Investigating the heterogeneity of selective pressures and driver events in cancer

Nulsen, Joel

Awarding institution:
King's College London

Joel Robert Nulsen

Investigating the heterogeneity of selective pressures and driver events in cancer

Submitted in application for a Doctor of Philosophy degree in Computational Biology

To my father, whose love got me this far.

Abstract

Inter-tumour heterogeneity is a significant barrier to the effective treatment of cancer. Heterogeneous molecular features of tumours cause patients to respond differently to therapies. The cancer community has addressed this problem to an extent by developing molecular stratifications and targeted therapies. However, formulating effective treatment strategies for individual patients remains challenging, and a deeper molecular understanding of inter-tumour heterogeneity is required to improve patient outcomes.

In this thesis, I investigate inter-tumour heterogeneity from two perspectives. First, I consider germline variation as a driving force behind inter-tumour heterogeneity. While heterogeneity is largely the result of stochastic processes, inherited genetic differences between patients can give rise to patient-specific selective pressures acting on somatic alterations during cancer evolution. However, this phenomenon is not yet fully understood. I analyse how germline variants that perturb the function of biological pathways affect the frequency of somatic driver alterations at the gene and pathway levels, using data from oesophageal adenocarcinoma. By addressing the methodological and statistical challenges involved in this analysis, I find evidence that *ATM* and its interactors play an important and as-yet unreported role in the biology of oesophageal adenocarcinoma. In particular, I find that perturbations to these genes can substitute for driver alterations in *TP53*, which is by far the most frequently altered gene in oesophageal adenocarcinoma. This analysis also uncovers evidence that *ATM* acts as a cancer predisposition gene and a tumour suppressor gene in oesophageal adenocarcinoma.

Second, I address the question of how to identify the aspects of inter-tumour heterogeneity that are most relevant to cancer biology and therapy, *i.e.* cancer drivers. The research community has identified many hundreds of driver genes across cancer types, and I describe the curation of a database, the Network of Cancer Genes (NCG), to capture this information. NCG also annotates the systems-level properties of reported cancer genes. I show that cancer genes are distinguished from other human genes by an array of these systems-level properties, and develop a machine learning method to use these properties to identify novel driver genes. This method (sysSVM2)

is capable of identifying driver genes at the level of individual patients, which overcomes the persistent problem of a portion of patients having too few driver alterations to explain the onset of cancer. It is also particularly useful in rare cancer types for which large-scale sequencing studies are infeasible. Using the properties of canonical driver genes to identify drivers in individual patients in this way can help to further the goals of precision oncology and overcome the challenge presented by inter-tumour heterogeneity.

Taken as a whole, this thesis presents novel research that sheds light on both the causes of inter-tumour heterogeneity and how to interpret the heterogeneous molecular landscape of cancer.

List of published research articles

Nulsen, J., Misetic, H., Yau, C. & Ciccarelli, F. D. Pan-cancer detection of driver genes at the single-patient resolution. *Genome Medicine* **13** (2021).

Nulsen, J., Repana, D., Dressler, L., Bortolomeazzi, M., Venkata, S. K., *et al.* The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biology* **20** (2019).

Mourikis, T., Benedetti, L., Foxall, E., Temelkovski, D., Nulsen, J. *et al.* Patient-specific cancer genes contribute to recurrently perturbed pathways and establish therapeutic vulnerabilities in esophageal adenocarcinoma. *Nat. Comm.* **10** (2019).

Table of contents

List of supplementary tables (Chapter 5)

List of abbreviations

Abbreviation	Definition
ARI	Adjusted Rand Index
AUC	Area under the receiver operating characteristic curve
BRCA	Breast cancer
CGC	Cancer Gene Census
CNV	Copy number variant
CPG	Cancer predisposition gene
DDR	Discoidin domain receptor
DLBC	Diffuse large B-cell lymphoma
DSB	Double strand break
ECM	Extracellular matrix
EUR	European samples from the 1000 Genomes Project (503 individuals)
FDR	False discovery rate
FWER	Family-wise error rate
GoF	Gain of function
GI	Gastro-intestinal
GSEA	Gene set enrichment analysis
GSIT	Germline-somatic interaction *in trans*
GWAS	Genome-wide association study
ICGC	International Cancer Genome Consortium
Indel	Insertion or deletion
LoF	Loss of function
LRT	Likelihood ratio test
MAF	Minor allele frequency
NCG	Network of Cancer Genes
NN	Neural network
OAC	Oesophageal adenocarcinoma
OG	Oncogene
OGLLR	Overlapping group lasso logistic regression
OR	Odds ratio
PAM	Partitioning around medoids

PC	Principal component
PCA	Principal components analysis
PPIN	Protein-protein interaction network
PTV	Protein-truncating variant
QQ	Quantile-quantile
RBO	Rank-biased overlap
RDGV	Rare damaging germline variant
RF	Random forest
ROC	Receiver operator characteristic
RSS	Residual sum of squares
RTK	Receptor tyrosine kinase
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SV	Structural variant
SVM	Support vector machine
TCGA	The Cancer Genome Atlas
TMB	Total mutational burden
TPM	Transcripts per million
TSG	Tumour suppressor gene
UTR	Un-translated region
VAE	Variational autoencoder
vGWAS	Voxelwise genome wide association study
VWJI	Variance-weighted Jaccard index

Chapter 1. Introduction

1.1. Cancer as a heterogenous genetic disease

"All happy families are alike; each unhappy family is unhappy in its own way" – Leo Tolstoy, *Anna Karenina*

Over the past century, cancer has risen to become a leading cause of death in many parts of the world. In 2016, it accounted for 29.8% of global premature deaths (age 30-69 years) from noncommunicable diseases[1]. The picture is even more stark in countries with high Human Development Indexes such as those in Western Europe and North America, where in 2016 cancer was the single leading cause of premature death[1]. While therapeutic advances have improved patient survival in many cancer types in recent decades[1], more work needs to be done to continue to improve clinical outcomes.

Medical advances in oncology have largely followed our increased understanding of the molecular basis of cancer. It has been known since the 1970's that cancer is caused by genetic alterations acquired randomly by cells over the course of their life[2] (somatic alterations). While the majority of such alterations have little or no phenotypic effect on cells, some confer a selective advantage, allowing cells to proliferate faster than their neighbours (Figure 1-1). Over time as these cells divide, they acquire more somatic alterations, and evolution selects for those subpopulations whose alterations give them the greatest proliferative advantages. These advantages can take a number of different forms, labelled the 'hallmarks of cancer' by Hanahan & Weinberg in 2000[3], and expanded upon in 2011[4]. They include: genomic and mutational instability; sustained proliferative signalling; evading growth suppressors; resisting cell death; avoiding immune destruction; enabling replicative immortality; activating invasion and metastasis; inducing angiogenesis; tumour-promoting inflammation; and deregulating cellular energetics[4]. The somatic alterations that contribute to these hallmarks, and thus ultimately to the initiation and progression of cancer, are known as driver alterations. Most driver alterations contribute to the hallmarks of cancer by causing a gene to either gain or lose functionality. Genes whose gain-of-function (GoF) drives cancer are referred to as oncogenes, while genes whose loss-of-function (LoF) drives cancer are called tumour suppressor genes (TSGs). Conversely, the alterations that do not contribute to cancer are termed passenger alterations, and genes whose alteration does not drive cancer are called passenger genes[5].

Due to the stochastic nature of cancer evolution and the vast number of possible alterations to the human genome, every cancer is genetically distinct (Figure 1-1). This phenomenon is termed inter-tumour heterogeneity, and it represents a significant challenge to oncology, because cancers with different genetic features often respond differently to therapy. In some cancer types, inter-tumour heterogeneity has been partially addressed by stratifying patients into broad categories based on molecular biomarkers. For example, it has been known since 1987 that breast cancers overexpressing the cell-surface receptor and oncogene HER2 are particularly aggressive[6]. Today, breast cancer patients are stratified by the presence of absence of HER2[7,8], and the resulting subtypes have different treatment guidelines. However, such stratifications are lacking in many cancer types. Moreover, even within established subtypes (such as those in breast cancer[9]) there is substantial variation in response to therapy between individual patients.

Since the advent of economical omic-scale DNA and RNA sequencing technologies, precision oncology has become a focus the cancer research community as a way to comprehensively address inter-tumour heterogeneity[10,11]. The core concept of precision oncology is the tailoring of cancer treatment to the molecular characteristics of individual tumours. It encompasses a variety of strategies, including improved stratification and targeting altered genes or pathways with specific drugs. Moreover, the biomarkers used to distinguish tumours can come from an array of sources, including DNA, RNA, and proteins[11]. Returning to the example of breast cancer, tumours that overexpress HER2 can be effectively treated with anti-HER2 monoclonal antibodies (such as trastuzumab[12]), and the introduction of these precision medicines has resulted in much-improved prognoses[13]. This development was one of the early successes of precision oncology, driving subsequent efforts in other cancer types and with other biomarkers.

However, the example of trastuzumab also illustrates how a deep understanding of the biology hidden behind inter-tumour heterogeneity is necessary for progress in precision oncology. In that example, improving patient outcomes first required identifying overexpression of HER2 as a molecular driver of a subset of breast cancers, before HER2-specific drugs could be developed. More generally, the advancement of precision oncology first requires us to be able to identify what genetic alterations drive cancer in individual patients. Once that has been accomplished, a detailed mechanistic understanding of how driver alterations function may be required to enable effective drug development, particularly for alterations that change protein function, such as amino acid substitutions. Even once these questions have been addressed and new precision therapies have been developed, however, the onset of resistance is likely to hamper progress in patient outcomes, so any single drug is unlikely to be sufficient in any tumour[14]. These stages in the development of precision oncology treatments represent substantial challenges to the cancer research community.

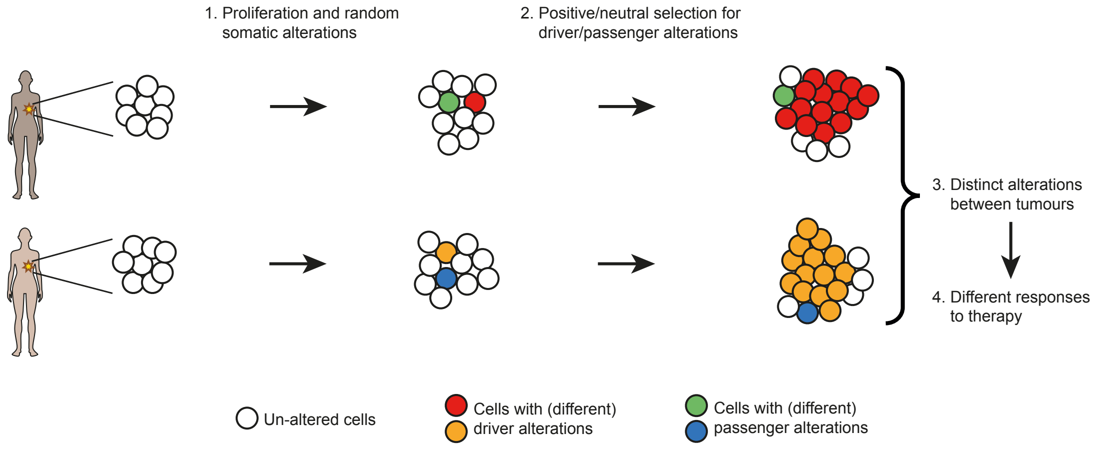
1. Proliferation and random somatic alterations
2. Positive/neutral selection for driver/passenger alterations
3. Distinct alterations between tumours
4. Different responses to therapy
○ Un-altered cells
● Cells with (different) driver alterations
● Cells with (different) passenger alterations

Figure 1-1: Inter-tumour heterogeneity

Cells randomly acquire somatic alterations as they divide. Driver alterations undergo positive selection, while passenger alterations are not selected for. As a result of the random process of acquiring somatic alterations, different tumours have distinct sets of somatic alterations, leading to different therapeutic requirements.

1.2. Lifting the mask of inter-tumour heterogeneity

In this thesis, I will address two primary research questions. Both are concerned with using cancer genomics data to investigate the heterogeneous nature of cancer driver events. The first (addressed in Chapters 2 and 3) asks how germline genetic variation promotes inter-tumour heterogeneity, by influencing the selection of somatic driver events. The second (addressed in Chapters 4 and 5) asks how to identify the aspects of inter-tumour heterogeneity that are the most relevant to cancer biology, *i.e.* the driver events, in individual tumours.

As discussed above, inter-tumour heterogeneity arises naturally from cancer evolution, as selection acts on randomly acquired somatic alterations. While the forces of selection are deterministic, the stochasticity of the acquisition of somatic alterations results in tumours having different sets of driver alterations. However, while much of this heterogeneity is simply due to the underlying stochastic process, there are context-specific factors that influence the selective pressures acting on driver alterations, and that thus constitute deterministic components of inter-tumour heterogeneity. Perhaps the most obvious example of such a factor is tissue type. Analysis of large pan-cancer cohorts has demonstrated very clearly that tumours arising from different tissues are characterised by different sets of recurring driver alterations[15,16]. For example, a study from The Cancer Genome Atlas (TCGA) found several driver genes that were recurrently altered only in gastrointestinal cancers, including the tumour suppressor gene *APC*[15]. This suggests that the selective pressures acting in different tissues favour different driver events. There is also evidence that environmental factors can influence the selection of driver events. For example, in lung adenocarcinoma tobacco smoking is positively associated with driver alterations in *KRAS*, and negatively associated with those in *EGFR*[17]. Inherited genetic (germline) variation represents a third possible factor in determining the selective pressures acting on driver alterations.

It is well-known that some germline variants can predispose individuals to developing certain types of cancer[18]. For example, inherited variants that lead to LoF of the *APC* gene are the main cause of familial adenomatous polyposis, a syndrome which strongly predisposes patients to colorectal cancer[19]. However, it has recently been proposed that germline variants might also act as oncogenic modifiers[20] that alter which driver events are positively selected for in the tumour. In Chapter 2, I will review the existing literature concerning the influence of germline variation on cancer evolution. I will then introduce a statistical framework to investigate this phenomenon in oesophageal adenocarcinoma, a cancer type with poor survival and minimal evidence for inherited predisposition[21-23]. In so doing, I will address some of the technical challenges posed by relating germline variants to somatic driver events, due to the high-dimensional nature of the data. In Chapter 3, I will describe a more sophisticated modelling approach that I applied to an expanded cohort, in order to address some of the limitations of the analysis in Chapter 2. In this way, I will uncover evidence of a hitherto unreported role for germline damage to *ATM* in oesophageal adenocarcinoma.

In addition to asking what factors influence the selection of driver events, I will investigate what somatic alterations can constitute driver events. Indeed, the search for genes whose alterations promote cancer (driver genes) has been a major focus of cancer genomics since its inception[15,16,24]. Many efforts to identify driver genes have focused on the recurrence of somatic alterations across cohorts of patients, since genes under strong selective pressure are likely to be altered at higher frequencies than expected by chance. Such approaches have successfully identified hundreds of driver genes[15,16,24]. However, when these genes are mapped back to patients, a fraction of patients are left with few or no driver events, even in the largest and most comprehensive studies[16]. Thus, these approaches are insufficient to identify the drivers in every patient, as is necessary for precision oncology. Bert Vogelstein famously described the landscape of selection in cancer as consisting of “mountains” and “hills”, where the hills are the numerous genes that exhibit only weak signs of selective pressure[25]. The inability of established methods to identify the genes driving every cancer suggests that other approaches are required to investigate the “hills” of the cancer genome. As a starting point to address this problem, in Chapter 4 I will introduce a database maintained by the Ciccarelli lab called the Network of Cancer Genes (NCG)[24,26]. NCG is a repository of cancer driver genes reported in the literature, including both well-established canonical drivers and candidate drivers found by cancer sequencing screens. NCG also annotates a wide range of gene properties that reflect genes’ systems-level biological role[24]. In Chapter 4, I will describe these systems-level properties and demonstrate that cancer genes are distinguished from other genes by these properties. I will also describe my contribution to the update of NCG to its sixth version. I will then use these properties of cancer genes to investigate the “hills” of the cancer genome in Chapter 5. To achieve this, I will present the development and benchmarking of a machine learning tool, sysSVM2, that uses systems-level properties to identify driver genes in individual patients. sysSVM2 is a new version of an existing cancer type-specific method[27], optimised for pan-cancer use.

In presenting this thesis, I hope to contribute to our understanding of the heterogeneous nature of cancer, and ultimately to how we might use this understanding to improve patient outcomes.

1.3. Intra-tumour heterogeneity

Inter-tumour heterogeneity is not the only type of heterogeneity seen in cancer. As molecular profiling of cancer has become more widespread in research, another important form of heterogeneity has come under increased scrutiny: intra-tumour heterogeneity[28,29]. Whereas inter-tumour heterogeneity refers to the differences between tumours, intra-tumour heterogeneity refers to the differences between cell populations within individual tumours.

These differences arise because the process of cancer evolution continues after the malignant transformation of the initial cell of origin of a tumour (Figure 1-2). The initial cell becomes cancerous after it has acquired sufficient driver alterations, as described above. These alterations are passed on to all daughter cells, and thus become common to all cancer cells in the tumour; such alterations are called clonal or truncal. However, as the daughter cancer cells proliferate, they continue to randomly acquire further somatic alterations (Figure 1-2). Most of these are passenger alterations that are not subject to selective pressures. However, a small subset of alterations confer additional selective advantages to already malignant cells, allowing them to proliferate faster than their neighbours. Over time, cell populations with these additional driver alterations can grow to make up a substantial proportion of the tumour (Figure 1-2). As a result, by the time that tumours present clinically, they often comprise multiple distinct subpopulations characterised by different sets of sub-clonal alterations[28,29].

Intra-tumour heterogeneity has important clinical implications in cancer. It is believed to be a major source of therapy resistance[29,30]. While the majority of cancer cells may be vulnerable to an initial therapy, a small minority may have somatic alterations that make them resistant. These cells can therefore survive the therapy, and go on to proliferate and form new tumours or metastases. Since all daughter cells then have the resistance alterations, these relapsed tumours are resistant to the initial therapy. Thus, resistance arises because the molecular heterogeneity between cancer cells is sufficient for some cells to be resistant to any given therapy[29,30]. Recent evidence also suggests that intra-tumour heterogeneity is relevant to response to immunotherapy, as responders to immune checkpoint blockade in non-small cell lung cancer have high burdens of clonal immunogenic mutations[31,32].

Motivated by the clinical importance of intra-tumour heterogeneity, research has considerably deepened our understanding of intra-tumour heterogeneity in recent years. For example, truncal driver alterations are relatively constrained compared to subsequently-acquired sub-clonal alterations, which are substantially more diverse across cancers[33]. Perhaps unsurprisingly, the extent of heterogeneity in an individual tumour is strongly predicted by genomic instability[34]. However, it appears that the combination of both chromosomal and mutational instability is particularly potent for generating intra-tumour heterogeneity[34]. There is also a large body of experimental evidence indicating that heterogeneous cell populations within a tumour can cooperate with each other to promote tumour growth and metastasis[35-37]. Interrogating intra-tumour heterogeneity presents technical challenges, however. Sub-clonal resolution in sequencing data is limited by sequencing depth, with high depths required to reliably identify alterations in small populations of cells[38,39]. Additional algorithms are also required to infer the clonal and/or phylogenetic makeup of tumours from sequencing data, but obtaining non-simulated ground truth data for benchmarking such algorithms is challenging[38,40,41]. Finally, inferring sub-clonal copy number alterations is particularly problematic, although algorithms have been developed for this purpose[42,43].

In this thesis, I will not focus on intra-tumour heterogeneity. While it is clearly of clinical and biological importance in cancer, incorporating intra-tumour heterogeneity adds complexity to any analysis. This is only beneficial if the data are sufficient to support this additional complexity. For example, it may be the case that the clonality of certain somatic alterations is influenced by particular germline variants. However, analysing clonality when investigating how the germline influences somatic evolution adds to model complexity in what is already a high-dimensional problem, as I will discuss in Chapter 2 and Chapter 3. In the absence of sufficiently large cohorts, considering intra-tumour heterogeneity in this context would therefore probably confound analyses due to lack of statistical power. By contrast, in identifying patient-level driver alterations, it may well be the case that certain drivers are more important in certain contexts when they are clonal or sub-clonal. An analysis of this hypothesis would represent an extension of the work described in Chapter 5, which instead seeks to develop a framework for driver gene identification that is applicable to any cancer type. Only once such a framework has been developed would it be appropriate to consider the additional complexity of intra-tumour heterogeneity. Thus, I will instead consider the problem of inter-tumour heterogeneity in this thesis.

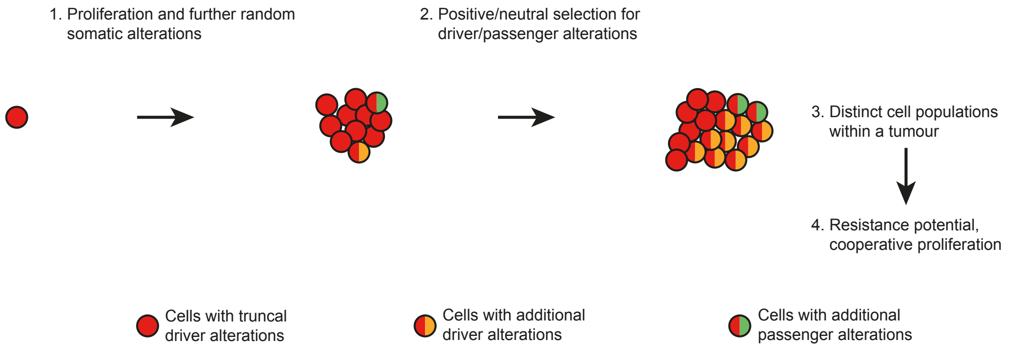
1. Proliferation and further random somatic alterations
2. Positive/neutral selection for driver/passenger alterations
3. Distinct cell populations within a tumour
4. Resistance potential, cooperative proliferation
Cells with truncal driver alterations
Cells with additional driver alterations
Cells with additional passenger alterations

Figure 1-2: Intra-tumour heterogeneity

An initial cancer cell with (truncal) driver alterations proliferates, and daughter cells randomly acquire further somatic alterations. Cells with additional driver alterations can out-proliferate their neighbours and go on to make up substantial portions of the tumour. As a result, individual tumours are composed of cell populations with distinct somatic alterations, leading to potential resistance to therapy and cooperation between populations to driver proliferation.

Chapter 2. A co-occurrence model for germline influence on cancer evolution

Inter-tumour heterogeneity is a natural consequence of the stochastic nature of cancer evolution. Although certain genomic loci are altered at a higher frequency than others due to chromatin state, transcription-coupled repair and residue-biased mutational processes, somatic alterations to the genome are largely acquired at random[44,45]. Since they are only subsequently subjected to selective pressure, it is unsurprising that each cancer evolves differently. However, there are some factors which drive this heterogeneity by modifying the selective pressures themselves. In this chapter, I will undertake an initial investigation of the hypothesis that perturbative germline variation (*i.e.* variation that damages a biological process) can influence selective pressures on somatic driver alterations in cancer.

I will first review the existing literature concerning this and related hypotheses, and describe how my research relates to the existing body of work (Section 2.1). I will then detail the preparation of germline and somatic data for a cohort of 260 oesophageal adenocarcinomas (OACs, 2.3). Given these data, I will describe the application of a co-occurrence-based method for identifying germline-somatic interactions that attempts to manage the available statistical power (2.4). This approach yields two distinct results, which I will investigate further using orthogonal approaches. Analysis in a validation cohort of 140 additional OACs does not reproduce these results, and I will discuss possible reasons why this is the case. I will end this chapter with a discussion of the limitations of the methods and data used (2.5), motivating the larger-scale analysis undertaken with more complex statistical modelling in Chapter 3.

2.1. Introduction and literature review

In this section, I will first review the theoretical basis for germline-somatic interactions in cancer (2.1.1). I will then review recent studies that systematically investigate germline-somatic interactions, and compare their approaches to those that I will describe in Chapters 2 and 3 of this thesis (2.1.2). Finally, I will briefly review several studies that investigate germline-interactions from a perspective that is different but complementary to the one taken in this thesis (2.1.3).

2.1.1. A developing view of germline-somatic interactions in cancer

Germline variation has long been known to play a role in cancer, but this has mostly been understood in terms of cancer predisposition. Predisposition to cancer can be carried either by high-penetrance variants whose effects are strong enough to identify them individually, or by diffuse low-impact variants whose effects can only be inferred from family studies[18,46,47]. Estimates of cancer heritability from twin studies indicate that low-impact variants explain approximately 30% of cancer incidence[46,47]. However, throughout this work I will focus on variants with a putatively damaging effect on protein function, since these can often be better understood mechanistically. Many high-penetrance cancer predisposition variants fall in cancer predisposition genes[18,48] (CPGs). Understanding cancer predisposition is clearly of clinical value in its own right, for example for improved monitoring and early detection in affected individuals[18,49]. However, the influence of germline variation on somatic evolution in cancer is less well-studied.

Aside from germline variation in general, cancer predisposition has nevertheless been known to be linked with the somatic evolution of cancer for some time. Indirect evidence for this link comes from the substantial overlap between CPGs and somatic driver genes. Indeed, Rahman[18] identified 114 CPGs for a variety of cancer types from the literature, of which 49 (43%) were also established somatic driver genes. Wang *et al.*[50] investigated this overlap at the pathway level for lung cancer, finding that both CPGs and somatic driver genes were enriched in cell cycle, TP53 signalling and TGF-β signalling pathways, among others. Zhang *et al.*[51] carried out a similar analysis in gastric cancer, finding common enrichment in pathways including insulin signalling and the PI3K cascade. These results suggest that germline predisposition to cancer and somatic driver events share some common mechanisms.

Direct evidence for the narrow case of damaging germline variants influencing somatic evolution *in cis* (where both germline and somatic variants are in the same gene) is also well-established. Indeed, the earliest example of this came from Alfred Knudson's two-hit hypothesis in 1971[52]. Based on observations in retinoblastoma, Knudson hypothesised that the disease was caused by two mutations, the first of which was carried in the germline of some individuals. In 1983, Godbout *et al.*[53] showed that the germline and somatic mutations both inactivated the same gene (*RB1*). While this explained the mechanism of predisposition and led to *RB1* being classified as a tumour suppressor gene[54] (TSG), it also illustrated that the germline could influence somatic evolutionary pressures. In this case, the germline hit to *RB1* greatly increased the selective advantage of the subsequent somatic hit to the same gene. Similar second hit effects have since been observed in many TSGs[49,55]. However, germline-somatic interactions *in cis* are not restricted to TSGs, with effects also reported in oncogenes including *JAK2*[56], *EGFR*[57] and *ERBB4*[58]. This hints at relatively complex genetic interplay between germline variants and somatic alterations in cancer.

More recently, there have been systematic efforts to identify germline-somatic interactions *in trans* (GSITs), a selection of which is shown in Table 2-1. The hypothesis investigated by these studies is that certain germline variants can affect the selective pressures acting on somatic alterations in completely different genes. These investigations have become feasible with the availability of large cohorts of cancer patients with genomic sequencing data, such as those provided by The Cancer Genome Atlas (TCGA, https://www.cancer.gov/tcga) and the International Cancer Genome Consortium (ICGC, https://icgc.org). Lu *et al.*[55] first used TCGA data to investigate GSITs in 2015, and more studies have followed since. However, even with the availability of large cohorts, relating germline variation to somatic alterations is non-trivial. Both the germline and somatic landscapes of cancer are complex[15,16,18,25,49,55], so relating the two to each other suffers from a multiplicity of complexity. Researchers must choose how to characterise both germline and somatic variation, *i.e.* what aspects of germline and somatic variation to investigate, in any study. For example, in the germline one might only look at cancer predisposing variants or genes, and at the somatic level one might consider genes with putative driver alterations. Table 2-1 summarises these choices made in existing GSIT studies, as well as in this thesis. I will now discuss these choices, and how they relate to my research in Chapters 2 and 3.

First author	**Year**	**Germline characterisation**	**Somatic characterisation**	**Cancer type**
Lu	2015	Rare truncations in predisposition genes	Driver genes	Pan-cancer
Carter	2017	Common SNPs	Driver genes	Pan-cancer
Agarwal	2017	Damaging SNPs	Driver genes	Breast
Puzone	2017	Predisposition SNPs	Driver genes	Breast
Wang	2018	Predisposition SNPs	Driver genes/pathways	Lung
Zhang	2018	Predisposition SNPs	Driver genes/pathways	Gastric
This thesis		Rare damaging variants at pathway level	Driver genes/pathways	OAC

Table 2-1: Existing studies of germline-somatic interactions *in trans*
SNP: single nucleotide polymorphism. OAC: oesophageal adenocarcinoma.

2.1.2. Current research into germline-somatic interactions *in trans*

Many GSIT studies have characterised the germline in terms of variants or genes that are known to predispose to cancer. Indeed, several have only investigated known predisposition single nucleotide polymorphisms (SNPs)[50,51,59], while Lu *et al.*[55] analysed truncating variants in CPGs (Table 2-1). However, focusing on predisposition is a restrictive choice. While predisposition variants are more likely to play a role in cancer development than others, many potentially relevant germline variants will be excluded from a study. Some researchers have looked at broader scopes of germline variation. Carter *et al.*[60] used all common SNPs (minor allele frequency, MAF, >1%) in a genome-wide association study (GWAS)-style approach. Interestingly they found many results that were unrelated to CPGs, confirming that restricting analysis to known CPGs discards relevant information. The GWAS-style approach taken by Carter *et al.* can be considered to be the opposite extreme of restricting to predisposition variants. A middle ground was taken by Agarwal *et al.*[20], who used CADD[61] annotations to restrict their analysis to the top 1% most damaging germline variants. As described in Section 2.3, I will take a similar approach in this thesis, focusing on rare damaging germline variants. This represents a balance between being overly restrictive by only using known predisposition variants, and overly permissive as in GWAS. Additionally, I will characterise germline variation at the pathway level. Despite not being done in any GSIT studies to date, pathway-level analysis is frequently used in case-control studies to characterise germline variation[22,62]. I will show in Section 2.4 that aggregating rare damaging variants into pathways substantially reduces the burden of multiple hypothesis testing and increases statistical power.

Somatic landscapes in GSIT studies have almost exclusively been characterised in terms of gene-level driver alterations (Table 2-1), in contrast to the range of approaches taken to germline characterisation. This is a sensible restriction, since the selective pressure on a gene is only likely to be influenced by the germline if that gene contributes to cancer. Moreover, driver genes are of the greatest clinical interest for patient stratification and providing potential drug targets. Wang *et al.*[50] and Zhang *et al.*[51] both extended this characterisation to the pathway level, although they restricted their analysis to truncating somatic mutations. This restriction is counter-productive, since it removes gain-of-function (GoF) alterations in oncogenes, and thus excludes many true drivers from their pathway analysis. I will investigate somatic driver alterations both at the pathway level (Chapter 2) and at the gene level (Chapter 3), integrating both mutation and copy number data to obtain as comprehensive a characterisation as possible.

A range of cancer types has been investigated by GSIT studies (Table 2-1). Both Lu *et al.*[55] and Carter *et al.*[60] carried out pan-cancer analyses. However, Lu *et al.* argued that a pan-cancer analysis can be confounding, since the frequencies of germline predisposition variants and somatic drivers differ across cancer types. The authors therefore carried out cancer type-specific analysis to mitigate this effect. Carter *et al.* instead controlled for cancer type as a covariate in a logistic regression model. Other studies have looked at individual cancer types, including breast, gastric and lung cancers. I will focus on oesophageal adenocarcinoma (OAC), a cancer type with poor clinical outcomes and increasing incidence in the UK[22,63,64]. In contrast to breast and gastric cancers, which are associated with highly penetrant predisposition variants and their associated familial syndromes[65,66], OAC does not have well-established mechanisms of predisposition[23]. Indeed, even GWAS have identified very few predisposing variants[21-23] (Table 2-2). OAC thus represents a particularly appropriate cancer type to investigate GSITs using damaging germline variants, rather than only predisposition variants. I will discuss OAC and its suitability to this study in more detail in Section 2.3.1.

The statistical challenge of identifying GSITs is reflected in the results of existing studies. This difficulty arises because of the complexity of both the germline and somatic landscapes in cancer. Wang *et al.*[50] and Carter *et al.*[60] reported only one0and two SNP-gene associations with a false discovery rate (FDR) <0.1, respectively. Meanwhile, Zhang *et al.*[51] found no results with FDR below 1, and Lu *et al.*[55] and Puzone *et al.*[59] made no attempt to correct for multiple hypothesis tests. I will investigate this challenge in detail in Section 2.4, and propose a method for managing statistical power in this context.

However, the results that have been reported by GSIT studies do indicate that useful biological insights can be obtained from this analysis. Carter *et al.*[60] experimentally validated a genetic interaction, elucidating details of the role of the oncogene *GNA11* in mTOR signalling. Puzone *et al.*[59] showed that several SNPs, equivalently to certain somatic mutations, were associated with elevated expression of the driver gene *MAP3K1*, increasing the mutation rate of another driver gene, *PI3KCA*. These results show that GSIT analyses can further our understanding of genetic interactions in cancer.

SNP	Odds ratio	Gene(s)	Variant effect
rs75783973	1.33	TPPP / CEP72	Intronic / Downstream
rs2188554	1.23	ASZ1	Intronic
rs76014404	1.21	KHDRBS2	Intronic
rs10419226	1.18	CRTC1	Intronic
rs9823696	1.17	-	-
rs7255	1.17	GDF7	3' UTR

Table 2-2: Known predisposition variants in oesophageal adenocarcinoma

Only SNPs that reached genome-wide significance for OAC predisposition in original GWAS or meta-analyses[21-23] are shown. Does not include SNPs found to jointly predispose to Barrett's oesophagus and OAC, as the majority of cases in these analyses are Barrett's. Genes and variant consequences are taken from dbSNP[67] (https://www.ncbi.nlm.nih.gov/SNP). UTR: un-translated region.

2.1.3. Complementary lines of investigation into germline-somatic interactions

In addition to investigating how damaging germline variants can influence selective pressures acting on somatic driver alterations, several complementary lines of inquiry have been undertaken in the literature. These studies look at germline-somatic interactions from different points of view. While they are less directly relevant to the hypothesis investigated in Chapters 2 and 3 than those listed in Table 2-1, they provide a broader context for this thesis.

Common germline variants are unlikely to damage protein function (at least in ways that have phenotypic consequences), and largely instead encode population structure[68,69]. Yuan *et al.*[70] investigated the effect of genetic ancestry on the frequency of somatic alterations in known driver genes. They found that African Americans had significantly different rates of alteration in five driver genes compared to patients of European ancestry. However, interpretation of these results is problematic. First, there is no clear genetic mechanism that can be investigated. Second, there are numerous potentially confounding factors, including environmental and socioeconomic correlates. While Yuan *et al.* did control for environmental factors including smoking and alcohol exposures, they admit that socioeconomic data were unavailable for their study. Nevertheless, this work suggests a need to control for genetic ancestry in GSIT studies. One approach is to restrict to studying a particular population, as done by Carter *et al.*[60] and in Chapter 2 of this thesis. Another alternative is to include principal components of genetic variation in a regression model, as in Wang *et al.*[50], Zhang *et al.*[51], and Chapter 3.

The somatic landscape of cancer can be characterised features besides driver alterations. For example, mutational burden indicates the overall rate of mutation, and mutational signatures can be used to measure the activity of individual mutational processes[45]. Some studies have investigated the effect of germline variation on such somatic features. Wang *et al.*[50] and Zhang *et al.*[51] used mutational signatures to complement their analyses of GSITs. More recently, the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium investigated this comprehensively[16]. They found several germline loci associated with signatures of APOBEC mutagenesis, as well as demonstrating that germline truncating variants in *BRCA1* and *BRCA2* increased the burden of certain types of somatic structural variants. I will incorporate both mutational burden and mutational signatures into my analysis in Section 3.5.

Research has also shown that the germline can shape somatic evolution in cancer by altering the tumour microenvironment. McGranahan *et al.*[71] showed that loss of heterozygosity of HLA genes is common in non-small-cell lung cancer, constituting an immunogenic germline-somatic interaction *in cis*. An investigation *in trans* by Marty *et al.*[72] found that HLA genotypes influenced the frequency of certain recurrent somatic driver alterations. While the tumour microenvironment clearly represents an important mechanism by which the germline can influence somatic evolution in cancer, the necessary focus on immunology in these studies is beyond the scope of this thesis.

2.2. Methods

2.2.1. Annotation of germline data

VCF files for germline variant calls were obtained for 260 OAC samples from the OCCAMS consortium[73], as well as for 2,504 samples from the 1000 Genomes project[68] (Phase III, v5a). Samples from 1000 Genomes were filtered to retain only unrelated European individuals, giving 503 in total (the EUR cohort). VCFs were annotated with ANNOVAR[74] (July 2017) to: (1) map variants to genes and predict variant effects (using the RefSeq[75] Gene hg19 database release 81); (2) provide MAFs in reference European populations (taken from the 1000 Genomes Project[68] Phase III v5a and ExAC[76] v0.3); and (3) annotate predicted measures of variant deleteriousness (taken from dbNSFP[77] v3.3 and dbscSNV[78] v1.1). Genes were then intersected with an internally curated set of 19,014 human genes, obtained by aligning protein sequences to the human genome[26].

2.2.2. Germline variant filters

All OAC germline variant calls were filtered based on reads supporting the alternative allele (‘alternative reads’). Exonic and splicing variants were further filtered based on their MAF and their deviation from Hardy-Weinberg equilibrium. These filters were manually calibrated and did not rely on statistical tests.

First, variant calls in OAC that had <2 reads alternative reads were removed. For each variant call, the percentage of alternative reads was calculated as the number of alternative reads divided by the total number of reads at that locus. Heterozygous calls were retained if they had between 36.5% and 62.3% alternative reads, and homozygous calls were retained if they had 95% or greater alternative reads. After this filtering, only variants annotated as exonic or splicing were retained for downstream analysis.

Next, MAFs for each variant were obtained from reference European populations, provided by the 1000 Genomes Project[68] and ExAC[76]. Since ExAC represented a much larger population (approximately 30,000 individuals, compared to just over 500 in the 1000 Genomes Project), the reference MAF was taken from ExAC where possible. Variants that were not annotated with a MAF in either database were assigned a reference MAF of 0. The MAF of each variant was then compared between the OAC cohort ($\mathrm{MAF_{OAC}}$) and the reference populations ($\mathrm{MAF_{ref}}$). Variants were removed if they had $\mathrm{MAF_{OAC}} > \sqrt{10 \times \mathrm{MAF_{ref}}}$ and $\mathrm{MAF_{OAC}} > 0.1$.

Finally, variants were filtered out if they exhibited an excess of heterozygotes. The expected proportion of heterozygotes ($\mathrm{het_{exp}}$) was calculated from Hardy-Weinberg equilibrium based on the observed MAF of each variant as $\mathrm{het_{exp}} = 2 \times \mathrm{MAF_{OAC}} \times (1 - \mathrm{MAF_{OAC}})$. Variants were then removed if the observed proportion of heterozygotes ($\mathrm{het_{OAC}}$) had $\mathrm{het_{OAC}} > 1.04 - \sqrt{1 - 1.87 \times \mathrm{het_{exp}}}$.

2.2.3. Definition of rare damaging germline variants

Four sets of predicted damaging germline variants were identified: Truncating variants, the Ensemble set, the Consensus set, and the High-confidence set. The High-Confidence set was used for downstream analysis, and was constructed from the other sets.

Truncating variants were defined as exonic SNPs with predicted stopgain or stoploss effect as well as frameshift indels, and included 7,583 unique variants. The Ensemble set included 4,158 missense variants predicted as damaging by both CADD[61] (phred score >25) and MetaLR[79].

The Consensus set included missense SNPs supported by at least five out of seven function-based deleteriousness prediction methods (SIFT[80], PolyPhen2-HDIV[81], PolyPhen2-HVAR, MutationTaster[82], MutationAssessor[83], LRT[84], and FATHMM[85]), as well as splicing variants supported by at least one of two splice site-specific deleteriousness prediction methods from dbscSNV[78] (ADA and RF). In total, the Consensus set consisted of 17,258 variants. In addition, 28,313 missense SNPs supported by at least two of three conservation-based methods (PhyloP[86], SiPhy[87], and GERP++[88]) were considered, but only those supported by function-based or splicing predictors were retained.

The High-Confidence set was then defined as the intersection of the Ensemble and Consensus sets, combined with the Truncating set, for a total of 11,383 variants. For downstream analysis, only 6,393 rare variants in the High-Confidence set were retained, defined as those with MAF <1% in reference European populations.

2.2.4. Gene expression and GSEA analyses

For gene expression analyses, 92 out of the 260 OAC samples had available matched RNA-Seq data provided by the OCCAMS consortium. Readcount values were normalised into transcripts per million (TPM). Only genes expressed across the cohort with median TPM >1 were considered for analysis (11,104 genes). To identify differential expression associated with germline-somatic interactions, samples were stratified into four groups according to their germline/somatic status, as described in Section 2.4.3. For each gene, the median TPM value in each group was calculated, and the fold-change was measured comparing the G+S+ group to each of the other three groups. Genes for which all three fold-changes were >1 were considered to be upregulated in the G+S+ group, and the smallest of the three fold-changes was taken as a conservative measure of upregulation. Conversely, genes for which the fold-changes were all <1 were considered to be downregulated, and the largest fold-change was taken.

In order to establish pathway enrichment of upregulated and downregulated genes, gene sets with different thresholds of dysregulation (fold-change <0.5, <0.67, >1.5, >2) were first identified. Hypergeometric tests were then used to measure enrichment of these genes in 2,828 pathways from Reactome and KEGG. P-values were then corrected for multiple hypothesis testing using the Benjamini-Hochberg FDR method[89], with all pathways considered together.

2.3. Preparing a high-quality dataset of 260 oesophageal adenocarcinomas

Investigating germline-somatic interactions requires a cohort with high-quality and well-characterised germline and somatic data. In this section, I will describe the preparation of a cohort of 260 oesophageal adenocarcinoma (OAC) samples for this purpose. First, I will give a brief overview of the clinical and genomic characteristics of OAC, and discuss its suitability as a model cancer type to investigate germline-somatic interactions (2.3.1). I will then describe the annotation and quality control of germline variants in this cohort (2.3.2). Given this, I will discuss damaging germline variants as potential effectors of germline-somatic interactions, and describe a relatively stringent approach for identifying them in this cohort (2.3.3). I will then investigate the distribution of damaging germline variants at the gene and pathway levels, identifying several genes to be excluded from subsequent analysis (2.3.4). Finally, I will briefly describe the work of a colleague to characterise somatic driver events at the pathway level, completing the curation of this cohort (2.3.5).

2.3.1. Features and suitability of oesophageal adenocarcinoma

The incidence of oesophageal adenocarcinoma in Western countries has increased dramatically in the past 40 years, with Europe and North America accounting for 34% of global cases in 2012[90,91] (Figure 2-1). Nevertheless, it remains a relatively rare cancer type, accounting for an estimated 52,000 of 14.1 million global cancer cases[90,92] (0.4%). OAC has poor clinical outcomes and a five-year survival rate of approximately 20%[93], largely due to the fact that most cases present at an advanced stage. Research into possible early detection strategies has revealed a number of risk factors for OAC. Gender has a clear effect, with men being up to 9 times more likely to develop the disease than women[90,91] (Figure 2-1). In addition, gastro-oesophageal reflux disease (associated with obesity) is a predictor for OAC, with chronic sufferers at approximately six-fold increased risk[94]. Long-term exposure of the oesophageal epithelium to acid reflux induces a metaplastic condition called Barrett's oesophagus, which is thought to desensitise cells to acid exposure[94]. Barrett's oesophagus is a precursor lesion to OAC, and in epidemiological studies the two conditions are often pooled together.

Our understanding of the somatic genomic landscape of OAC has increased substantially in the last ten years. There have been several OAC sequencing efforts globally. The Oesophageal Cancer Clinical and Molecular Stratification consortium (OCCAMS, part of ICGC) is the largest such effort and is my primary source of OAC data. Many OACs are genomically unstable[64,73,95] with high rates of focal amplifications and chromothripsis. Indeed, a recent pan-cancer study found that OAC has the fourth highest rate of genomic catastrophes across all cancer types[16]. Perhaps unsurprisingly given this instability, the most prevalent somatic driver gene in OAC is *TP53*, which is altered in approximately 72% of cases[96]. In total over 100 OAC driver genes have been identified by genomic screenings[73,96-100]. Analysis of mutational signatures in OAC has revealed a hallmark signature (COSMIC Signature 17), which is thought to reflect mutagenesis arising from acid reflux exposure[98,100]. Other mutational processes have also been identified in OAC by their signatures, including APOBEC mutagenesis and defective homologous recombination[73].

There are several reasons why I have chosen to use OAC to investigate how perturbative germline variation can influence the somatic evolution of cancer. As mentioned above, OAC has a poor five-year survival rate, so greater understanding of the disease in general is required to improve clinical outcomes. In addition, while results from GWAS indicate that 25% of OAC cases are determined by common germline variants[101], no high-penetrance predisposition genes or variants have been identified[23]. Indeed, a recent study showed that even the previously identified predisposition variants did not improve the ability to predict development of OAC compared to clinical factors[102]. This means that any germline influence on somatic evolution in OAC is unlikely to be dominated or biased by the effects of predisposition. It also reflects the fact that little is currently understood about the role of the germline in OAC, so there is a greater potential for results from this study to be of clinical use, for example in early detection. Finally, our laboratory (*i.e.* the Ciccarelli lab) is a member of the OCCAMS consortium, giving us particular access to data and expertise to learn more about this disease. The recent increased interest in germline-somatic interactions across cancer (Section 2.1.2) has coincided with the availability of sequencing data for large OAC cohorts, so this is the first time that such a study has been feasible.

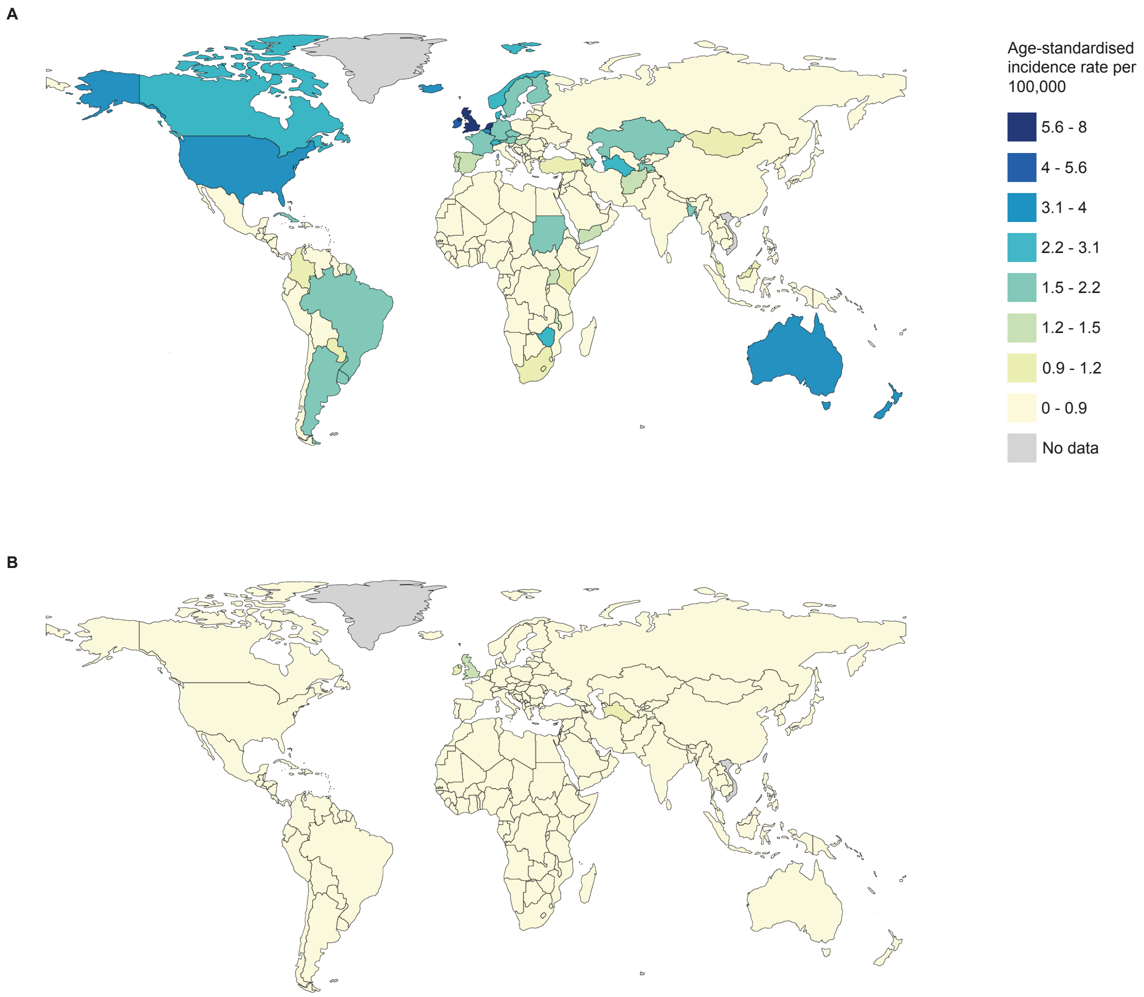

Figure 2-1: Global incidence of oesophageal adenocarcinoma

Age-standardised incidence rates of OAC in 2012 for males (**A**) and females (**B**). Data were taken from Arnold *et al.* 2015[90].

2.3.2. Quality control of germline variants

Data for 260 OAC patients were provided by the OCCAMS consortium. The germline data consisted of small variant calls (single nucleotide polymorphisms, SNPs, and insertions and deletions, indels) from whole genome sequencing (WGS). Since the data had not undergone benchmarking (the consortium had primarily focused on somatic data), and false positive germline calls could lead to spurious results in later analyses, I undertook ensure the quality of the data. Given that the vast majority of the 260 patients were White British (99% of those with information available, Figure 2-2A), I used European samples from the 1000 Genomes Project [68] for comparison. In total, this control cohort (which I will refer to hereafter as EUR) consisted of 503 unrelated individuals.

I first annotated the germline variants in the OAC and EUR cohorts to provide information on genes, allele frequencies in reference populations, and predicted variant effects (Methods 2.2.1). Across the whole genome, the average number of SNPs per sample was in very good concordance between OAC and EUR (Table 2-3). However, there was an excess of 33% more indels per sample in OAC (Table 2-3), suggesting that additional filtering would be required to remove potential false positive indel calls. Since I was primarily interested in variants that had damaging effects on protein sequences, I restricted subsequent analysis to exonic variants. The concordance of SNP burden and excess of indel burden were both also present at the exonic level, with indels inflated by 38% in OAC (Figure 2-2B, Table 2-3). Filtering was also motivated by the presence of a clear technical artefact in the OAC data that had a mean proportion of alternative reads of 34%, had an allele frequency over 150 times larger than in EUR, and was heterozygous in all 260 samples.

The percentage of alternative reads supporting any germline variant call should be near to 50% or 100%, depending on whether the variant is heterozygous or homozygous. The data did reflect this in large part (Figure 2-2C), however there were still many calls that were outliers to the expected distribution, particularly among indels. Removing these outlier variant calls filtered out 10% of SNP calls and 18% of indel calls (Figure 2-2C, Methods 2.2.2). This was appropriate given that indels were inflated in OAC compared to the EUR cohort.

I applied two other filters that relied on population-level measures. First, I compared the minor allele frequency (MAF) of each variant between the OAC cohort and reference European populations (Methods 2.2.2). While the vast majority of variants had very similar MAFs, a large number of variants deviated from this (Figure 2-2D). Since I was aiming to remove possible false positive germline calls from the OAC data, I filtered out a subset of variants that were common in OAC despite having near-zero MAF in reference populations (Figure 2-2D, Methods 2.2.2). The lack of high-penetrance predisposition variants in OAC[23] meant that such enrichment was most likely to represent technical artefacts. This filter removed only 0.15% of the unique variants in OAC. The second population-level filter removed 0.22% of unique variant that exhibited a clear excess of heterozygotes, as compared to expectation from Hardy-Weinberg equilibrium (Figure 2-2E, Methods 2.2.2). Excess heterozygosity has been reported as a characteristic of certain technical artefacts[103].

After applying these three filters, the final set of germline variants in the 260 OACs contained just over five million exonic variant calls (Table 2-4). Overall, samples in the OAC cohort now had 12% fewer exonic SNPs and 12% more indels than EUR samples, compared to 38% more indels before filtering (Figure 2-2F, *cf.* Figure 2-2B). Since indels are much more likely to have damaging effects on protein function than SNPs (resulting from a shift in reading frame), this reflected the removal of many potential false positive damaging germline variant calls. Analysing germline variants by their predicted effect, there was good concordance of per-patient numbers of variants between OAC and EUR (Table 2-4). This indicated that the filtered germline variants represent a high-quality dataset for further analysis.

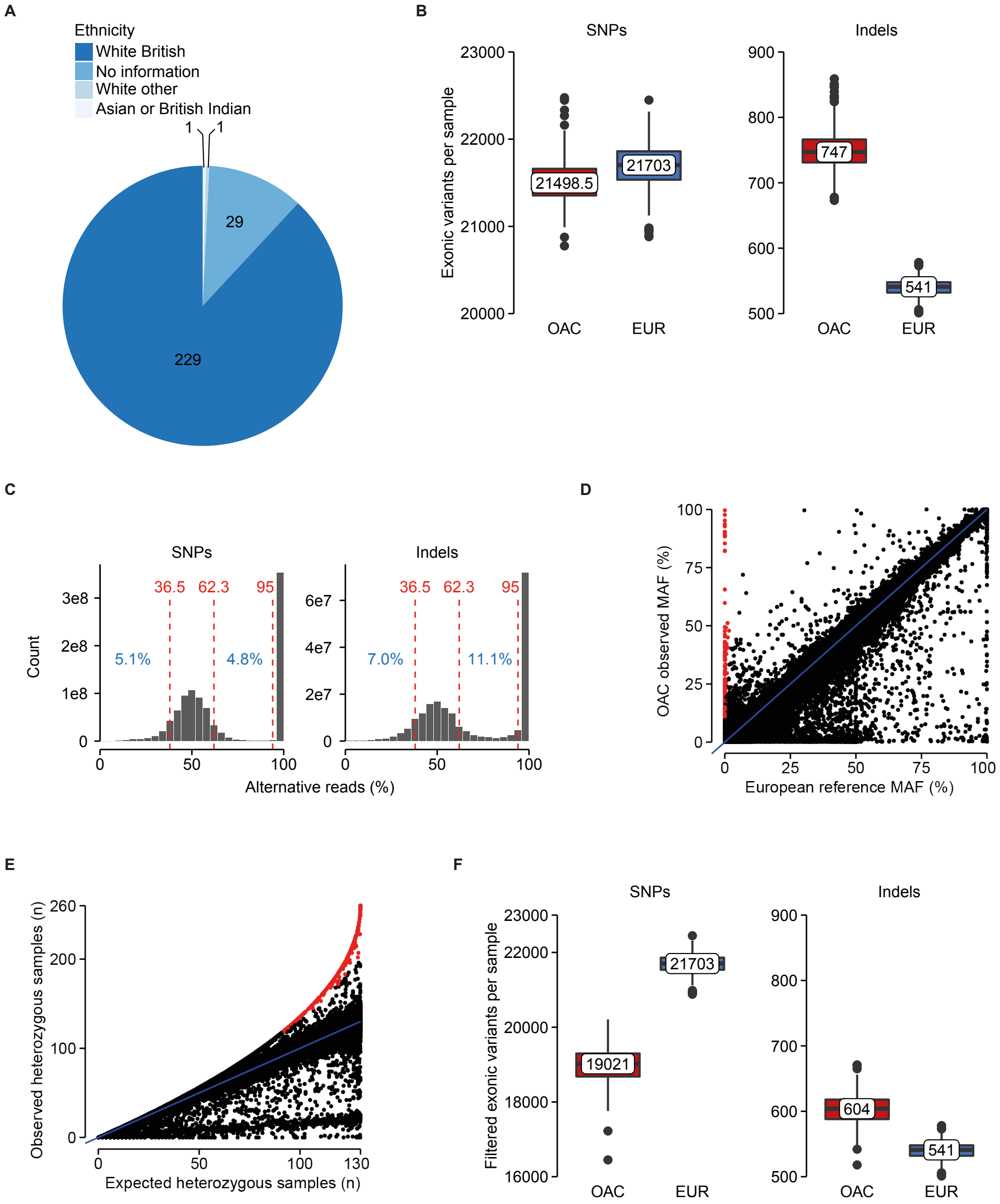

Figure 2-2: Quality control of germline variants

A. Reported ethnicity of 260 OAC samples, with numbers of samples indicated.

B. Distributions of exonic SNPs (left) and indels (right) per sample in the OAC and EUR cohorts, before filtering. Median values are indicated.

C. Percentage of alternative reads supporting variant calls in the OAC cohort. Heterozygous calls were retained if they had between 36.5% and 62.3% alternative reads, and homozygous calls were retained if they had 95% or greater alternative reads, as indicated by dashed red lines. Percentages in blue indicate the proportion of variant calls removed by this filter.

D. Minor allele frequencies (MAFs) of exonic variants in the OAC (y-axis) and EUR (x-axis) cohorts. Variants in red were filtered out as overly-enriched in OAC.

E. Observed (y-axis) and expected (x-axis) numbers of heterozygous individuals for each exonic variant. Expectation was derived from Hardy-Weinberg equilibrium (Methods 2.2.2). Variants in red were filtered out as exhibiting an excess of heterozygotes.

F. Distributions of exonic SNPs and indels per sample in the OAC and EUR cohorts, after filtering variants in OAC. Median values are indicated.

		Median per sample		**Total**	
		OAC	**EUR**	**OAC**	**EUR**
Genomic	**SNPs**	3.66 x10^{6}	3.66 x10^{6}	9.53 x10^{8}	1.84 x10^{9}
	Indels	7.45 x10^{5}	5.60 x10^{5}	1.95 x10^{8}	2.82 x10^{8}
Exonic	**SNPs**	2.15 x10^{4}	2.17 x10^{4}	5.59 x10^{6}	1.09 x10^{7}
	Indels	7.47 x10^{2}	5.41 x10^{2}	1.95 x10^{5}	2.72 x10^{5}

Table 2-3: Unfiltered germline variants in the OAC and EUR cohorts

	Median		**Total**	
Variant type	**OAC**	**EUR**	**OAC**	**EUR**
All exonic	19635.5	22248	5098472	11185436
SNP	19021	21703	4941555	10913734
Missense	8716.5	10079	2266680	5068068

Stopgain	64	78	16513	39219
Stoploss	10.5	11	2710	5638
Indel	604	541	156917	271702
Frameshift	158	137	40919	68742

Table 2-4: Filtered exonic germline variants by type in the OAC and EUR cohorts

2.3.3. Rare damaging germline variants as effectors of germline-somatic interactions in OAC

In order to investigate how germline variation can influence somatic evolution in cancer, it is first necessary to characterise germline variation appropriately. Studies that have systematically investigated germline-somatic interactions have mostly focused on cancer predisposing or putatively deleterious variants[50,51,55,59] (Table 2-1, Section 2.1.1). There are very few known cancer predisposition variants for OAC, and they are not in protein-coding regions so lack a clear mechanistic explanation for their effect[21-23] (Table 2-2, Section 2.1.2). As a balance between being overly restrictive and the permissive GWAS approach of Carter *et al.*[60], I chose to investigate germline variants that had a putatively damaging effect on a protein sequence, and that thus might be relevant for OAC biology. However, in order to avoid potential false positives, I adopted a relatively stringent approach for identifying these variants, as detailed below.

Protein-truncating variants (PTVs) represent an obvious source of damaging variants. In the OAC cohort of 260 patients, I included all 3,783 stopgain, stoploss and frameshift variants as PTVs. Samples in OAC and EUR had very similar numbers of PTVs, with OAC samples having 3% more on average (Figure 2-3A). While some studies of damaging germline variants have only focused on PTVs[55,69], this is restrictive. For example, among missense variants there is a proportion that are likely to have damaging effects on protein function[77].

Identifying damaging non-truncating variants is an open but well-studied problem. Many algorithms have been developed to predict the impact of variants, using features such as the chemical change of amino acid type[80-85] (“function”), or the evolutionary constraint at a locus across different species[86-88] (“conservation”). There are also ensemble methods that use combinations of other algorithms to make predictions[61,79]. Our lab has previously found that taking a consensus of multiple algorithms can successfully identify damaging variants, both at the germline[104] and somatic[27] levels.

To identify damaging non-truncating germline variants in the OAC cohort, I started from an approach previously published by the Ciccarelli lab[27]. This approach took a consensus of function-based and conservation-based annotations for missense variants, as well as specific annotations for splicing variants (Methods 2.2.3). However, I found that nearly half (48%) of the unique predicted damaging variants were only supported by conservation-based methods (Figure 2-3B). In addition, the majority (78%) of those supported by function-based methods were also supported by conservation-based methods (Figure 2-3B). This was also reflected in the numbers of variants per sample, where on average 79% of predicted damaging variants were only supported by conservation-based methods (Figure 2-3C). These observations suggested that the conservation-based methods were inflating the numbers of damaging variants and potentially introducing many false positives. Thus, I removed the conservation-based methods from the consensus approach to define a Consensus set of 17,258 unique variants. As a benchmark, I compared the overlap between the Consensus set and predicted damaging variants from two ensemble methods, CADD[61] and MetaLR[79]. Each of the three sets had many variants that were not identified by the other approaches (Figure 2-3D). In order to remove as many potential false positives as possible, I used the 3,800 variants common to all three methods as a high-confidence set of damaging non-truncating germline variants.

The combined PTVs and high-confidence non-truncating damaging variants totalled 7,583 unique variants. As an additional check to remove potential false positives, I investigated the frequency of these variants, since deleterious variants are likely to be selected against, and therefore to be rare[79]. The PTVs included many common variants, while the non-truncating damaging variants were highly skewed towards rarer variants (Figure 2-3E). This likely reflected the stringent approach taken to identifying the non-truncating damaging variants. As further support for this, the variants in the Consensus set that were not supported by ensemble methods were substantially more common than the high-confidence variants (mean MAF 24% compared to 6%, Figure 2-3E). In order to remove potential false positive damaging variants, particularly from among the PTVs, I retained only rare variants, with a MAF <1% in reference European populations (Methods 2.2.3).

The final set of rare damaging germline variants (RDGVs) included 6,393 unique variants, occurring 9,334 times across the cohort and present in all 260 OACs. Compared to the reference EUR cohort, OAC samples had 23% fewer RDGVs on average (Figure 2-3F). Since this discrepancy was mostly introduced by restricting to rare variants (OAC had only 3% fewer damaging variants overall, Figure 2-3F) it may have reflected a bias stemming from the use of samples from the 1000 Genomes Project in the definition of reference MAFs. Nevertheless, the RDGVs represented a high-confidence set of damaging variants for further analysis.

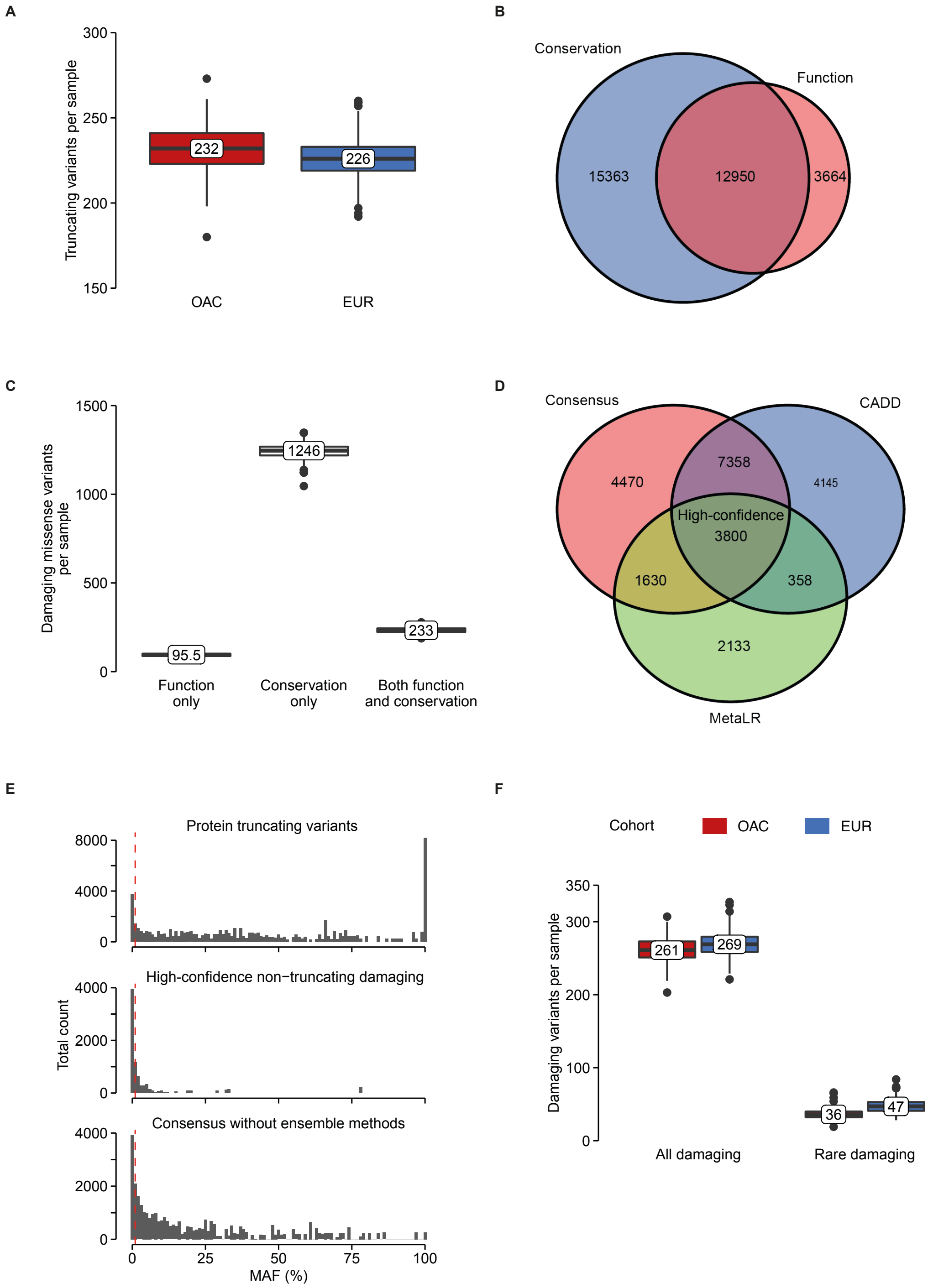

Figure 2-3: Identification of high-confidence damaging germline variants

A. Numbers of protein-truncating variants per sample in the OAC and EUR cohorts. Median values are indicated.

B. Overlap of unique missense variants labelled as damaging by conservation-based methods (blue) and function-based methods (red).

C. Distributions of missense variants per sample labelled as damaging by function-based methods only (left), conservation-based methods only (middle) or both (right). Median values are indicated. Conservation-based methods were removed from the Consensus set.

D. Overlap of unique missense variants predicted as damaging by the Consensus approach (red), CADD[61] (blue) and MetaLR[79] (green). The intersection of all three was carried forward as damaging missense variants.

E. Distributions of minor allele frequencies (MAFs) for protein-truncating variants (top), high-confidence damaging missense variants (middle) and missense variants in the Consensus set but not supported by CADD or MetaLR (bottom). Rare variants with MAF <1% (dotted red line) were retained.

F. Numbers of damaging and rare damaging germline variants per sample in the OAC and EUR cohorts. Median values are indicated.

2.3.4. Germline perturbations at the gene and pathway levels

The RDGVs constituted a high-confidence set of damaging variants that were likely to perturb biological processes. However, analysis at the variant level would have suffered from the rarity of individual variants, since each RDGV was damaged in a median of just one patient. By aggregating RDGVs into genes, and then into pathways, this problem could be alleviated. Across the OAC cohort, 4,230 genes had at least one RDGV.

As described in Section 2.3.1, no high-penetrance predisposition variants or genes are known for OAC[23]. However, most efforts to date have investigated germline data from SNP arrays, rather than high-throughput sequencing[21-23]. Thus, as a preliminary analysis I investigated the prevalence of RDGVs in cancer predisposition genes (CPGs) in the OAC cohort. Out of 132 CPGs obtained from the literature[18,48] for a range of cancer types, 47 had at least one RDGV in the OAC cohort. Fisher tests did not reveal any of these genes to be significantly enriched in OAC compared to the reference EUR cohort, and indeed the frequencies of the most commonly damaged CPGs did not differ substantially between the two cohorts (Figure 2-4A).

The most frequent CPG in both cohorts was *BRCA2*, damaged in 8 and 14 samples in OAC and EUR respectively. These *BRCA2*-mutated samples presented an opportunity to investigate whether RDGVs were impacting mutational processes in

OAC. *BRCA2* is an integral part of the DNA homologous recombination pathway, and its mutation is associated with COSMIC signature S3 in cancer[45]. Moreover, signature S3 has been shown to be prevalent in OAC, and to be potentially useful for clinical stratification[73]. Using mutational signatures extracted for the 260 OAC samples by the OCCAMS consortium, I compared the prevalence of S3 between samples with and without a RDGV in *BRCA2*. While the *BRCA2*-mutated samples had a 55% higher prevalence of signature S3 on average, this trend did not reach statistical significance (p =0.4, Wilcoxon rank-sum test, Figure 2-4B). This suggested that the germline might be playing a role in OAC mutational processes, but a larger sample size would be needed to verify this due to the rarity of *BRCA2* RDGVs.

Overall, most of the 4,230 genes were only rarely damaged, with the majority (60%) damaged in only one sample (Figure 2-4C). However, some genes were damaged in large numbers of samples, suggesting that there might still be false positive RDGVs. For example, *HLA-DRB5* was damaged in 113 samples (43%). In order to assess this systematically, I used Fisher tests to measure whether any genes were enriched compared to the reference EUR cohort. A total of 16 genes were enriched at FDR <0.1 (Figure 2-4D, Table 2-5). The lack of high-penetrance predisposition genes in OAC[23] meant that these enrichments were more likely to constitute false positives than true biological effects. Indeed, several of the enriched genes were obvious candidates for technical artefacts, including genes from the large paralogous HLA[105] and zinc finger families. HLA genes are also known to be highly polymorphic in order to ensure that immune systems can process diverse foreign peptides at the population-level[106]. I therefore flagged these genes as potential confounders for downstream analysis.

Finally, I aggregated RDGVs from genes into pathways. This is the level at which I will investigate germline-somatic interactions in this thesis, because it substantially improves statistical power when dealing with rare variants, as I will discuss in detail in Section 2.4. I obtained 2,828 pathways containing 11,199 genes from the union of three pathway databases (Reactome[107], KEGG[108] and BioCarta[109]). Of these 2,444 pathways had at least one RDGV, coming from 2,722 genes. Because of this aggregation, pathways were damaged substantially more frequently than genes. While genes were damaged in a median of only one sample, pathways were damaged in 11 samples on average (Figure 2-4E, *cf.* Figure 2-4C). This mitigated the statistical challenge of identifying germline-somatic interactions to some extent. In order to assess possible enrichment of RDGVs in OAC at the pathway level, I used MEGA-V[110] to compare OAC samples to reference EUR samples. MEGA-V uses Wilcoxon rank-sum and Kolmogorov-Smirnov tests to detect the enrichment of genetic alterations in biological processes in a cohort of interest. A total of 94 pathways were enriched with FDR <0.1. Of these pathways, 83 (88%) contained one of the 16 enriched genes (Figure 2-4F). Among the 11 pathways that did not (Table 2-6), there was no clear relationship to OAC biology. Moreover, their statistical significance was not robust, as removing all pathways that contained an enriched gene led to no pathways being enriched (Table 2-6). Since the 16 enriched genes were disproportionately affecting pathway-level variation, I removed them from all subsequent analysis. The resulting set of 2,442 damaged pathways represented a high-quality germline dataset with which to investigate germline-somatic interactions.

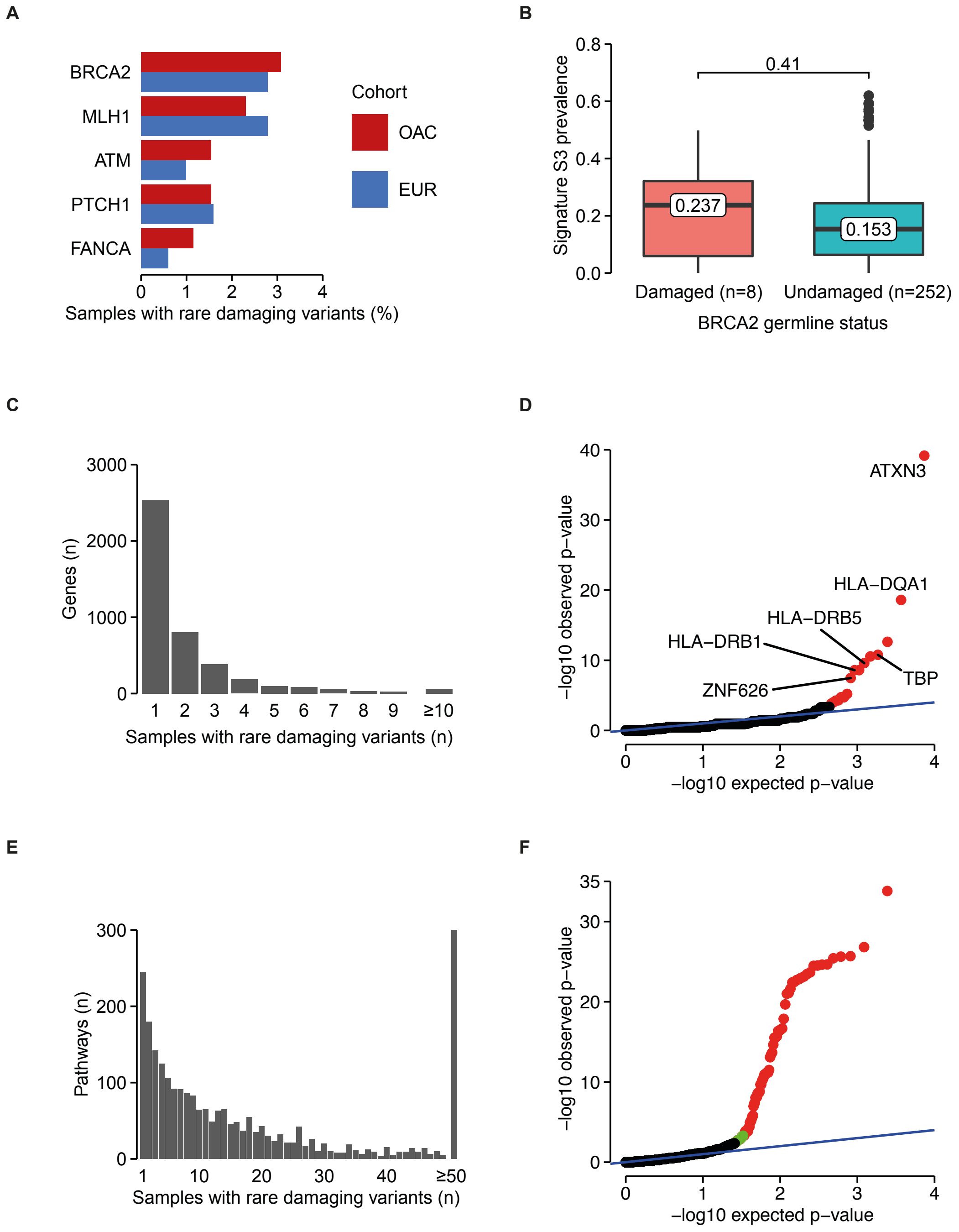

Figure 2-4: Rare damaging germline variants at the gene and pathway levels

A. Damaged cancer predisposition genes in the OAC and EUR cohorts. No genes were significantly enriched in OAC (Fisher's exact test).

B. BRCA-associated mutational signature (COSMIC S3) prevalence extracted from somatic mutation data, in samples with damaged (left) and undamaged (right) germline *BRCA2*. Median values are indicated. P-value from Wilcoxon rank-sum test.

C. Number of samples in which each of the 4,230 genes is damaged.

D. Quantile-quantile (QQ) plot showing enrichment of damaged genes in the OAC cohort compared to EUR (Fisher's exact test). Genes with FDR <0.1 are coloured red.
E. Number of samples in which each of the 2,444 pathways is damaged.
F. QQ plot showing enrichment of damaged pathways in the OAC cohort compared to EUR (p-values from MEGA-V[110]). Pathways with FDR <0.1 are coloured green if they do not contain an enriched gene, and red if they do.

Gene	OAC (260)	EUR (503)	P-value	FDR
ATXN3	76	0	6.95E-40	5.12E-36
HLA-DQA1	38	0	2.54E-19	9.36E-16
HRCT1	31	2	2.37E-13	5.81E-10
TBP	25	1	1.67E-11	3.08E-08
CD24	22	0	2.82E-11	4.15E-08
HLA-DRB5	114	109	2.55E-10	3.14E-07
ANKRD20A3	18	0	2.57E-09	2.37E-06
HLA-DRB1	18	0	2.57E-09	2.37E-06
ZNF626	18	1	3.38E-08	2.77E-05
FAM209B	11	0	6.24E-06	4.60E-03
GOLGA6L2	10	0	1.88E-05	1.15E-02
KRTAP5-4	10	0	1.88E-05	1.15E-02
COL18A1	14	3	5.10E-05	2.89E-02
MAGEF1	9	0	5.65E-05	2.97E-02
CCDC40	18	7	9.43E-05	4.63E-02
ANKLE1	10	1	1.44E-04	6.65E-02

Table 2-5: Enriched damaged genes in the OAC cohort

Numbers of samples with each damaged gene in the OAC and EUR cohorts are shown. P-values of enrichment in the OAC cohort were calculate using Fisher's exact test and corrected for false discovery rate (FDR). The 16 genes with FDR <0.1 are listed.

Pathway	**Source**	**OAC (260)**	**EUR (503)**	**P-value**	**FDR**	**Filtered FDR**
Golgi Cisternae Pericentriolar Stack Reorganization	Reactome	10	3	5.11E-04	0.03	0.46
Tryptophan catabolism	Reactome	11	4	5.98E-04	0.04	0.46
Phosphate bond hydrolysis by NTPDase proteins	Reactome	11	4	6.04E-04	0.04	0.46
Sperm Motility And Taxes	Reactome	8	2	1.03E-03	0.06	0.55
Serine biosynthesis	Reactome	10	4	1.47E-03	0.08	0.55
Zinc transporters	Reactome	6	1	1.90E-03	0.09	0.55
Zinc influx into cells by the SLC39 gene family	Reactome	6	1	1.90E-03	0.09	0.55
Catabolism of glucuronate to xylulose-5-phosphate	Reactome	6	1	1.91E-03	0.09	0.55
Biotin metabolism	KEGG	4	0	2.65E-03	0.03	0.58
D-Arginine and D-ornithine metabolism	KEGG	3	0	7.91E-03	0.08	0.58
Measles	KEGG	43	53	9.22E-03	0.09	0.62

Table 2-6: Enriched damaged pathways in the OAC cohort

Numbers of samples with each damaged pathway in the OAC and EUR cohorts are shown. P-values of enrichment in the OAC cohort were calculate using MEGA-V[110] and corrected for false discovery rate. Pathways with FDR <0.1 and that did not contain one of the enriched genes in Table 2-5 are listed. The filtered FDR was calculated by performing the FDR correction on p-values only for pathways that did not contain an enriched gene.

2.3.5. Somatic driver events at the pathway level

Having characterised germline variation in terms of RDGVs in pathways, I characterised somatic driver alterations in the OAC cohort at the pathway level as well. By associating pathways containing RDGVs with pathways containing somatic driver events, I could investigate how germline perturbations can influence selective pressures in cancer evolution (Section 2.4). This section describes work done by Thanos Mourikis in the Ciccarelli lab, described in Mourikis *et al.*[27], to identify pathways with somatic drivers in the OAC cohort. I have included a brief overview here for completeness.

Mutation and copy number data for the same cohort of 260 OACs were used to identify somatic driver events. Within these data, truncating, non-truncating damaging, and hotspot somatic mutations were identified, as well as CNVs in genes. These somatic alterations were then mapped to 202 cancer genes from the Cancer Gene Census[111] that were annotated as either oncogenes (OGs) or tumour suppressor genes (TSGs). Driver alterations included putative gain-of-function alterations in OGs and loss-of-function alterations in TSGs, and were identified in 259 of the 260 samples. As expected for OAC, the most common driver by far was *TP53*[96], altered in 197 samples (75%). Other common driver genes included *FHIT* (30%), *CDKN2A* (28%) and *MYC* (21%).

With somatic driver alterations having been identified at the gene-level, they were mapped to cancer-related processes that were identified through a gene set enrichment analysis. Among pathways from Reactome[107] and KEGG[108], 297 and 188 pathways were enriched for altered driver genes at FDR <0.05, respectively. As with the germline data, aggregating genes into pathways substantially increased the frequency of observations. While driver genes were altered in a median of 4.5 samples, pathways were altered in a median of 66 samples (Figures 2-5A, B). This increased frequency was statistically advantageous in finding germline-somatic interactions, as discussed in the next section.

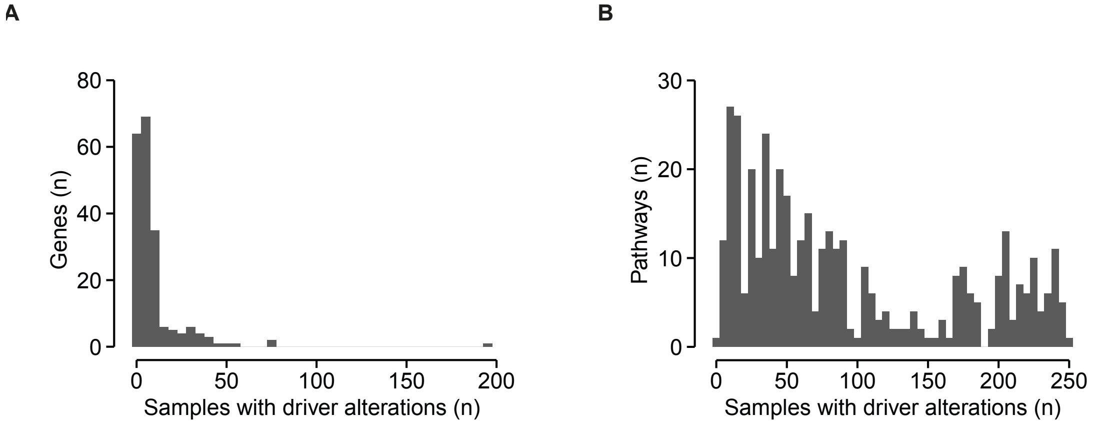

Figure 2-5: Frequencies of somatic drivers at the gene and pathway levels

A. Number of samples in which each of the 202 oncogenes and tumour suppressor genes had a somatic driver alteration.

B. Number of samples in which each of the 485 driver gene-enriched pathways had a somatic driver alteration.

2.4. Relating germline perturbations to somatic driver events

Having curated and quality-controlled the data for the cohort of 260 OAC samples, I proceeded to identify associations between germline perturbations and somatic driver alterations at the pathway level. In this section, I will first discuss a simple model to achieve this, as well as practical measures to manage the limited statistical power available (2.4.1). I will then describe two distinct germline-somatic associations revealed by this analysis (2.4.2), and investigate them in detail using both gene expression and clinical data (2.4.3, 2.4.4). Finally, I will discuss the results in an independent validation cohort of 140 OAC samples that became available at a later stage (2.4.5).

2.4.1. Measuring co-occurrence and managing statistical power

The hypothesis in question in Chapters 2 and 3 is that germline perturbations to biological processes could influence the selective pressures on somatic driver events in OAC. In theory, a germline perturbation could either increase or decrease the selective advantage of a particular driver alteration, but here I will restrict my focus to increased advantage. Such an increased advantage could be considered an 'inherited vulnerability' to a particular driver alteration. On the other hand, a decreased selective advantage might prove more difficult to interpret. Both increased and decreased selective advantages are considered in Chapter 3.

In order to identify cases of inherited vulnerability, I needed to measure co-occurrence of germline perturbations with somatic driver alterations. A simple model for measuring co-occurrence is Fisher's exact test, and this provided a useful first-pass approach to the problem. Agarwal *et al.*[20] used a Fisher test approach for their exploratory analysis of germline-somatic interactions. While considering both germline and somatic pathways as binary categorical variables is admittedly a highly granular approach, it has the advantage of making relatively few assumptions. More sophisticated statistical models are explored in Chapter 3.

Measuring co-occurrence between germline and somatic pathways requires large numbers of hypothesis tests to be carried out, demanding stringent thresholds for statistical significance. For the purposes of multiple hypothesis correction in this analysis, I considered pathways from different source databases as independent but complementary, and thus corrected tests pertaining to different databases separately. The largest database of both germline and somatic pathways was Reactome, for which there were 1,787 and 297 pathways characterising the germline and somatic data respectively (Table 2-7). To test every pair of germline and somatic pathways in Reactome would require 5.3x10^5 hypothesis tests. Using Bonferonni correction as a mathematically simple guide, this would require a stringent p-value threshold of 1.9x10^{-7} in order to achieve a family-wise error rate (FWER) <0.1. Given the available sample size and lack of highly penetrant germline events in OAC, it was a priori unlikely that any effect would be strong enough to reach this threshold. Thus, I needed to manage the available statistical power.

It is worth noting that the aggregation of germline variants into pathways had already reduced the multiple testing burden. In the OAC cohort, there were a total of 6,393 unique RDGVs, which would have required more than three times as many hypothesis tests as the pathway-level data. Moreover, most of these variants (82%) were present in a single sample, which would have made association with somatic drivers highly unreliable. By contrast, pathways were perturbed in a median of 11 samples.

An obvious way to further alleviate the multiple testing burden was to exclude some tests from consideration. Ideally, those tests whose statistical power was greatest would be retained. In order to achieve this, I considered the factors affecting the power of an individual Fisher test. These were: (1) the sample size N; (2) the test size α; (3) the effect size, as measured by the odds ratio (OR); (4) the frequency of the independent (*i.e.* germline) binary variable $f_{germline}$; and (5) the frequency of the dependent variable $f_{somatic}$. In this setting, N was fixed at 260 samples. The test size α depended on the number of hypothesis tests N_{hyp}; to achieve a FWER of 0.1, the significance threshold for individual tests was α=0.1/ N_{hyp}. However, in practice I used the less stringent FDR method[89], so this was a conservative estimate of power. I used simulations to quantify how the power of tests depended on the effect size and the germline and somatic pathway frequencies. This analysis showed that statistical power was greatest when $f_{germline}$ and $f_{somatic}$ were both nearest to 0.5 (*i.e.* altered in 50% of samples), and that the region around this centre with adequate statistical power expanded as the OR increased (Figure 2-6). This suggested a frequency restriction approach, retaining germline and somatic pathways whose frequencies were between θ and 1-θ, for some threshold $0<\theta<0.5$. I selected $\theta=0.25$, which reduced the number of tests for Reactome pathways from approximately 500,000 to 7,000. Moreover, the retained tests had 80% power to detect associations with OR ≥5 (Figure 2-6). Thus I proceeded to test for co-occurrence with pathways damaged in between 25% and 75% of samples (Table 2-7).

Another consideration for maximising statistical power was the method of multiple hypothesis test correction. As mentioned above, the FDR method of Benjamini & Hochberg[89] is less stringent than the FWER method of Bonferroni, and thus provides a more powerful alternative. However, the standard FDR method assumes that hypothesis tests are independent, and this assumption was violated in this analysis due to the high degree of overlap among pathways. Ji & Li[112] developed a method to calculate the effective number of independent hypothesis tests, and use this in FDR correction. Originally applied to overlapping multilocus SNP data, it takes the correlation of variables into account to relax p-value correction, increasing statistical power. It thus provided a potentially more appropriate way to calculate FDRs than the standard method in this context. Applying this method to the already reduced number of tests further reduced the burden of multiple hypothesis testing, giving 1,449 independent tests for Reactome pathways (Table 2-7). By combining frequency restriction and correcting for the effective number of independent tests, I had reduced the total number of tests to correct for by 180-fold, thus substantially alleviating the burden on statistical power.

Count type	**Germline**			**Somatic**	
	Reactome	KEGG	BioCarta	Reactome	KEGG
≥1 sample	1787	370	285	297	110
25% to 75% of samples	87	73	0	81	53
Effective independent	63	55	0	23	24

Table 2-7: Numbers of somatic and germline pathways in the OAC cohort

Only pathways altered in between 25% and 75% of samples were retained for germline-somatic interaction analysis. Using these pathways, the effective number of independent pathways was calculated using the method of Ji & Li[112].

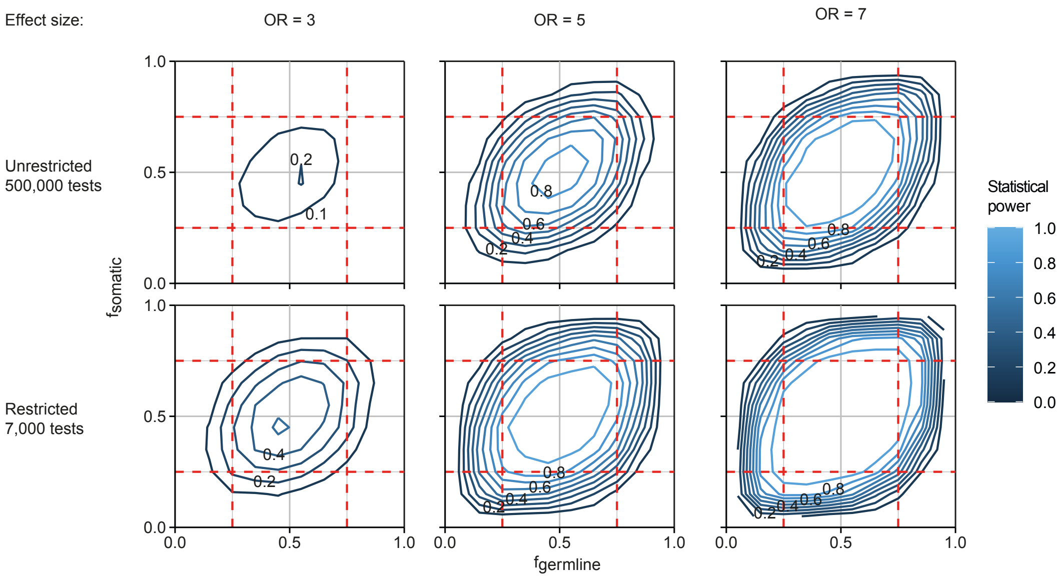

Figure 2-6: Statistical power of Fisher tests varies with observation frequencies Contour plots show the statistical power of a Fisher test to detect a germline-somatic association given the frequencies of the germline pathway ($f_{germline}$, x-axis) and the somatic pathway ($f_{somatic}$, y-axis). Analysis considered a range of effect sizes (odds ratios, ORs), indicated at the top. The test size was fixed using the Bonferroni method at 0.1 divided by the number of hypothesis tests. In the unrestricted (top) case, a total of 500,000 tests were considered. In the restricted (bottom) case, only pathways altered in between 25% and 75% of samples were considered (dotted red lines), giving 7,000 tests.

2.4.2. Two novel examples of inherited vulnerability in OAC

Applying the above approach resulted in 25 significant associations between damaged germline pathways and somatic driver pathways with FDR <0.1 (Table 2-8, Figure 2-7A). However, considering the overlap of the pathways involved, these 25 hits represented two distinct associations. This could be seen from the inter-correlations of the significant germline (Figure 2-7B) and somatic (Figure 2-7C) pathways.

The majority of the hits (23 out of 25) associated germline perturbations to digestive pathways with somatic driver alterations in FGFR signalling. However, closer investigation of gene-level data suggested that this in fact represented an association between extracellular matrix (ECM) germline perturbations with receptor tyrosine kinase (RTK) signalling somatic drivers, as discussed in Section 2.4.3. The other two hits associated germline damage to DNA replication and repair with driver alterations in genes that were downstream of the RTK signalling, as discussed in Section 2.4.4.

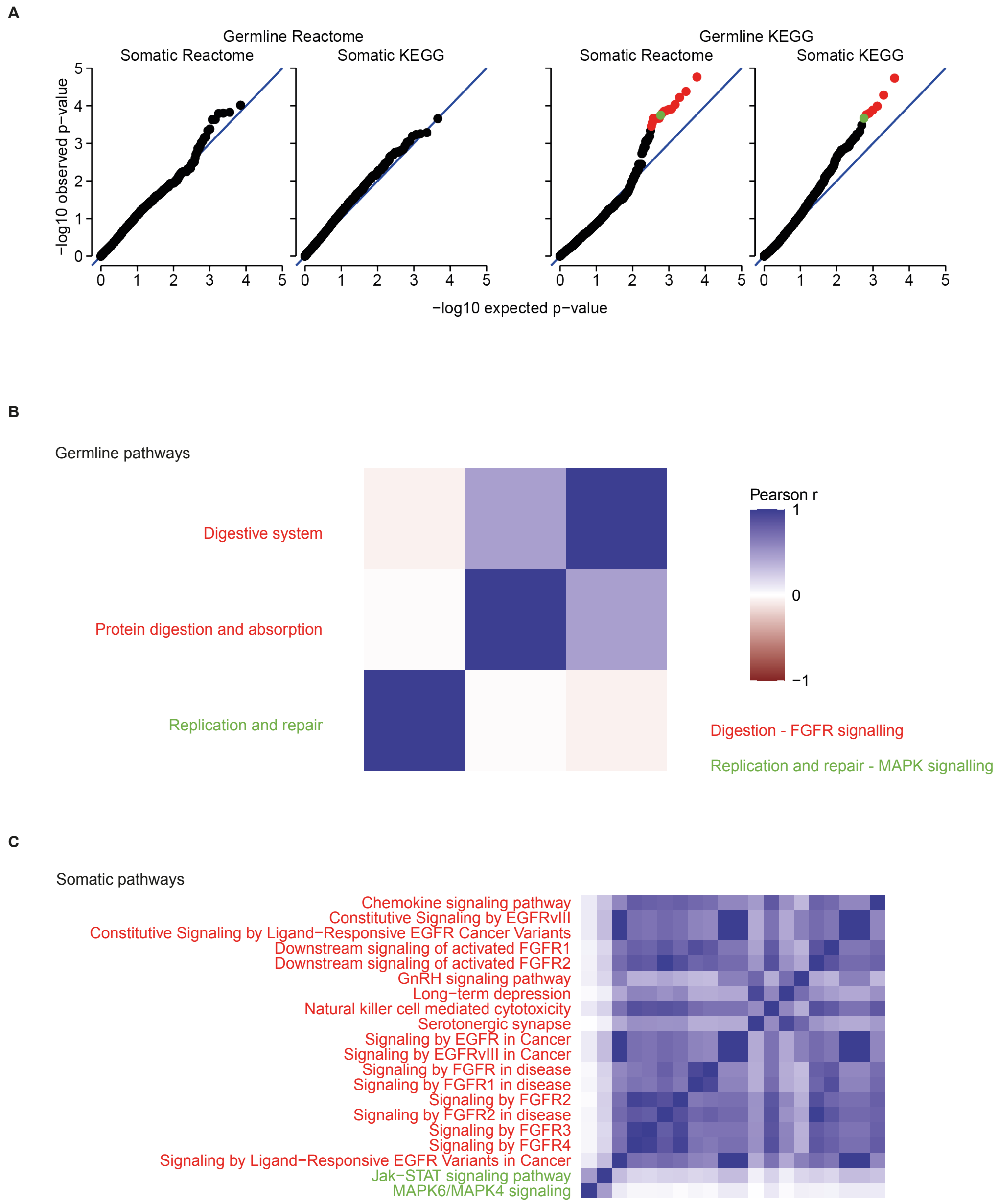

Figure 2-7: Results of the germline-somatic association analysis

A. QQ plots showing the statistical significance of Fisher tests, stratified by the source database of the germline and somatic pathways. Significant results (FDR <0.1) are coloured red if they related germline perturbations to digestive pathways with somatic drivers in FGFR signalling pathways, and green if they related germline perturbations to DNA replication and repair with somatic drivers in MAPK signalling pathways.

B. Inter-correlation of the three significant germline pathways. Each pathway was treated as a binary variable, and the Pearson correlation coefficient was calculated between each pair of pathways. Pathways are coloured as in **A.**

C. Inter-correlation of the 18 significant somatic pathways. Pathways are coloured as in **A.**

Germline pathway	Somatic pathway	OR	FDR	Germline samples	Somatic samples	Intersection
Protein digestion and absorption*	Signalling by FGFR in disease*	3.19	0.022	89	91	47
Digestive system	Serotonergic synapse	4.43	0.024	188	85	76
Digestive system	Signalling by FGFR2	4.16	0.043	188	80	71
Digestive system	Natural killer cell mediated cytotoxicity	4.58	0.051	188	72	65
Protein digestion and absorption	Downstream signalling of activated FGFR2	3.17	0.053	89	68	37
Digestive system	Downstream signalling of activated FGFR2	4.37	0.065	188	68	61
Protein digestion and absorption	Signalling by FGFR2 in disease	3.03	0.065	89	67	36
Digestive system	Signalling by FGFR2 in disease	4.27	0.065	188	67	60
Protein digestion and absorption	Signalling by FGFR2	2.86	0.065	89	80	41
Digestive system	Signalling by FGFR4	3.80	0.065	188	76	67
Protein digestion and absorption	Signalling by FGFR1 in disease	2.80	0.065	89	83	42
Replication and repair*	MAPK6/MAPK4 signalling*	2.75	0.065	85	90	43
Digestive system	Signalling by FGFR3	3.52	0.065	188	79	69
Digestive system	Constitutive Signalling by EGFRvIII	3.22	0.065	188	92	79
Digestive system	Constitutive Signalling by Ligand-Responsive EGFR Cancer Variants	3.22	0.065	188	92	79
Digestive system	Signalling by EGFR in Cancer	3.22	0.065	188	92	79
Digestive system	Signalling by EGFRvIII in Cancer	3.22	0.065	188	92	79
Digestive system	Signalling by Ligand-Responsive EGFR Variants in Cancer	3.22	0.065	188	92	79
Digestive system	Long-term depression	3.75	0.081	188	84	74
Digestive system	Signalling by FGFR in disease	3.15	0.082	188	91	78
Digestive system	GnRH signalling pathway	3.23	0.085	188	105	90
Digestive system	Chemokine signalling pathway	3.50	0.085	188	87	76
Protein digestion and absorption	Serotonergic synapse	2.77	0.085	89	85	43
Replication and repair	Jak-STAT signalling pathway	2.89	0.094	83	157	64
Protein digestion and absorption	Downstream signalling of activated FGFR1	2.73	0.096	89	72	37

Table 2-8: Results of the germline-somatic association analysis

*Indicates germline-somatic pathway pairs that were used for downstream analysis.

2.4.3. Germline ECM perturbations associate with RTK signalling driver events

The first distinct result from the germline-somatic association analysis associated germline perturbations to digestive pathways with somatic driver alterations in FGFR signalling-related pathways (Table 2-8). The two germline pathways involved (Digestive System and Protein Digestion & Absorption) were directly related in the KEGG hierarchy, with Digestive System containing Protein Digestion & Absorption. By contrast, the 18 somatic pathways were drawn from both Reactome and KEGG and were not directly adjacent to each other in the database hierarchies. However, many of these pathways clearly overlapped with each other, with 8 of them pertaining to FGFR signalling. The most significant hit was between the Protein Digestion & Absorption pathway and FGFR Signalling in Disease (FDR = 0.02). In the analysis in this section, I will take this as representative of the more general result.

In order to probe this association, I stratified the cohort of 260 OACs into four groups (Figure 2-8A): those with both the germline and somatic pathways altered ('G+S+', n=47); those with the germline pathway damaged only ('G+S-', 42); those with somatic only ('G-S+', 44); and those with neither ('G-S-', 127). I then investigated the genes most recurrently involved in both the germline and somatic pathways. The germline was strikingly dominated by collagen genes (Figure 2-8B), which accounted for 66 of the 89 G+S+ and G+S- samples (74%). Collagens are not obviously associated with digestion, and the KEGG map of the Protein Digestion & Absorption pathway lists collagens under the heading of 'Nondigestible proteins'. This suggested that the result may have been driven by germline perturbations to the ECM more than to digestive processes. Indeed, retaining only the collagen genes in the pathway still gave a significant result (FDR =0.04). A notable feature of the somatic driver genes in the G+S+ and G-S+ samples was that, despite the relevant pathway being FGFR Signalling in Disease, the most commonly altered genes were not specific to FGFR signalling. Instead, they were genes that are more widely important in RTK signalling, namely *KRAS*[113,114] and *PIK3CA* (Figure 2-8B). Thus, the association could be interpreted as germline perturbations to the ECM conferring an inherited vulnerability to RTK driver alterations.

As an orthogonal investigation, I compared the G+S+ samples to the other groups in the cohort using clinical and gene expression data. None of the available clinical indicators (including tumour stage, gender, age at diagnosis and survival time) significantly differentiated the G+S+ samples. Analysis of the 92 samples that had available RNA-Seq data revealed 80 genes that were overexpressed (fold-change >1.5) in the G+S+ group, compared to each of the other three groups separately (Methods 2.2.4). A gene set enrichment analysis (GSEA) of these 80 genes revealed 9 enriched pathways with FDR $<10^{-3}$ (Methods 2.2.4), all of which were associated with the ECM (Table 2-9). In particular, collagen pathways appeared several times. Indeed, 15 of the 80 upregulated genes (19%) were part of Reactome's Extracellular matrix organisation pathway, and this included five collagens (Figure 2-8C). This also provided orthogonal evidence that the RDGVs in collagens represented genuine biological perturbations. Moreover, among the 66 genes that were downregulated (fold-change <0.67) in the G+S+ samples, there was no clear enrichment in any biological processes. This suggested that samples with both germline ECM perturbations and RTK signalling somatic drivers had more transcriptional activity related to the ECM.

There are at least two potential biological connections between ECM perturbations and RTK signalling. One possibility is that the association could be mediated via discoidin domain receptors (DDRs). There are two human DDR genes (DDR1 and *DDR2*), both of which are RTKs that are triggered by collagen[115]. Moreover, there is evidence that they are involved that cancer progression, particularly in invasion and metastasis[115-117]. Interestingly, *DDR2* was expressed most highly in the G+S+ group (fold-change =1.4), although this difference only reached significance in comparison to G-S- samples (the largest group, Figure 2-8D). This could suggest that germline perturbations to collagens were leading to increased *DDR2* signalling, which increased the advantage of RTK signalling driver alterations. Another connection could be via integrins, which bind to ECM components and are known to cooperate with several RTKs, including *EGFR* and *VEGFR*[118]. Integrin inhibitors have been explored as potential anticancer drugs[119]. Interestingly *ITAV*, which codes for a component of the proposed targets (integrins $\alpha v\beta 3$ and $\alpha v\beta 5$), showed a trend of increased expression in the G+S+ group. However, as with *DDR2* this difference was only statistically significant compared to the G-S- group (Figure 2-8E). These connections are speculative, and a fuller investigation into potential mechanisms for mediating the interaction between germline ECM perturbations and RTK driver alterations would require larger sample sizes and experimental evidence.

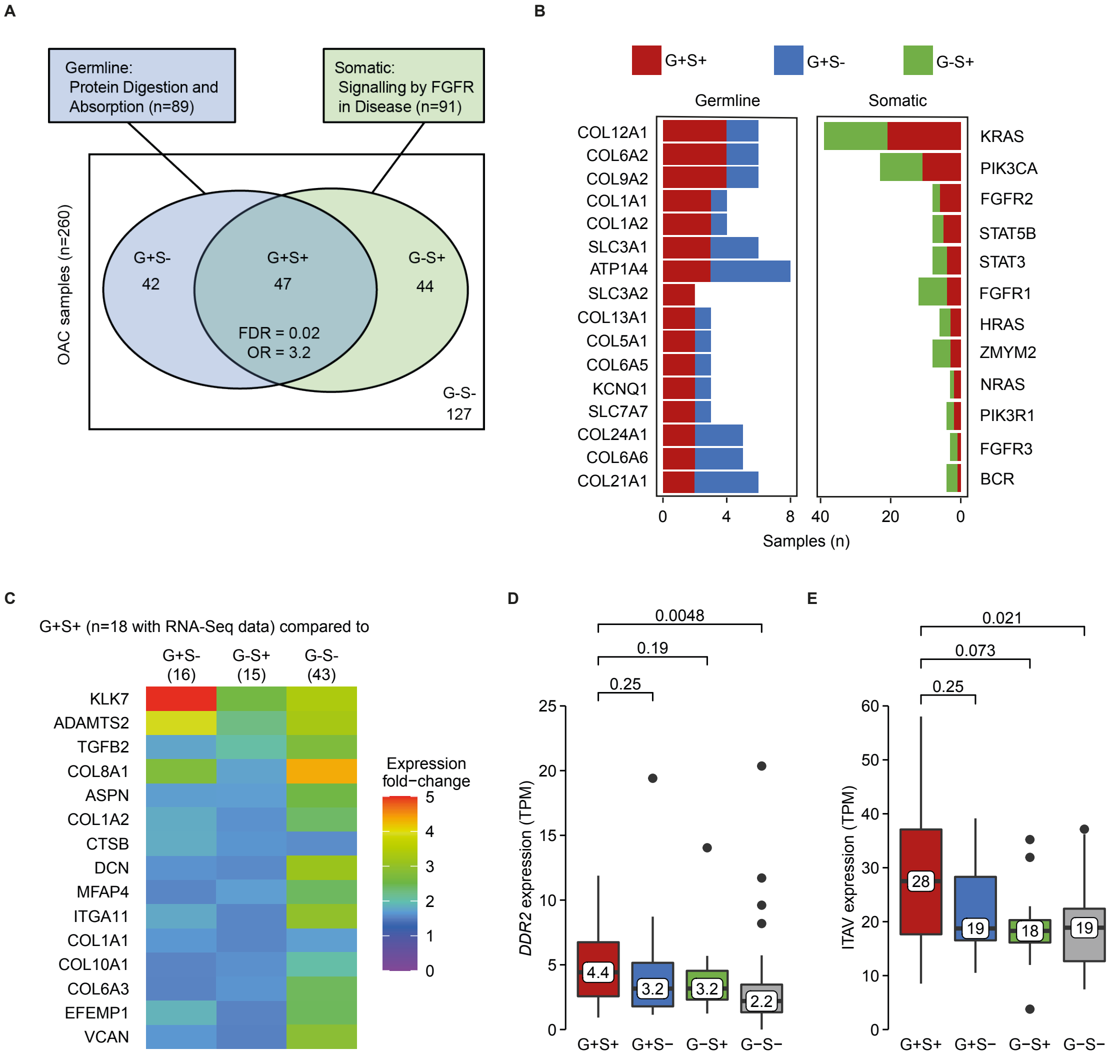

Figure 2-8: Germline ECM perturbations associate with RTK driver alterations

A. Stratification of the OAC cohort by germline status (G+/- indicates samples with/without RDGVs in the Protein Digestion and Absorption pathway) and somatic status (S+/- indicates samples with/without somatic driver alterations in the Signalling by FGFR in Disease pathway).

B. Genes involved in the germline-somatic interaction. For each gene, the number of samples of samples with a RDGV (left) or somatic driver alteration (right) is shown, coloured by the germline/somatic status of each sample.

C. Genes in the Extracellular matrix organisation pathway that were upregulated in G+S+ samples with >1.5 fold-change of median TPM. Numbers of samples with available RNA-Seq data are shown in brackets.

D. Expression of *DDR2* in OAC samples, stratified by germline/somatic status. P-values from Wilcoxon rank-sum test.

E. Expression of *ITAV*, stratified by germline/somatic status.

Pathway	Source	FDR	OR
Extracellular matrix organisation	Reactome	4.26E-09	15.26
ECM proteoglycans	Reactome	3.98E-05	26.58
Degradation of the extracellular matrix	Reactome	7.40E-05	16.31
Collagen formation	Reactome	7.40E-05	21.75
Assembly of collagen fibrils and other multimeric structures	Reactome	1.20E-04	27.81
Collagen degradation	Reactome	1.33E-04	26.37
Collagen biosynthesis and modifying enzymes	Reactome	1.51E-04	25.08
Collagen chain trimerisation	Reactome	3.92E-04	32.26
Integrin cell surface interactions	Reactome	4.84E-04	19.34

Table 2-9: Pathways enriched in genes overexpressed in G+S+ samples
A total of 80 genes with 1.5-fold overexpression were identified. Pathways enriched with FDR <0.001 are shown.

2.4.4. Germline DNA repair defects associate with driver events downstream of RTK signalling

The second distinct result found by the germline-somatic interaction analysis related germline damage in the DNA replication and repair pathway to driver events in the MAPK6/MAPK4 and Jak-STAT signalling pathways (Table 2-8). As in Section 2.4.3, I took the most significant of these hits (MAPK6/MAPK4) as representative of the general result, and stratified the cohort into four groups depending on germline and somatic status (Figure 2-9A).

In the germline, most of the damaged genes were involved in DNA repair rather than replication, and they included a number of CPGs, including *MLH1*, *PMS2*, *BRCA2* and the Fanconi Anaemia genes *FANCA/L/M* (Figure 2-9B). However, testing for drivers associated with germline damage to CPGs in general did not yield significant results, suggesting that this hit was not driven simply by CPGs. The somatic driver alterations were predominantly in *MYC*, *FOXO1* and *CCND3* (Figure 2-9B). Interestingly, these are all downstream of RTK signalling[120-122], suggesting that different germline backgrounds (*i.e.* damage to the ECM or DNA repair) could give rise to somatic evolutionary trajectories that converged to different but closely related aspects of cancer biology.

Analysis of gene expression revealed 98 genes with 2-fold overexpression in G+S+ samples, compared to the other three groups. These genes were strongly enriched in Keratinisation, and its sub-pathway Formation of the cornified envelope (Table 2-10). This pathway contained 16 of the overexpressed genes, the majority of which were either keratins (n=6) or small proline-rich proteins (n=5, Figure 2-9C). The two most strongly overexpressed genes, *KRT5* and *KRT14*, are the main keratins expressed by keratinocytes in the stratified squamous epithelium[123]. Intriguingly, while the normal oesophagus is not keratinised, oesophageal keratinisation is associated with oesophageal squamous cell carcinoma, rather than with OAC[124]. However, it is unclear why this particular germline and somatic background would be associated with a squamous morphology.

Stratified analysis of clinical data revealed that the G+S+ samples had a significantly lower age at diagnosis compared to other samples (Figure 2-9D). This could be related to the presence of CPGs in the germline pathway, since predisposition leads to earlier onset of many cancer types[125-128]. However, the G+S- samples, which had germline damage to DNA repair but not somatic drivers in the MAPK6/MAPK4 signalling pathway, did not exhibit an earlier onset of OAC (Figure 2-9D). In addition, samples with RDGVs in CPGs in general did not show earlier onset of OAC compared to other samples (p =0.26, Wilcoxon rank-sum test). This suggested that the combination of germline perturbations and somatic drivers, rather than simply effects from CPGs, were responsible for the younger age at diagnosis.

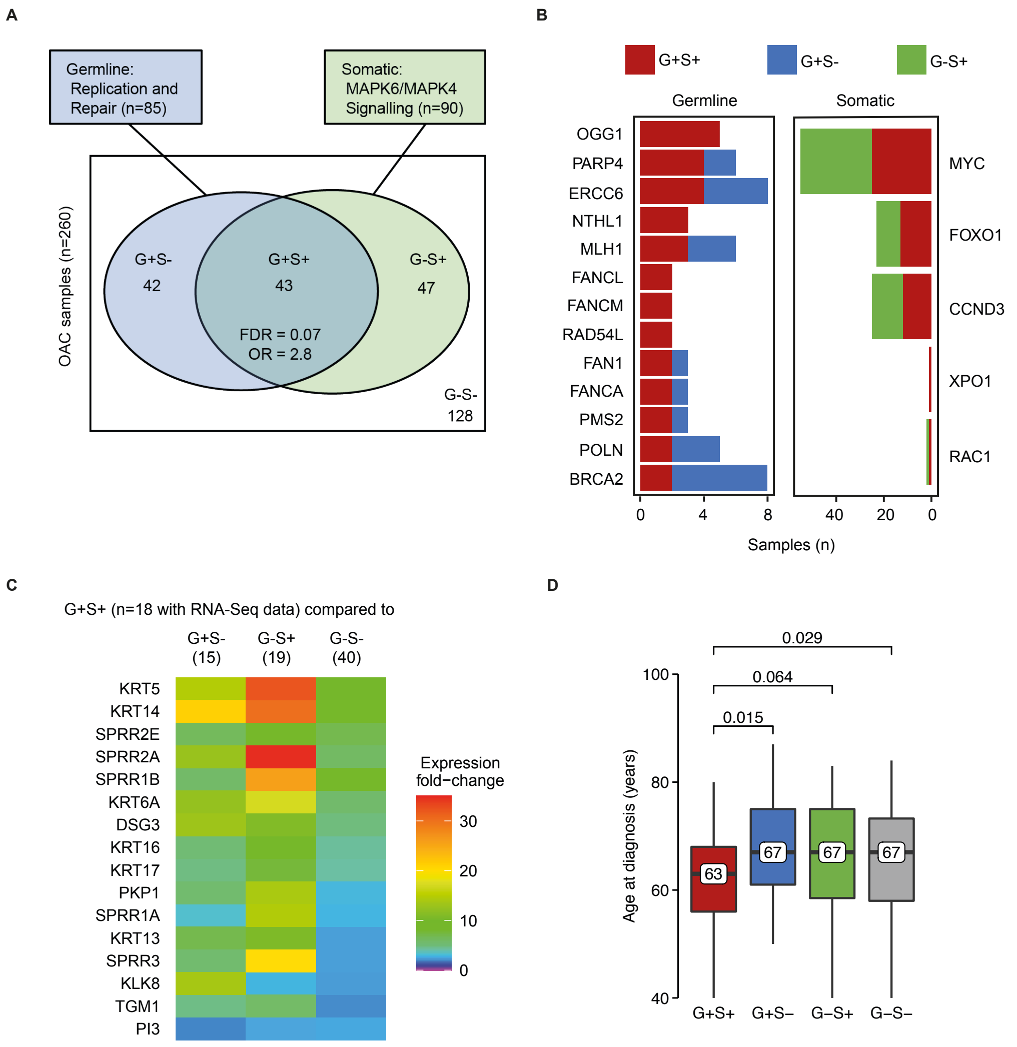

Figure 2-9: Germline DNA repair perturbations associate with driver alterations downstream of RTK signalling

A. Stratification of the OAC cohort by germline/somatic status.

B. Genes involved in the germline-somatic interaction.

C. Genes in the Formation of the cornified envelope pathway that were upregulated in G+S+ samples with >2 fold-change of median TPM.

D. Age at diagnosis, stratified by germline/somatic status. P-values from Wilcoxon rank-sum test.

Pathway	Source	FDR	OR
Formation of the cornified envelope	Reactome	5.17E-15	34.58
Keratinisation	Reactome	1.72E-12	22.03
Developmental Biology	Reactome	1.38E-04	4.70

Chemokine receptors bind chemokines	Reactome	3.33E-03	23.56

Table 2-10: Pathways enriched in genes overexpressed in G+S+ samples
A total of 98 genes with 2-fold overexpression were identified. Pathways enriched with FDR <0.01 are shown.

2.4.5. Results in an independent validation cohort

In order to see if the results of the germline-somatic association analysis were robust, I tested them in an independent validation cohort of an additional 140 OAC samples that became available from the OCCAMS consortium. I processed the data for these samples in the same way as described in Section 2.3, identifying pathways with rare damaging germline variants and pathways with somatic driver alterations. I then tested the two representative germline-somatic pathway pairs from the initial analysis (Table 2-8) in the validation cohort. Unfortunately, these results could not be reproduced (Table 2-11). This suggested that the results of the initial analysis were not statistically robust. Although it is difficult to be certain why this was the case, there are a number of possible reasons.

One explanation for the lack of validation lay with the method of multiple p-value correction. I had calculated FDRs in the initial analysis using the method of Ji and Li[112] to account for inter-correlations, and this relaxation of correction may have been too lenient. However, this choice did not have a substantial impact, since using the standard Benjamini-Hochberg correction[89] instead, both of the results discussed in Sections 2.4.3 and 2.4.4 would still have had FDR <0.08. Additionally, since different pathway databases provide similar but complementary information, I had corrected for multiple hypothesis tests treating associations between pathways from different databases separately. This may have been overly permissive. Indeed, correcting p-values from all tested germline-somatic pathway pairs together regardless of database, no results had FDR <0.15. As an additional benchmark, using the more stringent Bonferroni correction for FWER would have only given two out of the 25 results in Table 2-8 with FWER <0.5. It is also worth noting that the QQ plots in Figure 2-7A showed features simiar to genomic inflation[129], suggesting that some of the p-values themselves may not have reflected the true statistical significance of the results. These considerations suggested that the approach taken to multiple hypothesis test correction may have been too lenient.

Another explanation could be that, due to the basic nature of the Fisher test, I was unable to account for covariates. In particular, clinical features such as gender and environmental exposures, as well as genetic features such as population structure, may have had confounding effects. In order to account for these factors, a more sophisticated model would be required to analyse germline-somatic interactions.

Germline pathway	**Somatic pathway**	**OR**	**P-value**	**Germline samples**	**Somatic samples**	**Intersection**
Protein digestion and absorption	Signalling by FGFR in disease	0.94	0.63	46	50	16
Replication and repair	MAPK6/MAPK4 signalling	0.85	0.73	57	37	14

Table 2-11: Results of validation testing

2.5. Discussion

In this chapter, I have described a simple systematic investigation into germline-somatic interactions. Using OAC as a model cancer type, I first curated the data for a cohort of 260 samples, ensuring the quality of germline variant calls in particular. I then identified high-confidence RDGVs, and aggregated these into pathways. By measuring co-occurrence between RDGVs in pathways and somatic driver events in cancer-related pathways (identified by a colleague), I found two distinct results suggestive of inherited vulnerability to particular driver alterations. Further investigation of these associations using gene expression and clinical data was promising. In the first association, inherited defects in ECM genes, and particularly in collagens, led to an increased rate of driver alterations in RTK signalling genes, particularly *KRAS* and *PIK3CA*. Discoidin-domain receptors and integrins both provided potential mechanisms to mediate this association, and showed interesting trends of gene expression supporting the result. In the second germline-somatic association, germline perturbations to DNA repair co-occurred with driver events downstream of RTK signalling, particularly in *MYC*, *FOXO1* and *CCND3*. While potential mechanisms for this association were lacking, expression analysis suggested a link with tumour morphology, as germline/somatic status associated very strongly with upregulation of keratinisation genes. In addition, a statistically significant association with earlier age at diagnosis hinted at the possible clinical relevance of this association.

Both of these results demonstrated the potential for analysis of germline-somatic associations to further our understanding of cancer biology. Ultimately, however, neither of these associations could be reproduced in a validation cohort of 140 OAC samples, suggesting that the results were not statistically robust. In order to improve future analysis, both the data preparation and characterisation process and the applied method for identifying germline-somatic interactions needed to be critically evaluated.

Germline variants were extensively quality controlled in terms of their numbers per sample, read supporting alternative alleles, MAFs and deviation from Hardy-Weinberg equilibrium (Figure 2-2, Section 2.3.2). Benchmarking variants against reference samples from the 1000 Genomes Project[68] indicated that false positive variant calls were not prevalent in the OAC cohort. However, only rare variants (MAF <1%) were used in the germline-somatic association analysis, and identifying false positive calls for rare variants is challenging. Moreover, rare variants are proportionally enriched for false positives compared to common variants[69]. However, since truly deleterious variants are very unlikely to be common in populations, the restriction to rare variants was appropriate on balance. The approach for identifying deleterious variants was restrictive, requiring the support of a large number of prediction algorithms for non-truncating variants. Thus, the final set of RDGVs were high-confidence, and orthogonal evidence of their deleteriousness was provided by gene expression analysis, which revealed that germline perturbations to the ECM led to dysregulation of ECM genes in tumours (Figure 2-8, Section 2.4.3). However, the restrictive nature of this approach left relatively few RDGVs per patient (median 36, Figure 2F, Section 2.4.3). A slightly more permissive approach may have given more variants for analysis without compromising the integrity of the RDGV set.

The identification of somatic driver events could be improved in several ways. Drivers were identified using a list of known cancer genes[25,26,111], but this was a pan-cancer list, and was not specific to OAC. This was because the original purpose of these driver events was to characterise the features of cancer genes[27], as discussed in Chapter 4. However, for the germline-somatic association analysis, a more tailored approach could be beneficial. Using only driver genes shown to play a role in OAC could remove some false positive driver alterations. In addition, the driver events considered in this chapter included structural variants (SVs) such as inversions and breakends[27]. While SVs are prevalent in OAC[73], it is unclear to what extent they constitute driver events *per se*, or are simply the by-product of genomic instability. Thus, considering only mutation and copy number data could further refine the identification of drivers. Finally, while driver events were considered at the level of pathways in this chapter, the driver components of the results of the germline-somatic association analysis were dominated by individual highly recurrent genes. This suggests that treating drivers at the gene level might provide a more focused approach without losing relevant information. By contrast, the germline contribution to the results was diffuse and spread across many genes, indicating that considering germline perturbations at the pathway level was a useful approach.

The method used to identify germline-somatic interactions in this chapter was a simple first-pass approach to the problem, and could be improved in various ways. The Fisher test has the advantage of making very few assumptions. However, treating damage to germline pathways as a binary categorical event is crude. Instead, using counts of damaging variants could capture more information about samples. Additionally, as discussed in Section 2.4.5, the Fisher test does not allow potentially confounding covariates to be accounted for. A more sophisticated model would be able to address both of these issues. A retrospective investigation in Section 2.4.5 showed that the method of multiple p-value correction employed may have been overly permissive. In particular, treating different pathway databases separately had a substantial impact on FDRs. Moreover, the overlap between pathways complicated the interpretation of results and may have inflated measures of statistical significance. Pruning pathways before analysis to obtain a less redundant set of variables could mitigate these issues. The frequency restriction approach described in Section 2.4.1 substantially alleviated the burden of multiple hypothesis testing and removed tests of association between very rare or ubiquitous events. A similar approach should ideally be applied to any future analysis. Finally, using one-sided tests to focus on co-occurrence and the identification of inherited vulnerability to somatic driver events may have been overly restrictive. Germline perturbations being mutually exclusive with somatic driver events could indicate either compensatory or synthetic lethal interactions, both of which would be of interest. Moreover, using two-sided tests would only reduce the allowable test size by half, which represents only a marginal loss in comparison to the gains from frequency restriction.

While the results of the germline-somatic association analysis in this chapter could not be validated, an improved analysis may still provide useful insights into cancer biology. In Chapter 3, I will continue to investigate germline-somatic interactions in oesophageal adenocarcinoma, addressing many of the methodological limitations discussed here. I will also analyse a larger cohort of OAC samples, lending greater sensitivity and robustness to the analysis therein.

Chapter 3. Statistical modelling of germline-somatic interactions *in trans*

3.1. Introduction

As I discussed in Chapter 1, germline variation is a potential contributor to inter-tumour heterogeneity, with inherited genetic differences affecting evolutionary pressures acting on somatic alterations. In Chapter 2 I illustrated a co-occurrence-based approach for analysing germline-somatic interactions *in trans* (GSITs), although the results could not be validated in an external cohort. This suggested that improved methodologies and larger sample sizes might be required to robustly identify the influence of germline variants on somatic evolution.

In this chapter, I will attempt to address some of the limitations of the analysis in Chapter 2, which I discussed in detail in Section 2.5. Since the methods used to identify rare damaging germline variants (RDGVs) in Chapter 2 were restrictive, I will use a more permissive approach here. For somatic alterations, I observed that perturbed pathways were dominated by the same highly recurrent driver genes. Thus, in this chapter I will begin by focusing on individual driver genes instead. Later, I will also combine contributions from individual driver genes in a data-driven clustering approach. To improve upon the robustness of the Fisher testing approach for uncovering GSITs, I will instead use logistic regression modelling to account for potentially confounding covariates. Finally, I will use an expanded OAC cohort of 470 samples, allowing for greater statistical power. However, I will continue to focus on OAC as a model cancer type with an absence of highly penetrant predisposition genes[23] and poor clinical outcomes[93]. Nonetheless, the methodological questions of how best to analyse GSITs could apply to any cancer type.

In Section 3.3 I will describe the data preparation for an expanded cohort of 470 OAC patients, including germline data, somatic data, and potentially confounding covariates for downstream analysis. I will then describe the application of a logistic regression approach to associate RDGVs in pathways with the driver status of OAC driver genes in Section 3.4. This analysis uncovers two statistically significant results, which I investigate in more detail in Section 3.5. I will then characterise somatic variation with a more comprehensive set of descriptors, including multiple driver genes and other descriptors of somatic alterations, by adopting a sample clustering approach in Section 3.6, and attempt to relate RDGVs in pathways to the resulting clusters. Finally, I will discuss the results from this chapter in Section 3.7.

3.2. Methods

3.2.1. Clinical and demographic data

Clinical and demographic data were provided by ICGC for 400 OAC patients, and were extracted from the Genomic Data Commons (https://portal.gdc.cancer.gov/, downloaded December 2018) for 185 TCGA oesophageal cancer patients. Of these, 89 patients had a histologic diagnosis of "Esophagus Adenocarcinoma" and were retained. Intersecting TCGA patients with those with available somatic[15] and germline[49] sequencing data gave 70 patients.

Missing data were imputed for features to be used in downstream analysis using the median value across the combined ICGC and TCGA cohort.

3.2.2. Germline variant filtering

Germline variant calls were provided by ICGC for 400 OAC patients, and obtained for 70 TCGA OAC patients from the release of Huang *et al.*[49] via the Genomic Data Commons portal. VCF files were annotated with ANNOVAR[74] (downloaded November 2018), and only variants annotated as either exonic or splicing were retained. Annotated genes were intersected with a set of 19,549 genes derived from the alignment of reference protein sequences to the human genome[24], leaving 11.1x10^6 germline variants.

Variants were first filtered using their zygosity (as labelled in VCF files) and supporting read counts. Only variants supported by five or more alternative reads were retained. Heterozygous variants were removed if their variant allele frequency (VAF) satisfied either $VAF < 25\%$ or $VAF > 75\%$, where $VAF = \frac{\#\,\text{alternative reads}}{\#\,\text{total reads at locus}}$. Homozygous variants were removed if $VAF < 90\%$. Combined, these filters removed 387,003 exonic and splicing germline variants.

The remaining variants were then filtered to remove variants present at higher frequencies in the OAC dataset compared to reference populations. Minor allele frequencies (MAFs) of variants were annotated with ANNOVAR, using data from the 1000 Genomes[68] and ExAC projects[76]. The observed MAF of each variant in OAC was then compared to the reference MAFs using a beta-binomial test. First, the parameters of a beta distribution were fit to the empirical distribution of observed MAFs in the OAC cohort using maximum likelihood estimation. The resulting parameters a_{prior} and b_{prior} thus determined a prior distribution for MAFs. Second, for each observed variant, a posterior beta distribution was fit using the MAFs in reference populations. Posterior beta parameters were calculated as $a_{post} = a_{prior} + 2N_{ref}MAF_{ref}$, $b_{post} = b_{prior} + 2N_{ref}(1 - MAF_{ref})$, where MAF_{ref} is the MAF of the variant in either ExAC or 1000 Genomes, and N_{ref} is the number of individuals in the reference population (60,706 for ExAC or 2,504 for 1000 Genomes). If a MAF was available in ExAC, this was used preferentially over 1000 Genomes. If a MAF was not available from either source, MAF_{ref} was set to zero. For each variant, the posterior beta distribution was then used to parametrise a beta-binomial distribution, from which p-values of enrichment in OAC were calculated. A total of 274,056 variants with $p < 10^{-12}$ and observed MAF>10% in OAC were removed.

3.2.3. RDGV identification

The filtered exonic and splicing germline variants were annotated with deleteriousness predictions, similarly to Section 2.2.3. Truncating variants included stopgain, stoploss and frameshift variants. Splicing variants were considered damaging if they were predicted as such by either the ADA or RF methods from dbscSNV[78].

Missense SNPs were annotated with CADD[61] (phred score >25) and a set of seven function-based predictors (SIFT[80], PolyPhen2-HDIV[81], PolyPhen2-HVAR, MutationTaster[82], MutationAssessor[83], LRT[84], and FATHMM[85]). Missense variants supported by both CADD (phred score >25) and at least five of the seven function-based predictors were considered as damaging for downstream analysis. Annotations from MetaLR were not used in the definition of damaging germline variants. RDGVs consisted of damaging germline variants with $MAF_{ref} < 1\%$.

3.2.4. Variance-weighted Jaccard index

The variance-weighted Jaccard index (VWJI) was used to assess the functionally relevant overlap of pathways. First, the variance v_g of the RDGV allele count across the OAC cohort was calculated for each gene g. For two pathways p and q, the VWJI was calculated as

$$\mathrm{VWJI}_{pq} = \frac{\sum_{g \in p \cap q} v_g}{\sum_{g' \in p \cup q} v_{g'}}.$$

Thus, two pathways sharing a gene with high variance (or equivalently for Poisson-distributed RDGV allele counts, high mean) across the cohort was given more weight than sharing a gene with low variance.

The VWJI can be cast in a linear algebra form for fast computation with large numbers of genes and pathways. First, define the gene-pathway membership matrix $\boldsymbol{M}$ as $M_{gp} = \mathbb{I}[g \in p]$, where $\mathbb{I}$ denotes an indicator function taking the value 1 when its argument is true, and zero otherwise. From $\boldsymbol{M}$**,** the matrices $\overline{\boldsymbol{M}}$ with $\overline{M}_{gp} = 1 - M_{gp}$ (complement membership matrix) and $\widetilde{\boldsymbol{M}}$ with $\widetilde{M}_{gp} = v_g M_{gp}$ (variance-weighted membership matrix) can be constructed. It can then be shown that

$$\mathrm{VWJI}_{pq} = \frac{\left(\widetilde{\boldsymbol{M}}^T \boldsymbol{M}\right)_{pq}}{\left(\widetilde{\boldsymbol{M}}^T \boldsymbol{M}\right)_{pq} + \left(\widetilde{\boldsymbol{M}}^T \overline{\boldsymbol{M}}\right)_{pq} + \left(\overline{\boldsymbol{M}}^T \widetilde{\boldsymbol{M}}\right)_{pq}}.$$

The VWJI was calculated between all pairs of pathways. In order to reduce pathway overlap, groups of pathways for which the VWJI exceeded 0.8 in a transient manner between all pathway pairs were first identified (pathways p and q were considered to be transiently overlapping if $\mathrm{VWJI}_{pp'} > 0.8$ and $\mathrm{VWJI}_{p'q} > 0.8$ for some pathway p'). Within these groups, a single representative pathway was chosen as the pathway with the highest mean VWJI relative to all other pathways in the group.

3.2.5. Genetic ancestry

A total of 45,398 SNPs common SNPs with MAF >5% in reference populations were first identified. The genotypes of these SNPs (encoded as 0 for homozygous reference, 1 for heterozygous and 2 for homozygous alternative) were used as in the input for principal components analysis (without scaling).

3.2.6. Somatic driver alterations

Somatic small variant calls (SNVs and indels) were obtained from the MC3 TCGA release[15] and annotated with ANNOVAR[74]. Hotspot mutations were identified with OncodriveCLUST v1.0 with default settings. Non-truncating damaging mutations included missense SNVs labelled as deleterious according to either at least five of seven function-based predictors (SIFT[80], PolyPhen2-HDIV[81], PolyPhen2-HVAR, MutationTaster[82], MutationAssessor[83], LRT[84], and FATHMM[85]) or at least two of three conservation-based predictors (GERP++[88], SiPhy[87] and PhyloP[86]), as well as splicing mutations predicted as deleterious by one or both of the ADA and RF methods from dbSCSNV[77]. Truncating mutations included stopgain, stoploss and frameshift mutations.

Copy number segments were obtained by running ASCAT[130] on SNP array data from the Genomic Data Commons. Segments were intersected with the coordinates of 19,549 human genes[24], and genes were considered to have a CNV if more than 25% of their length overlapped with a copy number segment. Ploidy values for each sample were also obtained from ASCAT. Gene amplifications included all genes with copy number > 2 x ploidy, and deletions included genes with copy number zero.

A list of 76 OAC-specific driver genes was obtained from Frankell *et al.*[96] and intersected with lists of canonical driver genes across cancer types [24] to give a total of 40 driver genes with at least one damaging somatic alteration in the OAC cohort (hotspot, truncating and non-truncating damaging mutations, amplifications and deletions). The OAC driver genes were divided in Frankell *et al.* into those driven by mutations, amplifications and deletions, based on what types of somatic alterations were recurrent in each gene according to a combination of statistical tools[96]. To map driver events in individual samples, damaging mutations in mutation-driven genes, amplifications in amplification-driven genes, and deletions in deletion-driven genes were considered to be driver alterations.

3.2.7. Mutational signatures

All exonic and splicing somatic SNVs were used to extract mutational signatures. The MutationalPatterns[131] R package was run with default settings for *de novo* signature discovery. For interpretation, the profiles of 30 pan-cancer mutational signatures were obtained from COSMIC[45]. The contributions of each of the 96 possible trinucleotide contexts were compared between pairs of de novo and COSMIC signatures using the cosine similarity. For two signatures x and y, the similarity S_{xy} was calculated according to the formula

$$S_{xy} = \frac{\sum_{i=1}^{96} x_i \times y_i}{\sqrt{\sum_{j=1}^{96} x_j^2} \times \sqrt{\sum_{k=1}^{96} y_k^2}},$$

where x_i is the contribution of the ith trinucleotide context to signature x.

3.2.8. Logistic regression power calculations

Simulated data were used to calculate the power of likelihood ratio tests (LRTs) in the logistic regression modelling framework. For simplicity, no covariates were included in these calculations. Predictor variables were generated using Poisson distributions with rate λ, and outcome variables were then generated by Bernoulli trials parametrised to reflect both the desired effect size β and the overall frequency $\bar{y}$ of the outcome. For given values of λ, β and $\bar{y}$, this process was iterated 1000 times, and the power was calculated as the percentage of iterations for with the LRT returned a p-value smaller than the Bonferonni significance threshold.

Power curves with λ on the x-axis were generated with adaptive test sizes. For each value of λ, the test size was taken as 0.05 divided by the number of germline pathways with rate parameter greater than λ. For a given driver gene g or a random forest cluster c (*i.e.* a given value of $\bar{y}$) and effect size $\beta = \pm 1$, the appropriate minimum rate parameter λ_{min}^{g} or λ_{min}^{c} to achieve 80% statistical power was then calculated using linear interpolation on the appropriate power curve.

3.2.9. Random forest clustering

Cancer-enriched pathways were identified by performing GSEA on OAC-specific driver genes from Frankell *et al.*[96] in 1,338 pathways from Reactome[107] with between 10 and 500 genes, and excluding level one pathways. Hypergeometric tests gave 115 pathways enriched with FDR <0.01. Duplicated pathways containing the same driver genes were removed, leaving a total of 64 distinct enriched pathways.

Random forest clustering was performed with the randomForest R package[132]. The data used to cluster samples included: counts of OAC driver genes with driver alterations in each enriched pathway; somatic exonic SNV burden; somatic exonic indel burden; ploidy; and the contributions of four *de novo* mutational signatures. Forests with 2,000 trees were used, and proximity between samples was calculated as the default, *i.e.* the proportion of trees in which a pair of samples ended up in same terminal node. For samples i and j with proximity p_{ij}, a distance was calculated as $d_{ij} = \left(1 - p_{ij}^{0.25}\right)$, with the exponent chosen to sensitise clustering to small proximities since many proximities were near-zero. Partitioning around medoids (PAM) clustering was performed with the cluster R package[133]. For each number of medoids between 2 and 15, 10 random forests were implemented to calculate silhouette widths (included in the PAM clustering routine) for each forest and the Adjusted Rand Index (using the mclust R package[134]) between each pair of forests.

3.3. Germline, somatic and clinical data preparation for a cohort of 470 OACs

In order to investigate GSITs in OAC, I first needed a suitable cohort of patients. Given this cohort, I then needed to characterise the relevant aspects of germline and somatic variation, as in Section 2.3. In this case, I also wanted to curate a set of covariates that could potentially confound downstream GSIT analysis, to be used as controls and to improve the robustness of the analysis. I divided covariates into three categories: clinical and demographic; germline; and somatic.

In this section, I will describe the preparation of data for a cohort of 470 OAC samples from ICGC[73,96] and TCGA[15,49]. In Section 3.3.1, I will describe the clinical and demographic covariates for these patients. I will then go on to characterise the landscape of deleterious germline variants and derive potentially confounding germline covariates in Sections 3.3.3 and 3.3.3. Similarly, I will characterise the landscape of somatic driver alterations and identify somatic covariates in Sections 3.3.4 and 3.3.5. This dataset will be the basis of the vast majority of the analysis in the remainder of this chapter.

3.3.1. Clinical and demographic covariates

I obtained clinical and demographic data for 400 ICGC and 70 TCGA OAC patients (Methods 3.2.1). However, I reasoned that not all clinical or demographic annotations would be useful as covariates for GSIT analyses. I therefore imposed two criteria for annotations to be included in downstream analysis. First, they had to plausibly relate to either germline variation, somatic driver alterations, or both. Since the main role of covariates in the analysis was to control for potentially confounding effects, including annotations that were *a priori* irrelevant would simply have added to model complexity without introducing useful information. Second, they had to have adequate data completeness. While small numbers of missing values could be handled by imputation (Methods 3.2.1), I required annotations to be at least 80% complete across the combined ICGC and TCGA cohort to avoid substantially biasing analyses.

Based on these criteria, I included four clinical and demographic covariates: gender; smoking status; treatment with neoadjuvant chemotherapy; and tumour stage. The distributions of these covariates across the OAC cohort are described in Table 3-1, and their relevance to GSIT analysis is described below. Importantly, race and ethnicity have potentially confounding effects on both germline and somatic cancer landscapes[70]. However, since these data were not sufficiently complete in demographic annotations (74%), I instead derived genetic ancestry directly from germline data (Section 3.3.3).

Gender is strongly linked to OAC incidence, with men being roughly nine times more likely to develop OAC than women[135]. Both the ICGC and TCGA cohorts reflected the male prevalence of the disease, with 86% male patients overall (Table 3-1). It is therefore reasonable to assume that gender has a significant influence on OAC biology. Indeed, it has been found that among OAC patients, men are substantially more likely than women to have a *TP53* driver mutation (roughly 80% for men compared to 60% for women)[96], for example. Thus, gender was a clear potential confounder for GSIT analysis.

The relevance of tobacco smoking was less clear. Unlike oesophageal squamous cell carcinoma, it is not certain to what extent smoking can predispose to OAC, although it does seem to have a weak positive correlation with OAC incidence[93,136,137]. In my data, the proportion of current smokers in the UK-based ICGC cohort (15%) was not dissimilar from the UK national smoking rate in 2019 (14.1%)[138]. However, smoking has been shown to strongly affect somatic driver alterations in other cancer types. For example, in lung cancer current smokers and never-smokers have *EGFR* mutation rates of 5% and 43% respectively[17]. On balance, I decided to include smoking data in my analyses to control for possible predisposing and driver gene effects in OAC.

Some clinical trials have suggested that neoadjuvant chemotherapy is an effective treatment choice for OAC, but the evidence so far is inconclusive[139]. A large proportion (62%) of the 470 OAC samples in my dataset received neoadjuvant chemotherapy. Since many of the samples for sequencing were taken at surgical resection, this meant that chemotherapy was administered before the tumour genomes were sequenced. It is known that chemotherapy can induce somatic mutations, and it has been associated with specific mutational signatures[140], although the extent of this effect in OAC may be small[141]. Moreover, the rate of neoadjuvant chemotherapy was strikingly different between ICGC (60%) and TCGA (0%) samples. This might be explained by different countries of origin (UK for ICGC, mostly US for TCGA) and different inclusion criteria. Including an indicator of neoadjuvant therapy in subsequent analysis allowed me to control for any potential confounding affects arising from therapy-related clinical differences, or from the effects of chemotherapy on somatic mutations.

Finally, tumour stage indicates how advanced tumours are, and it is an important prognostic indicator. In addition, more advanced tumours typically have more somatic alterations, with metastases having particularly marked increases in mutational burden[142]. Since complete group staging data were unavailable for this cohort, I used separate tumour (T), node (N) and metastasis (M) stage information. In both the ICGC and TCGA samples, more than half of patients were diagnosed with highly invasive primary tumours (T3/4), in line with previous reports[93] (Table 3-1).

Overall, the combined ICGC and TCGA cohort had the expected clinical and demographic characteristics for OAC patients. Moreover, there were few differences between the two data sources, as seen in Table 3-1.

Characteristic	**Category**	**ICGC (400)**		**TCGA (70)**	
		N	**%**	**N**	**%**
Gender	Female	57	14.2	7	10
	Male	343	85.8	63	90
Smoking status	Never	99	24.8	17	24.3
	Former	173	43.2	35	50
	Current	59	14.8	7	10
	Missing/Unknown	69	17.2	11	15.7
Neoadjuvant chemotherapy	Yes	237	59.2	0	0
	No	78	19.5	70	100
	Missing/Unknown	85	21.2	0	0
T stage	0	1	0.2	1	1.4
	1	25	6.2	14	20
	2	56	14	9	12.9
	3	254	63.5	43	61.4
	4	16	4	0	0
	Missing/Unknown	48	12	3	4.3
N stage	0	110	27.5	19	27.1
	1	160	40	40	57.1
	2	77	19.2	5	7.1
	3	18	4.5	3	4.3

	Missing/Unknown	35	8.8	3	4.3
M stage	0	315	78.8	49	70
	1	21	5.2	10	14.3
	Missing/Unknown	64	16	11	15.7

Table 3-1: Distributions of clinical and demographic covariates in 470 OACs

3.3.2. Identifying pathways with deleterious germline variants

I obtained germline variant calls (SNPs and indels) for the 400 OAC patients from ICGC[96] and 70 OAC patients from TCGA[49]. In a similar vein to Section 2.3.3, I chose to characterise the germline using putatively deleterious variants, as potential effectors of GSITs. Therefore, I retained a total of 11.1M exonic and splicing germline variants in an internally curated set of 19,549 human genes[24] (Methods 3.2.2, Table 3-2).

In order to remove possible technical artefacts, I applied two filters to the germline variants. First, I removed 387,003 variants that substantially deviated from the expected distribution of variant allele frequencies for homozygous or heterozygous variants (Methods 3.2.2, Table 3-2). This removed a particularly large number of variant calls (131,209) with a low proportion of reads supporting the variant (<25%) in TCGA samples (Figure 3-1A). The vast majority of these variant calls were SNPs as opposed to indels (97%), suggesting that these may have been artefacts arising from sample contamination, rather than problems with read alignment. Second, I removed a further 274,056 variants that were significantly enriched in the combined ICGC and TCGA OAC cohort, as compared to reference populations ($p < 10^{-12}$, beta-binomial test, Methods 3.2.2, Table 3-2). Among the variants removed were 33,293 that were not present in any reference populations but had high MAFs in the OAC cohort (Figure 3-1B). Both the ICGC and TCGA cohorts contributed to these variants (96% and 4% respectively), suggesting that many of them may have been false negatives in the reference populations. However, any false positives in the OAC cohort could have had substantial impacts on downstream analyses, since I went on to characterise germline variation only using variants that were rare in reference populations. Therefore, I removed these variants to avoid spurious results at a later stage.

As in Section 2.3.3, I used a combination of computational tools to identify putatively deleterious germline variants (Methods 3.2.3). In addition to protein-truncating and splicing variants, I had previously identified deleterious missense variants by requiring support from three composite sources: a consensus of seven function-based annotations (Methods 2.2.3); CADD[61]; and MetaLR[79]. In the combined ICGC and TCGA cohort MetaLR was particularly restrictive, only retaining 35% of variants that were deleterious according to both the consensus method and CADD (Figure 3-1C). Moreover, while variants with MetaLR support were rarer in reference populations than those only supported by the consensus and CADD, this difference was not substantial (median MAF 0.007% compared to 0.01%, Figure 3-1D). Deleterious variants are expected to be subject to negative selection and therefore rare in populations, so MAF is a useful readout of true positive deleteriousness. However, I judged that the reduction in MAFs achieved by requiring support from MetaLR was not substantial enough to warrant removing 65% of potentially deleterious missense variants. Therefore, I proceeded with deleterious missense variants that were supported by the consensus method and CADD. In order to filter out likely false positives, I retained 46,329 damaging variants that had a MAF <1% in reference populations (RDGVs, Table 3-2). Overall, this reduced the number of germline variants per sample being analysed from around 22,000 exonic and splicing variants to roughly 100 RDGVs, with good agreement between the ICGC, TCGA and reference European cohorts (Figure 3-1E).

Filtering stage	**ICGC (400)**		**TCGA (70)**	
	Total variants	**Unique variants**	**Total variants**	**Unique variants**
Total exonic/ splicing	9,415,541	233,338	1,736,793	113,464
VAF filtering	9,172,460	223,524	1,592,871	106,568
MAF filtering	8,929,218	222,383	1,562,057	105,778
Rare damaging	38,745	23,032	7,584	5,463

Table 3-2: Germline variant processing

Numbers of germline variants at different processing stages in the ICGC and TCGA cohorts of OAC patients. VAF and MAF filtering are described in detail in the Section 3.7.

In order to more usefully leverage the RDGVs, I aggregated them into 10,941 genes. As a further filter against possible technical artefacts, I removed genes that showed significant enrichment for RDGVs between cohorts. First, I compared the OAC cohort to 503 reference European samples from the 1000 Genomes Project[68] (EUR, Figure 3-1F). Second, I compared the ICGC and TCGA OAC cohorts to each other, in order to account for possible differences in sample preparation and sequencing methods. In total, I removed 29 genes that showed highly significant enrichment in any of these comparisons (Tables 3-3 and 3-4, $p<10^{-10}$ Fisher's exact test). As in Section 2.3.4, these included several genes that are well-known to be highly polymorphic or problematic for read alignment, such as the HLA genes *HLA-A* and *HLA-DRB1*, and those encoding the zinc finger proteins 806 and 880. Thus, it seemed plausible that these enrichments were due to technical artefacts rather than genuine biological differences between the cohorts.

Gene	**OAC (n)**	**EUR (n)**	**P-value**
HLA-DRB1	123	0	1.65E-43
ZNF806	72	0	1.10E-24
CACNA1B	69	3	1.31E-19
MUC6	57	1	5.93E-18
TBP	55	1	2.78E-17
TCAF2	47	0	4.75E-16
BCLAF1	44	0	4.89E-15
CD24	44	0	4.89E-15
CNOT1	40	0	1.07E-13
HLA-A	38	0	5.00E-13
AP3S1	36	0	2.32E-12
ADGRV1	35	0	4.98E-12
ZNF880	32	0	4.90E-11
LFNG	35	1	9.82E-11

Table 3-3: Genes removed due to strong enrichment in the OAC cohort (n=470) compared to reference European samples from 1000 Genomes (n=503)

Gene	**ICGC (n)**	**TCGA (n)**	**P-value**

ZNF806	2	70	3.81E-82
CACNA1B	4	65	1.13E-68
MUC6	1	56	3.63E-58
BCLAF1	0	44	4.74E-44
AP3S1	0	36	8.07E-35
CNOT1	2	38	3.19E-34
PTCHD3	13	44	5.18E-32
TCAF2	8	39	6.14E-30
HLA-DRB1	64	59	2.20E-29
LFNG	2	33	7.35E-29
HLA-A	4	34	6.51E-28
CDCP2	3	31	8.81E-26
FOXD4L4	0	24	2.51E-22
PMS1	5	28	3.59E-21
CHST15	5	25	2.20E-18
GNRH2	2	22	5.02E-18
CBWD6	0	18	1.47E-16
ZRANB1	0	17	1.26E-15
SULT1A1	1	17	2.02E-14
CCDC144NL	2	18	2.22E-14
TMEM254	0	15	8.90E-14
ARSD	3	18	1.38E-13
PSORS1C1	1	16	1.61E-13
CAMKK2	3	17	1.01E-12
AGAP4	0	12	4.68E-11

Table 3-4: Genes removed due to strong enrichment in either the ICGC OAC cohort (n=400) or the TCGA OAC cohort (n=70), compared to each other

Finally, to overcome the fact that most genes were only very rarely damaged across the cohort, I aggregated the RDGVs into 2,737 pathways taken from Reactome[107], KEGG[108] and BioCarta[109]. Having already removed genes that were enriched for RDGVs in any of the cohorts, there was no sign of pathway-level enrichment of RDGVs in OAC compared to EUR (Figure 3-1G). Some pathways were clearly depleted for RDGVs in OAC. However, this indicated that some genuine RDGVs in the OAC cohort had been missed, which was unlikely to lead to spurious results in downstream analysis.

One of the difficulties of the analysis of GSITs in Chapter 2 was the substantial overlap of pathways with RDGVs. Therefore, I developed a method to identify groups of pathways with substantial functionally-relevant overlap, which I called the variance-weighted Jaccard index (VWJI, Methods 3.2.4). The VWJI measured to what extent a pair of pathways shared genes, giving higher weight to genes that had more RDGVs across the cohort. Many pathways (497) had a high VWJI (>0.8) relative to at least one other pathway, and 203 of these had exactly the same damaged genes as another pathway (Figure 3-1H). In order to reduce the burden of pathway overlap, I identified sets of pathways that had a high VWJI relative to each other, and chose a single representative pathway for each of these sets (Methods 3.2.4). This gave a total of 2,240 distinct pathways with at least one RDGV, to be used in downstream analysis of GSITs.

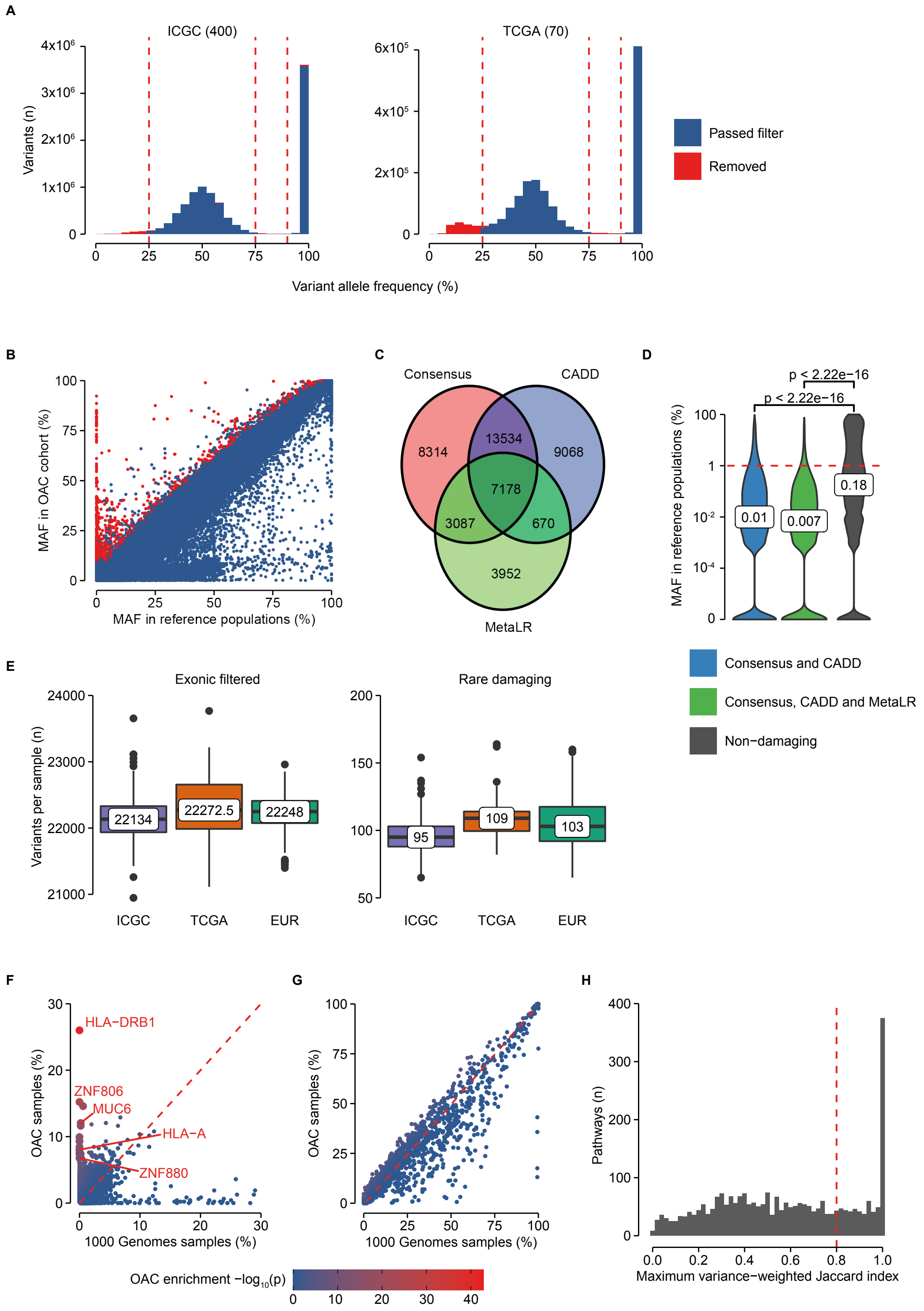

Figure 3-1: Preparation of germline data for 470 OAC samples

A. Histograms of variant allele frequencies (VAFs) of exonic and splicing germline variant calls in ICGC samples (left) and TCGA samples (right). Dotted red lines indicate VAF thresholds used to filter out variants.

B. Minor allele frequencies (MAFs) of exonic and splicing germline variants in the combined ICGC and TCGA OAC cohort (y-axis) and reference populations (x-axis).

C. Venn diagram of unique missense variants identified as deleterious by the consensus method, CADD and MetaLR.

D. MAF distributions of missense variants in reference populations. P-values from Wilcoxon's exact test. Median values are shown.

E. Numbers of filtered exonic and splicing germline variants (left) and RDGVs (right) per sample. Median values are shown.

F. Percentage of samples in which each gene has at least one RDGV in the OAC cohort (y-axis) and the EUR cohort (x-axis). Colour indicates level of enrichment in the OAC cohort, calculated using Fisher's exact test. A subset of the 29 genes that were removed from downstream analysis are indicated.

G. Percentage of samples in which each pathway has at least one RDGV in the OAC cohort (y-axis) and the EUR cohort (x-axis). Colour is on the same scale as panel F.

H. Distribution of the maximum value of the variance-weighted Jaccard index for each pathway with a different pathway.

3.3.3. Germline covariates

In addition to identifying RDGVs, I used the germline data for the OAC cohort to derive two potentially confounding covariates for downstream analysis of GSITs. These included principal components of common SNPs, and a total count of RDGVs in known cancer predisposition genes.

While I was using rare variants to characterise functional germline variation, common variants also encode information that could potentially affect analyses of GSITs, including batch effects and genetic ancestry[143]. I therefore used principal components (PC) analysis to reduce the dimensionality of common SNPs in the OAC cohort (Methods 3.2.5). The first PC clearly segregated the ICGC and TCGA cohorts from each other, suggesting a strong batch effect (Figure 3-2A). Since there may have been other batch effects in the data (for example in the calling of somatic alterations), PC1 served as a useful control for all cohort-related batch effects. PCs four, five and six appeared in part to capture the ethnicity of patients, judging by the reported ethnicities that were available in demographic annotations of the dataset (Figures 3-2B, C). In particular, Asian patients were outliers from the majority white cohort in terms of these PCs. By including PCs 4-6, I could thus account for genetic ancestry for all patients. It is also possible that, even among white patients, controlling for these PCs would account for subtle population structure. This was important since it has been shown that genetic ancestry can affect the frequency of somatic driver alterations in certain cancer genes[70]. Further PCs seemed to be uninformative, so I did not retain them as germline covariates (Figure 3-2D).

In the analysis of GSITs, I also wanted to account for possible effects coming from damaged cancer predisposition genes (CPGs). While CPGs are good candidates for genes that might influence somatic evolution, as a set they are not unified by a single biological process. In addition, recent work has shown that the presence of deleterious germline variants across all CPGs influences both the age at diagnosis and somatic mutational burden of patients at the pan-cancer level[144]. Thus, I wanted to account for the combined effect of CPGs. In order to do this, I derived the total allele count of RDGVs in a list of 151 pan-cancer CPGs from Huang *et al.*[49] in each OAC sample. There were 332 samples with at least one RDGV in a CPG, and there was a median of one damaged CPG per sample (Figure 3-2E). The most frequently damaged CPGs were *PMS1* (7%), *BRCA1* (6%) and *WRN* (5%). These genes are involved in different aspects of the maintenance of DNA integrity, carrying out mismatch repair[145], double-strand break repair[146] and helicase activities[147] respectively. This illustrates their potential to affect somatic alterations in cancer as a group.

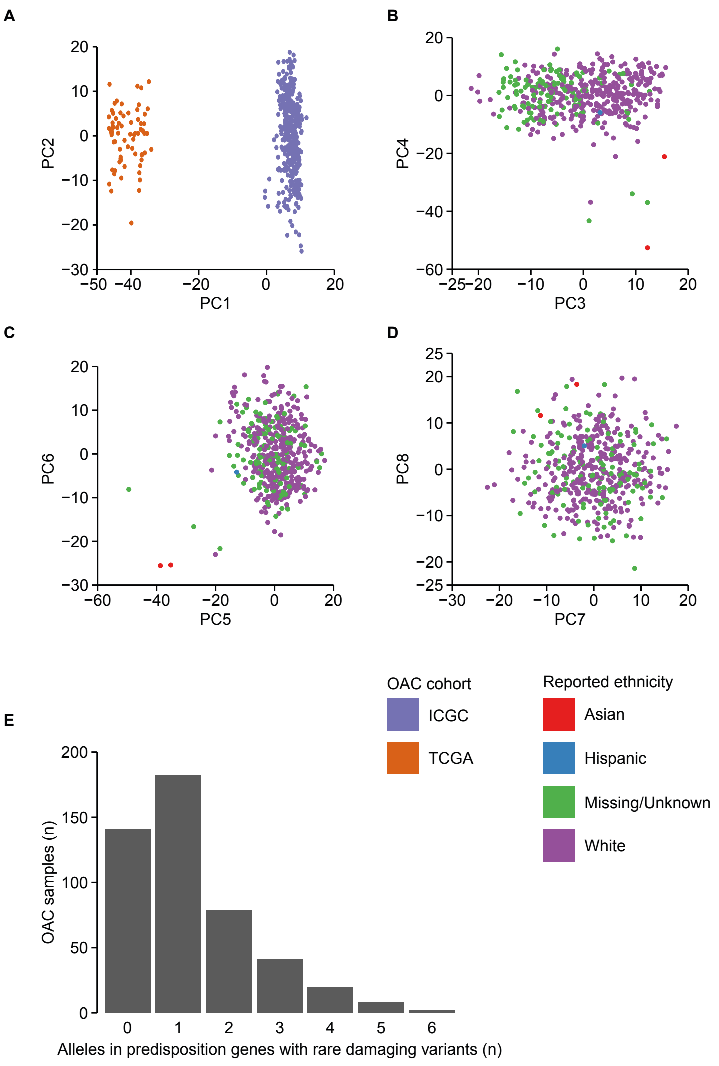

Figure 3-2: Germline covariates in the OAC cohort

Principal components analysis of samples using common SNPs, coloured by OAC cohort (**A**) or reported patient ethnicity (**B**-**D**).

E. Distribution of total allele count of RDGVs in 151 CPGs across samples.

3.3.4. Identifying OAC-specific driver alterations

As in Chapter 2, I chose to characterise the outcome of somatic evolution in cancer using driver alterations. However, in this case I tailored the driver identification approach to OAC, starting from a list of 76 OAC-specific driver genes[96]. By mapping genes to types of somatic alterations that were recurrent in OAC according to a combination of statistical methods (*e.g.* amplifications in recurrently amplified genes, Methods 3.2.6), I identified driver events in 40 genes. As expected, the most common driver by far was *TP53*, altered in 73% of samples. Other common driver genes included *CDKN2A* (29%), *ERBB2* (19%), *MYC* (19%) and *KRAS* (17%) (Figure 3-3A), reflecting previous observations in OAC[73,96,148]. There was a median of three driver events per sample, with 454 samples (97%) having at least one driver event (Figure 3-3B).

In Chapter 2, I aggregated somatic driver alterations into pathways. However, in the subsequent analysis of GSITs, the contributions to most pathways seemed to be coming from a small number of highly recurrent driver genes. Therefore, I did not aggregate the 40 OAC driver genes into pathways for downstream analysis. This also had the advantage of removing the difficulties associated with overlapping pathways. Nonetheless, it was still possible that significant inter-dependence between driver genes could affect downstream analysis. To assess this possibility, I measured patterns of co-occurrence and mutual exclusivity between each of the 780 pairs of driver genes with Fisher's exact test. Only four pairs of driver genes exhibited significant correlation (FDR <0.01, Figure 3-3C), suggesting that inter-dependence between drivers was not a widespread issue. Moreover, on inspection of these significant pairs, three of them represented co-occurrence between three genes (*ACVR2A*, *RNF43* and *RPL22*) that had acquired driver alterations in 10, 20 and seven samples respectively. The small numbers of samples affected by these genes (out of a total of 470 OACs) suggested that this pattern of co-occurrence would not be a substantial confounding factor. The other significant gene pair was *MDM2* and *TP53*, which tended towards mutual exclusivity in line with previous reports and *MDM2*'s role as an inhibitor of *TP53*[96]. However, since this was only a single gene pair, and *MDM2* was only altered in 28 samples, I did not explicitly account for this effect in downstream analysis.

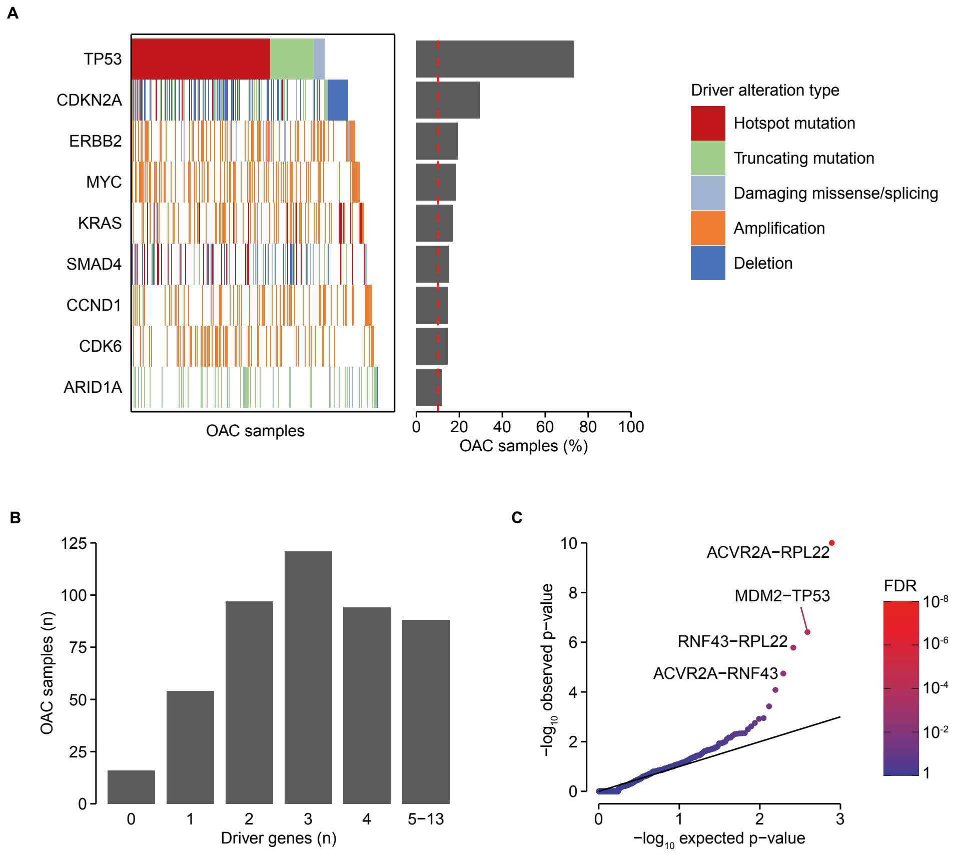

Figure 3-3: Somatic driver alterations in 470 OACs

A. Waterfall plot of somatic driver alterations across the OAC cohort, coloured by the type of somatic alteration. The bar plot shows the percentage of samples in which each gene has a driver alteration. The nine genes with driver alterations in ≥10% of samples are shown.

B. Distribution of the number of driver genes altered per sample.

C. Quantile-quantile plot showing p-values from Fisher's exact tests of 780 pairs of driver genes. The four gene pairs with FDR <0.01 are indicated.

3.3.5. Derivation of somatic covariates

As the final step in preparing data for analysis of GSITs, I extracted somatic covariates. These were quantities that either could confound later analyses, or that could be used along with driver alterations to characterise the somatic landscape of OAC in more detail. I used total mutational burden (TMB), ploidy and mutational signatures as somatic covariates.

Damaging germline variants in certain genes (such as mismatch repair genes) are known to increase somatic TMB in some cancer types[149,150]. Thus, including TMB as a covariate would help to ensure that any germline-somatic interactions found by subsequent analysis were truly due to the germline affecting selective pressures in cancer, rather than simply causing an excess of somatic mutations. I extracted TMB as the total number of exonic and splicing somatic SNVs and indels in each sample. TMB approximately followed a log-normal distribution across the OAC cohort, with a median of 159 mutations per sample (Figure 3-4A).

Ploidy served as an indicator of possible whole-genome doubling events, as well as a high-level descriptor of the extent and type of copy number variations in each sample, and could thus help to characterise the somatic landscape of the OAC cohort. Unsurprisingly given the prevalence of gene amplifications in OAC[73,96], many samples were polyploid, with 66% of samples having a ploidy of 2.5 or greater (Figure 3-4B).

Mutational signatures allowed me to further characterise the somatic landscape by the type of active mutational process in each sample. It has been shown that mutational signatures (particularly COSMIC signatures 2, 3, 17 and 18) are biologically and clinically relevant in OAC[73]. Moreover, germline variation influences a wide range of somatic mutational signatures across cancer types[16]. For a more parsimonious approach than using all 30 COSMIC signatures[45] in each sample, I extracted *de novo* mutational signatures (Methods 3.2.7). To assess the appropriate number of signatures to use, I calculated the residual sum of squares (RSS) when reconstructing mutations in the cohort with different numbers of signatures. Based on the presence of an inflexion point in the resulting RSS graph, I determined that four signatures would be appropriate for this cohort (Figure 3-5C). These four *de novo* signatures contributed roughly equally to mutations across the cohort (Figure 3-5D). To help interpret the biological significance of the mutational signatures, I compared each of them to the 30 COSMIC signatures using cosine similarity (Methods 3.2.7). The *de novo* Signature C was clearly analogous to COSMIC Signature 17, which is the hallmark signature of OAC and is believed to represent mutations arising from exposure of the oesophagus to stomach acid during acid reflux[73] (Figure 3-5E). The other three *de novo* signatures were more diffuse, but showed the strongest similarity to COSMIC Signatures 1 (aging, *de novo* Signatures B and D), 5 (unknown aetiology, A) and 6 (mismatch repair, D) (Figure 3-5E). COSMIC Signatures 2 (APOBEC) and 3 (*BRCA*-related) have previously been reported to be prevalent in OAC[73], but were not clearly represented by the *de novo* signatures. Nevertheless, these signatures provided an additional characterisation of the somatic landscape of the OAC cohort.

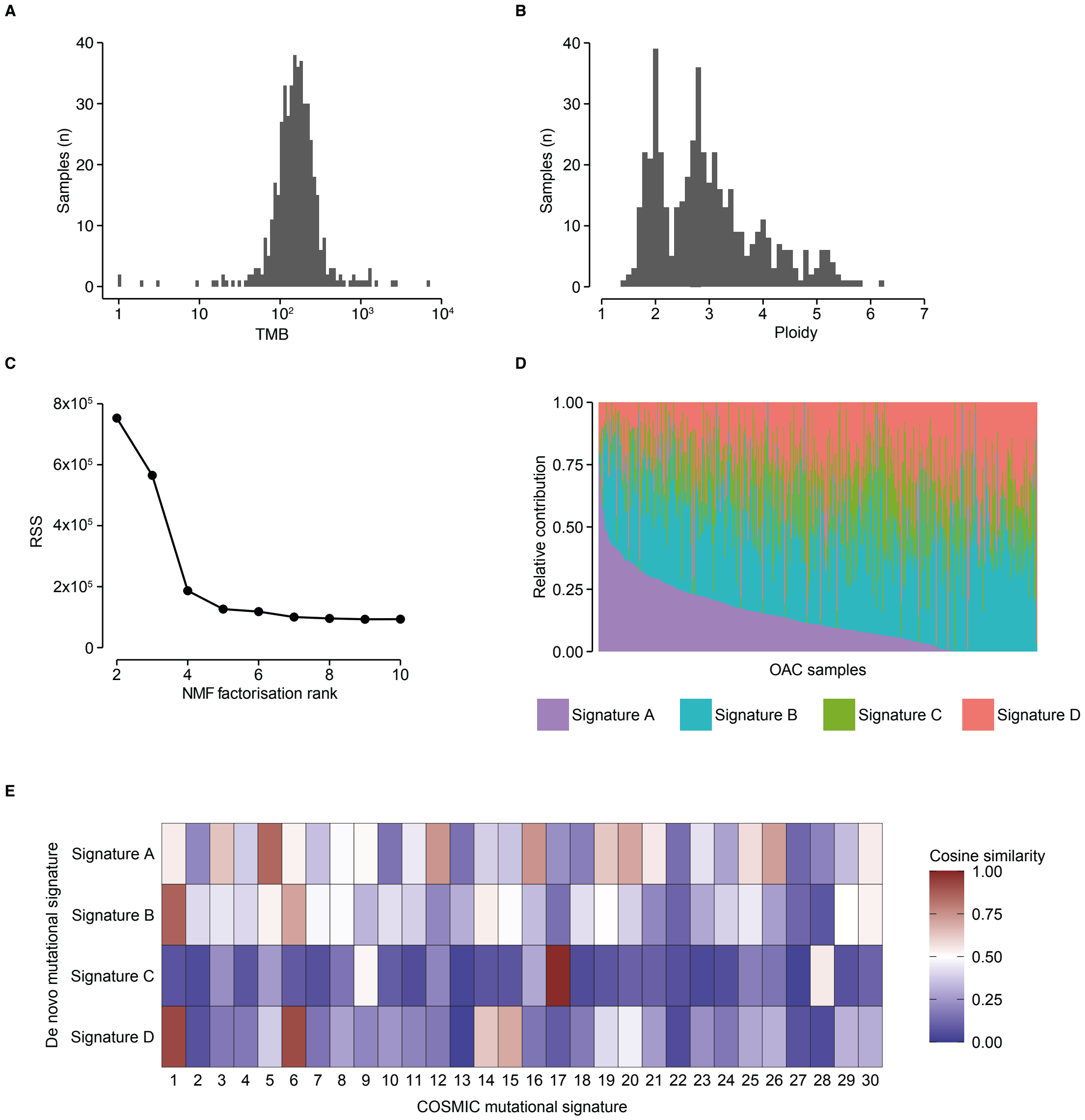

Figure 3-4: Somatic covariates in the OAC cohort

A. Distribution of total somatic mutational burden (TMB) across samples, calculated from exonic and splicing SNVs and indels.

B. Distribution of ploidy across samples.

C. Residual sum of squares (RSS) of extracted de novo mutational signatures for different values of the non-negative matrix factorisation (NMF) rank.

D. Relative contribution of four de novo mutational signatures to mutations in different samples.

E. Cosine similarity of *de novo* signatures to 30 mutational signatures from COSMIC.

3.4. Multivariate modelling of germline-somatic interactions *in trans*

Having prepared the germline, somatic and clinical data for a cohort of 470 OACs in Section 3.3, I incorporated the data into a statistical model to investigate the influence of germline perturbations on somatic evolution. I chose logistic regression as a model (discussed in Section 3.4.1), and then went on to consider how best to manage the available statistical power (3.4.2). Finally, I carried out the modelling analysis, finding statistically significant results involving driver alterations in *TP53* and *SMAD4* (3.4.3).

3.4.1. Logistic regression as a model for GSITs

The statistical modelling problem at hand involved three main components: RDGVs assembled into genes and pathways; somatic driver alterations in known OAC driver genes; and potentially confounding covariates. Given these, there were a number of suitable modelling options, each with their own strengths and weaknesses. I wanted any model to include a sense of causality, in which RDGVs and other covariates could influence which driver alterations became fixed in a tumour. Since regression models are well-studied and directional by design, and driver alterations are binary outcomes, I opted for logistic regression. I considered two possibilities: standard logistic regression, and overlapping group lasso logistic regression.

Standard logistic regression provided a simple, robust and well-established modelling approach while at the same time allowing for a full investigation of the effects of germline perturbations to pathways on somatic evolution. Because standard logistic regression lacks any form of parameter regularisation, I could not simultaneously model the effects of all germline pathways (n=2,240) given the cohort size (n=470). Instead, I chose to generate a single model for each germline pathway p and each somatic driver gene g. The resulting models took the form

$$\mathrm{Prob}(\text{driver alteration of gene } g \text{ in sample } i) = \sigma\left(\beta_{gp}x_{pi} + \sum_{k}\gamma_{gk}Z_{ki}\right),$$

where σ is the logistic sigmoid function, x_{pi} is the RDGV allele count in pathway p in sample i, Z_{ki} is the value of the k^{th} covariate in sample i, and β_{gp} and γ_{gk} are regression coefficients. By first fitting a "null model" for each driver gene g in which only covariates were used as predictor variables, I could calculate a p-value for the effect of p on g using a likelihood ratio test (LRT). Such a readily obtainable measure of statistical confidence was a major benefit of standard logistic regression.

As an alternative modelling strategy, I also considered overlapping group lasso logistic regression (OGLLR)[151,152]. Like other lasso methods, OGLLR uses an L_1 regularisation term to prevent model overfitting and select a small proportion of independent variables as having predictive value[153]. This allows for high-dimensional models to be fitted, where the number of independent variables can be many times more than the number of samples. The particular form of the OGLLR penalty is designed to capture group structures among the predictive variables. In the problem at hand, this would allow gene-level germline perturbations to be used as predictive variables, while still encoding the group structure of pathways into the model. For a particular outcome (somatic driver gene g), the probabilistic model would be very similar to that of standard logistic regression, *i.e.*

$$\text{Prob}(\text{driver alteration of gene } g \text{ in sample } i) = \sigma\left(\sum_h \beta_{gh} x_{hi} + \sum_k \gamma_{gk} Z_{ki}\right),$$

where x_{hi} is now the RDGV allele count in gene h. However, rather than applying vanilla maximum-likelihood estimation, the target function to be minimised to fit the OGLLR model would be

$$Q_\lambda(\beta, \gamma) = -l(\beta, \gamma) + \lambda \sum_p \sqrt{|p|}\, \|\boldsymbol{\beta}_p\|_1,$$

where $l(\beta, \gamma)$ is the log-likelihood of the standard logistic regression model, λ is a regularisation parameter, p indexes germline pathways, $|p|$ is the number of genes in p, and $\|\boldsymbol{\beta}_p\|_1 = \sum_{g \in p} |\beta_g|$ is the L_1-norm of the regression coefficients for all germline genes in pathway p. The effect of the penalty term (scaled by λ) is to select entire groups of predictive variables (in this case, pathways)[151,152]. Further 'elastic net-like' extensions to OGLLR have also been proposed which additionally apply an L_2-norm to shrink the non-zero parameters[154].

Such simultaneous modelling of genes and pathways was appealing, since some germline effects on somatic evolution could be mediated either at the gene or pathway level. However, there were two main practical drawbacks of the OGLLR approach for the problem at hand. First, as with all penalised regression methods, the regularisation parameter λ would need to be tuned to the data. This tuning process makes controlling false discovery rates challenging, and makes statistical inferences about the regression coefficients problematic[155,156]. Thus, while analysis with OGLLR may have selected individual genes and pathways, it would have been difficult to be confident in the validity of any models without a sizeable external validation cohort. Second, RDGVs were extremely sparse at the gene level compared to pathway-level data, with 3,774 genes (35%) damaged only in one sample (Figure 3-5). Due to the discrete nature of the data and relatively small cohort size, this would have encumbered predictive variables with very low signal to noise ratios.

On balance, I chose to proceed with standard logistic regression, because it allowed for robust statistical inference. The OGLLR might be considered to be an 'ideal' model to be used given the availability of much larger datasets. In what follows, I included all of the covariates described in Section 3.3, except for ploidy and mutational signatures, since these served as additional descriptors of the somatic landscape of samples rather than potentially confounding covariates.

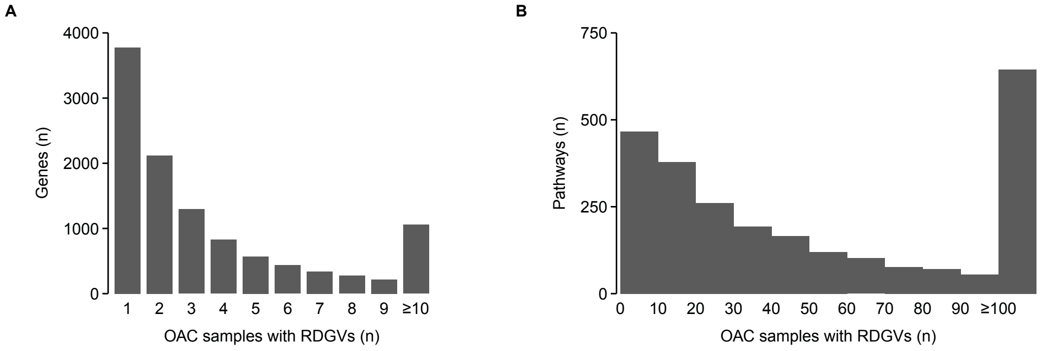

Figure 3-5: Sparsity of gene- and pathway-level germline data in 470 OACs
Distributions of the number of samples with RDGVs in each gene (**A**) and pathway (**B**).

3.4.2. Managing the statistical power of likelihood ratio tests

As with the Fisher testing approach in Section 2.4, the logistic regression modelling approach involved testing pairs of germline pathways and somatic drivers against one another. Even after reducing the burden of overlapping pathways and using only driver genes specific to OAC, there were 2,240 germline pathways and 40 driver genes, giving a total of 89,600 hypothesis tests to be carried out. To address this substantial multiplicity of hypothesis testing, I adopted two approaches. First, similar to Section 2.4, I restricted tests to those genes and pathways that occurred frequently enough in the cohort to have sufficient statistical power. Second, I made use of a method for calculating FDRs that was particularly well-suited to the structure of the problem[157].

Consider the statistical power of an individual LRT to detect the effect of a germline pathway on a somatic driver gene. The power depends on: the distribution of RDGVs in the pathway; the frequency of the driver alterations; the effect size; the test size; and the cohort size. In this setting, the cohort size was fixed at 470. The test size depended on the number of tests being carried out, and could be conservatively estimated using the Bonferroni method to achieve a particular overall family-wide error rate (FWER). Since I was treating driver alterations as binary outcomes, their distributions could be parametrised solely by the probability of observing a driver alteration in that gene in a given sample, denoted by $\bar{y}$. Conversely, since I was measuring RDGVs in pathways as allele counts, it seemed reasonable to suppose that these variables would have Poisson distributions across the cohort. In support of this, the mean and variance for all pathways were very similar (Figure 3-6A), and individual pathways appeared to fit Poisson distributions very well (Figure 3-6B). Thus, the distribution of RDGVs in a pathway could be captured by a single rate parameter λ, which corresponded to the mean number of RDGV alleles per sample. The question thus became how to restrict $\bar{y}$ and λ to optimally reduce the number of tests and conserve statistical power.

The effect size (regression coefficient β) in logistic regression can be interpreted as the log-odds ratio, *i.e.* $\beta = \ln(\text{odds ratio})$[158]. Thus, effect sizes of 1 and 1.5 would correspond to the plausible odds ratios of 2.7 and 4.5, respectively. Since I was interested in both positive and negative effects (where an RDGV would decrease the likelihood of a particular driver alteration), I used simulated data to estimate the power of LRT tests to find interactions with effect size ±1 and ±1.5, for different values of $\bar{y}$ and λ, with an overall FWER of q=0.05 (Methods 3.2.8, Figures 3-6C, D). The resulting power curves indicated that I was under-powered to detect effect sizes of ±1. They also suggested that driver genes altered in <10% of samples (*i.e.* $\bar{y} < 0.1$) would not be sufficiently powered to detect any associations. The direction of the effect was important, with more power to detect negative associations with drivers altered in many samples, and more power to detect positive associations with drivers altered in fewer samples (Figure 3-6D). This analysis suggested that for each driver gene g, a threshold λ_{min}^{g} could be calculated, and only pathways with $\lambda > \lambda_{min}^{g}$ tested for that gene, since only these pathways would confer sufficient power to the LRT. I calculated λ_{min}^{g} thresholds for each of the nine somatic driver genes with $\bar{y} \geq 0.1$ to achieve 80%

power (Methods 3.2.8), which left a total of 7,940 hypothesis tests to be carried out (median 885 per driver gene, Table 3-5).

The structure of the problem could now be thought of as involving nine independent families of hypothesis tests, *i.e.* one family per somatic driver gene. A reasonable approach for identifying GSITs might be to first select driver genes exhibiting some significant results, and then to identify the significant tests within these selected families. Controlling error rates in problems like this was previously explored by Benjamini and Bogomolov[157], who were motivated by the extremely high-dimensional case of voxel-wise genome wide association studies (vGWAS). In vGWAS, hundreds of thousands of SNPs are tested against tens of thousands of voxels from fMRI images to identify significant SNP-voxel associations[159]. The authors devised a method for controlling the average error rate across selected families of tests, valid for a wide variety of family selection procedures and error rate definitions. In the problem at hand, I could implement this method to control the average FDR at level q using the following procedure:

1. For each of the nine driver genes g, use logistic regression to calculate a p-value for the interaction of g with each germline pathway satisfying $\lambda > \lambda_{min}^{g}$
2. For each driver gene, assess the presence of one or more statistically significant germline pathways using Simes' test[160]
3. Perform FDR correction on the nine Simes p-values, and select the driver genes that have Simes FDR below a certain threshold
4. For each of the selected driver genes, calculate the FDR across germline pathways, but control the FDR at $q \times {}^{N_{selected}}\!/_{9}$.

By combining the approach of selecting only sufficiently frequent driver genes and germline pathways with the Benjamini-Bogomolov selective inference method, I could effectively manage the available statistical power in the problem at hand.

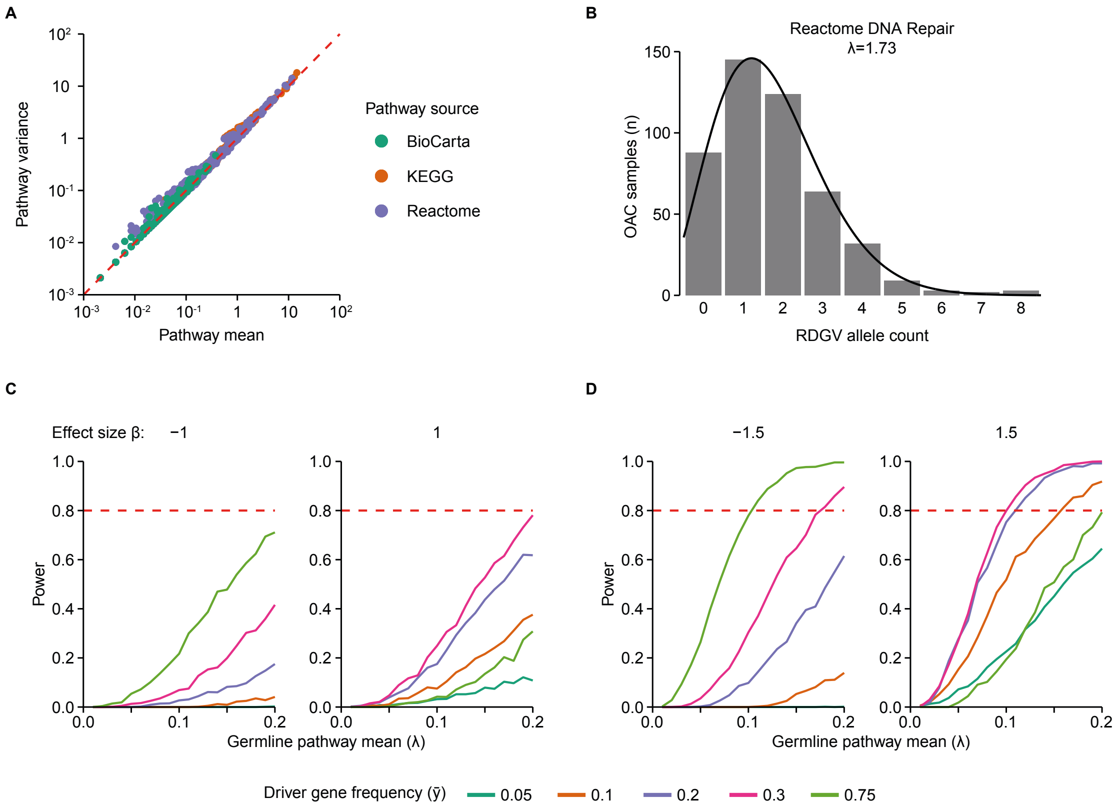

Figure 3-6: Managing statistical power in logistic regression modelling

A. Scatter plot of the means and variances of allele counts of RDGVs for 2,240 germline pathways.

B. Distribution of allele counts of RDGVs in Reactome's DNA Repair pathway. The mean allele count was 1.73, which was used to parametrise a Poisson distribution whose probability density is indicated by the black curve.

Power curves for different values of the driver gene frequency $\bar{y}$, the mean RDGV allele count of the germline pathway λ, for effect sizes $\beta = \pm 1$ (**C**) and $\beta = \pm 1.5$ (**D**). The dotted red line indicates 80% statistical power.

Driver gene	**Driver frequency $\bar{y}$**	**Effect size β**	**Threshold λ^{g}_{min}**	**Pathways to test**
TP53	0.734	-1.5	0.101	951
CDKN2A	0.294	1.5	0.098	965
ERBB2	0.191	1.5	0.113	885
MYC	0.185	1.5	0.110	912
KRAS	0.17	1.5	0.110	912

SMAD4	0.153	1.5	0.118	862
CCND1	0.149	1.5	0.123	843
CDK6	0.145	1.5	0.124	834
ARID1A	0.119	1.5	0.139	776

Table 3-5: Germline pathway frequency thresholds

Thresholds for the minimum value of the mean RDGV allele count of germline pathways λ_{min}^{g}, calculated for the nine somatic driver genes altered in >10% of OAC samples. Thresholds were derived assuming effect sizes of $\beta = \pm 1.5$, with positive β for genes altered in <50% of samples, and negative β for *TP53* (altered in >50% of samples). The number of pathways with $\lambda > \lambda_{min}^{g}$ is shown for each gene.

3.4.3. Results of likelihood ratio tests in nine driver genes

I carried out the 7,940 LRTs, associating nine somatic driver genes with between 776 and 965 germline pathways each (Table 3-5). Two of the driver genes (*TP53* and *SMAD4*) exhibited possible associations with one or more germline pathways (Simes FDR <0.15, Table 3-6). This was further confirmed by visualisation with a quantile-quantile plot, in which *TP53* and *SMAD4* showed the clearest deviation from the diagonal (Figure 3-7). I therefore selected these two genes and proceeded with the Benjamini-Bogomolov FDR corrections.

Among the 1,813 germline pathways tested for these two genes, three had FDR <0.2 (Table 3-7). There were two pathways in which RDGVs were positively associated with driver alterations in *SMAD4* (Regulation of insulin secretion, and Signalling by ROBO receptors). These pathways had a small overlap (three damaged genes out of a total of 113), so may or may not have been acting independently. Additionally, RDGVs in the *ATM* signalling pathway were negatively associated with driver alterations in *TP53*. I will explore these results in more detail in Section 3.5.

Driver gene	**Simes p-value**	**FDR**	**Selected**
SMAD4	0.0207	0.111	Yes
TP53	0.0246	0.111	Yes
CDKN2A	0.198	0.446	No

ARID1A	0.163	0.446	No
KRAS	0.525	0.675	No
CCND1	0.326	0.587	No
CDK6	0.657	0.687	No
MYC	0.459	0.675	No
ERBB2	0.687	0.687	No

Table 3-6: Driver gene test family selection

Selection of driver genes showing evidence of significant interactions with germline pathways. The FDR was calculated using the Benjamini-Hochberg method using the nine p-values from Simes' test shown.

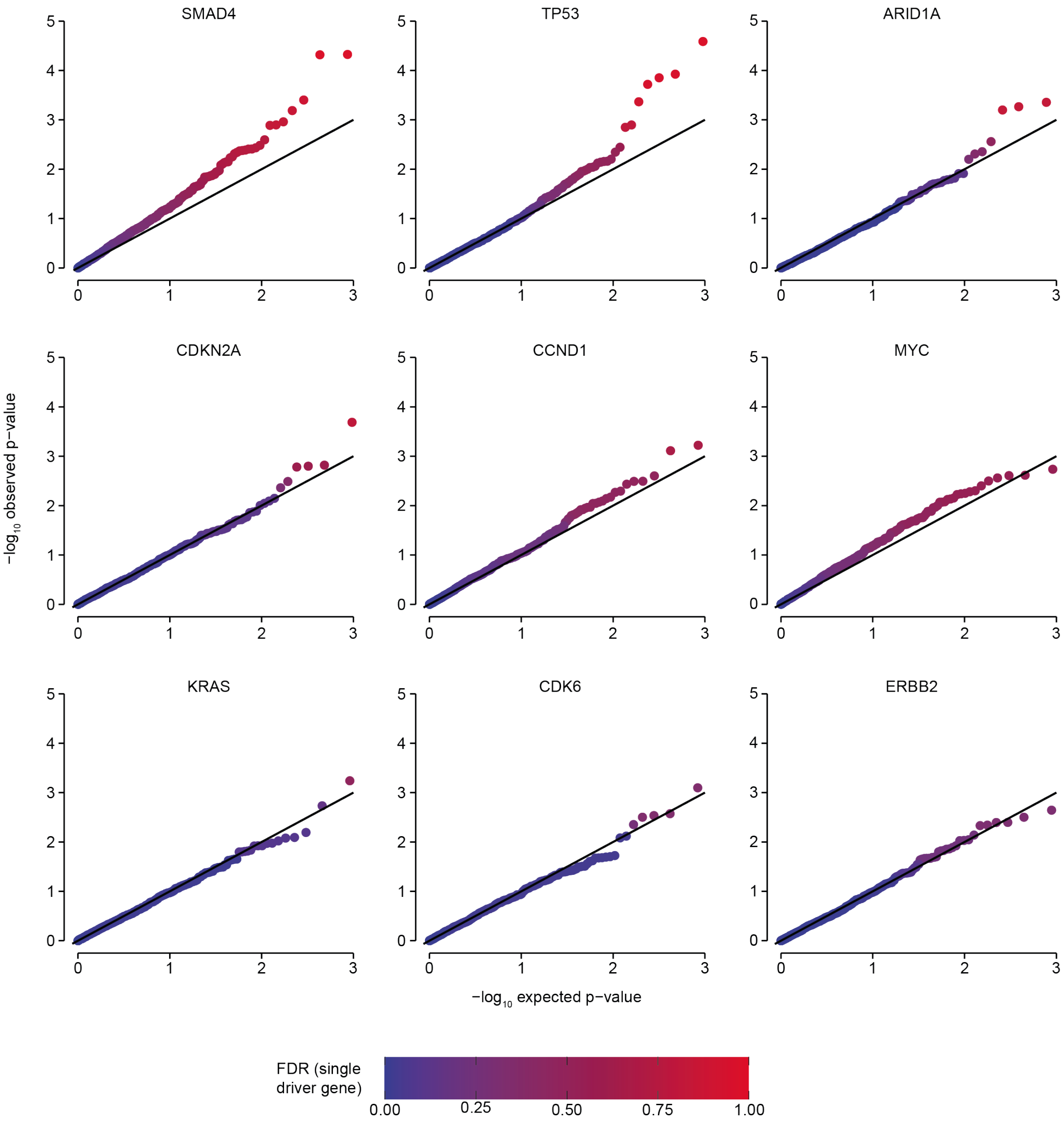

Figure 3-7: Logistic regression modelling results

Quantile-quantile plots of p-values from LRTs evaluating the association between somatic driver genes and RDGV allele counts in germline pathways. Each panel represents the hypothesis tests for a single driver gene (labelled above). The colour scale represents the FDR calculated for each driver gene separately.

Germline pathway	Driver gene	P-value	Estimated effect size	Benjamini-Bogomolov FDR

Regulation of insulin secretion	*SMAD4*	4.74E-05	0.807	0.093
Signalling by ROBO receptors	*SMAD4*	4.81E-05	0.757	0.093
ATM Signalling Pathway	*TP53*	2.59E-05	-1.25	0.11

Table 3-7: Significant results from logistic regression modelling

Significant associations between RDGVs in pathways and somatic driver alterations in cancer genes, with Benjamini-Bogomolov FDR <0.2.

3.5. Exploratory analysis of germline influence on *SMAD4* and *TP53* in OAC

In Section 3.4 I found three statistically significant associations between RDGVs in pathways and driver alterations in genes. In this section, I will explore these results in more detail, with a view to assessing both their robustness and possible implications. First, I will explore a positive association between two partially overlapping pathways (Regulation of insulin secretion and Signalling by ROBO receptors) with *SMAD4* somatic driver alterations (3.5.1). Second, I will describe how RDGVs in the *ATM* signalling pathway, and in particular in the *ATM* gene itself, reduce the likelihood of *TP53* driver alterations (3.5.2).

3.5.1. Associations with *SMAD4* are likely statistical artefacts

The logistic regression modelling approach also identified positive associations between RDGVs in two pathways (Regulation of insulin secretion and Signalling by ROBO receptors) and *SMAD4* driver alterations (Table 3-7). However, there were a number of reasons to suspect that these associations were not genuine, including the diffusivity of gene-level contributions in the germline, a lack of supporting results from orthogonal data sources, and the lack of a clear biological connection between these germline pathways and the biological function of *SMAD4*.

I first assessed the contributions of individual genes to the statistical results for both pathways. For each damaged gene in each pathway, I removed that gene from the pathway and re-calculated the significance of the association. If removing the gene led to a substantial increase of the p-value (*i.e.* a decrease of the significance of the result), I determined that this gene was contributing to the result. In both pathways, the distribution of gene-level contributions was diffuse, with a continuous gradient of gene contributions and many genes detracting from the result (Figures 3-8A, B). No single gene had a penetrant effect, and there were no clear unifying features of the genes that contributed to the result versus those that detracted from it. Finally, the vast majority (93 out of 112, 83%) of the genes in both pathways were damaged in fewer than five samples across the cohort. Thus, it was neither the case that a clear subset of genes was driving the results, nor that the pathway as a whole was contributing.

In order to investigate possible molecular or clinical mechanisms of the germline associations with *SMAD4*, I used orthogonal data sources. I stratified the cohort into four groups by the presence of germline RDGVs and *SMAD4* driver alterations, and looked for differences between these strata. First, I identified genes that were differentially expressed in samples with both germline RDGVs (in the insulin secretion or ROBO signalling pathways) as well as *SMAD4* driver alterations, compared to other samples. By performing GSEAs on these differentially expressed genes, I hoped to gain mechanistic insights into the association. However, no pathways were significantly enriched for genes differentially expressed at a range of different thresholds (>4x, >3x, <0.5x). Second, I searched for clinical differences between the strata in terms of annotations that might be relevant to the germline pathways, and in particular to insulin secretion. However, there were no significant differences in diabetes incidence, body mass index or gastro-oesophageal reflux symptom duration between the strata.

Finally, I searched the literature for possible connections between insulin secretion or ROBO receptor signalling and *SMAD4*. It is well-known that *SMAD4* is a critical component of TGF-beta signalling[161], and more recently evidence has suggested that insulin can sensitise cells to TGF-beta signalling[162]. Regulatory interactions between ROBO signalling genes and both TGF-beta and *SMAD4* activity have been reported[163], and this interaction seems to play a particularly important role in pancreatic cancer[164,165]. While these connections were promising, they were not reflected in the OAC data. For example, RDGVs in either pathway did not affect the expression of TGF-beta genes (Figures 3-8C, D). Additionally, some of the most important genes in ROBO signalling (*ROBO1/2/3* and *SLIT1/2/3*[165]) were on the whole not contributing to the association with *SMAD4* driver status.

Given the diffuse nature of gene-level contributions, the lack of mechanistic insights from orthogonal data sources and the discordance between potential explanations from the literature with gene expression data, I concluded that the germline associations with *SMAD4* driver alterations were either statistical artefacts, or were so weak as to require much larger sample sizes for insightful interrogation.

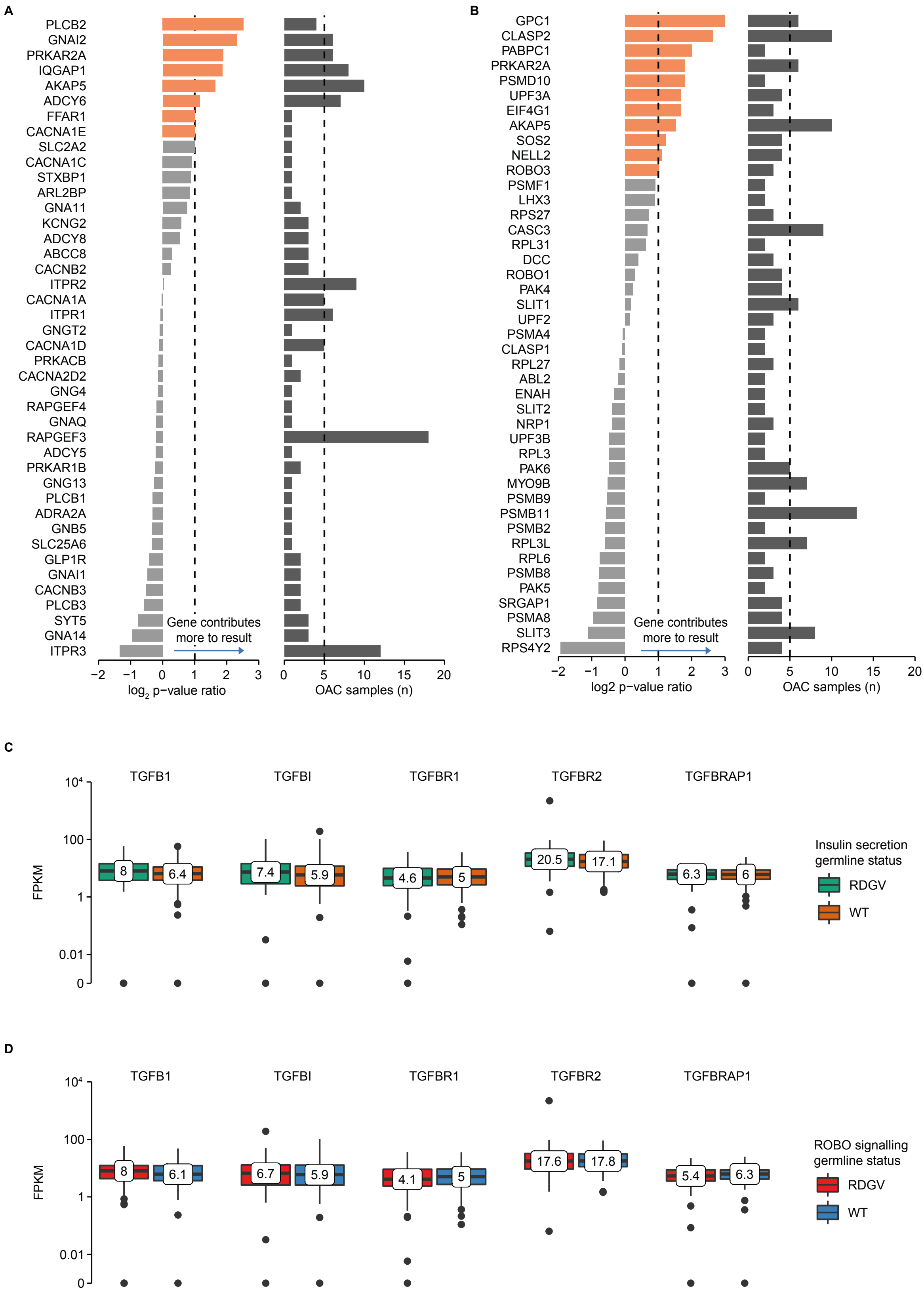

Figure 3-8: Interrogation of germline associations with *SMAD4* driver status

Importance and frequency of genes in the Regulation of insulin secretion (**A**) and Signalling by ROBO receptors (**B**) pathways. Gene importance is shown on the left, while the number of OAC samples with RDGVs in each gene is shown on the right. In (**B**), only genes with RDGVs in at least two samples are shown.

Expression levels of TGF-beta genes across samples, stratified by the presence of RDGVs versus wild type (WT) in the Regulation of insulin secretion (**C**) and Signalling by ROBO receptors (**D**) pathways. Median FPKM values are shown. Genes in the TGF-beta family with median expression >1 FPKM across the whole OAC cohort are shown.

3.5.2. Germline perturbations to *ATM* signalling influence the evolutionary trajectory of OAC

Logistic regression modelling revealed a significant negative association between RDGVs in the *ATM* signalling pathway and *TP53* driver alterations (Table 3-7), illustrated clearly in Figure 3-9A. Two covariates in the model were also found to be predictive of *TP53* driver status (Figure 3-9B). Female gender was negatively associated with *TP53*, with women less likely to have *TP53* driver alterations than men ($p=1.7x10^{-3}$) in line with previous studies[96]. In addition, there was a weak positive association between *TP53* driver alterations and PC3 derived from common SNPs (p=0.01). However, the *ATM* signalling pathway was the most significant predictor in the model ($p=2.6x10^{-5}$, Figure 3-9B) suggesting that it was not confounded by other factors.

In order to understand this association in greater detail, I investigated the contributions of individual genes within the *ATM* signalling pathway to the association with *TP53* as in Section 3.5.1. There were five genes in the *ATM* signalling pathway whose removal more than doubled the p-value (*i.e.* substantially weakened the association): *ATM*, *BRCA1*, *NBN*, *RAD50* and *RAD51* (Figure 3-9C). A total of 36 OAC patients had RDGVs in these ‘core’ *ATM* signalling genes, of whom only 13 (36%) had *TP53* driver alterations (compared with 73% across the whole OAC cohort). While each of these genes contributed to the negative association with *TP53*, it was clear upon visual inspection that germline protein truncating variants (PTVs) in *ATM* exhibited the strongest effect (Figure 3-9D). Of the five patients with *ATM* germline PTVs, none had a *TP53* driver alteration ($p=1.3x10^{-3}$, Fisher’s exact test). Moreover, three of these patients exhibited a second somatic hit to *ATM*, consistent with a tumour suppressive and possibly predisposing role for *ATM* in OAC. There was also a weak exclusivity between *TP53* driver alterations and *ATM* somatic mutations (p=0.013, OR=0.37, Fisher's exact test), indicating that the negative association was not purely mediated through the germline. This somatic exclusivity with *TP53* provided evidence that *ATM* should be considered to be a driver gene in OAC, despite the fact that previous studies have not identified it as such[73,96-100].

There were several reasons to believe that this observed negative association was genuine. *ATM* and its interactors are established driver genes in many cancer types[25,166], and germline *ATM* variants have reported predisposition roles in breast cancer[167], lymphoma and acute leukaemia[168]. The five core *ATM* signalling genes found to be contributing most to the negative association with *TP53* drivers in OAC are all involved in the sensing and repair of DNA double strand breaks[169] (DSBs). In addition, *ATM* interacts directly with *TP53*, phosphorylating it to activate an apoptosis transcriptional programme in response to failed DSB repair[169,170]. It is therefore plausible that damage to *ATM* (and to a lesser extent, its interactors) can act as a substitute for *TP53* driver alterations, permitting DSBs to go unrepaired and lead to further DNA damage, without inducing apoptosis. This pattern can also be observed in other cancer types. An analysis of other cancer types from TCGA revealed a negative association between germline RDGVs in the five core *ATM* signalling genes and *TP53* driver alterations in stomach adenocarcinoma (Figure 3-9E, p=0.013, LRT). Notably, stomach adenocarcinoma is molecularly similar to OAC[171], hinting at a common mechanism between these cancer types. Moreover, mutual exclusivity between *ATM* and *TP53* mutations has previously been reported in breast cancer[172], lung adenocarcinoma[55], T-cell leukaemia[173] and B-cell lymphoma[174]. Together, these observations provided strong evidence that the observed negative association in OAC was a genuine biological effect, rather than a statistical artefact.

An obvious avenue for investigating possible mechanisms of this association would have been to look at transcriptional differences between samples. Unfortunately however, the available RNA-Seq data for the OAC cohort proved inadequate for this analysis. First, only one sample with an *ATM* germline truncation had matched RNA-Seq data. Second, the nine samples with both *ATM* RDGVs and RNA-Seq data had substantially fewer total read counts across all human genes than other samples

(Figure 3-9F). These limitations made any transcriptional comparisons between samples unreliable.

I was particularly interested in the possibility that *ATM* variants could predispose to OAC, given the current lack of highly penetrant predisposition genes for OAC. As noted above, there was a perfect mutual exclusivity between *ATM* germline PTVs and somatic *TP53* driver alterations in this data. Moreover, the somatic second hits to *ATM* were consistent with the two-hit hypothesis for tumour suppressive cancer predisposition genes[52]. These observations suggested that *ATM* germline PTVs might predispose to OAC, and that *ATM* mutations could substitute for *TP53* drivers in these patients. If *ATM* could indeed predispose to OAC, one might expect that (i) patients with *ATM* germline PTVs would be diagnosed with OAC earlier than other patients, and (ii) that *ATM* germline PTVs would be enriched in OAC patients compared to reference healthy populations. Both of these predictions were borne out as weakly statistically significant. Patients with *ATM* germline truncations were diagnosed with OAC at a median age of 61, compared to the cohort-wide median of 67 (p=0.074, one-sided Wilcoxon rank-sum test, Figure 3-9G). Moreover, *ATM* germline PTVs were present in 1.06% of OAC patients, compared to 0.2% of Europeans from the 1000 Genomes Project (p=0.094, one-sided Fisher's exact test, Figure 3-9H). The borderline significance of these results may well be due to small sample sizes, as only five OAC patients had a germline *ATM* truncation.

In summary, the negative association between germline RDGVs in the *ATM* signalling pathway and *TP53* driver alterations was largely driven by five genes involved in the sensing and repair of DNA DSBs. Exclusivity with *TP53* was strongest for truncations of the *ATM* gene, consistent with reports in other cancer types. This exclusivity with OAC's most frequent driver gene, *TP53*, suggests that *ATM* plays an important role in the cancers of these patients. Moreover, the prevalence of second hits to *ATM*, as well as trends of younger age at diagnosis and enrichment compared to reference populations, strongly suggest that *ATM* truncations can predispose to the development of OAC and that *ATM* can act as a tumour suppressor gene in OAC.

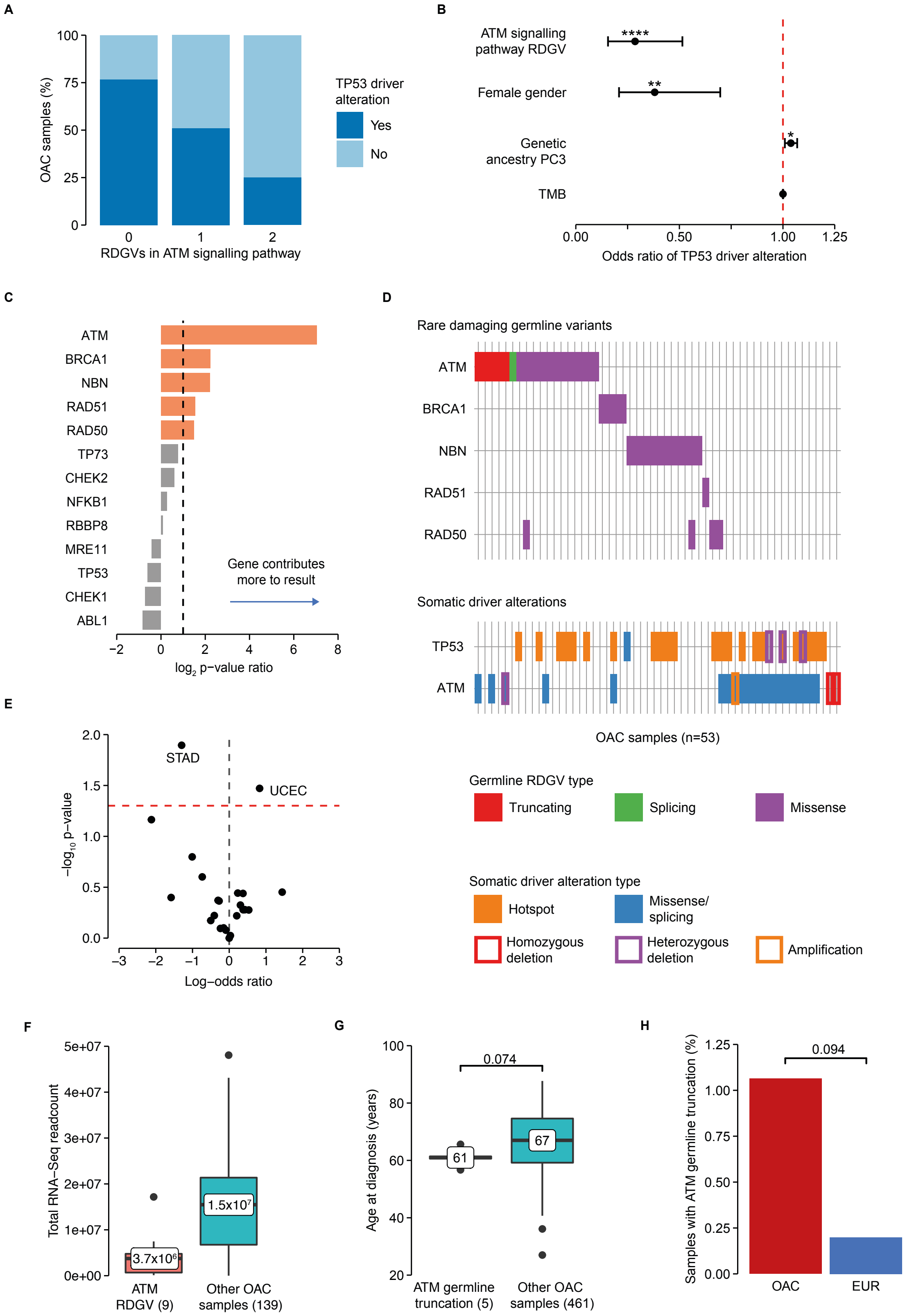

Figure 3-9: Negative association between germline RDGVs in the *ATM* signalling pathway and *TP53* driver alterations in OAC

A. Distribution of RDGVs in the *ATM* signalling pathway and the associated *TP53* driver status.

B. Odds ratios of statistically significant (p<0.05) predictive variables (as well as TMB) for *TP53* driver status from logistic regression modelling. Horizontal bars indicate 95% confidence intervals. *=p<0.05; **=p<0.01; ****=p<0.0001, p-values from likelihood ratio tests.

C. Contribution of individual *ATM* signalling genes to the negative association. The ratio between the p-value for the full pathway and the p-value for the pathway with each individual gene removed was calculated. Genes with higher p-value ratios contribute more to the association. The five genes with p-value ratio >2 are highlighted in orange.

D. Visualisation of co-occurrence and mutual exclusivity of germline RDGVs in the *ATM* signalling pathway with somatic driver alterations in *TP53* and *ATM*. The 53 samples with at least one RDGV in the *ATM* signalling pathway or a somatic *ATM* alteration are ordered on the x-axis.

E. Volcano plot showing the results of TCGA pan-cancer logistic regression modelling, measuring the association between RDGVs in the five core *ATM* signalling genes with *TP53* driver alterations. The dotted red line indicates p=0.05.

F. Distribution of total RNA-Seq read counts across all genes for OAC samples with and without *ATM* germline RDGVs, for the 148 samples with available transcriptomic data.

G. Age at diagnosis for OAC patients with and without *ATM* germline truncations. P-value from one-sided Wilcoxon rank-sum test.

H. Proportion of individuals with *ATM* germline truncations in the OAC cohort and 503 Europeans from the 1000 Genomes Project (EUR). P-value from one-sided Fisher's exact test.

3.6. Mixed data type modelling for germline-somatic associations

Having used logistic regression modelling to identify associations between RDGVs in pathways and driver alterations in genes, I next sought to use a more comprehensive characterisation of somatic variation in the OAC cohort to identify GSITs. In this section I will describe the application of a random forest (RF) clustering approach. I will first discuss the need for mixed data type analysis of somatic variation and the suitability of RF clustering for this task (3.6.1). I will then apply RF clustering to the OAC cohort, and describe the resulting partitions of OAC samples (3.6.2). Finally, I will apply logistic regression modelling to the results of the RF clustering in order to search for possible complex GSITs (3.6.3).

3.6.1. Random forests allow for mixed data type analyses

In Sections 3.4 and 3.5, logistic regression modelling had proven to be capable of identifying associations between RDGVs in pathways and driver alterations in individual genes. While the inclusion of both clinical and germline covariates in the model was useful, the characterisation of somatic variation with individual driver genes was not comprehensive. First, restricting the analysis to one driver gene at a time meant that complex patterns involving groups of driver genes could not be identified. However, in Chapter 2 it was clear that using the presence of driver alterations at the pathway level was not an optimal approach either, with most perturbed pathways being dominated by very few highly recurrent driver genes. Second, somatic covariates such as ploidy and mutational signatures, which also reflect the effects of cancer evolution, could not be simultaneously modelled with driver alterations as outcomes by logistic regression. I therefore sought a method for characterising samples based on both multiple driver genes and somatic covariates.

A clustering approach would enable simple logistic regression modelling to capture complex outcomes via cluster assignments, since clusters could be based upon arbitrarily complex patterns. One challenge of any such approach would be to incorporate both somatic driver data and somatic covariates, since these were naturally binary and continuous data types, respectively. Clustering mixed data types is currently an open area of research[175].

Random forests[176] (RFs) are most commonly used for classification problems, where they have been successfully used to handle mixed data types[177,178]. However, RFs have also been adapted to perform clustering[179]. Given a dataset with N samples and p features, clustering with RFs works in a three-stage process:

1. Simulated data are generated with N samples and the same p features by sampling each feature empirically and independently from the real dataset. Therefore, while each feature has the same marginal distribution in the real and simulated data, any inter-feature correlations are lost in the simulated data.
2. A standard RF is trained to classify the real and simulated data points. In order to do this, it must learn the correlations that are present between features in the real dataset. This process allows for a notion of distance between real samples to be defined, based on the number of trees for which samples end up in the same leaf nodes during classification.
3. Any clustering method can be applied using the distance metric calculated from the RF training and prediction process.

I decided to apply RF clustering to the OAC cohort of 470 samples to characterise the somatic landscape using multiple driver genes and somatic covariates.

3.6.2. Clustering OAC samples with mixed data type descriptors of somatic landscape

In addition to the RF itself, I needed to choose a clustering algorithm and an appropriate number of clusters. I decided to use partitioning around medoids (PAM) as a more robust alternative to standard k-means clustering[180]. In order to choose the optimal number of clusters, I used silhouette analysis, which is a standard method for assessing the robustness of clusters[181]. A high silhouette width indicates that samples are closer to other samples in the same cluster than they are to other clusters, and therefore that the clustering is robust. In addition to this metric of robustness, I also used the Adjusted Rand Index[182] (ARI). The ARI measures how similar two sets of cluster assignments are. It can thus be used to assess the reproducibility of stochastic clustering methods such as the RF.

I implemented ten RFs to calculate pairwise distances between samples based on driver alterations in cancer gene-enriched pathways and somatic covariates (Methods 3.2.9). I then clustered samples using PAM, and measured the silhouette width and ARI for a numbers of clusters ranging from two to 15. The silhouette analysis had inflexion points at three and nine clusters (Figure 3-10A), although three clusters would likely have been too few to be useful. The ARI had inflexion points at five and nine clusters (Figure 3-10B). I thus concluded that the most appropriate number of clusters to use was nine.

The resulting clusters had distinct combinations of driver genes and somatic features, although they were primarily driven by the most frequent driver genes (Figure 3-10C). For example, cluster 3 was comprised primarily of *TP53*-wild type samples, and had correspondingly lower burdens of somatic SNVs and indels, as well as lower ploidy, than other clusters (Figures 3-10D, E). Cluster 1 instead had *TP53* driver alterations but lacked other common drivers, while cluster 2 was mainly characterised by the combination of *TP53* and *MYC* drivers in the absence of *KRAS* alterations (Figure 3-10C). The remaining clusters were similarly characterised by different combinations of common driver genes. There was in general little difference in the relative contributions of mutational signatures to the nine clusters (Figure 3-10F).

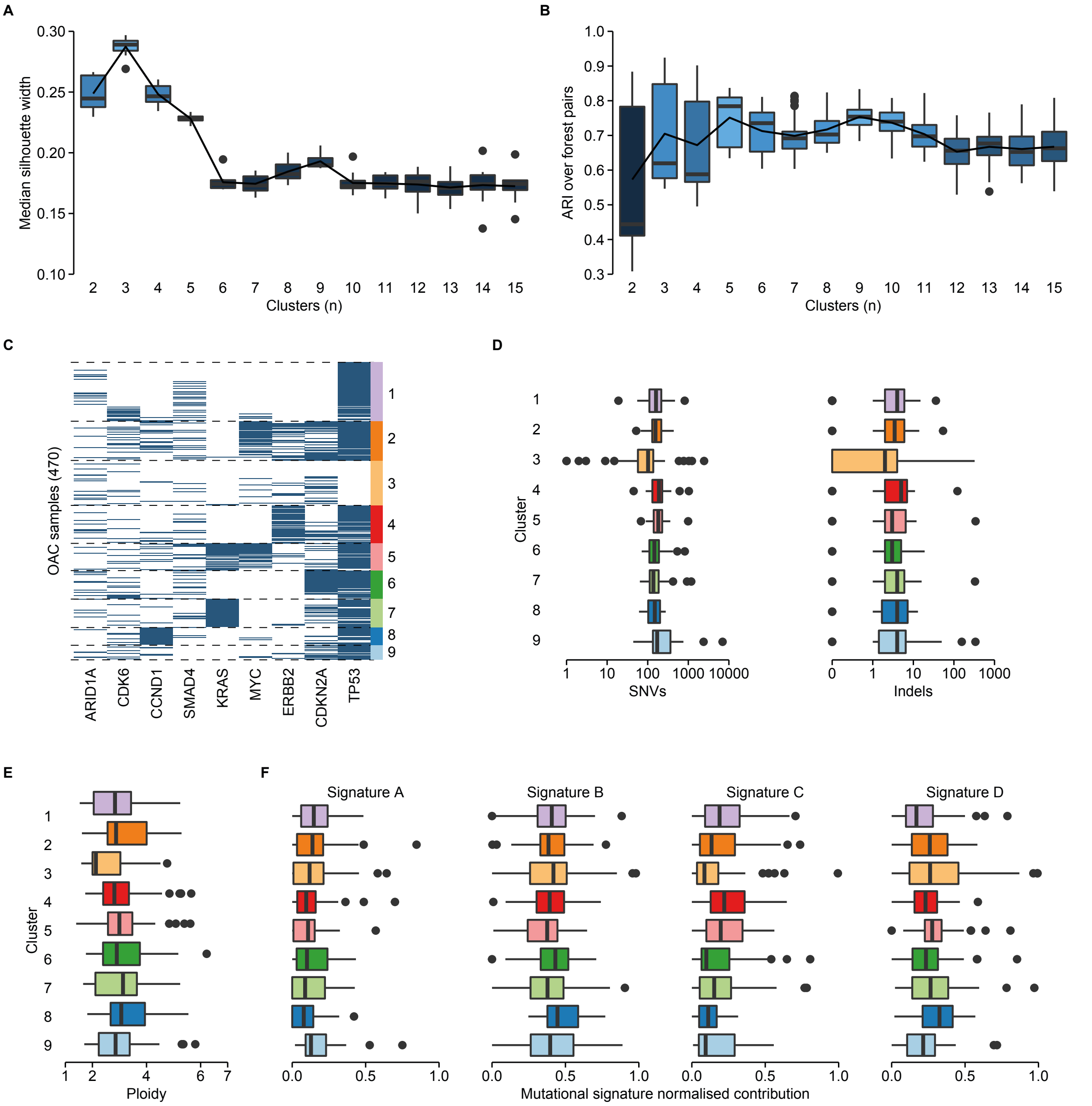

Figure 3-10: Random forest clustering of 470 OACs

A. Median silhouette width of clusters, for different numbers of PAM clusters. Boxplots show distributions over ten RF implementations.

B. Adjusted Rand Index (ARI) for different numbers of PAM clusters. The ARI was calculated between 45 distinct pairs of ten RF implementations. Boxplots show distributions of the ARI over RF pairs.

C. Incidence of common OAC driver genes in nine clusters. Dotted lines separate clusters of samples, which are labelled on the right-hand side. The nine driver genes altered in >10% of the OAC cohort are shown.

D. Exonic mutational burden of samples in each cluster, broken down into SNVs (left) and indels (right).

E. Ploidy distributions across clusters.

F. Contributions of four *de novo* mutational signatures (see Section 3.3) to samples in each cluster. The contribution of each signature was normalised to range between 0 and 1 across the entire cohort for each signature.

3.6.3. Germline associations with somatic clusters

Having partitioned the OAC samples into nine clusters, I used logistic regression modelling to associate RDGVs in pathways with individual clusters. Analogously to Section 3.4, for each cluster I restricted analysis to pathways with an average RDGV allele count (λ) that was sufficiently high to detect associations with an effect size $\beta = 1.5$ with 80% power (Methods 3.2.8). Given the cluster sizes ranging from 23 (cluster 9) to 93 (cluster 1), this involved testing between 504 and 896 germline pathways against each somatic cluster (Table 3-8).

Only clusters 1 and 2 showed potential signs of association with RDGVs (Simes p-value <0.1, Table 3-8, Figure 3-11). However, after FDR corrections using the Benjamini-Bogomolov method as in Section 3.4, the most significant pathways driving these results had relatively high FDRs (0.3<FDR<0.4, Table 3-9). The germline pathways (Chromatin organisation, and Activation of *ATR* in response to replication stress) were interesting, since they were both related to cancer biology[183,184]. However, orthogonal analyses using clinical and transcriptomic data did not yield any significant results that could give mechanistic insights into their associations with the somatic clusters. I therefore concluded that either these were not true associations of germline RDGVs with somatic features, or that substantially larger cohorts would be required to investigate them fully.

Interestingly, cluster 3 (characterised by a lack of *TP53* somatic driver alterations) was only weakly associated with RDGVs in the *ATM* signalling pathway (p=0.02, effect size =0.83). However, while all cluster 3 samples were *TP53*-wild type, cluster 3 only contained 57% of the total *TP53*-wild type samples across the cohort. The fact that the *ATM*-*TP53* association obtained from individual driver gene analysis in Section 3.5.2 was not recapitulated here indicated that stratifying samples in this way diluted the statistical strength of the original result.

Cluster	Samples (n)	Threshold λ_{min}^{c}	Pathways to test	Simes p-value
1	93	0.111	896	0.084
2	62	0.126	821	0.075
3	71	0.123	843	0.49
4	60	0.135	791	0.56
5	43	0.171	688	0.47
6	45	0.15	749	0.87
7	45	0.15	749	0.18
8	28	0.23	550	0.7
9	23	0.268	504	0.88

Table 3-8: GSIT analysis of RF clusters

RF clusters and their size, corresponding minimum RDGV allele count for pathways (λ_{min}^{c}), total numbers of pathways to test for association and resulting Simes p-values.

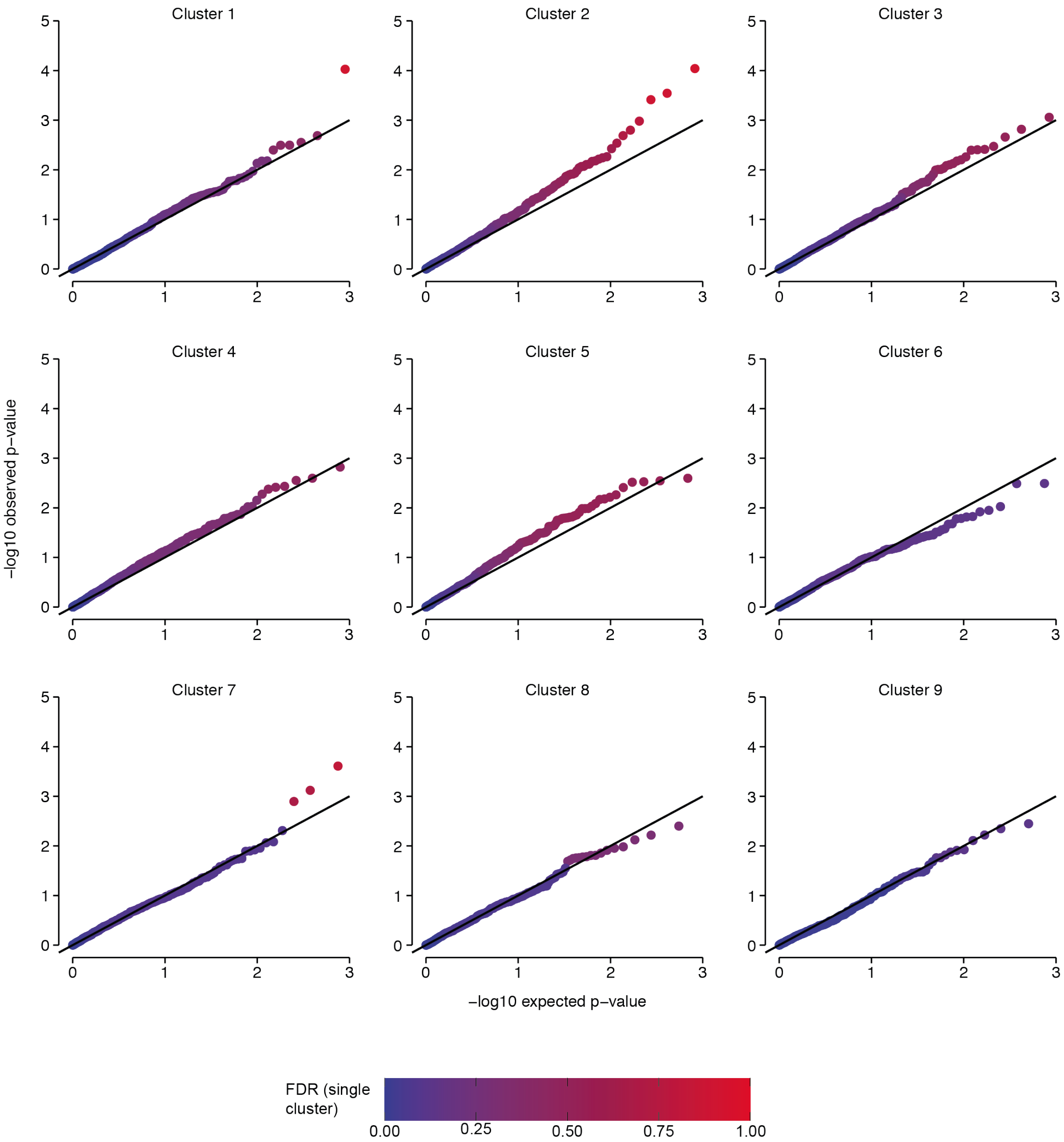

Figure 3-11: Logistic regression modelling results with RF clusters

QQ plot showing the results of testing for association between RDGVs in pathways with assignments of samples to each of the nine RF clusters. The colour scale corresponds to the standard FDR calculated for each cluster separately.

Germline pathway	Cluster	P-value	Estimated effect size	Benjamini-Bogomolov FDR
Chromatin organisation	2	9.1×10^{-5}	0.611	0.337

Activation of *ATR* in response to replication stress	1	$9.4x10^{-5}$	1.04	0.379

Table 3-9: Significant results from RF clustering and logistic regression modelling

Cluster-pathway pairs with Benjamini-Bogomolov FDR <0.4 are shown.

3.7. Discussion

In this chapter, I have attempted to identify the effects of germline variation on the acquisition of cancer drivers. I did so first by implementing a logistic regression approach to associate burden of RDGV alleles in pathways with driver alterations in individual genes. In an OAC cohort of 470 patients, this revealed two statistically significant results (FDR < 0.2). However, on closer inspection and analysis with orthogonal data sources, only one of these was readily interpretable, namely a negative interaction between germline *ATM* variants and somatic *TP53* driver alterations. I also attempted to characterise the landscape of somatic variation of OAC samples with a multiple driver genes and other features including mutational burden, ploidy and mutational signatures. While this analysis clustered samples into potentially interesting groups based primarily on recurrent driver genes, there was no interpretable significant association between somatic clusters and RDGVs in pathways.

This work has shown that there is value in analysing how germline perturbations influence somatic cancer evolution. In particular, the analysis in this chapter found that patients with RDGVs in *ATM* were significantly less likely than other OAC patients to have somatic *TP53* driver alterations. There were a number of pieces of evidence to support this being a genuine biological effect:

1. The result was found by treating all RDGVs equally, but was observed to be strongest for PTVs
2. *ATM* and *TP53* are direct interactors in a process that is relevant to cancer
3. *ATM* somatic alterations were also negatively associated with *TP53* drivers
4. Exclusivity between *ATM* and *TP53* mutations has previously been reported in several other cancer types.

This negative interaction alone is an interesting result in OAC, where *TP53* is by far most common somatic driver gene. The evolutionary pressures acting on *TP53* alterations in OAC are clearly mediated at least in part via *TP53*'s interactions with *ATM*. However, this result also raised the possibility that PTVs in *ATM* can predispose to developing OAC, albeit accounting for only ~1% of OAC cases (five out of 470 samples). The possibility of predisposition was also supported by several observations:

1. There were several somatic second hits to *ATM* for patients with germline PTVs
2. *ATM* germline PTVs predispose to individuals to developing other cancer types

3. There was a trend for younger age at diagnosis among OAC patients with germline *ATM* PTVs
4. *ATM* germline PTVs were observed to be more frequent in OAC patients compared to healthy controls.

If germline *ATM* variants do indeed predispose to OAC, then aside from any clinical implications for disease management, they further demonstrate the value of germline-somatic analyses to uncover functional roles of germline variation in cancer. It is important to note that case-control studies have not previously identified *ATM* as a predisposition gene for OAC. Indeed, OAC predisposition has only previously been identified from GWAS that have proven difficult to functionally interpret[93]. Thus, analyses of germline-somatic interactions may be able to detect germline effects in cancer that are missed by other study designs.

This work has also highlighted the difficulties involved in analysing interactions between germline variation and somatic drivers. The task is inescapably high-dimensional, with both germline variation and somatic mutational landscapes being complex, resulting in a multiplicity of complexity. In the approaches I chose, this burden was reflected by large numbers of hypothesis tests, despite characterising variation at the relatively high levels of pathways and genes. Combined with the fact that many effects of the germline are not highly penetrant in cancer, this placed a large burden on statistical power. Even with careful management of the available statistical power (for example, by frequency restriction and using non-standard corrections for multiple tests), it is likely that this work suffered from a small sample size. Indeed, I performed power calculations to estimate how many samples would be required to comprehensively assess GSITs using logistic regression, assuming that variants would be seen only in relatively few samples (such as *ATM* germline PTVs in 1% of OACs). In order to detect an interaction with effect size $\beta = 1$ between germline variants of interest present in 1% of samples and a driver gene altered in 5% of samples with 90% power, maintaining a FWER of 0.05 with a total of 10^4 hypothesis tests, roughly 40,000 samples would be required. Unfortunately, cohorts of this size from individual cancer types with paired germline and somatic sequencing data may not be available in the near future.

The analysis in this chapter relied on more sophisticated modelling approaches than those undertaken in Chapter 2. In particular, accounting for potentially confounding covariates was a substantial improvement. However, it is worth noting that, taking the association between *ATM* and *TP53* mutations as a benchmark, several of the methodological changes seemed to have little impact on the results. For instance, treating germline RDGVs as counts rather than as binary variables had little bearing on the main result, which was found to be largely driven by heterozygous variants in a single gene. The focus on identifying putatively deleterious missense variants may also have been unnecessary, since ultimately it was truncating variants in *ATM* that exhibited the strongest effect. Characterising germline variation at the level of pathways also appears not to have helped, since the result centred around a single gene in the germline. However, four other genes in the *ATM* signalling pathway also appeared to contribute to the association, suggesting that a balance between focusing on individual genes and on entire pathways could be useful. Finally, the fact that this gene (*i.e. ATM*) was a known cancer predisposition gene suggests that focusing on known cancer-related genes in the germline may be a more useful way of analysing germline-somatic interactions than using all biological processes, given currently available sample sizes.

Similarly, the fact that the RF clustering analysis failed to identify interesting germline-somatic interactions suggests that individual driver genes may be the most useful way to characterise somatic evolution in GSIT analyses. It may also be the case that any clustering driven solely by somatic data is unlikely to identify features relevant to germline variation. Moreover, it is important to acknowledge the fact that, although RF clustering was chosen to handle the mixed data types of binary somatic driver alterations continuous somatic covariates, the resulting clustering was driven primarily by driver genes alone. This imbalance may be rectified by finer tuning of the RF approach, or by using a different clustering method altogether.

Overall, the observations arising from the work in this chapter suggest that the approach taken by Lu *et al.*[55] of associating germline PTVs in cancer genes with somatic alterations in driver genes may well be the most effective GSIT analysis approach currently available. Two strands of future work may be of value. First, further investigation into the potential role of *ATM* as a predisposition gene for OAC may have important clinical implications. For example, it may be helpful for patients with germline *ATM* PTVs to be monitored with regular diagnostic tests, particularly if they also have Barrett's syndrome. Second, further analysis of the effect of germline variation on somatic evolution across cancer types will likely give greater insights into the role of the germline in cancer. However, it would seem that any such analysis should attempt to relate germline PTVs in cancer genes with somatic alterations in driver genes, while accounting for covariates, and sample sizes for such analysis should aim to be around 40,000.

Chapter 4. Literature support and systems-level properties of driver genes

The second broad research question of this thesis is how to identify the aspects of inter-tumour heterogeneity that are most relevant to cancer biology and therapy, *i.e.* driver genes. Reliably identifying driver genes is challenging because of the enormous repertoire of passenger alterations in cancer[25]. An obvious starting point for new driver gene identification efforts is existing knowledge about previously-identified driver genes. In this chapter, I will first introduce a database curated by the Ciccarelli lab, the Network of Cancer Genes[26,185-187] (NCG), that collects driver genes reported in the literature and annotates their properties (Section 4.1). I will then analyse the breadth of literature evidence for the driver genes curated in NCG and how this can inform focuses of future research (4.3), before describing the systems-level properties that distinguish driver genes from other human genes (4.4). Finally, I will discuss the results of this chapter in Section 4.5.

This chapter describes results from a published study of which I was a lead co-author[24]. I will present an overview of the main results of this publication, as well as detailed accounts of my own contributions.

4.1. Introduction

Genetic inter-tumour heterogeneity is the variation in somatically altered genes (including drivers) between tumours. The identification and further study of driver genes has allowed researchers and clinicians to overcome this heterogeneity to an extent, through the use of targeted therapies that have improved patient outcomes in some cancer types[188,189]. Much work has been done to identify driver genes, with hundreds of studies reporting on driver genes in many cancer types[15,16,24]. In order to leverage this body of work in ongoing and future research, it is imperative for researchers to have a resource where knowledge about driver genes identified in the literature is collected.

The Cancer Gene Census[111] (CGC) is a well-known example of such a resource. The CGC records genes with established roles in cancer, *i.e.* canonical driver genes, and it is constantly curated to maintain up-to-date information. The CGC identifies canonical drivers based on a combination of experimental evidence that genes contribute to the hallmarks of cancer[3,4], and computational evidence that somatic alterations are consistent with the roles of genes in cancer[111]. For example, genes with experimental evidence of a tumour suppressor role should exhibit a prevalence of loss-of-function alterations in cohorts, while oncogenes should show recurrent missense mutations or gene amplifications. As of version 86 (August 2018)[111], the CGC is divided into two tiers, with Tier 1 including genes supported by both experimental and computational evidence, and Tier 2 containing genes with a single line of evidence.

Many cancer sequencing screens have been conducted by the cancer research community to identify driver genes. However, many drivers reported by these screens are not included in either tier of the CGC, as they do not meet the stringent inclusion criteria. In particular, while cancer sequencing screens can provide computational evidence that genes drive cancer, they often lack comprehensive experimental validation. Genes reported by sequencing screens that are not canonical driver genes can be considered to be 'candidate' drivers. It is likely that these candidate drivers are in reality a mix of true drivers and false positives, but in the absence of definitive experimental evidence it is impossible to establish their role in cancer with certainty. Other databases of driver genes tend to be more restrictive than the CGC, focusing only on tumour suppressor genes[190], oncogenes[191], or genes that drive a particular cancer type[192]. As such, in order for the research community to have a complete picture, there is a need for a resource that collects driver genes as they are reported in the literature, including candidate drivers.

In addition to the identity of driver genes, the properties of driver genes have been studied extensively. Driver genes are distinguished from other human genes by an array of systems-level properties that do not directly relate to their role in cancer, but that instead describe them as a set. For example, canonical drivers, and in particular tumour suppressor genes, are less likely than other human genes to have duplicated loci elsewhere in the genome[193,194]. Conversely, oncogenes are more likely to have evolved through whole-genome duplication events that occurred at the basis of vertebrates (*i.e.* to be ohnologs)[193]. Canonical drivers are essential genes more often and in a higher proportion of cell lines than other genes[24]. They are expressed in a wider range of healthy human tissues, both at the gene[26,195] and protein[24] levels. Proteins encoded by canonical drivers have distinct topological features in the protein-protein interaction network (PPIN), namely they have higher PPIN degree, betweenness and clustering coefficients than other proteins[194]. Canonical driver proteins also participate in more complexes[26] and their genes are targeted by more miRNAs[193], suggesting that the expression of driver genes is tightly regulated. Tumour suppressor genes tend to be old genes with a pre-metazoan origin, while oncogenes tend to originate in metazoans and cancer drivers in general are depleted in post-vertebrate genes[193]. Finally, canonical drivers encode longer proteins with more domains than other human genes[26,195]. An holistic view of the properties of cancer genes can lead to a deeper understanding of cancer biology, and even the identification of new cancer genes[195]. Thus, it would be valuable to have a resource that collected the systems-level properties of driver genes identified in the literature.

The Network of Cancer Genes (NCG) is a project started in 2010[26,185-187] that aims to provide such a resource. NCG is a curated database that collects driver genes reported by cancer sequencing screens, extracted from the literature by a manual expert review. NCG also annotates the systems-level properties of these driver genes, with data extracted from external sources before being processed and analysed. Since many sequencing screens identifying driver genes are published each year, and the available data on the systems-level properties of driver genes continues to grow, the NCG database is updated regularly. In this chapter, I will describe the update to the sixth version of NCG, which was the subject of Repana *et al.*[24]. In Section 4.3 I will briefly outline the results of the literature review (conducted by several colleagues), and investigate the available literature evidence for driver genes (conducted by me) with a view to informing future research efforts. In Section 4.4, I will describe the annotation and analysis of the systems-level properties of cancer genes, focusing on those properties for which I was carried out the work. In Section 4.5, I will discuss NCG 6 and its implications both for the rest of this thesis and for wider cancer research.

4.2. Methods

For a full description of the methods used to extract driver genes from the literature for NCG 6, refer to Repana *et al.*[24].

4.2.1. Expression in healthy human tissues

RNA-Seq data from healthy human tissues for 18,984 human genes (including all 2,372 cancer genes) were derived from the non-redundant union of Protein Atlas[196] v.18 and GTEx[197] v.7. Protein Atlas reported the average transcripts per million (TPM) values in 37 tissues, and genes were considered to be expressed in a tissue if their expression value was ≥ 1 TPM. GTEx reported the distribution of TPM values for individual genes in 11,688 samples across 30 tissue types. In this case, genes were considered to be expressed if they had a median expression value ≥ 1 TPM. Tissues were matched between the GTEx and Protein Atlas databases, giving 43 unique human tissue types. In tissues for which both sources provided expression data, genes were considered to be expressed only if both databases supported this.

Protein expression was derived from immunohistochemistry assays of healthy human tissues, obtained from Protein Atlas v.18. Data were available for 13,001 human proteins including 1,799 cancer proteins. Proteins were categorised as not detected or as having low, medium, or high expression in 44 tissues on the basis of staining intensity and fraction of stained cells. In Protein Atlas, expression levels were reported in multiple cell types for each tissue. The highest reported value was retained as the expression level for that tissue. For the expression analyses, proteins were considered to be expressed in a tissue if they were recorded as having low, medium or high levels of expression in Protein Atlas.

4.2.2. Protein function

Data on functional categories (pathways) were collected from Reactome[107] v.63 and KEGG[108] v.85.1. All levels of Reactome were included, and level 1 and 2 pathways from KEGG were added separately. Overall, functional annotations were available for 11,344 human proteins, including 1,750 cancer proteins assigned to 2,318 pathways in total. Enrichment and depletion of cancer genes in pathways (for level 1 of Reactome and level 2 of KEGG) was assessed using Fisher's exact tests, using the FDR correction for multiple hypothesis testing.

4.3.3. Other systems-level properties

Other systems-level properties of human genes were obtained from various sources by my colleagues in the Ciccarelli lab. Duplicated gene loci (genes with 60% protein sequence shared with another locus) were identified by aligning protein sequences from RefSeq[75] release 85 to the human genome. Data on gene essentiality in cell lines were obtained from the PICKLES (September 2017)[198] and OGEE v2[199] databases. Genes in PICKLES were considered to be essential in a cell line if they had Bayes factor >3, while original annotations of essentiality from OGEE were retained. The protein-protein interaction network was constructed from the union of BioGRID v3.4.157[200], MIntAct v4.2.10[201], DIP (February 2018)[202] and HPRD v9[203], resulting in 16,322 proteins and 289,368 interactions supported by at least one original publication. Topological properties (PPIN degree, betweenness and centrality) were calculated using custom scripts. Genes were identified as participating in complexes using data taken from CORUM (July 2017)[204], HPRD v9[203] and Reactome v63[107]. The number of miRNAs targeting a gene was calculated using data from miRTarBase v7.0[205] and miRecords v4.0[206]. The evolutionary origin (age) of genes was identified as previously described[193], using data from EggNOG v4.5.1[207].

Other systems-level properties not included in NCG 6, including information on ohnolog genes, protein length and protein domain composition, were processed by me. A list of ohnolog genes was obtained from Nakatani *et al.*[208]. Protein length was obtained from RefSeq[75] release 94 using the longest isoform for each gene. The number of domains in each protein was extracted from InterPro[209] annotations in UniProt[210] release 2019_07.

4.3. Analysis of the available evidence from cancer sequencing screens

The core offering of the NCG database is a collection of driver genes, derived from published cancer sequencing screens as well as two separate sources of canonical drivers. The literature review to extract these genes was led by a colleague of mine, Dimitra Repana. Numerous other authors of the study[24] also contributed, both by identifying and curating cancer sequencing screens, as well as by reviewing the initial curations of others. The literature review encompassed two primary sources of canonical drivers (the CGC version 84[211] and a list from Vogelstein *et al.*[25]) and 273 cancer sequencing screens, published between 2008 and March 2018. Driver genes were extracted as reported in the original publications, for a total of 2,372 drivers, comprising 711 canonical drivers (of which 239 were identified as tumour suppressor genes and 239 were identified as oncogenes) and 1,661 candidate drivers (*i.e.* genes reported by sequencing screens and not included in the sources of canonical drivers). Compared to the previous version of NCG[26], this represented a more than 1.5-fold increase in both the number of publications and the number of driver genes. The 273 sequencing screens included six pan-cancer studies, as well as screens of 119 cancer types from 31 primary anatomical sites.

In order to help guide future research in driver gene identification, we sought to understand how driver genes were being identified and reported in the literature. I conducted analyses of how extensively different cancer types were studied in terms of sequencing screens, and of the breadth of literature support for different driver genes, described in the remainder of this section.

The number of reported driver genes varied greatly across primary sites (Figure 4-1A), ranging from one (vascular system) to 513 (blood). This stark variation was in large part accounted for by the total number of cancer donors sequenced for each primary site. Primary sites with more donors had a clear tendency to have more driver genes reported in the literature (Figure 4-1B, rho=0.9, p=$6.5\text{x}10^{-12}$, Spearman correlation). This suggested that the variation in the number of driver genes per primary site may have been due to differences in how extensively cancer types had been studied, rather than due to biological differences between cancer types. This was unsurprising, since most widely used methods for identifying driver genes rely on recurrence across samples, and thus have strong dependencies on cohort size for statistical power. Thus, it is reasonable to suppose that if researchers analysed more patients from as-yet under-studied cancer types, they might discover many more driver genes in these cancers. Alternatively, if sequencing larger cohorts in these cancer types is not feasible, other driver detection methods are required.

The proportion of canonical to candidate driver genes also showed substantial variation between primary sites (Figure 4-1A). More than 75% of driver genes in cancers of the prostate, soft tissues, bone, ovary, cervix, thymus, and retina were canonical. By contrast, more than 75% of driver genes in cancers of the penis, testis, and vascular system were candidate drivers. The proportion of candidate drivers per primary site showed a weak positive correlation with the total number of driver genes (rho=0.37, p=0.04, Spearman correlation). This suggested that, as more drivers were found in a given primary site, a greater proportion of these genes lacked established experimental support. Thus, in extensively-studied cancers like those of the blood, brain and breast, efforts may be better spent experimentally validating cancer genes, rather than performing more sequencing screens. It also indicated that as more patients in a given cancer type were sequenced, a greater number of rare or poorly-studied driver genes were reported. This was consistent with the dependence on cohort size of many established driver detection methods for statistical power. It may also be the case that such methods are ill-suited for analysing the long tail of rare driver genes, and therefore that other driver gene detection approaches are required to fully overcome inter-tumour heterogeneity and identify all driver genes.

In the absence of experimental evidence, the consensus of different studies to identify the same driver genes could be a useful measure of how reliable driver gene predictions were. I assessed this for the driver genes in NCG 6 by investigating how many of the 273 sequencing screens reported each gene as a driver. As expected, highly penetrant drivers such as *TP53*, *PIK3CA* and *KRAS* were reported by many screens (Figure 4-1C). Strikingly however, the vast majority of candidate driver genes (80%) were supported by only a single screen (Figure 4-1C). This suggested that most driver genes reported in the literature had a poor level of support, and so were potential false positives. It was also possible that methodological limitations were causing a lack of reproducibility between studies. To assess this, I restricted analysis to the 108 sequencing screens that used widely accepted gold-standard driver gene detection methods (MutSig[44] and MuSiC[212]). These screens reported a total of 875 candidate drivers, of which 89% were reported a single one of the 108 screens. This suggested that even gold-standard methods were limited in their ability to reproducibly identify driver genes. However, the number of screens alone may not be a good predictor of a gene's true role in cancer. This was evidenced by the fact that the majority of canonical driver genes (57%) were only supported by either one or even zero sequencing screens (outside of the two primary sources of canonical drivers). Indeed, this called the veracity of these canonical drivers into question, since it indicated that they were rarely somatically altered across studied cancer cohorts, if at all. Furthermore, the known false positive driver gene *TTN*[44] had broader support than any other candidate driver (nine screens, Figure 4-1C). This suggested that a combination of methodological limitations and extensive inter-tumour heterogeneity may have been driving a lack of consensus between cancer sequencing screens.

In order to account for biological differences between cancer types, I investigated the consensus between studies to identify the same driver genes in a single cancer type, skin melanoma. NCG 6 included nine skin melanoma sequencing screens. Despite the fact that these screens investigated the same cancer type over a four-year period (2011 to 2015), there was a clear lack of consensus between the screens. In seven of the nine skin melanoma studies, over half of the reported driver genes were not identified by any of the other screens (Figure 4-1D). This was unaffected by cohort size (p=0.6, Spearman correlation), and over 50% of the driver genes from two out of the three studies with 200 or more patients were study-specific. It may be the case that consensus can be obtained by both having large cohort sizes and by using gold-standard driver detection methods. However, in the two studies for which this was the case, the proportions of study-specific driver genes were 71% and 23%, suggesting that other factors may influence the consensus between cancer sequencing screens. An obvious possible factor is the heterogeneity between patients, which represents a biological barrier to identifying driver genes, particularly when using recurrence-based methods. Indeed, skin melanoma has particularly high levels of mutations[16] and genetic heterogeneity[213], which may partially explain the poor levels of consensus between cancer screenings in this cancer type. Nonetheless, further methodological improvements in driver gene detection could help to improve the consensus between cancer sequencing screens.

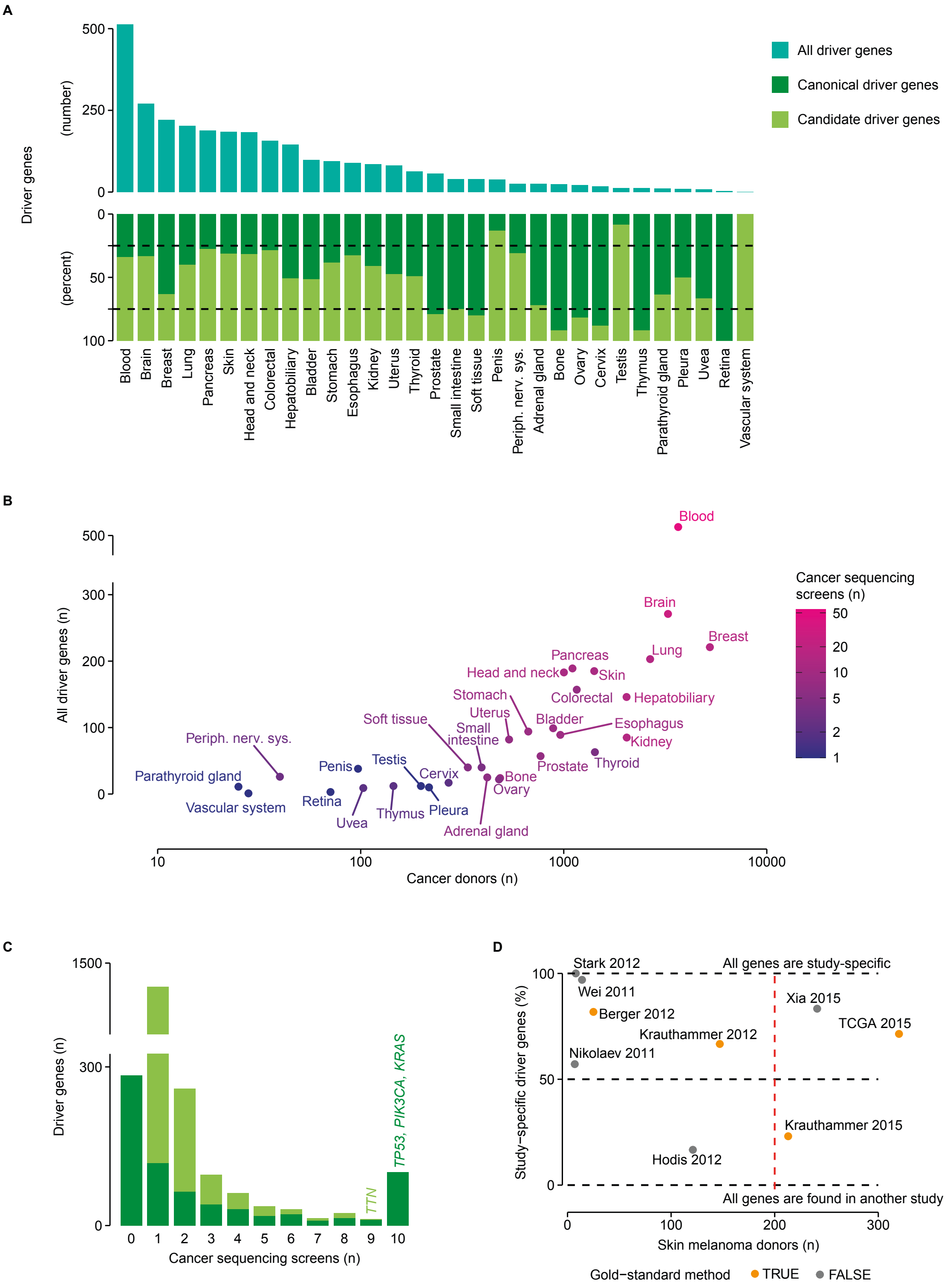

Figure 4-1: Analysis of driver genes reported in the literature

A. Total number of unique cancer genes (top) and proportion of canonical to candidate driver genes (bottom), reported for each primary anatomical site.

B. Relationship between the number of sequenced cancer donors and the number of unique reported driver genes for each primary site.

C. Breadth of support for driver genes in the literature, as measured by the number of publications supporting each canonical and candidate driver gene. The two primary sources of canonical driver genes are not included in this count, so canonical driver genes that were not reported in any of the cancer sequencing screens are listed as having zero screens.

D. Proportion of driver genes that are study-specific, for each of the nine studies investigating driver genes in skin melanoma. The overlap of reported driver genes was only assessed within these nine studies. Gold-standard methods include all versions of MutSig[44] and MuSiC[212].

4.4. Systems-level properties of driver genes

In addition to curating lists of driver genes reported in the literature, NCG annotates the systems-level properties that distinguish driver genes from other human genes. Two systems-level properties were new to the sixth version of NCG, namely breadth of protein expression and gene essentiality. Different authors of the study[24] were responsible for the analysis of different systems-level properties. I collected and analysed data on gene and protein expression in healthy human tissues and protein function, which I will describe in Sections 4.4.1 and 4.4.2. I will also provide an overview of the remaining systems-level properties of driver genes in Section 4.4.3.

4.4.1. Driver genes are widely expressed across healthy tissues

Driver genes tended to be more widely expressed across healthy human tissues than the rest of human genes. This was evident for canonical drivers both at the gene ($p<2.2\times10^{-16}$, Wilcoxon rank-sum test, Figure 4-2A) and protein ($p=9.5\times10^{-11}$, Figure 4-2B) levels (Methods 4.2.1). The trend did not hold for candidate drivers, which were expressed in similar numbers of tissues to the rest of human genes. In particular, a sizeable proportion of candidate drivers (21%) were expressed in few tissues (≤6 tissues for gene expression, ≤8 tissues for protein expression). This could be because candidate drivers include many false positives, and they therefore do not have the properties of true driver genes. However, it is also possible that many candidates are true drivers, but they are more rarely active in cancer than canonical drivers and are more tissue-specific in their cancer-promoting role.

Of note, the pattern of expression was markedly different between tumour suppressor genes (TSGs) and oncogenes (OGs, $p<0.006$, Figures 4-2A, B). TSGs showed the strongest tendency of any gene category to be widely expressed, indicating that they were active in almost all healthy human tissues. By contrast, OGs were less widely expressed, indicating less ubiquitous functions in normal physiology. This difference between TSGs and OGs lended support to the possibility that some candidate drivers were not widely expressed due to their particular role in cancer, rather than because they were simply false positive cancer genes.

4.4.2. Driver genes are enriched in certain biological processes

Driver genes also differed from the rest of human genes by their function in normal physiology. I performed gene set enrichment analyses using high-level pathways from Reactome and KEGG to assess the function of cancer genes (Methods 4.2.2). As expected, driver genes were consistently enriched in functional categories such as signal transduction, chromatin reorganisation, and cell cycle (Figures 4-2C, D). Candidate drivers generally showed weaker enrichment than canonical drivers, most notably in DNA repair. Interestingly, however, extracellular matrix (ECM) reorganisation showed a specific enrichment for candidate drivers. This might indicate that pathways that are not commonly associated with cancer, like ECM reorganisation, would be worth investigating more comprehensively, particularly from the point of view of experimental validation of their role in cancer.

TSGs and OGs showed some differences in their patterns of functional enrichment. Most notably and as expected, TSGs were selectively enriched in DNA repair. This may help to explain why TSGs were more widely expressed in human tissues than OGs, since DNA repair is a core biological function that is common to all tissue types.

Interestingly, all driver genes were depleted in metabolic pathways (Figures 4-2C, D). Given that metabolism is known to play a key role in cancer, this may reflect the fact that most metabolism genes promote cancer through transcriptional dysregulation, rather than through acquiring somatic mutations[214]. Thus, these genes would not be identified as cancer genes by DNA sequencing screens. This supports the developing trend to interrogate cancer biology by incorporating multi-omic data, rather than using genomic data alone[215].

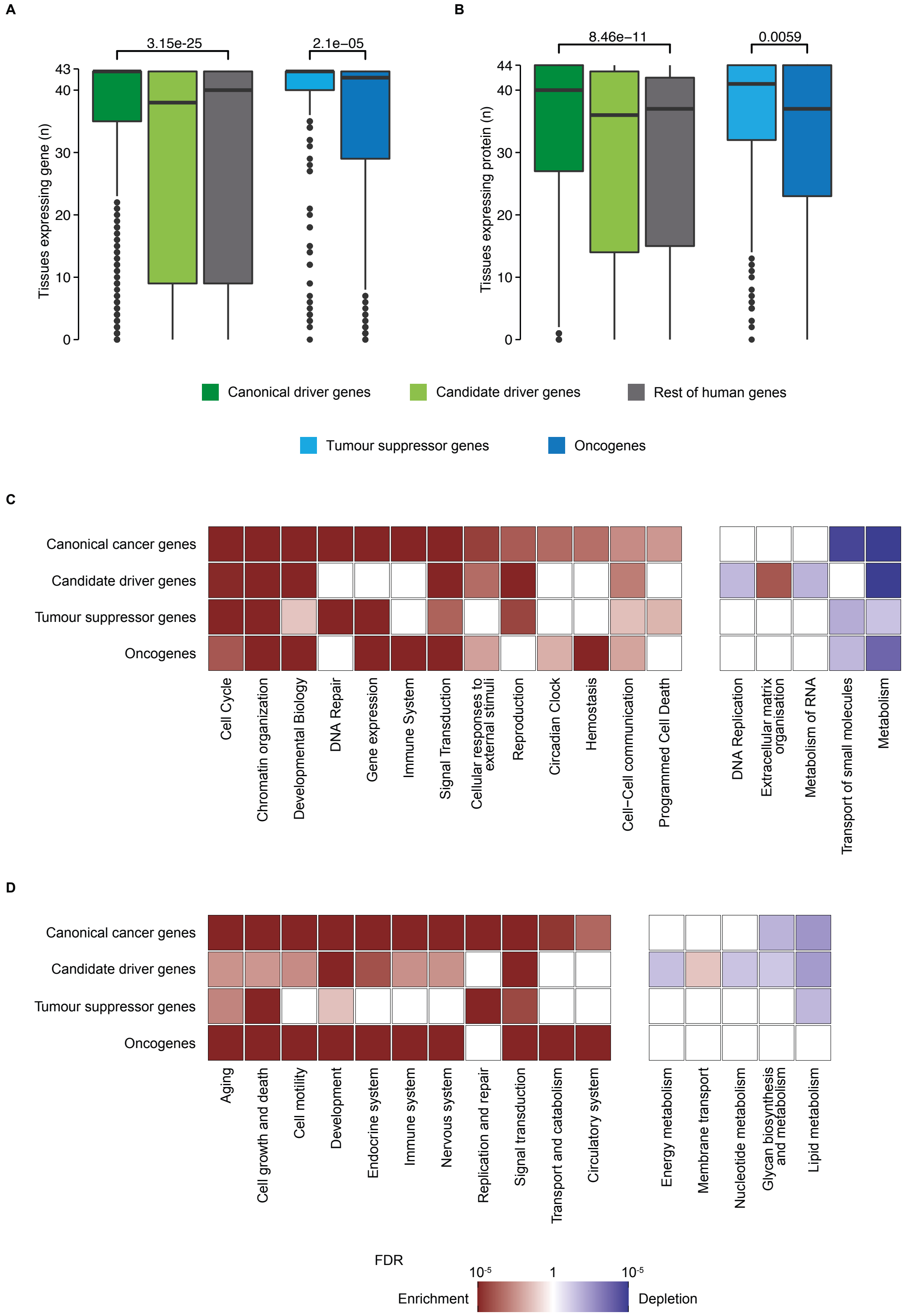

Figure 4-2: Breadth of expression and functional enrichment of driver genes

Number of healthy human tissues in which genes are expressed at the mRNA (**A**) and protein (**B**) levels. P-values were calculated with the Wilcoxon rank-sum test.

Enrichment and depletion of driver gene categories in pathways from level 1 of Reactome (**C**) and level 2 of KEGG (**D**), assessed with Fisher's exact test. Tile colour corresponds to the false discovery rate (FDR) of enrichment (red) or depletion (blue), with non-significant tiles (FDR ≥0.05) coloured white.

4.4.3. Other systems-level properties of driver genes

The sixth version of NCG annotated other systems-level properties of driver genes that were analysed by other authors of the study[24] (Methods 4.3.3). A smaller proportion of driver genes than other human genes had loci duplicated at ≥60% coverage elsewhere in the genome and this was particularly pronounced for TSGs, with OGs showing the opposite trend (Figure 4-3A). Canonical drivers, and in particular TSGs, were essential in more cell lines than the rest of human genes (Figure 4-3B). Proteins encoded by canonical and candidate driver genes had distinct topological features in the protein-protein interaction network, having higher degree (Figure 4-3C), betweenness (Figure 4-3D) and clustering coefficient (Figure 4-3E) than other genes. They also participated in higher numbers of protein complexes (Figure 4-3F). Canonical driver genes were targeted by a higher number of miRNAs than other human genes (Figure 4-3G). Finally, candidate driver genes and TSGs were also more likely than other genes to have a pre-metazoan evolutionary origin (Figure 4-3H).

In addition, driver genes were distinguished by several other systems-level properties that have been previously described[26,193,195] but were not included in NCG 6. I annotated these (Methods 4.3.3) and include them here for completeness, as they will be discussed further in Chapter 5. Canonical and candidate drivers, and particularly OGs, tended to have evolved through whole-genome amplification events, *i.e.* to be ohnologs (Figure 4-3I). They also encoded proteins with distinct structural features, being particularly long (Figure 4-3J) and having a large number of identifiable domains (Figure 4-3K). Together, these results demonstrated that driver genes possess distinct systems-level properties that set them apart from the rest of human genes.

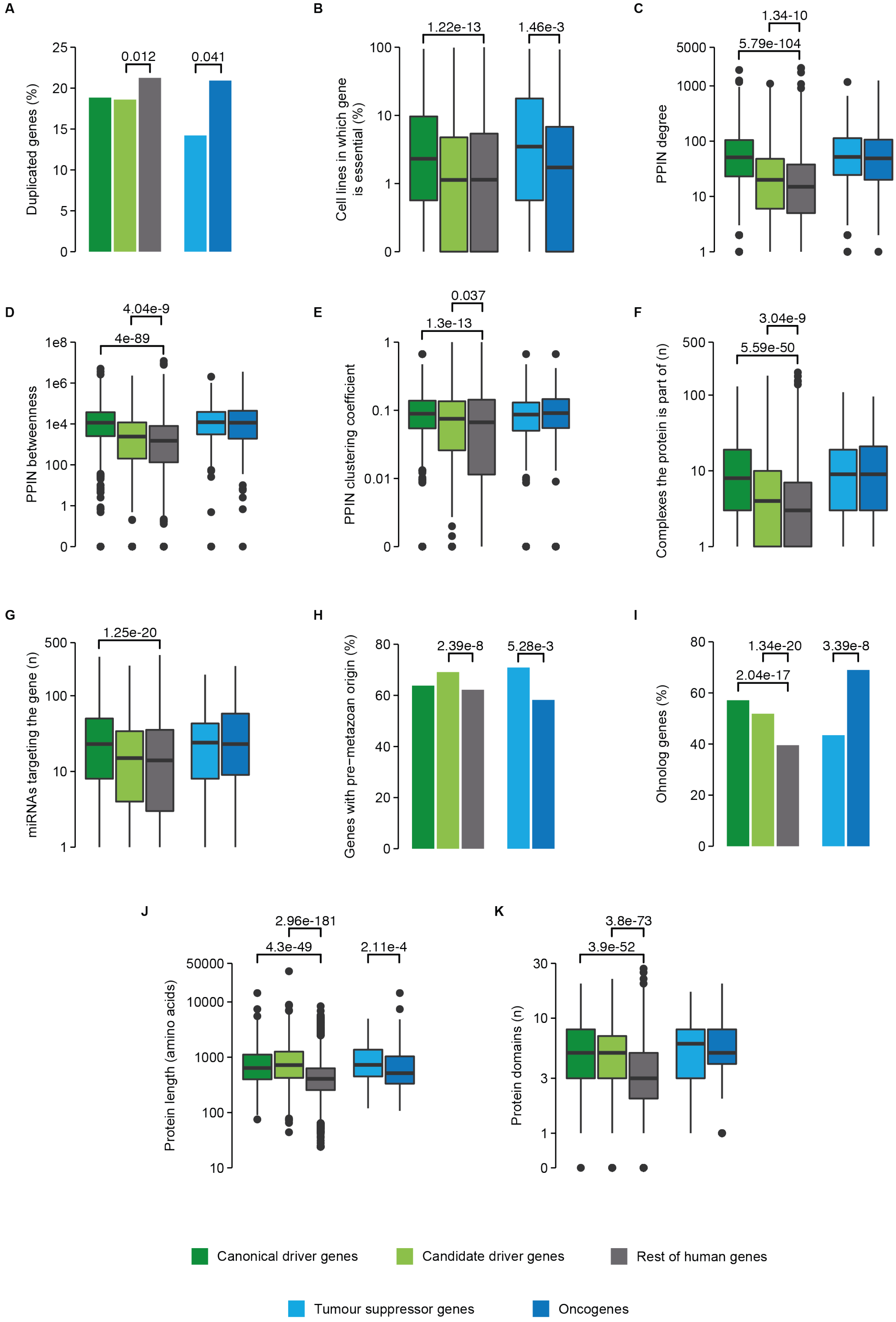

Figure 4-3: Other systems-level properties of driver genes

A. Percentage of genes with ≥ 1 gene duplicate covering ≥ 60% of the protein sequence.

B. Percentage of cell lines in which each gene is essential.

Degree (**C**), betweenness (**D**) and clustering coefficient (**E**) of each protein in the protein-protein interaction network (PPIN).

F. Number of complexes each protein is a part of.

G. Degree of the target genes in the miRNA-target interaction network.

H. Proportion of genes originating in pre-metazoan species.

I. Percentage of genes that are ohnologs, *i.e.* that have evolved through whole-genome duplication events that occurred at the basis of vertebrates.

J. Protein length in amino acids.

K. Number of identified domains making up each protein.

Significance was calculated using a two-sided Fisher test (**A**, **H**, **I**) or Wilcoxon test (**B**-**G**, **J**, **K**). Canonical drivers and candidate drivers were both compared to the rest of human genes, and TSGs and OGs were compared to each other. Only comparisons with $p < 0.05$ are shown.

4.5. Discussion

The Network of Cancer Genes (NCG) is a valuable resource for the cancer research community that collects driver genes reported by published sequencing screens, and annotates their systems-level properties. In the update to the sixth version of NCG, the database content increased more than 1.5-fold. Analysis of the breadth of support of driver genes showed that most reported driver genes were specific to a single study, and that studies could even fail to reach a consensus on the identity of driver genes in a single cancer type. Finally, driver genes were distinguished from other human genes by a range of systems-level properties.

The primary use for NCG is as a repository of driver genes, and at the time of writing it has been used at least 29 times for this purpose in published research articles. However, the analysis of how driver genes were identified in the literature also highlighted some of the ways in which future efforts for driver gene identification could be improved. For instance, there was a strong link between the number of donors sequenced in a given cancer type, and the number of driver genes reported in that cancer type. Some rarer cancers such as those of the vascular system, retina and parathyroid gland had very low numbers of both donors and driver genes, suggesting that larger sequencing efforts in these cancer types could identify new drivers and better inform molecular therapy. In addition, the observation of a poor consensus between sequencing screens supports the increasing adoption of gold-standard cancer gene identification methods, rather than bespoke algorithms. However, the fact that even existing gold-standard methods failed to identify consensuses of driver genes suggested that other driver detection methods may be required to robustly investigate the long tail of rare or even patient-specific driver genes. Methods tailored to identifying these genes may also provide an alternative to sequencing large numbers of patients from rare cancer types, which can be challenging to undertake.

As an example use-case of NCG, I made extensive use of the database throughout my work in this thesis. I used lists of canonical driver genes to identify somatic driver alterations in Chapters 2 and 3, to investigate how germline variation influenced somatic evolution in cancer. I will also use the systems-level properties of canonical driver genes as the basis for a patient-level driver gene detection algorithm, sysSVM2, in Chapter 5. This algorithm learns the properties of canonical drivers to identify other genes with similar properties, thus enabling the identification of new putative driver genes in individual patients. Thus, the NCG database is useful as a collection of reported driver genes, its literature review process can help to guide future studies of driver genes, and its annotation of the properties of driver genes can be used for new driver detection approaches.

Chapter 5. Patient-level driver gene detection

As discussed in Chapters 1 and 4, identifying cancer driver genes is necessary for a full understanding of cancer biology and for advancing cancer therapeutics. However, it is challenging to distinguish drivers from other human genes due to the large and heterogeneous array of passenger alterations to the genome that occur in cancer. Identifying driver genes in individual patients is an even greater challenge than doing so across cohorts, but also holds greater potential for advancing precision oncology. In this chapter, I will first introduce the problem of patient-level driver gene detection and review existing driver detection methods (Section 5.1). I will then describe the development of a machine-learning approach, sysSVM2, that uses the systems-level properties of canonical driver genes (discussed in Chapter 4), to identify drivers in individual patients (5.3). I will then assess the performance of sysSVM2 on real and simulated pan-cancer data (5.4), before exploring possible alternatives and extensions of the method (5.5). Finally, I will discuss the findings of this chapter in Section 5.6.

This chapter extensively incorporates material from Nulsen *et al.* Genome Medicine **13** (2021). I am the sole first author of this article, but I would like to thank the other authors for their contributions: Hrvoje Misetic, Prof. Christopher Yau and Prof. Francesca D. Ciccarelli. In particular, H.M. carried out the pan-cancer TCGA analysis (5.4.4), C.Y. supervised the machine learning aspects of the work, and F.D.C. conceived and directed the study.

5.1. Introduction

5.1.1. The challenge of patient-level driver detection

The majority of somatic alterations to the cancer genome are thought to have little or no phenotypic consequence for the development of the disease. Only a small proportion of somatic alterations are believed to play a role in driving cancer. The repertoire of observed somatic alterations differs greatly between tumours, giving rise to differences in prognosis and response to therapy that represent a significant challenge to oncology. The clinical response to inter-tumour heterogeneity has been the development of precision oncology, in which driver events specific to patients are targeted[10,11]. This of course requires driver events, and the genes that they effect (driver genes), to first be identified.

Identifying driver genes amounts to sorting through the heterogeneous cancer genomic landscape, finding the drivers among the passengers[25] (*i.e.* the somatic alterations with no phenotypic consequence). Driver alterations undergo positive selection during cancer development because of the advantages they confer to malignant cells[4]. As such, many approaches for identifying driver genes have looked for evidence of positive selection, most commonly in the form of recurrence across cohorts of patients [24]. These cohort-level methods have been of great value, leading to the identification of more than 2,000 well-established (canonical) or candidate cancer driver genes[24,111], as we saw in Chapter 4. However, cohort-level approaches fail to reliably identify rare driver events that occur in small cohorts or even in single patients because of low statistical power[216]. Even the largest and most comprehensive cohort-level efforts to identify driver genes typically leave a fraction of patients with few or no drivers[16]. Moreover, these methods are not ideal for application in the clinical setting because they return lists of drivers in entire cohorts, rather than predictions in individual patients.

In order to fully realise precision oncology, the driver events in every tumour must be found, including rare and patient-specific events. Some patient-level driver detection methods have been developed for this purpose, but they are more challenging to implement than cohort-level methods for a number of reasons. First, there is a lack of available ground truth. Gold-standard sets of rare cancer drivers are scarce, since these genes are challenging to identify in the first instance. In addition, outside of a few very well-studied genes, little is known about why certain alterations in driver genes might or might not contribute to cancer[16,217]. Thus, it is difficult to formulate models in order to systematise patient-level driver gene detection.

5.1.2. Review of existing driver detection methods

Numerous cohort-level driver detection methods have been developed to date. For example, recurrence-based methods such as MutSigCV[44] and MuSiC[212] search for genes whose mutation rate (single nucleotide variants (SNVs) and small insertions or deletions (indels) per nucleotide) is above the background level. This is because mutations in cancer drivers are more likely to become fixed and recur across samples than those in non-driver genes. GISTIC2[218] adopts a similar approach for recurrent copy number variants (CNVs). OncodriveCLUST[219] and ActiveDriver[220] look specifically for mutations clustering in hotspot positions or encoding post-translational modification sites. TUSON[221] and 20/20+[222] predict new drivers based on features of canonical oncogenes and tumour suppressors, including the proportion of missense or loss-of-function to silent mutations occurring across patients. dNdScv[217] computes the nonsilent to silent mutation ratio to identify gene mutations under positive selection, while OncodriveFM[223] focuses on biases towards variants of high functional impact. Finally, network-based methods like HotNet2[224] incorporate gene interaction networks to identify significantly altered modules of genes within the cohort. Albeit with different approaches, all these methods rely on the comparison of alterations and/or altered genes across patients.

Patient-level methods are less well-established and less numerous than cohort-level methods. A few attempts such as OncoIMPACT[225], DriverNet[226] and DawnRank[227] combine transcriptomic and genomic data to identify gene network deregulations in individual samples. Such methods require user-specified gene networks and deregulation thresholds, which can affect their results[225]. In addition, matched exome and transcriptome data from the same sample are not always available, especially in clinical settings where shotgun transcriptomic sequencing is still rare. Alternative approaches such as PHIAL[228] match the patient mutations with databases of known clinically actionable or driver alterations but have a limited capacity to identify as-yet unknown driver alterations. To overcome this limitation, iCAGES[229] combines deleteriousness predictions and curated database annotations to learn features of true positive and true negative driver alterations.

The Ciccarelli lab recently developed sysSVM, a patient-level driver detection method based on one-class support vector machines (SVMs)[27,230]. sysSVM learns the distinct systems-level features (gene properties) and molecular features (damaging somatic alterations) of canonical drivers. It then predicts as drivers the altered genes in individual patients that best resemble these features. When applied to 261 patients with oesophageal adenocarcinomas, sysSVM successfully identified the driver events in every patient[27].

5.2. Methods

5.2.1. The sysSVM algorithm

As previously described[27], sysSVM consists of four one-class Support Vector Machines[230] (SVMs) trained on the molecular and systems-level properties of the canonical cancer drivers damaged in a cohort of patients. It then ranks damaged genes outside the training set based on how similar their properties are to those of canonical cancer drivers. sysSVM is implemented in R using the e1071 package[231].

The four one-class SVMs use different kernels, which control how each one learns from the training set. A kernel k measures how similar the features of two genes are. Let x and y denote the features of two genes. Then an SVM measures their similarity as $k(x, y)$. The kernels used in sysSVM are:

- Linear: $k(x, y) = x \cdot y$
- Polynomial: $k(x, y) = (x \cdot y)^d$
- Radial: $k(x, y) = \exp -\gamma|x - y|^2$
- Sigmoid: $k(x, y) = \tanh\,(\gamma\, x \cdot y)$

Here $x \cdot y$ denotes the standard dot product, d and γ are parameters, and $\tanh$ is the hyperbolic tangent function. sysSVM combines the outputs of the four kernels into a single score to use for prediction. The sysSVM algorithm consists of three stages: feature mapping; model selection; and training and prediction.

Step 1. Feature mapping

In feature mapping, molecular and systems-level properties are mapped to the damaged genes of the training cohort. These include seven molecular features (relating to mutation and copy number status) and 19 systems-level features. An additional six systems-level features are available but are excluded from the model by default, since their inclusion was seen to worsen performance. The full list of features is given in Supplementary Table 1.

Step 2. Model selection

Model parameters are then selected. The four one-class SVMs are controlled by certain parameters, and a grid search is implemented to select the best parameters for each kernel separately. These parameters and their default grid ranges are:

- Nu (ν, all kernels): represents an upper bound on the proportion of the training set that can be classed as outliers. Values range from 0.05 to 0.35 in steps of 0.05.

- Gamma (γ, radial and sigmoid kernels): controls the level of influence of individual training points on the model. Values assessed are $\gamma = 2^x$, where $x \in \{-7, -6, \dots, 4\}$
- Degree (d, polynomial kernel): the degree of the polynomial kernel function, chosen from the set $\{3, 4, 5\}$.

Kernel parameters are tuned separately for each kernel. Therefore, the total number of kernel parameter combinations to be assessed is 7 (linear) + 7x12 (radial) + 7x12 (sigmoid) + 7x3 (polynomial) = 196.

In the first implementation of sysSVM[27] an additional parameter was tuned: γ in the polynomial kernel. In this setting, the polynomial kernel had the form $k(x, y) = (\gamma\, x \cdot y)^d$. However, in this case γ simply controls an overall scaling of the kernel, since $(\gamma\, x \cdot y)^d = \gamma^d (x \cdot y)^d$. Constant scalings such as this do not change the behaviour of SVMs, and so the γ parameter is redundant. Thus, in sysSVM2 γ is fixed to 1 for the polynomial kernel by default.

In sysSVM2, model selection was updated from the original sysSVM formulation to improve convergence. A parameter grid search was carried out for each kernel separately, for a total of 196 kernel-parameter combinations. The aim was to select parameters that resulted in a model with a high sensitivity and stability (low standard deviation of sensitivity). The model sensitivity for each parameter combination was assessed on the simulated training set using three-fold cross-validation for 5,000 iterations. For each kernel k and parameter combination i, the mean μ_{ki} and standard deviation σ_{ki} of the sensitivity were calculated across the cross-validation iterations. These were then converted into z-scores $z_{ki}^{(\mu)}$ and $z_{ki}^{(\sigma)}$, which measured the relative values of mean and standard deviation between the different parameter combinations such that:

$$\sum_i z_{ki}^{(\mu)} = \sum_i z_{ki}^{(\sigma)} = 0$$

and

$$\mathrm{Variance}_i\left(z_{ki}^{(\mu)}\right) = \mathrm{Variance}_i\left(z_{ki}^{(\sigma)}\right) = 1.$$

Finally, we defined the Δ_z score as:

$$\Delta_z = z_{ki}^{(\mu)} - z_{ki}^{(\sigma)}.$$

High Δ_z scores corresponded to parameter combinations that had high mean sensitivity and low standard deviation relative to the other combinations for that kernel.

The four parameter combinations (one per kernel) with the highest Δ_z scores were selected and used to train the four kernels on the entire training set.

When measuring the performance of various implementations of sysSVM using the Area Under the Curve (AUC) as in Figure 5-4A-D, we used a reduced range of parameter combinations. This assessment was based on the performance of the full model combining all four kernels, and thus the full range would have been prohibitively costly ($1{,}037{,}232$ parameter combinations, as described in Section 5.3.3). Instead, we considered $\nu \in \{0.05, 2\}$, $\gamma = 2^x$ where $x \in \{-7, 3, 0, 4\}$, and $d \in \{2, 4\}$. This resulted in a total of $(2) \times (2 \times 4) \times (2 \times 4) \times (2 \times 2) = 512$ parameter combinations to assess, where the brackets enclose the number of parameters for the linear, radial, sigmoid and polynomial kernels, respectively. We chose these ranges to provide a sparse coverage of the parameter grid used in the standard model selection step of sysSVM.

Step 3. Training and prediction

Once the parameters for each kernel have been selected, the four SVMs are trained using the entire training set of canonical drivers.

The trained sysSVM model can then be used for prediction in individual samples. To combine the outputs of the four kernels, a combined score S_{gs} is calculated for each gene g in sample s. S_{gs} measures the similarity of the features of gene g to those of the training set. It combines the rank of g in sample s according to each of the four kernels, in such a way that the final score is normalised between 0 and 1. High ranks in each kernel are given exponential weighting and the kernels are weighted according to their sensitivity, with more sensitive kernels contributing more to the score. If R_{kgs} is the rank of g in sample s according to the decision value of kernel k, N_s is the total number of damaged genes in sample s and μ_k is the mean sensitivity of kernel k as assessed by cross-validation iterations, then the score is

$$S_{gs} = \frac{\sum_{k=1}^{4}\left(-\log_{10}\left(\frac{R_{kgs}}{N_s}\right) \times \mu_k\right)}{4 \times \log_{10}(N_s)}.$$

5.2.2. Annotation of systems-level properties

Systems-level properties of human genes were obtained as described in Section 4.2. From these properties, 25 systems-level features were derived to be used for gene classification, of which 19 were retained in sysSVM2 after feature selection (Supplementary Table 1).

For each feature, missing values were imputed using the median or mode (for numeric and categorical features, respectively) of available data for canonical drivers and the rest of genes separately. All features used in the model significantly differentiated cancer drivers from other human genes (Supplementary Table 1).

Systems-level properties were encoded in sysSVM and sysSVM2 as either continuous or binary features, depending on their nature. To account for the fact that some SVM kernels learn more efficiently from binary features[27] and to more comprehensively describe the underlying properties, additional binary features were derived for PPIN properties as well as for gene and protein expression. For PPIN features, hubs and central proteins were defined as those in the top 25% of degree and betweenness distributions, respectively. These thresholds were chosen because they marked out the high tails of the continuous distributions while describing sufficiently large numbers of proteins to be informative (Figures 5-1A, B). For gene and protein expression, the distributions of tissues expressing the gene/protein are broken down into four distinct sections (Figures 5-1C, D). For protein expression, two binary features (9-34 tissues, and 35-40 tissues) were not statistically different between cancer proteins and the rest of proteins. All systems-level features considered for the optimisation of sysSVM for pan-cancer use differed significantly either between canonical drivers, oncogenes or tumour suppressor genes and the rest of genes. This resulted in 25 systems-level features (Supplementary Table 1), which underwent further feature selection to obtain the 19 final features used in sysSVM2 (Table 5-1).

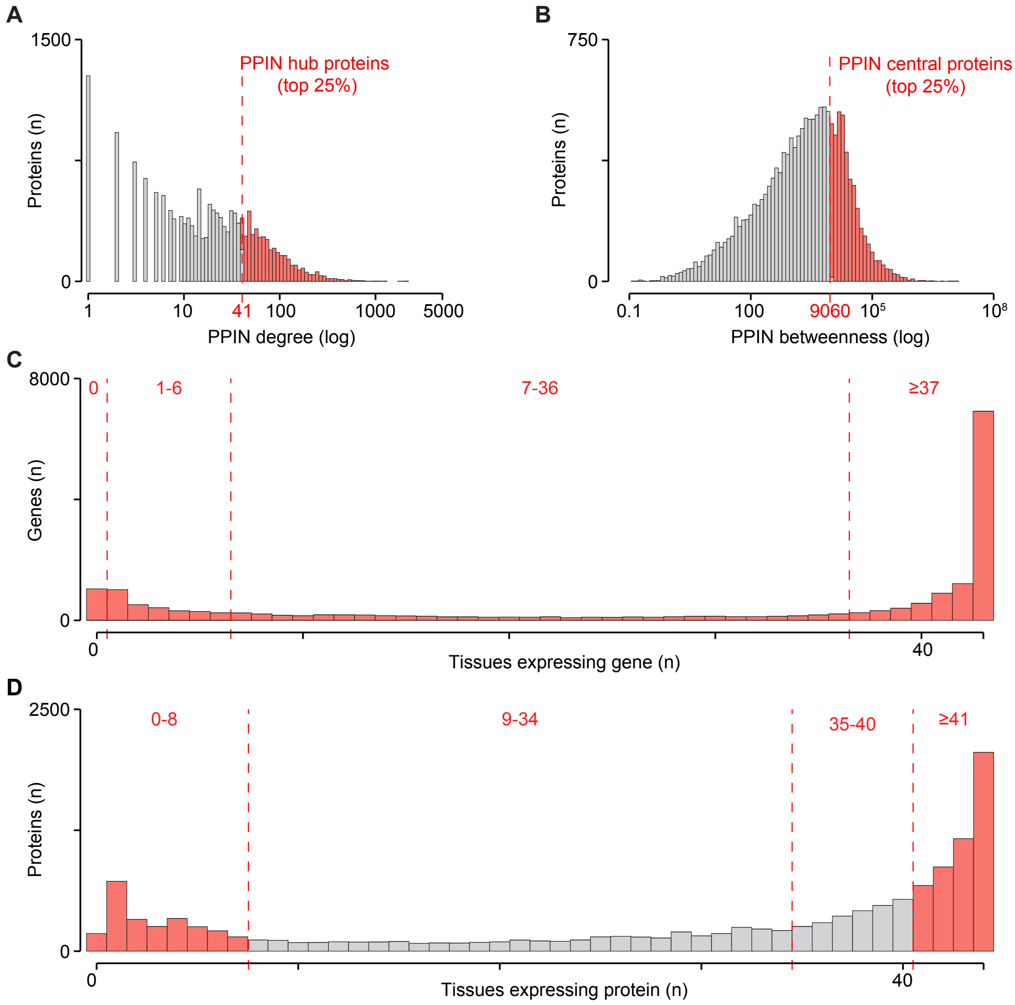

Figure 5-1: Construction of binary features for PPIN and tissue expression properties

Distributions of protein-protein interaction network (PPIN) degree (**A**) and betweenness (**B**) for 16,322 human proteins. Proteins in the top 25% for degree and betweenness values were designated as hubs and central proteins, respectively.

Distributions of the number of tissues expressing 18,641 genes (**C**) and 13,001 proteins (**D**). Both distributions were divided into four sections based on their shape to derive binary expression features.

Features highlighted in red were statistically different between canonical drivers and the rest of genes (Supplementary Table 1).

5.2.3. Pan-cancer TCGA data annotation and simulation

Sequence mutations (SNVs and indels) for 9,079 samples were obtained from the MC3 release of TCGA[15] and annotated with ANNOVAR[74] (downloaded April 2018) and dbNSFP v3.0[77]. Only mutations identified as exonic or splicing were retained.

Damaging mutations included (1) truncating (stopgain, stoploss, frameshift) mutations; (2) missense mutations predicted by at least five out of seven functional prediction methods (SIFT[80], PolyPhen-2 HDIV[81], PolyPhen-2 HVAR, MutationTaster[82], MutationAssessor[83], LRT[84] and FATHMM[85]), at least two out of three conservation-based methods (PhyloP[86], GERP++RS[88] and SiPhy[87]); (3) splicing mutations predicted by one of two splicing-specific methods (ADA[77] and RF) and (4) hotspot mutations identified by OncodriveCLUST[219] v1.0.0.

Copy number data (array intensities) for 11,379 samples were obtained from the Genomic Data Commons portal (https://portal.gdc.cancer.gov/). Copy Number Variant (CNV) segments, sample ploidy and sample purity values were obtained using ASCAT[130] v2.5.2, with 9,873 samples passing quality control. Segments were intersected with the exonic coordinates of 19,549 human genes that were derived as previously described[24] in the reference genome hg38, and converted to hg19 coordinates using the UCSC liftOver tool (http://genome.ucsc.edu/). Genes were considered to have undergone a CNV if at least 25% of their transcribed length was covered by a segment. RNA-Seq data were used to filter out false positive CNVs. Fragments per kilobase million (FPKM) values were obtained for 10,974 samples and only samples with matched copy number and expression data were retained. Damaging CNVs included homozygous gene losses (copy number 0 and FPKM <1 over mean cancer type purity) and gene amplifications (copy number >2 x sample ploidy).

Considering only one sample per patient, a total of 7,630 samples had matched mutation, CNV and RNA-Seq data that passed ASCAT quality controls and had at least one damaged gene. These comprised 535,615 damaging mutations and 2,041,598 damaging CNVs in 18,784 genes (Supplementary Table 2).

To simulate samples that reproduced the molecular features of real TCGA samples, the damaging mutation burden and ploidy were measured for the whole TCGA cohort. Then, 1,000 random combinations of damaging mutation burden and ploidy were extracted and assigned to the simulated samples. To preserve the same frequency of damaging alterations, damaged genes were extracted from within TCGA samples that had similar values of damaging mutation burden (+/-10%, for mutations) and ploidy (+/-0.1, for CNVs) and assigned to the simulated samples. Overall, the final simulated dataset of 1,000 samples contained 69,269 damaging mutations and 252,409 damaging CNVs in 18,455 genes.

As a training set, 220 tumour suppressors with 2,433 loss-of-function alterations (truncating mutations, missense or splicing damaging mutations, homozygous deletions, or double hits) and 236 oncogenes with 5,539 gain-of-function alterations (hotspot mutations, missense or splicing damaging mutations or gene amplifications) were retained. For *TP53* both loss- and gain-of-function alterations (n=352) were considered to be driver events and included in the training set. 301,103 damaging alterations in the remaining 17,998 human genes not present in the training set were used for prediction (Supplementary Table 2). Somatic alterations in tumour suppressors and oncogenes that were not of the appropriate types (i.e. gain-of-function in tumour suppressors and loss-of-function in oncogenes) were discarded from further analysis.

5.2.4. Performance assessment

The performances of all driver prediction models tested in sysSVM2 were measured using five metrics: Area Under the Curve (AUC); composition score; Observed/Expected (O/E) ratios; Rank-Biased Overlap (RBO) score and overlap of the top five predictions between models.

For the AUC, Receiver Operating Characteristic (ROC) curves were derived for each sample individually by comparing the ranks of canonical drivers not used for training to known false positive genes and to the rest of human genes. For both of these comparisons, the median AUC was then measured across samples. The median ROC curve across the cohort was also derived by calculating the median true positive rate for each value of a false positive rate.

The composition score assessed the top five predictions in each sample, and measured the prevalence and ranks of different types of genes. The score *S* was calculated as a weighted sum according to the following formula:

$$S = \sum_{g=1}^{5} w_g \times t_g.$$

The weight w_g of each gene *g* in the top five was such that higher-ranked genes were assigned greater weight. Specifically, $w_g = 6 - r_g$ where r_g is the rank of gene g (with 1 being the highest). The type contribution t_g of gene g was defined for different gene categories as follows: cancer-specific canonical drivers ($t_g = 3$); other canonical drivers (2); cancer-specific candidate genes (1.5); other candidate cancer genes (1); false positives (-1); other genes (0).

Ratios between observed and expected numbers of canonical drivers and false positives in the top five predictions (O/E ratios) were calculated as follows:

$$\text{Canonical driver O/E ratio} = \frac{\text{Canonical drivers in top 5 genes}}{\text{Total canonical drivers in sample}} \times \frac{\text{Total damaged genes in sample}}{5}$$

$$\text{False positive O/E ratio} = \frac{\text{False positives in top 5 genes}}{\text{Total false positives in sample}} \times \frac{\text{Total damaged genes in sample}}{5}$$

These formulas accounted for the fact that the difficulty of ranking canonical drivers (or false positives) in the top five predictions depends on how many damaged genes there are in each sample. The percentages of samples where the observed number of canonical drivers in the top five predictions was higher than twice, five and ten times the expected numbers, and where the observed number of false positives was lower than expected, were calculated.

The RBO score[232] was used to assess the similarity of the top five predictions from pairs of models. It measures the overlap of ranked lists at incrementally increasing depths using a convergent series. Including a correction for finite lists of length five, it was calculated according to the formula:

$$\text{RBO} = \frac{1-p}{1-p^5} \sum_{d=1}^{5} p^{d-1} \times A_d.$$

where *p* was set to 0.9[232], *d* indicates depth in the rankings (starting from the top-ranked elements), and A_d is the overlap of the two lists, restricted to depth *d*.

5.2.5. TCGA sample analysis

Starting from the TCGA samples with matched mutation, copy number and RNA-Seq data, more stringent filters were applied for homozygous deletions and gene amplifications analysis of real samples (Sections 5.4.3 and 5.4.4). Genes with CN=0 that either (1) had one or more mutations in their sequence, or (2) were expressed at greater than 1 FPKM over sample purity (as opposed to mean cancer type purity) were re-annotated as having CN=1. Amplifications were filtered by applying Wilcoxon tests to compare expression levels between amplified and non-amplified cases, for each gene separately. Only genes for which amplification was associated with overexpression at FDR <0.05 were retained as having damaging amplifications. The resulting dataset comprised 7,651 samples at least one damaging alteration. To prevent sysSVM2 training sets from being biased by individual samples, five samples each accounting for more than 50% of the damaged genes in their respective cancer types were removed, leaving 7,646 samples for further analysis. For the purposes of measuring performance and overlap of the cancer-specific and pan-cancer sysSVM2 settings, 752 samples with fewer than ten damaged genes were removed.

To tune the kernel parameters of sysSVM2, 10,000 iterations of three-fold cross-validation were run in each cancer type. In cases where parameters did not converge, an additional 15,000 iterations were run only for the non-converging parameters, with all other parameters fixed at their converged values. This was done for the gamma parameter of the sigmoid kernel in adrenocortical carcinoma (ACC), oesophageal adenocarcinoma (OAC), colon adenocarcinoma (COAD), kidney renal clear cell carcinoma (KIRC) and stomach adenocarcinoma (STAD), as well as for the degree parameter of the polynomial kernel in lung squamous cell carcinoma (LUSC). Predictions were assessed using the same metrics as for simulated data (AUC, composition score, RBO score and overlap of top-five predictions).

Lists of cancer driver genes in individual TCGA samples were identified using a top-up procedure as follows. First, a list of canonical driver genes for each of the 34 cancer types in TCGA was obtained from NCG[24]. Cancer-specific canonical drivers damaged in each sample were considered as cancer drivers for that sample. In samples with five or more such drivers, no further prediction was done. Otherwise, the highest-ranked genes from sysSVM2 were added so that there were five drivers in total. Samples with five or fewer damaged genes overall were not considered for the pan-cancer analysis. For the purpose of comparing sysSVM2 to other driver detection methods on gastro-intestinal cancer data one sample with less than five damaged genes (TCGA-FP-8210, stomach cancer) was included for completeness.

sysSVM2 predictions in 657 gastro-intestinal (GI) cancer samples were compared to those of PanSoftware[15]; dNdScv[217]; OncoIMPACT[225]; and DriverNet[226]. Forty genes identified as GI cancer drivers by PanSoftware was taken from the original publication[15]. These genes were considered as drivers in every sample in which they were damaged. dNdScv was implemented with default parameters, taking all mutations as input. Genes were considered to exhibit significant signs of positive selection if they had a FDR <0.05. OncoIMPACT was implemented with default parameters, using the gene network provided. Nonsilent mutations were used as input for mutations. Copy number amplifications were filtered as described above, and copy number losses included heterozygous and homozygous deletions. In each sample, all genes that shared the highest score were taken as the predicted drivers. DriverNet was implemented with default parameters, using the same gene network and data inputs as for OncoIMPACT. The events identified in each sample were taken as the predicted drivers.

Human pathways for gene set enrichment analysis were obtained from Reactome v72[107]. Before testing, pathways were restricted to those of level 2 or higher, and with between 10 and 500 genes (total of 1,429 pathways containing 10,178 genes). In each cancer type separately, the unique set of top-up predictions from sysSVM2 was tested for enrichment in pathways containing at least one prediction, using one-sided hypergeometric tests. The resulting p-values across all cancer types were corrected for False Discovery Rate (FDR) using the Benjamini-Hochberg method[89], as a single set.

5.2.6. Annotation of PCAWG osteosarcoma data

Variant call VCF files (SNVs, indels) and BED files (copy number segments) were downloaded for 36 PCAWG osteosarcoma samples from the ICGC Data Portal (https://dcc.icgc.org/). Variants were annotated as described for TCGA samples, except for gene amplifications that could not be filtered based on overexpression since matched gene expression data were unavailable. Instead, a more stringent threshold of copy number >2.5 x sample ploidy was used. This resulted in a total of 4,969 damaged genes across the cohort, comprising 3,270 unique genes (Supplementary Table 2).

5.2.7. Autoencoder implementations

Both the variational autoencoder (VAE) and the augmented autoencoder (AAE) were implemented in Python using the Tensorflow package[233] and the Keras Functional API (GitHub repository https://github.com/fchollet/keras).

The same reference simulated cohort used for sysSVM2 development was used for VAE training and prediction. The VAE was composed of an encoder, a bottleneck (Reduced Dimensionality Representation, RDR), and a decoder. The encoder consisted of three fully-connected intermediate layers, before a RDR with three dimensions. This architecture was repeated in an inverted form for the decoder, giving a total of 796 trainable parameters. All layers were connected using the hyperbolic tangent activation function for smoothly nonlinear behaviour.

During training, the accuracy of feature reconstruction was assessed using the Mean Squared Error (MSE). Each of the features was first normalised to range between -0.8 and 0.8 to be compatible with the tanh activation function. Then, the MSE was measured as:

$$\mathrm{MSE} = \frac{1}{26}\sum_{g}\sum_{f=1}^{26}\left(y_{gf} - x_{gf}\right)^2$$

where y_{gf} and x_{gf} are the real and reconstructed values for feature f in gene g.

To minimise the MSE for canonical drivers, the VAE was trained using backpropagation for 10,000 epoch iterations. The resulting trained model was then used for prediction on the 18,471 remaining damaged genes. In this case, the MSE was used to measure the similarity of the features of genes in the prediction set to those of canonical drivers and genes with low MSE were more similar to canonical drivers. Since VAEs are stochastic models, MSE values were averaged over 100 prediction iterations. The final score S_g was calculated as:

$$S_g = -\frac{1}{100}\sum_{i=1}^{100}\frac{1}{26}\sum_{f=1}^{26}\left(y_{gf} - x_{gf}^{(i)}\right)^2$$

where i labels the prediction iteration.

An AAE was developed to combine sysSVM2 with NN and incorporate additional samples into a pre-trained model. The AAE consisted of a standard auto-encoder and an extension that produced an additional output for the gene score (Figure 5-11A). The aim of the AAE was to reconstruct the sysSVM2 gene scores for initial samples and predict scores for additional samples, while learning from the features of both cohorts. Thus, unlike the VAE, the AAE was trained on all damaged genes, not only canonical drivers. Because of the substantially larger training set, the AAE had a larger number of trainable parameters and a higher learning capacity than the VAE. To reflect this, the encoder consisted of a linear pre-layer and four fully-connected intermediate layers. The RDR had seven dimensions, and the decoder had the inverted architecture of the encoder minus the linear pre-layer, for a total of 2,671 trainable parameters. The gene score was derived from the RDR through an additional three fully-connected layers.

The AAE was trained using molecular and systems-level features of all genes in initial and additional samples. During training, the accuracy of feature reconstruction was assessed with MSE. The accuracy of gene score reconstruction was measured with the sum of MSE and Mean Absolute Error, to penalise large numbers of small errors. The AAE was trained using backpropagation for 10,000 iterations. Gene scores of additional samples were predicted directly by the AAE from molecular and systems-level features.

5.3. Developing a systems-level approach for patient-level driver detection

In this section we introduce and further develop our method for patient-level driver detection, sysSVM. sysSVM was initially developed for use in oesophageal adenocarcinoma[27], and here we optimise it for pan-cancer use. We first describe the rationale behind sysSVM and give an overview of the algorithm (5.3.1). We then construct a cancer-agnostic simulated dataset to use for optimisation (5.3.2). Finally, we use this data to optimise sysSVM for pan-cancer use in terms of data normalisation, parameter tuning and feature selection (5.3.3).

5.3.1. The systems-level support vector machine (sysSVM)

We saw in Chapter 4 that driver genes, and particularly canonical driver genes, are distinguished from other human genes by an array of systems-level properties (Figures 4-2 and 4-3). The sysSVM approach to driver detection prioritises somatically damaged genes in individual samples with features similar to those of canonical cancer drivers[27]. In addition to the systems-level properties discussed in Sections 4.1 and 4.4, driver genes are also characterised by molecular properties. In sysSVM, both types of property are encoded as continuous or binary features for a Support Vector Machine (SVM) classifier (Methods 5.2.1).

For the optimisation of sysSVM for pan-cancer use, systems-level properties of genes are encoded by 25 features (Supplementary Table 1, Methods 5.3.2). Molecular properties describe the somatic alterations that genes undergo in cancer tumours (sequence and copy number alterations). Molecular properties change in each analysed cohort because they derive from patient cancer sequencing data. The molecular properties used in optimising sysSVM for pan-cancer use are encoded by five continuous features (total exonic mutational load, non-truncating damaging mutations, truncating mutations, hotspot mutations, gene copy number) and two binary features that indicate whether the gene is amplified or deleted (Supplementary Table 1). Details on how damaging mutations, gene amplifications and gene deletions are derived from cancer sequencing data are provided in Methods 5.2.3.

To leverage the systems-level and molecular properties of canonical drivers, sysSVM first identifies a set of true positive canonical drivers damaged within a cohort of patients (Figure 5-2). It then uses the features of this positive set to train one-class SVMs based on four kernels (linear, radial, sigmoid, polynomial). Finally, it ranks the remaining damaged genes in individual cancer patients with a combined score that weights the kernels based on their sensitivity (Methods 5.2.1). Highly ranked genes have the most similar properties to those of canonical drivers and will be then considered the cancer drivers for that patient. We use one-class SVMs for sysSVM because, while canonical drivers represent a reliable set of true positives, identifying a true negative set of non-cancer genes is not possible. For example, possible negative genes could be known false positives of driver gene detection methods[26,44]. However, these genes are representative of false positives rather than true negatives, so training a classifier on them is likely to introduce unwanted bias. A one-class support vector machine for novelty detection is therefore an optimal way to solve this issue.

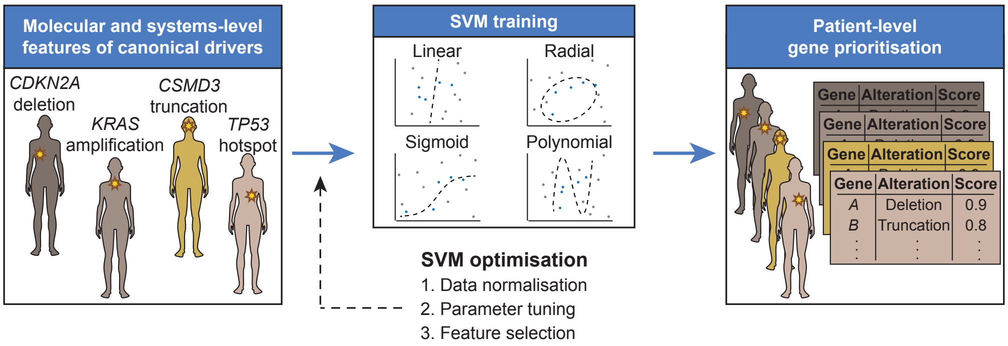

Figure 5-2: Overview of sysSVM

Molecular (somatic SNVs, indels and mutation burden) and systems-level features (Supplementary Table 1) of damaged canonical drivers in the analysed samples are used for training. The best models of support vector machines (SVMs) with four kernels are selected using cross-validation and trained on the whole set of damaged canonical drivers. Finally, a combined weighted score is used to prioritise driver genes in individual patients. The SVM implementation was generalised for optimal performance on a simulated cancer-agnostic dataset through data normalisation, parameter tuning and feature selection.

5.3.2. Preparing a cancer-agnostic simulated dataset

In order to optimise the use of sysSVM for any cancer type in a controlled an unbiased way, we simulated 1,000 cancer-agnostic samples starting from all TCGA tumours with matched mutation, CNV and gene expression data (Methods 5.2.3). We ensured that the tumour mutation and copy number burdens were similar between real and simulated samples (Figure 5-3A). The frequency of damaging alterations in known oncogenes and tumour suppressors was comparable between the two datasets, with *TP53*, *PIK3CA* and *CDKN2A* among the most frequently altered genes in both (Figure 5-3B). We further verified that the molecular features of individual damaged genes did not deviate significantly between the real TCGA and simulated reference cohorts Figure 5-3C). Finally, we found that gene alteration frequencies in the simulated data were not significantly biased by cancer types with large cohort sizes in TCGA (Figure 5-3D), confirming the suitability of the simulated data as a representative pan-cancer cohort.

The simulated cohort for sysSVM pan-cancer optimisation (hereafter referred to as the reference cohort) was composed of 1,000 samples with 18,455 genes damaged 309,427 times. Of these, 686 were canonical drivers with an experimentally proven role in cancer[25,111], 1,605 were candidate cancer genes from 273 cancer screens[24], 43 were known false positive predictions of driver detection methods[44,186] and 16,121 were the remaining damaged genes (hereafter referred to as the rest of genes; Figure 5-3E, Supplementary Table 2). We annotated the seven molecular and 25 systems-level features of all damaged genes (Supplementary Table 1) and used these features for training and prediction. As a training set, we selected 457 of the 686 canonical drivers with proven roles as oncogenes (236) or tumour suppressors (221). We restricted somatic alterations of oncogenes and tumour suppressors to gain-of-function or loss-of-function alterations, respectively (Methods 5.2.3). Since we could not reliably define the remaining 229 damaged canonical drivers as either oncogenes or tumour suppressors, we could not restrict their somatic alterations to the appropriate type. Therefore, we did not use them for training but could still use them for prediction and performance assessment (Figure 5-3E), together with 43 false positives and 16,121 the rest of genes.

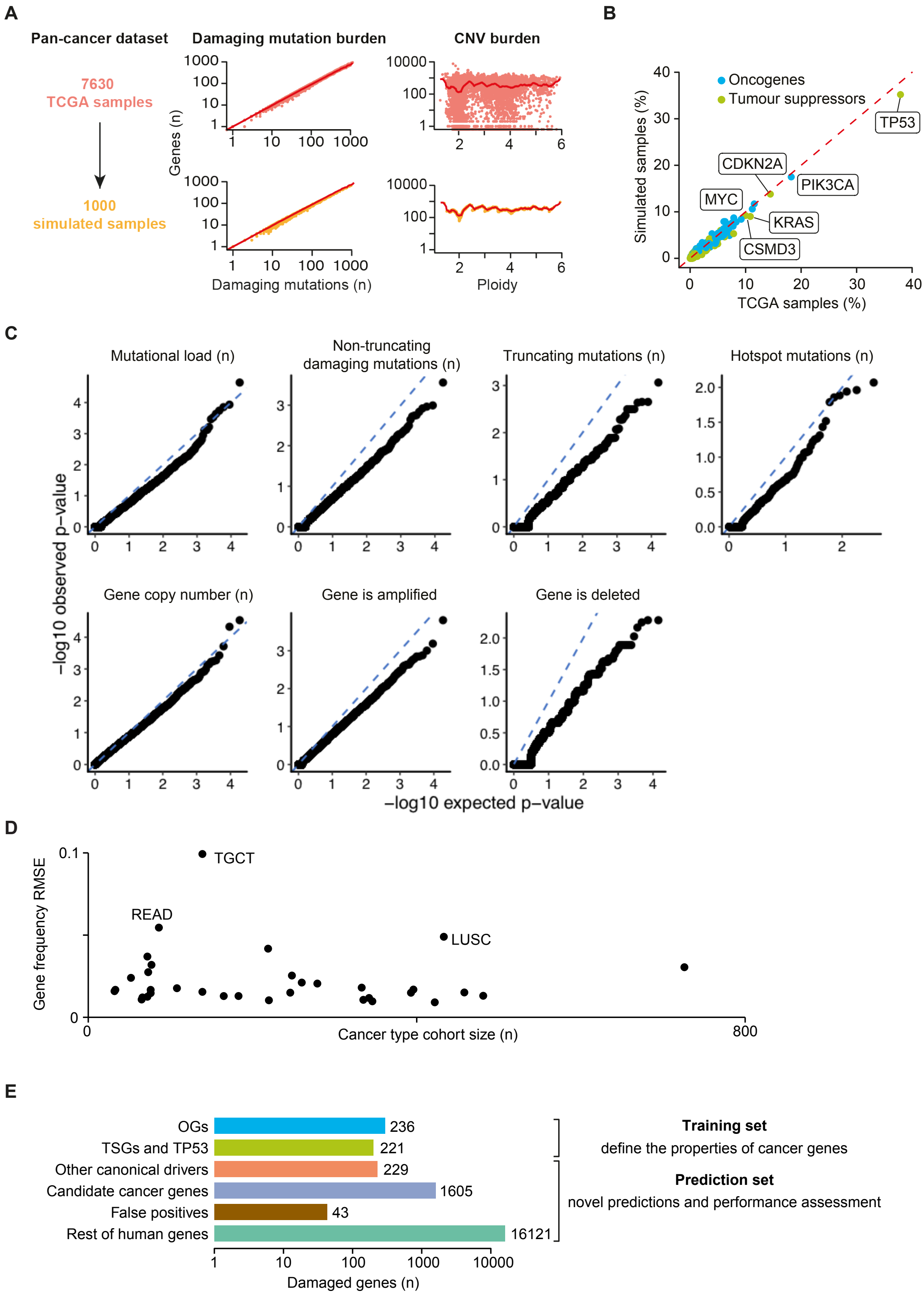

Figure 5-3: Constructing a simulated pan-cancer reference cohort

A. Generation of a simulated reference cohort from TCGA data. Values of damaging mutation burden and ploidy were randomly assigned to samples. Damaged genes were then extracted from real samples with similar values of damaging mutation burden (+/-10% for mutations) and ploidy (+/-0.1 for CNVs). Dots represent individual TCGA (orange) or simulated (yellow) samples. Red lines indicate average numbers of genes with damaging mutations or CNVs in TCGA samples, for each given values of damaging mutation burden or ploidy.

B. Frequencies of canonical drivers in real and simulated samples. Oncogene gain-of-function, tumour suppressor loss-of-function and both types of *TP53* alterations were considered.

C. Quantile-quantile (QQ) plots illustrating comparisons of molecular features of genes between TCGA (n=7,630) and simulated reference (n=1,000) cohorts. Features were compared using Poisson tests (mutation features), Wilcoxon tests (copy number) and Fisher tests (amplifications and deletions).

D. For each cancer type in TCGA, the gene alteration frequency profile was calculated as the proportion of samples with a damaging alteration in each gene. This was then compared to the gene alteration frequency profile of the simulated reference cohort of 1,000 samples using the root mean-squared error (RMSE). This was calculated as $RMSE = \sqrt{\frac{1}{19549}\sum_{g=1}^{19549}\left(f_g^{TCGA} - f_g^{sim}\right)^2}$, where f_g^{TCGA} and f_g^{sim} are the frequencies of gene g being damaged in the TCGA and simulated cohorts, and 19549 is the total number of human genes[24]. Higher values of the RMSE indicate greater differences between the simulated reference cohort and the TCGA cohort for that cancer type. Cancer types with RMSE >0.05 are labelled. There was no significant correlation between cohort size and RMSE (Spearman correlation p=0.54, rho=-0.11).

E. Gene sets used for sysSVM optimisation. The training set included oncogenes (OGs) and tumour suppressor genes (TSGs), as well as *TP53*. All other damaged genes were used for prediction and assessment. These included other canonical drivers (without a proven OG or TSG role), candidate cancer genes from published cancer sequencing screens, known false positives of established driver detection methods and the remaining damaged genes. Bars indicate the number of unique damaged genes across the reference cohort of 1,000 simulated samples.

5.3.3. Optimising sysSVM for pan-cancer use

Using the simulated reference cohort, we optimised sysSVM for pan-cancer use in terms of data normalisation, parameter tuning and feature selection (Figure 5-2).

Each kernel in sysSVM (linear, radial, sigmoid, polynomial) is controlled by one or more (hyper-)parameters. We tune these parameters to the training set of canonical drivers by performing a grid search (Methods 5.2.1). One of the challenges of implementing a multi-kernel approach like sysSVM is the combinatorial explosion of possible kernel parameter combinations. For example, in our standard grid search, the linear kernel has 7 parameter combinations, the radial and sigmoid kernels have 84 combinations each, and the polynomial kernel has 21 combinations. If we were to tune these parameter combinations simultaneously for all kernels, we would have to investigate a total of $7 \times 84 \times 84 \times 21 = 1{,}037{,}232$ parameter combinations. This would be prohibitively costly. By tuning parameters separately for each kernel, we only need to assess $7 + 84 + 84 + 21 = 196$ total parameter combinations, which is much more tractable.

However, for the purposes of optimising sysSVM on the pan-cancer simulated dataset, we wanted not to bias our approach with a particular set of pre-tuned kernel parameters. Therefore, we implemented 512 multi-kernel models with parameter combinations representing a sparse coverage of the $1{,}037{,}232$ possible models from a standard grid search (Methods 5.2.1). We then measured the ability of each of these 512 models to prioritise the 229 canonical drivers not used for training over the rest of damaged genes or the false positives. We did this by computing the Area Under the Curve (AUC) in each sample individually, and taking the median AUC as representative of the whole cohort (Methods 5.2.2).

First, we derived the optimal settings for data normalisation in terms of centered and un-centered data. While it is common practice in machine learning approaches to use centered data, there are theoretical considerations that suggest that this might not be optimal for the one-class SVM (Appendix 1). All 512 models robustly prioritised canonical drivers above the rest using either centered or un-centered data, but showed lower performance in distinguishing canonical drivers from false positives (Figure 5-4A). We reasoned that false positives from recurrence-based driver detection methods[44] shared some features with canonical drivers. For example, they encoded long and multi-domain proteins. Removing the protein length and number of protein domains features from the SVMs (Supplementary Table 2) substantially improved performance, particularly for un-centred data (Figure 5-4B). We therefore excluded these features from the model in further analysis.

Second, we selected the optimal sets of parameters in each kernel. Hyper-parameter choice is known to have substantial impacts on classification and it is an open problem for one-class SVMs[234]. Since the parameters for each kernel needed to be tuned separately, we could not use AUC of the combined multi-kernel model for tuning in general. Instead, we used the sensitivity of each kernel to predict canonical drivers calculated from three-fold cross-validation on the training set (Methods 5.2.1). Average sensitivity across kernels was indeed a good predictor of the overall AUC of canonical drivers over the rest of genes (Figure 5-4C) and false positives (Figure 5-4D). We therefore developed an approach to select the parameters that conferred the highest sensitivity for each kernel in multiple iterations of cross-validation (Methods 5.2.1). In the reference cohort, parameters chosen in this way converged within 2,000 cross-validation iterations for all kernels (Figure 5-4D).

Finally, since the presence of highly correlated features can hinder SVM performance[235], we performed systematic feature selection by assessing the pairwise correlations between all 25 systems-level features. Four features (gene expression in 1≤ tissues ≤6 and in ≥37 tissues; protein expression in 0≤ tissues ≤8 and central position in the protein-protein interaction network) exhibited a significant degree of inter-correlation (Pearson |r| >0.5, FDR<0.05, Figure 5-4E). Removing them led to faster convergence of kernel parameters (Figure 5-4D) and improved performance overall (Figure 5-4F).

Based on these results, we chose the default settings for the cancer agnostic SVM classifier, which we named sysSVM2. By default, data are un-centered but scaled to have unit standard deviation. Six of the original systems-level features are excluded resulting in a total of seven molecular and 19 systems-level features (Table 5-1). Finally, kernel parameters optimised on the simulated reference cohort are provided as a default (Figure 5-4D), although users may perform specific cross-validation iterations on their own cohorts.

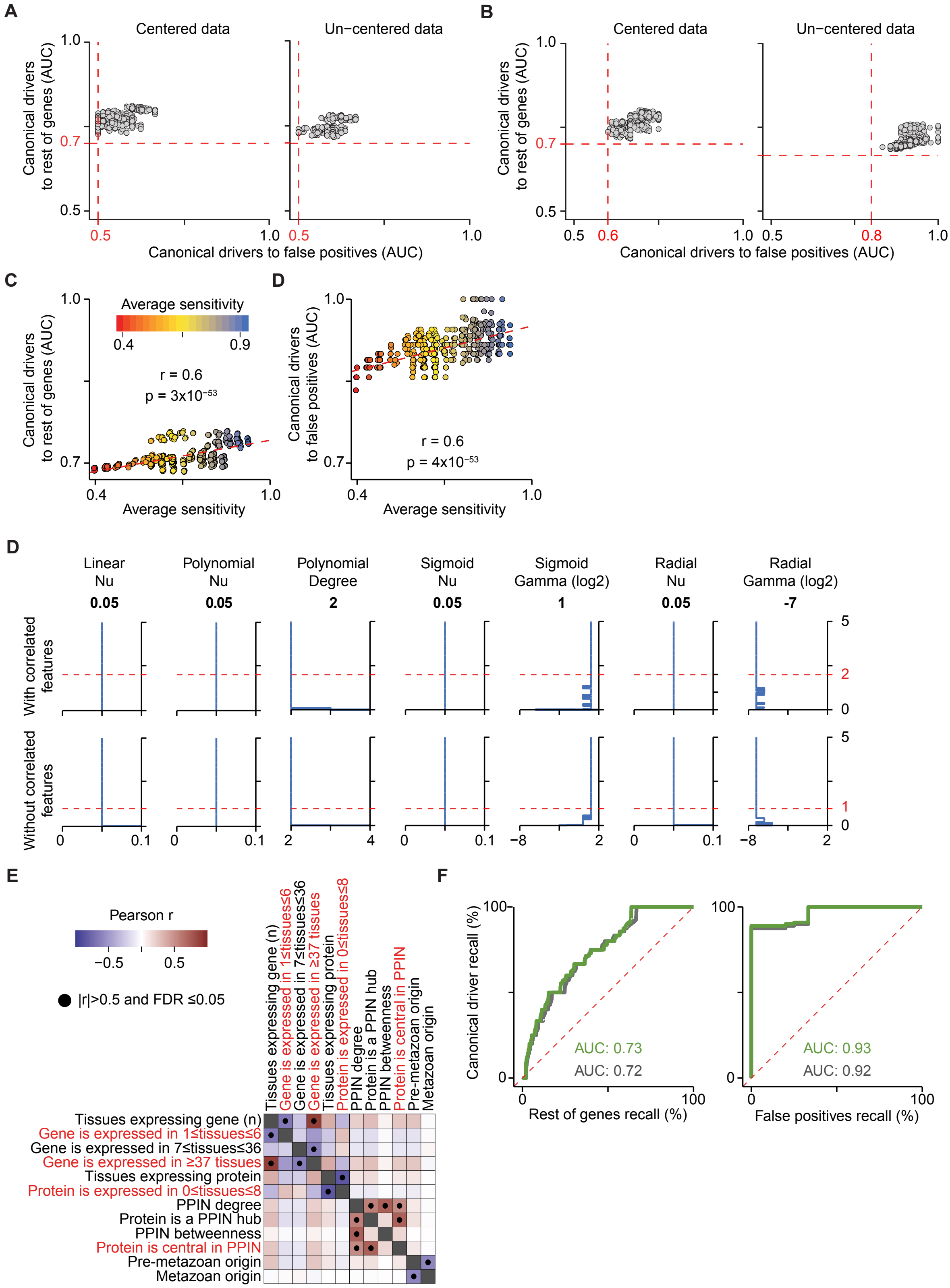

Figure 5-4: Optimisation of sysSVM for pan-cancer use

Model performances on the reference cohort using centered (left) and un-centered (right) data with all 25 systems-level features (**A**) or excluding protein length and number of protein domains (**B**). A sparse grid of 512 parameter combinations in the four kernels was tested. The performance of each model was measured using the Area Under the Curve (AUC), comparing the ranks of canonical drivers to the rest of genes and false positives. Median AUC values across all samples were plotted. Red dotted lines represent the minimum AUC values.

Correlation between model average sensitivity and AUCs of canonical drivers over the rest of genes (**C**) or false positives (**D**). The sensitivity of each kernel was measured on the training set over 100 three-fold cross-validation iterations. The median values over the four kernels are plotted. R and p-values from Pearson's correlation test are reported. Dotted red lines indicate the linear regression curves of best fit.

D. Parameter convergence on simulated data, with (top) and without (middle) correlated features. The four kernels had a total of seven parameters chosen from a grid search. Cumulative parameter selections are indicated over a series of 5,000 cross-validation iterations, but parameter choices converged within 2,000 iterations. The final selected parameter values are indicated at the top and were the same for both settings (with and without correlated features).

E. Correlation between systems-level features. Pearson correlation coefficient was measured between all possible pairs of features reported in Supplementary Table 1 considering all genes with available data. Only features with at least one significant correlation (Pearson |r|>0.5, FDR<0.05) are shown. Three correlated feature pairs (Protein-Protein Interaction Network (PPIN) degree – PPIN hub, PPIN degree – PPIN betweenness, and pre-metazoan and metazoan origin) were retained because they were complementary in their description of gene properties. In particular, PPIN betweenness is a distinct topological property from degree, and hubs included proteins with a large range of degrees (41 to 2,116). The evolutionary origin features instead described the two most common epochs for the origin of human genes (11,624 pre-metazoan, 3,076 metazoan, Supplementary Table 1). Final features removed are in red.

F. Comparison of model performances with (green) and without (grey) correlated features. Median Receiver Operating Characteristic (ROC) curves across samples are plotted, comparing the ranks of canonical drivers to the rest of genes (left) and to false positives (right). The median Areas Under the Curve (AUCs) are also indicated.

Category	Property	Feature	Type
Molecular	Gene mutation	Mutational load (n)	Continuous
		Non-truncating damaging mutations (n)	Continuous
		Truncating mutations (n)	Continuous
		Hotspot mutations (n)	Continuous
	Gene copy number	Gene copy number (n)	Continuous
		Gene is amplified	Binary
		Gene is deleted	Binary
Systems-level	Gene duplication	Gene is duplicated	Binary
		Gene is an ohnolog	Binary
	Gene essentiality	Cell lines in which gene is essential (%)	Continuous
		Gene is essential	Binary
	Gene expression	Tissues expressing gene (n)	Continuous
		Gene is expressed in 0 tissues	Binary
		Gene is expressed in 7≤ tissues≤ 36	Binary
	Protein expression	Tissues expressing protein (n)	Continuous
		Protein is expressed in ≥41 tissues	Binary
	Protein-protein interaction network (PPIN)	PPIN degree	Continuous
		Protein is a PPIN hub	Binary
		PPIN betweenness	Continuous
		PPIN clustering coefficient	Continuous
	Protein complexes	Complexes the protein is part of (n)	Continuous
	miRNA interactions	miRNAs targeting the gene (n)	Continuous
	Gene evolutionary origin	Pre-metazoan origin	Binary
		Metazoan origin	Binary
		Vertebrate origin	Binary

		Post-vertebrate origin	Binary

Table 5-1: Molecular and systems-level properties of genes used to prioritise cancer genes in sysSVM2. Molecular properties describe gene alterations in individual cancer samples. Systems-level properties are global gene properties and are not related to cancer (see also Supplementary Table 1). PPIN: Protein-protein interaction network. miRNA: micro RNA.

5.4. Assessment of sysSVM2 on real and simulated data

Having optimised sysSVM2 for pan-cancer use, we sought to comprehensively assess its performance. In this section, we will start with simulated data, first running sysSVM2 on the simulated reference cohort (5.4.1), and then assessing the effect of training cohort size on performance (5.4.2). We will then benchmark sysSVM2 against existing driver detection methods in gastro-intestinal samples from TCGA, both at the cohort-level and at the sample-level (5.4.3). Having found that sysSVM2 is stable and has a low false positive rate, we will then apply sysSVM2 to all cancer types available in TCGA (5.4.4). Finally, we will validate the utility of sysSVM2 by applying a model trained on pan-cancer data to an external cohort of osteosarcomas, a rare cancer type with few known drivers (5.4.5).

5.4.1. Performance in the simulated reference cohort

We first assessed the performance of sysSVM2 in prioritising cancer drivers over other genes using the same simulated reference cohort on which its performance had been optimised in Section 5.3. We confirmed that, overall, the prediction scores of 229 canonical drivers outside the training set were significantly higher than those of any other gene category (Figure 5-5A). Candidate cancer genes also scored significantly higher than the rest of genes, indicating that they were also in top ranking positions.

We also measured the relative ranks of genes in individual samples using Receiver Operating Characteristic (ROC) curves. Comparing canonical drivers to the rest of genes and to false positives gave AUCs of 0.73 and 0.93, respectively (Figure 5-5B), demonstrating that canonical drivers were prioritised above the rest of genes and especially above false positives. This was not surprising since canonical drivers differ more from false positives than they do from the rest of human genes by their systems-level properties (Figure 5-5C). This observation also supported the decision not to use two-class classification, since known false positives are not representative of non-cancer genes in general.

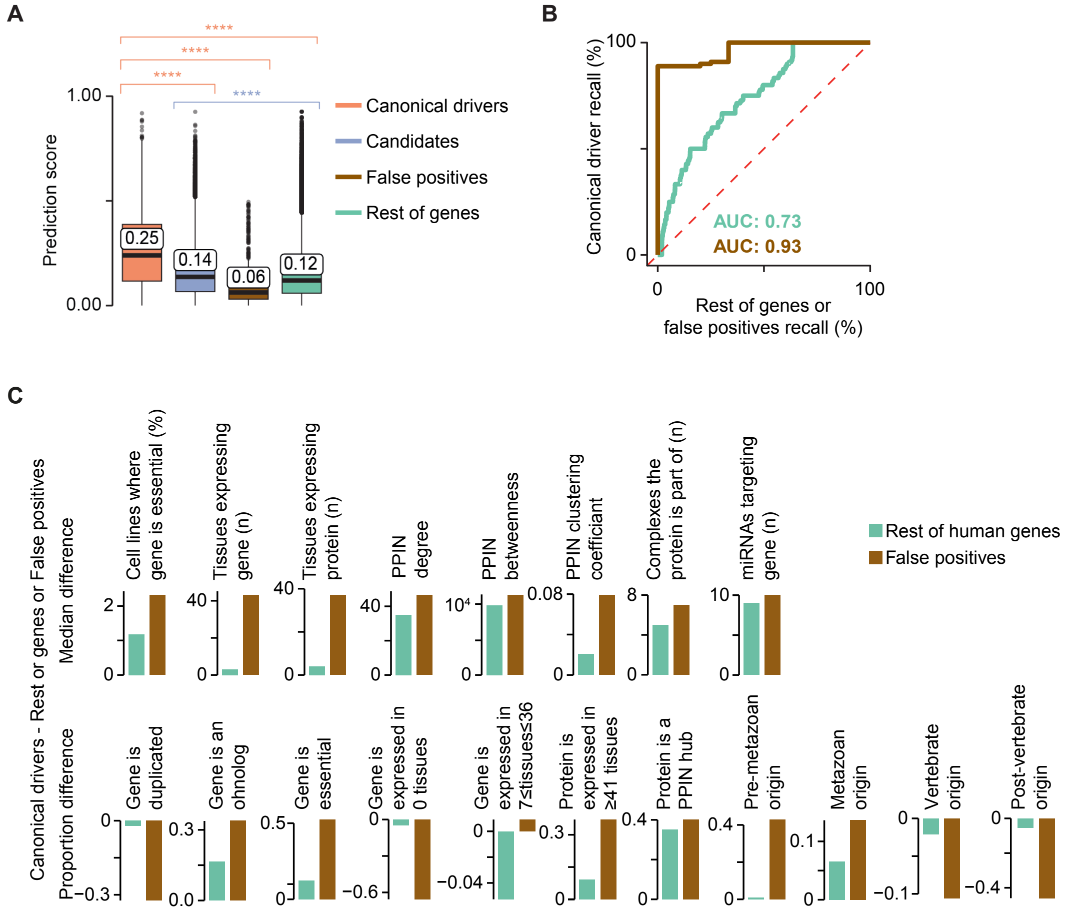

Figure 5-5: Performance of sysSVM2 on the simulated reference cohort

A. Distributions of sysSVM2 prediction scores for different types of damaged genes in the reference cohort. Whiskers extend to 1.5 times the Inter-Quartile Range (IQR). Statistical significance was measured using two-sided Wilcoxon tests. The median values of the distributions are labelled. **** = p <2.2x10^{-16}.

B. Receiver Operating Characteristic (ROC) curves, comparing canonical drivers to the rest of genes (green) and to false positives (brown). Recall rates were calculated for each sample separately and the median ROC curve across samples was plotted. Median Areas Under the Curve (AUCs) for both comparisons are also indicated.

C. Difference in average features between canonical drivers and the rest of human genes (green) and false positives (brown). For each of the three gene sets and each of the 19 systems-level features selected for sysSVM2, either the median value (continuous features, top row) or the proportion of genes for which the feature was positive (binary features, bottom row) was calculated. The difference in these values between canonical drivers and both the rest of genes and false positives is shown for each feature. For all features except Gene is expressed in 7≤tissues≤36, the difference between canonical drivers and false positives is greater than the difference between canonical drivers and rest of genes, while the direction of the difference is the same.

5.4.2. Effect of training cohort size on sysSVM2 performance

The sample size of patient cohorts can highly vary across cancer types. For example, in TCGA it ranges from 32 samples for diffuse large B-cell lymphoma (DLBC) to 726 for breast cancer (BRCA, Supplementary Table 3), with a median of 201 samples. We therefore sought to address how the sample size of the training cohort affected sysSVM2 performance.

Starting from all TCGA samples and using the previously described approach (Methods 5.2.3), we simulated 40 training cohorts, ten of which were composed of ten samples, ten of 100 samples, ten of 200 samples and ten of 1,000 samples. We then trained sysSVM2 on each of these 40 cohorts independently and used the resulting models to rank damaged genes in the reference cohort and compared their performance.

The distributions of AUCs of canonical drivers over the rest of genes or false positives were high for all four cohort sizes (Figure 5-6A). This suggested that sysSVM2 was overall very effective in prioritising cancer genes independently of the training cohort size. We then compared the composition of the prioritised gene list in each sample across models of a given size. We measured a composition score of the top five genes that accounted for the number and position of canonical drivers, candidate cancer genes and false positive genes (Methods 5.2.4). Similar to the AUC, the composition score of the top five genes was also very similar across training cohorts (Figure 5-6B). However, a few models trained on ten or 100 samples returned false positives in the top five positions while no false positives were predicted by models trained on larger cohorts of 200 or 1,000 samples. Finally, we measured the ratio between observed and expected canonical drivers and false positives in the top five genes (Figure 5-6C, Methods 5.2.4). Independently of the training cohort size, false positives in the top five genes were always lower than expected, confirming that sysSVM2 efficiently distinguished false positives from drivers. The number canonical drivers in the top five genes was more than twice the expected number in >85% of samples and more than five times the expected value in around 65% of samples. As with the other metrics, the performance of sysSVM2 did not change substantially with the size of the training cohort.

Since we used the same reference cohort for prediction, we could directly compare the gene ranks in each patient across models, thus assessing their prediction stability. To do so, we measured the Rank-Biased Overlap (RBO) score that compares two ranked lists giving greater weight to the higher-ranked positions[232] (Methods 5.2.4). The distributions of RBO scores of the top five genes were significantly higher for large training cohorts compared to those composed of ten samples (Figure 5-6D). Moreover, models trained on large cohorts showed overall higher gene overlap in the top five genes (Figure 5-6E).

These results showed that, although sysSVM2 successfully separates canonical drivers from other genes independently of the training cohort size, small cohorts lead to occasional false positive predictions and to unstable gene ranking. Since the median cohort size of TCGA cancers is 201 samples, sysSVM2 is likely to separate canonical drivers from the rest of genes with a very low false positive rate and stable gene rankings for most cancer cohorts.

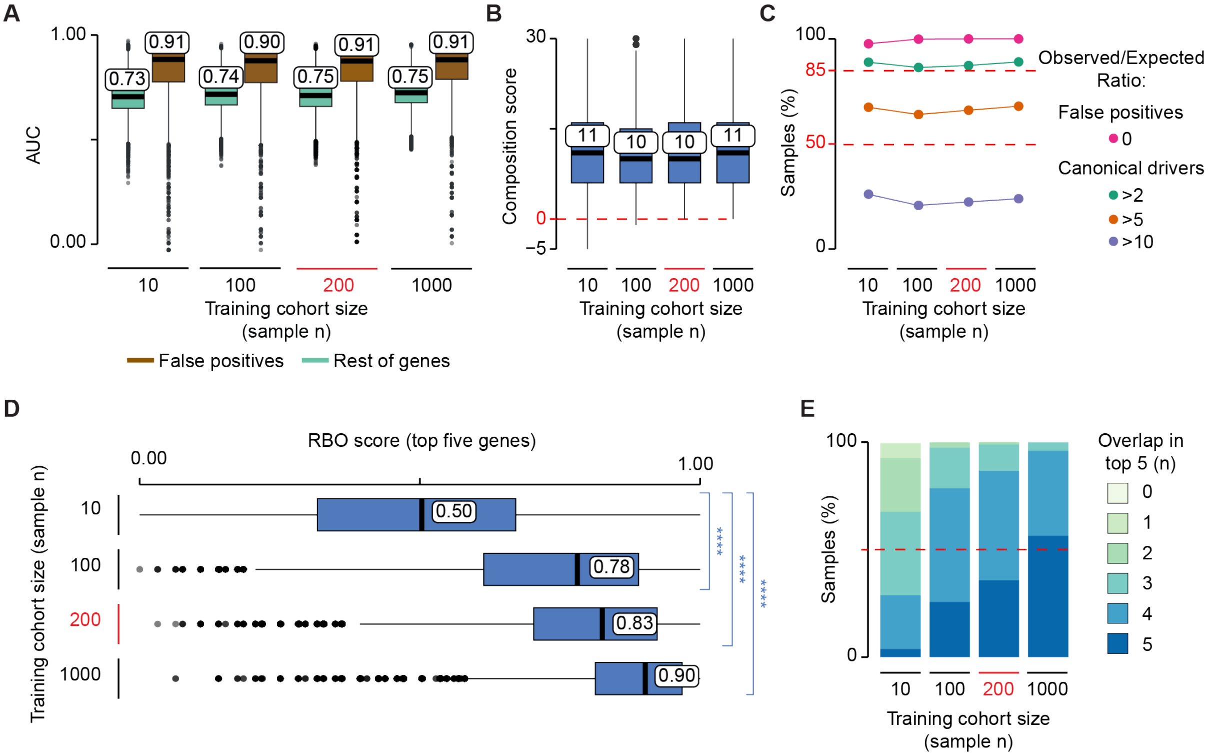

Figure 5-6: Effect of training cohort size on sysSVM2 performance

A. Distributions of AUCs comparing the ranks of canonical drivers to the rest of genes (green) and False Positives (brown). Models were trained on ten simulated cohorts composed of ten, 100, 200 and 1,000, for a total of 40 simulated cohorts. These were then used to predict on the same reference cohort of 1,000 samples. The AUC was measured for each set of predictions in each sample.

B. Distributions of composition scores of the top five predictions in terms of canonical drivers, candidate cancer genes, false positives and rest of genes (Methods 5.2.4). The composition score was measured for each set of predictions in each sample. Six training cohorts of size ten and two cohorts of size 100 gave negative composition scores in at least one sample, indicating highly ranked false positive genes.

C. Ratios between observed and expected numbers of canonical drivers and false positives in the top five predictions (O/E ratios). For each size of the training cohort, the percentages of samples with a false positive O/E ratio of zero and canonical driver O/E ratios greater that 2, 5 and 10 are shown (Methods 5.2.4).

D. Rank-Biased Overlap (RBO) score of the top five predictions in each sample (Methods 5.2.4). RBO scores measured the similarity between the predictions from every possible pair of models trained on cohorts of a particular sample size. Statistical significance was measured using two-sided Wilcoxon tests. **** = p <2.2×10^{-16}.

E. Distribution of the number of top five predictions shared between models trained with the same cohort size. The overlap was calculated between each pair of predictions in each sample.

5.4.3. Benchmark of sysSVM2 against existing methods

Next, we sought to compare the predictions of sysSVM2 on real cancer data to those of other driver detection methods. To do this, we used 657 Gastro-Intestinal (GI) adenocarcinomas from TCGA (73 oesophageal, 279 stomach, 219 colon and 86 rectal cancers, Supplementary Table 3). Overall, this cohort had 17,122 unique damaged genes, including 438 tumour suppressors and oncogenes used for sysSVM2 training (Supplementary Table 2). After ranking the remaining 16,684 damaged genes, we confirmed the overall ability of sysSVM2 to prioritise the 228 canonical drivers not used for training over the rest of damaged genes and false positives also in real data (Figure 5-7A).

To identify the list of cancer drivers in each patient, we adopted a top-up approach. Starting from the GI canonical drivers[24] damaged in each sample, we added sysSVM2 predictions progressively based on their rank to reach five drivers per patient (Methods 5.2.5). This was based on the assumption that each cancer requires at least five driver events to fully develop, in concordance with recent quantifications of the amount of excess mutations arising from positive selection in cancer[16,217]. While 154 patients had damaging alterations in five or more GI canonical drivers, 503 patients (77%) needed at least one prediction (Figure 5-7B), highlighting the need for additional cancer driver predictions. This resulted in 564 unique sysSVM2 drivers.

We then predicted the drivers in the same GI samples using two cohort-level (PanSoftware[15] and dNdScv[217]) and two patient-level (OncoIMPACT[225] and DriverNet[226]) detection methods. PanSoftware integrated 26 computational driver prediction tools and we took the list of 40 damaged drivers directly from the original publication[15], given that we used a large subset (87%) of the same TCGA GI samples. We ran the other three methods with default parameters (Methods 5.2.5) and obtained 25 predicted drivers with dNdScv, 607 with DriverNet and 1,345 with OncoIMPACT.

We compared sysSVM2 to the four other methods in terms of recall rates of canonical drivers or false positives, proportion of novel predictions and patient driver coverage. Overall, cohort-level methods had higher recall rates of GI canonical drivers, fewer novel predictions and a comparably low false positive recall than sysSVM2 (Figure 5-7C). However, unlike sysSVM2, neither cohort-level method predicted drivers in all patients, leaving the vast majority of them with less than five predictions and some with no predictions (Figure 5-7D).

Compared to sysSVM2, the other two patient-level methods had higher recall rates of the 228 canonical drivers, a comparably high proportion of novel predictions but higher false positive rate (Figure 5-7C). Namely, sysSVM2 made only one false positive prediction in one patient while DriverNet and OncoIMPACT predicted four and seven false positives in 124 and 306 patients, respectively (Figure 5-7E). Overall, all three methods had high patient driver coverage, but sysSVM2 outperformed the other two with only one sample where it predicted less than five drivers (Figure 5-7D). Interestingly, the overlap of predictions between sysSVM2 and the other patient-level methods was statistically significant (Figure 5-7E) even when only top-up predictions were considered (Figure 5-7F). This suggested that the majority of predictions converged to the same genes.

These results showed that cohort-level methods have high specificity and sensitivity to identify cancer-specific canonical drivers but often fail to find drivers in a substantial subset of patients. Compared to other patient-level detection methods, sysSVM2 outperforms them in terms of specificity and patient coverage.

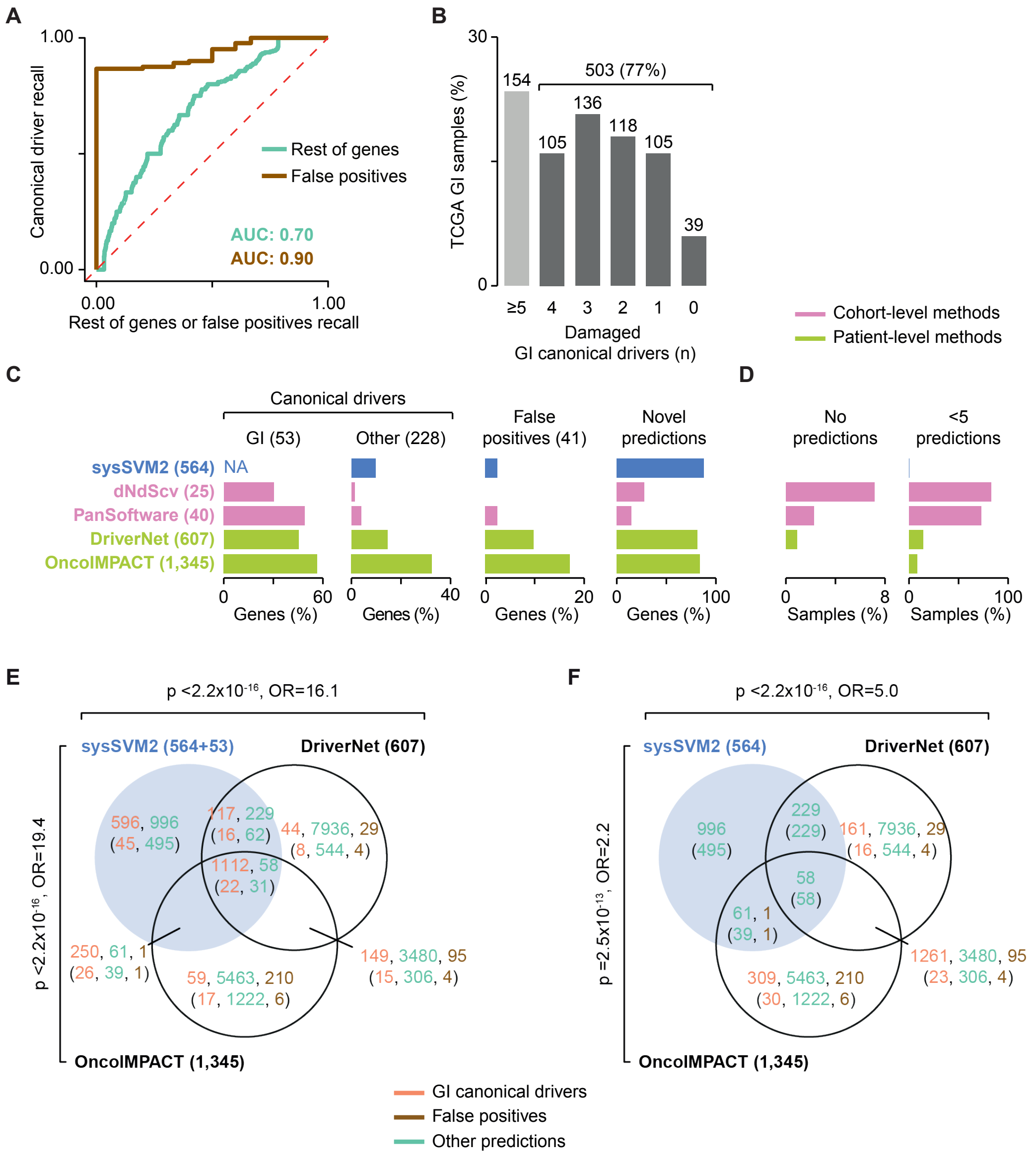

Figure 5-7: sysSVM2 benchmark on TCGA gastro-intestinal cancers

A. Median Receiver Operating Characteristic (ROC) curves across 657 Gastro-Intestinal (GI) samples from TCGA. Curves compare the ranks of canonical drivers to the rest of genes or to false positives. The median Areas Under the Curve (AUCs) are also indicated.

B. Distribution of GI canonical drivers across the GI cohort. Lists of canonical drivers for each GI cancer type were obtained from NCG6[24] and mapped to samples of the corresponding cancer type where they were damaged. Numbers of samples are indicated above each bar. Samples with five or more GI drivers did not require additional driver predictions.

C. Comparison of performance between sysSVM2 and four other driver detection methods. The set of unique drivers predicted by each approach were compared in terms of recall of GI canonical drivers, other canonical drivers (non-GI and outside the sysSVM2 training set) and false positives and proportion of novel predictions not previously associated with a cancer driver role. The number of genes in each category is reported in brackets. The recall of GI canonical drivers could not be assessed for sysSVM2 because these were part of the training set. They were however considered as drivers by default, rather than predicted by the algorithm. NA = Not Applicable.

D. Proportions of 657 GI samples left with no predicted drivers (left) or fewer than 5 predictions. The one sample left with fewer than 5 predictions by sysSVM2 (TCGA-FP-8210, stomach cancer) had four damaged genes overall.

Overlap of driver predictions in individual samples, between sysSVM2, DriverNet[226] and OncoIMPACT[225]. The sysSVM2 predictions were considered both with (**E**) and without (**F**) the GI canonical drivers. The numbers of overlapping patient-level predictions are indicated, along with the number of unique genes in each set in brackets. P-values and Odds Ratios (ORs) for pairwise overlap were calculated using Fisher's exact test, taking into account all damaged genes in all samples.

5.4.4. Patient-level drivers in 34 cancer types

In order to provide a comprehensive resource of trained models and patient-level drivers, we sought to apply sysSVM2 to 7,646 TCGA samples of 34 cancer types with at least one somatically damaged gene (Methods 5.2.5). Training, prediction and assessment of sysSVM2 in these 34 cancer types was conducted by a colleague of mine, Hrvoje Misetic.

To find the best training setting for the algorithm on real cancer samples, we compared the performance of sysSVM2 trained on the whole pan-cancer cohort as well as on the 34 cancer types separately. In the pan-cancer setting, we used all 477 tumour suppressors and oncogenes damaged across the whole cohort. In the cancer-specific setting, we used instead only the subsets of these genes damaged in each cancer type (Supplementary Table 3). We then predicted on the remaining damaged genes and applied the top-up approach as described above, starting from the cancer-specific canonical drivers damaged in each patient (Supplementary Table 3). We found that 6,067 samples (79%) required at least one sysSVM2 prediction in order to reach five drivers (Figure 5-8A). These corresponded to 4,369 and 4,548 unique genes in the pan-cancer and cancer-specific settings, respectively, with a significant overlap of predictions (3,896, $p < 2.2 \times 10^{-16}$, two-sided Fisher's exact test).

We then compared the performance of pan-cancer and cancer-specific settings of sysSVM2 in prioritising canonical drivers over rest of genes or false positives. The AUCs differed significantly (FDR <0.05) and substantially (|difference in medians| >0.05) in only five cancer types (Figure 5-8B). All of them were composed of small cohorts with <200 samples and in all cases the pan-cancer setting showed better performance than the cancer-specific setting. The composition score of the top five predictions also differed significantly and substantially (|difference in medians| >1) in only three cancer types (Figure 5-8C). All these cancer types were again characterised by small training cohorts and showed higher performance in the pan-cancer setting. Predictions of cancer-specific models and the pan-cancer model were mostly similar, with the exception of cancer types with small training cohorts (Figures 5-8D, E). Overall, these results confirmed the trend observed in the simulated data and indicated that the pan-cancer and cancer-specific settings performed similarly well in most cases, except for small cohorts where the pan-cancer model performed better.

Based on these results, we used the pan-cancer setting for cancer types with small cohorts (N <200) and the cancer-specific setting for the others, as this could reflect cancer-type specific biology without jeopardising performance or stability. The final list of patient-specific predictions in 34 cancer types was composed of 4,470 unique genes, the vast majority of which (93%) were rare (<10 patients) or patient-specific (Figure 5-8F, Supplementary Table 4). A gene set enrichment analysis on these genes revealed 984 enriched pathways overall (Reactome level 2 or above, FDR <0.01, Methods 5.2.5, Supplementary Table 5). Interestingly, when mapping these pathways to broader biological processes (Reactome level 1), a few processes were widely enriched in almost all cancer types (Figure 5-8G). These included well-known cancer-related processes such as chromatin organisation[236], DNA repair[237], cell cycle[238] and signal transduction[239]. Therefore, although not recurring across patients, sysSVM2 predictions converged to perturb similar biological processes that are known to contribute to cancer.

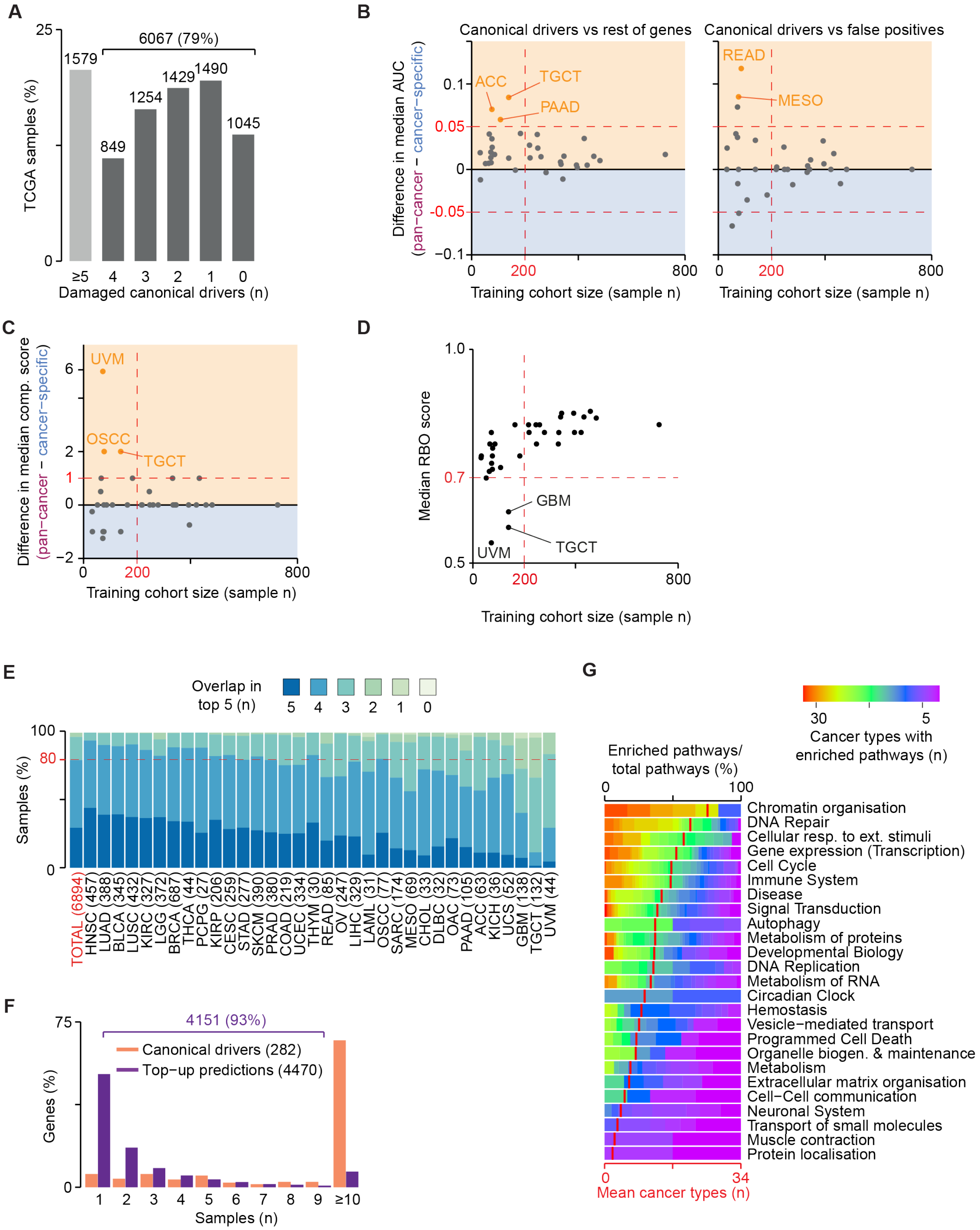

Figure 5-8: sysSVM2 predictions in 34 cancer types

A. Number of damaged canonical drivers per sample. Lists of canonical drivers for each cancer type were obtained from NCG[24] and mapped to samples of the corresponding cancer type. 6,067 samples with less than five canonical drivers damaged underwent the top-up procedure to reach five drivers.

Difference in Areas Under the Curve (AUCs) between the pan-cancer and cancer-specific settings in ranking canonical drivers over the rest of human genes and false positives (**B**) and in the composition score of the top five predictions (**C**). The median values of the distributions in each cancer type were used for comparison, with the yellow and blue regions indicating better performance in the pan-cancer and cancer-specific settings, respectively. The number of samples used for training is indicated on the x-axis. Colour dots represent cancer types where the two settings differ both significantly (FDR <0.05, Wilcoxon rank-sum test) and substantially (|difference in medians| >0.05 for AUCs, >1 for composition score). ACC, adrenocortical carcinoma; TGCT, testicular germ cell tumours; PAAD, pancreatic adenocarcinoma; READ, rectum adenocarcinoma; MESO, mesothelioma; UVM, uveal melanoma; and OSCC, oesophageal squamous cell carcinoma.

D. Rank-Biased Overlap (RBO) scores measuring the similarity between the top five predictions in the pan-cancer and cancer-specific settings. Median RBO scores for each cancer type are shown on the y-axis, and the number of samples used for training is shown on the x-axis. UVM, uveal melanoma; GBM, glioblastoma multiforme; TGCT, testicular germ cell tumours.

E. Overlap in the top five predictions between the pan-cancer and cancer-specific settings, for each cancer type. The numbers of samples whose predictions were considered are indicated in brackets.

F. Recurrence of damaging alterations in 282 canonical driver genes and 4,470 sysSVM2 top-up predictions across 7,646 samples.

G. Gene set enrichment analysis of sysSVM2 top-up genes, grouped in broad biological processes (Reactome level 1). Numbers of pathways enriched in at least one cancer type out of the total pathways tested are reported in brackets. Red vertical strokes indicate the mean number of cancer types that pathways from each broad process are enriched in (bottom x-axis).

5.4.5. sysSVM2 predictions in an independent cancer cohort

We finally sought to assess whether the sysSVM2 models trained on TCGA could be applied for driver prediction in a cancer type not included in TCGA. We therefore analysed 36 osteosarcomas from the Pan-Cancer Analysis of Whole Genomes

(PCAWG) consortium[16]. Osteosarcoma is a rare, genetically heterogeneous bone cancer with poor prognosis and only six well-established canonical drivers[240,241].

We annotated the genomic data of the PCAWG cohort finding 4,969 damaged genes overall with a median of 93 damaged genes per sample (Methods 5.2.6, Supplementary Table 2). Only two of these samples had three damaged osteosarcoma canonical drivers while 19 (53%) of them had no canonical driver (Figure 5-9A), highlighting the need for further predictions. Given the small cohort size, we used the TCGA pan-cancer setting to rank the damaged genes in each osteosarcoma. Considering the top five predictions per sample, we got 129 unique genes (Supplementary Table 6), which were poorly recurrent across samples (Figure 5-9B), reflecting again the genetic heterogeneity of osteosarcoma.

At the cohort level, sysSVM2 predictions included five of the six (83%) osteosarcoma canonical drivers[240,241]. At the patient level, the six osteosarcoma canonical drivers were damaged 27 times and in 14 of these cases (53%) they were in the top five predictions (Figure 5-9C). This proportion rose to 81% when considering the top ten predictions. In addition to osteosarcoma canonical drivers, 26 sysSVM2 predictions were canonical drivers in other cancer types, 16 were candidate cancer driver genes and 81 had no previously known involvement in cancer (Supplementary Table 6). Despite this, these 81 genes were enriched in eight pathways (FDR <0.1), most of which have a known role in cancer (Figure 5-9D). Moreover, they included genes known to promote osteogenesis such as *YAP1* and *YES1*[242,243].

These results showed that sysSVM2 is able to identify reliable cancer drivers in individual patients even for cancer types not used for training. This has relevant implications particularly in the case of rare cancers that are poorly studied and have little genomic data available.

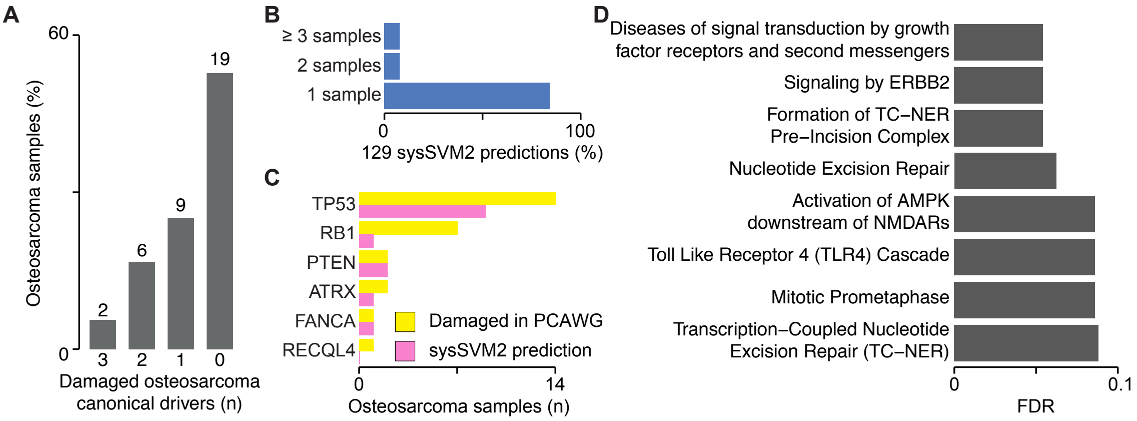

Figure 5-9: Validation of sysSVM2 in osteosarcoma

A. Distribution of osteosarcoma canonical drivers across the PCAWG osteosarcoma cohort. Lists of canonical drivers for osteosarcoma derived from the literature[240,241] and mapped to samples where they were damaged. Numbers of samples are indicated above each bar.

B. Recurrence of the 129 sysSVM2 predictions across the PCAWG osteosarcoma cohort. The percentages of genes that are predicted in 1, 2 and ≥3 are shown.

C. Patient-level predictions of osteosarcoma canonical drivers by sysSVM2 when considering the top five genes. The number of samples in which each canonical driver is damaged (yellow) and predicted as a driver by sysSVM2 (pink) is shown.

D. Gene set enrichment analysis of 81 sysSVM2 predictions with no previously reported involvement in cancer. Reactome level 2 and above were considered and pathways with FDR <0.1 are shown.

5.5. Alternative and extended models for driver gene identification

In order to identify drivers in individual patients using molecular and systems-level properties, a one-class learning approach is necessary due to the lack of an appropriate set of true negative driver genes for training a two-class model. However, one-class SVMs[230] are not the only available machine learning algorithms for this task. In particular, neural networks (NNs) provide flexible alternatives for one-class learning that have been shown to perform well in other settings[244], and that can benefit from flexible learning strategies[245]. We therefore sought to investigate whether NNs could perform better than sysSVM2 for patient-level driver gene identification.

In this section, we will first construct a NN-based alternative to sysSVM2, and show that it has a worse performance than the one-class SVM (5.5.1). We will then explore a fusion method that combines the performance of the SVM with the ability of NNs to learn from new data after an initial model has already been trained (5.5.2).

5.5.1. A variational autoencoder alternative to sysSVM2

Autoencoders are a class of NN algorithms that can be used for one-class learning, as well as for dimensionality reduction[246-248]. They have an hourglass-like architecture, with the output layer having the same dimension as the input layer. In between, one or more hidden layers are used first to 'encode' the inputs into a reduced-dimensionality representation (RDR), and subsequently to 'decode' the RDR to reproduce the original inputs as closely as possible. The principle behind using autoencoders for one-class applications is that they learn how to accurately encode and decode only the data included in the training set[247,248]. At prediction time, when they see a data point that differs substantially from the training set, the decoder should not accurately reconstruct the input. Thus, data with high or low reconstruction error are determined to be dissimilar or similar to the training set, respectively. Variational autoencoders (VAEs) are a particular type of autoencoder that, in addition to reconstructing the input data, attempt to produce an RDR that is similar to a Gaussian distribution[244]. This regularisation makes VAEs particularly stable, and makes their output probabilistically interpretable, as well as enabling them to perform generative modelling.

We implemented a VAE to perform the same task as sysSVM2 (Methods 5.2.7). Specifically, we used as input the same 26 molecular and systems-level features that sysSVM2 used to characterise genes (Figure 5-10A). The VAE encoded these 26 features into an RDR of just three dimensions before attempting to reconstruct the input data. We trained the VAE using the same simulated reference pan-cancer training set as sysSVM2, *i.e.* 457 unique tumour suppressor genes and oncogenes damaged 8324 times across 1000 simulated samples. When predicting on the remaining damaged genes, we considered low reconstruction errors to indicate high similarity to the training set of canonical drivers, and ranked genes accordingly. The training of the VAE converged within 10,000 epochs (Figure 5-10B).

Before assessing the performance of the VAE, we sought to ensure that it was meaningfully encoding the gene molecular and systems-level features. On inspection, genes with different features were indeed distributed differently within the RDR. For example, there was clear separation of genes according to amplification status (Figure 5-10C), evolutionary origin (Figure 5-10D) and degree in the protein-protein interaction network (PPIN, Figure 5-10E). Some features, such as expression across healthy tissues at the mRNA level, showed less clear separation in the RDR (Figure 5-10F). However, there was still evident structure in the RDR relating to these features. The combination of an average training set reconstruction MSE of 7% across all features, as well as the obvious structure of the RDR, indicated that the VAE had indeed learned how to encode and decode the molecular and systems-level properties of genes.

Finally, we used receiver operator characteristic (ROC) curves to assess the ability of the VAE to rank canonical driver genes above the rest of human genes. The VAE had substantially worse performance than sysSVM2 on simulated TCGA data. Comparing the ranks of canonical drivers to the rest of genes and false positives, the VAE had areas under the curve (AUCs) of 0.59 and 0.67, compared to 0.73 and 0.93 for sysSVM2 (Figures 5-10G, H). Since the VAE had a relatively high number of tuneable parameters, we reasoned that this might be due to a small training set size. However, even when training the VAE with data from the entire TCGA (7,630 samples), it was still outperformed by sysSVM2 (AUCs of 0.61 and 0.65, Figures 5-10G, H).

The VAE provided an alternative NN-based algorithm to carry out the same task as sysSVM2. However, it appeared that, despite any other advantages that NNs may have over SVMs, the VAE was unable to perform as well as sysSVM2.

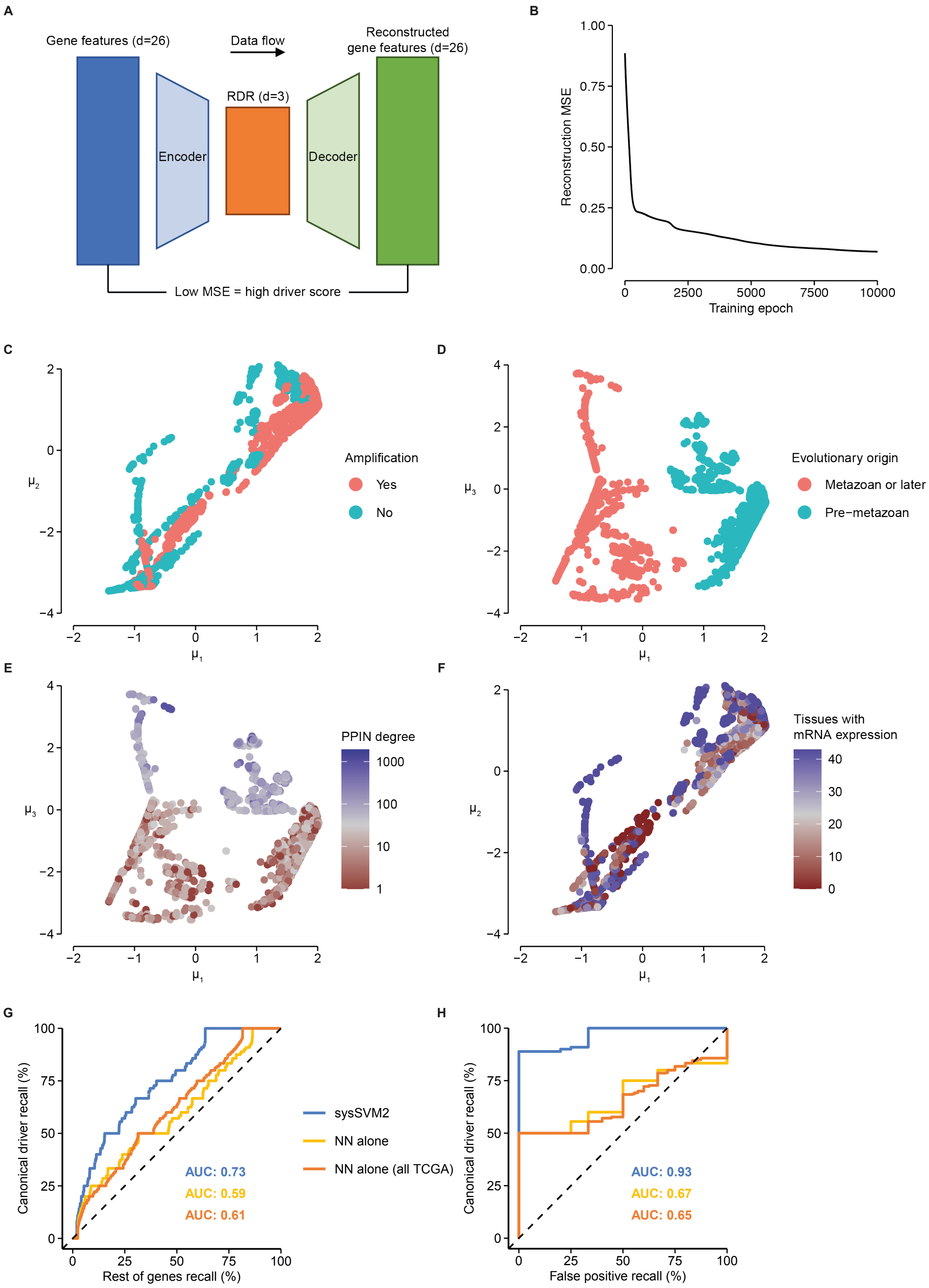

Figure 5-10: Implementation of a neural network alternative to sysSVM2

A. Architecture of the VAE. Mean squared error (MSE) was used to measure the similarity between the inputs and the reconstructed outputs. High reconstruction MSE indicated that a gene was dissimilar to the training set of canonical drivers, and therefore corresponded with a low driver score.

B. Convergence of VAE training. Each epoch represents one iteration of the training process.

Visualisations of a random sample of 10,000 damaged genes in the RDR, coloured according to binary (**C**, **D**) and continuous (**E**, **F**) properties. Each panel only shows two of the three dimensions in the RDR, labelled μ_1, μ_2 and μ_3. PPIN: protein-protein interaction network.

Performance of the VAE trained on simulated data (yellow) and all available TCGA data (orange), compared to sysSVM2 trained on simulated data (blue). The median ROC curve comparing ranks of canonical drivers to the rest of genes (**G**) or false positives (**H**) for each algorithm is shown, along with the median area under the curve (AUC).

5.5.2. Fusion of SVM and NN for high performance and flexible learning

Static models trained in a single instance on a given cohort can be used for prediction on new samples. However, they cannot incorporate information from additional samples without being re-trained *de novo*. This has drawbacks in both research and clinical settings. In research, it is common for new samples to be progressively added to large sequencing efforts such as the ICGC. Conversely, in clinical settings new data often arrive sporadically. For example, targeted and whole genome sequencing are rapidly becoming part of the diagnostic routine of new cancer patients. In both scenarios, it would be useful to dynamically extract data from these additional samples and use them to update a pre-trained driver detection model. This would allow a continuous learning of the model without re-training it *de novo* on the extended cohort. The one-class SVM framework behind sysSVM2 is not amenable to such a dynamic expansion. By contrast, the flexibility of NNs presents a possible solution[245]. We therefore investigated whether it would be possible to use NNs to address this drawback of sysSVM2 without compromising performance.

We constructed a NN to allow a trained sysSVM2 model to learn from additional samples without starting *de novo*. We used a multi-task architecture consisting of an autoencoder augmented with a secondary output that explicitly produced scores for genes (Methods 5.2.7, Figure 5-11A). We termed this NN an augmented autoencoder (AAE). The principle of the AAE was that information learned by the NN for the task of feature encoding and decoding could also help to improve explicit score prediction. Using one task to improve another is the basis of multi-task learning[249]. The AAE is trained using (1) a set of initial samples, on which sysSVM2 has already been trained, and (2) a set of additional samples that may have become available at a later time. During training, the AAE learns to encode and decode the features of all damaged genes from all samples, as in a standard autoencoder. However, it also learns to reproduce the scores that sysSVM2 assigned to genes damaged in the initial cohort. This is achieved with extra hidden layers branching off from the RDR of the autoencoder. During prediction on additional samples, the explicit scores predicted by this secondary output are used to score and rank genes as patient-specific drivers. Because the additional samples are used in training the autoencoder tasks of feature encoding and decoding, they can drive the AAE to learn a better-generalised RDR than by using initial samples alone. Since the explicit score prediction uses the RDR, this task can indirectly learn from the additional samples as well.

In order to assess the utility of the AAE, we first generated additional simulated pan-cancer cohorts of a range of sizes (from one to 1000), such that for each size we had explored a total of 6000 samples. We verified that the training of the AAE converged within 10,000 epochs (Figure 5-11B). Notably, for cohorts of 1000 samples the average wall clock training time was only 12% of that required to train sysSVM2 with 1000 cross-validation iterations (Figure 5-11C). This confirmed that the AAE could allow additional samples to be analysed more easily than with sysSVM2 alone.

To assess the impact of the AAE on the ability to detect driver genes, we made predictions in the additional samples with both the initial sysSVM2 model and the extended AAE model. For each set of predictions in each sample, we measured the AUC comparing the ranks of canonical drivers to both the rest of genes and to false positives. We then assessed whether the AUCs had increased or decreased as a result of applying the AAE. Overall, both types of AUC measures tended to increase with the application of the AAE (Figure 5-11C). However, while we had expected performance to improve more as larger additional cohorts were used to train the model, the opposite was true. For example, while the AUC comparing canonical drivers to the rest of genes increased in 69% of samples when using additional cohorts of size one, it only increased in 58.7% of samples with additional cohorts of 1000 samples. This decrease in performance with larger training cohorts indicated that changes in the AUCs may have been due to how the AAE learned from the initial samples, rather than from new information brought in by the additional samples. While the performance did increase form sysSVM2 alone overall, this counter-intuitive result suggested that the AAE did not fulfil its intended purpose.

Due to these results, we did not proceed further with development or testing of the AAE framework. However, the issue of continuous learning as more data become available is one that remains to be adequately addressed in the cancer research community. Moreover, the need for continuous learning techniques will only become more pressing as larger data sets and more complicated model-based methods are adopted.

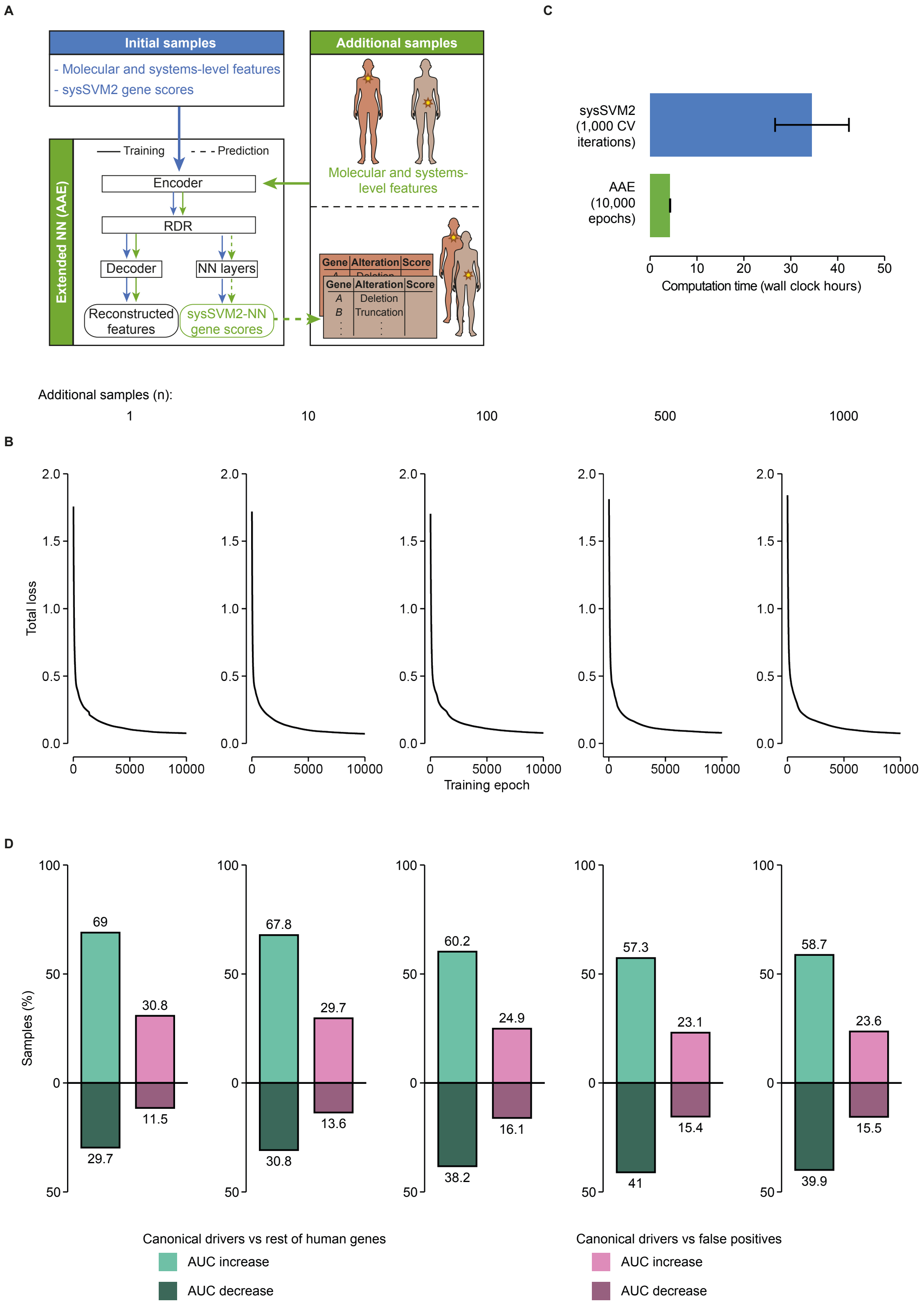

Figure 5-11: Augmented autoencoder extension to the sysSVM2 algorithm

A. Schematic illustrating the augmented autoencoder (AAE). First, sysSVM2 is trained on an initial cohort. The gene features and scores from the initial cohort are then combined with the gene features from additional samples. The NN attempts to simultaneously predict the scores for the initial cohort, and reconstruct the gene features of both initial and additional samples. The trained algorithm then outputs scores for genes in additional samples.

B. Wall-clock computation time for sysSVM2 with 1,000 cross-validation (CV) iterations (blue), and the AAE with 10,000 training epochs (green). All computations were carried out on a high-performance computing cluster.

C. Convergence of the AAE algorithm trained with different numbers of additional samples (left to right), starting from the sysSVM2 model trained on 1,000 simulated initial samples. The total loss was calculated as the sum of the feature reconstruction MSE (all samples) and the score error (initial samples only). The score error function was the sum of MSE and mean absolute error to penalise both small and large errors.

D. Performance change resulting from the AAE implementation. For each number of additional samples used in training, the percentage of initial samples for which the performance increased or decreased is shown. Performance was measured using the per-sample AUC comparing the ranks of canonical drivers to the rest of human genes (green) and to known false positive genes (pink). Light shades indicate improved performance while dark shades indicate worse performance, compared to sysSVM2.

5.6. Discussion

In this chapter, we developed a cancer-agnostic algorithm, sysSVM2, for identifying cancer drivers in individual patients. By refining the machine learning approach upon which the original algorithm was built[27], we broadened its applicability to the pan-cancer range of malignancies represented in TCGA. sysSVM2 successfully and stably prioritises canonical driver genes for most publicly available cancer cohorts. For those composed of fewer samples, the models optimised on the whole pan-cancer dataset offer a valid alternative. Moreover, compared to other patient-level driver detection methods, sysSVM2 has better patient coverage and a particularly low rate of predicting established false positives. We also found that sysSVM2 outperforms alternative models based on neural networks. The genes identified by sysSVM2 in pan-cancer data converged to well-known cancer-related biological processes. sysSVM2 can therefore be used to identify driver alterations in individual patients and rare cancer types where canonical drivers are insufficient to explain the onset of disease, as we have validated in osteosarcoma.

This work potentially opens up further research and therapeutic opportunities. Identifying the complete repertoire of driver events in each cancer patient holds great potential for furthering the molecular understanding of cancer and ultimately for precision oncology. While many recurrent driver genes have now been identified, the highly heterogeneous long tail of rare drivers still poses great challenges for detection and validation. However, in overcoming inter-tumour heterogeneity by identifying these rare drivers, sysSVM2 and other patient-level driver detection methods can help researchers develop new therapeutic interventions for patients whose somatic alterations are not currently considered in precision oncology.

Further studies could potentially use our driver predictions to investigate particular aspects of cancer biology, such as driver clonality and the progressive acquisition of drivers during cancer evolution. Extending the algorithm with additional sources of data is another avenue for future work. For example, transcriptomic and epigenomic data could enhance the ability of sysSVM2 to identify driver events. Additionally, recent efforts have identified a large number of driver events in non-coding genomic elements[16]. Given such a training set of true positives, sysSVM2 could be further developed to identify non-coding drivers in individual patients, as long as appropriate features could be identified. The general approach of identifying drivers using a combination of molecular and systems-level properties affords great flexibility for such developments.

It is increasingly common for sequencing studies to integrate multiple tools for driver detection[15], since building a consensus can make results robust to the weaknesses of individual methods. sysSVM2 also has its weaknesses. For example, while systems-level properties distinguish cancer genes as a set, there are some cancer genes that do not follow this trend[24] and are thus likely to be missed by the algorithm. In addition, our reliance on canonical drivers (most of which are highly recurrent) for training and assessment of sysSVM2 was imperfect, but it illustrated the lack of currently available ground truth for developing patient-level driver detection methods. Our approach in the current work of topping up known driver genes with predictions from sysSVM2 is a simple example of how sysSVM2 can be used in conjunction with other approaches. More broadly, it is likely the case that patient-level driver detection will eventually rely on an entire ecosystem of different methods. In this work, we have demonstrated that there is a place for sysSVM2 in such an ecosystem.

Chapter 6. Discussion

6.1. Summary and conclusions

In this thesis, I have investigated how the biology of cancer differs from one patient to another. First, I described how germline variation can give rise to patient-specific selective pressures acting on somatic alterations, thus contributing to inter-tumour heterogeneity. I investigated this effect in oesophageal adenocarcinoma (OAC) in Chapter 2 with a co-occurrence-based approach, and then again in Chapter 3 with a logistic regression modelling approach and an expanded OAC cohort. Driver genes are the most relevant aspect of genetic inter-tumour heterogeneity for cancer biology and therapy. I therefore introduced the Network of Cancer Genes, a repository of reported driver genes and their systems-level properties, in Chapter 4. Finally, I used these properties to develop a method, sysSVM2, for identifying driver genes (including potentially rare or patient-specific drivers) at the patient-level in Chapter 5.

My initial investigation of the germline influence on somatic evolution in Chapter 2 found two patterns of co-occurrence between deleterious germline variants and somatic driver alterations at the pathway-level in 260 OACs. In the first, deleterious germline variants in extracellular matrix genes co-occurred with drivers in RTK signalling genes, most notably *KRAS* and *PIK3CA*. Literature evidence suggested that this effect could have been mediated by discoidin-domain receptors[115-117] or integrins[118,119], both of which had patterns of expression that suggested a genuine effect. In the second result, germline perturbations to DNA repair genes led to increased frequencies of driver events downstream of RTK signalling, including in *MYC*, *FOXO1* and *CCND3*. Differential expression of keratinisation genes suggested that this association may have been related to differences in tumour morphology between samples. In addition, a statistically significant association with earlier age at diagnosis hinted at a possible clinical relevance of this result. However, when I analysed an independent cohort of 140 OACs, neither of these germline-somatic associations could be validated. I concluded that there were a number of methodological limitations of this analysis, including sample size and the statistical approaches used.

I attempted to address these limitations in Chapter 3. Using a logistic regression modelling approach and an expanded OAC cohort, I identified a negative association between germline perturbations to the *ATM* signalling pathway and *TP53* driver alterations. This signal was strongest for truncating variants in the *ATM* gene itself. This result was plausible because *ATM* mutations had previously been found to substitute for *TP53* driver alterations in breast cancer[172], lung adenocarcinoma[55], T-cell leukaemia[173] and B-cell lymphoma[174]. In addition, *ATM* and *TP53* interact directly to activate DNA repair and apoptosis programmes in response to DNA double-strand breaks[169,170]. The interaction between *ATM* and *TP53* in my data also strongly suggested that *ATM* had two roles in OAC that have not been reported to date. First, the fact that *ATM* truncations could substitute for *TP53* driver alterations (the most common driver in OAC) indicated that it acted as a tumour suppressor gene, as it does in other cancer types[166]. Second, *ATM* germline truncations exhibited frequent loss of heterozygosity, enrichment in OAC compared to healthy controls, and an association with younger age at diagnosis, suggesting that these variants could predispose individuals to developing OAC. While *ATM* is a known predisposition gene in other cancer types[18,166], this was of note because to date no highly penetrant predisposition gene for OAC has been found[23]. Thus, *ATM* warrants further scrutiny for its role in OAC.

In Chapter 4 I focused on the available knowledge regarding somatic driver genes as a starting point for future driver gene identification efforts. I studied the breadth of literature evidence for drivers, and found that the number of canonical driver genes in each cancer type depended strongly on the number of patients whose cancers had been sequenced. In particular, some rare cancer types had very few canonical drivers as a result of small cohort sizes. By contrast, extensively-studied cancer types tended to have large numbers of candidate drivers which had not been experimentally validated. I also found a general lack of consensus between studies to reproducibly find the same driver genes, even when using widely accepted gold-standard methods for driver identification. These results suggested that larger sequencing efforts would be valuable in rare cancer types. In addition, they indicated that existing methods were not well-suited for interrogating the long tail of rare and possibly patient-specific driver genes in cancer. Different driver detection methods were required to overcome this issue. Moreover, such methods could provide a more feasible alternative to sequencing large cohorts of patients from rare cancer types. I also introduced the systems-level properties of driver genes, which distinguish drivers from other human genes. It is possible to use these properties for driver gene identification[195].

In Chapter 5, I optimised sysSVM, a method previously developed by the Ciccarelli lab for patient-level driver detection in OAC[27], for pan-cancer use. The resulting method, sysSVM2, learned the molecular and systems-level properties of canonical driver genes, and used these to predict drivers in individual patients. I found that, compared to other patient-level driver prediction methods, sysSVM2 had a low false positive rate and better patient coverage. Moreover, despite being heterogeneous at the gene-level, its predictions in pan-cancer data converged to well-known cancer-related processes. It was also able to recover established driver genes in osteosarcoma, a rare cancer type. These results demonstrated that sysSVM2 is potentially useful in identifying drivers in individual patients. This is necessary to explain the onset of cancer in patients where no canonical drivers are altered, and for whom there is therefore a significant shortfall in our understanding of their tumours' biology.

These investigations helped to elucidate the ways in which the genetic components of cancer biology are specific to individual patients. For example, the *ATM* result in Chapter 3 showed that, in a small proportion of patients, inherited genetic variants strongly influence the evolutionary trajectories of OACs, preventing the fixation of the disease's most recurrent driver gene. The results in Chapters 4 and 5 also indicated that that many driver genes were rare or patient-specific. It is tempting to speculate that the reason for the rarity or patient-specificity for some drivers could be genetic, *i.e.* that some genes will only drive cancer in particular germline contexts. Indeed, in a previous study our group found that OAC patients with deleterious germline variants in cancer predisposition genes were less likely to have patient-specific driver alterations in DNA repair and cell cycle genes in general[27]. This could imply that germline variants affecting DNA repair pathways make further somatic alterations in these pathways either to confer no selective advantage on cells, or even to become lethal. In either case, the selective advantage for these somatic alterations would be greatly reduced by the presence of the germline perturbations. Similarly, inherited vulnerabilities in which selective pressures on somatic alterations are increased by certain germline variants could exist, although I was unable to find robust evidence of such effects in this thesis.

6.2. Limitations of current approaches

There are a number of limitations of currently-available approaches that impacted the research carried out in this thesis. These include the availability of patient genomic sequencing data, the lack of suitable experimental validation methods, a lack of available multi-omic data, and difficulties of systematically incorporating important aspects of tumour biology such as the tumour microenvironment and intra-tumour heterogeneity into analysis.

My analysis of the germline influence on somatic evolution in cancer used cohorts of OAC patients that were almost as large as was possible at the time, given the requirement for matched germline and somatic sequencing data (n=470 in Chapter 3). By comparison, only one of the cohorts in TCGA (breast cancer) was more than 10% larger than this cohort. However, power calculations showed that a fully statistically-powered study using the same approaches would require approximately 40,000 samples. Given that such analyses need to be conducted for individual cancer types to avoid confounding affects[55], it is not likely that sufficiently large cohorts will be available in the near future. Thus, researchers do not yet have the data required to fully understand the effect of the germline on somatic evolution in cancer. OAC served as a useful model cancer type with relatively good data available and a lack of known penetrant predisposition genes that could strongly influence any results. Interestingly, the associations between germline *ATM* signalling and *TP53* somatic drivers seemed to hold in stomach adenocarcinoma, suggesting a common mechanism between the two cancer types. However, analyses of other cancer types with well-established predisposition genes or less available data may not add to current knowledge. Larger cohorts than are currently available may be required to fruitfully investigate the role of the germline in these cancer types. By contrast, I used approximately 7500 pan-cancer samples for the development and assessment of sysSVM2 in Chapter 5. Here however, differences in cohort size between individual cancer types were consequential, with smaller cohorts giving rise to less stable models. In addition, TCGA does not include all cancer types. Thus, in the future more comprehensive cohorts could be used as a basis for patient-level driver gene prediction in the future. It should be noted however, that the validation of the pan-cancer model in osteosarcoma (not included in TCGA) suggested that information learned in one group of cancer types could be applied to others.

The results throughout this thesis would be challenging to systematically validate in experimental settings. The effect of germline variants on somatic evolution might be validated by experiments aimed at interrogating proposed mechanisms for germline-somatic interactions. However, it is probably infeasible to fully explore the effects of germline variants in a controlled *in vivo* experimental setting. The patient-level driver predictions of sysSVM2 are similarly challenging to validate. Agreement with other prediction methods and convergence of predictions to cancer-related processes provided indirect support of their validity, but these did not constitute conclusive evidence. Some genes predicted as drivers by the earlier version of the algorithm in OAC were experimentally shown to increase proliferation in OAC cell lines[27]. However, given the sheer number of total genes predicted by this approach in large cohorts, it is unlikely that all predictions can be validated in this way. Moreover, it is important to consider why a driver event might be rare or patient-specific. If its driver role depends on a particular biological context (*e.g.* genetic makeup or environmental factor) then it is unlikely to be possible to validate it *in vitro*. Experimental validation of the patient-specific molecular features of tumours is likely to become a greater challenge as more such features are discovered. However, patient-derived organoid technology provides a possible avenue for such work, as recently demonstrated with patient-specific drug screens[250].

Due to data availability, the discovery approaches used in this thesis relied almost exclusively on genomic sequencing data. A systematic integration of multi-omic data, such as from RNA-Seq and methylation bisulphite sequencing, could improve the ability of statistical analyses to identify biological effects. The value of multi-omic approaches is increasingly appreciated by the research community, particularly in identifying driver alterations. For example, two other patient-level driver prediction algorithms, DriverNet[226] and OncoIMPACT[225], rely on model-based integration of genomic and transcriptomic data. Gaining a more complete picture of driver events from multi-omic data could also enhance investigations of germline contributions to somatic evolution. For example, it is likely that my analysis missed somatic driver events due to methylation or other transcriptional deregulation of canonical drivers. An integrative multi-omic approach could refine the statistical signals of biological effects by reducing the prevalence of similar false negatives in the data. However, matched multi-omic data is often unavailable for large numbers of samples, so these approaches might be applied in the future as multi-omic sequencing becomes more common.

Throughout this thesis, I did not consider the effect of the tumour microenvironment on the role of the germline in somatic evolution and the driver impact of genes. For example, immune editing is known to affect somatic evolution of the cancer genome in a manner that is dependent on patients' germline HLA types[72]. Moreover, inherited immune defects are likely to modify the role of the immune system in cancer[251]. In a similar way, certain somatic driver genes are known to modulate tumour-immune interactions[252]. Understanding these relationships will likely be important for reliably identifying driver genes in individual patients. However, while my analysis of the germline in Chapters 2 and 3 did include immune system pathways, these did not associate with somatic driver genes, suggesting that larger cohorts may be required to analyse the impact of germline immune perturbations on somatic evolution. In addition, it is not clear how to usefully incorporate immune effects into driver gene detection methods. For example, including HLA types or immune cell population estimates as features in a classifier such as sysSVM2 may not improve performance if immune features are too heterogeneous between patients, subject to technical biases or simply not relevant for most driver genes. A deeper understanding of the role of the tumour microenvironment and greater data availability is required to systematically incorporate immune effects into the research questions of this thesis.

Finally, in order to balance model complexity with data availability I did not consider the effects of intra-tumour heterogeneity in my analyses. By considering both clonal and sub-clonal somatic alterations, it would be possible to characterise somatic evolution at a finer resolution than I have done. This could allow us to find more subtle effects of the germline on cancer evolution, such as on the ordering of driver events. In addition, the patient-specificity of driver alterations could in part depend on clonality, with rare drivers acquired preferentially either early or late in evolution. However, it is important to note that incorporating clonality into these analyses would add a degree of freedom to already complex problems. Thus, additional insights could likely only be gained from considering intra-tumour heterogeneity if sufficiently large cohorts were available.

6.3. Perspectives

I will conclude this thesis with my perspectives on inter-tumour heterogeneity and some outstanding challenges for the cancer genomics field.

The extent to which inter-tumour heterogeneity is random versus deterministic remains to be fully addressed. A simple conceptual model is that randomly acquired somatic alterations undergo selective pressures that are common across all cancers of a particular type. Results in this thesis and elsewhere indicate however that selective forces can vary from patient to patient due to inherited and environmental factors[17,50,51,55,60]. It would be interesting to know the extent to which selective pressures are patient-specific across cancer types. However, at least with currently available methods and data, the impact of these patient-specific selective pressures appears to be small. Thus, it is possible that inter-tumour heterogeneity is primarily the result of the inescapably stochastic nature of the underlying process of somatic alteration to the genome. This suggests, unfortunately, that we may never be able to predict the evolutionary trajectories of cancer with complete certainty, no matter how deeply we understand cancer biology. However, a possible alternative is that selective pressures in cancer are indeed largely common between cancer patients and lead to highly convergent evolution, but at high functional levels. For example, it is well-documented that many driver alterations converge to the same pathways[15,16,96], despite occurring in different genes. Describing these higher levels of convergent evolution is challenging, however. Moreover, cancer therapies with low toxicity are likely to target highly specific components of tumour biology, rather than high-level processes. Indeed, chemotherapy could be considered to be an extremely high-level therapy targeted at cell division. Thus, even being able to fully describe convergent cancer evolution at a high level may have limited clinical utility.

Inter-tumour heterogeneity presents one of the main obstacles for identifying potentially targetable, specific, components of tumour biology. For example, the exclusivity between germline *ATM* truncations and *TP53* driver alterations was a weak signal at the cohort-level. Only about 1% of samples were affected, and overall *ATM* truncations explained only 3% of the variance in *TP53* status. However, in those few samples where *ATM* was truncated, the effect appeared to be very strong, with perfect mutual exclusivity with *TP53*. This is a good illustration of how inter-tumour heterogeneity buries the signals of strong biological effects that rarely occur. Patient-level driver prediction faces a similar obstacle. The fact that gold-standard methods fail to identify drivers in some patients strongly suggests that many driver events are very rare. Nevertheless, by their very nature as drivers, they are crucial to the biology of the tumours where they have a driver role. Again, signals that appear weak at the cohort-level in fact reflect biology of great importance to a small number of cancers. This highlights the apparent paradox of inter-tumour heterogeneity. We know that every tumour is molecularly distinct, but in order to learn more about the disease we look for commonalities between tumours. For example, we look for recurrence of somatic alterations (at the mutation, gene or pathway levels), common gene properties (as in sysSVM2) or for patterns of germline and somatic co-occurrence and exclusivity. Nevertheless, all of these efforts have the ultimate aim of bringing more personalised therapies to the clinic. In order to better understand and treat the individual tumour, it is necessary to study thousands of tumours.

In order to fully realise precision oncology therefore, there is a pressing need for more sequencing data to become available to researchers. As already alluded to, widespread inclusion of multi-omic data will be vital to gain a complete picture of tumour biology for discovery. While addition of multi-omic data may appear at first to add complexity to the molecular landscape of cancer, it will hopefully lead to identifying more common patterns across tumours by allowing them to be characterised at higher functional levels, rather than purely in terms of genomic alterations. Indeed, by obtaining stronger signals of functional features of cancers in this way, it is possible that fewer samples overall will be required to identify common patterns. In the germline, there is also a particular need for sequencing data to replace array data. Many results in cancer predisposition, for instance, rely on SNPs that have no functional interpretation[23,253]. Leveraging these associations for functional insights is therefore often not possible.

Incorporating previous functional insights into discovery is non-trivial. A general methodological challenge for cancer researchers is appropriately leveraging prior knowledge in systematic analyses. For example, I aggregated damaging germline variants into known pathways in order to characterise them at a functional level. I might have also used interaction networks to prioritise tests between germline pathways and somatic drivers that are known to interact. I also leveraged prior knowledge by training sysSVM2 to learn the properties of canonical driver genes, in order to identify new drivers. Of course, all of these approaches are potentially biased by the very prior knowledge they leverage. Unbiased approaches might be ideal if sufficient data are available for discovery, and they often appeal to researchers' entirely correct desire to be impartial. However, prior knowledge should be used as a resource for more than just validation and assessment in computational analyses. How best to leverage prior knowledge while not just finding expected results, however, is a major challenge that will hopefully continue to be addressed in the coming years.

The 2020's are likely to be an exciting time for cancer research. Previous decades have brought a flurry of increased molecular understanding of cancer, widespread sequencing of the cancer genome, and the long-awaited development of targeted therapies. Substantial challenges remain, however, and cancer persists as the leading cause of premature death in the developed world[1]. We can hope that by adapting true multi-omic cancer medicine at scale, we might improve the outlook of millions of patients.

Appendix 1. Data normalisation and the one-class Support Vector Machine

In this section we discuss the theory behind why data centering is not optimal for the one-class Support Vector Machines in sysSVM.

The one-class SVM algorithm of Schölkopf *et al.*[230] aims to construct a decision boundary which encloses the region of input space $\mathcal{X}$ where the training set lies. It then classes points inside the decision boundary as similar to the training set (cancer genes), and points outside as outliers (non-cancer genes). In order to draw complicated decision boundaries efficiently, it uses a kernel function k to (implicitly) map the training set to a feature space, $\mathcal{F}$. In $\mathcal{F}$, the algorithm aims to separate the training set from the origin, using a linear decision boundary (a hyperplane). The final decision boundary is the result of mapping this hyperplane back to the input space $\mathcal{X}$.

The problem with data centering is most clear for the linear kernel. With this choice of kernel, feature space $\mathcal{F}$ corresponds exactly to input space $\mathcal{X}$. If data are centered around the origin (*i.e.* zero), then the algorithm attempts to separate the data from their own centre, which is clearly inappropriate. Indeed, experiments with a toy example indicate that this can lead to extreme sensitivity to small changes in the training set (data not shown).

On the other hand, centering is not an issue for the radial kernel, since it is translationally invariant: if gene features x and y are shifted by some constant value c, then it can be seen from the radial kernel function that $k(x - c, y - c) = k(x, y)$.

For the polynomial and sigmoid kernels, the effect of centering is less clear. They both preserve the origin – that is, the origin in $\mathcal{X}$ corresponds to the origin in $\mathcal{F}$. This can be seen from the fact that $k(0, x) = 0$ for any set of gene features x, with either kernel. This might suggest that centering should be avoided with these kernels, and this is further borne out in toy model experiments.

Appendix 2. References

1 IARC. *World cancer report: cancer research for cancer prevention*. (WHO, 2020).

2 Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23-28, doi:10.1126/science.959840 (1976).

3 Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57-70, doi:10.1016/s0092-8674(00)81683-9 (2000).

4 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).

5 Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719-724, doi:10.1038/nature07943 (2009).

6 Slamon, D. J. *et al.* Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **235**, 177-182, doi:10.1126/science.3798106 (1987).

7 Cheang, M. C. *et al.* Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *J Natl Cancer Inst* **101**, 736-750, doi:10.1093/jnci/djp082 (2009).

8 Goldhirsch, A. *et al.* Strategies for subtypes--dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Ann Oncol* **22**, 1736-1747, doi:10.1093/annonc/mdr304 (2011).

9 Waks, A. G. & Winer, E. P. Breast Cancer Treatment: A Review. *JAMA* **321**, 288-300, doi:10.1001/jama.2018.19323 (2019).

10 Kalia, M. Personalized oncology: recent advances and future challenges. *Metabolism* **62 Suppl 1**, S11-14, doi:10.1016/j.metabol.2012.08.016 (2013).

11 Senft, D., Leiserson, M. D. M., Ruppin, E. & Ronai, Z. A. Precision oncology: the road ahead. *Trends Mol. Med.* **23** (2017).

12 Piccart-Gebhart, M. J. *et al.* Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *N Engl J Med* **353**, 1659-1672, doi:10.1056/NEJMoa052306 (2005).

13 Wang, J. & Xu, B. Targeted therapeutic options and future perspectives for HER2-positive breast cancer. *Signal Transduct Target Ther* **4**, 34, doi:10.1038/s41392-019-0069-2 (2019).

14 Sabnis, A. J. & Bivona, T. G. Principles of Resistance to Targeted Cancer Therapy: Lessons from Basic and Translational Cancer Biology. *Trends Mol Med* **25**, 185-197, doi:10.1016/j.molmed.2018.12.009 (2019).

15 Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385 e318, doi:10.1016/j.cell.2018.02.060 (2018).

16 Consortium, I. T. P.-C. A. o. W. G. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93, doi:10.1038/s41586-020-1969-6 (2020).

17 Dogan, S. *et al.* Molecular epidemiology of EGFR and KRAS mutations in 3,026 lung adenocarcinomas: higher susceptibility of women to smoking-related KRAS-mutant cancers. *Clin Cancer Res* **18**, 6169-6177, doi:10.1158/1078-0432.CCR-11-3265 (2012).

18 Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **505**, 302-308, doi:10.1038/nature12981 (2014).

19 Half, E., Bercovich, D. & Rozen, P. Familial adenomatous polyposis. *Orphanet J Rare Dis* **4**, 22, doi:10.1186/1750-1172-4-22 (2009).

20 Agarwal, D., Nowak, C., Zhang, N. R., Pusztai, L. & Hatzis, C. Functional germline variants as potential co-oncogenes. *NPJ Breast Cancer* **3**, 46, doi:10.1038/s41523-017-0051-5 (2017).

21 Dong, J. *et al.* Determining Risk of Barrett's Esophagus and Esophageal Adenocarcinoma Based on Epidemiologic Factors and Genetic Variants. *Gastroenterology* **154**, 1273-1281 e1273, doi:10.1053/j.gastro.2017.12.003 (2018).

22 Gharahkhani, P. *et al.* Genome-wide association studies in oesophageal adenocarcinoma and Barrett's oesophagus: a large-scale meta-analysis. *Lancet Oncol* **17**, 1363-1373, doi:10.1016/S1470-2045(16)30240-6 (2016).

23 Callahan, Z. M., Shi, Z., Su, B., Xu, J. & Ujiki, M. Genetic variants in Barrett's esophagus and esophageal adenocarcinoma: a literature review. *Dis Esophagus* **32**, doi:10.1093/dote/doz017 (2019).

24 Repana, D. *et al.* The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol* **20**, 1, doi:10.1186/s13059-018-1612-0 (2019).

25 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).

26 An, O., Dall'Olio, G. M., Mourikis, T. P. & Ciccarelli, F. D. NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings. *Nucleic Acids Res* **44**, D992-999, doi:10.1093/nar/gkv1123 (2016).

27 Mourikis, T. P. *et al.* Patient-specific cancer genes contribute to recurrently perturbed pathways and establish therapeutic vulnerabilities in esophageal adenocarcinoma. *Nat Commun* **10**, 3101, doi:10.1038/s41467-019-10898-3 (2019).

28 Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer* **12**, 323-334, doi:10.1038/nrc3261 (2012).

29 Ramon, Y. C. S. *et al.* Clinical implications of intratumor heterogeneity: challenges and opportunities. *J Mol Med (Berl)* **98**, 161-177, doi:10.1007/s00109-020-01874-2 (2020).

30 Lim, Z. F. & Ma, P. C. Emerging insights of tumor heterogeneity and drug resistance mechanisms in lung cancer targeted therapy. *J Hematol Oncol* **12**, 134, doi:10.1186/s13045-019-0818-2 (2019).

31 McGranahan, N. *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351**, 1463-1469, doi:10.1126/science.aaf1490 (2016).

32 Caswell, D. R. & Swanton, C. The role of tumour heterogeneity and clonal cooperativity in metastasis, immune evasion and clinical outcome. *BMC Med* **15**, 133, doi:10.1186/s12916-017-0900-y (2017).

33 Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122-128, doi:10.1038/s41586-019-1907-7 (2020).

34 Raynaud, F., Mina, M., Tavernari, D. & Ciriello, G. Pan-cancer inference of intra-tumor heterogeneity reveals associations with different forms of genomic instability. *PLoS Genet* **14**, e1007669, doi:10.1371/journal.pgen.1007669 (2018).

35 Marusyk, A. *et al.* Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature* **514**, 54-58, doi:10.1038/nature13556 (2014).

36 Tabassum, D. P. & Polyak, K. Tumorigenesis: it takes a village. *Nat Rev Cancer* **15**, 473-483, doi:10.1038/nrc3971 (2015).

37 Cleary, A. S., Leonard, T. L., Gestl, S. A. & Gunther, E. J. Tumour cell heterogeneity maintained by cooperating subclones in Wnt-driven mammary cancers. *Nature* **508**, 113-117, doi:10.1038/nature13187 (2014).

38 Oesper, L., Mahmoody, A. & Raphael, B. J. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol* **14**, R80, doi:10.1186/gb-2013-14-7-r80 (2013).

39 Oh, B. Y. *et al.* Intratumor heterogeneity inferred from targeted deep sequencing as a prognostic indicator. *Sci Rep* **9**, 4542, doi:10.1038/s41598-019-41098-0 (2019).

40 Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat Methods* **11**, 396-398, doi:10.1038/nmeth.2883 (2014).

41 Dentro, S. C., Wedge, D. C. & Van Loo, P. Principles of Reconstructing the Subclonal Architecture of Cancers. *Cold Spring Harb Perspect Med* **7**, doi:10.1101/cshperspect.a026625 (2017).

42 Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007, doi:10.1016/j.cell.2012.04.023 (2012).

43 Cmero, M. *et al.* Inferring structural variant cancer cell fraction. *Nat Commun* **11**, 730, doi:10.1038/s41467-020-14351-8 (2020).

44 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).

45 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

46 Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* **343**, 78-85, doi:10.1056/NEJM200007133430201 (2000).

47 Mucci, L. A. *et al.* Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA* **315**, 68-76, doi:10.1001/jama.2015.17703 (2016).

48 Zhang, J. *et al.* Germline Mutations in Predisposition Genes in Pediatric Cancer. *N Engl J Med* **373**, 2336-2346, doi:10.1056/NEJMoa1508054 (2015).

49 Huang, K. L. *et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* **173**, 355-370 e314, doi:10.1016/j.cell.2018.03.039 (2018).

50 Wang, Y. *et al.* Interaction analysis between germline susceptibility loci and somatic alterations in lung cancer. *Int J Cancer* **143**, 878-885, doi:10.1002/ijc.31351 (2018).

51 Zhang, X., Wang, Y., Tian, T., Zhou, G. & Jin, G. Germline genetic variants were interactively associated with somatic alterations in gastric cancer. *Cancer Med* **7**, 3912-3920, doi:10.1002/cam4.1612 (2018).

52 Knudson, A. G., Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 820-823, doi:10.1073/pnas.68.4.820 (1971).

53 Godbout, R., Dryja, T. P., Squire, J., Gallie, B. L. & Phillips, R. A. Somatic inactivation of genes on chromosome 13 is a common event in retinoblastoma. *Nature* **304**, 451-453, doi:10.1038/304451a0 (1983).

54 Klein, G. The approaching era of the tumor suppressor genes. *Science* **238**, 1539-1545, doi:10.1126/science.3317834 (1987).

55 Lu, C. *et al.* Patterns and functional implications of rare germline variants across 12 cancer types. *Nat Commun* **6**, 10086, doi:10.1038/ncomms10086 (2015).

56 Campbell, P. J. Somatic and germline genetics at the JAK2 locus. *Nat Genet* **41**, 385-386, doi:10.1038/ng0409-385 (2009).

57 Liu, W. *et al.* Functional EGFR germline polymorphisms may confer risk for EGFR somatic mutations in non-small cell lung cancer, with a predominant effect on exon 19 microdeletions. *Cancer Res* **71**, 2423-2427, doi:10.1158/0008-5472.CAN-10-2689 (2011).

58 Shu, X. *et al.* Germline genetic variants in somatically significantly mutated genes in tumors are associated with renal cell carcinoma risk and outcome. *Carcinogenesis* **39**, 752-757, doi:10.1093/carcin/bgy021 (2018).

59 Puzone, R. & Pfeffer, U. SNP variants at the MAP3K1/SETD9 locus 5q11.2 associate with somatic PIK3CA variants in breast cancers. *Eur J Hum Genet* **25**, 384-387, doi:10.1038/ejhg.2016.179 (2017).

60 Carter, H. *et al.* Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer. *Cancer Discov* **7**, 410-423, doi:10.1158/2159-8290.CD-16-1045 (2017).

61 Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886-D894, doi:10.1093/nar/gky1016 (2019).

62 Kao, P. Y., Leung, K. H., Chan, L. W., Yip, S. P. & Yap, M. K. Pathway analysis of complex diseases for GWAS, extending to consider rare variants, multi-omics and interactions. *Biochim Biophys Acta Gen Subj* **1861**, 335-353, doi:10.1016/j.bbagen.2016.11.030 (2017).

63 Cancer Genome Atlas Research, N. *et al.* Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169-175, doi:10.1038/nature20805 (2017).

64 Contino, G., Vaughan, T. L., Whiteman, D. & Fitzgerald, R. C. The Evolving Genomic Landscape of Barrett's Esophagus and Esophageal Adenocarcinoma. *Gastroenterology* **153**, 657-673 e651, doi:10.1053/j.gastro.2017.07.007 (2017).

65 Lynch, H. T. *et al.* Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clin Genet* **76**, 1-18, doi:10.1111/j.1399-0004.2009.01230.x (2009).

66 Petrucelli, N., Daly, M. B. & Feldman, G. L. Hereditary breast and ovarian cancer due to mutations in BRCA1 and BRCA2. *Genet Med* **12**, 245-259, doi:10.1097/GIM.0b013e3181d38f2f (2010).

67 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311, doi:10.1093/nar/29.1.308 (2001).

68 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

69 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443, doi:10.1038/s41586-020-2308-7 (2020).

70 Yuan, J. *et al.* Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers. *Cancer Cell* **34**, 549-560 e549, doi:10.1016/j.ccell.2018.08.019 (2018).

71 McGranahan, N. *et al.* Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell* **171**, 1259-1271 e1211, doi:10.1016/j.cell.2017.10.001 (2017).

72 Marty, R. *et al.* MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell* **171**, 1272-1283 e1215, doi:10.1016/j.cell.2017.09.050 (2017).

73 Secrier, M. *et al.* Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nat Genet* **48**, 1131-1141, doi:10.1038/ng.3659 (2016).

74 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164, doi:10.1093/nar/gkq603 (2010).

75 O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745, doi:10.1093/nar/gkv1189 (2016).

76 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).

77 Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* **37**, 235-241, doi:10.1002/humu.22932 (2016).

78 Jian, X., Boerwinkle, E. & Liu, X. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res* **42**, 13534-13544, doi:10.1093/nar/gku1206 (2014).

79 Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* **24**, 2125-2137, doi:10.1093/hmg/ddu733 (2015).

80 Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-3814, doi:10.1093/nar/gkg509 (2003).

81 Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20, doi:10.1002/0471142905.hg0720s76 (2013).

82 Schwarz, J. M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**, 575-576, doi:10.1038/nmeth0810-575 (2010).

83 Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* **39**, e118, doi:10.1093/nar/gkr407 (2011).

84 Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res* **19**, 1553-1561, doi:10.1101/gr.092619.109 (2009).

85 Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* **34**, 57-65, doi:10.1002/humu.22225 (2013).

86 Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-121, doi:10.1101/gr.097857.109 (2010).

87 Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54-62, doi:10.1093/bioinformatics/btp190 (2009).

88 Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025, doi:10.1371/journal.pcbi.1001025 (2010).

89 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B* **57** (1995).

90 Arnold, M., Soerjomataram, I., Ferlay, J. & Forman, D. Global incidence of oesophageal cancer by histological subtype in 2012. *Gut* **64**, 381-387, doi:10.1136/gutjnl-2014-308124 (2015).

91 Xie, S. H. & Lagergren, J. The Male Predominance in Esophageal Adenocarcinoma. *Clin Gastroenterol Hepatol* **14**, 338-347 e331, doi:10.1016/j.cgh.2015.10.005 (2016).

92 Torre, L. A. *et al.* Global cancer statistics, 2012. *CA Cancer J Clin* **65**, 87-108, doi:10.3322/caac.21262 (2015).

93 Coleman, H. G., Xie, S. H. & Lagergren, J. The Epidemiology of Esophageal Adenocarcinoma. *Gastroenterology* **154**, 390-405, doi:10.1053/j.gastro.2017.07.046 (2018).

94 Cook, M. B. *et al.* Gastroesophageal reflux in relation to adenocarcinomas of the esophagus: a pooled analysis from the Barrett's and Esophageal Adenocarcinoma Consortium (BEACON). *PLoS One* **9**, e103508, doi:10.1371/journal.pone.0103508 (2014).

95 Cortes-Ciriano, I. *et al.* Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet* **52**, 331-341, doi:10.1038/s41588-019-0576-7 (2020).

96 Frankell, A. M. *et al.* The landscape of selection in 551 esophageal adenocarcinomas defines genomic biomarkers for the clinic. *Nat Genet* **51**, 506-516, doi:10.1038/s41588-018-0331-5 (2019).

97 Agrawal, N. *et al.* Comparative genomic analysis of esophageal adenocarcinoma and squamous cell carcinoma. *Cancer Discov* **2**, 899-905, doi:10.1158/2159-8290.CD-12-0189 (2012).

98 Dulak, A. M. *et al.* Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat Genet* **45**, 478-486, doi:10.1038/ng.2591 (2013).

99 Fels Elliott, D. R. *et al.* Impact of mutations in Toll-like receptor pathway genes on esophageal carcinogenesis. *PLoS Genet* **13**, e1006808, doi:10.1371/journal.pgen.1006808 (2017).

100 Weaver, J. M. J. *et al.* Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nat Genet* **46**, 837-843, doi:10.1038/ng.3013 (2014).

101 Ek, W. E. *et al.* Germline genetic contributions to risk for esophageal adenocarcinoma, Barrett's esophagus, and gastroesophageal reflux. *J Natl Cancer Inst* **105**, 1711-1718, doi:10.1093/jnci/djt303 (2013).

102 Kunzmann, A. T. *et al.* Information on Genetic Variants Does Not Increase Identification of Individuals at Risk of Esophageal Adenocarcinoma Compared to Clinical Risk Factors. *Gastroenterology* **156**, 43-45, doi:10.1053/j.gastro.2018.09.038 (2019).

103 Graffelman, J., Jain, D. & Weir, B. A genome-wide study of Hardy-Weinberg equilibrium with next generation sequence data. *Hum Genet* **136**, 727-741, doi:10.1007/s00439-017-1786-7 (2017).

104 Cereda, M. *et al.* Patients with genetically heterogeneous synchronous colorectal cancer carry rare damaging germline mutations in immune-related genes. *Nat Commun* **7**, 12072, doi:10.1038/ncomms12072 (2016).

105 Ka, S. *et al.* HLAscan: genotyping of the HLA region using next-generation sequencing data. *BMC Bioinformatics* **18**, 258, doi:10.1186/s12859-017-1671-3 (2017).

106 Williams, T. M. Human leukocyte antigen gene polymorphism and the histocompatibility laboratory. *J Mol Diagn* **3**, 98-104, doi:10.1016/S1525-1578(10)60658-7 (2001).

107 Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* **48**, D498-D503, doi:10.1093/nar/gkz1031 (2020).

108 Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353-D361, doi:10.1093/nar/gkw1092 (2017).

109 Nishimura, D. BioCarta. *Biotech Softw. Internet Rep.* **2** (2001).

110 Gambardella, G., Cereda, M., Benedetti, L. & Ciccarelli, F. D. MEGA-V: detection of variant gene sets in patient cohorts. *Bioinformatics* **33**, 1248-1249, doi:10.1093/bioinformatics/btw809 (2017).

111 Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**, 696-705, doi:10.1038/s41568-018-0060-1 (2018).

112 Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)* **95**, 221-227, doi:10.1038/sj.hdy.6800717 (2005).

113 Downward, J. Targeting RAS and PI3K in lung cancer. *Nat Med* **14**, 1315-1316, doi:10.1038/nm1208-1315 (2008).

114 Lemmon, M. A. & Schlessinger, J. Cell signaling by receptor tyrosine kinases. *Cell* **141**, 1117-1134, doi:10.1016/j.cell.2010.06.011 (2010).

115 Valiathan, R. R., Marco, M., Leitinger, B., Kleer, C. G. & Fridman, R. Discoidin domain receptor tyrosine kinases: new players in cancer progression. *Cancer Metastasis Rev* **31**, 295-321, doi:10.1007/s10555-012-9346-z (2012).

116 Wasinski, B. *et al.* Discoidin Domain Receptors, DDR1b and DDR2, Promote Tumour Growth within Collagen but DDR1b Suppresses Experimental Lung Metastasis in HT1080 Xenografts. *Sci Rep* **10**, 2309, doi:10.1038/s41598-020-59028-w (2020).

117 Lafitte, M., Sirvent, A. & Roche, S. Collagen Kinase Receptors as Potential Therapeutic Targets in Metastatic Colon Cancer. *Front Oncol* **10**, 125, doi:10.3389/fonc.2020.00125 (2020).

118 Cruz da Silva, E., Dontenwill, M., Choulier, L. & Lehmann, M. Role of Integrins in Resistance to Therapies Targeting Growth Factor Receptors in Cancer. *Cancers* **11** (2019).

119 Desgrosellier, J. S. & Cheresh, D. A. Integrins in cancer: biological implications and therapeutic opportunities. *Nat Rev Cancer* **10**, 9-22, doi:10.1038/nrc2748 (2010).

120 Bromann, P. A., Korkaya, H. & Courtneidge, S. A. The interplay between Src family kinases and receptor tyrosine kinases. *Oncogene* **23**, 7957-7968, doi:10.1038/sj.onc.1208079 (2004).

121 Hornsveld, M., Dansen, T. B., Derksen, P. W. & Burgering, B. M. T. Re-evaluating the role of FOXOs in cancer. *Semin Cancer Biol* **50**, 90-100, doi:10.1016/j.semcancer.2017.11.017 (2018).

122 Musgrove, E. A., Caldon, C. E., Barraclough, J., Stone, A. & Sutherland, R. L. Cyclin D as a therapeutic target in cancer. *Nat Rev Cancer* **11**, 558-572, doi:10.1038/nrc3090 (2011).

123 Karantza, V. Keratins in health and cancer: more than mere epithelial cell markers. *Oncogene* **30**, 127-138, doi:10.1038/onc.2010.456 (2011).

124 Jain, S. & Dhingra, S. Pathology of esophageal cancer and Barrett's esophagus. *Ann Cardiothorac Surg* **6**, 99-109, doi:10.21037/acs.2017.03.06 (2017).

125 Armes, J. E. *et al.* The histologic phenotypes of breast carcinoma occurring before age 40 years in women with and without BRCA1 or BRCA2 germline mutations: a population-based study. *Cancer* **83**, 2335-2345 (1998).

126 Bannon, S. A. *et al.* High Prevalence of Hereditary Cancer Syndromes and Outcomes in Adults with Early-Onset Pancreatic Cancer. *Cancer Prev Res (Phila)* **11**, 679-686, doi:10.1158/1940-6207.CAPR-18-0014 (2018).

127 Kharazmi, E., Fallah, M., Sundquist, K. & Hemminki, K. Familial risk of early and late onset cancer: nationwide prospective cohort study. *BMJ* **345**, e8076, doi:10.1136/bmj.e8076 (2012).

128 Mauri, G. *et al.* Early-onset colorectal cancer in young individuals. *Mol Oncol* **13**, 109-131, doi:10.1002/1878-0261.12417 (2019).

129 Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet* **19**, 807-812, doi:10.1038/ejhg.2011.39 (2011).

130 Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-16915, doi:10.1073/pnas.1009843107 (2010).

131 Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med* **10**, 33, doi:10.1186/s13073-018-0539-0 (2018).

132 Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2**, 18-22 (2002).

133 cluster: Cluster analysis basics and extensions (R package version 2.1.0, 2019).

134 L., S., M., F., T., M. & A., R. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal* **8**, 205-233 (2016).

135 Ireland, A. P. *et al.* Clinical significance of p53 mutations in adenocarcinoma of the esophagus and cardia. *Ann Surg* **231**, 179-187, doi:10.1097/00000658-200002000-00005 (2000).

136 Freedman, N. D. *et al.* A prospective study of tobacco, alcohol, and the risk of esophageal and gastric cancer subtypes. *Am J Epidemiol* **165**, 1424-1433, doi:10.1093/aje/kwm051 (2007).

137 Lagergren, J., Bergstrom, R., Lindgren, A. & Nyren, O. Symptomatic gastroesophageal reflux as a risk factor for esophageal adenocarcinoma. *N Engl J Med* **340**, 825-831, doi:10.1056/NEJM199903183401101 (1999).

138 Statistics, O. f. N. *Adult smoking habits in the UK: 2019*, <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/bulletins/adultsmokinghabitsingreatbritain/2019> (2019).

139 Altorki, N. & Harrison, S. What is the role of neoadjuvant chemotherapy, radiation, and adjuvant treatment in resectable esophageal cancer? *Ann Cardiothorac Surg* **6**, 167-174, doi:10.21037/acs.2017.03.16 (2017).

140 Pich, O. *et al.* The mutational footprints of cancer therapies. *Nat Genet* **51**, 1732-1740, doi:10.1038/s41588-019-0525-5 (2019).

141 Noorani, A. *et al.* A comparative analysis of whole genome sequencing of esophageal adenocarcinoma pre- and post-chemotherapy. *Genome Res* **27**, 902-912, doi:10.1101/gr.214296.116 (2017).

142 Schnidrig, D., Turajlic, S. & Litchfield, K. Tumour mutational burden: primary versus metastatic tissue creates systematic bias. *Immuno-Oncology Tech.* **4**, 8-14 (2019).

143 Peterson, R. E. *et al.* The utility of empirically assigning ancestry groups in cross-population genetic studies of addiction. *Am J Addict* **26**, 494-501, doi:10.1111/ajad.12586 (2017).

144 Qing, T. *et al.* Germline variant burden in cancer genes correlates with age at diagnosis and somatic mutation burden. *Nat Commun* **11**, 2438, doi:10.1038/s41467-020-16293-7 (2020).

145 Li, G. M. Mechanisms and functions of DNA mismatch repair. *Cell Res* **18**, 85-98, doi:10.1038/cr.2007.115 (2008).

146 Powell, S. N. & Kachnic, L. A. Roles of BRCA1 and BRCA2 in homologous recombination, DNA replication fidelity and the cellular response to ionizing radiation. *Oncogene* **22**, 5784-5791, doi:10.1038/sj.onc.1206678 (2003).

147 Aggarwal, M., Sommers, J. A., Shoemaker, R. H. & Brosh, R. M., Jr. Inhibition of helicase activity by a small molecule impairs Werner syndrome helicase (WRN) function in the cellular response to DNA damage or replication stress. *Proc Natl Acad Sci U S A* **108**, 1525-1530, doi:10.1073/pnas.1006423108 (2011).

148 Testa, U., Castelli, G. & Pelosi, E. Esophageal Cancer: Genomic and Molecular Characterization, Stem Cell Compartment and Clonal Evolution. *Medicines (Basel)* **4** (2017).

149 Urbina-Jara, L. K. *et al.* Landscape of Germline Mutations in DNA Repair Genes for Breast Cancer in Latin America: Opportunities for PARP-Like Inhibitors and Immunotherapy. *Genes (Basel)* **10**, doi:10.3390/genes10100786 (2019).

150 Zhao, P., Li, L., Jiang, X. & Li, Q. Mismatch repair deficiency/microsatellite instability-high as a predictor for anti-PD-1/PD-L1 immunotherapy efficacy. *J Hematol Oncol* **12**, 54, doi:10.1186/s13045-019-0738-1 (2019).

151 Jacob, L., Obozinski, G. & Vert, J. P. in *26th International Conference on Machine Learning.*

152 Zeng, Y. & Breheny, P. Overlapping Group Logistic Regression with Applications to Genetic Pathway Selection. *Cancer Inform* **15**, 179-187, doi:10.4137/CIN.S40043 (2016).

153 Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. B* **58**, 267-288 (1996).

154 Munch, M. M., Peeters, C. F. W., Van Der Vaart, A. W. & Van De Wiel, M. A. Adaptive group-regularized logistic elastic net regression. *Biostatistics*, doi:10.1093/biostatistics/kxz062 (2019).

155 Cule, E., Vineis, P. & De Iorio, M. Significance testing in ridge regression for genetic data. *BMC Bioinformatics* **12**, 372, doi:10.1186/1471-2105-12-372 (2011).

156 Lockhart, R., Taylor, J., Tibshirani, R. J. & Tibshirani, R. A Significance Test for the Lasso. *Ann Stat* **42**, 413-468, doi:10.1214/13-AOS1175 (2014).

157 Benjamini, Y. & Bogomolov, M. Selective inference on multiple families of hypotheses. *J. Royal Stat. Soc. B* **76**, 297-318 (2013).

158 Szumilas, M. Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry* **19**, 227-229 (2010).

159 Stein, J. L. *et al.* Voxelwise genome-wide association study (vGWAS). *Neuroimage* **53**, 1160-1174, doi:10.1016/j.neuroimage.2010.02.032 (2010).

160 Simes, R. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751-754 (1986).

161 Zhao, M., Mishra, L. & Deng, C. X. The role of TGF-beta/SMAD4 signaling in cancer. *Int J Biol Sci* **14**, 111-123, doi:10.7150/ijbs.23230 (2018).

162 Budi, E. H., Muthusamy, B. P. & Derynck, R. The insulin response integrates increased TGF-beta signaling through Akt-induced enhancement of cell surface delivery of TGF-beta receptors. *Sci Signal* **8**, ra96, doi:10.1126/scisignal.aaa9432 (2015).

163 Yuen, D. A. *et al.* Recombinant N-Terminal Slit2 Inhibits TGF-beta-Induced Fibroblast Activation and Renal Fibrosis. *J Am Soc Nephrol* **27**, 2609-2615, doi:10.1681/ASN.2015040356 (2016).

164 Pinho, A. V. *et al.* ROBO2 is a stroma suppressor gene in the pancreas and acts via TGF-beta signalling. *Nat Commun* **9**, 5083, doi:10.1038/s41467-018-07497-z (2018).

165 Rooman, I., Pinho, A., Phillips, P., Biankin, A. V. & Wu, J. Altered Slit-Robo signalling activates TGF-beta signalling and stroma remodelling in the pancreas. *Pancreatology* **17**, S6-S7 (2017).

166 Cremona, C. A. & Behrens, A. ATM signalling and cancer. *Oncogene* **33**, 3351-3360, doi:10.1038/onc.2013.275 (2014).

167 Ahmed, M. & Rahman, N. ATM and breast cancer susceptibility. *Oncogene* **25**, 5906-5911, doi:10.1038/sj.onc.1209873 (2006).

168 Gronbaek, K. *et al.* ATM mutations are associated with inactivation of the ARF-TP53 tumor suppressor pathway in diffuse large B-cell lymphoma. *Blood* **100**, 1430-1437, doi:10.1182/blood-2002-02-0382 (2002).

169 Roos, W. P. & Kaina, B. DNA damage-induced cell death: from specific DNA lesions to the DNA damage response and apoptosis. *Cancer Lett* **332**, 237-248, doi:10.1016/j.canlet.2012.01.007 (2013).

170 Norbury, C. J. & Zhivotovsky, B. DNA damage-induced apoptosis. *Oncogene* **23**, 2797-2808, doi:10.1038/sj.onc.1207532 (2004).

171 Hayakawa, Y., Sethi, N., Sepulveda, A. R., Bass, A. J. & Wang, T. C. Oesophageal adenocarcinoma and gastric cancer: should we mind the gap? *Nat Rev Cancer* **16**, 305-318, doi:10.1038/nrc.2016.24 (2016).

172 Weigelt, B. *et al.* The Landscape of Somatic Genetic Alterations in Breast Cancers From ATM Germline Mutation Carriers. *J Natl Cancer Inst* **110**, 1030-1034, doi:10.1093/jnci/djy028 (2018).

173 Ehrlich, L. A., Yang-Iott, K., DeMicco, A. & Bassing, C. H. Somatic inactivation of ATM in hematopoietic cells predisposes mice to cyclin D3 dependent T cell acute lymphoblastic leukemia. *Cell Cycle* **14**, 388-398, doi:10.4161/15384101.2014.988020 (2015).

174 Menter, T. *et al.* Mutational landscape of B-cell post-transplant lymphoproliferative disorders. *Br J Haematol* **178**, 48-56, doi:10.1111/bjh.14633 (2017).

175 Ahmad, A. & Khan, S. Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access* **7** (2019).

176 Ho, T. K. in *3rd International Conference on Document Analysis and Recognition.* (IEEE).

177 Chen, X. & Ishwaran, H. Random forests for genomic data analysis. *Genomics* **99**, 323-329, doi:10.1016/j.ygeno.2012.04.003 (2012).

178 Liu, S., Chen, Y. & Wilkins, D. Large margin classifiers and Random Forests for integrated biological prediction. *Int J Bioinform Res Appl* **8**, 38-53, doi:10.1504/IJBRA.2012.045975 (2012).

179 Shi, T. & Horvath, S. Unsupervized learning with random forest predictors. *J. Comp. Graph. Stat.* **15**, 118-138 (2006).

180 Mushtaq, H. *et al.* A Parallel Architecture for the Partitioning Around Medoids (PAM) Algorithm for Scalable Multi-Core Processor Implementation with Applications in Healthcare. *Sensors (Basel)* **18**, doi:10.3390/s18124129 (2018).

181 Rousseeuw, P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comp. Appl. Math.* **20**, 53-65 (1987).

182 Santos, J. & Embrechts, M. in *International Conference on Artificial Neural Networks.* (Springer).

183 Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360-364, doi:10.1038/nature14221 (2015).

184 Lecona, E. & Fernandez-Capetillo, O. Targeting ATR in cancer. *Nat Rev Cancer* **18**, 586-595, doi:10.1038/s41568-018-0034-3 (2018).

185 Syed, A. S., D'Antonio, M. & Ciccarelli, F. D. Network of Cancer Genes: a web resource to analyze duplicability, orthology and network properties of cancer genes. *Nucleic Acids Res* **38**, D670-675, doi:10.1093/nar/gkp957 (2010).

186 An, O. *et al.* NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. *Database (Oxford)* **2014**, bau015, doi:10.1093/database/bau015 (2014).

187 D'Antonio, M., Pendino, V., Sinha, S. & Ciccarelli, F. D. Network of Cancer Genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes. *Nucleic Acids Res* **40**, D978-983, doi:10.1093/nar/gkr952 (2012).

188 Cameron, D. *et al.* 11 years' follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive early breast cancer: final analysis of the HERceptin Adjuvant (HERA) trial. *Lancet* **389**, 1195-1205, doi:10.1016/S0140-6736(16)32616-2 (2017).

189 Hochhaus, A. *et al.* European LeukemiaNet 2020 recommendations for treating chronic myeloid leukemia. *Leukemia* **34**, 966-984, doi:10.1038/s41375-020-0776-2 (2020).

190 Zhao, M., Kim, P., Mitra, R., Zhao, J. & Zhao, Z. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Res* **44**, D1023-1031, doi:10.1093/nar/gkv1268 (2016).

191 Liu, Y., Sun, J. & Zhao, M. ONGene: A literature-based database for human oncogenes. *J Genet Genomics* **44**, 119-121, doi:10.1016/j.jgg.2016.12.004 (2017).

192 Agarwal, R., Kumar, B., Jayadev, M., Raghav, D. & Singh, A. CoReCG: a comprehensive database of genes associated with colon-rectal cancer. *Database (Oxford)* **2016**, doi:10.1093/database/baw059 (2016).

193 D'Antonio, M. & Ciccarelli, F. D. Modification of gene duplicability during the evolution of protein interaction network. *PLoS Comput Biol* **7**, e1002029, doi:10.1371/journal.pcbi.1002029 (2011).

194 Rambaldi, D., Giorgi, F. M., Capuani, F., Ciliberto, A. & Ciccarelli, F. D. Low duplicability and network fragility of cancer genes. *Trends Genet* **24**, 427-430, doi:10.1016/j.tig.2008.06.003 (2008).

195 D'Antonio, M. & Ciccarelli, F. D. Integrated analysis of recurrent properties of cancer genes to identify novel drivers. *Genome Biol* **14**, R52, doi:10.1186/gb-2013-14-5-r52 (2013).

196 Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419, doi:10.1126/science.1260419 (2015).

197 Consortium, G. T. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213, doi:10.1038/nature24277 (2017).

198 Lenoir, W. F., Lim, T. L. & Hart, T. PICKLES: the database of pooled in-vitro CRISPR knockout library essentiality screens. *Nucleic Acids Res* **46**, D776-D780, doi:10.1093/nar/gkx993 (2018).

199 Chen, W. H., Lu, G., Chen, X., Zhao, X. M. & Bork, P. OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Res* **45**, D940-D944, doi:10.1093/nar/gkw1013 (2017).

200 Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2017 update. *Nucleic Acids Res* **45**, D369-D379, doi:10.1093/nar/gkw1102 (2017).

201 Orchard, S. *et al.* The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* **42**, D358-363, doi:10.1093/nar/gkt1115 (2014).

202 Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* **32**, D449-451, doi:10.1093/nar/gkh086 (2004).

203 Keshava Prasad, T. S. *et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Res* **37**, D767-772, doi:10.1093/nar/gkn892 (2009).

204 Ruepp, A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes--2009. *Nucleic Acids Res* **38**, D497-501, doi:10.1093/nar/gkp914 (2010).

205 Chou, C. H. *et al.* miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res* **46**, D296-D302, doi:10.1093/nar/gkx1067 (2018).

206 Xiao, F. *et al.* miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res* **37**, D105-110, doi:10.1093/nar/gkn851 (2009).

207 Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* **44**, D286-293, doi:10.1093/nar/gkv1248 (2016).

208 Nakatani, Y., Takeda, H., Kohara, Y. & Morishita, S. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res* **17**, 1254-1265, doi:10.1101/gr.6316407 (2007).

209 Mitchell, A. L. *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res* **47**, D351-D360, doi:10.1093/nar/gky1100 (2019).

210 The UniProt, C. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**, D158-D169, doi:10.1093/nar/gkw1099 (2017).

211 Futreal, P. A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177-183, doi:10.1038/nrc1299 (2004).

212 Dees, N. D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Res* **22**, 1589-1598, doi:10.1101/gr.134635.111 (2012).

213 Grzywa, T. M., Paskal, W. & Wlodarski, P. K. Intratumor and Intertumor Heterogeneity in Melanoma. *Transl Oncol* **10**, 956-975, doi:10.1016/j.tranon.2017.09.007 (2017).

214 Phan, L. M., Yeung, S. C. & Lee, M. H. Cancer metabolic reprogramming: importance, main features, and potentials for precise targeted anti-cancer therapies. *Cancer Biol Med* **11**, 1-19, doi:10.7497/j.issn.2095-3941.2014.01.001 (2014).

215 Lu, M. & Zhan, X. The crucial role of multiomic approach in cancer research and clinically relevant outcomes. *EPMA J* **9**, 77-102, doi:10.1007/s13167-018-0128-8 (2018).

216 Armenia, J. *et al.* The long tail of oncogenic drivers in prostate cancer. *Nat Genet* **50**, 645-651, doi:10.1038/s41588-018-0078-z (2018).

217 Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041 e1021, doi:10.1016/j.cell.2017.09.042 (2017).

218 Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41, doi:10.1186/gb-2011-12-4-r41 (2011).

219 Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238-2244, doi:10.1093/bioinformatics/btt395 (2013).

220 Reimand, J. & Bader, G. D. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol Syst Biol* **9**, 637, doi:10.1038/msb.2012.68 (2013).

221 Davoli, T. *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948-962, doi:10.1016/j.cell.2013.10.011 (2013).

222 Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci U S A* **113**, 14330-14335, doi:10.1073/pnas.1616440113 (2016).

223 Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res* **40**, e169, doi:10.1093/nar/gks743 (2012).

224 Leiserson, M. D. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* **47**, 106-114, doi:10.1038/ng.3168 (2015).

225 Bertrand, D. *et al.* Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Res* **43**, e44, doi:10.1093/nar/gku1393 (2015).

226 Bashashati, A. *et al.* DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol* **13**, R124, doi:10.1186/gb-2012-13-12-r124 (2012).

227 Hou, J. & Ma, J. DawnRank: discovering personalized driver genes in cancer. *Genome Medicine* **6**, doi:10.1186/s13073-014-0056-8 (2014).

228 Van Allen, E. M. *et al.* Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med* **20**, 682-688, doi:10.1038/nm.3559 (2014).

229 Dong, C. *et al.* iCAGES: integrated CAncer GEnome Score for comprehensively prioritizing driver genes in personal cancer genomes. *Genome Med* **8**, 135, doi:10.1186/s13073-016-0390-0 (2016).

230 Scholkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J. & Williamson, R. C. Estimating the support of a high-dimensional distribution. *Neural Comput* **13**, 1443-1471, doi:10.1162/089976601750264965 (2001).

231 e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071) (Vienna University of Technology, Vienna, Austria, 2019).

232 Webber, W., Moffat, A. & Zobel, J. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems* **28**, doi:10.1145/1852102.1852106. (2010).

233 Abadi, M. *et al.* in *12th USENIX Symposium on Operating Systems Design and Implementation* (PeerJ, Savannah, GA, USA, 2016).

234 Wang, S., Liu, Q., Zhu, E., Porikli, F. & Yin, J. Hyperparameter selection of one-class support vector machine by self-adaptive data shifting. *Pattern Recognition* **74**, 198-211, doi:10.1016/j.patcog.2017.09.012 (2018).

235 Weston, J. *et al.* in *Conference on Neural Information Processing Systems* (2001).

236 Morgan, M. A. & Shilatifard, A. Chromatin signatures of cancer. *Genes Dev* **29**, 238-249, doi:10.1101/gad.255182.114 (2015).

237 Helleday, T., Petermann, E., Lundin, C., Hodgson, B. & Sharma, R. A. DNA repair pathways as targets for cancer therapy. *Nat Rev Cancer* **8**, 193-204, doi:10.1038/nrc2342 (2008).

238 Otto, T. & Sicinski, P. Cell cycle proteins as promising targets in cancer therapy. *Nat Rev Cancer* **17**, 93-115, doi:10.1038/nrc.2016.138 (2017).

239 Sever, R. & Brugge, J. S. Signal transduction in cancer. *Cold Spring Harb Perspect Med* **5**, doi:10.1101/cshperspect.a006098 (2015).

240 Chen, X. *et al.* Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma. *Cell Rep* **7**, 104-112, doi:10.1016/j.celrep.2014.03.003 (2014).

241 Kovac, M. *et al.* Exome sequencing of osteosarcoma reveals mutation signatures reminiscent of BRCA deficiency. *Nat Commun* **6**, 8940, doi:10.1038/ncomms9940 (2015).

242 Kegelman, C. D. *et al.* Skeletal cell YAP and TAZ combinatorially promote bone development. *FASEB J* **32**, 2706-2721, doi:10.1096/fj.201700872R (2018).

243 Pan, J. X. *et al.* YAP promotes osteogenesis and suppresses adipogenic differentiation by regulating beta-catenin signaling. *Bone Res* **6**, 18, doi:10.1038/s41413-018-0018-7 (2018).

244 Kingma, D. & Welling, M. in *International Conference on Learning Representations* (2014).

245 Kaeding, C., Rodner, E., Freytag, A. & Denzler, J. Active and Continuous Exploration with Deep Neural Networks and Expected Model Output Changes. *arXiv:1612.06129 [cs.CV]* (2016).

246 Kramer, A. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.* **37**, 223-243 (1991).

247 Xia, Y., Cao, X., Wen, F., Hua, G. & Sun, J. in *International Conference on Computer Vision* (IEEE, Santiago, Chile, 2015).

248 Sabokrou, M., Fathy, M. & Hoseini, M. Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. *Electronics Letters* **52**, 1122-1124 (2016).

249 Zhang, Y. & Yang, Q. A Survey on Multi-Task Learning. *arXiv:1707.08114v2 [cs.LG]* (2018).

250 Driehuis, E. *et al.* Pancreatic cancer organoids recapitulate disease and allow personalized drug screening. *Proc Natl Acad Sci U S A*, doi:10.1073/pnas.1911273116 (2019).

251 Franks, A. L. & Slansky, J. E. Multiple associations between a broad spectrum of autoimmune diseases, chronic inflammatory diseases and cancer. *Anticancer Res* **32**, 1119-1136 (2012).

252 Wang, G. *et al.* Oncogenic driver genes and tumor microenvironment determine the type of liver cancer. *Cell Death Dis* **11**, 313, doi:10.1038/s41419-020-2509-x (2020).

253 Walter, M. J. *et al.* Next-generation sequencing of cancer genomes: back to the future. *Per Med* **6**, 653, doi:10.2217/pme.09.52 (2009).