David Miguel Ferreira Francisco

Licenciado em Biologia Humana

Familial Genetics of Primary Spontaneous Pneumothorax

Dissertação para obtenção do Grau de Mestre em Genética Molecular e Biomedicina

Orientador: Sofia A. Oliveira, PhD, Instituto Medicina Molecular

Co-orientador: Inês Sousa, PhD, Instituto Medicina Molecular

Júri:

Presidente: Prof. Doutora Paula Gonçalves
Arguente: Prof. Doutora Ana Rita Grosso

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Setembro, 2015

David Miguel Ferreira Francisco

Licenciado em Biologia Humana

Familial Genetics of Primary Spontaneous Pneumothorax

Dissertação para obtenção do Grau de Mestre em Genética Molecular e Biomedicina

Orientador: Sofia A. Oliveira, PhD, Instituto Medicina Molecular

Co-orientador: Inês Sousa, PhD, Instituto Medicina Molecular

Júri:

Presidente: Prof. Doutora Paula Gonçalves
Arguente: Prof. Doutora Ana Rita Grosso

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Setembro, 2015

Familial Genetics of Primary Spontaneous Pneumothorax

“If we knew what it was we were doing, it would not be called research, would it?”

- (Probably) not Albert Einstein

Acknowledgments

First, I would like to thank Doctor Sofia Oliveira for giving me the opportunity to perform my masters in her group at *Instituto de Medicina Molecular* (IMM), and also for all the support, help and advices without which this project would have been impossible to carry out.

I would also like to thank Doctor Inês Sousa, my *"Sensei"*, for everything. Her help, supervision, assistance, friendship, limitless patience and above all for teaching me (no matter how stubborn I was being!) and for providing me with the best learning experience of my life, both academically and personally.

To Vânia Francisco, MSc, for being my north star (the person that provides me all guidance in life), for always being there when I needed, and for pushing me when necessary, while reminding me that sometimes the best you can do is to stop and rest.

To Maria João Rocha, for the help with the initial DNA extractions.

A special thanks to Doctor Daniel Sobral and his bioinformatics team (Renato Alves, Paulo Almeida and Isabel Marques) from *Instituto Gulbenkian de Ciência*, for giving me the opportunity to carry out all the bioinformatics analysis in their unit. Especially to Msc Patrícia Santos for all the support, advice, teaching and help with the testing of some tools.

To everyone at IMM for making this such a pleasant experience and in particular to Joana Tavares, MSc, for helping me test some bioinformatics tools. To Alice Melão, MSc, for helping me with the sequencing codes. Additionally, to everyone in the Computational Biology and Bioinformatics Seminars for constantly inspiring me and for the advice. Finally, to the IT department, specifically Pedro Eleutério for all the help with the technical issues of the IMM servers used.

To Dr. Salvato Feijó and his team for collecting the patients information and blood samples as well as their respective family members who agreed to participate in this project.

To everyone in the 4th NGS Course in Bertinoro from the European Society of Human Genetics for the amazing experience and advices. A special thanks to Doctor Peter N. Robinson, Doctor Joris Veltman and Doctor Christian Glissen for all the suggestions that helped to complement this project.

To the organization of the Week of Sciences and Technologies from the University of Évora for inviting me to speak about this project.

To all of my friends from "Monte K", Évora, Lisboa and Torres Vedras. Particularly, a special thanks to Liliana Capinha, Francisca Pereira, Mariana Dias, Luís Ribeiro, Ricardo Santos, Wilson Patrocínio, Miguel Leal, Carlos Narciso, Adneusa Vieira, João Sabino, Rafael Coelho, Francisco Calado, Aurélio Ruel, Mafalda Azevedo Gomes, Liliana Piegas and Isabela Corina for bringing joy and fun into my life and putting up with me no matter how stressed and ill-humored I was.

Lastly, a special thanks to all of my family and specially my mother for always supporting me, while giving me the liberty to forge my path.

Without all of you this year would not have been the amazing experience that it was. Thank you!

Abstract

Primary Spontaneous Pneumothorax (PSP) is a complex disorder characterized by the presence of air in the pleural cavity that occurs without preceding trauma or known cause in individuals with no lung disease. Until now, very few variants have been described as causing familial PSP. Unraveling PSP genetic underpinnings is essential to understand its pathogenesis, and might contribute to future advances in research and therapy. Our main goal with this study was to identify genetic variants causing PSP in two Portuguese multiplex families.

We performed whole exome sequencing (WES) in a total of eight individuals (six affected, one unaffected carrier and one unaffected) using the SureSelect Target Enrichment System kit to capture DNA and the Hiseq2000 (Illumina's Solexa) platform for sequencing. Bioinformatics analysis was carried out on a local server with an optimized version of the DNAseq Genome Analysis Tool Kit (GATK) workflow, designed by us. All steps in the analysis were subject to conservative and extensive quality control stages to ensure the maximum confidence in our results. We selected nine variants from family 1 for technical validation using Sanger sequencing: c.10855G>C in *DYNC2H1*, c.1440G>C in *STOX2*, c.*106G>A in *DCLK2*, c.*162delT in *TRAPPC11*, c.1067+3A>C in *TDO2*, c.1015G>A in *MAP2K3*, c.280T>C in *RFPL1*, c.582_583dupAC in *NOP16* and the intergenic variant g.34729259_34729260insACCAGC. Seven were successfully validated (c.10855G>C, c.1440G>C, c.*106G>A, c.*162delT, c.1067+3A>C, c.280T>C, and g.34729259_34729260insACCAGC) and confirmed our initial findings. Family 2 was used for replication of validated variants/genes from family 1. One variant (c.-124C>A in *STOX2*) was selected and successfully validated in family 2.

In conclusion, we found that variants in *STOX2* (Storkhead Box 2) may cause PSP in these families, but additional studies using a larger number of families are now warranted to investigate the pathogenic potential of the remaining variants.

Keywords: Primary Spontaneous Pneumothorax (PSP); lung; Whole Exome Sequencing (WES); Bioinformatics analysis; Sanger sequencing; *STOX2.*

Resumo

O Pneumotórax Espontâneo Primário (PEP) é uma doença complexa caracterizada pela presença de ar na cavidade pleural que ocorre sem trauma prévio ou causas conhecidas em indivíduos sem doença pulmonar. Até agora, poucas variantes foram descritas como sendo causais para PEP. Desvendar os factores genéticos subjacentes ao PEP é essencial para perceber os seus mecanismos patogénicos, e pode contribuir para futuros avanços na investigação e tratamento. O principal objetivo deste estudo foi identificar variantes genéticas que possam causar PEP em duas famílias multiplex portuguesas.

Realizámos *whole exome sequencing* (WES) de oito indivíduos (seis doentes, um portador não afectado e um controlo familiar) usando o *SureSelect Target Enrichment System* kit para capturar DNA e a plataforma Hiseq2000 (Solexa, Illumina) para sequenciação. A análise bioinformática foi efectuada num servidor local usando uma versão otimizada do *Genome Analysis Tool Kit* (GATK) para DNAseq desenvolvida por nós. Todos os passos da análise foram sujeitos a rigorosas etapas de controlo de qualidade para assegurar a máxima confiança nos nossos resultados. Selecionámos nove variantes da família 1 para validação técnica por sequenciação de Sanger: c.10855G>C no *DYNC2H1*, c.1440G>C no *STOX2*, c.*106G>A no *DCLK2*, c.*162delT no *TRAPPC11*, c.1067+3A>C no *TDO2*, c.1015G>A no *MAP2K3*, c.280T>C no *RFPL1*, c.582_583dupAC no *NOP16* e a variante intergénica g.34729259_34729260insACCAGC. Sete destas variantes foram validadas por esta segunda técnica (c.10855G>C, c.1440G>C, c.*106G>A, c.*162delT, c.1067+3A>C, c.280T>C, e g.34729259_34729260insACCAGC), confirmando os nossos resultados iniciais. A família 2 foi utilizada para replicação das variantes/genes validados na família 1. Selecionámos uma variante (c.-124C>A no *STOX2*) que foi validada com sucesso.

Concluindo, foram encontradas variantes no *STOX2 (Storkhead Box 2)* que podem causar PSP nestas famílias, no entanto, estudos adicionais com um maior número de famílias serão necessários de forma a investigar a patogenicidade das restantes variantes.

Palavras-chave: Pneumotórax Espontâneo Primário (PSP); Pulmão; *Whole Exome Sequencing* (WES); Análise bioinformática; Sequenciação de Sanger; *STOX2.*

Table of Contents

List of Figures

List of Tables

List of abbreviations

1000 GP	1000 Genomes Project
A	Adenine
BAM	Binary Alignment Files
BHD	Birt-Hogg-Dubé syndrome
BGI	Beijing Genomics Institute
BLAT	Blast-Like Alignment Tool
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
BWA	Burrows-Wheeler Aligner
BWT	Burrows-Wheeler Transform
C	Citosine
cDNA	Complementary DNA
CDS	Coding DNA Sequence
CNV	Copy Number Variant
CTCF	CCCTC-Binding Factor
dbSNP	Single Nucleotide Polymorphism Database
DCLK2	Doublecortin-Like Kinase 2 gene
ddATP	Dideoxyadenosine triphosphate
ddCTP	Dideoxycytidine triphosphate
ddGTP	Dideoxyguanosine triphosphate
ddNTP	Dideoxyribonucleotide triphosphate
ddTTP	Dideoxy thymidine triphosphate
DNA	Deoxyribonucleic Acid
dNTPs	Deoxyribonucleotide triphosphates
DP	Read Depth
DYNC2H1	Dynein, Cytoplasmic 2, Heavy Chain 1 gene
EDTA	Ethylenediaminetetraacetic Acid
FET	Fisher's Exact Test
FLCN	Folliculin gene
FS	Fisher Strand
FSB	Fisher's Strand Bias
g	Relative centrifugal force

G	Guanine
GAT-I	GABA transporter-I
GATK	Genome Analysis Toolkit
GEO	Gene Expression Omnibus
GQ	Genotype Quality
GRC	Genome Reference Consortium
GRCh37	Genome Reference Consortium Human 37 (same as hg19)
GRCh38	Genome Reference Consortium Human 38 (same as hg38)
GWAS	Genome-Wide Association Study
hg19	Human Genome 19 (same as GRCh37)
hg38	Human Genome 38 (same as GRCh38)
HSM	Hospital Santa Maria
HGVS	Human Genome Variation Society
IFT	Intraflagellar transport
IGC	Instituto Gulbenkian de Ciência
IGV	Integrative Genomics Viewer
IMM	Instituto de Medicinal Molecular
InDel	Insertion or Deletion of bases in the DNA
LD	Linkage Disequilibrium
LINC00824	long intergenic non-protein coding RNA 824 gene
LM-PCR	Ligation Mediated Polymerase Chain Reaction
MAF	Minor Allele Frequency
MAP2K3	MItogen-Activated Protein Kinase 3 gene
MAPQ	Mapping Qualiy
MGI	Mouse Genome Informatics
MIM	OMIM number (see OMIM)
MIR1208	MicroRNA 1208 gene
MQRankSum	Mapping Quality Rank Sum Test
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
NOP16	NOP16 nucleolar protein gene
NSV	Non-Synonymous Variants
OMIM	Online Mendelian Inheritance in Man
PCR	Polymerase Chain Reaction

PolyPhen-2	Polymorphism Phenotyping version 2
PSP	Primary Spontaneous Pneumothorax
Q	Base quality value
QC	Quality Control
QD	Quality by Depth
QR code	Quick Response code
QUAL	Quality score for SNVs and InDels
RBC	Red Blood Cells
RFPL1	Ret Finger Protein-Like 1
RNA	Ribonucleic acid
RPM	Revolutions Per Minute
SAM	Sequence Alignment/Map
SD	Standard Deviation
SIFT	Sorting Intolerant From Tolerant
SLC1A3	Solute carrier family 1 (glial high affinity glutamate transporter), member 3
SLC6A1	Solute carrier family 6 member 1 gene
SLC8A1	Solute carrier family 8 (sodium/calcium exchanger), member 1
SNP	Single-Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SO	Sequence Onthology
SRTD	Short-rib thoracic dysplasia
SSP	Secondary Spontaneous Pneumothorax
STOX2	Storkhead Box 2 gene
STR	Short Tandem Repeat
T	Timine
TAE	Tris-Acetate-EDTA
TDO2	Tryptophan 2,3-Dioxygenase gene
Ti	Transitions
Tm	Melting temperature
TRAPPC11	Trafficking Protein Particle Complex, Subunit 11 gene
Tv	Transversions
UTR	Untranslated Region
UV	Ultra Violet light
VEP	Variant Effect Predictor

WES Whole-Exome Sequencing

WGS Whole-Genome Sequencing

1. Introduction

1.1 – Primary Spontaneous Pneumothorax

A pneumothorax an abnormal presence of air or gas in the pleural cavity (Figure 1) first recognized by Jean-Marc Gaspard Itard in 1803 (Itard, 1803) and first described by René Laennec, his tutor, in 1819 (Laennec, 1819).

Air does not normally enter the pleural space because while the intra-pleural pressures are negative through most of the respiratory cycle, the necessary pressure for this to happen is much lower than what is usually measured ,-54mmHg (Noppen and Schramel, 2002). To the best of our knowledge, for the air to be present in the pleural space, one of three events must occur: communication between alveolar spaces and pleura; direct or indirect communication between the atmosphere and the pleural space and/or presence of gas-producing organisms in the pleural space.

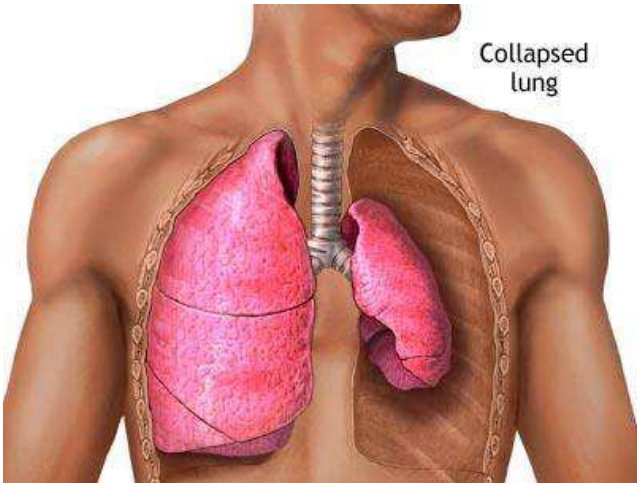

Figure 1 – Schematic example of a normal (right) and collapsed (left) lung (http://www.pennmedicine.org/encyclopedia/em_PrintImage.aspx?gcid=19589&ptid=2). Air enters in the pleural space and compresses the lung (left) occupying its space.

Pneumothoraces are divided into spontaneous pneumothoraces (occurring without trauma or any known preceding event), traumatic pneumothoraces (due to visceral pleural penetration, to alveolar rupture due to sudden compression of the chest, to penetrating or blunt chest injury, with air entering the pleural space directly through the chest wall), and iatrogenic pneumothoraces resulting from diagnostic or therapeutic medical intervention (Noppen and Schramel, 2002; Baumann and Noppen, 2004; Sharma and Jindal, 2008).

Spontaneous Pneumothorax can be divided into primary and secondary. Secondary spontaneous pneumothorax (SSP) are associated with underlying pulmonary pathology (see Table 1) (Noppen and De Keukeleire, 2008).

Table 1 - Typical causes of secondary spontaneous pneumothoraces (Noppen and De Keukeleire, 2008)**.**

Airway diseases	Infectious lung diseases	Interstitial lung disease	Connective tissue disease	Malignant Disease
Emphysema	*Pneumocystis carinii* pneumonia	Idiopathic pulmonary fibrosis	Rheumatoid arthritis	Lung cancer
Cystic fibrosis	Tuberculosis	Sarcoidosis	Scleroderma	Sarcoma
Severe asthma	Necrotising pneumonia	Histiocytosis X	Ankylosing spondylitis	
		Lymphangioleiomyomatosis	Marfan syndrome	
			Ehlers-Danlos syndrome	

Primary Spontaneous Pneumothorax (PSP) [MIM 173600], the focus of the present study, is characterized by presence of air in the pleural cavity that occurs without preceding trauma or known cause in individuals with no lung disease (Figure 2). When Laennec described pneumothoraces he found that some of them occurred in patients with apparently healthy lungs. He named these "*pneumothorax simple*" (Laennec, 1819). Nevertheless, the official description of PSP as a distinct disorder, and the one currently in use, was accomplished in 1932 based on Kjaergaard's definition (Kjaergaard, 1932).

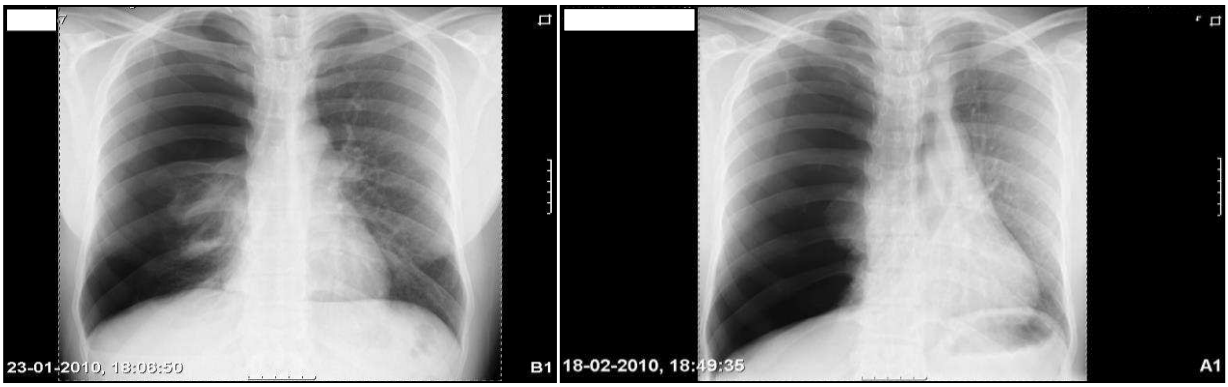

Figure 2 – Chest radiographs of two PSP patients with, respectively, 90% (left, B1) and 100% (right, A1) collapse of their right lung. In both patients the left lung presents no collapse.

Although PSP has been recognized and treated for almost 200 years, very little is known regarding its etiology. Blebs (Figure 3) and bullae or emphysema-like changes (bullae being equal to blebs but with over 2 cm of diameter) are frequently found during thoracoscopy or thoracotomy, and may play a role in PSP pathogenesis (Henry *et al,* 2003; Haynes and Baumann, 2010). The mechanism of bleb/bullae formation remains speculative, though a plausible explanation is the degradation of elastic fibers in the lung by chronic inflammation, oxidative stress and hypoxia, amongst others. Macroscopic stages can be determined when thoracoscopy is performed and categorized in four stages: I – normal visceral pleura, II – some pleural adhesions, III – blebs or bullae (emphysema-like changes) < 2 cm in diameter, and IV – bullae > 2 cm in diameter (Vanderschueren, 1981).

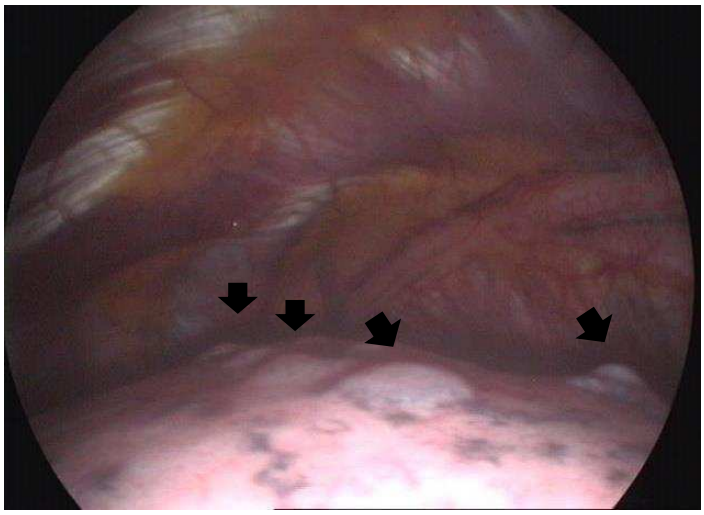

Figure 3 – Example of blebs or bullae/emphysema-like changes, typically found in the lungs of PSP patients. The presence and size of these emphysema-like changes, that define the macroscopic stage, were only accessed in our patients when a thoracoscopy was carried out.

Interestingly, most PSP episodes occur while the patient is at rest, and virtually all patients present chest pain or acute dyspnea (Bense *et al.,* 1987; Seremetis, 1970). Chest pain may be minimal or severe and, at onset, has been described as “sharp” and later as a “steady ache.” Patients with a small pneumothorax (one involving <15 percent of the hemithorax) may have a normal physical examination. Tachycardia is the most common physical finding. In patients with a larger pneumothorax, the findings on examination may include decreased movement of the chest wall, a hyperresonant percussion note, diminished fremitus, and decreased or absent breath sounds on the affected side. Tachycardia of more than 135 beats per minute, hypotension, or cyanosis should raise the suspicion of a tension pneumothorax.

PSP predominantly occurs in young, tall, thin, males between the ages of 10 and 30 and rarely occurs in persons over the age of 40 (Henry *et al,* 2003; Primrose, 1984). The reported annual incidence is 7-28 and 1-6 cases per 100,000 individuals, among men and women respectively (Gupta *et al.* 2000; Henry *et al,* 2003; Melton 1979). Smoking increases the risk of primary spontaneous pneumothorax in men and women by as much as a factor of 20 and 9, respectively, in a dose-

dependent manner. The life span risk of contracting PSP among lifelong heavily smoking men is roughly estimated to be 12% but only 1/1000 among never smokers (Bense *et al.,* 1987).

Several studies indicate a 30% average recurrence rate, usually occurring within six months to two years after the initial pneumothorax (Henry *et al,* 2003; Bense *et al.,* 1987). Radiographic evidence of pulmonary fibrosis, asthenic habitus, a history of smoking, and younger age have been reported to be independent risk factors for recurrence (Adesina *et al.* 1991). In contrast, computed tomographic (Baronofsky *et al.,* 1957) or thoracoscopic (Robbins *et al.* 1990) evidence of bullae during the evaluation of an initial spontaneous pneumothorax is not predictive of recurrence. Therefore, the presence of bullae by themselves should not form the basis of decisions regarding the prevention of recurrence. Hence, new factors need to be taken into account as basis of decisions regarding the prevention of recurrence.

The management of PSP currently consists in extracting air from the pleural space and prevention of recurrent episodes (Haynes and Baumann, 2010). Available therapeutic options vary from simple observation and aspiration with a catheter to complex video-assisted thoracoscopic surgery. The selection of an approach depends on the size of the pneumothorax, the severity of symptoms, whether there is a persistent air leak, and whether the pneumothorax is primary or secondary. Despite existing clinical guidelines, optimal management of PSP remains controversial, resulting in extensive variations in Europe and North America (Henry *et al,* 2003). PSP is the only condition in which young patients are discharged, after a first episode, having approximately 30% chance of recurrence and no effective secondary prevention measures.

1.2 – PSP genetic factors

PSP is thought to have a complex architecture where multiple genetic and environmental factors contribute to the clinical phenotype.

Familial PSP was first reported by Faber in 1921 and is rarer, as most PSP cases are sporadic (Faber, 1921). A retrospective family survey by Abolnik *et al*. in 286 males with PSP found positive family history in approximately 10% of these patients, which supports a genetic predisposition in this cases (Abolnik *et al.,* 1991). Known genetic causes of pneumothoraces in the context of other diseases include alpha1-antitrypsin deficiency, Marfan syndrome, Ehlers-Danlos syndrome and Birt-Hogg-Dubé syndrome (BHD), amongst others. Family case reports in which multiple family members have PSP, without clinical evidence of the above-cited syndromes, have been often described (Abolnik *et al.,* 1991).

PSP is genetically heterogeneous, most likely polygenic in its familial form and multifactorial in its sporadic form. The scarce published literature mainly suggests autosomal dominant inheritance with incomplete penetrance, polygenic, or X-linked recessive inheritance (Abolnik *et al.,* 1991).

Mutations in the folliculin gene (*FLCN*) have been identified in individuals with familial PSP (Lim *et al.,* 2010). Only one whole-genome linkage screen performed in one large Finnish family with dominantly inherited PSP has been published so far, in which both linkage to 17p11 and a heterozygous 4bp *FLCN* deletion were found (Painter *et al.,* 2005). *FLCN* mutations were known

previously to cause a rare skin disease, BHD syndrome, an autosomal dominantly inherited disorder characterized by benign skin tumors, diverse types of renal cancer, pulmonary cysts and spontaneous pneumothorax. All carriers of the deletion had bullous lung lesions as the only manifestation with 100% penetrance. Due to the strong correlation between PSP and BHD syndrome, it was suggested that patients with familial PSP might be at greater risk of developing renal cancer (Painter *et al.,* 2005). Furthermore, Gunji *et al.* identified heterozygous germline *FLCN* mutations in five patients with recurrent PSP and multiple lung cysts, whose family members had pneumothoraces (Gunji *et al.,* 2007). Graham *et al.* identified premature *FLCN* stop codon mutations in two PSP families (Graham *et al.,* 2005) and Fröhlich *et al.* reported a premature stop codon mutation in three affected members of one family and a heterozygous missense mutation in a sporadic case (Frohlich *et al.,* 2008). Thus, although some PSP cases are caused by *FLCN* mutations and may be milder forms of BHD, in the vast majority of cases the genetic aetiology of PSP remains unknown.

Despite its high recurrence, PSP has been investigated mainly from the clinical perspective. To date there is no published genetic association study for PSP and no pharmacological treatment available, to the best of our knowledge. This may be due to the low PSP prevalence, which makes the gathering of large clinically homogeneous datasets of patients an extremely hard task to carry out. Hence, our group conducted the first genome-wide association study (GWAS) for PSP, as well as its technical validation and replication (*paper submitted*). The GWAS was designed to the involvement of common genetic variations in sporadic PSP risk. The high-throughput genotyping queried 906,600 single nucleotide polymorphisms (SNPs) in the Human Affymetrix SNP6.0 Array and a total of 192 cases and 554 Portuguese controls were used in the discovery and replication phases. Two genetic risk factors were identified for sporadic PSP - the intronic rs11708202 polymorphism in the solute carrier family 6 member 1 gene (*SLC6A1*) and the intergenic rs4733649 SNP.

SLC6A1 encodes for the GABA transporter-I (GAT-I), which removes GABA from the synaptic cleft. Genetic variants in *SLC6A1* have been associated with attention deficit and hyperactivity disorder (Lasky-Su *et al.,* 2008) and anxiety disorders (Thoeringer *et al.,* 2009). Even though SLC6A1 was not reported to be differentially expressed in a gene expression study (conducted on lung tissue biopsies from 4 controls and 7 PSP patients) pointing for a possible role for hypoxia, inflammation and apoptosis in PSP pathogenesis (Fang *et al.* 2010), two family members (SLC1A3 - solute carrier family 1 (glial high affinity glutamate transporter), member 3, and SLC8A1 - solute carrier family 8 (sodium/calcium exchanger), member 1) were up-regulated in PSP patients. Further research is therefore warranted to replicate the association of *SLC6A1* with PSP, to understand how a GAT-I transporter may be involved in PSP pathogenesis or how a specific inhibitor may be helpful in recurrence prevention.

rs4733649 maps to 8q24.21, approximately 358 kb and 636 kb downstream from nearest genes LINC00824 (long intergenic non-protein coding RNA 824) and MIR1208 (microRNA 1208), respectively. As observed in most GWAS published to date, the top findings localize to non-coding genomic regions and do not have an immediate functional relevance (Edwards *et al.,* 2013). Once again, the first and most important step to follow up the association of rs4733649 with PSP is to confirm this finding in a dataset collected by independent researchers.

The project described herein aims to complement the GWAS study by investigating the role of both novel and lowfrequency/rare variants in familial PSP instead of sporadic PSP cases.

1.3 – Genetic Variants

It is believed that genetic variation is the major factor that contributes to human diversity (Wu and Jiang, 2013). A genetic variation or polymorphism is a variation in the deoxyribonucleic acid (DNA) sequence, being classified into several categories, including SNPs, insertions and deletions (InDels), structural variants, amongst others (Mullaney *et al.,* 2010).

A SNP is a common genetic variation that occurs in a single base of a DNA sequence, and where one of the four nucleotides (adenines, tymines, guanines or cytosines) is substituted for another. This is the most frequent type of variation in the human genome (Wu and Jiang, 2013; Carlson *et al.,* 2004).

Insertions and deletions (InDels), another type of genetic variant, are defined as an insertion or deletion of at least one base from the reference DNA sequence. Several InDels have been detected over the past decade in the human genome, however their number is smaller when comparing to the number of SNPs. Moreover, small InDels (ranging from 1 to 10 000 base pairs) are thought to cause similar phenotypic effects as SNPs in some cases (Mullaney *et al.,* 2010).

SNPs and InDels sometimes map to functionally important *loci*, likely to cause structurally damage to proteins and/or their regulation, ultimately leading to human traits and diseases. InDels are not only less frequent in the genome but also more difficult to detect, validate and genotype, when compared to SNPs and so they tend to be less studied (Mullaney *et al.,* 2010).

In the scope of this project, variants were also categorized depending on their predicted functional consequence according to the Sequence Onthology (SO [Eilbeck *et al.,* 2005]) terminology (Figure 4). The most important of these include:

- missense - a sequence variant that changes one or more bases resulting in a different amino acid, but where the length is preserved;

- splice region - a sequence variant in which a change has occurred within the splice site region, either within 1-3 bases of the exon or 3-8 bases of the intron;

- frameshift - a sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three;

- 3 prime UTR – a variant within the 3’ untranslated region (UTR) of the respective gene;

- 5 prime UTR – a variant within the 5’ UTR of the respective gene;

- regulatory region – a variant that is predicted to affect gene regulation.

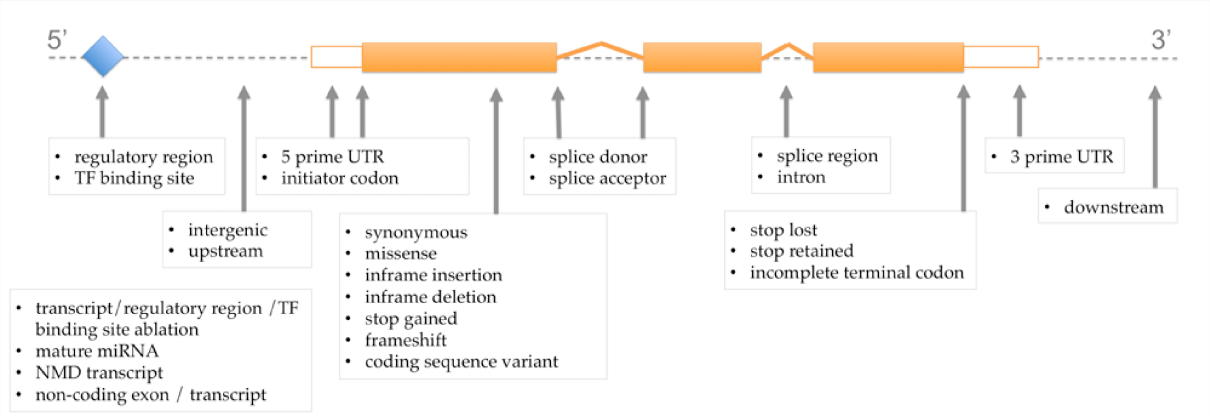

Figure 4 – Variants according to Sequence Ontology (SO) terminology (http://www.ensembl.org/info/genome/variation/predicted_data.html). This diagram illustrates the location of different types of variants in the genome sequence (dashed line) in relation to a gene (here represented by the orange boxes [exons], orange lines [introns] and respective white boxes [5' and 3' UTR regions]), with a 5' to 3' orientation. The blue diamond represents an example of a regulatory region, with a transcription factor-binding site.

1.4 – Next-Generation Sequencing

In the complex disorders field, tools available until recently only allowed the discovery of either common susceptibility variants (e.g. GWAS) or rare variants of major effect (e.g. linkage). Rare variants of small/moderate effect, which are expected to help explain the missing heritability after the GWAS and linkage era, are now under intense investigation using Next-Generation Sequencing (NGS) approaches. These rare variants may be independent of each other and confer a detectable risk of developing a disease. Several disorders are thought to be triggered by a combination of highly penetrant rare variants and common variants. Latest NGS approaches result in higher genomic coverage, easier mapping of reads, and therefore a greater ability to detect new variants with higher accuracy. Ultimately, whole genome sequencing (WGS) of many populations worldwide will facilitate the development of a genomic variation map that includes SNPs, CNVs, InDels and other genomic features, assisting a greater understanding of human disorders' etiologies.

NGS is a high-throughput parallel-sequencing approach that produces hundreds of thousands/millions of short-reads (approximately 100-500 bp) in a short time. These short reads are then aligned against a reference genome to identify variations. The first successful cases of NGS carried out in rare monogenic disorders were with the Miller syndrome (Ng *et al.* 2010), the Freeman-Sheldon syndrome (Ng *et al.* 2009) and the Schinzel-Giedion syndrome (Hoischen *et al.,* 2010) and since then, numerous novel variants have been identified for previously unresolved families and diseases.

Many diseases are thought to be caused by a combination of highly penetrant rare variants and common variants. Rare, protein-altering variants, such as non-synonymous variants (NSV) or stop-gain/loss single-base substitutions or small InDels, are predicted to have functional consequences and/or to be deleterious (Bamshad *et al.,* 2011) and to have an important role in complex diseases (Asimit and Zeggini, 2011). NGS enables the detection of these rare variants, complementing the investigation in the complex diseases field (Lee *et al.,* 2014).

There are three main types of NGS: gene panels, Gene panels, WGS and Whole-exome sequencing (WES). The first one is mainly used for diagnostics since the information obtained is limited by the chosen genes or locations. The second one has a greater cost and an inherent higher complexity in the analysis. WES is a robust strategy for discovering both novel and rare variants associated to complex diseases since the exome (1-2% of the genome, which encodes for proteins) represents a highly enriched subset of the genome, in which to search for variants with large effect sizes (Bamshad *et al.,* 2011; Lee *et al.,* 2014; Vasli and Laporte, 2013; Sun *et al.* 2015; Boycott *et al.,* 2013).

There are an increasing number of studies using WES to identify genes associated with complex diseases (Lee *et al.,* 2014; Jia *et al.* 2012) and, although WES is a robust approach to search for disease-related variants, it can also have a high error rate (Smith *et al.,* 2011). Moreover, WES analysis has several technical constrains when identifying risk alleles, depending on their mode of inheritance. Genes that cause fully-penetrant recessive and de novo dominant diseases are easier to identify due to the relatively low number of variants shared between affected individuals, decreasing the number of possible candidate genes (Boycott *et al.,* 2013)..

To minimize the possible inherent errors associated with WES, a strict bioinformatics analysis is of the upmost importance since the concordance rate between different bioinformatics pipelines is as low as 57,4% for SNPs and 26,8% for InDels with up to 5,1% unique variants to each pipeline (O'Rawe *et al.,* 2013). To address this issue, a number of optimizations and troubleshooting steps were introduced that will be further described in Chapter 2 and further discussed in Chapter 4.

1.5 – Objectives

The main goal of this thesis was to identify both novel and low frequency or rare genetic variants (and respective genes) that contribute to PSP risk in two multiplex families of Portuguese ancestry.

In order to achieve this, the project was divided in the following tasks:

1. To perform WES in all selected individuals;

2. To carry out the bioinformatics analysis of the WES data using an optimized pipeline, based on the Genome Analysis Toolkit (GATK) Best Practices workflow from the Broad Institute;

3. To select a final list of candidate variants;

4. To technically validate the variants of interest by Sanger sequencing and compare them between the families under study.

Unraveling PSP genetic underpinnings will be essential to understand its pathogenesis, and might contribute to future advances in research and therapy.

2. Methods

2.1 – Study participants

This study was carried out using two Caucasian multiplex families of Portuguese ancestry (families 01 and 02). Multiplex families are families with multiple affected individuals, usually present in more than one generation. All subjects were ascertained and collected at Hospital de Santa Maria (HSM) by Doctor Salvato Feijó and patients were diagnosed with idiopathic PSP according to the conventional criteria (Henry *et al,* 2003). Exclusion criteria (Appendix A) were applied to guarantee that the etiology of the pneumothorax episode was unknown. A thorough medical history was performed for all study participants, including information on PSP family history, pulmonary disorders, smoking, physical activity, medication, anthropometric measures, and a detailed characterization of PSP episodes for the patients. All clinical and demographic information was collected using a standardized questionnaire and the data was subsequently stored in a secure database (BCgene - https://bcgene.igc.gulbenkian.pt/bcos/index.html). This study was approved by HSM's ethical committee (Figure B1, Appendix B) and all participants provided written informed consent.

2.2 – Sample preparation

7.5 mL of whole blood was collected using sodium Ethylenediaminetetraacetic Acid (EDTA) tubes. For each individual, the medical team collected two blood EDTA tubes, one for DNA extraction and another for pellet isolation. This was the preferred DNA source for the study performed since it was the source already available, and also it allows the extraction of large amounts of DNA required to perform several genomic assays (Haines and Pericak-Vance, 1998).

At the start of this thesis work, the DNA samples from the two families that where sent for WES to Beijing Genomics Institute (BGI) had already been collected and extracted. However, I participated in the collection and DNA extraction of additional individuals from family 01 that were then used for the validation by Sanger Sequencing.

DNA extraction was carried out using Real Genomics: Genomic DNA Extraction kit (RBC Bioscience Corp., New Taipei City, Taiwan) using a slightly modified version of the "Fresh Blood 300 µL - 3 ml" protocol was used, described bellow:

<u>Red blood cell (RBC) Lysis:</u>

1. Pipet up to 400 µL of whole blood to a 2.0 mL microcentrifuge tube;
2. Add three times the sample volume of RBC lysis buffer (to remove red blood cells) and mix by inversion 10-15 times (no vortex);
3. Incubate for 5 minutes at room temperature;
4. Centrifuge at 2000 g for 5 minutes;

5. Remove the supernatant, but retain about 50 µL of residual buffer to re-suspend the white cell pellet by vortexing;

Cell lysis:

6. Add up to 600 µL of cell lysis buffer (to lyse the cell membrane) to the tube and mix by vortexing;

7. Incubate at 60ºC for 15 minutes, until the sample lysate is clear. During incubation invert the tube every 3 minutes;

Protein removal:

8. Add 250 µL of protein removal buffer (to digest proteins) to the sample lysate and mix immediately by vortexing for 10 seconds;

9. Incubate on ice for 5 minutes;

10. Centrifuge at 13000 rpm for 5 minutes;

DNA precipitation:

11. Transfer the supernatant from the previous step to a 1.5 mL microcentrifuge tube;

12. Add 300-600 µL of isopropanol (for DNA precipitation) and mix well by inverting;

13. Centrifuge at 13000 rpm for 3 minutes;

14. Discard the supernatant and add 300 µL of 70% ethanol (to improve DNA precipitation), washing the pellet;

15. Centrifuge at 13000 rpm for 1 minute;

16. Discard the supernatant and air-dry the pellet for as long as necessary (30 minutes approximately);

DNA Rehydration:

17. Add 50-100 µL of water and incubate at 60ºC for 30-60 minutes to dissolve the DNA pellet. During incubation, tap the bottom of tube to promote DNA rehydration;

Troubleshooting:

18) If after quantification the 260/230 or the 260/280 ratios were deviated from the expected values, samples were incubated again at 60ºC for 60 minutes (15 minutes with the lid off and 45 minutes with the lid on)

DNA concentration and quality were assessed using the NanoDrop 2000. Briefly, 1 µL of MilliQ water is used as a blank measurement. After that, 1 µL of each sample was pipetted onto an optical pedestal (receiving fiber). Another fiber is then brought into contact with the first one by closing the NanoDrop's arm, forcing the 1 µL solution to fill the gap between the two fiber optic ends. A source light is then directed to the sample, and the spectrometer analyses it after passing through the solution based on the amount of radiation absorbed. The results are then converted to analyzable data using the

NanoDrop 2000/2000c software. This spectrophotometer uses a spectra range from 190 to 840 nm, giving accurate estimates of the template concentration in ng/μL. DNA has optimal absorption at 260 nm, due to the structure of the aromatic rings of their bases (Clark *et al.,* 2005). Thus, DNA concentration can be estimated based on the amount of absorbed radiation at 260 nm (Cavaluzzi and Borer, 2004). To correlate the absorbance with concentration the Lambert-Beer equation is applied:

$$A = \varepsilon.l.C$$

"A" is the absorbance, "ε" is the wavelength-dependent molar absorptivity coefficient (ng−1cm−1μL), "l" is the path length (cm), and "C" is the nucleic acid concentration (ngμL−1). Standard values for the molar absorptivity coefficient of double-stranded and single-stranded DNA (0.020 ng−1cm−1μL and 0.027 ng−1cm−1μL respectively, when exposed to Ultra-Violet light (UV) at 260nm) are applied for the concentration estimation (Cavaluzzi and Borer, 2004).

The purity of the sample can then be roughly determined using the A260/A280 ratio. The DNA is considered pure if this ratio is 1.8. If this ratio is significantly different from 1.8, then the DNA sample is contaminated, either by Ribonucleic Acid (RNA), if the ratio is closer to 2.0, or by proteins or other contaminants, if it is below 1.8, since the latter will absorb radiation at around 280 nm (Clark *et al.,* 2005; Haines and Pericak-Vance, 1998).

For the Whole Exome Sequencing (WES) samples were diluted to a final concentration of 40 ng/uL (minimum concentration required by BGI) and to 10 ng/uL (working solutions) for the validation assays.

2.3 – Whole exome sequencing

WES was outsourced to BGI and performed using the Illumina's Solexa Hiseq2000. This is a complex, multi-step and highly informative technique (Crona *et al.,* 2013) that is briefly described as follows.

2.3.1 – Genomic enrichment

Each DNA sample was randomly sonicated into fragments with approximately 200 bp (in our case). These fragments were then bound to adapters and the resulting templates are then purified by AgencourtAMPure SPRI beads. The purified DNA fragments were hybridized with the SureSelectBiotinylated RNA library (BAITS). This library targeted not only the exonic region but also the regulatory and intronic regions was selected since that could be potentially interesting. After the hybridization, the fragments that hybridized are bound to streptavidin beads and magnetically captured while the ones that did not are discarded (Vasli and Laporte, 2013). After washing the beads and digesting the RNA, the captured fragments are amplified by ligation-mediated polymerase chain reaction (LM-PCR) and can subsequently be loaded for sequencing (Figure 5).

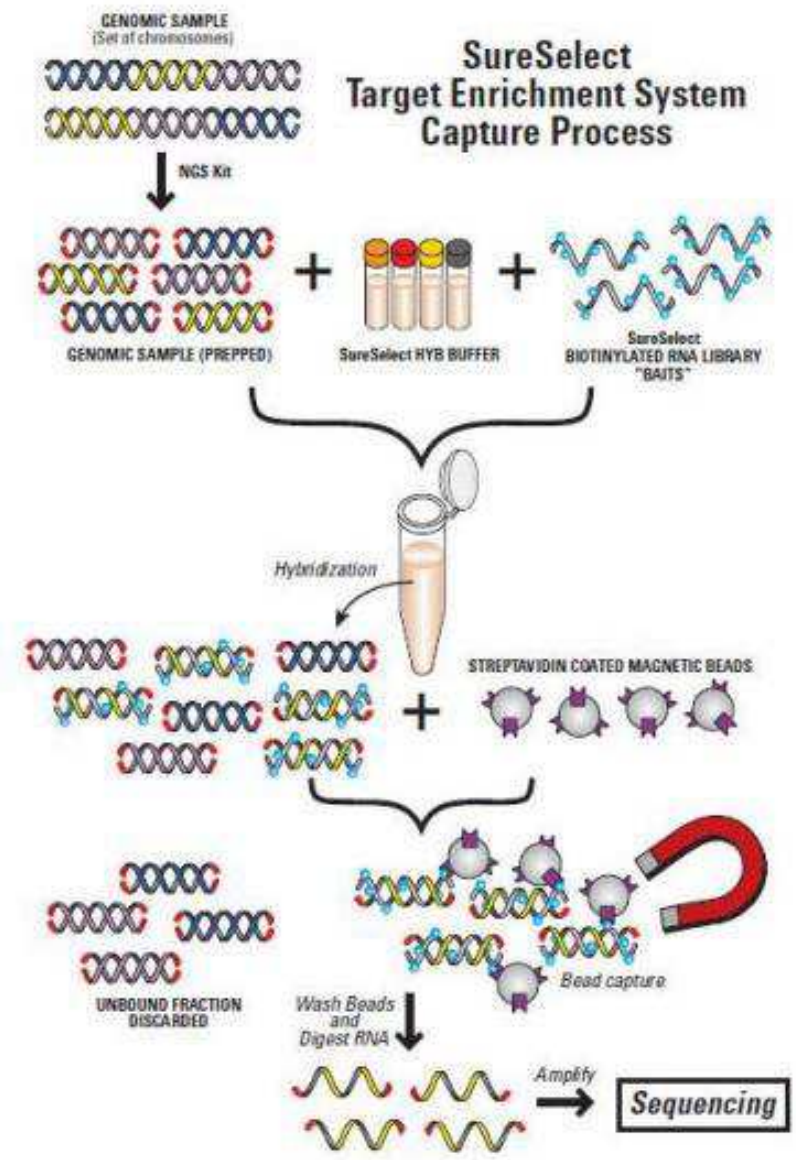

Figure 5 – Genomic enrichment steps - This illustration highlights the major steps of WES genomic enrichment performed by BGI, prior to the sequencing steps (http://ncifrederick.cancer.gov/atp/genetics-and-genomics/laboratory-of-molecular-technology/lmt-protocols-and-resources/sureselect/background/).

2.3.2 – Sequencing

Illumina Hiseq 2000 was used since it is one of the most robust platforms with lower false positive rates at the time. The amount of required starting DNA is lower than usual, so there is no PCR amplification step previous to the sonication that could create artificial mutations and AT or GC-rich amplification bias (one of the biggest problems with traditional WES and one of the most associated with traditional Illumina techniques).

The sequences of each individual were generated as 200 bp paired-end reads. Paired-end sequencing allows sequencing both ends of a strand by the ligation of adapters, containing attachment

sequences, and sequencing primer sites (forward and reverse) into exonic DNA. Single-end sequencing on the other hand only allows sequencing of one strand because it only uses one sequencing primer (forward or reverse). By using paired-end reads one can improve the confidence in our results and it also allows a better quality control (QC) during the bioinformatics analysis.

Illumina´s technology is based on clonally amplified templates coupled with cyclic reversible termination method with four fluorescent colors (one for each different base). First, one fluorescently modified nucleotide complementary to the template sequence is incorporated. After washing the non-bound nucleotides and imaging for detection of the incorporated nucleotide, a cleavage step removes the fluorescent dye and a novel incorporation step is performed. These steps are done in a cyclic manner, 72 or 100 times (Vasli and Laporte, 2013).

Sequencing was performed with an average fold-coverage of approximately 60x which is the recommended coverage value (Sims *et al.*, 2014; Weber-Lehmann *et al.,* 2014). Raw image files are processed by Illumina base calling software 1.7. The base quality value (denoted as Q) allows for the calculation of the sequencing error rate (denoted as E, $sQ = -10 \log_{10} E$, Table 2). This parameter is of extreme importance to the QC and variant calling steps. In our project only variants with a Q equal or superior to 30 were taken into consideration.

Table 2 – Base quality value (Q) and sequencing error rate (E) values correspondence, calculated by BGI for our assay. The error rate is showed in percentages and calculated using the formula mentioned above.

Q	E (%)
13	5
20	1
30	0.01

Nowadays, WES is already a useful alternative or complementary technique for molecular diagnosis. Its routine use leads to a rapid screening and fast identification of mutations in rare genetic disorders through sequencing of the coding region. This technique provides a significantly higher amount of information when compared to a multi-gene panel analysis. In our project, as we are studying idiopathic PSP we were specially looking for novel and unrelated rare variants, so WES was the most powerful, cost-effective and unbiased choice (Sun *et al.* 2015).

2.4 – Bioinformatics Analysis

The bioinformatics pipeline used in this project (Figure 6) was an upgraded and modified version the one based on the Genome Analysis Toolkit (GATK) Best Practices workflow, from the Broad Institute (Figure C1, Appendix C).

As the GATK best practices pipeline, this one is also divided in three main phases (Figure 6): Data pre-processing, Variant discovery and Variant annotation. An optional (but in our case essential) step was also performed - variant prioritization. All of our data has been stored redundantly at IMM.

Some tools were added to tentatively optimize the analysis (such as Freebayes and Samtools mpileup used for the variant calling). Additionally, we were in general stricter with our QC parameters than required as we used a combined variant calling (instead of a simple variant calling) and our annotation and prioritization phases were also vastly optimized for novel variants (all scripts and hardware/software specifications in Appendix D). In addition, we also used the automated bcbio-nextgen pipeline, developed by BCBIO, with our data to compare the results (https://github.com/chapmanb/bcbio-nextgen/tree/master/docs/contents and https://bcbio-nextgen.readthedocs.org/en/latest/).

Our bioinformatics analysis is described in detail bellow.

2.4.1 – Data pre-processing

The bioinformatics analysis begins with the raw data that results from the sequencing step. The first pre-processing steps were carried out by BGI and consist of eliminating 1) reads that contained the sequence of the adapter; 2) the low-quality reads which have long stretches of "N"s; 3) reads in which unknown bases are more than 10%. This step produced the "clean data", originating "clean" FASTQ files that where subsequently used in the bioinformatics analysis. After receiving this data our first step is to evaluate its quality. If it passes QC it can then proceed to the alignment step.

2.4.1.1 – Alignment

The FASTQ files were aligned to the human reference genome GRCh37/hg19 (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/) using the Burrows-Wheeler Aligner tool (BWA) originating binary alignment files (BAM) (http://bio-bwa.sourceforge.net/). GRCh38 was, at the time, already available but the multiple ancillary files were not, so GRCh37/hg19 was used.

BWA identifies the best position where the reads match and then does a linear scan through all the potential hits for both paired-end reads. BWA estimates the insert size distribution of both paired-end reads mapped and pairs them. After that, it aligns the unmapped reads whose mate pairs are already respectively aligned with higher confidence (Li and Durbin, 2009). For each mapped read, BWA generates a mapping quality score (MAPQ) which varies between 0 and 60. The higher the MAPQ score is, the smaller the probability of the alignment being incorrect. Finally, BWA records the results for each read as a CIGAR string (Figure E1, Appendix E).

We tested three other alignment tools (CGAP-Aligner, HPG alignment tool and Bowtie2), all of them based on the BWT (Burrows-Wheeler Transform) algorithm and compared them with BWA. For this purpose, we used data from a previous lab project where we already also had validation results by Sanger sequencing. In particular, we used data from a case and a control and searched for differences in the called variants (apart from the alignment, all the other pipeline phases were performed in the same way). Both C-GAP Alignment Tool and HPG-Aligner failed and we suspected that this happened since they are very recent tools and still have several bugs. However, Bowtie2 succeeded and presented no significant differences to BWA.

2.4.1.2 – Mark duplicates

There are multiple types of duplicate reads. The problematic ones are the PCR duplicates that result from the LM-PCR step because they are artifacts of the amplification that introduce a bias in our results (artificial duplicates). To solve this problem, all duplicated reads were marked using Picard (http://picard.sourceforge.net/command-line overview.shtml). This tool finds the 5' coordinates and mapping orientations of each read pair. When doing this, it takes into account all clipping that has occurred as well as any gaps or jumps in the alignment. It then matches all read pairs that have identical 5' coordinates and orientations and marks as duplicates all but the "best" pair. "Best" is defined as the read pair having the highest sum of base qualities as bases with Q >= 15. It attributes a score to these duplicates - the higher the score the higher the probability of being a true duplicate (exactly overlapping reads that naturally occur within deep sequence samples), which does not introduce bias, and the lower the score the higher the probability of being a PCR artifact, which introduces bias to the data (Bainbridge *et al.,* 2010).

2.4.1.3 – InDel realignment

BAM files are the re-aligned using the GATK InDelRealigner. This step is crucial because InDels could lead to errors in the alignment itself originating mismatches. Local realignment around InDels reduces the number of mismatching bases.

2.4.1.4 – Base recalibration

Base recalibration is the last step of the data pre-processing and was performed using the GATK base quality recalibration tool. This tool models these errors empirically and adjusts the quality scores accordingly. This results in more accurate base qualities, which in turn improves the accuracy of the variant calls. The base recalibration process involves two key steps: first the program builds a model of covariation based on the data and a set of known variants, then it adjusts the base quality scores in the data based on the model. The BAM file originated from this step is ready for the Variant Calling step.

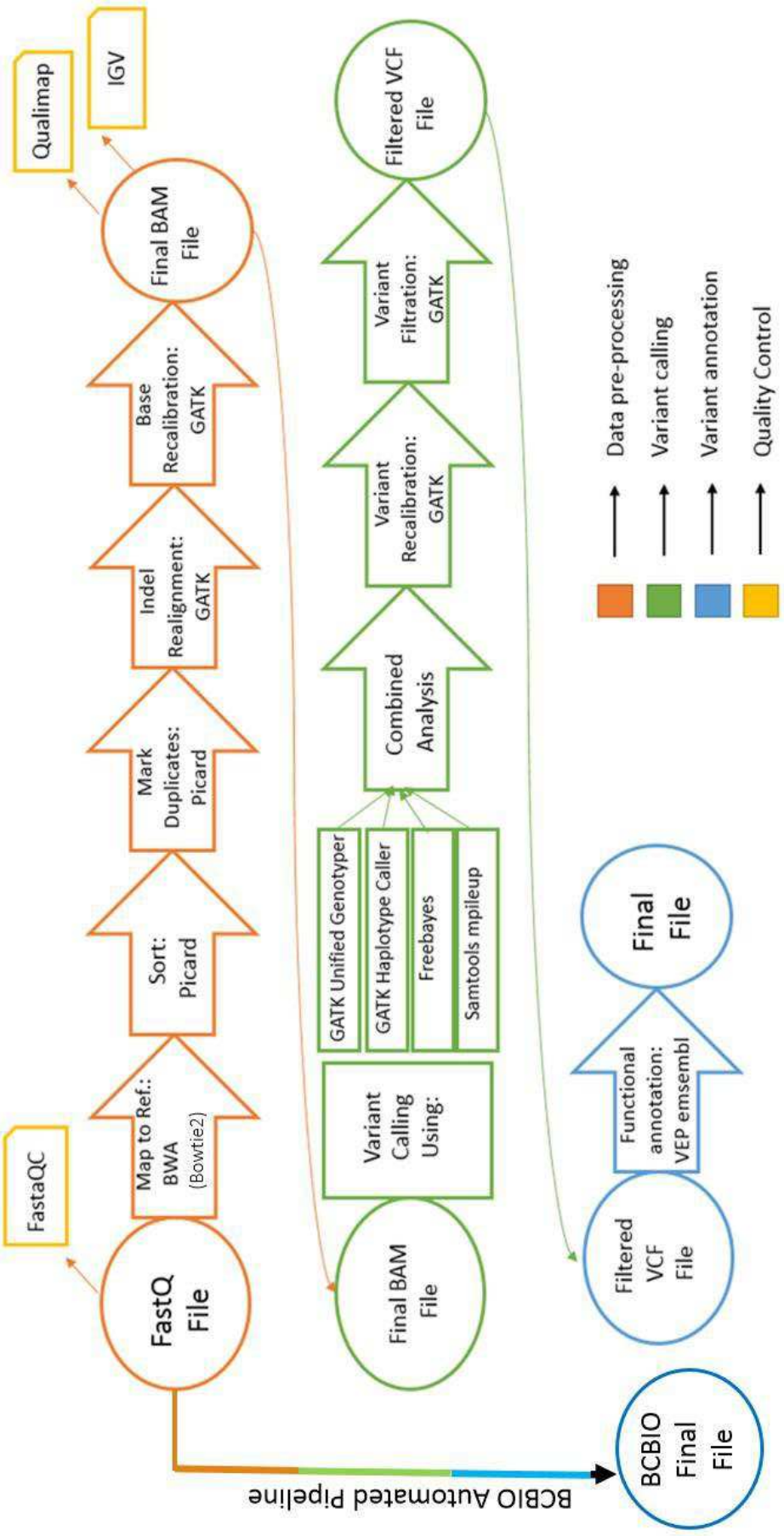

Figure 6 – Optimized bioinformatics pipeline used for our analysis. Includes the main pipeline used (based on the GATK Best Practices workflow, Broad Institute), the use of Bowtie2 (to test the alignment) and the parallel use of the BCBIO automated pipeline. The first phase, Data pre-processing (orange), starts with a FastQ file from BGI and ends with a final BAM file. The second phase, Variant Discovery (green), starts with the final BAM file and ends with a filtered VCF file. The final phase, Variant Annotation (blue), may either include the variant prioritization or not. The Quality control (yellow) carried out, consists not only of FastQC, Qualimap and IGV, but also quality parameters within the other phases. As for Bowtie2 testing, only this tool was changed in the pipeline, replacing BWA. Regarding, the BCBIO automated pipeline, it starts with the same FastQ file and goes through the same 3 phases, although using different parameters and tools.

2.4.1.5 – Quality control

Throughout all of the data pre-processing phase, there are multiple QC steps that were undertaken to certify that the data is trustworthy.

The first QC tool is the FASTQC program. For this we used the FASTQ files that we received from BGI. This allowed to evaluate the quality of the data received from BGI. We obtained data of approximately the length of the sequence (important to see if the minus strand and the plus strand match), the GC percentage, the per sequence quality scores, per base sequence quality, the duplication levels, amongst other parameters. This is a crucial step since the quality of the input data will critically impact the quality of the results coming out of the pipeline.

To QC the BAM files, we used Qualimap (http://qualimap.bioinfo.cipf.es/). This program provides partially the same information than FASTQC but additionally provides further information, such as insert sizes, InDel percentages, and information per chromosome amongst other. It can also be used in all of the generated BAM files (for each step).

Finally, the quality was also visually inspected using IGV (Integrative Genomics Viewer). This software allows the visualization and exploration of the data using the BAM files (Thorvaldsdóttir *et al.,* 2013). It is possible to compare regions, to compare the genome of two or more individuals, to compare with the reference genome, amongst many others. It was also used later when choosing which variants to validate.

2.4.2 – Variant discovery

In the Variant Discovery phase the major difference between our pipeline and GATK Best practices pipeline is the use of a combined variant calling with four different variant calling programs (Yi *et al.,* 2014).

2.4.2.1 – Combined variant calling

We used four different tools in the variant calling step in order to minimize the error rate associated with any of the tools individually. The tools used were the GATK Unified Genotyper (UG) (which is the one recommended in the Best Practices Pipeline), the GATK Haplotype Caller (HC), Freebayes (FB), and Samtools mpileup (SM).

The main difference between the two GATK tools is that UG calls SNPs and InDels separately by considering each variant locus independently, whereas HC calls SNPs and InDels with a local *de novo* assembly. Only bases that fulfill the criteria specified in the program are included, namely a minimum base quality value required to consider a base for calling and a good confidence call (http://www.broadinstitute.org/gatk/events/3391/GATKw1310-BP-5-Variant_calling.pdf.).

FB calls the variants using short-read alignments for any individual and a reference genome (GRCh37/hg19) to determine the most likely combination of genotypes at each position in the genome (https://wiki.gacrc.uga.edu/wiki/Freebayes).

Finally, SM needs bcftools to make the variant call because it is only capable of collecting the information of BAM files, computing the likelihoods of data given each possible genotype and stores the scores in the binary format (.bcf file). Then, Bcftools (from Samtools) makes the variant calling applying the prior likelihoods obtained by Samtools and performing the actual calling (http://samtools.sourceforge.net/mpileup.shtml).

All of these four tools applied use a Bayesian formulation for picking the base that maximizes the posterior probability with the highest Phred quality score, the previously mentioned Q score (McKenna *et al.,* 2010).

After running the four tools for each individual, a combined analysis was performed, where the variants that are not common between the four variant calling tools were filtered out. This way, only the variants that all different tools called were selected, providing more confidence in the results of this other decisive step.

2.4.2.2 – Ti/Tv ratio

The Ti/Tv (Transitions/Tranversions, also known as Ts/Tv) ratio uses the amount of Transitions (Ti - A–G, C–T) and transversions (Tv - A–C, G–T, A–T, C–G) found to determine the probability of false positives. Usually Ti are twice as frequent as Tv thus false positive SNPs should have Ti/Tv ratios of around 0.5 for randomly assigned variations. These could be the consequence of systematic sequencing errors, alignment artifacts and data processing failures. The expected Ti/Tv ratio for exonic regions data in the literature varies between 2.8 and 3.3 (Ebersberger *et al.,* 2002; Buetow *et al.,* 1999). A combination of residual false positive leads to a lower Ti/Tv ratio at new sites when compared to known sites. For the whole genome the expected ratio should be approximately 2.1 (DePristo *et al.,* 2011). The Ti/Tv ratios were calculated using VCFtools (version 0.1.12a) for each of the four variant calling tools as well as for the combined analysis in order to decide which approach had a lower false positive rate.

2.4.2.3 – Variant recalibration

The GATK VariantRecalibrator tool was used to evaluate the probability that each call is real. This tool uses known sites (from databases such as dbSNP, Mills, 1000 Genomes Project [1000GP] gold standard and HapMap) to estimate the relationship between SNP call annotations and the probability that a SNP is a true genetic variant versus a sequencing error or data processing artifact (http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_variantrecalibration_VariantRecalibrator.html.).

2.4.2.4 – Variant filtration

The last step of the variant discovery phase is the variant filtration, in which all variants in the call set were QC evaluated based on seven annotations parameters. These are:

1) Quality by depth (QD), which indicates the confidence in a variant at a given site (based on the Q score), over the coverage of the reads that do not match a given site. This parameter is calculated by GATK and is normalized by read length (the higher the length, the more confident is the QD score). Low scores (below 10) are indicative of false positives calls and artifacts, being subsequently discarded (http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_annotator_QualByDepth.html).

2) Mapping quality (MAPQ) is a parameter calculated by the alignment tool that measures the quality of the alignment for the sequencing reads. When BWA aligns the sequencing reads with a reference genome it calculates a score between 0 and 60, in which 0 indicates low mapping quality (less confidence in the alignment) and 60 represents the maximum confidence in the alignment (all bases are correctly mapped with the reference). When no cut-off is available for the MAPQ score it is recommended to draw the distribution of mapping quality scores and examine the distribution of outliers (http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_annotator_RMSMappingQuality.html).

3) Read depth (DP), also called Coverage, indicates the number of times that each region was sequenced. However, this could be easily skewed by high-depth regions during exome sequencing, in a phenomenon called unspecific binding where certain regions of the genome have much higher depth than usual. In order to obtain a more realistic value, we used the DP parameter calculated by GATK that describes the total depth of the reads, for each variant, that passed the caller's internal QC metrics such as MQ > 10, as seen above (http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_annotator_Coverage.html).

4) Fisher Strand (FS) detects the strand bias in the reads, identifying the number of reference and/or alternative alleles seen on the forward and reverse strand. This score is calculated with the Fisher's exact test to obtain the *P*-value and calculate the Phred-scale score ($FS = -10 \log_{10}(P\text{−value})$). The *P*-value is calculated based on a contingency table (2 x 2) with the number of strands (negative and positive) for the reference and alternate alleles. Higher scores are indicative of false positive calls. The recommended value for this parameter for SNPs is below 60 and for InDels variants is below 200.

5) Mapping quality rank sum test (MQRankSum) measures the mapping quality of heterozygous calls (reads with reference allele versus alternate allele). It uses the Wilcoxon Rank Sum test to assess whether two samples are the same or not. This score indicates how many times the standard deviation is distant from the mean. The recommended value for this parameter for SNP variants is above -12.5.

6) Haplotype score measures the consistency of the site with two segregating haplotypes. This score is calculated for each consensus haplotypes against one of the consensus haplotype found in a prior set, for each read. The haplotype score is the mean of haplotype scores of all reads for each

locus (in a window of 21 bp). It generates one or two consensus haplotypes with the best quality scores (the lowest score). The recommended value for this parameter for SNPs variants is below 13.

7) ReadPosRankSum measures the distance from the end of the read, for reads with the alternate allele. If the alternate allele is only seen near the extremities of the read it is indicative of an error. It uses the Wilcoxon Rank Sum test to test if two reads are the same or not. The score indicates how many times the standard deviation is distant from the mean. The recommended value for this parameter is above -8 for SNPs and above -20 for InDels.

The VCF files that result from this step contain the list of variants that will be considered for validation but first they need to be annotated.

2.4.3 – Variant Annotation

There are multiple annotation tools such as ANNOVAR, SNPSift and Gemini, amongst others. However, we opted to use the VEP Ensembl tool because it is not only one of the most widely used and most trusted for human data, but it is also simple to customize (important for novel variants annotation), it provides extensive information and at the start of this annotation process, a new version (version 79) which allowed the use of the most recent versions of the databases dbSNP 142 (http://www.ncbi.nlm.nih.gov/SNP/) and 1000 GP phase 3 (http://www.1000genomes.org/) with our reference genome, was released.

We ran both a version with non-autossomal chromosomes and mitochondrial information and the version without them to see if we were losing important information by eliminating them. This could happen since this process can take many hours and huge server power and it has to be constantly refined and repeated, the less the amount of initial data the better. We found no significant data in those tests for the X, Y and mitochondrial chromosomes so in later versions of the annotation (including the final one) they were not included.

The connection to the databases was done by running the tool online. In the first attempts we ran it offline with a downloaded file for the databases but constant problems with the files emerged (as files needed to be in very specific formats and tabixed in a very specific way) as well as the fact that dbSNP 142 was so recent that constant updates were being released. These obstacles made us switch to the online approach that ran much more efficiently.

The main aims of the annotation are:

1) to identify the variants that are present in the database;

2) to carry out a functional prediction when applicable, using both Sorting Tolerant From Intolerant (SIFT) and PolyPhen-2 tools. Both of the output scores range between 0 and 1, with damaging scores being the lowest value for SIFT and the higher value for Polyphen-2 (Kumar *et al.,* 2009; Adzhubei *et al.,* 2010; Wu and Jiang, 2013);

3) to obtain a conservation score, using PhyloP and PhastCons tools. PhyloP measures acceleration as well as conservation at a specific base pair, with negative values indicating faster

evolution and positive values indicating slower evolution. PhastCons measures the probability of negative selection for a small region, ranging from 0 to 1 (https://genome.ucsc.edu/cgi-bin/hgTables);

4) to perform gene-based annotation and region-based annotation, identifying the gene of interest and the transcript. We also obtained further important information such as minor allele frequencies (MAF) (global and population specific), the SNP ID number, amongst others.

Most annotation tools are optimized for diagnostics. They are designed to facilitate the identification of specific variation in an individual. This created additional complications in our project. For instance, variants that are flagged (for instance, multi-allelic variants) are not annotated with a SNP ID, so when filtered they are labeled as novel variants. Moreover, when comparing a VCF file to the databases the only fields that are compared are the chromosome and the base pair position. So if a different novel variant (a one bp deletion, for example) is in the same position of a known SNV it will be annotated as the SNV variant. To solve this we contacted the VEP Ensembl help team and with their assistance we were able to use a command line with arguments that solved both of these problems.

The annotation process generates two (possibly three) output files: a VCF output file with all the information described above, an HTML file with the general statistics of the run (total missense variants or per chromosome intronic variants, for example) and, when necessary, a warning file.

2.4.4 – Variant Prioritization

To decide which variants to prioritize for validation, there are several possible strategies depending on the mode of inheritance and on the number of affected/non-affected individuals sequenced. Data was already divided into two VCF files: SNPs and InDels. Both files contain several types of alterations in the coding DNA sequence (CDS) region and regulatory region.

The first step was to prioritize the variants assuming an autosomal mode of inheritance. Therefore we selected all the variants shared between the affected individuals and then we removed all of the variants which were present in the non-affected individual. Sequencing several affected and non-affected individuals from the same family dramatically improves the filtering process.

Ranking of the previously selected variants was then made according to the type of variant. For SNPs, non-synonymous variants in the coding region were analyzed first since cause amino acids changes which could affect protein function, thus a probable cause for disease, as well as stop-gain/stop-loss variants. For InDels, we analyzed frameshift insertions/deletions, non-frameshift insertions/deletions, frameshift/non-frameshift block substitution and stop-gain/stop-loss variants. For both SNPs and InDels, the search was expanded for other regulatory regions (3'UTR, 5'UTR, upstream, downstream, intronic, intergenic and regulatory). Regulatory and 3'UTR variants were also prioritized.

Next, we wanted to prioritize novel variants, so we filtered out variants described in public databases - dbSNP142 and 1000 GP phase 3.

Assuming that this variant is most likely rare, or at least of low frequency, variants with EUR_MAF (MAF in European populations) values below 5% were selected (even if the global MAF

was not also below 5%, but we preferred those that were). Some variants where there was no information regarding the EUR_MAF if the remaining parameters applied were also kept.

Thereafter, SIFT (http://sift.jcvi.org/) and Polyphen-2 (http://genetics.bwh.harvard.edu/pph2/) scores were used to assess if the variants are possibly damaging to the protein function and to prioritize those that were (Wang *et al.,* 2010).

Then, variants that were similarly present in the BCBIO automated pipeline results were also selected. This pipeline starts with the same FastQ file, however it uses different parameters and tools (Appendix D)

Lastly we used the Exomiser HIPHIVE tool (http://www.sanger.ac.uk/resources/software/exomiser/submit/). Here we inserted the pre-annotation file and described the phenotype of the disease (Appendix D), and afterwards the tool compares our results with human, mouse, zebrafish and other animal data, with identical or similar symptoms, attributing a score per variant. Variants that scored higher were also prioritized.

The standard nomenclature recommendations of the HGVS (Human Genome Variation Society) were followed to name variants (Ogino *et al.,* 2007).

2.5 – Validation by Sanger sequencing

2.5.1 – Primer design

PCR and sequencing primers were designed using Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/). This program takes into consideration parameters such as sequence specificity, melting temperatures (Tm) of primer pairs, GC content and the possibility of secondary structure formation, such as loops (Untergasser *et al.,* 2012).. Primer3 provides an output with the possible combinations for reverse and forward primers so one can select the best primer pairs to fit the project design. In order to obtain the best possible pairs, we specified a "Product Size Range" between 350 and 500 bp, used a minimum primer size of 20 bp and an optimal size of 22 bp and we identified the region of interest when submitting the sequence. The chosen primers varied slightly in size, GC content and Tm's but were consistent overall.

To confirm the uniqueness of each primer sequence we used both the BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=OGP__9606__9558&LINK_LOC=blasthome) and BLAT (https://genome.ucsc.edu/cgi-bin/hgBlat?command=start) tools. These tools compare each DNA sequence against the whole human genome and score them with a percentage of identity. For a similar purpose we also used the "*in silico* PCR" (http://genome.ucsc.edu/cgi-bin/hgPcr?command=start), which searches how specific each pair of primers is, as well as calculates the PCR product and its characteristics. Lyophilized primers (NZYTech, Lisboa, Portugal) were re-suspended in MilliQ water, making final 100 µM stock solutions that were stored at -20^{o}C. the primer sequences used for pcr amplification are shown in Table F1 (Appendix F).

2.5.2 – PCR

Before the sequencing itself, it is necessary to amplify our regions of interest using the PCR. PCR is an enzymatic process in which a specific DNA region is replicated exponentially to yield many copies of a particular sequence of interest. This process involves heating and cooling samples in a precise thermal cycling pattern over a number of cycles (usually around 30 cycles). During each cycle, a copy of the target DNA sequence is generated for every molecule containing the target sequence. The oligonucleotide primers, that are complementary to the 5´- and 3′-ends of the sequence of interest, define the boundaries of the amplified product. Theoretically, after 30 cycles, approximately one billion copies of the target region (DNA template) have been generated. The PCR product is called the amplicon and can then be used for sequencing (Butler, 2005.).

Two different kits were used in the PCR reactions: the NZYTaq 2x Colourless Master Mix (hereby referred as NZYTaq Colourless - NZYTech, Lisboa, Portugal) and the KAPA HiFi HotStart ReadyMix PCR Kit (hereby referred as KAPA RM – KAPA Biosystems, Wilmington, Massachusetts, USA). The PCR conditions used to obtain each amplicon can be found in Table G1 (Appendix G). PCR reactions were performed on a PCR 2720 Thermal Cycler (Applied Biosystems, Waltham, USA).

The NZYTaq Colourless mix consists of a premixed, ready to use solution containing dNTPs, reaction buffer, $MgCl_2$ and a recombinant modified *Taq* DNA polymerase purified from *Escherichia Coli (*all at the optimal concentrations for an efficient amplification).

KAPA RM is also a premixed, ready to use solution but it has an increased affinity for DNA, resulting in yield improvements, speed and PCR sensitivity. It contains dNTPs, reaction buffer with stabilizers, $MgCl_2$, and a Hotstart DNA 5'->3' polymerase with a strong 3'->5' exonuclease (for proofreading) activity but no 5'->3' exonuclease activity that results in a superior accuracy and, consequently, a lower error rate when compared to other kits.

2.5.3 – Electrophoresis

Gel electrophoresis was mainly used to examine DNA quality, fragment size and PCR product specificity. The DNA fragments always migrate from the cathode (negative pole) to the anode (positive pole), because DNA carries a negative charge due to its phosphodiester backbone. For optimal resolution the agarose concentration used was 0.8-1% for total DNA (post extraction) and 2% for amplicons. The gels were made by dissolving agarose powder (NZYTech, Lisboa, Portugal) in 1x Tris-Acetate-EDTA (TAE), diluted from 50x TAE (Sigma, Missouri, United States of America). GreenSafe Premium (NZYTech, Lisboa, Portugal) was added as a nucleic acid stain (allowing for the vizualization latter on) at a final volume of 4 µL in a 150 mL agarose gel. The whole solution was poured into a plate and allowed to cool and set. Once the gel was set, it was immersed in a 1x TAE buffer solution. Loading buffer (3 µL) was added to each PCR product (6 µL) and these were loaded into wells with a molecular weight ladder running in parallel (4 µL of 1x NZYDNA Ladder VI [NZYTech, Lisboa, Portugal]). This ladder provided an approximate quantification of the fragment's size. A negative control (prepared and run through the PCR process at the same time of the samples and using all PCR reagents but instead

of DNA, it had water) to guarantee that there was no contamination and a blank control (no PCR reagents, only water) were also loaded on the gel. The voltage was set at a constant value (between 90-120 V) for the length of time required, which can be visually checked by the migration of the loading buffer. As the DNA migration in the gel matrix occurs, the fragments are separated according to size - small fragments migrate faster and run further in the gel compared to larger ones. During the migration the GreenSafe intercalates with the DNA allowing its visualization using an UV transilluminator (GenoSmart gel documentation system, VWR International, Radnor, Pennsylvania) to be photographed and analyzed.

2.5.4 – Amplicon purification

To remove the excess of primers and dNTPs before the sequencing reaction, PCR products were purified using illustra ExoStar 1-step (GE Healthcare Life Sciences, Buckinghamshire, UK) that contains a mix of illustra Alkaline Phosphatase and Exonuclease 1. The samples were first quantified and diluted the GATC requirements. Then 2 µl of ExoStar were added to 5 µl of the sample, and the mix was then incubated at 37º C for 15 minutes (the active period of the enzyme) immediately followed by 15 minutes at 80º C (to inactivate the enzyme). Six µl of one of the primers (either forward or reverse primer) was then added to this mix and shipped to GATC (Konstanz, Germany).

2.5.5 – Sanger sequencing

Sanger sequencing was outsourced to GATC. This technique is considered the gold standard of sequencing, hence why it was chosen to validate our NGS results. Briefly, it is described by the incorporation of dideoxyribonucleotide triphosphates (ddNTPs) as chain terminators followed by a separation step capable of single nucleotide resolution. This is possible since there is no hydroxyl group at the 3′-end of the ddNTPs and, therefore, the chain growth terminates when the polymerase incorporates a ddNTP into the synthesized strand. Extendable dNTP and ddNTP terminators are both present in the reaction mix so that some portions of the DNA molecules are extended. At the end of the sequencing reaction a series of molecules are present that differ by one base from another (Butler, 2005).

Each DNA strand was sequenced in separate reactions with a single primer. Either the forward or reverse PCR primers are used for this purpose. Four different colored fluorescent dyes are attached to each of the four ddNTP. Thus, ddTTP (thymine) is labeled with a red dye, ddCTP (cytosine) is labeled with a blue dye, ddATP (adenine) is labeled with a green dye, and ddGTP (guanine) is labeled with a yellow dye, although it is typically displayed in black for easier visualization.

2.5.6 – Sequence Analysis

Sequence files generated at GATC were imported into the Staden package (Pregap4, Trev and Gap4 tools) and the presence or absence of variation was checked in each individual separately or together with the remaining individuals of interest (Staden, 1996).

Both primer pairs were aligned and compared to the sequences and the position of the relevant variants was highlighted. This program also provides a confidence measure for each allele call per position.

2.6 – Reagents and buffers

- MilliQ water
- Isopropanol
- EtOH 70% and 95%
- RBC Lysis Buffer (RBC Bioscience Corp., New Taipei City, Taiwan)
- Cell Lysis Buffer (RBC Bioscience Corp., New Taipei City, Taiwan)
- Protein Remove Buffer (RBC Bioscience Corp., New Taipei City, Taiwan)
- NZYTaq 2x Colourless Master Mix (NZYTech, Lisboa, Portugal)
- TAE 1x (Sigma-Aldrich, Sintra, Portugal)
- Agarose (NZYTech, Lisboa, Portugal)
- Loading dye 1x (NZYTech, Lisboa, Portugal)
- GreenSafe Premium (NZYTech, Lisboa, Portugal)
- 1x NZYDNA Ladder VI (NZYTech, Lisboa, Portugal)
- KAPA HiFi HotStart ReadyMix PCR Kit (Kapa Biosystems, Woburn, USA)
- illustra ExoStar 1-Step (GE Healthcare Life Sciences, Buckinghamshire, UK)

3. Results

3.1 – Families and individuals

All affected individuals were diagnosed by Doctor Salvato Feijó and his team at HSM. The families' medical history, key lifestyle and demographic information are summarized in Table 3. Diagnosis of idiopathic PSP was performed as described in Chapter 2.

The use of pedigree information can substantially narrow the genomic search space for candidate causal alleles. Depending on the frequency of the disease-causing allele and the nature of the relationship between the individuals within the family under study we chose the most informative individuals to further analyze (Bamshad *et al.* 2011). The genogram of families 1 and 2 are depicted in Figures 7 and 8, respectively.

Sequencing multiple affected individuals and one unaffected control from within one family allowed us to significantly reduce the number of candidate variants, since we could select the variants that are present in all of the affected individuals and absent in the unaffected one. This allows a higher causal variant prioritization precision.

The selection of which unaffected individual to use as the control was crucial. In family 1, individual I.1-110842 was chosen as she is well above the expected age of onset (usually around 30 years of age for the first episode), while also being a direct family member for all the affected individuals and had never been diagnosed with PSP.

For family 2, two individuals were selected for the WES assay, one affected (II.3–110875) and one unaffected carrier (II.8–080312). The unaffected individual (carrier) was chosen for being the sibling of one of the affected individuals and the parent of another. The remaining individuals from family 2 were not available for the WES, but became later accessible for the validation phase. Considering that only the bioinformatics data from individuals II.3–110875 and II.8–080312 were available in the prioritization stage, only variants of interest present in the same genes as variants prioritized and validated in family 1 were selected.

A total of eight informative DNA samples were therefore sent to BGI for WES, and we later received back their respective raw sequencing data.

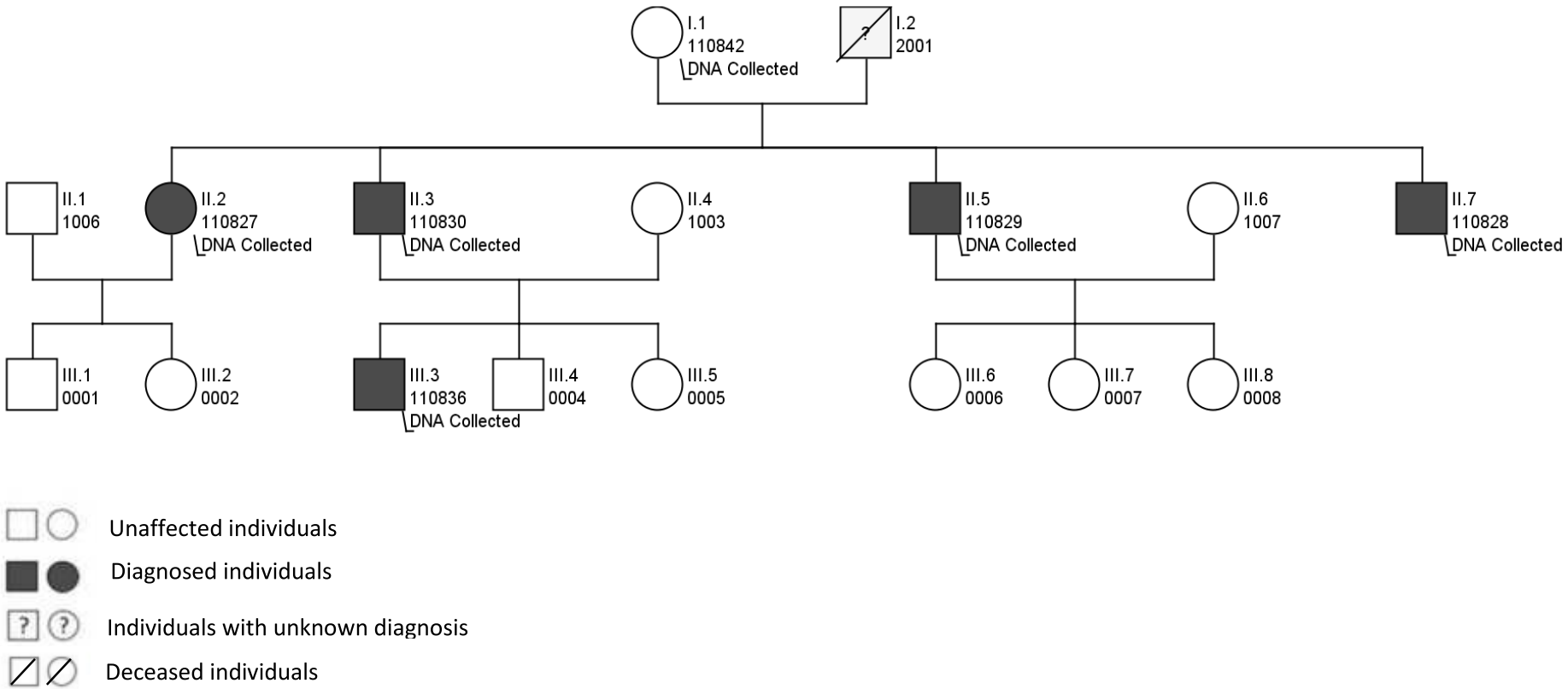

Figure 7. Family 1 genogram. All individuals from whom DNA was available (I.1-110842, II.2-110827, II.3-110830, II.5-110829, II.7-110828 and III.3-110836) were whole-exome sequenced and were also used for Sanger sequencing validation. This genogram was constructed using CeGat Pedigree Chart Designer (http://www.cegat.de/en/diagnostics/pedigree-chart-designer/).

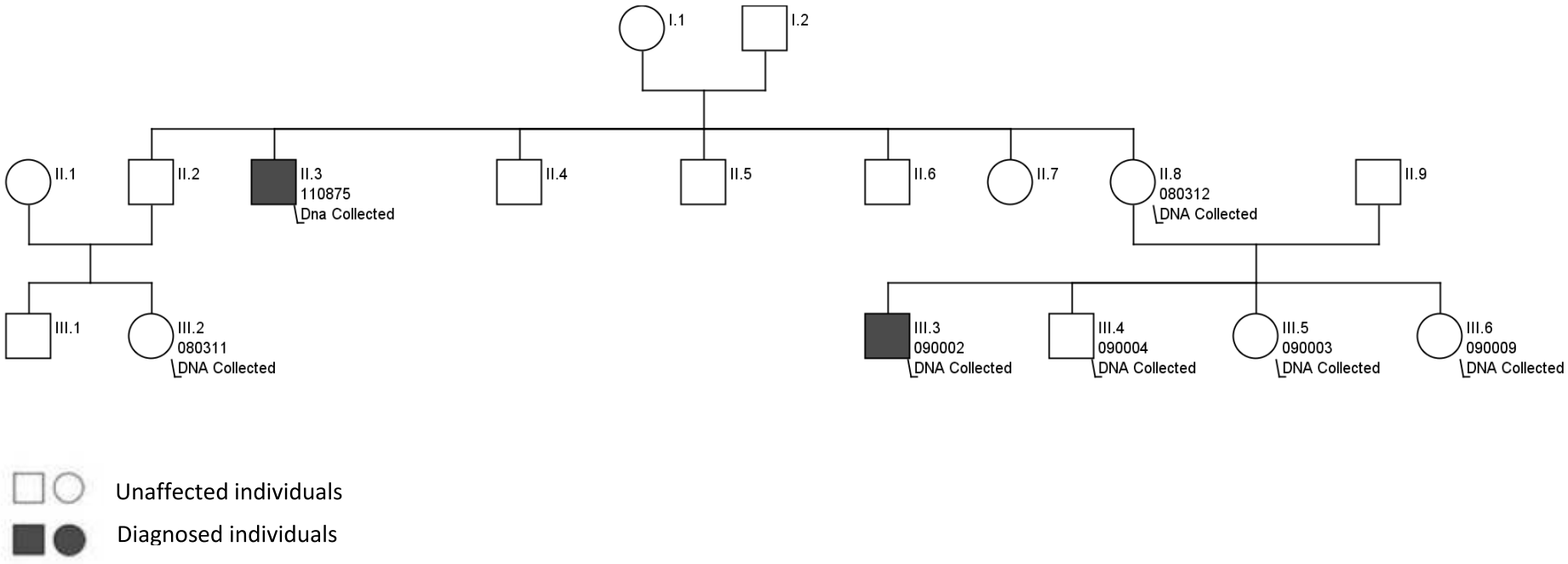

Figure 8. Family 2 genogram. All individuals from whom DNA was available were used for Sanger sequencing validation, but only individuals II.3–110875 and II.8–080312 were whole-exome sequenced. This genogram was constructed using CeGat Pedigree Chart Designer (http://www.cegat.de/en/diagnostics/pedigree-chart-designer/).

Table 3. Key clinical and demographic characteristics of families 1 and 2.

Individual ID	Family	Affection status	Gender	Age at diagnosis	Year of birth	Smoker	Nº of episodes and characteristics of the 1st episode	Symptoms at onset	Other observations
I.1–110842	1	Unaffected	F	-	1946	Yes	-	-	History of pleurisy, high blood pressure and high cholesterol
II.2–110827	1	Affected	F	34	1972	Yes	1 episode; Unilateral (L); 33% of collapse	Localized pain; Tachycardia	History of asthma
II.3–110830	1	Affected	M	24	1970	Yes	4 episodes; Unilateral (R);	Localized pain; Dyspnea	History of asthma and high blood pressure
II.5–110829	1	Affected	M	16	1969	No	2 episodes; Bilateral	Localized pain; Dyspnea	History of high cholesterol, high blood pressure and pneumonia
II.7–110828	1	Affected	M	Unknown	1972	Yes	4 episodes; Unilateral (R)	Dyspnea	History of asthma and high cholesterol
III.3–110836	1	Affected	M	14	1998	No	1 episode; Unilateral (R)	Localized pain; Dyspnea	History of pneumonia (after pneumothorax)
II.3–110875	2	Affected	M	18	1951	No	2 episodes; Unilateral (L)	Localized pain	History of asthma, high blood pressure, high cholesterol and pneumonia
II.8–080312	2	Unaffected	F	-	1949	No	-	-	History of high blood pressure
III.2–080311	2	Unaffected	F	-	1986	No	-	-	_
III.3–090002	2	Affected	M	26	1983	No	1 episode; Unilateral (L); 33% of collapse; Stage I	Localized pain; Dyspnea; Tachycardia	_
III.4–090004	2	Unaffected	M	-	1991	No	-	-	History of asthma
III.5–090003	2	Unaffected	F	-	1972	No	-	-	History of high cholesterol
III.6-090009	2	Unaffected	F	-	1979	No	-	-	_

Abbreviations: F – Female; M – Male; L – Left lung; R – Right lung.

3.2 – Bioinformatics analysis

3.2.1 – Quality control

The first step in the bioinformatics pipeline consists in the QC of the raw data obtained from BGI. FastQC was the first to be used for the fastq files (from BGI) and Qualimap for our BAM files. Both tools were used to obtain the number of reads per individual, GC percentage, mapping, duplication and error rates, and the average insert size (Table 4).

The GC, mapping rate and duplicate rate were selected as the most important parameters since they most dramatically affect the subsequent analysis. These values allowed us to decide if it was viable to continue with the bioinformatics pipeline and which was the best way to proceed. For instance, if the duplicate rate was low (< 10%) it was not strictly necessary to mark or remove the duplicates, if it was in the 20-30% range duplicates should be marked or removed.

Table 4 - WES data and alignment statistics. This information was generated with both the FastQC and Qualimap programs.

Family	Individual	Nº of reads (N)	GC content (%)	Mapping rate (%)	Duplication rate (%)	Error rate (%)	Average insert size (bp)
1	**I.1– 110842**	63036246	45,85	99,1	24,43	0,28	199
	II.2– 110827	53186523	47,14	99,36	25,19	0,3	196
	II.3– 110830	47099586	46,96	99,57	27,29	0,24	202
	II.5– 110829	69753870	45,62	98,78	19,4	0,3	193
	II.7– 110828	60542266	47,43	98,58	22,5	0,25	188
	III.3– 110836	52038795	46,98	99,48	26,48	0,25	195
2	**II.3–110875**	53056603	48,56	99,37	27,61	0,29	208
	II.8– 080312	55137472	47,15	99,41	24,37	0,27	213

3.2.2 – Ti/Tv ratios

The second QC step were the Ti/Tv ratios which were calculated for all SNP files using all four variant calling tools (described previously in Chapter 2). Ti/Tv ratio reference values in the literature vary between ~ 2,0 – 2,1 for the whole genome and ~ 3,0 – 3,3 for coding regions (DePristo *et al.,* 2011).

A combined analysis was then performed in which all variants that do not match between all four tools were filtered out in an attempt to keep only true variants. The combined analysis achieved a higher Ti/Tv score (2,33) than any of the variant calling tools individually (1,84-2,24) suggesting a lower false positive rate with this combined analysis (Table 5). While this Ti/Tv ratio is lower than expected for a WES assay in the literature, our RNA library included exonic, intronic and intergenic regions and a significant portion of our called variants (> 50%) were from intronic and intergenic regions, which

explains a significant lower Ti/Tv ratio. With this in consideration, the variants obtained with the combined analysis were the ones used in the subsequent SNPs/InDels analyses.

Table 5 – Ti/Tv ratios. The ratios were calculated for each individual using the four variant calling tools individually and in combination (shaded in dark grey).

Tool	I.1–110842	II.2–110827	II.3–110830	II.5–110829	II.7–110828	III.3–110836	II.3–110875	II.8–080312	Average	SD
UG	1,92	1,82	1,94	1,88	1,86	1,89	1,94	1,96	1,90	0,05
HC	2,20	2,27	2,28	2,15	2,23	2,24	2,29	2,25	2,24	0,05
FB	2,01	1,94	2,13	1,95	1,91	2,07	2,09	2,13	2,03	0,09
SAM	1,84	1,75	1,90	1,81	1,74	1,85	1,88	1,94	1,84	0,07
CA	2,29	2,36	2,37	2,23	2,31	2,33	2,38	2,33	2,33	0,05

Abbreviations: UG – UnifiedGenotyper GATK; HC – HaplotypeCaller GATK; FB – Freebayes; SAM – Samtools mpileup; CA – Combined analysis.

3.2.3 – Variant prioritization

Assuming an autosomal dominant with complete penetrance mode of transmission of PSP in family 1, the first step in variant prioritization was to identify those variants which were present in all affected individuals and absent in the unaffected individual. A single file for SNPs (with a total of 3231 variants) and another for InDels (with a total of 228 variants) was obtained. Variants were then prioritized in two different groups based on how they scored in selected parameters.

The first group had a stricter prioritization and resulted in five variants prioritized: c.10855G>C (located in *DYNC2H1*), c.1440G>C (*STOX2*), c.*106G>A (*DCLK2*), c.*162delT (*TRAPPC11*) and c.1067+3A>C (*TDO2*). Since in the pre-annotation phase these variants had already passed the quality filtering parameters (Chapter 2), the main factors used here for the ranking in the first group were:

- Novel variants (absent in the 1000 GP and dbSNP142 databases);
- Type of variant: missense, 3'-UTR (untranslated region) and splice regions were given priority while intronic and intergenic regions were not prioritized;
- EUR_MAF (minor allele frequency for European populations) at least lower than 5% (and, preferentially, lower than 0.5%);
- SIFT and Polyphen-2 values considered damaging;
- If the variants were also present with the parallel automated pipeline BCBIO;
- Exomiser, that while not compulsory was considered as a good indication (Figure 9).

The second group, resulted in four filtered variants: c.1015G>A (*MAP2K3*), c.280T>C (*RFPL1*), c.582_583dupAC (*NOP16*) and g.34729259_34729260insACCAGC. The idea behind this second group was to include variants for validation that were not such good candidates as those in the first group but were also considered worthy of further investigation. In fact, the biggest difference between the first and second group was that the first group had defined EUR_MAF values while the second one did not. Other differences included unknown or not damaging SIFT and Polyphen-2

values, different types of variants (regulatory region and frameshift), and IGV results contrary to the results of our bioinformatics pipeline. Despite this, these variants presented some good indicators. For example, variant c.1015G>A had unknown frequency values but was predicted to be deleterious/damaging by SIFT and Polyphen-2..

A total of eight variants were selected in family 1 (Table 6).

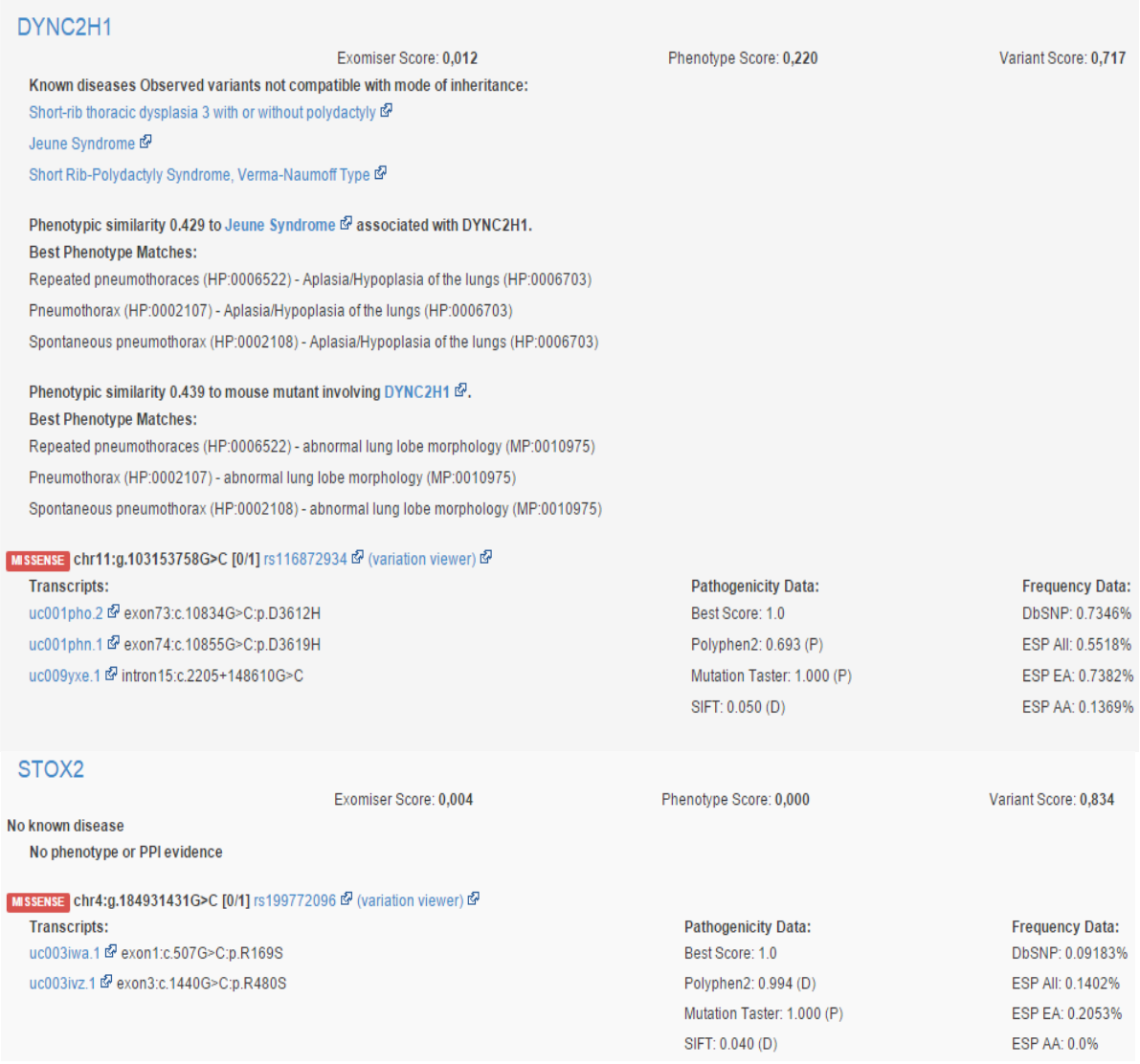

Figure 9 – Exomiser results for two prioritized variants (c.10855G>C and c.1440G>C) and their respective genes (*DYNC2H1* and *STOX2*). Exomiser matches requested phenotypes (in this case: spontaneous pneumothorax, pneumothorax and repeated pneumothorax) with known phenotypes and associated variants across all available species. It then uses these matches to attribute a “Phenotype Score”, calculates a “Variant Score” based on pathogenicity and frequency data, and an overall “Exomiser Score”. The input file and parameters used are described further in Appendix D.

Table 6 – Variants prioritized in family 1 for Sanger sequencing validation. Novel and rare variants were ranked according to several parameters, from which the most important was the resulting type of change and, when applicable, the frequency. The order in which they are presented in this table is indicative of the final ranking order.

Variant ID[A]	Alleles	dbSNP ID	Chr	Position[B]	Gene	Exon/Intron	Type of change	EUR_MAF	SIFT	Polyphen-2	Exomiser	BCBIO
c.10855G>C	G/C	rs116872934	11	103153758	*DYNC2H1*	Exon - 74 (in 90)	Missense	0,007	0,050	0,693	Yes	Yes
c.1440G>C	G/C	rs199772096	4	184931431	*STOX2*	Exon - 3 (in 4)	Missense	0,000	0,020	0,997	Yes	Yes
c.*106G>A	G/A	rs553352770	4	151177505	*DCLK2*	Exon - 17 (in 17)	3' UTR	0,001	-	-	No	Yes
c.*162delT	T/-	Novel Variant	4	184633959	*TRAPPC11*	Exon - 30 (in 30)	3' UTR	-	-	-	-	Yes
c.1067+3A>C	A/C	Novel Variant	4	156839394	*TDO2*	Intron - 11 (in 11)	Splice region	-	-	-	-	Yes
c.1015G>A	G/A	rs2363198	17	21217513	*MAP2K3*	Exon - 12 (in 12)	Missense	N/K	0,010	0,870	No	Yes
c.280T>C	T/C	rs16987627	22	29835060	*RFPL1*	Exon - 1 (in 2)	Missense	N/K	0,350	0,000	No	Yes
c.582_583dupAC	- /GT or -/AT	rs56989856	5	175811094-175811095	*NOP16*	Exon - 5 (in 5)	Frameshift	N/K	-	-	No	Yes
g.34729259_34729260insACCAGC	/ACCAGC	rs146569082	22	34729259-34729260	-	-	Intergenic regulatory region	N/K	-	-	No	Yes

Abbreviations: N/K – Unknown; Chr – Chromosome; UTR – Untranslated Region; EUR_MAF – Minor allele frequency for European populations;

[A] * - Indicative of 3' UTR variant; del – Deletion; dup – Duplication; ins – Insertion.

[B] Genomic position according to GRCh37/hg19.

Although it was not used to differentiate between the two groups of prioritized variants, conservation scores were also calculated using both PhyloP and PhastCons (Table 7). The majority of the variants were in conserved regions but c.*106G>A, c.280T>C and g.34729259_34729260insACCAGC appears to lie in less conserved regions. Since these tools are only predictive, we proceeded to consider these variants for validation.

Finally, Fisher Strand Bias (FSB) was calculated for all selected variants (Tables 8 and 9). While the annotation gives us an FSB value for the majority of the variants, it was not available for all of them. Even for the variants with an FSB value, it is only an approximation, so we decided to calculate this parameter ourselves. This is done by constructing a 2x2 table with the number of reference and alternate alleles for the positive and the negative strands visualized in IGV. With this 2x2 table we can calculate the result for a Fisher's Exact Test (FET) from which we can then calculate the FSB [FSB=-10*LOG_{10}(FET)]. An example is described in Appendix H.

FET values range between 0 and 1, 0 being the highest possible strand bias and 1 corresponding to no strand bias. In turn, FSB values range between 0 and infinity, 0 corresponding to no strand bias while the higher the value obtained higher the strand bias will be. For a significance of 5%, the FET and FSB values for variant c.1015G>A in individual II.3–110830 indicate a possible strand bias, but considering that in this variant for the remaining individuals the values were not significant, we decided to use the results from this individual nonetheless.

Table 7 – PhyloP and PhastCons results for family 1 variants. PhyloP measures acceleration (faster evolution than expected under neutral drift) as well as conservation (slower than expected evolution) ranging from - ∞ to +∞. PhastCons is a hidden Markov model-based method that estimates the probability that each nucleotide belongs to a conserved element, based on the multiple alignment, measuring the probability of negative selection for a small region, ranging from 0 to 1.

Variant ID	Position[a]	PhyloP	PhastCons
c.10855G>C	chr11: 103153758	6,565	1
c.1440G>C	chr4: 184931431	0,569	1
c.*106G>A	chr4: 151177505	-1,646	0
c.*162delT	chr4: 184633959	2,471	1
c.1067+3A>C	chr4: 156839394	2,471	1
c.1015G>A	chr17: 21217513	7,165	1
c.280T>C	chr22: 29835060	-2,514	0,134
c.582_583dupAC	chr5: 175811094-175811095	-1,318	0
g.34729259_34729260insACCAGC	chr22: 34729259-34729260	0,182	0,075

[a] Genomic position according to GRCh37/hg19.

Table 8 – Fisher strand bias results for family 1 variants in individuals I.1–110842, II.2–110827 and II.3–110830.

		I.1–110842				II.2–110827				II.3–110830			
		PLUS	MINUS	FET	FSB	PLUS	MINUS	FET	FSB	PLUS	MINUS	FET	FSB
c.10855G>C	REF	13	15	-	-	11	9	0,746	1,274	6	6	0,716	1,449
	ALT	0	0			8	9			10	7		
c.1440G>C	REF	44	43	-	-	23	24	0,836	0,776	24	28	1	0
	ALT	0	0			21	25			20	25		
c.*106G>A	REF	6	7	-	-	1	5	1	0	2	4	0,604	2,187
	ALT	0	0			1	4			2	8		
c.*162delT	REF	4	14	1	0	0	8	0,536	2,708	3	16	1	0
	ALT	0	1			3	17			2	13		
c.1067+3A>C	REF	5	14	-	-	3	8	0,642	1,922	6	6	0,243	6,152
	ALT	0	0			4	5			4	12		
c.1015G>A	REF	45	23	0,347	4,597	27	15	0,079	11,051	25	17	0,025	16,035
	ALT	24	7			29	6			32	6		
c.280T>C	REF	6	7	-	-	1	5	1	0	11	0	1	0
	ALT	0	0			1	4			8	0		
c.582_583dupAC	REF	7	12	0,592	2,275	2	5	0,674	1,716	3	7	1	0
	ALT	22	50			14	45			17	40		
g.34729259_34729260insACCAGC	REF	13	10	-	-	13	6	1	0	15	8	1	0
	ALT	0	0			6	3			8	5		

Abbreviations: REF – Reference strands; ALT – Alternative strands; PLUS – Positive strands; MINUS – Negative strands; FET – Fisher's Exact Test result; FSB – Fisher's Strand Bias values.

Table 9 – Fisher strand bias results for family 1 variants in individuals II.5–110829, II.7–110828 and III.3–110836.

		II.5 – 110829				II.7 – 110828				III.3 – 110836			
		PLUS	MINUS	FET	FSB	PLUS	MINUS	FET	FSB	PLUS	MINUS	FET	FSB
c.10855G>C	REF	9	8	0,545	2,636	7	7	1	0	6	7	1	0
	ALT	11	15			8	7			4	4		
c.1440G>C	REF	21	24	0,393	4,058	33	26	0,849	0,714	24	25	0,511	2,914
	ALT	24	18			29	21			20	15		
c.*106G>A	REF	5	4	0,187	7,282	3	7	1	0	3	1	0,242	6,154
	ALT	3	10			3	5			2	5		
c.*162delT	REF	2	14	1	0	0	6	1	0	0	16	0,486	3,136
	ALT	1	12			0	12			2	7		
c.1067+3A>C	REF	2	11	0,403	3,953	4	8	1	0	6	8	0,216	6,648
	ALT	4	9			2	4			2	10		
c.1015G>A	REF	41	14	1	0	50	12	1	0	26	7	0,783	1,0623
	ALT	30	11			41	10			29	10		
c.280T>C	REF	13	2	0,546	2,632	6	1	0,35	4,559	9	1	1	0
	ALT	7	0			13	0			12	1		
c.582_583dupAC	REF	2	13	0,331	4,807	6	4	0,311	5,077	5	5	0,480	3,189
	ALT	19	44			36	56			20	38		
g.34729259_34729260insACCAGC	REF	10	7	0,411	3,862	11	14	1	0	12	6	1	0
	ALT	3	5			8	11			9	5		

Abbreviations: REF – Reference strands; ALT – Alternative strands; PLUS – Positive strands; MINUS – Negative strands; FET – Fisher's Exact Test result; FSB – Fisher's Strand Bias values.

3.3 – Variants validation

The gold-standard for validation of WES data results is confirmation by Sanger sequencing. Genomic regions encompassing the variants of interest were PCR-amplified (Figures 10 and 11) and the resulting PCR products were sequenced as described in the methods section. The selected variants were also visualized for all family members using IGV to verify if they were present in the BAM files (obtained at the end of the data pre-processing phase of our bioinformatics pipeline) – Figures 12A, 13A, 14A, 15A, 16A, 17A, 18A, 19A and 20A. We were expecting that the affected individuals presented each variant, while the unaffected would not. The sequencing data was then analyzed and chromatograms for one strand for each variant are depicted in Figures 12B, 13B, 14B, 15B, 16B, 17B, 18B, 19B and 20B. Complementary strands were also analyzed and the same results were obtained (*data not shown*).

For variants c.10855G>C, c.1440G>C, c.*106G>A, c.*162delT, c.1067+3A>C, c.280T>C, and g.34729259_34729260insACCAGC, all affected individuals are heterozygous, having the reference and alternate alleles at their respective positions. However, the unaffected individual (I.1–110842) has the reference genotype, supporting the results of our bioinformatics analysis (Figures 12B, 13B, 14B, 15B, 16B, 18B and 20B).

Regarding the known c.1015G>A variant, all affected individuals are heterozygous having the reference allele (G) and alternate (A) alleles at the 21217513 bp position in chromosome 17, again supporting the results obtained with the WES analysis (Figure 17B). However, the unaffected individual (I.1–110842) also presented the variation for this position, which contradicts our bioinformatics analysis but agrees with the IGV results (Figure 17A). Possible reasons explanations for this inconsistency are discussed in Chapter 4.

For variant c.582_583dupAC, individuals II.3–110830, III.3–110836 and II.5–110829 are homozygous having the expected GT duplication at chr5:175811094-175811095 (GRCh37/hg19) positions in both strands. However, individuals I.1–110842, II.2–110827 and II.7–110828 are heterozygous presenting the expected duplication (GT) in one strand but a different insertion (AT) in the same position for the other strand (Figure 19A). Both the GT duplication and AT insertion (chr5:175811094-175811095, GRCh37/hg19) are described in databases (eg. db142). IGV presented the variant for all the individuals, confirming the validation results (Figure 19B). It also offered additional information about which duplication or insertion to expect, so we decided to calculate the strand bias for the heterozygous individuals (Table 11). Furthermore, the insertion was also present in the unaffected individual, which contradicts our bioinformatics results.

Table 10 – Fisher strand bias results for the c.582_583dupAC variant in individuals I.1–110842, II.2–110827 and II.7–110828.

	I.1–110842				II.2–110827				II.7–110828			
	PLUS	MINUS	FET	FSB	PLUS	MINUS	FET	FSB	PLUS	MINUS	FET	FSB
-/GT	11	11	0,802	0,959	7	19	0,759	1,194	15	21	0,827	1,825
-/AT	23	27			7	26			21	35		

Abbreviations: -/GT – Strands with [GT] duplication; -/AT – Strands with [AT] insertion; PLUS – Positive Strands; MINUS – Negative strands; FET – Fisher's Exact Test result; FSB – Fisher's Strand Bias values.

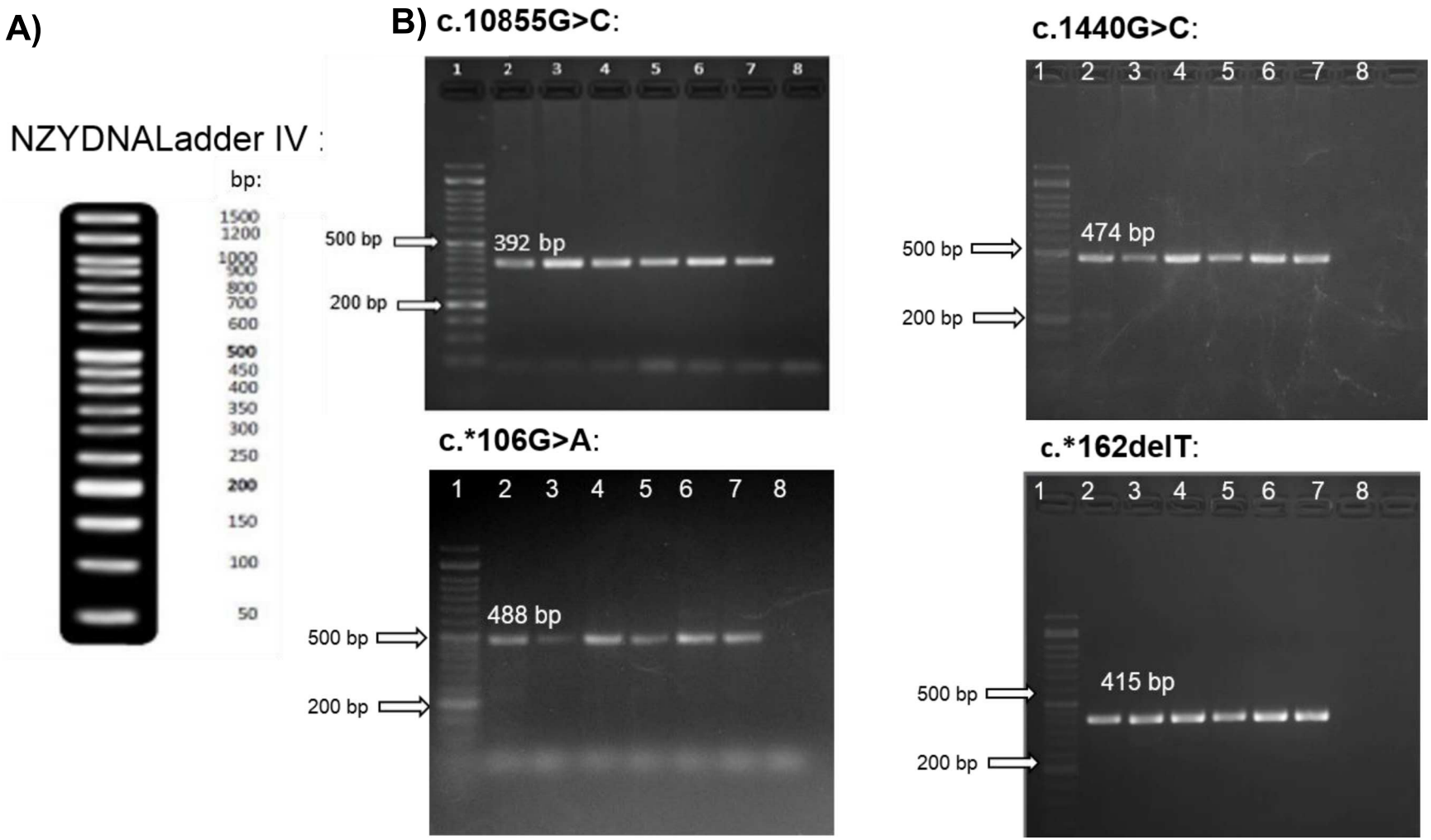

Figure 10 - A) NZYDNALadder VI B) PCR amplification for the c.10855G>C, c.1440G>C, c.*106G>A and c.*162delT variants in individuals from family 1. Lane 1 contains the NZYDNALadder VI. In lanes 2 to 7, the PCR products from individuals II.3–110830, I.1–110842, III.3–110836, II.2–110827, II.7–110828 and II.5–110829, respectively, were loaded, and lane 8 is the negative control. PCR products were run on 2% agarose gel and had the expected size of 392, 474, 488 and 415 bp, respectively.

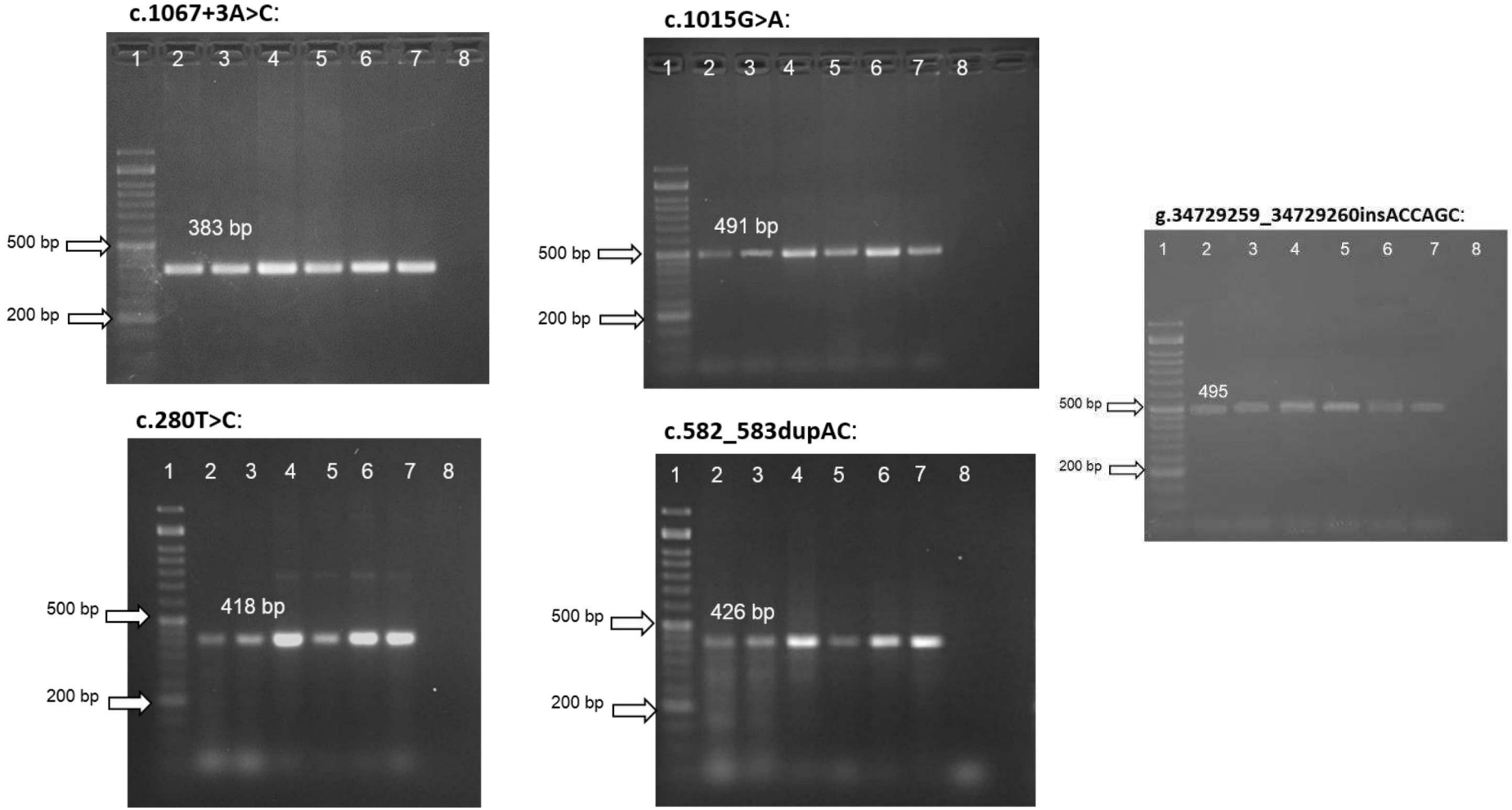

Figure 11 - PCR amplification of the c.1067+3A>C, c.1015G>A, c.280T>C, c.582_583dupAC and g.34729259_34729260insACCAGC variants in individuals from family 1. Lane 1 contains the NZYDNALadder VI, in lanes 2 to 7 are the PCR products of individuals II.3–110830, I.1–110842, III.3–110836, II.2–110827, II.7–110828 and II.5–110829, respectively, and lane 8 has the negative control. PCR products were run on 2% agarose gel and had the expected size of 383, 491, 418, 426 and 495 bp, respectively.

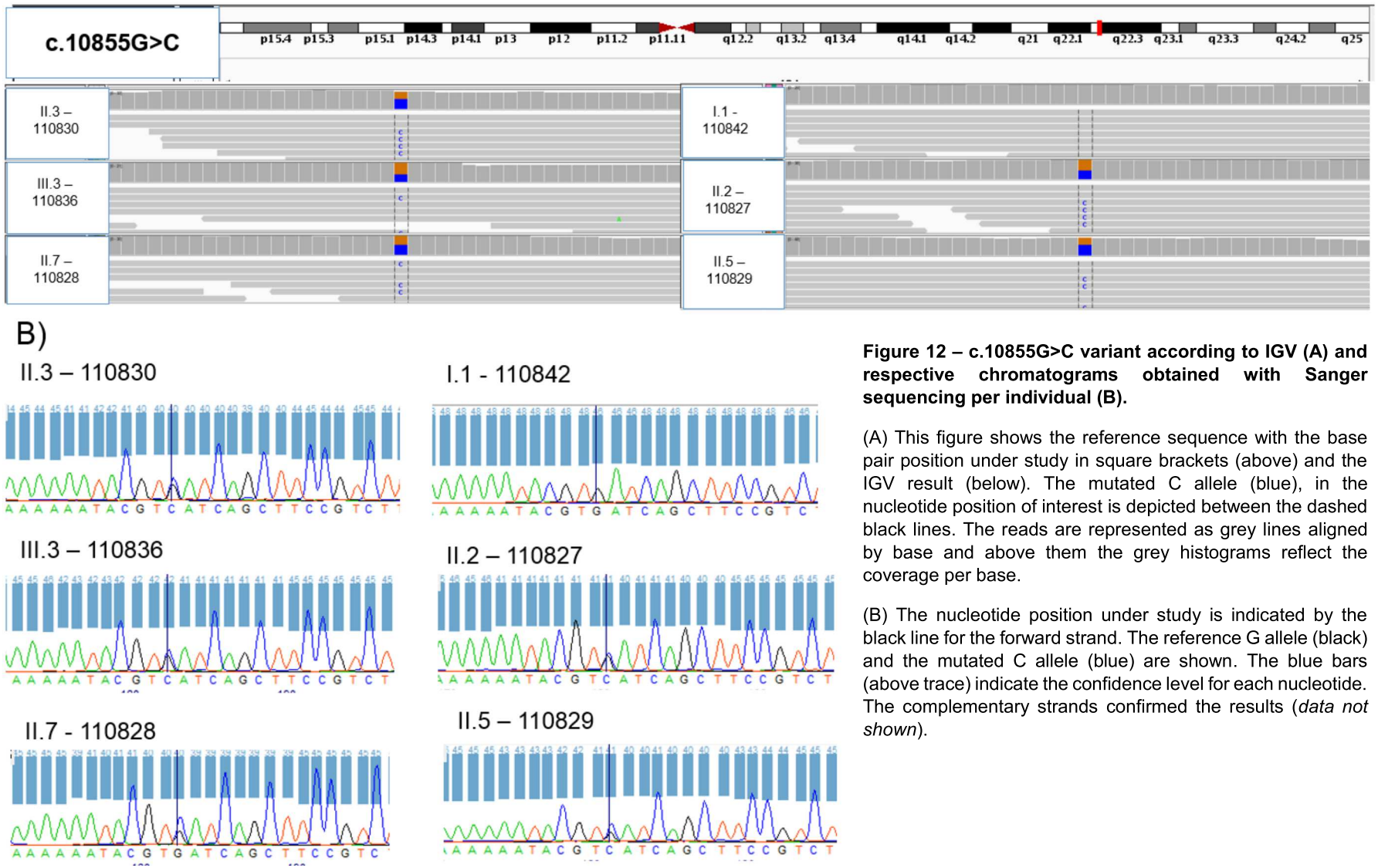

Figure 12 – c.10855G>C variant according to IGV (A) and respective chromatograms obtained with Sanger sequencing per individual (B).

(A) This figure shows the reference sequence with the base pair position under study in square brackets (above) and the IGV result (below). The mutated C allele (blue), in the nucleotide position of interest is depicted between the dashed black lines. The reads are represented as grey lines aligned by base and above them the grey histograms reflect the coverage per base.

(B) The nucleotide position under study is indicated by the black line for the forward strand. The reference G allele (black) and the mutated C allele (blue) are shown. The blue bars (above trace) indicate the confidence level for each nucleotide. The complementary strands confirmed the results (*data not shown*).

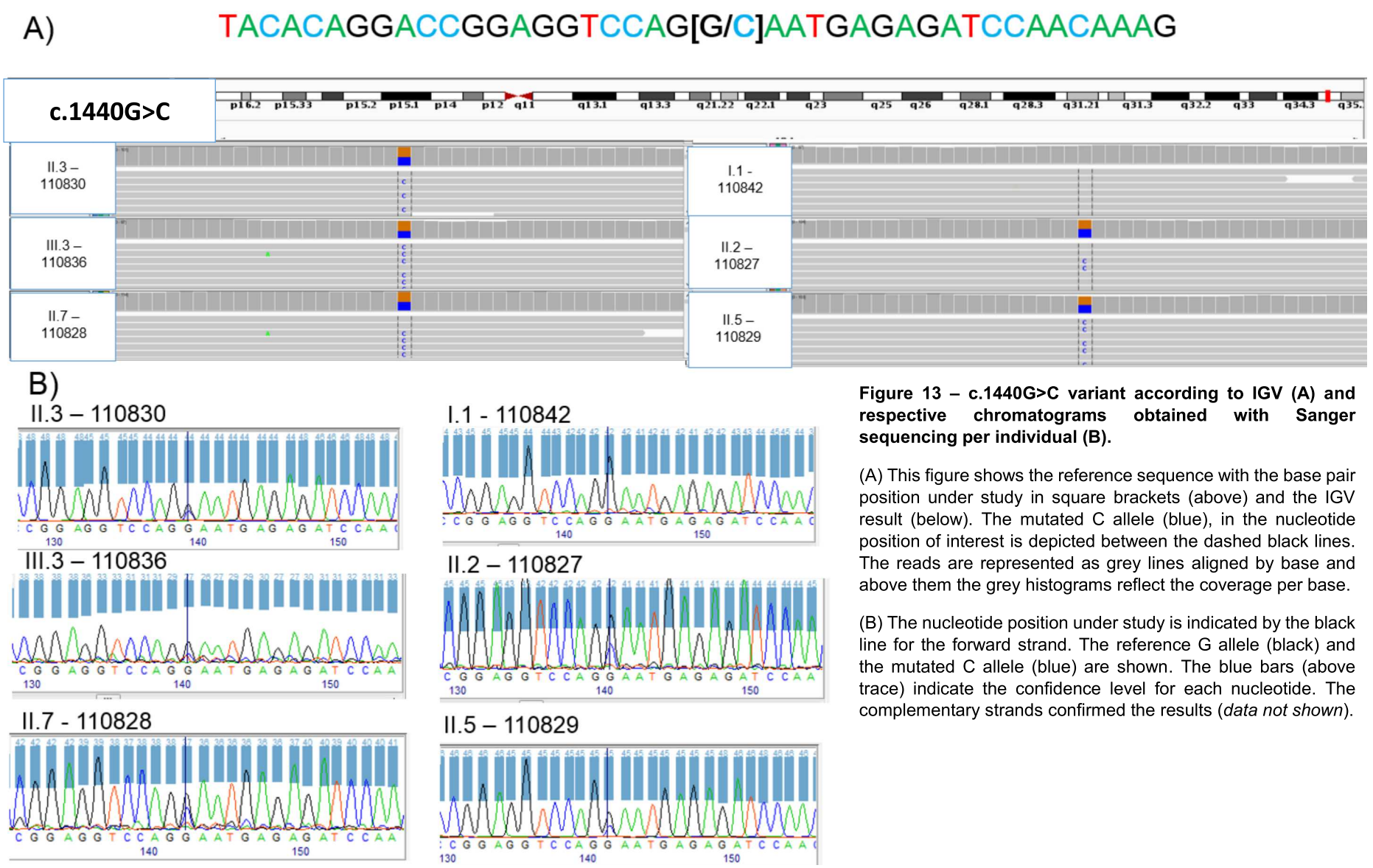

Figure 13 – c.1440G>C variant according to IGV (A) and respective chromatograms obtained with Sanger sequencing per individual (B).

(A) This figure shows the reference sequence with the base pair position under study in square brackets (above) and the IGV result (below). The mutated C allele (blue), in the nucleotide position of interest is depicted between the dashed black lines. The reads are represented as grey lines aligned by base and above them the grey histograms reflect the coverage per base.

(B) The nucleotide position under study is indicated by the black line for the forward strand. The reference G allele (black) and the mutated C allele (blue) are shown. The blue bars (above trace) indicate the confidence level for each nucleotide. The complementary strands confirmed the results (*data not shown*).

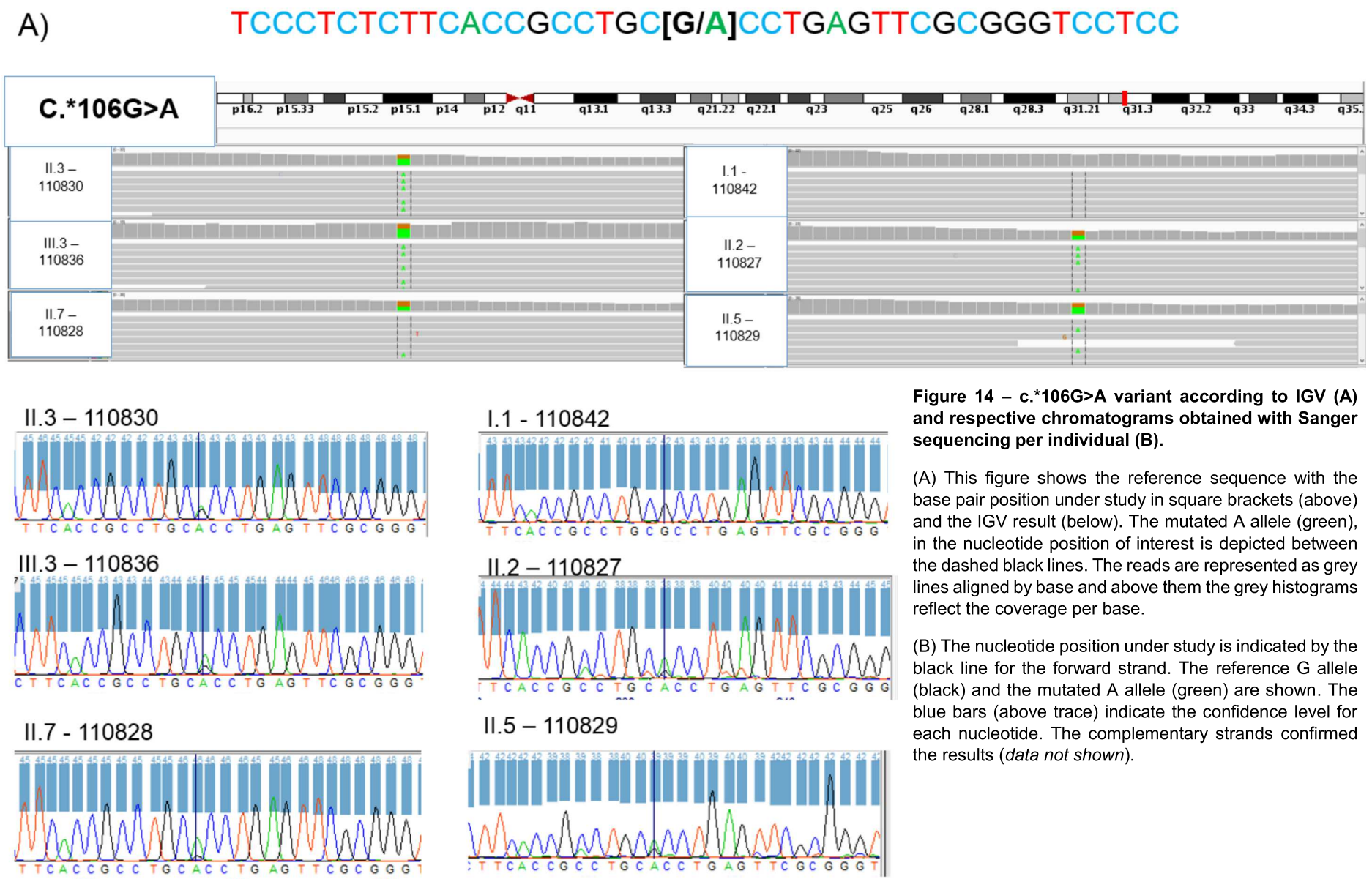

Figure 14 – c.*106G>A variant according to IGV (A) and respective chromatograms obtained with Sanger sequencing per individual (B).

(A) This figure shows the reference sequence with the base pair position under study in square brackets (above) and the IGV result (below). The mutated A allele (green), in the nucleotide position of interest is depicted between the dashed black lines. The reads are represented as grey lines aligned by base and above them the grey histograms reflect the coverage per base.

(B) The nucleotide position under study is indicated by the black line for the forward strand. The reference G allele (black) and the mutated A allele (green) are shown. The blue bars (above trace) indicate the confidence level for each nucleotide. The complementary strands confirmed the results (*data not shown*).

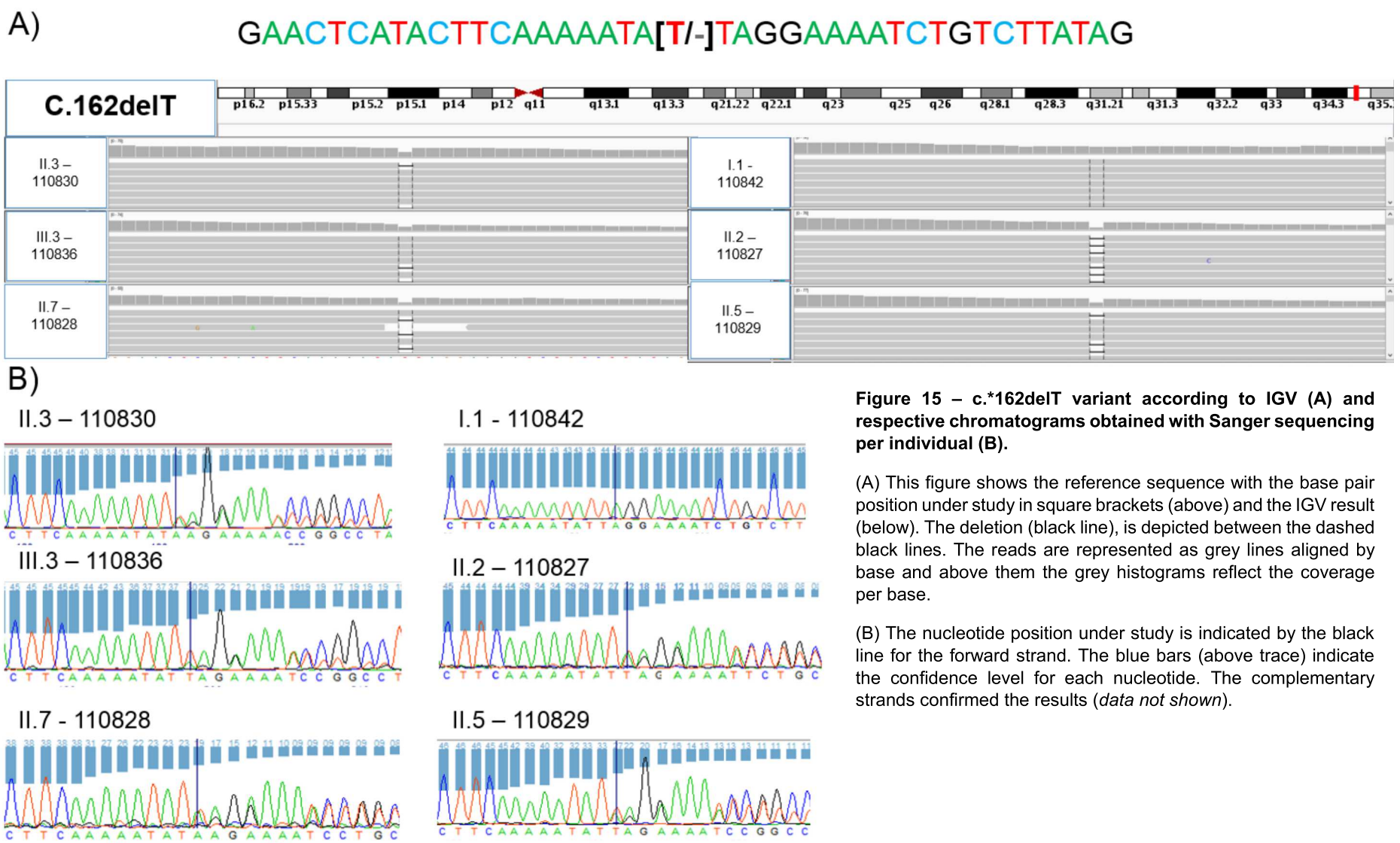

Figure 15 – c.*162delT variant according to IGV (A) and respective chromatograms obtained with Sanger sequencing per individual (B).

(A) This figure shows the reference sequence with the base pair position under study in square brackets (above) and the IGV result (below). The deletion (black line), is depicted between the dashed black lines. The reads are represented as grey lines aligned by base and above them the grey histograms reflect the coverage per base.

(B) The nucleotide position under study is indicated by the black line for the forward strand. The blue bars (above trace) indicate the confidence level for each nucleotide. The complementary strands confirmed the results (*data not shown*).

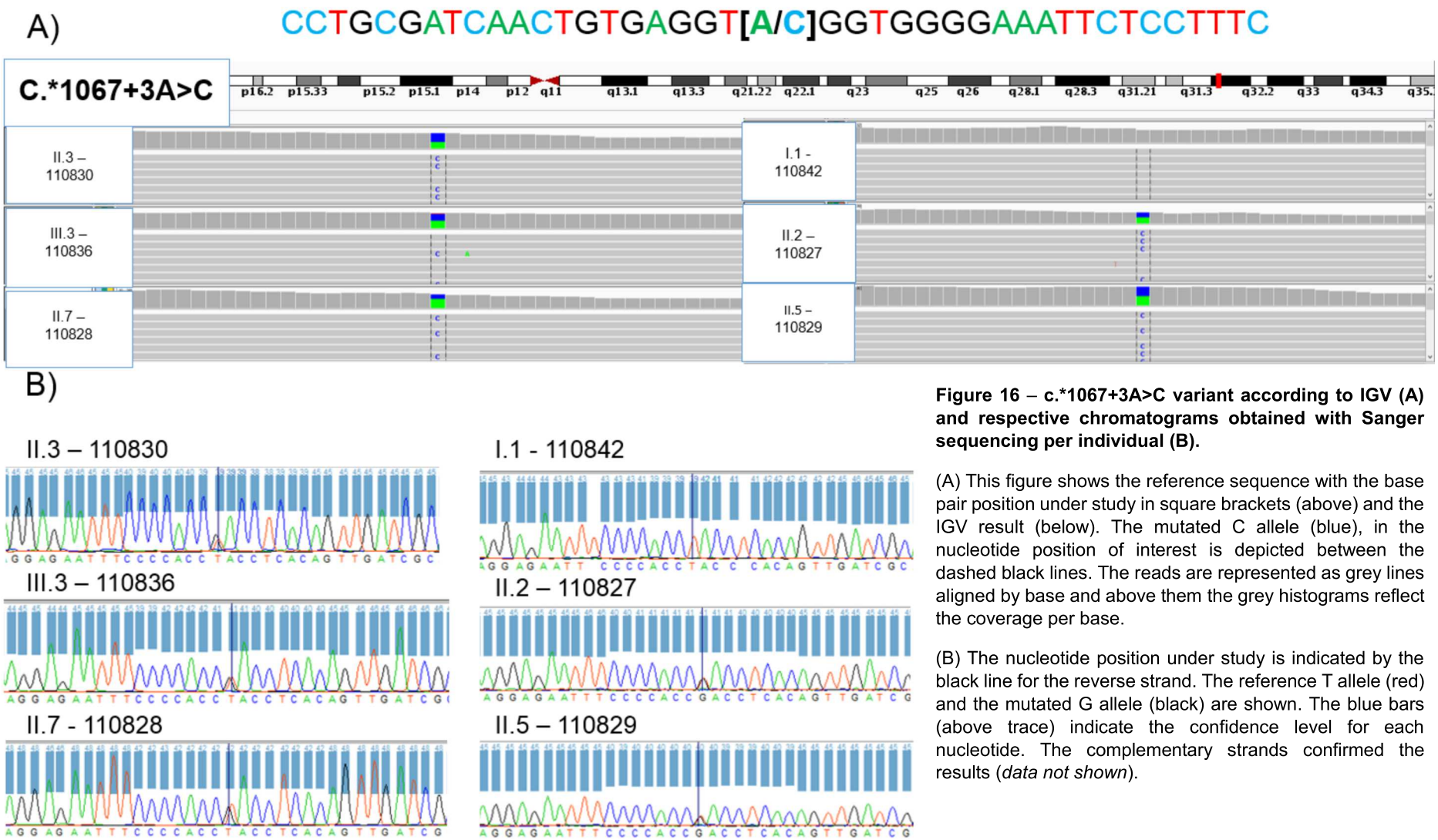

Figure 16 – c.*1067+3A>C variant according to IGV (A) and respective chromatograms obtained with Sanger sequencing per individual (B).

(A) This figure shows the reference sequence with the base pair position under study in square brackets (above) and the IGV result (below). The mutated C allele (blue), in the nucleotide position of interest is depicted between the dashed black lines. The reads are represented as grey lines aligned by base and above them the grey histograms reflect the coverage per base.

(B) The nucleotide position under study is indicated by the black line for the reverse strand. The reference T allele (red) and the mutated G allele (black) are shown. The blue bars (above trace) indicate the confidence level for each nucleotide. The complementary strands confirmed the results (*data not shown*).

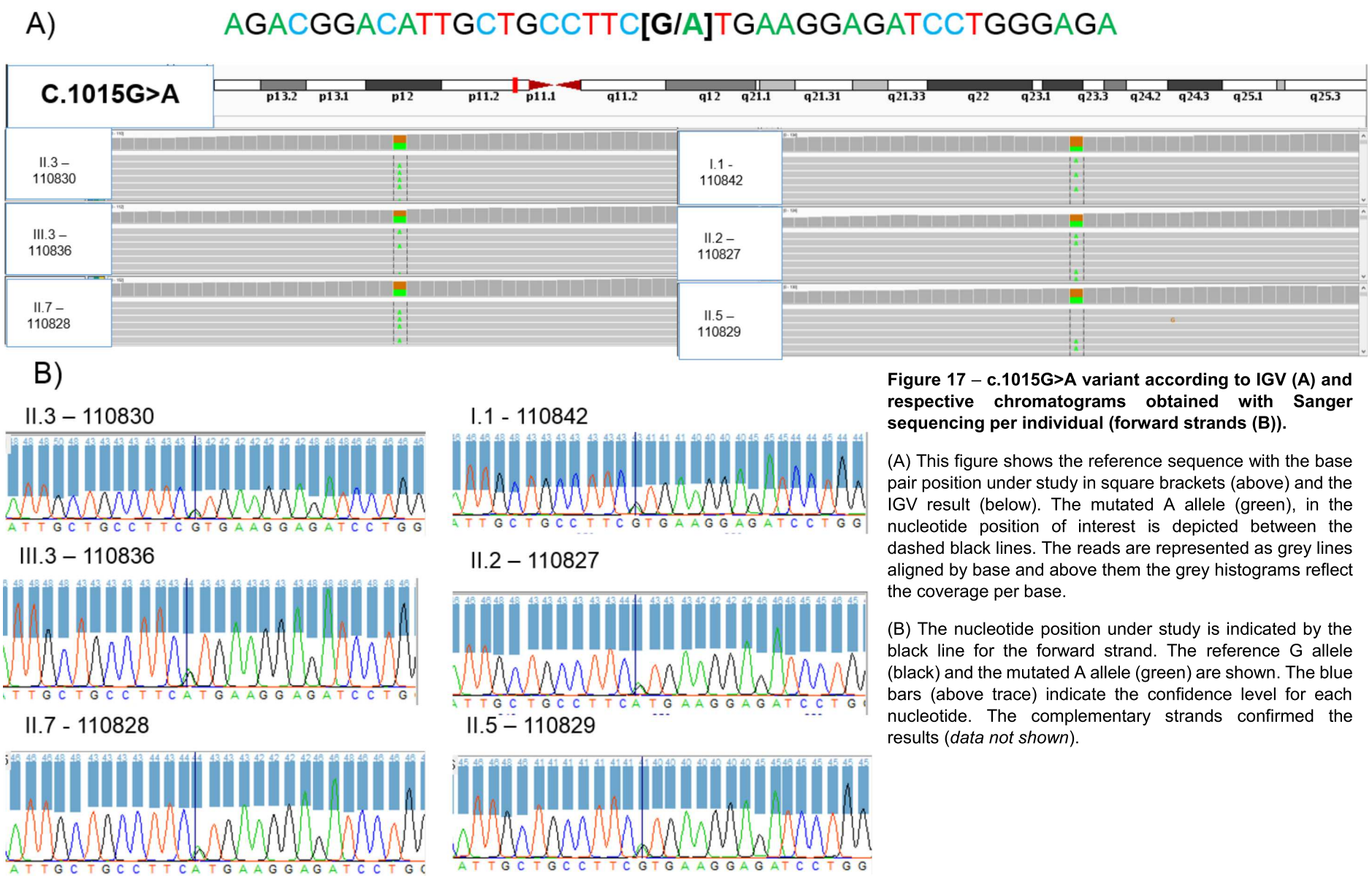

Figure 17 – c.1015G>A variant according to IGV (A) and respective chromatograms obtained with Sanger sequencing per individual (forward strands (B)).

(A) This figure shows the reference sequence with the base pair position under study in square brackets (above) and the IGV result (below). The mutated A allele (green), in the nucleotide position of interest is depicted between the dashed black lines. The reads are represented as grey lines aligned by base and above them the grey histograms reflect the coverage per base.

(B) The nucleotide position under study is indicated by the black line for the forward strand. The reference G allele (black) and the mutated A allele (green) are shown. The blue bars (above trace) indicate the confidence level for each nucleotide. The complementary strands confirmed the results (*data not shown*).

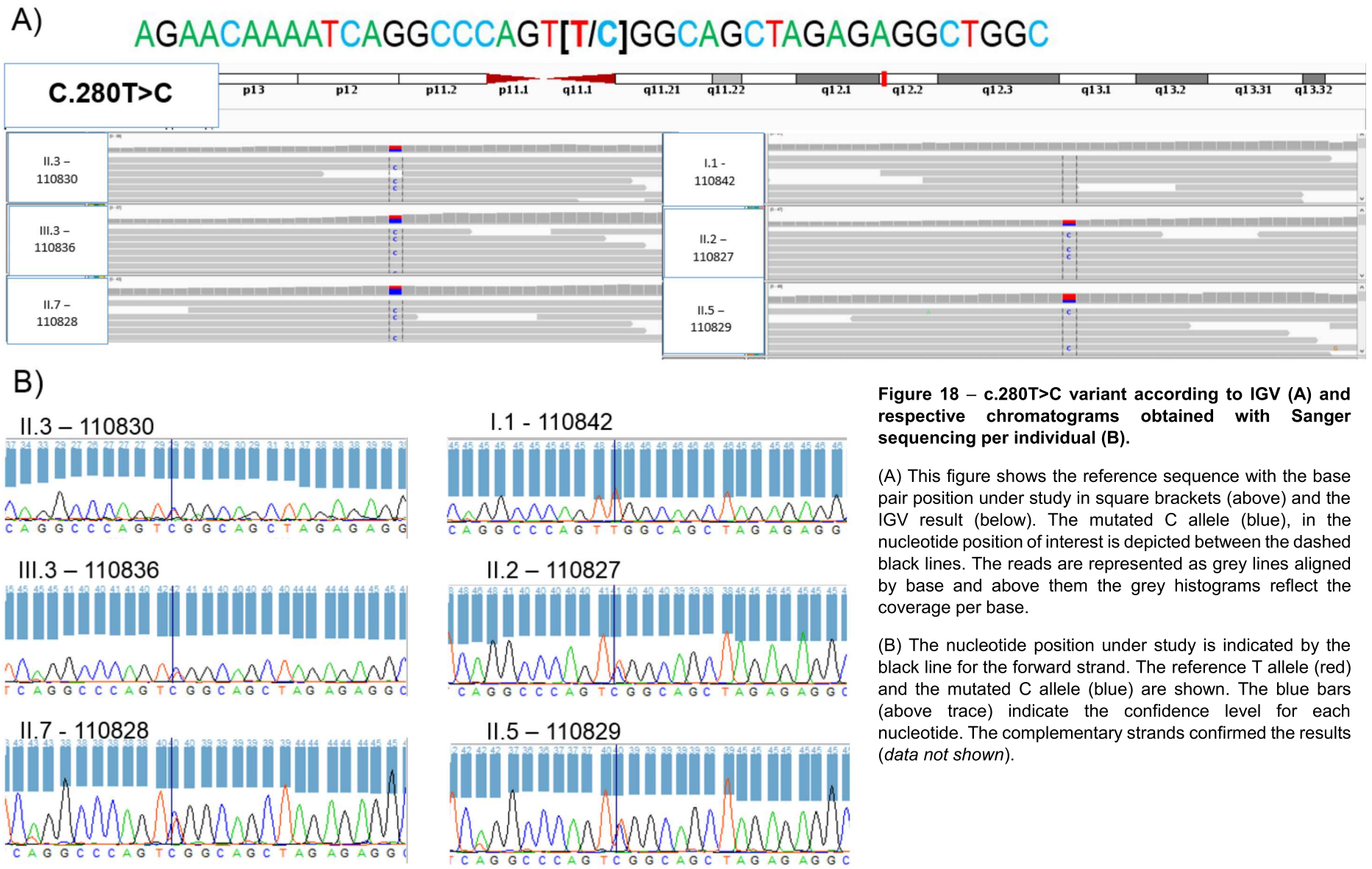

Figure 18 – c.280T>C variant according to IGV (A) and respective chromatograms obtained with Sanger sequencing per individual (B).

(A) This figure shows the reference sequence with the base pair position under study in square brackets (above) and the IGV result (below). The mutated C allele (blue), in the nucleotide position of interest is depicted between the dashed black lines. The reads are represented as grey lines aligned by base and above them the grey histograms reflect the coverage per base.

(B) The nucleotide position under study is indicated by the black line for the forward strand. The reference T allele (red) and the mutated C allele (blue) are shown. The blue bars (above trace) indicate the confidence level for each nucleotide. The complementary strands confirmed the results (*data not shown*).

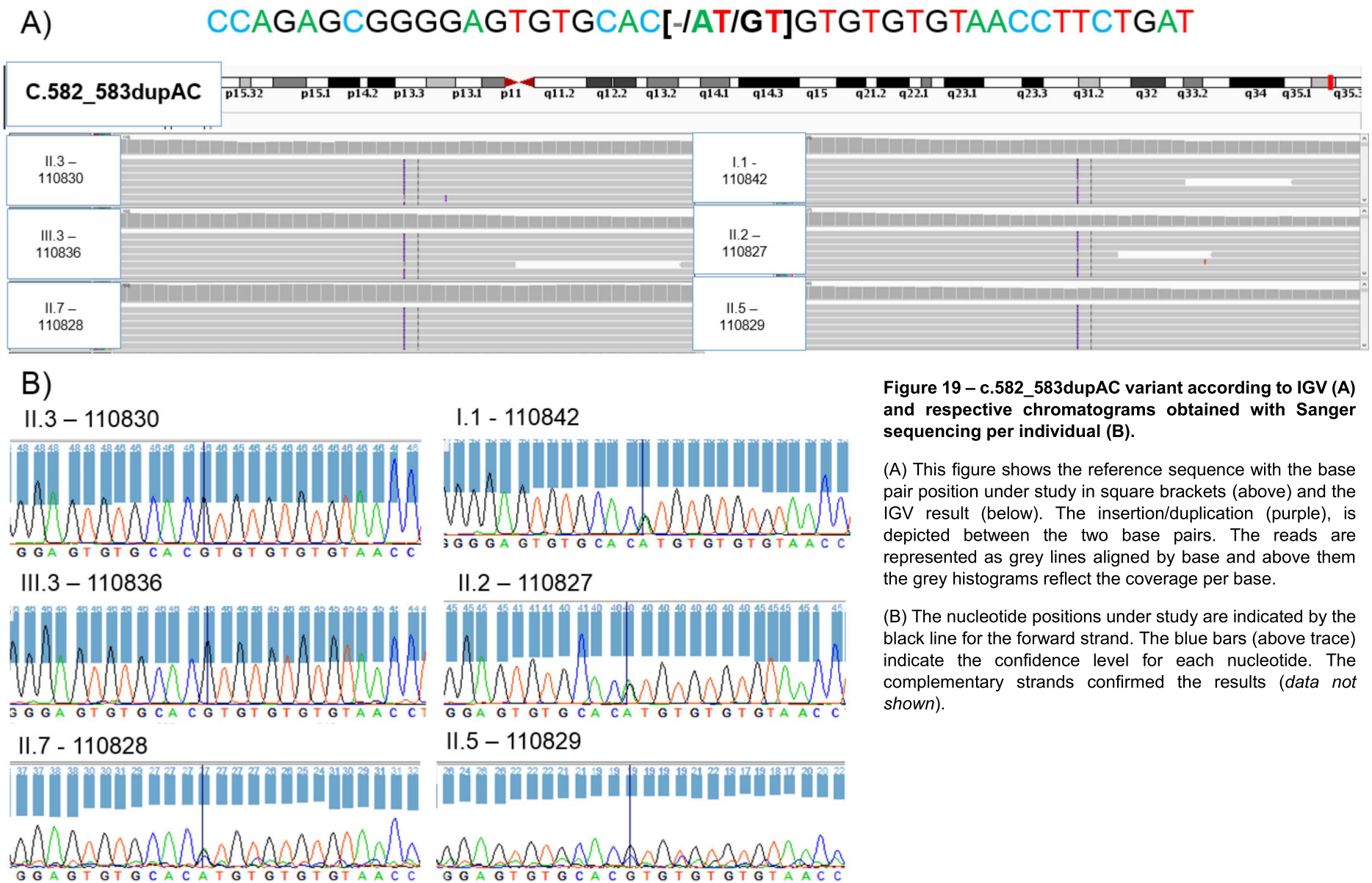

Figure 19 – c.582_583dupAC variant according to IGV (A) and respective chromatograms obtained with Sanger sequencing per individual (B).

(A) This figure shows the reference sequence with the base pair position under study in square brackets (above) and the IGV result (below). The insertion/duplication (purple), is depicted between the two base pairs. The reads are represented as grey lines aligned by base and above them the grey histograms reflect the coverage per base.

(B) The nucleotide positions under study are indicated by the black line for the forward strand. The blue bars (above trace) indicate the confidence level for each nucleotide. The complementary strands confirmed the results (*data not shown*).

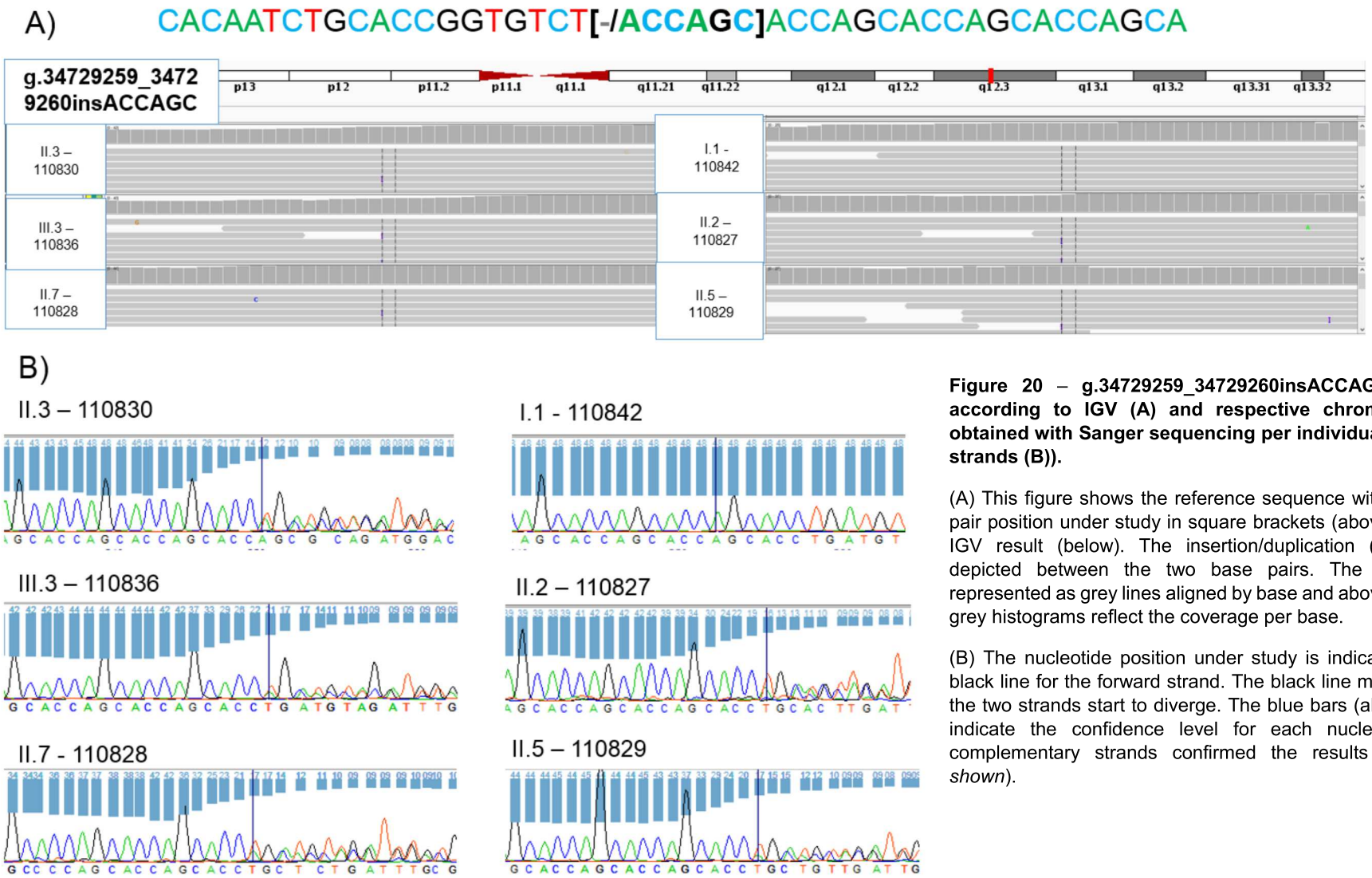

Figure 20 – g.34729259_34729260insACCAGC variant according to IGV (A) and respective chromatograms obtained with Sanger sequencing per individual (forward strands (B)).

(A) This figure shows the reference sequence with the base pair position under study in square brackets (above) and the IGV result (below). The insertion/duplication (purple), is depicted between the two base pairs. The reads are represented as grey lines aligned by base and above them the grey histograms reflect the coverage per base.

(B) The nucleotide position under study is indicated by the black line for the forward strand. The black line marks where the two strands start to diverge. The blue bars (above trace) indicate the confidence level for each nucleotide. The complementary strands confirmed the results (*data not shown*).

Sanger sequencing therefore confirmed the expected co-segregation of seven variants (c.10855G>C, c.1440G>C, c.*106G>A, c.*162delT, c.1067+3A>C, c.280T>C, and g.34729259_34729260insACCAGC) with PSP affection status in family 1. Given that the co-segregation of these variants with PSP could occur by chance, identification mutations in the same gene in other multiplex PSP families is considered the gold-standard to identify pathogenic mutations. Therefore, we used family 2 to search for mutations in these seven genomic regions.. Since we only had WES and bioinformatics data for two individuals (the affected II.3-110875 and the unaffected carrier II.8-080312), this family could only be used for results confirmation (and not in the screening phase), meaning that only variants in genes that were validated for family 1 were selected for further investigation.

Variants prioritization was similar to the one conducted for family 1. First, only variants present both in the affected individual (II.3-110875) and in the unaffected carrier (II.8-080312) were chosen. Then, only variants within or around the genes from family 1 validated variants were selected. Lastly, the same parameters previously described for family 1 were applied (such as EUR_MAF and type of change). No variants fulfilled this last step of prioritization.

Considering that, as already observed for family 1 validation, the bioinformatics analysis can discard variants that are real, the order of prioritization was switched for family 2. First, variants within or around genes of interest that fulfilled the previous described parameters (eg. variant frequency) were chosen for both individuals. One variant, in the 5' UTR of *STOX2*, in individual II.3-110875, passed this criteria (Table 11). Next, we searched its location, using IGV, in individual II.8-080312. This variant was present, albeit in a reduced number of strands (Figure 22A) and, as such, after passing a FSB test (Table 12) it was selected for validation under the HGVS name of c.-124G>A.

Table 11 – c.-124G>A variant prioritized for validation in family 2.

Variant ID	c.-124G>A
Alleles	G/A
dbSNP ID	rs114230413
Chromosome	4
Position[A]	184827820
Gene	*STOX2*
Exon/Intron	Exon - 1 (in 4)
Type of change	5' UTR
EUR_MAF	0,02
SIFT	-
Polyphen-2	-
Exomiser	No
BCBIO	Yes
PhyloP	2,228
PhastCons	1

Abbreviations: UTR – Untranslated Region; EUR_MAF – Minor allele frequency for European populations; - – Indicative of 5' UTR variant.
[A] Genomic position according to GRCh37/hg19.

Table 12 – Fisher strand bias results for the family 2 variant in individuals II.3–110875 and II.8–080312.

		II.3–110875				II.8–080312			
		PLUS	MINUS	FET	FSB	PLUS	MINUS	FET	FSB
c.-124G>A	REF	2	0	1	0	6	0	1	0
	ALT	6	0			2	0		

Abbreviations: REF – Reference strands; ALT – Alternative strands; PLUS – Positive strands; MINUS – Negative strands; FET – Fisher's Exact Test result; FSB – Fisher's Strand Bias values.

Firstly, the genomic region encompassing variant c.-124G>A was amplified in all individuals from family 2 who participated in this study (Figure 21) and PCR products were Sanger sequenced. For this variant, both individuals (II.3–110875 and II.8–080312) were heterozygous according to IGV (Figure 22A), contradicting our bioinformatics results. The technical validation presented the same results as IGV for the two individuals for which bioinformatics data was available (Figure 22B). Also, it demonstrated that the unaffected III.5–090003 presented the variant, while the remaining unaffected individuals (II.2–080311, III.4–090004 and III.6–090009) did not. Moreover, the affected individual III.3–090002 is predicted to be homozygous for the variant according to our validation (Figure 22B).

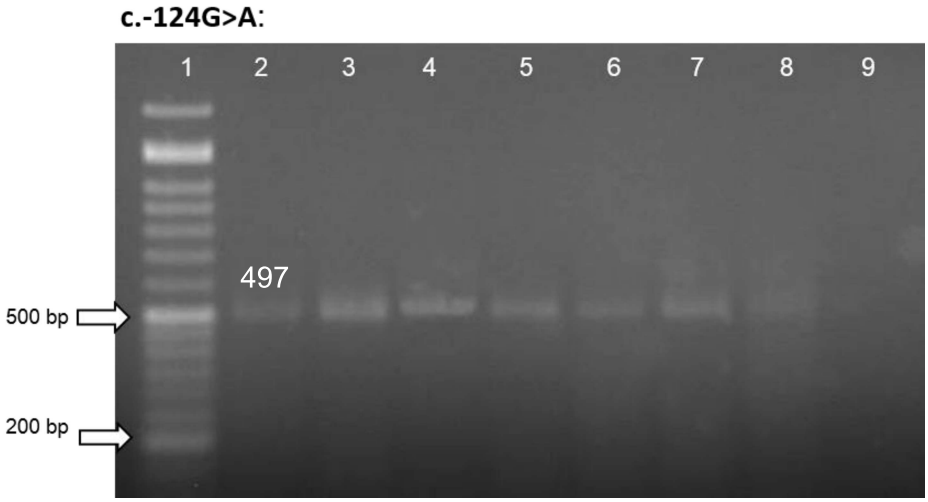

Figure 21 - PCR amplification of the c.-124G>A variant in individuals from family 2. Lane 1 contains the NZYDNALadder VI, lanes from 2 to 8 comprised respectively individuals II.2–080311, II.8–080312, III.3–090002, III.5–090003, III.4–090004, III.6–090009 and II.3 –110875 and lane 9 represents the negative control. PCR products were run on a 2% agarose gel and presented the expected size of 497 bp.

The presence of the homozygous individual was surprising considering that his parents are not known to be consanguineous and that this variant has a low frequency. To further understand this apparent Mendelian inconsistency, Sanger sequencing data was used to identify all heterozygous variants present in the region of interest. As the homozygous affected individual presented no neighboring heterozygous variants, it was possible to phase the variant of interest relative to neighboring variants and to construct the most likely haplotypes of sequenced individuals in family 2 in the region of interest (Figure 23). Hypotheses for the latter are further discussed in Chapter 4.

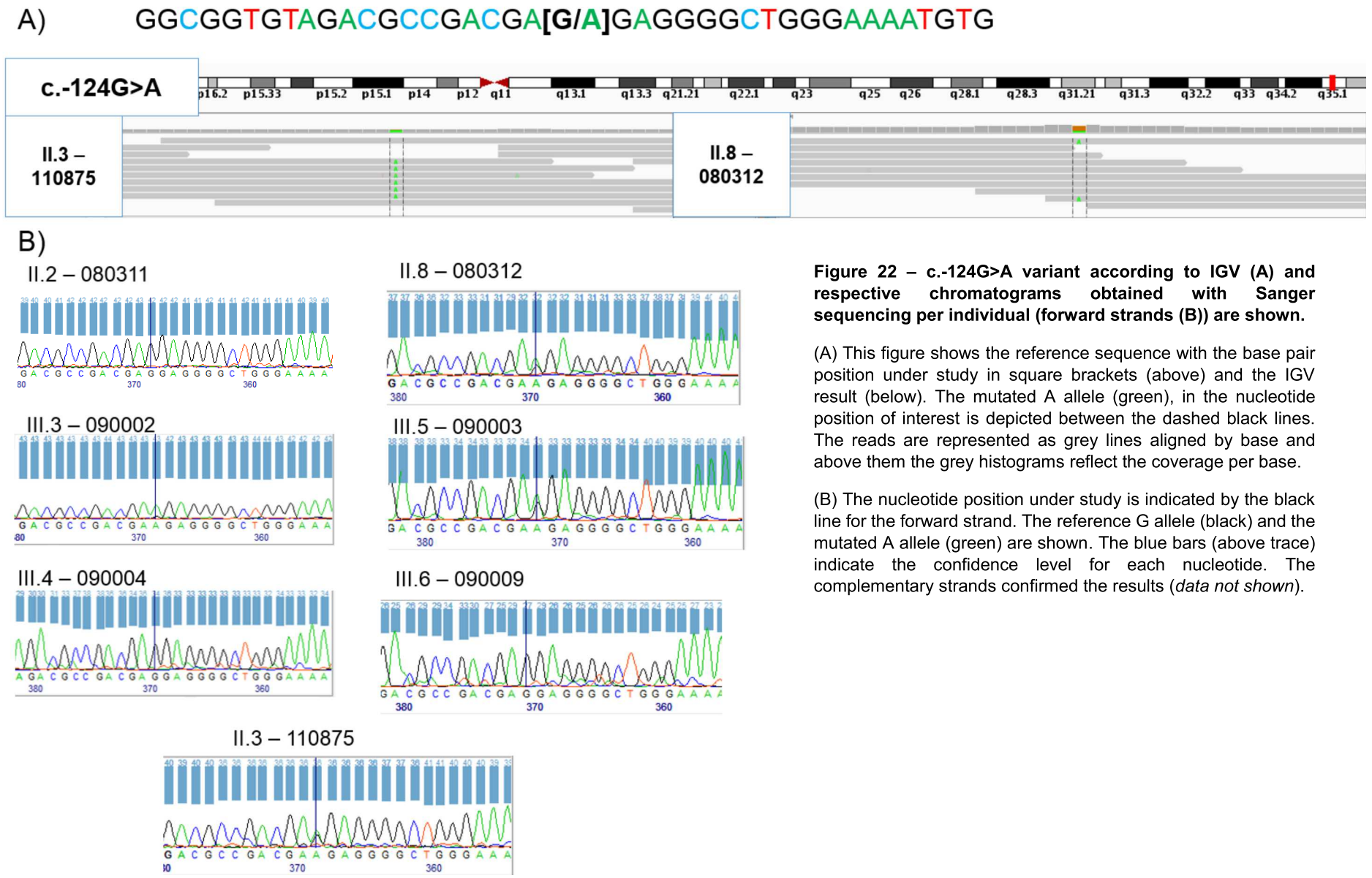

Figure 22 – c.-124G>A variant according to IGV (A) and respective chromatograms obtained with Sanger sequencing per individual (forward strands (B)) are shown.

(A) This figure shows the reference sequence with the base pair position under study in square brackets (above) and the IGV result (below). The mutated A allele (green), in the nucleotide position of interest is depicted between the dashed black lines. The reads are represented as grey lines aligned by base and above them the grey histograms reflect the coverage per base.

(B) The nucleotide position under study is indicated by the black line for the forward strand. The reference G allele (black) and the mutated A allele (green) are shown. The blue bars (above trace) indicate the confidence level for each nucleotide. The complementary strands confirmed the results (*data not shown*).

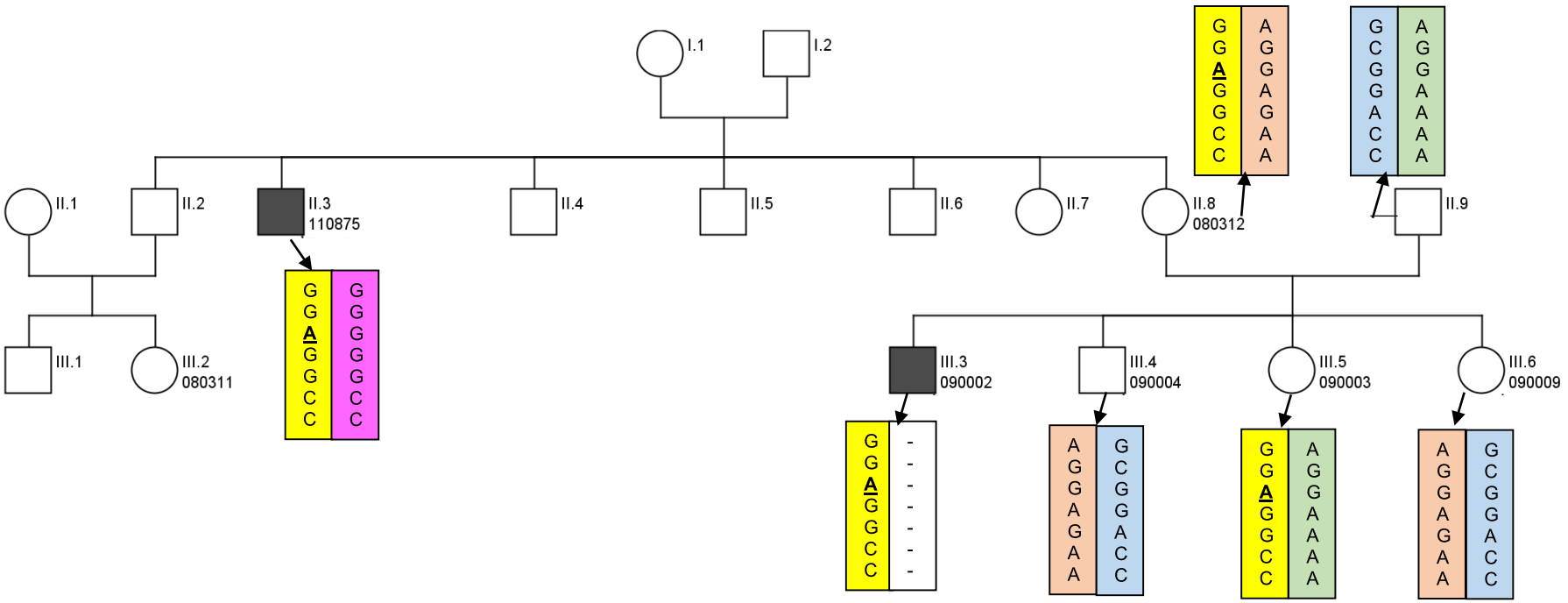

Figure 23. Family 2 genogram with predicted haplotypes around the c.-124G>A variant. Each colored box represents a different haplotype for the region of interest. The haplotype with the variant of interest is highlighted in yellow and the variant is underlined and in bold.

As a summary, a total of ten variants, nine from family 1 and one from family 2, were selected for Sanger sequencing validation. Of those, eight were successfully validated. However, there were discordant results for three variants when comparing IGV, bioinformatics analysis and validation outputs (Table 13). The possibilities for these incongruences will be further discussed in Chapter 4.

Table 13 - Summary of the results obtained with IGV, bioinformatics and sequencing analyses. This table shows the genotypes for affected and unaffected individuals when we visualized the BAM files in IGV, bioinformatics analysis and Sanger sequencing. Discordant results are highlighted in bold and underlined.

Variant ID	Gene	Individuals	IGV	Bioinformatics analysis	Sanger sequencing
c.10855G>C	*DYNC2H1*	A	G/C	G/C	G/C
		U	N/V	N/V	N/V
c.1440G>C	*STOX2*	A	G/C	G/C	G/C
		U	N/V	N/V	N/V
c.*106G>A	*DCLK2*	A	G/A	G/A	G/A
		U	N/V	N/V	N/V
c.*162delT	*TRAPC11*	A	T/-	T/-	T/-
		U	**T/- ***	N/V	N/V
c.1067+3A>C	*TDO2*	A	A/C	A/C	A/C
		U	N/V	N/V	N/V
c.1015G>A	*MAP2K3*	A	G/A	G/A	G/A
		U	G/A	**N/V**	G/A
c.280T>C	*RFPL1*	A	T/C	T/C	T/C
		U	N/V	N/V	N/V
c.582_583dupAC	*NOP16*	A	-/GT or -/At	-/GT or -/At	-/GT or -/At
		U	-/GT or -/At	**N/V**	-/GT or -/At
g.34729259_34729260insACCAGC		A	-/ACCAGC	-/ACCAGC	-/ACCAGC
		U	N/V	N/V	N/V
c.-124G>A	*STOX2*	A	G/A	G/A	G/A
		UC	G/A *	**N/V**	G/A

Abbreviations: A – Affected individual; U – Unaffected individual; UC – Unaffected carrier individual; N/V – No variant.
* For all variants (especially heterozygous ones) not all reads presented the variant in IGV, but in these variants only one or two reads had the variant in question (Tables 8 and 12).

4. Discussion

4.1 – Families and affection status

Since we only had bioinformatics results for one affected and one unaffected carrier individual in family 2, variant screening was performed in family 1. Variants of interest in family 2 belonging to the same gene as those previously confirmed in family 1 were also selected.

As pointed out in Table 1 (Chapter 1), asthma (MIM: 600807) is usually considered an exclusion factor for PSP diagnosis since it is one of the diseases which may cause SSP. We opted not to include asthma as an exclusion criteria (Appendix A) for this study since only severe asthma is described as being associated with pneumothorax and, while some affected individuals from both families 1 and 2 had moderate cases of asthma (Table 3, Chapter 3) other affected individuals (II.5–110829 and III.3–110836 from family 1, and III.3–090002 from family 2) did not.

Lastly, we must point out that while the unaffected individual I.1–110842 is considered as a good familial control for the previously mentioned reasons (Chapter 3), a history of pleurisy (inflammation of the pleura) was mentioned. Although cases of spontaneous pneumothorax have been described in individuals as a possible consequence of pleurisy, the pneumothorax only manifests at an elderly age (Hoover *et al.,* 1989) and the opposite (pneumothorax as a cause for pleurisy) has never been reported to the best of our knowledge. I.1–110842 referred to have pleurisy at the age of 9/10, which is a very young age when compared to the typical PSP age of onset. I.1–110842 is presently 69 years old and never had a clinically-diagnosed PSP episode. In contrast, not only the affected individuals in this family were never diagnosed with pleurisy, but also manifested PSP at a much younger age when compared to PSP cases in the literature that also present pleurisy (Hoover *et al.,* 1989). Additionally, pleurisy itself has never been described as a hereditary disorder. So, taken altogether, and since this control is a direct relative for all affected individuals studied (Figure 7), we then decided to use this individual as a control.

Conversely, in family 2, individual II.8–080312 has never been diagnosed with PSP but is most likely an unaffected carrier of an incompletely penetrant mutation since III.3–090002 (her son) had PSP and II.9 (her husband) has no known history of PSP. This family has a mode of inheritance for PSP consistent with an autosomal dominant transmission with incomplete penetrance.

4.2 – Bioinformatics analysis

Most of the traditional QC parameters analyzed (Table 4, Chapter 3) were within the expected ranges: GC content percentage between 45% and 50%, a mapping rate of approximately 99%, an error rate equal or below 0,3% and an average insert size of approximately 200 bp (the expected value for our chosen RNA library). The only quality parameter with a value outside the ideal range was the

duplication rate which presented values between 19% and 28%. Values above 10% for this parameter indicate the presence of amplification artifacts as discussed below.

As previously explained in Chapters 2 and 3, marking the duplicates instead of removing them was the chosen method. This is a still controversial topic amongst the NGS community. Some defend that the possibility of introducing a bias, originated from the artificial "False" duplicates that are byproducts of the amplification, should require the removal of the duplicates, while others defend that by simply removing the duplicates, valid information is being unused since there are naturally occurring "True" duplicates and, as such, one should mark them. Considering our pipeline, the decision was to mark the duplicates instead of removing them since, briefly, this allowed us to ignore the "False" duplicates while in theory maintaining the "True" duplicates (the "best" pairs) that the "Variant Calling" tools can then, based on the marking, either use to call the variant (selecting the "best" pairs) or discard it (ignoring the remaining ones). This way the loss of valid information is, theoretically, mostly avoided and also allowed us not to insert duplicate bias in our analysis (Bainbridge *et al.,* 2010).

A less controversial topic is when to either mark or remove these duplicates. In theory, they can be removed at any stage of the first phase (data pre-processing, Figure 6, Chapter 2), but it is advisable to do this after the alignment (especially if the option is to mark the duplicates). There is some debate around the notion of doing the marking/removal before or after the InDel realignment, but since this is such a recent field, no definite conclusions to the advantages of either approach have yet surfaced. In our study, we opted to mark the duplicates before the InDel realignment, since we were optimizing our pipeline from the GATK best practices workflow that marks/removes the duplicates at this stage (https://www.broadinstitute.org/gatk/guide/bp_step.php?p=1).

Considering that our final analysis produced no false positives variants called by the bioinformatics analysis that were not present in the validation) we are confident with our decision of marking the duplicates instead of removing them and doing so after the alignment, but before the InDel realignment.

The calculated Ti/Tv ratio for SNPs in the combined analysis obtained (used to gage the rate of false positive in our samples [Table 5, Chapter 3]), was of 2,33 (average of all eight individuals) with an average standard deviation (SD) of 0,05. These values for the combined analysis were than any individual tool per se. The closest was HaplotypeCaller with a 2,24 average (and SD of 0,05), with the less optimal being SAMTools mpileup with a 1,84 average (and a SD of 0,07). Interestingly UnifiedGenotyper, the most commonly used of these tools in the literature and also the one recommended by GATK best practices workflow, had the second worst performance (average of 1,90 with SD of 0,05) and Freebayes, the less commonly used of these tools, had the second best performance (average of 2,03 with SD of 0,09).

Comparing our combined analysis with the literature scores (Ti/Tv between 3,0 and 3,3 for WES [DePristo *et al.,* 2011]) we verify that our rate of false positives is potentially higher. This can be due to the novel SNPs that can reduce the global Ti/Tv ratio obtained for all the variants in the sample but, most probably, this score is lower because our chosen RNA library does not capture only the exonic region. Annotation results (after the Variant Filtration step) show that more than half the variants present in our final file are from intronic and intergenic regions. So, it is expected the Ti/Tv ratio to vary

between 2,1 and 3,0 for SNPs inside these target regions, with the value depending on the fraction of exons inside the target regions (Guo *et al.,* 2012).

The combined analysis was also performed for the InDels since selecting only the variants that were called by the four “variant calling” tools allowed for a greater confidence in each call.

4.3 – Variants of interest

Throughout the bioinformatics analysis a total of ten variants were selected for validation, using the prioritization parameters (Chapters 2 and 3), and divided into 3 distinct groups.

The first group was composed by the variants of family 1 of which we had greater pre-validation confidence (c.10855G>C, c.1440G>C, c.*106G>A, c.*162delT and c.1067+3A>C) as described in Chapter 3.

The second group was composed by family 1 variants that lacked information about certain parameters or presented less encouraging values but were still deemed interesting for validation (c.1015G>A, c.280T>C, c.582_583dupAC and g.34729259_34729260insACCAGC).

The final group was composed by one variant from family 2 (c.-124G>A) that presented encouraging parameters while simultaneously mapping within *STOX2*, a gene of an already validated family 1 variant (c.1440G>C).

Of the ten selected variants, two were novel (c.*162delT and c.1067+3A>C) and the remaining ones previously known (with either rare or unknown frequency) and described in dbSNP142.

These prioritized variants were subjected to validation by Sanger sequencing to have a confirmation with an independent method. This is a requirement since NGS techniques still have a higher error rate when compared to the traditional sequencing (Crona *et al.,* 2013). Moreover, NGS differs from Sanger sequencing in several aspects, including producing much more data in a shorter time (high throughput) and depending on the detection of pyrophosphate release on nucleotide incorporation, rather than chain termination with dideoxynucleotides. Furthermore, the NGS technique produces shorter reads (~200bp in our case) with potentially lower quality when compared to Sanger sequencing (Pareek *et al.,* 2011) and its alignment with the reference is still a much more complex endeavor (Day-Williams and Zeggini, 2011).

One of the variants was homozygous for three individuals (c.582_583dupAC) and another was homozygous for one affected individual (individual III.3-090002 for the c.-124G>A variant). All the other variants were heterozygous.

c.1440G>C is a rare (EUR_MAF=0), known (rs199772096) missense variant selected in family 1. It is a chromosome 4 G/C SNP in position 184931431 (Grch37/hg19) and is located in exon 3 (of 4) of the *STOX2* ENST00000308497 transcript. The allele change corresponds to the third position of the codon producing an AGG/AGC modification that in turn corresponds to an amino acid substitution from Arginine to Serine. Both SIFT and Polyphen-2 presented predictive damaging values for this variant (0,020 and 0,997 respectively) and both PhyloP and PhastCons presented values that indicated that this variant is conserved (0,569 and 1 respectively). Exomiser prioritized this variant as

the second pick in our analysis for the matching phenotypes (Figure 9, Chapter 3). FET and FSB values (Tables 8 and 9, Chapter3) revealed that there was no strand bias in our analysis. It is also worth mentioning that a variant of interest located in the same gene was also found in family 2.

c.-124G>A is a low frequency (EUR_MAF=0,02), known (rs114230413) 5' UTR variant of unknown frequency selected in family 2. It is a chromosome 4 G/C SNP in position 184827820 (hg19) located in exon 1 (of 4) of the *STOX2* (Storkhead Box 2 gene) ENST00000308497 transcript. Both PhyloP and PhastCons indicated that this variant is conserved (2,228 and 1, respectively). FET and FSB values (Table 10, Chapter 3) revealed that there was no strand bias in our analysis.

STOX2 has been identified as the only human paralog of *STOX1 (*Van Dijk, *et al.,* 2005) that encodes a winged-helix domain-containing transcription factor. It is believed to play a role in the differentiation of trophoblast cells. Although not much is known about *STOX2* in the literature*,* there is evidence it is a component of a unique molecular profile, globally characteristic of uncommitted stem cells (Thomas *et al.,* 2008), and is included in a transcriptional profile observed with increased inflammatory response to air pollutants (Fenstad *et al.,* 2010; Fedulov *et al.,* 2008). Moreover, *STOX2* is expressed in lung tissues (AceView and GEO profiles, NCBI).

Unexpectedly, individual III.3–090002 from family 2 was found to be homozygous at this variant by Sanger sequencing. Analysis of SNPs around the c.-124G>A variant led to the most likely haplotype results shown in Figure 23. Several hypotheses may explain the homozygosity at all SNPs around the c.-124G>A variant in individual III.3–090002: Consanguinity between parents, false paternity, *de novo* variant, uniparental isodisomy or a large deletion, with the most probable being the last one. However, further studies are required to test these possibilities. For instance, to confirm the presence of a deletion, quantitative real-time PCR may be performed. To assess uniparental isodisomy, the presence of homozygosity along chromosome 4 should be observed. Chances of false paternity could be estimated by analyzing mendelian inconsistencies in other polymorphisms throughout the genome.

c.10855G>C is a low frequency (EUR_MAF=0,007), known (rs116872934) missense variant selected in family 1. It is a chromosome 11 G/C SNP (position 103153758, Grch37/hg19) and it maps within exon 74 (of 90) of *DYNC2H1* (Dynein, Cytoplasmic 2, Heavy Chain 1 gene, MIM: 603297) ENST00000398093 transcript. The allele change occurs in the first position of the codon, producing a GAT/CAT change, which in turn corresponds to an Aspartic Acid to Histidine modification. Both SIFT and Polyphen-2 (Table 6, Chapter 3) presented predictive damaging values (0,050 and 0,693 respectively) and both PhyloP and PhastCons (Table 7, Chapter 3) indicated that this variant is significantly conserved in humans (6,565 and 1 respectively). In addition, Exomiser prioritized c.10855G>C as the top pick in our analysis (Figure 9, Chapter 3). FET and FSB values (Tables 8 and 9, Chapter 3) also revealed that there was no strand bias in our analysis. .

DYNC2H1 maps to 11q22.3 and encodes a protein involved in ciliary intraflagellar transport (IFT), an evolutionarily conserved process that is essential for ciliogenesis and plays a role in cell signaling events. DYNC2H1 is the central ATPase subunit of the IFT dynein-2 complex, the principal minus-end directed microtubule motor that drives retrograde transport of the IFT-A protein complex that regulates tip-to-base transport in cilia.

Variations in *DYNC2H1* have been associated with short-rib thoracic dysplasia, type III with or without polydactyly (MIM: 613091 [Dagoneau *et al.*, 2009; McInerney-Leo *et al.,* 2014]). Short-rib thoracic dysplasia (SRTD), with or without polydactyly, refers to a group of skeletal ciliopathies of unclear inheritance model (but predicted to be autossomal recessive [Schmidts *et al.,* 2013]) that are characterized by a constricted thoracic cage, short ribs, shortened tubular bones, and a 'trident' appearance of the acetabular roof. SRTD encompasses Ellis-van Creveld syndrome, the disorders previously designated as Jeune syndrome (or asphyxiating thoracic dystrophy), the short rib-polydactyly syndrome, and the Mainzer-Saldino syndrome. Interestingly, typical symptoms of patients with SRTD include lung aplasia and lung hypoplasia in 7,5% of the diagnosed individuals (Köhler *et al.,* 2014). Additionally, cases of pneumothorax, recurrent respiratory infections and asthma (phenotypes prevalent in family 1) have been reported in individuals afflicted by Sensenbrenner Syndrome, a genetically allelic disorder related to SRTD (Arts and Knoers, 2013). Moreover, *DYNC2H1* is expressed in the lung and pleural tissues (AceView and GEO profiles, NCBI). Considering that the inheritance model of SRTD, requires variants in both strands, it is possible that when the variation is present in only one strand (as is our case) variable expression can affect the phenotype. Finally, it is also worth mentioning that mice mutants for *DYNC2H1* presented abnormal lung lobe morphology (MGI).

c.1067+3A>C is a novel splice region variant found in family 1. It is a chromosome 4 A/C SNP (156839394, Grch37/hg19), that maps within intron 11 (of 11) of the *TDO2* (Tryptophan 2,3-Dioxygenase gene, MIM: 191070) ENST00000536354 transcript. Both PhyloP and PhastCons pointed out that this variant is conserved (2,471 and 1 respectively), while FET and FSB values (Tables 8 and 9, Chapter3) revealed that there was no strand bias in our analysis.

TDO2 maps at 4q32.1 and encodes a heme enzyme that plays a role in catalyzing the first and rate-limiting step in the kynurenine pathway, the major pathway of tryptophan metabolism. The tryptophan catabolite kynurenine is an endogenous ligand of the human aryl hydrocarbon receptor that is constitutively generated by human tumor cells via TDO2. Tryptophan catabolite kynurenine suppresses antitumor immune responses and promotes tumor cell survival and motility through the human aryl hydrocarbon in an autocrine/paracrine fashion. This pathway is active in human brain tumors and is associated with malignant progression and poor survival.

TDO2 mutations are associated with pellagra and *haemophilus influenza*, and the latter is itself associated with pneumonia (GeneCards and MalaCards databases). In addition, this gene was selected as a candidate gene for nicotine dependency for its interaction with both niacin and nicotidine (Sullivan *et al.*, 2004; Miller, 2011). *TDO2* is also expressed in lung tissues (AceView, NCBI).

c.*106G>A is a rare (EUR_MAF=0,001), known (rs553352770) 3' UTR variant selected in family 1. It is a chromosome 4 G/A SNP (151177505, GrcH37/hg19), that maps within exon 17 (of 17) of the *DCLK2* (Doublecortin-Like Kinase 2 gene, MIM: 613166) ENST00000302176 transcript. Both PhyloP and PhastCons highlighted that this variant is not conserved (-1,649 and 0, respectively). Moreover, FET and FSB values (Tables 8 and 9, Chapter 3) revealed that there was no strand bias in our analysis.

DCLK2 maps to 4q31.2-q31.3 and encodes a protein which is both a member of the kinase superfamily and the doublecortin family. It contains two N-terminal doublecortin domains (which in turn bind microtubules and regulate microtubule polymerization), a C-terminal serine/threonine protein kinase domain (which shows homology to Ca2+/calmodulin-dependent protein kinase), and a serine/proline-rich domain (between the two family domains, which mediate multiple protein-protein interactions). The microtubule-polymerizing activity of the encoded protein is independent of its protein kinase activity.

The doublecortin gene family is associated with subcortical band heterotopia, lissencephaly, epilepsy, developmental dyslexia and retinitis pigmentosa (Dijkmans *et al.,* 2010). *DCLK2* has been shown to enhance dendritic remodeling and to negatively regulate synapse maturation. Animal studies showed that this gene has a function in the establishment of hippocampal organization and the mouse knockout leads to a severe epileptic phenotype and lethality, as described in human patients with lissencephaly (Tuy *et al.*, 2008; Kerjan *et al.,* 2009.). In addition, *DCLK2* is expressed in lung and pleural tissues (AceView and GEO profiles, NCBI).

c.*162delT is a novel 3' UTR variant. It is a single base deletion of a Thymine (T) in position 184633959 from chromosome 4 (Grch37/hg19). This variant is located within exon 30 (of 30) of the *TRAPPC11* (Trafficking Protein Particle Complex, Subunit 11 gene, MIM: 614138) ENST00000334690 transcript. Both PhyloP and PhastCons indicated that this variant is conserved (2,471 and 1 respectively) and FET and FSB values (Tables 8 and 9, Chapter 3) revealed that there was no strand bias in our analysis.

TRAPPC11 maps to 4q35.1 and encodes a component of the TRAPP multi-subunit tethering complex involved in intracellular vesicle trafficking. *TRAPPC11* depletion in HeLa cells via interfering RNA resulted in partial disassembly of the TRAPP complex and Golgi fragmentation into dispersed punctae and arrested anterograde trafficking, suggesting that has a function in early trafficking events between the endoplasmic reticulum and Golgi (Scrivens *et al.,* 2011).

TRAPPC11 is described in the literature as being associated with limb girdle muscular dystrophy (MIM: 615356) and to a lesser extent with myopathy movement disorder and intellectual disability (Bögershausen *et al.,* 2013). Moreover, it is expressed in lung and pleural tissues (GEO profiles, NCBI). Furthermore, considering that the inheritance model of as limb girdle muscular dystrophy (predicted to be autosomal recessive), might require variants in both strands to express the phenotype. It is possible that when the variation is present in only one strand (as is our case), variable expression can affect the phenotype.

The unaffected individual I.1-110842 presented the deletion in one single read in IGV. However, our bioinformatics analysis did not call the variant for this individual and validation confirmed it. It is probable that the deletion in that read was an artifact from WES originating a False Positive that our analysis eliminated.

g.34729259_34729260ACCAGC is a known (rs146569082) intergenic variant of unknown frequency in a regulatory region selected in family 1. It is a chromosome 22 insertion, between positions 34729259 and 34729260 (Grch37/hg19), located in a regulatory region of the *CTCF* (CCCTC-Binding Factor gene, MIM: 604167) . PhyloP and PhastCons presented contradictory values,

with PhyloP indicating that this variant is conserved (0,182) and PhastCons indicating that the variant is not conserved (0,075). FET and FSB values (Tables 8 and 9, Chapter3) revealed that there was no strand bias in our analysis.

CTCF maps at 16q22.1 and encodes a transcriptional regulator protein. The CTCF has eleven highly conserved zinc finger domains, and is able to use different combinations of the zinc finger domains to bind different DNA target sequences and proteins. Depending upon the context of the site, the protein can bind a histone acetyltransferase complex and function as a transcriptional activator, or bind a histone deacetylase complex and function as a transcriptional repressor. If the protein is bound to a transcriptional insulator element, it can block communication between enhancers and upstream promoters, thereby regulating imprinted expression. CTCF plays a critical role in oocyte and preimplantation embryo development by activating or repressing transcription in either epigenetic regulation, gene silencing over considerable distances in the genome, and chromatin remodeling. Also, it acts as tumor suppressor and it seems to be essential for homologous X-chromosome pairing (Whitfield *et al*., 2012).

CTCF is significantly associated with intellectual disability, and its specific forms (eg. autosomal dominant 21 and intellectual disability-feeding difficulties-developmental delay-microcephaly syndrome (Court *et al.,* 2014; Watson *et al.,* 2014). Moreover, *CTCF* is highly expressed in lung and pleural tissues (AceView and GEO profiles, NCBI).

c.280T>C is a known (rs16987627) missense variant of unknown frequency selected in family 1. It is a chromosome 22 G/C SNP (position 29835060, Grch37/hg19), that maps within exon 1 (of 2) of the *RFLP1* ENST00000354373 transcript. The allele change corresponds to the first position of the codon producing a Tgg/Cgg modification that in turn corresponds to an amino acid change from Tryptophan to Arginine. SIFT and Polyphen-2 present contradictory results, while SIFT points out for a predictive damaging value (0,350), Polyphen-2 presented a predictive benign value (0,000). Both PhyloP and PhastCons indicated that this variant is not conserved (-2,514 and 0,134 respectively). Furthermore, FET and FSB values (Tables 8 and 9, Chapter 3) revealed that there was no strand bias in our analysis.

RFPL1 (Ret Finger Protein-Like 1 gene, MIM: 605968) maps at 22q12.2 and is a primate-specific gene that controls cell-cycle progression through cyclin B1/Cdc2 (Bonnefont *et al.,* 2011). It also has an evolutionary role in the neocortex development, being highly expressed in embryonic stem cell-derived neurogenesis and in fetal neocortex (Bonnefont *et al.,* 2008).

c.1015G>A is a known (rs2363198) missense variant of unknown frequency selected in family 1. It is G/C SNP in position 21217513 (Grch37/hg19) from chromosome 17, and maps within exon 12 (of 12) of the *MAP2K3* (Mitogen-Activated Protein Kinase 3 gene, MIM: 602315) ENST00000354373 transcript. The allele change corresponds to the first position of the codon producing a Gtg/Atg modification that in turn corresponds to an amino acid substitution from Valine to Methionine. Both SIFT and Polyphen-2 presented predictive damaging values for this variant (0,010 and 0,870 respectively) and both PhyloP and PhastCons presented values that indicate that this variant is conserved (7,165 and 1 respectively). FET and FSB values (Tables 8 and 9, Chapter3) revealed that

there was a possible strand bias in our analysis for II.3-110830, but our validation discarded this hypothesis.

MAP2K3 maps to 17p11.2 and encodes a specific MAP2K. MAP2Ks phosphorylate specific classes of MAKs, activating them so they can act in cellular signal transduction pathways in response to extracellular signals. This kinase can be activated by insulin, and is necessary for the expression of glucose transporter. Interestingly, the inhibition of this kinase is involved in the pathogenesis of *Yersina pseudotuberculosis (*Orth *et al.,* 1999). Also, although this variant passed our validation for the affected individuals it failed for the unaffected individual.

c.582_583dupAC is a known (rs56989856) frameshift variant of unknown frequency selected in family 1. Specifically, it is either an AT insertion or a GT duplication, between positions 175811094 and 175811095 from chromosome 5 (Grch37/hg19). This variant is located within exon 5 (of 5) of the *NOP16* (NOP16 nucleolar protein gene, MIM: 612861) ENST00000510123 transcript. Moreover, both PhyloP and PhastCons highlighted that this variant is not conserved (-1,318 and 0 respectively). Also, FET and FSB values (Tables 8 and 9, Chapter3) showed no strand bias in our analysis.

NOP16 maps to 5q35.2 and is transcriptionally regulated by c-Myc. Very little is known about this gene in the literature, but it has been described to be upregulated in breast cancer, with its overexpression being associated with poor patient survival (Butt *et al.,* 2008).

This variant was validated for the affected individuals but not for the unaffected individual, since the latter presented the variant. Additionally, this variant differed from the remaining in three main aspects:

1) c.582_583dupAC maps in a gene located in the -1 strand (all the other variants are from genes in the +1 strand). This is also why the HGVS name (given considering the gene location) for the variant presents an AC duplication and the variant is described in the literature as either a GT duplication (reverse complement of AC) or as an AT insertion.

2) Regarding c.582_583dupAC homozygosity, while all six individuals presented the insertion/duplication in both strands, three presented the GT duplication in both strands and the remaining three presented the GT duplication in one strand and the AT insertion in the other. FET and FSB values for these last three individuals (Table 11, Chapter 3) revealed that there was no strand bias between the two forms of this variant in our analysis. It is possible that this could be a potential case of allelic heterogeneity but further studies and analysis are required.

3) Lastly, c.582_583dupAC was not only flagged by Ensembl but also the Genome Reference Consortium refers that both forms of this variant are not validated and decided to eliminate it from Grch38/hg38 assembly (GRC, NCBI).

4.4 – Bioinformatics considerations

There was a lack of concordance between results from IGV, Sanger sequencing and our bioinformatics analysis (Table 13) for the c.1015G>A, c.582_583dupAC, c.-124G>A and c.*162delT variants in respect to the unaffected individual (Table 12). Considering that the IGV results were observed using the final BAM file (pre Variant Calling, Figure 6) three hypotheses were formulated:

1st) The variant was a false positive from the WES and our analysis eliminated it;

2nd) Our quality filters in the Variant Filtration stage were too conservative and discarded this variant because one or more of the parameters was not being met;

3rd) One or more of the Variant Calling tools failed to call this variant. As only the variants that passed the combined analysis for all four Variant Calling tools were selected, if this was not the case, the variant was lost.

To test these hypotheses, all VCF files (between the final BAM files and the final prioritized VCF files) were analyzed with the intention of discovering at which point of the analysis the variant in question was discarded.

For c.1015G>A, c.582_583dupAC and c.-124G>A the first hypothesis was disregarded since there was a concordance between IGV and Sanger sequencing (that is still considered the gold standard for sequencing and for result validation). As for c.*162delT, this is most probably a false positive from the WES (1st hypothesis) since it was not called by any of the Variant Calling tools, hence the deletion observed in IGV is possibly an artifact.

For c.1015G>A, and c.-124G>A, the 3rd hypothesis was confirmed with the variant being eliminated in the Variant Calling SAMtools mpileup file. For a still undetermined reason this tool did not call these variants for the unaffected individual. If SAMtools mpileup had a high Ti/Tv ratio it could indicate that, in order to be conservative regarding false positives, it should be considered keeping the tool but since this tool: 1) had the lowest average ratio, 2) takes the longest time to run, demanding more processing power, 3) gives no indication of the time remaining and 4) is responsible for two false negatives in the bioinformatics analysis, it is advisable for this tool to be eliminated from the pipeline. In this case the combined analysis would consist of the remaining three tools (without SAMTools mpileup the average Ti/Tv ratio for the combined analysis decreases less than one centesimal point).

Regarding c.582_583dupAC, this variant was discarded for the unaffected individual in the annotated file. However, when recently testing the annotation, and using the same pre-annotation file as before, the variant was not eliminated. Since we annotated our files using the most recent VEP version (79) in online mode, and it was launched just a few days before we used it, it is possible that the removal of the variant from the unaffected individual was a bug that was, since then, fixed. This variant, as previously discussed, is also problematic being flagged by Ensembl (also VEP creators). This only happened for this variant in our study. It is, however, impossible at this moment with the available data to give a better answer to this problem, and we hope to be able to address this issue in the future.

4.5 – Conclusions and future work

In this study we selected ten variants for validation, nine from family 1 and one from family 2. Of those variants, eight were successfully validated by Sanger sequencing (c.10855G>C, c.1440G>C, c.*106G>A, c.*162delT, c.1067+3A>C, c.280T>C, g.34729259_34729260insACCAGC, and c.-124G>A).

Of the two non-validated variants, we were able to troubleshoot and identify the responsible tool within our pipeline for one of them (c.1015G>A), and formulated a hypothesis to explain the second one (c.582_583dupAC).

The seven most interesting variants and their respective genes (c.1440G>C and c.-124G>A in *STOX2*, c.10855G>C in *DYNC2H1*, c.*106G>A in *DCLK2*, c.*162delT in *TRAPPC11*, c.1067+3A>C in *TDO2*, and g.34729259_34729260insACCAGC in *CTCF*) may cause PSP in multiplex families. From these, taking together the information published until today and our results, we suggest that c.1440G>C and c.-124G>A (both from *STOX2* for family 1 and 2, respectively), c.10855G>C (low frequency variant from *DYNC2H1*), and c.1067+3A>C (*TDO2* novel variant) are the most interesting findings and warrant additional studies.

Of the initial group of prioritized variants (c.10855G>C, c.1440G>C, c.*106G>A, c.*162delT and c.1067+3A>C), we had a 100% validation success. Having this in consideration, we believe that our optimized bioinformatics pipeline was successful and can, with constant updating and the previously proposed modifications, be used in similar studies in the future.

In the future, multiple paths may be followed regarding our findings and to the familial genetics of PSP. Functional studies, such as *in silico* 3D modeling of the proteins with the alterations, luciferase reporter assays, mRNA expression assays, animal model studies, amongst others, could be carried out.

Replication studies are also mandatory, including a larger number of multiplex families. It is likely that, as a complex disorder, familial PSP is a genetically heterogeneous disease and as such a greater number of families will be mandatory.

Lastly, WGS should also be considered for future studies. The issues associated to WGS are quickly dissipating with the price becoming more and more competitive when compared to WES, new and better bioinformatics tools are being developed and the move from local sever-based analysis to cloud-based analysis is reducing the hardware issue of dealing with the massive data size. WGS is not only more informative (covering the entire genome) and differs from WES from a technical standpoint, eliminating some concerns regarding low InDel validation rates and amplification bias. Sequencing with long reads with 3rd generation sequencing (for instance, Pacific Biosciences single-molecule real time sequencing) is also a currently available option which is better for haplotype determination, larger InDel detection and for regions with high GC content, and which would have been very useful to resolve the apparent Mendelian inconsistency of variant c.-124G>A in family 2.

Overall, in the study described herein, we identified both novel and rare genetic variants that may contribute to increased PSP risk in two multiplex families, accomplishing our main goal and the respective aims initially proposed. We hope that with this work we were able to provide another piece of the PSP etiology puzzle, contributing to further understanding this pathology.

5. References

Abolnik I.Z., Lossos I.S., Zlotogora J., Brauer R. 1991. On the inheritance of primary spontaneous pneumothorax.AmJMedGenet,40:155-58

Adesina A.M., Vallyathan V., McQuillen E.N. *et al.* 1991. Bronchiolar inflammation and fibrosis associated with smoking: a morphologic cross-sectional population analysis. Am Rev Respir Dis, 143:144–149.

Adzhubei, I., Schmidt, S., Peshkin,L. *et al.* 2010. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–9.

Asimit, J. and Zeggini, E. 2011. Testing for rare variant associations in complex diseases. Genome Med. 1, 24.

Arts, H., and Knoers, N. 2013. Cranioectodermal Dysplasia. GeneReviews® [Internet].

Bainbridge M.N., Wang M., Burgess D.L., *et al.* 2010. Whole exome capture in solution with 3 Gbp of data. Genome Biol.

Bamshad, M., Ng, S., Bigwam, A. *et al.* 2011. Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. 12, 745–55.

Baronofsky ID, Warden HG, Kaufman JL, *et al.,* 1957. Bilateral therapy for unilateral spontaneous pneumothorax. J Thorac Surg. Sep;34(3):310–322

Baumann, M.H., Noppen, M., 2004. Pneumothorax. Respirology; 9: 157–164.

Bense L, Eklund G, Odont D, Wiman LG. 1987. Smoking and the increased risk of contracting spontaneous pneumothorax. Chest, 92:1009-12.

Bögershausen, N., Shahrzad, N., Chong, J.X., *et al.* 2013. Recessive TRAPPC11 mutations cause a disease spectrum of limb girdle muscular dystrophy and myopathy with movement disorder and intellectual disability. Am J Hum Genet. 2013 Jul 11; 93(1):181-90.

Bonnefont, J., Laforge, T., Plastre, O., *et al.* 2011. Primate-specific RFPL1 gene controls cell-cycle progression through cyclin B1/Cdc2 degradation. Cell death and differentiation. 18:293–303.

Bonnefont, J., Nikolaev, S.I., Perrier, A.L., *et al.* 2008. Evolutionary forces shape the human RFPL1,2,3 genes toward a role in neocortex development. American Journal of Human Genetics. 83:208–218.

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

Buetow K., Edmonson M., Cassidy A. 1999. Reliable identification of large numbers of candidate SNPs from public EST data. Nat. Genet 21: 323–325.

Butler, J. M. 2005. Forensic DNA Typing - Biology Technology and genetics of STR Markers.

Butt, A. J., Sergio, C. M., Inman, C. K., *et al.* 2008. The estrogen and c-Myc target gene HSPC111 is over-expressed in breast cancer and associated with poor patient outcome. Breast Cancer Res. 10: R28.

Carlson, C. S., Eberle, M. a, Kruglyak, L. and Nickerson, D. a. 2004. Mapping complex disease loci in whole-genome association studies. Nature 429, 446–52.

Cavaluzzi, M. J. and Borer, P. N. 2004. Revised UV extinction coefficients for nucleoside-5′-monophosphates and unpaired DNA and RNA. Nucleic Acids Research 32.

Clark, D. P., 2005. Molecular Biology. Elsevier Academic Press Publications. chap. 21, 567–598.

Court, F., Camprubi, C., Garcia, C.V., *et al.* 2014. The PEG13-DMR and brain-specific enhancers dictate imprinted expression within the 8q24 intellectual disability risk locus. Epigenetics Chromatin. 7:5.

Crona, J., Verdugo, A., Granberg, D., *et al.* 2013. Next-generation sequencing in the clinical genetic screening of patients with pheochromocytoma and paraganglioma. Endocr. Connect. 2, 104–111.

Dagoneau, N., Goulet, M., Genevieve, D., *et al.* 2009. DYNC2H1 mutations cause asphyxiating thoracic dystrophy and short rib-polydactyly syndrome, type III Am J Hum Genet, 84, pp. 706–711

Day-Williams, A. G. and Zeggini, E. 2011. The effect of next-generation sequencing technology on complex trait research. Eur. J. Clin. Invest. 41, 561–7.

DePristo, M., Banks, E., Poplin, R. *et al.* 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–8.

Dijkmans T. F., Van Hooijdonk L. W. A., Fitzsimons C. P., Vreugdenhil E. 2010. The doublecortin gene family and disorders of neuronal structure. Cent. Nerv. Syst. Agents Med. Chem. 10, 32–46.

Ebersberger I., Metzler D., Schwarz C., Pääbo S., 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. Am J Hum Genet 70: 1490–1497

Edwards S.L., Beesley J., French J.D., Dunning M. 2013. Beyond GWASs: Illuminating the dark road from association to function. Am J Hum Genet. 93(5):779-797.

Eilbeck K., Lewis S.E., Mungall C.J. *et al.,* 2005. The Sequence Ontology: A tool for the unification of genome annotations. Genome Biology 6:R44

Faber E. 1921. Spontaneous pneumothorax in 2 siblings. Hospitalstid; 64: 573–574.

Fang H-Y, Lin C-Y, Chow K-C, *et al.* 2010. Microarray detection of gene overexpression in primary spontaneous pneumothorax. Exp Lung Res.;36(6):323-30.

Fedulov, A.V., Leme, A., Yang, Z. *et al.* 2008. Pulmonary exposure to particles during pregnancy causes increased neonatal asthma susceptibility. Am J Respir Cell Mol Biol. 38:57–67

Fenstad, M.H., Johnson, M.P., Løset, M. *et al.* 2010. STOX2 but not STOX1 is differentially expressed in decidua from pre-eclamptic women: data from the Second Nord-Trondelag Health Study. Mol. Hum. Reprod. 16 960–96810.1093

Frohlich B. A., Zeitz C., Matyas G., 2008. Novel mutations in the folliculin gene associated with spontaneous pneumothorax. Eur. Respir. J.; 32:1316-1320.

Graham R.B., Nolasco M., Peterlin B., Garcia C.K.. 2005. Nonsense mutations in folliculin presenting as isolated familial spontaneous pneumothorax in adults. Am J Respir Crit Care Med, 172:39-44.

Gunji Y., Akiyoshi T., Sato T., 2007. Mutations of the Birt-Hogg-Dube gene in patients with multiple lung cysts and recurrent pneumothorax. (Letter) J Med Genet, 44:588-93.

Guo, Y., Li, J., Li, C. *et al.* 2012. The effect of strand bias in Illumina short-read sequencing data. BMC Genomics 13:666.

Gupta D., Hansell A., Nichols T. *et al.* 2000. Epidemiology of pneumothorax in England. Thorax, 55:666-71

Haines, J. and Pericak-Vance, M. A. 1998. Approaches to Gene Mapping in Complex Human Diseases.

Haynes D, Baumann MH. 2010. Management of pneumothorax. Semin Respir Crit Care Med;31:769-80.

Henry M, Arnold T and Harvey J. 2003. BTS guidelines for the management of spontaneous pneumothorax. Thorax , 58: ii39-ii52.

Hoischen, A., van Bon, B., Glissen, C., *et al.* 2010. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat. Genet. 42, 483–5.

Hoover EL, Hsu HK, Mkandawire K, Minnard E. 1989. Adhesions as a cause of persistent spontaneous pneumothorax. Journal of the National Medical Association.;81(4):464-466.

Itard J. 1803. Dissertation sur le pneumothorax ou les congestions gazeuses qui déforment dans la poitrine. Paris

Jia, P. Li, F., Xia, J. *et al.* 2012. Consensus rules in variant detection from next-generation sequencing data. PLoS One 7, e38470.

Kerjan, G., Koizumi, H., Han, E.B., *et al.* 2009. Mice lacking doublecortin and doublecortin-like kinase 2 display altered hippocampal neuronal maturation and spontaneous seizures. Proc Natl Acad Sci U S A. 106(16):6766-71

Kjaergaard H. 1932. Spontaneous pneumothorax in the apparently healthy. Acta Med Scand Suppl;43:1-159.

Köhler, S., Doelken, S., Mungall, C., *et al.* 2014. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucl. Acids Res. 42 (D1): D966-D974

Kumar, P., Henikoff, S. and Ng, P., 2009. Predicting the effects of coding nonsynonymous variants on protein function using the SIFT algorithm. Nat. Protoc. 4, 1073–81.

Laennec RT. 1819. De l'Auscultation Médiate ou Traité du Diagnostic des Maladies des Poumons et du Coeur. Brosson and Chaudé. Paris

Lasky-Su J., Neale B.M., Franke B., *et al.* 2008. Genome-wide association scan of quantitative traits for attention deficit hyperactivity disorder identified novel associations and confirms candidate gene associations. Am J Med Genet Part B Neuropsychiatr Genet.;147B(8):1345-1354.

Lee, S., Abecasis, G. R., Boehnke, M. and Lin, X. 2014. Rare-Variant Association Analysis: Study Designs and Statistical Tests. Am. J. Hum. Genet. 95, 5–23.

Li, H. and Durbin, R. 2009. Fast and accurate short read alignment with Burrows Wheeler transform. Bioinformatics 25, 1754–60.

Lim D.H., Rehal P.K., Nahorski M.S., *et al.* 2010. A new locus-specific database (LSDB) for mutations in the folliculin (FLCN) gene. Hum. Mutat. 31:E1043–E1051.

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

McInerney-Leo, A.M., Harris, J.E., Leo, P.J. *et al.,* 2014. Whole exome sequencing is an efficient, sensitive and specific method for determining the genetic cause of short-rib thoracic dystrophies. Clin Genet. 1399-0004.

Melton LJ 3rd, Hepper NG, Offord KP. 1979. Incidence of spontaneous pneumothorax in Olmsted County, Minnesota: 1950 to 1974. Am Rev Respir Dis. 120(6):1379–1382.

Miller, C. L., 2011. The kynurenine pathway: modeling the interaction between genes and nutrition in schizophrenia. J. Orthomol. Med. 26, 59–84

Mullaney, J. M., Mills, R. E., Pittard, W. S. and Devine, S. E. 2010. Small insertions and deletions (INDELs) in human genomes. Hum. Mol. Genet. 19, R131–6.

Ng, S. B., Buckingham, K., Lee, C., *et al.* 2010. Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet. 42, 30–5.

Ng, S. B., Turner, E., Roberstson, D. *et al.* 2009. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461, 272–6.

Noppen M, De Keukeleire T. 2008. Pneumothorax. Respiration;76:121-7

Noppen M and Schramel F. 2002. Pneumothorax. Eur Respir Mon; 22: 279–296.

O'Rawe J, Jiang T, Sun G, *et al.* 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med.

Ogino, S., Gulley, M. L., den Dunnen, J. T. and Wilson, R. B. 2007. Standard mutation nomenclature in molecular diagnostics: practical and educational challenges. J. Mol. Diagn. 9, 1–6.

Orth, K., Palmer, L. E., Bao, Z. Q., *et al.* 1999. Inhibition of the mitogen-activated protein kinase kinase superfamily by a Yersinia effector. Science 285: 1920-1923.

Painter J.N., Tapanainen H., Somer M. 2005. A 4-bp Deletion in the BirtHogg-Dubé Gene (FLCN) Causes Dominantly Inherited Spontaneous Pneumothorax. Am J Hum Genet, 76:522-27.

Pareek, C. S., Smoczynski, R. and Tretyn, A. 2011. Sequencing technologies and genome sequencing. J. Appl. Genet. 52, 413–35.

Primrose WR. 1984. Spontaneous pneumothorax: a retrospective review of aetiology, pathogenesis and management. Scott Med J.;29(1):15–20

Robbins R.A., Gossman G.L., Nelson K.J. *et al.* 1990. Inactivation of chemotactic factor inactivator by cigarette smoke: a potential mechanism of modulating neutrophil recruitment to the lung. Am Rev Respir Dis, 142:763–768.

Schmidts, M., Arts, H., Bongers, E., *et al.* 2013. Exome sequencing identifies DYNC2H1 mutations as a common cause of asphyxiating thoracic dystrophy (Jeune syndrome) without major polydactyly, renal or retinal involvement. J. Med. Genet. 50: 309-323.

Scrivens, P. J., Noueihed, B., Shahrzad, N., *et al.* 2011. C4orf41 and TTC-15 are mammalian TRAPP components with a role at an early stage in ER-to-Golgi trafficking. Molec. Biol. Cell 22: 2083-2093.

Seremetis MG. 1970. The Management Of Spontaneous Pneumothorax, Chest.;57(1):65-68.

Sharma A, Jindal P. 2008. Principles of diagnosis and management of traumatic pneumothorax. Journal of Emergencies, Trauma and Shock.;1(1):34-41. doi:10.4103/0974-2700.41789.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A. and Ponting, C. P. 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 15, 121–32.

Smith, K. R., Bromhead, C., Hildebrand, M. *et al.* 2011. Reducing the exome search space for mendelian diseases using genetic linkage analysis of exome genotypes. Genome Biol. 12, R85.

Staden, R. 1996. The Staden sequence analysis package. Mol Biotechnol 5, 233–41.

Sullivan, P. F., Neale, B. M., van den Oord, E., *et al*. 2004. Candidate genes for nicotine dependence via linkage, epistasis, and bioinformatics. Am. J. Med. Genet., 126B: 23–36.

Sun, Y., Ruivenkamp, CA, Hoffer, MJ *et al.* 2015. Next-Generation Diagnostics: Gene Panel, Exome, or Whole Genome? Hum. Mutat. 36, 648-55

Thoeringer C.K., Ripke S., Unschuld P.G., *et al.* 2009. TheGABA transporter 1 (SLC6A1): a novel candidate gene for anxiety disorders. J Neural Transm. 116(6):649-657.

Thomas, S., Thomas, M., Wincker, P. *et al.* 2008. Human neural crest cells display molecular and phenotypic hallmarks of stem cells. Hum Mol Genet.17:3411–3425

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

Tuy, F.P., Saillour Y., Kappeler C., et al. 2008. Alternative transcripts of Dclk1 and Dclk2 and their expression in doublecortin knockout mice. Dev Neurosci. 30(1-3):171-86.

Untergasser, A., Cutcutache, I., Koressaar, T., et al. 2012. Primer3--new capabilities and interfaces. Nucleic Acids Res. 40, e115.

Van Dijk, M., Mulders, J., Poutsma, A., *et al.* 2005. Maternal segregation of the Dutch preeclampsia locus at 10q22 with a new member of the winged helix gene family. Nat Genet. 37:514–519

Vanderschueren RG. 1981. Pleural talcage in patients with spontaneous pneumothorax. Poumon Coeur; 37:273-6.

Vasli, N. and Laporte, J. 2013. Impacts of massively parallel sequencing for genetic diagnosis of neuromuscular disorders. Acta Neuropathol. 125, 173–85.

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

Watson, L.A., Wang, X., Elbert, A., *et al.* 2014. Dual Effect of CTCF Loss on Neuroprogenitor Differentiation and Survival. The Journal of Neuroscience, 34(8): 2860-2870.

Weber-Lehmann, J., Schilling, E. *et al*. 2014. Finding the needle in the haystack: Differentiating “identical” twins in paternity testing and forensics by ultra-deep next generation sequencing. Forensic Science International: Genetics, Volume 9, 42 – 46

Whitfield, T.W., Wang, J., Collins, P.J., *et al*. 2012. Functional analysis of transcription factor binding sites in human promoters. Genome Biol 13, R50

Wu, J. and Jiang, R. 2013. Prediction of deleterious nonsynonymous single-nucleotide polymorphism for human diseases. Scientific World Journal. 2013, 675851.

Yi, M., Zhao, Y., Jia, L, *et al.* 2014. Performance comparison of SNP detection tools with illumina exome sequencing data-an assessment using both family pedigree information and sample-matched SNP array data. Nucleic Acids Res. 42, e101.

5.1 – Web links in order of appearance

http://www.pennmedicine.org/encyclopedia/em_PrintImage.aspx?gcid=19589&ptid=2

http://www.ensembl.org/info/genome/variation/predicted_data.html

https://bcgene.igc.gulbenkian.pt/bcos/index.html

http://ncifrederick.cancer.gov/atp/genetics-and-genomics/laboratory-of-molecular-technology/lmt-protocols-and-resources/sureselect/background/

https://github.com/chapmanb/bcbio-nextgen/tree/master/docs/contents

https://bcbio-nextgen.readthedocs.org/en/latest/

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/

http://bio-bwa.sourceforge.net/

http://picard.sourceforge.net/command-line overview.shtml

http://qualimap.bioinfo.cipf.es/

http://www.broadinstitute.org/gatk/events/3391/GATKw1310-BP-5-Variant_calling.pdf

https://wiki.gacrc.uga.edu/wiki/Freebayes

http://samtools.sourceforge.net/mpileup.shtml.

http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_variantrecalibration_VariantRecalibrator.html

http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_annotator_QualByDepth.html

http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_annotator_RMSMappingQuality.html

http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_annotator_Coverage.html

http://www.ncbi.nlm.nih.gov/SNP/

http://www.1000genomes.org/

https://genome.ucsc.edu/cgi-bin/hgTables

http://sift.jcvi.org/

http://genetics.bwh.harvard.edu/pph2/

http://www.sanger.ac.uk/resources/software/exomiser/

http://bioinfo.ut.ee/primer3-0.4.0/

http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=OGP__9606__9558&LINK_LOC=blasthome

https://genome.ucsc.edu/cgi-bin/hgBlat?command=start

http://genome.ucsc.edu/cgi-bin/hgPcr?command=start

http://www.cegat.de/en/diagnostics/pedigree-chart-designer/

https://www.broadinstitute.org/gatk/guide/bp_step.php?p=1

https://www.broadinstitute.org/gatk/img/cartoon-blackbox-workflow-web-blackblue.png

http://cdn.vanillaforums.com/gatk.vanillaforums.com/FileUpload/eb/44f317f8850ba74b64ba47b02d1bae.png

APPENDICES

Appendix A

Individuals with at least one of these items were excluded as patients from the PSP project:

- Cardiopulmonary disorder (inflammatory or malignant);
- Pregnant and breastfeeding women;
- Individual unable to give informed consent;
- Family history of known related disorders:
  - ✓ Alpha1-antitrypsin deficiency;
  - ✓ Marfan syndrome ;
  - ✓ Ehlers-Danlos syndrome;
  - ✓ Histiocytosis X;
  - ✓ Cystic fibrosis:
  - ✓ Birt-Hogg-Dubé syndrome;
  - ✓ Lymphangioleiomyomatosis;
  - ✓ Sarcoidosis;
  - ✓ Tuberculosis;
  - ✓ Chronic obstructive lung disease.

Appendix B

CENTRO HOSPITALAR LISBOA NORTE, EPE

HOSPITAL DE SANTA MARIA

Hospital Pulido Valente

Presidente
Prof. Doutor João Lobo Antunes
Vice-Presidente
Profª. Doutora Maria Luísa Figueira
Membros
Dr. Mário Miguel Rosa
Prof. Doutor Carlos Calhaz Jorge
Dra. Leonor Carvalho
Dra. Graça Nogueira
Dra. Gabriela Martins Mendes
Mestre Enfª. Isabel Côrte-Real
Padre Fernando Sampaio

Exmo. Senhor
Dr. Salvato Feijó
Serviço de Pneumologia I
Centro Hospitalar Lisboa Norte, E.P.E.

Lisboa, 19 de Setembro de 2008

Assunto: Estudo multicêntrico intitulado *"Estudo Genético de Pneumotórax Espontâneo Primário"*

Pela presente, informamos que o Projecto citado em epígrafe, obteve na reunião realizada em 17 de Setembro de 2008, parecer favorável da Comissão de Ética.

Mais se informa que o referido Projecto foi enviado ao Director Clínico, Dr. Correia da Cunha, a fim de obter a autorização final para a sua realização.

Com os melhores cumprimentos,

O Presidente da Comissão de Ética para a Saúde

Prof. Doutor João Lobo Antunes

Figure B1 - HSM's ethical committee approval for the PSP project.

APPENDIX C

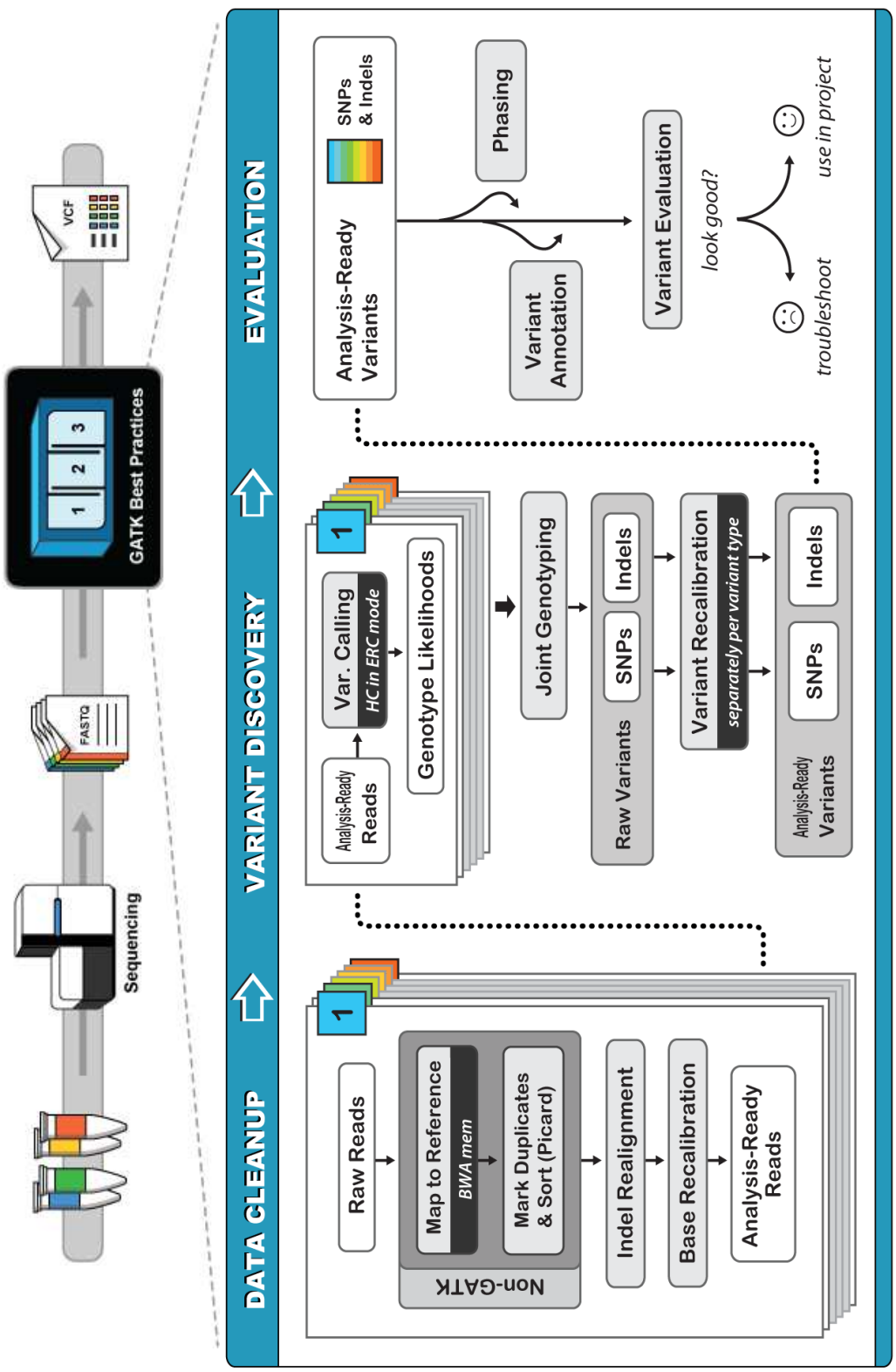

Figure C1 - Genome Analysis Toolkit (GATK) Best Practices workflow from the Broad Institute for human exome sequencing analysis. This image represents the general steps used to perform NGS (https://www.broadinstitute.org/gatk/img/cartoon-blackbox-workflow-web-blackblue.png) as well as the workflow (http://cdn.vanillaforums.com/gatk.vanillaforums.com/FileUpload/eb/44f317f8850ba74b64ba47b02d1bae.png) used as the basis of our bioinformatics optimized pipeline (Figure 6).

Appendix D

This appendix contains the hardware, software and command lines information pertaining to this thesis. To visualize this appendix choose one of the following hypotheses:

1) Go to the web link http://tiny.cc/THESIS_PSP_APPENDIX_D in your chosen browser;

2) Scan the following QR code:

3) Open the file “Appendix D” in the CD/DVD.

Appendix E

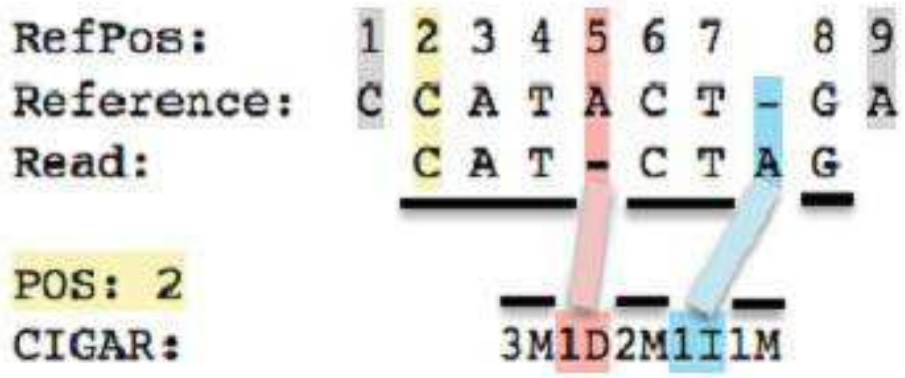

Figure E1 - CIGAR string example. The POS indicates that the read aligns starting at position 2 in the reference. The CIGAR string indicates that the first three bases in the read sequence align with the reference. The next base in the reference does not exist in the read (deletion), followed by two bases that align with the reference. The next base in the read does not exist in the reference (insertion) and afterwards we have one more base that aligns with the reference.

APPENDIX F

Table F1 – Primers used for the validation step of the ten selected variants and respective information.

Variant	Family	Orientation	Primer sequence (5' --> 3')	Primer lenght (bp)	PCR product (bp)	Tm (ºC)	GC content (%)
c.10855G>C	1	FWD	CCTAAATGGTTCTCAGTCCCTTA	23	392	58,70	43,48
		RVS	ACTGGGGGAACTTACATCCT	20		57,43	50,00
c.1440G>C	1	FWD	AACTCTCGGGTGGTTTTCCT	20	474	59,95	50,00
		RVS	ACAGCAACACAAACGTGCTC	20		59,97	50,00
c.*106G>A	1	FWD	TCTCTCCAGTTCCTCCCTCA	20	488	59,91	55,00
		RVS	TGCATCGAATGCTGGAATAA	20		60,18	40,00
c.*162delT	1	FWD	TCCCACTTTAAAGCCACAGG	20	415	60,10	50,00
		RVS	TGGTGCATCTGCTCTCTTTG	20		60,14	50,00
c.1067+3A>C	1	FWD	TTGCCTAGATAACCATGTGTGC	22	383	60,02	45,45
		RVS	CTGCCTGGATGTAAGCACTG	20		59,47	55,00
c.1015G>A	1	FWD	ACTGAGCCTGGAAGGATGAATA	22	491	60,10	45,45
		RVS	CAGTAGCTTATGGTGTGGGTGA	22		60,05	50,00
c.280T>C	1	FWD	TGACAAAGCTGGGACACAATAC	22	418	60,04	45,45
		RVS	ATACAGATCCCTCACCTTGGAA	22		59,83	45,45
c.582_583dupAC	1	FWD	TGGCTGAGAGCAGAAAAATACA	22	426	60,02	40,91
		RVS	CCCAAAACAGATTCGGAGTAAG	22		59,99	45,45
g.34729259_34729260insACCAGC	1	FWD	GAATTCTCCCACCATGATTTGT	22	495	60,06	40,91
		RVS	GCAATTCCAGAAAGTACCCAAG	22		60,00	45,45
c.-124G>A	2	FWD	CATTGTGGGACGTGTGTAAAAT	22	497	59,65	40,91
		RVS	ACTTATCGCATCCATCACTGC	21		60,12	47,62

Abbreviations: FWD – Forward strand; RVS – Reverse Strand; Tm – melting temperature.

Appendix G

Table G1 – PCR conditions used to amplify each variant of interest.

Variant	Family	Used kit	Stepdown	Nº cycles	Temperature (ºC) - Running time (s)	Reagents per individual (µl)
c.10855G>C	1	NZYTaq Colourless	No	1	95 – 120	DNA – 4 Kit – 12,5 Primer FWD – 1 Primer RVS – 1 H_2O – 6,5
				30	95 – 45 60 – 45 72 – 60	
				1	72 – 360	
c.1440G>C	1	KAPA RM	No	1	95 – 180	DNA – 2,5 Kit – 6,25 Primer FWD – 0,75 Primer RVS – 0,75 H_2O – 14,75
				30	98 – 20 60 – 15 72 – 60	
				1	72 – 180	
c.*106G>A	1	NZYTaq Colourless	Yes	1	95 - 120	DNA – 4 Kit – 12,5 Primer FWD – 1 Primer RVS – 1 H_2O – 6,5
				15	95 - 30	
					[70-55]* - 40	
					72 - 60	
				30	95 - 20	
					60 – 45	
					72 – 60	
				1	72 - 360	
c.*162delT	1	KAPA RM	No	1	95 – 180	DNA – 5 Kit – 12,5 Primer FWD – 1,5 Primer RVS – 1,5 H_2O – 4,5
				30	98 – 20 60 – 15 72 – 60	
				1	72 – 180	
c.1067+3A>C	1	KAPA RM	No	1	95 – 180	DNA – 5 Kit – 12,5 Primer FWD – 1,5 Primer RVS – 1,5 H_2O – 4,5
				30	98 – 20 60 – 15 72 – 60	
				1	72 – 180	
c.1015G>A	1	NZYTaq Colourless	Yes	1	95 – 120	DNA – 4 Kit – 12,5 Primer FWD – 1 Primer RVS – 1 H_2O – 6,5
				15	95 – 30	
					[70-55]* - 40	
					72 – 60	
				30	95 – 20	
					60 – 45	
					72 – 60	
				1	72 - 360	
c.280T>C	1	NZYTaq Colourless	No	1	95 – 120	DNA – 4 Kit – 12,5 Primer FWD – 1 Primer RVS – 1 H_2O – 6,5
				30	95 – 45 60 – 45 72 – 60	
				1	72 – 360	
c.582_583dupAC	1	NZYTaq Colourless	No	1	95 – 120	DNA – 4 Kit – 12,5 Primer FWD – 1 Primer RVS – 1 H_2O – 6,5
				30	95 – 45 60 – 45 72 – 60	
				1	72 – 360	
g.34729259_34729260insACCAGC	1	NZYTaq Colourless	No	1	95 – 120	DNA – 4 µl Kit – 12,5 µl Primer FWD – 1 Primer RVS – 1 H_2O – 6,5
				40	95 – 45 60 – 45 72 – 60	
				1	72 – 360	

Variant	Family	Used kit	Stepdown	Nº cycles	Temperature (ºC) - Running time (s)	Reagents per individual (µl)
c.-124G>A	2	NZYTaq Colourless	Yes	1	95 - 120	DNA – 5 Kit – 12,5 Primer FWD – 1 Primer RVS – 1 H2O – 5,5
				15	95 – 30	
					[70-55]* – 40	
					72 – 60	
				45	95 – 20	
					62 - 45	
					72 – 60	
				1	72 - 360	

Abbreviations: FWD – Forward; RVS – Reverse.

* The first cycle is run at the highest temperature and each successive cycle is run one degree below the previous one (*stepdown*), for the total number of cycles.

APPENDIX H

In order to calculate the Fisher Strand Bias (FSB) one has:

1) To select a variant;

2) To select an individual;

3) To observe the total number of reads (depth), the reads with the reference allele (REF), the reads with the alternate allele (ALT). Afterwards, one has to verify which ones are in the positive strand (PLUS) and which ones are in the negative strand (MINUS);

4) To construct a 2x2 contingency table per variant and per individual. For example, for c.10855G>C and II.3-110830:

c.10855G>C	**PLUS**	**MINUS**	***Sum***
REF	6	6	12
ALT	10	7	17
Sum	16	13	29 (*Grand* Total)

5) To calculate the Fisher's probability using the following formula:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ \ b!\ \ c!\ \ d!\ \ n!}$$

, where a = intersection between REF and PLUS, b = intersection between ALT and PLUS, c = intersection between REF and MINUS, d = intersection between ALT and MINUS and, n = *Grand* Total.

6) To calculate contingency tables (tables with the same sums) and their respective Fisher's probabilities and select the ones with lower Fisher's probability than the original;

7) To sum up the Fisher probabilities of all the selected tables to obtain the Fisher's Exact Test value (FET);

8) To use FET to calculate the FSB values using the following formula:

$$FSB = -10\log_{10}[FET]$$