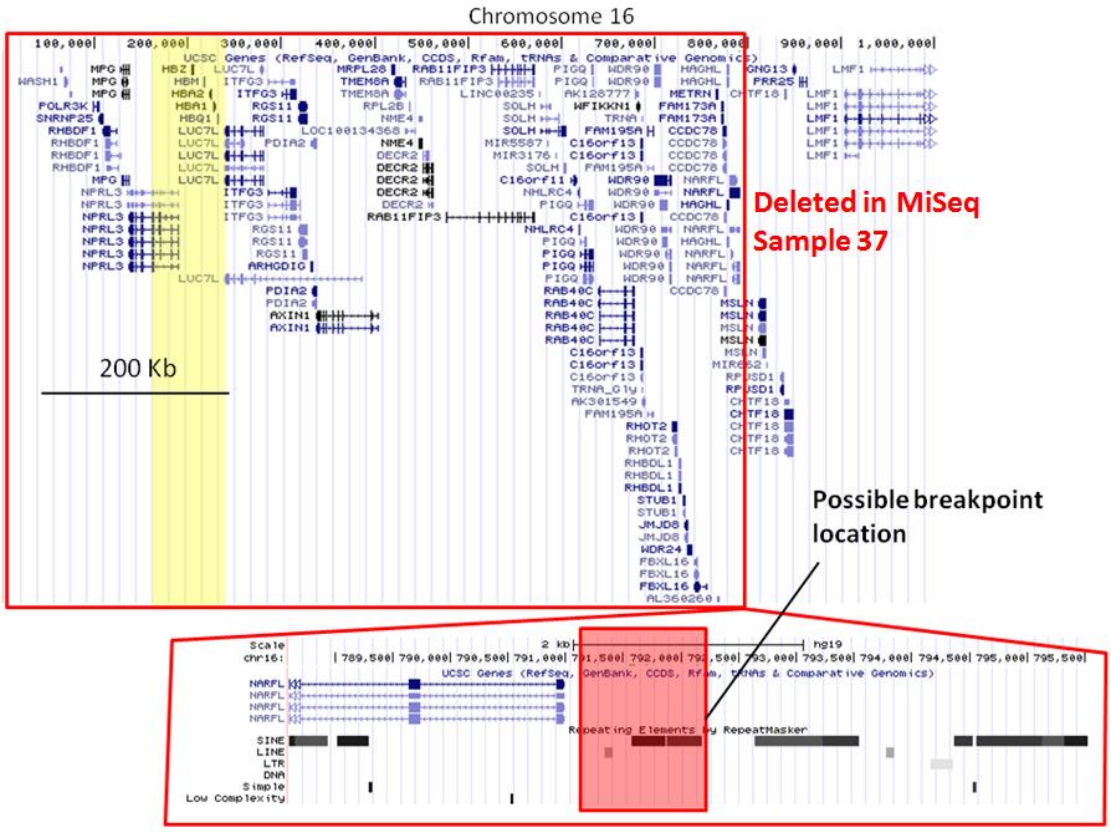


Figure 95 Approximate Region of Chromosome 16 Removed By Large Deletion in MiSeq Sample 37. Upper image: Approximate region and genes removed by large shown in UCSC Genome Browser (red box shows deleted region). Lower image: enlarged image of region of 5' breakpoint with suspected breakpoint position highlighted in red. The region is repetitive and not included in the bait design.

MiSeq Sample 38, Figure 93.2 (Test)

MiSeq Sample 38 was from a patient with reported red cell indices of 2.7% HbA_2, 0.2% HbF, RBC 5.57, Hb12, MCV 68.2, MCH 21.5. The RPKM plot for MiSeq Sample 38 identified a deletion extending from Chr16:164,041 – 248,193, removing 84,152 bp of

sequence. Reads containing break point sequence could be derived from inspection of the read pile-up at the start and end points for the deletion indicated by the RPKM plot in the NextGene Viewer. BLAT query of these reads showed that these sequences matched the break points suggested by the RPKM plot. Additionally, the read sequences revealed that an additional 341 bp deletion was present 71 bp upstream of the large deletion (Chr16:248,263-248,605) (Figure 96A). The presence of both deletions was confirmed via Gap PCR and dye-terminator sequencing analysis (Figure 96B).

The 5' breakpoint of the large deletion is situated in unique sequence. The 3' breakpoint of the large deletion is situated in an Alu repeat, AluY, which has no homology with the 5' breakpoint region. The large deletion removes *NPRL3* exon 1-3/12, the alpha globin gene cluster, and *LUC7L* exons 7-9/9. The small deletion removes part of intron 6/8 of *LUC7L*. The 5' breakpoint of the small deletion is situated in the same Alu repeat as the 3' breakpoint of the large deletion. The 3' end of the small deletion is situated in unique sequence, but adjacent to another Alu repeat, AluSx. The homology between AluY and AluSx is extremely high. There is a novel insertion ('TCTCGC') at the small deletion breakpoint.

Novel Deletion in MiSeq Sample 38

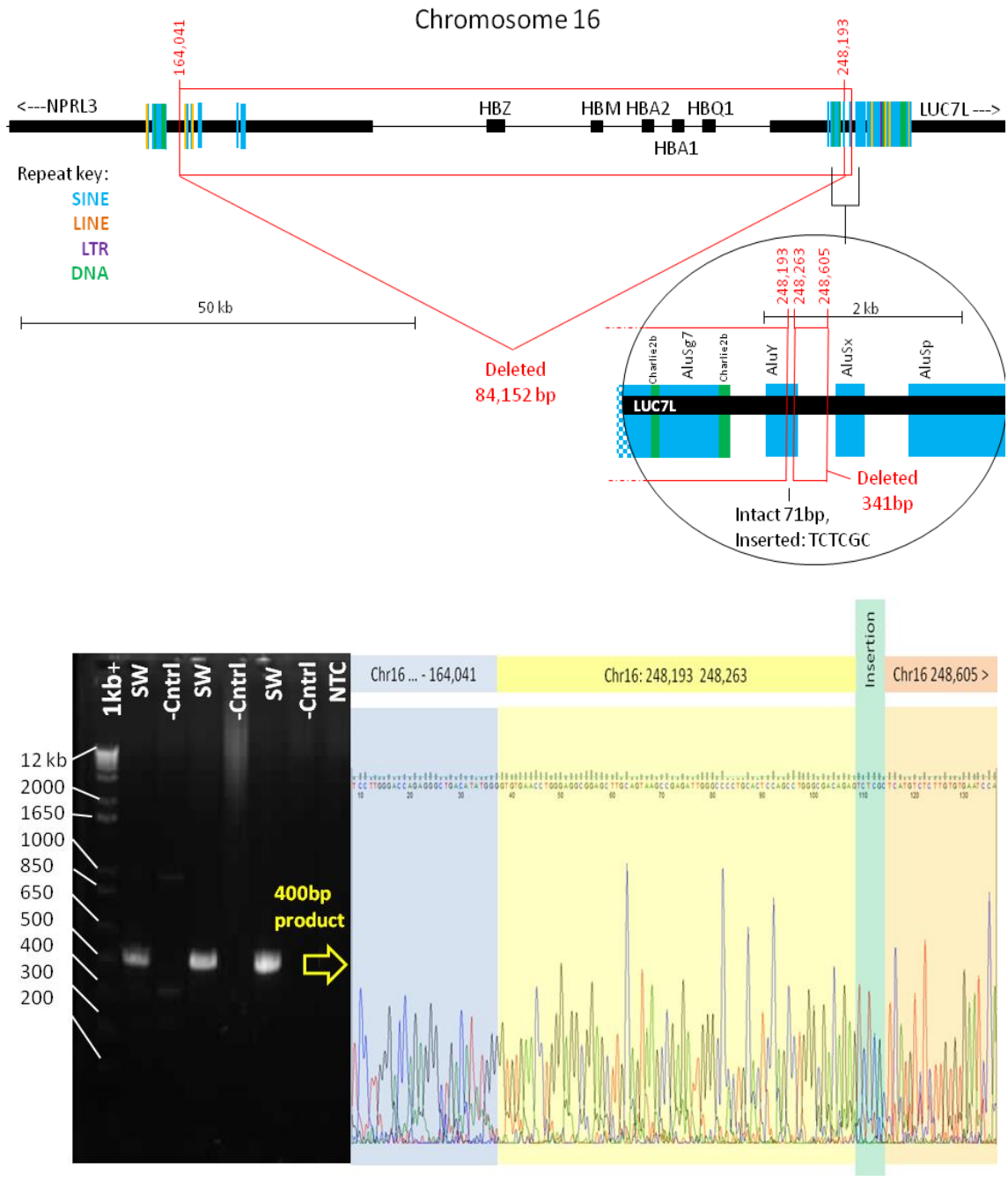


Figure 96: Novel Deletion in MiSeq Sample 38. (A) A schematic of the deletions on chromosome 16: A deletion of 84 Kb removes the alpha globin gene cluster. 71 bp upstream of this, an additional deletion removes 341 bp of sequence. Repetitive elements in the region are represented by coloured bars – blue (SINE), orange (LINE), purple (LTR) and DNA (green). The positions of primers designed to confirm the presence of these deletions are included on the schematic (not to scale). (B) Confirmation of the two deletions by Gap PCR: Left panel shows gel image of products amplified by primer pair Pr1 and Pr2 (L-R: 1. 1 Kb+ ladder 2. proband 3. negative control (58°C annealing temperature) 4. Proband 5. negative control (60°C annealing temperature) 6. Proband 7. negative control (63°C annealing temperature) 8. no template control). A 400 bp PCR product is present in the proband and absent in the negative control. Dye-terminator sequencing of the PCR product reveals the breakpoints of the two deletions, matching the co-ordinates predicted by NGS.

MiSeq Sample 39 Figure 93.3 (Test)

The RPKM data showed MiSeq Sample 39 was heterozygous for the (α-$^{3.7}$/) deletion. The number of negative controls included in the Run was sufficient to clearly identify this rearrangement in a heterozygous state from the RPKM data (Figure 97). No concurrent deletions affecting the globin gene loci were included in the mutation report and the beta globin gene locus on chromosome 11 appeared balanced.

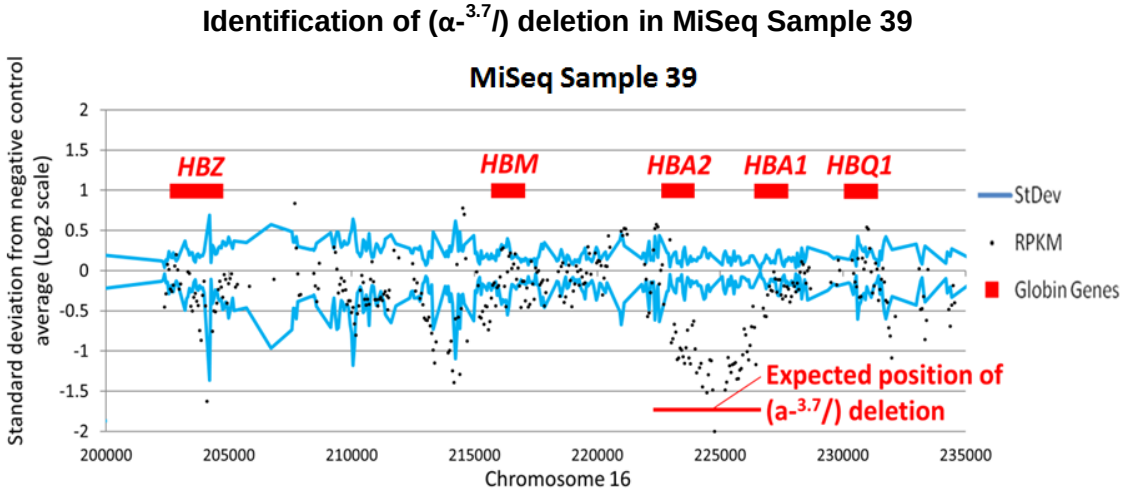


Figure 97: Identification of (α-$^{3.7}$/) deletion in MiSeq Sample 39. RPKM data shows a region corresponding to the size and position of the (a-$^{3.7}$/) deletion as listed in HbVar (which is indicated by a red bar).

MiSeq Samples 40, 41, 42, Figure 92.2,3,4 (Test)

Samples 40, 41 and 42 are a family trio (Sample 40 = mother, Sample 41 = father, Sample 42 = proband). The mother was a carrier for HbS (HbS 33.3%). The father was a carrier for HbS with an abnormally high HbS fraction of 41.3%.The proband had a majority fraction of HbS, with some HbF and no HbA. A duplication affecting the beta globin gene loci was suspected based on these abnormal HbS fractions. The mutation report identified the sickle cell variant in all three samples. The mutant allele frequency was 1:1 in the mother indicating that she was heterozygous for the variant. The father showed a 2:1 ratio of the normal allele to the variant, and the proband showed a 1:2 ratio of the normal allele to the variant (Table 57). RPKM plots showed that both the father (Sample 41) and the proband (Sample 42) had a large duplication encompassing the beta globin gene cluster. Inspection of the read pile-up at the start and end points of the duplicated region as indicated by the RPKM plots showed reads containing breakpoint sequences (Figure 98). While BLAT query of these sequences showed multiple matches to repetitive regions in the genome, complimentary matches to the expected break point regions according to the RPKM data were assigned the highest scores. The duplication break points were confirmed by Gap-PCR and dye-terminator sequencing analysis. Comparison of the mutant allele frequencies between the father

and the proband showed that the duplication included one copy of β^{-S} and one copy of β^{-A}. The proband was surmised to have inherited the father's duplicated HBB allele with its copies of HBB^{S} and HBB^{A}, plus a β^{-S} allele from the mother.

Table 57 Mutation report listing for rs334 in MiSeq Run 6 Samples 40, 41 and 42

Sample	Position	Reference Nucleotide	Coverage	Score	Mutation Call	Mutant Allele Frequency
Sample 40	Chr11: 5248232	T	101	16	T>AT	45.54
Sample 41	Chr11: 5248232	T	258	18.3	T>AT	23.64
Sample 42	Chr11: 5248232	T	226	18.2	T>AT	61.06

The inheritance of a 2:1 ratio of β^{S}:β^{A} genes did not fully explain the red cell indices in the sample. Alone, this would be expected to give a HbS fraction of up to 66%. It was noted that the duplicated region did not include the beta globin gene cluster locus control region (LCR), meaning that one of the two copies of the globin gene cluster on the duplicated allele was missing an immediately upstream LCR on the chromosome (Figure 99). We hypothesise that the HBB^{A} gene that was located further away from the common LCR, and hence produced less product that the more closely located HBB^{S} gene. The duplication showed two notable regions where the dosage did not indicate three copies of the sequence: the first was attributed to copy number variation producing additional copies of the sequence between positions chr11:4,967,214 and 4,976,583. Multiple CNVs affecting this region are recorded in dbVAR, entry '*dgv187e199'* most closely matching the loci of the rearrangement seen in the samples. A small balanced region with normal dosage also appears to interrupt the duplication between positions 5,200,463 and 5,244,326. This is seen in both the proband and father, suggesting it is a part of the duplicated allele, rather than a concurrent deletion on the other chromosome inherited by the proband. The break points of this deletion were both situated in repeats, and were not identified. Gap-PCR to resolve the breakpoints of this interruption to the duplication was performed by the referring laboratory at the University of Utah.

The breakpoints of the duplication were confirmed by Gap-PCR. Primer design assumed that the duplication was top-to-tail in orientation, based on the sequence of the break point reads. The primer sequences were PrF:GGTTCCCTCTCTCCCTGTTT and PrR:ACAAATGGGTACAGCGAGGT. Standard reagents and cycling settings (as detailed in Methods) were used to amplify the product, with a 60°C annealing temperature. The primers produced a 227 bp product in the proband and the father. No product was amplified in either a negative control sample or in the mother, who was concluded to be negative for the rearrangement. Dye-terminator sequencing analysis of the Gap PCR product in the proband and the father confirmed the breakpoints of the duplication as identified by NGS (Figure 100).

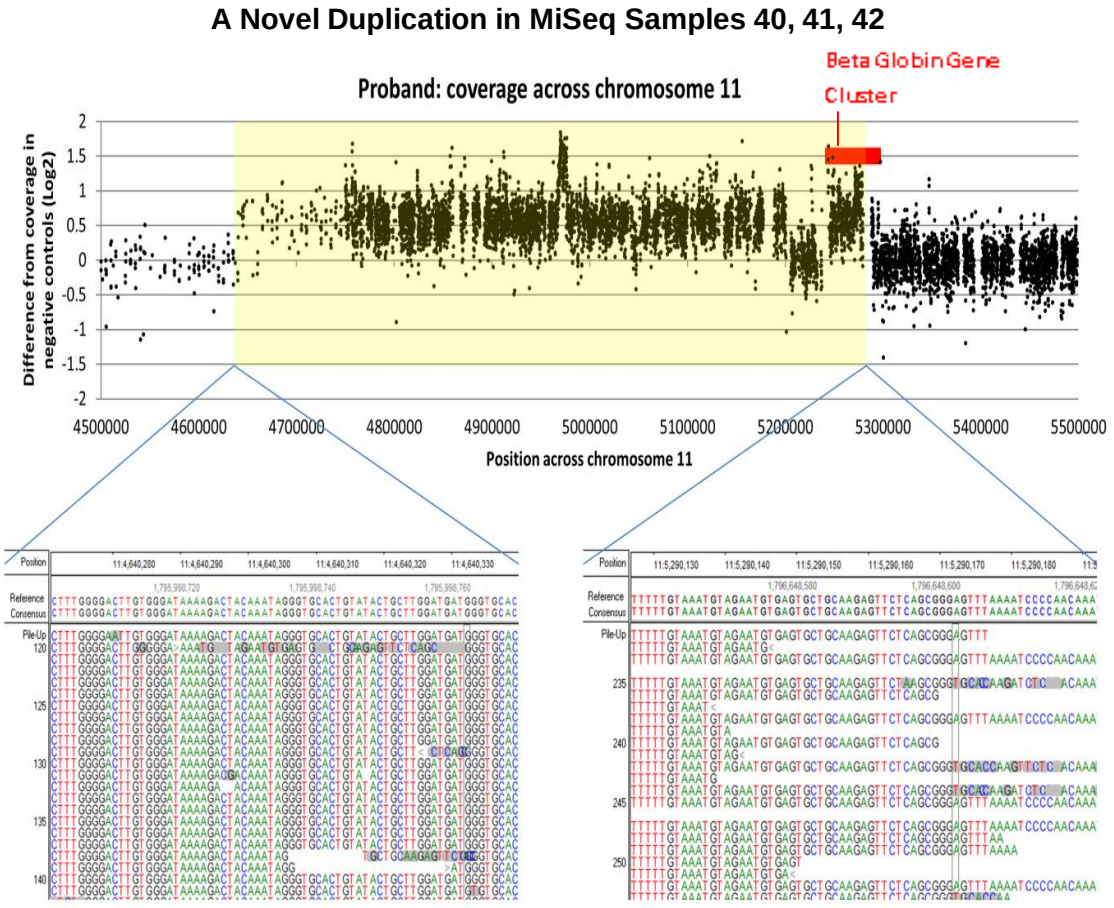


Figure 98: A Novel Duplication in MiSeq Samples 40, 41, 42. Upper: RPKM Plot Showing Duplicated Region in Proband and Breakpoint Reads in NextGene Viewer. Note quadruplicated CNV from 4,967,214-4,976,583 and balanced region from 5,200,463-5,244,326. Lower: Reads aligning to the start and end positions of the duplication contain break point sequences.

Schematic of the Alleles Inherited By the Proband (MiSeq Sample 42)

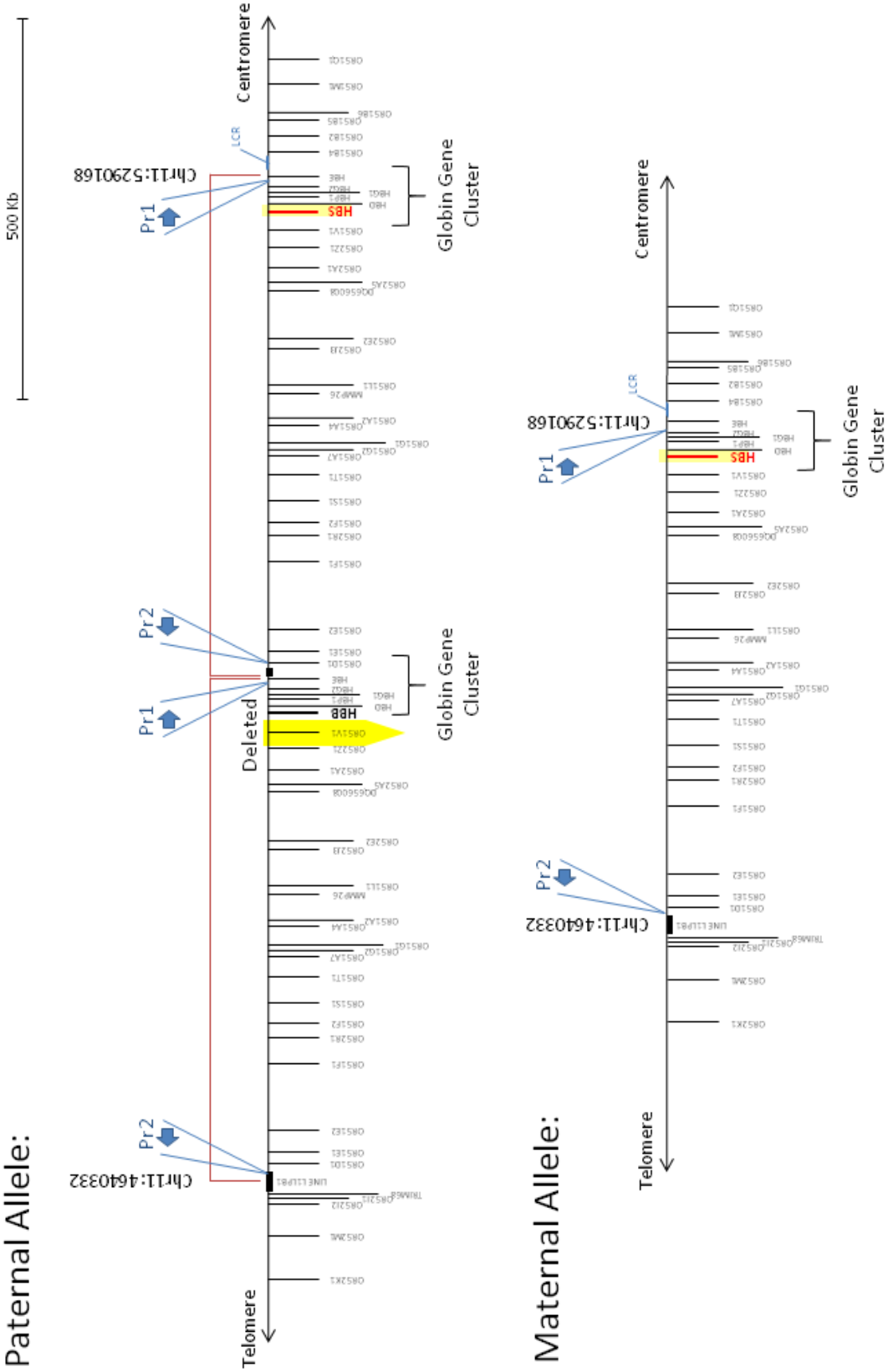


Figure 99: Schematic of the Alleles Inherited By the Proband (Sample 42). Primer locations represented by blue arrows Proband inherited an allele with a novel duplication from their father (Sample 41) and a balanced allele with the sickle cell variant from their mother (Sample 40). Note suspected positions of β^S and β^A on duplicated genes).

Breakpoint Confirmation for a Novel Deletion in MiSeq Samples 41 and 42

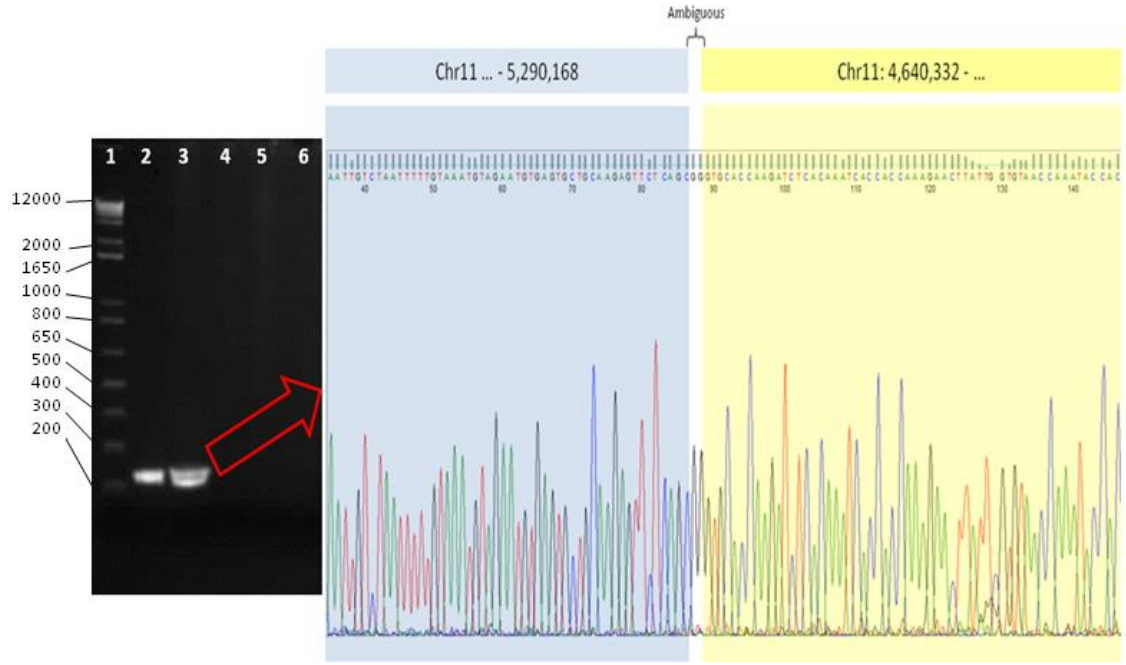


Figure 100: Breakpoint Confirmation for a Novel Deletion in MiSeq Samples 41/42. Left panel shows gel image of products amplified by Gap-PCR Primers Pr1 and Pr2 (L-R: 1. 1 Kb+ ladder 2. proband (sample 42) 3. Father (sample 41) 4. Mother (sample 40) 5. negative control 6. no template control). Right panel shows chromatogram of PCR product. Sequence confirms breakpoints as determined by NGS. The reference sequence expects two 'A' bases at each break point, thus the origin of the two centre bases is ambiguous.

Sample 43, Figure 92.5 (Test)

Sample 43 had reported red cell indices of 3.5% HbA_2, 23.9% HbF, RBC 4.56%, Hb 126, MCV 84.6, MCH 27.6. Sample 43 exhibited a deletion removing the *HBB* gene. Inspection of the read pile-up at the positions indicated by the RPKM data revealed the breakpoint sequence. The deletion removed 27,412 bp between chr11:5,222,877 and 5,250,289 (Figure 101). Primers were designed to confirm the variant, PrF:TCAAAAGCCTCATGGTAGCAG and PrR:GCATTTTCTTTGACCCAGGA. The PCR was performed using the standard reagent mix and cycling parameters detailed in Methods (with an annealing temperature of 60°C). Dye-terminator sequencing analysis of the 150 bp product confirmed the breakpoints of the deletion as identified by NGS (Figure 101).

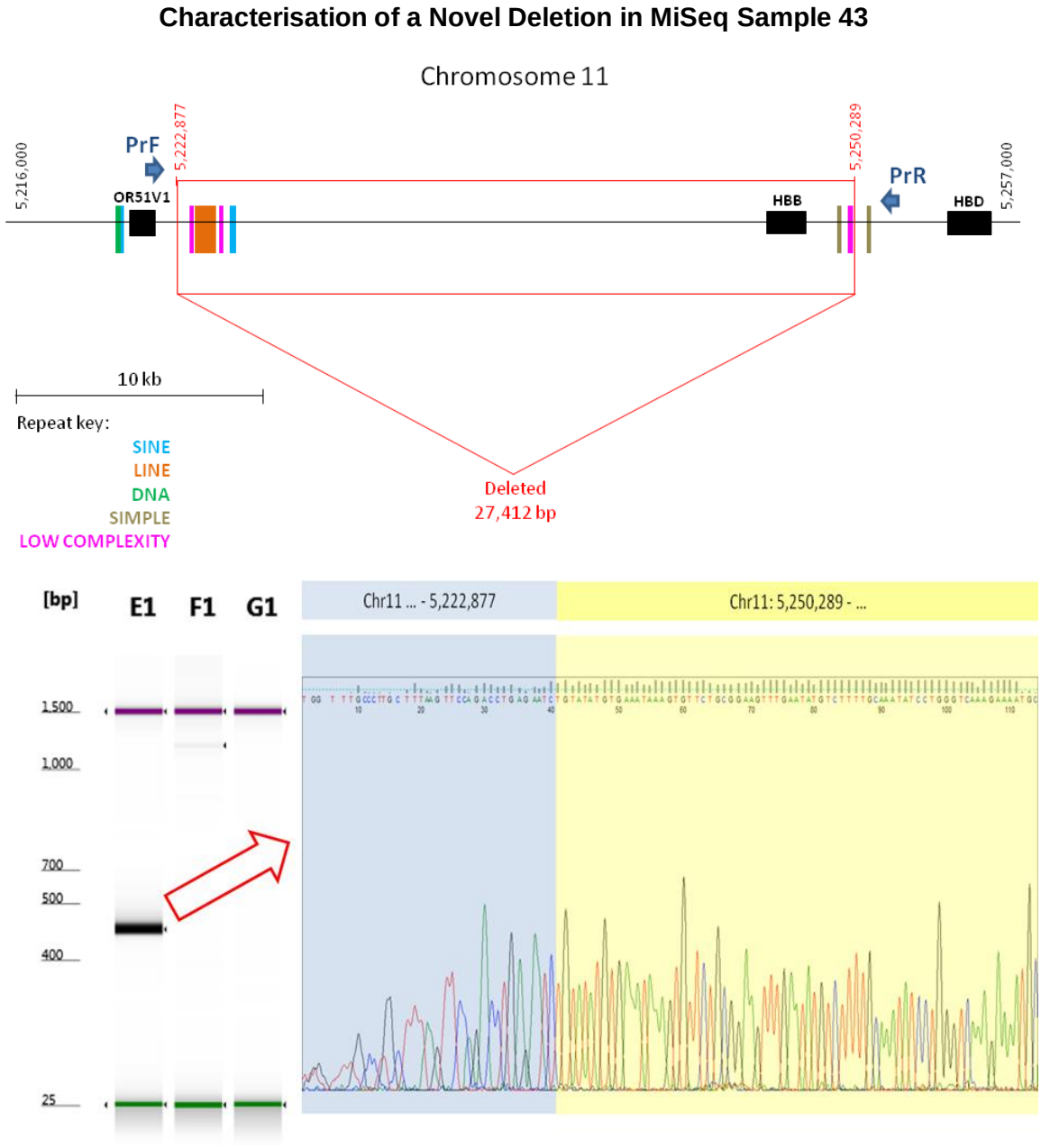


Figure 101: Characterisation of a Novel Deletion in MiSeq Sample 43. Upper panel: schematic of deletion, which removes *HBB*. Repetitive elements in close proximity to the breakpoints are denoted by coloured bars: blue (SINE), orange (LINE), green (DNA), brown (Simple), fuchsia (low complexity). Approximate positions of the primers PrF and PrR are indicated on the schematic. Lower panel: digital gel image of PCR product taken on Tapestation (Left to right: 25-1,500 bp ladder, E1: MiSeq Samples 43, F1: negative control, G1: no template control). The PCR product was sequenced, confirming the breakpoints indicated by NGS sequencing.

MiSeq Sample 44 Figure 93.3 (Test)

MiSeq Sample 44 was one of a sibling pair with a discordant phenotype that were prepared for sequencing. Both samples were known carriers of the codon 39 beta thalassaemia variant. The sample for the sibling of this individual failed to complete sample preparation, resulting from an unknown error occurring between hybridization and the final AMPure bead clean-up. The known deletion was included in the mutation report for MiSeq Sample 44. Both the alpha and beta globin gene clusters appeared to

be balanced, with no evidence of rearrangements visible in the RPKM data or opposite/same direction read reports. It was presumed that the discordant phenotype between this individual and their sibling was produced by a variant in the sibling sample which was not sequenced due to failure during the sample preparation stage.

Table 58 Mutation Report Listing for Rs63750945 in MiSeq Sample 44.

Position	Reference Nucleotide	Coverage	Score	Mutation Call	Mutant Allele Frequency
Chr11: 5248004	G	320	19.3	G>AG	48.53

MiSeq Sample 45 Figure 93.10 (Test)

The RPKM plot for MiSeq Sample 45 showed a deletion of approximately 11 Kb removing two exons of *HBZ*. No opposite or same direction reads showed any clustering at either position.

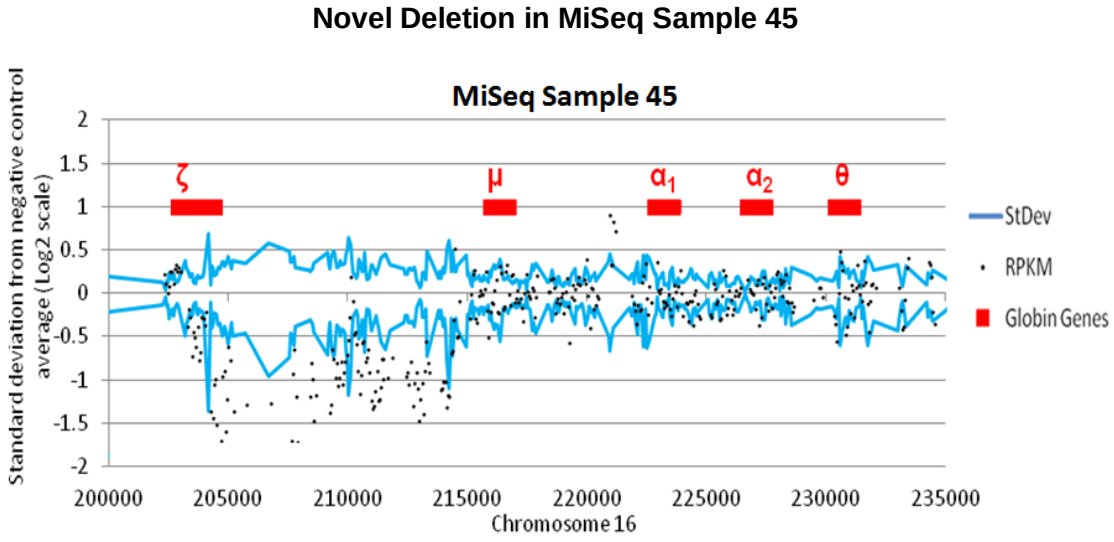


Figure 102: Novel Deletion in MiSeq Sample 45. RPKM plot shows a deletion of approximately 11 Kb removing two exons of *HBZ*

A small number of reads were identified in the alignment that showed strings of mismatched bases that had potentially crossed the break point of the deletion. BLAT query of the reads showed that the regions had extremely high homology to one another. Gap PCR primers were designed against the co-ordinates indicated by the RPKM data and BLAT data to be the potential breakpoints of the rearrangement: PrF (5'-CCTGTAAGGCCACAGGAGAG-3') and PrR (5'-CTCAGCTTCTCCCCTCCTTC-3'). Multiple attempts to amplify the breakpoint product were unsuccessful.

Positions of BLAT Hits From Partially Aligning Reads Found At The Break Point Regions

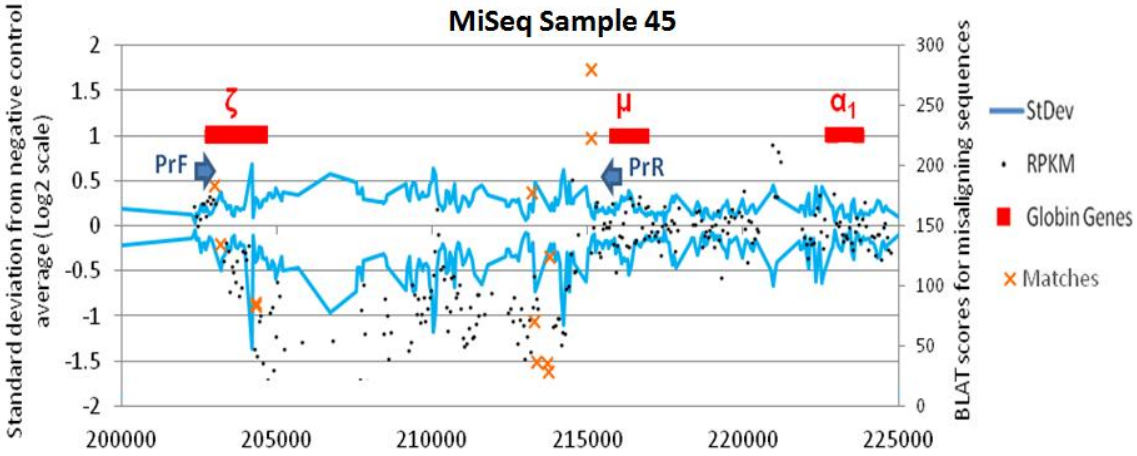


Figure 103 Positions of BLAT Hits from Partially Aligning Reads Found at the Break Point Regions. Secondary Y axis shows BLAT match score for each hit, where two positions attract hits at the 5' end of the deleted region. Approximate positions and orientation or PrF and PrR are indicated by blue arrows.

Unresolved Rearrangements: MiSeq Sample 37, 45.

Rearrangements affecting the alpha globin gene loci proved exceedingly difficult to resolve: homology between the alpha globin genes and their surroundings, large numbers of repeats, frequent CNVs in samples and controls, and the low mapability of the tip of chromosome 16 were all problematic. These features caused issues with (i) successfully aligning reads to accurately identify breakpoints (ii) capturing and aligning sequences containing breakpoints (iii) identifying significant reads against high background noise in opposite and same direction read data (iv) designing unique primers against the estimated breakpoints of rearrangements. Two novel rearrangements sequenced in MiSeq Run 5 (MiSeq Samples 22, 35) could not be resolved with to-the-base accuracy. Another two rearrangements sequenced in MiSeq Run 6 (MiSeq Samples 37, 45) are also unresolved. Attempts were made to resolve all the rearrangements that did not remove the telomere via Gap-PCR, but no unique products were successfully amplified.

Understanding the factors commonly associated with these rearrangements may help improve our ability to target them specifically, and detect them with a greater level of sensitivity.

Evaluating the New Capture Library

A new capture library was successfully used for sequencing samples in MiSeq Runs 5 and 6 (See Methods for details of new design). The new design intended to improve upon the first capture library in the following ways:

- **Reduce Variation between Samples and Provide More Reliable Data**

The new design increased the bait tiling density of the core regions of interest (chr11: 5,000,000-6,000,000 and chr16:0-1,000,000) from 1x to 2x, with the hope that this would increase the coverage depth across this region and also reduce the coverage variability seen between different samples in these locations. We also employed the 'max performance' boosting strategy offered in the Agilent SureDesign suite in order to ensure that baits with a high GC content performed as well as baits with GC <50%. It was not possible to precisely evaluate the difference in performance between these two libraries with the data included in this study. MiSeq runs performed using the old capture library contained less samples, used different shearing strategies and were prepared manually rather than using the BioMek^{FXP}. The different bait densities, regions covered and boosting parameters also made a detailed comparison different. Dr David Brawand (Bioinformatician, Dept. Molecular Pathology, King's College Hospital) made a preliminary analysis of the differences between the coverage variability of four samples from MiSeq Run 5b and four samples from MiSeq Run 2. Random samples of one million read pairs from each sample were used in the analysis in order to control for the different read depth in the two runs.

The regions where the bait tiling density in Bait Capture Library 2 was increased to 2x tiling (where 1x tiling had been used before) was covered at a higher depth than both the 1x region in the old design and the 1x region in the new design. As the coverage increased, the inter-sample variability also increased. We had expected that the increased bait density at regions with high GC content would decrease the level of variation across the covered region, and also the inter-sample variation. This was not the case: inter-sample variability increased with depth of coverage, and the difference in coverage across the region remained broadly the same (NB: The samples used to calculate this data were primarily samples with structural variants affecting the beta globin gene locus. Consequently, the region of the beta globin gene loci (outlined in red) appears to show more inter-sample variability than the rest of the region.).

Inter-Sample Variation between Bait Capture Library 1 and Bait Capture Library 2

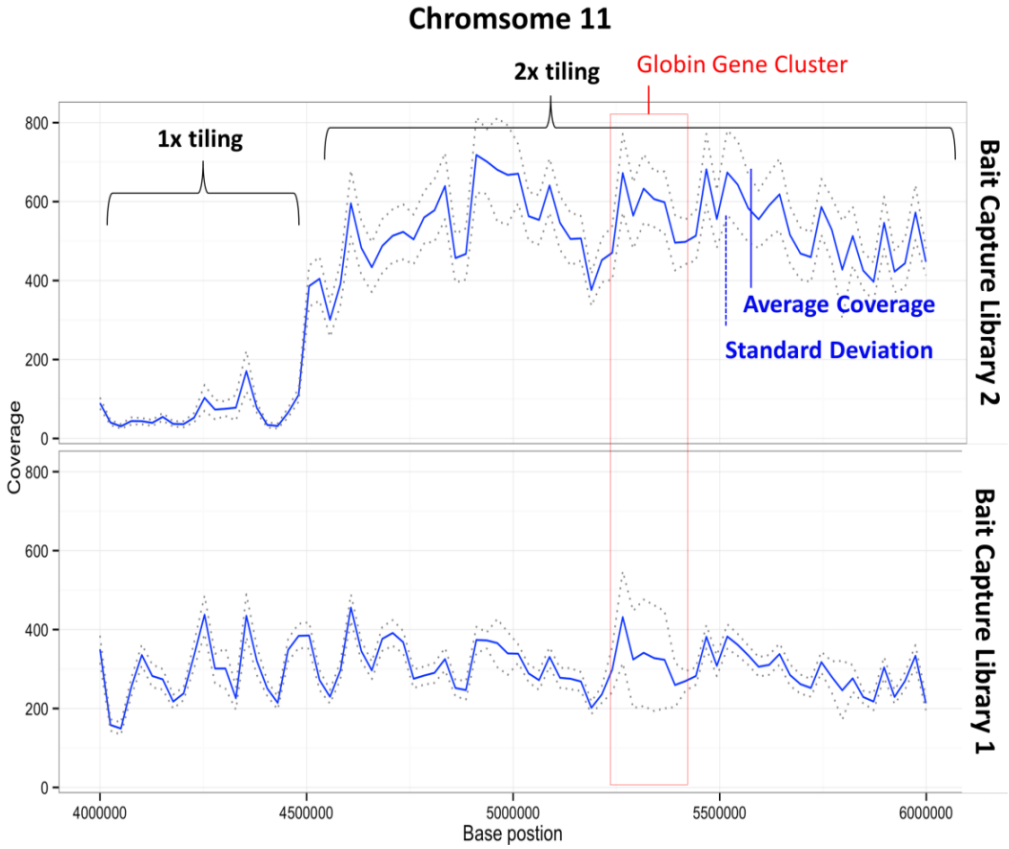


Figure 104: Inter-Sample Variation between Bait Capture Library 1 and Bait Capture Library 2. Upper panel shows part of the region covered by both designs on chromosome 11 (including a region that is covered with 1x tiling in both libraries) in new design. Lower panel shows the same region covered by the Bait Capture Library 1. Solid line shows average coverage and dotted lines show standard deviation from this.

Inter-Sample Variation between Bait Capture Library 1 and Bait Capture Library 2

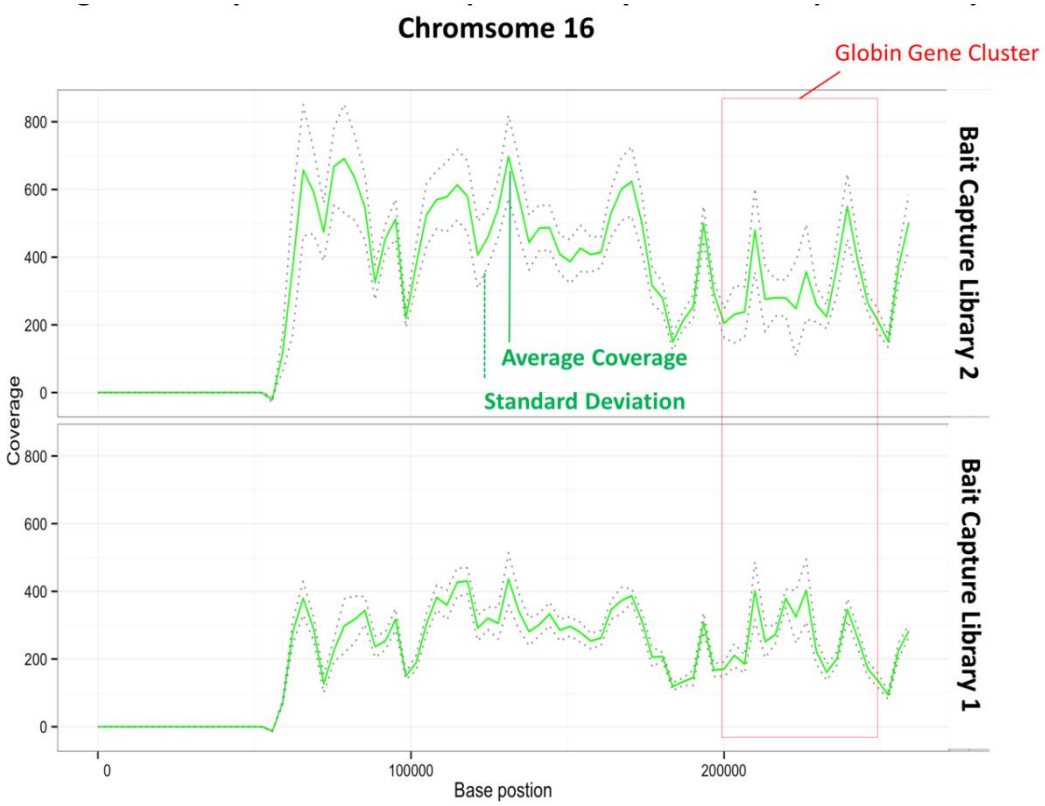


Figure 105: Inter-sample variability between Bait Capture Library 1 and Bait Capture Library 2. The entire region covered by Bait Capture Library 1 is shown in both panels. Upper panel shows variation in this region in Bait Capture Library 2, lower panel shows variation in this region in Bait Capture Library 1. Solid line shows average coverage and dotted line shows standard deviation.

- **Cover a Larger Area of the Regions of Interest**

The new capture library extended the region of chromosome 16 covered by baits from 260,000 bp to 2,000,000 bp. This permitted to-the-base characterisation for MiSeq Samples 34 and 35. These samples could not have been characterised using the old design, as no breakpoint sequence was captured so it was necessary to know the approximate locations of both breakpoints to design Gap-PCR primers. Both of these samples had rearrangements that extended beyond the covered region of the first design.

- **Increase Functionality**

The new capture library included baits that covered larger regions of chromosome 11 and 16. This allowed improved detection of large rearrangements, particularly on chromosome 16. The library also included regions on chromosomes X and Y, which made sex determination possible through this assay. The inclusion of regions on chromosomes 2 and 6 (where variants that modify disease phenotype in haemoglobinopathies have been reported) is yet to be used. The inclusion of these

regions is useful, as determining whether these regions are balanced can differentiate between imbalances on chromosome 11 and 16 that are due to extremely large variants, versus just the result of lower coverage from a particular sample as a whole. This was the case for MiSeq Sample 22, where the entire bait covered region of chromosome 16 appeared to be duplicated. To ensure that this was not due to the sample being overloaded and thus obtaining comparatively more sequence than all negative controls, the RPKM data from the covered regions on chr2, chr6 and chr11 was inspected. These regions were determined to be balanced, indicating that the duplication indices on chromosome 16 were genuine.

Due to the inclusion of regions of the X and Y chromosomes into Bait Capture Library 2, sex determination was possible with this assay. Sex could be successfully determined for every sample where gender was known: females showed an RPKM ratio of 0:1 Y:X, and males showed a ratio of 1:1 Y:X (Table 59).

Table 59: Sex Determination in the New Design Run. The average RPKM value achieved across the covered region of the X and Y chromosomes is shown for 16 samples from female subjects and 12 samples from male subjects. The Y:X ratio in the female samples is 1:1 and in the males is ~1:1.

Gender		F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	M	M	M	M	M	M	M	M	M	M	M	M
Average RPKM	Y	0.68	0	0	0	1.48	1.45	0	0	0	0	0	0	3.29	2.7	0	1.06	175	184	139	141	132	163	96.4	112	153	142	146	127
	X	204	193	231	199	239	171	150	174	216	178	192	165	177	204	196	206	108	108	79.8	112	109	98.2	76.8	72.1	105	98.3	91.2	96.8
Y : X Ratio		0.00 : 0.99	0.00 : 1.00	0.00 : 1.00	0.00 : 1.00	0.00 : 0.99	0.00 : 0.99	0.00 : 1.00	0.00 : 1.00	0.00 : 1.00	0.00 : 1.00	0.00 : 1.00	0.00 : 1.00	0.01 : 0.98	0.01 : 0.98	0.00 : 1.00	0.00 : 0.99	0.61 : 0.38	0.62 : 0.37	0.63 : 0.36	0.55 : 0.44	0.54 : 0.45	0.62 : 0.37	0.55 : 0.44	0.60 : 0.39	0.59 : 0.40	0.59 : 0.40	0.61 : 0.38	0.56 : 0.43

Comparing the Sensitivity of NGS Analysis to MLPA

MLPA data was available for a number of samples included in this study, where the assay had been performed as part of routine diagnosis when these samples were received in the diagnostic laboratory at KCH. Figure 106 shows RPKM plots for four samples included in this study where MLPA data was also available. The positions of probes in the MLPA assay used are depicted by green squares. The MLPA probes effectively identify both duplications and deletions occurring at the alpha globin gene cluster (indicated by a value >1.3 on the secondary vertical axis for a duplication or a value of < 0.5 for a deletion). However, the number of probes and distance between the probes limits the resolution of the assay. A comparison between the resolution of NGS data and MLPA data is given in Table 60.

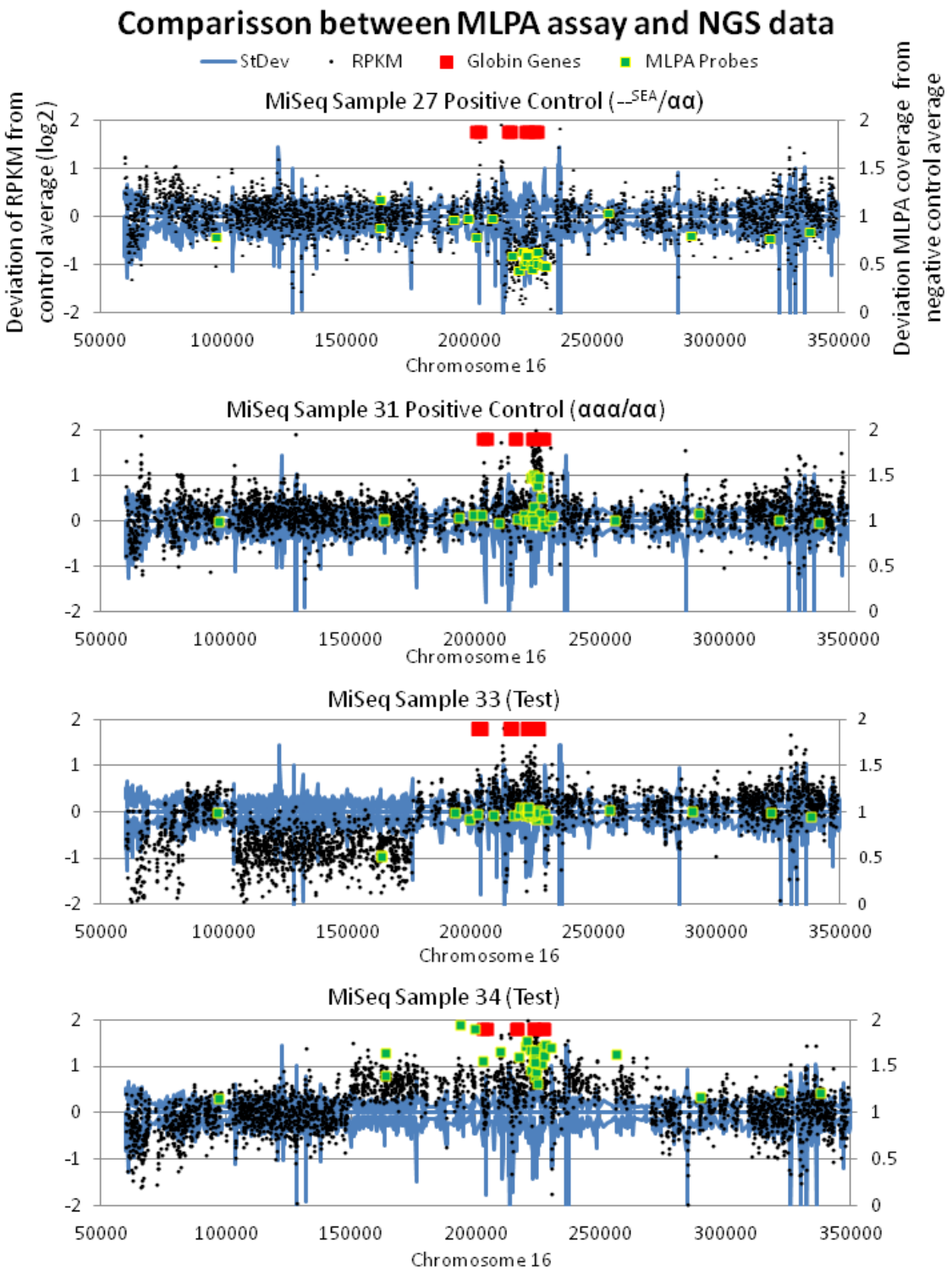


Figure 106 Comparison between MLPA and NGS data for four samples. RPKM plots are shown for four samples analysed in this study, including two positive controls and two novel rearrangements characterised successfully as a result of NGS sequencing. Deviation of RPKM values per bait from the negative control average are plotted on a Log2 scale on the primary vertical axis and represented by black dots on the charts. NegC STDev is plotted in blue on the same scale. The positions of probes included in the Mol. Pathology lab diagnostic assay are shown as green squares with a yellow outline. The deviation of these values in each sample from a negative control average is plotted on the secondary vertical axis. A value > 1.3 indicated a duplication, and a value < 0.5 indicates a deletion in the MLPA data.

Table 60 Resolution achieved for variants through MLPA probe dosage data versus NGS RPKM data (based on distance between first position to show a dosage change and last probe to show balanced dosage at either end of each variant).

Variant	Resolution achieved by MLPA	Resolution achieved by NGS
Positive control (--SEA/)	+/- 16 Kb	+/- 600 bp
Positive control (/)	+/- 264 bp	+/- 120 bp
MiSeq Sample 33	+/- 48 Kb	+/- 1.5 Kb
MiSeq Sample 34	+/- 50 Kb	+/- 7.5 Kb

Resolution is calculated as the distance between probes or baits that indicate balanced sequence, versus probes or baits that indicate a dosage change has occurred. The size of this region is the length of the sequence where the dosage is not known, based on these metrics. While both assays achieve a good resolution of the (ααα/) insertion (albeit at the cost of numerous controls for the NGS assay – see previous section), the resolution achieved by NGS sequencing and RPKM data analysis is significantly higher than the resolution for MLPA for the other three samples. NGS analysis can also be furthered by inspecting the sequences of captured DNA fragments, which is not possible via MLPA analysis.

Chapter Discussion

Over the course of these sequencing experiments, many steps were taken to improve the utility of this assay for detecting and resolving variants for the diagnosis of haemoglobinopathies. The key steps taken to improve the assay were:

- Redesigning the Capture Library
- Increasing DNA fragment size
- Increasing read length
- Increasing the number of negative controls
- Automating sample preparation

Over the course of this study, 25 structural variants were sequenced and analysed. All variants could be detected through RPKM data analysis (Table 61). In 15 of these cases, the rearrangement could be resolved with to-the-base accuracy. This was achieved either by identification of breakpoint sequences in the reads aligning to the reference, or through Gap-PCR and sanger sequencing, where primers were designed

based on the RPKM data analysis. All novel variants identified will be submitted to HbVar.

Table 61 Summary of success of variant characterisation in this study

Variant	Identified in RPKM Data	Breakpoint Identified
Known Alpha English Deletion	Yes	Yes
Known (α^{-MED}/) Deletion	Yes	Yes
Known (α^{-FIL}/) Deletion	Yes	No
Known (α^{-THAI}/) Deletion	Yes	No
Known (α^{-SEA}/) Deletion	Yes	Yes
Known ($\alpha^{-20.5}$/) Deletion	Yes	Yes
Known ($\alpha^{-4.2}$/) Deletion	Yes	No
Known ($\alpha^{-3.7}$/) Deletion	Yes	No
Known (ααα/) Insertion	Yes	No
Novel alpha duplication HiSeq Samples 1, 8	Yes	Yes
Novel alpha duplication HiSeq Sample 3	Yes (but size unknown)	No
Novel alpha deletion MiSeq Sample 33	Yes	Yes
Novel alpha duplication MiSeq Sample 34	Yes	Yes
Novel Alpha Duplication MiSeq Sample 35	Yes	No
Novel Alpha Deletion MiSeq Sample 37	Yes	No
Novel Alpha Deletion MiSeq Sample 38	Yes	Yes
Novel Alpha Deletion MiSeq Sample 45	Yes	No
Known Beta Asian-Indian Inversion Deletion	Yes	Yes
Known Beta HPFH1 Deletion	Yes	Yes
Known Beta 619bp Deletion	Yes	Yes
Novel Beta Duplication HiSeq Sample 7	Yes	Yes
Novel Beta Deletion HiSeq Sample 11	Yes	No
Novel Beta Inversion Deletion HiSeq Samples 5, 10 MiSeq Samples 2, 3	Yes	Yes
Novel Beta Duplication MiSeq Samples 41, 42	Yes	Yes
Novel Beta Deletion MiSeq Sample 43	Yes	Yes

Redesigning the capture library allowed larger variants on the alpha globin gene loci to be characterised with to-the-base accuracy. The redesign also allowed the inclusion of other genomic regions with potential diagnostic importance, such as positions on chromosome 2 and chromosome 6 that may contain variants that modify disease severity of beta thalassaemia and sickle cell disease. Additionally, regions of the X and

Y chromosomes were included in the study, which allowed sex determination. Contrary to our expectations, employing the manufacturer recommended boosting strategy did not reduce variation across the region of interest or between samples compared to our first design. Furthermore, increasing the bait tiling density of this region increased the level of inter-sample variability in the assay, as a result of the increase in coverage. While this did not help to reduce the noise associated with the assay, the increased bait density in the core locations for the assay increases the amount of sequence captured in these regions, allowing a higher read depth per sample.

Increasing DNA fragment size and read length improved the capability of this assay to sequence into repetitive regions and increased the alignment specificity for captured sequences. The increased read length also allowed longer strings of mismatched sequences to be included in the alignment, which improved break-point characterisation. There were still issues with correctly aligning reads containing small indels to the reference sequence, with many being discounted as the software tried to 'shoe-horn' reads into alignments that matched the reference sequence. Where indels were included in the mutation report, they were indexed incorrectly, resulting in multiple entries in the report for each base affected by the indel, rather than a single call for the entire variant. Single nucleotide changes that were known to occur in certain samples were successfully identified in the variant reports for both capture libraries, where the most severe impact to their confidence scores came from low read depth.

Increasing the number of negative controls used in the assay was crucial to successful characterisation of the structural variants. Negative control samples can be used to reduce the noise in the assay even when they are not prepared or included in the same run. As more sequencing runs are carried out, it would be useful to assemble a 'bank' of negative controls that can be used to analyse data in new runs. This will increase the number of test samples that can be included per run, although some negative controls should still be prepared and sequenced with each batch of test samples for quality control purposes, to check inter-run variation.

The utility of these negative controls from other runs is dependent on minimising the inter-run variation. Inter-run variation (between runs on the same sequencing platform) appears to be produced during sample preparation and hybridization. In order to streamline the time consuming sample preparation process, and also to reduce variability between runs, a robotic sample preparation platform was introduced. There were extensive problems with programming the platform to reliably carry out sample preparation. For MiSeq Run 5 and Run 6, sample preparation was carried out on the

robotic platform, but the sensitive hybridization stage was still performed manually. It was possible to use negative controls from one run to identify structural variants occurring in the other run thanks to the uniformity in the sequencing data achieved by this method. An automated hybridization step could further reduce the inter-run variability of this assay, providing it could be engineered to work reliably.

Future Directions: Moving Away From the NextGene Software Package and on to a Custom Analysis Pipeline

The NextGene software package was an essential tool to develop this assay. The department had no experience with manipulating and analysing NGS data at the beginning of this project, and an off-the-shelf toolkit was instrumental in determining whether this technology was able to fulfil the requirements for a diagnostic assay, to optimise the sample preparation procedure and the bait capture library for our purposes, and to evaluate our data while developing a routine SOP for using this technology in the diagnostic laboratory. The platform was found to have some limitations which meant that not all variants in this study could be reliably characterised, particularly where the breakpoints occurred within repetitive or homologous regions. Variant characterisation with this software was performed on a time-consuming, case-by-case basis. The structural variant detection tool offered within the NextGene package was found to be extremely poor, calling large numbers of false positive results and missing known variants within the samples. Additionally, while excellent at reporting SNPS, NextGene was extremely poor at reporting small indels in the mutation report, where it called each base affected by an indel as a distinct variant, which impeded efficient analysis of the data because these redundant calls needed to be separated from the clinically relevant data.

Developing a Bespoke Sequencing Analysis Pipeline for Diagnostic Use

A bioinformatician was recruited into the diagnostic laboratory to develop a custom analysis pipeline for analysing NGS data from this assay, as well as others that were being concurrently developed on site after the acquisition of the MiSeq platform. Moving to a custom pipeline will provide multiple advantages in variant analysis. NextGene is an expensive tool which was not built specifically for the purposes of this study. This is exemplified by the use of RPKM throughout this study as a normalised measure of coverage, which is a tool intended for use in RNA sequencing. Identifying and characterising both known and novel structural variants in NextGene requires the user to manually compile data from up to five separate sources (Expression Report, Mutation Report, Opposite Direction Reads Report, Same Direction Reads Report,

NextGene Viewer) and even go back into the raw FASTA data. A purpose-built tool can avoid work-arounds of this nature, and also has the advantage that the inner workings of the tool are precisely understood, and can be manipulated for optimal efficacy.

The pipeline uses multiple publicly available and widely used SNP callers together to create extremely accurate consensus variant calls. This includes haplotype-aware SNP calling programs that evaluate variant calls based on known information about the allelic backgrounds on which different variants commonly occur.

The pipeline also identifies structural variants with high sensitivity using multiple tools. A CNV calling algorithm called ExomeDepth is the primary tool used to detect dosage changes (Plagnol, Curtis et al. 2012). This tool controls for technical variability between samples by creating an optimized reference set of controls for comparison with sample data. The tool aligns read data to a reference using a beta binomial distribution (i.e. determining the best alignment of each read based on maximum likelihood, assuming the data has a binomial distribution where the probability of success is random). A hidden Markov model is used to calculate the probability of three scenarios (deleted, balanced, duplicated) for each covered exon in the design and identifies the scenario which best fits the coverage data, corroborated by pre-existing estimates of the likelihood of each scenario in each exon, based on previous data. The model also applies penalties for sequencing depth, GC content and Phred score to the calls it makes. This tool is most sensitive when calling rare CNVs that are unlikely to occur in a normal reference dataset, so is well-suited to detecting novel or rare rearrangements causing thalassaemia and distinguishing them from the frequent copy number variants occurring in these regions. Additional tools are also used to increase the power of the analysis, including ‘Delly’ and ‘Lumpy’. Lumpy identifies rearrangements in alignment data based on similar attributes to those compared manually from NextGene data in this study, including coverage, discordant read mapping (pairs of reads mapping at unexpected distances or orientations to one another) and split read mapping (where one read from a pair crosses a rearrangement breakpoint and therefore maps partially to two normally non-adjacent sequences or with a change in orientation) (Layer, Chiang et al. 2014). Delly operates in a similar manner, focussing on discordant read mapping. The combined use of all these programs allows sensitive variant detection by the analysis pipeline (Rausch, Zichner et al. 2012).

The implementation of this pipeline is expected to call both single nucleotide variants and structural rearrangements with superior accuracy to NextGene, and to require less user input to do so. This tool will allow the sample preparation techniques optimised

during this study to be used for routine diagnosis, rather than for case-by-case analysis developed in NextGene.

Rapid Detection of known breakpoints in FASTA data

One novel feature that could be incorporated into this pipeline which was developed in this study is a search for sequences within the FASTA file that are unique to known variants. This could provide rapid and reliable detection of certain variants without the need to perform alignment and other downstream analysis steps. As proof of principle for this concept, we devised a script that could search FASTA files for sequences containing the unique breakpoint sequences produced by several variants (the sickle cell variant, the Asian-Indian inversion-deletion, the 619 bp deletion, and English V inversion-deletion). The script also searches for the wildtype counterparts to these sequences. It then returns the result of “Negative”, “Homozygous” or “Heterozygous” calls for each variant. The output of the script for HiSeq Sample 6 (Positive control for the Asian-Indian inversion-deletion) is shown in Figure 107. The results of running this script for read files from several variants are shown in Table 62. The script itself is included in Appendix 5. In future versions of the Bait Capture Library, it could be possible to include additional baits which target these sequences. However, as many breakpoints occur within repetitive regions, including baits to target them could have the consequence of increasing off-target sequence captured at these regions.

Variant Finder Script in Python Shell

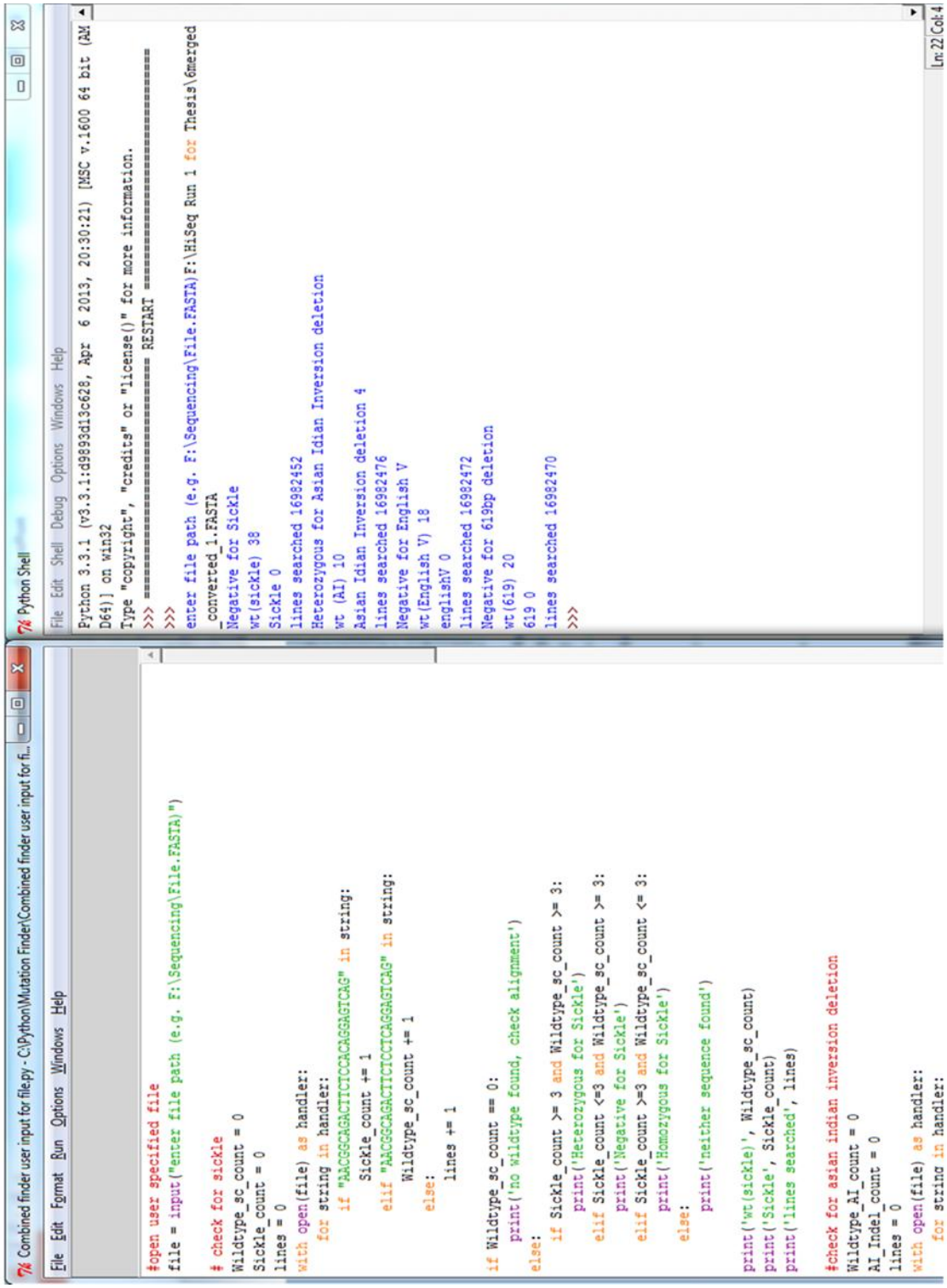


Figure 107: Variant Finder script applied to HiSeq Sample 6. Left window shows script, right window shows output in python shell.

This is an extremely basic example of how this strategy could be incorporated into data analysis. The only stringency measure included in this script is that a minimum of three reads including either the mutant or wild type sequence are necessary to form a diagnosis. If neither sequence is identified in >3 reads, the script returns “neither

sequence found". In two of the cases shown here, a reads that contain the unique sequence for the sickle cell variant are picked up in samples which are not carriers of this variant. This demonstrates the necessity for employing a minimum read requirement to provide accurate variant calling, an appropriate threshold for which would need to be investigated. In order to be suitable for diagnostic applications, this script would need to be heavily modified to provide more robust results. Requirements of a script for diagnostic use would include a score calculation for each call, tolerance for SNPs or miscalls within the variant and normal allele-identifying sequences and detection of the reverse-compliment of the sequences. A provision could also be included for unusual allele frequency, as in the case of HiSeq Sample 7 (in which the sample has a duplication of the beta globin gene cluster resulting in an unusual mutant allele frequency for the sickle cell variant). Determination of reliable cut-offs for false positive and false negative read numbers would need to be calculated, and provisions would also have to be made for the possibility of mosaicism and contamination in the sample.

Table 62 Output of Variant Finder script for multiple samples

*** "Neg" = Negative "Het" = Heterozygous "Hom" = Homozygous	HiSeq Sample 2 (Negative Control)	HiSeq Sample 7 (Sickle Cell Het)	HiSeq Sample 6 (AI Indel Het)	MiSeq Sample 2 (English V Deletion Het)	MiSeq Sample 3 (English V Deletion Het)	MiSeq Sample 5 (619bp Deletion Het)
Sickle Cell Test: Call	Neg	**Het**	Neg	Neg	Neg	Neg
Normal Sequence Count	46	**78**	38	110	72	65
HbS Sequence Count	0	**41**	0	0	1	1
Asian Indian Inversion Deletion: Call	Neg	Neg	**Het**	Neg	Neg	Neg
Normal Sequence Count	21	60	**10**	80	55	44
AI Indel Sequence Count	0	0	**4**	0	0	0
English V: Call	Neg	Neg	Neg	**Het**	**Het**	Neg
Normal Sequence Count	14	28	18	**37**	**35**	38
English V Sequence Count	0	0	0	**22**	**30**	0
619 bp deletion: Call	Neg	Neg	Neg	Neg	Neg	**Het**
Normal Sequence Count	32	85	20	76	82	**24**
619 bp Sequence Count	0	0	0	0	0	**32**

Results Chapter 3: The Basis and Characteristics of Structural Rearrangements Affecting the Alpha and Beta Globin Gene Loci

Introduction

At the alpha and beta globin gene loci, a large number of structural variants of different kinds have been reported. Different parts of the genome are subjected to different types of structural rearrangement, and the events occur at different rates. Understanding what factors contribute to the formation of these rearrangements is of scientific interest, and is potentially useful for diagnosis. This chapter will examine how recorded variants at the globin gene loci relate to known drivers of structural rearrangement to see if they have a common cause.

Interactions between features of the DNA sequence and the pathways that repair DNA breakages – particularly double stranded breaks (DSBs) – lead to the formation of structural rearrangements. Some of these interactions have been described in detail to explain how certain rearrangements arise. Other interactions are not yet fully understood.

Features of the DNA Sequence that are Associated with Structural Rearrangements

DNA bases are charged molecules, and the sequence of bases within a section of DNA can influence its 3D structure: complimentary sequences can contort the entire molecule because of their attraction to one another. Certain sequences of bases can cause tight loops to form, or for the molecule to take on non-double helix conformations such as hairpins and quadruplexes. Various DNA features have been reported to be associated with the formation of structural rearrangements and DNA mutation (Lupski 1998, Bacolla, Jaworski et al. 2004, Gu 2008, Bacolla 2009).

Repeats

Non-coding, repetitive DNA sequences make up over half of the human genome (Richard, Kerrest et al. 2008). These repeats can be divided into several classes based on their sequence, or the pattern which their sequence takes. Different classes may have different functional properties in the genome. The classes of repeats found at the

globin gene loci on chromosomes 11 and 16 are described in Table 63. Some repetitive DNA sequences are highly homologous and the breakpoints of structural rearrangements often fall within, or near to, these repeats (Lovett 2004). Repetitive elements arrive and propagate through DNA via different mechanisms, so they are not evenly dispersed throughout the genome (Katti, Ranjekar et al. 2001). LINE repeats are extremely common in the region of the beta globin gene cluster on chromosome 11 and less frequent around the alpha globin gene cluster on chromosome 16. Conversely, Alu-type SINE repeats – the most common class of repeats in the human genome – are more abundant on chromosome 16 than chromosome 11 (Richard, Kerrest et al. 2008).

Table 63 Major repeat types found at the globin gene loci (data from (Katti, Ranjekar et al. 2001, Antonarakis 2010)

Repeat type	**Description**
Long Terminal Repeat (LTR)	Long terminal repeats are vestiges of proviral DNA formed by reverse transcription of retroviral RNA
Tandem Repeat	Repeated iterations of the same sequence of bases. Depending on the number of bases forming the unit of sequence which is repeated, these are subdivided into dinucleotide repeats, minisatellite repeats, microsatellite repeats, and short tandem repeats (STRs).
DNA Transposons	Cut-and-paste repeats, or Class II Transposons are a common type of repeat left behind by viruses which multiply through the genome by replicative transposition
Non-LTR Retrotransposons ● Short Interspersed Nuclear Elements (SINE) ● Long Interspersed Nuclear Elements (LINE)	In the human genome most SINE repeats are of the subclass ‘Alu’ The complete SINE repeat sequence is 280 bp long In the human genome most LINE repeats are of the subclass ‘L1’ The full LINE element is 6 Kb in length
RNA repeats	Repeated RNA sequences, often including pseudogenes

Segmental Duplications

Segmental duplications (also known as low copy repeats) are 1 – 400 Kb highly homologous (>95%) and region-specific repeats. Particular segmental duplications are often confined to a single chromosome or chromosomal region, so are classed differently from the other types of repeats which occur many thousands of times throughout the genome (Samonte and Eichler 2002, Gu 2008). Segmental duplications make up approximately 5% of the genome (Samonte and Eichler 2002). They have been found to coincide with the breakpoints of recurrent deletions and duplications where their homology mediates non-allelic homologous recombination (NAHR). Segmental duplications are frequent in the region of the alpha globin gene locus.

Non-B DNA

Non-B DNA conformations are 3D structures that double stranded DNA can adopt, other than the Watson-Crick double helix (Bacolla 2009). Non-B DNA conformations include triplex and tetraplex helixes, cruciforms and hairpins (Table 64). The ability of DNA to adopt these structures is dictated by sequence motifs, such as mirror repeats, G Quadruplexes and Z DNA. Some non-B DNA conformations are associated with rearrangements and sequences that take on non-B DNA conformations are more vulnerable to double stranded breaks (DSBs) than DNA that retains the double helix formation (Stankiewicz and Lupski 2002, Bacolla 2009). When DNA breakages occur, the physical shape of the DNA and the sequence occurring at the breakpoint both affect the manner in which it is repaired, giving rise to different types of DNA rearrangement. Tetraplex DNA is associated specifically with duplications, the breakpoints of which occur in G-Quadruplex DNA repeats. Slipped hairpin and cruciform structures have been associated with sequence inversions (Bacolla 2009, Lam, Beraldi et al. 2013).

Table 64 Non-B DNA motifs from (Bacolla 2009)

Non-B DNA Motif	Composition	Resulting non-B conformation
G-Quadruplex	> Four iterations of the sequences GG, GGG or GGG separated by 1-7 other nucleotides	Tetraplex
Z DNA	Dinucleotide repeats of CG CG CG ... and occasionally TG CA TG CA TG ...	Left-handed Z DNA
Mirror Repeat	TTGGACTTCAGGTT	Triplex
Inverted Repeat	TCGGTTACCGA AGCCATTGGCT	Cruciform
Direct Repeat	TCGGT·TCGGT·TCGGT	Slipped hairpin
Short tandem repeats	Repeated iterations of a 2-13 nucleotide sequence	Slipped DNA
A phased repeats	Repeats of alpha nucleotide sequences interspersing normal sequence at regular (~10bp) intervals	Static bending

Origin of Replication

During DNA replication many small sections of the DNA are concurrently 'unzipped' creating replication forks, both strands of which are then copied simultaneously (Slack, Thornton et al. 2006). The DNA segments replicated from all the forks along the chromosome are then eventually joined together to make a complete copy. The fork start positions are 'origin of replication' (ORI) points (Tiller 2000). Structural rearrangements can arise from errors that occur during fork replication. In a study of replication in plasmids transformed into *E. Coli*, deletion breakpoints were found to frequently occur within 150bp of a ORI (Bacolla, Jaworski et al. 2004). For this reason, the location of ORI in relation to the breakpoints of rearrangements in thalassaemia is of interest. In budding yeast, where ORI have been studied most closely, these points are marked by the same 11 bp sequence. In humans, multiple sequences are associated with ORI. A database of replication origin points mapped in humans to date has been created (DeORI) (Gao, Luo et al. 2012), which lists ORI that have been identified by replication initiation point mapping in cell lines as described in (Karnani et al 2009). The database currently includes 433 entries from human cell lines. A common feature of replication origin sites is their high AT content. The database also includes

replication origin sites from cancer cell lines, which are extremely abundant due to interruptions to the cell replication process.

DNA Repair Pathways

DNA molecules can experience breaks, interstrand crosslinks and other issues during replication. Several mechanisms are in place to repair these breaks. A successful repair may completely restore the original sequence. Alternatively, homologous recombination may occur, in which the two copies of a chromosome exchange allelic segments with one another (Figure 108). Structural rearrangements result from incomplete repair of a DNA break via one of the numerous non-homologous repair mechanisms available. In these instances, repairing the break results in the loss, insertion or translocation of sequence.

Homologous Recombination Hotspots

Positions where homologous recombination has taken place can be identified by changes to the chromosome haplotype – the pattern of SNPs that occur along the chromosome. These patterns change when homologous recombination has occurred, swapping part of the haplotype from one allele with another. The variety of haplotypes that occur across a region can be used to calculate its recombination rate. This value is usually given in centimorgans (cM). The HapMap Project has identified large numbers of haplotypes in the population due to the frequent homologous recombination events at certain positions. These regions are called recombination hotspots. There are two recombination hotspots at the site of the beta globin gene cluster on chromosome 11, one situated within the cluster and another in close proximity to it (Paigen and Petkov 2010). The large number of haplotypes reported across the beta globin gene cluster indicate that homologous recombination events are frequent at this position (Heyer, Ehmsen et al. 2010).

Homologous recombination involves equal exchange of genomic material and is mediated by different repair mechanisms than events which lead to non-homologous recombination. In spite of this, there is some correlation between the locations of segmental duplications and homologous recombination hotspots. As such, recombination rate will be evaluated as a potentially contributing factor in the generation of thalassaemia-associated structural rearrangements (Gu 2008).

Non-homologous recombination

Several mechanisms repair DNA breaks via means other than reciprocal homologous recombination: Non-allelic homologous (NAHR), Single Strand Annealing (SSA), non-

homologous end joining (NHEJ), and microhomology-mediated end joining (MMEJ). These repair mechanisms leave deletions, duplications, insertions and inversions of various sizes in their wake, which sometimes include specific DNA signatures identifying the repair mechanism that created them.

Non-allelic Homologous Recombination (NAHR)

Non-allelic homologous recombination (NAHR) is a DNA repair mechanism that occurs between two homologous sequences that are not allelic counterparts (Figure 108). Recombination via NAHR is closely associated with segmental duplications.

Homologous Recombination and Non-Allelic Homologous Recombination

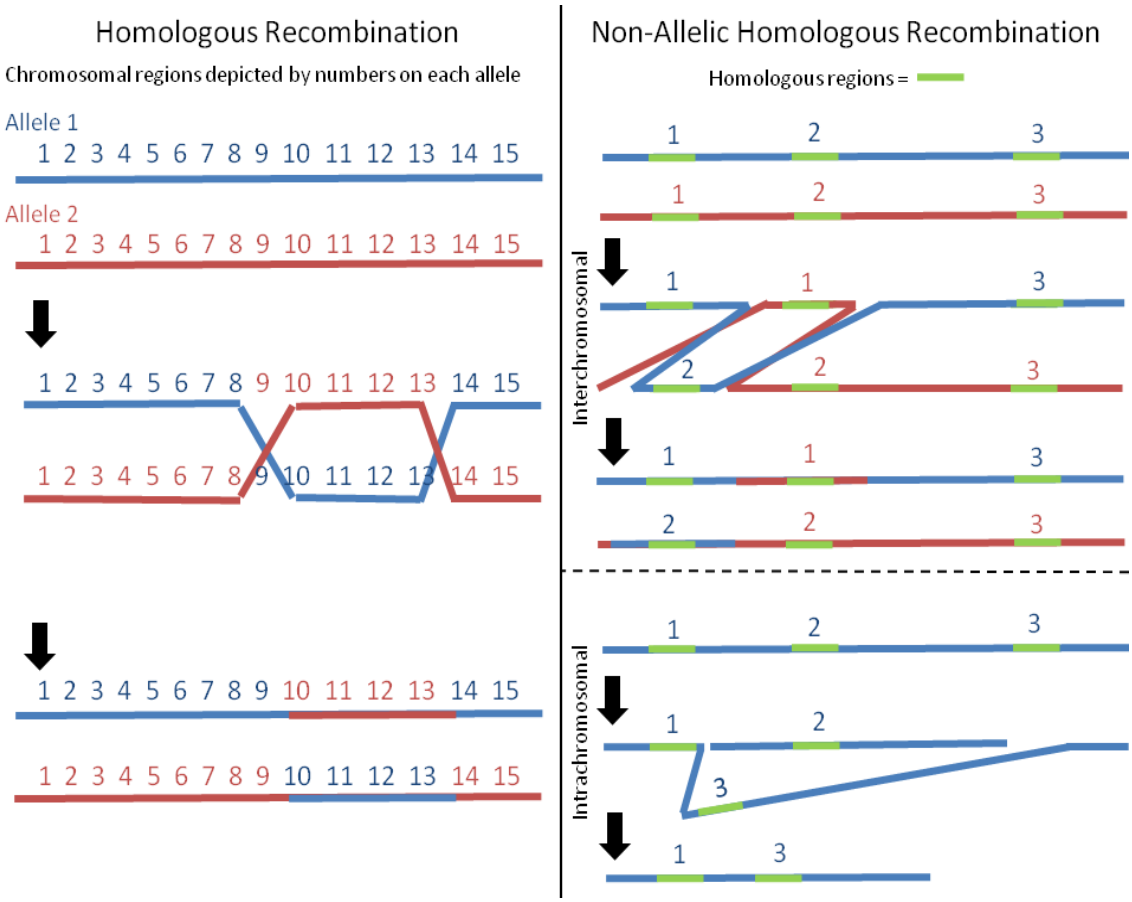


Figure 108 Homologous Recombination and Non-Allelic Homologous Recombination (NAHR). Two alleles are depicted in red and blue, with the homologous sequence along each allele represented by numbers. In left panel, reciprocal crossover between the homologous and allelic sequences results in an equal exchange of homologous sequence. In the right panel, three low copy repeats (green) which are all highly homologous but NOT allelic are situated in this region. Homologous recombination between these regions can result in interchromosomal or intrachromosomal NAHR.

Examples of NAHR include the frequently recurring duplications causing Charcott-Marie Tooth Syndrome (*CMT1A*) and the haemoglobin alpha (α-$^{3.7}$/) and (α-$^{4.2}$/) deletions and their counterparts. Non-recurrent recombination events can also occur via NAHR, primarily between repetitive and highly homologous Alu or LINE sequences (Gu 2008).

Charcott-Marie Tooth Syndrome is caused by a duplication on chromosome 17 in 70% of cases, 10% of which are de novo events (Blair, Nash et al. 1996). The duplication break points invariably occur in two discreet regions that are segmental duplications with a high level of homology. In the alpha globin gene cluster, the ($\alpha^{-3.7}$/) and ($\alpha^{-4.2}$/) deletions, plus the (ααα) triplication complex that is the counterpart of the ($\alpha^{-3.7}$/) deletion all occur between low copy repeats with high homology (Emmanuel 2001, Samonte and Eichler 2002). A study by Lam et al. investigated the frequency with which these rearrangements occurred during meiosis and mitosis (using assays to detect the rearrangements in sperm and blood). In two non-carrier men, the deletions arose spontaneously at a rate of $6.7/6.8 \times 10^{-6}$ per haploid genome in blood and in sperm at a rate of $16 \times 10^{-6}/68 \times 10^{-6}$ per sperm. Four percent of these rearrangements were the leftward 4.2 conformation and 96% were the rightward 3.7 conformation (Lam and Jeffreys 2006). The (ααα/) triplication did not occur at the same frequency as its counterpart, ($\alpha^{-3.7}$). This indicates that NAHR can occur intrachromosomally between the two points on a single chromosome - creating a single deletion allele - as well as interchromosomal NAHR - creating both alleles by removing the segment from one chromosome and inserting it into the other (Lam and Jeffreys 2006, Gu 2008). Studies of recurrent duplications in other regions noted that minimal efficient processing segments (MEPs) were required for recurrent segmental duplication mediated recombination events to occur. This is the length in nucleotides and percentage of homology required for recombination. MEPs were estimated to be 300-500bp in length based on studies of several genomic regions. The segmental duplications in the alpha globin gene region, however, were found by this study to have substantially shorter MEPs: recombination could occur between homologous stretches of as little as 37 nucleotides, owing to the close proximity of the alpha segmental duplications to one another (Gu 2008, Liu, Lacaria et al. 2011).

Single Strand Annealing (SSA)

Single strand annealing digests DNA on either side of a break point until a microhomology between the ends is reached. The DNA is trimmed to these regions and the two ends are then ligated back together. Single strand annealing principally occurs at tandem repeat sequences, with the general result that one of the two homologous repeats is deleted, along with the intervening sequence (Lee 2014). This mechanism of repair is considered to be distinct from NHEJ and MMEJ because it is mediated by Rad51-family proteins which are associated with homologous recombination, rather than Ku and PARP which govern NHEJ and MMEJ (Sancar, Lindsey-Boltz et al. 2004).

Non-Homologous End Joining (NHEJ) and Microhomology-Mediated End Joining (MMEJ)

NHEJ is frequently used to repair unequal DSBs, and evidence of NHEJ activity can be identified at DNA break points by the 'signatures' it leaves behind. MMEJ appears to be a less frequently used mechanism, and is considered by some to be a subtype of SSA repair, although it appears to follow a different pathway and have a less destructive impact on the sequences it operates on (McVey and Lee 2008).

When a DSB is detected, molecular bridging from both broken strands is employed to make them compatible and ligatable. The two sequences need to have some homology to join correctly. This can happen between regions with existing microhomologies (5-25 bp) via MMEJ, frequently leading to the loss of any sequence between the microhomologous regions (Boulton and Jackson 1996, McVey and Lee 2008). Alternatively, the DSB can be repaired via NHEJ which either joins regions with small microhomologies (1-5 bp), or otherwise creates small sequence insertions, deletions or duplications to produce microhomology (Roth and Wilson 1986, Lupski 1998). The breakpoints of NHEJ and MMEJ events are frequently within or in close proximity to LTR, LINE, Alu, MIR and MIR2 sequences. As these process can used to mediate repair of DSBs created by non-B DNA conformations, breakpoints are also often associated with these DNA motifs, particularly inverted repeats. Both deletions and duplications have been attributed to mistakes by NHEJ repair of DSBs. Non-recurrent tandem duplications have been described that show junction microhomology and small insertions characteristic of NHEJ repair. These duplications were believed to be the result of NHEJ and homologous recombination (HR) being employed in combination to repair a DSB, and in doing so invaded and copied the sister chromatid (Gu 2008). MMEJ is always associated with deletion events, removing bases until a microhomology is found (McVey and Lee 2008).

Fork Stalling and Template Switching (FoSTeS)

As described previously, DNA replication is initiated at multiple points across the chromosome simultaneously, at ORIs. At these locations, the DNA opens into fork structures, and both strands of open chromatin are replicated simultaneously. During this process it is possible for replication to stall on one strand of a fork, causing it to lag behind the replication of the complimentary strand. The lagging strand can then become disengaged from the DNA it is replicating and anneal to another fork that is open and has microhomology to the lagging sequence. The replicating strand continues to extend along the new fork and may disengage and jump to other forks

before returning to its original position (Michel 2000, Jennifer A. Lee 2007). These forks could be from other regions of the same chromosome, or other chromosomes currently undergoing replication. FoSTeS can create high complexity rearrangements, such as inversion-deletions. FoSTeS events are associated with segmental duplications, Alu repeats and high GC sequences. Some non-B DNA conformations such as palindromic sequences and loop structures may provoke a stalling event (Jennifer A. Lee 2007, Gu 2008).

The genomic landscape at the alpha and beta globin gene loci

The alpha and beta globin gene clusters are situated on different chromosomes. The two gene clusters exhibit many similarities, and the genes are believed to have evolved from a single common ancestor via a series of duplication and gene conversion events (Czelusniak J 1982, Zhang 2003). Conversely, there are also some marked differences in the features of the two clusters (Table 65)

Table 65 The alpha and beta globin gene clusters: similarities and differences

Alpha globin gene cluster	**Beta globin gene cluster**
Similarities	
• Arose by duplication • Both clusters include 5 active genes, two of which are close homologues (*HBA1, HBA2, HBG1, HBG2*) • Each globin gene consists of three exons and two introns • Organised in order of expression • Structural variants of the genes are common: 800 variants with no clinical significance have been recorded to date • Situated on the short arms of their respective chromosomes near the telomere • Duplications recorded to date are top-to-tail in orientation • Both clusters include pseudogenes	
Differences	
• Situated on chromosome 16 • The first location where the globin	• Situated on chromosome 11 • Thalassaemia is mainly caused by point mutations and deletions are rare

genes evolved (Hardison 2012) • Thalassaemia is mostly caused by deletions and point mutations are rare • No thalassaemia causing variants in the promoter have been recorded • Cluster is spread over 28Kb • Genes cover 800-1.6Kb of sequence	• Cluster spread out over 45Kb • Genes cover 1.5-2 Kb of sequence • Two recombination hotspots are situated within the cluster • The cluster is near to an 'origin of replication' point.

Variants that are associated with thalassaemia differ between the two globin gene clusters: Most reported variants on chromosome 16 are deletions, while those affecting chromosome 11 are mainly point mutations (Trent 2006). On a more subtle level, duplications of the alpha globin genes that can modify the severity of beta thalassaemia disease phenotype have been recorded with a higher frequency than duplications of the beta globin cluster, and inversion-deletions have only been recorded on the beta globin gene cluster. Duplications that have been identified in the beta globin cluster to date have been phenotypically silent, unless inherited with another globin variant or thalassaemia. These co-inheritance incidents are the only occasions when these duplications have been detected, so the frequency of phenotypically silent duplications at this locus will not be known until we start screening DNA on a wider scale (Harteveld, 2008).

This study has shown that features of the DNA sequence can impact the ability of NGS to resolve some rearrangements:

- The (α-$^{3.7}$/) deletion and insertion are difficult to detect because they affect one of two highly homologous regions and has no identifiable breakpoint sequence (See Results: 'Detection of the 3.7 Kb Deletion and Insertion')
- The English V inversion-deletion first characterised in this study was difficult to resolve because it interrupted a DNA palindrome that impeded sequencing
- Rearrangement breakpoints frequently fell in repetitive regions that had not been sequenced.

Identifying DNA features that associate with rearrangements at the globin gene clusters could improve our ability to locate likely breakpoints positions that are not readily apparent from NGS sequencing data. This could improve diagnosis in thalassaemia and be scientifically valuable.

Previous studies of structural rearrangements at the globin gene loci

A previous attempt to identify common factors causing rearrangements at the alpha globin gene cluster was undertaken by Nicholls et al (1987). The study looked for features that united 12 known deletions causing alpha thalassaemia. They found that five of these deletions removed a similar region of sequence (20-30 Kb) and at least one breakpoint in each deletion mapped within, or close to, an Alu repeat (Nicholls, Fischel-Ghodsian et al. 1987). Despite this observation, this study was not able to draw conclusions about the significance of the relationship between Alu repeats and rearrangements on chromosome 16 due to the small sample size used and the extremely high frequency of Alu repeats throughout the genome. Regardless, they posited that mechanisms by which Alu repeats could induce deletion events were that their homology created misalignments in meiotic chromosomes, or that Alu repeats may act as origins of replication.

One deletion, ($\alpha\alpha^{RA}$/) was found to seamlessly join two homologous Alu repeats, resulting in a breakpoint that was indistinguishable from normal sequence. They concluded that this indicated the deletion mechanism was a linear chromosomal deletion, representing the first example of a deletion resulting from homologous recombination between two members of a family of interspersed repeats to be reported in higher eukaryotes. Other features noted in the study were a GC rich inverted repeat adjacent to one deletion breakpoint, and short region of mild homology between the break points of another deletion including a 6-8bp direct repeat.

In their analysis of the ($-^{Med}$/) deletion they found a 134 bp insertion at the deletion break points. This was identified as an Alu repeat fragment originating from 37 Kb upstream of the affected region, which had be reinserted in an inverted orientation at the break point position.

Nicholls et al. noted that the deletions tended conform to a distinct size range of 20-30 Kb. They hypothesised that this could be attributed to chance and the small sample cohort, or that these regions has characteristics making them particularly prone to recombination. Further, they could be a consequence of the way chromatin at this position loops during the cell cycle, with the loop between these two points being removed. This is supported by the chromatin loop model proposed by Vanin et al (1983) of rearrangements on the beta globin gene cluster. The insertion of novel inverted sequence in the ($-^{Med}$/) deletion, originating from a position a similar distance away, could also be explained by this model (Vanin, Henthorn et al., 1983).

The study by Vanin et al examined four deletions impacting the beta globin gene cluster, also aiming to identify a common underlying feature of these rearrangements (Vanin, Henthom et al. 1983). The DNA breakpoint sequences of these deletions were studied via restriction digest mapping and Maxam-Gilbert sequencing of the resulting fragments. Vanin et al did not find any homology occurring at the break points in the four variants they examined at the beta globin gene cluster, although one of the eight breakpoints examined was in close proximity to an Alu repeat, which had also been reported at the break points of other deletions causing beta thalassaemia. The principle finding reported by this study was that the deletions analysed were all approximately the same length, removing different but equally sized sections of sequence (Figure 109).

Deletions at the Beta Globin Gene Cluster from Vanin et al. (1983)

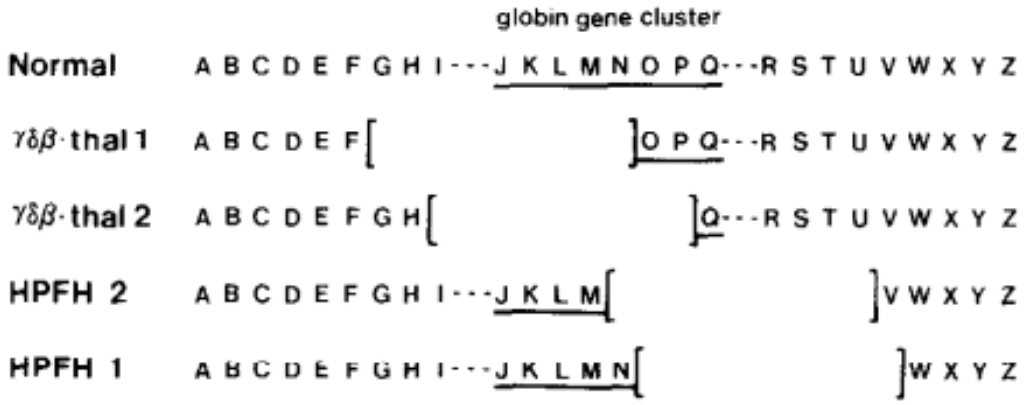


Figure 109: Deletions at the Beta Globin Gene Cluster Studied by Vanin et al. "*Schematic representation of normal DNA and deletions in γδβ-thalassaemia samples 1 and 2 and HPFH deletions 1 and 2. Letters A through Z represent normal DNA; the β-globin gene cluster is underlined. The brackets surround the regions known or predicted to be deleted in the respective haemoglobinopathies*" (Image and quoted legend taken from Vanin et al, 1983).

Vanin et al suggested that this consistency in size, along with the lack of homology at the 5' and 3' break points in these deletions implied that the they had come about as a result of recombination between DNA segments that are physically close within the nucleus due to the looping of chromatin during replication. (Vanin, Henthom et al. 1983). Nicholls et al. suggest that this mechanism is also in play at the alpha globin gene cluster, but involving smaller chromatin loop sizes. There is no universal marker for loop size, and this size varies between different cell types, in different stages of cell division and also between men and women, hence there is no standard for their comparison.

Summary and Section Aim

Structural rearrangements can be caused by many different processes that occur during DNA replication. The locations of these events are dictated by the DNA sequence itself and the influence it has on the 3D structure of the DNA molecule. Next generation sequencing has the capability to characterise rearrangements with to-the-base accuracy, but in some circumstances this is difficult and time consuming, or not possible. In these situations, it could be useful to identify sequence motifs that are likely to cause structural rearrangements, and target them specifically when looking for rearrangement breakpoints, for example when designing Gap PCR primers. Understanding the mechanics of structural rearrangements could also be scientifically valuable.

The locations of known breakpoints of rearrangements affecting the globin gene loci (from HbVar) will be examined to determine if they are commonly associated with any known DNA features that provoke DNA rearrangement. Several additional variants were also identified from a literature search and included, along with novel variants from this study where breakpoints were identified. The alpha and beta globin gene clusters are situated on different sequence backgrounds, and the different characteristics of the clusters may mean that rearrangements affecting these loci have different causes.

This chapter will investigate whether any trends exist between the locations of motifs associated with DNA rearrangements and the rearrangements reported at the globin gene loci. This may reveal particular motifs that associate strongly with different types of rearrangement which could help direct the characterisation of novel rearrangements.

Methods

A list of variants affecting the alpha and beta globin gene loci was obtained from the globin gene server at HbVar.org. The list was filtered to all rearrangements affecting >50 bp of sequence with known breakpoints. Rearrangements characterised in this investigation were added to the list, as were duplications reported in the literature that were not recorded in the globin gene server database. The resulting list comprised of 56 pairs of breakpoint co-ordinates for rearrangements on chromosome 11 and 29 pairs of breakpoint co-ordinates on chromosome 16. This included several separate entries for breakpoint co-ordinates of two complex rearrangements on chromosome 11 – the Asian-Indian inversion deletion and English V deletion – which are the culmination of several different rearrangement events (See Results: HiSeq Sample 6 and MiSeq Samples 2 and 3 for details). The alpha and beta globin gene clusters were investigated separately, as rearrangements at the two loci may be driven by different mechanisms.

Sequence features identified in the literature as associated with structural variation were obtained from various online databases. Some databases were extremely large and data collected from them were restricted to a region of interest for each locus. These regions were defined according to the broad locations of the most 3' and most 5' breakpoints of rearrangements used in the investigation. The two regions of interest were:

Chr16: 0-1,000,000

Chr11: 4,500,000-6,00,000

The data collected for these regions of interest were:

- Repeats identified by RepeatMasker along each region downloaded from http://www.dfam.org/search/hits.
- Non-B DNA motifs downloaded from the non-B database (National Cancer Institute at Fredrick https://nonb-abcc.ncifcrf.gov/apps/site/default)
- Recombination Rate data for chromosomes 11 and 16 downloaded from HapMap.org
- The locations of segmental duplications downloaded from the Segmental Duplication Database (University of Washington, http://humanparalogy.gs.washington.edu/SDD/)

- The locations of origins of replication identified in the HeLa cell line were downloaded from OriDB (http://cerevisiae.oridb.org/)

Where breakpoint sequence data was available (primarily from rearrangements characterised in this study via NGS), they were inspected for novel sequences and other features associated with different modes of DNA rearrangement.

The proximity of deletion and duplication breakpoints recorded at each cluster were compared to the locations of DNA features associated with DNA rearrangement. Breakpoints that fell (a) directly within or (b) within 500bp of these features were counted. The proportion of breakpoints meeting condition (a) or (b) were compared to an equal sized group of randomly selected genomic co-ordinates from within the region defined by the most 5' and most 3' breakpoints recorded on each chromosome in the HbVar database. The random co-ordinate group was regenerated a total of 20 times, and the average count for conditions (a) and (b) was used. These comparisons were included in the analysis because of the limited amount of breakpoint data available for study. We felt that there may be some mild associations between these datasets and the sequence motifs examined that could have some biological importance may be overlooked during simple statistical examination due to lack of power.

Genome Annotation Tester (GAT) is a publically available tool for comparing whether sets of genomic intervals overlap significantly (Heger, Webber et al. 2013). In addition to the experiments outline above, Dr David Brawand (Bioinformatician, Dept. Molecular Pathology, King's College Hospital) used GAT to calculate whether any relationships between breakpoint locations and sequence features were strong enough to reach statistical significance. Co-ordinates of the breakpoints and the sequence features in BED format were analysed by the tool, and P and Q values for the overlap between the datasets (in both conditions a and b as described above) were returned.

Note: two entries in the chromosome 16 rearrangement cohort are telomere tip duplications, effectively beginning at position chr16:1. To avoid generating random control co-ordinates within the poorly mapped telomere region of the chromosome, the next most telomeric breakpoint in the dataset was used to define the random control cohort (location chr16:80,000)

Results

Chromosome 16

A total of 29 rearrangements on chromosome 16 with known breakpoints were collated from the HbVar database, sequencing experiments in this study and the literature. The 29 rearrangements provide 58 breakpoints for investigation. Twenty-three rearrangements were deletions and six were duplications. A control group of 58 co-ordinates not associated with rearrangements were randomly generated between the most 3' and most 5' locations of the breakpoints in the known rearrangement cohort.

The locations of these variants and of sequence motifs associated with structural rearrangement across the region of interest on chromosome 16 are shown below (Figure 110), and the described in detail in the following sections.

The Region of Interest on Chromosome 16 in UCSC Genome Browser

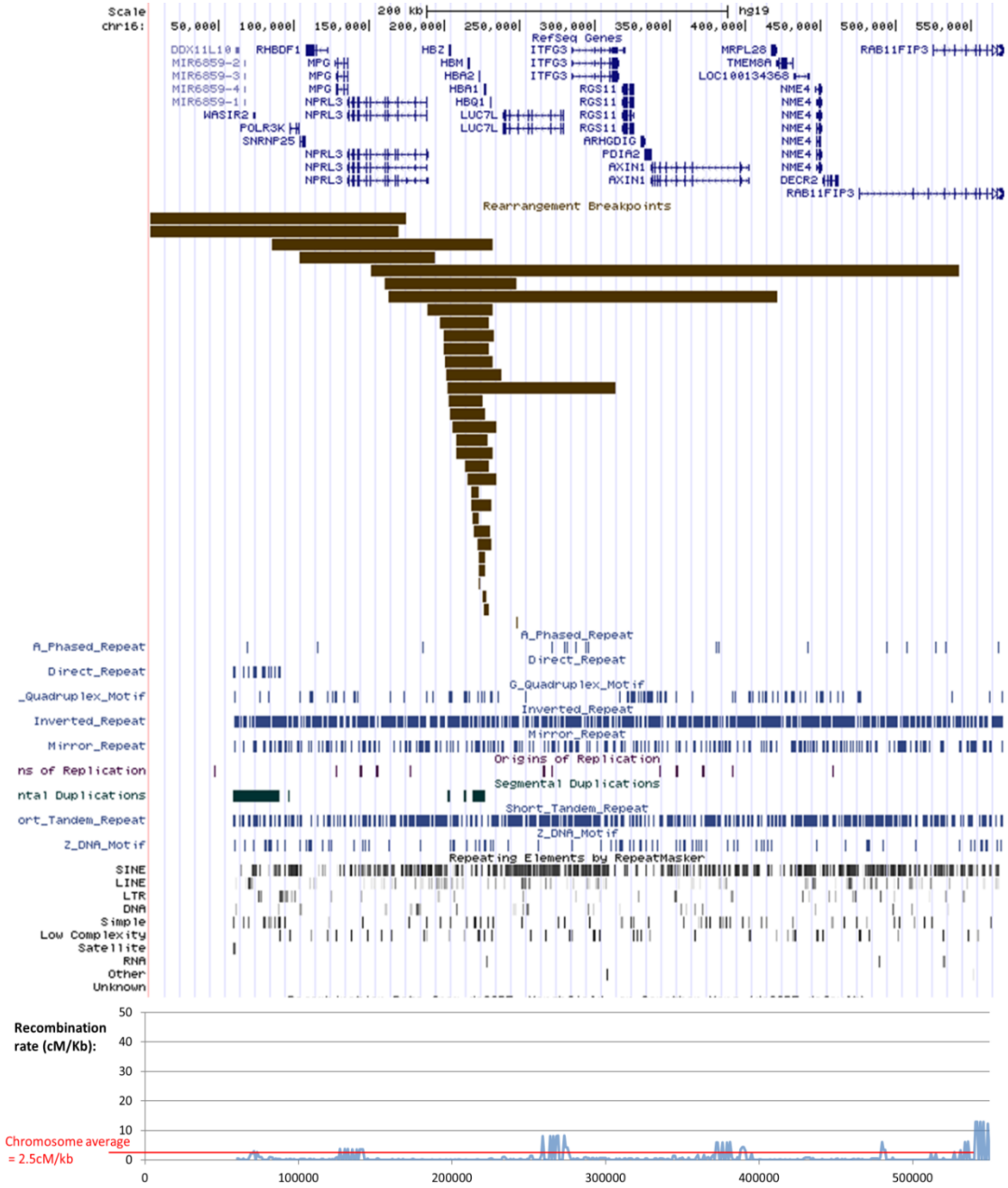


Figure 110 The Region of Interest on Chromosome 16 in UCSC Genome Browser . The locations of known thalassaemia related rearrangements included in this study are shown, along with the relative positions of genes and sequence features associated with structural rearrangements.

Repeats According to RepeatMasker

Repetitive elements that occurred within the region of interest on chr16 were separated according to their class. The region contained five classes of repetitive element: Cut and Paste, LINE, LTR, SINE and RNA. In Condition (a) there was no substantial difference in the number of breakpoint co-ordinates that occurred within Cut and Paste, LINE, LTR or RNA repeats and the random control cohort. Compared to the random control cohort, rearrangement breakpoints were 50% more likely to occur within SINE

repeats. In Condition (b) the difference between rearrangement points and randomly generated points that were adjacent (+/-500bp) to a SINE repeat was less pronounced (Table 66).

Table 66 Co-localisation of rearrangement breakpoints and repetitive elements on chromosome 16 compared to randomly generated control co-ordinates

Repetitive Elements											
		Condition (a) Number of breakpoints falling within repetitive elements					**Condition (b)** Number of breakpoints falling within or near (+/-500bp) repetitive elements				
Repeat type		**Cut and Paste**	**LINE**	**LTR**	**SINE**	**RNA**	**Cut and Paste**	**LINE**	**LTR**	**SINE**	**RNA**
Thalassaemia variants	Total (n=58)	2 (3.45%)	2 (3.45%)	0 (0%)	**20 (34.48%)**	2 (3.45%)	3 (5.17%)	11 (18.97%)	1 (1.72%)	**46 (79.31%)**	0 0%
	Deletions (n=46)	2 (4.35%)	2 (4.35%)	0 (0%)	14 (30.43%)	2 (4.35%)	3 (6.52%)	11 (23.91%)	1 (2.17%)	40 (86.96%)	0 (0%)
	Duplications (n=12)	0 (0%)	0 (0%)	0 (0%)	6 (50%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	6 (50%)	0 (0%)
Randomly generated co-ordinates	Chr16: 80,000-542,000 (n=58)	0.6 (1.03%)	1.25 (2.16%)	0.35 (0.6%)	**8.6 (14.83%)**	0.05 (0.09%)	2.95 (5.09%)	11.4 (19.66%)	3.15 (5.43%)	**36.6 (63.1%)**	0.45 (0.78%)

Replication Origin Points

The DeORI database records 16 replication origin points on chromosome 16 (Table 67). Only a single ORI on chromosome 16 is situated within the region of interest, and none of the breakpoints being investigated fell (a) directly within or (b) within 500 bp of this location.

Table 67 Replication origin points on chromosome 16 as listed in DeORI. The single ORI within the region of interest for this study is highlighted in yellow.

#NO	DeOri AC	Organism	Cell Type	Chromosome	Location	Length	GC content
1	eori000400193	Human_1	Hela	chr16	47415-48530	1116	0.544803
2	eori000400194	Human_1	Hela	chr16	128359-129221	863	0.650058
3	eori000400195	Human_1	Hela	chr16	144166-145603	1438	0.596662
4	eori000400196	Human_1	Hela	chr16	154879-156172	1294	0.697836
5	eori000400197	Human_1	Hela	chr16	176977-177951	975	0.587692
6	eori000400198	Human_1	Hela	chr16	265752-267214	1463	0.667122
7	eori000400199	Human_1	Hela	chr16	271684-272766	1083	0.65374
8	eori000400200	Human_1	Hela	chr16	343083-344285	1203	0.677473
9	eori000400201	Human_1	Hela	chr16	353862-355393	1532	0.610313
10	eori000400202	Human_1	Hela	chr16	371926-373475	1550	0.619355
11	eori000400203	Human_1	Hela	chr16	391235-392389	1155	0.664935
12	eori000400204	Human_1	Hela	chr16	458042-459289	1248	0.503205
13	eori000400205	Human_1	Hela	chr16	60988052-60989443	1392	0.369253
1	eori001000121	Human_2	Hela	chr16	26017327-26018351	1025	0.45561
2	eori001000122	Human_2	Hela	chr16	26191886-26193321	1436	0.401811
3	eori001000123	Human_2	Hela	chr16	60896848-60898068	1221	0.365274

Non-B DNA Motifs

Rearrangement breakpoints were not clearly associated with any non-B DNA motif compared to randomly generated controls in either condition A or condition B (Table 68).

Table 68 Co-localisation of rearrangement breakpoints and Non-B DNA motifs on chromosome 16 compared to randomly generated control co-ordinates

Non-B DNA Motifs Chromosome 16															
		Condition (a) Number of breakpoints falling within non-B DNA motifs							**Condition (b)** Number of breakpoints falling within or near (+/- 500bp) non-B DNA motifs						
Repeat type		Z DNA	STR	Mirror	A Phase	Direct Repeat	G Quadruplex	Inverted Repeats	Z DNA	STR	Mirror	A Phase	Direct Repeat	G Quadruplex	Inverted Repeats
Thalassaemia variants	Total (n=58)	0	0	0	0	0	1 1.72%	2 3.45%	16 27.59%	17 29.31%	21 36.21%	1 1.72%	12 20.69%	25 43.1%	32 55.17%
	Deletions (n=46)	0	0	0	0	0	0	0	11 23.91%	12 26.09%	14 30.43%	1 2.17%	7 15.22%	17 36.96%	21 45.65%
	Duplications (n=12)	0	0	0	0	0	1 8.33%	2 16.67%	5 41.67%	5 41.67%	7 58.33%	0	5 41.67%	8 66.67%	11 91.67%
Randomly generated co-ordinates	Chr16 (n=58): 80,000-542,000	0.14 0.24%	1.06 1.83%	1.64 2.83%	0.05 0.09%	1.88 3.24%	1.16 2%	2.51 4.33%	9.32 16.07%	39.2 67.59%	25.88 44.62%	2.22 3.83%	24.28 41.86%	17.57 30.29%	49.39 85.16%

Recombination Rate

Recombination rate is a continuous variable, and as such it was not analysed using the same 'within' or 'adjacent' conditions as other variables. The recombination rates calculated across chromosome 16 were downloaded from HapMap. An average value for the recombination rate was calculated per 1 Kb of the region of interest. The average recombination rate across the region of interest was 1.0231cM. The number of breakpoints occurring per 1 Kb of the region of interest was also calculated. The two datasets showed a weak negative correlation with an R^2 value of 0.0076 (Figure 111).

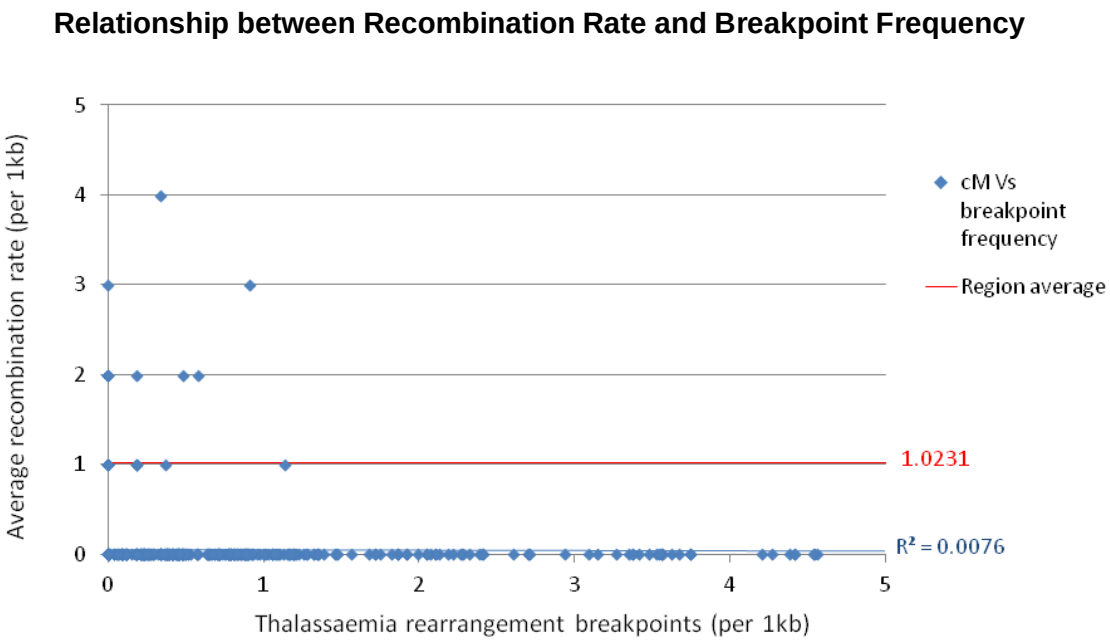


Figure 111 Relationship between Recombination Rate and Breakpoint Frequency per 1000 bp of sequence on chromosome 16

Segmental Duplications

Thirty-one percent of rearrangement breakpoints were situated within segmental duplications compared to 5% of randomly generated controls. Thirty-six percent of rearrangement breakpoints were adjacent (<+/-500bp) to segmental duplications compared to 5.7% of randomly generated controls (Table 69).

Table 69 Co-localisation of rearrangement breakpoints and Segmental Duplications on chromosome 16 compared to randomly generated control co-ordinates

Segmental Duplications			
		Condition (a) Number of breakpoints falling within segmental duplications	**Condition (b)** Number of breakpoints falling within or near (+/-500bp) segmental duplications
Thalassaemia variants	Total (n=58)	18 (31%)	21 (36%)
	Deletions (n=46)	14 (30%)	16 (34%)
	Duplications (n=12)	4 (33%)	5 (45%)
Randomly generated co-ordinates	Chr16: 80,000-542,000 (n=58)	3.2 (5.5%) 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 3 2 1 2 7 2 3 3 4 3 2 7 3 5 8 3 1 2 2 1	3.35 (5.7%) 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 5 5 2 5 2 1 3 1 4 6 1 4 1 3 3 2 4 1 9 5

Similarities in Rearrangement Size Indicative of Recombination between Chromatin Loops

Previous work has suggested that rearrangements on chromosomes 11 and 16 may fall into certain size groups dictated by the chromatin loop length across these regions. Rearrangements in the cohort on chromosome 16 were plotted according to their size and median position on the chromosome. The reported common rearrangement size on chromosome 16 was 20-30 Kb. The distribution of rearrangements in this study by size and median position is shown in Figure 112. Two scales are used to clearly show the range of variant sizes. The majority of variants are <5 Kb and appear to occur at an evenly dispersed range of sizes below this length.

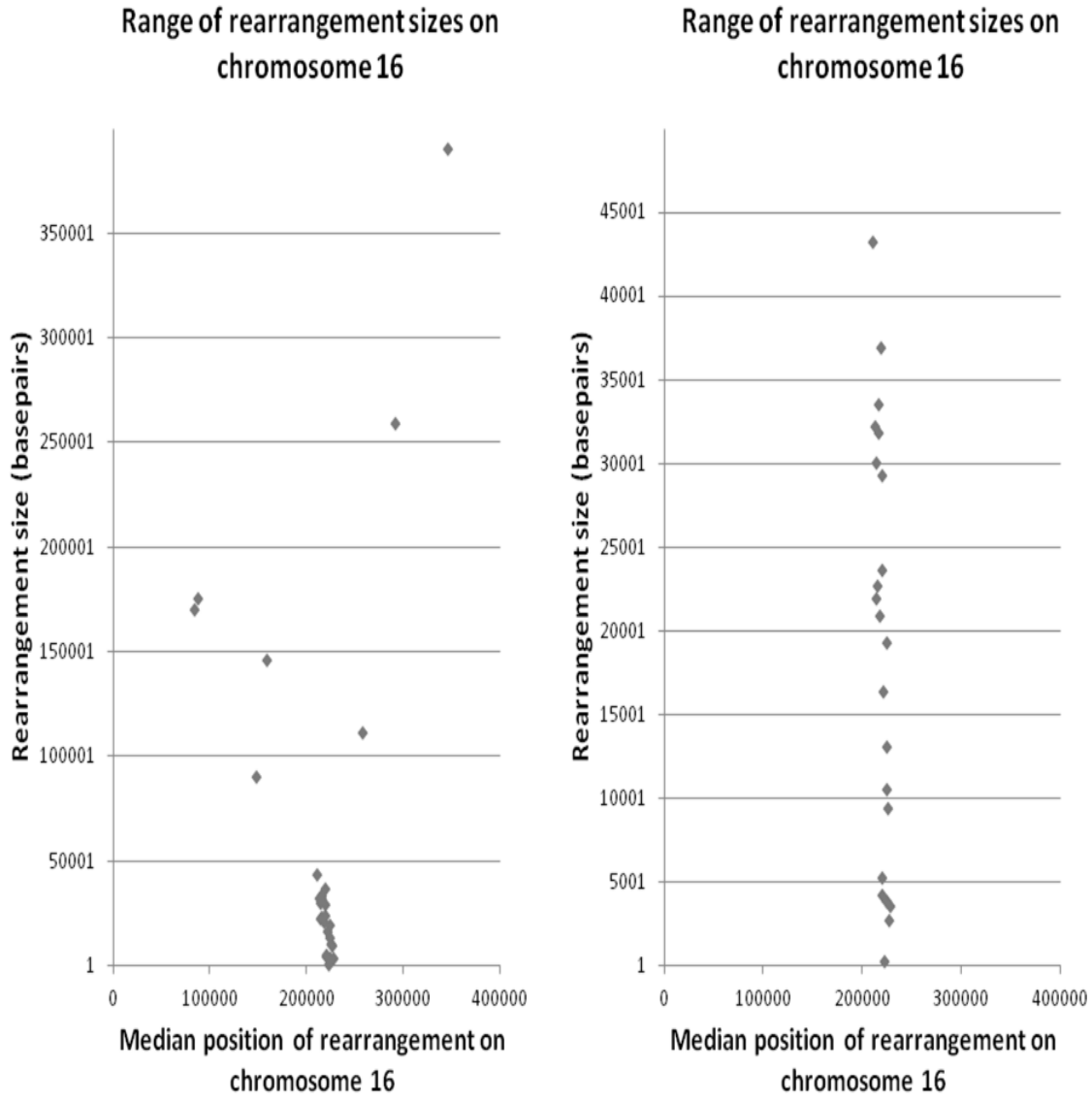


Figure 112: Range of Variant Sizes on Chromosome 16. Rearrangements on chromosome 16 analysed in this study, plotted by median affected position (X-axis) and rearrangement size (Y-axis). Two scales are provided for clear illustration of the size ranges.

Notable Features of Breakpoint Sequences

Rearrangements that had been mapped using Next Generation Sequencing were studied in closer detail, as permitted by our knowledge of the exact breakpoint sequences. Details of the breakpoint sequences were examined to identify common features of the rearrangements. This included features that caused the initial double stranded break and mutational signatures left behind in the rearrangement, indicative of the repair mechanism that sutured the chromosome (Table 70). For detailed description of the rearrangements described in Table 70 please see sections describing their characterisation in Results Chapters 1 and 2.

Table 70 Notable features of breakpoint sequences on chromosome 16

Rearrangement	Features of break point sequence	Rearrangement mechanism
MiSeq Sample 33 Deletion Breakpoint 1 = 104,934-104,943 Deletion Breakpoint 2 = 177,954-177,966	◦ Breakpoint shares 12bp of microhomology (hence ranges given as breakpoints) ◦ Breakpoint 1 and breakpoint 2 are both situated in Alu repeats with high homology ◦ Neither breakpoint is within a NonB DNA motif	◦ Microhomology at break point could indicate a DSB that has been repaired by MMEJ ◦ Involvement of homologous Alu repeats suggests NAHR
HiSeq Sample 7 Duplication Breakpoint 1=16:112,164- Breakpoint2=16:300,927	◦ Duplication is top-to-tail ◦ No novel bases ◦ No ambiguous bases ◦ Breakpoint 1 is <500bp from Cut-And-Paste and Alu (SINE) repeat regions ◦ Breakpoint 2 is in <500bp from an Alu (SINE) repeat region ◦ Neither breakpoint within Non-B DNA motif ◦ Breakpoint 1 is <500bp from an inverted repeat ◦ Breakpoint 2 is <500bp from STR, Mirror, Direct and G-Quadruplex repeats ◦ Neither breakpoint is <500bp from a known segmental duplication ◦ Neither breakpoint is <500bp from a known segmental duplication	◦ Non recurrent, non allelic homologous recombination (NAHR) between repetitive elements ◦ Proximity to a G-Quadruplex motif, which are known to be associated with tetraplex structures that can cause duplications
MiSeq Sample 34 Duplication Breakpoint1=16:148,451 Breakpoint2=16:269,041	◦ Duplication is top-to-tail ◦ 3 base pair microhomology at breakpoint ◦ Breakpoint 2 is situated within a Alu (SINE) repeat. Breakpoint 1 is 200bp upstream of a Alu (SINE) repeat with high homology to the repeat at breakpoint 2. ◦ Neither breakpoint within Non-B DNA motif ◦ Breakpoint 1 is <500bp from STR inverted and direct repeat motifs ◦ Breakpoint 2 is <500bp from STR, mirror and direct repeat motifs ◦ Neither breakpoint is <500bp from a known segmental duplication	◦ Microhomology at break point could indicate a DSB that has been repaired by MMEJ ◦ Inverted and direct repeats are both associated with NHEJ ◦ Involvement of Alu repeats is associated with NAHR
MiSeq Sample 38 Breakpoint1=16:161,049 Breakpoint2=16:248,193 Breakpoint3=16:248,263 Breakpoint4=16:248,605	◦ Rearrangement consists of two deletions separated by 71bp intact sequence ◦ Intact sequence includes a novel insertion of 6bp (TCTCGC) ◦ Breakpoint 1 is situated in unique sequence ◦ Breakpoints 2 and 3 are situated in a Alu (SINE) repeat ◦ Breakpoint 4 100bp upstream of an Alu (SINE) repeat with high homology to the repeat encompassing breakpoints 2 and 3 ◦ No breakpoints within Non-B DNA motif ◦ Neither breakpoint is <500bp from a known segmental duplication	◦ Insertion of novel sequence at deletion breakpoint suggests NHEJ ◦ Involvement of homologous Alu repeats suggests NAHR ◦ Complex rearrangements are attributed primarily to the FoSTeS mechanism, but it is not known whether the different elements of the rearrangement arose simutaneously or during different events.

Chromosome 11

A total of 57 rearrangements on chromosome 11 with known breakpoints were collated from the HbVar database and from sequencing experiments in this study. The 57 rearrangements provide 114 breakpoints for investigation. Forty-eight rearrangements were deletions, seven were duplications and two were inversions (Table 71). A control group of 114 co-ordinates not associated with rearrangements were randomly generated between the most 3' and most 5' locations of the breakpoints in the known rearrangement cohort.

Table 71 Breakpoint pairs included in the study on chromosome 11 by type

Chr11	Count
Deletions	48
Duplications	7
Inversions	2
Total	**57**

The locations of these variants and of sequence motifs associated with structural rearrangement across the region of interest on chromosome 11 are shown below (Figure 113), and described in detail in the following sections.

The Region of Interest on Chromosome 11 in UCSC Genome Browser

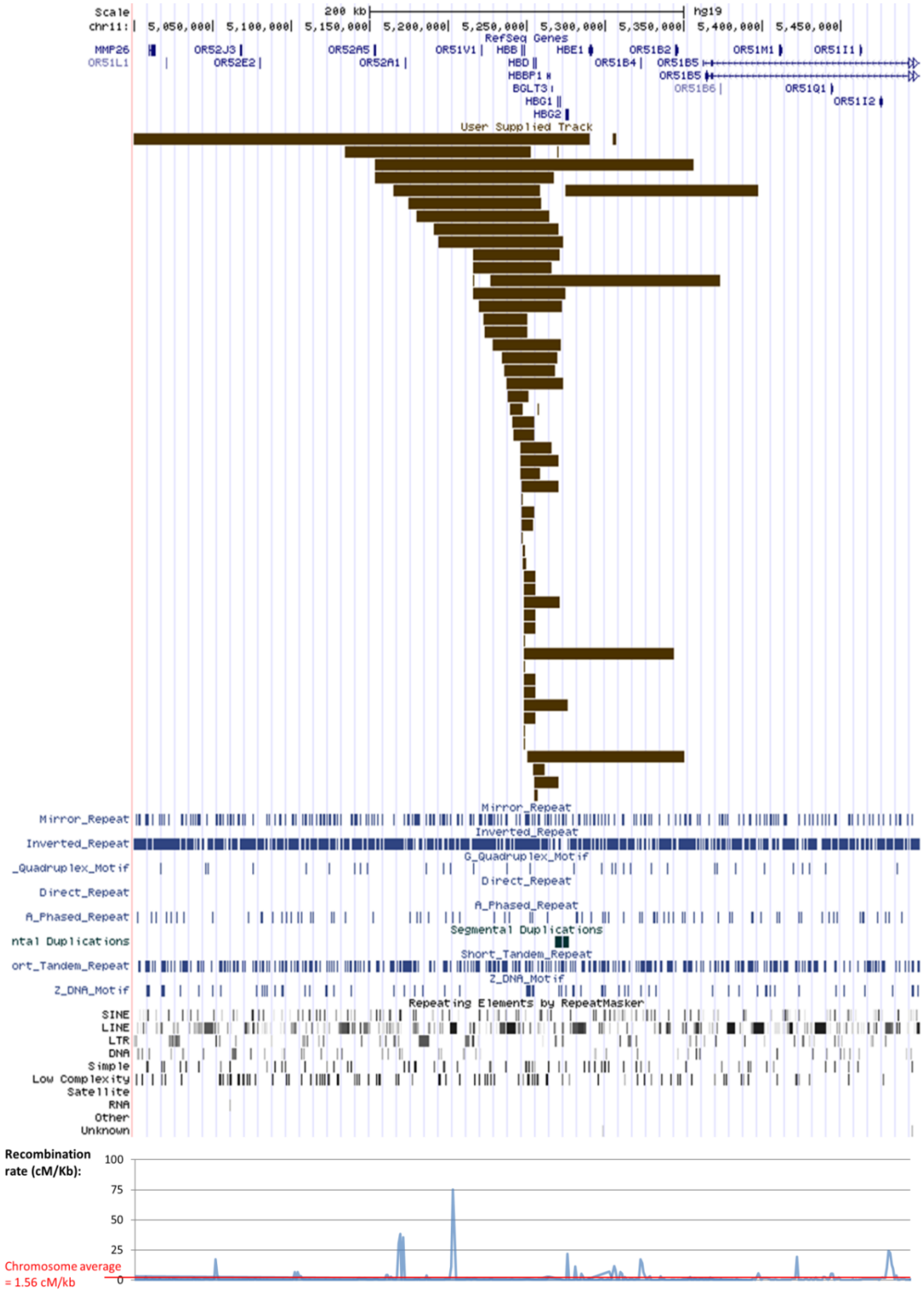


Figure 113 The Region of Interest on Chromosome 11 in UCSC Genome Browser . The locations of known thalassaemia related rearrangements included in this study are shown, along with the relative positions of genes and sequence features associated with structural rearrangements..

Repeats According to RepeatMasker

No duplication or inversion breakpoints are within a recorded repetitive region. In 17% of deletions the breakpoint occurred within a repeat of some kind. The majority (13.16% of deletion breakpoints) occur within LINE repeats, compared with 3.15% of randomly selected co-ordinates. There was no difference in the frequency of adjacent (+/-500bp) repeats to break point positions and randomly selected positions for any repeat class (Table 72).

Table 72 Co-localisation of rearrangement breakpoints and repetitive elements on chromosome 16 compared to randomly generated control co-ordinates

Repetitive Elements Chromosome 11													
		Condition (a) Number of breakpoints falling within repetitive elements						**Condition (b)** Number of breakpoints falling within or near (+/- 500bp) repetitive elements					
Repeat type		**Cut and Paste**	**DNA**	**LINE**	**LTR**	**RNA**	**SINE**	**Cut and Paste**	**DNA**	**LINE**	**LTR**	**SINE**	**RNA**
Thalassaemia variants	Total (n=114)	2 (1.75%)	0	15 (13.16%)	0	0	3 (2.63%)	11 (9.65%)	0	33 (28.95%)	8 (7.02%)	17 (14.91%)	0
	Deletions (n=96)	2 (2.08%)	0	15 (15.63%)	0	0	3 (3.13%)	10 (10.42%)	0	30 (31.25%)	8 (8.33%)	15 (15.63%)	0
	Duplications (n=14)	0	0	0	0	0	0	1 (7.14%)	0	2 (14.29%)	0	1 (7.14%)	0
	Inversions (n=4)	0	0	0	0	0	0	0	0	1 (25%)	0	1 (25%)	0
Randomly generated co-ordinates	Chr11: 4640332-5397195 (n=114)	5 (4.39%)	0	4 (3.51%)	2 (1.75%)	1 (0.88%)	0	15.7 (13.77%)	0	43.25 (37.94%)	16.2 (14.21%)	24.3 (21.32%)	0

Replication Origin Points

The DeORI database lists 18 replication origin points identified on chromosome 11 (Table 73), of which five are within the region of interest for beta thalassaemia related rearrangements. No rearrangement breakpoints fell within, or adjacent to (+/- 500bp) an origin of replication point.

Table 73 Origins of replication on chromosome 11 from the DeORI database. Origins of replication within the region of interest for thalassaemia rearrangements are highlighted in yellow

#NO	DeOri AC	Organism	Chr	Cell Type	Location	Length	GC content
1	eori001000078	Human_2	chr11	Hela	1861642-1863584	1943	0.504375
2	eori001000079	Human_2	chr11	Hela	1891643-1893782	2140	0.515888
3	eori001000080	Human_2	chr11	Hela	1920771-1921948	1178	0.504245
4	eori001000081	Human_2	chr11	Hela	5032960-5034446	1487	0.321453
5	eori001000082	Human_2	chr11	Hela	5099957-5101447	1491	0.382294
6	eori001000083	Human_2	chr11	Hela	5298535-5299559	1025	0.345366
7	eori001000084	Human_2	chr11	Hela	5422465-5423600	1136	0.328345
8	eori001000085	Human_2	chr11	Hela	5451216-5452829	1614	0.314126
1	eori000400142	Human_1	chr11	Hela	1715184-1716695	1512	0.474868
2	eori000400143	Human_1	chr11	Hela	2127225-2128288	1064	0.523496
3	eori000400144	Human_1	chr11	Hela	2148231-2149404	1174	0.599659
4	eori000400145	Human_1	chr11	Hela	64131110-64132987	1878	0.592119
5	eori000400146	Human_1	chr11	Hela	64154782-64156726	1945	0.570694
6	eori000400147	Human_1	chr11	Hela	64160072-64161187	1116	0.586918
7	eori000400148	Human_1	chr11	Hela	64167296-64168971	1676	0.629475
8	eori000400149	Human_1	chr11	Hela	64249506-64251467	1962	0.567788
9	eori000400150	Human_1	chr11	Hela	64289674-64293005	3332	0.573229
10	eori000400151	Human_1	chr11	Hela	64302035-64303808	1774	0.630214
11	eori000400152	Human_1	chr11	Hela	64326897-64328689	1793	0.63469
12	eori000400153	Human_1	chr11	Hela	64367486-64368648	1163	0.657782
13	eori000400154	Human_1	chr11	Hela	64368631-64369884	1254	0.62201
14	eori000400155	Human_1	chr11	Hela	64412337-64413999	1663	0.566446
15	eori000400156	Human_1	chr11	Hela	64438527-64439948	1422	0.552743
16	eori000400157	Human_1	chr11	Hela	130855483-130856606	1124	0.449288
17	eori000400158	Human_1	chr11	Hela	130932556-130933663	1108	0.456679
18	eori000400159	Human_1	chr11	Hela	131047767-131049230	1464	0.504781

Non-B DNA Motifs

Breakpoints occurred within mirror repeats at a greater frequency than randomly generated controls. Breakpoints were more likely to be situated adjacent to (<+/-500bp) Z DNA motifs than randomly generated controls. Breakpoints were less likely to be adjacent to A phase or G-quadruplex repeats than randomly generated controls. All four inversion breakpoints were adjacent to mirror repeats, and 75% were adjacent to direct repeats (Table 74).

Table 74 Co-localisation of rearrangement breakpoints and Non-B DNA Motifs on chromosome 11 compared to randomly generated control co-ordinates

Non-B DNA Motifs Chromosome 11															
		Condition (a) Number of breakpoints falling within non-B DNA motifs							**Condition (b)** Number of breakpoints falling within or near (+/-500bp) non-B DNA motifs						
Repeat type		**Z DNA**	**STR**	**Mirror**	**A Phase**	**Direct Repeat**	**G Quadruplex**	**Inverted Repeats**	**Z DNA**	**STR**	**Mirror**	**A Phase**	**Direct Repeat**	**G Quadruplex**	**Inverted Repeats**
Thalassaemia variants	Total (n=114)	0	2 1.75%	5 4.4%	0	1 0.88%	0	2 1.75%	15 20%	27 36%	36 47%	1 1%	21 28%	2 3%	31 41%
	Deletions (n=96)	0	2 2.08%	5 5.21%	0	1 1.04%	0	2 2.08%	14 14.6%	25 26%	31 32.3%	1 1.04%	19 19.8%	2 2.08%	24 25%
	Duplications (n=14)	0	0	0	0	0	0	0	1 7.14%	2 14.3%	5 35.7%	0	2 14.3%	0	7 50%
	Inversions (n=4)	0	0	0	0	0	0	0	3 75%	2 50%	4 100%	0	3 75%	1 25%	1 25%
Randomly generated co-ordinates	Chr11 (n=114): 4640332 - 5397195	0.1 0.1%	1.2 1%	3 2.6%	0.7 0.6%	1 0.9%	0.4 0.3%	2.4 2.1%	6.3 5.3%	35.2 30.9%	53.1 46.6%	34.2 30%	23.2 20.4%	21 18%	53.4 46.8%

Recombination Rate

Recombination rate is a continuous variable, and as such it was not analysed using the same 'within' or 'adjacent' conditions as other variables. The recombination rates calculated across chromosome 11 were downloaded from HapMap. An average value for the recombination rate was calculated per 1 Kb of the region of interest. The average recombination rate across the region of interest was 1.69 cM. The number of breakpoints occurring per 1 Kb of the region of interest was also calculated. The two datasets showed a weak negative correlation with an R^2 value of 0.0108 (Figure 114).

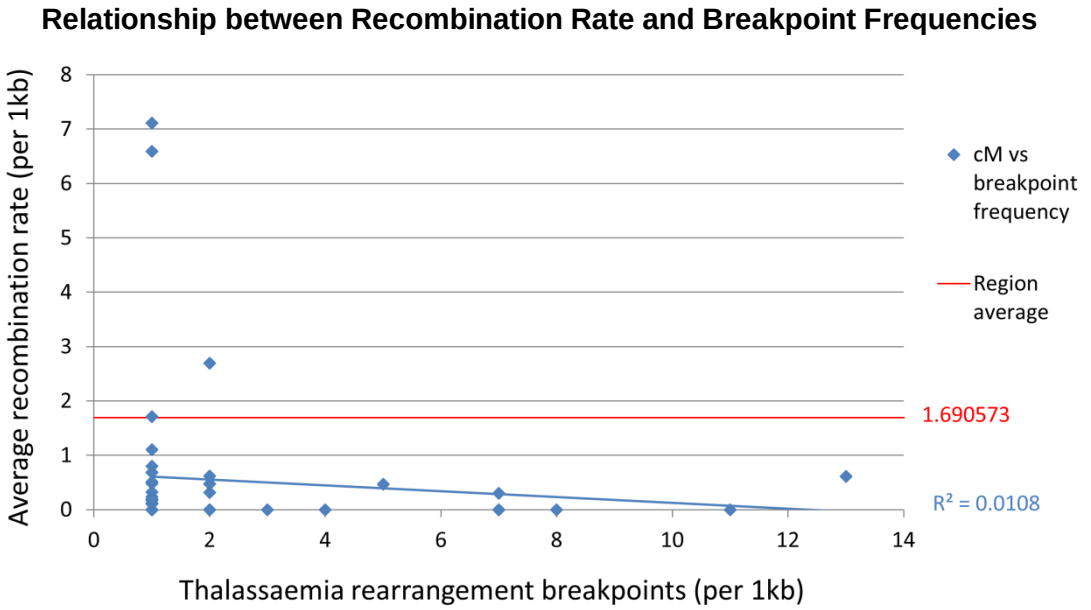


Figure 114: Relationship between Recombination Rate and Breakpoint Frequencies per 1000bp of Sequence on Chromosome 11

Segmental Duplications

Rearrangement breakpoints were more likely than randomly generated control co-ordinates to occur within segmental duplications. No additional breakpoints were situated adjacent to segmental duplications (Table 75).

Table 75 Co-localisation of rearrangement breakpoints and segmental duplications on chromosome 11 compared to randomly generated control co-ordinates

Segmental Duplications			
		Condition (a) Number of breakpoints falling within Segmental Duplications	**Condition (b)** Number of breakpoints falling within or near (+/-500bp) Segmental Duplications
Thalassaemia variants	Total (n=114)	16 (14.04%)	16 14.04%)
	Deletions (n=96)	14 (14.58%)	14 (14.48%)
	Duplications (n=14)	0	0
	Inversions (n=4)	2 (50%	2 (50%)
Randomly generated co-ordinates	Chr11: 4640332-5397195 (n=114)	2.95 (2.59%)	4.45 (3.9%)

Similarities in Rearrangement Size Indicative Of Recombination Between Chromatin Loops

Rearrangements in the cohort on chromosome 11 were plotted according to their size and median affected position (Figure 115). Previous research had suggested rearrangements affecting 40 Kb of sequence may be the most common at this locus. The majority of rearrangements affected less than 50,000 bp of sequence, so the graph is shown on two scales to provide clearer illustration of the size distribution of the smaller variants. The size range of variants < 50,000 bp shows small clusters of rearrangements 6-8 Kb in size and 12- 14 Kb in size. No pronounced grouping of rearrangement lengths into blocks is visible on the larger scale.

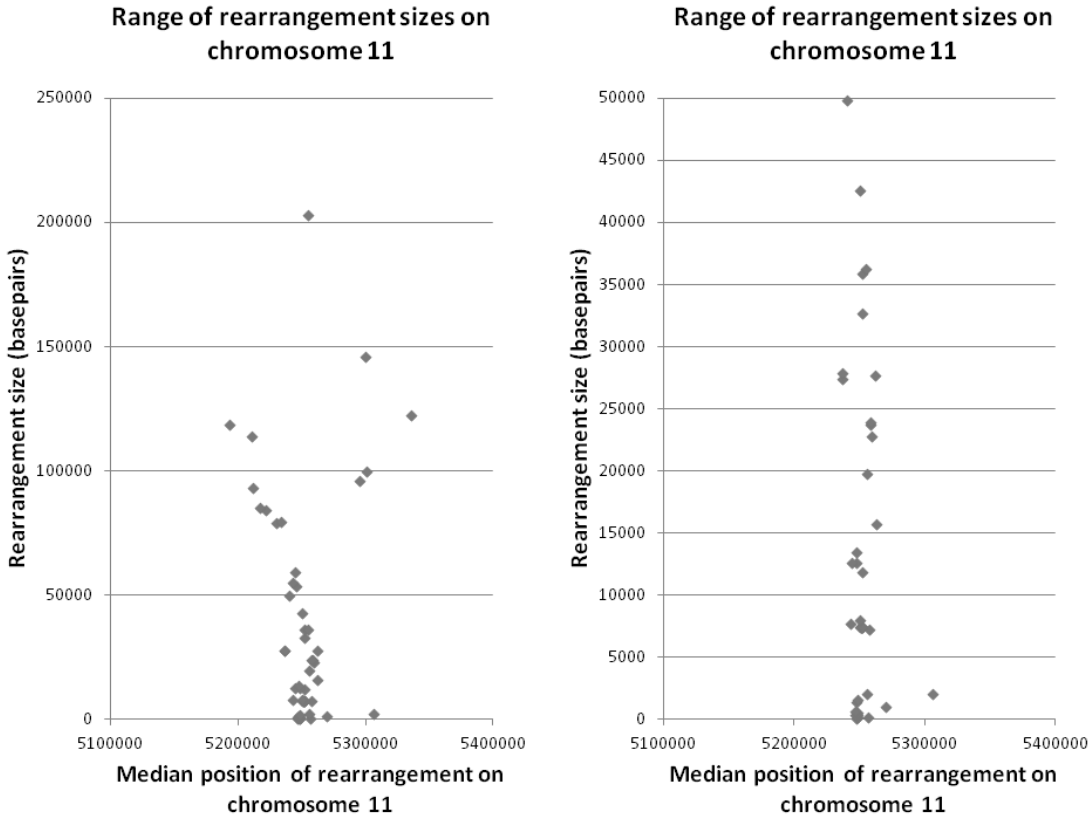


Figure 115: Range of Variant Sizes on Chromosome 16. Rearrangements on chromosome 11 analysed in this study, plotted by median affected position (X-axis) and rearrangement size (Y-axis). Two scales are provided for clear illustration of the size ranges.

Notable Features of Breakpoint Sequences

Table 76 describes the findings of an in-depth inspection of the features and sequences of breakpoints in rearrangements described with to-the-base accuracy on chromosome 11 during this study. For detailed description of the rearrangements described please see appropriate results sections in Results Chapters 1 and 2.

Table 76 Notable features of breakpoint sequences in DNA samples characterised via NGS

Rearrangement	Features of break point sequence	Rearrangement mechanism
619bp Deletion (MiSeq Sample 5) Breakpoint 1 = 5246619 Breakpoint 2 = 5247237	◦ Breakpoint 2 adjacent to L1MA6 repeat ◦ Breakpoint 2 within inverted repeat ◦ 7bp insertion (TCTACTT) at breakpoint	◦ Insertions and inverted repeats at the breakpoint are associated with NHEJ ◦ Line repeats are often found at or adjacent to breakpoint sites where double stranded breaks have been repaired by NHEJ or NAHR
English V Deletion (MiSeq Sample 3, 4, 6) Breakpoint 1 (inversion, deletion) = 5215690 Breakpoint 2 (deletion) = 5215722 Breakpoint 3 (inversion, deletion) = 5274684 Breakpoint 4 (deletion) =5397195	◦ Breakpoint 1 situated at centre of 158bp palindrome ◦ Breakpoint 2 situated adjacent to repeat LTR12D ◦ Breakpoint 3 situated in unique sequence with high homology to region of Breakpoint 1 ◦ Breakpoint 4 within 6kb LINE repeat	◦ Complex rearrangements are most likely to be formed by FoSTes events. ◦ Palindrome sequences and the hairpin structures they can create may cause fork stalling during replication leading to FoSTes. ◦ NAHR is associated with joining regions with high homology to one another
African 1 Duplication (MiSeq Sample 7) Breakpoint 1 = 5226885 Breakpoint 2 = 5372677	◦ Breakpoint 1 adjacent to low complexity AT rich repeat ◦ 11bp insertion (CACCTCCACTT) creates 13bp mirror repeat at breakpoint	◦ NHEJ has been identified as the causative mechanism of non-recurrent tandem duplications with insertions, although the literature specifies that these insertions are small, i.e. 1-6bp
Novel Duplication 1 (MiSeq Samples 41, 42) Breakpoint 1 = 4640322 Breakpoint 2 = 5290168 Duplication is interrupted by a balanced region with unknown breakpoints. Approximate breakpoint 3 = 5201000 Approximate breakpoint 4 = 5244500	◦ Ambiguous 'GG' at breakpoint ◦ Breakpoint 1 within LINE L1PB1 repeat ◦ Approximate breakpoint 3 of the balanced region that interrupts the duplication is within a LINE repeat ◦ Approximate breakpoint 4 of the balanced region that interrupts the duplication is within a SINE repeat	◦ Microhomology is associated with FoSTeS, MMEJ ◦ LINE repeats are associated with NHEJ and NAHR ◦ The interruption to the duplication classes this as a complex rearrangement, most commonly associated with FoSTeS
Novel Deletion 1 (MiSeq Sample 43) Breakpoint 1 = 5222877 Breakpoint 2 = 5250289	◦ Breakpoint 2 adjacent to low complexity AT rich repeat	◦ AT rich repeats are associated with replication origin sites
Asian-Indian Inversion Deletion (HiSeq Sample 6) Breakpoint 1 (deletion) =5246800 Breakpoint 2 (Deletion/Inversion) = 5254778 Breakpoint 3 (deletion/inversion) = 5270446 Breakpoint 4 (deletion) = 5269408	◦ Breakpoint 2 adjacent to simple repeat ◦ Breakpoint 3 within inverted repeat ◦ Breakpoint 3 adjacent to simple repeat ◦ Microhomology ('CC') at breakpoint ◦ Novel insertion ('AGTTGT') at breakpoint	◦ FoSTes is associated with complex rearrangements ◦ Novel insertions associated with NHEJ ◦ Microhomology associated with FoSTeS, NHEJ
HPFH1 Deletion (MiSeq Sample 1) Breakpoint 1 = 5174451 Breakpoint 2 = 5259369	◦ Breakpoint 1 within inverted repeat ◦ Breakpoint 2 within aluSq2 repeat ◦ Novel insertion ('TATTT') at breakpoint	◦ Inverted repeats associated with NHEJ, cruciform structures and deletions ◦ Novel insertions associated with NHEJ

Summary of Significance of Structural Variant Association

Genome Association Tester (GAT) is a tool to determine whether two sets of genomic intervals overlap one another more than just by chance alone (Heger, Webber et al. 2013). To return quantitative values for the data described above, the co-ordinates of the structural variant sets from chromosomes 11 and 16 were compared to co-ordinates lists of genome features potentially associated with them. A second dataset tested the association between genome features and broader regions (+/- 500bp) containing a structural variant breakpoint. For each genome feature, the significance with which they are associated with structural variants is returned as a P value and a Q value, based on the *Expected* versus *Observed* overlap between the two datasets. The data generated by GAT is summarized in Table 77.

Chromosome 11 variants were associated with CutandPaste type repeats and mirror repeats with a P value of P=<0.05. Chromosome 16 variants were associated with LINE repeats with a P value of P=<0.05 and SINE type repeats and segmental duplications at a P value of P=0.01. Widening the structural rearrangement co-ordinates from the breakpoint locations to the breakpoint regions (breakpoint +/- 500 bp) did not increase the association between any genomic features and the list of variants. From this we conclude that there is a significant association between the breakpoints of recorded structural variants on chromosome 16 and the locations of SINE repeats and segmental duplication regions. On chromosome 11, only a mild association between sequence motifs and rearrangement breakpoints is found. This suggests either that a different mechanism is responsible for the rearrangements that occur on the two chromosomes, or that there are issues with the validity of this data. SINE repeats are far more frequent on chromosome 16 than chromosome 11. The apparent involvement of these repeats with rearrangements at this locus may explain why there are a larger number of rearrangements reported here than on chromosome 11.

Table 77: Identification of significant overlaps between DNA sequence features and thalassaemia breakpoint co-ordinates. Findings calculated by GAT. Relationships that are significant at P<0.05 are highlighted in red. Relationships that are significant at P<0.001 are bold, and highlighted in red.

Association of genomic features with:		Structural Variant Breakpoint Co-ordinates				Breakpoint Regions (+/-500 bp)	
Chr	Feature	Observed	Expected	P value	Q value	P value	Q value
Chr 11	Cutandpaste	40	27.64	0.2610	0.5840	0.0470	0.4416
Chr 16		16	3.978	0.0680	0.2720	0.0580	0.1952
Chr 11	DNA	0	1.245	0.8830	1.0000	0.3220	0.4873
Chr 16		0	0	1.0000	1.0000	1.0000	1.0000
Chr 11	LINE	178	190.241	0.3890	0.6916	0.1380	0.4416
Chr 16		20	11.457	0.2900	0.5800	0.0160	0.0853
Chr 11	Long Tandem Repeat	50	71.554	0.2500	0.5840	1.0000	1.0000
Chr 16		0	4.756	0.6100	0.8133	0.3920	0.5400
Chr 11	SINE	71	59.674	0.2850	0.5840	0.0910	0.4416
Chr 16		**269**	**79.033**	**0.0010**	**0.0080**	**0.0010**	**0.0080**
Chr 11	RNA	0	0.292	0.9690	1.0000	0.8940	0.9536
Chr 16		0	0.182	0.9800	1.0000	0.8490	0.9703
Chr 11	DeORI	0	2.794	0.7630	0.9391	0.8140	0.9536
Chr 16		0	7.619	0.4500	0.6836	0.4050	0.5400
Chr 11	Z DNA Motif	0	1.842	0.7590	0.9391	0.1340	0.4416
Chr 16		10	2.272	0.0990	0.3168	0.6150	0.7569
Chr 11	Short Tandem Repeat	19	12.398	0.2540	0.5840	0.3350	0.4873
Chr 16		0	8.572	0.2400	0.5486	0.2570	0.5400
Chr 11	Mirror Repeat	52	24.127	0.0360	0.2880	0.3880	0.5173
Chr 16		1	16.07	0.1550	0.4133	0.3410	0.5400
Chr 11	A Phased Repeat	0	3.579	0.6070	0.9391	0.8590	0.9536
Chr 16		0	0.566	0.9310	1.0000	0.3350	0.5400
Chr 11	Direct Repeat	13	9.803	0.2920	0.5840	0.2010	0.4680
Chr 16		0	23.115	0.0620	0.2720	0.2860	0.5400
Chr 11	G Quadruplex Motif	0	2.288	0.7500	0.9391	0.2310	0.4680
Chr 16		9	11.654	0.3880	0.6836	0.0610	0.1952
Chr 11	Inverted Repeat	65	44.271	0.1570	0.5840	0.2700	0.4800
Chr 16		20	23.507	0.4700	0.6836	0.1970	0.5253
Chr 11	Segmental Duplication	140	75.602	0.0150	0.2400	0.0870	0.4416
Chr 16		**161**	**28.803**	**0.0010**	**0.0080**	**0.0010**	**0.0080**

Chapter Discussion

The positions of known structural variants on chromosome 11 and 16 were compared to data from multiple databases of genome sequence features. Genome sequence features that had significant, mild or no association with break point positions were identified. There were clear differences between the factors involved with rearrangements affecting the alpha and beta globin gene loci.

Previous studies have implicated segmental duplications in the creation of structural rearrangements. This is attributed to their high homology to one another, which allows non-allelic homologous recombination (NAHR) to occur between these points. There was some localization of structural rearrangement breakpoints to segmental duplication regions at both loci, although this was only statistically significant for the cohort of rearrangements on chromosome 16. The fact that segmental duplications are associated with variants on one locus, but not the other, suggests that a different process may cause rearrangements at these two loci. Alu-type SINE repeats and LINE repeats have also been associated with structural rearrangements. These are also believed to facilitate NAHR due to their high level of homology to one another. SINE repeats were found to be significantly associated with the structural rearrangement co-ordinates on chromosome 16, but not on chromosome 11. LINE repeats were also significantly associated with rearrangements on chromosome 16 at $P<0.05$.

An investigation by Nicholls et al identified Alu repeats and homologous sequences as sequence features associated with alpha thalassaemia breakpoints (Nicholls, Fischel-Ghodsian et al. 1987). The study noted that all four of the variants they surveyed fell into a broadly similar size range. This led to the suggestion that rearrangements at the alpha globin gene locus happened between DNA sequences brought into close physical proximity by chromatin looping during replication. The rearrangements investigated in this study did not appear to show any grouping into distinct size brackets, or into the particular size bracket (20-30 Kb) suggested by Nicholls et al. Therefore, our findings do not support their suggestion.

Nicholls et al described an inverted insertion at the breakpoints of the ($--^{MED}/$) deletion. Although this variant was sequenced as a positive control sample in this study, the breakpoint sequences of this variant could not be identified in the read data. This was concluded to be a result of high homology at the breakpoint sequences, along with their location within unmapped repetitive regions (low complexity and SINE). This inversion is not included in the description of this variant HbVar or mentioned elsewhere in the literature, and so the case found by Nicholls et al may be a rare sub-variant of ($--^{MED}$).

Without further information about this inversion deletion, we will not comment on what process that may have caused it.

The breakpoint locations of rearrangements that had been characterised by NGS were examined closely. On chromosome 16, four rearrangements that had been characterised with to-the-base accuracy were examined. Three of these showed homology at the breakpoint and association with Alu repeats, which implicated NAHR in the generation of these variants. One of the four rearrangements had an insertion of novel sequence at the breakpoint, which is associated with DNA repair by non-homologous end joining (NHEJ). An additional variant which had not been characterised with to-the-base accuracy at the time of writing (MiSeq Sample 35) also appeared to occur between two Alu repeats with high homology to one another. NAHR-mediated structural rearrangements have been previously implicated in numerous diseases caused by a variety of structural variants (Amos-Landgraf, Ji et al. , Nagamani, Erez et al. 2011). Our findings support the notion that this is also the case in alpha thalassaemia.

Two rearrangements of the alpha globin gene cluster that were included in this study included the telomere. We are unable to design baits against the telomere sequence, so this represents a limitation of the study where breakpoints within the telomere may have significant and unknown properties.

Rearrangements on the beta globin gene cluster were not significantly associated with any sequence features at a significance of $P<0.01$. Mirror repeats were associated with rearrangements at a significance of $P<0.05$, as were proximal (+/- 500 bp) Cut-and-Paste repeats. Unlike the alpha globin gene cluster, there was no significant association between LINE or SINE repeats or segmental duplications and breakpoint locations. This confers with previously reported findings by Henthorn et al (Henthorn, Smithies et al. 1990). In their study, of twenty beta thalassaemia rearrangements investigated, 15 showed microhomology at the breakpoint, six had novel insertions, and four were associated with inverted repeats. Only two rearrangement breakpoints of the cohort in our study were directly within inverted repeats, although thirty breakpoints were within close proximity to one. However, inverted repeats are extremely frequent across this region of the genome, and this association was not statistically significant: a set of randomly generated co-ordinates showed proximity to inverted repeats with a similar frequency to the genuine rearrangement breakpoints. Henthorn et al. conducted their study before detailed databases of such DNA features were available, so the association they reported may have been a false positive result.

In the six NGS characterised rearrangements on chromosome 11 studied in close detail, five showed microhomology or a small novel insertion at the breakpoint. Microhomology at the breakpoint is indicative that the microhomology-mediated end joining (MMEJ) repair mechanism was employed to repair a DSB at this position, while small insertions indicate the NHEJ mechanism was used. This is in clear contrast to the rearrangements associated with alpha thalassaemia which occur between homologous regions, implicating the NAHR mechanism in repairing DSBs at that location.

The lack of significant association between rearrangement breakpoints on chromosome 11 and any of the sequence features analysed suggests more than one mechanism may contribute to rearrangements occurring at this location. Alternatively, they may coincide with another sequence feature that is yet to be linked to the generation of structural variants. Inversion-deletions – which have been reported in beta thalassaemia but not in alpha thalassaemia – may occur via a different mechanism to simple deletions and duplications. A larger cohort of inversion breakpoints would be necessary to draw any conclusions regarding this. The factors that determine which repair pathway is used to repair double stranded breaks are not yet fully understood, but may be influenced by chromatin structure. Differences in chromatin folding at the alpha and beta globin gene loci may be the reason that different varieties and rates of structural rearrangement are recorded at the two loci (Henthorn, Smithies et al. 1990, Yu and McVey 2010). The sizes of rearrangements in this study did not conform to discreet size intervals, but showed an even distribution between 1 and 50Kb, with some variants much larger. This does not support the suggestion by a previous study that rearrangements at this locus are confined to certain size range dictated by the size of chromatin loops at this location (Vanin, Henthom et al. 1983).

Replication origin points showed no association with rearrangements at either locus. However, the FoSTeS mechanism of recombination is likely to occur inside replication forks rather than at their start or end points, so a lack of association of the replication fork origin point with structural variants does not suggest that recombination events are not mediated by FoSTes. There may be more replication origin points in both clusters that have not yet been identified and included in the DeORI database.

There was no association between homologous recombination rate and frequency of structural rearrangement breakpoints. This result is in line with previous findings in the literature that homologous and non-homologous recombination events are governed by different mechanisms, so an association between the two was not likely.

This study has multiple limitations: The breakpoint cohorts on both loci are small, due to the lack of breakpoint data for many recorded variants affecting the globin gene loci. Furthermore, the data that are available could be affected by a sampling bias, in that their breakpoints were in DNA regions that allowed them to be characterised in the first place, rather than in large repeats or highly homologous sequences. Small numbers of inversions and duplications were included in the study, and the numbers may not have been sufficient to identify any trends specific to these rearrangements. Imperfect repeats that have an impact on DNA structure or mutability may not have been detected by the algorithms used to populate the non-B DNA database. The segmental duplications database appeared to contain some inconsistencies, where multiple entries were identified that referred to the same single location. Different mechanisms may contribute to the formation of small and large rearrangements. By analysing all rearrangements >50 bp as one sample cohort, we may have overlooked significant relationships affecting different sized variants formed by different mechanisms.

The sequence motif databases used in this study may not be complete – the DeORI database only represents a fraction of the expected ORI sites in the genome. The relationship between ORI and structural rearrangement is difficult to comment on. The requirements for a region to become an ORI are unclear and there may be additional unreported ORI occurring within the studied in this investigation. The manner in which ORI may influence structural rearrangement is also not fully understood: it is possible that breakpoints are more likely to occur somewhere within replication forks, e.g. at the midpoint of the fork furthest from ORI positions. If this were the case, our analysis, which only examined whether rearrangement breakpoints were associated with ORI's themselves, would not identify that relationship.

Previous attempts to determine a common cause of rearrangements at either of the globin gene clusters have been limited by the small number of rearrangements for which the breakpoint sequences had been determined. With the advent of NGS, the number of fully characterised rearrangements available for study may shed new light on the underlying mechanisms of rearrangements. However, as demonstrated in this study, NGS also has limitations that may prevent to-the-base characterisation of rearrangements under some circumstances. This is a limiting factor of the conclusions that can be drawn from studies that rely on this technology, as the rearrangements that cannot be characterised may display some significant common element that we are unable to detect. Furthermore, many other rearrangements affecting these loci may be unreported, as they are frequently silent when inherited in a heterozygous form, so variants identified from phenotypic screening may only represent a small subset of the

genetic variants present in the general population. This may change as variant screening becomes faster, cheaper and more commonplace with the increased availability of NGS platforms. To make the best use of this data, an improved standard for describing structural variants is needed with a single reference sequence for both loci, large enough to cover all the structural variants in thalassaemia. Reporting these variants should be encouraged, and a high standard of accuracy for doing so maintained. An example of an issue with the current practise is that the HbVar database entry for the 619bp deletion omits the insertion of 7 nucleotides at the deletion breakpoint, despite this being published in the literature. The presence of such insertions may have scientific value. This highlights the need for manual curation of variant databases. However, as variant calling becomes an increasingly automated process and submissions become more frequent, this may become less of an issue.

Discussion

Study Aim

Haemoglobinopathies are a common group of highly heterogeneous disorders in both their genetic cause and phenotypic consequences. Increasingly, the diagnostic laboratory at King's College Hospital (KCH) is required to provide diagnostic testing for a large range of variants causing haemoglobinopathies, as global migration increases. The two key requirements of the diagnostic service are to provide results with accuracy and with speed. The current standard laboratory protocol can require up to five tests and may or may not arrive at a diagnosis. This process is time consuming, particularly when rare or novel rearrangements are encountered. The current standard practise could be streamlined by adopting a NGS approach. Targeted NGS has advantages over other methods such as SNP arrays and CGH array for detecting structural variants, as the dataset provides variant calls, coverage and phase across the length of the sequence fragment. Its main drawback is that it only works for a targeted area and if the structural variant is bigger than the ROI it may not detect the breakpoints.

This study set out to evaluate NGS as a tool for the diagnosis of haemoglobinopathies. In order to improve upon the current standard of diagnosis, we intended to create an assay that was fast, medium to high throughput, comprehensive and highly accurate. This technology has already been proved to be a highly accurate means of detecting single nucleotide changes, which are a frequent cause of haemoglobinopathies. Structural variants also cause haemoglobinopathies and therefore we needed to evaluate the capability of NGS to detect these variants before it could be claimed that a single technology could detect all pathogenic genetic variation.

The development of this assay for diagnostic use can be divided into three stages: sample preparation, target enrichment and analysing sequencing data.

The Sample Preparation Process

There are multiple methodologies available for preparing DNA samples for sequencing and enriching for target regions, as well as a variety of sequencing platforms. At the start of this investigation, we opted to use Agilent SureSelect library preparation in combination with in-solution bait capture target enrichment with Illumina sequencing technology. Illumina sequencing platforms are widely available, with an increasingly diverse technological portfolio, catering for different sequencing requirements, and

associated with a low error rate, making them ideal for our investigation (Harismendy, Ng et al. 2009).

The first sequencing run we performed in this study was on the Illumina HiSeq 2000 platform. A cohort of 12 samples were prepared for sequencing manually according to the Agilent SureSelect protocol. The run achieved high read depth per sample as expected given the sequencing capacity of this platform. This allowed dosage changing rearrangements to be identified based on coverage (RPKM) data, and known SNPs in the cohort were accurately reported. Shortcomings of this method were the length of time required for sequencing (the HiSeq 2000 takes two weeks to generate FASTQ data), the small DNA fragment (starting size ~150 bp), and the small read size available on the HiSeq 2000 (2x97 bp). The small fragment and read sizes used in this assay prevented complete characterisation of several of the sequenced variants (HiSeq Samples 3, 5, 7, 9, 10, 11). A false negative result for the presence of an indel was also identified in one sample, which may have been detected correctly if a longer read length had been used (HiSeq Sample 6). This sequencing run showed that high sample throughput was possible on the HiSeq 2000, but the assay was not fast or accurate enough for diagnosis and did not provide comprehensive variant detection.

Subsequent runs were performed on the MiSeq platform. In MiSeq Run 1 and MiSeq Run 2, the starting DNA fragment size was increased to ~500 bp. Smaller batches of samples were prepared for sequencing on the MiSeq platform, which has a lower capacity than the HiSeq 2000 but affords faster sequencing and longer read lengths. It was found that the sample preparation process preferentially amplified smaller DNA fragments, and that over the course of sample preparation the average fragment size was reduced from 500 bp to 300 bp. The increased read length and fragment length in these assays improved sequence alignment and allowed a rearrangement that could not be characterised on the HiSeq 2000 (HiSeq Sample 10) to be resolved with to-the-base accuracy (MiSeq Sample 3). The rearrangement involved a deletion which had been visible in the HiSeq 2000 data, and also a previously unidentified inversion which could be detected with to-the-base accuracy in the MiSeq data. With the identification of this breakpoint, Gap-PCR primers against the inversion-deletion breakpoints were designed and completely resolved the sequence at both positions. Without the knowledge that an inversion had occurred, previous attempts to create a Gap PCR assay for this rearrangement had failed, as primers had only been designed to target the deleted region. Of the four novel rearrangements sequenced in these two runs, three were found to have this rearrangement (MiSeq Samples 2, 3 and 6). Two positive controls for known deletions were also sequenced and the breakpoints as reported in

the literature were successfully identified (MiSeq Samples 1 and 5). With the inclusion of only a single negative control in each of these runs, the sensitivity of the assay in differentiating genuine structural variants from background noise was reduced compared to the HiSeq 2000 platform.

An automated sample preparation workstation, the BioMek FX^P was brought into the laboratory. The intention of this was to reduce the man-hours associated with sample preparation and make the process more uniform, reducing the noise in this assay associated with inter-sample variation. Implementing the complex sample preparation workflow on this platform was extremely challenging. Several attempts to prepare samples for sequencing using this workstation resulted in failures, including MiSeq Runs 3 and 4, in which the sample preparation process was found to have completed successfully, but after sequence alignment, it was found that the hybridization stage of the process had failed.

MiSeq Runs 5 and 6 were prepared on the BioMek FX^P platform, apart from the hybridization stage which was performed manually. The average fragment length of the input DNA in these samples was 600 bp. An additional bead-based size selection step was introduced prior to the SureSelect library preparation step. This tightened the fragment size range, removing the short fragments that had been preferentially amplified in previous experiments. The input DNA volume was increased from 3 µg to 5 µg to offset the increased DNA loss from the double size selection process. Based on the read depth achieved in previous runs, we increased the number of samples prepared and sequenced per batch to 14. Automated sample preparation increased the number of samples that could be prepared for sequencing simultaneously and reduced the number of man-hours required for sample preparation. MiSeq Run 5 did not include sufficient negative controls for reliable variant detection. It was necessary to use additional negative control data from MiSeq Run 6 to reliably identify structural variants in this run. In both runs, dosage changing rearrangements could be identified from coverage (RPKM) data. Some rearrangements could be resolved with to-the-base accuracy and were confirmed by Gap-PCR. Other rearrangements could be broadly identified, and then resolved with to-the-base accuracy by Gap-PCR with primers designed based on the approximate breakpoint locations. Some samples could be broadly identified, but Gap-PCR primers designed against the approximate break points did not amplify a break-point product.

The Bait Capture Library

Two bait capture libraries were designed to enrich prepared DNA samples for fragments originating from regions of interest. Bait Capture Library 1 was designed using eArray (Agilent). This library covered the regions of the globin gene clusters on chromosome 11 and 16 using 1x tiling, plus boosting for orphan baits, or baits with a high GC content. This library did not cover a large enough portion of chromosome 16 to fully capture some structural rearrangements sequenced in this study (HiSeq Samples 1, 3 and 8). Sequencing experiments performed using this bait capture library also showed multiple regions where coverage was highly variable between samples and variation between samples across the bait covered region.

Bait Capture Library 2 was designed to succeed Bait Capture Library 1, extending the region of bait coverage on chromosome 16. This extension meant that all but one of the rearrangements sequenced at the chromosome 16 locus were completely covered by the bait design. The exception was a sample that showed a duplication >2Mb in length (MiSeq Sample 22) exceeding the region of coverage. This library increased the number of baits covering core regions of chromosomes 11 and 16 where the majority of variants were expected to occur. The increased number of baits allowed improved resolution of rearrangements (as each 120 bp bait overlaps another by 60 bp, the resolution of RPKM data plots increases from +/- 120 bp to +/- 60 bp). Sparse tiling was also extended across an even wider region of each locus, increasing its potential to characterise large rearrangements. The increased tiling density across the core region of interest increased the amount of DNA fragments originating from this region which were captured per sample, improving the level of coverage achieved and proportion of reads that were on target. Increasing the tiling density and employing the boosting parameters recommended by Agilent to maximise bait performance did not reduce the variation in coverage across the target region within samples, or the variation seen between samples.

Agilent's library design tools (eArray and SureDesign) were not able to design baits against the first 60 Kb of chromosome 16. This is the telomeric region of the chromosome, and the sequence of this region is poorly understood. Several variants removed some or all of this region, and because it was not included in the library design, the size of these variants and their breakpoint sequences could not be determined. Bait Capture Library 2 also included regions of chromosome X and chromosome Y so that the assay could also be used for sex determination during prenatal diagnosis. Two regions on chromosome 6 and chromosome 2 were also

included as they are known to contain genetic variants controlling foetal haemoglobin, which influence the clinical severity of beta thalassaemia and sickle cell disease. Such genetic variants may also be diagnostically relevant in the future to increase disease severity prediction once the role of these loci in modifying disease state is better understood. A limitation of Bait Capture Library 2 is that it does not include the ATRX region on chromosome X. This is believed to be a regulatory region for multiple genes, including the alpha globin loci on chromosome 16. Inherited variants at this locus can result in alpha thalassaemia/mental retardation syndrome (Gibbons, Wada et al. 2008, Lugtenberg, de Brouwer et al. 2009). Some cases of acquired thalassaemia in patients with MDS have also been attributed to somatic mutations at this locus (Steensma, Gibbons et al. 2005). Although these conditions are extremely rare, including the region would have increased the diagnostic scope of this assay.

Data Analysis

Data analysis was performed using an off-the shelf package developed by SoftGenetics called NextGene. This package included tools for quality filtering data from the sequencing platform, converting it into FASTA format and aligning it to a reference sequence of the human genome with user-defined stringency. The program included multiple tools for data analysis which were used with varying degrees of success. The mutation report successfully identified all sequence variants previously reported in all the sequenced samples. It also provided quality score and mutant allele frequency information which were both useful metrics for assessment and clinical interpretation. Opposite and same direction read reports were able to isolate break-point spanning reads in several of the rearrangements that were sequenced, but they were not always informative (See results: detection of 3.7 Kb deletion in opposite direction read data). In some cases, breakpoint reads were included in the alignment and not displayed in the opposite direction read report. In the same direction read report, a genuine inversion (English V) was only present in one entry in the report (MiSeq Sample 2), although this can be attributed in part to the presence of a non-functional bait covering one of the break point positions which reduced the number of reads captured at the breakpoint location. Both these reports were hampered by a large amount of ‘noise’ stemming from misalignment between repetitive, off-target or highly homologous sequences. This was an issue that affected the most homologous of the globin genes (*HBA1* and *HBA2*; *HBGA* and *HBGG*) and regions that were repetitive. The expression report was extremely useful for calculating coverage across a user-defined set of regions. RPKM plots based on this information provided the basis of structural variant detection in this study. This method of analysis proved to be

extremely accurate in identifying dosage changing variants. Where the breakpoints of rearrangements were identified in the read reports or in the NextGene Viewer, the breakpoint sequences conformed absolutely to the breakpoint sequences determined through Gap-PCR and Sanger sequencing.

The software had several shortcomings, using the multiple results above to characterise structural variants was awkward and time consuming. The 'structural variant' tool (included in an updated version of the software introduced part way through this study) failed to detect any of the rearrangements that were known to occur in the samples, and generated multiple false positive calls across the genome based on off-target and misaligned sequences. Calling of indels yielded at least one false negative result and they were listed in a cumbersome manner in the mutation report, with multiple entries giving different scores for the same variant. The software frequently produced misalignments of reads from regions with high homology, such as the *HBGA* and *HBGG* genes and the *HBA1* and *HBA2* genes. This impaired variant detection in several cases, including the Asian-Indian inversion-deletion and the English V rearrangement variants in the *HBB* complex, and the (α-$^{3.7}$/) deletion in the *HBA* complex. A bioinformatician was recruited into the diagnostic department in order to design a bespoke pipeline for NGS data analysis. This pipeline is expected to address these issues by employing a range of tools specifically designed for each of the required tasks to provide more accurate, consensus based calls (Nielsen, Paul et al. 2011) presented in a more streamlined fashion. Preliminary tests of this program showed that it was able to confidently detect the (α-$^{3.7}$/) deletion due to the superior precision in sequence alignment capabilities. This pipeline is still under development.

Conclusions Regarding Assay Potential

This technique was capable of characterising multiple structural variants. All dosage changing variants could be identified with an accuracy comparable to CGH array and greater than MLPA, both techniques that are part of the current standards in diagnostics. In many cases, the variants could be identified with to-the-base accuracy. Thirteen novel variants were sequenced in this study, which had eluded characterisation by other techniques. Eight of these variants were resolved with to-the-base accuracy. For the remaining five variants, it was possible to identify dosage changes that affected the globin genes and estimate the size of the rearrangement (except in the instance of the large duplication in MiSeq Sample 27 which exceeded the size of the target region). Twelve positive controls for known structural variants were also sequenced. All of these could be detected with accuracy comparable to CGH

array, and seven could be resolved with to-the-base accuracy. In addition identifying structural variants, the technique is capable of determining the location and orientation of duplicated sequences and identifying novel inversion events. The presence of concurrent sequence variants affecting the globin genes could be also be detected without the need to perform additional tests. None of the techniques currently used in the standard diagnostic protocol have this capability. In theory, the technique is also capable of detecting translocations, although none were identified in this study.

The two principle limiting factors of characterising variants in haemoglobinopathies using this technique are the inability of the assay to sequence repetitive regions, and shortcomings of the alignment software. It is necessary to employ repeat-masking when designing the bait capture library in order to avoid capturing huge amounts of off-target sequence. In bait capture library 1, baits were allowed to extend up to 20 bp into repetitive regions. A DNA fragment must share 40bp of homology with a bait in order to be selected for during hybridization. In spite of the specificity for the target region that this should confer on the captured fragments, a large amount of off-target sequence (An average of 39% of total reads in MiSeq Run 6) was still captured. The amount of off-target sequence fell over the course of the runs performed in this study, so it is possible further improvements will be seen as the persons performing library preparation become more experienced. The inclusion of these fragments which limited read depth of the target regions and contributed to noise in both RPKM data and opposite/same direction read reports. In spite of this, break-point sequences for multiple variants were not captured when they fell within repeat sequences that were too large to be covered by fragments ligating to baits surrounding these regions. The alignment software was unable to detect any of the structural rearrangements sequenced automatically, and this required work-arounds including using the expression report to detect coverage changes, opposite and same direction read reports which contained a large number of false positive results, and having to delve into the FASTA data files to retrieve the sequences of read pairs that appeared to contain rearrangement break point sequences. Although we did not have a large amount of data concerning indels and SNPs in the sequenced samples, a clear indel was noticed in one sample that had not been called (HiSeq Sample 6). This indel was present in a large proportion of reads in a region with a high level of coverage, and was also included in dbSNP. NextGene appeared to have issues with aligning indels that it did detect, which provided cause for concern in using this tool for accurate diagnosis. Several samples had known single nucleotide changes. All of these were identified with satisfactory accuracy. For this reason, although the sample preparation and target

enrichment strategies were clearly capable of indicating the presence of variants (if not capturing the breakpoints), it was necessary to develop a bespoke in-house pipeline for data analysis rather than relying on the NextGene package. Balanced rearrangements between repetitive regions not included in the bait capture library could be undetectable using this technique. There is a possibility that some of the variants that could not be resolved in this study included balanced rearrangements that were not identified, meaning that Gap-PCR primers were not designed against the correct breakpoint positions for these variants. This represents a theoretical blind spot in this assay, although cases where thalassaemia is caused by a rearrangement that doesn't increase or decrease the globin gene dosage (by deletion or duplication) are likely to be extremely rare. This is one instance where third generation sequencing technology, such as single molecule sequencing could prove to be a significant improvement on this assay.

The current diagnostic strategy involves a series of tests, each followed by a review of the results and then if necessary, progression to another test. The turnaround time for an individual sample following this algorithm may take up to a month for comprehensive testing in the case of a rare variant. The cost for characterisation of novel rearrangements by MLPA/CGH array resolution may reach £700 per patient. Complete characterisation of novel variants with to-the-base accuracy is not considered to be viable in terms of time and cost in a diagnostic laboratory and hence in the past characterisation has been carried out by research groups. By contrast, this technique represents a single assay that can achieve MLPA/CGH array level resolution with a turnaround time of less than two weeks. The cost of the assay per patient is approximately £500 and sequencing costs will continue to fall. The technique has the potential to detect novel balanced rearrangements and provide to-the-base characterisation of novel variants. Accurate diagnosis is essential if individuals will use these results for prenatal diagnosis and in these tests gap PCR is preferred rather than NGS, due to the speed and accuracy. Gap PCR is only possible when the breakpoints have been characterised, or are at least approximately known.

To surmise, CGH array and MLPA add to the current diagnostic repertoire, patching diagnostic holes where specific types of variants need to be identified. NGS promises to be a single technique replacing all that has gone before with increased diagnostic yield.

It would have been advantageous to re-sequence some samples to assess how changes we made to the assay improved its ability to characterise rearrangements.

The false negative indel identified in the Asian Indian inversion deletion (HiSeq Sample 6) may have been detectable with the longer fragment and read lengths used in MiSeq sequencing runs. This could also have benefitted analysis of the British deletion sample (HiSeq Sample 9). Time constraints due to extensive issues first with off-site sample preparation and then with the setup of the BioMek FX^P platform prevented re-runs of these samples. Two variants were not resolved due to insufficient DNA being available to complete the investigation (HiSeq Sample 11; and HiSeq Sample 3, later MiSeq Sample 22).

Novel Rearrangements of the Globin Gene Cluster

Many of the novel 'test' variants sequenced in this study were selected for NGS because they had eluded diagnosis by other means. Eight of the variants were resolved with to-the-base accuracy in this study, where other techniques had failed to determine their breakpoints. Some of these rearrangements had interesting molecular features and phenotypic consequences.

The English V Deletion (MiSeq Samples 2, 3 and 6)

The proband for this deletion was an English Anglo-Saxon female who had hypochromic microcytic anaemia that was unresponsive to oral iron supplements. Her haematological profile showed an imbalanced α/β globin chain synthesis (reduced β globin) with normal HbA_2 and HbF fractions. This phenotype is consistent with heterozygous εγδβ thalassaemia (Shooter, Rooks et al. 2015). Southern blotting had identified a deletion removing the β globin gene cluster LCR, but the breakpoints of the deletion could not be resolved.

Next Generation Sequencing revealed that the rearrangement consisted of two events: the deletion identified in southern blotting which removed the LCR, *HBE* and the third exon of *HBG2*, and additionally an inversion of 59 Kb encompassing the rest of the beta globin gene cluster. The inversion had occurred first, bringing together two sequences that were highly homologous to one another. This was followed by a deletion that removed 122 Kb of upstream sequence and also 82 bp of the adjoining inverted sequence. The inversion breakpoint was at the centre of a 156 bp palindrome which may have played a role in forming this rearrangement.

Despite rearranging *HBB*, *HBD*, *HBG1* and both deleting and inverting *HBG2*, the inversion portion of this variant is phenotypically silent, as the LCR removed with the deletion prevented any of these genes from being transcribed. Only one other inversion-deletion causing thalassaemia has been reported in HbVar. No

rearrangements causing εγδβ thalassaemia that have been reported to date include inversions. All but one εγδβ thalassaemia rearrangements reported to date have been confined to a single family. This rearrangement was found in 3 different families, all of Caucasian Anglo-Saxon origin. Given the complexity of the rearrangement, it seems likely that the rearrangement is of a single origin. This is supported by the similarities between variant calls in the region around the rearrangement in all three cases sequenced. In one family, the proband was adopted. Undetected inversion events may underlie other εγδβ rearrangements that have been reported in the past where break point products were resistant to PCR amplification. A previously reported εγδβ deletion, English I, is reported to delete approximately 100 Kb of sequence including the LCR, extending from the third exon of *HBG2* (Curtin, Pirastu et al. 1985). This bears a close resemblance to the deletion portion of English V, and may be another example of this variant. In that case, the breakpoint was never fully characterised.

The African I Duplication (MiSeq Sample 7)

The African I duplication is the first reported duplication of the entire beta globin gene cluster. Although this rearrangement displaces 80Kb of sequence, the additional globin genes are expressed and would be phenotypically silent during screening. It was only identified due to the dilution affect it had on the expression level of the HbS (13% total not 40%) variant on the other chromosome, seen by the screening laboratory.

The duplication is top-to-tail in orientation, as all other reported duplications in the alpha and beta globin gene cluster have been to date. This may suggest a common mechanism is responsible for duplications affecting the globin gene clusters. Throughout the duplicated region, only two genotypes were recorded (with mutant allele ratios of 70% or 30%). This suggested that the rearrangement that caused the duplication had been intrachromosomal, creating an identical copy of the original sequence immediately adjacent to it.

Novel Beta Globin Duplication 2 (Not yet named) in MiSeq Samples 41 and 42

A second duplication affecting the beta globin gene cluster was also characterised in a family trio, where it was also identified by an abnormal HbS fraction in the proband, who had inherited the duplication from the father. In this case, the duplication incorporated sequences from two different alleles into the same chromosome, one of which carried the normal *HBB* gene and the other the HbS variant. The proband had inherited this duplication from the father and also inherited another copy of the HbS

allele from the mother (MiSeq Sample 40). Thus, the proband genotype was HbS/HbS/HbA, the father's genotype was HbS/HbA/HbA and the mother's genotype was HbS/HbA.

Like the African 1 duplication, the rearrangement inserted the duplicated sequence immediately after its first iteration on chromosome 11 and in top-to-tail orientation. Unlike African 1, the duplication resulted from inter-chromosomal recombination resulting in two different genotypes in the duplicated region.

One breakpoint of the rearrangement was in a LINE repeat and the other within unique sequence. The breakpoints had a 2 bp microhomology. The duplication did not include the beta globin LCR, meaning that the two copies of the globin gene cluster present on the duplicated allele shared a single LCR. We were unable to determine the order of the two duplicated regions on the chromosome based on the NGS data, however, the HbS fractions reported in the father (HbS 41.3%) and the proband (no detectable HbA) implied that the HbS gene was transcribed at a higher level than the HbA gene. We suspect from this that the HbS gene is in the copy of the duplicated region closest to the LCR, and is thus preferentially expressed.

Novel Deletion of the HS40 Regulatory Region of the Alpha Globin Gene Cluster (MiSeq Sample 33)

A novel 73 Kb deletion on chromosome 16 was identified which removed the HS40 locus that controls alpha globin gene expression. This variant conferred an alpha thalassaemia phenotype similar to an alpha zero thalassaemia variant (both HBA genes deleted) despite leaving the entire alpha globin gene cluster intact. Several similar deletions removing this region and resulting in alpha thalassaemia have been reported previously (Hatton, Wilkie et al. 1990, Viprakasit, Kidd et al. 2003, Viprakasit, Harteveld et al. 2006). Like these deletions, the breakpoints of this rearrangement were both situated in highly homologous Alu repeat sequences. This demonstrates the importance of including regions other than the globin genes themselves in assays for diagnosing haemoglobinopathies. This sample had previously been analysed on a targeted CGH array but the results from the array did not show a clear breakpoint at both ends making it difficult to accurately design GAP PCR primers. This may be due to the GC content and repetitive nature of the genome at the tip of chromosome 16.

The Molecular Basis of Structural Rearrangements Causing Haemoglobinopathies

A wide range of variants causing haemoglobinopathies have been identified (Gibbons, Wada et al. 2008, Lugtenberg, de Brouwer et al. 2009). These can range from a single base change to the rearrangement of kilobases or even megabases of DNA. This is in contrast to the majority of inherited diseases, which are caused by a single variant in the majority of cases. The range of variants that can cause thalassaemia presents a challenge to diagnosis.

Understanding the molecular basis of these rearrangements could be beneficial to diagnosis by informing break point characterisation. Many rearrangements occur within repetitive regions that cannot be effectively selected for by target enrichment for NGS. Identifying features of these sequences that contribute to rearrangements could help in cases where one or more breakpoints are within repetitive regions and no breakpoint spanning sequences are captured by the NGS design. By understanding what types of sequences are involved may allow a simplified approach to the design of Gap-PCR assays allowing confirmation and to-the-base characterisation of the rearrangement.

In the case of the novel duplication identified in MiSeq Sample 34, both breakpoints of the rearrangement were situated in unmapped repeat regions. Based on findings from two other duplications characterised in this study (African 1: MiSeq Sample 7 and HiSeq Samples 1/8), we hypothesised that the duplicated sequence would be adjacent to the original sequence, and in head-to-tail orientation. We designed primers according to this hypothesis, and were successfully able to produce a Gap-PCR product and achieve to-the-base characterisation of this rearrangement. The English V inversion-deletion initially evaded characterisation in part because the inversion breakpoint was within a palindromic sequence. The palindrome rendered the bait that covered this position non-functional, which made coverage at this position appear balanced in relation to the negative control (when in reality coverage in both samples was zero), confounding diagnosis. Identifying such features within the target region during bait design could allow improved positioning of baits that cover this region to improve their performance and reduce the risk of false-negative results in the read coverage data. In the second design no baits were tiled in this region.

We found that many rearrangements affecting the alpha globin gene locus began and/or ended in SINE repeat or segmental duplication regions. These can result in a breakpoint that is indistinguishable from the normal sequence expected at either

location. Therefore, we recommend that when designing PCR primers for rearrangements based on suspected but unknown breakpoint locations, nearby segmental duplication or SINE repeat regions should be treated as likely breakpoint positions and included in the target region for the rearrangement breakpoint.

Better understanding of how these rearrangements occur is scientifically interesting: It could help illuminate what mechanisms cause mutation and what elements in the DNA sequence contribute to its propensity to mutate. In particular, it could identify features of the regions in the alpha and beta globin gene loci that make them more susceptible to mutation, or whether the variety of variants we see at these locations are predominantly due to other factors, such as selection pressures (Lam and Jeffreys 2006).

Previous studies had investigated small numbers of variants causing thalassaemia at both the alpha and beta globin gene loci (Vanin, Henthom et al. 1983, Nicholls, Fischel-Ghodsian et al. 1987). These studies had shown a relationship between rearrangements of the alpha globin gene cluster and both segmental duplications and SINE repeats (an association also reported in numerous other diseases caused by duplications (Potocki, Chen et al. 2000)). Rearrangements at the beta globin gene cluster had been noted to show some association with inverted repeats. In both studies, only a small number of fully characterised rearrangements were available for analysis. Since then, many rearrangements causing thalassaemia have been identified with to-the-base accuracy and recorded online in the HbVar database. We compared the locations of known breakpoints at the globin gene clusters to databases of DNA features that have been implicated in causing structural rearrangements. The dataset available for this study was still small and was limited by sampling bias. Principally, rearrangements in which breakpoints have been successfully identified may represent an inherently different subgroup of rearrangements at from those that have not been characterised. A common underlying cause of the uncharacterised rearrangements may also make them difficult to sequence. Furthermore, several of the datasets of DNA features may not be comprehensive at this point. As the catalogue of fully characterised variants causing thalassaemia grows with the expansion of DNA screening, we hope that further studies will examine their relationship with more comprehensive lists of DNA features. Nonetheless, we used GAT, a tool for comparing the overlap of genomic co-ordinates to identify associations between our small variant dataset and DNA features and identified some significant trends. Variants on chromosome 16 were significantly associated with SINE repeats ($P<0.001$) and segmental duplications ($P<0.001$). A mild association with proximity to LINE repeats

was also uncovered at this locus. Variants on chromosome 11 showed a mild association with mirror repeats and proximity to cut-and-paste repeats. Although many rearrangement breakpoints on chromosome 11 coincided with inverted repeats as previously suggested, this relationship was not statistically significant and may result from the high frequency of these features on chromosome 11.

Examination of the break points of rearrangements sequenced in this study revealed that on chromosome 11 most variants showed either novel insertions (4/6) or microhomology (2/6) at the breakpoint. These features are associated with two related DNA DSB repair mechanisms – NHEJ (non-homologous end joining) with insertions and MMEJ (microhomology-mediated end joining) with microhomology. This is in line with the findings of the previous study of rearrangements on chromosome 11 (Vanin, Henthom et al. 1983). Rearrangements on chromosome 16 primarily brought together highly homologous SINE or segmental duplication regions. This is associated with the NAHR (non-allelic homologous recombination) DSB repair mechanism, and also echoes previous study of rearrangements at this locus (Nicholls, Fischel-Ghodsian et al. 1987, Emmanuel 2001, Liu, Lacaria et al. 2011). There is no clear rule governing all rearrangements seen at either of the globin gene clusters. Multiple mechanisms clearly contribute to rearrangements at both locations, but both loci also appear to be governed by different factors. This may be due to either the features of the DNA sequence at these locations, or perhaps with the 3D conformation of the chromosomes.

Interestingly, the two duplications on chromosome 11 appear to have been formed by different processes, one by intra-chromosomal and the other inter-chromosomal recombination. Inter-chromosomal recombination is believed to be responsible for one duplication because only two haplotypes are detectable in the sample, and it is unlikely that two different chromosomes would have identical sequence over the entire 80 Kb duplicated region.

The polymorphic feature of the globin genes has been noted for some time. Homologous recombination occurs at a high rate on chromosome 11, and de novo non-allelic homologous recombination events, such as those causing the (α-$^{3.7}$/) deletion, occur frequently on chromosome 16. It has been suggested that the globin loci are exposed to multiple gene conversion events that may be exchanging small (<100 bp) sections of genetic material during meiosis and that the genes and their protein products tolerate this well. Hence, there are over 800 globin gene variants and some of the more common variants are on multiple backgrounds (Law, Luo et al. 2006,

Borg, Georgitsi et al. 2009). The sickle variant is believed to have arisen on 5 separate occasions, according to founder theory, due to the haplotypes with which it is associated.

Applications of This Work and Future Directions

This work has shown that next generation sequencing can capture the entire range of variants associated with haemoglobinopathies to at least the accuracy of the techniques that are currently in use. For rare pathogenic variants, this can be delivered with higher speed, lower cost per patient and fewer man-hours of work. The technique also has the added advantage that it can detect balanced rearrangements. While the technology is capable of capturing these rearrangements, it can be let down by the analysis software. For NGS-led diagnosis to be delivered with the accuracy necessary for diagnostic assays, it is imperative that the correct analysis software is used. As a result of the findings of this study, a red cell diagnosis panel has been developed in the Department of Molecular Pathology at King's College Hospital to be used for the rapid diagnosis of multiple haematological disorders affecting erythrocytes. This panel will be used to quickly and efficiently identify the majority of variants in a number of diseases that the laboratory is tasked with diagnosing. The bait library developed during the study reported here will be used to characterise novel structural variants that are identified by this panel, which can then be resolved more rapidly than via the previous diagnostic workflow, and with the potential to definitively determine the break points. The techniques used in this study could be applied to the diagnosis of other diseases with the same accuracy and constraints.

Efforts are currently being made to introduce NGS to multiple domains within the NHS, most notably through the 100k Genomics England project. The 100k project intends to sequence 100,000 genomes from around 70,000 NHS patients with cancer or rare diseases and their family members. This project intends to pave the way towards personalised medicine by identifying new or improved treatment strategies for these diseases which could become available for the wider population in the future as screening becomes more commonplace. Developing new treatments based on data generated by the 100k genome project is likely to be a slow process: a new drug takes many years to progress from target identification to approval for clinical use. This raises the issue that many of the individuals who initially provide the data for the 100k genome project are unlikely to themselves benefit from its findings. In some cases, however, the project has already permitted better treatment of current subjects, by achieving a more accurate diagnosis of rare or poorly characterised conditions.

Introducing genomics as an aspect of routine care will require a huge amount of work on the part of the NHS in terms of training, cost and logistics. New ethical guidelines and data protection policies may need to be implemented to ensure public support of the initiative. Many individual laboratories are also introducing disease targeted gene panels or exome sequencing, replacing the low-throughput high-cost sanger sequencing that was previously the gold standard in many of them. In the UK many laboratories are participating in the NHS Laboratory England Redesign scheme to help prepare laboratories to provide genomic services. New tests are being commissioned by the NHS for many diseases. In the USA the ACMG (American College of Medical Geneticists) has issued a list of recommendations that should be considered before patients are submitted for genetic testing. A major aim of these guidelines is to limit testing when it is unnecessary, rather than the practise being given an inaccurate 'cure-all' reputation. They also recommend limitations to discourage 'hypothesis free testing' which is likely to confound results without proper controls being in place. Concerns also include the use of markers that have not been fully validated to inform patient care, highlighting the particular examples of APOE and MTHFR (in Alzheimer's Disease and hereditary thrombophilia, respectively), the clinical significance of both of which are not yet fully understood.

Future work should be performed to evaluate bioinformatic strategies that allow accurate diagnosis and reporting of indels, SNPs and structural variants automatically. A bespoke pipeline for this is currently being developed in the Molecular Pathology laboratory at King's College Hospital. Variants affecting the regions on chromosome 2 and 6 included in this assay should be recorded and compared with clinical data from patients wherever possible in order to discern the roles of these regions in modifying haemoglobinopathy disease phenotype. Attempts should be made to characterise any novel variants identified by this assay in the future with to-the-base accuracy. Successfully characterised variants should be submitted to HbVar, where this data could lead to a better understanding of how rearrangements at the alpha and beta globin gene loci are formed. Furthermore, creating a comprehensive library of breakpoint sequences for known thalassaemia variants could speed diagnosis of such variants in the future. As demonstrated here, it is possible to deliver a diagnosis by querying FASTA data for a list of known unique breakpoint sequences. This could be incorporated as a preliminary step in future analysis pipelines that could speedily identify previously recognised rearrangements.

Applications of this assay that were not investigated in this study include the detection of translocations, genetic mosaicism, and non-invasive prenatal diagnosis.

Translocations are balanced rearrangements that excise DNA from one chromosome and move it to another location. Translocations affecting the alpha globin genes have been reported, but are extremely rare. These incidents have been phenotypically silent, as the enhancer regions for the genes were also translocated, meaning there was no effect on the expression of the alpha globin genes (Ledbetter 1992). Such events still have potentially serious consequences as the balanced nature of these rearrangements could mask the potential of carriers to pass on alpha-zero chromosomes and with them the risk of Bart's Hydrops Fetalis, if the partner has alpha-zero thalassaemia. The current diagnostic standard is unable to detect translocations, but in theory the assay developed in this study would have this capability: the chromosome from which the sequence had been removed would show opposite direction reads spanning the excised region, and the transplanted sequence would also be captured on the recipient chromosome, with opposite direction reads occurring at the points where chromosome 16 sequence (still captured by the bait library) gave way to the sequence of the original chromosome.

Genetic mosaicism and acquired alpha thalassaemia are known but relatively rare phenomena (Harteveld, Refaldi et al. 2013). Acquired alpha thalassaemia is occasionally reported in MDS sufferers (Steensma, Gibbons et al. 2005)[1]. NGS has been widely reported to amplify all the input DNA from a sample at an equal rate, meaning that the final allele frequencies in analysis are representative of the sample makeup (as seen in this study where the frequencies of heterozygous SNPs in duplicated regions have allele frequencies of 30% or 70%). Therefore, subpopulations of the sequenced DNA with a different genotype to the sample consensus that are indicative of mosaicism can be identified by this method.

NGS has great potential in non-invasive prenatal diagnosis of haemoglobinopathies as it is capable of identifying the small fraction of foetal DNA that circulates in the maternal blood stream during pregnancy (Lo, Hjelm et al. 1998, Lo, Tein et al. 1998). This technique is still in early development, but could one day be applied to haemoglobinopathies. An additional boon to prenatal diagnosis offered by NGS is that it can be used to identify mosaic abnormalities that can occur in the foetus during pregnancy without the need for re-sampling via CVS or amniocentesis, which is

[1] The majority of cases of acquired alpha thalassaemia in MDS are X-linked and associated with acquired clonal mutations inactivating the ATRX gene that regulates alpha globin gene expression. The current bait capture library Bait Capture Library 2 does not include this region. However, it should be added into future versions of the assay to ensure comprehensive diagnosis.

necessary using the currently used diagnostic techniques (Mendilcioglu, Yakut et al. 2011). We are still exploring the diagnostic potential of NGS but it is clear it will transform how healthcare is delivered. For haemoglobinopathies and many other conditions it may be logical to sequence first and then follow up with phenotypic assessments, reversing the current diagnostic paradigm. How this technology is integrated into UK healthcare is the current focus of NHS England.

References

Albitar, M., C. Peschle and S. A. Liebhaber (1989). "Theta, zeta, and epsilon globin messenger RNAs are expressed in adults." Blood **74**(2): 629-637.

Ali, S. A. (1970). "Milder variant of sickle-cell disease in Arabs in Kuwait associated with unusually high level of foetal haemoglobin." British journal of haematology **19**(5): 613-619.

Amos-Landgraf, J. M., Y. Ji, W. Gottlieb, T. Depinet, A. E. Wandstrat, S. B. Cassidy, D. J. Driscoll, P. K. Rogan, S. Schwartz and R. D. Nicholls "Chromosome Breakage in the Prader-Willi and Angelman Syndromes Involves Recombination between Large, Transcribed Repeats at Proximal and Distal Breakpoints." The American Journal of Human Genetics **65**(2): 370-386.

Antonarakis, S. E. (2010). Human genome sequence and variation. Vogel and Motulsky's Human Genetics, Springer**:** 31-53.

Ardlie, K. G., L. Kruglyak and M. Seielstad (2002). "Patterns of linkage disequilibrium in the human genome." Nature Reviews Genetics **3**(4): 299-309.

Ayi, K., F. Turrini, A. Piga and P. Arese (2004). "Enhanced phagocytosis of ring-parasitized mutant erythrocytes: a common mechanism that may explain protection against falciparum malaria in sickle trait and beta-thalassemia trait." Blood **104**(10): 3364-3371.

Bacolla, A., A. Jaworski, J. E. Larson, J. P. Jakupciak, N. Chuzhanova, S. S. Abeysinghe, C. D. O'Connell, D. N. Cooper and R. D. Wells (2004). "Breakpoints of gross deletions coincide with non-B DNA conformations." Proceedings of the National Academy of Sciences of the United States of America **101**(39): 14162-14167.

Bacolla, A., Wells, R.D. (2009). "Non-B DNA conformations as determinants of mutagenesis and human disease." Molecular Carcinogenesis.

Baird, M., C. Driscoll, H. Schreiner, G. V. Sciarratta, G. Sansone, G. Niazi, F. Ramirez and A. Bank (1981). "A nucleotide change at a splice junction in the human beta-globin gene is associated with beta 0-thalassemia." Proceedings of the National Academy of Sciences **78**(7): 4218-4221.

Benesch, R. and R. Benesch (1961). "The chemistry of the Bohr effect." J Biol Chem **236**: 405-410.

Berg JM, T. J., Stryer L., I. L. Berg, R. Neumann, S. Sarbajna, L. Odenthal-Hesse, N. J. Butler and A. J. Jeffreys (2002). "Hemoglobin Transports Oxygen Efficiently by Binding Oxygen Cooperatively

Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations." Biochemistry.

Blair, I. P., J. Nash, M. J. Gordon and G. A. Nicholson (1996). "Prevalence and origin of de novo duplications in Charcot-Marie-Tooth disease type 1A: first report of a de novo duplication with a maternal origin." American journal of human genetics **58**(3): 472.

Borg, J., M. Georgitsi, V. Aleporou-Marinou, P. Kollia and G. P. Patrinos (2009). "Genetic recombination as a major cause of mutagenesis in the human globin gene clusters." Clinical biochemistry **42**(18): 1839-1850.

Borgna-Pignatti, C., M. Cappellini, P. Stefano, G. Vecchio, G. Forni, M. Gamberini, R. Ghilardi, R. Origa, A. Piga and M. Romeo (2005). "Survival and complications in thalassemia." Annals of the New York Academy of Sciences **1054**(1): 40-47.

Boulton, S. J. and S. P. Jackson (1996). "Identification of a Saccharomyces cerevisiae Ku80 homologue: roles in DNA double strand break rejoining and in telomeric maintenance." Nucleic acids research **24**(23): 4639-4648.

Bouva, M. J., C. L. Harteveld, P. van Delft and P. C. Giordano (2006). "Known and new delta globin gene mutations and their diagnostic significance." haematologica **91**(1): 129-132.

Bunn, H. F. (1997). "Pathogenesis and Treatment of Sickle Cell Disease." New England Journal of Medicine **337**(11): 762-769.

Cao, A. and R. Galanello (2010). "Beta-thalassemia." Genetics in Medicine **12**(2): 61-76.

Cao A, G. R. (2000). "Beta-Thalassaemia." GeneReviews, from http://www.ncbi.nlm.nih.gov/books/NBK1426/.

Chakravarti, A., K. Buetow, S. Antonarakis, P. Waber, C. Boehm and H. Kazazian (1984). "Nonuniform recombination within the human beta-globin gene cluster." American journal of human genetics **36**(6): 1239.

Cherf, G. M., K. R. Lieberman, H. Rashid, C. E. Lam, K. Karplus and M. Akeson (2012). "Automated forward and reverse ratcheting of DNA in a nanopore at 5-A precision." Nature biotechnology **30**(4): 344-348.

Choi, O.-R. B. and J. D. Engel (1988). "Developmental regulation of β-globin gene switching." Cell **55**(1): 17-26.

Chui, D. H., S. Fucharoen and V. Chan (2003). "Hemoglobin H disease: not necessarily a benign disorder." Blood **101**(3): 791-800.

Chui, D. H. and J. S. Waye (1998). "Hydrops fetalis caused by α-thalassemia: an emerging health care problem." Blood **91**(7): 2213-2222.

Clark, B. E., Thein, S.L. (2004). "Molecular Diagnosis of Haemoglobin Disorders." Clinical and Laboratory Haematology **26**: 159-176.

Clegg, J. B. and D. J. Weatherall (1967). "Haemoglobin Synthesis in [alpha]-Thalassaemia (Haemoglobin H Disease)." Nature **215**(5107): 1241-1243.

Conesa, A., S. Götz, J. M. García-Gómez, J. Terol, M. Talón and M. Robles (2005). "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research." Bioinformatics **21**(18): 3674-3676.

Craig, J. (1996). "Dissecting the loci controlling fetal haemoglobin production on chromosomes 11p and 6q by the regressive approach." Nature Genetics: 1061-4036.

Curtin, P., M. Pirastu, Y. W. Kan, J. A. Gobert-Jones, A. D. Stephens and H. Lehmann (1985). "A distant gene deletion affects beta-globin gene function in an atypical gamma delta beta-thalassemia." J Clin Invest **76**(4): 1554-1558.

Cyrklaff, M., C. P. Sanchez, N. Kilian, C. Bisseye, J. Simpore, F. Frischknecht and M. Lanzer (2011). "Hemoglobins S and C interfere with actin remodeling in Plasmodium falciparum–infected erythrocytes." Science **334**(6060): 1283-1286.

Czelusniak J, G. M., Hewett-Emmett D, Weiss ML, Venta PJ, Tashian RE (1982). "Phylogenetic origins and adaptive evolution of avian and mammalian haemoglobin genes." Nature **298**(297).

Donnall Thomas, E., J. E. Sanders, C. D. Buckner, T. Papayannopoulou, C. Borgna-Pignatti, P. De Stefano, R. Clift, K. Sullivan and R. Storb (1982). "Marrow transplantation for thalassaemia." The Lancet **320**(8292): 227-229.

Emmanuel, B. S., Shaikh, T.H. (2001). "Segmental Duplications: An Expanding Role in Genomic Instability and Disease." Nature Reviews Genetics **2**: 791-801.

Fenstermacher, D. (2005). "Introduction to bioinformatics." Journal of the American Society for Information Science and Technology **56**(5): 440-446.

Firth, H. V., S. M. Richards, A. P. Bevan, S. Clayton, M. Corpas, D. Rajan, S. V. Vooren, Y. Moreau, R. M. Pettett and N. P. Carter (2009). "DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources." The American Journal of Human Genetics **84**(4): 524-533.

Fischel-Ghodsian, N. (1987). Long range genome structure around the human alpha-globin complex analysed byPFGE. R. D. Nicholls. Nucleic acids research**:** 6197-6207.

Fischel-Ghodsian, N., R. D. Nicholls and D. R. Higgs (1987). "Unusual features of CpG-rich (HTF) Islands in the human α globin complex: association with non-functional pseudogenes and presence within the 3′ portion of the ζ gene." Nucleic Acids Research **15**(22): 9215-9225.

França, L. T., E. Carrilho and T. B. Kist (2002). "A review of DNA sequencing techniques." Quarterly reviews of biophysics **35**(02): 169-200.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly and D. Altshuler (2002). "The Structure of Haplotype Blocks in the Human Genome." Science **296**(5576): 2225-2229.

Galanello, R. and R. Origa (2010). "Review: Beta-thalassemia." Orphanet J Rare Dis **5**(11).

Gao, F., H. Luo and C.-T. Zhang (2012). "DeOri: a database of eukaryotic DNA replication origins." Bioinformatics **28**(11): 1551-1552.

Gaston, M. H., J. I. Verter, G. Woods, C. Pegelow, J. Kelleher, G. Presbury, H. Zarkowsky, E. Vichinsky, R. Iyer and J. S. Lobel (1989). "Prophylaxis with oral penicillin in children with sickle cell anemia: A randomized trial." Pediatrics **83**(5): 835-835.

Gibbons, R., G. Suthers, A. Wilkie, V. Buckle and D. Higgs (1992). "X-linked alpha-thalassemia/mental retardation (ATR-X) syndrome: localization to Xq12-q21.31 by X inactivation and linkage analysis." American Journal of Human Genetics **51**(5): 1136-1149.

Gibbons, R. J., T. Wada, C. A. Fisher, N. Malik, M. J. Mitson, D. P. Steensma, A. Fryer, D. R. Goudie, I. D. Krantz and J. Traeger-Synodinos (2008). "Mutations in the chromatin-associated protein ATRX." Hum Mutat **29**(6): 796-802.

Gu, W., Zhang, F., Lupski, J.R. (2008). "Mechanisms for Human Genomic Rearrangements." Pathogenetics.

Hardies, S., M. Edgell and C. Hutchison (1984). "Evolution of the mammalian beta-globin gene cluster." Journal of Biological Chemistry **259**(6): 3748-3756.

Hardison, R. C. (2012). "Evolution of Hemoglobin and its Genes." Cold Spring Harb Perspect Med **2**(12).

Harismendy, O., P. C. Ng, R. L. Strausberg, X. Wang, T. B. Stockwell, K. Y. Beeson, N. J. Schork, S. S. Murray, E. J. Topol and S. Levy (2009). "Evaluation of next generation sequencing platforms for population targeted sequencing studies." Genome biol **10**(3): R32.

Harteveld, C. and D. Higgs (2010). "alpha-thalassaemia." Orphanet Journal of Rare Diseases **5**(1): 13.

Harteveld, C., C. Refaldi, E. Cassinerio, M. Cappellini and P. Giordano (2008). "Segmental duplications involving the α-globin gene cluster are causing β-thalassemia intermedia phenotypes in β-thalassemia heterozygous patients." Blood Cells, Molecules, and Diseases **40**(3): 312-316.

Harteveld, C. L., C. Refaldi, A. Giambona, C. A. L. Ruivenkamp, M. J. V. Hoffer, J. Pijpe, P. De Knijff, C. Borgna-Pignatti, A. Maggio, M. D. Cappellini and P. C. Giordano (2013). "Mosaic segmental uniparental isodisomy and progressive clonal selection: a common mechanism of late onset β-thalassemia major." Haematologica **98**(5): 691-695.

Hatton, C., A. Wilkie, H. Drysdale, W. Wood, M. Vickers, J. Sharpe, H. Ayyub, I. Pretorius, V. Buckle and D. Higgs (1990). "Alpha-thalassemia caused by a large (62 Kb) deletion upstream of the human alpha globin gene cluster." Blood **76**(1): 221-227.

Heger, A., C. Webber, M. Goodson, C. P. Ponting and G. Lunter (2013). "GAT: a simulation framework for testing the association of genomic intervals." Bioinformatics **29**(16): 2046-2048.

Henthorn, P. S., O. Smithies and D. L. Mager (1990). "Molecular analysis of deletions in the human β-globin gene cluster: deletion junctions and locations of breakpoints." Genomics **6**(2): 226-237.

Henthorn, P. S., O. Smithies and D. L. Mager (1990). "Molecular analysis of deletions in the human beta-globin gene cluster: deletion junctions and locations of breakpoints." Genomics **6**(2): 226-237.

Heyer, W. D., K. T. Ehmsen and J. Liu (2010). "Regulation of homologous recombination in eukaryotes." Annu Rev Genet **44**: 113-139.

Huehns, E. and E. M. Shooter (1965). "Human haemoglobins." Journal of medical genetics **2**(1): 48.

Humphries RK, L. T., Turner P, Moulton AD, Nienhuis AW (1982). "Differences in human a-, 8- and &-globin gene expression in monkey kidney cells." Cell **30**(173).

Hutchison, C. A. (2007). "DNA sequencing: bench to bedside and beyond." Nucleic Acids Research **35**(18): 6227-6237.

Ingram, V. M. (1957). "Gene mutations in human hæmoglobin: The chemical difference between normal and sickle cell hæmoglobin." Nature **180**(4581): 326-328.

Jennifer A. Lee, C. M. B. C., James R. Lupski (2007). "A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders." Cell **131**: 1235-1247.

Katti, M. V., P. K. Ranjekar and V. S. Gupta (2001). "Differential Distribution of Simple Sequence Repeats in Eukaryotic Genome Sequences." Molecular Biology and Evolution **18**(7): 1161-1167.

Kihm, A. J., Y. Kong, W. Hong, J. E. Russell, S. Rouda, K. Adachi, M. C. Simon, G. A. Blobel and M. J. Weiss (2002). "An abundant erythroid protein that stabilizes free α-haemoglobin." Nature **417**(6890): 758-763.

Konings, D., P. Hogeweg and B. Hesper (1987). "Evolution of the primary and secondary structures of the E1a mRNAs of the adenovirus." Molecular biology and evolution **4**(3): 300-314.

Kosche K, D. C., Bank A (1984). "The role of intervening sequences (IVS) in human p globin gene expression." Blood **64**(58a).

Krawetz, S. A. and D. D. Womble (2003). Introduction to bioinformatics: a theoretical and practical approach, Springer Science & Business Media.

Lam, E. Y. N., D. Beraldi, D. Tannahill and S. Balasubramanian (2013). "G-quadruplex structures are stable and detectable in human genomic DNA." Nat Commun **4**: 1796.

Lam, K.-W. G. and A. J. Jeffreys (2006). "Processes of copy-number change in human DNA: the dynamics of α-globin gene deletion." Proceedings of the National Academy of Sciences **103**(24): 8921-8927.

Law, H.-Y., H.-Y. Luo, W. Wang, J. Ho, H. Najmabadi, I. Ng, M. H. Steinberg, D. Chui and S. S. Chong (2006). "Determining the cause of patchwork HBA1 and HBA2 genes: recurrent gene conversion or crossing over fixation events." haematologica **91**(3): 297-302.

Layer, R. M., C. Chiang, A. R. Quinlan and I. M. Hall (2014). "LUMPY: a probabilistic framework for structural variant discovery." Genome biology **15**(6): R84.

Ledbetter, D. H. (1992). "Minireview: cryptic translocations and telomere integrity." American journal of human genetics **51**(3): 451.

Lee, S. (2014). Single-Strand Annealing. Molecular Life Sciences. E. Bell, Springer New York**:** 1-4.

Lehmann, H. (1970). "DIFFERENT TYPES OF ALPHA-THALASSÆMIA AND SIGNIFICANCE OF HÆMOGLOBIN BART'S IN NEONATES." The Lancet **296**(7663): 78-80.

Li, R., C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, K. Kristiansen and J. Wang (2009). "SOAP2: an improved ultrafast tool for short read alignment." Bioinformatics **25**(15): 1966-1967.

Liebhaber, S. A., F. E. Cash, et al. (1986). "Human alpha-globin gene expression. The dominant role of the alpha 2-locus in mRNA and protein synthesis." J Biol Chem **261**(32): 15327-15333.

Liu, P., M. Lacaria, F. Zhang, M. Withers, P. J. Hastings and James R. Lupski (2011). "Frequency of Nonallelic Homologous Recombination Is Correlated with Length of Homology: Evidence that Ectopic Synapsis Precedes Ectopic Crossing-Over." American Journal of Human Genetics **89**(4): 580-588.

Lo, Y. M. D., N. M. Hjelm, C. Fidler, I. L. Sargent, M. F. Murphy, P. F. Chamberlain, P. M. K. Poon, C. W. G. Redman and J. S. Wainscoat (1998). "Prenatal Diagnosis of Fetal RhD Status by Molecular Analysis of Maternal Plasma." New England Journal of Medicine **339**(24): 1734-1738.

Lo, Y. M. D., M. S. C. Tein, T. K. Lau, C. J. Haines, T. N. Leung, P. M. K. Poon, J. S. Wainscoat, P. J. Johnson, A. M. Z. Chang and N. M. Hjelm (1998). "Quantitative Analysis of Fetal DNA in Maternal Plasma and Serum: Implications for Noninvasive Prenatal Diagnosis." The American Journal of Human Genetics **62**(4): 768-775.

Lovett, S. T. (2004). "Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences." Molecular Microbiology **52**(5): 1243-1253.

Lugtenberg, D., A. P. de Brouwer, A. R. Oudakker, R. Pfundt, B. C. Hamel, H. van Bokhoven and E. M. Bongers (2009). "Xq13.2q21.1 duplication encompassing the ATRX gene in a man with mental retardation, minor facial and genital anomalies, short stature and broad thorax." Am J Med Genet A **149a**(4): 760-766.

Lupski, J. R. (1998). "Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits." Trends in Genetics **14**(10): 417-422.

Lupski, J. R. (2004). "Hotspots of homologous recombination in the human genome: not all homologous sequences are equal." Genome biology **5**: /2004/2005/2010/2242-/2004/2005/2010/2242.

Madoui, M.-A., S. Engelen, C. Cruaud, C. Belser, L. Bertrand, A. Alberti, A. Lemainque, P. Wincker and J.-M. Aury (2015). "Genome assembly using Nanopore-guided long and error-free DNA reads." BMC genomics **16**(1): 327.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel and M. Daly (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome research **20**(9): 1297-1303.

McVey, M. and S. E. Lee (2008). "MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings." Trends in Genetics **24**(11): 529-538.

Mendilcioglu, I., S. Yakut, I. Keser, M. Simsek, A. Yesilipek, G. Bagci and G. Luleci (2011). "Prenatal diagnosis of β-thalassemia and other hemoglobinopathies in southwestern Turkey." Hemoglobin **35**(1): 47-55.

Menzel, S., C. Garner, I. Gut, F. Matsuda, M. Yamaguchi, S. Heath, M. Foglio, D. Zelenika, A. Boland, H. Rooks, S. Best, T. D. Spector, M. Farrall, M. Lathrop and S. L. Thein (2007). "A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15." Nat Genet **39**(10): 1197-1199.

Michel, B. (2000). "Replication Fork Arrest and DNA Recombination." TIBS Reviews **25**: 173-179.

Michelson, A. and S. Orkin (1983). "Boundaries of gene conversion within the duplicated human alpha-globin genes. Concerted evolution by segmental recombination." Journal of Biological Chemistry **258**(24): 15245-15254.

Mills, G. B. (2012). "An emerging toolkit for targeted cancer therapies." Genome Research **22**(2): 177-182.

Mills, R. E., K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye and R. K. Cheetham (2011). "Mapping copy number variation by population-scale genome sequencing." Nature **470**(7332): 59-65.

Milne, R. (1991). "Penicillin prophylaxis in children with sickle cell disease." BMJ: British Medical Journal **302**(6789): 1402.

Modell, B., Darlison, M., (2008). "Global epidemiology of haemoglobin disorders and derived service indicators." Bulletin of the World Health Organization **86**: 417-496.

Nagamani, S. C. S., A. Erez, P. Bader, S. R. Lalani, D. A. Scott, F. Scaglia, S. E. Plon, C.-H. Tsai, T. Reimschisel, E. Roeder, A. D. Malphrus, P. A. Eng, P. M. Hixson, S.-H. L. Kang, P. Stankiewicz, A. Patel and S. W. Cheung (2011). "Phenotypic manifestations of copy number variation in chromosome 16p13.11." Eur J Hum Genet **19**(3): 280-286.

Najmabadi, H., R. Karimi-Nejad, S. Sahebjam, F. Pourfarzad, S. Teimourian, F. Sahebjam, N. Amirizadeh and M. H. Karimi-Nejad (2001). "THE β-THALASSEMIA MUTATION SPECTRUM IN THE IRANIAN POPULATION." Hemoglobin **25**(3): 285-296.

Nicholls, R., N. Fischel-Ghodsian and D. Higgs (1987). "Recombination at the human a-globin gene cluster: sequence features and topological constraints." Cell **49**: 369-378.

Nielsen, R., J. S. Paul, A. Albrechtsen and Y. S. Song (2011). "Genotype and SNP calling from next-generation sequencing data." Nat Rev Genet **12**(6): 443-451.

Nobrega, M. A., Y. Zhu, I. Plajzer-Frick, V. Afzal and E. M. Rubin (2004). "Megabase deletions of gene deserts result in viable mice." Nature **431**(7011): 988-993.

Pai, V. B. and M. C. Nahata (2000). "Duration of penicillin prophylaxis in sickle cell anemia: issues and controversies." Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy **20**(1): 110-117.

Paigen, K. and P. Petkov (2010). "Mammalian recombination hot spots: properties, control and evolution." Nature Reviews Genetics **11**(3): 221-233.

Paigen, K., J. P. Szatkiewicz, K. Sawyer, N. Leahy, E. D. Parvanov, S. H. Ng, J. H. Graber, K. W. Broman and P. M. Petkov (2008). "The recombinational anatomy of a mouse chromosome." PLoS genetics **4**(7): e1000119.

Patrinos, G. P., B. Giardine, C. Riemer, W. Miller, D. H. Chui, N. P. Anagnou, H. Wajcman and R. C. Hardison (2004). "Improvements in the HbVar database of human hemoglobin variants and thalassemia mutations for population and sequence variation studies." Nucleic acids research **32**(suppl 1): D537-D541.

Phylipsen, M., A. Chaibunruang, I. P. Vogelaar, J. R. Balak, R. A. Schaap, Y. Ariyurek, S. Fucharoen, J. T. den Dunnen, P. C. Giordano and E. Bakker (2012). "Fine-tiling array CGH to improve diagnostics for α-and β-thalassemia rearrangements." Human mutation **33**(1): 272-280.

Plagnol, V., J. Curtis, M. Epstein, K. Y. Mok, E. Stebbings, S. Grigoriadou, N. W. Wood, S. Hambleton, S. O. Burns and A. J. Thrasher (2012). "A robust model for read count data in exome sequencing experiments and implications for copy number variant calling." Bioinformatics **28**(21): 2747-2754.

Potocki, L., K.-S. Chen, S.-S. Park, D. E. Osterholm, M. A. Withers, V. Kimonis, A. M. Summers, W. S. Meschino, K. Anyane-Yeboa, C. D. Kashork, L. G. Shaffer and J. R. Lupski (2000). "Molecular mechanism for duplication 17p11.2[mdash] the homologous recombination reciprocal of the Smith-Magenis microdeletion." Nat Genet **24**(1): 84-87.

Quick, J., A. R. Quinlan and N. J. Loman (2014). "A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer." GigaScience **3**(1): 22.

Rausch, T., T. Zichner, A. Schlattl, A. M. Stütz, V. Benes and J. O. Korbel (2012). "DELLY: structural variant discovery by integrated paired-end and split-read analysis." Bioinformatics **28**(18): i333-i339.

Richard, G.-F., A. Kerrest and B. Dujon (2008). "Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes." Microbiology and Molecular Biology Reviews **72**(4): 686-727.

Rosatelli, M., A. Dozy, V. Faa, A. Meloni, R. Sardu, L. Saba, Y. Kan and A. Cao (1992). "Molecular characterisation of beta-thalassemia in the Sardinian population." American journal of human genetics **50**(2): 422.

Roth, D. B. and J. H. Wilson (1986). "Nonhomologous recombination in mammalian cells: role for short sequence homologies in the joining reaction." Molecular and cellular biology **6**(12): 4295-4304.

Samonte, R. V. and E. E. Eichler (2002). "Segmental duplications and the evolution of the primate genome." Nature Reviews Genetics **3**(1): 65-72.

Sancar, A., L. A. Lindsey-Boltz, K. Ünsal-Kaçmaz and S. Linn (2004). "Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints." Annual review of biochemistry **73**(1): 39-85.

Sanger, F., A. R. Coulson, G. Hong, D. Hill and G. d. Petersen (1982). "Nucleotide sequence of bacteriophage λ DNA." Journal of molecular biology **162**(4): 729-773.

Schleif, R. (1993). Genetics and Molecular Biology. Baltimore, USA, John Hopkins University Press.

Sharpe, J., R. Summerhill, P. Vyas, G. Gourdon, D. Higgs and W. Wood (1993). "Role of upstream DNase I hypersensitive sites in the regulation of human α globin gene expression." Blood **82**(5): 1666-1671.

Sheridan, C. (2014). "Erratum: Illumina claims $1,000 genome win." Nature biotechnology **32**(2): 115.

Sherry, S. T., M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski and K. Sirotkin (2001). "dbSNP: the NCBI database of genetic variation." Nucleic acids research **29**(1): 308-311.

Shinar E, R. E. (1990). "Oxidative denaturation of red blood cells in thalassemia." Seminars in Haematology **27**: 70-82.

Shooter, C., H. Rooks, S. L. Thein and B. Clark (2015). "Next Generation Sequencing Identifies a Novel Rearrangement in the HBB Cluster Permitting to-the-Base Characterisation." Human mutation **36**(1): 142-150.

Shooter, Claire, et al. "First Reported Duplication of the Entire Beta Globin Gene Cluster Causing an Unusual Sickle Cell Trait Phenotype." British journal of haematology (2014).

Slack, A., P. Thornton, D. B. Magner, S. M. Rosenberg and P. Hastings (2006). "On the mechanism of gene amplification induced under stress in Escherichia coli." PLoS Genet **2**(4): e48.

Stankiewicz, P. and J. R. Lupski (2002). "Genome architecture, rearrangements and genomic disorders." TRENDS in Genetics **18**(2): 74-82.

Statistics, O. f. N. (2011). 2011 Census: Aggregate data (England and Wales) G. UNIT. London DataStore, http://data.london.gov.uk/dataset/2011-census-key-findings-summaries/resource/6d7062d8-e8b1-4d86-b42d-e123b834afca.

Steensma, D. P., R. J. Gibbons and D. R. Higgs (2005). Acquired α-thalassemia in association with myelodysplastic syndrome and other hematologic malignancies.

Steinberg, M. and J. d. Adams (1991). Hemoglobin A2: origin, evolution, and aftermath.

Steinberg M.H., H. D. R., Nagel R.L., (2001). "Disorders of Haemoglobin, Genetics, Pathophysiology, and Clinical Management." Journal of the Royal Society of Medicine **94**(11): 602-603.

Stuart, M. J. and R. L. Nagel "Sickle-cell disease." The Lancet **364**(9442): 1343-1360.

Taylor, S. M., C. M. Parobek and R. M. Fairhurst (2012). "Haemoglobinopathies and the clinical epidemiology of malaria: a systematic review and meta-analysis." The Lancet Infectious Diseases **12**(6): 457-468.

Thein, S. L. (1992). "Dominant β thalassaemia: molecular basis and pathophysiology." British journal of haematology **80**(3): 273-277.

Thein, S. L. (1993). "6 β-Thalassaemia." Baillière's clinical haematology **6**(1): 151-175.

Thein, S. L., Menzel, S., Lathrop, M., Garner, C. (2009). "Control of fetal hemoglobin: new insights emerging from genomics and clinical implications." Human Molecular Genetics **18**: R216-R223.

Tiller, E. R. M., Collins, R. (2000). "Genome rearrangement by replication-directed translocation." Nature America Letters **26**: 195-198.

Trent, R. J. (2006). "Diagnosis of the haemoglobinopathies." The Clinical biochemist. Reviews / Australian Association of Clinical Biochemists **27**(1): 27-38.

Vanin, E. F., P. S. Henthom, D. Kioussis, F. Grosveld and O. Smithies (1983). "Unexpected relationships between four large deletions in the human β-globin gene cluster." Cell **35**(3): 701-709.

Viprakasit, V., C. L. Harteveld, H. Ayyub, J. S. Stanley, P. C. Giordano, W. G. Wood and D. R. Higgs (2006). "A novel deletion causing α thalassemia clarifies the importance of the major human alpha globin regulatory element." Blood **107**(9): 3811-3812.

Viprakasit, V., A. M. Kidd, H. Ayyub, S. Horsley, J. Hughes and D. R. Higgs (2003). "De novo deletion within the telomeric region flanking the human α globin locus as a cause of α thalassaemia." British journal of haematology **120**(5): 867-875.

Wajcman, H., Moradkhani, K. (2011). "Abnormal haemoglobins: detection & characterisation." Indian J Med Res: 538-546 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3237254/.

Waye, J. S. and D. H. Chui (2001). "The alpha-globin gene cluster: genetics and disorders." Clinical and investigative medicine **24**(2): 103-109.

Weatherall, D. (2001). "Thalassaemias." Encyclopedia of Life Sciences, Nature Publishing Group: 1-3.

Weatherall, D. (2008). "Genetic variation and susceptibility to infection: the red cell and malaria." British journal of haematology **141**(3): 276-286.

Weatherall, D. J. (2008). "Hemoglobinopathies worldwide: present and future." Current molecular medicine **8**(7): 592-599.

Weatherall, D. J. and J. B. Clegg (2001). "Inherited haemoglobin disorders: an increasing global health problem." Bulletin of the World Health Organization **79**: 704-712.

Weischenfeldt, J., O. Symmons, F. Spitz and J. O. Korbel (2013). "Phenotypic impact of genomic structural variation: insights from and for human disease." Nature Reviews Genetics **14**(2): 125-138.

Wheeler, R. (2007). 1GZX Hemoglobin.png. en:pymol. G. Hemoglobin.png. Wikimedia Commons, Zephyris, English Language Wikipedia. **1600 x 1600:** By Richard Wheeler (Zephyris) 2007.

Created with en:pymol from en:PDB enzyme 2001GZX.

en:Category:Protein images.

Wilkie, A., H. Zeitlin, R. Lindenbaum, V. Buckle, N. Fischel-Ghodsian, D. Chui, D. Gardener-Medwin, M. MacGillivray, D. Weatherall and D. Higgs (1990). "Clinical features and molecular analysis of the alpha thalassaemia/mental retardation syndromes. II. Cases without detectable abnormality of the alpha globin gene complex." The American Journal of Human Genetics **46**(6): 1127-1140.

Yu, A. M. and M. McVey (2010). "Synthesis-dependent microhomology-mediated end joining accounts for multiple types of repair junctions." Nucleic Acids Research **38**(17): 5706-5717.

Zhang, J. (2003). "Evolution by gene duplication: an update." Trends in ecology & evolution **18**(6): 292-298.

Appendix 1: URLs

Bioanalyser DNA 1000 Quick Guide, Agilent, USA	http://www.chem.agilent.com/Library/usermanuals/Public/G2938-90015_DNA1000Assay_QSG.pdf
Covaris User Guide, Covaris, USA	http://covarisinc.com/wp-content/uploads/pn_0ten119.pdf
SureSelect - How it Works	http://www.genomics.agilent.com/article.jsp?pageId=3083
eArray	https://earray.chem.agilent.com/
SureDesign	https://earray.chem.agilent.com/suredesign/
SureSelect Sample Preparation protocol (2012 version)	http://www.genome.duke.edu/cores/microarray/services/ngs-library/documents/G7530-90000_SureSelect_IlluminaXTMultiplexed_141.pdf
dbSNP	http://www.ncbi.nlm.nih.gov/SNP/index.html
dbVar	http://www.ncbi.nlm.nih.gov/dbvar
deORI	http://cerevisiae.oridb.org/ Segmental Duplication Database http://humanparalogy.gs.washington.edu/SDD/)
The HapMap project	www.HapMap.org
DFam RepeatMasker Database	http://www.dfam.org/search/hits
Non-B DNA Motif Database	https://nonb-abcc.ncifcrf.gov/apps/site/default
Primer3Plus	http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi/
UCSC Bioinformatics Site	https://genome.ucsc.edu/
Clustal Omega	http://www.ebi.ac.uk/Tools/msa/clustalo/
HbVar (The Globin Gene Server)	http://globin.bx.psu.edu/cgi-bin/hbvar/query_vars3
PipMaker	http://pipmaker.bx.psu.edu/cgi-bin/pipmaker?basic
DNA fragment size selection using SPRIselect beads	http://www.broadinstitute.org/scientific-community/science/platforms/genome-sequencing/broadillumina-genome-analyser-boot-camp

Appendix 2: List of Solutions

1. Lysis Buffer (for 1L)
    - In a 1L autoclaved Duran, combine:
        - 10 ml 10 mM Tris HCL
        - 3.3 ml 10 mM NaCl
        - 20 ml 10 mM EDTA
        - 967 ml pure water
    - Mix until dissolved
2. Proteinase K
    - In a 1.5 ml microcentrifuge tube, combine:
        - 10 ml molecular grade water
        - 500 mg stock of proteinase K
    - Shake to mix and incubate at room temperature for 10 minutes (Unused Proteinase K was frozen in aliquots)
3. 5x ANE Buffer
    - In a falcon tube, combine:
        - 0.05 M NaOAc.3H2O pH6
        - 0.5 M NaCL
        - 0.005 M EDTA
        - 2.5 % SDS
    - Mix until dissolved
4. 20x SSC Buffer (for 500 ml)
    - In a 500 ml autoclaved Duran, combine:
        - 87.65g NaCL

        - 44.1g Na Nitrate
        - 400 ml pure water
    - Place on stirrer until all powder is dissolved
    - Measure pH and adjust to pH7 using 15% HCL as necessary
    - Autoclave
5. 0.5% NP40 (for 500 ml)
    - In a 500 ml autoclaved Duran, combine:
        - 497.5 ml pure water
        - 2.5 ml Nonidet P40
    - Stir until dissolved
    - Autoclave
6. Lysis buffer mix
    - In a 1.5 ml microcentrifuge tube, combine:
        - 1.5 ml Lysis Buffer
        - 50 µL Proteinase K
        - 80 µL 10% SDS
    - Invert tube several times to mix
7. 3M NaCl (for 1L)
    - In a 1L autoclaved Duran, combine:
        - 175.2g NaCl
        - 800ml water
    - Stir until dissolved
    - Autoclave
8. 1x TE (for 1L)

- In a 1L autoclaved Duran, combine:
    - 10 ml Tris-Cl pH 7.6
    - 2 ml EDTA pH 8
    - 988 ml pure water
- Stir until dissolved
- Autoclave

9. 1M Tris-HCL pH 7.6 (for 1L)
    - In a 1L autoclaved Duran, combine:
        - 121.1 g Tris base
        - 800 ml pure water
    - Stir until dissolved
    - Autoclave
10. 0.5M EDTA pH 8 (for 1L)
    - In a 1L autoclaved Duran, combine:
        - 186.1 g disodium EDTA 2H2O
        - 20 g NaOH pellets to bring pH to 8 (and 15% HCL to reduce pH to 8 if it became too high)
        - 800ml pure water
    - Stir until dissolved
    - Autoclave
11. 70% Ethanol (for 1L)
    - In a 1L autoclaved Duran, combine:
        - 300 ml pure water
        - 700 ml absolute ethanol

- Shake vigorously
- Aliquot into 50 ml falcon tubes
- Store at -20 ◦C

12. 1x TE (for 1L)
    - In a 1L autoclaved Duran, combine:
        -

13. 0.5M EDTA pH 8 (for 1L)
    - In a 1L autoclaved Duran, combine:
        - 186.1 g disodium EDTA 2H2O
        - 20 g NaOH pellets to bring pH to 8 (and 15% HCL to reduce pH to 8 if it became too high)
        - 800ml pure water
    - Stir until dissolved
    - Autoclave

14. End Repair Mix for one reaction:
    - Keep reagents and mix on ice until use
    - In a 1.5 ml microcentrifuge tube, combine:
        - 1.6 µL dNTP mix
        - 1 µL T4 DNA polymerase
        - 2 µL Klenow Polymerase
        - .2 µL T4 Polynucleotide Kinase
        - 35.2 µL Nuclease-free water
    - Vortex briefly to mix

15. Adenylation Mix for one reaction:

- In a 1.5 ml microcentrifuge tube, combine:
    - 5 µL 10 x Klenow Polymerase Buffer
    - 1 µL dATP
    - 3 µL Exo(-) Klenow
    - 11 µL Nuclease-free water

16. Adapter Ligation Mix for one reaction:
    - In a 1.5 ml microcentrifuge tube, combine:
        - 10 µL 5x T4 Ligase Buffer
        - 10 µL SureSelect Adaptor Oligo Mix
        - 1.5 µL T4 DNA Ligase
        - 15.5 µL Nuclease-Free water
    - Vortex briefly to mix
17. Pre-Hybridization PCR mix for one sample:
    - Keep reagents and mix on ice until use
    - In a 1.5 ml microcentrifuge tube, combine:
        - 1.25 µL SureSelect Primer
        - 1.25 µL SureSelect ILM Indexing Pre Capture PCR Reverse Primer
        - 10 µL 5x Herculase II Rxn Buffer
        - 0.5 µL 100 mM dNTP Mix
        - 1 µL Herculase II Fusion DNA Polymerase
    - Up to >21 µL Nuclease-free water to make a final volume of 50 µL after addition of 250ng DNA sample)
    - Vortex briefly to mix

    NB: *Herculase reagents and 100mM dNTPs are from a separate kit. No other dNTPs must be used with the Herculase reagents*

18. Hybridization Buffer Mix for one reaction:
    - Keep reagents and mix on ice until use
    - In a 1.5 ml microcentrifuge tube, combine:
        - 25 µL SureSelect Hyb #1
        - 1 µL SureSelect Hyb #2
        - 10 µL SureSelect Hyb #3
        - 13 µL SureSelect Hyb #4
    - Vortex briefly to mix
19. SureSelect Capture Library Mix:
    - Keep reagents and mix on ice until use
    - Different mixes are made for libraries greater than or equal to, or less than, 3Mb in size. Library design 1 was > than 3Mb and library design 2 was <3Mb. The appropriate mixes were made for both.
    - For a <3Mb Library, in a 1.5 ml microcentrifuge tube, combine:
        - 2 µL of capture library
        - 5 µL of RNase block dilution (10% block, 90% nuclease-free water)
    - For a ≥ 3Mb Library, in a 1.5 ml microcentrifuge tube, combine:
        - 5 µL capture library
        - 2 µL RNase block dilution (25% RNase block 75% Nuclease-free water).
    - Vortex briefly to mix
20. SureSelect Block Mix:
    - Keep reagents and mix on ice until use
    - In a 1.5 ml microcentrifuge tube, combine:
        - 2.5 µL SureSelect Indexing Block #1

        - 2.5 µL SureSelect Indexing Block #2
        - 0.6 µL SureSelect Indexing Block #3
    - Vortex briefly to mix
21. Indexing PCR mix:
    - Keep reagents and mix on ice until use
    - In a 1.5 ml microcentrifuge tube, combine:
        - 10 µL 5x Herculase II Rxn Buffer
        - 0.5 µL 100 mM dNTP mix
        - 1 µL Herculase II Fusion DNA Polymerase
        - 1 µL SureSelect ILM Indexing Post Capture Forward PCR Primer
        - 22.5 µL Nuclease-free water
    - Vortex briefly to mix
22. Wash Solution for HiSeq Instrument
    - In a 5L carboy, combine:
        - 750 ml laboratory grade water
        - 250 ml 10% tween
        - 1.5 ml ProClin 300
23. Dye-terminator sequencing PCR mix (for one reaction):
    - In a 1.5 ml microcentrifuge tube, combine:
        - 0.5 µL Dye-terminator ready mix
        - 0.5 µL Dye-terminator buffer
        - 3 µL nuclease free water
        - 4 µL template DNA for sequencing (approx 100ng/µL)
        - 1 µL 5pM PCR primer

- Pulse vortex to mix
- Centrifuge for one minute at 280 x g

24. Generic Gap PCR reaction mix (for one reaction):
    - In a 1.5 ml microcentrifuge tube, combine:
        - 10 µL Q Multiplex Mix (Qiagen)
        - 1 µL 10M forward primer
        - 1 µL 10M reverse primer
        - 1-8 µL Template DNA (50-500 ng DNA)
        - Nuclease free water to bring total reaction volume to 20 µL
    - Pulse vortex to mix
    - Centrifuge for one minute at 280 x g
25. Generic long amplicon Gap PCR reaction mix (for one reaction)
    - In a 1.5 ml microcentrifuge tube, combine:
        - LongAmp Taq Reaction Buffer (NEB, USA)
        - 1 µL 10mM forward/ reverse primer
        - 0.75 µL dNTPs
        - 1-2 µL 30-200ng/ µL template DNA
        - nuclease-free water to make a total volume of 25 µL
    - Pulse vortex to mix
    - Centrifuge for one minute at 280 x g

Appendix 3: Cluster Generation and Sequencing on the HiSeq 2000

Cluster generation and sequencing on the HiSeq 2000 was performed off site at the Department of Haematological Medicine, Rayne Institute, KCL by Dr Alex Smith. One 12 Sample Run was performed. Cluster generation was performed using reagents from the Illumina HiSeq PE Cluster Kit (v2) (Illumina part # PE-401-4001). Sequencing was performed using the HiSeq PE Sequencing Kit (v2) (Illumina # FC401-4001).Tris-Cl and NaOH were obtained from generic laboratory suppliers. The steps below are in accordance with the current version of the protocols supplied by Illumina.

Cluster Generation

Preparing the DNA:

Index-tagged DNA samples were pooled at 2nM. Reagent HT1 from the HiSeq Sequencing Kit was thawed and stored on ice until use.

The pooled library concentration was normalised to 2 nM using Tris-HCL 10 mM (pH 8.5) with 0.1% Tween. 10 µL of the pooled library was combined with 10 µL freshly made 0.1 N NaOH. The mix was briefly vortexed and centrifuged at 280 x g for one minute. The mix was incubated at room temperature for 5 minutes. 980 µL of chilled HT1 reagent was added to the mix which was then placed on ice. The library was diluted to 12 pM by adding 600 µL of the library to 400 µL HT1.

Preparing the PhiX control:

A PhiX control was prepared. Two µL 10 nM PhiX library was combined with 8 µL 10 mM Tris-Cl (pH 8.5) with 0.1% Tween 20. The mix was briefly centrifuged and 10 µL of 0.1 N NaOH was added to the tube. The tube was vortexed briefly, spun for one minute at 280 x g and incubated for 5 minutes at room temperature. 980 µL of HT1 were added to the Phi-X control. The Phi-X was diluted to 12 pM by combining 600 µL of the diluted Phi-X with 400 µL HT1 and kept on ice.

Spiking the DNA pool with Phi-X:

Ten µL of the Phi-X Control was added to 990 µL of the DNA library and mixed by pipetting.

Preparing clustering reagents:

The cBot reagent plate was thawed in a room temperature water bath containing deionised water. Once thawed, the plate was inverted 5 times to mix, vortexed for 10 seconds and then tapped firmly on a hard surface to collect all reagent droplets at the bottoms of the wells. The tubes were inspected for a secure seal and absence of air bubbles.

Clustering on the cBot:

The cBot wash reservoir was filled with 12 ml deionized water, closed and set to "**wash**" on the cBot software interface. Once the wash was complete, the reservoir was dried with lint free tissue.

The reagent plate was placed in the cBot and foil was removed from the wells in row 10 of the plate. The flow cell clamp was lifted and the adapter plate was washed with deionized water, and then wiped dry with lint free tissue. The flow cell was removed from storage using plastic forceps, rinsed using deionized water and dried with lint free tissue. The flow cell was then positioned on the thermal stage with the flow cell port holes facing upwards. The manifold was inspected to ensure no parts had been damaged and positioned over the flow cell, with the guide pins aligned on the thermal stage. The flow cell clamp was closed. The outlet manifold was connected to the wash reservoir and snapped shut. The run was initiated on the cBot screen and a pre-run check was performed by the machine.

Upon completion of the run, "**Unload**" was selected on the cBot screen. The manifold and wash reservoir were disconnected from one another and the flow cell and reagent plate were retrieved. The wash reservoir was then filled with 12ml deionized water and "**Wash**" was selected on the cBot.

Sequencing on the HiSeq 2000

Preparing Reagents for Sequencing Read 1:

Reagents from the SBS reagent kit were removed from cold storage and thawed. The SRE and CMR reagents were thawed in a room temperature waterbath for one hour. Each reagent was inverted five times to mix and then placed on ice until use. Reagent LFN36 was thawed in a beaker containing room temperature deionized water for 20 minutes and placed on ice until use.

Preparing ICB: ICB mix was prepared using reagents from the SBS reagent kit: One tube of LFN36 was added to one bottle of ICB-50. One tube of EDP-50 was then added to this mix. The bottle was then gently mixed several times by inversion, and placed on ice.

Reagents HP3, HT2 and HP8 were removed from cold storage and thawed in a beaker filled with room temperature deionized water for 20 minutes until thawed completely. Reagent HT2 was then inverted five times to mix and centrifuged at 280 x g for one minute. HP3 was inverted five times to mix and then pulse centrifuged. 325 µL PW1 was transferred into a 15 ml Sarstedt conical tube. 175 µL of HP3 was added to the tube, which was then inverted five times to mix, centrifuged at 280 x g for one minute and then set aside at room temperature. Reagent HP8 was inverted five times to mix, centrifuged at 280 x g for one minute and then set aside at room temperature.

Run parameters were then entered onto the Instrument. The appropriate storage route was selected, the type and ID of the flow cell were input and the experiment was named. On the Advanced screen, "**Confirm First Base**" and "**Align to PhiX**" were selected. One the Recipe screen, "**Single Index**", "**96 cycles**" and the "**Paired end cluster kit**" settings were selected.

The ID numbers of the reagents were scanned and "**prime reagents**" was selected. The reagent compartment of the HiSeq was then opened, sippers were removed from the reagent rack and the rack was retrieved. The SBS reagent mix was fitted with a funnel cap. Reagents directly from the kit, and reagent mixes described above were placed in the correct positions on the reagent rack. 25 ml of PW1 was added to the bottle in reagent position 2. The reagent rack replaced and the sippers were repositioned. Indexing reagents were loaded into the indexing reagents rack in the same manner. Any unused rack positions were filled with 15 ml conical tubes containing 10 ml laboratory grade water and the rack was replaced.

The flow cell holder was cleaned with a lint free tissue and a priming (used) flow cell was washed and placed inside. The vacuum was engaged and the flow cell lever was moved to position one. After five seconds, the flow cell lever was slowly moved to position two. A green light indicates that the flow cell is ready for use. The "**Vacuum engaged**" checkbox was checked.

Proper flow was then confirmed: "**Solution 2**" was selected from the drop down list on the HiSeq 2000 prompt screen. "**Pump**" was selected and the flow cell was inspected for bubbles or leaks. If no problems were found, position tubing and start prime was

initiated: waste tubing was placed in empty 15 ml tubes and placed in the waste container. “**Start prime**” was then selected from the on screen prompt. Once the step was complete, the volume of waste collected in each tube was checked to be > 1.75 mls.

The flow cell lever was returned to position 0 and disengaged from the vacuum. The used priming flow cell was removed from the flow cell holder, which was then cleaned with a lint free tissue. The run flow cell was inserted into the flow cell holder. The flow cell lever was slowly moved to position 1, and then to position 2. A green light indicates that the manifold are in position. Flow was confirmed as before, and “**vacuum engaged**” and “**door closed**” were selected on the instrument screen. The sequencing run was initiated by selecting “**start**”.

Preparing Reagents for Sequencing Read 2

Thaw reagents LMX2, BMX, AMX2, APM2, AT2, HP3, HT2, HP7 and RMR from the SBS kit in a beaker filled with room temperature deionized water for 20 minutes. Reagents RMR, LMX2, BMX and AMX2 were placed on ice. Reagents LMX2, BMC and AMX2 were inverted five times and centrifuged at 280 x g for one minute, and then set aside on ice. AMP2, AT2, HP3, HT2 and HP7 reagents were inverted five times to mix, centrifuged at 280 x g for one minute and set aside at room temperature.

Reagent HP3 was inverted five times to mix and pulse centrifuged. 2.85 ml of PW1 was added to a 15 ml Sarstedt conical tube followed by 150 µL HP3. The tube was inverted five times to mix, centrifuged at 280 x g for one minute and set aside at room temperature.

Reagent RMR was inverted five times to mix, centrifuge at 280 x g for one minute and set aside on ice.

Fresh ICB was prepared as described for Read 1.

Once Read 1 and index reading had completed, The sippers were removed from the reagent rack. The reagent tubes were placed at their labelled positions with the lids removed. The reagent compartment was replaced and the sippers lowered into the sippers into the reagent tubes.

Select “**Next**” on the instrument to continue the run.

Washing the Instrument

Five litres of Maintenance Wash Solution was prepared: 250 ml 10% tween was made (See Appendix: List of Reagents). In a 5 L carboy, combine (adding water first): 750 mL laboratory grade water, 250 ml 10% tween, 1.5 ml ProClin 300. Place on a stir plate until mixed.

"**Wash | Maintenance**" was selected on the instrument. Select "**Yes**" to wash PE reagent positions. Fill 8 SBS bottles with 250 ml maintenance wash solution and 10 PE tubes with 12 mls PE wash solution. Load bottles and tubes into the reagent racks in their assigned positions. Check the "Wash solution loaded" and "template loading station closed" checkboxes on the screen and select "Next".

Remove the flow cell, replace the front and rear gaskets and then reload the same flow cell. Check "Vacuum engaged" and select "Next".

Perform a fluids check as described above for Read 1.

Remove the eight waste tubes from the waste container and bundle the tubes with parafilm to direct them into a single 250 ml bottle. Select "Next" to start the wash. When the wash is complete, select "Return to start" and unwrap the waste container tubes from the parafilm.

Appendix 4: Phenol-Chloroform DNA Purification

The following procedure was performed by Mrs Helen Rooks, Department of Molecular Haematology, King's College London.

DNA samples obtained from the Molecular Haematology group at King's College London were extracted using the phenol chloroform method.

The following solutions (compositions of which are listed in Appendix 1: List of Solutions) were prepared in advance: Lysis Buffer, Proteinase K, 5x ANE Buffer, 20x SSC Buffer, 0.5% NP40, 3M NaCl, 1x TE, 1M Tris-HCL, 0.5M EDTA pH8, 70% Ethanol.

3-10 ml of whole blood was equilibrated to room temperature in a 15 ml Falcon tube by leaving them to stand for 30 minutes on the bench. The Falcon tubes were centrifuged briefly to remove liquid from the lid of tubes before making the volume up to 10 ml with 1x SSC and centrifugation (make model) at 3000 rpm for 10 minutes at 4°C. The top layer of blood was aspirated and discarded, leaving a pellet remaining in the bottom of the Falcon tube. 2 ml 0.5% NP40 was added to the tube. The pellet was broken up with a pastette and vortex briefly. The Falcon tube was centrifuged at 2000 rpm (930 x g)for 5 minutes at 4°C. The supernatant was poured away carefully, so as not to dislodge the pellet. Lysis buffer mix was added to the Falcon tube, vortexed briefly, and placed in a 37°C water bath to incubate overnight (16 hours). Lysis was complete when the pellet had disappeared and the liquid was an amber transparent colour.

Following the incubation, the Falcon tube was centrifuged briefly prior to opening. 5x ANE Buffer was equilibrated to 37°C in the water bath and 0.2x lysis volume of 5x ANE was added to the sample, followed by a 1:1 volume of phenol. The Falcon tube was inverted to mix and then centrifuged at 2110 x g for 10 minutes at room temperature. The aqueous, DNA containing, phase lay on top and was carefully transferred to a new Falcon tube with a pastette. The total volume of the aqueous phase was measured and 0.5x that volume of phenol and of chloroform was added to the sample. The Falcon tube was inverted to mix and centrifuged at 3000 rpm for 10 minutes at room temperature. The aqueous DNA was transferred into a new 15 ml Falcon tube. 10 % of the total volume of the tube of 3 M NaCl was added. The mixture was made up to 11 ml with ice cold absolute ethanol (100 %) and inverted to mix. The mix was incubated at -20°C for a minimum of 30 minutes to aid DNA precipitation.

The Falcon tube was centrifuged at 2110 x g for 10 minutes at 4°C. The ethanol was carefully poured off the pellet and washed with 5 ml 70% ethanol and again centrifuged at 3000 rpm for 10 minutes at 4°C. The pellet was washed a further 2 times before removing all ethanol and placing the Falcon tube in the fume hood to completely dry the pellet. The pellet was re-suspended in 500 µL TE buffer and left overnight on a turning roller. After 12-24 hours the DNA was inspected for full re-suspension and even viscosity. 5 µL of sample was diluted in 45 µL molecular grade water. The diluted sample was vortexed and left at room temperature for one hour. The concentration of the sample was recorded and it was stored at -20°C.

Appendix 5: Python Script for Identifying Known Rearrangements in FASTA Data

```
#open user specified file
file = input("enter file path (e.g. F:\Sequencing\File.FASTA)")

# check for sickle
Wildtype_sc_count = 0
Sickle_count = 0
lines = 0
with open(file) as handler:
    for string in handler:
        if "AACGGCAGACTTCTCCACAGGAGTCAG" in string:
            Sickle_count += 1
        elif "AACGGCAGACTTCTCCTCAGGAGTCAG" in string:
            Wildtype_sc_count += 1
        else:
            lines += 1

if Wildtype_sc_count == 0:
    print('no wildtype found, check alignment')
else:
    if Sickle_count >= 3 and Wildtype_sc_count >= 3:
        print('Heterozygous for Sickle')
    elif Sickle_count <=3 and Wildtype_sc_count >= 3:
        print('Negative for Sickle')
    elif Sickle_count >=3 and Wildtype_sc_count <= 3:
        print('Homozygous for Sickle')
    else:
        print('neither sequence found')

print('wt(sickle)', Wildtype_sc_count)
print('Sickle', Sickle_count)
print('lines searched', lines)

#check for Asian Indian inversion deletion
Wildtype_AI_count = 0
AI_Indel_count = 0
lines = 0
with open(file) as handler:
    for string in handler:
        if
"CAGAGGACTAACTGGGCTGAGACCAGTTGTCCAAAGTTGCGGGCCAGCACAC" in
string:
            AI_Indel_count += 1
        elif
"GGTGAATTCCTTGCCAAAGTTGCGGGCCAGCACACACACCAGCACATTGC" in
string:
            Wildtype_AI_count += 1
        else:
            lines += 1

if Wildtype_AI_count == 0:
```

```
    print('no wildtype found, check alignment')
else:
    if AI_Indel_count >= 3 and Wildtype_AI_count >= 3:
        print('Heterozygous for Asian Indian Inversion deletion')
    elif AI_Indel_count <=3 and Wildtype_AI_count >= 3:
        print('Negative for Asian Indian Inversion deletion')
    elif AI_Indel_count >=3 and Wildtype_AI_count <= 3:
        print('Homozygous for Asian Indian Inversion deletion')
    else:
        print('neither sequence found')

print('wt (AI)', Wildtype_AI_count)
print('Asian Indian Inversion deletion', AI_Indel_count)
print('lines searched', lines)

#check for english V inversion deletion
Wildtype_EV_count = 0
English_V_count = 0
lines = 0
with open(file) as handler:
    for string in handler:
        if "TAAAGATGAACCCATAGTGAGCTGAGATCCCCACTATATTCTTTGTTCCT"
in string:
            English_V_count += 1
        elif
"CAGAATCAAGCCTATGTTAACTTCCCTCAAAGCCTGAGATTTTGCTTTCCCATTAA
ATGCAGGTAGTTGTTCTTCTTGCAGC" in string:
            Wildtype_EV_count += 1
        else:
            lines += 1

if Wildtype_EV_count == 0:
    print('no wildtype found, check alignment')
else:
    if English_V_count >= 3 and Wildtype_EV_count >= 3:
        print('Heterozygous for English V')
    elif English_V_count <=3 and Wildtype_EV_count >= 3:
        print('Negative for English V')
    elif English_V_count >=3 and Wildtype_EV_count <= 3:
        print('Homozygous for English V')
    else:
        print('neither sequence found')

print('wt(English V)', Wildtype_EV_count)
print('englishV', English_V_count)
print('lines searched', lines)

#check for 619bp deletion
Wildtype_619_count = 0
count_619 = 0
lines = 0
with open(file) as handler:
    for string in handler:
        if "TCTACTTGAATCTCTACTTGTTA" in string:
            count_619 += 1
```

```
        elif "CTTACATCAGTTACAATTTATAT" in string:
            Wildtype_619_count += 1
        else:
            lines += 1

if Wildtype_619_count == 0:
    print('no wildtype found, check alignment')
else:
    if count_619 >= 3 and Wildtype_619_count >= 3:
        print('Heterozygous for 619bp deletion')
    elif count_619 <=3 and Wildtype_619_count >= 3:
        print('Negative for 619bp deletion')
    elif count_619 >=3 and Wildtype_619_count <= 3:
        print('Homozygous for 619bp deletion')
    else:
        print('neither sequence found')
print('wt(619)', Wildtype_619_count)
print('619', count_619)
print('lines searched', lines)
```

METHODS

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

Next Generation Sequencing Identifies a Novel Rearrangement in the *HBB* Cluster Permitting to-the-Base Characterization

Claire Shooter,[1] Helen Rooks,[1] Swee Lay Thein,[1,2]* and Barnaby Clark[1,3]

[1]*King's College London, Faculty of Life Sciences and Medicine, Molecular Haematology, London, UK;* [2]*Department of Haematology, King's College Hospital NHS Foundation Trust, London, UK;* [3]*Department of Molecular Pathology, Viapath at King's College Hospital NHS Foundation Trust, London, UK*

Communicated by Paolo Fortina

Received 24 June 2014; accepted revised manuscript 16 September 2014.
Published online 21 October 2014 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22707

ABSTRACT: Genetic testing for hemoglobinopathies is required for prenatal diagnosis, understanding complex cases where multiple pathogenic variants may be present or investigating cases of unexplained anemia. Characterization of disease causing variants that range from single base changes to large rearrangements may require several different labor-intensive methodologies. Multiplex ligation probe amplification analysis is the current method used to detect indels, but the technique does not characterize the breakpoints or detect balanced translocations. Here, we describe a next-generation sequencing (NGS) method that is able to identify and characterize a novel rearrangement of the *HBB* cluster responsible for $\varepsilon\gamma\delta\beta$ thalassemia in an English family. The structural variant involved a 59.0 kb inversion encompassing *HBG2* exon 3, *HBG1*, *HBD*, *HBB*, and *OR51V1*, juxtaposed by a deletion of 122.6 kb including 82 bp of the inverted sequence, *HBG2* exon 1 and 2, *HBE*, and the β-locus control region. Identification of reads spanning the breakpoints provided to-the-base resolution of the rearrangement, subsequently confirmed by gap-PCR and Sanger sequence analysis. The same rearrangement, termed Inv-Del English V $\varepsilon\gamma\delta\beta$ thalassemia (HbVar 2935), was identified in two other unrelated English individuals with a similar hematological phenotype. Our NGS approach should be applicable as a diagnostic tool for other disorders.
Hum Mutat 36:142–150, 2015. © 2014 Wiley Periodicals, Inc.

KEY WORDS: hemoglobinopathy; hemoglobin; inversion–deletion; $\varepsilon\gamma\delta\beta$ thalassemia; NGS, palindrome

Introduction

Hemoglobinopathies are a diverse group of disorders caused by genetic variants affecting the structure and abundance of the α-like and β-like globin chains that form the subunits of hemoglobin (Hb) [Forget and Bunn, 2013]. Non-synonymous changes result in globin chains with altered amino-acid sequences and give rise to Hb variants. Variants that lead to reduced synthesis of the globin chains cause thalassemia, with α and β thalassemia being the most common forms, characterized by a quantitative deficiency of the α and β globin chains, respectively. The variants underlying both α and β thalassemia are extremely heterogeneous ranging from single base changes to large deletions, and rearrangements of their respective gene clusters [Giardine et al., 2011; Higgs, 2013; Thein, 2013]. Some rare forms of both α and β thalassemia result from deletions of their respective upstream regulatory elements, but leave all of the downstream globin genes unaltered.

A phenotypic primary screen based on a complete blood count, Hb electrophoresis for Hb fractionation and quantification of HbA_2 and HbF, identifies a carrier for a thalassemia variant. DNA testing is required for definitive analysis and is a prerequisite for identifying genetic disease risk and prenatal diagnosis. Hb DNA diagnostics follow a sequential process that can be time consuming when the genetic variant is uncommon [Game et al., 2003; Rooks et al., 2005; Rooks et al., 2012]. A case in point is the large deletions and rearrangements of the *HBB* and *HBA* cluster. In current routine practice, these are identified by multiplex ligation probe amplification (MLPA) analysis of the suspected loci on chromosome 11 or 16 respectively, after negative gene sequencing analysis [Harteveld et al., 2005; Traeger-Synodinos and Harteveld, 2014]. Further characterization of the break-points involves refining the region of interest, followed by gap-PCR of the specific break-points [Craig et al., 1994] and sequence analysis of the break-point fragment. This intensive workup can be hampered by the difficulty in designing gap-PCR primers and obtaining a specific break-point fragment, leaving many cases still not completely characterized [Rooks et al., 2012; Shalev et al., 2013].

Here, we present to-the-base resolution of a complex inversion/deletion rearrangement of the *HBB* cluster, and use it as an example of how analysis of a single next-generation sequencing (NGS) dataset can detect all genetic variation in a targeted region of the genome.

We designed a SureSelect bait capture panel (Agilent, Santa Clara, CA) to cover both the α and β globin loci on chromosomes 16p and 11p, respectively, in two contiguous regions, and analyzed the sequence data in NextGene (Softgenetics, State College, PA) software. The software enabled us to identify small genetic variants (1–3 bases) and by comparing coverage between controls and patients, we were able to identify large deletions to within a single bait-covered position of 120 bp. Using this technique, a $\varepsilon\gamma\delta\beta$ thalassemia rearrangement previously identified by CGH array, qPCR, Southern blotting and MLPA, could be fully characterized.

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Swee Lay Thein, King's College London, James Black Centre, 125 Coldharbour Lane, London SE5 9NU, UK. E-mail: sl.thein@kcl.ac.uk, swee.thein@nhs.net

The deletion of 122.6 kb was consistent with previous findings, it also included 82 bp of an inversion at one side of the deletion adjacent to *HBG2* (MIM #142250) exon 2 extending to the upstream β-locus control region (β-LCR). The inversion of 59.0 kb was previously not detectable by CGH array or MLPA, a confounding factor in previous attempts to fully characterize this rearrangement. We identified the same rearrangement using gap-PCR and DNA sequence analysis in the proband's daughter and two other unrelated English individuals with the same hematological phenotype. The structural variant is named the English V $\varepsilon\gamma\delta\beta$ thalassemia (g.5215690_5274684invdel5215690_5215772del5274684_5397195, HbVar 2935 in Hb Var Database of Human Hemoglobin Variants and Thalassemias, http://globin.cse.psu.edu/hbvar). To our knowledge, this is the first inversion–deletion rearrangement causing $\varepsilon\gamma\delta\beta$ thalassemia and this diagnosis highlights the strengths of an NGS approach, being applicable to any inherited disease.

Materials and Methods

Patient Samples and DNA Extraction

Patient samples were submitted for clinical testing to The Red Cell Centre at King's College Hospital, London, a reference center for hemoglobinopathy diagnosis. Samples were collected in EDTA-containing tubes and DNA extracted from whole blood either by the QIASymphony (Qiagen, Venlo, The Netherlands) Midi kit or by phenol-chloroform extraction. DNA from the proband in Family 1, two negative controls (determined by MLPA), and two positive controls with known variants (619 bp *HBB* deletion, HbVar 979, NG_000007.3:g.71609_72227delinsAAGTAGA [Orkin et al., 1979; Pritchard et al., 2010] and 106 kb *HBB* deletion [Feingold and Forget, 1989] causing hereditary persistence of fetal hemoglobin1, HPFH1, HbVar 1021, NG_000007.3:g.59478_144395del84918) were selected for sequence analysis.

Target Enrichment Library Design

A SureSelect Custom Target Enrichment Library was designed using Agilent's in-browser tool eArray (Agilent). 120 bp biotinylated RNA baits were designed against a 4.26 Mb region of chromosomes 11 and 16 (genome build 37.3) covering both globin gene clusters (Chr11: 3250000–7250000, Chr16: 0–260000). One times tiling was used and the region was repeat masked allowing baits to extend up to 20 bp into repetitive regions (see bait library design parameters in Supporting Information, Supp. Table S1). Baits which were likely to perform poorly (those with a high GC content or 'orphans' which were >20 bp away from their neighbors) were "boosted" according to parameters listed in Supp. Table S2.

Sample Preparation and Sequencing

DNA samples (3 μg) were sonicated to a mean size of 500 bp using a Bioruptor (Diagenode, Ougrée, Belgium) and prepared for sequencing using the SureSelect Library Preparation Kit (Agilent) in accordance with the manufacturer's protocol (v1.4.1). DNA fragments underwent end repair, followed by the addition of an adenosine overhang allowing subsequent adapter ligation. Target enrichment was performed using the custom bait library and separated from non-hybridized fragments using streptavidin magnetic beads. The fragments were purified, amplified, and ligated to indexing tags. Each step was followed by clean-up using AMPure XP beads (Beckman Coulter). Samples were pooled at an equimolar concentration of 12 pMols and sequenced on an Illumina MiSeq using the v2 2 × 250 bp sequencing kit (Illumina, San Diego, CA). Sequencing parameters and quality statistics from the MiSeq run are listed in Supp. Table S3.

Data Analysis

All NGS data analysis was carried out using NextGene software (SoftGenetics). Quality filtering was performed by NextGene's Format Conversion tool using its standard parameters to generate FASTA files (see Supp. Table S4 for the format conversion settings).

FASTA files were aligned with high stringency to a reference sequence of the human genome (hg 19 build 37.3) as paired reads using NextGene (SoftGenetics: State College, PA). The alignment settings are listed in Supp. Table S5. The alignment produced a variant report, a coverage report, and opposite and same direction read reports. The uses of these reports in detecting mutations are outlined in Supp. Table S6. The in-browser tool BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat)—part of the University of California, Santa Cruz (UCSC) Genome Bioinformatics Site—was used to query sequences in the alignment that could represent either novel sequence or the misalignment of off-target sequence. The UCSC tool used the same human genome build (37.3) as the reference sequence in NextGene.

Characterization of the Deletion Breakpoints by gap-PCR and DNA Sequence Analysis

Genomic DNA encompassing the inversion points and deletion breakpoints was amplified by PCR using specific primers. For products <2 kb, PCR was carried out using the Qiagen Multiplex Mix reagent kit (Qiagen) in a 20 μl volume containing 10 pmol each of the forward and reverse primers, and 10 μl of the reagent mix. Cycling parameters were 95°C for 10 min, followed by 35 cycles of 95°C for 40 sec, 63°C for 40 sec, 72°C for 1 min, and a final extension at 72°C for 10 min. For PCR products >2 kb, PCR was carried out in a 25 μl volume using LongAmp Taq Polymerase (NEB, Ipswich, MA) with 200 ng DNA, 10 pmol each of the forward and reverse primers, and 5 μl of the reaction mix. The cycling parameters were 94°C for 30 sec, followed by 35 cycles of 94°C for 30 sec, 64°C for 30 sec, 65°C for 10 min, and a final extension at 65°C for 10 min.

The PCR products were resolved by electrophoresis in a 1% agarose gel and the specific fragments purified using AMPure beads (Beckman Coulter, Pasadena, CA) and then sequenced in triplicate using BigDye Terminator sequencing chemistry v3.1 (Thermo Fisher Scientific, Waltham, MA). Dye terminator products were purified using Cleanseq beads (Beckman Coulter) before being run on a 3130xl Genetic Analyzer (Thermo Fisher Scientific). Sequence data were analyzed with Sequencher Software version 4.1 (Gene Codes, Ann Arbor, MI).

Results

Identification of $\varepsilon\gamma\delta\beta$ Thalassemia in an English Family – Case 1

The proband was a 39-year-old woman of English Anglo-Saxon origin who had been noted to have hypochromic microcytic anemia since infancy, unresponsive to oral iron supplements. She was referred for further investigation at 14 years of age with a hematology profile of Hb 99 g/l, RBC 5.37 × 10^{12}/l, MCV 58.0 fl,

Table 1. Hematological Indices of Individuals Heterozygous for $\varepsilon\gamma\delta\beta$ Thalassemia

Individual	Age	Hb (gm/dl)	RBC ($\times 10^{12}$/l)	MCV (fl)	MCH (μg)	HbA_2 (%)	HbF (%)	β/α
Proband	14 y	9.9	5.37	58.0	18.4	2.7	0.9	0.26
	25 y	7.8	4.02	61.9	19.4	3.0	<1.0	0.59
Daughter	3½ y	9.3	5.05	56.6	18.4	3.2	1.0	
Mother	56 y	8.0	4.96	53.5	16.2	2.6	2.0	0.54
Case 2	52 y	6.8	4.53	50.0	15.0	2.5	2.1	
Case 3	24 y	9.7	5.01	62.7	19.4	2.6	1.0	

MCH 18.4 pg, reticulocytes 0.6%, HbA2 2.7%, HbF 0.9%, and serum ferritin 160 μg/l. Globin chain synthesis ratio was β:α = 0.26 (Table 1). Her mother had a similar hematological profile with hypochromic microcytic red blood cell indices, HbA_2 2.6%, HbF 2.0% and globin chain synthesis ratio β:α = 0.54 (Table 1), a phenotype typical of heterozygous $\varepsilon\gamma\delta\beta$ thalassemia. DNA sequence analysis excluded any causative variants in the *HBB* (MIM #141900) gene but Southern blotting indicated a deletion extending from *HBG2* upstream to the β-LCR leaving the *HBB* gene intact, that is, a diagnosis of Type II $\varepsilon\gamma\delta\beta$ thalassemia. Breakpoints of the deletion proved extremely difficult to characterize because of the difficulty in designing specific gap-PCR primers. The proband was referred for antenatal genetic counseling, nine years later, to King's College Hospital, when she was 25 years of age. Her hematological profile remained hypochromic microcytic with normal HbA_2 3.0%, HbF < 1.0% and globin chain synthesis ratio β:α = 0.59 (Table 1).

A fetal blood sample revealed a globin chain synthesis ratio of β:α = 0.07 indicating that her baby was heterozygous for $\varepsilon\gamma\delta\beta$ thalassemia. The newborn had received two intra-uterine blood transfusions followed by another blood transfusion at birth when she developed neonatal jaundice and anemia. Hematological profile of her daughter at 3½ years of age confirmed the thalassemic indices characteristic of $\varepsilon\gamma\delta\beta$ thalassemia (Table 1).

Cases 2 and 3

Case 2 was a 52-year-old English woman referred for further analysis of her hypochromic microcytic anemia as was Case 3 (also English) who was also pregnant. Both patients 2 and 3 had HbA_2 levels within the normal limits (Table 1) and were negative for the common alpha thalassemia deletions. Sequence analysis of their α and β globin genes did not detect any causative variants. In both cases 2 and 3, however, MLPA detected a deletion of >42 kb on chromosome 11p15.5, removing the β-LCR, causing $\varepsilon\gamma\delta\beta$ thalassemia consistent with the hematological phenotype.

NextGene Sequence Analysis

Sequence data from the five DNA samples generated on the Illumina MiSeq instrument were converted from FASTQ to FASTA format prior to alignment at high stringency to the reference sequence using NextGene. The proportion of fragments on the flow cell attributed to each sample were within acceptable margins of equality (with no sample being over or under represented by >10%). Significantly fewer reads were produced for the unknown patient sample compared with controls (Supp. Table S3) despite >99% of reads passing quality filtering steps during the conversion of the data from FASTQ to FASTA format. The lower number of reads for this sample resulted in a lower mean coverage compared to controls but this did not adversely affect the ability to characterize the rearrangement breakpoints.

Genotype and SNP Calling

As part of clinical diagnostics, a small number of variants had been previously identified in the alpha and beta globin genes by dye-terminator sequencing. Despite the high degree of homology between the different globin genes, all previously identified variants were correctly aligned by NextGene and were in the variant report, confirming its correct alignment and variant calling capability (data not shown).

RPKM Analysis

After sequence alignment NextGene produces an expression report for each sample, including a coverage statistic called the 'reads per kilobase exon model per million mapped reads' (RPKM). The RPKM value is an average coverage value for every 1 kb of the alignment, divided by the total number of mapped reads within the length of the reference region. This internal sample normalization allows coverage comparisons to be made between controls and patient samples, similar to MLPA analysis. By uploading the bait BED file, the average RPKM value for each 120 bp bait within each 1 kb region could be calculated. This permitted increased resolution and review of individual bait performance. Baits that covered the start and end points of large rearrangements could then be identified. Differences in coverage between controls and test samples that represent the presence or absence of insertions or deletions (indels) were represented as a ratio, by dividing the RPKM values of the patient by the mean RPKM values of the normal controls. These RPKM ratios were plotted on a log2 scale in Excel (Microsoft, Redmond, WA) allowing direct comparison to CGH array data (although this was not necessary for the fold differences observed). The known breakpoints of the positive controls and MLPA analysis of undetermined samples allowed quick adjustment of the *x*-axis (chromosomal location) to best display the indel position.

Normal and positive control samples were prepared, hybridized, and sequenced together with each batch of unknown samples. Including normal controls and test samples in the same batch during library preparation reduced variation seen between the RPKM values in balanced regions, giving clearer discrimination of the breakpoints in unbalanced regions. The variation is most likely due to the manual library preparation method, particularly the hybridization of the baits to the genomic fragments, which is temperature dependent. The RPKM data plots could be used to identify the location of the deletion start points and end points to within a single bait-covered position, giving this methodology an accuracy of ±240 bp, assuming that the region is tiled continuously. Where break points are situated >500 bp into repetitive regions that are not included in the bait design, this accuracy is reduced to the size of the repeat.

Most normal variation in the RPKM values of the patient was less than 0.5 from the values for the same bait in the control, when plotted on a log2 scale. Most bait positions involved in a heterozygous deletion, however, had RPKM values that deviated by >0.5 from the control. Two issues with this analysis method are identifying small changes that may only affect a few bait positions, and differentiating genuine changes from copy number variation and bait-performance variability, which can also produce variation of >0.5. However, both copy number variations and bait performance variability tended to be localized to the same regions in both samples and controls. Superimposing a plot of the standard deviation from

the average RPKM values for several controls over the RPKM data for specific samples provided an effective means of separating much of the spurious data from genuine variants (see Supp. Fig. S1). These same regions appeared to be consistently variable between batches and were not investigated during the analysis. The genome browser confirmed the presence of true copy number variants identified in the analysis.

The *HBB*-like globin genes cover a 50 kb region accounting for less than 1.25% of the entire bait-tiled region on chromosome 11. Although large deletions (>100 kb) may still be easy to identify at low resolution, it is also necessary to look more closely at the globin cluster region at higher resolution for smaller deletions. Using the 619 bp deletion as a positive control, we generated the comparative RPKM plot on three different scales focusing on 6 sequential baits with variation >0.5. Viewing the alignment in this region characterized the 619 bp deletion to-the-base, including the additional 7 bp insertion characteristic of this deletion [Pritchard et al., 2010], data not shown. Breakpoints for the other positive control (HPFH1 deletion) were also accurately resolved in this way.

Opposite and Same Direction Reads Data Analysis

Paired reads that did not map during the first alignment (either because of their orientation or their distance from one another) were separated into opposite and same direction read reports. Opposite direction reads are the expected norm from Illumina paired end data as each DNA strand is read in opposite directions. The opposite direction reads in the report are rejected from the alignment as the two paired ends map back at a gap distance of >600 bp (as specified in the sequence alignment settings) and may indicate an indel in the test sample. Same direction reads were recorded when both halves of the read aligned to the reference in the same orientation, which may indicate the break points of an inverted sequence. Both these reports are useful when trying to detect structural variation greater than the fragment size (mean 500 bp) as described in Supp. Table S6.

Using Excel (Microsoft), reads in the opposite and same direction reports were filtered, removing sequences where neither of the two reads aligned within the bait-tiled region of interest (chr11: 3500000–7500000, chr16: 0–260000). Because of the large number of opposite direction reads produced for all samples, this report was filtered again to remove pairs where neither end aligned near the approximate break points of a region indicated as deleted by the RPKM plots. This was not necessary for the less-frequent same direction reads. The reads were plotted, being overlaid on the RPKM data plot with both halves of each pair being given the same arbitrary *y*-axis value. Regions where opposite or same-direction reads "piled up" indicated the potential break points for rearrangements. Spurious pile-up was frequently recorded between highly similar or repetitive regions such as between the two gamma globin genes. These were excluded as evidence of genuine variation as similar results were obtained using normal control sample data. Visual inspection of the reads aligning to these regions in NextGene and BLAT query of the sequences of the rejected opposite or same direction reads also removed false positives.

Mismatched Base String Analysis

NextGene visually indicates deviations in the sample sequence from the reference in three ways: (1) deviations that do not meet the criteria of a variant are indicated by gray bars in the NextGene viewer; (2) deviations that are recorded as variants as listed in dbSNP are indicated by purple bars and, (3) deviations which are recorded as variants that are novel are indicated by blue bars. A string of novel mismatched bases that are blue in a large number of records can suggest the break-points of a large rearrangement. The mismatched string of bases in the read indicates a break in the continuation of the normal reference. NextGene does not currently have a method for identifying these regions. Identifying the regions in Excel is also difficult because, as NextGene attempts to make the sequence fit its expected alignment, it breaks the mismatched sequence into smaller chunks around any bases which are similar enough to the reference to be aligned successfully. Furthermore, many of the mismatched bases may not meet the criteria to be recorded as a genuine variation and will be excluded from the report. As such, it is necessary to visually inspect the sequence aligning around the suspected deletion break points identified by the RPKM and opposite/same direction read plots and look for mismatching bases in the sequence alignment viewer.

Once these regions are identified the names of reads containing these break-point sequences can be recorded, and their two halves can be identified in the original FASTA files. The entire read sequence can then be queried in BLAT (UCSC). The mismatched base sequence string can represent misaligned data but if the sequence produces partial high fidelity alignment to two distinct regions it is likely to represent a genuine break point, particularly if they align to the same regions identified by RPKM analysis. The break point can be confirmed by Gap-PCR and sequencing of the breakpoint spanning amplicon.

Characterization of the $(\varepsilon\gamma\delta\beta)^0$ Rearrangement

An RPKM plot showed that the unknown variant deviated from a normal control indicating a large deleted region which included the β-LCR, *HBE* (MIM #142100) and part of *HBG2* (Fig. 1A). This was consistent with previous data from MLPA and CGH array analyses. Mismatched base strings were identified in the region of the breakpoint identified by RPKM analysis. After BLAT query the mismatched region was found to align 59.0 kb upstream (telomeric to the deletion) indicating the presence of an unsuspected inversion (Fig. 1B). A same direction read also mapped to these locations (Fig. 1B).

PCR primers in the same direction were designed to confirm the inversion deletion breakpoints (Fig. 2A–C). A unique gap-PCR fragment of 1,031 bp was successfully amplified using same direction primers, P1: 5′-AGCTGGTTGGTCCGTTTTGG-3′ (from UCSC 5215416 to 5215435), and P3: 5′-CTCTGCATCATGGGCAGTGAG-3′ (from UCSC 5274485 to 5274505) (Supp. Fig. S2). Sequence analysis of this unique PCR product confirmed the inversion, consistent with the NGS data.

Confirmation of the deletion breakpoint at the centromeric end, however, required a longer amplicon to accommodate the 6 kb LINE repeat (Fig. 2C and Supp. Fig. S2). A unique gap-PCR product of 4,499 bp was successfully amplified using primers P2: 5′-AGTGCAAAGGATGCCAGGAC-3′ (from UCSC 5216446 to 5216465), and P4: 5′-GAGCAAGTGCATGCAAGGAGA-3′ (from UCSC 5401092 to 5401112) (Supp. Fig. S2). Sequence analysis of the unique fragment confirmed the breakpoint within the LINE repeat as suspected and revealed that the deletion had also removed 82 bp of the now adjacent inverted sequence (Supp. Fig. S3). A thymidine base at the exact break point of the inversion–deletion may belong to either the pre-deletion or post-deletion sequence as a thymidine is expected at either position. The coordinates used assume the thymidine base is post-deletion, position 5397195. The new rearrangement in this

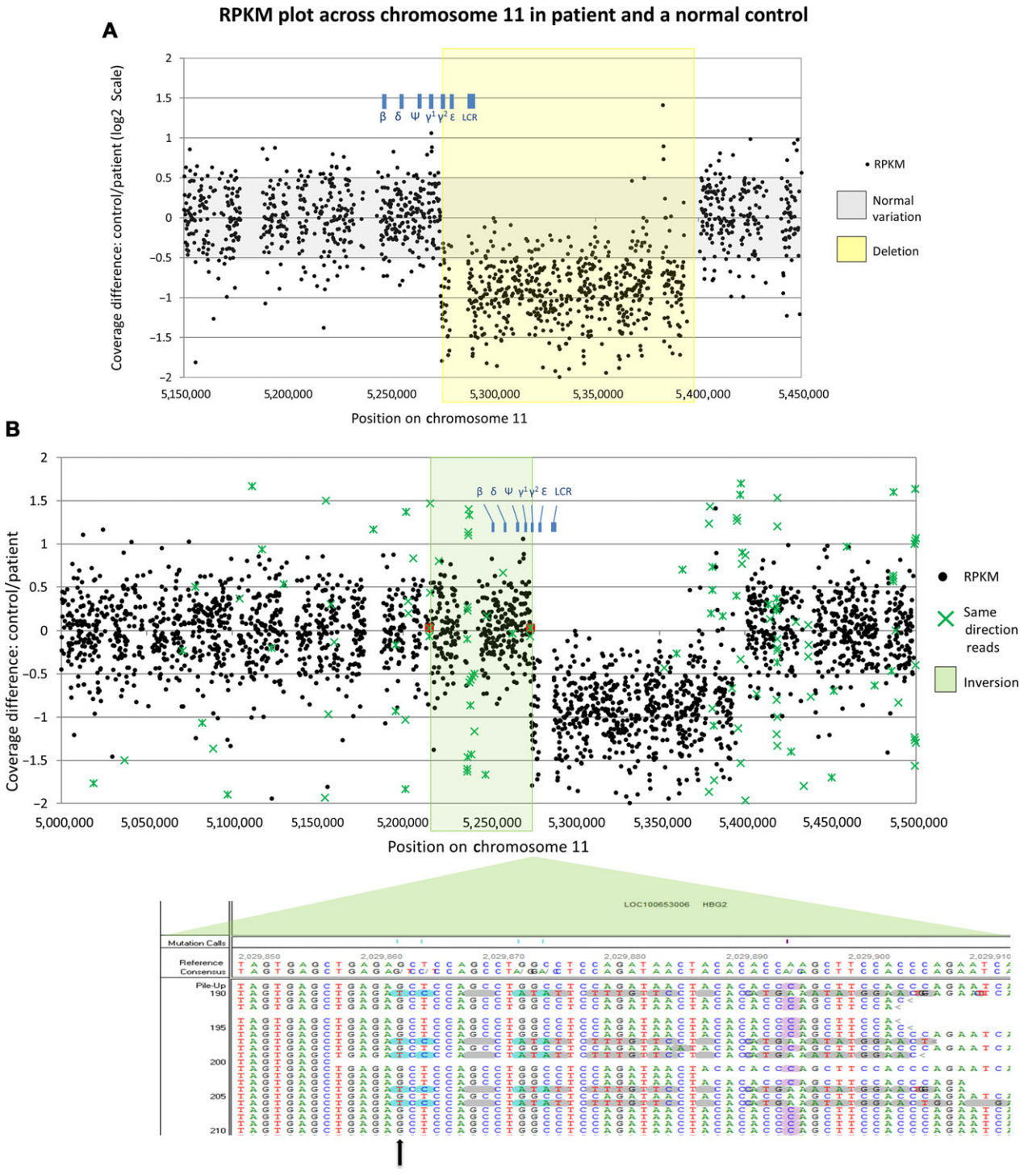
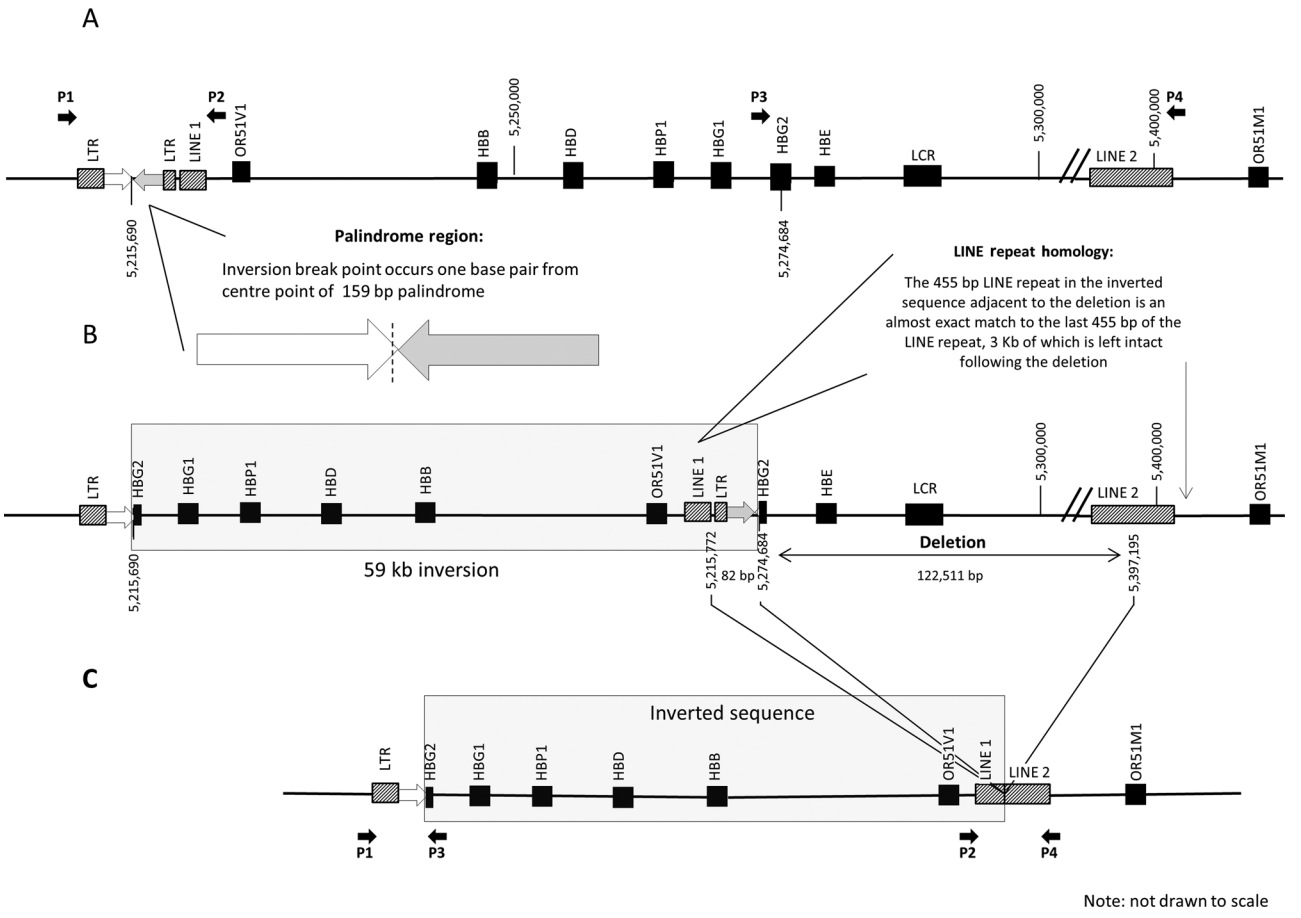


Figure 1. **A: RPKM plot across chromosome 11 in patient and a normal control.** RPKM plot showing variation of sample RPKM values from control on a Log2 scale. A continuous region (highlighted in yellow) of ~120 kb yields significantly lower coverage—producing values of >–0.5 equating to 50% less sequences aligning to each bait position—in the test sample compared with control. This indicates the presence of a heterozygous deletion of this region in the test sample. The deletion removes part of the *HBB* globin cluster including the β-locus control region (LCR). **B: Positions of same direction reads across chromosome 11. Upper panel:** The positions of same direction reads are overlaid on the RPKM data plot and are useful in detecting inverted sequences. The positions at which the two same-direction reads from each fragment align (marked green cross) have the same *y*-axis value. Two same-direction reads of a single fragment are boxed in red, indicating the break points of an inversion of the sequence (shaded green region) between them. **Lower panel:** Inspection of these positions (green cross) at one potential inversion point in the NextGene viewer reveals reads with strings of misaligned bases—the bases highlighted in gray, blue, and purple do not match the reference sequence—indicating mutated allele, whereas some reads are perfect match to the reference sequence (indicating normal allele). Reading the sequence from left to right, all the aligned sequences match the reference sequence up until the black vertical arrow. From the black arrow onwards, five of the reads do not match the reference at this location and correspond to the point at which the inversion occurs. The reads that do not contain any mismatched bases, compared with the reference, are from the normal allele and confirm that the individual is heterozygous for the inversion.



Note: not drawn to scale

Figure 2. **Line representation of the rearrangement events on chromosome 11p. A:** Representation of the normal region encompassing the *HBB* cluster on chromosome 11p. Downstream of the OR51V1 gene is a 159 bp palindrome between two LTRs (details shown in Supp. Fig. S4). **B and C:** A 59.0 kb inversion (gray box) occurs between positions chr11: 5215690–5274684 encompassing *HBB*, *HBD*, *HBP1*, *HBG1*, and two exons of *HBG2*. The start point for the inversion is in the 159 bp palindrome sequence, one base-pair from its central point. The inversion event is followed by a 122.6 kb deletion, which removes 82 bp of the newly inverted sequence, and the adjoining uninverted third exon of *HBG2*, *HBE*, and the β-LCR; chr11 5274684–5397195. It should be noted that the inversion deletion fuses two previously distant LINE repeats. The first LINE repeat, which is flipped during the inversion, is 98% homologous to the 5′ end of the second line repeat, of which 3.5 kb is intact after the deletion. Solid black arrows (P1, P2, P3, and P4) in **A** and **C** indicate the location of the primers used in gap-PCR to confirm the rearrangement.

individual is therefore described as Chr11 Hg19 (build 37.3) g.5215690_5274684invdel5215690_5215772del5274684_5397195 and has been submitted to the HbVar database (http://globin.cse.psu.edu/hbvar), HbVar ID 2935. The rearrangement involves a single 122.6 kb deletion, but because this affects two normally distant areas, it is listed here as two separate deletions (122.5 kb and 82 bp) to describe both the affected regions, respectively, interrupted by an inversion of 59.0 kb. The inversion–deletion mutation is likely to have been created in two events, the first being the 59.0 kb inversion followed by a single deletion event. The start point of the inversion is one base from the center of a 159 bp palindrome sequence which abuts an LTR/LINE repeat region which is inverted (Fig. 2A and Supp. Fig. S4). Following the inversion, a LINE repeat that shares a high level of homology with a 6 kb LINE repeat upstream of the β-LCR, is now adjacent to HBG2 exon 1. The inverted region of the palindrome and the abutting LTR, and part of the upstream 6 kb LINE repeats were removed, leaving one inverted LINE repeat and 3.5 kb of the 6 kb LINE repeat intact after the deletion (Fig. 2C). Identical Gap-PCR products using primer pairs P1/P3 and P2/P4 were amplified in the daughter of the proband and the other two patients, all of which had a similar hematological phenotype, confirming that they all had identical DNA rearrangements (Supp. Fig. S2).

Discussion

An antenatal patient was referred to the molecular pathology laboratory at King's College Hospital for thalassemia variant identification. The hypochromic microcytic anemia with normal HbA_2 level suggested that she could be heterozygous for α thalassemia, but common alpha thalassemia deletion variants were excluded by gap-PCR, and sequence analyses of the alpha globin genes did not identify a causative variant. These negative results together with results of the globin chain synthesis ratios, prompted MLPA analysis of the *HBB* loci. The test identified a deletion which removed the β-LCR, causative of the phenotype of an $\varepsilon\gamma\delta\beta$ thalassemia carrier. Various other techniques including Southern blotting, quantitative PCR and CGH array were employed in an attempt to characterize this mutation, but generating a gap-PCR product spanning the breakpoint was not possible. One end of the deletion in *HBG2* exon 2 appears similar to that previously reported in 1985 and 1988, that was named English I [Curtin et al., 1985; Curtin and

Kan, 1988]. However, as we do not have DNA to confirm if English I $\varepsilon\gamma\delta\beta$ thalassaemia is identical to the present case, we have termed the newly characterized rearrangement as English V $\varepsilon\gamma\delta\beta$ thalassemia.

We sequenced the patient sample on the MiSeq platform (Illumina) using an in-solution target enrichment chemistry (Agilent) covering two contiguous regions which included the *HBA* and *HBB* globin gene loci, respectively. Two negative controls with no structural rearrangements and two positive controls with known thalassemia deletions were prepared and sequenced in parallel. After sequence alignment the coverage per bait was normalized using total reads for the sample, allowing coverage between normal samples and the test case to be compared. The coverage comparison plot allowed identification of the deleted region and highlighted specific areas where the breakpoint spanning sequences could be located. Inspection of the aligned sequence surrounding the deleted region showed a mismatched string of bases that, when analyzed in BLAT, indicated an inversion of 59.0 kb adjacent to one end of the deletion. Once the inversion was confirmed by gap-PCR, the deletion could also be confirmed by long range PCR. As the 122.6 kb deletion includes a part (82 bp) of the abutting inverted sequence, we suggest that the inversion event occurred first, followed by the deletion. Sequence analysis of the Gap-PCR amplicons allowed to-the-base resolution of the inversion–deletion mutation and further screening of previously uncharacterized historical samples with a similar phenotype.

This sequencing and analysis methodology allows identification of all known categories of variants causing not only hemoglobinopathies but also other diseases. More importantly, the approach detects deletions to a greater resolution than MLPA and Southern blotting. A one times tiling density across the region allows a resolution of ±240 bp (2 × 120 bp) where both breakpoints were in bait-covered regions. It is likely that even greater resolution could be achieved with increased tiling density, or by taking averages of smaller regions of each bait position. Our experience suggests that the technique is able to detect single base changes and indels up to the length of the fragment size being analyzed as well as larger structural variants of over a megabase. As long as one end of the breakpoint is less than 500 bp into a repetitive sequence it can also identify breakpoint sequences without resorting to gap-PCR analysis.

Key to the success of the methodology was being able to shear the DNA into large fragments, (mean 500 bp) allowing a long read length (2 × 250 bp), in combination with flexibility of alignment in NextGene. The small base match percentage necessary for successful bait hybridization further allowed reads with large amounts of novel sequence to be aligned to a reference sequence with a high degree of accuracy, permitting sensitive detection of small mutations and, uniquely, the detection of inversions. Although not demonstrated here, the methodology should also be applicable to characterization of other forms of rearrangement mutations, such as balanced translocations and insertions. Using an in-solution bait capture chemistry with minimal PCR amplification permitted relative comparison of coverage and the capture of breakpoint sequences. It is unlikely that a PCR based capture approach such as Ampliseq or Raindance [Cheng et al., 2014; Tewhey et al., 2009; Zhang et al., 2014], would reduce noise in the assay, being affected by variants in the primer regions and differences in amplification efficiency. The PCR approach would not identify novel breakpoint sequences as this would require prior knowledge of the breakpoint trying to be detected. Hence, the bait capture approach is probably unique in its ability to identify and characterize unknown breakpoints in a single assay.

Characterization of this novel rearrangement also highlighted some major considerations for a routine diagnostic laboratory. Problems were posed by the repetitive nature of both ends of the deletion which prevented baits being tiled into these regions, such that sequences which were picked up at these points were mostly misalignments of off-target sequence. Another problem was that the inversion was hard to identify against background noise in the same-direction paired reads report. This was partly because of the low number of baits covering the region, and partly a consequence of NextGene's increased tolerance for mismatching bases resulting from the large read length of the MiSeq. Finally, as many breakpoint reads were included in the alignment rather than being rejected as same direction reads, their mismatched base strings were dismissed as errors by the software and were excluded from the variant report. The consequence of these problems was that breakpoint sequence detection was largely manual which slowed down the diagnosis. Gap-PCR and dye-terminator sequencing was still necessary to establish the deletion breakpoints when it fell >500 bp within a non-tiled repeat.

Large rearrangements are a common feature of alpha thalassemia and to a lesser extent in beta thalassemia so a technique that can handle these in addition to sensitive SNP detection is highly useful [Giardine et al., 2011; Higgs et al., 2012]. Currently, five separate tests—multiplex gap-PCR, sequence analysis of specifically amplified alpha and beta globin gene products, and MPLA of the alpha and beta globin loci are routinely used in a diagnostic laboratory for complete characterization of all variants [Traeger-Synodinos and Harteveld, 2014]. NGS represents a comprehensive single methodology that can fully characterize all the types of variant by analysis of a single data set [Berglund et al., 2011; Nielsen et al., 2011]. If this process can be honed and applied to routine diagnostics it will have a dramatic effect on molecular diagnostic laboratories [Kassahn et al., 2014]. NextGene can identify SNPs automatically but there is currently no reliable way of detecting large rearrangements. Dosage changing deletions and duplications can be detected by comparing the coverage of a test subject to a normal control but this is still associated with high background noise and inter-sample variability which could impact detection accuracy. Increasing the bait tiling density, automating the sample preparation process to produce less variable results and the use of multiple negative controls for comparison to test subjects are steps that are likely to improve the methodology. A comprehensive bioinformatic setup that includes indel detection (currently requires a Linux environment), would automate data analysis and reduce the need for Excel. Non-dosage changing rearrangements will require more work for robust routine identification. Reducing the noise associated with same direction reads is challenging when dealing with such large fragment sizes, so the best method for identifying these is likely to be by identifying the strings of mismatched bases with which they are associated. These strings are also frequently associated with the misalignment of off-target or otherwise repetitive sequences, so it is likely to be a challenge for software developers.

e$\gamma\delta\beta$ thalassemias describe a rare sub-group of thalassemias characterized by downregulation of the genes on the *HBB* cluster, that is, *HBE*, *HBG2*, *HBG1* (MIM #142200), *HBD* (MIM #142000) and *HBB* itself. They are rare and caused by large deletions, classified molecularly into two categories: group I deletes all or most of the *HBB* cluster, including the *HBB* gene, and group II, which removes the upstream β-LCR, leading to inactivation of the downstream globin genes. We summarize the total number of $\varepsilon\gamma\delta\beta$ thalassemias reported to date (Fig. 3) that includes 17 type I (the longest being 1.78 Mb in a Bedouin family [Shalev et al., 2013]), and 13 type II deletions. The inversion–deletion mutation described here is first

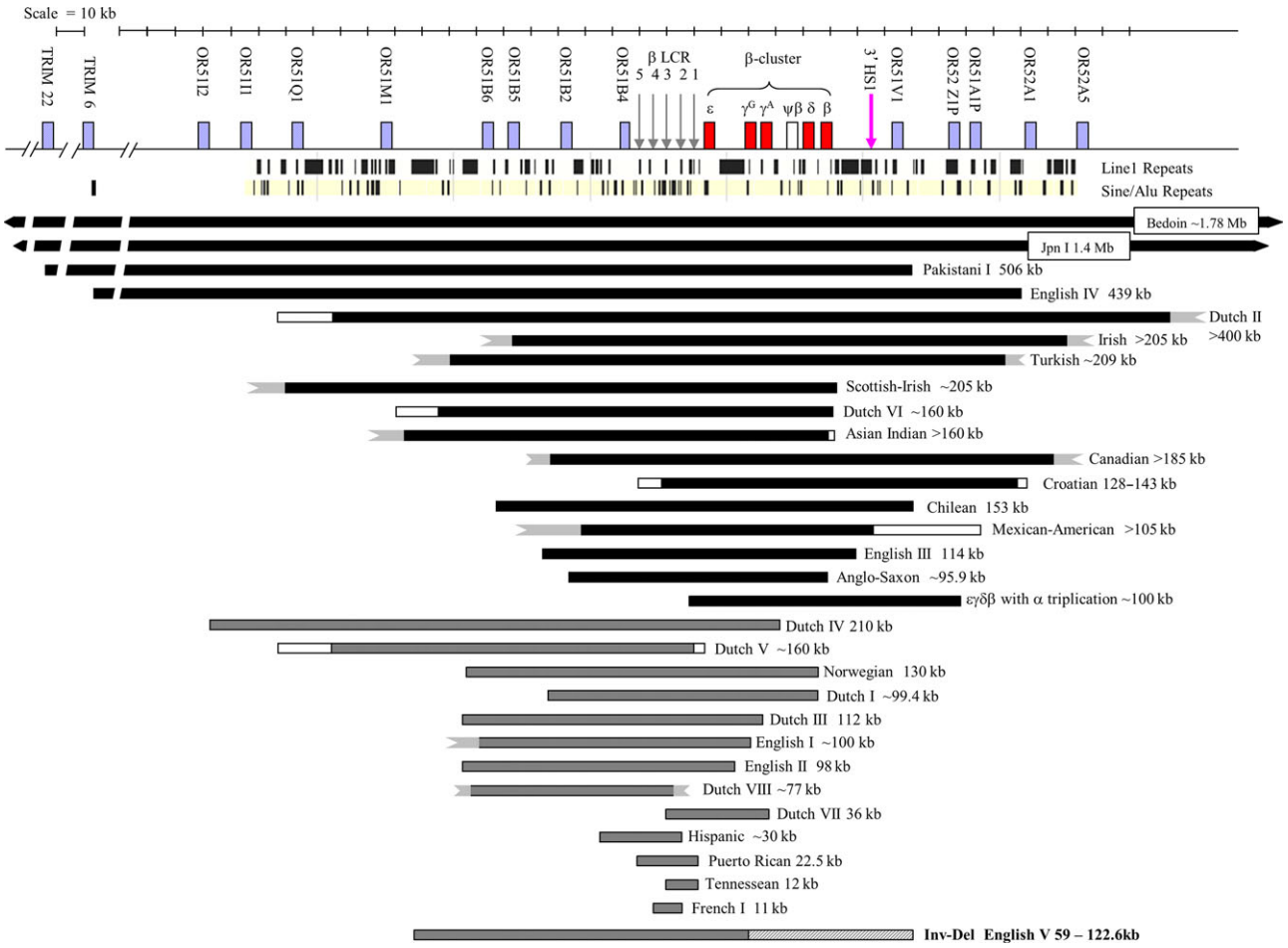


Figure 3. Summary of 31 $\varepsilon\gamma\delta\beta$ thalassemias reported to date. The newly reported inversion–deletion English V mutation (highlighted in bold, bottom of panel) is first of its type. The hashed box indicates the inverted sequence and the gray filled box the deleted sequence. For details of all the other mutations, refer to Rooks et al. (2012).

of its type to cause $\varepsilon\gamma\delta\beta$ thalassemia; the deletion removes the upstream β-LCR but leaves *HBG1*, *HBD*, and *HBB* intact, although the sequences are inverted. We suggest that the inversion event is initiated by the inverted palindromic sequence at the breakpoint, as was also noted in a previously reported structural variation in the *HBB* region [Rooks et al., 2012]. It is also worth noting that palindromic sequences are associated with genomic regions that have undergone genome duplication [Butler et al., 2002] and this is consistent with the globin gene loci. In many cases where the breakpoint of the deletions have been characterized, they occurred within regions of repeat sequences (LI, LINE, or Alu) containing short regions of direct homology between the flanking sequences [Rooks et al., 2005], a feature that is likely to have contributed to the illegitimate recombination in this case.

$\varepsilon\gamma\delta\beta$ thalassemias have only been found in the heterozygous form; presumably homozygotes for such deletions are not compatible with fetal survival because of the inactivation of all the β-like globin genes. Heterozygotes may have severe anemia at birth, and in some cases (as in the daughter of the proband in Case 1), intrauterine and perinatal blood transfusions are required to tide them over the neonatal period. All except for one $\varepsilon\gamma\delta\beta$ deletion (Scottish-Irish 205 kb [Pirastu et al., 1983; Trent et al., 1990]) characterized to date, have been unique to the families described, and several mutations appear de novo. It is remarkable that all three cases of $\varepsilon\gamma\delta\beta$ thalassemia reported here had the same inversion–deletion rearrangement, raising questions about the frequency of this variant. This case illustrates the power of NGS that has allowed us to fully characterize the mutation, and design gap-PCR primers to confirm the same mutation in two other non-related individuals with a similar hematological phenotype. We have detailed our NGS approach that should be applicable as a diagnostic tool for other diseases. Automation of sample preparation and data analyses should allow the methodology to eventually be applied in routine diagnostics.

Acknowledgments

We than Professor Chris Shaw, Dr. Bradley Smith, and Athinia Gkazi at The Institute of Psychiatry, King's College London, for help and advice in library preparation and access to facilities. We thank Claire Steward for help in preparation of the manuscript, and the King's College Hospital Charity for support (CS is KCHC-funded).

Disclosure statement: The authors declare no conflict of interest.

Author Contributions

C.S., H.R., and B.C. performed the experiments; C.S., H.R., S.L.T., and B.C. analyzed the data; C.S., S.L.T., and B.C.

wrote the manuscript. All coauthors gave feedback on the manuscript.

References

Berglund EC, Kiialainen A, Syvanen AC. 2011. Next-generation sequencing technologies and applications for human genetic history and forensics. Investig Genet 2:23.

Butler DK, Gillespie D, Steele B. 2002. Formation of large palindromic DNA by homologous recombination of short inverted repeat sequences in Saccharomyces cerevisiae. Genetics 161:1065–1075.

Cheng DT, Cheng J, Mitchell TN, Syed A, Zehir A, Mensah NY, Oultache A, Nafa K, Levine RL, Arcila ME, Berger MF, Hedvat CV. 2014. Detection of mutations in myeloid malignancies through paired-sample analysis of microdroplet-PCR deep sequencing data. J Mol Diagn 16:504–518.

Craig JE, Barnetson RA, Prior J, Raven JL, Thein SL. 1994. Rapid detection of deletions causing delta beta thalassemia and hereditary persistence of fetal hemoglobin by enzymatic amplification. Blood 83:1673–1682.

Curtin P, Pirastu M, Kan YW, Gobert-Jones JA, Stephens AD, Lehmann H. 1985. A distant gene deletion affects β-globin gene function in an atypical $\gamma\delta\beta$-thalassemia. J Clin Invest 76:1554–1558.

Curtin PT, Kan YW. 1988. The inactive β globin gene on a $\gamma\delta\beta$ thalassaemia chromosome has a normal structure and functions normally *in vitro*. Blood 71:766–770.

Feingold EA, Forget BG. 1989. The breakpoint of a large deletion causing hereditary persistence of fetal hemoglobin occurs within an erythroid DNA domain remote from the β-globin gene cluster. Blood 74:2178–2186.

Forget BG, Bunn HF. 2013. Classification of the disorders of hemoglobin. Cold Spring Harb Perspect Med 3:a011684.

Game L, Bergounioux J, Close JP, Marzouka BE, Thein SL. 2003. A novel deletion causing $(\varepsilon\gamma\delta\beta)^0$ thalassemia in a Chilean family. Br J Haematol 123:154–159.

Giardine B, Borg J, Higgs DR, Peterson KR, Philipsen S, Maglott D, Singleton BK, Anstee DJ, Basak AN, Clark B, Costa FC, Faustino P, et al. 2011. Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. Nat Genet 43:295–301.

Harteveld CL, Voskamp A, Phylipsen M, Akkermans N, Dunnen JT, White SJ, Giordano PC. 2005. Nine unknown rearrangements in 16p13.3 and 11p15.4 causing α- and β-thalassaemia characterised by high resolution multiplex ligation-dependent probe amplification. J Med Genet 42:922–931.

Higgs DR. 2013. The molecular basis of alpha-thalassemia. Cold Spring Harb Perspect Med 3:a011718.

Higgs DR, Engel JD, Stamatoyannopoulos G. 2012. Thalassaemia. Lancet 379:373–383.

Kassahn KS, Scott HS, Caramins MC. 2014. Integrating massively parallel sequencing into diagnostic workflows and managing the annotation and clinical interpretation challenge. Hum Mutat 35:413–423.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12:443–451.

Orkin SH, Old JM, Weatherall DJ, Nathan DG. 1979. Partial deletion of β-globin gene DNA in certain patients with β^0-thalassemia. Proc Natl Acad Sci USA 76:2400.

Pirastu M, Kan YW, Lin CC, Baine RM, Holbrook CT. 1983. Hemolytic disease of the newborn caused by a new deletion of the entire β-globin cluster. J Clin Investig 72:602–609.

Pritchard CC, Tait JF, Buller-Burckle AM, Mikula M. 2010. Annotation error of a common beta degrees -thalassemia mutation (619 bp-deletion) has implications for molecular diagnosis. Am J Hematol 85:978.

Rooks H, Bergounioux J, Game L, Close JP, Osborne C, Best S, Senior T, Height S, Thompson R, Hadzic N, Fraser P, Bolton-Maggs P, et al. 2005. Heterogeneity of the egdb thalassaemias: characterisation of 3 novel English deletions. Br J Haematol 128:722–729.

Rooks H, Clark B, Best S, Rushton P, Oakley M, Thein OS, Cuthbert AC, Britland A, Ruf A, Thein SL. 2012. A novel 506kb deletion causing $\varepsilon\gamma\delta\beta$ thalassemia. Blood Cells Mol Dis.

Shalev H, Landau D, Pissard S, Krasnov T, Kapelushnik J, Gilad O, Broides A, Dgany O, Tamary H. 2013. A novel epsilon gamma delta beta thalassemia presenting with pregnancy complications and severe neonatal anemia. Eur J Haematol 90:127–133.

Tewhey R, Warner JB, Nakano M, Libby B, Medkova M, David PH, Kotsopoulos SK, Samuels ML, Hutchison JB, Larson JW, Topol EJ, Weiner MP, et al. 2009. Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat Biotechnol 27:1025–1031.

Thein SL. 2013. The Molecular Basis of beta-Thalassemia. Cold Spring Harb Perspect Med 3:3/5/a011700 [pii] 011710.011101/cshperspect.a011700.

Traeger-Synodinos J, Harteveld CL. 2014. Advances in technologies for screening and diagnosis of hemoglobinopathies. Biomark Med 8:119–131.

Trent RJ, Williams BG, Kearney A, Wilkinson T, Harris PC. 1990. Molecular and hematologic characterization of Scottish-Irish type $(\varepsilon\gamma\delta\beta)^0$ thalassemia. Blood 76:2132–2138.

Zhang JD, Schindler T, Kung E, Ebeling M, Certa U. 2014. Highly sensitive amplicon-based transcript quantification by semiconductor sequencing. BMC Genomics 15:565.

lymphocytosis to those without del(17p), suggesting that the effect of T12 on lymphocytosis pattern is dominant; this is consistent with the prior observation that up-regulation of integrin signalling in T12 is not modulated by additional del (11q) or del(17p) (Riches *et al*, 2014). We did not have confirmatory integrin expression data. In contrast to the effect on lymphocytosis pattern, the presence of T12 in addition to del (17p) did not attenuate the adverse impact of del(17p) on outcomes. Further studies are required to ascertain the mechanisms underlying the abbreviated lymphocytosis in T12 CLL and its relationship, if any, to therapeutic efficacy.

Acknowledgements

The authors would like to acknowledge Ms. Susan Smith for assistance in data gathering and management.

Authorship

PT provided clinical care to patients, collected and analysed data, performed statistical analysis and wrote the paper. AF, SOB, WGW, MJK and JAB provided clinical care to patients, assisted in the analysis of data and development of critical themes and coauthored the paper.

Conflict of interest

SOB and JAB received research funding from Pharmacyclics.

Philip A. Thompson
Alessandra Ferrajoli
Susan O'Brien
William G. Wierda
Michael J. Keating
Jan A. Burger
Department of Leukemia, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.
E-mail: pathompson2@mdanderson.org

Keywords: chronic lymphocytic leukaemia, ibrutinib, lymphocytosis, trisomy 12, integrins

First published online 18 December 2014
doi: 10.1111/bjh.13269

References

Burger, J.A. (2011) Nurture versus nature: the microenvironment in chronic lymphocytic leukemia. *Hematology/The Education Program of the American Society of Hematology. American Society of Hematology. Education Program*, **2011**, 96–103.

Byrd, J.C., Furman, R.R., Coutre, S.E., Flinn, I.W., Burger, J.A., Blum, K.A., Grant, B., Sharman, J.P., Coleman, M., Wierda, W.G., Jones, J.A., Zhao, W., Heerema, N.A., Johnson, A.J., Sukbuntherng, J., Chang, B.Y., Clow, F., Hedrick, E., Buggy, J.J., James, D.F. & O'Brien, S. (2013) Targeting BTK with ibrutinib in relapsed chronic lymphocytic leukemia. *New England Journal of Medicine*, **369**, 32–42.

de Gorter, D.J., Beuling, E.A., Kersseboom, R., Middendorp, S., van Gils, J.M., Hendriks, R.W., Pals, S.T. & Spaargaren, M. (2007) Bruton's tyrosine kinase and phospholipase Cgamma2 mediate chemokine-controlled B cell migration and homing. *Immunity*, **26**, 93–104.

Furman, R.R., Sharman, J.P., Coutre, S.E., Cheson, B.D., Pagel, J.M., Hillmen, P., Barrientos, J.C., Zelenetz, A.D., Kipps, T.J., Flinn, I., Ghia, P., Eradat, H., Ervin, T., Lamanna, N., Coiffier, B., Pettitt, A.R., Ma, S., Stilgenbauer, S., Cramer, P., Aiello, M., Johnson, D.M., Miller, L.L., Li, D., Jahn, T.M., Dansey, R.D., Hallek, M. & O'Brien, S.M. (2014) Idelalisib and rituximab in relapsed chronic lymphocytic leukemia. *New England Journal of Medicine*, **370**, 997–1006.

Herishanu, Y., Perez-Galan, P., Liu, D., Biancotto, A., Pittaluga, S., Vire, B., Gibellini, F., Njuguna, N., Lee, E., Stennett, L., Raghavachari, N., Liu, P., McCoy, J.P., Raffeld, M., Stetler-Stevenson, M., Yuan, C., Sherry, R., Arthur, D.C., Maric, I., White, T., Marti, G.E., Munson, P., Wilson, W.H. & Wiestner, A. (2011) The lymph node microenvironment promotes B-cell receptor signaling, NF-kappaB activation, and tumor proliferation in chronic lymphocytic leukemia. *Blood*, **117**, 563–574.

Ponader, S., Chen, S.S., Buggy, J.J., Balakrishnan, K., Gandhi, V., Wierda, W.G., Keating, M.J., O'Brien, S., Chiorazzi, N. & Burger, J.A. (2012) The Bruton tyrosine kinase inhibitor PCI-32765 thwarts chronic lymphocytic leukemia cell survival and tissue homing in vitro and in vivo. *Blood*, **119**, 1182–1189.

Riches, J.C., O'Donovan, C.J., Kingdon, S.J., McClanahan, F., Clear, A.J., Neuberg, D.S., Werner, L., Croce, C.M., Ramsay, A.G., Rassenti, L.Z., Kipps, T.J. & Gribben, J.G. (2014) Trisomy 12 chronic lymphocytic leukemia cells exhibit upregulation of integrin signaling that is modulated by NOTCH1 mutations. *Blood*, **123**, 4101–4110.

Woyach, J.A., Johnson, A.J. & Byrd, J.C. (2012) The B-cell receptor signaling pathway as a therapeutic target in CLL. *Blood*, **120**, 1175–1184.

Woyach, J.A., Smucker, K., Smith, L.L., Lozanski, A., Zhong, Y., Ruppert, A.S., Lucas, D., Williams, K., Zhao, W., Rassenti, L., Ghia, E., Kipps, T.J., Mantel, R., Jones, J., Flynn, J., Maddocks, K., O'Brien, S., Furman, R.R., James, D.F., Clow, F., Lozanski, G., Johnson, A.J. & Byrd, J.C. (2014) Prolonged lymphocytosis during ibrutinib therapy is associated with distinct molecular characteristics and does not indicate a suboptimal response to therapy. *Blood*, **123**, 1810–1817.

First reported duplication of the entire beta globin gene cluster causing an unusual sickle cell trait phenotype

Heterozygotes for the *HBB* p.Glu7Val mutation (β^S gene) typically have ~40% haemoglobin S (HbS, $\alpha_2\beta_2^S$), and ~55% HbA ($\alpha_2\beta_2$) with normochromic normocytic red blood cells (RBCs). Co-inheritance of α thalassaemia with *HBB* p.Glu7Val reduces the amount of HbS in peripheral blood, as the normal β^A chains compete more effectively than β^S chains

for the limiting number of α globin chains. The relative amount of HbS reduces with the severity of co-existing α thalassaemia, paralleled by the reduction in mean cell volume (MCV) and mean cell haemoglobin (MCH) of the RBCs. Thus, AS individuals with (αα/α−) genotype have 30–35% HbS, and those with (α−/α−) genotype, have 25–30% HbS (Steinberg, 2001).

A 26-year-old West African woman underwent routine antenatal screening for haemoglobinopathies. Her blood counts showed Hb 119 g/l, RBC $4{\cdot}45 \times 10^9$/l, MCV 77·2 fl, MCH 26·8 pg. High performance liquid chromatography (BioRad variant II, Hercules, CA, USA) screen identified a haemoglobin variant in the HbS position comprising 13·5%, HbF 0·2% plus normal adult haemoglobins. When separated at acid pH the unknown variant migrated to the HbS position but this could not be confirmed by a solubility assay due to the low HbS percentage.

DNA was extracted from EDTA whole blood using a Qiagen Symphony midi kit (Qiagen, Hilden, Germany) and from a saliva sample (Oragene kit, Genotek, Ontario, Canada) using a Qiagen Virus kit (v2) on an EZ1 Biorobot (Qiagen). Genomic DNA was analysed for variants in the *HBA* and *HBB* genes, as described (Clark & Thein, 2004). Genetic testing showed that the individual had the $\alpha\alpha/\alpha-^{3{\cdot}7}$ genotype and no alpha thalassaemia variants. Sequence analysis of the *HBB* genes did not identify any β thalassaemia mutations that could have been *in cis* with the *HBB* p.Glu7Val mutation, which would have explained the unusually low HbS percentage. We confirmed the *HBB* p.Glu7Val mutation by TaqMan analysis but the relative proportions of the sickle and normal alleles in the gene sequencing and TaqMan assay suggested the presence of an additional functioning normal *HBB* allele.

A next generation sequencing (NGS) assay and bioinformatic strategy was used to identify any potential rearrangement in the *HBA* and *HBB* globin gene loci that could explain the unusually low HbS (Shooter *et al*, 2014). Leucocyte-extracted DNA (3 μg) was sheared to 500 bp using a Bioruptor (Diagenode, Denville, NJ, USA) and used to construct a sequencing library with minimal polymerase chain reaction (PCR) amplification. In-solution bait capture (Agilent, Santa Clara, CA, USA) target enrichment isolated DNA fragments from two genomic regions, 4 Mb on chromosome 11 encompassing the *HBB* cluster, and 260 kb on chromosome 16, covering the *HBA* globin loci. Sequencing was carried out on a MiSeq using a 500 cycle sequencing kit (V2) (Illumina, San Diego, CA, USA). Normal and positive controls with known genetic rearrangements were analysed in parallel with the test samples. All data analyses were carried out using NextGene (SoftGenetics, State College, PA, USA) as described (Shooter *et al*, 2014). A normalized measure of the coverage across the sequenced region, known as repeats per kilobase exon model per million mapped reads (RPKM), was calculated for each 120 bp region covered in the in-solution bait capture. By comparing the RPKM values between the patient sample and a normal control, duplicated regions are suggested by a 33% increase in aligned sequence, and heterozygous deletions, a 50% decrease.

Comparing the RPKM values of the patient sample to a normal control across the sequenced region on chromosome 16 confirmed the 3·7 kb deletion affecting one copy of the *HBA* cluster. A comparison on chromosome 11 showed a 145·7 kb duplication encompassing the entire *HBB* cluster (Fig 1A). Heterozygous single nucleotide polymorphisms (SNPs) in this region showed allele frequencies of ~33% or ~66% (reflecting dosage in one or two of three alleles respectively), as opposed to ~50% of each, which would be expected with two alleles present. The *HBB* p.Glu7Val mutation (rs334) was identified in a third of the reads, indicating that it is one of three functional *HBB* genes, consistent with the HbS level of 13·5%.

The RPKM coverage data indicated the approximate start and end points of the duplication (Fig 1A). Reads aligning to the reference sequence at these positions showed that the duplication was head to tail; the sequence repeating itself immediately after it finished, interrupted by 11 bp of novel sequence. The duplication break-points were confirmed by Gap-PCR and dye-terminator Sanger sequencing of the gap-PCR amplicon (Fig 1B). This sequence matched the NGS data exactly, confirming a novel duplication of the beta globin region, Hg19.Chr11 g.5372677_5372678insCACCTCCACTTdup5226885_5372677 (HbVar ID 2961). Gap PCR on the saliva DNA sample showed amplification of the duplication-specific PCR product (Fig 1B,C), confirming that this was a germline mutation, and excludes any haemopoietic mosaicism or other non-Mendelian inheritance, such as uniparental disomy (Joly *et al*, 2013).

Haemoglobinopathies are caused mainly by single base substitutions or deletions. Triplicated or quadruplicated *HBA* genes are the best described duplications and most clinically relevant (Origa *et al*, 2014). Recently, duplication of the whole *HBA* cluster has also been described (Harteveld *et al*, 2008). Although these duplicated genes are expressed, this is only clinically significant when co-inherited with beta thalassaemia when the extra α globin chains exacerbate the chain imbalance, increasing the ineffective erythropoiesis (Thein, 2008). Duplications in the *HBB* cluster are rare; the gene encoding anti-Lepore haemoglobin (e.g. Hb Miyada) which creates fusion products of the 5′ end of *HBB* gene and 3′ sequences of HBD (i.e. *HBD-HBB/HBD-HBB*) could be considered as a partial *HBB* duplication (Weatherall & Clegg, 2001).

This is the first description of the entire *HBB* gene cluster being duplicated and would have gone undetected during screening, if it had not been for the unusually low sickle percentage. Definitive analysis was helpful as it informed genetic counselling. This report demonstrates the importance of accurate interpretation of haemoglobin electrophoresis results in combination with genetic testing. A dataset

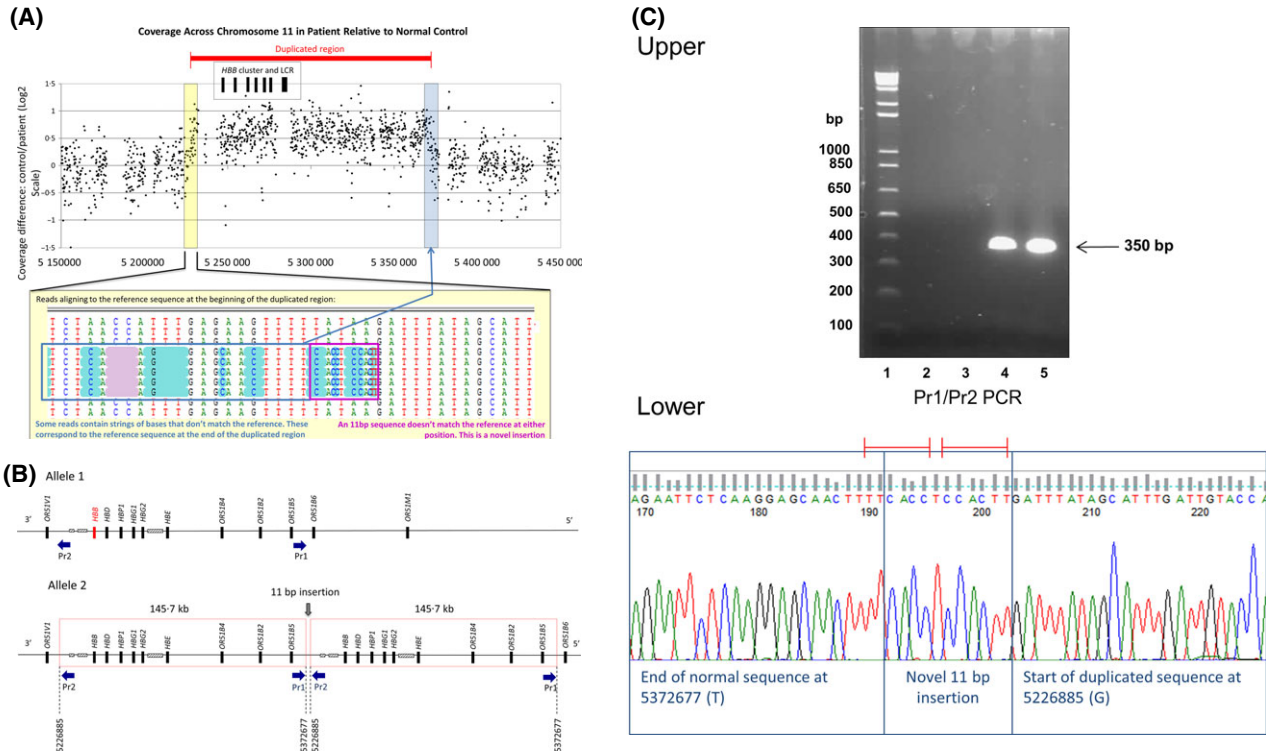


Fig 1. (A) Coverage across chromosome 11 in patient relative to normal control. Each black dot represents a 120 bp segment of chromosome 11 that is covered in our bait capture library. Gaps in the coverage arise due to repetitive regions which cannot be included in the bait library design. The difference in coverage between the proband and a normal control is represented on a Log2 scale on the *Y*-axis. A distinct region of ~140 kb is overrepresented in the sequence of the patient compared to the control (indicated by the horizontal red bar), which implies this area – which contains the entire beta globin gene cluster – has been duplicated in the patient. The duplicated region includes the whole *HBB* gene cluster [*HBB, HBD, HBBP1, HBG1, HBG2, HBE1* genes and beta LCR (locus control region)] as indicated below the red bar. The read pile-up in the NextGene Alignment Viewer shows the reference sequence at the top and aligned sequences below with strings of mismatching bases in some of the reads at this position (within the blue box). The mismatch begins at the same point and contains the same sequence in all affected reads. BLAT [Basic Local Alignment Search Tool (BLAST)-like Alignment Tool] query of the sequence from the point where this mismatch begins reveals these sequences align to the region at the end of the duplication indicated by the repeats per kilobase exon model per million mapped reads (RPKM). At this location, similar mismatch is seen in reads in the NextGene Viewer, where the misaligning sequences match to the duplication start point. This suggests that these reads have captured the duplication break point and that the duplication is 'head to tail' in orientation. Eleven bp of mismatching bases showed up in reads at both the start and end of the duplication (purple box). BLAT query could not match these 11 bp to anywhere in the genome, which appears to be a novel insertion at the start of the duplication, creating a 13 bp mirror repeat. (B) Line representation of the rearrangement on chromosome 11p in proband. Allele 1 represents the chromosome that contains the *HBB* cluster with the *HBB* p.Glu7Val mutation (*HBB* gene in red), and allele 2, the other chromosome, with a 145·7 kb duplication encompassing the entire normal *HBB* cluster. Solid blue arrows (Pr1 and Pr2) indicate the position and direction of the primers used to confirm the head-to-tail duplication. Amplification using Pr1 (5′-GCTGTATCCCCAACGTAGTATCCA-3′) at the end of the duplicated region, and Pr2 (5′-CTGCTTGGTTTTCTTCTTTCATTCAG-3′), at the start of the duplicated region produces a unique 350 bp PCR amplicon.(C) Upper panel - Gel picture showing the unique Pr1-Pr2 PCR product of 350 bp. The lanes are represented by (1) 1 kb ladder; (2) Blank; (3) Normal control; (4) Proband genomic DNA from peripheral blood; (5) Proband genomic DNA from saliva. The presence of the product in DNA extracted from both blood and saliva indicates that this is a germline mutation. Lower panel - Part of the chromatogram of the Pr1/Pr2 PCR amplicon. Dye-terminator sequencing of the unique PCR product confirmed the breakpoints indicated by NGS analysis and also the insertion of 11 bp of novel sequence situated between the duplicated regions. Red bars above the DNA sequence show the mirror repeat created by the novel insertion, centering around a 'T' base.

of haplotypes for the duplicated region (HapMap, www.hapmap.ncbi.nim.nih.gov) were compared to the haplotype of the proband at 49 SNP loci. Four individuals (one of Nigerian and three of Kenyan origin) shared the proband haplotype across the 122 kb region, but none carried the duplication. We suggest that such *HBB* duplications are rare, but more examples could be identified as multiplex ligation probe amplification and comparative genome hybridization are applied in screening methods (Traeger-Synodinos & Harteveld, 2014). However, this specific duplication is likely to be very rare, and perhaps unique to the proband.

Our NGS approach, which permitted to-the-base resolution of the breakpoint and orientation of the *HBB* duplication, should be applicable as a diagnostic tool for other disorders caused by structural rearrangements.

Acknowledgements

We thank Professor Chris Shaw, Ms Athinia Gkazi and Dr Bradley Smith at the Institute of Psychiatry, King's College London, for their assistance with the sequencing protocol. We also thank Claire Steward and Robin Swabey for help in preparation of the manuscript, and King's College Hospital Charity (KCHC) for support (CS is KCHC funded).

Authorship contribution

CS, TSM, MO, TJ and BC performed experiments; CS, BC and SLT analysed data and wrote the manuscript. All co-authors gave feedback on manuscript.

Disclosure of conflicts of interest

All authors declare no competing financial interests.

Claire Shooter[1]
Tania Senior McKenzie[2]
Matthew Oakley[2]
Tracey Jacques[2]
Barnaby Clark[1,2]
Swee Lay Thein[1,3]
[1]Molecular Haematology, Division of Cancer Studies, King's College London Faculty of Life Sciences & Medicine, [2]Department of Molecular Pathology, Viapath at King's College Hospital NHS Foundation Trust, and [3]Department of Haematology, King's College Hospital NHS Foundation Trust, London, UK
E-mail: sl.thein@kcl.ac.uk; swee.thein@nhs.net

Keywords: next generation sequencing, globin genes, *HBB* duplication

First published online 18 December 2014
doi: 10.1111/bjh.13274

References

Clark, B.E. & Thein, S.L. (2004) Molecular diagnosis of haemoglobin disorders. *Clinical and Laboratory Haematology*, **26**, 159–176.

Harteveld, C.L., Refaldi, C., Cassinerio, E., Cappellini, M.D. & Giordano, P.C. (2008) Segmental duplications involving the alpha-globin gene cluster are causing beta-thalassemia intermedia phenotypes in beta-thalassemia heterozygous patients. *Blood Cells, Molecules, & Diseases*, **40**, 312–316.

Joly, P., Schluth-Bolard, C., Lacan, P., Barro, C., Pissard, S., Labalme, A., Sanlaville, D. & Badens, C. (2013) HBB loss of heterozygosity in the hemopoietic lineage gives rise to an unusual sickle-cell trait phenotype. *Haematologica*, **98**, e7–e8.

Origa, R., Sollaino, M.C., Borgna-Pignatti, C., Piga, A., Feliu Torres, A., Masile, V. & Galanello, R. (2014) alpha-globin gene quadruplication and heterozygous beta-thalassemia: a not so rare cause of thalassemia intermedia. *Acta Haematologica*, **131**, 162–164.

Shooter, C., Rooks, H., Thein, S.L. & Clark, B.. (2014) Next Generation Sequencing Identifies a Novel Rearrangement in the HBB Cluster Permitting to-the-Base Characterization. *Human Mutation*. doi: 10.1002/humu.22707 [Epub ahead of print]

Steinberg, M.H. (2001) Sickle cell trait. In: *Disorders of Hemoglobin: Genetics, Pathophysiology, and Clinical Management* (eds. by M.H. Steinberg, B.G. Forget, D.R. Higgs & R.L. Nagel), pp. 811–830. Cambridge University Press, Cambridge, UK.

Thein, S.L. (2008) Genetic modifiers of the beta-haemoglobinopathies. *British Journal of Haematology*, **141**, 357–366.

Traeger-Synodinos, J. & Harteveld, C.L. (2014) Advances in technologies for screening and diagnosis of hemoglobinopathies. *Biomarkers in Medicine*, **8**, 119–131.

Weatherall, D.J. & Clegg, J.B. (2001) *The Thalassaemia Syndromes*. Blackwell Science, Oxford.

Ibrutinib-associated lymphocytosis corresponds to bone marrow involvement in mantle cell lymphoma

Bruton tyrosine kinase (BTK) is involved in the signalling pathway of the B-cell receptor (BCR). Signal regulation within this pathway is critical for the maturation and functioning of normal B cells (Gauld *et al*, 2002). Functional BCR signalling has been shown to be necessary for maintaining the viability of malignant B cells in varied non-Hodgkin lymphoma subtypes (Davis *et al*, 2010; Cinar *et al*, 2013). Ibrutinib is a selective inhibitor of the BTK protein that irreversibly binds to cysteine-481 within the active site (Honigberg *et al*, 2010). *In vitro*, ibrutinib blocks B-cell activation, arrests cell growth and induces apoptosis in human B-lymphocyte cell lines (Davis *et al*, 2010; Honigberg *et al*, 2010; Cinar *et al*, 2013). In clinical trials, ibrutinib has exhibited significant activity in certain B-cell malignancies, including chronic lymphocytic leukaemia (CLL) and mantle cell lymphoma (MCL) (Advani *et al*, 2013; Byrd *et al*, 2013; Wang *et al*, 2013).

Peripheral blood lymphocytosis was a significant feature observed with ibrutinib therapy during phase 1 and 2 studies in CLL (Byrd *et al*, 2013) and MCL (Wang *et al*, 2013). In a phase 2 study of single-agent ibrutinib (PCYC-1104-CA; NCT01236391), 34% of the enrolled patients with MCL ($n = 111$) experienced a transient increase (to $\geq$50% of baseline and $>5 \times 10^9$/l) in peripheral absolute lymphocyte count (ALC). This lymphocytosis peaked at a median of 4 weeks of treatment and gradually decreased over 15–18 weeks of therapy (Wang *et al*, 2013). The circulating lymphocytes were characterized as