Identification of a new genetic cause of cholestatic liver disease

Sambrotta, Melissa

Awarding institution:
King's College London

Identification of a new genetic cause of cholestatic liver disease

Melissa Sambrotta

Thesis submitted to the School of Medicine
at King's College London for the degree of Doctor of Philosophy

Institute of Liver Studies
Division of Transplantation Immunology & Mucosal Biology
King's College London School of Medicine

Acknowledgements

This PhD project is a result of a challenging journey, in which many people have contributed and given support.

Foremost, I would like to thank my supervisor Dr Richard Thompson, for his constant and valuable guidance received during these years. I feel that I have achieved a great deal, and it would have not been possible without his trust and motivation. He has been and will always be a great mentor for me. I would also like to thank my second supervisor, Prof Giorgina Mieli-Vergani, for giving me the opportunity to come, in the first place, to the Institute of Liver Studies as an MSc student. I have greatly appreciated the good advice she has provided in completing my PhD thesis. I would like to thank the Alex Mowat studentship that supported my PhD fees and living.

I want to express my gratitude to Dr Efterpi Papouli for helping me during the library preparation for next-generation sequencing; Dr Alex Knisely for his histopathological consultancy and my colleague, housemate and friend Dr Lucy Newbury for teaching me the immunoblotting and for our several dinner conversations about real-time PCR. In addition, I would like to thank the past and present members of the Liver Molecular Genetics lab, including Sandra, Pierre, Laura and Jeid; my collaborators, Dr Barnaby Clark and Peter Rushton from the Department of Molecular Haematology at King's College Hospital and Dr Michael Simpson from the Department of Molecular Genetics at King's College London.

During these years, I have had the pleasure to meet numerous colleagues and friends. My sincere thanks goes to you all for being part of this journey.

Finally, I would like to dedicate this thesis to my family and my partner for their unconditional support that let me follow my dream.

Abstract

Cholestatic liver diseases are defined by impairment of bile flow or bile formation. Progressive familial intrahepatic cholestasis (PFIC) is a group of cholestatic disorders, so far associated with three genes encoding canalicular membrane transporters; nevertheless, one third of patients with progressive intrahepatic cholestasis remain without an aetiology. Targeted and/or whole-exome sequencing was undertaken in 83 families, mostly consanguineous, with no mutations in known PFIC genes.

Homozygous mutations in tight junction protein 2 (*TJP2*) were identified in 21 individuals of 15 families. Most were predicted to be protein-truncating, and most patients had severe liver disease requiring liver transplantation. Some also had extrahepatic manifestations. Four individuals were found to carry the same homozygous missense mutation, three had late-onset and remittent cholestasis, one was asymptomatic.

TJP2, also known as zona occludens-2 (ZO-2), encodes a cytosolic component of cell-cell junctional structures. Patients with severe disease had no ZO-2 protein. The presence, and distribution of integral tight junction protein, claudin-1, was found to be disrupted. Tight junction structure was abnormal on transmission electron microscopy.

The absence of ZO-2 might have been compensated by other junctional components. The expression of tight-junction-related genes was analysed in TJP2 deficiency patients; however significant, biological-relevant, changes were not identified.

It appears, therefore, that the complete absence of TJP2 causes disruption of tight junction structures and severe cholestatic liver disease, whilst missense mutations in TJP2 lead to less severe phenotypes.

In addition whole-exome sequencing analysis revealed one patient with a novel homozygous missense mutation in the gene α-methylacyl-CoA racemase (*AMACR*). The change was predicted to be deleterious for the encoded protein. The finding was also supported by the clinical manifestation of this rare metabolic disorder with early-onset cholestasis.

A failure in the identification of the causative mutations occurred for the other 5 patients where whole-exome sequencing was performed, possibly due to limitations in the methods used.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A1AT	α-1-antritrypsin
AATD	A1AT deficiency
ABC	ATP-binding cassette
ABCB11	ATP-binding cassette subfamily B member 11
ABCB4	ATP-binding cassette subfamily B member 4
ABCC2	ATP-binding cassette subfamily C member 2
ABCC7	ATP-binding cassette subfamily C member 7
ABCG5	ATP-binding cassette subfamily G member 5
ABCG8	ATP-binding cassette subfamily G member 8
ACBT	β-actin
ADAM	a disintegrin and metalloproteinase
ADNSHL	autosomal dominant non-syndromic hearing loss
AE2	anion exchanger 2
AFA	Adaptive Focused Acoustics
AGS	Alagille syndrome
AKR1D1	δ (4)-3-oxosteroid 5-β-reductase
ALT	alanine aminotransferase
AMACR	α-methylacyl-CoA racemase
AP	alkaline phosphatase
ARC	arthrogryposis renal dysfunction and cholestasis
AST	aspartate aminotransferase
ATP	adenosine triphosphate
ATP11C	ATPase type 11c
ATP8B1	phospholipid-transporting ATPase type 8b member 1

BA	biliary atresia
BAAT	bile acid CoA: amino acid N-acyltransferase
BAM	Binary Alignment Map
BF	Bayes factor
bp	base pairs
BR	broad-range
BRIC	benign recurrent intrahepatic cholestasis
BSA	bovine serum albumin
BSEP	bile salt export pump
CA	cholic acid
cAMP	cyclic adenosine monophosphate
CASAVA	Consensus Assessment of Sequence and Variation
CCD	charge couple device
CD81	cluster of differentiation 81
CDC	cell division cycle
CDCA	chenodeoxycholic acid
CDH	cadherin
cDNA	complementary DNA
CDS	coding DNA sequence
CEACAM	carcinoembryonic antigen-related cell adhesion molecule
CF	cystic fibrosis
CFTR	cystic fibrosis transmembrane conductance regulator
CGN	cingulin
Chr	chromosome
Cl	chloride
CLDN	claudin
CNV	copy number variation

COPD	chronic obstructive pulmonary disease
CSL	CBF1/Suppressor of Hairless/Lag-1
Ct	cycle threshold
CTNL2	citrullinemia II
CTNNA1	α-Catenin
CTNNB	β-Catenin
CX43	connexin 43
CYP27A	sterol 27 hydroxylase
CYP7A1	cholesterol-7α-hydroxylase
db	database
DBD	DNA binding domain
dbSNP	single nucleotide polymorphism database
ddNTP	dideoxynucleotide triphosphates
DFNA51	autosomal dominant deafness-51
Dig	discs-large
DJS	Dubin-Johnson syndrome
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
dNTP	deoxynucleotide triphosphates
DSL	Delta/Serrate/Lag-2
EB	elution buffer
EBT	elution buffer with Tris
EDTA	ethylenediaminetetraacetic acid
EGF	epithelial grow factor
EJC	exon junction complex
EL	extracellular loop
ELM	extension-ligation mix

ER	endoplasmic reticulum
ESP	Exome Sequencing Project
FAM	fluorescein
FFPE	formalin fixed and paraffin-embedded
FHC	familial hypercholanemia
FIC1	familial intrahepatic cholestasis 1
FXR	farnesoid X receptor
GAPDH	glyceraldehyde 3-phosphate dehydrogenase
GATK	genome analysis toolkit
Gb	gigabase
GGT	gamma-glutamyl transferase
GJA	gap junction α-1
GJB2	gap junction β-2
GK	guanylate kinase
GMP	guanosine monophosphate
GRCh37	Genome Reference Consortium human genome build 37.
GSK-3β	glycogen synthase kinase-3β
GTP	guanosine triphosphate
HCC	hepatocellular carcinoma
HCO_3	bicarbonate
Het	heterozygous
hg19	human genome 19
Hom	homozygous
HPLC	high-performance liquid chromatography
HRP	horseradish peroxidase
HSD3B7	3-β-hydroxy-δ-5-c27-steroid dehydrogenases
HT	hybridisation buffer

HGNC	HUGO Gene Nomenclature Committee
ICP	intrahepatic cholestasis in pregnancy
IEF	isoelectric focusing electrophoresis
indel	insertions and deletions
JAG1	gene encoding Jagged-1
JAM	junctional adhesion molecules
kb	kilobases
kDa	kilodalton
L	ladder
LBD	ligand binding domain
LNA	locked nucleic acids
LNA1	library normalisation additives
LNB	library normalisation buffer
LNRs	lin-12 Notch repeats
LNS	library normalisation storage buffer
LNW	library normalisation wash buffer
LPAC	low phospholipid-associated cholelithiasis
MAF	minor allele frequency
MAGUK	membrane-associated guanylate kinase
Mb	megabases
MDCK	Madin-Derby canine kidney
MDR3	multidrug p-glycoprotein resistance 3
Mrd1	multidrug resistance1
mRNA	messanger RNA
MRP2	multidrug resistance-associated protein 2
NAD	nicotinamide adenine dinucleotide
NBD	nucleotide-binding domain

NCID	Notch intracellular domain
NES	nuclear export signal
NGS	next-generation sequencing
NICCD	neonatal-onset citrin deficiency
NISCH	neonatal sclerosing cholangitis associated with ichthyosis
NLS	nuclear localisation signal
NMD	nonsense mediated messenger ribonucleic acid decay
NR1H4	nuclear receptor subfamily 1 group H member 4
NSF	N-ethylmaleimide-sensitive factor
OCLN	occludin
OLT	orthotropic liver transplantation
OMIM	online Mendelian inheritance in man
PC	phosphatidylcholine
PCR	polymerase chain reaction
PDZ	Psd-95/Dig/Zo-1
PFIC	progressive familial intrahepatic cholestasis
PI	protease inhibitor
PMM	PCR master mix
PMSF	phe-nylmethylsulfonyl fluoride
PS	phosphatidylserine
Psd-95	post synaptic density protein 95
PTC	premature terminator codon
Q-score	Phred quality score
rcf	relative centrifugal force
RCL	reactive centre loop
RIN	RNA integrity number
RNA	ribonucleic acid

ROI	regions of interest
rpm	revolution per minute
RTA	real time analysis
RT-PCR	reverse transcription polymerase chain reaction
RXR	retinoid X receptor
SAF-B	scaffold attachment factor B
SAGE	serial analysis of gene expression
SAM	sequence alignment/map
SAV	sequencing analysis viewer
SDS	sodium dodecyl sulphate
SEM	standard error of the mean
SH_3	Src domain homology 3
SHP	small heterodimer partner
SLC25A13	solute carrier family 25, member 13
SM	Sec1/Munc18
SNAP	Soluble NSF Attachment Protein
SNARE	SNAP receptor
SNP	single nucleotide polymorphism
SNV	single nucleotide variation
SNX27	sortin nexin 27
SS	Sanger sequencing
SW	stringent wash
TBE	Tris/borate/EDTA
TDP1	TruSeq DNA polymerase
TEM	transmission electron microscopy
TGP	1000 Genome Project
TJP2	tight junction protein 2

TMD	transmembrane domain
TMEM	transmembrane
TRS	target resequencing
TSCA	TruSeq custom amplicon
t-SNARE	target-membrane SNARE
U	unique
UB	universal buffer
UCSC	University of California Santa Cruz
UniProt	Universal Protein Resource
UPL	Universal Probe Library
VIPAS39	VPS33B interacting protein
VPS33B	vacuolar protein sorting 33
v-SNARE	Vesicle SNARE
WES	whole-exome sequencing
WGS	whole-genome sequencing
ZO	zona occluden

Statement of works

The majority of the research described here was independently optimised and performed; however, the collaboration with experts was needed for the application of some techniques.

For the initial cohort of patients selected for target resequencing (TRS-21) and whole-exome sequencing (WES), the sequencing itself was performed by Dr Efterpi Papouli using an Illumina HiSeq 2000. This high throughput sequencing instrument belongs to the Genomics Facility at Guy's and St Thomas' NHS Foundation Trust and King's College London. Afterwards, the data analysis was conducted using different methods, including a Linux-based pipeline developed by Dr Michael Simpson from the division of Genetics and Molecular Medicine at King's College London. This specific method was used for the analysis of the TRS-21 and the WES data.

The biological consequences of genetic mutations were investigated using immunohistochemistry and transmission electron microscopy (TEM). The former was performed in the liver histopathology laboratory, at King's College Hospital, under the supervision of Dr Alex Knisely; whilst the ultrastructural investigations were carried out by Dr Bart E. Wagner in the histopathology Department at Royal Hallamshire Hospital in Sheffield, United Kingdom

1 General Background

Bile formation and secretion is a vital function of the liver. Since the first description of the physiological basis of bile formation (Sperber, 1959), the knowledge of biliary functions has progressed rapidly. Over the years, bile pathophysiology has been intensively investigated, demonstrating that a functional impairment is responsible for a severe spectrum of conditions known as cholestatic liver diseases (Trauner *et al.*, 1998).

1.1 Bile

Bile is a complex hepatic aqueous solution containing bile acids, phospholipids, proteins, cholesterol and bilirubin. The importance to the human body of bile formation lies in two main functions: i) bile regulates the homeostasis of cholesterol, elimination of bilirubin, and excretion of xenobiotics and endotoxins, such as bacterial lipid products; ii) bile acids, which represent approximately 65% of the total bile solutes, contribute to the digestion and absorption of fats and fat-soluble vitamins in the upper small intestine (Kurbegov & Karpen, 2008). Hepatocytes constitute the main cell type of the liver, responsible for most of the major hepatic functions including bile production and secretion. As epithelial cells, they are highly polarised and face two distinct environments: the blood sinusoids on the basolateral side, and the canalicular bile lumen on the apical side. Cell polarity is evidenced by different subsets of protein transporters and lipids distributed on the two plasma membranes, but it is also manifest with an asymmetric intracellular architecture. Acquisition and maintenance of these

subcellular zones require a complex intracellular mechanism of cell-cell interaction and protein trafficking and recycling (Treyer & Müsch, 2013).

The formation of bile is an active adenosine triphosphate (ATP)-dependent process; organic and inorganic solutes are secreted by hepatocytes into the canalicular lumen against steep concentration gradients followed by osmotic movement of water. Water represents more than 95% of bile solution. The passive movement of water can follow a paracellular pathway through tight junctions of two neighbouring cells or a transcellular pathway across hepatocytes mediated by water channels, known as aquaporins. Ten members, numbered from 0 to 9, are included in the aquaporin family. Aquaporin 8 has been demonstrated to be the main route of water entry into the biliary lumen. The channel is usually found inside cytoplasmic vesicles, until a cyclic adenosine monophosphate (cAMP) stimulus triggers the localisation of this water channel into the canalicular membrane, therefore increasing the water permeability (Huebert *et al.*, 2002).

Within the organic solutes, bile acids are the most abundant. In humans, the biosynthesis of bile acids is a multi-enzymatic process that takes place into the hepatocytes, from a molecule of cholesterol. This process can occur via two distinct pathways: the classical (neutral) pathway initiated by the enzyme cholesterol-7α-hydroxylase (CYP7A1), and alternative (acidic) pathway trigged by the mitochondrial sterol 27 hydroxylase (CYP27A1) (Setchell *et al.*, 2008). In addition to the liver, the alternative pathway takes place in the macrophages and in other tissues such as fibroblasts and vascular endothelium; although this mechanism contributes a minimal fraction of the bile acid pool (approximately 5%), it has been suggested as having an important role as an auxiliary pathway when the classical bile acid biosynthesis is compromised (Crosignani *et al.*, 2007). Both mechanisms drive the final production of cholic acid (CA) and chenodeoxycholic acid (CDCA), known as primary bile acids. Before their secretion into the canalicular lumen, the

carboxyl terminus of CA and CDCA are linked to the amino group of glycine or taurine. This modification, called amidation, generates highly stable molecules resistant to the cleavage of the pancreatic carboxypeptidase. Also, it increases their amphipathic nature becoming less toxic to the apical cell membrane of the hepatocytes and cholangiocytes, and more soluble at acidic pH environment. In fact, the double nature of bile acids consists of a hydrophilic α-face and a hydrophobic β-face, which allows the self-aggregation of bile acids in the canalicular lumen together with insoluble lipids such as phosphatidylcholine (PC), generating mixed micelles. Conjugated bile acids are exclusively excreted through ATP-dependent bile salt export pump (BSEP) (Gerloff *et al.*, 1998). This membrane transporter belongs to the ATP-binding cassette (ABC) superfamily, the most represented transporter proteins expressed on the apical surface of hepatocytes facing toward the lumen of canaliculi. They are structurally composed of two transmembrane domains (TMDs) with 6-11 membrane-spanning hydrophobic α-helices, and two cytoplasmic ATP-binding domains or nucleotide-binding domains (NBDs), which include the characteristic Walker A and Walker B motifs, and the ABC signature or C motif (Jones & George, 2004). After their canalicular secretion via BSEP and subsequently small intestine delivery, >95% of the bile acids are reabsorbed by the intestine enterocytes. A small portion of conjugated bile acids are subjected to a process of deconjugation and dehydroxilation by bacterial enzymes during the transit through the colon, forming deoxycholate and lithocholate, known as secondary bile acids. They represent the <5% of the total pool of bile acids.

In the canalicular lumen, free bile acids could severely damage the integrity of plasma membranes; the formation of mixed micelles has a critical role in bile formation and in the protection of the plasma membrane against the detergent property of bile acids. Phosphatidylcholines (PC) are the most abundant class of phospholipids in the biological membranes and they play a fundamental role in the mixed micelles formation. Hepatocytes move PC from the inner to the outer leaflet

of the canalicular membrane with an active ATP-dependent process. This transport is mediated by the multidrug p-glycoprotein resistance 3 (MDR3), another member of the ABC transport family. Along with PC, mixed micelles also contain cholesterol. Biliary cholesterol secretion is an ATP-dependent process predominantly dependent upon the heterodimerization of ABCG5 and ABCG8 proteins, further members of the ABC transporter family (Graf *et al.*, 2003).

An additional important function of bile is the excretion of bilirubin. Produced in the spleen from erythrocyte haemoglobin catabolism, bilirubin is a lipid soluble molecule. In the blood it is bound to serum albumin in order to prevent its passive diffusion across membranes causing cell toxicity. Bilirubin is, however, actively transported across the basolateral surface of hepatocytes, where it is converted into water-soluble compounds, bilirubin mono- and di-glucuronide. Conjugated bilirubin is then excreted into bile by an active transport, mediated by another member of the ABC transporter family identified as ATP-dependent multidrug resistance-associated protein 2 (MRP2). In addition, MRP2 is involved in the excretion of other multi organic anions, such as divalent bile salts, glutathione, glucuronate and sulphate conjugates (Erlinger *et al.*, 2014).

Along with the transport of the organic components of bile, into the lumen of the canaliculi there is also a canalicular bile salt independent flow. This highly regulated system actively excretes bicarbonate (HCO_3^-) from the hepatocytes in exchange for biliary chloride (Cl^-); in association with this, an osmotic movement of water occurs through the transcellular pathway. Anion exchanger 2 (AE2) is the main HCO_3^-/Cl^- exchanger expressed on the canalicular membrane as well as cholangiocyte membrane (Banales *et al.*, 2006).

1.2 Cholestatic liver disease

Cholestatic liver diseases are a group of heterogeneous hepatobiliary disorders where either the formation of the bile or its normal flow are compromised (Popper, 1981). Biliary flow goes from hepatocytes to the intestine through the network of the biliary tree, so a possible interference can occur at any level. Retention of biliary constituents inside the hepatocytes and bile ducts along with regurgitation into blood are responsible for the major manifestations of the syndrome of cholestasis (Hofmann, 2002). Pruritus is a common sign of clinical presentation of cholestasis probably due to accumulation of bile acids in body tissues (Bergasa, 2014). Jaundice and growth failure in children are additional features of biliary abnormality, caused by conjugated bilirubin circulating in the blood and absence of bile secretion in the intestinal duodenum resulting in fat and fat-soluble vitamin malabsorption. Biochemical diagnosis is based on the evaluation of the increased concentration of the serum bile acids, serum alkaline phosphatase (AP) and hyperbilirubinaemia. Gamma-glutamyl transferase (GGT) is a microsomal enzyme that catalyses the transfer of a gamma-glutamyl residue to an amino acid acceptor. The most significant glutamyl donor is glutathione. Expressed on the cell surface of bile canaliculi, it has been extensively adopted as a diagnostic marker of cell damage. In cholestasis, the serum concentration of GGT is dependent on the aetiology of the disease. High serum concentration of 50-100 times the upper limit of normal is detected, for example, in primary sclerosing cholangitis, a chronic inflammatory disorder affecting the intra and/or the extrahepatic bile ducts. On the other hand, normal or low serum GGT activity is found in a subtype of progressive intrahepatic cholestasis, in which the expression of the major transporter of bile acids is altered (Cabrera-Abreu & Green, 2002).

Cholestasis can be classified on the basis of the location of the impairment. Extrahepatic cholestasis occurs with physical obstruction of bile ducts outside the

liver; it is commonly caused by stones in the bile ducts, but also cholangiocarcinoma and more rarely primary sclerosis cholangitis and chronic inflammation. Intrahepatic cholestasis, however, arises inside the liver; it can be due to an obstruction of the smallest bile ducts caused by adenoma, liver fibrosis, or to chronic inflammation, both primary and secondary such as sclerosing cholangitis and chronic hepatitis, or to acquired or congenital determinants that can affect the metabolism of the bile acids and the transport of the bile constituents into the biliary lumen (Miethke & Balistreri, 2008). Congenital disorders may be caused by genetic abnormalities and occur before birth manifesting during the first months of life. A heterogeneous group of congenital disorders is represented by genetic intrahepatic cholestasis, in which genetic defects cause alteration in the normal biliary physiology. To date, numerous monogenic disorders in cholestasis have been identified; however, the understanding of the physiology of these diseases is still only partially known (Balistreri *et al.*, 2005). Chronic intrahepatic cholestasis has mainly been described associated with genetic alterations in ABC transporter family members, such as MDR3, BSEP and MRP2, which function as active canalicular transporters of phospholipids, bile acids and bilirubin. Given the complexity of bile formation, several different genomic regions could be mutated and could lead to cholestasis. In fact, inborn defects of bile formation, bile acid synthesis, embryonic development of biliary system and cell-cell organisation have been identified both in inherited cholestatic liver diseases and in other complex inherited disorders, such as Alagille syndrome, where cholestasis is one of the several features (van Mil *et al.*, 2005). However, a significant percentage of familial intrahepatic cholestatic liver disease still has an unclear aetiology, remaining the focus of the further investigation.

1.3 Progressive familial intrahepatic cholestasis

Progressive familial intrahepatic cholestasis (PFIC) represents a rare genetic condition with an estimated incidence of approximately 1 in 100,000 births (Davit-Spraul *et al.*, 2009). It represents a heterogeneous group of inherited disorders caused by genetic defects in *ATP8B1* (PFIC1 also known FIC1 deficiency), *ABCB11* (PFIC2 also known BSEP deficiency) and *ABCB4* (PFIC3 also known MDR3 deficiency). The three genes encode different ATP-dependent canalicular transporter proteins involved, respectively, in maintaining the asymmetric lipid composition of canalicular membrane (FIC1), in the transport of conjugated bile acids into the canalicular lumen (BSEP) and in the translocation of molecules of phosphatidylcholines into the canalicular lumen for the formation of mixed micelles (MDR3) (Figure 2.1.1). The biological functions of these three hepatocellular transporters are described in more details in the following session 2.1.1.1.

PFIC have been historically subdivided into two subgroups based on the concentration of serum GGT, a biochemical marker of canalicular membrane damage. Normal and/or low concentration of serum GGT has been identified in patients with BSEP and FIC1 deficiencies, while high concentration in patients with MDR3 deficiency (van Mil *et al.*, 2005). Clinically, the three types of PFIC show some differences. In FIC1 deficiency, cholestatic features such as jaundice, severe pruritus and failure to thrive have been reported with high degree of severity during the first year of life; however, intermittent cholestatic episodes have been identified at any age (van Mil *et al.*, 2001). The latter represents a less severe manifestation of PFIC, therefore it is known as benign recurrent intrahepatic cholestasis (BRIC). Extrahepatic manifestations are prominent characteristics of FIC1 deficiency including deafness, diarrhoea and pancreatitis, suggesting that the expression of *ATP8B1* is not restricted to liver tissues (Lykavieris *et al.*, 2003). For example, in the human intestinal epithelial Caco-2 cells, loss of *ATP8B1* expression

was shown to cause a perturbation of apical membrane composition and a defect in the brush border formation (Verhulst *et al.*, 2010). Histologically, canalicular cholestasis is usually observed, with retention of biliary constituents in the canalicular lumen, portal area infiltration and periportal fibrosis.

Similar to FIC1 deficiency, BSEP deficiency occurs, in severe cases, with early-onset including persistent jaundice, pruritus and malnutrition, and sometimes with hepatocellular carcinoma (Strautnieks *et al.*, 2008). Liver histology is characterised by severe alteration of the liver architecture, hepatocellular necrosis and giant cell transformation. Milder manifestations have also been identified after the first decade with recurrent episodes of cholestasis (BRIC-2) or in adults, possible trigged by drugs or pregnancy, respectively named as drug-induced cholestasis (DIC) and intrahepatic cholestasis in pregnancy (ICP) (Pauli-Magnus *et al.*, 2004; Pauli-Magnus & Meier, 2006). The third PFIC condition can also manifest at every age, from infancy to adulthood. Patients with no MDR3 function from either allele present with early onset progressive liver disease, whilst those with a single missense mutation can have cholestasis isolated to pregnancy, with no progression. Patients with every degree of severity in between have been described, and late presentation does not necessarily confer a good prognosis. Histological characteristics include portal fibrosis and bile ductular proliferation.

Nevertheless, there is a proportion of patients that does not harbour any pathological mutations in those three genes and remain with an unidentified genetic diagnosis; this suggests that other genes might been involved in the aetiology of familial cholestatic liver diseases.

Genetic disorders rarely have a “cure” that can restore the normal physiology; however, symptoms are usually treated or managed. The most common medical management of all type of chronic cholestasis, including PFIC, consists in the administration of ursodeoxycholic acid (UDCA), a non-toxic bear bile acid (Jacquemin *et al.*, 1997). The therapy is intended to displace some of the endogenous bile acid pool with a less detergent compound, which will therefore reduce the liver injury during cholestasis. Beneficial effects has been demonstrated

in MDR deficiency and in particular, children having missense mutation showed a better response compared to those carrying protein-truncating mutations (Jacquemin *et al.*, 2001). It has been speculated that those patients having a missense mutations have a possible residual transport activity that in combination with UDCA therapy could create a less detergent environment. On the other hand, most of the patients with BSEP deficiency and FIC deficiency seem not to response to UDCA therapy and therefore, surgical procedures are necessary, especially in patients with severe cholestatic manifestation (Ismail *et al.*, 1999). Two surgical interventions are commonly used, partial external biliary diversion (PEBD) and liver transplantation. In PEBD, bile is externally excreted by a conduit between the gallbladder and the abdominal wall, partially reducing the percentage of bile flow entering the enterohepatic circulation. This procedure has demonstrated clinical and biochemical improvement in majority of patients with chronic cholestasis (Kurbegov *et al.*, 2003). PBED is recommended as an early surgical intervention or as a bridge before liver transplantation. In cases with end-stage PFIC, with development of biliary cirrhosis, liver transplantation is essential. The treatment has been associated with good outcome and survival rate after 1 year of age (Aydogdu *et al.*, 2007). Unfortunately, neither form of surgery resolves the extrahepatic manifestations characteristics of FIC1 deficiency, such as the exacerbation of diarrhoea. In addition other treatment options have emerged. One of the potential targets for future treatment of cholestasis is represented by the farnesoid X nuclear receptor (FXR). Bile acids synthesis is a finely regulated mechanism, mediated by bile acids cholic acid and chenodeoxycholic acid. Bile acids bind the nuclear receptor FXR that indirectly repress the transcription of the *CYP7A1*, which is the enzymatic initiator of bile acid synthesis. The mechanism of action is described in more details in section 2.1.2.1. Understanding the powerful action of this bile acids/FXR affinity has led researchers to develop bile acid analogues, such as obeticholic acid (OCA), as potential FXR agonist and investigate their function in the treatment of cholestatic conditions (Pellicciari *et al.*, 2002). OCA is currently tested in phase II trial for treatment of primary

sclerosing cholangitis (PSC) and in phase III trial for the treatment of primary biliary cirrhosis (PBC). The beneficial effects of this drug could also improve the cholestatic condition of patients affected by PFIC, but unfortunately no clinical data are available yet. An additional target is the apical sodium bile acid transporter (ASBT), expressed on the luminal surface of the ileum enterocytes (Balakrishnan & Polli, 2006). ASBT plays a critical role in the enterohepatic circulation of bile acids, and in particular in the bile acid re-absorption in the small intestine. Therefore, chemical compounds that inhibit ASBT could reduce the bile acid pool and give a beneficial effect in cholestatic patients. However, as yet there are no clinical data in support of this theory.

Several biological aspects of progressive familial intrahepatic cholestasis are now understood, allowing researchers also to investigate new potential targets to improve the clinical management. Nevertheless, a proportion of patients still remain with idiopathic causes, suggesting that other possible biological pathways could have been altered and be the origin of cholestasis. In this research project, investigation is focused on disclosing these idiopathic cases.

2 Genetics of cholestatic liver disease

2.1 Introduction to genetics of cholestasis

Several cellular mechanisms have been described associated with inherited forms of cholestasis. The majority of them are due to genetic defects in bile acid synthesis and bile transport and formation, leading to liver damage. Progressive familial intrahepatic cholestasis (PFIC) represents a heterogeneous group of autosomal recessive conditions due to alterations in three genes encoding three canalicular transporter proteins (section 1.3). In addition other syndromes, such as Alagille syndrome and arthrogryposis renal dysfunction cholestasis (ARC) syndrome, are characterised by multi-systemic features, including cholestasis. A panel of genes associated with different cholestatic conditions was selected for the initial part of this research project (Table 2.1.1). The molecular functions of every selected gene and their roles in the manifestation of the cholestatic disease are described in this chapter, categorised by their respective physiological functions.

*OMIM	Gene name	Protein name	Disease/Function
603201	*ABCB11*	BSEP	PFIC
602397	*ATP8B1*	FIC1	PFIC
171060	*ABCB4*	MDR3	PFIC
601920	*JAG1*	Jagged 1	Alagille syndrome
600275	*NOTCH2*	Notch2	Alagille syndrome
607709	*TJP2*	ZO-2	FHC
602938	*BAAT*	BAT	FHC
608552	*VPS33B*	VPS33B	ARC
613401	*VIPAS39*	SPE-39	ARC
601107	*ABCC2*	MRP2	Dubin-Johnson syndrome
603718	*CLDN1*	Claudin-1	NISCH
107400	*SERPINA1*	α-1-antytripsin	A1AT deficiency
602421	*ABCC7*	CFTR	Cystic fibrosis
603859	*SLC25A13*	Citrin	Citrin deficiency
607764	*HSD3B7*	3-β-HSD VII	HSD3B7 deficiency
604741	*AKR1D1*	3-oxo-5-β-steroid 4-dehydrogenase	AKR1D1 deficiency
611028	*TMEM30A*	CDC50A	FIC1 chaperone
611029	*TMEM30B*	CDC50B	FIC1 chaperone
611030	*TMEM30C*	CDC50C	CDC50 family member
603826	*NR1H4*	FXR	Nuclear receptor
300516	*ATP11C*	ATP11C	Cholestasis (mouse model)

Table 2.1.1 Summary of selected genes for target resequencing (TRS) analysis.

The 21 cholestatic genes were sequenced in the initial cohort of 18 patients (TRS-21); the first seven genes on the list were selected for the sequencing of the subsequent expanded cohort (TRS-7). The gene names were defined in accordance to the HUGO Gene Nomenclature Committee (HGNC). The list was also enhanced with information regarding the entry number in the online Mendelian inheritance in man (OMIM) database, the protein name in accordance to universal protein resource (UniProt) and the known diseases associated (section 2.1). The full gene names are listed in **List of Abbreviations**

2.1.1 **Membrane transporters**

Membrane transporters are a class of integral proteins that mediate the passage of ions, metals, macromolecules and drugs across the phospholipid bilayer. They play a crucial role as gatekeepers, controlling the efflux and influx from one compartment to another of cells and organelles (Hediger *et al.*, 2004). They can be classified into passive and active transporters. During passive transport, these anchored proteins aid the spontaneous diffusion of polar molecules or charged ions through the membrane without energy consumption; in contrast active transport requires chemical energy, such as adenosine triphosphate (ATP) or an electrochemical gradient, to allow and control the translocation of a variety of substances across the phospholipid bilayer from a region with low concentration to a region with high concentration (against concentration gradient). In normal bile secretion, these mechanisms are essential and their alteration plays a crucial role in the molecular pathogenesis of cholestasis. In Figure 2.1.1 the membrane transporters selected for this study are represented in their specific cellular localisation. Bile acids and other biliary constituents are actively transported from the hepatocytes into canaliculi by different classes of membrane transporters; subsequently, they transit into the bile ducts where they undergo electrolyte modifications mediated by anion channels located in the cholangiocyte membrane. The final product then is excreted into the upper part of the small intestine (Esteller, 2008).

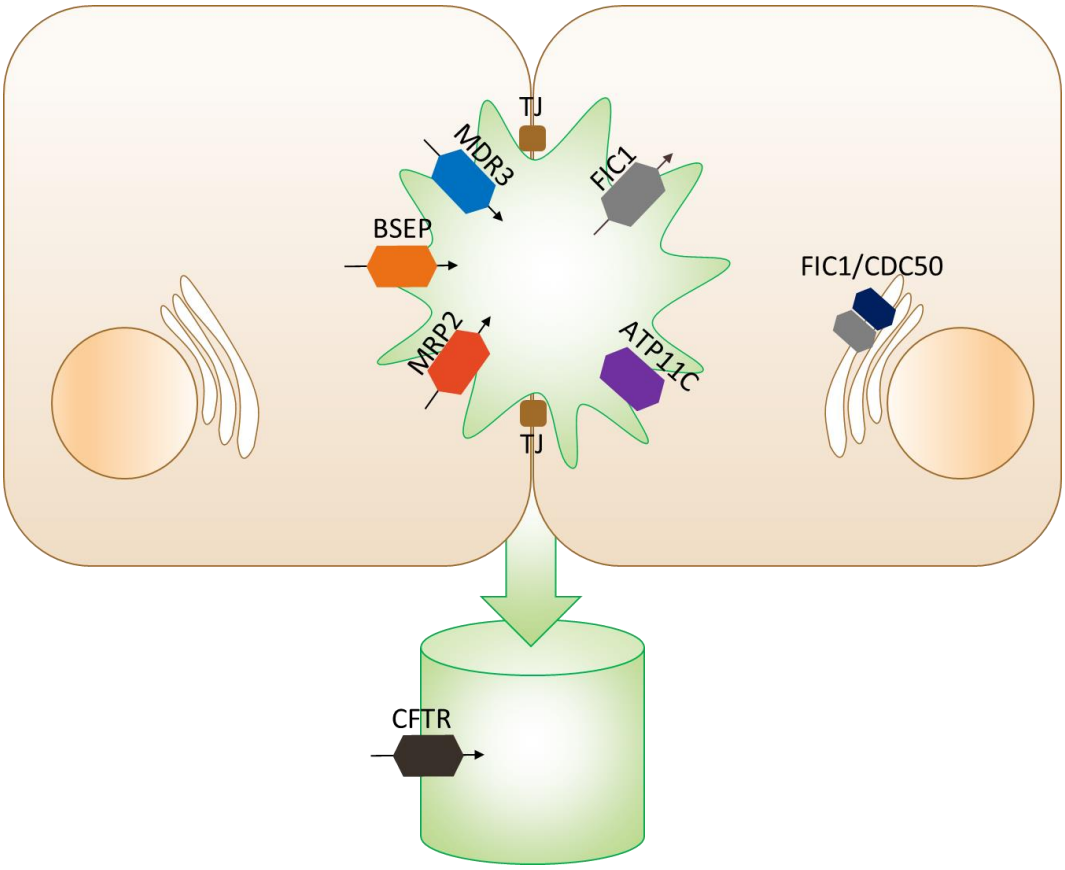

Figure 2.1.1 Schematic representation of selected membrane proteins involved in bile formation.

The image shows two adjacent hepatocytes sealed by tight junction structures (TJ) creating the canalicular lumen and a portion of bile duct. Bile formation is a complex mechanism that involves several membrane transporters located on the apical surface of hepatocytes facing toward the bile canaliculi, such as MDR3, BSEP, MRP2, FIC1 and ATP11C, and on the apical membrane of cholangiocytes forming bile ducts, such as CFTR. The transmembrane protein CDC50 is instead an integral protein located on the endoplasmic reticulum co-expressing with FIC1; it is involved in translocation of FIC1 on the canalicular membrane. Mutations in the respective genes have been associated to the cholestatic liver diseases. Abbreviations are listed in **List of Abbreviations**.

2.1.1.1 Hepatocellular transporters

The apical surface of the hepatocytes facing toward the lumen of canaliculi is enriched in different classes of active transporter proteins. One of the most represented groups are the ATP-binding cassette (ABC) transporters (Figure 2.1.2). These proteins are expressed ubiquitously in human tissues, highly in gut, liver and kidney, and are involved in many cellular pathways, like nutrient uptake, stem cell differentiation, lipid trafficking and multidrug resistance; therefore their dysfunction causes a variety of genetic diseases, such as cystic fibrosis, Dubin-Johnson syndrome and progressive familial intrahepatic cholestasis (Gottesman & Ambudkar, 2001).

Bile acids represent the major solute of bile. The active secretion of conjugated bile acids into the biliary canaliculi occurs through a bile acid-specific ABC transporter, ATP-Binding cassette subfamily B member 11 (*ABCB11*), which is known as bile salt export pump (BSEP) (Figure 2.1.2). Exclusively expressed in the liver, the *ABCB11* gene was discovered in 1998 by homozygosity mapping of a potential disease-causing locus of PFIC (Strautnieks *et al.*, 1998). PFIC caused by genetic variations of *ABCB11* is commonly known as PFIC type 2. Homozygous mutations and compound heterozygous mutations in *ABCB11* were identified as causing reduced or complete loss of BSEP expression on the canalicular membrane, leading to impaired bile acid flow from the inner to the outer side of the hepatocytes (Strautnieks *et al.*, 2008). The resulting increase in concentration of bile acids inside the cytoplasm might also promote the malignant transformation of hepatocytes. A high risk of hepatocellular carcinoma (HCC) in early childhood was demonstrated as consequence of BSEP deficiency (Knisely *et al.*, 2006). In addition, *ABCB11* mutations have been described in a milder form of intrahepatic cholestasis, manifesting with remittent episodes of cholestasis at highly variable age of onset (van Mil *et al.*, 2004) and in the increase in susceptibility to intrahepatic cholestasis in pregnancy (ICP) (Dixon *et al.*, 2009), highlighting BSEP

deficiency as a spectrum of disease. High predisposition to drug-induced cholestasis (Lang *et al.*, 2007) and ICP (Dixon *et al.*, 2009) was identified in association with V444A polymorphism. Although, this amino acid substitution is present with an allele frequency of 50% in the general population, low protein expression was exhibited with alanine in position 444 (Meier *et al.*, 2006). Biochemical markers typically show normal or low serum GGT concentration, suggesting no membrane damage.

In bile formation, phosphatidylcholine (PC) is important in creating mixed micelles of bile acids together with molecules of cholesterol. This compact structure is important for the protection of biliary epithelium from the detergent property of bile acids. PC translocates from the inner to the outer leaflet of the canalicular membrane by the action of ATP-binding cassette subfamily B member 4 (*ABCB4*), an active transporter protein (Oude Elferink & Paulusma, 2007) (Figure 2.1.2). *ABCB4* was cloned for the first time in 1987 and revealed high homology with the hamster multidrug resistance1 (*Mrd1*) gene. The human homologous was named *MDR3 (*Van der Bliek et al., 1987). Homozygous mutations, that cause complete loss of function, are associated with PFIC type 3. The absence of MDR3 transporter produces deleterious consequences: the failed secretion of PC inside the canaliculi does not allow the formation of mixed micelles, leading to high concentrations of free bile acids in the biliary lumen and exposure of canalicular and cholangiocyte membrane to their toxic properties. Portal inflammation, portal fibrosis and bile duct proliferation are the consequences, in association with rise of serum level of GGT. However, a few patients with MDR3 deficiency and low GGT have been identified (unpublished data). MDR3 deficiency is characterised by a wide phenotypic spectrum; cholestasis caused by alteration in *MDR3* is also associated with benign recurrent intrahepatic cholestasis (BRIC) type 3, low phospholipid-associated cholelithiasis (LPAC) (Rosmorduc *et al.*, 2001) and ICP (Jacquemin *et al.*, 1999).

Bilirubin is a major constituent of bile. After intracellular conjugation, water-soluble bilirubin is actively excreted by a canalicular membrane transporter. Multidrug resistant protein 2 (MRP2), encoded by the gene ATP-binding cassette subfamily C member 2 (*ABCC2*), belongs to a subgroup of the ABC transporter family, involved in ion and drug secretion activity (Paulusma *et al.*, 1996) (Figure 2.1.2). Homozygous or compound heterozygous mutations of *ABCC2* cause the abnormal function of MRP2 protein identified in associated with Dubin-Johnson syndrome (DJS). DJS is a rare autosomal recessive disorder, characterised by persistent or intermittent increase of serum conjugated bilirubin leading to chronic jaundice, abdominal pain and fatigue. Adolescents and young adults are mainly affected, however neonatal-onset cases presenting with severe cholestasis have also been reported (Lee *et al.*, 2006). Histologically, sedimentation of melanin-like pigments within hepatocytes is a distinctive sign; it is due to a retention and polymerisation of anion metabolites followed by tissue deposition (Paulusma *et al.*, 1997). Bilirubin is not the unique substrate of MRP2; in fact, it is involved in the clearance of several different drugs, such antibiotics, leukotrienes, toxins and heavy metals. Being expressed as it is on the apical surface of hepatocytes and on the renal proximal tubular cells, abnormal MRP2 function can also impair renal drug excretion and subsequently cause renal toxicity (Hulot *et al.*, 2005).

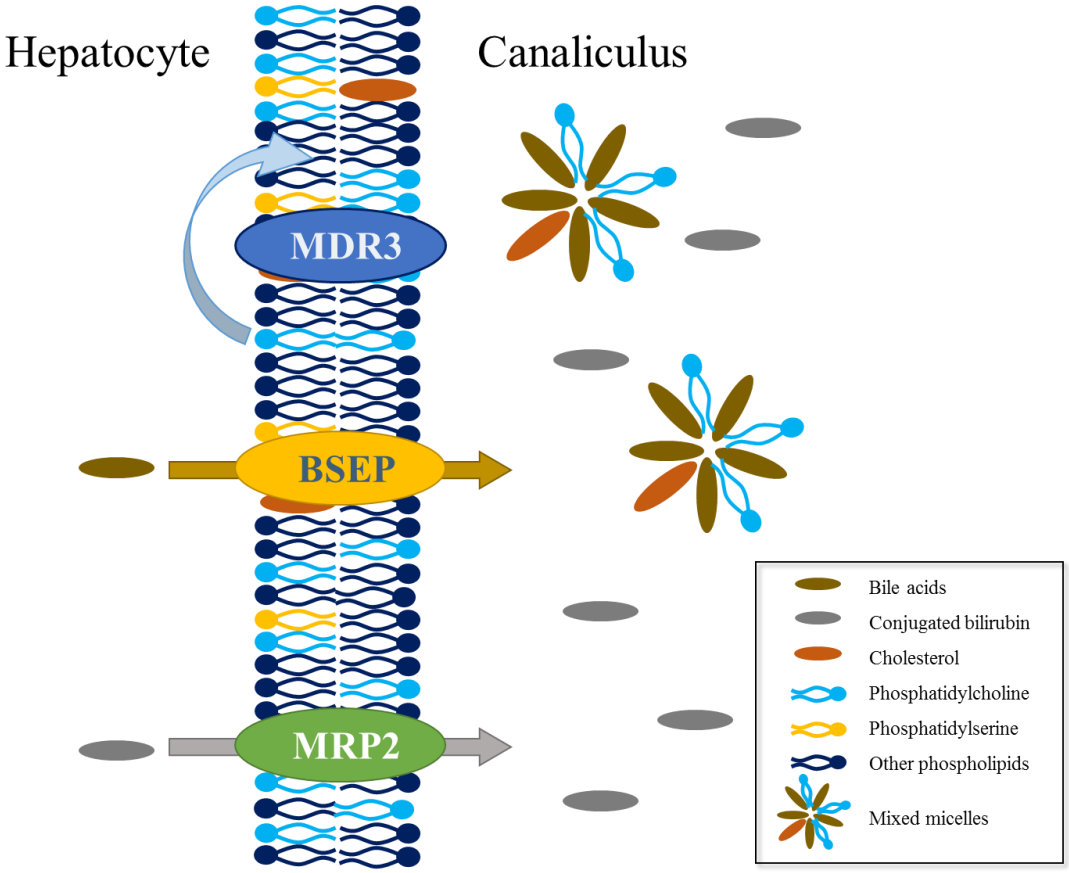

Figure 2.1.2 Hepatocellular transporters involved in bile formation

On the apical surface of the hepatocytes, facing the canalicular lumen, several ATP-dependent transporter proteins are involved in bile formation. BSEP is responsible for the transport of bile acids, while MRP2 is responsible for the transport of bilirubin. They represent the major organic solutes of bile. As bile acids are characterised by detergent properties able to seriously damage the plasma membrane, to prevent injury, molecules of phosphatidylcholine translocate from the inner to the outer leaflet of the bilayer of the membrane where they create harmless mixed micelles with bile acids and cholesterol. This active transport is mediated by the MDR3 protein

The phospholipid bilayer of the canalicular membrane has an asymmetric distribution: PC and sphingolipids are enriched in the outer side, while phosphatidylethanolamine, phosphatidylinositol and phosphatidylserine (PS) are enriched in the inner side toward the cytoplasm. During bile acid and phospholipid transport, PS can flip spontaneously into the biliary lumen surface. These aminophospholipids on the outer side of the canalicular membrane change the

viscosity of the plasma membrane leading to higher susceptibility to the bile toxicity. In 1998, a member of the p-type ATPase transporter family was identified to be associated with PFIC type 1 (Bull *et al.*, 1998) (Figure 2.1.3). The P-type ATPase is a large family of ATP-dependent integral proteins driving ions and lipids across the cellular membrane. This ATP-dependent phospholipid transporter gene (*ATP8B1*) encodes the familial intrahepatic cholestasis 1 protein (FIC1), whose function is to maintain a non-random distribution of phospholipids across the membrane bilayer, flipping PS from the outer to the inner leaflets (Paulusma & Elferink, 2010). Depending on the severity of the mutations and the likelihood to alter protein function, *ATP8B1* mutations have been identified associated with the severe manifestation of familial intrahepatic cholestasis PFIC type 1, but also to the milder phenotype of BRIC-1 (Bull *et al.*, 1998). In addition, extrahepatic features have been described, such as diarrhoea, liver steatosis, and rarely hearing loss and pancreatitis (Lykavieris *et al.*, 2003). Biochemical analysis shows normal serum concentration of GGT.

In vivo studies have identified that FIC1 is co-expressed with CDC50A and CDC50B proteins in the endoplasmic reticulum (ER) (Figure 2.1.3). The heteroduplex complex formation is a crucial mechanism within hepatocytes that allows the translocation of FIC1 from ER to its localisation on the plasma membrane. Genetic defects that cause an alteration in the FIC1-CDC50 complex assembly may affect the FIC1 stability and lead to ER retention followed by degradation, phenotypically manifesting with PFIC1 (Paulusma *et al.*, 2008). The CDC50 family was initially identified in mammalian homologs of *Saccharomyces Cerevisiae* by *in silico* gene expression analysis; it includes three members CDC50A (*TMEM30A*), CDC50B (*TMEM30B*) and CDC50C (*TMEM30C*). CDC50C is the less well characterised integral membrane protein compared with the other paralogue members. Structurally, they are all composed by two transmembrane segments and an exoplasmic domain stabilised by one or more disulphide bonds (Katoh & Katoh, 2004).

In 2004 *ATP11C* gene, which belongs to the same p-type ATPases and encodes ATP11C protein, was cloned on the X chromosome (Andrew Nesbit *et al.*, 2004). It was identified as being associated with a mild form of cholestasis and a B-cell lymphopenia syndrome; however, its role has not been elucidated. ATPase IQ might be involved in the symmetry of the canalicular membrane, a critical point for bile formation and bile flow, but how it can affect B-cell development remains to be determined (Siggs *et al.*, 2011) (Figure 2.1.3).

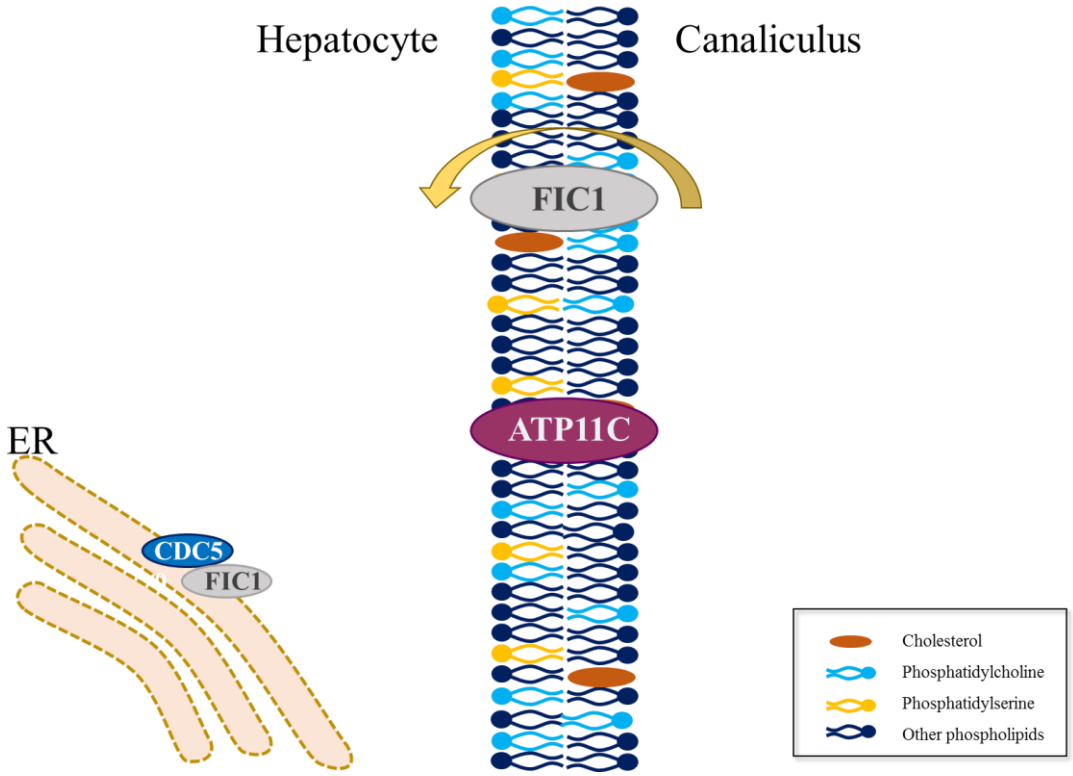

Figure 2.1.3 Hepatocellular proteins involved in the symmetry of the canalicular membrane

The lipid composition of the canalicular membrane is asymmetric. The hepatocellular transporter FIC1 is actively involved in maintaining phosphatidylserine in the inner side of the membrane facing the cytoplasm. The localisation of FIC1 on the membrane is mediated by the formation of the heteroduplex complex with CDC50 in the endoplasmic reticulum (ER). A role in the membrane symmetry has also been suggested for ATP11C.

2.1.1.2 Cholangiocyte transporters

After excretion into canaliculi, bile proceeds through the wide network of the biliary tree. There, a series of processes of reabsorption and secretion contribute to the electrolyte composition of bile. Within these processes an important role is played by the cystic fibrosis transmembrane conductance regulator (CFTR) protein. In the liver, this ATP-dependent chloride channel is expressed on the apical membrane of polarised intra and extrahepatic bile duct epithelial cells or cholangiocytes. It is also expressed on lung and pancreatic epithelial cells, and on non-epithelial cells, such as erythrocytes, myocardiocytes, and immune cells. In the bile formation, the chloride excretory function is mediated by an increase intracellular concentration of cAMP stimulated by gastrointestinal hormones and in particular by secretin. The high concentration of chloride in the ductal lumen stimulates the apical Cl^-/HCO_3^- exchanger to secrete HCO_3^- and reabsorb Cl^- (Lenzen *et al.*, 1992). Furthermore, CFTR has a regulatory role in maintaining the chloride balance through the interaction with other channels, such as epithelial Na^+ channels, and through controlling the movement of water across membranes (Lubamba *et al.*, 2012). *In vitro* studies have shown that CFTR deficient cholangiocytes have impaired chloride transport that causes a less alkaline and hydrated bile with higher toxicity. Consequently biliary epithelium can be damaged, which leads to focal biliary cirrhosis. However, an improvement in the chloride excretory activity and consequently HCO_3^- secretion has been reported after Ca^{2+} stimulation (Zsembery *et al.*, 2002). CFTR belongs to the ABC transporter family and it is also named ATP-binding cassette subfamily C member 7 (*ABCC7*). According to the CFTR mutation database (http://www.genet.sickkids.on.ca/cftr/app) to date 1,950 different variants have been discovered. On the basis of the biological consequence of the mutation on the protein function, six different classes of CFTR mutations have been created. The first group compromises all mutations that lead to a total or a partial absence of CFTR protein, such as frameshift insertions or deletions and nonsense. Mutations

that alter protein folding in the ER have been grouped in the second class, including p.Phe508del, the most frequent and most studied cystic fibrosis (CF) mutation. The third class of CFTR mutations usually affect the ATP-binding domain, coding for a protein that is structurally stable and able to locate on the plasma membrane, but has no functional activity. Impairment in the channel conductance has been shown for the proteins of the fourth class, which have mutations in the membrane spanning domain, an important site for the structure of the channel. The fifth class includes all mutations that cause reduction of the total amount of CFTR protein, while the last class of mutations are encoding for CFTR protein highly unstable on the plasma membrane and therefore highly subject to degradation (Fanen *et al.*, 2014). CF is an autosomal recessive disorder manifest with a variable degree of phenotype, depending of the nature of the genotype. The expression of CFTR is wide in the human body, so numerous organs are potentially involved in the CF pathology. Lung and pancreas are affected in 90% of the cases; in addition, manifestation in sweat gland, small and large bowel, liver and biliary tree can also be identified. Clinically, liver disease is present in only one third of CF patients. Focal biliary cirrhosis is the most typical liver pathology associated with CF and is probably a consequence of a deficient Cl^-/HCO_3^- exchanging mechanism, that leads to biliary damage and progressive fibrosis (Colombo, 2007). Neonatal cholestasis is a rare manifestation of CF, having an incidence lower that 1%. A good prognosis has been described in a series of CF paediatric patients with early-onset of cholestasis (Shapira *et al.*, 1999).

2.1.2 **Metabolism**

Numerous metabolic processes occur in the liver; each of them requires the functional activity of highly specific enzymes. Genetic defects causing an alteration of these enzymes are responsible for a group of genetic disorders known as inborn errors of metabolism. Enzymatic deficiency decreases or abolishes the catalysis of specific substrates, which then causes disease through several different mechanisms. Liver, and other organs, dysfunction can occur through accumulation of precursors or intermediate metabolites, or a lack of product.

2.1.2.1 Bile acid synthesis and regulation

Bile acids are a group of steroid acids biosynthesised in the liver by the catabolism of cholesterol. This process involves a cascade of enzymatic reactions and it occurs in the cytosol, mitochondria, ER, and peroxisomes of hepatocytes (Figure 2.1.4). In humans, the major bile acids are cholic acid (CA) and chenodeoxycholic acid (CDCA); in the final step of bile acid synthesis, they undergo a glycine or taurine conjugation that confers a less detergent property (section 1.1). It is estimated that defects in the enzymes involved in the bile acid synthesis are present in 1-2% of cholestatic disorders in children. Because of accumulation of toxic bile acid intermediates in the hepatocytes, these inherited autosomal recessive disorders can cause a rapid development of progressive cholestasis in children, but also congenital liver disease at birth or neonatal hepatitis (Monte *et al.*, 2009).

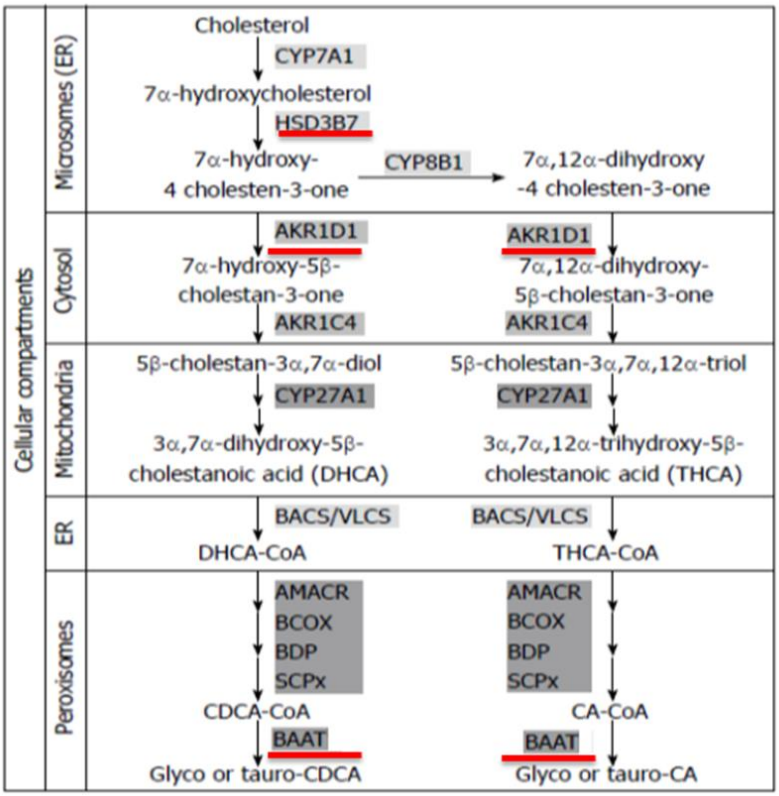

Figure 2.1.4 Representation of the bile acid synthesis

The biosynthesis of bile acids is derived from the catabolism of cholesterol by a cascade of enzymatic reactions that takes place in different cellular compartments of hepatocytes. In red are highlighted three particular enzymes, HSD3B7, AKR1D1 and BAAT, which are associated with the aetiology of cholestatic liver diseases. This figure was edited from (Monte *et al.*, 2009).

In the initial phase of the classical bile acid biosynthesis, 3-β-hydroxy-δ-5-c27-steroid dehydrogenases, also called 3-β-HSD VII, plays a catalytic role against the 7-α hydroxylated sterol substrates (Figure 2.1.4). In this enzymatic activity, nicotinamide adenine dinucleotide (NAD^+) is used as cofactor. 3-β-HSD VII is predominantly expressed in the liver and is anchored to the ER membrane. Different homozygous and compound heterozygous mutations in the 3-β-hydroxy-δ-5-c27-steroid dehydrogenase gene (*HSD3B7*) code for inactive enzyme resulting in an impairment in bile acid production and accumulation of intermediate sterols that lead to liver dysfunction. HSD3B7 deficiency is associated with an autosomal

recessive disorder leading to neonatal progressive cholestasis. Clinically, it manifests with features similar to those of other inherited disorders due, for example, to loss of function of canalicular transporter proteins or other bile acid enzymes; namely jaundice, hyperbilirubinaemia and malabsorption of fat and lipid-soluble vitamins. Histopathologically, liver biopsies reveal hepatocyte inflammation with giant cell hepatitis, fibrosis and canalicular and hepatocellular cholestasis (Cheng *et al.*, 2003).

An additional bile acid biosynthesis enzyme identified to be associated with disease is δ (4)-3-oxosteroid 5-β-reductase (AKR1D1). This cytosolic enzyme catalyses a crucial intermediate step in the cholesterol catabolism during the formation of bile acids (Figure 2.1.4). Homozygous point mutations have been identified that cause instability and abnormal kinetic parameters, resulting in AKR1D1 deficiency. The lack of this enzyme is associated with reduction of bile acid synthesis and accumulation of hepatotoxic δ.3.oxo and 5α-reduced bile acids (AKR1D1 substrate) that phenotypically manifest with neonatal hepatitis and cholestasis (Drury *et al.*, 2010).

The last step of bile acid synthesis is carried out in hepatocyte peroxisomes by bile acid CoA: amino acid N-acyltransferase (BAAT) (Figure 2.1.4). The bile acid moiety, derived from the cholesterol catabolism and first step of conjugation, is ready to be transferred from the acyl-CoA thioester to the amino acid glycine or taurine to complete the biosynthesis. Conjugation of bile acids represents a crucial mechanism in this process, important to obtain the detergent property required for the absorption of lipids and lipid-soluble vitamins. Familial hypercholanemia (FHC) is a rare genetic disorder identified in the Amish population. It manifests with hyperbilirubinaemia, itching, fat malabsorption, and increased serum alanine aminotransferase activity. Serum GGT shows normal activity. Genetic analysis identified a single missense mutation c.226A>G in *BAAT*, sometimes in

combination with the missense mutation c.143T>C in the tight junction protein 2 (*TJP2)*. The authors suggested that there was evidence of reduced penetrance and most probably oligogenic inheritance. A possible third locus has been proposed for those individuals where no mutations in *BAAT* and *TJP2* were discovered (Carlton *et al.*, 2003). Fat-soluble vitamin deficiency was also described as a common feature in other cases of BAAT deficiency, caused by different pathogenic homozygous mutations. The degree of cholestasis is although variable (Setchell *et al.*, 2013).

Accumulation of bile acids can be cytotoxic in the liver and in the intestine, so the human body has evolved a finely regulated mechanism to control the size and distribution of the bile acid pool. Bile acids themselves are involved in the control of this regulation acting as natural ligands for the farnesoid X receptor (FXR). FXR is also called nuclear receptor subfamily 1 group H member 4 (NR1H4). Nuclear receptors represent a superfamily of intracellular ligand-activated proteins that regulate the transcription of genes involved in development, proliferation, reproduction and metabolism, through the interaction with specific DNA transcriptional binding sites. Structurally all nuclear receptors contain six highly conserved domains (A-F). Of these, the DNA binding domain (DBD) or C domain is involved in the recognition of specific DNA sequences known as hormone response elements, while the ligand binding domain (LBD) or E domain is required for dimerization and transactivation of the receptor after interaction with hormonal or non-hormonal ligands (Olefsky, 2001). On the basis of the ligand specificity and the intracellular mechanism of action, which includes homo or heterodimerization followed by DNA binding, nuclear receptors can be classified into four different groups: homodimeric steroid receptors, retinoid X receptor (RXR) heterodimers, homodimeric orphan receptors and monomeric orphan receptors (Mangelsdorf *et al.*, 1995). FXR is a member of the second group of nuclear receptor, so its functional activity is trigged by the interaction with RXR. FXR is highly expressed in liver and intestine, but moderate expression can be also identified in adrenal

gland and kidney (Forman *et al.*, 1995). After forming the heterodimer FXR/RXR, it migrates into the nucleus where it binds with high specificity to a consensus IR-1 sequence proximal to gene promoters (Laffitte *et al.*, 2000). In the bile acid synthesis regulation, CDCA, the major component of the bile acids, shows strong affinity for FXR; the activation of FXR causes a transcriptional repression of *CYP7A1*, which encodes the crucial enzyme involved in conversion of cholesterol in the first step of the classical bile acid pathway. However, specific binding sites for FXR were not initially identified in that gene promoter (Chiang *et al.*, 2000). In fact, studies have demonstrated that the repression of *CYP7A1* transcription is not directly mediated by FXR, but it is induced by a small orphan nuclear receptor SHP-1, whose expression in turn depends on FXR (Goodwin *et al.*, 2000). Nevertheless, the transcription of several different genes has been identified to be under the action of this nuclear receptor, including genes encoding for the main canalicular transporter proteins (*ABCB11*, *ABCB4*, *ATP8B1* and *ABCC2)* (Eloranta & Kullak-Ublick, 2008). Phenotypic association with cholestatic liver disease characterised by elevated serum bile acids, abnormal bile salt canalicular secretion and failure to thrive have been shown in the *FXR* knock-out mouse (Sinal *et al.*, 2000). Reduction of FXR mRNA expression has been also identified in patients with compound heterozygous missense and nonsense mutations in *ATP8B1* gene and homozygous missense mutations in *BSEP* gene, indicating a contributory role in the inherited cholestatic liver disorders (Alvarez *et al.*, 2004). In addition, three genetic variants in *FXR* are significantly associated with ICP: the variable penetrance of these changes, however, suggests that modifiers or environmental factors may be necessary to promote the disease (Van Mil *et al.*, 2007).

2.1.2.2 Other metabolic pathways

A-1-antritrypsin (A1AT) is an enzyme primarily synthesised in the liver. It represents the most abundant serum serpin protein in the human body. Serine protease inhibitors or serpins constitute a large superfamily of proteins able to inhibit members of a particular class of enzymes, the proteases. To date over 1000 serpins have been discovered in a variety of organisms, of which 36 are human. Involved in numerous biological processes, they play a crucial role in blood coagulation, fibrinolysis, apoptosis and inflammation; therefore they are implicated in the aetiology of several human diseases (Gettins, 2002). A1AT acts as protease inhibitor by binding covalently and irreversibly to enzymes such as trypsin, elastase, chymotrypsin, thrombin and bacterial proteases. Two distinct promoter regions are responsible for the transcription of the two A1AT isoforms: the first isoform is expressed in liver, lung, and intestine due to the activation of a proximal promoter, while the second isoform is activated in macrophages, monocytes and cornea by a distal promoter (Marsden & Fournier, 2005). Correctly known as *SERPINA1*, the gene encodes a 394 amino acid protein that structurally forms three β-sheets and eight α-helices exposing a reactive centre loop (RCL) as inhibitor domain. The protease inhibitor (PI) locus is highly polymorphic; PI phenotypes have been characterised by the speed of migration of different protein variants on isoelectric focusing electrophoresis (IEF). With a medium rate of migration, the most frequent allele in the general population is the PiM allele expressing a normal protein level of A1AT. The PiS allele (resulting from Glu264Val amino acid substitution) and PiZ allele (resulting from Glu342Lys amino acid substitution) have been described as rare variants and associated with A1AT deficiency (AATD). The majority of AATD cases are homozygous for the PiZ allele; the mutation causes a conformational change of the protein enhancing a spontaneous polymerization inside the ER of the hepatocytes. Inside the ER the mutant protein is degraded, the end result being an 85% reduction of normal A1AT expression (Lomas *et al.*, 1992). The clinical presentation of AATD depends on the age of

onset. Chronic obstructive pulmonary disease (COPD) is frequently manifested in adults, while liver disorders can affect children and adults. Neonatal hepatitis and cholestatic jaundice are early symptoms of AATD associated-liver disease, while in adults AATD is diagnosed at the mean age of fifty and it is commonly associated with chronic liver disease, where cirrhosis is the most common manifestation. (Fairbanks & Tavill, 2008). Development of hepatocellular carcinoma is associated with cirrhosis. Hepatocellular injury due to accumulation of A1AT mutated Z protein can activate cell cycle progression and therefore could increase the mutation rate in dividing cells and lead to tumourigenesis (Marcus *et al.*, 2010).

Another inherited metabolic disease that affects liver functionality is caused by mutations in the solute carrier family 25 member 13 (*SLC25A13*). Also named aspartate/glutamate carrier or citrin, it is a calcium dependent inner mitochondrial membrane transporter that mediates the exchange of aspartate for glutamate in the urea cycle. Through a homozygosity mapping strategy, *SLC25A13* gene was mapped for the first time to the adult-onset citrullinemia type II (CTNL2) locus on the long arm of chromosome 7, in a consanguineous family of 18 individuals affected by citrullinemia II (Kobayashi *et al.*, 1999). As with other members of the mitochondrial carrier family, it is structurally composed of three repeat regions and six α-helices segments in the transmembrane domain; in addition the protein exhibits four EF-hand calcium binding sites on the N-terminal. CTLN2 is a liver-specific autosomal recessive disease that manifests with recurrent episodes of hyperammonaemia and psychiatric symptoms such as delirium, loss of memory, and disorientation usually provoked by infection, alcohol/drug intake and surgical intervention during adulthood. In newborns and infants abnormal SLC25A13 synthesis is associated with neonatal-onset citrin deficiency (NICCD) with recurrent episodes of intrahepatic cholestasis, growth failure, jaundice, and hepatomegaly, but with resolution within the first year of life with specific treatment including lactose-free formula and fat-soluble vitamin supplements (Saheki *et al.*, 2002).

2.1.3 **Cell fate determination**

The interaction between ligand and receptor is a vital biological function, crucial for example in the determination of cell fate during development. Small molecules, such as hormones, neurotransmitters, drugs, but also ions and macromolecules create reversible complexes with proteins anchored on the membrane, or localised in the cytoplasm or in the nucleus. Ligand/receptor complexes lead to activation of an intracellular pathway that triggers intracellular events via a direct interaction with transcription binding domains in genomic DNA.

2.1.3.1 Notch signalling pathway

The Notch signalling pathway is a conserved molecular mechanism that controls and regulates a wide range of biological processes, such as neuronal, cardiac and other cell fate decisions. (Fortini, 2009). It is involved in the aetiology of numerous inherited genetic diseases and in cancer. The activation of this pathway occurs through the direct interaction between a Notch receptor expressed on the cellular membrane and a Notch ligand located on the surface of an adjacent cell (Figure 2.1.5). Jagged proteins represent a family of Notch ligands. The Jagged/Notch interaction changes the receptor conformation by exposing the Notch protein to the cleavages of both the extracellular region mediated by the ADAM metalloproteins, and the intracellular region mediated by the presenilin-dependent γ secretase. This interaction releases the Notch intracellular domain (NCID), which translocates into the nucleus and binds with high affinity the transcription factor CSL (also known as CBF1/RBP-J in mammals, Suppressor of Hairless in *Drosophila,* and Lag-1 in *Caenorhabditis elegans*), therefore promoting the transcription of target genes. One of the five members of the Jagged family proteins is Jagged protein 1 (Jagged1). Structurally, it is composed of a single α-helix transmembrane protein and a large extracellular region formed by four domains. The Cys-rich region, two

epithelial grow factor (EGF) regions with potential calcium-binding EGF sites, the Jagged1 binding site for Notch receptor named DSL (Delta/Serrate/Lag-2) domain are found from the nearest membrane boundary to the N-terminus (Pintar *et al.*, 2009). In 1997, genetic defects were discovered in *JAG1*, associated with an autosomal dominant inherited syndrome named Alagille syndrome (AGS) (Li *et al.*, 1997; Oda *et al.*, 1997). Of the mutations identified in *JAG1*, 72% are predicted to cause protein truncation, while in the remaining fraction missense and splice site mutations have been discovered (Spinner *et al.*, 2001). AGS is a complex multi-systemic disorder presenting with variable clinical manifestations involving five biological systems. On the basis of these systems five classical criteria for the diagnosis of Alagille syndrome have been established: early-onset of cholestasis due to paucity of bile ducts, congenital cardiac defects, skeletal abnormalities, ophthalmologic defects and classical facial features (Turnpenny & Ellard, 2012). Additional renal failure and renovascular hypertension can be present (Harendza *et al.*, 2005).

In a high percentage (94%) of patients diagnosed with AGS, pathogenic mutations in *JAG1* have been described; however a small subset of cases have been identified having mutations in one of the four Notch receptors, *NOTCH2* (Figure 2.1.5) (McDaniell *et al.*, 2006). Structurally, Notch2 contains an N-terminal extracellular domain with 35 EGF-like repeats and three lin-12 Notch repeats (LNRs), and the Notch intracellular domain (NCID) (Tien *et al.*, 2009). Heterozygous mutations in the gene include missense, frame shift and nonsense mutations, and have been found mainly in the EGF repeats and in one portion of the NCID. To date, *NOTCH2* represents the second most frequent mutated gene associated with Alagille syndrome (Kamath *et al.*, 2012).

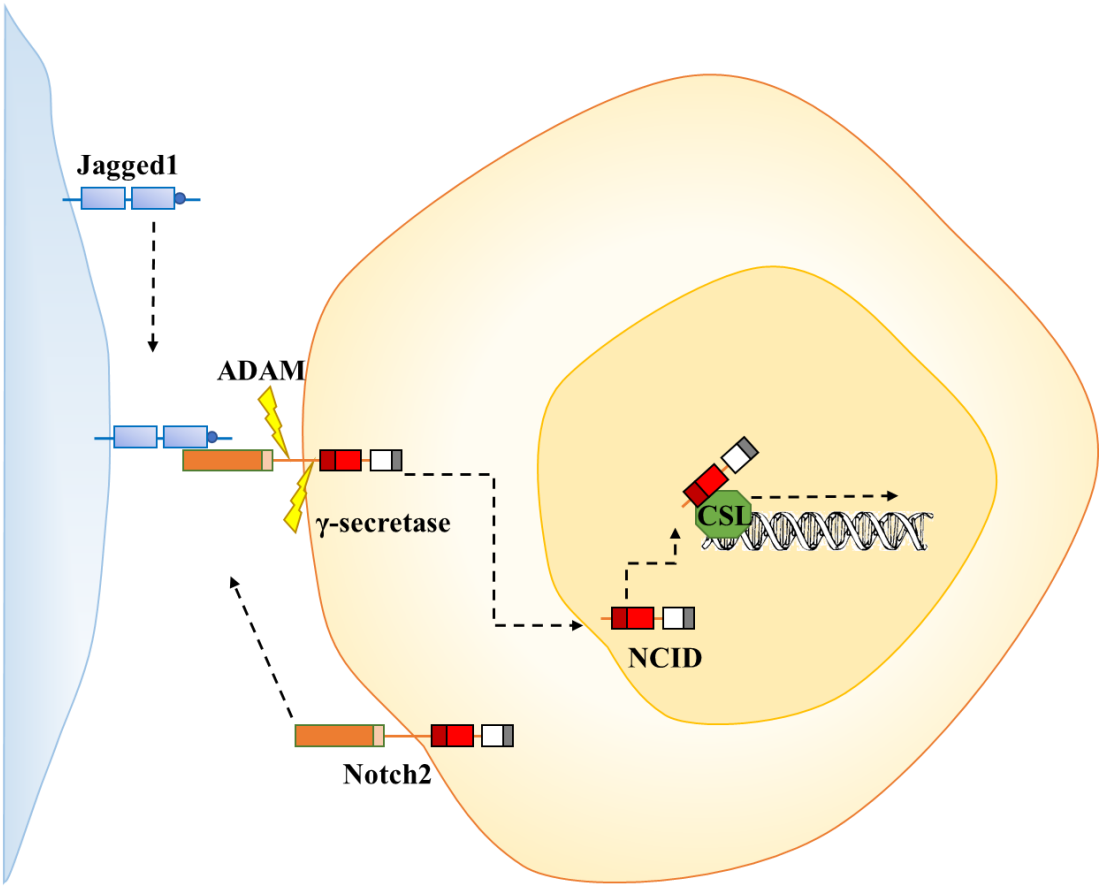

Figure 2.1.5 Jagged1/Notch2 signalling pathway

Jagged1/Notch2 signalling is activated when a physical interaction occurs between the Jagged ligand and the Notch receptor located on the surface of the adjacent cell. This triggers the activation of a double cleavage on the intracellular and extracellular side of the Notch receptor and the release of the Notch intracellular region (NCID). NCID then translocates to the nucleus where it binds the transcription factor CSL (CBF1, Suppressor of Hairless, and Lag-1) which activates the transcription of target genes.

2.1.4 **Cell-cell junctions: tight junctions**

The interaction between neighbouring cells or between cell and extracellular matrix is mediated by multi-protein complexes, known as cell junction proteins (Figure 2.1.6). On the basis of their function, they can be categorised into three different groups: 1) occludin junctions or tight junctions, important in sealing adjacent epithelial cells, determining cell polarity, and creating a selective barrier, that prevents and regulates the paracellular diffusion of water or small proteins; 2) adherens junctions, that play a role in the physical attachment between cell-cell and cell-extracellular matrix; 3) gap junctions, that, through the formation of channels, mediate the transcellular passage of small molecules or chemical compounds between adjacent cells (Alberts *et al.*, 2002a).

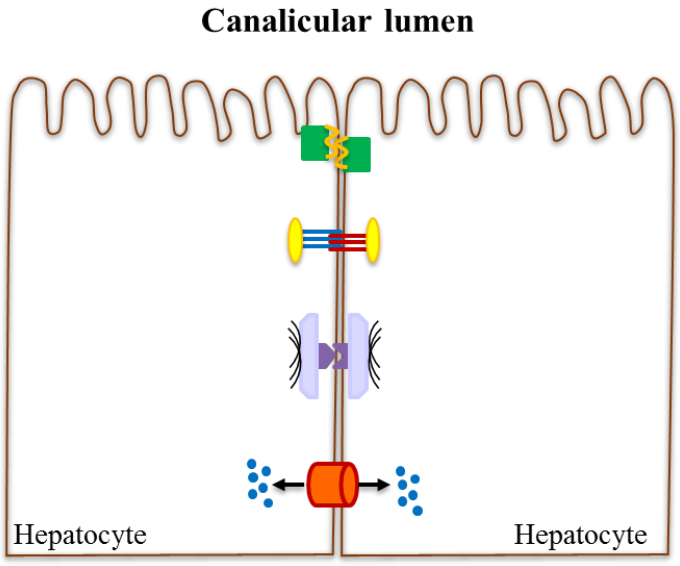

Figure 2.1.6 Representation of cell-cell junctions in hepatic epithelial cells

From the apical surface facing towards the canalicular lumen to the basolateral surface of hepatic epithelial cells are localised tight junctions, adherens junctions, desmosomes and gap junctions. Genetic alteration of component of tight junctions have been associated to cholestatic conditions. In the section 2.1.4, the different functions are described.

Different organs in the human body are composed of epithelial cells, including liver (hepatocytes) and biliary tree (cholangiocytes). The most apical junctional complex identified in epithelial cells are the tight junctions. This multi-protein complex structure is constituted of transmembrane proteins such as claudins, occludins and junctional adhesion molecules (JAM), and cytoplasmic proteins such as tight junction proteins (TJP) (Sawada, 2013) As mentioned above, in addition to the property of tethering two neighbouring cells, tight junctions confer cell polarity (fence function) and create a barrier (gate function). Different human diseases have been shown to be related to disturbance of tight junction functions. Alteration in the gate function, for example, have been identified in various inherited disorders, such as deafness, neonatal sclerosing cholangitis associated with ichthyosis (NISCH) and FHC.

Claudins are a large family of 24 human integral membrane proteins that play an important role in the tight junction structure (Angelow *et al.*, 2008) (Figure 2.1.6). They are formed of four membrane-spanning domains, two extracellular loops and a cytoplasmic COOH-terminal motif. Through the extracellular domain, claudins undergo homo or heterodimerization followed by trans-interaction between the small extracellular loops of two claudins, located on the opposite cell membranes (Piontek *et al.*, 2008). This process results in a network of continuous strands that seal the paracellular space between two adjacent cells. On the other side of the plasma membrane, the intracellular domain of the claudins binds directly the tight junction-associated proteins, including the scaffolding tight junction proteins (TJP), otherwise called zona occludens proteins (ZO). These create a link between the transmembrane tight junction proteins and the actin cytoskeleton. Claudin expression varies amongst different tissues; in the liver, claudin-1 is most represented (Furuse *et al.*, 1998). Homozygous deletion of 2 bp in the first exon of the claudin-1 gene (*CLDN1)* has been described, which leads to an absence of claudin-1 protein, and is associated with the pathogenesis of NISCH (Hadj-Rabia *et al.*, 2004). This rare genetic condition is characterised by chronic inflammation

and fibrosis in the intra and extrahepatic bile ducts, associated with hypotrichosis and ichthyosis. It leads to liver failure. In addition, the larger extracellular loop (EL1) of claudin-1 has been identified as a co-receptor with CD81 in the hepatitis C virus entry pathway (Evans *et al.*, 2007). Worldwide hepatitis C is currently the leading cause of cirrhosis and hepatocellular carcinoma.

The cytoplasmic plaque of tight junctions consists of numerous proteins, most of which share the common PDZ (PSD-95/Dig/Zo-1) domain in their structure. The PDZ domain is a highly conserved region that binds specifically the PDZ-binding site located on the intracellular COOH-terminal of integral proteins, such as claudins (Itoh *et al.*, 1999). The major group of proteins in the tight junction plaque are the tight junction proteins. They belong to the membrane-associated guanylate kinase (MAGUK) family, composed of three functionally essential domains – PDZ domain, Src domain homology 3 (SH_3) and guanylate kinase (GK)-like domain - and a C-terminal proline rich region (Balda & Matter, 2008). These proteins are identified as scaffolding proteins, due to their double action of binding on one side the transmembrane proteins and on the opposite side the actin cytoskeleton. Zona occludens 2 (ZO-2) is one of the three members of the ZO protein family; as it is a tight junction associated protein, it is also known as TJP2. TJP2 was characterised as a cytoplasmic component of the tight junction plaque after a co-immunoprecipitation study, where it was discovered to have a close interaction with ZO-1 (Gumbiner *et al.*, 1991). A single homozygous mutation c.143T>C was identified associated with FHC, as described in section 2.1.2.1. This mutation affects the first PDZ domain, decreasing the binding affinity to the C-terminal of the claudins. However, because the *TJP2* mutation did not account for all FHC patients, a possible oligogenic inheritance was identified with a single change in *BAAT* (Carlton *et al.*, 2003). In addition, a tandem inverted genomic duplication that includes the entire *TJP2* has been described in progressive non-syndromic hearing loss (Walsh *et al.*, 2010). This finding suggests that overexpression of ZO-2 leads to the activation of the apoptosis pathways in the inner ear cells. Recently,

two heterozygous variants, c.334G>A (p.Ala112Thr) and c.3562A>G (p.Thr1188Ala), were identified in Korean families with autosomal dominant non-syndromic hearing loss. The pathogenicity role of this finding has been proven only with *in silico* prediction, and no further functional studies have been undertaken (Kim *et al.*, 2014).

2.1.5 **Protein trafficking**

Protein trafficking is a crucial post-translational mechanism for several biological functions, including cell compartment development and regulation, neurotransmitter and hormone secretion and cell growth. Nascent polypeptides are targeted for selected intracellular compartments - such as plasma membrane, ER, Golgi complex, lysosomes, mitochondria or extracellular compartments - where they are able to perform their specific functions. This protein localisation is mediated by membrane-enclosed vesicles (Carr & Rizo, 2010). Vesicle formation requires a complex set of cytosolic proteins classified into four major types: 1) Soluble NSF Attachment Protein (SNAP) receptors (SNAREs), which represent a large class of vesicle (v-SNARE) or target-membrane (t-SNARE) anchored proteins that play a primary role on vesicle membrane fusion; 2) N-ethylmaleimide-sensitive factor (NSF), which is an ATP dependent enzyme involved in the disassembly and recycling of SNARE complex proteins after membrane fusion and vesicle formation; 3) Rab/family GTPases, that generate the energy utilised in vesicle attachment and SNARE complex assembly; 4) Sec1/Munc18 (SM) proteins that through the direct interaction with the SNARE proteins, regulate the vesicle attachment and membrane fusion (Alberts *et al.*, 2002b).

Vacuolar protein sorting 33 yeast homolog B (VPS33B) belongs to the SM proteins. Binding directly to the N-terminus of synthaxin proteins, which are members of target-membrane anchored SNAREs, it aids the mechanism of

membrane fusion during vesicle formation (Gissen *et al.*, 2004). VPS33B is expressed ubiquitously in human tissues (Huizing *et al.*, 2001). Like the other SM factors, it is composed of three domains which interact creating an arch shape protein with a large cavity on the opposite side, where SNARE proteins are bound. A rare autosomal recessive multisystem disorder, known as ARC (arthrogryposis, renal dysfunction and cholestasis) syndrome, is associated with protein-truncating mutations in *VSP33B* leading to dysfunction in the protein trafficking (Gissen *et al.*, 2004). Phenotypically, in addition to the three cardinal features described above, this neonatal syndrome can manifest with failure to thrive, neurological symptoms, platelet-function abnormality, nephrogenic diabetes insipidus, and deafness. Biochemical analysis of the liver disease shows a low concentration of serum GGT and a near normal activity of aspartate aminotransferase (AST) and alanine aminotransferase (ALT) (Jang *et al.*, 2009). In addition, an incomplete ARC phenotype has been identified in a patient diagnosed with a single-base homozygous deletion leading to absence of the VSP33B protein. In this patient, all other ARC features were present, but not arthrogryposis (Bull *et al.*, 2006).

In 2010, mutations in another related gene were identified in a cohort of individuals with ARC syndrome having different ethnic backgrounds and no genetic defects in *VPS33B* (Cullinane *et al.*, 2010). VPS33B interacting protein, apical-basolateral polarity regulator, spe-39 homolog (VIPAS39) forms a complex with VPS33B. Then, the VPS33B-VIPAS39 complex interacts with the apical membrane protein RAB11A, a small GTPase protein. RAB11A belongs to the Rab family and is primarily involved in the apical membrane recycling pathway but also plays an important role in the early intracellular mechanism of vesicle trafficking (Wilcke *et al.*, 2000). In the liver of ARC individuals, defects in the VPS33B-VIPAS39 complex lead to an apical membrane mis-localisation of BSEP, MRP2 and the adhesion molecule CEACAM5, highlighting the importance of this complex in the intracellular trafficking and regulation of epithelial polarization (Cullinane *et al.*, 2010).

2.2 Research hypothesis and aim

Progressive familial intrahepatic cholestasis represents a heterogeneous group of genetic conditions, which have an early-onset in paediatric patients. To date, several genes have been implicated in this pathology, such as genes involved in the formation of the bile. In addition, cholestatic features are present in other genetic disorders, like cystic fibrosis, Alagille syndrome, neonatal sclerosing cholangitis, ARC syndrome and several enzyme deficiencies. However, at least a third of paediatric patients with a clinical diagnosis of progressive familial intrahepatic cholestasis do not have a defined genetic cause, suggesting that other biological pathways might have been involved in the aetiology of cholestasis. Therefore, the aim of this first part of the project is to investigate new genetic causes of cholestatic liver disorders, in the subset of idiopathic paediatric patients, through a combination of next-generation sequencing (NGS) technology approaches.

2.3 Materials and Methods

2.3.1 Patients

For the initial part of this study, 25 paediatric patients, including three siblings, were selected from King's College Hospital Paediatric Liver Tissue Bank database. All were clinically diagnosed with early-onset congenital intrahepatic cholestasis. Biochemical analysis showed normal, or upper limit of normal, activity of serum concentration of GGT, and raised concentration of serum alanine transaminase (ALT), aspartate transaminase (AST) and/or serum bile acids. Genetic testing for *ABCB11* and *ATP8B1* had been previously undertaken as part of the diagnostic protocol and two mutated alleles had not been identified. The majority of the individuals belonged to consanguineous families. Due to low DNA concentration, four patients were excluded from the analysis. Eighteen children, one individual per family, were selected to undergo the genetic analysis through NGS; clinical information is summarised in Table 2.3.1.

After analysing the NGS data and identifying the disease-causing gene, the cohort for study was enlarged. Seventy more paediatric patients were selected by following the previous selection criteria. Individuals with a milder phenotype, such as later age of onset and recurrent episodes of cholestasis, were also included. However, five children were excluded to the data analysis due to the failure in the sequencing. The analysis was therefore carried out for 65 paediatric patients.

Family number	Case number	Sex	Consanguinity	Early, post neonatal GGT (UI/l)	Age of GGT measurement (months)	Bilirubin (µmol/l)	ALT (IU/l)	AST (IU/l)	Serum Bile acids (µmol/l)
1	1	M	Y	84	23	591	74	109	670
2	2	M	Y	15	6	15	43	61	132
3	3	M	Y	98	13	7	142	87	16
4	4a	M	Y	2	74	64	94	-	-
	4b	F	Y	58	4	31	156	-	-
5	5	M	N	44	86	11	95	63	106
6	6	F	N	15	36	43	92	75	max 20
7	7	F	Y	41	4	112	65	163	219
8	8	M	Y	65	2	203	-	479	6
9	9	F	Y	44	3	129	-	112	-
10	10a	M	Y	45	26	120	-	227	165
	10b	M	Y	75	17	241	94	143	-
11	11a	F	Y	22	21	284	-	156	-
	11b	M	Y	27	21	397	-	234	-
12	12a	F	Y	63	24	264	-	285	261
	12b	F	Y	304	0.75	139	-	43	100
13	13	F	Y	45	7	178	-	107	99
14	14	M	Y	-	-	-	-	-	-
15	15	M	N	-	-	-	-	-	-
16	16	F	N	-	-	-	-	-	-
17	17	F	Y	-	-	-	-	-	-
18	18	M	N	361	9	154	55	80	-

Table 2.3.1 Clinical characteristics of the 18 patients selected for the first targeted resequencing (TRS-21)

At 2 months the normal activity of GGT is <200 IU/l; by 9 months is <55 IU/l. Normal bilirubin is <21 µmol/l. ALT normal range is 5-55 IU/l. AST normal range in infants is <75 IU/l, in older children is <36 IU/l. Serum bile acids are normally <14 µmol/l. A dash indicates information not available. M: male; F: female; Y: yes; N: no; GGT: gamma-glutamyl transferase. ALT: alanine aminotransferase; AST: aspartate aminotransferase.

2.3.2 **DNA isolation from whole blood**

Genomic DNA was isolated from whole blood by QIAamp DNA Blood Mini kit, a solid-phase extraction method provided by Qiagen Ltd, Manchester, UK. In order to yield from 3 μg to 12 μg of DNA, 200 μl of patients' blood sample preserved with EDTA buffer were used 20 μl of Qiagen protease and 200 μl of lysis Buffer AL were added to the sample volume, mixed by vortexing and then incubated on a heat block for 10 minutes at 56°C. Afterwards, 200 μl of ethanol (96-100%) were added to each sample and the solution was loaded into QIAamp Mini spin columns and centrifuged in a Eppendorf Microcentrifuge 5417R (Eppendorf UK Limited, Stevenage, UK) at 8,000 rpm (revolution per minute) for 1 minute. These chromatographic columns are formed of a stationary phase of mainly silica gel, which binds nucleic acids. To improve the purity, the DNA bound on the QIAamp Mini spin column was washed with 500 μl of buffer AW1, centrifuged at 8,000 rpm for 1 minute, and subsequently washed with 500 μl of buffer AW2, centrifuged at 14,000 rpm for 3 minutes. The collection tube with the filtrate was discarded and the QIAamp Mini spin column was placed in a clean 1.5 ml tube. Two hundred μl of elution buffer AE were applied to the column and incubated at room temperature for 5 minutes; the isolated DNA was then eluted in the clean tube after a final spinning at 8,000 rpm for 1 minute.

2.3.3 **DNA quantification**

Eluted DNA was quantified using Qubit 2.0 Fluorometer provided by Invitrogen (Life Technology, Paisley, UK). The Qubit assay uses dyes specific for each biological molecule, such as DNA, RNA or protein that are able to emit fluorescence only after binding the target. The amount of fluorescence is directly

proportional to the concentration of the target analysed. The fluorescent signal emitted by each sample is captured by Qubit software and converted into a measurement using a specific standard curve. The Qubit broad-range (BR) assay kit is used to quantify the dsDNA concentration between 2 ng and 1000 ng starting from a small amount of 1 µl of DNA. A Quanti-IT Working Solution was prepared for each sample and for the two dsDNA BR standards provided by Qubit dsDNA BR assay kit adding 199 µl Quanti-IT buffer and 1 µl Quanti-IT reagent (dilution 1:200) in the test tubes to reach a final volume of 200 µl. Afterwards, for the standard assay 190 µl of Quanti-IT Working Solution were added to 10 µl of each standard, while for the sample assay 199 µl of Quanti-IT Working Solution were added to 1 µl of each sample. All tubes were mixed by vortexing and incubated for 2 minutes at room temperature. The reading was performed in the Qubit 2.0 Fluorometer selecting the appropriate programme on the home screen. After selecting the Qubit dsDNA BR programme, the quantification was performed firstly for the standards, in order to define the standard curve, and secondly for the samples, following the manufacturer's guidelines.

2.3.4 **Next-generation sequencing**

Since 2004, the advent of NGS has completely changed the approach to genetic research (Shendure & Ji, 2008). This high throughput sequencing technology offers the possibility of studying millions of DNA fragments in parallel, derived, for example from a few kilobases (kb) of targeted regions of interest (ROI) to 50 megabases (Mb) of whole-exome, and up to 4 gigabases (Gb) of the whole human genome. Currently several platforms and applications have been produced and improved, allowing a higher performance in less time. For NGS study, the genetic material needs to undergo a series of modifications in order to construct a library of DNA suitable for the following sequencing stage. During this process, the specific regions of interest are captured. In this study, two separate hybridisation

systems were adopted. Subsequently, the library is amplified to generate cloned clusters and massively parallel sequenced. Illumina's sequencing technology was selected as NGS technology approach.

2.3.4.1 Probe design for capturing ROI

The selected ROI includes genes known to be involved in the aetiology of cholestatic liver diseases, or other inherited disorders in which cholestasis is a clinical feature. The molecular mechanisms and the diseases associated with these genes are described in section 2.1. For the initial group of 18 patients, 21 genes were selected covering approximately 150 kb. Within this gene panel, seven were analysed in the subsequent enlarged cohort. Gene annotation was obtained from Ensembl and University of California Santa Cruz (UCSC) datasets, using the human genome assembly GRCh37.p7/hg19, released in February 2009. The summary of the genes selected is described in Table 2.1.1. The DNA samples from the initial group of 18 patients were processed using Agilent SureSelectXT Target Enrichment kit (Agilent Technology UK Limited, Cheshire, UK) for Illumina technology. This system requires a library of biotinylated single strand RNA probes, called baits, to capture the genomic ROI, which are subsequently enriched and sequenced using a high throughput Illumina sequencing platform (Illumina, Chesterford Research Park, Essex, UK). The bait library was designed using Agilent eArray web portal (https://earray.chem.agilent.com/earray). Each bait was designed of 120-mer in length centred to the ROI with 40 base pairs (bp) of nucleotides overlapping with the boundaries. In addition, the design strategy was optimised for covering each base at least 4 times. Typically, poor sequence coverage and consequently inadequate base-calling detection has been demonstrated in the area of GC content compared with the rest of the genome (Sulonen *et al.*, 2011; Wang *et al.*, 2011). Thus, using *SureSelect Bin Orphan and High GC Baits Workflow* from the Galaxy web-based platform the designed bait library was divided into six different groups according to the amount of GC

content: less than 50%, between 50% and 55%, between 55% and 60%, between 60% and 65%, between 65% and 70% and greater than 70%. Then, for each group the number of copies of each bait was increased a number of times in relation to the percentage of GC content, respectively 1x, 2x, 4x, 6x, 9x and 16x. The capture library was designed in conjunction with a research project within the Department of Molecular Haematology including a panel of 102 genes with a final size of 1.05 Mb. The custom probes were submitted to the manufacturer.

Whole-exome sequencing (WES) was also performed. The library of probes was pre-manufactured by Agilent Technologies (SureSelectXT Human All Exon V4 kit) and was designed to capture all known exons covering a target region of 51 Mb. WES was performed on 7 patients, selected from the initial 18, in whom no mutations were found, all belonging to consanguineous families.

The DNA from the expanded cohort of 70 patients was processed using TruSeq Custom Amplicon kit (Illumina). This system requires oligo probes upstream and downstream flanking the specific ROI (Table 2.1.1), which are extended and ligated across the ROI. The amplicons are then PCR (polymerase chain reaction)-enriched and sequenced using the Illumina sequencing platform. The design of the oligo probes was undertaken using the Illumina web-based sequencing assay design tool DesignStudio (http://designstudio.illumina.com/). Seven genes, included in the previous target enrichment panel, were selected with a total coverage of 60 kb. Oligonucleotides were designed to generate a total of 475 amplicons, each of them having a length of approximately 250 bp.

2.3.4.2 Library preparation: Agilent SureSelect Target Enrichment System

The SureSelect Target Enrichment workflow was used for the initial target resequencing of the 21 gene panel described above (TRS-21), as well as for the whole-exome sequencing (WES). Both libraries were sequenced by Illumina HiSeq 2000. The workflow is divided into three different phases: pre-hybridisation, hybridisation and post-hybridisation (Figure 2.3.1).

In the pre-hybridisation step the genomic DNA is fragmented by Covaris (Marlow Buckinghamshire, UK). This technology uses the Adaptive Focused Acoustics (AFA) system; it generates precise and controlled ultrasonic acoustic waves in an isothermal water bath that are able to shear genetic material in fragments of a specific size range. Three µg of DNA of each sample were diluted in 130 µl of elution buffer (EB) and sheared into fragments between 150 to 200 bp in length. The Covaris shear setting was in accordance with SureSelect Target Enrichment protocol v1.3.1, February 2012.

The quality control and the length size of each sample were assessed using Agilent 2200 TapeStation System; an automated electrophoresis method. D1K ScreenTape with D1K reagents (ladder and sample buffer) were chosen to analyse the fragment length of the sheared DNA library. In optical strip tubes, 3 µl of ladder were placed in the first tube and 3 µl of sample buffer were mixed with 1 µl of each sample in the sequential tubes. After vortexing and spinning down, the optical tube strips were placed on the TapeStation instrument together with the specific loading tips and the D1K ScreenTape. The TapeStation Controller software was launched and the required sample panel was selected on the controller framework before starting the electrophoresis run. The distribution of the target peak height was investigated using TapeStation analysis software (Figure 2.3.2a).

Afterwards, the sheared DNA library was purified using Agencourt AMPure XP beads (Beckman Coulter Ltd, High Wycombe, UK), paramagnetic beads that can

recover DNA fragments greater than 100 bp. One hundred and thirty µl of sheared DNA library of each sample were mixed with 180 µl of homogenous AMPure XP beads in a 1.5 ml tube. After a 5-minute incubation the solution was placed on a magnetic rack to separate the beads binding the DNA fragments from impurities. The beads/fragments were washed with fresh 70% ethanol twice and the DNA fragments eluted with 50 µl of nuclease-free water.

Next, the purified sheared DNA library was end-repaired, dATP tailed, ligated with indexing-specific paired-end adapters and amplified as described in the Agilent SureSelect Target Enrichment workflow (Figure 2.3.1). The reaction mix for each step was prepared in accordance with the SureSelect Target Enrichment protocol and the adapter-ligated library amplification was optimised for 6 cycles. The DNA library was purified after all steps using Agencourt AMPure XP beads. The quality of adapter-ligated library was assessed using the Agilent 2200 TapeStation with D1K ScreenTape with D1K reagents (ladder and sample buffer). Clear evidence of adapter ligation was shown with a shift of peak distribution of approximately 100 bp (size range of 250 to 275 bp) (Figure 2.3.2b).

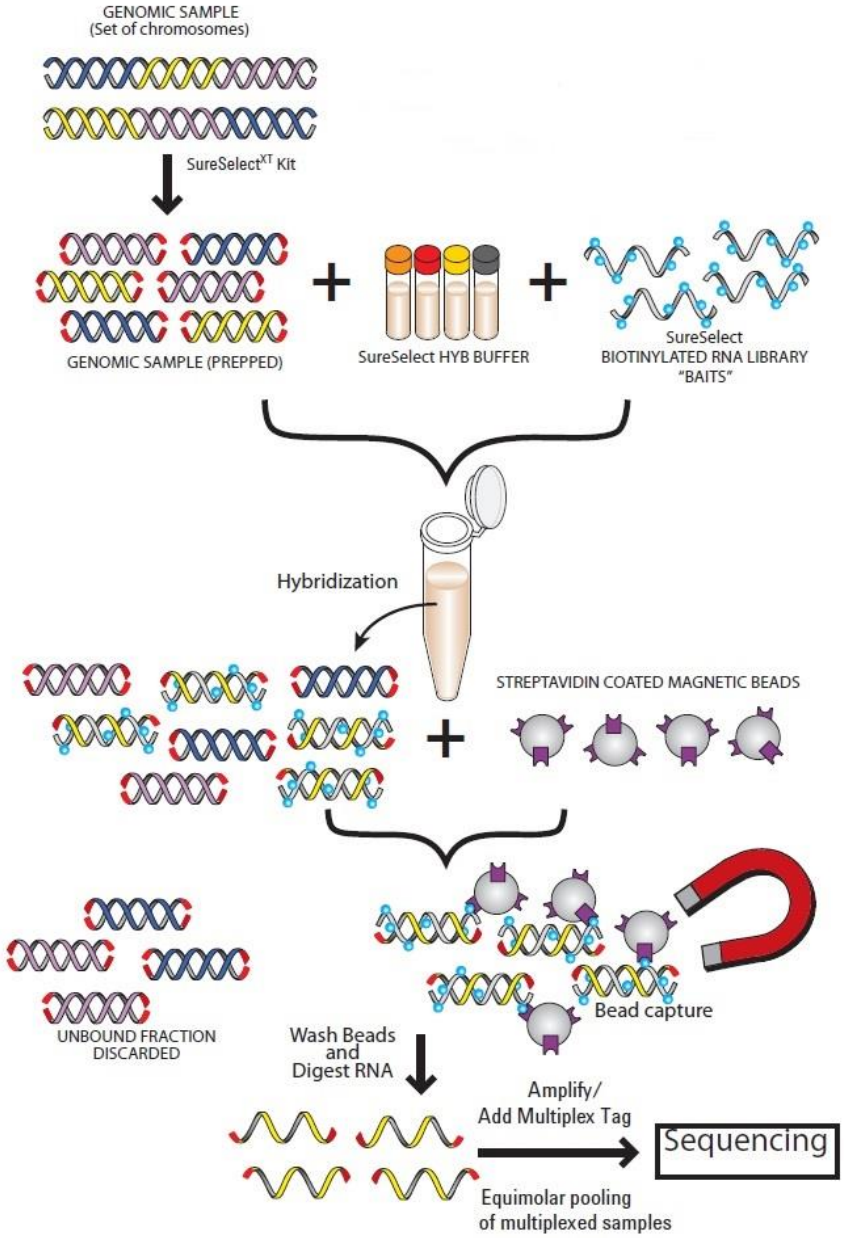

Figure 2.3.1 Agilent SureSelect Target Enrichment workflow for Illumina platform.

Representation of the workflow used for the library preparation of the targeted resequencing (TRS-21) and whole-exome sequencing (WES) analysis, including the three main stage-pre-hybridisation, hybridisation and post-hybridisation- described in section 2.3.4.2. The figure is provided by Agilent Technology.

During the second phase of the SureSelect Target Enrichment workflow the previously prepared library was hybridised with the designed capture library for the 21 panel genes or the whole-exome described in section 2.3.4.1. For each sample 500 ng of amplified adapter-ligated library were required in a maximum volume of 3.4 µl. The DNA concentration was quantified using Qubit 2.0 Fluorometer with Qubit broad-range system assay kit (discussed in section 2.3.3). For the samples with a concentration below 147 ng/µl, aliquots of 500 ng were prepared and concentrated down to 3.4 µl using a vacuum concentrator at ≤ 45°C. Then, the DNA library of each sample was added to the "B" row in a PCR plate with 5.6 µl of SureSelect Block Mix previously prepared in accordance with the SureSelect Target Enrichment protocol. The PCR plate was placed on a thermal cycler at 95°C for 5 minutes and then lowered to 65°C with a heated lid at 105°C. During this process the prepared DNA library is denatured. Maintaining the PCR plate on the thermal cycler at 65°C, 40 µl of hybridisation buffer for each sample, previously prepared in accordance with the SureSelect Target Enrichment protocol and warmed at 65°C for 5 minutes, were loaded on the "A" row. The SureSelect capture library solution was prepared by adding 2 µl of custom biotinylated bait library and 5 µl of diluted RNase Block, on the basis of the capture size for each sample; one part of RNase Block was diluted with nine parts of nuclease-free water (dilution 1:9). Maintaining the PCR plate on the thermal cycler at 65°C, the SureSelect Capture Library solution was added on row "C". After 2-minute incubation the two solutions in row "A" and row "B" were combined with the SureSelect capture library mix the row "C" using a multichannel pipette. The hybridisation mixture in the row C was then well-sealed to reduce the evaporation, and incubated for 24 hours at 65°C with the heated lid at 105°C. Excessive evaporation, greater than 10 µl, may interfere with the capture performance; therefore, special strip caps were used to reduce evaporation rates.

The post-hybridisation stage was performed after the one-day incubation. Dynal MyOne Streptavidin T1 provided by Invitrogen (Life Technology) was used to pull

out the heteroduplexes, formed by the biotinylated RNA bait and the complementary DNA fragment, from the unbound DNA fraction. This technique combines the magnetic properties of the Dynabeads with the strong affinity between biotin and streptavidin. The Dynabeads were washed and re-suspended in 200 µl of SureSelect binding buffer and then added to the hybridisation mixture, which might have lost volume due to evaporation after the 24-hour incubation. The selective RNA:DNA capture was performed in accordance with the SureSelect Target Enrichment protocol. Subsequently, the capture library was purified using Agencourt AMPure XP beads. Fourteen µl of each sample were enriched and tagged with primer indexes or barcodes. Inserting an end-tag specific for each sample enables single-run sequencing of different samples and to discriminate the data of each patient during the analysis. The PCR reaction was prepared following the manufacturer's protocol with an optimal cycle number of 14. The assay is manufactured with 12 different indexes to allow the simultaneous sequencing of 12 samples. So, the 18 samples were divided into two batches, one with 12 samples and one with 6 samples and tagged respectively with the 12 index primers and with the first 6 index primers. The amplified library was purified using 90 µl of homogenous AMPure beads. The quality was assessed with High Sensitivity D1K ScreenTape and reagents for Agilent 2200 TapeStation, and the quantity with Qubit high-sensitivity DNA assay. The size range of the amplified capture library was approximately between 300 and 400 nucleotides (Figure 2.3.2c) and the DNA concentration between 1.6 ng/µl and 1.8 ng/µl. The 12 samples in one tube and the 6 samples in another tube were pooled in equimolar proportions totalling 1 nM. The final concentration of the pooled library was measured by Qubit high-sensitivity DNA assay. The same protocol was used for WES. The 7 patients were divided into two groups composed of 4 and 3 samples respectively, in order to obtain an estimated final coverage of each base (read depth) of at least 20 times. The estimation was calculated on the basis of the targeted region and the maximum system data output (37.5 Gb per lane)

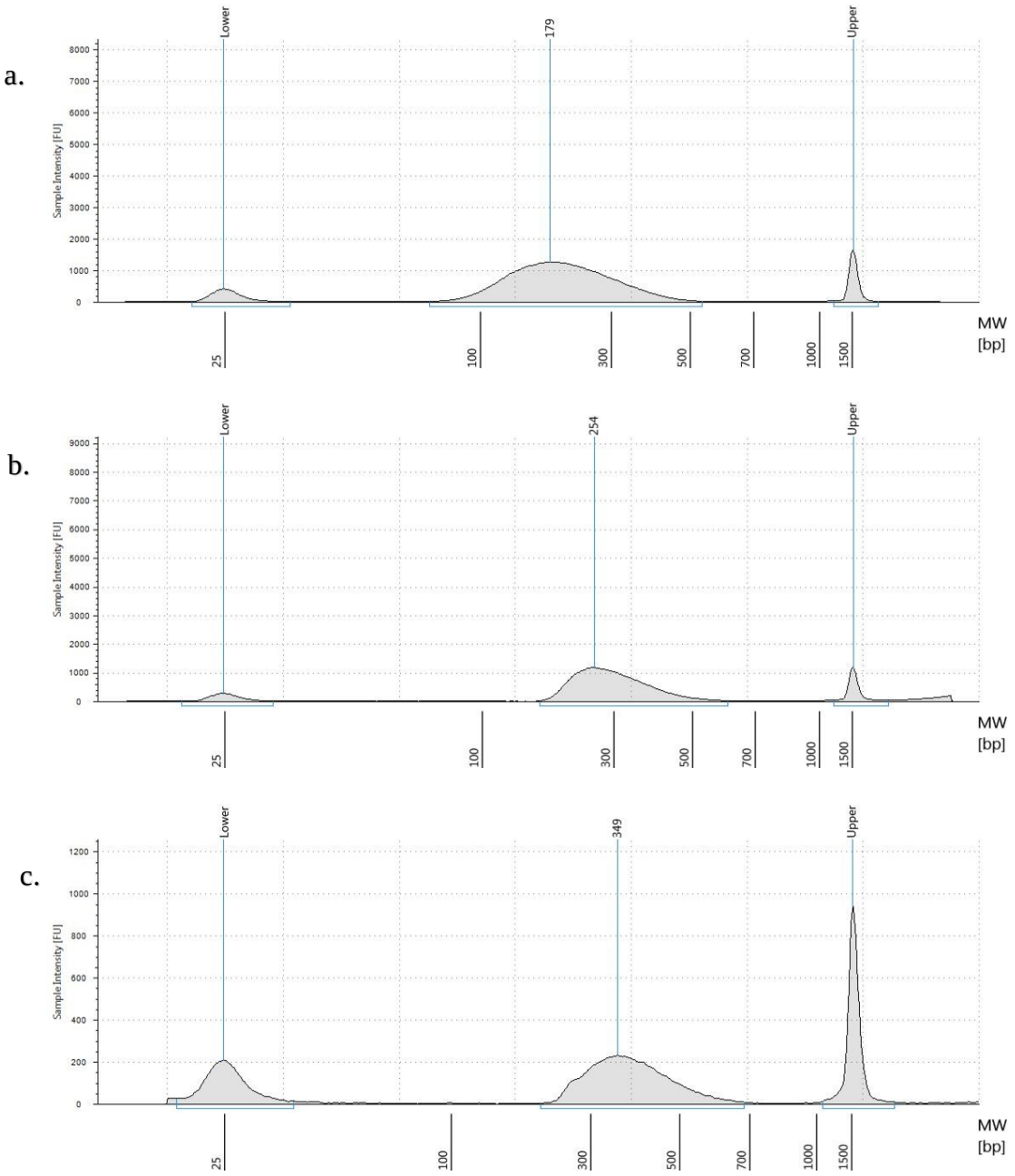

Figure 2.3.2 Representative electropherograms of the DNA library from three stages of the library preparation

a) Distribution of the sheared DNA with a mean peak size of 179 bp (optimal range: 150-200 nt); b)distribution of DNA library after adapter ligation. The mean peak size of the fragments increased by approximately 100 bp; c) distribution of the DNA library after index tagging, reaching a final peak size of 349 bp (optimal range: 300-400 nt).

2.3.4.3 Library preparation: Illumina TruSeq Custom Amplicon

Illumina TruSeq Custom Amplicon kit (TSCA) was used for the library preparation of the enlarged cohort of 70 DNA samples targeting 7 genes (TRS-7). All samples prepared by TSCA were sequenced on the MiSeq system. Before starting, a plate layout of the samples was created. This step is important during the subsequent PCR amplification where index primers are used. The illustration of the workflow is shown in Figure 2.3.3.

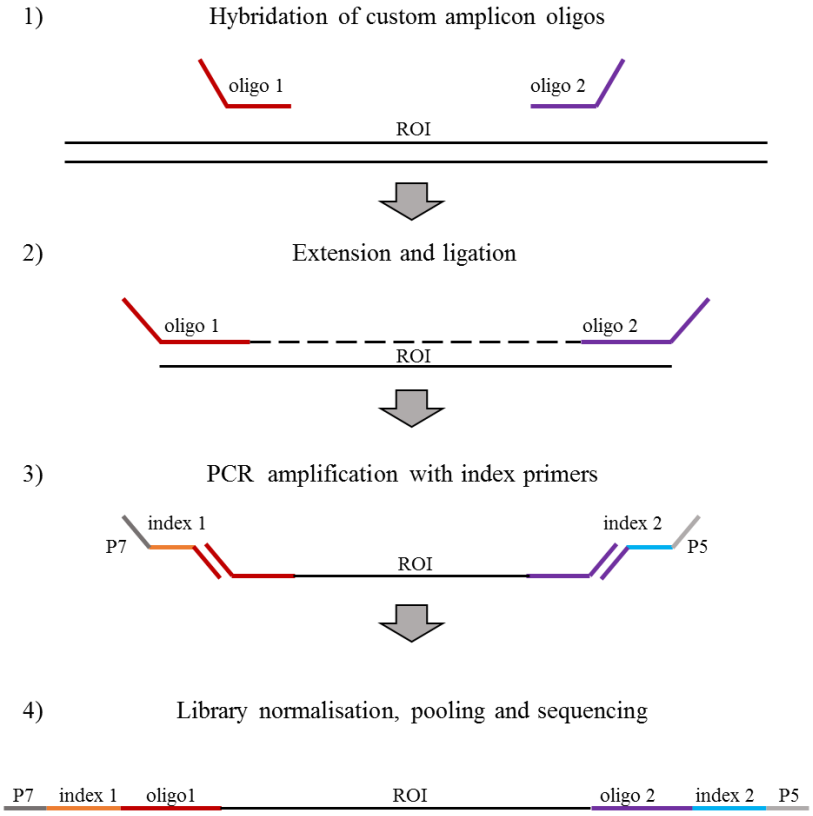

Figure 2.3.3 TruSeq Custom Amplicon workflow for Illumina platform

Simplified workflow adopted for the library preparation of the larger cohort of samples selected for TRS-7. The three main stages are here represented: 1) hybridisation of the oligos customised for specific regions of interest (ROI) to the DNA; 2) extension and ligation of the oligos having specific DNA fragments; 3-4) PCR amplification using primers with two indexes or barcodes for patient. The adapters P5 and P7 were also included in the primers. These sequences are important to allow the binding between the library fragments and the flow cell surface. The figure has been adapted from the original in the TSCA Illumina user guide.

The TSCA system is designed for up to 96 samples for one sequencing run; however in practice batches of no more than 16 samples were prepared. Five µl of DNA at 50 ng/µl are required as a starting material for each sample. On a PCR plate, 5 µl of control DNA, provided by the kit, were added in the first well to 5 µl of control oligo pool and in the remaining wells, 5 µl of patient DNA with 5 µl of custom amplicon oligos (section 2.3.4.1). Using a multichannel pipet, 35 µl of hybridisation buffer were gently mixed to each sample pipetting up and down for 5 times. After sealing with an aluminium adhesive sealer, the plate was placed on a pre-heated block at 95°C for 1 minute; then, the temperature was set at 40°C and the plate was left to incubate for approximately 80 minutes, until the lower temperature was reached. This gradual cooling is a key part of the DNA:oligo hybridisation. In the meantime, a filter plate unit was assembled in accordance with the manufacturer's protocol; it was pre-washed with 45 µl of wash buffer SW1 and centrifuged in a microcentrifuge at 4100 rpm at 20°C for 10 minutes. At the end of the 80-minute incubation, the hybridisation plate was spun down and each well-content transferred into the pre-washed filter plate unit maintaining the plate layout created at the beginning of the experiment. The filter plate was then covered with the appropriate lid and centrifuged at 4100 rpm for 5 minutes. Using a multichannel pipette, 45 µl of SW1 were added to each well and then the plate centrifuged at 4100 rpm for 5 minutes. The washing step was repeated for a second time and the flow-through discarded. In order to remove all the unbound oligos, 45 µl of universal buffer UB1 were lastly added and then the plate centrifuged at 4100 rpm for 5 minutes. At this point, the oligos are bound upstream and downstream of the ROI and the extension-ligation action is required. Forty-five µl of extension-ligation mix ELM4 were added to each sample-well using a multichannel pipet; the plate was covered with an adhesive seal and incubated for 45 minutes at 37°C. During the incubation period, materials and reagents for the following amplification stage were prepared. In accordance with the Illumina Experiment Manager, the index primers were determined and recorded on the plate layout. In deciding of the index, it is important to ensure that at each sequencing cycle at least

one of two nucleotides for each channel colour is read. The green laser is used to sequence G/T and the red laser to sequence A/C. The image registration can be affected if there is a colour imbalance. The TSCA method requires two index primers (i5 and i7) for each sample. Using TruSeq Index Plate Fixture provided by Illumina, the i5 index primers were arranged vertically, the i7 index primers were arranged horizontally and a clean PCR plate for the index amplification stage was placed in the middle empty space in accordance with the manufacturer's protocol. Using a multichannel pipette, 4 µl of i5 primers were dispensed to each column and 4 µl of i7 primers to each row following the plate layout. After 45-minute incubation, the extension-ligation plate was spun down at 4100 rpm for 5 minutes and 25 µl of 50 nM NaOH were added to each well. To ensure that NaOH was interacting with the full well-content, the solution was mixed by pipetting up and down 5 times and then incubated at room temperature for 5 minutes. During the 5-minute incubation, the PCR mix was prepared with 0.6 µl of DNA polymerase TDP1 and 29 µl of PCR master mix PMM2 per sample. After inverting and spinning down the solution, 22 µl of PCR mix were added to each well of the index amplification plate previously placed with the index primers on the TruSeq Index Plate Fixture. From the extension-ligation plate, 20 µl of the each well-content were then transferred to the index amplification plate in accordance with the plate layout designed. The plate was covered with PCR adhesive sealer, spun down and placed in the thermal cycler. The PCR reaction was set following the manufacturer's protocol. The plate was left to thermal cycle overnight at a hold temperature of 4°C. The experiment was continued the following day. The success of the DNA library amplification was confirmed by running 5 µl of control sample and 5 µl of one patient sample for each column through a 4% agarose gel. The gel was made adding 8 mg of agarose powder (Melford Laboratories Ltd, Suffolk, UK) in 200 ml of 1X TBE buffer (Sigma-Aldrich Ltd, Dorset, UK) and heating the solution in a microwave at the full power temperature for 10 minutes. Afterwards, the solution was allowed to cool down for 10 minutes at room temperature, then 2 µl of Nancy-520 (Sigma-Aldrich Ltd) were added. The usage of Nancy-520 gel stain is

important to visualise the DNA bands under ultraviolet light. After mixing, the solution was poured into a gel tray with a 24 well comb in place and left to solidify for at least 30 minutes. Once solidified, the gel was placed in an electrophoresis unit covered with 1X TBE buffer. The samples were loaded into each well of the lane created by the comb. Along with the samples, 1 µl of 1kb plus DNA ladder (Invitrogen, Life Technologies) and 1 µl of 100 bp ladder, (Thermo Scientific, Waltham, MA USA) both diluted in 4 µl of water, were loaded. The gel was run for 20 minutes at 250 V. Using a Transilluminator device (Syngene, Cambridge, UK) the bands were evaluated. The expected PCR product size was approximately 350 bp. Using a bead-base system, the excess reagents from the PCR reaction are removed. Sixty µl of AMPure XD beads were added to each well-sample; each relevant well was then covered with MicroAmp 8 cap strip (Life Technologies) and the whole plate placed on a microplate shaker at 1200 rpm for 2 minutes. Afterwards, the plate was incubated at room temperature for 10 minutes and placed on a 96-wells magnetic stand, where the supernatant was removed. The beads/fragments were washed with 100 µl of 80% ethanol and eluted with 30 µl of elution buffer EBT1. On a clean PCR plate labelled “library normalisation plate” 20 µl of each eluted sample were then transferred. To ensure that an equal amount of each sample was represented in the final pooled library, samples were normalised using the Illumina bead-based system. Library normalisation beads LNB1 provided by the kit were re-suspended in the library normalisation additive LNA1 and 45 µl of the mixture added to each well-sample. After making sure that each well was tightly covered, the plate was placed on the microplate shaker for 30 minutes at 1200 rpm. Afterwards, the plate was placed on a 96-well magnetic rack and the supernatant removed. Each well-sample, containing DNA fragments bound to the beads, was washed with 45 µl of library normalisation wash buffer LNW1 twice and eluted with 30 µl of 0.1 N NaOH. Thirty µl of each sample were then transferred to a clean PCR plate and mixed with 30 µl of library normalisation storage buffer LNS2. In the last step, each sample was mixed in a single tube in order to perform a multiplex single sequencing run. Five µl of each sample were

mixed in a single tube and 10 µl of the mixture diluted in 590 µl of hybridisation buffer HT1. After vortexing and spinning down, the pooled library was denatured in a heat block at 96°C for 2 minutes and then placed on ice before loading the full content into the indicated reservoir of the Illumina reagent cartridge.

2.3.4.4 Cluster generation

After undergoing the preparation process, the DNA library was amplified. The amplification or so called cluster generation is an automated process performed on a flow cell, an optical transparent solid surface divided into 8 individual lanes in the Illumina HiSeq 2000 system and one lane in the Illumina MiSeq benchtop sequencer. For HiSeq 2000, the pooled DNA library was diluted and denatured with NaOH in accordance with the manufacturer's protocol. The TruSeq Cluster Generation kit (Illumina, Inc.) for the C-Bot cluster generation automated system was used. In addition, this sequencing system requires a positive control library for the estimation of the error rate during the clustering, sequencing and alignment processes. PhiX control library was added at low concentration of 1%. For MiSeq, the cluster generation is performed on-board; all the reagents required are included in the MiSeq reagent cartridge. On the plane surface of each lane, millions of oligonucleotides have been covalently anchored by the manufacturer. These oligonucleotides are complementary to the adapters ligated to both ends of the template, previously described in the library preparation section 2.3.4.2 and 2.3.4.3. After denaturation, strand templates are immobilised on the surface by adapter:oligonucleotide hybridisation and clonally amplified by an isothermal "bridge" amplification method. In this method, the annealing, the extension and the denaturation steps take place on the "arching" strand. From each fragment up to 1,000 copies are produced forming a unique cluster. On average between 500 and 800 k/mm^2 of cluster density was obtained.

2.3.4.5 Sequencing by synthesis

On the Illumina platform each DNA fragment, among the millions of clusters generated on the flow cell, is linearized and then sequenced simultaneously. In each cycle, four fluorescent-labelled reversible terminator nucleotides compete to bind the template. During the first part of the sequencing reaction the complementary 3'-blockage reversible nucleotide is added to the template by DNA polymerase; because of the obstruction of the elongation site, the sequencing reaction is interrupted. Thereafter, the fluorescent dyes are excited by a laser emitting a unique colour corresponding to the base incorporated, which is detected by a high sensitivity camera. The flow cell is divided into tiles and in each lane there are 48 tiles for the HiSeq System and 28 for the MiSeq. The detection occurs in each tile where the image is acquired four times per cycle, one image for each base. Then, the fluorescent label and the blocking group are cleaved and washed away allowing the second round of sequencing. This process is then repeated depending on the number of cycles in the run. To increase the accuracy and to facilitate sequence assembly a paired-end approach for Illumina technology was chosen (Roach *et al.*, 1995). This method generates sequences starting from both ends of the fragments during the same sequencing run. After completing the first round of sequencing the templates undergo a second strand regeneration. The complementary strand is denatured and stripped out; this is followed by primer annealing and sequencing. A second read starting from the opposite end of the fragments is then enabled (Holt & Jones, 2008). This approach is named paired-end module, and two reads, forward and reverse, were generated. The sequencing was performed using a multiplexing assay, allowing the running of multiple samples in a single lane. For HiSeq 2000, all the reagents, including reversible terminator nucleotides mixed with DNA polymerase and cleavage reaction mix, were prepared before the sequencing and loaded on. The flow cell was stationed on the sequencer tray and then the reaction was started. For the paired end module the run was set up with 100 cycles for read 1 and other 100 cycles for read 2. The indexes or barcodes were sequenced after

read 1using 6 cycles. The sequencing was performed at the Genomics Facility, Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and King's College London. For the MiSeq, all the reagents were already included and prepared in the MiSeq reagent cartridge. The flow cell was cleaned and loaded on the flow cell tray. The paired end module was set up with 150 cycles for read 1 and 150 cycles for read 2. The barcodes were sequenced after read 1 using 8 cycles. In the TSCA library preparation, two indexed or barcodes were used (Figure 2.3.4). The second barcode was sequenced immediately after the second strand was generated. The MiSeq machine is owned by our department and the sequencing was performed in the Liver Molecular Genetics' laboratory.

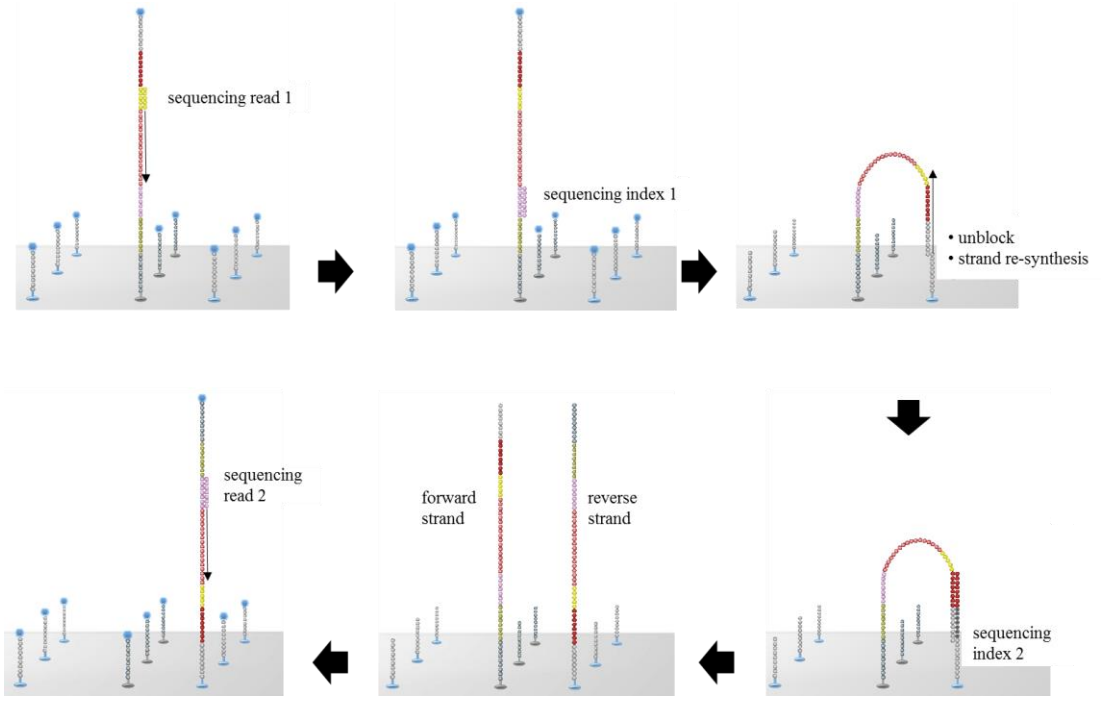

Figure 2.3.4 Paired-end sequencing by synthesis using Illumina platform

Paired-end sequencing was adopted as protocol to sequence DNA fragments from both ends. The prepared (sections 2.3.4.2 and 2.3.4.3) and bridge amplified (section 2.3.4.4) reads are firstly linearized and the 3'end blocked. Then each strand is submitted to the first round of sequencing, as described in section 2.3.4.5. After the completion of the first read, the read is washed away and the index 1 is sequenced. Subsequently, the 3'end block is removed, the read folds to bind the complementary oligo on the flow cell and the index 2 is sequenced. Then, a DNA polymerase re-synthesis the fragment. The double strand is linearized and the 3'end blocked one more time. After removing the original strand, the reverse strand.is sequenced. The image was adapted from Illumina.

2.3.5 **Sequence alignment, variant calling and annotation**

NGS sequencers typically create short sequences of less than 200 bp, defined as reads. The reads for each patient were firstly separated through the specific index/barcode sequence with the process of de-multiplexing; secondly, each of them was converted from the primary sequencing output BCL into a FASTQ file format. This initial process was undertaken using CASAVA (Consensus Assessment of Sequence and Variation). CASAVA is part of the Illumina's sequencing analysis software pre-installed in every Illumina machine. For each sample, two FASTQ files were generated corresponding to subsequent rounds of sequence in accordance to the paired-end module selected (section 2.3.4.5). So, after the sequencing, the FASTQ data were generated and exported for further processes. FASTQ file is a text-based format that includes the nucleotide sequence and the Phred quality score for each base; the latter represents the probability that the nucleotide base incorporated is incorrect. FastQC is an open source software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and was used to check the quality of each read before undergoing the sequencing analysis. Three main steps were included in the NGS data analysis: 1) sequence alignment or mapping: all reads for each sample were compared to the GRCh37/hg19 human reference genome in order to identify region of similarities; 2) variant calling: the genetic variations to the reference were detected; 3) variant annotation: the variations identified were then linked to biological information such as gene name, intronic, exonic or amino acid position, variant impact prediction and minor allele frequency (MAF). Several different software have been built to achieve these tasks, both commercial such as NextGENe (Softgenetics, State College, PA, USA) and CLCbio (Qiagen) described in sections 2.3.5.1 and 2.3.5.3 respectively, or open source Linux-based pipeline described in section 2.3.5.2.

2.3.5.1 NextGENe Software analysis

The targeted resequencing data from the initial cohort of 18 patients were analysed using NextGENe software. This Windows-based platform provides the complete bioinformatics tools designed to examine NGS data from targeted resequencing, RNA-Seq and many other NGS applications. Before starting the analysis, the input files of each patient were converted to FASTA format using the NextGENe Format Conversion tool, where reads with a median quality score of less than 15 and a size length of less than 25 were filtered out. Subsequently, the reads for each sample were aligned to the human reference genome using a mapping algorithm preloaded in the software. NextGENe employs a modified Burrows-Wheeler transformation alignment algorithm (Burrows & Wheeler, 1994), which identifies the best location of each read through a multi-step process. In the initial step, reads with a perfect similarity are matched, followed by those with a defined number of mismatches. For the remaining reads, where no match is found, a small part of the full length of the sequence, called seed, is used to identify the best position within the genome and then extend the alignment for the whole read. So, for each sample a new project was started. Using the project wizard, the alignment settings were defined in the alignment setting page as following: a) the entire reads with one mismatched base could still be aligned to the reference; b) the reads were allowed to map to multiple locations; 3) for the remaining reads with no match, seeds with a length of 37 bases where used for subsequent alignment; 4) overall, 85% of the entire read were matched exactly with the reference. Parameters for the variant calling were specified in the same panel. The variations were detected only if they met the following values: a) the mutation frequency of a variation at a given position occurred with a percentage greater than 20; b) reads with single nucleotide polymorphism (SNP) allele at a given position were greater than 5; c) the total number of reads at a given position had a coverage greater than 5. The variants were subsequently annotated using a database pre-installed in NextGENe software

where information, such as gene name, mRNA and coding DNA sequence (CDS) were retrieved. The single nucleotide polymorphism database (dbSNP) 137 dataset (ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b141_GRCh37p13/) was also preloaded in the "DataBase" tab and used to link additional information to the variants. The data processing was applied to each patient independently. At the end of the full analysis, a mutation report was generated and the variations underwent the subsequent filtering process described in section 2.3.6.

2.3.5.2 NGS Pipeline

NGS Pipeline is a Linux-based workflow, composed of a set of open-source tools developed for the analysis of next-generation sequencing data. The NGS Pipeline used in this study was optimised by Dr Michael Simpson in collaboration with Genomics Facility, Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and King's College London. Data from the TRS-21 of the cohort of 18 samples and from the following WES of 7 individuals were analysed using the following workflow (Figure 2.3.5). The sequences of each patient were mapped against the human reference genome using Novoalign (Novocraft Technologies, Selangor, Malaysia) as alignment tool. The format of each output file was then converted from a human-readable SAM format in a compressed and binary equivalent BAM format using SAMtools (Li *et al.*, 2009). This conversion is an important step to have more suitable and efficient input files for the subsequent tools. During the alignment, there is a high probability that reads containing insertions and deletions (indel) could not be matched perfectly, therefore, creating alignment artefacts and false positive SNP. Using IndelRealigner from the Genome Analysis Toolkit (GATK) package (DePristo *et al.*, 2011; McKenna *et al.*, 2010), reads with indel were correctly relocated. Afterwards, using another function of SAMtools, PCR duplicates were removed. SNP and indel were then identified using GATK UnifiedGenotyper and GATK IndelGenotyper from the GATK package and then filtered for quality and coverage. Variants passed the filter when

they had a quality score greater than 20 and a minimum coverage of 4 reads. The changes identified were finally annotated with known genetic information; this process was undertaken using ANNOVAR (Wang *et al.*, 2010). In addition, they were also annotated with data related to their minor allele frequency (MAF) in the general population, in accordance with several different databases, such as dbSNP135, 1000 Genome Project (http://www.1000genomes.org), and NHLBI GO Exome Sequencing Project (ESP) (http://evs.gs.washington.edu/EVS/), and approximately 1000 control exomes processed through the same bioinformatics pipeline. A score based on the effect of the change on the protein function was also added using SIFT (http://sift.bii.a-star.edu.sg/). The variation table generated from ANNOVAR was then processed through the filtering workflow described in section 2.3.6. As an additional tool, the BAM file generated from each sample underwent the ExomeDepth package in order to identified copy number variations (CNV) (Plagnol *et al.*, 2012). This tool utilises a read depth approach in which the numbers of reads are compared to the number of reads expected for every specific area of the exome. The number of reads expected were defined using a reference dataset, generated by the combination of all exome data belonging to the same sequencing run. CNVs are defined a discrepancy between the reads observed and the reads expected, therefore giving a read count ratio value. A ratio equal to 1 indicates no difference of coverage between the two dataset. Heterozygous deletions were identified with a ratio value between 0.2 and 0.8, whilst heterozygous duplications ranged between 1.2 and 1.7. CNVs with a ratio less than 0.2 and greater than 1.7 were considered homozygous deletions and duplications. In addition, the CNVs identified were annotated with a set of common CNVs and flagged as known in the output file (Conrad *et al.*, 2010). The likelihood for each CNV was defined by the Bayes factor (BF). The statistical confidence of a real CNV call was indicated with a higher BF value. CNVs were also flagged when occurring on segmental duplications. Integrative Genomics Viewer (Thorvaldsdottir *et al.*, 2013) was used as a NGS data visualisation tool. The data for each sample were processed independently with the described method.

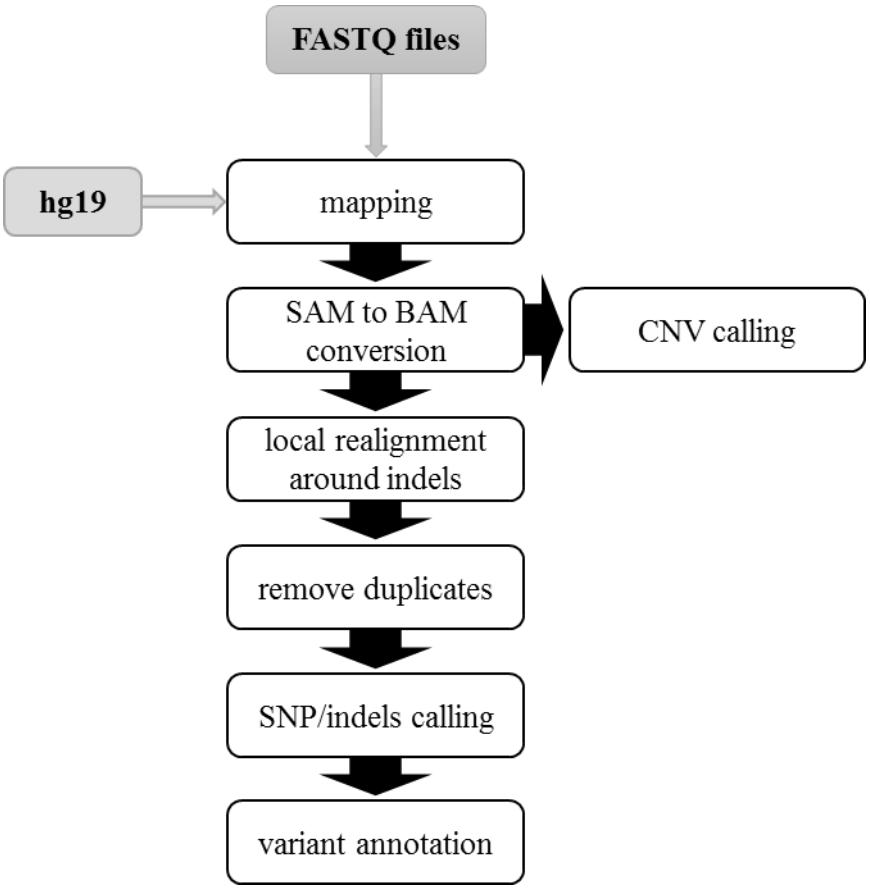

Figure 2.3.5 NGS Linux-based Pipeline flow chart

The FASTQ files of each sample first underwent a process of alignment against the reference genome hg19. Then, though several steps of data processing, the single nucleotide polymorphism (SNP), insertions and deletions (indels) and copy number variations (CNV) were identified. Different open source software were adopted at each stage as described in section 2.3.5.2.

2.3.5.3 CLCbio analysis

From the expanded cohort of 70 patients, five patients were excluded because of a failure of sequencing. Data from 65 individuals were analysed using CLCbio Genomic workbench. This bioinformatics platform is composed of numerous tools applied for the analysis of different next-generation sequencing approaches, such as targeted resequencing, RNA-Seq, Chip-Seq, etc. Through a series of connected tools, this software allows the creation and the installation of a personal workflow. The workflow for the analysis of TRS-7 data is shown in Figure 2.3.6. After importing FASTQ files of each sample, the sequence alignment was undertaken by the "map reads to reference" alignment tool. The configuration was used as default, varying only the similarity fraction to 0.9 that represents the 90% of identity between the read and the references. As mentioned in the NGS Pipeline, indels can creates misalignment. In CLCbio, the "local realignment" tool was used to resolve this issue. After the alignment, variants were called using the "probabilistic variant detection" tool. At a given position, the minimum coverage of 20 and variant probability of 80 were selected, which represents the probability of the specific variant of being different to the reference, as a percentage. Afterwards, the changes identified were filtered according to the quality and the frequency of occurrence using the "filter marginal variant call" tool. All changes having a second allele read frequency of less than 10% and a quality Phred score of less than 20 were removed. The variants identified were then annotated using "annotate from known variants" from dbSNP and the 1000 Genome Project, and "amino acid change tools" from Ensembl. The data of each sample were transferred to the CLCbio workflow independently. Subsequently, the output from this analysis with the complete table of variants identified was filtered in accordance to section 2.3.6. For CNVs detection, an in-house method was created, in which the read counts of each region were compared to the mean reads counts of the selected region within the all samples in the same sequencing run.

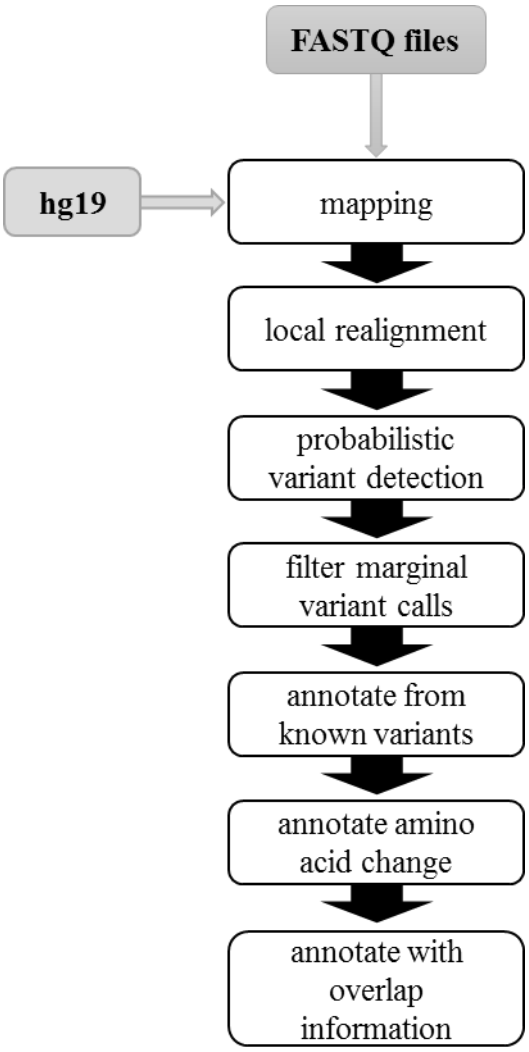

Figure 2.3.6 CLCbio workflow

The CLCbio workflow was created using the data processing tools pre-installed in the software. After aligning the FASTQ files against the reference genome hg19, the variations were identified and annotated in accordance to the method described in section 2.3.5.3 and here schematically represented. The data of each patient underwent the same processing analysis independently.

2.3.6 **Variant filtering strategy**

A final variant report was created as the output from every NGS data analysis software utilised. A filtering process was then employed to identify the potential disease-causing mutations and it was applied to the TRS-21, TRS-7 and WES data. The variant filtering flow chart is shown below (Figure 2.3.7). Using MAF information derived from several different databases, such as dbSNP, 1000 Genome Project and ESP, rare variations with minor allele frequency less than 2% in the population were selected. When data were available from internal databases, rare variants were kept in, if found a maximum of once as a homozygote or 20 times as a heterozygote. Afterwards, synonymous variations likely to be non-pathogenic were removed. To identify the possible impact of the change to the amino acid function and structure, each variant was analysed using several different web-based prediction tools, such as MutationTaster (http://www.mutationtaster.org/), PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/), SIFT and Ensembl's SNP Effect Predictor (http://www.ensembl.org/info/docs/tools/vep/index.html). Variations that were flagged as damaging were selected. Because of the high probability of cholestatic liver disease having an autosomal recessive inheritance, homozygous and compound heterozygous variations were selected in the first instance, and underwent biological interpretation. For WES data, only homozygous variations were initially considered, since patients were exclusively selected from consanguineous families, where the probability of inheriting the same mutated allele from both parents is substantially increased.

For copy number variations (CNV), the ExomeDepth package provided an annotated output file for each sample, which then underwent the following filtering process. CNVs affecting the sex chromosomes were firstly filtered out, together with those common between populations (Conrad *et al.*, 2010). Then, CNVs

located on segmental duplications were not considered due to the high chance of being false positive. Because the investigation of new disease-causing mutations in cholestasis was mainly focused on variations leading to loss of function, homozygous CNV deletions and duplication were selected using a ratio value less than 0.2 and greater than 1.7 (section 2.3.5.2). The CNVs which passed the previous data filtering were then sorted by the Bayes factor (BF) value from the highest to lowest value; CNVs with BF lower than 20 were excluded as this suggests as lack of statistical evidence (Johnson, 2013)

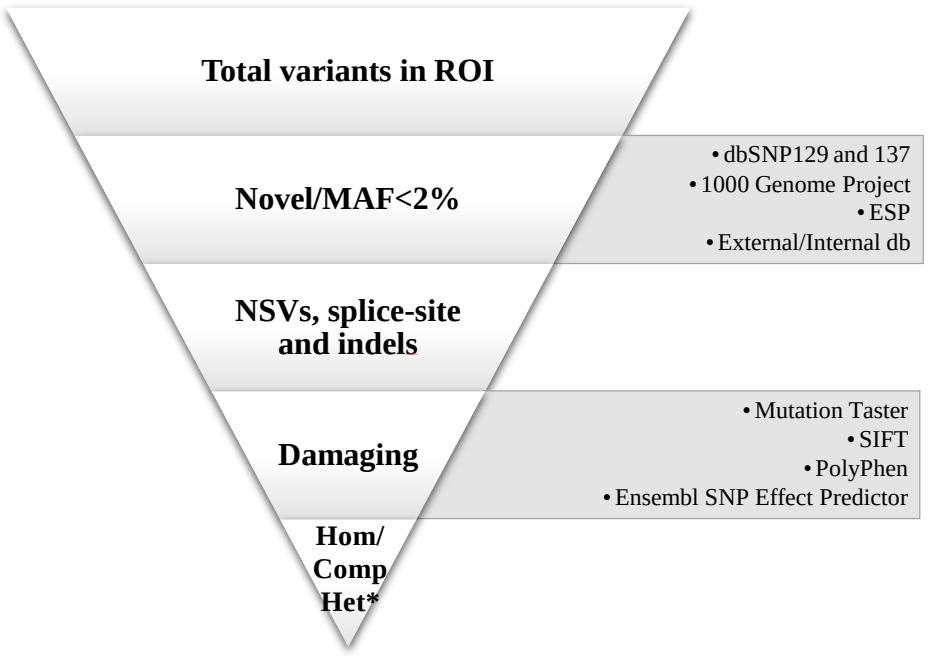

Figure 2.3.7 Variant filtering chart

Schematic representation of the filtering process adopted for the identification of the disease-causing mutations in both targeted resequencing (TRS) and whole-exome sequencing (WES) data.

* Compound heterozygous were considered in the TRS data. In the WES data, due to the selection of patients belonging to consanguineous families, only homozygous variations were considered in the initial approach.

2.3.7 **Sanger sequencing**

Developed in 1977 (Sanger *et al.*, 1977), Sanger sequencing was the most commonly used sequencing method until the advent of NGS. Through cyclic reactions of denaturation, annealing and extension, where a mixture of nucleotides (dNTPs) and terminator nucleotides (ddNTPs) are used, this technique is able to analyse targeted previously amplified DNA sequences with read lengths up to 1000 bp. Sanger sequencing was performed to sequence regions missed from the NGS investigation. Especially using TSCA library preparation, several areas were identified that lacked aligned reads or had low coverage; this issue was mainly due either to the presence of polymorphisms in the oligo-binding site that compromised the efficiency of the amplification reaction, or due to the high GC content. Additionally, Sanger sequencing was adopted for the validation of identified mutations and to analyse the effect of splice-site mutations and frameshift deletions on mRNA splicing. Initially, selected regions were amplified by polymerase chain reaction (PCR). Forward and reverse primers were manually designed for the flanking regions and are listed in Appendix Table I.1, Appendix Table I.2. Primers were manufactured by Integrated DNA Technologies (Leuven, Belgium). PCR amplification was optimised for each specific primer set and then performed using GC-rich PCR system (Roche Applied Science, West Sussex, UK) in accordance to the PCR protocol (Table 2.3.2).

	Volume (µl)	Final concentration
10X PCR buffer with 20 mM $MgCl_2$	2	1X
5X GC rich buffer	4	1X
dNTPs mix (10mM)	0.4	200 µM
Taq polymerase (5U/µl)	0.25	1.25 U
Forward primer (10 µM)	1	500 nM
Reverse primer (10 µM)	1	500 nM
Sterile water	9.35	

Table 2.3.2 Protocol used for an individual PCR amplification

Two µl of genomic DNA at approximately 50 ng/µl were added to the mix reaching a final volume of 20 µl. The PCR reaction was as follows:

Step	Temperature	Time
1	96°C	8 minutes
2	96°C	1 minute
3	50-60°C	20 seconds
4	72°C	30 seconds
5	*Repeat from Step 2 to Step 4 for 40 times*	
6	72°C	7 minutes
Hold	4°C	∞

Table 2.3.3 Programme used for the PCR amplification

The optimal temperature of each primer set in Step 3 is shows in Appendix Table I.1 and Appendix Table I.2. Successful PCR amplification was evaluated by electrophoresing 5 µl of each sample mixed with 5 µl of loading buffer through a 2% agarose gel, as described in section 2.3.4.3, using in this circumstance the lower concentration of agarose. Four grams of agarose were added to 200 ml of 1X TBE buffer. Later, PCR products were purified of excess dNTPs, primers, salts and all reagents that can interfere with the sequencing reaction. Thirty µl of AMPure XD

were used for each sample following the method described in section 2.3.4.2. The purified PCR products were washed with 70% ethanol and eluted with 40 µl of high-performance liquid chromatography (HPLC) water. For each sequencing reaction, 2 µl of product were added to 8 µl of master mix, which included the following reagents:

	Volume (µl)	**Final concentration**
5X BigDye Terminator v3.1 sequencing buffer	1.5	1.5X
BigDye® Terminator v3.1 (2.5X)	1	0.25X
Forward or reverse primer (10 µM)	1	1 µM
Sterile water	4.5	

Table 2.3.4 Protocol used for the preparation of one sequencing reaction

Sequencing buffer and dye were provided by Applied Biosystems, Life Technologies. The reaction was set as follows:

Step	**Temperature**	**Time**
1	96°C	5 minutes
2	96°C	1 minute
3	54°C	20 seconds
4	60°C	30 seconds
5	*Repeat from Step 2 to Step 4 for 40 times*	
6	60°C	5 minutes
Hold	4°C	∞

Table 2.3.5 PCR programme used for the sequencing reaction

The sequencing products from unincorporated primers or dye terminators were purified by the Sephadex spin column (Millipore, Watford, Hertfordshire, UK). Before purifying the sequencing products, the Sephadex plates were washed with

150 µl of HPLC water and centrifugation at 1200 rpm for 5 minutes. This action was repeated twice. Each sample was added of 10 µl of sterile water and then carefully loaded into the middle of each filtered column. The samples were recovered by a new centrifugation at 1200 rpm for 5 minutes in a new 96-well plate. Sequencing was undertaken using an automated ABI Prism 3130xl Genetic Analyzer (Applied Biosystems, Life Technologies). After separation by length size in a polyacrylamide gel capillary, each ddNTP fluorescent dye was detected using a charge couple device (CCD) camera and analysed by ABI Prism Sequencing Analyzer software. The data were exported and aligned to a reference gene using Sequencher 4.8 (Gene Codes Corporation, Ann Arbor, MI, USA) and CLC Main Workbench (Qiagen). Each electropherogram was also viewed using Chromas Lite 2.1.1 (Technelysium, South Brisbane, Australia).

2.3.8 RNA isolation from liver tissue

Frozen liver tissues from 5 patients with severe cholestatic liver disorder and 5 healthy liver donors were used. The specimens were products of hepatectomy and stored a -80°C for further clinical or research purposes. Total RNA was isolated using TRIzol® reagents (Ambion, Life Technologies). One ml of TRIzol reagent per 100 mg of tissue was added and the tissue homogenised using plastic pestles. After incubating the solution for 5 minutes at room temperature, 200 µl of chloroform were added and the tubes vigorously inverted for 15 times. To allow a better separation, the solution was incubated at room temperature for 3 minutes and then centrifuged at 12,000 rcf (relative centrifugal force) for 10 minutes. Following centrifugation, two layers were created with RNA present in the upper aqueous phase. This phase was carefully removed and transferred into a new 1.5 ml tube, where 500 µl of isopropyl alcohol were added for the RNA precipitation step. The solution was then incubated at room temperature for 10 minutes and centrifuged at 12,000 rcf for 10 minutes. Afterwards, the RNA formed a white pellet on the

bottom of the tube, often visible to the naked eye. The RNA pellet was then washed with 1.2 ml of 75% ethanol and then dissolved in 100 µl of RNase-free water (Ambion, Life Technologies). To allow a better RNA re-suspension, the solution was mixed by pipetting several times up and down and then incubated for 10 minutes at 55°C for 8 minutes. The RNA solution was temporarily stored on ice for further quantity and quality controls. The quantification was undertaken using the Qubit 2.0 Fluorometer as described in section 2.3.3, using in this case the Qubit RNA assay kit. The RNA quality control was performed using Agilent 2200 TapeStation system, as described in section 2.3.4.2, with Agilent RNA ScreenTape and RNA ScreenTape sample buffer and ladder. The RNA integrity was evaluated by the RNA integrity number (RIN) (Figure 2.3.8).

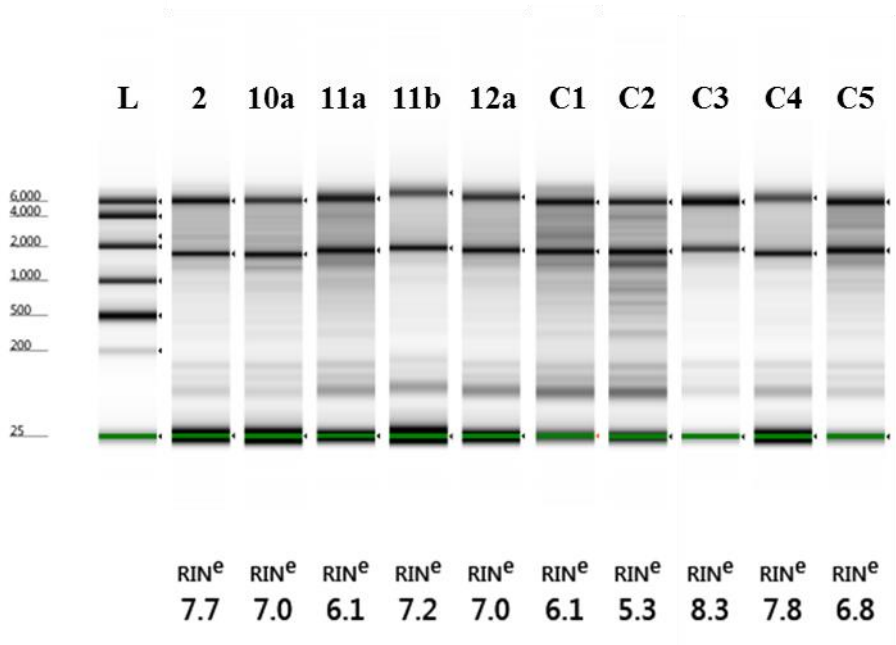

Figure 2.3.8 Gel image of the RNA analysis

The gel image shows the electrophoretic separation of the RNA isolated from the frozen liver tissues of five patients (2,10a,11a,11b,12a) and five healthy liver donors (C1-5) using Agilent 2000 TapeStation system (method in section 2.3.4.2). A ladder (L) was added to the gel. The higher band represents the ribosomal subunit 28S, followed by the ribosomal subunit 18S and the lower marker. The RNA quality was established assigning a RIN number, which estimates the integrity of the RNA based on the peak ratio between the 28S and 18S and the presence or absence of degraded material.

2.3.9 Reverse transcription polymerase chain reaction (RT-PCR)

First-strand cDNA synthesis was performed to investigate the effect of splice-site mutations and frameshift deletions on mRNA splicing. Before starting, all samples were normalised to an RNA concentration of 1 µg/µl. High Capacity RNA-to-cDNA kit (Applied Biosystems, Life Technologies) was adopted as reverse transcription kit. The reaction was set in accordance to the manufacturer's protocol.

2.3.10 Long-range polymerase chain reaction

DNA fragments from 5kb up to 20 kb were amplified using a long-range PCR. The Expand Long Template dNTPack kit (Roche) was adopted, which combines the most common PCR enzyme, the Taq DNA polymerase enzyme, and a thermostable DNA polymerase enzyme with proofreading activity for obtaining a greater efficiency and accurate amplification. Forward and reverse primers were designed for selected regions and manufactured by Integrated DNA Technologies (Leuven, Belgium). Each pair, summarised in Appendix Table I.3 was tested using a temperature gradient from 52 up to 58 °C. An optimal concentration of 4% of dimethyl sulfoxide (DMSO) was also added to the PCR mix. The inclusion of this PCR additive is recommended for the denaturation of regions of high GC contents. Individual PCR reactions were prepared in accordance to the protocol detailed in Table 2.3.6.

	Volume (µl)	Final concentration
5X Expand Long Range buffer with $MgCl_2$	10	1X
PCR nucleotide mix (10mM)	2.5	500 µM
Forward primer (10 µM)	2	400 nM
Reverse primer (10 µM)	2	400 nM
DMSO	2	4%
Expand Long Range enzyme mix (5 U/µl)	0.7	3.5 U
Sterile water	28.8	

Table 2.3.6 Long range PCR mix protocol

Two µl of genomic DNA at approximately 50 ng/µl were added to the reaction mix to a final volume of 50 µl. The reaction was then conducted using the following PCR programme in Table 2.3.7.

Step	Temperature	Time	Cycle
Denaturation	92°C	2 minutes	1X
Denaturation Annealing Elongation	92°C 52 to 58 °C* 68°C	10 seconds 15 seconds 5 minutes**	10X
Denaturation Annealing Elongation	92°C 52 to 58 °C* 68°C	10 seconds 15 seconds 5 minutesΔ	20X
Final Elongation	60°C	5 minutes	1X
Hold	8°C	∞	

Table 2.3.7 Long range PCR thermal cycler programme

* A temperature gradient of 52 -54-56 and 58 °C was used for each pair of primers
** The long-range PCR system recommends an elongation time of 1 minute per 1 kb. For the extension of a DNA fragment of approximately 5 kb an extension time of 5 minutes was used.
Δ The Auto Delta option was set up to increase of 20 seconds the elongation step for each successive cycle

The PCR amplification was validated by electrophoresis as described in section 2.3.4.3. Because the expected size was approximately 5 kb, a 1% of agarose gel was prepared for the evaluation of the large DNA fragments. Two grams of agarose were added to 200 ml of 1X TBE buffer. If specific bands were visible, PCR products were purified from excess of dNTPs, primers and all unincorporated reagents, and prepared for the sequencing reaction (section 2.3.7).

2.4 Results

2.4.1 NGS run metrics for TRS-21

For the initial part of this study, specific regions of interest were sequenced using the Agilent SureSelect method for Illumina platform, as described in section 2.3.4.2. The method allowed sequencing of up to 12 samples simultaneously in one single lane of the flow cell. Each patient was distinguished amongst the others by individual barcodes ligated to the reads during the preparation of the DNA library; one specific barcode was assigned to each patient. The 18 patients selected for this project were divided into two batches. The DNA library of the first batch of 12 samples was prepared in accordance to section 2.3.4.2 and sequenced in one lane of the flow cell using Illumina HiSeq 2000 technology (section 2.3.4.5). The quality metrics were acquired by the Real Time Analysis (RTA) software pre-installed in the Illumina HiSeq system, and by the Illumina Sequencing Analysis Viewer v.1.8 (SAV), installed in a personal desktop computer. This sequencing run generated a cluster density of 584 ± 88 k/mm^2. The total number of reads obtained were 161.59 million. However, the system is set up to remove the clusters having a poor quality in fluorescent intensity, which could result from a low signal or high noise ratio due, for example, by two overlapping clusters. Thus, the percentage of clusters passing this quality control filter was 93.97 ± 1.31, with a total number of reads passing filter of 151.57 million. In addition, the reads underwent a further quality control based on the Phred quality (Q) score. A threshold of Q score equal or greater of 30, representing the chance to have an incorrect base of 1 in 1000 or less, was utilised. In the first TRS-21 sequencing run, 91.4% of the bases have a Q score equal or greater than 30, yielding 28.7 Gb of sequence (Figure 2.4.1).

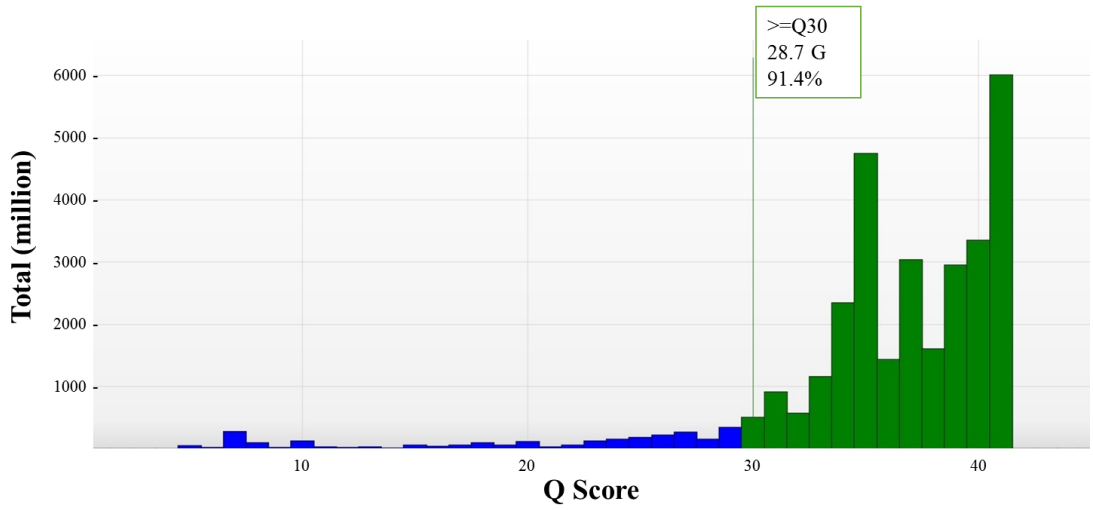

Figure 2.4.1 Quality distribution of the first TRS-21 sequencing run

Display of the distribution of the number of reads in million by their quality score (Q score). A Q score of 30 was used as a threshold. Represented in blue are the reads with a Q score less than 30, and in green the reads with a Q score equal of higher than 30. In total, 91.4% of the reads had a Q score of 30 or higher, with a total number of bases sequenced of 28.7 Gb.

From the 18 samples, the second batch of 6 was multiplexed and sequenced in a second lane. Similar to the previous run, the cluster density generated was 581 ± 57 k/mm^2, producing a total of 160.66 million of reads. However, 87 ± 15.88% of clusters passed the quality control based on the intensity of the base-fluorescent signal, with a total number of reads passing filter of 141.64 million. Considering the Phred score, 89.3% of bases had a Q score of 30 or higher with a total yield of 26.2 Gb, as shown in Figure 2.4.2.

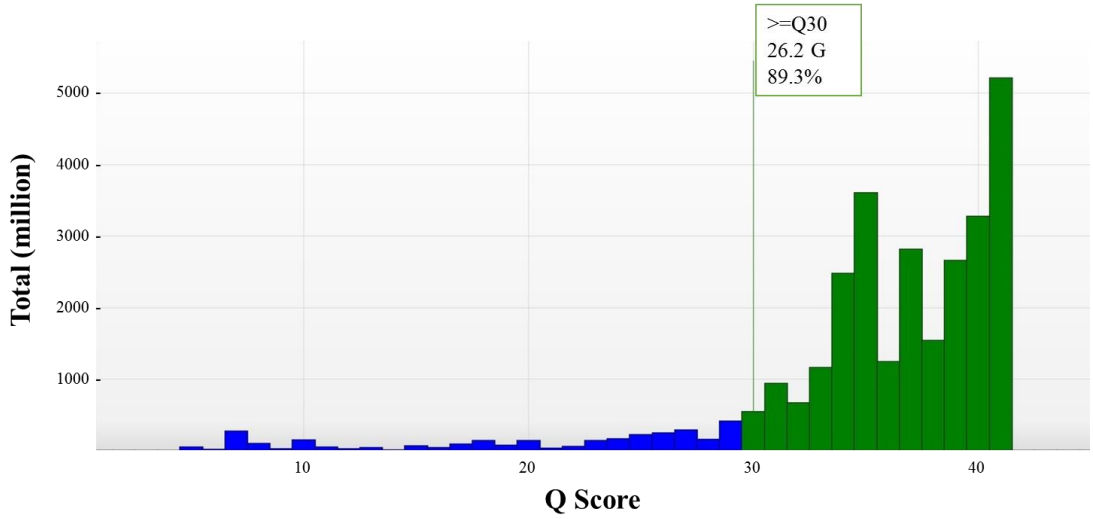

Figure 2.4.2 Quality distribution of the second TRS-21 sequencing run

Distribution of the number of reads in million by their quality score (Q score). A score of 30 was used as a threshold. Represented in blue are the reads with a Q score less than 30, and in green the reads with a Q score equal of higher than 30. In total, the second TRS-21 sequencing run yields 26.2 Gb with 89.3% of the reads with a Q score of 30 or higher.

2.4.2 Targeted resequencing variant detection

2.4.2.1 Variant detection after filtering strategy

The data from TRS-21 were analysed initially by NextGENe software and subsequently by NGS Linux-based pipeline following the methods described in sections 2.3.5.1 and 2.3.5.2 respectively. For each patient a mutation report was created, which includes the single nucleotide variations (SNVs), indels and splice-site variants identified. To increase the likelihood of discovering disease-causing mutations, the mutation reports from the 18 TRS-21 patients were combined and the variants filtered through the variant filtering process described in section 2.3.6. As shown in Table 2.4.1, a total number of 582 variations were collected from the data analysed by NextGENe software. These resulted from the sequencing data of 21 genes of the 18 patients, with an average of approximately 30 variations per patient. Afterwards, novel and rare variations were selected; the latter were considered rare when having a minor allele frequency (MAF) less the 2% in the general population, according to the different public databases, such as dbSNP, 1000 Genome Project and ESP. As shown in Table 2.4.1, the majority of variations were commonly present in the general population and circa 10% were positively selected for further filtering steps. Synonymous variations were then discarded because of their usual benign effect on the translated protein, and SNVs, indels and variants affecting the canonical 5'-donor splice and 3'-acceptor splice sites were maintained. Subsequently, the changes were investigated for their impact on the protein structure and function. Using different damage prediction tools (section 2.3.6), only 11 variations within the total of 26 were likely to cause abnormalities at the protein level. The genetic inheritance of cholestatic liver diseases usually displays a Mendelian autosomal recessive pattern and therefore only homozygous or compound heterozygous variations were considered. An exception was however made for any variations affecting *JAG1* or *NOTCH2*, where mutations show

autosomal dominant inheritance and where only one mutated allele is sufficient to cause the disease. No pathological mutations in those genes were discovered. After following the process of variant filtering, five homozygous potential disease-causing mutations were identified, all affecting the same gene.

Total variants	582
Novel/MAF<0.02	42
SNVs, indels, splice-site variants	26
Damaging	11
Homo/compound Het	5

Table 2.4.1 Number of variants in TRS-21 sequencing data analysed by NextGENe software

The total number of variant results from the merged NextGENe sequencing data output of all 18 patients were analysed. The variants underwent a process of filtering as described in section 2.3.6. Five variants were satisfying the criteria and were described as potential disease-causing mutations.

In addition, a re-analysis was undertaken using a NGS Linux-base Pipeline (section 2.3.5.2). The same analysis process was applied; in fact the variants from all patients were merged and filtered in accordance to section 2.3.6. Comparing to the previous method, a similar total number of variants was detected (Table 2.4.2), and five homozygous potential disease-causing mutations were in the end identified.

Total variants	569
Novel/MAF<0.02	35
SNVs, indels, splice-site variants	23
Damaging	10
Homo/compound Het	5

Table 2.4.2 Number of variants in the TRS-21 data analysed by NGS Linux-base Pipeline

The total number of variants results from the NGS Pipeline sequencing data output of all 18 patients were analysed. The variants underwent a process of filtering as described in section 2.3.6. Five variants were satisfying the criteria and were described as potential disease-causing mutations.

2.4.2.2 Description and interpretation of the findings

Amongst the selected cohort of 18 patients, targeted resequencing discovered five individuals having five distinctive homozygous pathological mutations in the tight junction protein 2 (*TJP2)* (Table 2.4.3). A 4-bp deletion on exon 5 was identified in patient 2, causing a shift of the reading frame of the amino acid sequence. The reads covering that specific segment were 1873 (NextGENe software output) and 1185 (NGS Pipeline output), where the deletion was present in 99.89% and 100% respectively. These differences in coverage were due to the different thresholds of quality score adopted: reads were removed when having a Q score less than 15 by NextGENe and 20 by NGS Pipeline. Targeted resequencing was carried out in one affected individual per family (section 2.3.1). The selection between affected siblings was made on the basis of the quality and quantity of the genetic material available. Different 1-bp deletions were identified in one affected sibling of family 4 and in patient 9. The first was present in 97.92% of 867 reads in the analysis undertaken by NextGENe software and 99.47% of 659 reads in the NGS Pipeline analysis; whereas the second 1-bp deletion occurred in 99.79% of 1405 reads and 99.89% of 1039 reads respectively. Both mutations were located on exon 5 of *TJP2*. In the three cases having deletions in exon 5, in each case a novel amino acid

sequence was generated starting from the respective deleted base and in all ending with the same premature terminator codon (PTC). The fourth homozygous potential disease-causing mutation identified was discovered in one of the affected siblings of the family 10 and was another 1-bp deletion, this time in exon 9 of *TJP2*. As with the previous cases, the high read depth demonstrated by both analysis methods (99.53% of 2146 reads in NextGENe analysis method and 99.91% of 1233 reads in NGS Pipeline analysis methods) confirms the veracity of the finding. The frameshift affected the amino acid sequence leading to a downstream PTC. Differently, the fifth mutation identified in one affected sibling of the family 11 caused an alteration in the 3' splice acceptor site of intron 13 of *TJP2*. The mutated nucleotide sequence was detected in 99.81% of 1032 reads by NextGENe software and by the totality of the 697 reads by NGS Pipeline. The consequence of that change on the splicing mechanism is revealed with the sequencing analysis of the transcribed mRNA, described in the following section 2.4.3.

			NextGENe software		**NGS Pipeline**	
Case number	**Nucleotide change**	**Amino acid change**	**Coverage (reads)**	**Frequency (%)**	**Coverage (reads)**	**Frequency (%)**
2	c.766_769del	p.Ala256Thrfs*54	1873	99.89	1185	100
4a	c.885delC	p.Ser296Alafs*15	867	97.92	659	99.46
9	c.782delA	p.Tyr261Serfs*50	1405	99.79	1039	99.89
10a	c.1361delC	p.Ala454Glyfs*60	2146	99.53	1233	99.91
11a	c.1992-2A>G	p.?	1032	99.81	697	100

Table 2.4.3 Cases identified by targeted resequencing-21

Mutations are described using the reference transcript NM_004817.

2.4.3 Sanger sequencing validation and splicing consequences for TRS-21 findings

Sanger sequencing technology (section 2.3.7) was used to validate the mutations discovered. They were all confirmed to be real. In addition, the affected siblings (cases 4b, 10b, 11b) were genotyped and identified as having the same disease-causing mutations as their siblings. The summary of all the individuals with their respective genetic information is shown in Table 2.4.4. In family 11, the investigation of the consequence of the mutation on the RNA splicing mechanism, such as possible intron retention, exon skipping or an alternative acceptor site, was carried out by cDNA Sanger sequencing technology. cDNA was reverse transcribed from isolated RNA as described in sections 2.3.8 and 2.3.9. Specific primers were designed to exon 12 (forward) and exon 16 (reverse) in order to amplify and then sequence the cDNA region containing the mutation (Appendix Table I.4). The analysis revealed the occurrence of an alternative splicing mechanism due to the creation of an alternative acceptor site. The loss of the normal splice acceptor site on the intron 13 was replaced by the first two adjacent AG bases located at the 5'-end of the exon 14. Therefore, 2-bp were deleted from the mRNA and consequently led to an alteration of the reading frame of the coding sequence and the generation of a premature terminator codon (Figure 2.4.3). In addition, the investigation was carried out on the affected sibling (case 11b) where, not surprisingly, the same cDNA change was confirmed to be present (Figure 2.4.3a).

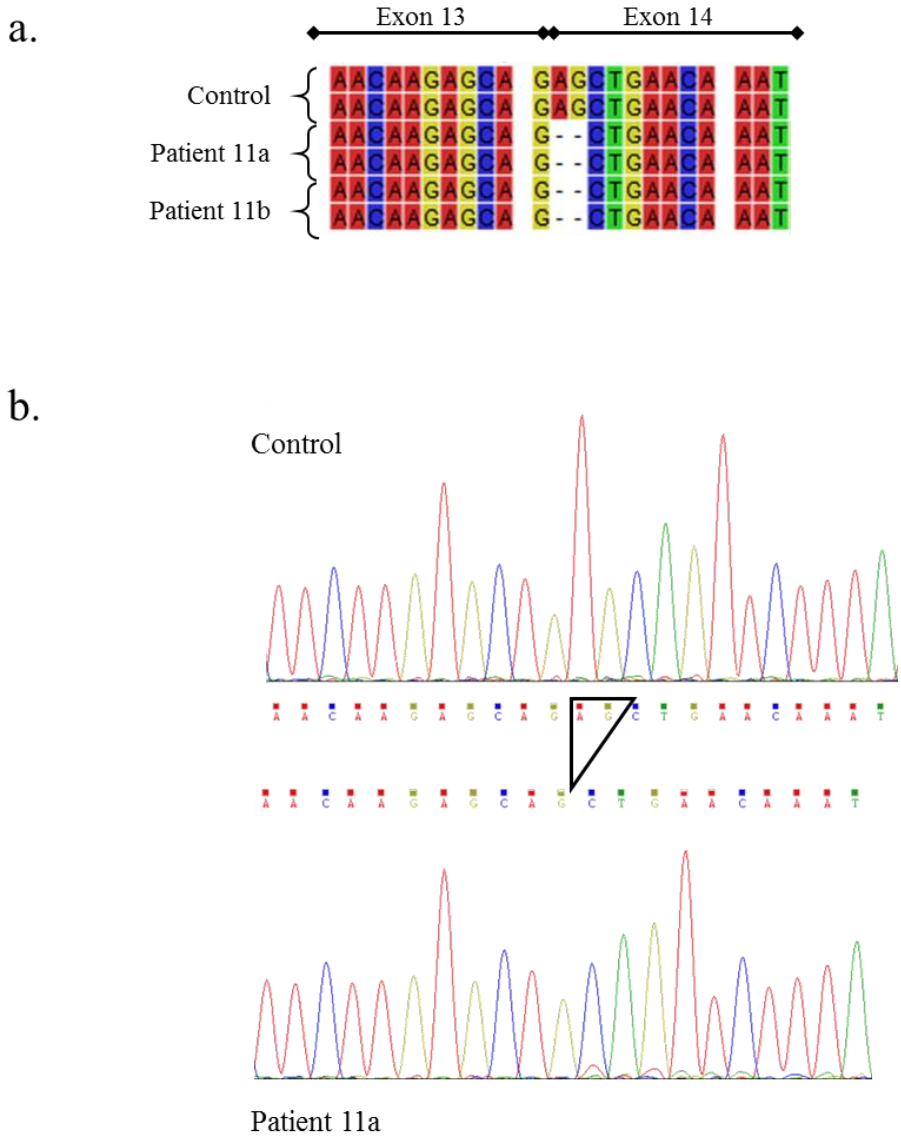

Figure 2.4.3 Consequence of splice site mutation on family 11

a) cDNA sequence alignment of a control sample (healthy liver donor) and the two affected siblings of family 11. The splice site mutation caused the generation of an alternative acceptor site using the first 2 (AG) bases of exon 14. b) Representation of the 2 bases deletion on the electropherograms of the control sample and the affected patient, in each case obtained by cDNA sequencing

Family number	Case number	Chromosome position (start)	Chromosome position (end)	Exon-intron number	Nucleotide change	Amino acid change	Mutation discovery
2	2	71836226	71836229	Exon 5	c.766_769del	p.Ala256Thrfs*54	TRS-21
4	4a	71836345	71836345	Exon 5	c.885delC	p.Ser296Alafs*15	TRS-21
	4b						SS
9	9	71836242	71836242	Exon 5	c.782delA	p.Tyr261Serfs*50	TRS-21
10	10a	71842938	71842938	Exon 9	c.1361delC	p.Ala454Glyfs*60	TRS-21
	10b						SS
11	11a	71851863	71851863	Intron 13	c.1992-2A>G	p.Arg664Serfs*2	TRS-21
	11b						SS

Table 2.4.4 Summary of the mutations discovered by TRS-21 in *TJP2*

All mutations discovered by TRS-21 in five families were genotyped by Sanger Sequencing (SS). Mutations are described using the reference transcript NM_004817.

2.4.4 **NGS run metrics for WES**

To obtain a predicted read depth of at least 20X per base, the samples of seven patients selected to undergo the WES analysis were separated into batches of four and three and sequenced into two different lanes of the flow cell. From the first run, a cluster density of 695 ± 99 k/mm^2 was generated with a total number of 192.25 million of reads. However, after applying the quality filter based on the intensity of the fluorescent signal, 91.54 ± 2.15% of the clusters passed with a total number of reads passing filter of 175.64 million. Considering the Phred score, 90% of the bases had a Q score of 30 or higher. The number of bases sequenced was 32.7 Gb in total for the first WES run (Figure 2.4.4).

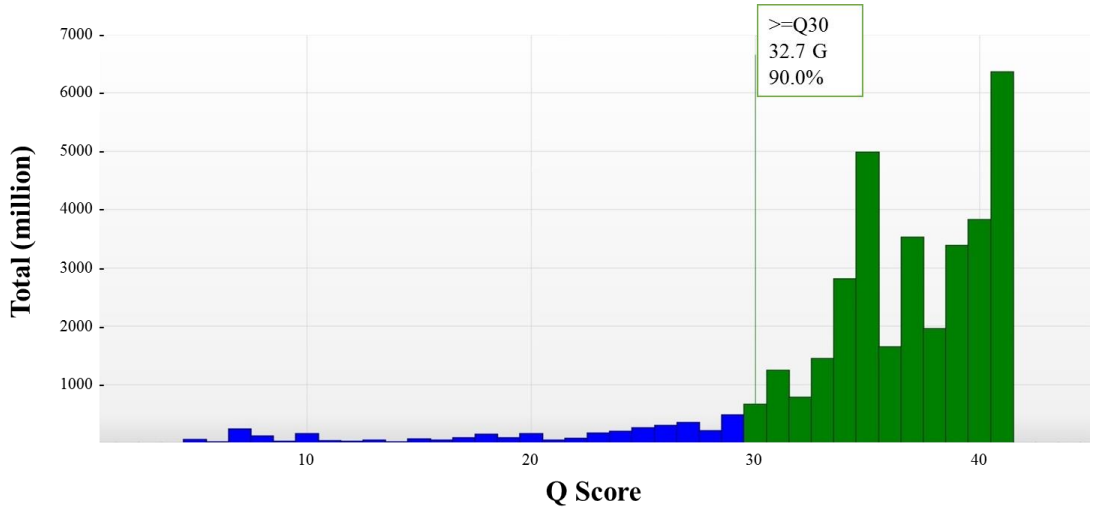

Figure 2.4.4 Quality distribution of the first WES sequencing run

Distribution of the number of reads in million by their quality score (Q score). The Q score threshold was 30. Reads with a Q score less than 30 are represented in blue, those with a Q score equal or higher than 30 in green. In total, the first WES sequencing run yielded 32.7 Gb with 90% of the reads with a Q score of 30 or higher.

The second batch of samples to undergo WES generated a cluster density of 668 ± 98 k/mm^2, with a total number of 184.54 million of reads. After passing the quality filter, the percentage of clusters was 91.97 ± 2.65 with a total number of reads passing filter of 169.3 million. However, within this group of reads 90.5% of bases had a Q score equal or greater than 30, yielding 31.7 Gb (Figure 2.4.5).

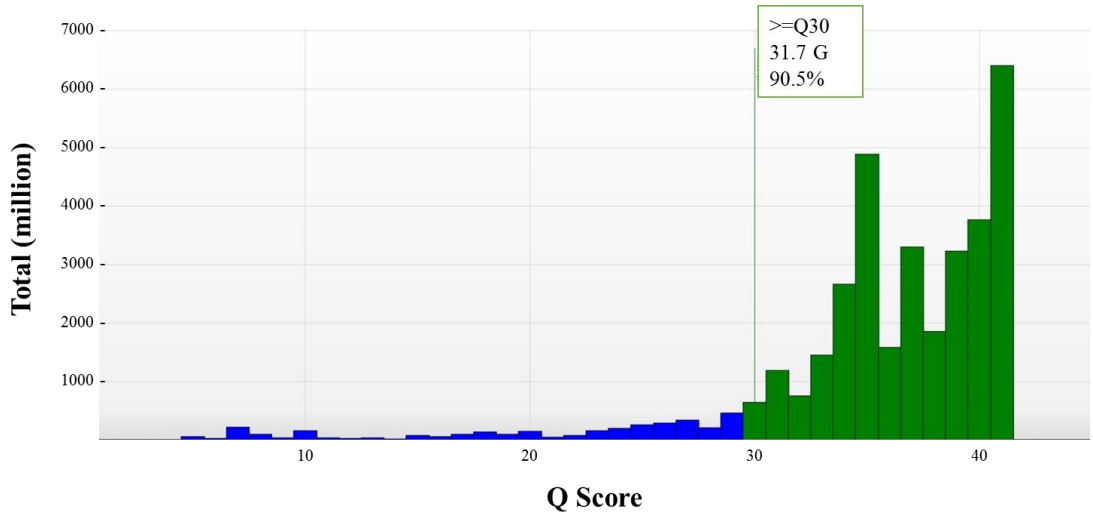

Figure 2.4.5 Quality distribution of the second WES run

Distribution of the number of reads in million by their quality score (Q score). The Q score threshold was 30. Reads with a Q score less than 30 are represented in blue, while the reads with a Q score equal of higher than 30 are in green. In total, the first WES sequencing run yielded 31.7 Gb with 90.5% of the reads with a Q score of 30 or higher.

2.4.5 **Whole-exome sequencing variant detection**

2.4.5.1 Variant detection after filtering strategy

WES was performed on 7 patients in whom no mutations were identified in the previous NGS analysis. Because the inheritance of recessive genetic conditions has a higher frequency in consanguineous marriages, patients belonging to consanguineous families were specifically selected (Table 2.3.1). WES data were then processed using the Linux-based NGS Pipeline as described in section 2.3.5.2. The output file, including SNVs, indels and splice-site variants for each patient, was then filtered using the workflow schematised in Figure 2.3.7. The exome represents 1% of the human genome, and includes approximately 20,000 coding genes. The analysis of the whole-exome data identified a mean of 24,704 variations per patient (Table 2.4.5). However, a small percentage of them, (6% on average), proved to be novel or rare with MAF<2% in the general population, according to several variation databases, including dbSNP, 1000 genome Project and ESP. Synonymous changes were not considered because of their neutral impact on the protein function and structure, and splice-site variations were selected only when affecting the canonical GT and AG donor and acceptor splice sites. As described in section 2.3.6, in this initial stage of analysis the attention was focused exclusively on homozygous variations. The variations then were submitted to the last phase of the filtering process, in which approximately 30% were predicted to cause a severe impairment of the protein function and structure. As shown in Table 2.4.5, in the 7 patients analysed, a variable number from 6 to 25 variations was identified as possible disease-causing mutations. Further investigations were undertaken considering their biological significance in the aetiology of cholestatic liver disorders.

	Patient 1	Patient 3	Patient 8	Patient 12a	Patient 13	Patient 17	Patient 18
A	24336	24977	23981	24832	25716	24344	24747
B	1537	1766	1325	1456	1681	1151	1259
C	947	1090	844	914	1026	737	811
D	60	62	62	36	24	35	22
E	22	14	25	14	7	18	6

Table 2.4.5 Numbers of variants identified in the 7 patients analysed by WES

For the identification of disease-causing mutations, WES data of each patient underwent a filtering process described in section 2.3.6. From the total number of variations identified (A), novel variations and those with MAF<2% in the general population (B) were selected. Then synonymous variations were filtered out (C) and only homozygous changes (D) were at this stage analysed due to the high probability to occur in consanguineous families. Afterwards, the variations causing damage to the function and structure of the protein were selected (E) for biological interpretation.

Copy number variations (CNV) are defined as the major class of genomic structural variations caused by an alteration in the number of copies of a variable DNA length sequence (Strachan & Read, 2011). The NGS Linux-based Pipeline has the limitation of failing to identify these genetic variants. For this reason, the WES data were submitted to the analysis of CNVs using the ExomeDepth package as described in section 2.3.5.2. Then, for each sample a filtering strategy was applied (section 2.3.6). Table 2.4.6 summarises the number of CNVs identified at each filtering stage. Approximately 100 variations were detected in all patients except for patient 18, where an over calling of CNVs occurred. False positives were identified and removed after applying the Bayes factor value greater than 20. The threshold of 20 was determined as the requirement to strongly support the evidence statistically (section 2.3.6). In patient 18, the majority of variations had a BF less than 20 with a mean value of 7.3, except one homozygous CNV deletion that had a BF of 27 and read ration of 0.159 (152 reads observed in 955 reads expected) (Appendix Table II.1). Regarding the other cases, only two structural variations were identified after using the filter strategy. Patient 12a and patient 17 had

homozygous deletions with a BF of 212 and 201 respectively. The read ratio in both cases was null, due to no reads being observed in the specific regions (Appendix Table II.1).

	Patient 1	**Patient 3**	**Patient 8**	**Patient 12a**	**Patient 13**	**Patient 17**	**Patient 18**
a	113	90	37	108	148	86	1446
b	112	78	24	70	64	77	1428
c	28	31	1	7	2	7	709
d	9	1	0	2	0	1	96
e	0	0	0	1	0	1	1

Table 2.4.6 Total number of CNVs identified in the 7 patient analysed by WES

Copy number variations (CNV) identified in each patient by ExomeDepth analysis. From the total number of CNVs (a), genetic variants affecting the sex-chromosome were filtered out (b) and subsequently those known to be common in the population and located on segmental duplications (c) were removed. Then, homozygous CNVs deletions and duplications were selected having a ratio of observed reads and expected reads <0.2 and > 1.7 respectively (d). Bayes factor value of 20 was used as the threshold to eliminate false positive calls (e).

The interpretation of three CNVs, together with the previous filtered variations discovered, was undertaken on the basis of their biological function and their relevance in the pathology of cholestatic liver disorders.

2.4.5.2 Description and interpretation of the findings

For the initial investigation, variants common amongst the patients or those affecting the same gene were examined. Variations belonging to each patient were conveniently barcoded and recorded. Then, all changes, including SNVs, indels, splice-site and CNVs, which had passed the whole filtering process, were merged in a single file and then sorted by gene name or chromosome position. No causative gene or variants shared between all patients were identified. However, one patient (patient 12a) with a large homozygous deletion in *TJP2* was discovered. The

deletion affected 11 exons, from exon 6 to exon 16 inclusive, so eliminating half of the entire gene. No reads were observed in that segment compared to the 1233 reads expected to be present (Figure 2.4.6). Very strong evidence of the occurrence of the deletion was also suggested by the high statistical Bayes factor value (equal 212) (Appendix Table II.1).

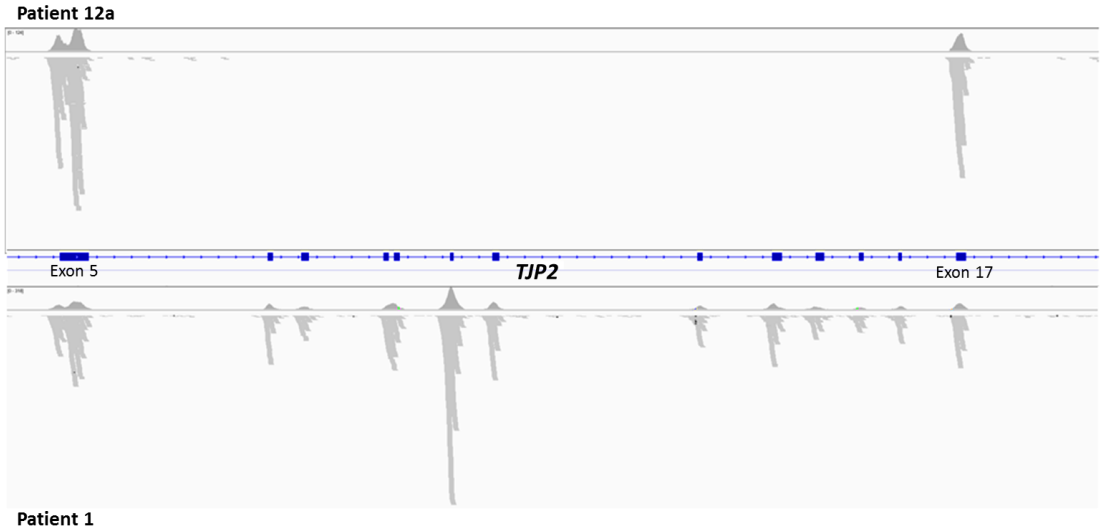

Figure 2.4.6 Visualisation of the large homozygous deletion in patient 12a

The visualisation of the alignment of each read to the human exome is shown by IGV tool. On the top panel, the patient 12a displays no reads aligned from exon 6 to exon 16 of *TJP2* compared to the patient 1 on the bottom panel, in whom no structural variations in that gene were identified.

Subsequently, the analysis of the WES filtered data from the other five individuals focused on the identification of disease-causing mutations on the basis of their biological significance, using gene databases such as GeneCards (http://www.genecards.org/) and OMIM (http://www.ncbi.nlm.nih.gov/omim). Clinical information was also re-evaluated. Within the WES filtered variation data, a novel missense mutation in patient 3 was discovered, which affects the gene α-methylacyl-CoA racemase (*AMACR)* (Appendix Table II.3). A transition of a thymine to a cytosine occurred in position 877 (NM_014324), predicted to cause

an amino acid substitution of cysteine in position 293 for arginine. The mutated allele showed a read depth of 77, while no reads for the wild-type allele were shown. The Q score was 222. In addition, the substitution was described as disease-causing by every prediction tool utilised. The gene encodes a racemase enzyme involved in the oxidation of methyl-branched fatty acids and in the biosynthesis of bile acids in the peroxisomes (Savolainen *et al.*, 2004). AMACR deficiency is an extremely rare genetic disorder with sensorimotor neuropathy in adults and with cholestatic liver disease in neonates (Setchell *et al.*, 2003). The latter is characterised by mild elevation of liver enzymatic activity, vitamin D malabsorption and usually with high concentrations of serum bile acids. In patient 3, this congenital abnormality of bile acid synthesis was confirmed biochemically. For the other individuals, the summary of the SNVs, indels and splice-site variations that passed the filters are shown in Appendix Table II.2 (patient 1), Appendix Table II.4 (patient 8), Appendix Table II.6 (patient 13), Appendix Table II.7 (patient 17) and Appendix Table II.8 (patient 18). From this first analysis, no obvious disease-causing mutations were identified. In this high stringency examination it looks likely that the filtering has removed the true genetic causes in the remaining patients. Hence, subsequent investigation will include, for example, examination of compound heterozygous changes. In addition, whole-exome sequencing will be performed on siblings affected by the same chronic cholestasis.

2.4.6 Sanger sequencing validation and splicing consequences for WES findings

As described in the previous section, the analysis of the CNV output from WES data identified a large homozygous deletion of 11 exons (from exon 6 to exon16) of *TJP2* in one affected individual from family 12 (patient 12a). To validate the deletion it was necessary to characterise the breakpoint location. Several primers were designed on the 3' of exon 5 and intron 6 (forward) and on the 5' of exon 17 and intron 16 (reverse) (Appendix Table I.3). Each pair amplified a DNA fragment with a predicted length size greater than 5 kb if the deletion was present. Initially, long range polymerase chain reaction (PCR) was adopted (section 2.3.10), because this is suitable for the amplification of target DNA sequence from 5 kb up to 20 kb in length. However, no result was obtained. Subsequently, end-point PCR was tested using the same combination of primers with an extension time of 6 minutes (section 2.3.7). As shown in Figure 2.4.7, the pair of primers located in exon 5 for the forward primer and in exon 17 for the reverse primer (primer pair number 4 in Appendix Table I.3), amplified the specific sequence only in patient 12a, who had the mutated allele.

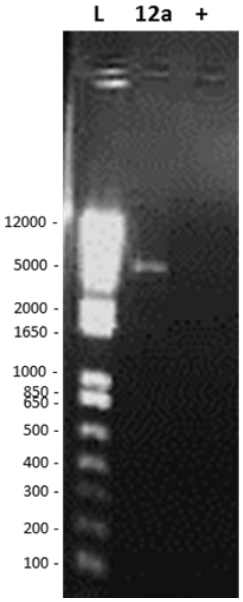

Figure 2.4.7 Agarose gel image of the large deletion present in patient 12a

The image reveals a positive amplification in the patient 12a, previously identified as having 11 exons deleted. The primers were designed for exon 5 (forward) and exon 17 (reverse). A ladder (L) was also loaded in the gel for the identification of the size length. The PCR product for patient 12a resulted in a 5 kb band as expected; no amplification was seen in the wild type DNA control (+).

Afterwards, Sanger sequencing was carried out for the identification of the breakpoint. An additional nested pair of primers located on the 3' of the intron 6 (forward) and on the intron 16-17 exon boundary (reverse) were used for the sequencing reaction. The sequences were aligned to the reference gene (NG_016342) using CLC Main Workbench and the breakpoint localised 758 bases upstream the 5' of exon 6 and 254 bases upstream the 5' of exon 17 (c.953-758_2356-254del). The analysis of the sequence suggested that a possible rearrangement had occurred between two short interspersed nuclear elements localised on the intron 6 and on the intron 16.

Subsequently, a new pair of primers was designed in the proximity of the deletion breakpoint (Appendix Table I.2). The affected patient 12a was re-genotyped by

end-point PCR along with the affected sibling (12b), the unaffected sibling and both parents. Exon 9 was selected for the detection of wild-type allele amplification. As shown in Figure 2.4.8, DNA bands related to the mutated allele were present in all members of the family 12, while in the control DNA and in the negative template control no amplification was observed. This suggests that the entire family 12 carried the mutated allele containing the 11 deleted exons. Furthermore, the amplification of exon 9 revealed no DNA bands for the two affected siblings (12a and 12b), but clear bands for the unaffected siblings (U), both parents (M and F) and for the control DNA (+). This confirmed the presence of homozygous deletion in *TJP2* in the siblings 12a and 12b affected by chronic cholestatic liver disease, and of heterozygous deletion in the unaffected sibling and in both parents.

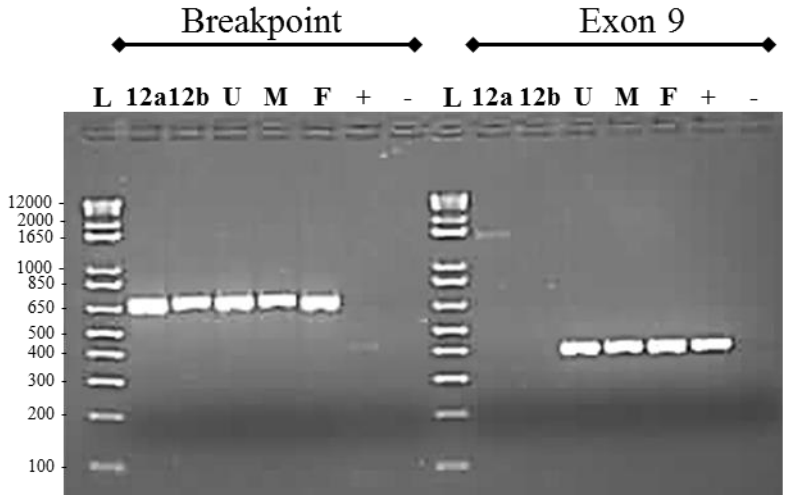

Figure 2.4.8 PCR genotyping of the large deletion in *TJP2* in family 12

The agarose gel electrophoresis represents the PCR amplification of the mutated allele due to a deletion of 11 exons (6-16) (left side) and the wild-type allele (right side) with the amplification of exon 9. The size of the bands is determined by a ladder (L). For the siblings 12a and 12b affected by chronic cholestatic liver disease clear bands for the mutated allele, but no bands for exon 9, are present, whilst, the unaffected sibling (U), the mother (M) and the father (F) showed amplification of both alleles. DNA control (+) was also tested, which shows a DNA band only for exon 9. No amplification is seen in the negative control (-).

2.4.6| Sanger sequencing validation and splicing consequences for WES findings

Subsequent investigation was carried out to evaluate the consequence of that large deletion on the splicing mechanism. RNA from patient 12a was isolated and reverse transcribed as described in sections 2.3.8 and 2.3.9. Primers for exon 5 (forward) and for the boundary between exon 17 and exon 18 (reverse) were designed and optimised. The cDNA of patient 12a was amplified at the optimal annealing temperature of 58°C. As shown in Figure 2.4.9a, the PCR amplification occurred in the patient 12a, who was previously genotyped as homozygous for the 11-exons deletion. As expected, the control cDNA synthesised from the RNA of a healthy donor showed no amplification. The PCR product of patient 12a was then purified and Sanger sequenced (section 2.3.7). The data from the forward and the reverse strands were then aligned (Figure 2.4.9b) and visualised (Figure 2.4.9c) for interpretation. The analysis revealed a splicing event between exon 5 and exon 17, without any intronic retention. However, examining the novel coding sequence, an alteration in the reading frame was discovered; the glutamic acid in position 318 was substituted with a glycine, which was then followed by a premature terminator codon (p.Glu318Glyfs*2). As the previous mutations identified by TRS-21, the large deletion in *TJP2*, identified through WES was predicted to be protein-truncating.

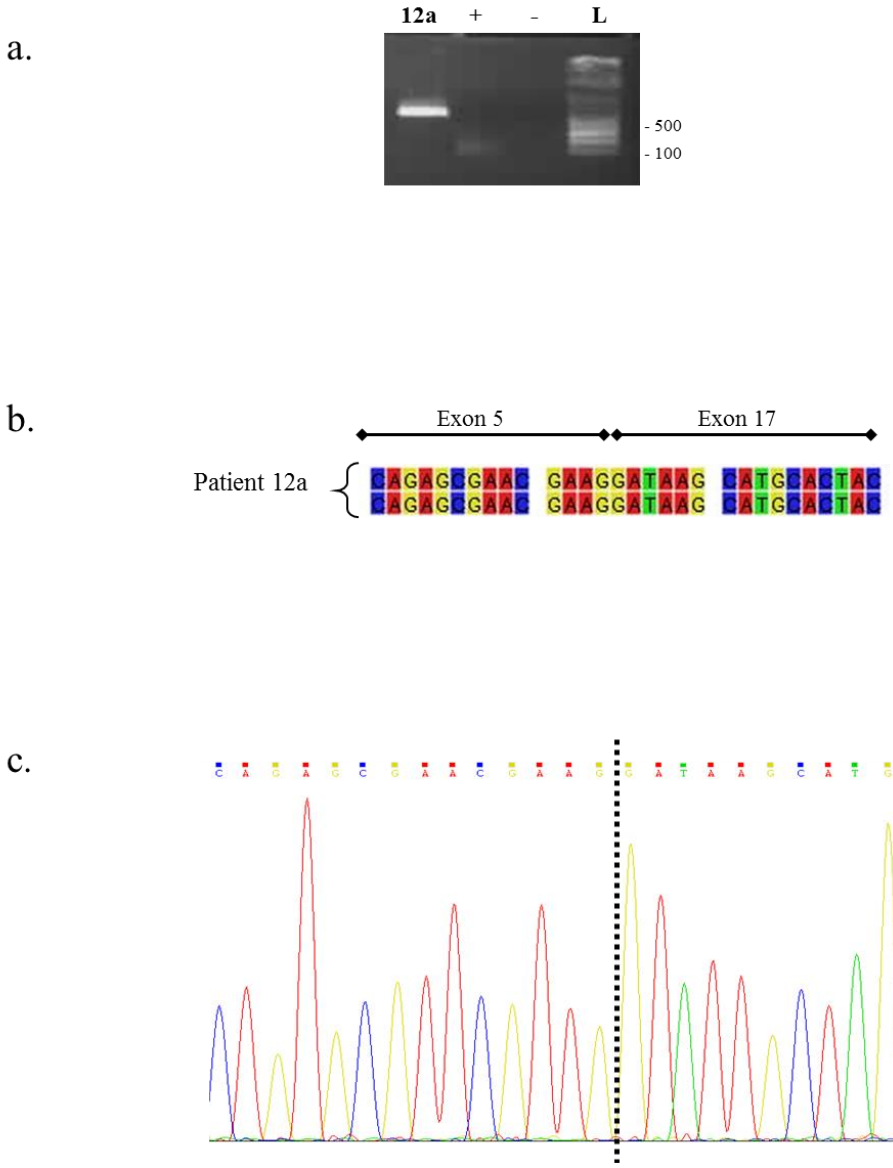

Figure 2.4.9 cDNA analysis of the breakpoint deletion in patient 12a

a) Electrophoresis showing the amplification of the cDNA fragment from exon 5 and exon 17. Whilst a band is seen in patient 12a, who has the mutated allele with 11 deleted exon (6-16), no amplification occurs using wild-type cDNA (+) and in the negative template control (-). A ladder (L) defines the size of the band; b) Alignement of the sequences from the forward and reverse strands showing a splicing event between exon 5 and exon 17; c) electropherogram of the nucleotide sequence of the junction between exon 5 and exon 17, indicated by the the dotted line.

2.4.7 **NGS run metrics for TRS-7**

Using TRS-21 and WES, six families were identified as having disease-causing mutations in the tight junction protein 2 gene. In light of this finding, an expanded cohort of patients was selected for target resequencing using a panel of 7 genes, as described in sections 2.3.4.1 and 2.3.4.3. The patients were selected using criteria similar to those applied for the previous study cohort, such as normal or low concentration of GGT activity and no mutations identified by routine genetic testing of *ABCB11* and *ATP8B1*. In addition, cases affected by milder phenotypes of cholestatic liver disease were also included. In this part, consanguinity was not adopted as a stringent factor. Genetic material was collected from the 70 patients meeting the selection criteria described above. The DNA library was prepared in accordance to section 2.3.4.3. Five different sequencing runs were undertaken multiplexing up to 16 samples per run. MiSeq benchtop sequencer from Illumina platform was used as the NGS platform. Analysing the metrics of each run, an intra-run variability was observed (Table 2.4.7). Different stages of the library preparation might affect the performance, such as a low quality DNA or an inadequate hybridisation condition, or operator errors. However, the millions of reads generated for the 7 regions of interest were sufficient to obtain good data quality. For the low coverage regions, Sanger sequencing was additionally performed. This was not applicable for five samples, in which the number of reads were <2 million and the cluster density of 89 k/mm^2 have produced unreadable data.

Millions of reads (mean)	9.32 (range 1.89 – 19.64)
Millions of reads PF (mean)	8.67
Mean >Q30	96.56
Cluster density (k/mm^2)	Range 89 ± 3 - 1043 ± 21
% Clusters PF	Range 90.9 ±0.33 – 94.79 ±0.32

Table 2.4.7 **Summary of the metrics of the TRS-7 sequencing runs**

2.4.8 Targeted resequencing variant detection in the larger cohort

The analysis was carried out for the TRS-7 data of 65 samples using CLCbio software (section 2.3.5.3); the data of five samples were excluded for further analysis. The software was set up for the detection of SNVs, indels and splice site variants, whereas for large deletions an in-house method was applied. Although the samples were exclusively selected on the basis of a lack of mutations in *ABCB11* and *ATP8B1*, these two genes were also included in the gene panel, due to the higher sensitivity of the NGS method. In the total cohort of 65 individuals, nine were identified having homozygous disease-causing mutations in *TJP2* (Table 2.4.8)

In Table 2.4.8 the summary of *TJP2* mutations are described. A one-base deletion was identified in exon 16 of patient 67; at the amino acid level the deletion was predicted to lead to a substitution of a leucine with a premature stop codon. A truncation of the protein was also the result of a homozygous nonsense mutation identified in patient 36a. In family 36, an additional sibling (36b) was diagnosed with cholestatic liver disease. Sanger sequencing confirmed that sibling 36b carried the same disease-causing mutation. One homozygous missense mutation was found on exon 15 of patient 38, which was causing the amino acid substitution of a glycine in position 737 for an arginine. In addition, three cases (19, 35 and 75a) were identified sharing the same homozygous missense mutation on exon 17. In this case, the amino acid substitution occurred between a histidine in position 788 and a leucine. In family 75, the unaffected sibling (75b) was also genotyped for the p.His788Leu change. Except for case 34 previously discussed, all changes showed a good read coverage with almost 100% frequency; they were all validated with the conventional sequencing technology. During the analysis, few areas of low or missing coverage were detected and Sanger sequenced. Due to the high GC content

and sequence repetitions, exon 5 is the major area affected. Two different one-base deletions were discovered in exon 5 in patients 34 and 39, both causing a shift of the reading frame. Interestingly, the novel protein-coding sequences ended at the same premature terminator codon of the patients described in the previous section 2.4.2.2; all having disease-causing mutations in exon 5.

Case number	Exon number	Nucleotide change	Amino acid change	Total coverage (reads)	Frequency variation (%)
19	17	c.2363A>T	p.His788Leu	95	100
34	5	c.590delG	p.Arg197Leufs*114	*Sanger Sequencing identification*	
35	17	c.2363A>T	p. His788Leu	1033	100
36a	17	c.2524 C>T	p.Gln842*	119	100
36b	17	c.2524 C>T	p.Gln842*	*Sanger Sequencing identification*	
38	15	c.2209G>C	p.Gly737Arg	73	100
39	5	c.782delA	p.Tyr261Serfs*50	*Sanger Sequencing identification*	
62	17	c.2509C>T	p.Arg837*	1138	99.56
67	16	c.2326delT	p.Leu776*	4446	99.73
75a	17	c.2363A>T	p.His788Leu	856	99.65
75b	17	c.2363A>T	p.His788Leu	*Sanger Sequencing identification*	

Table 2.4.8 Homozygous mutations identified in *TJP2* through TRS-7

Mutations identified in 11 individuals belonging to 9 families in *TJP2* affected by cholestatic liver disease with different age of onset and degree of severity. Mutations are described using the reference transcript NM_004817. Each change is described by the nucleotide change, the amino acid substitution, the total number of reads covering each specific position and the percentage of reads that differs from the reference nucleotide base.

2.4.9 **Phenotypic spectrum of TJP2 deficiency**

A total of 83 families was analysed using different next-generation sequencing strategies: targeting specific regions of interest (TRS-21 and TRS-7) and sequencing the entire human exome (WES). Most of the paediatric patients selected had a clinical diagnosis of chronic cholestatic liver disease; some mild cases were also included in the enlarged cohort and analysed by TRS-7. Mutations in tight junction protein 2 were identified in 15 families, including 20 affected individuals. All patients belonged to consanguineous families (Figure 2.4.10). *TJP2* was initially included in the panel of 21 genes due to its association to familial hypercholanemia (FHC) (Carlton *et al.*, 2003). In this study, the 17 individuals recruited were affected by FHC and were all descendants of the Old Order Amish in the Lancaster Country, Pennsylvania, US. FHC was manifest by variably elevated serum bile acid concentrations, which ranged between a minimum of 14 µg/ml up to a maximum of 217 µg/ml. The biochemical markers of liver injury were all normal except for the alkaline phosphatase, which showed an intermittent raised activity. Clinical symptoms, such as failure to thrive, vitamin-K deficiency associated coagulopathy, rickets, jaundice and itching, were not evident in all patients. In fact, some cases had elevated serum bile acids and no other related features. Phenotypically, FHC showed often a mild alteration in liver function, which was in some cases only evident if biochemical tests were undertaken. Genotypically, the disease was proposed to have an oligogenic inheritance (section 2.1.2.1 and section 2.1.4). In fact, six individuals were homozygous for the missense mutation c.143T>C in *TJP2*, 4 were homozygous for the missense mutation c.226A>G in *BAAT* and 6 were both homozygous c.143T>C in *TJP2* and heterozygous c.226A>G in *BAAT*. Mutations were not present in one individual with FHC. In addition, unaffected siblings of the six patients homozygous missense mutation c.143T>C in *TJP2* were also homozygous for the same mutation, suggesting incomplete penetrance. In contrast to FHC manifestations, within our

cohort of 21 individuals having mutations in *TJP2*, 12 families including 17 individuals presented severe intrahepatic cholestasis at early age of life (Table 2.4.9). Novel homozygous deletions (8 families) and novel homozygous nonsense mutations (3 families) in *TJP2* were discovered; all were presumed to eliminate protein expression. A homozygous missense mutation was identified in patient 38, who also presented with early-onset chronic cholestasis and required liver transplantation at the age 5. Liver transplantation was necessary for other 9 individuals. Biochemical markers showed a serum concentration of GGT slightly above the upper limit of normal. Liver injury was characterised by elevated serum bile acids, bilirubin, ALT and AST (Table 2.3.1). In family 12, chronic respiratory disease was also diagnosed in conjunction with chronic cholestasis. Three siblings from the same family unfortunately died; genetic material was available from only one of them, which was genotyped for the homozygous deletion of 11 exons in *TJP2*. Lung disease was also present in family 36, however, with different degrees of manifestation. Other extrahepatic features were seen in patient 9, who had neurological disease causing limited mobility and in patient 39 with hearing loss. A different phenotype was seen in the remaining three families, where the affected individuals were all homozygous for the same missense mutation (Table 2.4.9). In patient 19 cholestasis first manifested with jaundice and pruritus at 14 months, and then rapidly resolved. Two following episodes of jaundice with raised serum bile acids occurred at age of 4 and 5 years, probably triggered by antibiotic administrations. Patient 35 presented with cholestasis manifesting as jaundice and gallstones at the age of 13 years, which cleared after 3 months. No other episodes occurred. The last case of cholestasis due to abnormalities in tight junction protein 2 was patient 75a, in whom jaundice and pruritus manifested at the age of 9 years. The sibling 75b had the same homozygous missense mutation, however he is still asymptomatic with normal biochemical markers.

Family number	Case number	Nucleotide change	Amino acid change	Age of presentation (months)	Age of OLT (years)	Degree of cholestasis	Notes
2	2	c.766_769delGCCT	p.Ala256Thrfs*54	3	2.6	Chronic	
4	4a	c.885delC	p.Ser296Alafs*15	2	10	Chronic	Subdural hematomas
	4b			0.5	ND		Stable age 7
9	9	c.782delA	p.Tyr261Serfs*50	2	2	Chronic	Neurological disease
10	10a	c.1361delC	p.Ala454Glyfs*60	1	2.5	Chronic	
	10b			0.25	8.5		
11	11a	c.1992-2A>G	p.Arg664Serfs*2	2	6	Chronic	
	11b			2	4		
12	12a	c.953-758_2356-254del	p.Glu318Glyfs*2	3	4	Chronic	Chronic respiratory disease
	12b			0.75	ND		Died
19	19	c.2363A>T	p.His788Leu	14	ND	Remittent	Resolved
34	34	c.590delG	p.Arg197Leufs*114	24	NK	Chronic	
35	35	c.2363A>T	p.His788Leu	13 years	ND	Remittent	Resolved
36	36a	c.2524 C>T	p.Gln842*	36	NK	Chronic	Respiratory disease
	36b			60	NK		Respiratory disease
38	38	c.2209G>C	p.Gly737Arg	5	5	Chronic	
39	39	c.782delA	p.Tyr261Serfs*50	<6	ND	Chronic	Deafness
62	62	c.2509C>T	p.Arg837*	NK	NK	Chronic	
67	67	c.2326delT	p.Leu776*	7	Listed	Chronic	
75	75a	c.2363A>T	p.His788Leu	9 years	ND	Remittent	Resolved
	75b			Asymptomatic	ND	Asymptomatic	

Table 2.4.9 Summary of the genotype-phenotype association in *TJP2* deficiency patients

All patients identified with mutation in *TJP2* are here summarised. Mutations are described using the reference transcript NM_004817. OLT: orthotropic liver transplantation; ND: not done; NK: not known. A dash indicates information not available.

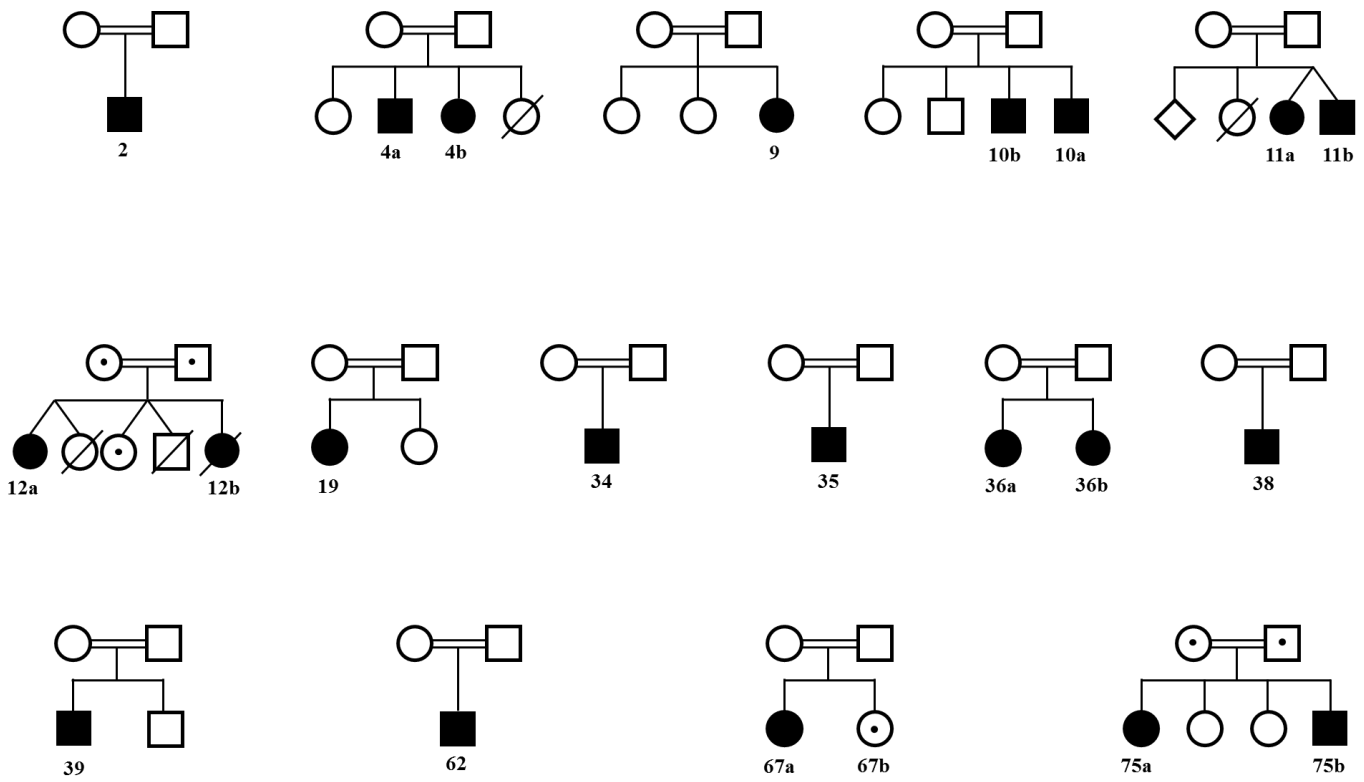

Figure 2.4.10 Pedigree of the families found with mutation in *TJP2*

Black filled shapes indicate individuals with mutations in *TJP2*. Untested relatives are indicated with unfilled shapes, while carriers of heterozygous mutations are marked with unfilled shapes with central dot. Genetic material was not available for every unaffected individual.

2.5 Conclusions of the genetic analysis

Genetic cause of cholestatic liver disease was identified in 5 patients, having homozygous protein-truncating mutations in tight junction protein 2 gene (*TJP2),* as described previously; no mutations in the selected panel of genes was found in 13 individuals. Within this group, 7 were selected for WES as they belonged to consanguineous families and therefore, had a higher change to inherit autosomal recessive disorder, and a homozygous mutation. WES analysis revealed an additional TJP2 deficiency patient. In light of this findings, an extended cohort of 70 patients was sequenced though NGS, targeting a panel of 7 genes, including *TJP2* (Figure 2.5.1b). Cholestasis in these patients was manifest at different age of onset and different degree of severity, from mild remittent episodes to severe cholestatic condition. Unfortunately, sequencing failed for 5 cases, maybe due to a possible degraded DNA; therefore data analysis was undertaken for the remaining 65 patients. Different mutations in TJP2 were found in 9 families. In summary, different methods of next-generation sequencing technology had identified 20 individuals belonging to 15 families with a diagnosis of idiopathic intrahepatic cholestatic disorder as having homozygous mutations in a single gene, *TJP2*. This gene encodes the tight junction protein “zona occludens (ZO)-2”, a cytoplasmic component of the tight junctional structures that mediate the linkage between the integral proteins, such as claudins, and the cytoskeleton of actin. Several sites in the gene were shown to be mutated, however a high susceptibility was present on exon 5 due to a great number of sequence repetitions. The majority of the mutations were predicted to be protein-truncating and were associated with a severe phenotype that in 9 cases required liver transplantation. A large deletion of 11 exons was discovered only through whole-exome sequencing analysis, in which a CNV detection tool was applied. Interestingly, extrahepatic features, such as severe respiratory disease and deafness were also manifest in some cases. Tight junction

complexes are ubiquitously present in epithelial cells, including hepatocytes and cholangiocytes, so these manifestations could have been triggered by the presumed deficiency of ZO-2 expression outside the liver. However, further investigations are required to address this question. In addition, three families including 4 individuals were found to carry the same homozygous missense mutations, which *in silico* analysis predicted would cause severe damage to the function and the structure of the translated protein. The substitution occurs on a guanylate kinase-like domain, which is involved in the folding of the nascent protein (section 3.1.1). A possible alteration of this mechanism has been proposed, but experimental validation studies are needed to better characterise the change at the structural level. Clinically, these three cases have a milder phenotype compared to the patients reported with protein-truncating mutations, manifesting with later disease-onset and remittent cholestasis. One sibling, however, carries the same homozygous missense mutation, but remains still asymptomatic with no alteration in liver function. In conclusion, the evaluation of the genetic results and the clinical phenotypes of the patients with mutations in *TJP2* has identified a genotype-phenotype correlation, showing that this novel entity has a wide spectrum of cholestatic liver disorder manifestations. In the following part of the project, the consequences of these mutations were investigated. Furthermore, WES discovered one individual to have a homozygous missense mutation in the gene α-methylacyl-CoA racemase (*AMACR*) (Figure 2.5.1a). This finding was in agreement with the clinical phenotype of a possible metabolic disorder, which was then confirmed by clinical mutation analysis elsewhere. For the other 5 patients, the failure in the identification of the causative mutations might be due to the usage of a stringent variant filtering process. Therefore, after an adjustment in the data analysis and the interpretation strategy, the WES data of these patients will be re-analysed. In addition, WES analysis of affected siblings will be generated and merged to the existing data, to increase the likelihood of discovering the genetic cause of the cholestatic liver disease in the other affected individuals.

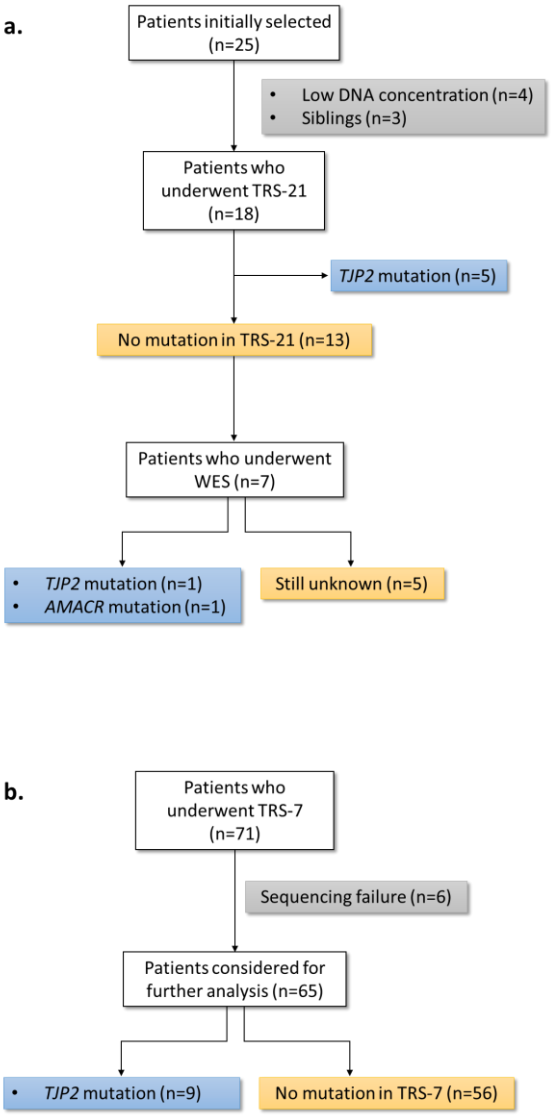

Figure 2.5.1 Flow diagram of findings identified through the different next-generation sequencing approaches

a) 25 patients were initially selected; however DNA concentration was suitable only for 21 individuals. Targeting resequencing for a panel of 21 genes (TRS-21) was performed in 18 individuals, one per family. Five patients were identified having protein-truncating mutations in tight junction protein 2 (*TJP2*). No mutations was found in 13 individuals. Within this group, 7 were selected for whole-exome sequencing (WES) as belonging to consanguineous families. One additional mutation was identified in TJP2 and a homozygous missense mutation in the gene α-methylacyl-CoA racemase (AMACR). In the remaining 5 cases, a failure on identification a disease-causing mutation occurred; b) extended cohort of 71 patients undergone targeting resequencing for a panel of 7 genes (TRS-7). Sequencing failure occurred for 6 samples and therefore data analysis was carried in the remaining 65 cases. Nine families were discovered having disease-causing mutations in TJP2.

3 Consequence of mutations in tight junction protein 2

3.1 Tight junction proteins

In epithelial tissues, the plasma membrane of two neighbouring cells is sealed together through cell-cell junctional complexes. Tight junctions represent the most apical of the junctional structures and are principally composed of transmembrane proteins, such as claudins and occludins, linked internally to cytoplasmic plaques. These intracytoplasmic complexes represent a crosslink between the integral proteins and the filaments of actins, whose importance is to maintain the solid and confluent structural characteristics of epithelial tissues. Multiple proteins are recruited in the formation of these peripheral membrane complexes, where zona occludens proteins carry out a key role.

3.1.1 Tight junction protein zona occludens 2 (TJP2/ZO-2)

When studying the components of tight junctions in the plasma membrane, a 160 kilodalton (kDa) polypeptide was discovered to be stably bound to the tight junction protein "zona occludens-1" (ZO-1) (Gumbiner *et al.*, 1991). This novel polypeptide was identified by co-immunoprecipitation assay performed in Madin-Derby canine kidney (MDCK) cells and it was proposed as an important member of the cell junctional structure. Subsequently, in 1994 a new gene *X104* was localised in the most centromeric region of chromosome 9 within the Friedreich's ataxia locus (Duclos *et al.*, 1994); however, no association to the Friederich's ataxia was identified. In addition, its expression was manifest with a similar pattern

in 8 different human tissues, showing, therefore, a biological role widely distributed in the human body. Afterwards, the comparison of the cDNA sequence of the canine ZO-2 to the sequence of *X104* identified homology (Beatch *et al.*, 1996). However, due to their interspecies difference, the canine sequence possessed an additional region of 69 amino acids and a terminal region with a low degree of similarity of 13%. On the contrary, a high degree of similarity was identified between the amino acid sequence of the canine ZO-2, the human ZO-1 and ZO-2 in the characteristic MAGUK domains. As described in section 2.1.4, three distinguishing functional regions were included in the MAGUK proteins: three PDZ domains, a single SH_3 and GK-like domains, described in section 2.1.4 (Figure 3.1.1)

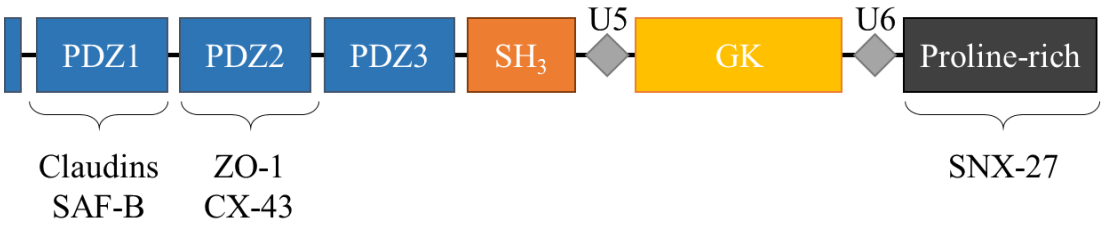

Figure 3.1.1 Structure of ZO-2 and protein-protein interaction

The characteristic structure of ZO-2 includes three PDZ regions, a SH_3 domain, a guanylate kinase (GK)-like domain and a proline-rich region in the C-terminus. In addition, two unique (U5-U6) regions are also identified. The interaction of ZO-2 with the claudin proteins, the DNA-binding protein scaffold attachment factor B1 (SAFB), the homologous zona occludens 1 (ZO-1), the gap junction protein connexin 43 (CX43) and the sortin nexin 27 (SNX27), are included in the figure and described in section 3.1.2. The figure is adapted from (Bauer *et al.*, 2010)

The PDZ domain is a highly conserved region of approximately 100 amino acid residues; its name derives from the first letter of the three proteins where it was discovered: the post synaptic density protein 95 (Psd-95), the *Drosophila* protein Discs-large (Dig) and the tight junction protein ZO-1(Harris & Lim, 2001). This domain is contained in numerous proteins and identified in different species including plants, bacteria, yeast and humans. Its major role was established as a

protein-protein interaction element, binding in particular a specific C-terminal motif of partner proteins. Multi-protein complexes are constructed through this interaction, and they are involved in a variety of functions such as intracellular signalling, cell adhesion and cell polarity. In ZO-2, the three PDZ modules are necessary to directly link to claudins, to the homologous ZO-1 and to other structural proteins described in section 3.1.2 below. The second MAGUK domain is represented by the small sequence SH_3. Known as the moderator of the tyrosine-protein kinase Src in the activation of the intracellular signalling transduction pathways, the role of the SH_3 domain in the MAGUK proteins has not been clarified as yet (Mayer, 2001). In ZO-1 protein the formation of a protein core is mediated by the interaction of the SH_3 to the GK-like domain (Lye *et al.*, 2010). In addition, U (unique) 5 and U6 regions were also identified in the core region. Interestingly, these two motifs appear to carry out two opposite roles: while the U5 are involved in the localisation of ZO-1 *in vivo* and in the binding of occluding proteins *in vitro*, the U6 motif delocalises the tight junction and inhibits the protein interaction. The third domain type of the MAGUK proteins is the guanylate kinase (GK)-like domain. The named is derived by the high sequence similarity to guanylate kinase, which is involved in the conversion of GMP to GDP; however, this GK-like domain has lost its catalytic property and it has evolved into a protein-protein interaction site and as part of the protein folding with SH_3 (Olsen & Bredt, 2003).

In normal epithelial cells, the two alternative promoters P_A and P_C transcribe two ZO-2 isoforms: the full length ZO-2A and the truncated ZO-2C, in which the first 23 amino acids in the N-terminal region are missing (Chlenski *et al.*, 1999). Using Northern Blot analysis, the expression of the two alternative isoforms was characterised in 8 different human tissues (Chlenski *et al.*, 2000). (Figure 3.1.2). ZO-2 was shown to have higher specificity for heart and brain tissues, while ZO-C was found in kidney, pancreas and placenta tissues. In the liver, both isoforms were observed, however ZO-A had a greater abundance.

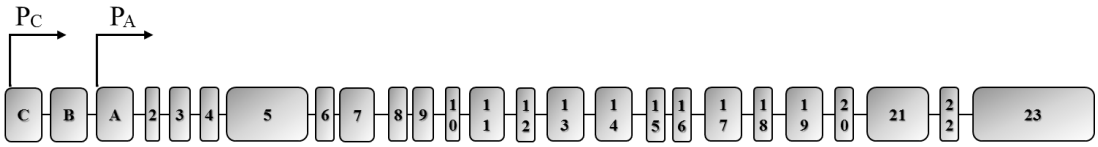

Figure 3.1.2 Representation of *TJP2* gene

Each exon is shown as a grey box. P_C and P_A identify the two promoters that transcribe for ZO-C and ZO-A respectively.

3.1.2 TJP2/ZO-2 protein-protein interaction

The structure of cell junctions is maintained by the interaction of numerous integral and cytoplasmic proteins. ZO-2 is a scaffolding protein responsible for the assembly of multi-protein complexes in the plasma membrane. As described above in section 2.1.4, the COOH-terminals of claudins are directly associated with ZO-2. Claudins represent a large family composed of 24 transmembrane proteins in humans, whose expression varies amongst tissues. In mice mRNA expression or protein level distribution revealed expression of claudin-1, -2, -3, -5 and -7 to be detectable in the liver (Mitic *et al.*, 2000). In addition, *in silico* analysis using SAGE (Serial analysis of gene expression) Genie database (http://cgap.nci.nih.gov/SAGE) provided an expanded quantitative analysis of claudins' tissue distribution, where, for example, claudin-10 was identified as expressed in the liver (Hewitt *et al.*, 2006). However, the complete tissue variability of claudins' expression has not been demonstrated, as protein analysis has not been performed.

Studies have reported that the PDZ-1 domain is also directly bound to the C-terminal region of the DNA binding scaffold attachment factor B (SAF-B) (Traweger *et al.*, 2003) (Figure 3.1.1). SAF-B acts as a structural molecule in the organisation of chromatin loops and in the transcriptional machinery. The ZO-2/SAF-B interaction has been suggested to occur in the nucleus and to be involved in the transcriptional regulation.

Through the second PDZ domain, ZO-2 forms heterodimers with its homologous ZO-1 (Wu *et al.*, 2007), but also directly binds the C-terminal cytoplasmic region of connexin 43 (CX43) (Singh *et al.*, 2005) (Figure 3.1.1). ZO-2/CX43 has been shown to be co-localised on the plasma membrane in gap junctional structures at different stage of cell cycles. The functional significance of this association has not been clarified yet.

An interaction between the junctional adhesion molecule JAM and the PDZ-3 of ZO-1 has been demonstrated at the tight junctional site (Ebnet *et al.*, 2000). Due to the high homology between ZO-1 and ZO-2, it has been suggested that there may also be an association between JAM and ZO-2. To date, no data have validated this hypothesis.

In vitro analysis has, however, demonstrated that ZO-2 directly binds also to integral occludin protein (Itoh *et al.*, 1999). Knock-out studies have revealed that ZO-2 remains concentrated at the tight junction in occludin-deficient cells, indicating that other transmembrane proteins such as claudins are recruited into the junctional structure.

During the disassembly of junctional structures, caused for example by cell division, ZO-2 temporarily resides inside early endosomes. Studies have shown that this intracellular trafficking is mediated by sortin nexin 27 (SNX27), which interacts via its PDZ-domain to the C-terminal PDZ-binding motif of the ZO-2 (Zimmerman *et al.*, 2013) (Figure 3.1.1). The re-localisation of ZO-2 to the tight junction might require an additional interaction. However, no proteins have been identified in ZO-2 recycling. The complexity of the junctional structures has been partially described in this section. The tight junction protein ZO-2 has been shown to interact with several components involved both in the cell-cell junctional structures and in the regulation of gene transcription. Nevertheless, more binding sites still remain to be identified.

3.1.3 **Localisation and functions of ZO-2**

Hepatic and bile duct epithelial cells have the essential role of creating an interface between two different environments. The maintenance of the two separate spaces is primarily regulated by the tight junctional complexes. These complexes are specialised to seal the paracellular space and select the passage of molecules on the basis of their charge or size (Rao & Samak, 2013). In fact, ZO-2 together with ZO-1 have shown an active role in the preservation of the junctional assembly and integrity (Umeda *et al.*, 2006) (Figure 3.1.3).

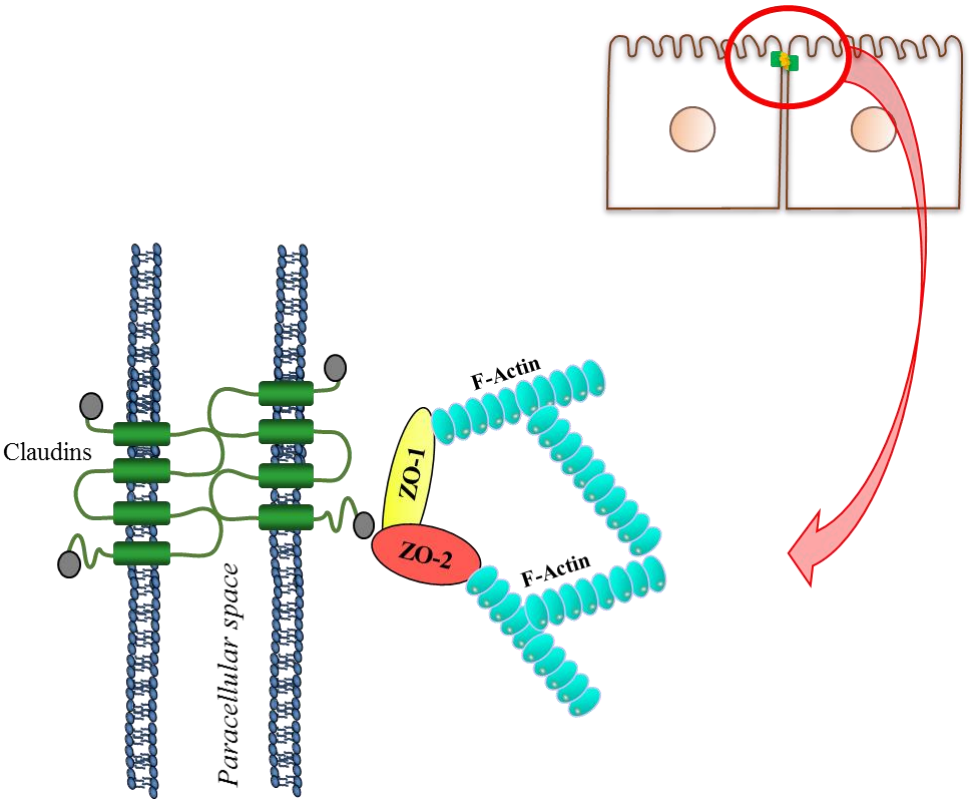

Figure 3.1.3 Simplified representation of tight junction structure in epithelial cells

Tight junctions are the most apical cell-cell junctions in epithelial cells (top image). They are composed of integral proteins such as claudins, which link to the actin cytoskeleton through a complex cytoplasmic structure, here extremely simplified. Tight junction proteins ZO-1 and ZO-2 are the main elements represented.

The importance of these proteins in tight junction formation and function has been demonstrated using ZO-1-deficient epithelial cells, in which the expression of ZO-2 was also supressed through RNA interference (Umeda *et al.*, 2006). Firstly, these cells were lacking completely the tight junction structure. Secondly, the polymerisation and the localisation of claudins at the site of polymerisation within the plasma membrane were dependent on ZO-1 and ZO-2. In the tight junction assembly no involvement of the third homologous ZO protein (ZO-3) has been identified. ZO-2 is principally involved in the construction of a selective gate at the apical end surface of epithelial cells. However, in sparse cells, an abundance of ZO-2 expression was identified at the nuclear level; this presence is therefore temporary, because the redistribution to the plasma membrane level was evident as soon as the cells were became confluent (Islas *et al.*, 2002). Due to its size, the transport of ZO-2 into the nucleus is controlled by the nuclear pore complexes. The migration of ZO-2 into the nucleus is mediated by the nuclear localisation signal (NLS), located in the end region of the first PDZ domain, while its translocation from the nucleus to the cytoplasm has been suggested to be via a nuclear export signal (NES), though the consensus sequences have not been found in humans yet (Gonzalez-Mariscal *et al.*, 2006). Nuclear ZO-2 has been suggested to be involved in cell proliferation, apoptosis and transcriptional regulation. The role of ZO-2 in cell proliferation and apoptosis is still controversial. Nuclear accumulation of ZO-2 has been demonstrated to lead to an increase in the cell proliferation rates, followed by delay in the formation of the tight junctional complexes (Traweger *et al.*, 2008). However, overexpression of ZO-2 was shown to reduce the inhibitory phosphorylation of glycogen synthase kinase-3β (GSK-3β) at the serine position 9, which led to decreased expression of cyclin D1 (Tapia *et al.*, 2009). This resulted in a block of the cell cycle at the G_0/G_1 phase and consequently to a decrease in cell proliferation. In contrast, cell proliferation and apoptosis rates were shown to be unchanged after silencing the expression of ZO-2 in epithelial cells (Hernandez *et al.*, 2007). In mice, knocking down the ZO-2 expression during the blastocyst formation caused a delay in the blastocoel cavity formation without altering the

cell proliferation or outgrowth morphogenesis (Sheth *et al.*, 2008). However, ZO-2 knock out embryos died after the implantation due to a decrease in cell proliferation at embryonic day 6.5 and an increase in apoptosis rate at embryonic day 7.5 (Xu *et al.*, 2008).

Regarding its role in the signal transduction, several proteins were identified as interacting with ZO-2 within the nucleus. For example, in epithelial cells ZO-2 was associated with Fos, Jun (Activator Protein-1 complex) and C/ERB transcriptional factors in the nuclei of sparse cells. Interestingly, when observing the sub-localisation of these proteins *in vivo*, ZO-2 was co-localised with Fos, Jun and C/ERB at the nuclear and plasma membrane region of sparse cells, but only at the plasma membrane region of confluent cells. The effect of this interaction has not been clearly determined as yet (Betanzos *et al.*, 2004).

3.1.4 **Disease associations**

A single missense mutation in *TJP2* was identified in members of the Amish population affected by a rare condition named familial hypercholanemia (FHC), described in sections 2.1.2.1 and 2.1.4 (Carlton *et al.*, 2003). Functional studies had demonstrated that the p.V48A change, located in the N-terminal of the PDZ1 domain, leads to protein instability. In addition, the binding affinity of the mutant TJP2-PDZ1 domain was reduced in respect to five different claudin C-terminal peptides (claudin-1, 2, 3, 5, 7) known to be expressed in the liver tissues. However, for the clinical phenotype to be manifest, a further mutation in an adjacent gene, *BAAT*, was identified in many patients (section 2.1.2.1), suggesting oligogenic inheritance. Mutations associated with FHC, however, were not discovered in all FHC individuals, suggesting the existence of a third locus. Until now, the p.V48A is considered a founder mutation appearing exclusively in the Amish. No other individuals belonging to different populations have been demonstrated to carry the same mutation.

The function of ZO-2 in other organs was clarified by its involvement in the genetic aetiology of progressive non-syndromic hearing loss DFNA51 (Walsh *et al.*, 2010). This autosomal dominant disorder manifests during adulthood with a hearing impairment, which progresses toward complete deafness. Using genome-wide linkage analysis and comparative genomic hybridization array, an inverted genomic duplication of part of chromosome 9 was discovered in a large affected Israeli kindred. This region included the full *TJP2* gene. Analysing the relative expression of *TJP2* mRNA in lymphoblast cells of affected individuals carrying the duplication, an increase of approximately 2-fold was identified compared to the wild-type controls. As described in section 3.1.1, the overexpression of TJP2 was previously associated with a decrease in GSK-3β inhibitory phosphorylation, which leads to an increase in GSK-3β activity. Active GSK-3β function has been

previously described promoting the overexpression of pro-apoptosis-related genes in lymphoblast cells (Beurel & Jope, 2006). It has been hypothesised that DNFA51 is caused by overexpression of ZO-2, which, in the inner ear cells, as in the lymphoblast cells, enhances GSK-3β activity and consequently cell apoptosis.

A genetic association of *TJP2* with autosomal dominant non-syndromic hearing loss (ADNSHL) was also recently described in a Korean population (Kim *et al.*, 2014). Two heterozygous missense mutations were identified as being involved in the genetic aetiology of the disease. The first p.A112T was diagnosed in one single ADNSHL patient, whose tone audiogram suggested moderate hearing loss. The alanine in position 112 located in the first PDZ domain is chemically characterised by apolarity; in this patient the wild-type amino acid had been substituted by a threonine, a polar and slightly bigger amino acid. It has been hypothesised the p.A112T change might cause protein instability and reduction in the claudin binding affinity with the first PDZ domain of the mutated ZO-2. In three additional ADNSHL individuals belonging to three different families, a different variation was identified. No correlation between the extent of their hearing loss and genetic defect has been found. *In silico* analysis revealed p.T1118A to be pathogenic and possibly causative of ADNSHL. Another gene, *CLDN14*, previously described to be involved in the hearing loss, has been studied (Wilcox *et al.*, 2001), but no linkage to ADNSHL was identified.

3.2 Research hypothesis and aims

In the previous part of this study, mutations in tight junction protein 2 were identified as being associated with the aetiology of a spectrum of cholestatic conditions. The majority of them were predicted to be protein-truncating and affect the normal physiology of ZO-2 protein. Phenotypically, these mutations caused chronic cholestatic disease. The lack of ZO-2 was hypothesised to disrupt the structure of tight junctions in hepatocytes, and consequently alter the liver function. In this part of the study, in order to investigate the consequences of these mutations, *TJP2* gene expression and ZO-2 protein were studied in liver tissues. In addition, the possible activation of a compensatory mechanism was studied and the changes in expression level of other genes associated with ZO-2 were evaluated.

3.3 Materials and methods

3.3.1 RNA isolation from liver tissue

For quantitative gene expression analysis, the cDNA synthesised from the 5 TJP2 deficiency patients with severe cholestasis and 5 healthy liver donors was used (section 2.3.8). In addition, disease controls were also selected. Total RNA was isolated from frozen liver tissue of 3 BSEP deficiency patients and 2 FIC1 deficiency patients who underwent hepatoctomy and from liver needle biopsies of 2 patients with biliary atresia (BA). Additional disease control samples were initially included in the study but liver tissues proved to be not suitable for the following procedures. The RNA isolation was performed using TRIzol® reagents (Ambion, Life Technologies), as described in section 2.3.8. The quantity and the quality of the RNA were evaluated using Qubit 2.0 Fluorometer (Life Technologies) and Agilent 2200 TapeStation system respectively. For the latter the RIN number was calculated (Figure 3.3.1).

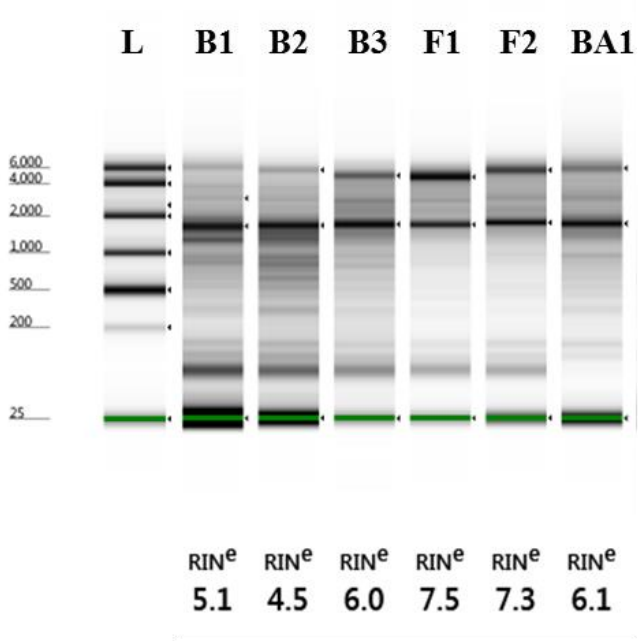

Figure 3.3.1 Analysis of the RNA quality in disease control patients

Electrophoretic separation of the RNA isolated from the liver biopsies of three BSEP deficiency patients (B1, B2, B3), two FIC1 deficiency patents (F1, F2) and one patient with biliary atresia (B1). For the second patient with BA, genetic material for quality assessment was not available. A ladder (L) was also loaded. The two bands represent the ribosomal subunit 28S (around 5 kb) and the ribosomal subunit 18S (around 1.9 kb). An RIN number was assigned to each sample depending on the 28S/18S ratio. The presence or absence of degraded material was also considered.

3.3.2 **Reverse transcription PCR**

The Omniscript Reverse Transcriptase kit (Qiagen) was used to synthesise the first strand of the cDNA from the RNA isolated from 5 TJP2 deficiency patients and 5 healthy donors (section 2.3.8), together with the RNA extracted from the liver biopsies of the patients described above (section 3.3.1). RNA was normalised to a concentration of 1 µg/µl. Oligo(dT)$_{20}$ primers, random hexamer primers and RNase inhibitor solution were provided all by Invitrogen (Life Technologies), as well as RNase inhibitor solution. Oligo(dT) primers were optimised at a final concentration of 0.5 µM, and random primers at a final concentration of 5 µM. The reaction was processed following the protocol described in Table 3.3.1 to a final volume of 20 µl. The reaction was then incubated at 37°C for 60 minutes.

	Volume (µl)	**Final concentration**
10X buffer RT	2	1X
dNTP mix (5 mM each)	2	0.5 mM each
Primers*	variable	
RNase inhibitor (10 units/µl)	0.25	10 units
Omniscript reverse transcriptase	1	4 units
Sterile water**	variable	

Table 3.3.1 Reverse-transcription reaction protocol

* Three different set of primers were employed to evaluate the nonsense mediated decay in TJP2 deficiency patients: 1) oligo(dT) was used at a volume of 1 µl with a final concentration of 0.5 µM; 2) random primers at a volume of 2 µl with a final concentration of 5 µM; 3) oligo(dT) and random primers at a volume of 3 µl. For other quantitative gene expression analysis, the primer mix was used.
** Sterile water was added to reach a volume of 20 µl

3.3.3 **Quantitative RT-PCR**

A quantitative gene expression assay was performed using the Universal Probe Library (UPL) method (Roche). The UPL technology is based on hydrolysis probes of 8 nucleotides in length, labelled with fluorescein (FAM) at the 5'-end and a dark quencher dye at the 3'. In addition, locked nucleic acids (LNA), a class of nucleic acid analogues, were inserted in the probe sequence, conferring higher thermal stability compared to conventional DNA sequence. The UPL technology provides a set of 165 probes specific for more than 7000 transcripts. In this study, a panel of genes known to interact with ZO-2, or potentially related to ZO-2 function, was selected as summarised in Table 3.3.2. The UPL technology was used also to evaluate the expression of the different *TJP2* transcripts in accordance with the Ensembl database (Figure 3.4.2). Using the probe finder web-based software tool (www.universalprobelibrary.com), the optimal assays for primers and probes were designed across the exon/exon boundaries common to the transcripts of each gene in accordance with the Ensembl database. For *CLDN14* (claudin-14) and *GJB2* (gap junction β-2 protein), where no suitable sets were available, TaqMan expression assays (Life Technologies) were purchased (Hs00377953_m1; Hs00955889_m1) and the reaction was prepared as described in Table 3.3.4. Primer sequences were then copied and purchased from Integrated DNA Technologies. Primers and probes are listed in Appendix Table III.1 and Appendix Table III.2.

Gene name	Protein name	Junctional component	Localisation
TJP1	ZO-1	tight junctions	cytoplasm
TJP3	ZO-3	tight junctions	cytoplasm
CLDN1	Claudin-1	tight junctions	transmembrane
CLDN2	Claudin-2	tight junctions	transmembrane
CLDN3	Claudin-3	tight junctions	transmembrane
CLDN7	Claudin-7	tight junctions	transmembrane
CLDN10	Claudin-10	tight junctions	transmembrane
CLDN11	Claudin-11	tight junctions	transmembrane
CLDN12	Claudin-12	tight junctions	transmembrane
CLDN14	Claudin-14	tight junctions	transmembrane
CLDN15	Claudin-15	tight junctions	transmembrane
OCLN	Occludin	tight junctions	transmembrane
F11R	Junctional adhesion molecule A (JAMA)	tight junctions	transmembrane
MARVELD2	Tricellulin	tight junctions	transmembrane
ACBT	Actin	cytoskeleton	cytoplasm
CGN	Cingulin	tight junctions	cytoplasm
CDH1	Cadherin-1	adherens junctions	transmembrane
CTNNA1	Catenin α-1	adherens junctions	cytoplasm
CTNNB1	Catenin β-1	adherens junctions	cytoplasm
GJA1	Connexin 43	gap junctions	transmembrane
GJB2	Connexin 26	gap junctions	transmembrane
RAB13	Ras-related protein Rab-13	non-structural	cytoplasm
SAFB	Scaffold attachment factor B1	non-structural	nucleus
SNX27	Sorting nexin-27	non-structural	cytoplasm

Table 3.3.2 Gene panel selected for quantitative analysis

The genes were selected on the basis of their known interaction or possible interaction with ZO-2. Gene and protein names are in accordance to HUGO Nomenclature Committee database (HGNC) and Universal Protein Resource (UniProt).

For each sample, FastStart Universal Master Mix with ROX (Roche) was added in accordance to the optimal protocol described below.

	Volume (µl)	Final concentration
2X FastStart MasterMix	5	1X
Forward primer (10 µM)	0.4	400 nM
Reverse primer (10 µM)	0.4	400 nM
UPL probe (10 µM)	0.2	200 nM
Sterile water**	3	

Table 3.3.3 PCR protocol for one reaction optimised for the UPL assay

	Volume (µl)	Final concentration
2X FastStart MasterMix	5	1X
20X TaqMan gene expression assay	0.5	1X
Sterile water**	3.5	

Table 3.3.4 PCR protocol for one reaction optimised for TaqMan assay

cDNA was diluted 1:5 and 1 µl added to each reaction to reach a final volume of 10 µl. For each sample, the reaction was performed in triplicate with the ViiA 7 real-time PCR system (Life Technologies), following the run module below.

Step	Temperature	Time
1	95°C	10 minutes
2	95°C	15 seconds
3	60°C	30 seconds
5	*Repeat from Step 2 to Step 3 for 40 times*	

Table 3.3.5 PCR programme used in the ViiA7 system

The relative expression of each target was determined in relation to an internal reference gene, ubiquitously expressed in every cell and in which the expression level remains unchanged during a disease state. Glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) was used as the internal reference gene. The ΔCt method, which represents a variation of the $2^{-\Delta\Delta Ct}$ or Livak method (Livak & Schmittgen, 2001), was chosen as the expression ratio for the reference sample, called the calibrator, being different to 1. Unlike the Livak method, the expression ratio between the Ct (cycle threshold) of *GAPDH* and the Ct of the selected target gene was initially calculated for each sample, as described below.

$$\text{Expression ratio (reference/target)} = 2^{Ct\ (reference)\ -\ Ct\ (target)}$$

Then, the mean of the expression ratio of healthy donor samples was used as calibrator. The relative gene expression for each patient sample was obtained by dividing each expression ratio by the expression ratio of the calibrator. Statistical analysis was performed using GraphPad Prism 6 software (GraphPad Software, Inc., LaJolla, CA, USA).). Normality testing was used to determine the normal distribution of the Ct values. If values were normally distributed, a Student's t-test was applied, otherwise the Mann-Whitney U-test was used as a non-parametric alternative. In addition, the Kruskal-Wallis one-way ANOVA was employed when comparing three or more unpaired groups. A p-value lower than 0.05 was considered significant.

3.3.4 **Protein isolation from liver tissue**

Proteins were isolated from frozen liver tissues of 5 patients with mutations in *TJP2* and one healthy liver donor. Up to 50 mg of tissue were homogenised in 200 µl of radioimmunoprecipitation assay buffer containing 1% nonyl phenoxypolyethoxylethanol, 0.5% sodium deoxycholate and 0.1% Sodium dodecyl sulphate (SDS) with protease inhibitors added just before use [10 µl/ml phenylmethylsulfonyl fluoride (PMSF), 30 µl/ml aprotinin, 10 µl/ml sodium orthovanadate]. Additionally 3 µl of PMSF were added to the homogenate, which was then vortexed and left to incubate for 1 hour. The entire procedure was undertaken at a temperature of 4°C. Subsequently, the homogenate was centrifuged at 15,000 rcf at 4°C for 20 minutes and the proteins isolated from the supernatant, which was transferred into a clean 1.5 ml tube. An aliquot was used for the protein quantification method and placed on ice, while the remaining solution was stored temporarily in an -80°C freezer for later immunoblot analysis.

3.3.5 **Protein quantification**

Protein quantification was determined using Lowry's method with bovine serum albumin (BSA) as protein standard (Lowry *et al.*, 1951). The 1X BSA working solution (1 mg/ml) was obtained by mixing 100 µl of 10X BSA stock solution with 900 µl distilled water. The BSA standard curve was prepared using seven BSA concentrations on a scale from 0 mg/ml to 1 mg/ml, as shown in the table below.

BSA final concentration (mg/ml)	1X BSA working solution (µl)	Distilled water (µl)
0	0	250
0.1	25	225
0.2	50	200
0.4	100	150
0.6	150	100
0.8	200	50
1	250	0

Table 3.3.6 Bovine serum albumin (BSA) concentrations for the standard curve preparation

In a 96-well plate (Thermo Scientific), 50 µl of each BSA standard and each patient sample were loaded in duplicate. In each well, 50 µl of alkaline copper reagent were then added and mixed using a microplate shaker. The plate was incubated at room temperature for 10 minutes. Afterwards, the Foolin & Ciocalteu's solution was prepared adding 0.5 ml of Folin & Ciocalteu's phenol reagent (Sigma-Aldrich Ltd) to 8 ml of distilled water; 200 µl were added to each well. The plate was mixed using a microplate shaker and incubated at room temperature for 15 minutes. The absorbance of each standard and sample was read at 650 nm using a Dynex MRX microplate reader. After plotting the BSA standard curve, the protein concentration of each sample was determined.

3.3.6 **Protein Immunoblot**

3.3.6.1 Sample preparation

After quantifying the protein concentration, each protein sample was normalised to 1-2 µg/µl. Ten µl of each sample were mixed with a solution composed of 3 µl of 10% western blot buffer (0.5 M Tris, pH 6.8, 0.35 M Sodium dodecyl sulphate (SDS), 30% glycerol, 0.6 M dithiothreitol, 0.175 M bromophenol blue) and 2 µl of

10% β-mercaptoethanol. In order to denature the proteins, the samples were heated for 5 min at 100°C and placed on ice until they were loaded into the electrophoretic gel.

3.3.6.2 SDS-polyacrylamide gel preparation

Mini gels were hand-cast using the Bio Rad Mini Gel system (Bio Rad, Hercules, CA USA). The gels were made in accordance with the molecular weight of the protein of interest. For the analysis of ZO-2, a 10% SDS-polyacrylamide gel was prepared, while for claudin-1 and claudin-2, which have a smaller molecular weight, a 15% SDS-polyacrylamide gel was used. The SDS-polyacrylamide gel includes two parts: the resolving gel on the bottom and the stacking gel on the top. While the first is involved in the migration through the gel pores and separation of the proteins depending on their size, the second allows the proteins to align and enter at the same time into the following running gel. Twenty ml of resolving gel solution were prepared by mixing water, polyacrylamide, SDS, 1.5 M Tris, 10% ammonium persulphate and Temed at different ratios depending on the percentage of the acrylamide concentration selected (Table 3.3.7)

	10% (ml)	**15% (ml)**
Water	7.9	4.6
30% acrylamide mix	6.7	10.0
1.5M Tris (pH 8.8)	5.0	5.0
10% SDS	0.2	0.2
10% ammonium persulfate	0.2	0.2
TEMED	0.008	0.008

Table 3.3.7 Protocol for 20 ml resolving gel solution at 10% and 15% acrylamide concentration

After preparing the resolving gel solution, it was poured into a glass cassette sandwich clamped to a casting stand, where it was left to solidify for approximately

30 minutes. Subsequently, 10 ml of 5% the stacking gel solution were prepared following the protocol described in Table 3.3.8 and poured on the resolving gel. A 24-well comb was then placed.

	5% (ml)
Water	6.8
30% acrylamide mix	1.7
1 M Tris (pH 6.8)	1.25
10% SDS	0.1
10% ammonium persulfate	0.1
TEMED	0.01

Table 3.3.8 Protocol for 10 ml stacking gel solution at 5% acrylamide concentration

An additional incubation of 30 minutes was allowed for the polymerisation of the stacking gel.

3.3.6.3 Electrophoresis

The hand-cast gel was subsequently unclamped from the casting place and transferred onto a gel support, where the inner chamber was filled with electrophoresis running buffer. The comb was then removed and the samples, prepared as described in section 3.3.6.1, were loaded. To monitor the progression of the protein electrophoresis, an Amersham ECL full-range rainbow molecular weight marker was also loaded (GE Healthcare Life Sciences, Amersham Place, Buckinghamshire, UK). To detect the molecular weight ladder when using the horseradish peroxidase (HRP) detection system, a biotin ladder was also loaded alongside the samples (Cell Signalling Technology, Danvers, MA, USA). Then, the gel was placed inside the Bio Rad running tank, where the outer chamber was filled with electrophoresis buffer. The tank was closed with a lid attached to a power supply with a constant voltage of 200 V for 1- 2 hours.

3.3.6.4 Membrane transfer

At the completion of the electrophoretic run, the gel was removed from the glass cassette sandwich. The transfer stack was prepared as follows: three layers of blotting paper with a size similar to the gel, the gel, a Hybond enhanced nitrocellulose membrane (GE Healthcare) with a size similar to the gel, three more layers of blotting paper. This structure was then positioned on a sponge soaked in Towbin buffer, placed into a Bio Rad gel holder cassette, which was then inserted in a transfer tank filled with Towbin buffer and an ice block to cool down. The transfer occurs from the negative to the positive electrode, so it is important that the tank nitrocellulose membrane faces the positive electrode. The lid was placed and attached to a power supply with a constant voltage of 100 V for 45 minutes.

3.3.6.5 Immunoblotting

The successful transfer was visualised firstly with the coloured markers on the nitrocellulose membrane and secondly with the bands stained with Ponceau S solution (Sigma-Aldrich). To prevent non-specific binding of the antibody, the membrane was blocked for 1 hour in using 5% BSA diluted in 0.1% TBS and 0.05% Tween-20 (TBST). Following that, primary immunoblotting was performed. For the 10% gel, several anti-ZO-2 antibodies binding different epitopes were tested (Table 3.3.9); for the 15% gel, anti-claudin-1 (18-7362, Zymed Laboratories, Invitrogen) and anti-claudin-2 (32-5600, Invitrogen) antibodies were used. The membranes were incubated overnight at 4°C. Subsequently, three washes with 0.1% TBST of 10 minutes; each was carried out to remove all excess antibodies. The membranes then were blotted with a horseradish peroxidase (HRP)-conjugated anti-rabbit secondary antibody (sc-3837, Santa Cruz Biotechnology, Heidelberg, Germany) or an HRP-conjugated anti-mouse secondary antibody (NA931V, GE Healthcare) for 1-hour incubation at room temperature. Subsequently, three further

washes with 0.1% TBST of 10 minutes; each was carried out to remove all the excess antibodies. The Amersham ECL Prime chemiluminescence detection system (GE Healthcare) was used to visualise protein bands. The membranes were exposed to X-rays films from 1 seconds to 5 minutes, depending on the intensity of the bands. Each film was converted to digital format to obtain the highest resolution. GADPH was used as housekeeping gene. After detecting the proteins of interest, the membranes were stripped of the previous blotting using the Abcam mild stripping protocol (Abcam, Cambridge, UK) and re-blocked with 5% of BSA. GADPH blotting (MAB374, Millipore) and detection was performed as previously described. Dilutions are shown in the table below.

Antibody name	Product number	Epitope	Primary antibody dilution	Secondary antibody dilution
Anti-ZO-2 Rabbit Polyclonal	HPA001813 (Sigma-Aldrich)	aa 306-452	1:5,000	1:10,000
Anti-ZO-2 Mouse Monoclonal	H00009414-M01 (Abnova)	aa 121-218	1:5,000	1:50,000
Anti-ZO-2 Mouse Polyclonal	Ab168667 (Abcam)	aa 1-1190	1:1,000	1:20,000
Anti-ZO-2 Rabbit Polyclonal	LS-B2185 (LS Source Biosciences)	C-terminus	1:5,000	1:50,000

Table 3.3.9 ZO-2 antibodies and dilutions used in the western blot analysis

Protein of interest	Antibody name	Primary antibody dilution	Secondary antibody dilution
Claudin-1	Anti-claudin-1 Rabbit Polyclonal	1:5,000	1:10,000
Claudin-2	Anti-claudin-2 Mouse Monoclonal	1:10,000	1:50,000
GAPDH	Anti-GAPDH Mouse Monoclonal	1:20,000	1:50,000

Table 3.3.10 Dilutions used in the western blot analysis for other proteins

3.4 Results

3.4.1 Analysis of *TJP2* expression

As described previously, mutations in tight junction protein 2 were discovered in 23 individuals with chronic cholestatic liver disease. The majority of them were predicted to truncate the translation of the protein, due to the generation of premature termination codons. Messenger RNAs containing PTC are usually substrates for a surveillance mechanism named nonsense mediated messenger ribonucleic acid (mRNA) decay (NMD) (Kervestin & Jacobson, 2012). This mechanism is proposed to be dependent on an exon junction complex (EJC) situated approximately 24 nucleotides upstream of the exon/exon boundary. If the translation terminates before releasing the EJC complex, the activation of NMD is trigged. The ribosomes are disassembled and recycled as is the entire translation machinery, while the abnormal mRNA is degraded. Therefore, the expression of the targeted gene is down-regulated.

In this part of the study the expression of *TJP2* was investigated. Total RNA was isolated from the five patients where specimens were available; all patients were identified as having protein-truncating mutations in *TJP2*. Five healthy donors were used as controls. An important step in the gene quantification is represented by the synthesis of cDNA and the priming method has been demonstrated to influence the efficiency of the real-time PCR (Resuehr & Spiess, 2003). To obtain a good representation of the *TJP2* expression, mRNA from patients and controls was reverse transcribed using three different primer strategies: oligo(dT), random hexamers and a combination of both (section 3.3.2). Real-time PCR was carried out using a set of primers and probes located on the exon/exon boundaries shared

between the different *TJP2* transcripts (Appendix Table III.1). The reactions were performed in triplicate and the samples from the three priming methods were amplified simultaneously, in order to reduce the inter-plate variability. Data were normalised using *GAPDH* as the internal reference gene and the relative expression ratios of *TJP2* were analysed. The Mann-Whitney nonparametric test was used to identify statistical differences between the two unpaired groups. The amplification performed using oligo(dT)-primed cDNA showed a significant reduction of approximately 80% in patients compared to controls, with a high statistical significance (p-value<0.01) (Figure 3.4.1a). The investigation was carried out using two additional priming strategies. Using random hexamer-primed cDNA or cDNA synthesised with a combination of oligo(dT) and random hexamer primers, the *TJP2* expression level decreased with statistical significance in patients compared to controls with a p-value lower than the threshold of 0.05 (Figure 3.4.1b and Figure 3.4.1c). The widespread usage of the oligo(dT) priming is related to their specificity to bind the poly(A) tail of the mRNA, always initiating the reverse transcription at the 3'-end of the mRNA. However, during the performance of the transcription, long transcripts could create secondary structures and lead to an incomplete cDNA synthesis. The stronger evidence for a decreased *TJP2* expression identified in patients against the control groups using the oligo(dT)-primed cDNA approach, suggested that this priming strategy is more susceptible to underestimating the real expression. Overall, using the different priming methods, the analysis has demonstrated that patients having protein-truncating mutations in *TJP2* had a reduced mRNA level compared to the group of controls, probably resulting from activation of a nonsense mediated mRNA decay pathway. For the further quantitative analysis described in sections 3.4.2 and 3.4.4, a combination of random primers and oligo(dT) was adopted in order to have a good estimation of the expression level of the targeted genes.

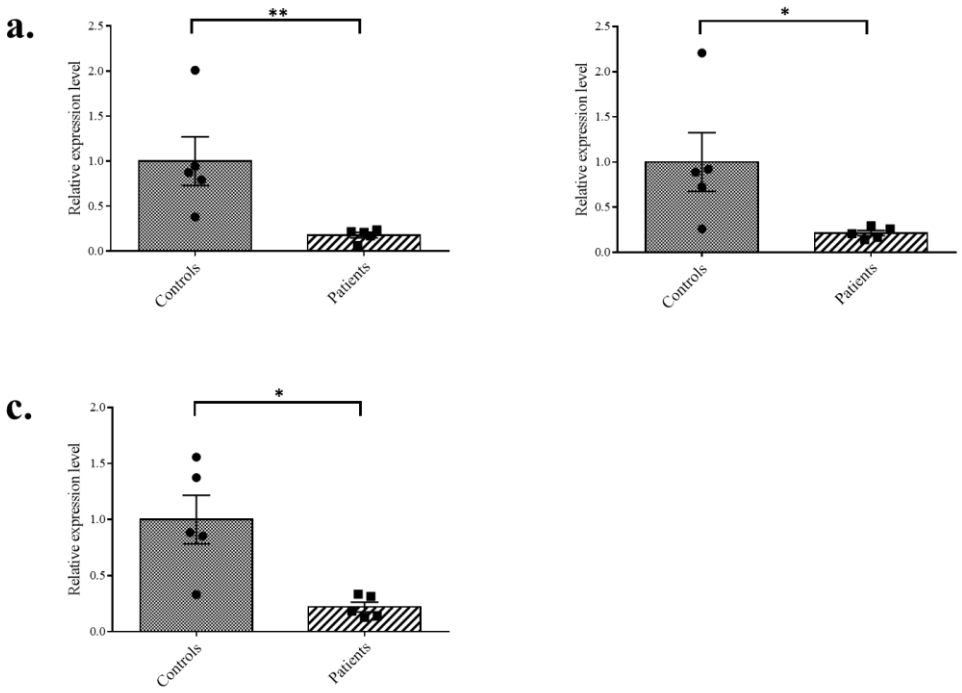

Figure 3.4.1 Liver *TJP2* expression in patients and controls

Quantitative RT-PCR was performed in a group of patients with protein-truncating mutations in *TJP2* (n=5) and a group of healthy donors selected as controls (n=5). Three priming strategy methods were used for the generation of cDNA. a) Expression analysis of the target gene using oligo(dT)-primed cDNA; b) expression analysis using random hexamer-primed cDNA; c) expression analysis of cDNA synthesised using a combination of oligo(dT) and random hexamers. The bars represent the mean ± the standard error of the mean (SEM). Statistics was carried out using the non-parametric Mann-Whitney test. * p-value <0.05; ** p-value <0.01.

3.4.2 Expression of different *TJP2* isoforms

Northern blotting expression analysis and RT-PCR have previously demonstrated that the expression of the two ZO-2 protein isoforms (ZO-2A and ZO-2C) was driven by two distinct promoters (P_A and P_C) and characterised by tissue-specificity. In liver tissue a clear higher expression of ZO-2A was shown, although a weaker band of ZO-2C was also visible (Chlenski *et al.*, 2000). The protein isoforms ZO-2A and ZO-2C are transcribed from the mRNAs *TJP2*-003 and *TJP2*-201 respectively. Alternative splicing of these two isoforms are also present (Figure 3.4.2). According to the Ensembl database, nine different alternative transcripts have been identified for the tight junction protein 2 gene. A quantitative RT-PCR assay was optimised for the relative quantification of the different loci of the gene isoforms in the patients and controls studied in section 3.4.1. Six sets of primers and probes were designed crossing the exon/exon boundary of specific loci of the different splicing *TJP2* variants; an exception was made for the analysis of the *TJP2*-004 transcript, in which primers and probe were located on its unique 3'-UTR (Figure 3.4.2).

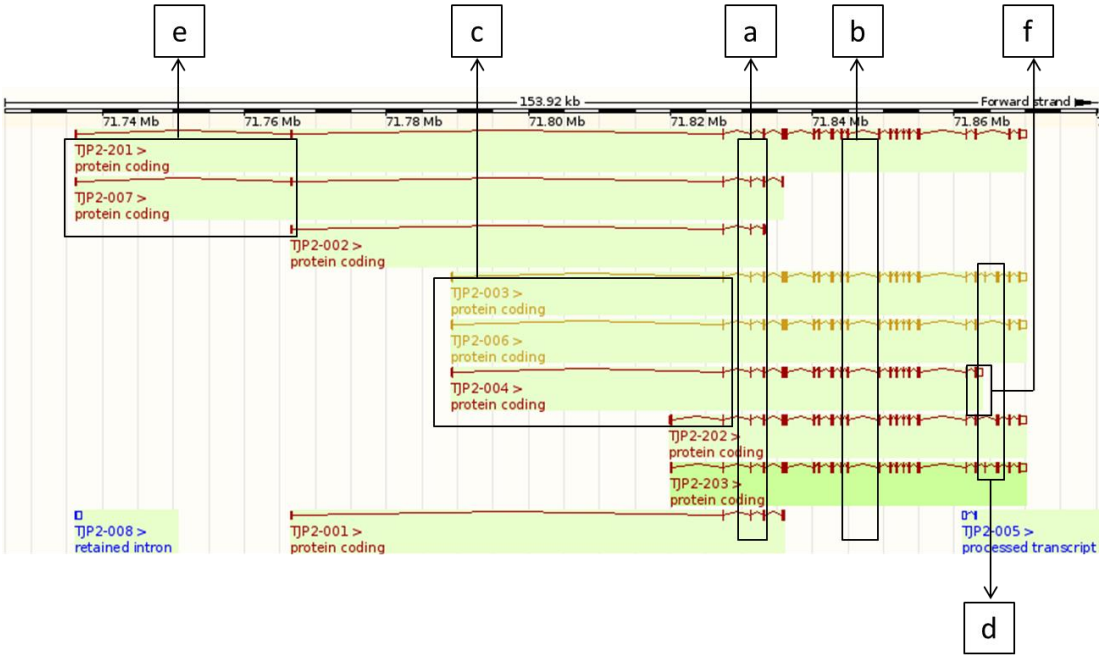

Figure 3.4.2 *TJP2* transcripts and specific UPL probes for quantitative analysis

Illustration of the assays optimised for different *TJP2* loci in accordance with the Ensembl database. Each outlined box represents the localisation of primers and probes; the lettering system correlates with those used in Figure **3.4.3**

For each sample the reaction was performed in triplicate and all target genes were carried out in the same PCR plate. The gene expression was determined relative to *GADPH* as the internal reference gene and the expression ratio calculated as described in section 3.3.3. Non parametric testing (Mann-Whitney U test) was performed to compare the two populations. The analysis revealed that the patients with protein-truncating mutations in *TJP2* have a significant reduction in mRNA expression compared to healthy donors. This was evident using the primers/probe set a, b, c and d (Figure 3.4.3). In contrast, no difference was found when using the primers/probe sets e or f (Figure 3.4.3). As shown in Figure 3.4.2, these two assays were specific for the gene isoforms *TJP2*-201, *TJP2*-007 (e) and *TJP2*-004 (f). On the basis of their Ct values, it was hypothesised that in liver tissue there was a scarcity of these specific gene transcripts both in controls and patients. However, the accuracy and the application efficiency of each probe have not been addressed

and therefore the quantification of different spliced isoforms could not be evaluated. The generation of an absolute standard curve would be required.

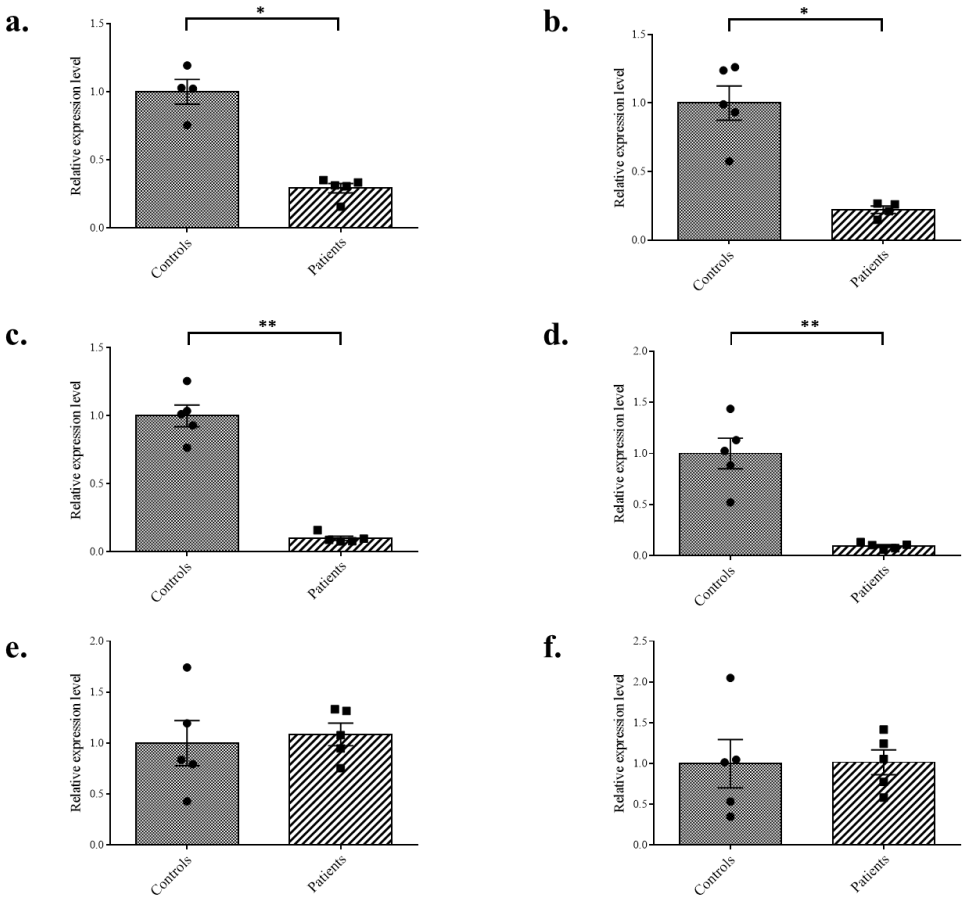

Figure 3.4.3 Expression of liver tissue *TJP2* transcripts in patients and controls

Quantitative RT-PCR was performed in patients identified with protein-truncating mutations in *TJP2* (n=5) and healthy donors as controls (n=5). Six sets of primers and probes were designed for different loci of the gene (a-f). The specific locations are illustrated in Figure **3.4.2**. The bars represent the mean ± standard error of the mean (SEM). Non-parametric Mann-Whitney test was used for statistical analysis.
* p-value <0.05; ** p-value <0.01.

3.4.3 Protein expression analysis of ZO-2 and ZO-2 interacting claudins

The expression of the mutant ZO-2 protein in patients having protein-truncating mutations in *TJP2* was investigated by western blotting. Four different antibodies, which recognise different epitopes of the protein were purchased, as described in Table 3.3.9. The first primary anti-ZO-2 polyclonal antibody from Sigma-Aldrich, raised in rabbit against the PDZ-2 domain, was initially tested in extracted proteins from healthy liver donor tissue (section 3.3.4) and the efficiency validated. Unfortunately, the resulting staining pattern was unclear with multiple bands and no precise discrimination between them. A second test was undertaken, but the same result was obtained; therefore, this antibody was considered not suitable for further protein expression analysis. Subsequently, western blotting was performed using a different primary anti-ZO-2 antibody. In contrast to the previous, it was purchased from Abnova and raised in mouse against a unique sequence located downstream of the first PDZ domain. The optimisation test was undertaken in the same protein extract from healthy donor liver tissue. Also in this case, no specific bands for ZO-2 protein were seen and, therefore, this antibody was considered inappropriate for the study. An additional polyclonal antibody was purchased from Abcam, raised in mouse against the full length of the protein (Table 3.3.9). The validation test was undertaken using proteins isolated from the same liver tissue of the healthy donor control. Five bands were visible at approximately 45, 60, 70, 100 and 160 kDa. Because of this distinctive staining pattern, the experiment was continued and ZO-2 expression was investigated in patients having protein-truncating mutations in *TJP2*. Although there was a background noise due to possible cross-reactivity between the HRP-conjugated anti-mouse secondary antibody and the blocking agent, the immunoblotting revealed that the patients lacked the bands of approximately 100 and 160 kDa when compared to the control sample (Figure 3.4.4a). This result, however, was different from the gel image

shown in the product datasheet where only one band at the predicted size of 134 kDa was visible. It may be due to the high specificity and purity of the protein lysate used for the immunoblotting test performed by the manufacturer; the protein lysate was isolated from *TJP2* transfected human embryonic kidney 293T cells.
In this study, a possible interaction with other proteins containing similar epitopes was suggested as non-specific bands at approximately 45, 60 and 70 kDa were detected in every tissue extract, both in patients and control. In the last ZO-2 protein analysis, western blotting was performed using a polyclonal antibody raised in rabbit against the C-terminus of the human protein. The immunoblotting showed a high background noise and a pattern of 5 distinctive bands in the isolated protein of healthy donor liver tissue, used as a control sample (Figure 3.4.4b) The patients identified as having protein-truncating mutations in *TJP2* lacked of the highest bands at approximately 100 and 160 kDa. GAPDH was used as the internal reference control and it showed similar intensity in every samples loaded and the expected size at 37 kDa (Figure 3.4.4c).

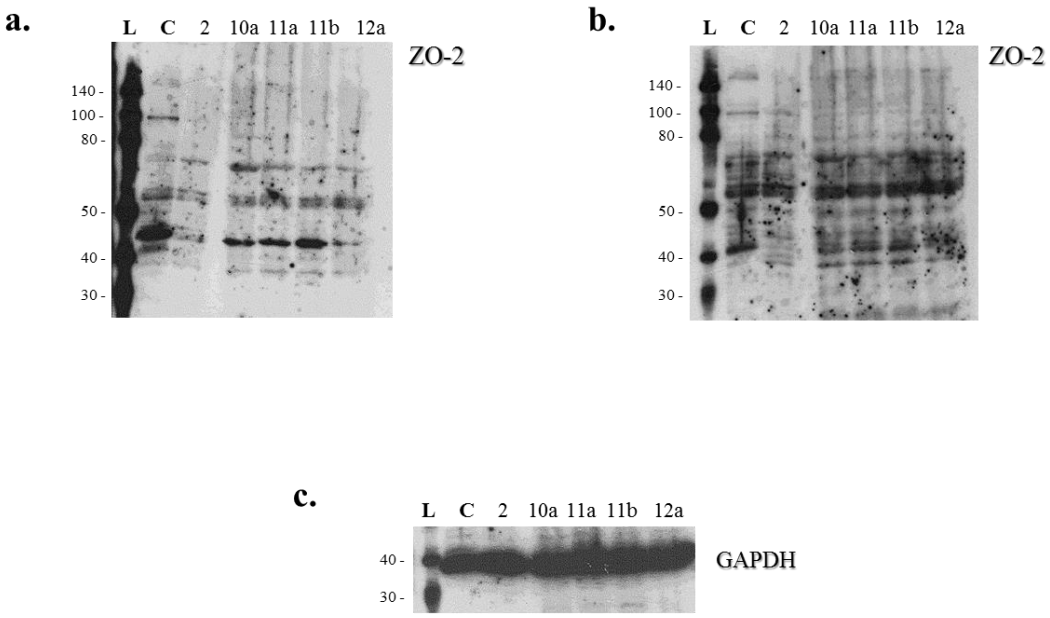

Figure 3.4.4 Western blotting for ZO-2

a) Polyclonal antibody raised in mouse against the full protein used for the identification of ZO-2 in the control (C) and in patients with protein-truncating mutations in *TJP2*. A biotinylated protein ladder (L) was also loaded on the gel; b) polyclonal antibody raised in rabbit against the C-terminus of ZO-2; c) GAPDH immunostaining used as internal control.

Subsequently, claudin-1 and claudin-2 were studied in order to investigate if the absence of ZO-2 expression could have caused a down-regulation of their expression. As described in section 3.1.2, claudin-1 and claudin-2 are integral tight junction proteins that bind directly ZO-2, and the expression of which have been identified in liver tissue. Western blotting analysis was undertaken in the same group of patients with protein-truncating mutations in *TJP2* and the healthy liver donor as a control sample. GAPDH was used as internal reference control. The expected size of 23 kDa for claudin-1 was detected both in patients and in the control (Figure 3.4.5). Then, for each sample the band intensity of the target protein was compared to the band intensity of the GAPDH internal control and, as seen in the figure below, no change in protein level was identified.

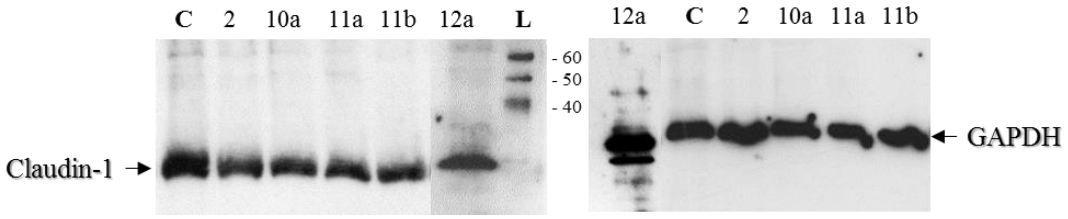

Figure 3.4.5 Western blotting for claudin-1

Left-hand gel image: claudin-1 immunostaining. In the control (C) and the patients a band at the expected size of 23 kDa is identified. A biotinylated protein ladder (L) is also included. Right the gel image of GAPDH immunostaining.

Western blotting was performed for claudin-2 and it showed the presence of bands at the expected size of 25 kDa in patients and control (Figure 3.4.6). In addition non-specific bands were detected at approximately 27 kDa as predicted by the antibody datasheet. The internal reference protein GAPDH was also visible on the same gel at the expected size of 37 kDa. Normalisation against the band intensity of GAPDH was undertaken for each sample and no alteration in protein expression level was identified in claudin-2 (Figure 3.4.6).

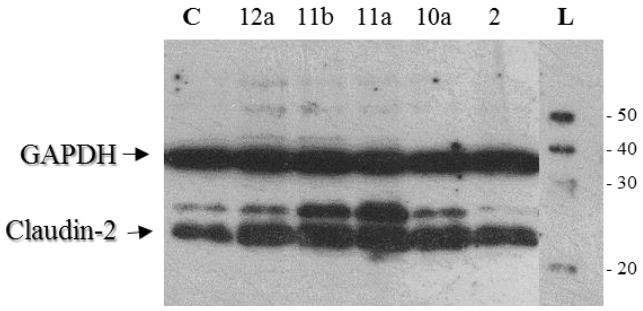

Figure 3.4.6 Western blotting for claudin-2

Proteins isolated from the liver tissues of a control (C) and five patients were immunostained with anti-claudin-2 and anti-GAPDH antibodies. A non-specific band stained by anti-claudin-2 antibody is also present at 27 kDa, in accordance to the manufacturer. A biotinylated ladder is included (L).

3.4.4 **Downstream pathway alteration**

In the previous sections, a down-regulation of *TJP2* causing an absence of ZO-2 protein was identified in patients having protein-truncating mutations. As described in section 3.1.3, ZO-2 is part of different junctional complexes, important in maintaining the organs' structure. The activation of possible compensatory mechanisms was hypothesised as a consequence of this alteration, as a physiological attempt to preserve the organ integrity. In addition, ZO-2 has been described as being involved in cell-cycle regulation after translocation into the nucleus. Therefore, other molecular pathways could have been affected. A panel of genes was therefore selected on the basis of their interaction or possible interaction to ZO-2, and their expression was investigated by quantitative RT-PCR analysis. In addition to the groups of patients and controls analysed in the previous studies (sections 3.4.1 and 3.4.2), individuals with other cholestatic conditions were included. Suitable genetic material from liver tissues was isolated in three patients having mutations in *ABCB11* (BSEP deficiency), two patients with mutation in *ATP8B1* (FIC1 deficiency) and two patients with biliary atresia (BA). The reactions were performed in triplicate. Gene expression was normalised using *GAPDH* as internal reference gene, and quantified as described in section 3.3.3. Statistical analysis was performed using nonparametric one way ANOVA test (Kruskal-Wallis test) with Dunn's correction for multiple comparisons as post hoc analysis. Initially, the gene expression of the three members of the tight junction protein family (*TJP*) was investigated (Figure 3.4.7). The statistical analysis showed significant differences amongst the five groups of subjects studied in *TJP1* and *TJP2* genes expression levels, but not in *TJP3*. Multiple comparisons identified a strong difference (p-value <0.05) to be present only for the level of *TJP2* expression between the ZO-2 deficiency patients and the BSEP deficiency patients. Regarding the ZO-2 deficiency patients, it seemed that the drastic down-regulation

of *TJP2* level, demonstrated in sections 3.4.1 and 3.4.2, was associated with an increase of *TJP1* but not *TJP3*; though *TJP1* up-regulation was detected also in the other cholestatic conditions.

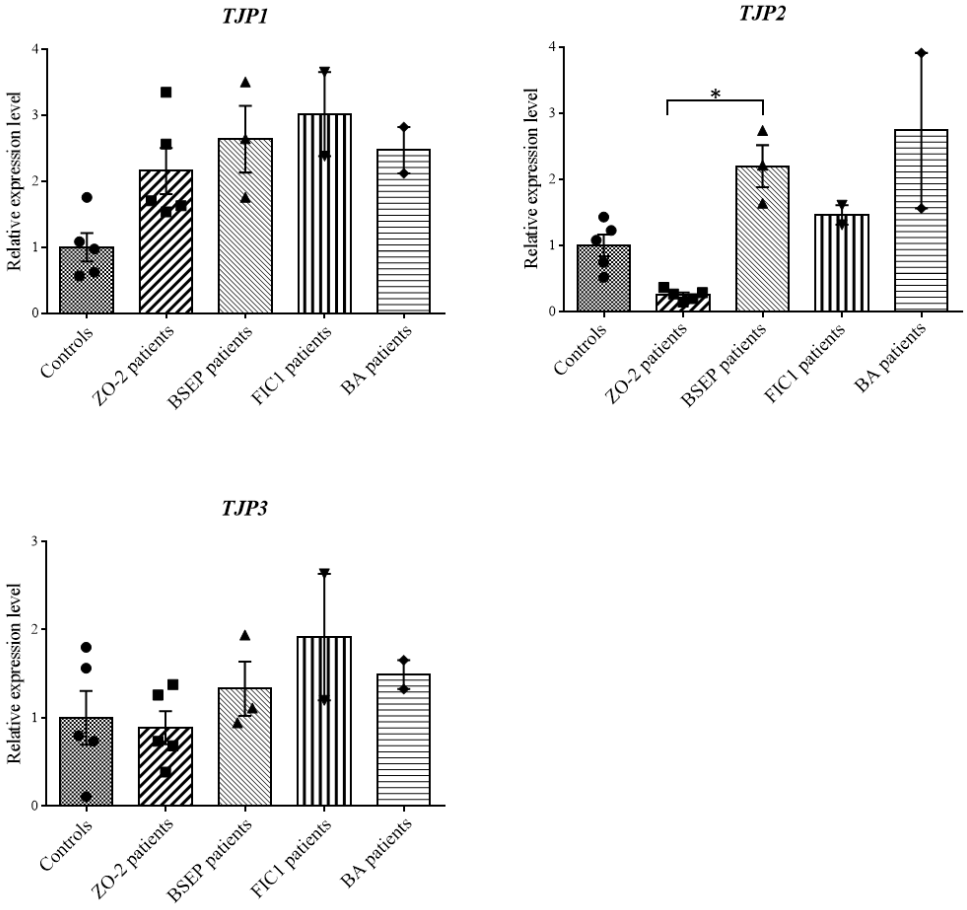

Figure 3.4.7 Expression level of tight junction protein genes in ZO-2 deficiency patients compared to healthy and pathological controls

Quantification of *TJP1*, *TJP2* and *TJP3* in controls (n=5), ZO-2 deficiency patients (n=5), BSEP deficiency patients (n=3), FIC1 deficiency patients (n=2) and patients affected with biliary atresia (n=2). *GAPDH* was used as internal reference gene. The bars represents the means ± standard error of the mean (SEM). Statistics was carried out using one way ANOVA test (Kruskal-Wallis test) with Dunn's correction for multiple comparisons as post hoc analysis. * p-value <0.05; ** p-value <0.01.

Subsequently, other genes involved in the tight junction structure were selected and analysed (Figure 3.4.8). Genes encoding integral tight junction proteins, such as claudin-1 (*CLDN1*), occludin (*OCLN*), JAMA (*F11R*) and tricellulin (*MARVELD2*), were included along with the cytoplasmic protein cingulin (*CGN*) and β-actin (*ACTB*). The Kruskal-Wallis test with Dunn's correction for multiple comparisons as post hoc analysis showed a variability between the control group and the BSEP deficiency patients for the expression level of *OCLN* (p-value <0.05), *F11R* (p-value <0.01), *MARVELD2* (p-value < 0.05) and *CGN* (p-value <0.01).

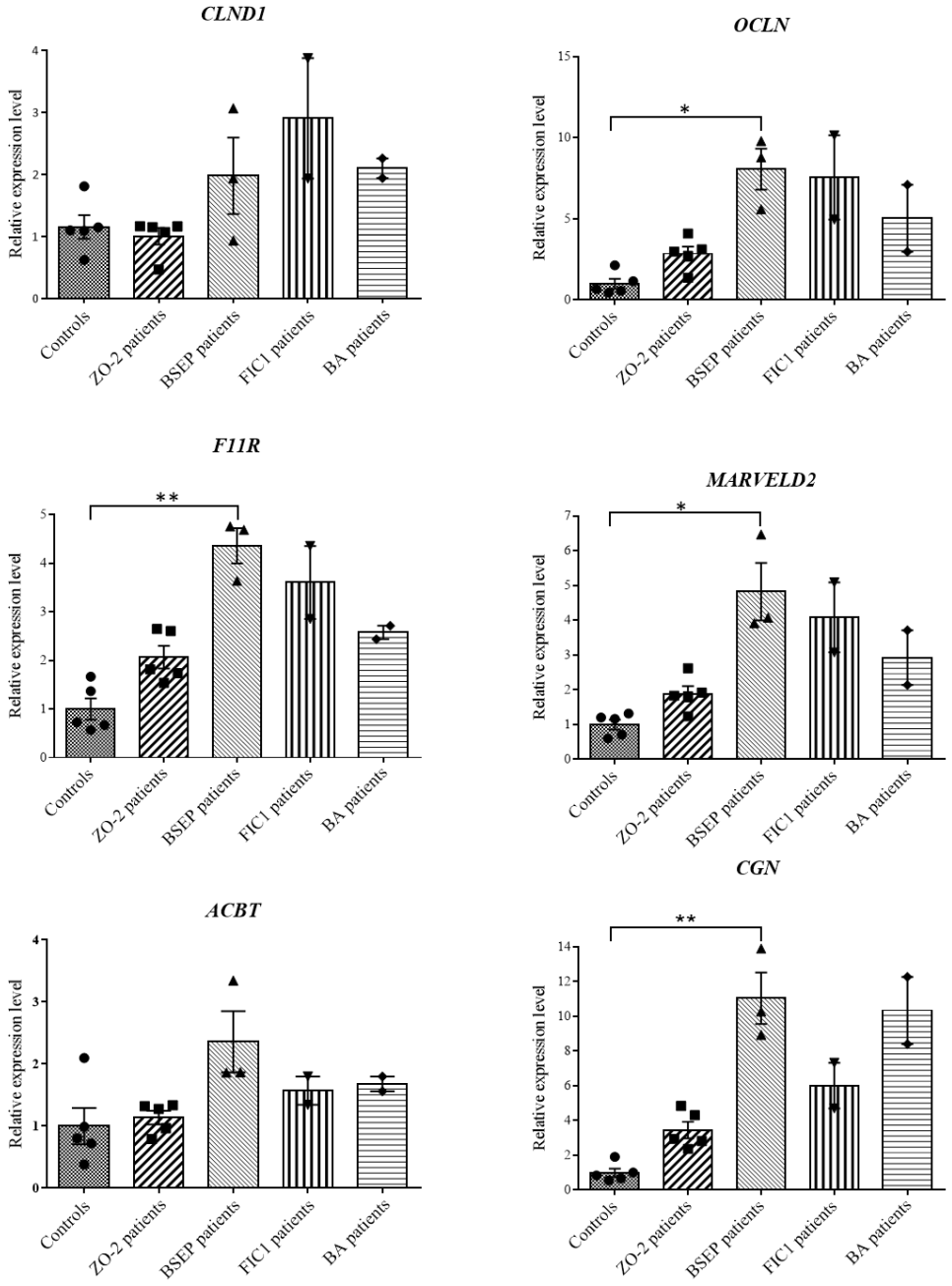

Figure 3.4.8 Expression level of genes involved in the tight junctions in ZO-2 deficiency patients compared to healthy and pathological controls

Quantification of *CLDN1*, *OCLDN*, *F11R*, *MARVELD2*, *ACTB* and *CGN* was performed in controls (n=5), ZO-2 deficiency patients (n=5), BSEP deficiency patients (n=3), FIC1 deficiency patients (n=2) and patients affected by biliary atresia (n=2). *GAPDH* was used as internal reference gene. The bars represents the means ± standard error of the mean (SEM). Statistics was carried out using one way ANOVA test (Kruskal-Wallis test) with Dunn's correction for multiple comparisons as post hoc analysis. * p-value <0.05; ** p-value <0.01.

As described in section 3.1.2, several claudins (*CLDN*) are present in humans and their expression varies from organ to organ. In this study, claudin-2, -3, -7, -10, -11, -12, -14 and -15 were evaluated (Figure 3.4.9 and Figure 3.4.10). Claudin-4, -5, -6, -8 and -9 were initially included in the gene panel, however no expression was observed in the healthy donor control group. Statistical tests were performed to search for evidence of variability amongst the groups of subjected studied for the expression of *CLDN2*, *CLDN7*, *CLDN11*, *CLDN12*, *CLDN14* and *CLDN1*5. Using multiple comparison post hoc analysis, with Dunn's correction, a difference was identified to be present only between the control group and the BSEP deficiency patients for *CLDN12* (p-value <0.05), *CLDN14* (p-value <0.05) and *CLDN15* (p-value <0.05) gene expression (Figure 3.4.10). In addition, *CLND11* gene expression was differently expressed between the control group and the BSEP deficiency patients (p-value <0.05), and between the control group and the ZO-2 deficiency patients (p-value <0.05). With regard to *CLDN2* expression, the post hoc analysis test identified a difference between patients with biliary atresia (BA) and healthy donors (p-value <0.05) (Figure 3.4.9).

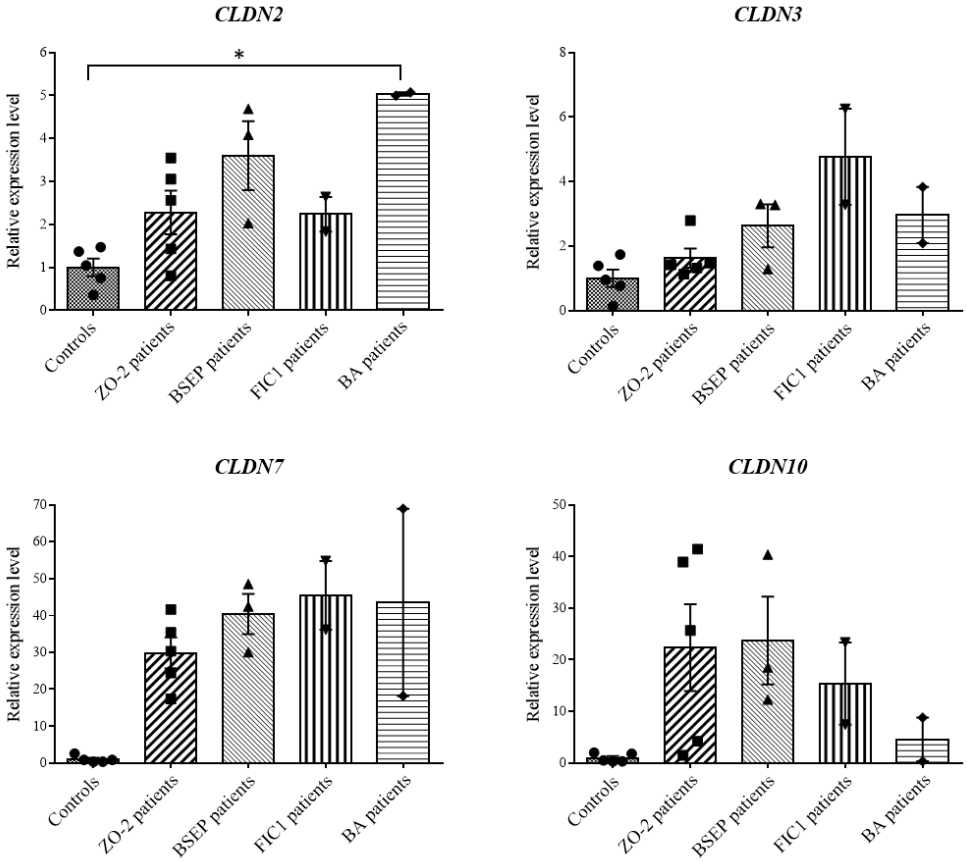

Figure 3.4.9 Expression level of claudin genes in ZO-2 deficiency patients compared to healthy and pathological controls

Quantification of *CLDN2*, *CLDN3*, *CLDN7* and *CLDN10* in controls (n=5), ZO-2 deficiency patients (n=5), BSEP deficiency patients (n=3), FIC1 deficiency patients (n=2) and patients affected by biliary atresia (n=2). *GAPDH* was used as internal reference gene. The bars represent the mean ± standard error of the mean (SEM). Statistics was carried out using one way ANOVA test (Kruskal-Wallis test) with Dunn's correction for multiple comparisons as post hoc analysis. * p-value <0.05; ** p-value <0.01.

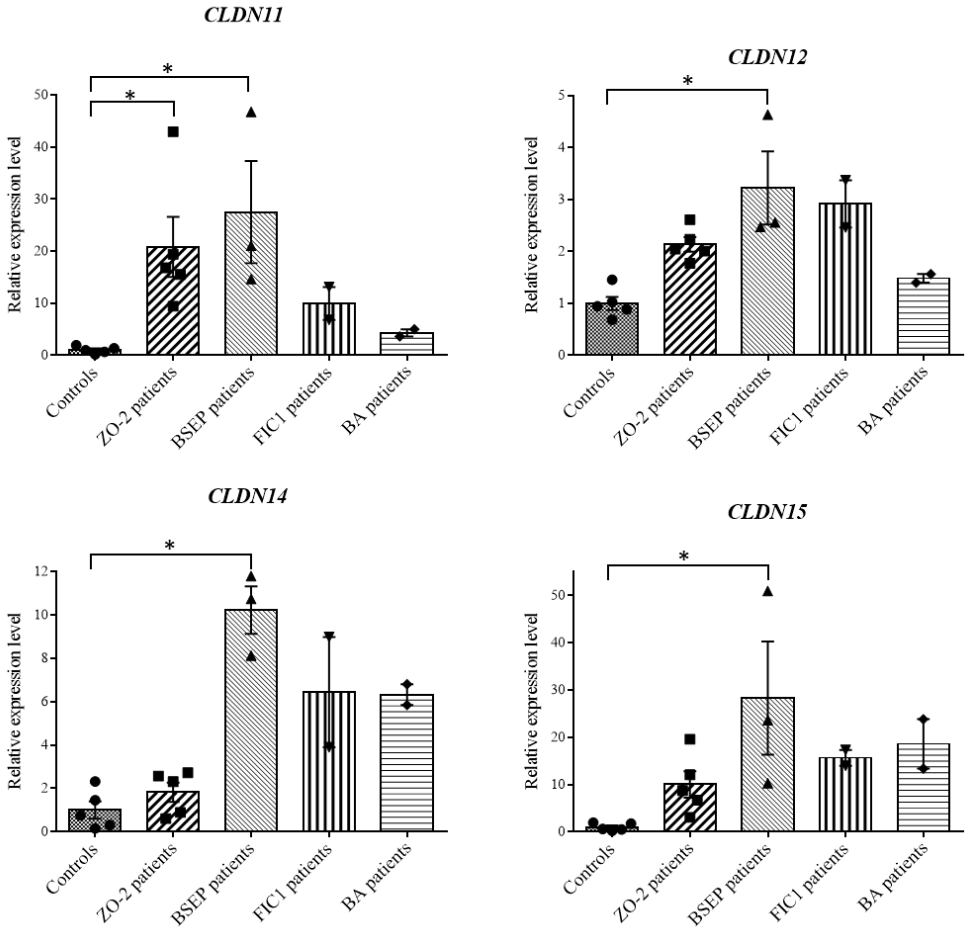

Figure 3.4.10 Expression level of additional claudin genes in ZO-2 deficiency patients compared to healthy and pathological controls

Quantitation of *CLDN11, CLDN12, CLDN14* and *CLDN15* in healthy controls (n=5), ZO-2 deficiency patients (n=5), BSEP deficiency patients (n=3), FIC1 deficiency patients (n=2) and patients affected by biliary atresia (n=2). *GAPDH* was used as internal reference gene. The bars represent the mean ± standard error of the mean (SEM). Statistics was carried out using one way ANOVA test (Kruskal-Wallis test) with Dunn's correction for multiple comparisons as post hoc analysis. * p-value <0.05; ** p-value <0.01.

The role of ZO-2 in the structure and formation of adherens junctions has been established (Tsukita *et al.*, 2009). Similar to tight junctions, adherens junctions are composed of integral membrane proteins, mainly cadherins (*CDH*). They are important in anchoring two adjacent cells, and have a cytoplasmic plaque mostly composed of α and β catenins (*CTNNA1* and *CTNNB1*). This cytoplasmic plaque is

involved in mediating the connection between the transmembrane proteins and the actin cytoskeleton. Gene expressions were evaluated in the control group, in the patients with mutations in *TJP2* and in three groups of other infantile cholestatic conditions (Figure 3.4.11). Differential expression was identified between the control group and the BSEP deficiency patients for *CDH1* (p-value <0.05) and *CTNNB1* (p-value <0.05), after multiple comparison testing.

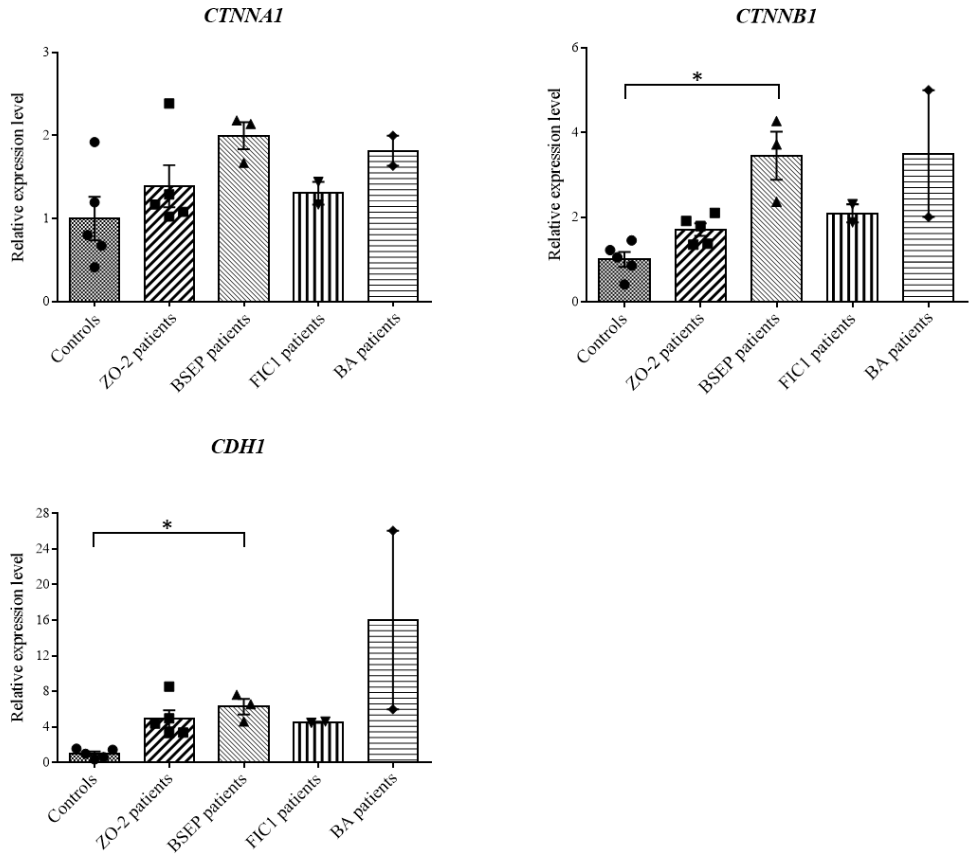

Figure 3.4.11 Expression level of genes encoding adherens junctional proteins in ZO-2 deficiency patients compared to healthy and pathological controls

Quantification of *CTNNA1*, *CTNNB1* and *CDH1* performed in controls (n=5), ZO-2 deficiency patients (n=5), BSEP deficiency patients (n=3), FIC1 deficiency patients (n=2) and patients affected by biliary atresia (n=2). *GAPDH* was used as internal reference gene. The bars represent the mean ± standard error of the mean (SEM). Statistics was carried out using one way ANOVA test (Kruskal-Wallis test) with Dunn's correction for multiple comparisons as post hoc analysis. * p-value <0.05; ** p-value <0.01.

In hepatocytes, as well as endothelial cells, gap junctions are involved in the transcellular communication between two neighbouring cells allowing the exchange of solutes, ions and water (Kojima *et al.*, 2003). Connexin 43 (*GJA1*) has been demonstrated to directly bind the second PDZ domain of ZO-2 via its carboxyl-terminal region, as described in section 3.1.2. In primary cultures of rat hepatocytes, connexin 26 (*GJB2*) and connexin 32 (*GJB1*) are highly expressed; co-localisation at the apical surface of rat hepatocytes have been observed only for connexin 32 and tight junction proteins, including ZO-1, ZO-2, occludin and claudin-1, suggesting connexin 32 is particularly involved in cell polarisation (Kojima *et al.*, 2001). Unfortunately, it was not possible to design suitable sets of primers or probes for the encoded gene *GJB1*. The gene expression of the additional two connexins present in liver was analysed and statistical difference was found (Figure 3.4.12). For *GJA1* expression level, a significant difference was found between the ZO-2 deficiency patients and the control group (p-value <0.05); while for *GJB2* expression, it was found between the BSEP deficiency patients and the control group (p-value <0.05).

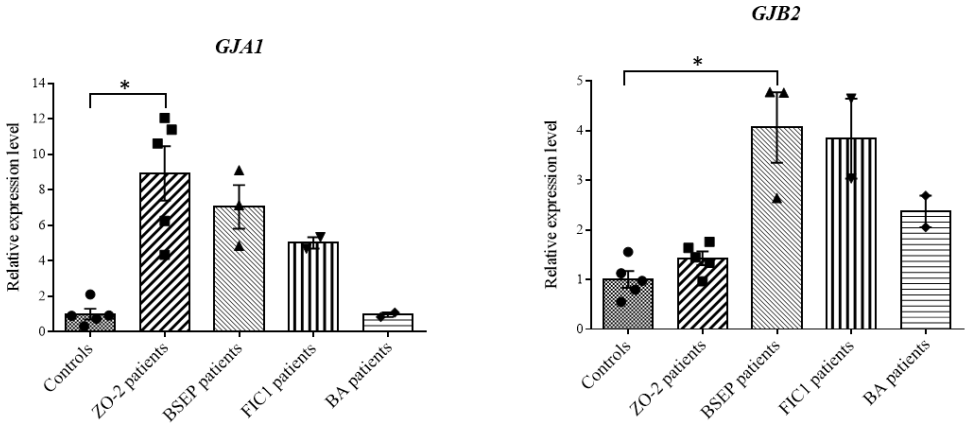

Figure 3.4.12 Expression level of connexin genes in ZO-2 deficiency patients compared to healthy and pathological controls

Quantification of *GJA1* and *GJB2* in controls (n=5), ZO-2 deficiency patients (n=5), BSEP deficiency patients (n=3), FIC1 deficiency patients (n=2) and patients affected by biliary atresia (n=2). *GAPDH* was used as internal reference gene. The bars represent the mean ± standard error of the mean (SEM). Statistics was carried out using one way ANOVA test (Kruskal-Wallis test) with Dunn's correction for multiple comparisons as post hoc analysis. * p-value <0.05; ** p-value <0.01.

As described in section 3.1.2, ZO-2 does not function only as a structural component. In fact, it has a role in transcriptional regulation after binding, in the nucleus, to the DNA-binding protein SAFB. The intracellular trafficking of ZO-2 is mediated by vesicle formation, in which SNX27 and the small GTPase RAB13 play an important role (Marzesco *et al.*, 2002; Zimmerman *et al.*, 2013). Their gene expression was then evaluated and a difference was identified in *RAB13* between controls and BSEP deficiency patients (p-value <0.05) (Figure 3.4.13).

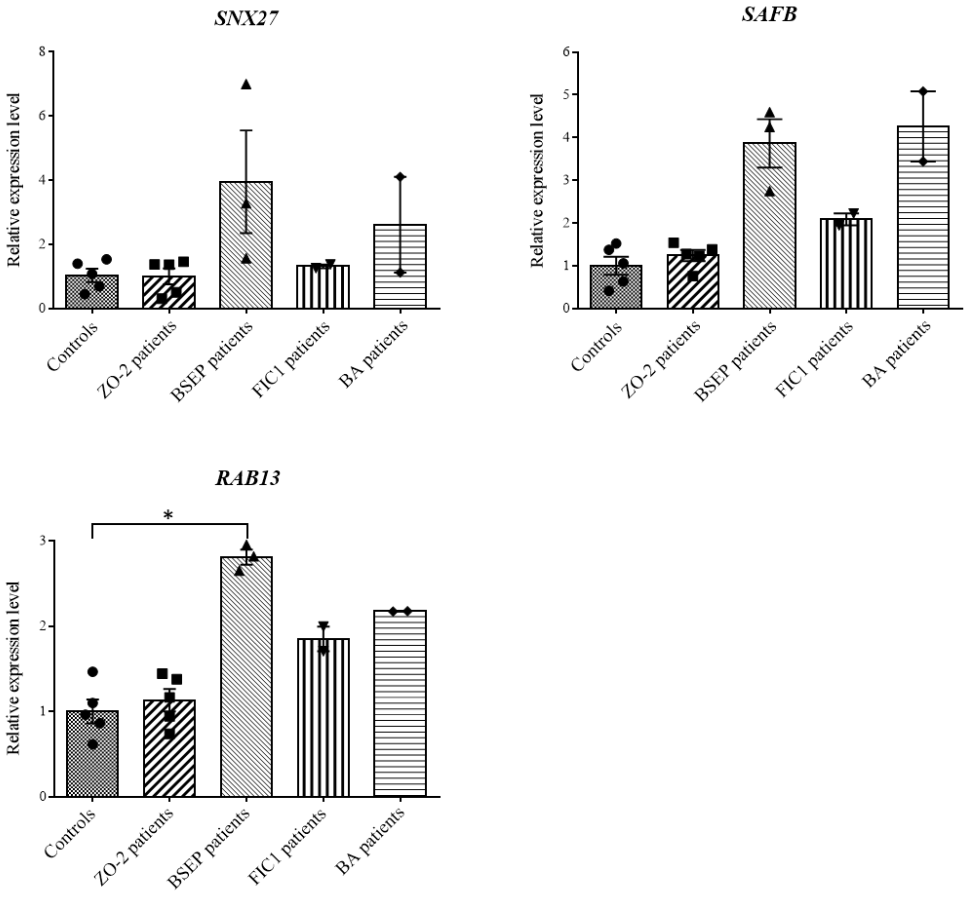

Figure 3.4.13 Expression level of genes involved in non-structural function of ZO-2 in ZO-2 deficiency patients compared to healthy and pathological controls

Quantification of *SNX27*, *SAFB* and *RAB13* in controls (n=5), ZO-2 deficiency patients (n=5), BSEP deficiency patients (n=3), FIC1 deficiency patients (n=2) and patients affected by biliary atresia (n=2). *GAPDH* was used as internal reference gene. The bars represent the mean ± standard error of the mean (SEM). Statistics was carried out using one way ANOVA test (Kruskal-Wallis test) with Dunn's correction for multiple comparisons as post hoc analysis. * p-value <0.05; ** p-value <0.01.

The gene expression analysis was performed for a panel of genes encoded different junctional components with known or possible association to ZO-2. A variety of genes were identified as being up-regulated in BSEP deficiency patients compared to the control group. These genes included: i) four members of the claudin family (*CLDN11*, *CLDN12*, *CLDN14*, *CLDN15*); ii) occludin (*OCLDN*), junctional adhesion molecule-A (*F11R*) and tricellulin (*MARVELD2*), encoding three tight

junction integral proteins; iii) the cytosolic tight junction protein cingulin (*CGN*); iv) two components of the adherens junction: the integral membrane cadherin (*CDH1*) and the cadherin-associated protein β-catenin (*CTNBB1*); v) the gene encoding the gap junction connexin 26 (*GJB2*); vi) a member of the Ras-associated small GTPase (*RAB13*). The biological relevance of these alterations is still not clear. However, these data showed that an overall up-regulation was observed in every cholestatic condition. Different claudin-2 expression was identified between patients with biliary atresia (BA) and healthy donors. As only two samples were included within this disease group, this finding would be confirmed only after analysing a greater cohort. In the patients with absence of ZO-2, up-regulation of claudin-11 (*CLDN11*) and the gap junction connexin 43 (*GJA1*) was found.

3.5 Conclusions of the functional studies

In the first part of this thesis, several patients were identified as having mutations in *TJP2*, most of which caused an alteration of the reading frame and consequently the generation of premature terminator codons (PTC). An important biological quality control system has evolved in eukaryotic cells for the degradation of these erroneous transcripts, named nonsense mediated mRNA decay (NMD) (Nicholson *et al.*, 2010). This mechanism prevents the translation of PTC-containing mRNAs into truncated proteins, which can result in lack of function, partially conserved function, and dominant-negative effect or even with a new gained activity. On one side NMD has the beneficial role of eliminating deleterious proteins, but on the other side it has the negative effect of potentially degrading proteins with residual function. Therefore, the association of NMD with genetic diseases can be highly variable. The investigation of the consequences of *TJP2* protein-truncation was undertaken initially by evaluating the quantitative expression of mRNA. The analysis showed a mean reduction of approximately 80% in patients with *TJP2* mutations compared to the control group of healthy liver donors, suggesting the activation of the nonsense mediated mRNA decay pathway. The same result was obtained for all the three different priming approaches utilised. The mutations affected the different *TJP2* transcripts. At protein level, after several attempts due the difficulties in obtaining a suitable antibody, the analysis of the ZO-2 protein identified, not surprisingly, the absence of the both protein isoforms by western blotting.

Along with ZO-1, ZO-2 is the main protein in the cytoplasmic plaque of tight junction structures (Bauer *et al.*, 2010). As a scaffolding protein, it connects the cytoskeleton of actin to transmembrane proteins, such as claudins, which in turn

bind opposite transmembrane proteins, closing the paracellular space between two adjacent cells. Claudin expression varies amongst organs; in the liver the most representatives are claudin-1 and claudin-2 (Mitic *et al.*, 2000). In patients with no expression of ZO-2, the protein expression levels of those claudins were studied. Claudin-1 and claudin-2 showed no alteration in the expression level compared to healthy and pathological controls. This analytical technique has established that the absence of ZO-2 does not interfere with the quantitative expression of claudin proteins; however, it does not exclude other possible biological effects. Further investigations will be discussed in the chapter 4.

Tight junction protein ZO-2 has been also demonstrated to have both structural and non-structural characteristics (section 3.1.2). Along with the different junctional components shown to directly interact with ZO-2, other proteins have been hypothesised to have a similar direct link, suggested by their homology with the tight junction protein ZO-1, the most studied member of the same family. In addition, ZO-2 has been shown to translocate into the nucleus and to be involved in chromatin remodelling and transcriptional regulation. Therefore, in the patients where the absence of ZO-2 has been previously identified, it was hypothesised that there might be a possible activation of a compensatory effect on the expression level of the other constituents. The gene expression was also studied in a group of disease control patients affected by other inherited forms of cholestatic liver disorders that manifest with a normal or low serum concentration of GGT, namely BSEP deficiency and FIC1 deficiency, and by another form of progressive cholestasis: biliary atresia (BA). A total number of 24 related genes was selected and analysed. The reduced expression of *TJP2* in the mutated patients appears to cause an increase in the expression of the homologous *TJP1*; this finding was however not specific for this group of patients. Increased expression of *CLDN11* and *GJA1* genes were also identified in this group of patients. *CLDN11* encodes the transmembrane tight junction protein claudin-11; neurological and reproductive defects were exhibited in the $CLDN11^{-/-}$ knock-out mice (Gow *et al.*, 1999). To date, alteration of claudin-11 in liver diseases has not been identified. In contrast,

connexin 43 is encoded by *GJA1* and it is abundantly expressed in non-parenchymal liver cells (Gonzalez *et al.*, 2002). Increased connexin-43 expression was seen during liver inflammation. Additional studies would be essential to understand better the biological consequence of these results and possible association with the absence of ZO-2.

In the three disease control groups, a general alteration was however observed, suggesting that genes involved in junctional complexes could be differentially expressed in cholestasis. During cholestatic episodes, structural alteration in the cell junctions of hepatocytes could have occurred; therefore, the expression of junctional components might be increased in order to maintain the structural integrity. In addition, the small sample size and the high variability within samples might have compromised the findings, so a larger cohort might be necessary for further analysis.

4 Collaborative work

4.1 Immunohistochemical studies

To understand better the overall expression pattern of tight junction proteins, immunohistochemical studies were performed by the liver histopathology laboratory at King's College Hospital, under the supervision of Dr Alex Knisely, as part of collaborative work. Formalin fixed and paraffin-embedded (FFPE) hepatoctomy specimens from individuals with mutations in *TJP2* were subjected to routine immunohistochemical procedures. In addition, FFPE normal liver samples were used as controls. The histology slides from each sample were mounted as 4 μm thick sections and inserted into Leica Bond-Max automated immunostaining system (Leica Biosystem, Newcastle Upon Tyne, UK), where they were dewaxed and rehydrated. Formalin fixation has the limitation of masking epitopes for immunostaining, causing false negative results. Therefore, epitopes were exposed using heat treatment in buffer solution at 100°C for 20-30 minutes. Afterwards, a peroxidase block treatment was added to inhibit the endogenous peroxidase activity and thereby to reduce the non-specific background signal, which can occur when using the horseradish peroxidase (HRP)-conjugated secondary antibodies. Primary antibodies specific for the proteins of interest, such as ZO-2, claudin-1 and claudin-2, were subsequently applied. All primary antibodies used for this study will be described in the following part of this section. Subsequently, the Bond Polymer Refine detection system (DS9800 for Leica Bond-Max automated immunostaining, Leica Biosystem) was applied. This kit included the HRP-conjugated secondary antibody against the mouse and rabbit IgG primary antibodies. The detection of the antigen was then visualised using 3,3′-diaminobenzidine (DAB) as HRP substrate, giving a brown stain. Haematoxylin

was finally used as a counterstaining, allowing the visualisation of cell nuclei (blue). In the first place, TJP2/ZO-2 expression was evaluated using the same primary antibody used for protein blotting, raised against the C-terminus of the protein (Table 3.3.9), at dilution of 1:40. The secondary antibody and detection system utilised are in accordance to the description stated in the beginning of this section. The liver tissues were then view and interpreted under the microscope.

For histological studies, the liver is divided into subunits called hepatic lobules; these are hexagonal structures constituted by a central vein in the centre of the lobule and portal tracts at the six vertices. Each portal tract includes a branch of bile duct, a branch of portal vein and a branch of hepatic artery. Bile canaliculi are formed of the apical membranes of adjacent hepatocytes within the parenchyma of the hepatic lobule; they then converge and form small bile ducts located in the portal tract areas. The histological examination of patients' liver specimens showed micronodular cirrhosis of the parenchyma with multiple regenerative nodules; hepatocellular cholestasis was predominant in the lobular area with lymphocytic inflammatory infiltration. Immunohistochemical staining for tight junction zona occludens 2 protein in normal liver tissue, showed clear marking at the canalicular margins in the parenchyma, as would be expected for a protein involved in junctional complexes (Figure 4.1.1a and 4.1.1b), and at the apical membrane of the bile duct located in the portal tract (Figure 4.1.1b). In contrast, in patients affected by early-onset severe cholestasis with homozygous protein-truncating mutations in *TJP2* (cases 10a, 10b, 11a, 12a) the marking of the protein was absent both in the parenchyma and in the bile ducts (Figure 4.1.1c-f). Patient 10b showed an intracanalicular accumulation of bile, noticeable by colour different to the dark brown of the immunohistochemical reaction product (Figure 4.1.1d). These findings are in accordance to the previous results showing no expression of ZO-2 protein by western blotting (3.4.3).

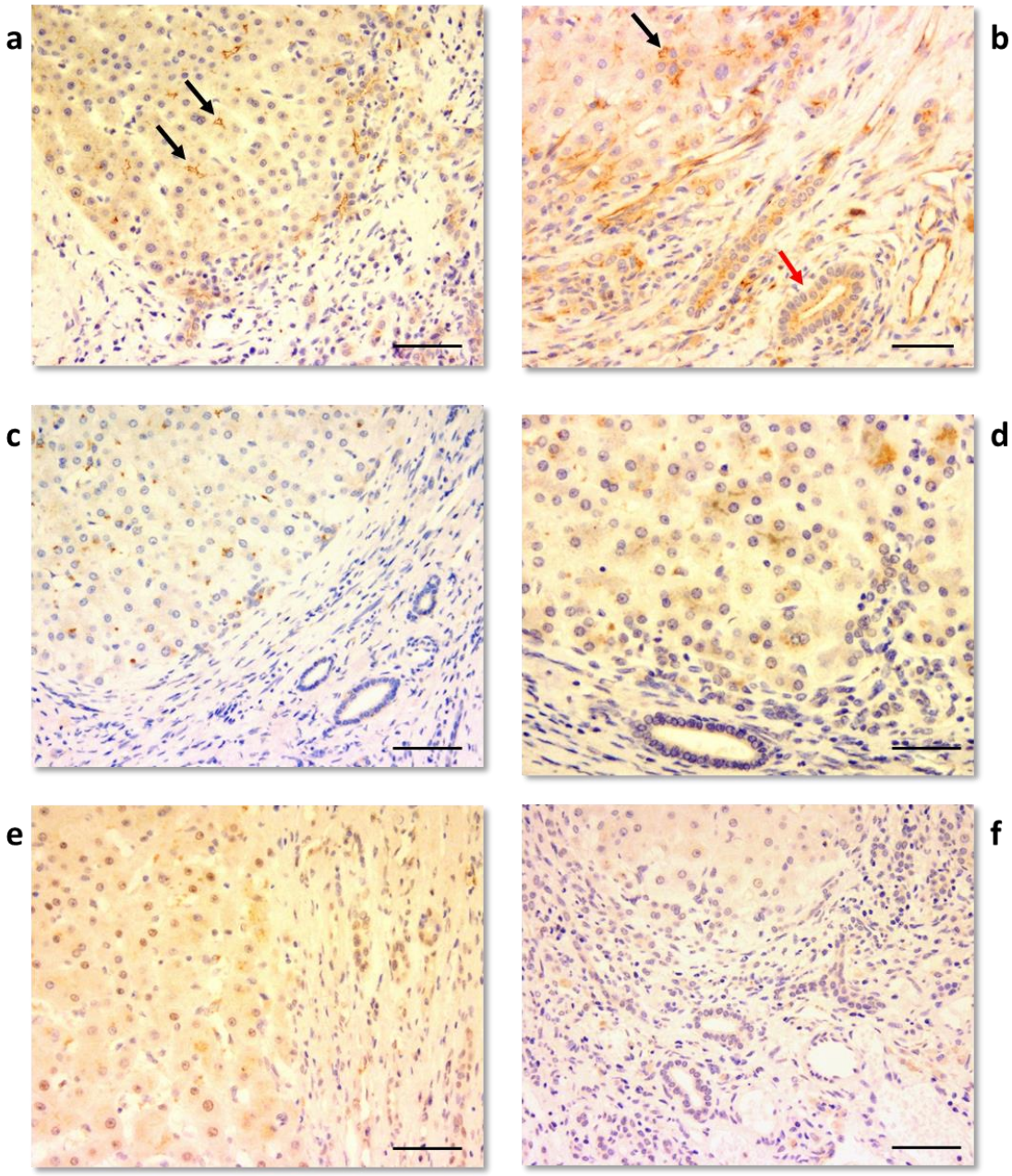

Figure 4.1.1 Immunohistochemical staining for ZO-2 in patients with protein-truncating mutations in *TJP2* and in controls

In normal liver tissue ZO-2 is identified as brown staining. The black arrows highlight few examples of the protein expression at the canalicular margins of the parenchyma (a, b), while the red arrow indicates ZO-2 staining at the apical membrane of the bile duct located at the portal tract (b). No ZO-2 marking is detected in patient 10a (c), 10b (d), 11a (e), and 12a (f) with protein-truncating mutation in *TJP2* leading severe cholestatic condition. Patient 10b shows an intracanalicular accumulation of bile. Scale bar: 100 µm

The expression of claudin-1 was next examined to investigate a possible alteration of its expression due to the absence of ZO-2. The same primary antibody used for the protein blotting was adopted (Table 3.3.10) with a dilution of 1:100. Secondary antibody and detection system were as described above. In the normal liver tissue, claudin-1 immunostaining showed clear marking at the canalicular borders and at the bile ducts (Figure 4.1.2a). In liver tissue of the patients 10a (Figure 4.1.2Figure 3.4.5b) and 12a (Figure 4.1.2e) a significant reduction of protein expression was observed both at the canalicular membranes in the parenchyma of the hepatic lobule and at the membranes of the bile ducts, whilst in liver tissues of the patients 10b (Figure 3.4.5c) and 11a (Figure 4.1.2d) claudin-1 expression level was clearly decreased only in the parenchyma. In addition, in patient 10b diffuse cytoplasmic marking was evident (Figure 4.1.2c). These results suggest that the absence of ZO-2 could affect the localisation of claudin-1 at the canalicular membrane. At the cholangiocyte borders, however, the variability in claudin-1 expression has no clear explanation; secondary events, such as inflammation, could have possibly contributed to the damage of the membrane in some patients rather than others.

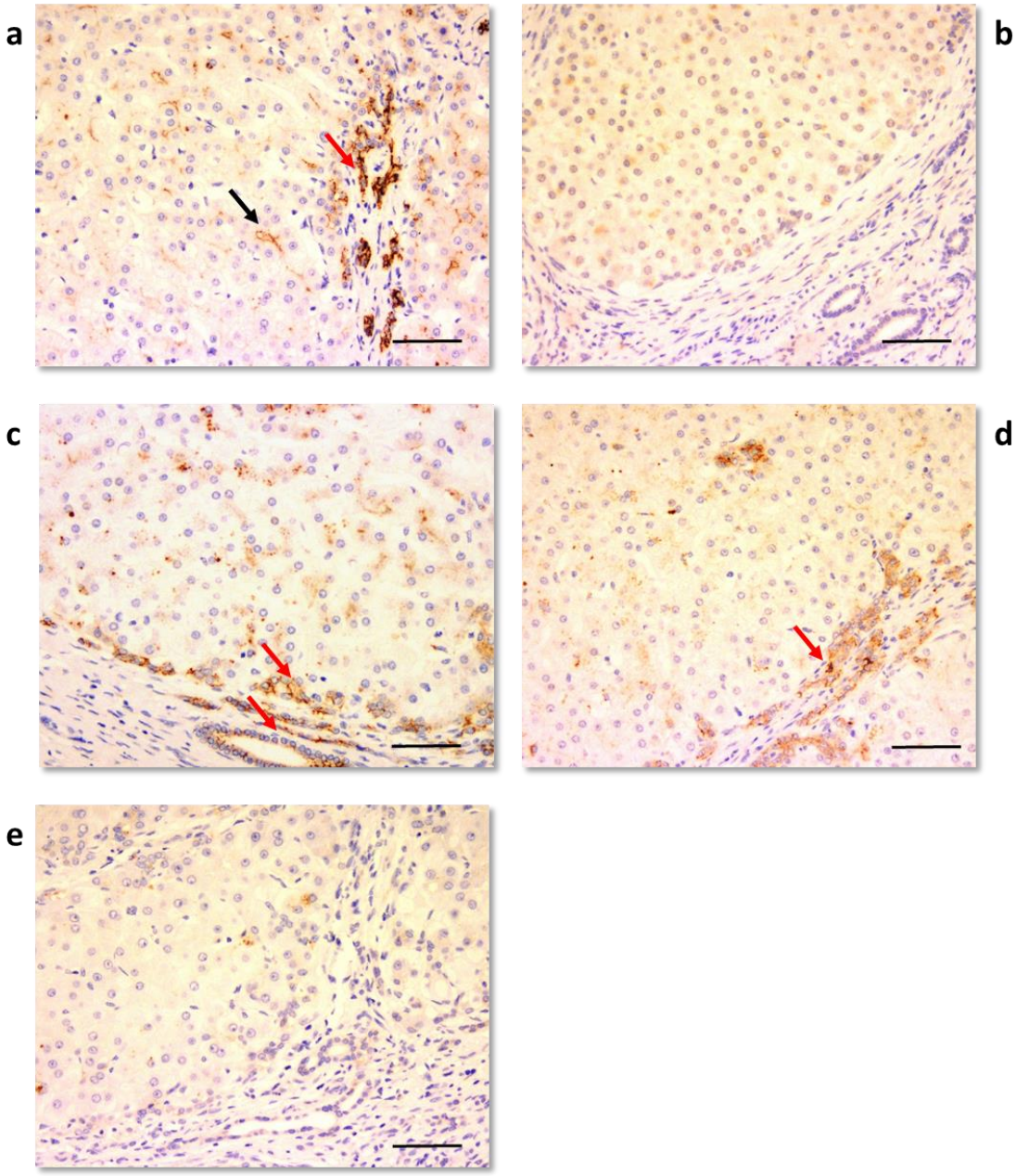

Figure 4.1.2 Immunohistochemical staining for claudin-1 in patients with protein-truncating mutations in *TJP2* and control

Claudin-1 expression is marked by brown staining. In normal liver tissue (a), claudin-1 is present at the canalicular membranes in the hepatic parenchyma (black arrow) and in the bile ducts (red arrow). In patient 10a (b) and patient 12a (e), the expression of claudin-1 is observed to be drastically reduced in every area of the liver tissue examined, whilst in patient 10b (c) and patient 11a (d) only at the canalicular membranes in the parenchyma. Bile ducts expression in patients is indicated by red arrows (c-d). Patient 11a also shows intense intracellular accumulation of claudin-1. Scale bar: 100 µm.

The second member of the claudin family was also studied. The same primary antibody used for immunoblotting (Table 3.3.10) was used at a 1:16,000 dilution. The presence of claudin-2 was detected by a brown staining using the technique described above. Claudin-2 is one of the main integral tight junction proteins present in the liver tissues that, at variance from claudin-1, is expressed within the cytoplasm along the pericanalicular margins (Holczbauer *et al.*, 2013). To distinguish the bile canaliculi, an anti-BSEP rabbit polyclonal primary antibody (HPA019035, Sigma-Aldrich) at a dilution of 1:1,500 was also used and detected with a red chromogen (Bond Polymer Refine Red detection system, DS9390 for Leica Bond-Max automated immunostaining, Leica Biosystem). As shown in Figure 4.1.3, claudin-2 was expressed both in liver tissues of the normal control and patient 11b with protein-truncating mutation in *TJP2*. Dilatation of bile canaliculi was observed in the patient; however, this is a histological characteristic of severe cholestasis. This finding indicted that in patient lacking ZO-2 protein there is no alteration in the expression of claudin-2.

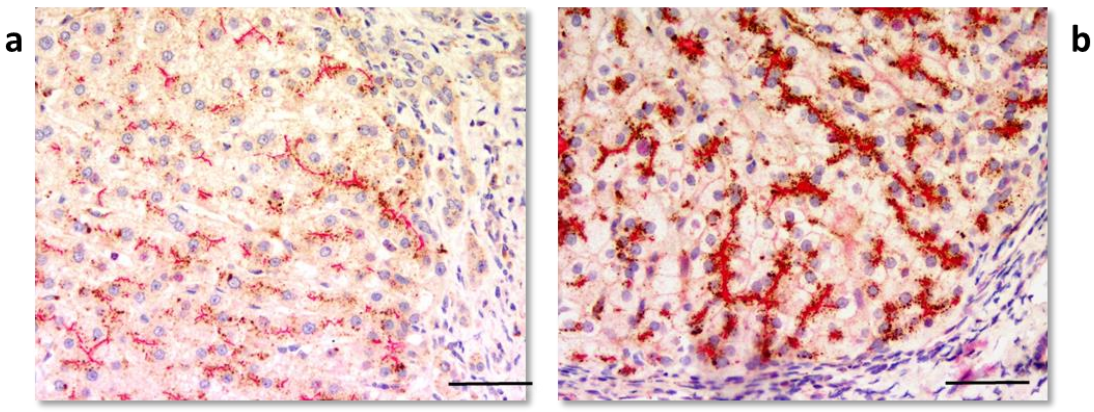

Figure 4.1.3 Immunohistochemical staining for claudin-2/BSEP

Claudin-2 expression (brown staining) is evaluated in the liver tissues of a normal control (a) and of patient 11a (b) with severe cholestatic disease and homozygous protein-truncating in *TJP2*. Pericanalicular marking is evident in liver parenchyma of both patient and control. BSEP (red staining) was co-stained to identify bile canaliculi. Dilatation of bile canaliculi in liver tissue of the patient represents a histological sign of severe cholestasis. Scale bar: 100 µm.

Subsequently, the expression of ZO-2 was investigated in the liver of the three patients (patient 19, 35 and 75 in Table 2.4.9), having the same homozygous missense mutation in *TJP2* (c.2363A>T; p.His788Leu). Phenotypically, they presented with remittent cholestasis at different ages of onset. The FFPE liver tissues were available from liver needle biopsies performed during cholestatic episodes. Mild centrilobular fibrosis was evident in all the three patients' specimens. ZO-2 staining was undertaken following the procedure described above. ZO-2 protein staining revealed different expression patterns in different patients (Figure 4.1.4). Whilst a lack of ZO-2 expression was identified in patient 19 at the canalicular membranes in the parenchyma of the hepatic lobule, as seen in the affected individuals with protein-truncating mutations (Figure 4.1.4b), patient 35 and 75 showed a reduction in the canalicular margin staining (Figure 4.1.4c-d). However, nuclear localisation of ZO-2 was evident in all three patients, dramatically so in patient 35 and patient 75.

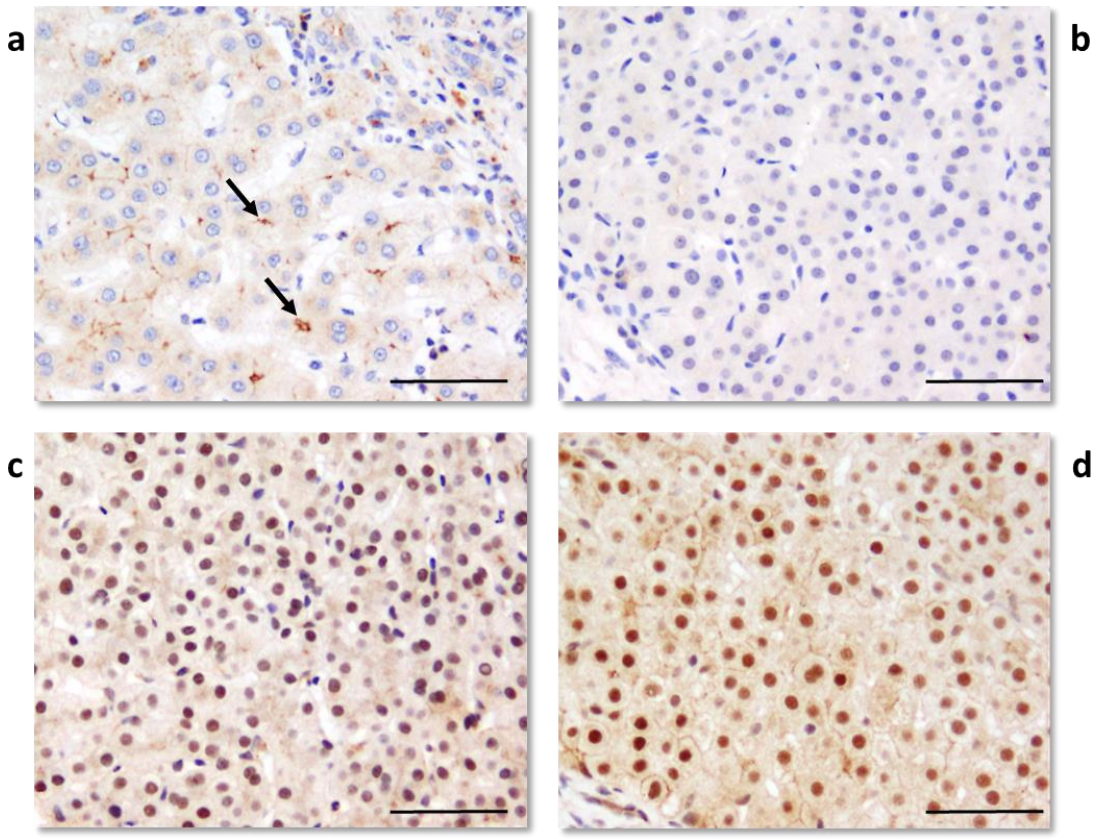

Figure 4.1.4 Immunohistochemical staining for ZO-2 in patients with missense mutation (p.His788Leu) in *TJP2* and control

ZO-2 expression (brown staining) is studied in liver tissues of control (a) and patient 19 (b), patient 35 (c) and patient 75 (d) during cholestatic episodes. Black arrows highlight the ZO-2 expression in some canalicular membranes in control liver tissue. In patient 19, absence of ZO-2 expression is evident at the canalicular margins; patient 35 and 75 show reduced protein expression. Nuclear staining is present in all three cases, with greater intensity in patient 35 and 75. Scale bar: 100 µm.

Lastly, the expression of claudin-1 was investigated in the liver of the same three patients. The sample preparation and the immunohistochemical procedure were as described above. While in the normal liver claudin-1 expression was localised at the canalicular borders and at the membrane of bile ducts (Figure 4.1.5a), marked cytoplasmic accumulation of the protein was identified in all three patients (Figure 4.1.5b-d). Therefore, although in these three patients with the same homozygous

missense mutation ZO-2 staining showed variable patterns, the consequence of its alteration appears to be the same with respect to claudin-1.

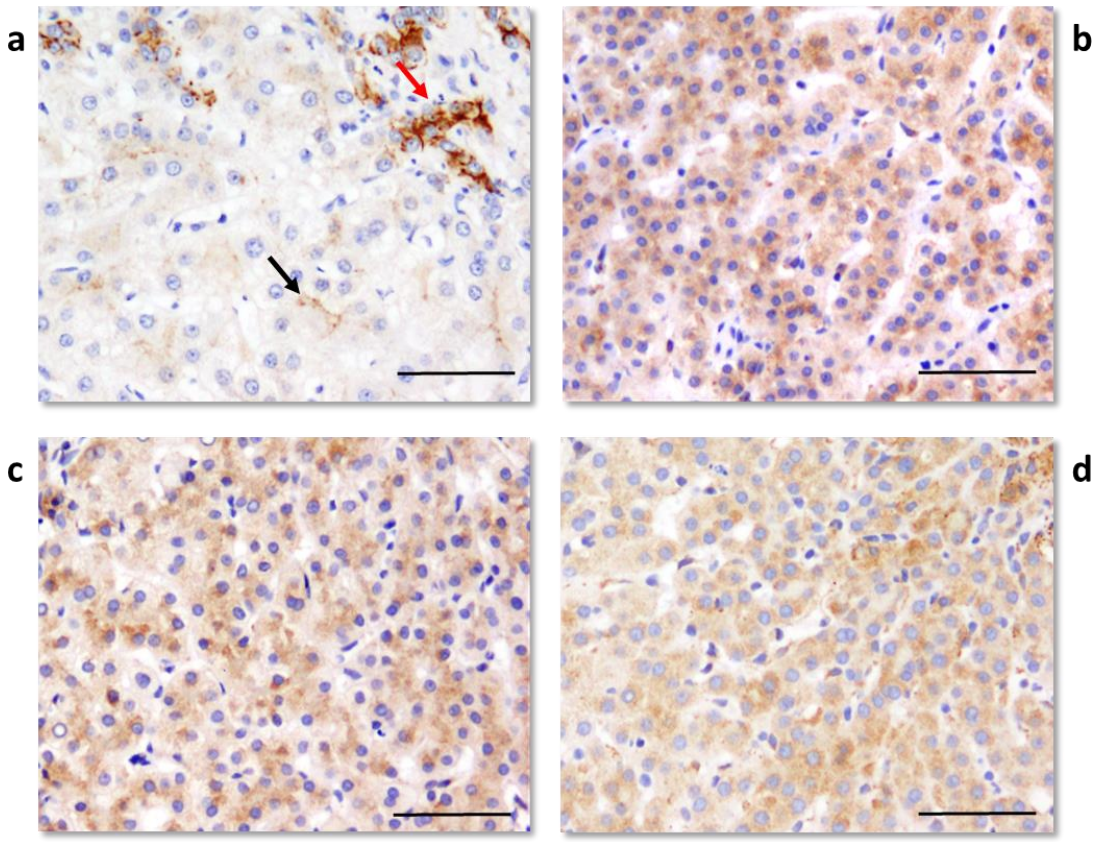

Figure 4.1.5 Immunohistochemical staining for claudins-1 in patients with missense mutation (p.His788Leu) in *TJP2* and control

The expression of claudin-1 (brown staining) in the liver tissues of a normal control (a) and of patients 19 (b), 35(c) and 75 (d). In the control samples (a), the red arrow indicates the claudin-1 expression in the bile duct area; the black arrow indicates an example of staining at the canalicular membrane. All patients have the same homozygous missense mutation in *TJP2* and were affected by remittent cholestasis. Cytoplasmic diffusion of claudin-1 is evident in the hepatic parenchyma of the three patients. Scale bar: 100 µm.

4.2 Transmission electron microscopy studies

The ultrastructure of tight junction was investigated through transmission electron microscopy (TEM), performed by Dr Bart E. Wagner in the histopathology Department at Royal Hallamshire Hospital in Sheffield, United Kingdom. The liver tissues were prepared in accordance to the following procedure: the samples were fixed in paraformaldehyde and glutaraldehyde, post-fixed in osmium tetroxide (OsO_4), dehydrated in ethanol and then embedded in resin. Ultrathin sections of 70-80 nm thickness were prepared for the TEM observation with uranyl acetate and lead citrate post-staining. Bile canaliculi of normal liver were examined initially (Figure 4.2.1). The cell-cell junctions of adjacent hepatocytes were identified by intense electron-dense plaques that sealed the paracellular space and created the canalicular space. Tight junctions were localised close to the apical surface of the canaliculus.

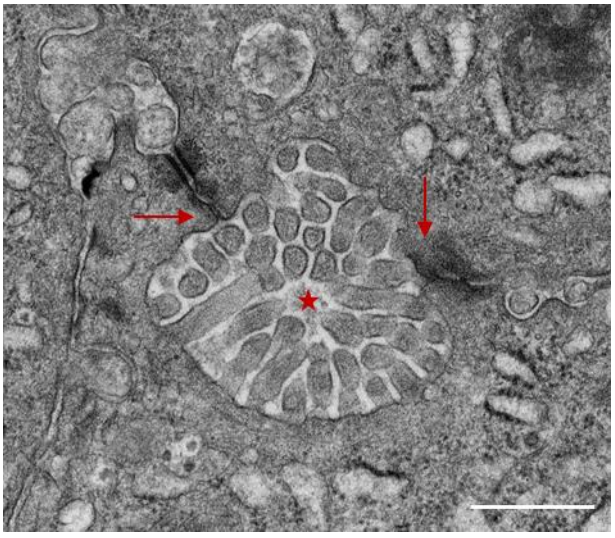

Figure 4.2.1 Transmission electron microscope image of a bile canaliculus of the liver

The bile canaliculus from a normal liver is highlighted by a red star. Tight junctions are indicated by red arrows. Scale bar: 500 nm

The analysis of tight junctions was undertaken in three groups of patients: two individuals (patient 2 and 11b) with severe cholestatic liver disease and protein-truncating mutations in *TJP2*; two (patient 35 and patient 75) with remittent cholestasis now resolving and with the same homozygous missense mutation in *TJP2*; one BSEP deficiency patient and one FIC1 deficiency patient, included in cholestatic disease control groups (Figure 4.2.2). As shown in Figure 4.2.1, normal tight junction appeared condensed in an electron-dense area close to the apical surface of the canalicular membrane. The same structure appeared to be preserved in disease control patients (Figure 4.2.2e-f), while all patients with mutations in *TJP2*, both those with severe cholestasis (Figure 4.2.2a-b) or a milder phenotype (Figure 4.2.2c-d), showed elongated junction structures and dispersed electron-dense area along the paracellular space.

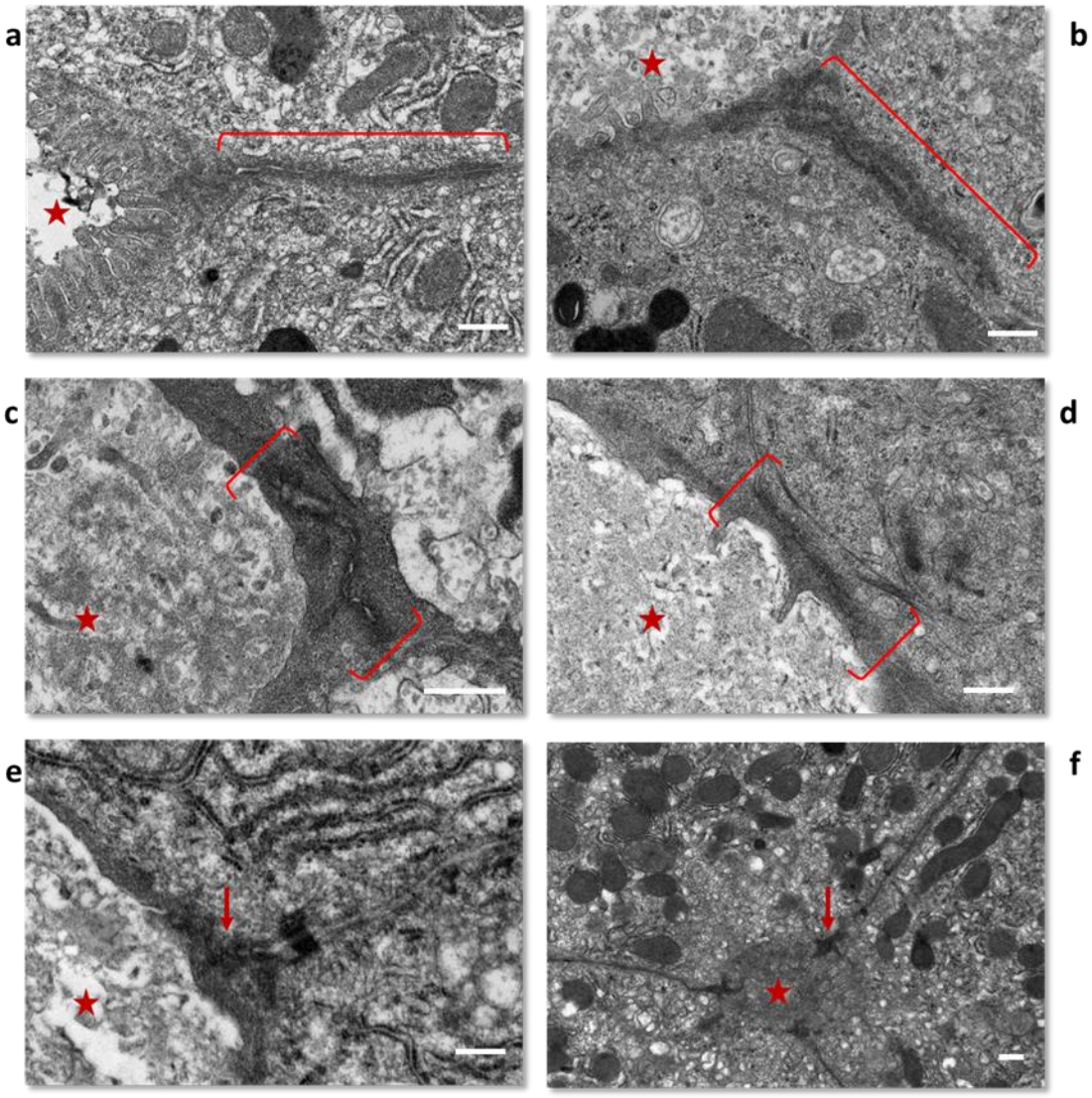

Figure 4.2.2 Transmission electron microscope images of tight junction structures in the liver biopsies

Tight junctional structures are shown in all panels. Bile canaliculi are indicated by red stars. Liver tissue from patients with severe cholestatic conditions having protein-truncating mutations in *TJP2* are in panel a (patient 2) and panel b (patient 11b). In panel c and d are liver biopsies from patient 35 and patient 75 respectively, both having the same homozygous missense mutation in *TJP2* and being affected by remittent cholestasis. Panels e and f show the liver from a BSEP deficiency patient and an FIC1 deficiency patient respectively. The electron-dense structure of tight junctions is indicated by red arrows in panel e and f; in patients elongated junctional structures are highlighted by red brackets in panel a-d. Scale bar: 500 nm

5 General Discussion and future works

The identification of genetic causes of rare Mendelian diseases has rapidly improved after the advent of next-generation sequencing technology (Boycott *et al.*, 2013). It has been estimated that over 115 novel causative genes of autosomal recessive disorders were identified during the initial stage of next-generation sequencing application in genetic research.

Progressive familial intrahepatic cholestasis is a heterogeneous group of rare autosomal recessive diseases, manifesting as an impairment of bile formation or bile transport from the hepatocytes into the bile canaliculi (Davit-Spraul *et al.*, 2009). Previously, mutations in three genes encoding three canalicular membrane transporter proteins (*ABCB11*, *ABCB4* and *ATP8B1)* have been identified as the primary genetic causes of the three major forms of this inherited cholestatic condition: BSEP (bile salt export pump) deficiency, MDR3 (multidrug P-glycoprotein resistance) deficiency and FIC1 (familial intrahepatic cholestasis 1 protein) deficiency. Congenital chronic cholestasis is also one of the features of other rare genetic disorders, such as Alagille syndrome, Dubin-Johnson syndrome and different enzyme deficiencies. As described in the first chapter, bile is composed of many components, including bile acids, cholesterol, bilirubin and toxic/pathogenic metabolites. Several molecular pathways are essential for normal bile synthesis, formation, and for the transport of biliary constituents from the basolateral membrane into the cytoplasm of hepatocytes, and across the apical membrane into the canalicular lumen. Genetic modifications that damage any of these mechanisms, could lead to impairment of bile production and retention of biliary components into the liver and, therefore, liver disease.

Although several rare genetic disorders have been described as being associated with cholestatic liver disease, one third of paediatric patients still remain idiopathic. Therefore, in this thesis, through targeted resequencing and whole-exome sequencing approaches for next-generation sequencing, different areas of the genome were simultaneously interrogated in a cohort of 83 cholestatic patients; one representative subject per family was included. Clinically, the patients manifested with different degrees of cholestasis with normal, or around the upper limit of normal, serum concentration of gamma-glutamyl transferase (GGT) and no mutations in *ABCB11* (BSEP) or *ATP8B1* (FIC1). The evaluation of GGT serum level is routinely used as a biochemical marker of liver disease; within progressive familial intrahepatic cholestasis the serum concentration of GGT is normal and/or low in FIC1 and BSEP deficiencies, and high in MDR3 deficiency (van Mil *et al.*, 2005). However, the recent application of next-generation sequencing technology for the genetic diagnosis of cholestasis has identified individuals with mutations in *ABCB4* (MDR3) with normal serum concentration of GGT, indicating that the utilisation of this biochemical marker will need to be reconsidered. In this study, only patients with normal serum GGT levels were investigated; future work should extend the investigation to all types of idiopathic infantile cholestasis.

5.1 TJP2 deficiency

The high throughput sequencing analysis of the 83 families revealed that 15 had homozygous mutations in the tight junction protein 2 (*TJP2*) gene. Siblings were subsequently Sanger sequenced, and the total of 20 affected children with pathogenic mutations in the gene were identified. All of these patients belonged to consanguineous families. In the majority of them, homozygous mutations were identified causing an alteration of the reading frame and a generation of a premature terminator codon (PTC). The consequent translation of these abnormal mRNA transcripts is likely to have a deleterious effect on the cells. A mechanism of quality

control, named nonsense mediated mRNA decay (NMD), however, has evolved in a wide variety of organisms, including humans, to recognise and eliminate these defectives mRNAs (Frischmeyer & Dietz, 1999). In this study, *TJP2* gene expression in liver tissue was significantly reduced in the group of patients with protein-truncating mutations compared to healthy liver donors, suggesting that these PTC-containing *TJP2* transcripts are subject to nonsense mediated mRNA decay. However, a residue of *TJP2* expression of 20% was still present, leaving open the possibility that a portion of transcripts could have escaped to the RNA surveillance and be translated into truncated *TJP2* encoded ZO-2 proteins. This hypothesis was tested through Western blotting. Despite the difficulties in the identification of an optimal antibody, due to a possible cross-reactivity between similar epitopes, no ZO-2 protein was shown to be expressed in these patients.

As the name suggests, tight junction zona occludens 2 (ZO-2) is a cytoplasmic component of the tight junctions located close to the apical membrane of polarised epithelial cells, such as hepatocytes and cholangiocytes (Bauer *et al.*, 2010). These structures form sealed channels between two neighbouring cells, which regulate the passage of specific molecules and ions through the paracellular space, preventing the entrance of toxic or pathogenic molecules. In normal liver tissues, the expression of ZO-2 was clearly evident along the margins of canaliculi and cholangiocytes. These markings were lost in the liver of patients with protein-truncating mutations.

In the tight junction, ZO-2 is a scaffolding protein; a mediator between the cytoskeleton of actin and the integral tight junction proteins, such as claudins. Therefore, the expression of claudin-1 and claudin-2 was studied. Although the levels of the proteins were normal by Western blotting, a marked reduction in the expression of claudin-1 was observed at the canalicular borders. It is possible to hypothesise that the absence of ZO-2 might have impaired the characteristic

compactness of the tight junctions, leading to a leakage of the biliary components through the paracellular space in the liver parenchyma. Tight junctions might not have clustered at the so called “kissing points” of the plasma membrane, and therefore might be detected with a weak signal when stained with anti-claudin-1 antibody. This hypothesis was supported by the finding on transmission electron microscopy that the tight junctions appeared elongated. However, this is merely an observation; to increase the sensitivity and the specificity of this observation, a different approach is proposed for future studies. Immunogold for transmission electron microscope could be used, for example, to detect claudin-1 proteins and to precisely localise them within the hepatocytes of patients affected by TJP2 deficiency. Differently, at the cholangiocytes-cholangiocytes borders, variability in claudin-1 expression was identified by immunohistochemistry, suggesting that potential independent events, secondary to the ZO-2 deficiency, might have disrupted the membrane integrity of the epithelial tight junctional barrier at these sites. Various studies have shown that pro-inflammatory cytokines, like interferon gamma and tumour necrosis factor alpha, can induce endocytosis of integral tight junction proteins, such as occludin, claudin-1, claudin-4 and junctional adhesion molecule A (JAMA), and consequently increase the paracellular permeability (Bruewer *et al.*, 2003). However, this mechanism has been tested only on intestinal epithelium.

In our study, functional investigations, in an attempt to clarify the consequence of the absence of ZO-2, have been limited to the better studied claudin-1 and claudin-2. However, since a multiplicity of proteins constitutes cell-cell junctions, other components could have been affected to a higher degree. Therefore, cell culture systems or animal models would be useful tools to investigate further the cellular mechanisms that form the basis of cholestasis caused by mutations in *TJP2*. These models will be discussed below.

The effect of TJP2/ZO-2 in mammalian development was studied by knocking-out the complete orthologous gene in mice (Xu *et al.*, 2008). It was shown that ZO-2$^{-/-}$

mice were unable to mature beyond the eight-cell embryonic stage, due to arrest of cell proliferation and activation of apoptosis. On transmission electron microscopy, the tight junction integrity was compromised, as shown by the absence of the electron-dense plaque. This feature was very similar to what seen in our patients. On the other hand, the postnatal survival of our patients with TJP2 deficiency highlights an important interspecies difference between humans and mice.

Cell-cell junctions represent essential biological structures for the development and the support of different organs. The absence of ZO-2 might have been compensated by other junctional components in human development, but not in mice, where functional redundancy was not evident. Using a quantification assay, the expressions of different tight-junction-related genes were analysed in the liver tissue of TJP2 deficiency patients, compared with a group of healthy liver donors and a group of disease controls, including other inherited forms of cholestasis. A general up-regulation of tight junction-related genes was identified in all cholestatic conditions. Due to the hostile environment represented by the detergent property of bile acids, junctional constituents, such as occludins, claudins, cadherin/β-catenin complex and gap-junctional proteins, might be enriched at the canalicular and cholangiocyte membranes during cholestatic episodes. This possible mechanism of self-preservation points out the importance of ZO-2 in the liver, as its absence caused liver disease. Its role, however, seems to be tissue-specific, since the absence of ZO-2 expression has not caused the same degree of damage in other organs. Unfortunately, gene expression data regarding other tissues were not available due to the difficulty of obtaining biological specimens other than the liver these patients.

Interestingly, extrahepatic features were identified in TJP2 deficiency patients, but their association to the mutations in *TJP2* gene is still not clear. It is conceivable that the diverse extrahepatic manifestations of this disease between patients may be caused by genetic modifiers. Using whole-exome sequencing (WES) data, candidate genes can undergo intensive screening for known potential modifiers of specific phenotypes. In the present study cohort, this investigation could be

performed in patient 12a with protein-truncating mutation in *TJP2,* who is affected by severe cholestasis in addition with a chronic respiratory disorder. An additional family has been diagnosed with a similar phenotype, but no WES data have been produced as yet. Though these data could be used for the identification of potential modifiers. WES generates sequencing information related only to the protein coding areas of the genome, which represents 1% of the entire genome. The association of genetic polymorphisms and diseases has been identified not only on the exonic area, but frequently within intronic and promoter regions of the chromosome. Sequencing the whole genome (WGS), therefore, might be a more productive approach.

A single missense mutation in *TJP2* was identified in 2003 in individuals belonging to an isolated Amish community, affected by familial hypercholanemia (FHC) (Carlton *et al.,* 2003). A variable serum concentration of bile acids and fat malabsorption associated with a fluctuating increase in alkaline phosphate activity define this rare liver condition. However, FHC did not occur in every individual with *TJP2* mutation; three subjects, even though homozygous, remained persistently asymptomatic. This diverse manifestation was explained by incomplete penetrance; however *in vitro* assays showed that this mutation affects the binding between the PDZ1 region of ZO-2 and the claudin proteins. Further functional studies to clarify the implication of the mutation in the liver disease have not been performed and other affected individuals have not been identified.
Within this research project, patients with a severe cholestatic condition with an early-onset, and often requiring liver transplantation, were shown to harbour protein-truncating mutations in *TJP2*. This represents the severe end of a disease spectrum contrasting with the very mild phenotype seen in the Amish individuals (Figure 5.1.1). In this thesis, the possible mechanism through which genetic alterations lead to the absence of the encoded protein ZO-2, are described. Absence of ZO-2 might cause a failure of localisation of claudin-1 in the canalicular

membrane, de-stabilising the junctional structure and consequently leading to a leakage of bile into the paracellular space of the liver parenchyma.

In addition, an intermediate manifestation of the spectrum of disease is represented by three patients with the same homozygous missense mutation having a later disease-onset and recurrent episodes of cholestasis (Figure 5.1.1). Interestingly, a sibling was genotyped and shown to be a carrier of the same homozygous missense mutation, but liver function was still normal. In view of the possible disease-onset during late childhood, routine follow ups are in place to observe any further development. *In silico* analysis has shown that this nucleotide change affects the guanylate kinase-like domain, important in the interaction with the adjacent SH_3 domain in the formation of the protein core (Lye *et al.*, 2010). Therefore, a possible alteration of this interaction could have occurred. Nevertheless, immunohistochemical staining showed an intense nuclear staining of ZO-2 in all the three patients tested, and a variable canalicular expression pattern. Interestingly, claudin-1 expression in all the cases was mainly cytoplasmic. Hence, as suggested previously, the abnormal function of ZO-2 protein could have caused a failure of claudin-1 to localise at the canalicular membrane. The impact of the missense change on the ZO-2 protein, however, is still unclear. Since two putative conserved nuclear export signal (NES) have been identified as being located in the canine sequence of ZO-2 at the guanylate kinase-like region, the homozygous missense mutation in *TJP2* could have altered the NES consensus sequence and therefore the encoded protein could have been retained inside the nucleus (Gonzalez-Mariscal *et al.*, 2006). On the other hand, the known interaction of ZO-2 with chaperone proteins might have facilitated a small amount of proteins to leave the nucleus and localise at the canalicular margin. Unfortunately, biological material was not available for other functional studies.

The identification of human disease genes has proven essential for the knowledge of human physiology and pathophysiology. With this study, a spectrum of cholestatic liver disease severity has been associated with mutations in *TJP2,* with

a direct phenotype-genotype correlation (Figure 5.1.1). In accordance with these data, it is estimated that the screening of patients with idiopathic cholestasis will identify at least 10% as having genetic alterations in *TJP2*. The discovery of more patients will then clarify the disease phenotype, enabling us to distinguish it from the other inherited forms of progressive familial intrahepatic cholestasis. However, the few cases identified so far have suggested a similarity with FIC1 deficiency. Previously, the absence of ZO-2 has been suggested to alter the tight junction structure, leading to the diffusion of bile into the paracellular space between adjacent hepatocytes. An alternative hypothesis has however been proposed. Abnormal biological function or absence of ZO-2/TJP2 proteins, resulting from genetic alterations, could lead to tight junction instability; as a result of which a compensatory mechanism could rescue the barrier functionality, provoking in doing so an alteration of the canalicular membrane lipid composition. The phospholipid asymmetry across the lipid bilayer has an important role in the protection of the canalicular plasma membrane from the detergent properties of bile acids, as described in FIC deficiency and in *Atp8b1*$^{G308V/\ G308V}$ mutant mice (Paulusma *et al.*, 2006). Therefore, it will be interesting to characterise the lipid composition of the canalicular membrane in TJP2 deficiency patients using biological models, such as the ones described below.

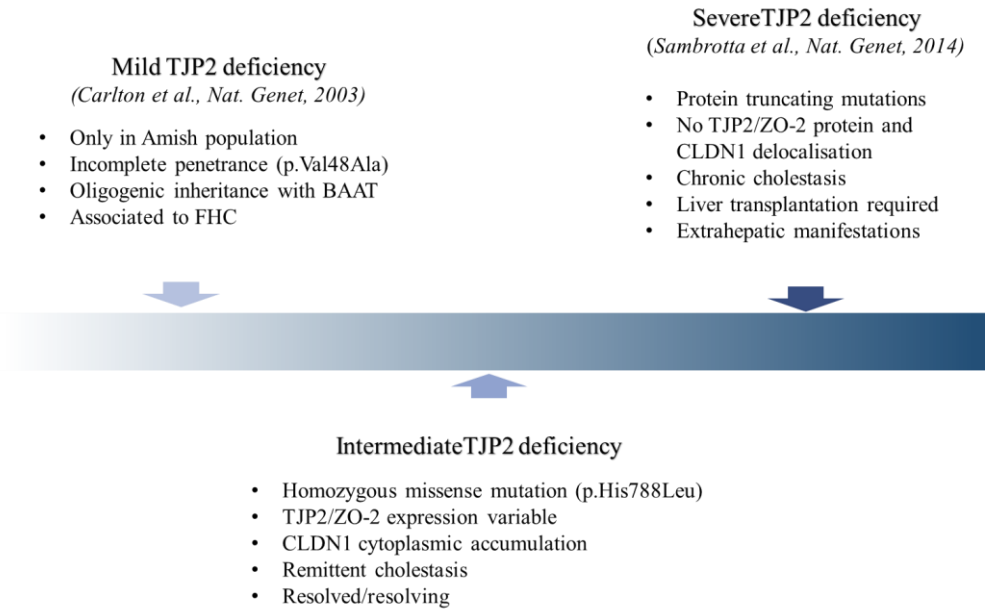

Figure 5.1.1 Representation of TJP2 deficiency as a spectrum of disease

The degree of severity of the TJP2 deficiency is shown with a gradient of blue, from mild to severe phenotype. On the left side of the image, a single missense mutation in *TJP2* was identified in 2003 in an Amish population affected by familial hypercholanemia (FHC) (Carlton *et al.*, 2003). Incomplete penetrance was exhibited, as this atypical liver condition did not occur in every individual with *TJP2* mutation. Within this project, two additional spectra of the disease were discovered, with a genotype-phenotype correlation. An intermediate phenotype with remittent episodes of cholestasis was caused by a single homozygous missense mutation. Immunohistochemical studies showed a variable membrane expression of the TJP2/ZO-2 protein on the membrane but intense nuclear staining and the directly-linked integral protein, claudin-1, with cytoplasmic accumulation. On the right side of the image, a severe cholestatic condition was identified in the majority of the cases with protein-truncating mutations. The protein was no expressed and claudin-1 showed delocalisation. Most of these patients required liver transplantation and few of them had extrahepatic manifestations, possibly due by lack of ZO-2 in other organs (Sambrotta *et al.*, 2014).

An *in vitro* TJP2 deficiency model can be used to understand better the different, as yet unclear, molecular aspects of the disease described above. Due to the poor *in vitro* longevity of primary human hepatocytes, a cell culture system model has been proposed using patient-specific hepatocyte-like cells derived from human induced pluripotent stem cells (iPSCs-Hep) (Rashid *et al.*, 2010). This system has been successfully validated for the investigation of different metabolic disorders,

such as α-1 antitrypsin deficiency and glycogen storage disease type 1a. The potential of these *in vitro* models lies in studying a given genetic condition in a controlled environment, allowing detailed analysis of the different molecular pathways. In TJP2 deficiency, this approach could be adopted, for example, for the investigation of the lipid composition of the plasma membrane; the detection and localisation of other tight junction proteins, using fluorescence electron microscopy and confocal microscopy; and for the examination of a possible alteration in cell polarity due to lack of the ZO-2 protein. However, cell culture system models can only partially represent the real nature of the disease, as different biological factors are likely to influence the pathogenesis and phenotype. Therefore, animal models could be a better research method. As described above, ZO-2 knock-out mice have shown embryonic lethality. An alternative biological model could be represented by a conditional ZO-2 knock-out mouse model, where the expression of the protein is ablated only in the liver. This model has allowed, for example, to study the bile duct proliferation in Alagille syndrome and the role of Jagged-1 protein in liver development, as the homozygous Jagged-1 knock-out mouse is an inappropriate model, being embryonic lethal (Loomes *et al.*, 2007).

5.2 AMACR deficiency

The application of next-generation sequencing in genetic research has enriched human molecular biology, as for every gene discovered known molecular pathways are re-evaluated or new pathways are elucidated. With the discovery of *TJP2* mutations, the role of tight junctions in the aetiology of progressive familial intrahepatic cholestasis has been highlighted. In addition, the high sensitivity of the method has led to a rapid discovery of novel mutations in known genes. As shown in this study, a novel missense mutation in *AMACR* was discovered in one patient affected by early-onset cholestasis. The findings were supported by the clinical manifestation of a possible metabolic disorder, manifest with mild

elevation of liver enzymatic activity, vitamin D malabsorption and urinary intermediate metabolites of bile acids. In addition, *in silico* analysis predicted the nucleotide substitution to be causative of functional damage to the encoded protein. No further investigations have been made, but the gene has been also associated with this rare infantile metabolic disease in an independent centre. The identification of a metabolic disorder in this study highlights that a permissive inclusion criteria was adopted for the recruitment of patients. In fact, they were selected on the basis of the manifestation of early-onset of severe cholestatic condition, with strong family history of liver diseases and with normal/low concentration of serum GGT; the other biochemical markers of liver dysfunction were variable. These criteria were therefore including a wide range of liver genetic conditions, including inborn errors of metabolism.

5.3 Limitations of experimental method

Despite these advances, several paediatric patients with cholestasis still have no known genetic diagnosis. As shown in Figure 2.5.1, targeted resequencing for the panel of 21 genes and for the panel of 7 genes failed the identification of disease-causing mutations in 6 and 26 paediatric patients respectively. Additionally, whole exome sequencing analysis missed the discovery of novel pathological genetic alterations in 5 individuals. The failure of identification could be explained by 1) technical limitations, 2) stringent analytical approach or 3) biological factors.

1) The major limitation of using a capture-based methods is the lack of detection of variants located in non-coding area, omitting in the window of investigation for example promotors or enhancer regions. In addition, there are areas on the human genome with high GC content and enriched in repetitive regions that are likely to not be captured during the library preparation. A solution for these issues has been identified in using a whole genome sequencing approach, in which the efficiency of the capture is not

an issue. (Belkadi *et al.*, 2015). However the cost is still higher than whole-exome sequencing and the routine application for the discovery of disease-causing mutations in Mendelian disorders is still not widely available.

2) In this research project and in particular in the analysis of whole-exome sequencing data, the lack of discovery could have been due to the usage of a stringent filtering procedure. The data were analysed with the assumption that homozygous mutations are most likely to have caused this autosomal recessive disease in those patients belonging to consanguineous families. However, two heterozygous disease-causing mutations could also have been the causative alterations. Therefore, compound heterozygous mutations will be investigated. Another limitation of the method is related to the possibility to miss the detection of structural variations, such as inversion and translocations, and repeat expansion. However progress in bioinformatics has led to specific computational tools to help overcome these issues.
3) Interfamilial locus heterogeneity could have been an explanation of failure of identification in those individuals where whole-exome sequencing was performed. Interfamilial locus heterogeneity is described when the same Mendelial trait is caused by mutations in different genes. Each family studied therefore could be mutated at a unique disease-causing locus. Using whole-exome sequencing data from affected siblings and/or parents will reduce this biological limitation and increase the power of discovery

As highlighted in this research project, different pathways are involved in the aetiology of cholestatic liver disease, as several molecules contribute to the formation and transport of bile. In the immediate future it is highly likely that more genes will be identified through next generation sequencing, especially when whole genome sequencing will be more accessible and the limitation in gene discovery will be reduced. This rapid identification of the molecular basis of liver disease will contribute further insight into the pathophysiology of these diseases and subsequently into the development of novel targeted treatments.

6 References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. &Walter, P. (2002a). Cell Junctions. Molecular Biology of the Cell fifth edition. Garland Science.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. &Walter, P. (2002b). The Molecular Mechanisms of Membrane Transport and the Maintenance of Compartmental Diversity. Molecular Biology of the Cell fifth edition. Garland Science.

Alvarez, L., Jara, P., Sanchez-Sabate, E., Hierro, L., Larrauri, J., Diaz, M. C.*, et al.* (2004). Reduced hepatic expression of farnesoid X receptor in hereditary cholestasis associated to mutation in ATP8B1. *Hum Mol Genet* **13,** 2451-60.

Andrew Nesbit, M., Bowl, M. R., Harding, B., Schlessinger, D., Whyte, M. P. &Thakker, R. V. (2004). X-linked hypoparathyroidism region on Xq27 is evolutionarily conserved with regions on 3q26 and 13q34 and contains a novel P-type ATPase. *Genomics* **84,** 1060-70.

Angelow, S., Ahlstrom, R. &Yu, A. S. (2008). Biology of claudins. *Am J Physiol Renal Physiol* **295,** F867-76.

Aydogdu, S., Cakir, M., Arikan, C., Tumgor, G., Yuksekkaya, H. A., Yilmaz, F.*, et al.* (2007). Liver transplantation for progressive familial intrahepatic cholestasis: clinical and histopathological findings, outcome and impact on growth. *Pediatr Transplant* **11,** 634-40.

Balakrishnan, A.Polli, J. E. (2006). Apical sodium dependent bile acid transporter (ASBT, SLC10A2): a potential prodrug target. *Mol Pharm* **3,** 223-30.

Balda, M. S.Matter, K. (2008). Tight junctions at a glance. *J Cell Sci* **121,** 3677-82.

Balistreri, W. F., Bezerra, J. A., Jansen, P., Karpen, S. J., Shneider, B. L. &Suchy, F. J. (2005). Intrahepatic cholestasis: summary of an American Association for the Study of Liver Diseases single-topic conference. *Hepatology* **42,** 222-35.

Banales, J. M., Prieto, J. &Medina, J. F. (2006). Cholangiocyte anion exchange and biliary bicarbonate excretion. *World J Gastroenterol* **12,** 3496-511.

Bauer, H., Zweimueller-Mayer, J., Steinbacher, P., Lametschwandtner, A. &Bauer, H. C. (2010). The dual role of zonula occludens (ZO) proteins. *J Biomed Biotechnol* **2010,** 402593.

Beatch, M., Jesaitis, L. A., Gallin, W. J., Goodenough, D. A. &Stevenson, B. R. (1996). The tight junction protein ZO-2 contains three PDZ (PSD-95/Discs-Large/ZO-1) domains and an alternatively spliced region. *J Biol Chem* **271,** 25723-6.

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A.*, et al.* (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A* **112,** 5473-8.

Bergasa, N. V. (2014). Pruritus of Cholestasis. Itch: Mechanisms and Treatment, Boca Raton (FL).

Betanzos, A., Huerta, M., Lopez-Bayghen, E., Azuara, E., Amerena, J. &Gonzalez-Mariscal, L. (2004). The tight junction protein ZO-2 associates with Jun, Fos and C/EBP transcription factors in epithelial cells. *Exp Cell Res* **292,** 51-66.

Beurel, E.Jope, R. S. (2006). The paradoxical pro- and anti-apoptotic actions of GSK3 in the intrinsic and extrinsic apoptosis signaling pathways. *Prog Neurobiol* **79,** 173-89.

Boycott, K. M., Vanstone, M. R., Bulman, D. E. &MacKenzie, A. E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet* **14,** 681-91.

Bruewer, M., Luegering, A., Kucharzik, T., Parkos, C. A., Madara, J. L., Hopkins, A. M., *et al.* (2003). Proinflammatory cytokines disrupt epithelial barrier function by apoptosis-independent mechanisms. *J Immunol* **171,** 6164-72.

Bull, L. N., Mahmoodi, V., Baker, A. J., Jones, R., Strautnieks, S. S., Thompson, R. J., *et al.* (2006). VPS33B mutation with ichthyosis, cholestasis, and renal dysfunction but without arthrogryposis: incomplete ARC syndrome phenotype. *J Pediatr* **148,** 269-71.

Bull, L. N., van Eijk, M. J., Pawlikowska, L., DeYoung, J. A., Juijn, J. A., Liao, M., *et al.* (1998). A gene encoding a P-type ATPase mutated in two forms of hereditary cholestasis. *Nat Genet* **18,** 219-24.

Burrows, M.Wheeler, D. J. (1994). A block-sorting lossless data compression algorithm. *In* "Technical report", Vol. 124. Digital Equipment Corporation, Palo Alto, CA, .

Cabrera-Abreu, J. C.Green, A. (2002). Gamma-glutamyltransferase: value of its measurement in paediatrics. *Ann Clin Biochem* **39,** 22-5.

Carlton, V. E., Harris, B. Z., Puffenberger, E. G., Batta, A. K., Knisely, A. S., Robinson, D. L., *et al.* (2003). Complex inheritance of familial hypercholanemia with associated mutations in TJP2 and BAAT. *Nat Genet* **34,** 91-6.

Carr, C. M.Rizo, J. (2010). At the junction of SNARE and SM protein function. *Curr Opin Cell Biol* **22,** 488-95.

Cheng, J. B., Jacquemin, E., Gerhardt, M., Nazer, H., Cresteil, D., Heubi, J. E., *et al.* (2003). Molecular genetics of 3beta-hydroxy-Delta5-C27-steroid oxidoreductase deficiency in 16 patients with loss of bile acid synthesis and liver disease. *J Clin Endocrinol Metab* **88,** 1833-41.

Chiang, J. Y., Kimmel, R., Weinberger, C. &Stroup, D. (2000). Farnesoid X receptor responds to bile acids and represses cholesterol 7alpha-hydroxylase gene (CYP7A1) transcription. *J Biol Chem* **275,** 10918-24.

Chlenski, A., Ketels, K. V., Engeriser, J. L., Talamonti, M. S., Tsao, M. S., Koutnikova, H., *et al.* (1999). zo-2 gene alternative promoters in normal and neoplastic human pancreatic duct cells. *Int J Cancer* **83,** 349-58.

Chlenski, A., Ketels, K. V., Korovaitseva, G. I., Talamonti, M. S., Oyasu, R. &Scarpelli, D. G. (2000). Organization and expression of the human zo-2 gene (tjp-2) in normal and neoplastic tissues. *Biochim Biophys Acta* **1493,** 319-24.

Colombo, C. (2007). Liver disease in cystic fibrosis. *Curr Opin Pulm Med* **13,** 529-36.

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., *et al.* (2010). Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704-12.

Crosignani, A., Del Puppo, M., Longo, M., De Fabiani, E., Caruso, D., Zuin, M., *et al.* (2007). Changes in classic and alternative pathways of bile acid synthesis in chronic liver disease. *Clin Chim Acta* **382,** 82-8.

Cullinane, A. R., Straatman-Iwanowska, A., Zaucker, A., Wakabayashi, Y., Bruce, C. K., Luo, G., *et al.* (2010). Mutations in VIPAR cause an arthrogryposis, renal dysfunction and cholestasis syndrome phenotype with defects in epithelial polarization. *Nat Genet* **42,** 303-12.

Davit-Spraul, A., Gonzales, E., Baussan, C. &Jacquemin, E. (2009). Progressive familial intrahepatic cholestasis. *Orphanet J Rare Dis* **4,** 1.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., *et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43,** 491-8.

Dixon, P. H., van Mil, S. W., Chambers, J., Strautnieks, S., Thompson, R. J., Lammert, F., *et al.* (2009). Contribution of variant alleles of ABCB11 to susceptibility to intrahepatic cholestasis of pregnancy. *Gut* **58,** 537-44.

Drury, J. E., Mindnich, R. &Penning, T. M. (2010). Characterization of disease-related 5beta-reductase (AKR1D1) mutations reveals their potential to cause bile acid deficiency. *J Biol Chem* **285,** 24529-37.

Duclos, F., Rodius, F., Wrogemann, K., Mandel, J. L. &Koenig, M. (1994). The Friedreich ataxia region: characterization of two novel genes and reduction of the critical region to 300 kb. *Hum Mol Genet* **3,** 909-14.

Ebnet, K., Schulz, C. U., Meyer Zu Brickwedde, M. K., Pendl, G. G. &Vestweber, D. (2000). Junctional adhesion molecule interacts with the PDZ domain-containing proteins AF-6 and ZO-1. *J Biol Chem* **275,** 27979-88.

Eloranta, J. J.Kullak-Ublick, G. A. (2008). The role of FXR in disorders of bile acid homeostasis. *Physiology (Bethesda)* **23,** 286-95.

Erlinger, S., Arias, I. M. &Dhumeaux, D. (2014). Inherited Disorders of Bilirubin Transport and Conjugation. *Gastroenterology*.

Esteller, A. (2008). Physiology of bile secretion. *World J Gastroenterol* **14,** 5641-9.

Evans, M. J., von Hahn, T., Tscherne, D. M., Syder, A. J., Panis, M., Wolk, B.*, et al.* (2007). Claudin-1 is a hepatitis C virus co-receptor required for a late step in entry. *Nature* **446,** 801-5.

Fairbanks, K. D.Tavill, A. S. (2008). Liver disease in alpha 1-antitrypsin deficiency: a review. *Am J Gastroenterol* **103,** 2136-41; quiz 2142.

Fanen, P., Wohlhuter-Haddad, A. &Hinzpeter, A. (2014). Genetics of cystic fibrosis: CFTR mutation classifications toward genotype-based CF therapies. *Int J Biochem Cell Biol* **52,** 94-102.

Forman, B. M., Goode, E., Chen, J., Oro, A. E., Bradley, D. J., Perlmann, T.*, et al.* (1995). Identification of a nuclear receptor that is activated by farnesol metabolites. *Cell* **81,** 687-93.

Fortini, M. E. (2009). Notch signaling: the core pathway and its posttranslational regulation. *Dev Cell* **16,** 633-47.

Frischmeyer, P. A.Dietz, H. C. (1999). Nonsense-mediated mRNA decay in health and disease. *Hum Mol Genet* **8,** 1893-900.

Furuse, M., Fujita, K., Hiiragi, T., Fujimoto, K. &Tsukita, S. (1998). Claudin-1 and -2: novel integral membrane proteins localizing at tight junctions with no sequence similarity to occludin. *J Cell Biol* **141,** 1539-50.

Gerloff, T., Stieger, B., Hagenbuch, B., Madon, J., Landmann, L., Roth, J.*, et al.* (1998). The sister of P-glycoprotein represents the canalicular bile salt export pump of mammalian liver. *J Biol Chem* **273,** 10046-50.

Gettins, P. G. (2002). Serpin structure, mechanism, and function. *Chem Rev* **102,** 4751-804.

Gissen, P., Johnson, C. A., Morgan, N. V., Stapelbroek, J. M., Forshew, T., Cooper, W. N.*, et al.* (2004). Mutations in VPS33B, encoding a regulator of SNARE-dependent membrane fusion, cause arthrogryposis-renal dysfunction-cholestasis (ARC) syndrome. *Nat Genet* **36,** 400-4.

Gonzalez-Mariscal, L., Ponce, A., Alarcon, L. &Jaramillo, B. E. (2006). The tight junction protein ZO-2 has several functional nuclear export signals. *Exp Cell Res* **312,** 3323-35.

Gonzalez, H. E., Eugenin, E. A., Garces, G., Solis, N., Pizarro, M., Accatino, L.*, et al.* (2002). Regulation of hepatic connexins in cholestasis: possible involvement of Kupffer cells and inflammatory mediators. *Am J Physiol Gastrointest Liver Physiol* **282,** G991-G1001.

Goodwin, B., Jones, S. A., Price, R. R., Watson, M. A., McKee, D. D., Moore, L. B.*, et al.* (2000). A regulatory cascade of the nuclear receptors FXR, SHP-1, and LRH-1 represses bile acid biosynthesis. *Mol Cell* **6,** 517-26.

Gottesman, M. M.Ambudkar, S. V. (2001). Overview: ABC transporters and human disease. *J Bioenerg Biomembr* **33,** 453-8.

Gow, A., Southwood, C. M., Li, J. S., Pariali, M., Riordan, G. P., Brodie, S. E.*, et al.* (1999). CNS myelin and sertoli cell tight junction strands are absent in Osp/claudin-11 null mice. *Cell* **99,** 649-59.

Graf, G. A., Yu, L., Li, W. P., Gerard, R., Tuma, P. L., Cohen, J. C.*, et al.* (2003). ABCG5 and ABCG8 are obligate heterodimers for protein trafficking and biliary cholesterol excretion. *J Biol Chem* **278,** 48275-82.

Gumbiner, B., Lowenkopf, T. &Apatira, D. (1991). Identification of a 160-kDa polypeptide that binds to the tight junction protein ZO-1. *Proc Natl Acad Sci U S A* **88,** 3460-4.

Hadj-Rabia, S., Baala, L., Vabres, P., Hamel-Teillac, D., Jacquemin, E., Fabre, M.*, et al.* (2004). Claudin-1 gene mutations in neonatal sclerosing cholangitis associated with ichthyosis: a tight junction disease. *Gastroenterology* **127,** 1386-90.

Harendza, S., Hubner, C. A., Glaser, C., Burdelski, M., Thaiss, F., Hansmann, I.*, et al.* (2005). Renal failure and hypertension in Alagille syndrome with a novel JAG1 mutation. *J Nephrol* **18,** 312-7.

Harris, B. Z.Lim, W. A. (2001). Mechanism and role of PDZ domains in signaling complex assembly. *J Cell Sci* **114,** 3219-31.

Hediger, M. A., Romero, M. F., Peng, J. B., Rolfs, A., Takanaga, H. &Bruford, E. A. (2004). The ABCs of solute carriers: physiological, pathological and therapeutic implications of human membrane transport proteinsIntroduction. *Pflugers Arch* **447,** 465-8.

Hernandez, S., Chavez Munguia, B. &Gonzalez-Mariscal, L. (2007). ZO-2 silencing in epithelial cells perturbs the gate and fence function of tight junctions and leads to an atypical monolayer architecture. *Exp Cell Res* **313,** 1533-47.

Hewitt, K. J., Agarwal, R. &Morin, P. J. (2006). The claudin gene family: expression in normal and neoplastic tissues. *BMC Cancer* **6,** 186.

Hofmann, A. F. (2002). Cholestatic liver disease: pathophysiology and therapeutic options. *Liver* **22 Suppl 2,** 14-9.

Holczbauer, A., Gyongyosi, B., Lotz, G., Szijarto, A., Kupcsulik, P., Schaff, Z.*, et al.* (2013). Distinct claudin expression profiles of hepatocellular carcinoma and metastatic colorectal and pancreatic carcinomas. *J Histochem Cytochem* **61,** 294-305.

Holt, R. A,Jones, S. J. (2008). The new paradigm of flow cell sequencing. *Genome Res* **18,** 839-46.

Huebert, R. C., Splinter, P. L., Garcia, F., Marinelli, R. A. &LaRusso, N. F. (2002). Expression and localization of aquaporin water channels in rat hepatocytes. Evidence for a role in canalicular bile secretion. *J Biol Chem* **277,** 22710-7.

Huizing, M., Didier, A., Walenta, J., Anikster, Y., Gahl, W. A. &Kramer, H. (2001). Molecular cloning and characterization of human VPS18, VPS 11, VPS16, and VPS33. *Gene* **264,** 241-7.

Hulot, J. S., Villard, E., Maguy, A., Morel, V., Mir, L., Tostivint, I.*, et al.* (2005). A mutation in the drug transporter gene ABCC2 associated with impaired methotrexate elimination. *Pharmacogenet Genomics* **15,** 277-85.

Islas, S., Vega, J., Ponce, L. &Gonzalez-Mariscal, L. (2002). Nuclear localization of the tight junction protein ZO-2 in epithelial cells. *Exp Cell Res* **274,** 138-48.

Ismail, H., Kalicinski, P., Markiewicz, M., Jankowska, I., Pawlowska, J., Kluge, P.*, et al.* (1999). Treatment of progressive familial intrahepatic cholestasis: liver transplantation or partial external biliary diversion. *Pediatr Transplant* **3,** 219-24.

Itoh, M., Furuse, M., Morita, K., Kubota, K., Saitou, M. &Tsukita, S. (1999). Direct binding of three tight junction-associated MAGUKs, ZO-1, ZO-2, and ZO-3, with the COOH termini of claudins. *J Cell Biol* **147,** 1351-63.

Jacquemin, E., Cresteil, D., Manouvrier, S., Boute, O. &Hadchouel, M. (1999). Heterozygous non-sense mutation of the MDR3 gene in familial intrahepatic cholestasis of pregnancy. *Lancet* **353,** 210-1.

Jacquemin, E., De Vree, J. M., Cresteil, D., Sokal, E. M., Sturm, E., Dumont, M.*, et al.* (2001). The wide spectrum of multidrug resistance 3 deficiency: from neonatal cholestasis to cirrhosis of adulthood. *Gastroenterology* **120,** 1448-58.

Jacquemin, E., Hermans, D., Myara, A., Habes, D., Debray, D., Hadchouel, M.*, et al.* (1997). Ursodeoxycholic acid therapy in pediatric patients with progressive familial intrahepatic cholestasis. *Hepatology* **25,** 519-23.

Jang, J. Y., Kim, K. M., Kim, G. H., Yu, E., Lee, J. J., Park, Y. S.*, et al.* (2009). Clinical characteristics and VPS33B mutations in patients with ARC syndrome. *J Pediatr Gastroenterol Nutr* **48,** 348-54.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proc Natl Acad Sci U S A* **110,** 19313-7.

Jones, P. M.George, A. M. (2004). The ABC transporter structure and mechanism: perspectives on recent research. *Cell Mol Life Sci* **61,** 682-99.

Kamath, B. M., Bauer, R. C., Loomes, K. M., Chao, G., Gerfen, J., Hutchinson, A.*, et al.* (2012). NOTCH2 mutations in Alagille syndrome. *J Med Genet* **49,** 138-44.

Katoh, Y.Katoh, M. (2004). Identification and characterization of CDC50A, CDC50B and CDC50C genes in silico. *Oncol Rep* **12,** 939-43.

Kervestin, S.Jacobson, A. (2012). NMD: a multifaceted response to premature translational termination. *Nat Rev Mol Cell Biol* **13,** 700-12.

Kim, M. A., Kim, Y. R., Sagong, B., Cho, H. J., Bae, J. W., Kim, J.*, et al.* (2014). Genetic analysis of genes related to tight junction function in the korean population with non-syndromic hearing loss. *PLoS One* **9,** e95646.

Knisely, A. S., Strautnieks, S. S., Meier, Y., Stieger, B., Byrne, J. A., Portmann, B. C.*, et al.* (2006). Hepatocellular carcinoma in ten children under five years of age with bile salt export pump deficiency. *Hepatology* **44,** 478-86.

Kobayashi, K., Sinasac, D. S., Iijima, M., Boright, A. P., Begum, L., Lee, J. R.*, et al.* (1999). The gene mutated in adult-onset type II citrullinaemia encodes a putative mitochondrial carrier protein. *Nat Genet* **22,** 159-63.

Kojima, T., Kokai, Y., Chiba, H., Yamamoto, M., Mochizuki, Y. &Sawada, N. (2001). Cx32 but not Cx26 is associated with tight junctions in primary cultures of rat hepatocytes. *Exp Cell Res* **263,** 193-201.

Kojima, T., Yamamoto, T., Murata, M., Chiba, H., Kokai, Y. &Sawada, N. (2003). Regulation of the blood-biliary barrier: interaction between gap and tight junctions in hepatocytes. *Med Electron Microsc* **36,** 157-64.

Kurbegov, A. C.Karpen, S. J. (2008). Bile Formation and Cholestasis. Walker's Pediatric Gastrointestinal Disease 5th edition.

Kurbegov, A. C., Setchell, K. D., Haas, J. E., Mierau, G. W., Narkewicz, M., Bancroft, J. D.*, et al.* (2003). Biliary diversion for progressive familial intrahepatic cholestasis: improved liver morphology and bile acid profile. *Gastroenterology* **125,** 1227-34.

Laffitte, B. A., Kast, H. R., Nguyen, C. M., Zavacki, A. M., Moore, D. D. &Edwards, P. A. (2000). Identification of the DNA binding specificity and potential target genes for the farnesoid X-activated receptor. *J Biol Chem* **275,** 10638-47.

Lang, C., Meier, Y., Stieger, B., Beuers, U., Lang, T., Kerb, R., *et al.* (2007). Mutations and polymorphisms in the bile salt export pump and the multidrug resistance protein 3 associated with drug-induced liver injury. *Pharmacogenet Genomics* **17,** 47-60.

Lee, J. H., Chen, H. L., Chen, H. L., Ni, Y. H., Hsu, H. Y. &Chang, M. H. (2006). Neonatal Dubin-Johnson syndrome: long-term follow-up and MRP2 mutations study. *Pediatr Res* **59,** 584-9.

Lenzen, R., Alpini, G. &Tavoloni, N. (1992). Secretin stimulates bile ductular secretory activity through the cAMP system. *Am J Physiol* **263,** G527-32.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., *et al.* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078-9.

Li, L., Krantz, I. D., Deng, Y., Genin, A., Banta, A. B., Collins, C. C., *et al.* (1997). Alagille syndrome is caused by mutations in human Jagged1, which encodes a ligand for Notch1. *Nat Genet* **16,** 243-51.

Livak, K. J.Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25,** 402-8.

Lomas, D. A., Evans, D. L., Finch, J. T. &Carrell, R. W. (1992). The mechanism of Z alpha 1-antitrypsin accumulation in the liver. *Nature* **357,** 605-7.

Loomes, K. M., Russo, P., Ryan, M., Nelson, A., Underkoffler, L., Glover, C., *et al.* (2007). Bile duct proliferation in liver-specific Jag1 conditional knockout mice: effects of gene dosage. *Hepatology* **45,** 323-30.

Lowry, O. H., Rosebrough, N. J., Farr, A. L. &Randall, R. J. (1951). Protein measurement with the Folin phenol reagent. *J Biol Chem* **193,** 265-75.

Lubamba, B., Dhooghe, B., Noel, S. &Leal, T. (2012). Cystic fibrosis: insight into CFTR pathophysiology and pharmacotherapy. *Clin Biochem* **45,** 1132-44.

Lucena, J. F., Herrero, J. I., Quiroga, J., Sangro, B., Garcia-Foncillas, J., Zabalegui, N.*, et al.* (2003). A multidrug resistance 3 gene mutation causing cholelithiasis, cholestasis of pregnancy, and adulthood biliary cirrhosis. *Gastroenterology* **124,** 1037-42.

Lye, M. F., Fanning, A. S., Su, Y., Anderson, J. M. &Lavie, A. (2010). Insights into regulated ligand binding sites from the structure of ZO-1 Src homology 3-guanylate kinase module. *J Biol Chem* **285,** 13907-17.

Lykavieris, P., van Mil, S., Cresteil, D., Fabre, M., Hadchouel, M., Klomp, L.*, et al.* (2003). Progressive familial intrahepatic cholestasis type 1 and extrahepatic features: no catch-up of stature growth, exacerbation of diarrhea, and appearance of liver steatosis after liver transplantation. *J Hepatol* **39,** 447-52.

Mangelsdorf, D. J., Thummel, C., Beato, M., Herrlich, P., Schutz, G., Umesono, K.*, et al.* (1995). The nuclear receptor superfamily: the second decade. *Cell* **83,** 835-9.

Marcus, N. Y., Brunt, E. M., Blomenkamp, K., Ali, F., Rudnick, D. A., Ahmad, M.*, et al.* (2010). Characteristics of hepatocellular carcinoma in a murine model of alpha-1-antitrypsin deficiency. *Hepatol Res* **40,** 641-53.

Marsden, M. D.Fournier, R. E. (2005). Organization and expression of the human serpin gene cluster at 14q32.1. *Front Biosci* **10,** 1768-78.

Marzesco, A. M., Dunia, I., Pandjaitan, R., Recouvreur, M., Dauzonne, D., Benedetti, E. L.*, et al.* (2002). The small GTPase Rab13 regulates assembly of functional tight junctions in epithelial cells. *Mol Biol Cell* **13,** 1819-31.

Mayer, B. J. (2001). SH3 domains: complexity in moderation. *J Cell Sci* **114,** 1253-63.

McDaniell, R., Warthen, D. M., Sanchez-Lara, P. A., Pai, A., Krantz, I. D., Piccoli, D. A., *et al.* (2006). NOTCH2 mutations cause Alagille syndrome, a heterogeneous disorder of the notch signaling pathway. *Am J Hum Genet* **79,** 169-73.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., *et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20,** 1297-303.

Meier, Y., Pauli-Magnus, C., Zanger, U. M., Klein, K., Schaeffeler, E., Nussler, A. K., *et al.* (2006). Interindividual variability of canalicular ATP-binding-cassette (ABC)–transporter expression in human liver. *Hepatology* **44,** 62-74.

Miethke, A. G.Balistreri, W. F. (2008). Approach to Neonatal Cholestasis. Walker's Pediatric Gastrointestinal Disease.

Mitic, L. L., Van Itallie, C. M. &Anderson, J. M. (2000). Molecular physiology and pathophysiology of tight junctions I. Tight junction structure and function: lessons from mutant animals and proteins. *Am J Physiol Gastrointest Liver Physiol* **279,** G250-4.

Monte, M. J., Marin, J. J., Antelo, A. &Vazquez-Tato, J. (2009). Bile acids: chemistry, physiology, and pathophysiology. *World J Gastroenterol* **15,** 804-16.

Nicholson, P., Yepiskoposyan, H., Metze, S., Zamudio Orozco, R., Kleinschmidt, N. &Muhlemann, O. (2010). Nonsense-mediated mRNA decay in human cells: mechanistic insights, functions beyond quality control and the double-life of NMD factors. *Cell Mol Life Sci* **67,** 677-700.

Oda, T., Elkahloun, A. G., Pike, B. L., Okajima, K., Krantz, I. D., Genin, A., *et al.* (1997). Mutations in the human Jagged1 gene are responsible for Alagille syndrome. *Nat Genet* **16,** 235-42.

Olefsky, J. M. (2001). Nuclear receptor minireview series. *J Biol Chem* **276,** 36863-4.

Olsen, O.Bredt, D. S. (2003). Functional analysis of the nucleotide binding domain of membrane-associated guanylate kinases. *J Biol Chem* **278,** 6873-8.

Oude Elferink, R. P.Paulusma, C. C. (2007). Function and pathophysiological importance of ABCB4 (MDR3 P-glycoprotein). *Pflugers Arch* **453,** 601-10.

Pauli-Magnus, C., Lang, T., Meier, Y., Zodan-Marin, T., Jung, D., Breymann, C.*, et al.* (2004). Sequence analysis of bile salt export pump (ABCB11) and multidrug resistance p-glycoprotein 3 (ABCB4, MDR3) in patients with intrahepatic cholestasis of pregnancy. *Pharmacogenetics* **14,** 91-102.

Pauli-Magnus, C.Meier, P. J. (2006). Hepatobiliary transporters and drug-induced cholestasis. *Hepatology* **44,** 778-87.

Paulusma, C. C., Bosma, P. J., Zaman, G. J., Bakker, C. T., Otter, M., Scheffer, G. L.*, et al.* (1996). Congenital jaundice in rats with a mutation in a multidrug resistance-associated protein gene. *Science* **271,** 1126-8.

Paulusma, C. C.Elferink, R. P. (2010). P4 ATPases--the physiological relevance of lipid flipping transporters. *FEBS Lett* **584,** 2708-16.

Paulusma, C. C., Folmer, D. E., Ho-Mok, K. S., de Waart, D. R., Hilarius, P. M., Verhoeven, A. J.*, et al.* (2008). ATP8B1 requires an accessory protein for endoplasmic reticulum exit and plasma membrane lipid flippase activity. *Hepatology* **47,** 268-78.

Paulusma, C. C., Groen, A., Kunne, C., Ho-Mok, K. S., Spijkerboer, A. L., Rudi de Waart, D.*, et al.* (2006). Atp8b1 deficiency in mice reduces resistance of the canalicular membrane to hydrophobic bile salts and impairs bile salt transport. *Hepatology* **44,** 195-204.

Paulusma, C. C., Kool, M., Bosma, P. J., Scheffer, G. L., ter Borg, F., Scheper, R. J.*, et al.* (1997). A mutation in the human canalicular multispecific organic anion transporter gene causes the Dubin-Johnson syndrome. *Hepatology* **25,** 1539-42.

Pellicciari, R., Fiorucci, S., Camaioni, E., Clerici, C., Costantino, G., Maloney, P. R., *et al.* (2002). 6alpha-ethyl-chenodeoxycholic acid (6-ECDCA), a potent and selective FXR agonist endowed with anticholestatic activity. *J Med Chem* **45,** 3569-72.

Pintar, A., Guarnaccia, C., Dhir, S. &Pongor, S. (2009). Exon 6 of human JAG1 encodes a conserved structural unit. *BMC Struct Biol* **9,** 43.

Piontek, J., Winkler, L., Wolburg, H., Muller, S. L., Zuleger, N., Piehl, C., *et al.* (2008). Formation of tight junction: determinants of homophilic interaction between classic claudins. *FASEB J* **22,** 146-58.

Plagnol, V., Curtis, J., Epstein, M., Mok, K. Y., Stebbings, E., Grigoriadou, S., *et al.* (2012). A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* **28,** 2747-54.

Popper, H. (1981). Cholestasis: the future of a past and present riddle. *Hepatology* **1,** 187-91.

Rao, R. K.Samak, G. (2013). Bile duct epithelial tight junctions and barrier function. *Tissue Barriers* **1,** e25718.

Rashid, S. T., Corbineau, S., Hannan, N., Marciniak, S. J., Miranda, E., Alexander, G., *et al.* (2010). Modeling inherited metabolic disorders of the liver using human induced pluripotent stem cells. *J Clin Invest* **120,** 3127-36.

Resuehr, D.Spiess, A. N. (2003). A real-time polymerase chain reaction-based evaluation of cDNA synthesis priming methods. *Anal Biochem* **322,** 287-91.

Roach, J. C., Boysen, C., Wang, K. &Hood, L. (1995). Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics* **26,** 345-53.

Rosmorduc, O., Hermelin, B. &Poupon, R. (2001). MDR3 gene defect in adults with symptomatic intrahepatic and gallbladder cholesterol cholelithiasis. *Gastroenterology* **120,** 1459-67.

Saheki, T., Kobayashi, K., Iijima, M., Nishi, I., Yasuda, T., Yamaguchi, N.*, et al.* (2002). Pathogenesis and pathophysiology of citrin (a mitochondrial aspartate glutamate carrier) deficiency. *Metab Brain Dis* **17,** 335-46.

Sambrotta, M., Strautnieks, S., Papouli, E., Rushton, P., Clark, B. E., Parry, D. A.*, et al.* (2014). Mutations in TJP2 cause progressive cholestatic liver disease. *Nat Genet* **46,** 326-8.

Sanger, F., Nicklen, S. &Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74,** 5463-7.

Savolainen, K., Kotti, T. J., Schmitz, W., Savolainen, T. I., Sormunen, R. T., Ilves, M.*, et al.* (2004). A mouse model for alpha-methylacyl-CoA racemase deficiency: adjustment of bile acid synthesis and intolerance to dietary methyl-branched lipids. *Hum Mol Genet* **13,** 955-65.

Sawada, N. (2013). Tight junction-related human diseases. *Pathol Int* **63,** 1-12.

Setchell, K. D., Heubi, J. E. &Bove, K. E. (2008). Bile Acid Synthesis and Metabolism. Walker's Pediatric Gastrointentestinal Disease 5th edition.

Setchell, K. D., Heubi, J. E., Bove, K. E., O'Connell, N. C., Brewsaugh, T., Steinberg, S. J.*, et al.* (2003). Liver disease caused by failure to racemize trihydroxycholestanoic acid: gene mutation and effect of bile acid therapy. *Gastroenterology* **124,** 217-32.

Setchell, K. D., Heubi, J. E., Shah, S., Lavine, J. E., Suskind, D., Al-Edreesi, M.*, et al.* (2013). Genetic defects in bile acid conjugation cause fat-soluble vitamin deficiency. *Gastroenterology* **144,** 945-955 e6; quiz e14-5.

Shapira, R., Hadzic, N., Francavilla, R., Koukulis, G., Price, J. F. &Mieli-Vergani, G. (1999). Retrospective review of cystic fibrosis presenting as infantile liver disease. *Arch Dis Child* **81,** 125-8.

Shendure, J.Ji, H. (2008). Next-generation DNA sequencing. *Nat Biotechnol* **26,** 1135-45.

Sheth, B., Nowak, R. L., Anderson, R., Kwong, W. Y., Papenbrock, T. &Fleming, T. P. (2008). Tight junction protein ZO-2 expression and relative function of ZO-1 and ZO-2 during mouse blastocyst formation. *Exp Cell Res* **314,** 3356-68.

Siggs, O. M., Schnabl, B., Webb, B. &Beutler, B. (2011). X-linked cholestasis in mouse due to mutations of the P4-ATPase ATP11C. *Proc Natl Acad Sci U S A* **108,** 7890-5.

Sinal, C. J., Tohkin, M., Miyata, M., Ward, J. M., Lambert, G. &Gonzalez, F. J. (2000). Targeted disruption of the nuclear receptor FXR/BAR impairs bile acid and lipid homeostasis. *Cell* **102,** 731-44.

Singh, D., Solan, J. L., Taffet, S. M., Javier, R. &Lampe, P. D. (2005). Connexin 43 interacts with zona occludens-1 and -2 proteins in a cell cycle stage-specific manner. *J Biol Chem* **280,** 30416-21.

Sperber, I. (1959). Secretion of organic anions in the formation of urine and bile. *Pharmacol Rev* **11,** 109-34.

Spinner, N. B., Colliton, R. P., Crosnier, C., Krantz, I. D., Hadchouel, M. &Meunier-Rotival, M. (2001). Jagged1 mutations in alagille syndrome. *Hum Mutat* **17,** 18-33.

Strachan, T.Read, A. (2011). Human Genetic Variability and Its Consequences. Human Molecular Genetcs, 4th edition.

Strautnieks, S. S., Bull, L. N., Knisely, A. S., Kocoshis, S. A., Dahl, N., Arnell, H., *et al.* (1998). A gene encoding a liver-specific ABC transporter is mutated in progressive familial intrahepatic cholestasis. *Nat Genet* **20,** 233-8.

Strautnieks, S. S., Byrne, J. A., Pawlikowska, L., Cebecauerova, D., Rayner, A., Dutton, L., *et al.* (2008). Severe bile salt export pump deficiency: 82 different ABCB11 mutations in 109 families. *Gastroenterology* **134,** 1203-14.

Sulonen, A. M., Ellonen, P., Almusa, H., Lepisto, M., Eldfors, S., Hannula, S., *et al.* (2011). Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol* **12,** R94.

Tapia, R., Huerta, M., Islas, S., Avila-Flores, A., Lopez-Bayghen, E., Weiske, J., *et al.* (2009). Zona occludens-2 inhibits cyclin D1 expression and cell proliferation and exhibits changes in localization along the cell cycle. *Mol Biol Cell* **20,** 1102-17.

Thorvaldsdottir, H., Robinson, J. T. &Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14,** 178-92.

Tien, A. C., Rajan, A. &Bellen, H. J. (2009). A Notch updated. *J Cell Biol* **184,** 621-9.

Trauner, M., Meier, P. J. &Boyer, J. L. (1998). Molecular pathogenesis of cholestasis. *N Engl J Med* **339,** 1217-27.

Traweger, A., Fuchs, R., Krizbai, I. A., Weiger, T. M., Bauer, H. C. &Bauer, H. (2003). The tight junction protein ZO-2 localizes to the nucleus and interacts with the heterogeneous nuclear ribonucleoprotein scaffold attachment factor-B. *J Biol Chem* **278,** 2692-700.

Traweger, A., Lehner, C., Farkas, A., Krizbai, I. A., Tempfer, H., Klement, E., *et al.* (2008). Nuclear Zonula occludens-2 alters gene expression and junctional stability in epithelial and endothelial cells. *Differentiation* **76,** 99-106.

Treyer, A.Müsch, A. (2013). Hepatocyte polarity. *Comprehensive Physiology* **3,** 243-287.

Tsukita, S., Katsuno, T., Yamazaki, Y., Umeda, K., Tamura, A. &Tsukita, S. (2009). Roles of ZO-1 and ZO-2 in establishment of the belt-like adherens and tight junctions with paracellular permselective barrier function. *Ann N Y Acad Sci* **1165,** 44-52.

Turnpenny, P. D.Ellard, S. (2012). Alagille syndrome: pathogenesis, diagnosis and management. *Eur J Hum Genet* **20,** 251-7.

Umeda, K., Ikenouchi, J., Katahira-Tayama, S., Furuse, K., Sasaki, H., Nakayama, M.*, et al.* (2006). ZO-1 and ZO-2 independently determine where claudins are polymerized in tight-junction strand formation. *Cell* **126,** 741-54.

Van der Bliek, A. M., Baas, F., Ten Houte de Lange, T., Kooiman, P. M., Van der Velde-Koerts, T. &Borst, P. (1987). The human mdr3 gene encodes a novel P-glycoprotein homologue and gives rise to alternatively spliced mRNAs in liver. *EMBO J* **6,** 3325-31.

van Mil, S. W., Houwen, R. H. &Klomp, L. W. (2005). Genetics of familial intrahepatic cholestasis syndromes. *J Med Genet* **42,** 449-63.

van Mil, S. W., Klomp, L. W., Bull, L. N. &Houwen, R. H. (2001). FIC1 disease: a spectrum of intrahepatic cholestatic disorders. *Semin Liver Dis* **21,** 535-44.

Van Mil, S. W., Milona, A., Dixon, P. H., Mullenbach, R., Geenes, V. L., Chambers, J.*, et al.* (2007). Functional variants of the central bile acid sensor FXR identified in intrahepatic cholestasis of pregnancy. *Gastroenterology* **133,** 507-16.

van Mil, S. W., van der Woerd, W. L., van der Brugge, G., Sturm, E., Jansen, P. L., Bull, L. N.*, et al.* (2004). Benign recurrent intrahepatic cholestasis type 2 is caused by mutations in ABCB11. *Gastroenterology* **127,** 379-84.

Verhulst, P. M., van der Velden, L. M., Oorschot, V., van Faassen, E. E., Klumperman, J., Houwen, R. H., *et al.* (2010). A flippase-independent function of ATP8B1, the protein affected in familial intrahepatic cholestasis type 1, is required for apical protein expression and microvillus formation in polarized epithelial cells. *Hepatology* **51,** 2049-60.

Walsh, T., Pierce, S. B., Lenz, D. R., Brownstein, Z., Dagan-Rosenfeld, O., Shahin, H., *et al.* (2010). Genomic duplication and overexpression of TJP2/ZO-2 leads to altered expression of apoptosis genes in progressive nonsyndromic hearing loss DFNA51. *Am J Hum Genet* **87,** 101-9.

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

Wang, W., Wei, Z., Lam, T. W. &Wang, J. (2011). Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. *Sci Rep* **1,** 55.

Wilcke, M., Johannes, L., Galli, T., Mayau, V., Goud, B. &Salamero, J. (2000). Rab11 regulates the compartmentalization of early endosomes required for efficient transport from early endosomes to the trans-golgi network. *J Cell Biol* **151,** 1207-20.

Wilcox, E. R., Burton, Q. L., Naz, S., Riazuddin, S., Smith, T. N., Ploplis, B., *et al.* (2001). Mutations in the gene encoding tight junction claudin-14 cause autosomal recessive deafness DFNB29. *Cell* **104,** 165-72.

Wu, J., Yang, Y., Zhang, J., Ji, P., Du, W., Jiang, P., *et al.* (2007). Domain-swapped dimerization of the second PDZ domain of ZO2 may provide a structural basis for the polymerization of claudins. *J Biol Chem* **282,** 35988-99.

Xu, J., Kausalya, P. J., Phua, D. C., Ali, S. M., Hossain, Z. &Hunziker, W. (2008). Early embryonic lethality of mice lacking ZO-2, but Not ZO-3, reveals critical and nonredundant roles for individual zonula occludens proteins in mammalian development. *Mol Cell Biol* **28,** 1669-78.

Zimmerman, S. P., Hueschen, C. L., Malide, D., Milgram, S. L. &Playford, M. P. (2013). Sorting nexin 27 (SNX27) associates with zonula occludens-2 (ZO-2) and modulates the epithelial tight junction. *Biochem J* **455,** 95-106.

Zsembery, A., Jessner, W., Sitter, G., Spirli, C., Strazzabosco, M. &Graf, J. (2002). Correction of CFTR malfunction and stimulation of Ca-activated Cl channels restore HCO3- secretion in cystic fibrosis bile ductular cells. *Hepatology* **35,** 95-104.

Appendix

Appendix I. Primer sequences

Gene name	No exon	Strand	Primer sequence (5' to 3')	Annealing Temperature for PCR (°C)
ABCB11	Exon 12	Forward	GATACATGCAAACTAAGAGGC	54
		Reverse	GCCTGAGATTAAGCCAACACC	
ABCB11	Exon 23	Forward	CCAGATGATGCATTCTCTGAT	54
		Reverse	AGCTATTGTAAGACACCAAGC	
ABCB11	Exon 28	Forward	ATCCTCTCTTATGTTGAGCC	54
		Reverse	CTGGTGCGTCATGTGTGTC	
ABCB4	Exon 7	Forward	TTAGGTGGGGAAGATGTTATTC	54
		Reverse	CTGGATGTAGTTTCAACTGAC	
ABCB4	Exon 16	Forward	AGCTATCCTTGATTGAGAAGC	54
		Reverse	GAACTGATTTCAATTCATTGTC	
ABCB4	Exon 25	Forward	TTGTCTAATCTCACCTATACC	54
		Reverse	CCAGATATGGTGCCAGTTG	
ATP8B1	Exon 18	Forward	GCAACCAGGATGTATAATTAG	54
		Reverse	GATCAGGAAAGGATGCAGAAG	
TJP2	Exon 5	1-Forward	GAACCTGAACCTTAGTGAG	56
		1-Reverse	TAGTCCTGGTCAATGCTC	
TJP2	Exon 5	2-Forward	CCTGGACCACGACTTTG	56
		2-Reverse*	CTGGATGACAGAGCAACAC	
TJP2	Exon 13	Forward	GTCTAGTACTGTAACTTGTAC	56
		Reverse	GACAAGTAATGAACTCATCATG	
TJP2	Exon 21	Forward	CCTGACATTCCTAATAGACTAG	58
		Reverse	GGCAATCTGAAGTGACATACAC	
NOTCH2	Exon 1	Forward	ACACACGAGGCTGCTTCGT	58
		Reverse	CCGGCGATGTCCAAACTCTT	
NOTCH2	Exon 2	Forward	AGAGAAATAAGAGCATCACTC	58
		Reverse	AGAGTGCAATGATGTGATG	
NOTCH2	Exon 3	Forward	CCTGCCAGGACTCAAAAGGA	58
		Reverse	TATCTGCTGAAGGTAGGGAAC	

Appendix Table I.1 Primer sequences and annealing temperature used for sequencing TruSeq custom amplicon (TSCA) gaps. * see Appendix Table I.2

No exon	Strand	Primer sequence (5' to 3')	Annealing Temperature for PCR (°C)
Exon 5	1-Forward	GAACCTGAACCTTAGTGAG	56
	1-Reverse	TAGTCCTGGTCAATGCTC	
Exon 5	2-Forward	CCTGGACCACGACTTTG	56
	2-Reverse*	CTGGATGACAGAGCAACAC	
Exon 9	Forward	CTCTTGTACATGTCATTGTG	54
	Reverse	AAAGTCAGAACATGTGCTG	
Exon 14	Forward	AGATTTCACAGAGACCCAG	54
	Reverse	GTTAGAACTTGACCATACAGTC	
Breakpoint (6-16)	Forward	TCACCCACTGAATTCCTTTC	54
	Reverse	TCTCATGGTTTTGACACCTTG	
Exon 15	Forward	GATAGTGAAGGCATATTCTTCAG	54
	Reverse	GGTGTTCATGATTCTTCCAAC	
Exon 16	Forward	CACTGAATCTTGTTAGGAGAC	54
	Reverse	GATTAAAAGTACAGATACTATCC	
Exon 17	Forward	AACAGAGCAAGAATCTATCCC	54
	Reverse	GAACACAATTTCAAATTCTC	

Appendix Table I.2 Primer sequences and annealing temperature for Sanger sequencing validation of *TJP2* mutations

*The 2-reverse primer was used only for PCR. Due to a long stretch of T, a nested reverse primer (5'- CACATCAAGCATGCCTAC -3') was adopted for Sanger sequencing.

No pair of primers	No exon	Strand	Primer sequence (5' to 3')
1	Exon 5	Forward	GAGCATTGACCAGGACTA
	Exon 17	Reverse	GAACACAATTTCAAATTCTC
2	Intron 6	Forward	GAGCGAACGAAGGTAGGCATG
	Exon 17	Reverse	CCACATCATGGATCTTTATGG
3	Intron 6	Forward	GTGTTGCTCTGTCATCCAGGC
	Exon 17	Reverse	GAACACAATTTCAAATTCTC
4	Exon 5	Forward	GAGCATTGACCAGGACTA
	Exon 17	Reverse	CCACATCATGGATCTTTATGG

Appendix Table I.3 Primer sequences used for *TJP2* breakpoint identification in family 12

Family number	No Exon	Strand	Primer sequence (5' to 3')	Annealing Temperature for PCR (°C)
11	Exon 5	Forward	CCTGGACCACGACTTTG	56
	Exon 17-18	Reverse	ATTTAGGTTGATTGTAGCTG	
12	Exon 12	Forward	GAACACACAGGATTTCAGAG	56
	Exon 16	Reverse	CTCCAGTGGATTTCTCAGATC	

Appendix Table I.4 Primer sequences for cDNA sequencing of *TJP2*

Appendix II Variations identified in WES data

Case number	Chr	Chr position (start)	Chr position (end)	Gene name	No exons	Read ratio	BF
12a	chr9	71840241	71853695	*TJP2*	11	0	212
17	chr4	120057688	120161076	*MYOZ2; USP53*	6	0	201
18	chr9	126135374	126136212	*CRB2*	2	0.159	27

Appendix Table II.1 Copy number variations identified in WES filtered data as possible disease-causing

After applying the filtering process described in section 2.3.6, three homozygous structural deletion were identified. Their details are here summarised by chromosome (chr) location, gene names in accordance to HUGO database Nomenclature Committee (HGNC) and number of exons involved. Read ratio represents the number of reads observed divided by the number of reads expected in that specific area. Statistical significance is defined by the Bayes factor (BF)

Gene	RefSeq ID	Nucleotide change	Predicted amino acid change	dbSNP
ADAM29	NM_001130705	c.1879C>T	p.Pro627Ser	rs78280171
AGBL1	NM_152336	c.3043C>G	p.Leu1015Val	novel
AKAP13	NM_007200	c.253G>T	p.Ala85Ser	rs116551873
ANKMY1	NM_016552	c.1228G>A	p.Glu410Lys	rs3796119
ANP32E	NM_030920	c.562_570del	p.188_190del	novel
COL8A2	NM_005202	c.464G>A	p.Arg155Gln	rs75864656
CRIPAK	NM_175918	c.599C>A	p.Ala200Asp	rs79550423
CXorf65	NM_001025265	c.85C>A	p.Arg29Ser	novel
ERN2	NM_033266	c.370G>A	p.Val124Ile	rs150200841
MXRA5	NM_015419	c.8212G>A	p.Glu2738Lys	novel
MYO10	NM_012334	c.5642G>A	p.Arg1881Gln	.
NAP1L2	NM_021963	c.640_641insGAG	p.Glu214delinsGluGlu	novel
PLIN1	NM_002666	c.1043C>T	p.Ser348Lue	rs8179071
RC3H1	NM_172071	c.1354C>T	p.Arg452Cys	novel
RGAG1	NM_020769	c.3062C>A	p.Ala1021Asp	novel
RHD	NM_016124	c.1189C>A	p.His397Asn	novel
SEMA4A	NM_001193300	c.2167C>T	p.Arg723Cys	rs199933282
SESTD1	NM_178123	c.724C>A	p.Leu242Ile	novel
SH3D21	NM_001162530	c.31G>A	p.Ala11Thr	novel
SPTA1	NM_003126	c.3472C>T	p.Arg1158Trp	.
TMEM200B	NM_001003682	c.881A>G	p.Try294Cys	.
TMEM54	NM_033504	c.248G>A	p.Arg83His	novel

Appendix Table II.2 Variants identified in patient 1 by WES filtered data

All the variations are homozygous. Gene names are in accordance to HUGO database Nomenclature Committee (HGNC).

Gene	RefSeq ID	Nucleotide change	Predicted amino acid change	dbSNP
AMACR	NM_014324	c.877T>C	p.Cys293Arg	novel
C5orf62	NM_032947	c.151C>T	p.Arg51WTrp	rs2278396
C6	NM_001115131	c.T848T>G	p.Ile283Ser	rs142653101
CES1	NM_001266	c.52+1G>A	p.?	rs139063675
CSF1R	NM_005211	c.2807_2808insGCAGCA	p.G936delinsGlySerser	novel
GPR151	NM_194251	c.T69T>G	p.Phe23Leu	rs144066680
PCDHB3	NM_018937	c.2257G>T	p.Val753Leu	novel
PIWIL3	NM_001008496	c.1921G>A	p.Val641Met	rs148034582
PLXNA3	NM_017514	c.5570G>A	p.Arg1857Gln	rs201267859
PRDM6	NM_001136239	c.784A>G	p.Ile262Val	rs139705595
SLC22A1	NM_003057	c.113G>A	p.Gly38Asp	rs35888596
SMTN	NM_001207018	c.1094C>A	p.Ala365Asp	rs201342047
TNFRSF11A	NM_003839	c.671C>G	p.Ala224Gly	novel
ZNF449	NM_152695	c.94C>T	p.Arg32Cys	novel

Appendix Table II.3 Variants identified in patient 3 by WES filtered data

All the variations are homozygous. Gene names are in accordance to HUGO database Nomenclature Committee (HGNC).

Gene	RefSeq ID	Nucleotide change	Predicted amino acid change	dbSNP
ABAT	NM_000663	c.910G>A	p.Ala304Thr	.
ABCA7	NM_019112:	c.1856T>A:	p.Ile619Asn	.
APPL1	NM_012096	c.2018C>G	p.Ser673Cys	rs138485817
BSN	NM_003458	c.10658A>G	p.Tyr3553Cys	novel
C2orf54	NM_001085437	c.175G>A	p.Val59Met	novel
COL6A5	NM_153264	c.3342C>G	p.Ile1114Met	rs1353613
COL6A6	NM_001102608	c.6025C>T	p.Arg2009Trp	.
DGKK	NM_001013742	c.1916C>T	p.Pro639Leu	novel
DNASE1L2	NM_001374	c.721G>A	p.Gly241Ser	novel
HEPH	NM_138737	c.74A>C	p.Lys25Thr	rs143028997
HGS	NM_004712	c.724C>G	p.Leu242Val	novel
HIST1H1C	NM_005319	c.512C>T	p.Ala171Val	rs79483116
MEFV	NM_000243	c.1105C>T	p.Pro369Ser	rs11466023
MMP15	NM_002428	c.1114G>A	p.Gly372Arg	novel
MRS2	NM_020662	c.1328G>A	p.Arg443His	rs114959453
OR4C11	NM_001004700	c.82A>G	p.Ile28Val	novel
PARP14	NM_017554	c.2623C>G	p.His875Asp	novel
PRKAR2A	NM_004157	c.401G>A	p.Cys134Tyr	rs199498651
PRSS21	NM_144956	c.692C>A	p.A231DAsp	rs139303479
RPP40	NM_006638	c.443T>C	p.Ile148Tyr	rs140492752
SPINK8	NM_001080525	c.126C>A	p.Cys42*	novel
SRRM2	NM_016333	c.8215C>T	p.Pro2739Ser	rs138495768
TBCD	NM_005993	c.1403G>C	p.Cys468Ser	novel
TNXB	NM_019105	c.11533G>A	p.Gly3845Ser	rs199688928
ZNF77	NM_021217	c.1025A>G	p.His342Arg	novel

Appendix Table II.4 Variants identified in patient 8 by WES filtered data

All the variations are homozygous. Gene names are in accordance to HUGO database Nomenclature Committee (HGNC).

Gene	RefSeq ID	Nucleotide change	Predicted amino acid change	dbSNP
ACAD11	NM_032169	c.1057C>T	p.Gln353*	NOVEL
BCAM	NM_001013257	c.1310C>G	p.Tyr437Ser	NOVEL
DMD	NM_004009	c.3257A>T	p.Gln1086Leu	NOVEL
HIF3A	NM_152794	c.163C>2A	p.Asp544Glu	NOVEL
MAMDC2	NM_153267	c.1085G>A	p.Arg362Gln	.
MEX3D	NM_203304	c.1531C>T	p.Pro511Ser	NOVEL
OR4S2	NM_001004059	c.605G>T	p.Gly202Val	NOVEL
PIKFYVE	NM_015040	c.373G>A	p.Ala125Thr	NOVEL
PLIN4	NM_001080400	c.3080C>T	p.Ala1027Val	RS145519622
PLK5	NM_001243079	c.569-1G>A	p.?	RS150328666
SH3KBP1	NM_031892	c.1621C>G	p.Gln541Glu	RS61761898
TMC1	NM_138691	c.247_249del	p.83_83del	.
ZNF222	NM_001129996	c.654_655del	p.218_219del	NOVEL
ZNF226	NM_001032372	c.1412G>A	p.Gly471Glu	RS61742481

Appendix Table II.5 Variants identified in patient 12 by WES filtered data

All the variations are homozygous. Gene names are in accordance to HUGO database Nomenclature Committee (HGNC).

Gene	RefSeq ID	Nucleotide change	Predicted amino acid change	dbSNP
CNTN4	NM_001206955	c.2990C>T	p.Ala997Val	rs184575174
KLF16	NM_031918	c.112G>A	p.Ala38Thr	.
MKL2	NM_014048	c.4G>T	p.Asp2Tyr	rs75549726
MNX1	NM_005515	c.373_374insGCCGCCGCC	p.Ala125delinsAlaAlaAlaAla	novel
NOXA1	NM_006647	c.409C>A	p.Leu137Ile	rs77546115
QRFP	NM_198180	c.260G>A	p.Arg87His	rs139997194
TBC1D3C	NM_001001418	c.1082-1G>A	p.?	novel

Appendix Table II.6 Variants identified in patient 13 by WES filtered data

All the variations are homozygous. Gene names are in accordance to HUGO database Nomenclature Committee (HGNC).

Gene	RefSeq ID	Nucleotide change	Predicted amino acid change	dbSNP
ATOH8	NM_032827	c.386_387insGCCGCC	p.Pro129delinsProProPro	.
FAM209B	NM_001013646	c.110C>A	p.Pro37Gln	RS200150839
FAM209B	NM_001013646	c.113G>A	p.Cys38YTyr	RS201542308
DOK5	NM_018431	c.599G>A	p.Arg200Lys	NOVEL
GK2	NM_033214	c.1396G>A	p.Ala466Thr	NOVEL
HECW1	NM_015052	c.3551T>C	p.Val1184Ala	RS116945469
LAMA3	NM_198129	c.1694G>C	p.Arg565Pro	.
MAP4K4	NM_145687	c.3063G>C	p.Met1021Ile	NOVEL
NAP1L5	NM_153757	c.427G>A	p.Glu143Lys	RS187220478
NOTCH3	NM_000435	c.1690G>A	p.Ala564Thr	.
NOTCH3	NM_000435	c.1487C>T	p.Pro496Leu	RS11670799
PDCD11	NM_014976	c.133A>G	p.Lys45Glu	RS150893869
RSBN1	NM_018364	c.266G>T	p.Arg89Leu	NOVEL
SH2D6	NM_198482	c.268G>A	p.Val90Met	RS149721029
WDR83	NM_001099737	c.832G>A	p.Gly278Ser	RS34373915
ZC3H6	NM_198581	c.1763A>G	p.Tyr588Cys	RS75832760
ZNF844	NM_001136501	c.1044_1048del	p.348_350del	.
ZSWIM4	NM_023072	c.824A>G	p.Tyr275Cys	RS149065965

Appendix Table II.7 Variants identified in patient 17 by WES filtered data

All the variations are homozygous. Gene names are in accordance to HUGO database Nomenclature Committee (HGNC).

Gene	RefSeq ID	Nucleotide change	Predicted amino acid change	dbSNP
CABLES2	NM_031215	c.215C>G	p.Pro72Arg	RS184691263
INF2	NM_022489	c.1280_1285del	p.427_429del	.
KPRP	NM_001025231	c.52G>A	p.Val18Ile	RS147807095
PODXL	NM_001018111	c.71_72insCGTCGT	p.Pro24delinsProSerser	NOVEL
RPGR	NM_001034853	c.2689_2691del	p.897_897del	NOVEL
SLC7A8	NM_012244	c.86C>T	p.Ser29Phe	RS149980964

Appendix Table II.8 Variants identified in patient 18 by WES filtered data

All the variations are homozygous. Gene names are in accordance to HUGO database Nomenclature Committee (HGNC).

Appendix III. Primers and probes for qPCR

Target Name	Forward primer (5' to 3')	Reverse primer (5' to 3')	No UPL probe
a	CAATTGTCATTTCTGATGTGCTC	TTGACCATGACCACTCTGTCA	40
b	TGTTGCTGGCATTCAAGAAG	CCTCTGAAATCCTGTGTGTTCA	15
c	CTACGCGGGACCTGTGTC	ACAGTGTACTGTTCCCATATCAGC	24
d	GGTGCAGCACGAGGAGAG	GCGGGCTATTGTCCCTAAG	21
e	GGAAGCACAGAGCTGTACCA	GGAATCCTTTTGTAGGGTCACA	57
f	CTCTCTGTACTCTTAGATGTCCTTGC	CAAGAGCTTTGGCAAGATTAGAA	60

Appendix Table III.1 Primers and probes used for quantitative analysis of *TJP2* transcripts

The exact location of each target is shown in Figure 3.4.2. The target name “a” was adopted also for TJP2 quantitative expression analysis described in section 3.4.1.

Gene Name	Forward primer (5' to 3')	Reverse primer (5' to 3')	No UPL probe
TJP1	TCAGACAGGCGGTCAGTG	ATATGGCTTGCCAATCGAAG	20
TJP3	TCTTCATCAAGCACATTACAGATTC	GGCTAGACACCCCGTTGAT	4
CLDN1	TTGACTCCTTGCTGAATCTGAG	GGCCACAAAGATTGCTATCAC	79
CLDN2-202	GCTCACAGGCCATTCAGG	GTCTCTCTGCCAGGCTGACT	63
CLDN3	CACTGCCACAGGACCTTCA	GTCCTGCACGCAGTTGGT	49
CLDN7	AAAATGTACGACTCGGTGCTC	CACTTCATGCCCATCGTG	72
CLDN10	TTGATCCTCTCTTTGTTGAGCA	AAAGCAAAATATGACACCACCA	51
CLDN11	CCCGGTGTGGCTAAGTACAG	CAACAAGGGCGCAGAGAG	11
CLDN12	AGCTGTTTTGGAACTGTCAGG	TTCCACACAGGAAGGAAAGG	5
CLDN15	CTCTCCAGGAAAGCCAAGC	TGATGTTGAAGGCGTACCAG	49
OCLN	CACTATGAGACAGACTACACAACTGG	TTGATCTGAAGTGATAGGTGGATATT	25
F11R	GCAGCCGTCCTTGTAACC	GGCTGGCTGTAAATCACCTT	2
MARVELD2	GGCTGTCCTGAGGAAGTTTG	GGATTCTTGAAATTCTCTCATGTTC	42
ACBT	ATTGGCAATGAGCGGTTC	CGTGGATGCCACAGGACT	11
CGN	GCAAGATGCAACCCAGGA	AGTCTCCTCCAGCTCCCTCT	29
CDH1	GGAGCCAGACACATTTATGGA	GTGGAAATGGCACCAGTGT	19
CTNNA1	CGTGATGACCGTCGTGAG	CTTCTTTCTTTACGTCCAGCATT	53
CTNNB1	TGTTAAATTCTTGGCTATTACGACA	CCACCACTAGCCAGTATGATGA	8
GJA1	CGTGACTTCACTACTTTTAAGCAAA	CAGGATTCGGAAAATGAAAAGTA	5
RAB13	CGCTTTGCAGAGGACAACTT	CCACAGTGCGGATCTTGA	22
SAFB	CGGCAACAAGAGCGTTTT	TCGTCAGGATTACCACCTTCA	41
SNX27	ACTGCATGCCTGCACTGA	TGCATCTCATCCCATTCAAA	78

Appendix Table III.2 Primers and probes used for the quantitative expression analysis of the selected panel of genes.

Gene names are in accordance to HUGO database Nomenclature Committee (HGNC).

Publication

Sambrotta M, Strautnieks S, Papouli E, Rushton P, Clark BE, Parry DA, Logan CV, Newbury LJ, Kamath BM, Ling S, Grammatikopoulos T, Wagner BE, Magee JC, Sokol RJ, Mieli-Vergani G; University of Washington Center for Mendelian Genomics, Smith JD, Johnson CA, McClean P, Simpson MA, Knisely AS, Bull LN, Thompson RJ. (2014). Mutations in TJP2 cause progressive cholestatic liver disease. *Nature Genetics* 46, 326-8.
doi: 10.1038/ng.2918.

Presentations to conferences

Oral presentation

Sambrotta, M., Strautnieks, S., Brett, L., Davison, S., & Thompson, R. Mutations in Tight Junction Protein 2 Underlie a Spectrum of Cholestatic Liver Disease.
Digestive Disease Week, May 2014, Chicago, IL, United States.

Sambrotta M., Strautnieks S., Papouli E., Rushton P., Clark B.E., Parry D.A., Brett L., Logan C.V., Newbury L.J., Kamath B.M., Ling S., Grammatikopoulos T., Wagner B.E., Magee J.C., Sokol R.J., Mieli-Vergani G., University of Washington Center for Mendelian Genomics, Smith J.D., Johnson C.A., Davison S., McClean P., Simpson M.A., Knisely A.S., Bull L.N., Thompson R.J. TJP2 deficiency: a new cholestatic liver disease.
European Society of Human Genetics Conference, June 2014, Milan, Italy.

Sambrotta M., Knisely A.S., Thompson R.J. Genetic, clinical, and histopathologic features of intermediate tight junction protein 2 deficiency.
British Association of the Study of the Liver, September 2014, Newcastle upon Tyne, United Kingdom.

Poster presentation

Sambrotta M., Knisely A.S., Thompson R.J. Genetic, clinical, and histopathologic features of intermediate tight junction protein 2 deficiency.
American Association of the Study of Liver Disease, November 2014, Boston, MA, United States