



Università degli Studi di Padova

DIPARTIMENTO DI BIOLOGIA
Corso di Laurea Magistrale in Biologia Sanitaria

**Hereditary non syndromic deafness: development of a methodology
for the analysis of genes with high homology and validation of
splicing variants with hybrid minigenes**

Laureando:
Emanuele Savino

Relatore:
Prof. Giovanni Vazza
Dipartimento di Biologia

Correlatore:
Dott. Matteo Cassina
**Dipartimento di Salute
della Donna e del Bambino**

to my parents

Abstract

Hearing loss is a complex and heterogeneous disease with more than 100 genes involved. NGS-based approaches are thus indicated, but they fail to discriminate between high homologous (>99%) genomic regions because reads are too short. In the first part of this study a new method for the accurate analysis and sequencing of high homology hear-related genes is presented. 3'-phosphorothioate primers are used to extend the principle of allele specific PCR to long amplicons. The template specificity is eventually checked through sequencing of regions with non-homologous bases. Moreover, a simple statistical test for the discovery of CNVs (particularly deletions) in the stereocilin (STRC) gene based on NGS base coverages is discussed and results were compared with SureCall (Agilent) pair analysis software and trio analysis. In the second part, five SNVs, potentially involved in pathological splicing alterations of hear-related genes, are analysed through hybrid minigene assay using HEK293 cells. The *COL1A1* (NM_000088.3: c.1515G>A) mutation is dominant negative and causes skipping of exon 22, leading to a lethal form of osteogenesis imperfecta. The other tested variants are likely benign and do not alter splicing.

Acknowledgments

Writing a M.D. thesis requires a lot of energy and time, and I am particularly grateful to my supervisors Prof. Giovanni Vazza and Dr. Matteo Cassina that showed a lot of patience. I particularly want to acknowledge Dott. Cassina for his constant trust in me and in my capacities and for his unconditioned help. I would also like to thank all the colleagues in the laboratory who welcomed me warmly since the first day as "Y-chromosome", "Bagigio", "Biological engineer" and "Serial killer". I appreciated the numerous friendly moments I had there with them. I am also indebted to Dr. Annica Barizza and M^o Gabriele Nuzzi who reviewed this thesis. I appreciate the time and thought that went into their comments.

I would also like to acknowledge my family which supported me during these years and made a lot of sacrifices out of love to allow my studies here at the University of Padua. I am deeply thankful for these unforgettable 5 years spent at the *Gregorianum* college during which I learnt to be a man and took some precious opportunities. My university career would not have been the same without this second family.

Contents

1	Introduction	1
1.	Hearing loss	1
1.1	Hearing loss is a complex and heterogeneous disease . . .	1
1.2	Genetic and environmental causes of hearing loss	2
1.3	Hearing loss diagnostics during the NGS era	3
1.4	Types of pseudogenes	4
1.5	Hearing physiology	5
1.6	Molecular biology of stereocilin and otoancorin	7
1.7	Implications of <i>STRC</i> and <i>OTOA</i> genetic structure in diagnostic tests	8
1.8	In some cases diagnostic tests need further studies	9
2.	Splicing	10
2.1	Core signals	10
2.2	The splicing reaction	11
2.3	Spliceosome assembly	11
2.4	Cis-acting sequences	12
2.5	Exonic regulatory elements	14
2.6	Intronic regulatory elements	15
2.7	Alternative splicing	15
2.8	Point mutations that affect splicing	15
2.9	RNA analysis: PROs and CONs	18
2.10	Hybrid Minigenes	19
2	Aim of the study	21
3	Materials and methods	23
1.	<i>STRC/pSTRC</i> and <i>OTOA/pOTOA</i> variants sequencing	23
1.1	Patient samples and suspected mutations	23
1.2	Experimental design	23
1.3	PCR primer design	24
1.4	Long range PCR	24
1.5	Nested PCR or direct PCR	27
1.6	Agarose gel	27
1.7	Purification and Sanger sequencing	27
2.	CNVs assessment	27
2.1	Statistical model for CNV detection	27
2.2	SureCall pair analysis software	28
3.	Hybrid minigene assays	28

3.1	DNA samples for hybrid minigene constructs	28
3.2	Bioinformatic analysis	29
3.3	PCR primer design and cycling conditions	29
3.4	Agarose gel and purification	30
3.5	pcDNA3.1 hygrob-globin vector	30
3.6	Plasmid miniprep	31
3.7	Digestion	31
3.8	Purification and quantification	32
3.9	Ligation	32
3.10	Competent bacteria transformation	33
3.11	Screening of colonies	33
3.12	Sequencing of positive colony PCRs	33
3.13	Miniprep	34
3.14	<i>COL1A1</i> hybrid minigene mutagenesis	34
3.15	HEK293 cell transfection	34
3.16	Total RNA extraction	35
3.17	RNA retrotranscription to cDNA	36
3.18	Selective PCR and agarose gel	36
3.19	Bands extraction and Sanger sequencing	36
4	Results	39
1.	<i>STRC/pSTRC</i> and <i>OTOA/pOTOA</i> variants sequencing	39
1.1	Mandelker's approach	39
1.2	Confounding variability	39
1.3	First and second IPCR primers redesign	42
1.4	Searching for a rationale	42
1.5	Tackling homology: third and final IPCR primer redesign	43
1.6	IPCR time and cost optimization	45
1.7	Validating NGS variants: before and after	46
2.	CNVs assessment	53
2.1	A statistical model for CNV detection	53
2.2	SureCall pair analysis	53
2.3	Trios analysis	53
3.	Hybrid minigene assays	56
3.1	Bioinformatic analyses	56
3.2	Hybrid minigene constructs	56
3.3	<i>COL1A1</i> hybrid minigene mutagenesis	58
3.4	cDNA analysis	58
3.5	<i>COL1A1</i> NM_000088.3: c.1515G>A p.(=)	58
3.6	<i>COL2A1</i> NM_001844.5: c.1734+3A>G	59
3.7	<i>COL11A2</i> NM_080680.2: c.1819-5T>C	60
3.8	<i>MYO15A</i> NM_016239.3: c.4779+9G>A	60
3.9	<i>OTOG</i> NM_001277269.1: c.7926C>T p.(=)	62
5	Discussion	65
1.	<i>STRC/pSTRC</i> and <i>OTOA/pOTOA</i> variants sequencing	65
2.	Hybrid minigene assays	69

References

73

Chapter 1

Introduction

1 Hearing loss

The sense of hearing had developed in vertebrates hundreds of millions of years ago, well before they acquired the ability to make sounds and exploited it to communicate. In fact, hearing was a natural complement to sight as it allowed to predict incoming hazards even when they were not in the visible range and still be able to reconstruct their positions in space. This ancient origin could explain why it seems that roughly 1% of human genes are needed for development of the hearing apparatus. Indeed, hearing is probably our most important social sense because suicide rates are higher among deaf people than among those who have lost their sight [1].

Furthermore, as it is commonly seen for important organs and functions, maturation of this apparatus is not completed at birth (although the foetus can already respond to sound at 25-28 weeks) but indeed continues, particularly for external and middle ear, for up to 2-3 years, thus having substantial effects on how sounds are absorbed, processed, filtered and transmitted to the auditory system after birth [2]. From this evolutionary and developmental complexity, one might esteem the very heterogeneous nature of human hearing loss (with a worldwide incidence of 1 per 1,000 new-born infants) which truly has still to be completely uncovered.

1.1 Hearing loss is a complex and heterogeneous disease

As a proof of its heterogeneity, hearing loss (HL) can be categorized by site of lesion, age of onset, mode of inheritance, presence or absence of progression, severity of loss, frequencies involved, the configuration of the audiogram and presence or absence of vestibular involvement [3]. We now focus only on the major categories. According to the site of the lesion one can distinguish HL as conductive, in which the outer or middle ear is affected, sensorineural, when the inner ear, auditory nerve or central auditory pathway is affected or mixed if it is both conductive and sensorineural. Common causes of conductive HL can range from an ear canal plugged with earwax, to fluid in the middle ear from an infection, to diseases or trauma that impede vibration of the malleus, incus or stapes in which cases reconstructive microsurgical techniques may

be resolvable. On the other hand, sensorineural hearing loss can be further subdivided into sensory hearing loss (when the hair cells are affected), central hearing loss (when the cause is located along the central auditory pathway) or auditory neuropathy spectrum disorder (ANSO). Currently the primary treatment for sensorineural HL is the use of hearing aids, but amazing results have been obtained with cochlear implants attached to tiny computers.

As regards the age of onset of hearing loss, two major distinctions are prelingual and postlingual HL which are made taking into account a delay or not in child's development of speech. The former can be further subdivided into congenital and early childhood forms while the latter comprises late-onset (which probably includes also Beethoven's deafness) and age-related HL. Both prelingual and postlingual HL include acquired forms which could be mainly due to noise or head trauma, but also drugs (particularly aminoglycosides and cyclophosphamides) and long-term otitis media [4].

1.2 Genetic and environmental causes of hearing loss

Furthermore, there are multiple genetic and environmental causes of HL with genetic factors accounting for more than 50% of all congenital forms, even though more than 95% of those with congenital HL are born to hearing parents [5]. Physiologically, hearing loss is classified to reflect the presence (syndromic hearing loss) or absence (non-syndromic or isolated hearing loss) of coexisting physical or laboratory findings.

Syndromic HL accounts for 30% of genetic cases with common physical findings including pre-auricular pits and tags, branchial cysts or fistulae or dystopia canthorum (the lateral displacement of the inner corners of the eyes, giving the appearance of a widened nasal bridge), heterochromia iridis and pigmentary abnormalities, though it is also possible an association with renal, cardiac, neurological/neuromuscular, endocrine, metabolic and dental disorders. Currently more than 400 of these syndromes have been described but Pendred, Usher, Waardenburg (which is also a classical example of variable phenotypic expression) and branchio-oto-renal syndromes are undoubtedly the most frequent. Fortunately for several of these the associated genes are known and genetic testing is available.

Isolated HL is indeed extremely heterogeneous with 118 genes identified to date (<https://hereditaryhearingloss.org/>). Of the 70% of isolated HL cases with a genetic origin, 15-24% are inherited in a dominant fashion, 75-85% are recessive and 1-2% are X-linked or have other (primarily mitochondrial) modes of inheritance [3, 5]. For each mode of inheritance, the associated genes or loci are classified with the mark DFN plus a group-specific letter (A for dominant, B for recessive, X for X-linked) and a number, indicating the chronological order of discovery. Except for *GJB2* which encodes for the gap junction β 2 protein and accounts for more than half of all recessive forms, the other 117 genes may very well have little to rare epidemiological significance, further contributing to the difficulty of a clear HL diagnosis.

In addition to genetic heterogeneity, there is also considerable variation in expression, with some non-syndromic genes producing dominant, recessive

and even some syndromic phenotypes. These variations have generally been attributed to the location and type of mutation within the gene. Fortunately, recent advances in sequencing techniques have facilitated identification of new genes to an extent that previous approaches (positional cloning, homozygosity mapping, cDNA libraries from cochlear tissues) have been mostly outpaced.

Among the environmental causes of congenital HL, CMV infection is truly the single most common non-genetic cause of nonsyndromic sensorineural hearing loss, which can be unilateral or bilateral and is often progressive. Although the virus is shed in bodily fluids, such as urine, saliva and blood, and exposure to it is most frequently encountered through both sexual contact or contact with bodily fluids, the risk of infection and congenital development of HL is largely dependent on socioeconomic status, the availability of prevention strategies or hygienic measures. Apart from CMV, other environmental causes of congenital HL can be toxoplasmosis and Rubella virus infection.

1.3 Hearing loss diagnostics during the NGS era

The extreme genetic heterogeneity of hearing loss has made gene-by-gene approaches time-consuming and really expensive, thus inapplicable. Fortunately, in the last fifteen years a revolution has taken place in the field of genomics thanks to the development of numerous platforms for NGS which have increased enormously the speed and throughput of sequencing data generation. Although those systems are slightly different from each other, it is beyond the scope of this thesis to cover in detail functioning, pros and drawbacks of each one. Instead, it might suffice to say that they all share this workflow: genomic DNA (gDNA) library preparation, clonal amplification of fragments, sequencing and bioinformatic analysis. As a unique example we focus on Illumina's Genome Analyzer not only because it is one of the most common platforms but also because all NGS analyses done in my laboratory prior to this study were based on this instrument.

First of all, it is necessary to fragmentate gDNA either by enzymatic cleavage or mechanical shearing as this allows to consider each region (i.e. exons) as independent. Then, depending on whether or not it is a whole genome sequencing project, further targeted genomic enrichment might be performed in order to selectively isolate only genomic regions of interest before NGS. Currently, there are both solid-phase and solution-based targeted enrichment approaches which give similar results provided there is enough DNA template available [6]. Following fragmentation, attachment of 5' and 3' specific adaptors to fragments allows amplification and barcoding for pooling of samples. Denaturation of such DNA fragments is required to permit hybridization on *flow-cell* lane where oligonucleotide probes complementary to adaptors are attached and all following reactions are performed.

Clonal amplification can now take place through a bridge PCR reaction leading to the formation of DNA microclusters of approximately 1,000 identical copies each. Once amplification is finished, first a denaturation step and then sequencing reaction occur. The sequencing principle used by Illumina and other companies is called *sequencing by synthesis* and it basically reproduces a

Sanger sequencing with fluorescently labelled modified dNTPs which instead of acting as being irreversible chain terminator are indeed reversible. A laser beam causes fluorophore emission at the end of each cycle and a software stores the information. In this way it is possible to decipher the sequence nucleotide by nucleotide as the sequencing reaction goes on.

More specifically, the custom platform developed in my laboratory to assess HL genetic factors is based on Agilent Haloplex technology. It takes advantage of solution-based targeted enrichment followed by NGS to provide, with a unique test, all currently relevant information on genes. Though, of course, its applicability is restricted to diagnostic purposes, it offers the promise of a paradigm shift that will make genetic testing an early and integral part of deaf patient management, thereby precluding other more expensive and invasive tests. Furthermore, pinpointing the genetic cause of hearing loss offers the possibility of personalized medical management of hearing impairment as novel therapies are developed to prevent the progression or remediate the loss of hearing.

Even though NGS is a really powerful tool to harness the complexity of hearing loss, the study of associated genetic loci is sometimes complicated by the presence, among other factors, of segmental duplications which can give rise to the formation of pseudogenes. This is the case, in particular, of *STRC* and *OTOA*, two genes involved in nonsyndromic HL presenting a single pseudogene.

1.4 Types of pseudogenes

Pseudogenes are nonfunctional copies of genes due to the presence of inactivating mutations in the promotor or coding sequence [7]. Although it is simpler to identify pseudogenes for protein coding genes, they can also exist for RNA genes and mitochondrial ones as well (though in the nuclear genome). In the first case there can be two types of pseudogenes: processed and unprocessed. The former ones derive from inverse transcription of mature mRNA into cDNA and later integration into genome, while the latter are originated mostly by tandem duplication. Usually duplication leads to promotor and upstream regulatory sequences to be included into the pseudogene along with exons and introns. Even though processed and unprocessed pseudogenes are not necessarily copies of the entire gene, they are distinct from genetic fragments by the fact of having more than a single exon.

In the human genome there are approximately 15,000 pseudogenes and this is confirmed by the tendency of eukaryotes of having more pseudogenes than prokaryotes (which also possess compact genomes). However, as one might suppose, not all genes have a copy and the more one gene is transcribed the higher the number of both processed and unprocessed pseudogenes it possesses. This is the case of the ribosome proteins' family which despite including 95 functional genes it also has more than 2,000 processed pseudogenes. A reason to explain this can be the fact that highly transcribed genes show also a high degree of chromatin accessibility thus increasing the chance of duplication and recombination events in that region.

Moreover, as a positive selective pressure is often conserved only for the original gene copy, the new one is free to accumulate mutations that not only rapidly inactivate it (forming a proper pseudogene) but also could later on give rise to new functions or the rescuing of previous functions or subfunctions. As such, this is currently regarded as one of the major evolutionary advantages of eukaryotes in having pseudogenes since they can be the silent workbench where hundreds of mutations are continuously tested without hazard until sudden new useful products come up. When this happens, the mutation rate decreases accordingly protecting the new gene from deleterious mutations.

1.5 Hearing physiology

Before we discuss about *STRC* (stereocilin) and *OTOA* (otoancorin), the main genes involved in this study, in order to better appreciate their functions and how their mutations can eventually impair hearing, it is the case to briefly recap the basic anatomy and physiology of the ear.

The ear is structurally and functionally subdivided into three sections (see fig. 1.1): external, middle and internal ear. The external ear includes the pinna (or outer ear) and the ear canal, both of which basically serve to gather sound waves from the environment and to amplify them. The ear canal is sealed at its proximal end by a thin membranous sheet of tissue called the tympanic membrane, which separates the external ear from the middle one. Here it comes the first of multiple energy transduction processes that make hearing possible and that is from sound waves to mechanical vibrations. In fact, the tympanic membrane is physically connected to a chain of ossicles (indeed the smallest ones we have), namely the malleus, incus and stapes, which, to a great extent, act as a unique rigid impedance-matching body together with the tympanic membrane. This impedance-matching system is required because the second energy transduction process is actually from mechanical vibrations to fluid waves within the cochlea and it is realized through stapes-oval window coupling at the edge of middle and internal ear. In other words, without this special device, most sounds reaching the ear would simply be reflected as are voices from shore when swimming underwater.

The cochlea contains sensory receptors for hearing and on external view is like a snail shell coiled tube but actually it is composed of three parallel, fluid-filled channels: the scala vestibuli, the scala media and the scala tympani. While the scalae vestibuli and tympani are continuous with each other at the tip of the cochlea, the scala media is obviously in the middle between them and contains the organ of Corti, the neural apparatus responsible for the third energy transduction process. It lies on the basilar membrane (separating the scala media from the scala tympani) and is partially covered by the acellular tectorial membrane (TM) where two sets of hair cells are aligned with. The first set is composed of one row of inner hair cells (IHC) while the second one has three graded height rows of outer hair cells (OHC).

As the waves travel through the cochlea, they displace basilar and tectorial membranes creating upward and downward oscillations that bend hair cells, eventually resulting in altered membrane potentials, neurotransmitter release

1. HEARING LOSS

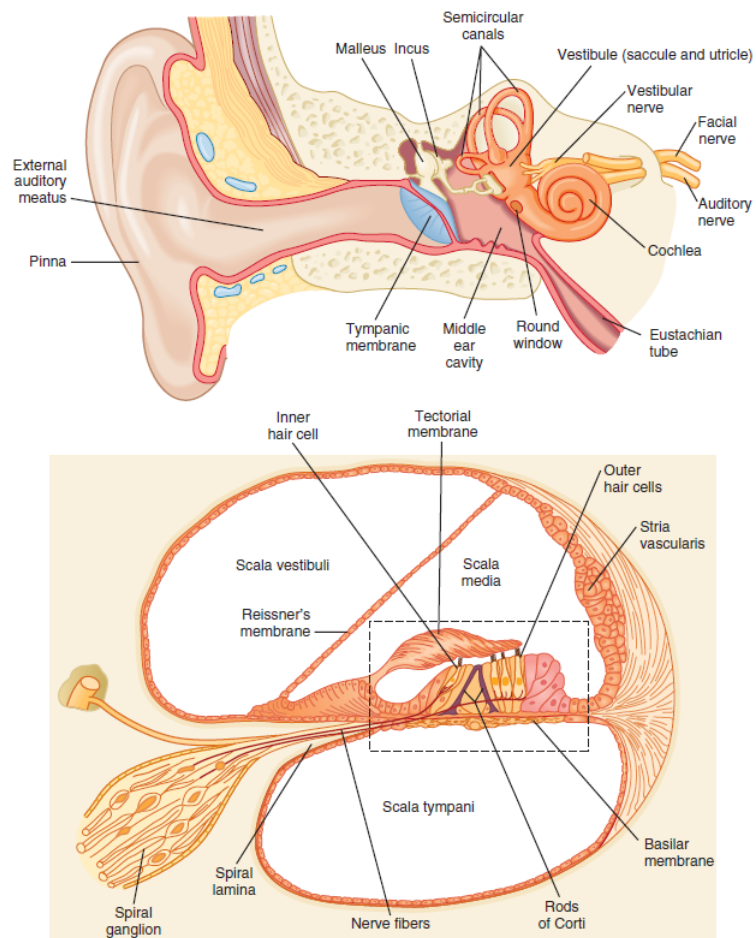


Figure 1.1: The upper image shows the anatomy of the ear. The lower image is a cross section of the cochlea where a dotted rectangle represents the organ of Corti.

and sensory neurons' action potentials to the brain. However, not all hair cells are bent equally during the process because the width and tension along the basilar membrane vary with distance from the base. Therefore, there is a distinct site of maximum displacement of the basilar membrane for any given sound frequency, even though the membrane moves as a whole in travelling waves. This results in a tonotopic map of the cochlea with the 100 μm -wide base vibrating to higher frequencies (up to 20,000 Hz) than the 500 μm -wide apex. Interestingly, this entire phenomenon, which takes only milliseconds and occurs in our ears every time we hear a sound, is mathematically equivalent to performing a Fourier transform of a harmonic sound wave, which is really a perfect example of how a physical property of nature shapes the evolution of biological structure.

1.6 Molecular biology of stereocilin and otoancorin

The key to understand the physiological and pathological implications of STRC and OTOA relies on the ultrastructure of outer hair cells' stereocilia bundles and the gelatinous tectorial membrane. This membrane is mainly formed by collagenous proteins (type II, IX, XI) and noncollagenous proteins (TECTA and TECTB among others). Furthermore, attachment of this membrane to the organ of Corti requires at least two proteins [8, 9] otoancorin, a GPI-anchored glycoprotein that mediates adhesion of the TM to the apical surface of spiral limbus (basically a region above the spiral lamina where nerve fibres leave the cochlea, see fig. 1.1) and stereocilin that forms top connectors at the tips of the tallest stereocilia (in the so called attachment crowns) in the hair bundles of the OHCs with the imprints of the membrane. Attachment of the TM to the OHCs may also involve Np55, a splice variant of neuroplastin that is only expressed in OHCs. Interestingly, these two proteins show sequence similarity and although the ligands for OTOA and STRC in the TM are unknown, there is evidence from knock-out mice that TECTA is a potential binding partner at least for OTOA.

Moreover, immature attachment crowns are found at the tips of all three rows of stereocilia during early postnatal development, but when STRC first appears in the stereocilia at P5, it localizes uniquely to the distal tips of the kinocilia or the tallest stereocilia. In fact, while the IHC stereocilia bundle is freestanding and is stimulated mostly by the motion of endolymphatic fluid (the fluid in scala media), only the tallest row of stereocilia in OHC is actually embedded into the tectorial membrane. This could explain why IHCs are genuine sensory cells that transmit information via the cochlear nerve fibres to the brainstem auditory nuclei, while in contrast, OHCs, which are endowed with electromotility, constitute the cochlear amplifiers that contribute to the detection of weak sound-induced vibrations [10].

Apart from attachment crowns, OHCs' stereocilia bundles present a variety of other inter-stereociliary protein connections (see fig. 1.2): tip links, top connectors, shaft connectors and ankle links. Without now giving too much detail about each of them, it may suffice to say that tip links are directly attached to ion channels responsible for the mechano-electrical transduction process, while

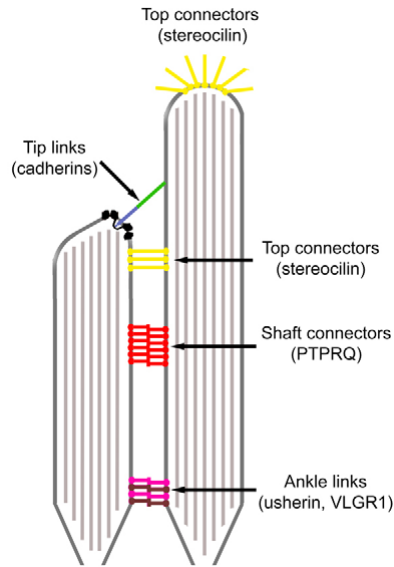


Figure 1.2: A schematic representation of all hair cell connectors. Top connectors are present only in the tallest stereocilia of OHCs and mediate adhesion to the tectorial membrane (not shown).

top connectors have a characteristic zipper-like structure and connect adjacent stereocilia both within and across rows, as well as shaft connectors and ankle links do. Top connectors are thought to have two essential functions: (i) the maintenance of bundle-cohesive architecture by bundling the stereocilia together to form a cohesive V-shape structure to minimize frictional drag and (ii) keeping the bundle as a coherent unit when moving dynamically [10].

Therefore it is not surprising that *STRC*^{-/-} mice show distinct phenotypes which include the distal tips of the stereocilia in all three rows being no longer aligned precisely and exhibiting a degree of disorganization (see fig. 1.3), but also evidence of coupling of TM to hair bundles of OHCs even though it is not possible to detect otoacoustic emissions (a key phenomenon of a hearing ear due to TM's physiological discontinuities in impedance). However, how the exact attachment of TM to hair bundles is realized and how otoacoustic emissions are linked to *STRC* are topics still to be cleared.

1.7 Implications of *STRC* and *OTOA* genetic structure in diagnostic tests

One can now eventually understand the relevance of both stereocilin and otoancorin in allowing to hear and, more importantly, their pathological implications. *STRC* is involved in autosomal recessive prelingual hearing loss (DFNB16 MIM# 603720) that may be severe to profound but it is frequently mild to moderate. Overall mutations in *STRC* accounted for about 6% of cases in *GJB2* negative children. Also *OTOA* is responsible, when mutated, of causing autosomal recessive prelingual hearing loss (DFNB22 MIM# 607039) but this time it can be moderate to severe. It should be noted that major burdens to the clinical and molecular research of these two types of hearing losses are of course the objective general difficulty of in vivo studies addressing the

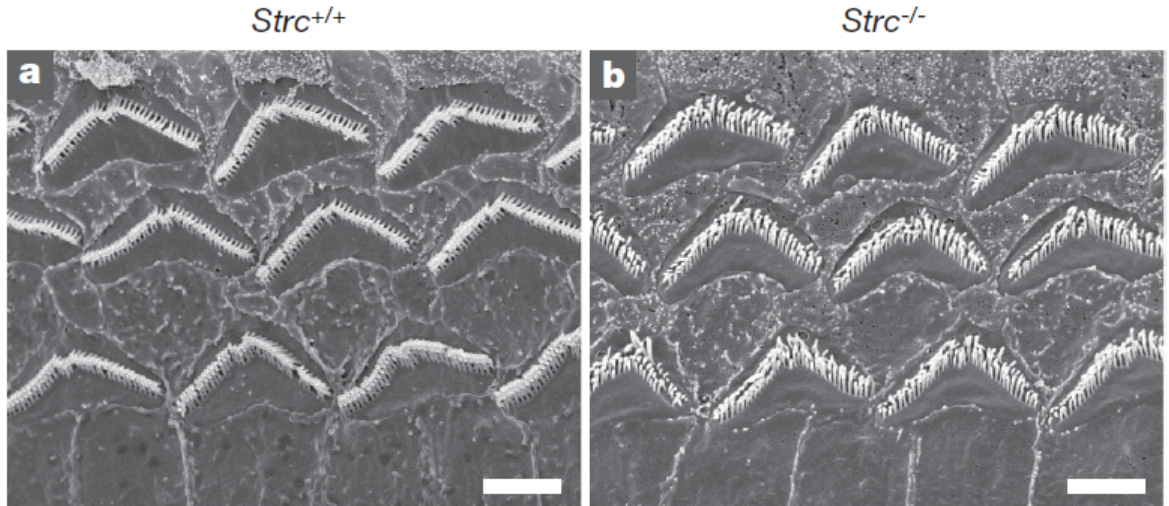


Figure 1.3: Ultramicrograph of OHCs in *STRC*^{+/+} and *STRC*^{-/-} mice seen from above. The cohesiveness of bundles is clearly lost from a) to b) .

internal ear and the lack of in vitro cell culture alternatives, but in particular the presence of an unprocessed pseudogene both for *STRC* and *OTOA*.

In the first case the pseudogene (*pSTRC*) is actually part of a 100 kb tandem duplication that includes also 3 other genes: *KIAA03777* (unknown function), *CKMT1B* (creatine kinase mitochondrial 1), and *CATSPER2* (cation channel, sperm-associated 2). In the second case the pseudogene (*pOTOA*) is not a copy of the entire gene but rather it only duplicates the last 32kb out of the almost 82kb of *OTOA* (although the duplication comprehends also downstream regions). The degree of homology between the pseudogene and the original copy of the gene is astonishing: 99.93% for the first 15 exons of *STRC* (which has a mean of 99.6% homology over the entire region) and 99.6% for *OTOA*, with a mean of 1 different base every 243 identical nucleotides.

These pose a severe constrain to diagnostic PCR- or NGS-based approaches which are usually capable to analyse accurately up to 1,000 (PCR) or 200bp (NGS) fragments. In fact, to my knowledge, there are only two published papers [11, 12] which assume to have found a way to specifically amplify only the expressed copy of the gene (i.e. not the pseudogene) and neither tackle *OTOA*. In addition, due to their particular genetic structures, these regions appear to be a hotspot for CNVs and, although it has not been reported yet, could in principle undergo non allelic homologous recombination (i.e. genic conversion) adding further complexity to diagnostic testing. It is of no surprise, then, that the amount of knowledge accumulated over the years about *STRC* and particularly *OTOA* is somewhat limited.

1.8 In some cases diagnostic tests need further studies

As it is common case in molecular biology when a new methodology is established or when the sequencing of a DNA results in the discovery of new variants, further studies may be necessary to fully comprehend the functional implications, if any, of that variant on cellular biology and physiology. This

is extremely true for the majority of diagnostic tests and in particular when splicing affecting variants are under hypothesis (excluding variants affecting the canonical ± 1 or ± 2 splice sites). In fact, there is no general rule that leads to an accurate prediction of these variants and bioinformatic tools cannot be considered as solid evidence as direct functional studies. Nevertheless, before we further elaborate on that, it is better to recap the framework of pre-mRNA splicing.

2 Splicing

Pre-mRNA splicing is an essential step in eukaryotic gene expression, since it allows 3' polyadenylation and 5' capping processes and the formation of the mature mRNA. Splicing removes from pre-mRNAs the non-coding sequences called introns (usually hundred to thousands base pairs long,) which usually account for $>90\%$ of the primary transcript length [13] and separate the shorter coding sequences called exons, typically 50-250 bp long. In addition to this constitutive splicing, some genes can undergo alternative splicing (AS) as well, which allows the generation of different transcripts with different combinations of exons. As one might expect, these transcripts can lead to the synthesis of very different proteins: for this reason, mutations affecting splicing regulatory elements can have deleterious effects on human health. In their essence, splicing reactions are sequential phosphodiester transfer reactions catalysed by large ribonucleoprotein complexes called spliceosomes, containing more than 1000 core proteins and five small nuclear RNAs (snRNA U1, U2, U4, U5 and U6). However, this number is actually greater, since additional regulatory proteins are involved in the splicing of particular pre-mRNAs.

2.1 Core signals

Core signals (see fig. 1.4) are extremely important in splicing reactions because they are recognized multiple times during spliceosome assembly. Each intron contains the following core splicing signals:

- the 5' splice site (5' ss, also called Donor Splicing Site), that marks the exon/intron junction at the 5' end of the intron and includes a GU dinucleotide;
- the 3' splice site (3' ss, also called Splicing Acceptor Site), formed by a terminal AG at the extreme 3' end of the intron following the polypyrimidine tract (PPT);
- the Branch Point Site (BPS), typically located 18-40 nt upstream of the 3'ss, in higher eukaryotes is followed by a PPT.

In order to function properly, splicing sites sequences are conserved and allow for an easy recognition of either exons (through exon definition) or introns (through intron definition) by the spliceosome.

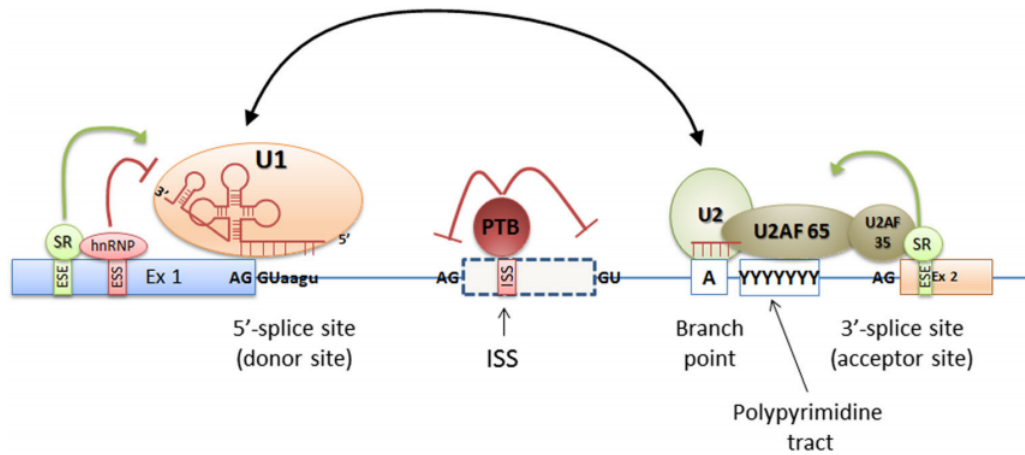


Figure 1.4: Schematic localization of cis and trans splicing elements. Cis elements are the donor (5') and acceptor (3') splice sites, the branch point, the polypyrimidine tract, splicing enhancers and silencers.

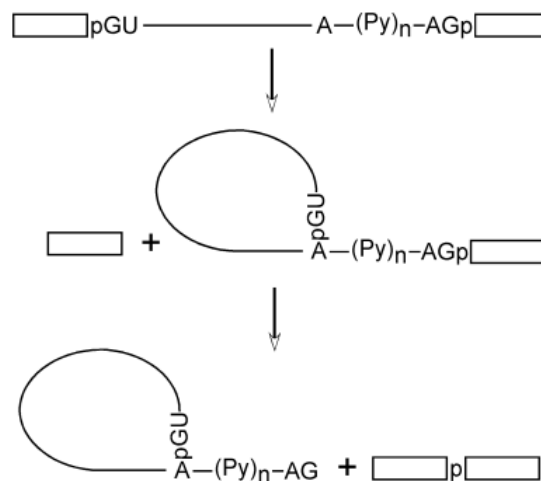


Figure 1.5: Schematic representation of the two splicing transesterification reactions.

2.2 The splicing reaction

Spliceosomes, actually, remove pre-mRNA introns through two consecutive transesterification reactions (see fig. 1.5). Firstly, a nucleophilic attack on the 5'ss phosphate group is carried out by the 2'OH group of the intron branch adenosine. This results in cleavage of this site and ligation of the 5' end of the intron to the branch adenosine, forming a typical lariat structure. Secondly, the other transesterification step is realized through a nucleophilic attack on the phosphate at the 3' intron end by the 3'OH group of the detached exon, resulting in the ligation of 5' and 3' exons. At this point, mRNA is formed, and the intron, in the lariat form, is released.

2.3 Spliceosome assembly

The spliceosome is a multimegadalton ribonucleoprotein (RNP) complex formed by a large number of trans-acting factors interacting with the pre-mRNA in order to spatially position the reactive groups for the catalysis [14]. Not only

it defines the native exon–intron boundaries, facilitating the splicing reactions, but also it is highly dynamic in both conformation and composition: this allows to obtain accuracy and flexibility at the same time. In particular, trans-acting factors includes five snRNPs and several non-snRNP proteins. Each snRNP is formed by a snRNA (except for the U4/U6 snRNP, which is composed by two snRNAs), seven Sm proteins, that are common between the snRNPs, and some particle-specific proteins.

At the beginning, the spliceosome assembly pathway can occur in two different ways, depending on how the recognition of functional splice sites is accomplished by the five snRNPs: the exon definition and the intron definition. These constrain the length of the element being defined to 200–250 nt. The exon definition mechanism probably evolved later than the intron one, and indeed it is the main mechanism in higher eukaryotes. For instance, in mammals most of pre-mRNAs contain introns ranging from hundreds to several thousand nt, while exons have more or less a fixed length (120 nt) [14]. Both the exon and intron definition processes, however, end with the formation of the commitment complex, also known as the E (early) complex (see fig. 1.6). Without now giving too much detail about how each particular definition step occurs, it may suffice to say that the E complex comprehends U1 snRNP bound to 5' splice site and stabilized by protein factors such as the branch point binding protein (BBP or SF1), U2AF (a heterodimer consisting of U2AF65 and U2AF35) and protein of the SR family.

At this point, the first ATP-dependent step is carried on when U2 snRNP associates with BPS and U1 snRNP, leading to the formation of the pre-spliceosome or A complex (see fig. 1.6). When a trimer containing U5 and U4/U6 snRNPs is added to the A complex, the B1 complex is formed and subsequently converted to the B2 complex by replacement of U1 snRNP with U6 snRNP and U4 snRNP release. After a series of RNA rearrangements is completed, the RNA helicase Prp2 is catalytically activated and is responsible for the formation of the spliceosome active site (complex C). In this way, cells make sure that the active site is formed only in correspondence of the right splicing sites. Finally, the two splicing transesterification reactions can take place releasing the lariat; then the spliceosome dissociates, mRNA is released and snRNPs are free to take part in a new splicing reaction [14].

2.4 Cis-acting sequences

Although core human splice site motifs are necessary to a correct splicing, they contain only about half of the information required to accurately define exon/intron boundaries. Therefore, the complementary information is given by cis and trans splicing regulatory elements (SREs). The activity of SREs is characterized by “context dependence” [13], which can be divided into two categories: location-dependent activity, that varies with the relative position of the SREs in the pre-mRNA sequence, and gene-dependent activity. However, SREs activity depends also on the presence of the trans-factors needed for the process, leading to tissue/cell-specific splicing regulation.

In particular, cis-regulatory elements serve as either splicing enhancers or

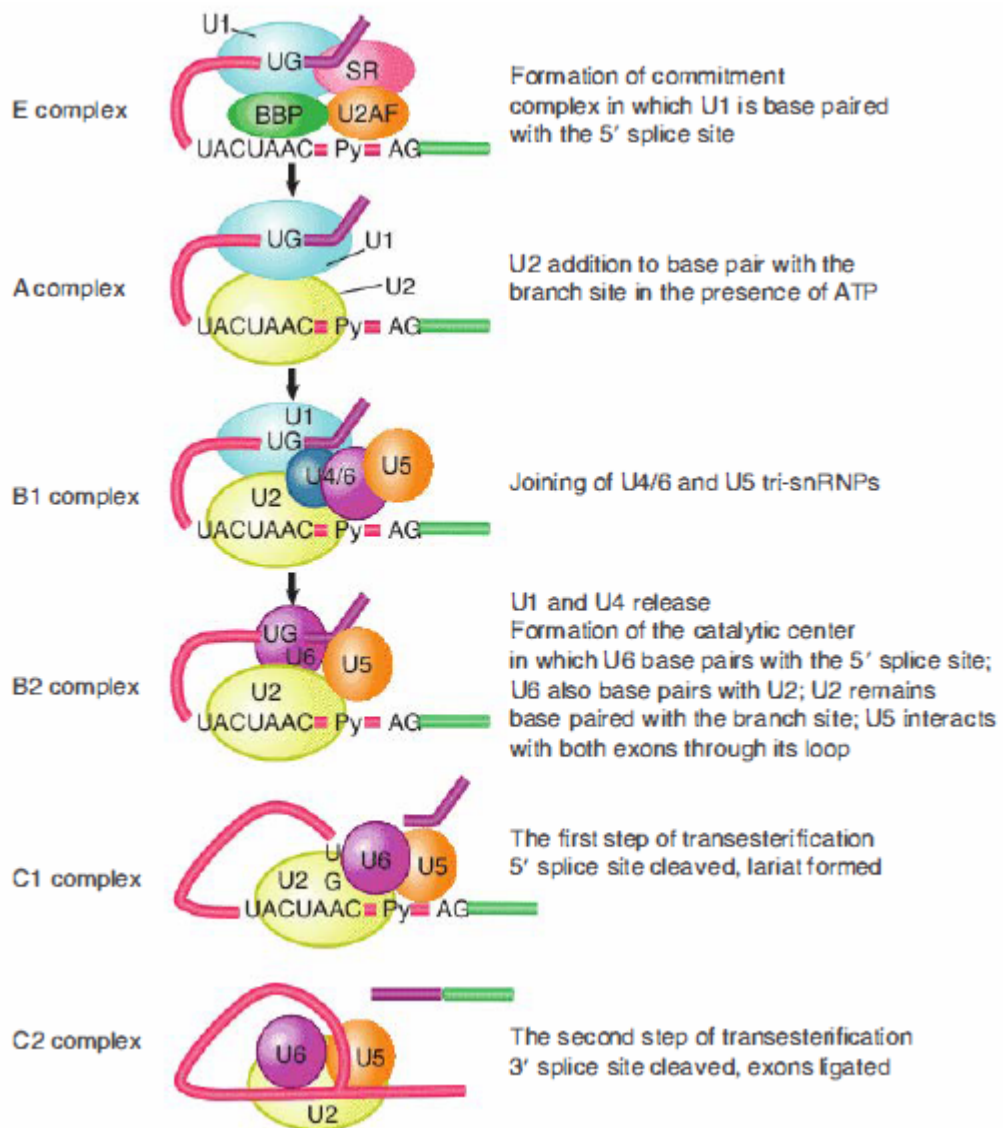


Figure 1.6: Schematic representation of the spliceosome assembly pathway starting from E complex.

silencers [13]. Based on their genetic location they can be classified as exonic splicing enhancers (ESEs) or silencers (ESSs) and intronic splicing enhancers (ISEs) or silencers (ISSs): their function is to modulate both constitutive and alternative splicing, by binding regulatory proteins. Moreover, while ESEs and ESSs promote or inhibit the inclusion of the exon in which they are contained, ISEs and ISSs enhance or inhibit the usage of adjacent splice sites or exons. On the other hand, trans-acting regulatory elements are recruited by the already described cis-acting regulatory elements, and cooperate in activating or suppressing the splice site recognition or spliceosome assembly. Finally, another level of splicing modulation is given by the additive function splicing enhancers and silencers often have[13]. In fact, additional copies increase their effect, because they allow to increase the affinity of the associated factor, or the copy number of recruited factors.

2.5 Exonic regulatory elements

As one might expect, exonic regulatory elements are often embedded within protein coding nucleotides. The difficult identification of these regulatory elements was overcome through the study of exon mutations that block splicing, through computational comparisons of exon sequences, and through the selection of sequences that activate splicing or bind to splicing regulatory proteins (especially SR proteins) [15].

Actually, most ESEs function by recruiting members of the SR protein family, which are responsible for protein–protein interactions, facilitating spliceosome assembly. They constitute the best-studied family of splicing regulators and have a common domain structure formed by one or two RNP-cs RNA binding domains (RRMs), followed by a carboxy-terminal RS domain containing repeated arginine/serine dipeptides, often with highly phosphorylated serines. While RRM domains mediate the sequence-specific binding to the RNA, the RS domain is involved in protein-protein interactions. For example, the unphosphorylated RS domain is highly positively charged and enhance the affinity of the protein for the RNA hybridization, acting as a counter ion. However, it was demonstrated that phosphorylation is required for the activity during splicing. Nevertheless, the binding of SR proteins to ESEs can promote exon definition in two different ways: either the RS domain can directly recruit the splicing machinery to the enhancer sequence, or the protein can antagonize the action of silencer elements that are found nearby. Furthermore, ESEs are often found in long or little clusters made by many ESEs able to be recognized by different proteins [16, 17].

Despite exonic splicing enhancers, ESSs are less well characterized. They are often bound by splicing repressors of the hnRNP (Heterogeneous Nuclear Ribonucleoprotein) class through RNA-binding domains. These hnRNP proteins have a modular structure formed by one or more RNA-binding domains associated with a domain responsible for protein-protein interactions. They have been identified by their association with unspliced mRNA precursors, and the most studied protein is hnRNP A1, which contains two RNP-cs RNA binding domains and a glycine-rich auxiliary domain [18].

2.6 Intronic regulatory elements

Many splicing regulatory sequences are found in introns and can act from distant positions, hundreds of nucleotides away from the regulated exon, or from sites within the polypyrimidine tract or immediately adjacent to the 5' splice site, where regulatory protein binding sites were identified. Although many ISEs, ISSs and the proteins that mediate their effects are still not identified [13, 19], a well characterized ISE is the G triplet, or G run, which is common in GC-rich introns, while intronic CA repeats often enhance splicing of upstream exons. However, when CA-rich intronic sequences are bound by hnRNP L, they act as ISSs. Furthermore, ISSs can be bound not only by SR proteins, demonstrating how these proteins can be either splicing activators or repressors, depending on where they bind to the pre-mRNA, but also by hnRNP A1, though in a different binding site. Finally, also intronic regulatory sequences are often found in clusters.

2.7 Alternative splicing

The major way through which living organisms can leverage their genomic information is through alternative splicing or the different inclusion or exclusion of a portion of the coding sequence in the mature mRNA, giving rise to different transcripts (see fig. 1.7). This mechanism leads, eventually, to the synthesis of different protein isoforms that are composed by different modular peptide sequences, that determine particular chemical and biological activities. Some pre-mRNAs often have multiple sites of alternative splicing: this allows to give rise to an entire family of related proteins, starting from a single gene. However, the small changes of the peptide sequence derived from AS mRNA sequences can often alter the final protein characteristic, such as enzymatic activity, ligand binding, allosteric regulation, and protein localization [18, 20].

In a typical mRNA, various exons are contained. Most of them are constitutive: this means that they are always spliced and included in the final mature transcript. The exons that are not constitutive can be regulated in several ways. The main one is the cassette exon, which can be either included or excluded from the final mRNA. Sometimes multiple cassette exons are mutually exclusive: in the final mRNA, only one of the possible exons is included. For this reason, there are mechanisms that enforce the choice of the exon to be included. Moreover, by altering the position of 5' or 3' splice sites, exons can be lengthened or shortened. However, 5' and 3' terminal exons are not fixed and can be switched with other exons by alternative promoters or polyadenylation sites, respectively. It is also possible that an intron fails to be removed, leading to intron retention.

2.8 Point mutations that affect splicing

Sometimes alternative splicing is not a physiological mechanism of the cell, but is an undesired outcome of alterations found in splicing regulation sequences. As we have seen, splicing modulation is subtle and complex, and requires an elaborate cross-talk between several cis-acting elements and trans-acting

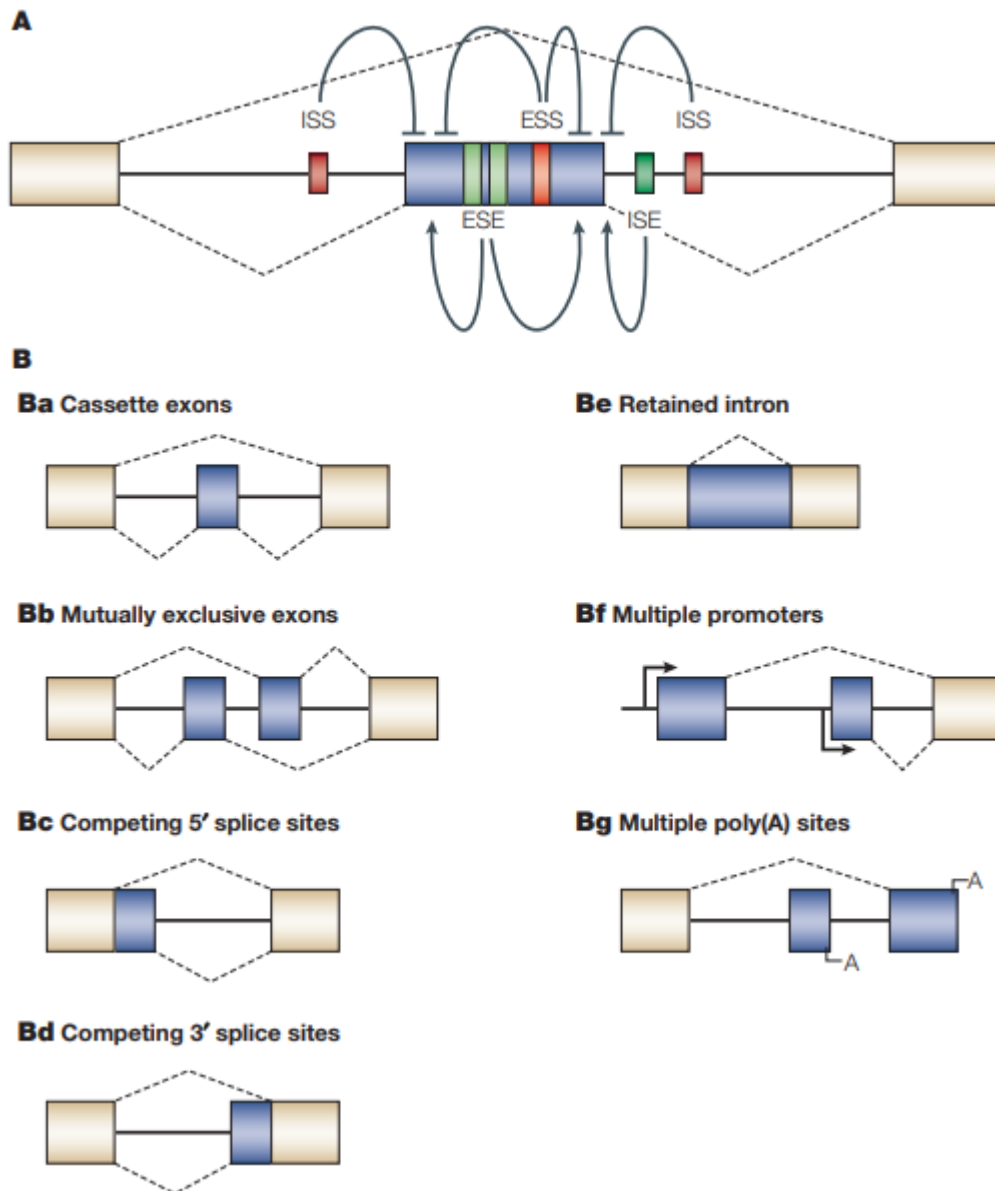


Figure 1.7: A: Splicing regulatory elements and their relation to gene structure. B: Schematic representation of the various alternative splicing patterns possible. Exons are shown in blue.

factors. Exon inclusion, for instance, depends on the intrinsic strength of the flanking splice sites and the combination of the effects of positive and negative regulatory elements. Therefore, if a mutation occurs at the level of these regulatory sequences, it can highly affect the extent of the inclusion of that exon.

Errors in splicing regulation have been implicated in a number of different disease states, such as genetic diseases and cancer. Furthermore, it has been estimated that at least 15% of the point mutations that cause human genetic diseases are splicing defects [16]. When an exon-intron boundary is not accurately recognized, or an intron fails to be removed, an aberrant mRNA is generated and can be unstable, or it can code for defective or deleterious protein isoforms. Of course, this is true as well for mutation in tissue- or cell-specific alternative splicing signals.

The extent of splicing mutations is known to be underestimated because many of the potential mutations located in cis-acting elements are also part of the coding sequence and are first commonly classified as missense, nonsense or silent. Missense mutations are assumed to change an amino acid of the final protein sequence, nonsense mutations are usually thought to produce truncated isoforms of the protein and silent mutations are considered to be neutral allelic polymorphisms. However, neutral mutations can have an important impact on the translated product, if the variation affects sequences with a splicing regulatory role. For this reason, point mutations that do not create ectopic splice-site consensus sequences should be confirmed experimentally to understand their possible pathogenetic role [21].

Mutations that affect splicing can be divided into five categories:

1. mutations affecting splicing donor site (GT) or splicing acceptor site (usually AG), usually causing exon skipping;
2. deep intronic variants, able to create novel acceptor or donor sites, or novel regulatory elements inside an intron;
3. changes in the exonic sequences that introduce a new 5' or 3' splice site, or activate a cryptic one that is stronger than the original one;
4. mutations leading to the activation of a cryptic splicing site, that otherwise would not be recognized;
5. exonic changes able to disrupt exonic splicing enhancers and silencers, often causing exon skipping.

As already noted, it is important to remember that the latter type is difficult to identify, since it can easily be misclassified as synonymous, missense, or nonsense variant [22]. In addition to these 5 types, there are also mutations affecting the branch point and the polypyrimidine tract. This last class of splicing affecting mutations are very rare and hard to identify when genomic DNA is analysed. In fact, the bioinformatic prediction of their exact localization and possible effect is difficult because consensus sequences of these motifs are usually degenerated [22].

Splice-site mutations can have different consequences such as exon skipping, truncation of the polypeptide (e.g. intron-less genes), intron retention, activation of cryptic splicing sites (i.e. sites that otherwise would not be recognized by spliceosome) and decreased mRNA stability (through Non-sense Mediated Decay). The common consequence of these splicing mutation mechanisms is the synthesis of abnormal proteins or the reduction (or even lack) of their synthesis. For these important implications, the development of computational tools able to predict with accuracy ESE or ESS elements as well as ISEs and ISSs, using the consensus recognition sequences for different splicing factors, can significantly help in distinguishing real neutral variants from mutations that severely affect splicing. The study of single-nucleotide changes and their effects on pre-mRNA splicing can have, in fact, a significant impact in the diagnosis and treatment of genetic diseases.

2.9 RNA analysis: PROs and CONs

A significant number of variants associated with Mendelian disorders, or mutations identified using Next Generation Sequencing (NGS) screening or exon sequencing are located near intron-exon boundaries, and are therefore suspected of causing aberrant splicing [23]. Obviously, the most straightforward and reliable method to study a splicing mutation is the direct analysis of patient's RNA [24], in order to study the produced mRNA isoforms and to predict the proteins synthesised from these transcripts. However, the study of the produced proteins is fundamental as well, since it allows to hypothesise the severity of the disease based on residual protein amount and how much activity is still retained.

Nevertheless, RNA analysis has several limitations and drawbacks, also due to RNA limited accessibility. First of all, it is difficult to storage and handle. In fact, RNA samples must reach the laboratory within 2 hours from the sampling procedure and they must be conserved in the right conditions in order to avoid not only RNA degradation but also the appearance of illegitimate splicing, that can interfere with the analysis, especially with long transcripts. Furthermore, the RNA extraction itself is complicated and tricky, due to contamination risks, to the presence of RNases and more importantly to RNA degradation. As a matter of fact, RNA is more susceptible to degradation than DNA and more sensitive to the action of specific lytic enzymes, since it has larger grooves that make it easier to be attacked by enzymes. In addition, while DNases require metal ions for their activity and are thermolabile, RNases do not need cofactors, resist to high temperatures and maintain their activity within a wide pH range. For these reasons, RNases are ubiquitous, thus both RNA extraction and analysis require the use of RNase free spaces and material, expertise, and attentions that DNA analysis do not require.

It is also important to remember that gene expression depends on cell type and tissues, therefore RNA should be extracted from cells expressing the abnormal transcript, but this is not possible in tissues that are barely accessible, like brain, pancreas or inner ear. Gene expression is also time-regulated, and some transcripts are produced only in precise time frames, for

example at particular stages of development, during which RNA sampling cannot be performed.

For these reasons, clinical genetic testing still relies largely on genomic DNA [24], that can be easily obtained from all human tissues. But using genomic DNA, it is not possible to directly assess the effect of splicing mutations, that can be seen only after the transcription process.

2.10 Hybrid Minigenes

In order to assess if DNA genetic variants can alter the splicing mechanism and to establish the related pathogenicity, bioinformatic analyses and in particular functional studies are required. The former can be used to predict if the effect of DNA variants could affect pre-mRNA splicing, or to identify exon-intron boundaries. These methods have intrinsic limitations due to the arbitrary choice of threshold levels and the lack of reliable standard guidelines for the interpretation [23]. Functional studies on the other hand are obviously more accurate than bioinformatic analyses but are more complex as they include splicing assays and the use of hybrid minigenes.

In vitro splicing assays consist of labelling preformed RNA molecules, transcribed using bacterial RNA polymerases, and incubating them with nuclear extracts in order to activate the splicing process. Splicing products are then resolved using polyacrylamide denaturing gels. This step allows also the visualization of splicing reaction intermediates. However, drawbacks of this technique include the relatively short sequences that can be analysed using this method, the not easy standardization, and most importantly the fact that the intimate connection between cell transcription and splicing is not taken into account.

Hybrid minigenes can help overcome some of these issues, and represent an in vitro functional assay that allows testing the splicing efficiency and effects without using the patient's RNA. Therefore, minigenes can turn extremely useful when the genes analysed are expressed in tissues difficult to reach like hear-related genes. However, it should always be clear that minigenes have several limitations due to the fact that they are artificial constructs and, as such, do not reflect a physiological condition, sometimes showing some degrees of illegitimate splicing.

Chapter 2

Aim of the study

Hearing loss is a common but complex and genetically heterogeneous disease which has received growing attention during the last decades thanks to the advent of NGS. Two genes, *STRC* and *OTOA*, involved in autosomal recessive hearing loss possess a high homologous pseudogene and are hotspot for CNVs. Screening for variants in these two genes is hardly approachable by NGS techniques which cannot correctly align the identified variants over neither the gene nor the pseudogene locus due to short read length.

Therefore, this study aims to:

1. develop a novel method for selectively amplifying either the gene or the pseudogene copy;
2. validate previous NGS-identified variants in *STRC/pSTRC* and *OTOA/pOTOA*;
3. develop a simple statistical test to predict CNVs based on NGS coverage at the genomic position of the variant;
4. assess CNVs in deaf patients for *STRC* and *OTOA* loci.

As a complement, this study also aims to:

- 5 validate splice site variants identified by NGS in other hear-related (*OTOG*, *MYO15A*, *COL2A1*, *COL11A2*) and not hear related (*COL1A1*) genes through hybrid minigene assay.

Chapter 3

Materials and methods

1 *STRC/pSTRC* and *OTOA/pOTOA* variants sequencing

1.1 Patient samples and suspected mutations

Genomic DNA samples were extracted from whole blood of deaf or suspected deaf patients using MagPurix® (ZINEXTS) which exploits magnetic silica beads for automation of the process. Qualitative and quantitative DNA analyses were then performed respectively by 1.5% agarose gel electrophoresis (see section 1.6 Agarose gel) and Nanodrop 2000. All samples had been investigated for diagnostic purposes and a panel of genes associated with hearing loss was analyzed. Written informed consent was provided by all patients.

The coding regions of the selected genes were isolated and captured using the HaloPlex Target Enrichment system (Agilent Technologies, Santa Clara, CA, USA); indexed DNA fragments libraries were generated according to the manufacturer's protocol and sequenced on a MiSeq Dx instrument (Illumina, San Diego, CA, USA), with 150 bp paired-end sequencing. Variant calling and bioinformatic analyses were performed using the SureCall software (Agilent Technologies). The remaining DNA was eventually stored at -20 °C.

A total of 28 rare, likely pathogenic variants identified in either *STRC/pSTRC* or *OTOA/pOTOA* loci were selected for further analyses. A protocol for the confirmation of these variants by Sanger sequencing was tested.

1.2 Experimental design

The main idea behind the experimental design of this study was to first perform a long-range PCR (lPCR) over either *STRC* or *OTOA* gene for each sample and then assess the genotype of variants previously identified by NGS with Sanger sequencing of a nested PCR (nPCR). Since gene and pseudogene are really highly homologous with only few sparse different bases, it was expected that, for the sake of simplicity and protocol standardization, the lPCR should have provided the specificity of the reaction to either gene or pseudogene, thus allowing nPCR to be more often than not non specific.

1.3 PCR primer design

If not already available in laboratory, all PCR primer pairs were designed taking into account several factors in order to maximize specificity and yield. These include melting temperature, GC content, absence of SNPs (particularly at 3' end) or common genomic repeated sequences (such as transposons) and prediction of both secondary structure and self-annealing.

As commonly suggested by most authors [7], melting temperature of forward and reverse primers should not differ for more than 5°C, as well as they should have a GC content between 40 and 60% in order to be readily denatured during PCR: these guidelines were strictly followed for every primer pair designed throughout this study. Moreover, since SNPs are known to be a major cause of allele dropout or aspecific contamination during PCR, each primer pair was checked via SNP CHECK website (<https://genetools.org/SNPCheck/snpcheck.htm>; `jsessionid=49AD61323CEC362C5DDFAB4388578DFA`) and BLAT tool (<https://genome.ucsc.edu/cgi-bin/hgBlat>) from UCSC Genome Browser. However, in few cases, where it was not possible to avoid SNPs, the primer site was slightly adjusted to put the SNP as far as possible from the 3' end, thus reducing its destabilizing effect on primer extension. At this point, the supposed primer site was verified through SMS primer stat tool (https://www.bioinformatics.org/sms2/pcr_primer_stats.html) in order to avoid both secondary structure (such as hairpin) and self annealing, which could greatly reduce the effective available primer amount and eventually result in PCR failure.

Although at first only those foresaid factors had been considered, it was found really useful in improving PCR quality to exclude as much as possible genomic repeated regions from designing primers, despite the fact that every primer pair was predicted to amplify a unique (or double in the case of gene/pseudogene amplification) sequence tagged site (STS) by UCSC Genome Browser's *In Silico* PCR tool. Sometimes this was possible only at a major cost as repeated sequences cover up to almost 45% of the Human genome.

In few cases it was of no convenience to redesign pairs already published in the literature and a complete list of all primer pairs used throughout this study is reported in tables 3.1 and 3.2.

1.4 Long range PCR

As a starting point for the *STRC* IPCR, the protocol described by Mandelker et al. was followed [11]. The reaction mix was set in 30 μ L final volume including roughly 100 ng gDNA, 1X buffer 1, 1 μ L dNTPs mix (10mM each), 1.5 μ L of forward and reverse primers (10mM each, see table 3.1), 3U of Taq Long and water. However, instead of the IPCR kit TAKARA LA kit v2.1 (Takara Bio Inc.), the ROCHE Expand Long Template PCR SYstem was used. Thermocycling conditions were as follows: 1 cycle of 94°C for 2 minutes, 36 cycles of 98°C for 10 seconds, 68°C for 12 minutes 10 seconds, 1 cycle of 68°C for 7 minutes. The reaction mix was slightly modified for *OTOA* IPCR as 1.7% DMSO was added to improve specificity. Annealing and final extension temperature of *OTOA* were of 64°C. Another similar protocol was published

1. STRC/PSTRC AND OTOA/POTOA VARIANTS SEQUENCING

Table 3.1: List of all IPCR primers used in this study. Highlighted in green are original primers of Mandelker’s approach[11], while in orange are primers that failed completely in achieving the expected amplicon.

locus	exon/intron	primer	5'-3' sequence	amplicon lenght (bp)	Tm (°C)
pSTRC/STRC	all	FOR REV	cagctcagagttttgataggccttca aggaagcagatcaagattagtgtccctt	20,343	66
pSTRC/STRC	all	FOR REV	tgatgggcttttaccta gagatcaagattagtgtccttc	20,323	55
pSTRC	all	FOR REV	cagctcagagttttgatgggct* t*t*t gaagcagagatcaagattagtgtcc* t*t*c	19,992	66
STRC	all	FOR REV	cagctcagagttttgataggct* t*t*c aggaagcagatcaagattagtgtcc* t*t	20,343	66
pOTOA 1/OTOA 1	ex 20-23	FOR REV	cctactgccagagataccag gttcagacatggacatcttc	15,764	51
pOTOA 1/OTOA 1	ex 20-23	FOR REV	gggcagacctcttagagatgt gggctcctcaactatcaggcac	17,718	64
pOTOA 1/OTOA 1	ex20-23	FOR REV	gggcagacctcttagagatgtc ggctcctcaactatcaggcacg	17,716	64
OTOA 1	ex 20-23	FOR REV	gggcagacctcttagaga* t*g*t gggctcctcaactatcagg* c*a*c	17,718	64
pOTOA 2/OTOA 2	ex 24-28	FOR REV	cattgcctgaagacacttt atgacttcagtaaagcagagtaag	12,359	50
pOTOA 2/OTOA 2	ex 24-28	FOR REV	gtaagagattcattgcctgaagacacttt aaacaatcagtaatcgggaagcgacc	13,738	64
pOTOA 2/OTOA 2	ex 24-28	FOR REV	gtaagagattcattgcctgaagacactttc gcatgacagaagttaagactattatcgg	14,910	64
OTOA 2	ex 24-28	FOR REV	gtaagagattcattgcctgaagacac* t*t*t gtgcatgacagaagttaagactatta* t*c*g	14,910	64

by Vona, Hofrichter, Neuner *et al.* [12]. While in *STRC* it was possible to amplify the entire gene with a unique IPCR, this was not the case with *OTOA* since the homologous region between gene and pseudogene was roughly 35kb long and thus required to split the analyzed region into two smaller IPCR amplicons (see table 3.1).

Since *STRC* IPCR results were not sufficiently reproducible following the Mandelker’s protocol, then some relevant modifications were made. The lack of reproducibility could be due to the different IPCR kit and to the use of different thermocyclers available in the laboratory. Constant reproducible results were obtained by using a single thermocycler (Agilent SureCycler 8800) and by lowering the annealing and final extension temperatures from 68°C to 66°C. Once IPCR variability was drastically reduced, further optimization up to 33% save in cost (2U of enzyme) and 22% save in time (28 cycles) were applied both to *STRC* and *OTOA* samples.

As it will be explained later (see section 1.5 Tackling homology: third and final IPCR redesign and Discussion), the modification of both primers with three 3' phosphorothioate bonds was necessary in order to finally achieve the requested specificity. However, before coming to this final solution different primer set were tried (see table 3.1).

1. STRC/PSTRC AND OTOA/POTOA VARIANTS SEQUENCING

Table 3.2: List of all nPCR primers used in this study. Highlighted in yellow are primers already present in the laboratory.

locus	exon/intron	primer	5'-3' sequence	amplicon length (bp)	T _m (°C)
pSTRC	ex 25	FOR REV	aggtaaaggtggacttgctg taaacacccctctcaggccca	415	59
pSTRC/STRC	int 18	FOR REV	cctctgatttcgggtaaaagg gaaattcgagaccacccctga	267	54
STRC	ex 15	FOR REV	gctttggtcctttccacc taaccacctgtcgtcctagc	471	60
pSTRC/STRC	ex 8-9	FOR REV	acagcagggtacagagg tcttcctagaacaccgaccc	609	60
pSTRC/STRC	ex 4.3	FOR REV	gccaatgcaggataagtcgt cactactcctctagatttcg	551	59
pSTRC/STRC	ex 29	FOR REV	ttggctgtctggctctta tcaggteggteggaaagcat	370	60
pSTRC/STRC	ex 23	FOR REV	tcctatttctcagtgctct acctcctactgtgacccaa	433	59
pSTRC/STRC	ex 1	FOR REV	tatccacacagtgagaat tcttttccagccacccctctc	179	59
pSTRC/STRC	ex 5-6	FOR REV	ccagcaagtgagctggaat gatgctcagatccctgccatt	476	63
pSTRC/STRC	ex 7	FOR REV	tggagcctagtgttcagagg gcacattgcctatctggc	398	58
pSTRC/STRC	ex 4.1	FOR REV	tcagggtcagaatcttcagc agcaggccagcacagag	543	58
pSTRC/STRC	ex 4.2	FOR REV	cacgaccagtttctctgatg ggatggtcccagtggtg	525	59
pSTRC/STRC	ex 10	FOR REV	tgtaccatacctctgctg caagttgacacaatgggaaag	295	58
STRC	ex 25	FOR REV	gatacccatactagtgcct atgggaccagacctctcatg	483	56
STRC	ex 20	FOR REV	gctcattgcttaagggaag aggggaccttaagaatattgg	453	53
STRC	ex 26	FOR REV	gaaggatcatgaaggtctggtc ccttaagaattgcagggcagt	367	58
pOTOA 1	ex 21	FOR REV	cctactgccagagataccag gttcagacatggacatctttc	703	50
pOTOA 2	int 27	FOR REV	actttgggaggctgaggcaa atttaagacagagtctcgtc	250	54
pOTOA 1/OTOA 1	ex 21	FOR REV	cacagtttgacagtcccacg cgcacctgaaccaaagtgt	619	56
pOTOA 2/OTOA 2	ex 28	FOR REV	caaggctctgcatcaagtgg cacatcgcctcttatca	470	56
pOTOA 1/OTOA 1	ex 20	FOR REV	ggaagtgggacctgtgattg gcttcagaacctatagacagc	330	55
pOTOA 1/OTOA 1	ex 22	FOR REV	gattcaaacgtgctccctgcc acatctcatcccattctctgc	569	56
pOTOA 1/OTOA 1	ex 23	FOR REV	aatcatctcccaaggcccaa tccacaatgctcatcaccca	496	58
pOTOA 2/OTOA 2	ex 24	FOR REV	gatggtttggagtggggatt ccaagtctcaggagtgga	398	57
pOTOA 2/OTOA 2	ex 25	FOR REV	gggctaaggataagctgg acgccactgatgttctac	392	56
pOTOA 2/OTOA 2	ex 26	FOR REV	ccaatcagggtcaatgtgacc aaattccaaggcccatcacc	385	58
pOTOA 2/OTOA 2	ex 27	FOR REV	atacacacaaggggatggc cttgtaggggaagcttgatagg	476	55

1.5 Nested PCR or direct PCR

All nested PCR reactions were conducted as they were normal PCR with the only exception of genomic DNA being substituted by long PCR amplicons. Mix reactions were set in 25 μL final volume with 2 μL of template DNA (for the sake of standardization), 1X buffer, 1.5 μL of MgCl_2 , 0.5 μL dNTPs mix (10mM each), 1.25 μL primers (10mM each), 0.15 μL Taq Gold and water. Thermocycling conditions were as follows: 1 cycle of 95°C for 10 minutes, 35 cycles of 95°C for 30 seconds, T_m (see table 3.2) for 40 seconds, 72°C for 1 minute and 1 cycle of 72°C for 10 minutes.

For each IPCR product, at least one aspecific nested PCR was designed for the amplification of regions with different bases between gene and pseudogene; this was used as a control of IPCR specificity. In addition, to assess the contribute of residual gDNA in each nested reaction, results were initially normalized through a unique STS outside each long amplicon.

1.6 Agarose gel

Before any subsequent analysis, all IPCR amplicons were quality checked on 0.7% agarose gel and run in 1X TBE, while a 1.5-2% agarose gel was used for nested amplicons (max 1000 bp long) and minigene analysis.

At least two lanes were always reserved for molecular marker and for negative control, in order to have an idea of the weight of samples and the possibility of contamination during PCR mix preparation. DNA bands were stained and UV visualized with 7.5% SYBR Safe which is reported as a safer alternative to ethidium bromide. Usually all gels were run at 120-130V for 30-40 minutes which gave a time effective resolution of bands without risking too much thermal degradation of DNA.

1.7 Purification and Sanger sequencing

Nested amplicons were subsequently sequenced by Sanger method. Before sequencing, nPCR products were purified from excess primers and unincorporated dNTPs with Illustra ExoProStar 1-Step (GE Healthcare Life Science) and then prepared with BigDyeTM Terminator v3.1 Cycle Sequencing Kit (ThermoFisher) with the same PCR primers. Following BigDye reaction purification using CENTRI-SEP columns (Princeton Separations), samples were run on ABI 3500 Genetic Analyzer with 8 capillaries.

2 CNVs assessment

2.1 Statistical model for CNV detection

Once all patients were genotyped, in order to further analyze results and try to predict CNVs (in particular deletions), previous data from NGS targeted exon sequencing were used. First, the mean allele frequency amongst both copies of the gene and the pseudogene was calculated for each tested variant

as follows: the sum of the reads with the variant mapping to the gene and the pseudogene was divided by the total depth of coverage (gene + pseudogene) at the level of the nucleotide with the genetic variation.

In order to compare and classify each patient by the number of mutated alleles based on the previous calculated estimate of mean allele frequency, it was necessary to gather a reference population of estimates of allele frequency from known real heterozygotes in autosomal genes with no significant homology with other genomic regions (which have an expected mean of 50%). A total of 21 mutations or SNPs were thus screened and mean and variance computed. Then, using mean and variance properties, this population was translated twice in such a way that new means were respectively of 33% and 25%, corresponding to one mutated allele over a total of 3 or 4 alleles. At this point, 95% and 99% confidence intervals were calculated for each distribution and estimate of real allele frequency matched with the nearest intervals. In this way, each patient's variant could be classified as being hemizygous, heterozygous or homozygous/double heterozygous (across gene and pseudogene). However, it is important to remember that this system was not built to predict duplications.

In addition, a slight variation to this model was done by comparing mean allele frequencies with intervals for each population (with mean 50%, 33% and 25%) based on the maximum difference (in percentage), with respect to the mean, seen in the 21 mutations or SNP previously screened. This was done to improve the sensitivity of the system towards heterozygous variants.

2.2 SureCall pair analysis software

The reference method for CNV assessment in the laboratory is SureCall (Agilent) built-in pair analysis tool. It performs a match between reads coverage for each base in a testing sample and a reference one. When reads coverage counts differ significantly from one another in a wide region, the testing sample is marked as a potential copy number gain or loss in that region and given a score between 0 and 1 (maximum likelihood). However, results are the most reliable when more than one adjacent region is found (ideally over the whole locus) since this tool is too sensitive and has a high background noise (which of course depends on the chosen reference sample, too).

3 Hybrid minigene assays

3.1 DNA samples for hybrid minigene constructs

In this study, a total of 5 mutations were analyzed and tested for pathogenicity through hybrid minigenes (see table 3.3). All but one DNA samples were both already present in the laboratory and genotyped as heterozygote by NGS. The latter one (*COL1A1*), which was actually seen by a fellow laboratory in a fetus died for *osteogenesis imperfecta*, was not stored in laboratory and required an additional mutagenesis step in order to recreate the correct haplotype.

3.2 Bioinformatic analysis

Before all subsequent steps, bioinformatic analyses were conducted in order to predict the position of 5' and 3' splice sites and to evaluate if any alteration of the expected splice site caused by the mutation would be detected or not. For this purpose, three different online software were used: Human Splicing Finder (<http://www.umd.be/HSF/>), NetGene2 (<http://www.cbs.dtu.dk/services/NetGene2/>) and NNSplice (https://www.fruitfly.org/seq_tools/splice.html). The first one is probably the most known and exploits a position-dependent logic to discover splice sites, while NetGene2 and NNSplice are both neural networks-based. As usually happens, there is not a perfect algorithm and each of them has its own strengths and weaknesses. However, it should be remembered that prediction combined from different *in silico* tools was considered in this study as a single piece of evidence because algorithms have similarities in their underlying basis [25].

3.3 PCR primer design and cycling conditions

Design of primer pairs for hybrid minigene assays required extra effort compared to foresaid ones, since it should be avoided as much as possible to disrupt any exonic or intronic regulatory sequences adjacent to tested mutations. Even though it is not possible, with current knowledge, to exactly predict where these sequences are located in genes, as a general rule it was tried to include at least 150-200 intronic nucleotides upstream and downstream exon of interest, which should give a high chance of retaining most of them. Indeed, when adjacent introns were too short (e.g. *COL2A1*, *COL11A2*, *MYO15A*) and this approach would have resulted in partial successive exon inclusion or incomplete (i.e. less than 150-200 nucleotides) 5'/3' intron retention, the previous rule was adjusted increasing or decreasing the region to be amplified. However, hybrid minigene construct poses a limit for the insert size (approximately 1,000bp) and there is the possibility that not all splicing regulatory sequences have been included, thus hindering result interpretation (see Discussion).

Moreover, each expected amplicon was checked with NEB cutter V2 tool (<http://nc2.neb.com/NEBcutter2/>) to detect restriction sites, in particular for HindIII, XhoI and NotI (for details see figg. 3.1, 3.2). When any of these were present, NotI and HindIII restriction sites were introduced as 3' tail of, respectively, forward and reverse primers (see table 3.3). If possible, the choice deliberately excluded XhoI for which there is not a commercial available high-fidelity version. The introduction of these sites would then turn useful for ligation of PCR products into the expression vector.

Cycling conditions were kept equal as much as possible for all samples in order to standardize protocol and reduce error rate. Therefore, all reactions were performed in 50 μ L final volume and contained 10 μ L of 5X Phusion HF or GC Buffer, 1 μ L of dNTPs (10 mM each), 2.5 μ L of forward and reverse primers (10 μ M each), 0.5 μ L (1U) of Phusion® High-Fidelity DNA Polymerase (NEB), 150ng of DNA and water. At first, Phusion HF Buffer was the favourite testing choice, but in three cases (variants in *COL1A1*, *COL2A1* and *COL11A2*), better results were achieved with GC buffer and up to 2%

3. HYBRID MINIGENE ASSAYS

Table 3.3: Primers used for hybrid minigene assays are shown below. Restriction sites are capitalized. Highlighted in yellow are PCRs that required also 2% DMSO with GC buffer, while in red and capitalized there is the mutated base for *COL1A1*.

Gene and exon	Refseq and variant	primer	5'-3' sequence	restriction site	amplicon length (bp)	Tm (°C)
OTOG ex 48	NM_001277269.1: c.7926C>T p.(=)	FOR	cttctGCGGCCGCTgaggaggccaatttaatgag	NotI	602	67
		REV	cttctAAGCTTaccgaaaccctcttggaaat	HindIII		
MYO15A ex 14-15	NM_016239.3: c.4779+9G>A	FOR	cttctGCGGCCGCTccccactactaccagcatt	NotI	832	67
		REV	cttctAAGCTTccagcccttgcataattctgtt	HindIII		
COL1A1 ex 21-23	NM_000088.3: c.1515G>A p.(=)	FOR	cttctGCGGCCGCTcctccttgcctctctctct	NotI	855	68
		REV	cttctAAGCTTtggctcatttccagcacagc	HindIII		
COL2A1 ex 24-26	NM_001844.5: c.1734+3A>G	FOR	cttctGCGGCCGCTcttcttccatccacaccg	NotI	942	68
		REV	cttctAAGCTTcccagaaagtgcacacacg	HindIII		
COL11A2 ex 19-21	NM_080680.2: c.1819-5T>C	FOR	cttctGCGGCCGCTgggtgtgtttgtaattggg	NotI	868	68
		REV	cttctAAGCTTcagacctccaatccatcca	HindIII		
COL1A1 mutagenesis	/	FOR	gtgttctgtgtcccaaAgtaacctctccttgcg	/	all vector	60
		REV	cgcaaggagaggttacTtgggaccagcaacac	/		
β -globin int 2	/	FOR	cagtctcctagtacattactatttgg	/	vary	55
		REV	ttccctgaaagaaagagatt	/		
β -globin ex 2-3	/	FOR	ttgagtccttgggatctg	/	vary	55
		REV	accagccaccacttctgat	/		

DMSO as the regions investigated have actually high levels of GC composition (around 60% or more). Thermocycling parameters were as suggested from manufacturer's protocol (for annealing temperatures see table 3.3).

3.4 Agarose gel and purification

In order to quality check PCR results, all amplicons were run on 1.5-2% agarose gel with standard conditions (see section 1.6 Agarose gel). Although, generally speaking, results were sufficiently good, in two cases (*COL2A1* and *COL11A2*) they were not considered as optimal as desired and thus slightly different conditions were unsuccessfully tried to improve PCR specificity; in these cases, band purifications through gel excision and DNA extraction using QIAquick Gel Extraction Kit (Qiagen) were performed. On the contrary, the purification step of the remaining samples was performed with Amicon Ultra 30K Centrifugal Filter Device (Merck Millipore). It should be noted that PCR purification from unextended primers and primer dimers is a key point for definitely improving digestion efficiency and ligation efficacy. Nevertheless, it is also important to remember that dimer or multimer formation between digested products or even self ligation, although rare, remain possible.

3.5 pcDNA3.1 hygro β -globin vector

The minigene vector pcDNA3.1 hygro β -globin was previously generated in my laboratory [26]. Briefly, the backbone of the hybrid minigene was obtained from a 1.8Kb fragment containing the entire human β -globin gene amplified from gDNA of a healthy volunteer. PCR primers allowed introduction of NheI and ApaI restriction sites in the product. Following digestion and cloning of this fragment into corresponding restriction sites of pcDNA3.1 hygro vector (Invitrogen), human β -globin gene expression was driven by human cytomegalovirus (CMV) immediate-early promoter (see fig. 3.1). Finally, BsrGI

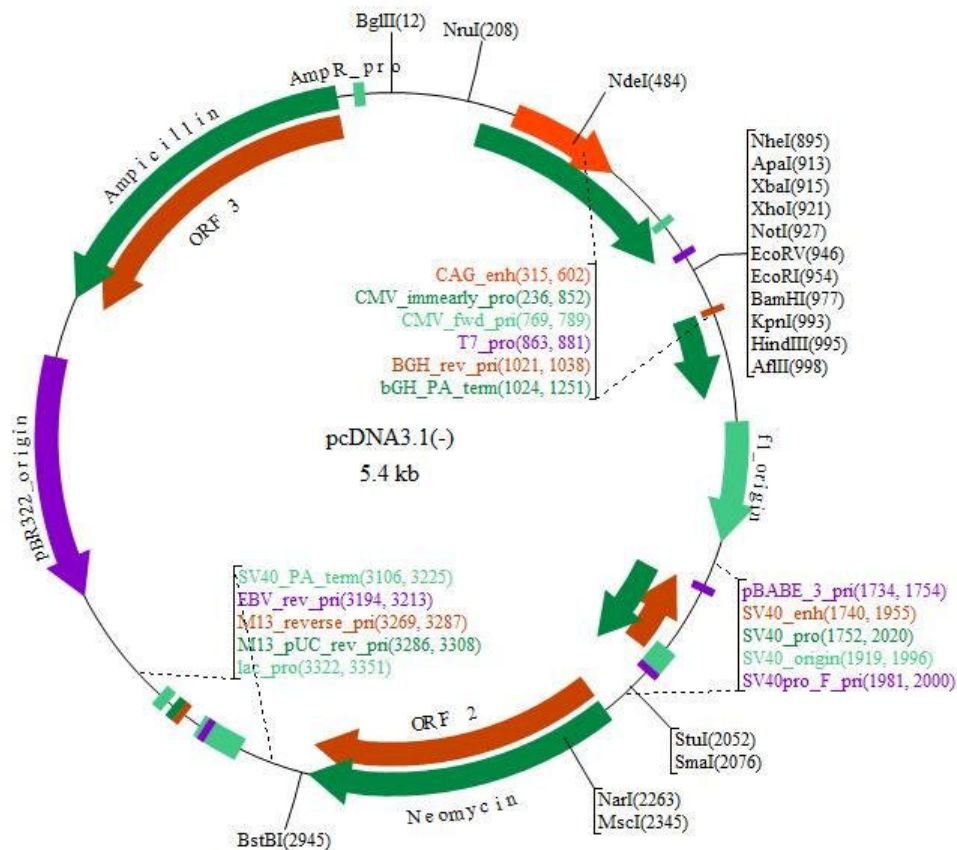


Figure 3.1: Schematic representation of pcDNA3.1 hygro vector (Invitrogen) with all restriction sites.

restriction site within intron 2 of human β -globin gene was exploited for the introduction of a multiple cloning site with XhoI, NotI and HindIII (see fig. 3.2).

3.6 Plasmid miniprep

In order to have sufficient DNA vector for all samples, bacteria previously transformed with pcDNA3.1 hygro β -globin vector and stocked in glycerol were inoculated into 2ml of LB liquid medium + 1X ampicillin and incubated overnight at 37°C with shaking at 230 rpm. The following day, plasmid DNA isolation was realized through either QIAprep Spin Miniprep Kit (Qiagen) or PureLink Quick Plasmid Miniprep Kit (Invitrogen) as described in manufacturers' instructions. Finally, DNA quantification was obtained using Nanodrop 2000 (Thermo Fisher Scientific).

3.7 Digestion

All PCR fragments and pcDNA3.1 hygro β -globin vector were digested in 50 μ L final volume reaction with 1X CutSmart Buffer (NEB), 10 U of restriction enzymes (NotI-HF and HindIII-HF), either 20 μ L of PCR product or 4 μ g of vector and water up to final volume. Following incubation at 37°C overnight,

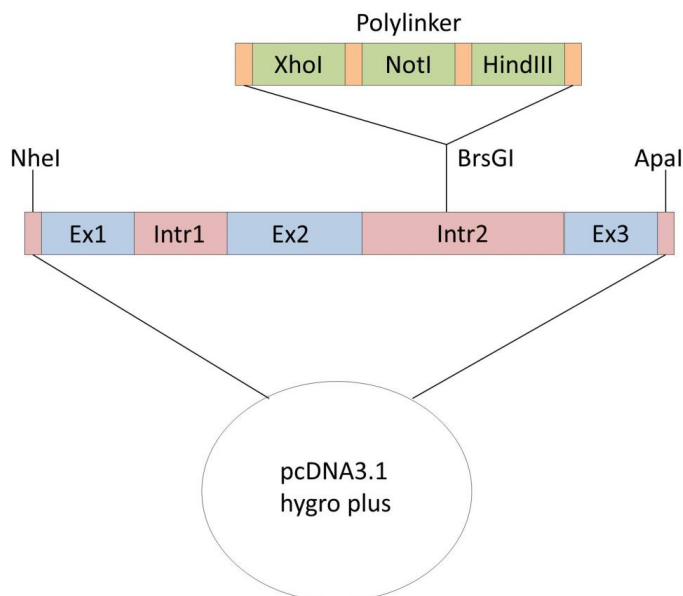


Figure 3.2: Schematic representation of pcDNA3.1 hygroβ-globin vector with highlighted the β-globin insert and the polylinker.

enzymes were then heat inactivated at 80°C for 20 minutes. Although samples were not quantified, from previous laboratory knowledge, it was thought 20 μL would be a sufficient amount to cope with subsequent steps. The day after, in order to further decrease the chance of self-ligation of digested fragments, which is extremely deleterious for the ligation efficiency between PCR fragments and vector, the latter was dephosphorilated with 2 μl (2 U) recombinant Shrimp Alkaline Phosphatase (rSAP, NEB) even though in theory such events should not have happened. Then, residual activity was stopped at 65°C for 5 minutes. The advantage of using such phosphatase is that it works as well in CutSmart Buffer and could even be added directly to the digestion mix.

3.8 Purification and quantification

To separate digested PCR products (from now on called inserts) from adaptors and check for quality of digested vector, all samples were run on 2% agarose gel and bands were extracted using QIAquick Gel Extraction Kit (Qiagen). As a negative control an undigested vector was run, too. Finally, DNA was quantified using Nanodrop 2000, although just to have a general idea of quantity since high contamination levels of carbohydrates were expected.

3.9 Ligation

The amount of insert to use in each ligation reaction was calculated using the NEBioCalculator tool (<https://nebiocalculator.neb.com/#!/ligation>) which exploits the following formula:

$$ng\ insert = \frac{ng\ insert * Kb\ size\ of\ insert}{Kb\ size\ of\ vector} * molar\ ratio$$

The higher the molar ratio, the higher the chance of a ligation event between insert and vector; therefore, it was chosen a molar ratio of 5:1 which was also compatible with actual DNA volume. Reactions were performed in 20 μL final volume with 100 ng of vector, calculated amount of insert, 1X T4 DNA Ligase Buffer, 1 μL of T4 DNA Ligase and water. Reactions were gently mixed by pipetting and incubated at room temperature (RT) for 2 hours (but it is also possible to incubate overnight at 16°C). The hybrid minigene construct should have been eventually formed at the end of this period.

3.10 Competent bacteria transformation

5 μL of the ligation reactions were used to chemically transform One ShotTM Mach1TM T10 Chemically Competent *E. coli* (Thermo Fisher Scientific). This strain is more competent than others and undergoes recombination less frequently, therefore has a really high transformation efficiency. In particular, the transformation protocol included at first thawing bacteria on ice for 30 minutes with each ligation aliquot. Then a sudden thermal shock was performed at 42°C for 45 seconds after which bacteria were chilled on ice for 2 minutes. Recovering was allowed by adding 250 μL of Luria Broth (LB) liquid medium to each vial and left at 37°C for 1 hour with shaking at 230 rpm. Finally, they were plated on LB agar Petri dishes containing ampicillin (for which pcDNA3.1 hygro β -globin vector confers resistance against) and incubated at 37°C overnight.

3.11 Screening of colonies

In order to identify positive colonies containing the minigene construct, screening colony PCR reactions were performed using published primers [26] designed over β -globin intron 2. Reaction mixes were prepared in 25 μL final volume with 0.5 μL of dNTPs (10 mM each), 5 μL of forward and reverse primers (1 μM), 1X Reaction Buffer with MgCl₂ (Lucigen), 0.25 μL of EconoTaq DNA Polymerase (Lucigen) and water.

Each colony was picked up using a sterile tip, plated in a new LB agar Petri dish containing ampicillin and left for a couple of minutes in a 0.2 ml microtube containing the PCR reaction mix. Tips were then discarded and colony PCR thermocycling conditions included one cycle at 94°C for 12 minutes (in order to disrupt bacterial membranes and release plasmid DNA), then 35 cycles at 94°C for 1 minute, 55°C for 1 minute and 72°C for 1 minute. Final extension was at 72°C for 7 minutes and results were run on 1.5% agarose gel with standard conditions.

3.12 Sequencing of positive colony PCRs

Even though everything was accurately set in order to maximize transformation success, a lot of positive bacterial colonies resulted to carry only the vector without the insert. To verify that construct were as expected in real positive colonies (i.e. with the entire hybrid minigene), samples were first purified from

excess primers and unincorporated dNTPs with Illustra ExoProStar 1-Step (GE Healthcare Life Science) and then prepared for Sanger sequencing with BigDye™ Terminator v3.1 Cycle Sequencing Kit (ThermoFisher) with the same PCR primers. Following BigDye reaction purification using CENTRI-SEP columns (Princeton Separations), samples were run on ABI 3500 Genetic Analyzer with 8 capillaries. This step was also useful since it allowed to correctly identify wild-type (wt) and mutated alleles for each insert. Moreover, although in the laboratory they were used to perform Sanger sequencing after hybrid minigene miniprep from real positives, it was really found convenient, time and cost-effective to directly sequence from colony PCR amplicons. In fact, only a slight increase in electrophoretic background noise was seen and sequences were readable with ease.

3.13 Miniprep

This way, only one wt and one mutated allele per hybrid minigene were furtherly cultured for plasmid miniprep following previous described protocol (see section 3.6 Plasmid miniprep). Finally, DNA quantification was obtained using Nanodrop 2000 (Thermo Fisher Scientific).

3.14 *COL1A1* hybrid minigene mutagenesis

In order to introduce the mutation carried from the patient in the *COL1A1* wt minigene construct, QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent Genomics) was used. A pair of specific primers (see table 3.3) was designed using the QuikChange Primer Design online software (<https://www.agilent.com/store/primerDesignProgram.jsp>). The reaction was set in 51 μL final volume including 1X QuikChange Lightning Buffer, 10 ng wt hybrid minigene, 125 ng for both forward and reverse primer, 1.5 μL QuikSolution reagent, 1 μL dNTPs mix (10 mM each), 1 μL QuikChange Lightning Enzyme and water. Thermocycling conditions were as follows: 1 step at 95°C for 2 minutes, 18 cycles of 95°C for 20 seconds, 60°C for 10 seconds and 68°C for 4 minutes. Final extension was performed at 68°C for 5 minutes.

At this point 2 μL of DpnI enzyme were added to the PCR mix, and digestion was carried out at 37°C for 5 minutes. This step was fundamental to digest the parental (wt) supercoiled dsDNA, since DpnI recognizes and digests only methylated DNA. Next, 2 μL of the digested PCR were used to transform One Shot™ Mach1™ T10 Chemically Competent *E. coli* and later plasmid DNA was extracted following foresaid instructions (see section 3.10 Competent bacteria transformation and section 3.13 Miniprep). Finally, it was Sanger sequenced to confirm the introduction of the desired mutation.

3.15 HEK293 cell transfection

700,000 HEK293 cells were plated in each well of a 6 multiwells plate in order to have them at approximately 90% confluence the following day. Trasfection was possible using Lipofectamine® 2000 Transfection Reagent (Invitrogen) which

exploits lipofection or lipid-based transfection. Its principle is to associate negatively charged nucleic acids with a cationic lipid formulation in order to facilitate the crossing of cell membrane.

First, two different solutions were prepared for each Petri dish with either DNA or Lipofectamine®. Both contained 250 μL of Opti-MEM (a reduced serum medium that positively affects transfection efficiency) but while the former also included 1 μg of hybrid minigene construct, the latter had 1 μL of Lipofectamine®. Unless the volume contribution of DNA aliquot is negligible, it was strongly recommended to concentrate it before. So, once prepared, these were incubated for 2 minutes, mixed together and let stand for another 20 minutes. This mix was then pooled with 1 mL of DMEM (for a total of 1.5mL) and both carefully and uniformly distributed to each Petri dish, having previously removed the old culture medium: this procedure maximized the probability that all cells were reached by DNA-Lipofectamine® complexes. After 5 hours, transfection medium was substituted with 2 mL of culture medium and Petri dishes were incubated o.n. at 37°C.

3.16 Total RNA extraction

The following day, cellular RNA was extracted using TRIzol (Invitrogen) reagent, a monophasic solution at acidic pH of phenol, guanidine isothiocyanate, and other proprietary components that maintain the integrity of nucleic acids but destroy cellular components during sample homogenization. In particular, guanidine isothiocyanate works as a protein denaturant, therefore acting as an effective RNases and DNases inhibitor. As such, TRIzol™ Reagent is an improvement to the single-step RNA isolation method developed by Chomczynski and Sacchi [27].

Firstly, following manufacturer's protocol, cultured cells were washed twice with PBS and detached from plate using 0.05% Trypsin-EDTA. Then, cells were centrifuged twice at 1500 rpm for 5 minutes, discarding supernatant, and resuspending them in 5 mL of 1X PBS. Once these steps were completed, the supernatant was discarded again and pellet was resuspended in 800 μL of TRIzol then freezing tubes for at least one hour at -80°C. 200 μL of chloroform were added to the thawed sample, effectively mixing the resulting solution for 1 minute and centrifuging at 12500 rpm for 15' at 4°C. From now on, homogenates were kept on ice, since at this point there were 3 distinct phases into the tube: a clear upper aqueous layer (containing RNA), a white ring-like interphase and a red lower organic layer (containing DNA and proteins). This is pretty much the same result one can obtain performing phenol-chloroform or organic DNA extraction, apart from the fact that in this last case chances are that RNA would be much more degraded.

Following transfer of the upper phase in another tube, carefully trying not to touch the intermediate DNA ring and thus achieving RNA separation from TRIzol, RNA was then precipitated from the aqueous solution with one volume cold isopropanol and stored at -80°C for at least 30 minutes. After that, samples were thawed at RT and centrifuged for 15' at 12500 rpm and 4°C. This time, the resulting supernatant was discarded and the RNA containing

pellet was resuspended in cold 70% ethanol solution prepared in diethylpyrocarbonate (DEPC) - treated water (which inactivates RNAses). Next, other two washing steps with ethanol were performed, and pellet was dried using Savant spin vacuum at low temperature (i.e. room temperature) for 5 minutes. Finally, it was resuspended in 60 μL of DEPC water and stocked at -80°C for at least 3 hours before being quantified with Nanodrop 2000 (carefully taking into account the slightly lower optical density with respect to DNA).

3.17 RNA retrotranscription to cDNA

Once total RNA was completely extracted and purified from HEK293 cells, in order to evaluate the effect of the supposed splicing mutation on hybrid minigene it was necessary to retrotranscribe it to cDNA as DNA is definitely more stable and easier to work with than RNA. Although this passage could in theory be solved using random primers or gene specific primers, the first ones were preferred because an optimized protocol had already been set in the laboratory.

As such, RNA retrotranscription was carried out using the SuperScriptTM II Reverse Transcriptase kit (Invitrogen). A total of 1 μg of RNA was diluted in 9.5 μL of DEPC water, pooled with 0.5 μL random primers (500 ng/ μL) and 1 μL of dNTPs mix (10 mM each), then incubated at 65°C for 5 minutes and quickly chilled on ice. Remaining steps and technicalities were as suggested by the manufacturer protocol. Finally, cDNA was quantified using Nanodrop 2000.

3.18 Selective PCR and agarose gel

After all cDNA synthesis was completed, it was necessary to selectively amplify the fragments of interest as suddenly performing an agarose gel would have resulted in a smear of bands. Therefore, a new primer pair was designed over flanking 2 and 3 β -globin exons (see table 3.3) in order to amplify all interesting transcripts. This approach has the advantage of reducing costs and time, since only this new primer pair could be used for all hybrid minigenes. PCR mix components were the same as for colony PCR (see section 3.11 Screening of colonies) but 100 ng of cDNA were used as template. Thermocycling conditions were also slightly different with first denaturation step at 94°C for 3 minutes, then 35 cycles at 94°C for 1 minute, 55°C for 1 minute and 72°C for 2 minutes. This time there was not any final extension. In the end, 1.5% agarose gel was prepared and run for 30-40 minutes at 120-130 V to visualize 5 μL of PCR samples.

3.19 Bands extraction and Sanger sequencing

Based on gel results, the patterns of bands obtained from wt and mutated minigenes were evaluated in order to understand the effects of the tested variants on splicing (see sections from 3.5 *COL1A1* NM_000088.3: c.1515G>A p.(=) to 3.9 *OTOG* NM_001277269.1: c.7926C>T p.(=)); after evaluation,

specific bands were selected for sequencing. So, another 2% agarose gel was prepared and loaded with all remaining PCR samples (approx. 45 μ L) for having the most reliable and readily distinguishable bands, that were cut. DNA was then extracted from the gel using QIAquick Gel Extraction Kit (Qiagen) following manufacturer's protocol, quantified and 35 ng were Sanger sequenced as previously described (see section 3.12 Sequencing of positive colony PCR and section 1.7 Purification and Sanger sequencing).

Chapter 4

Results

1 *STRC/pSTRC* and *OTOA/pOTOA* variants sequencing

1.1 Mandelker's approach

The simple idea of doing a lPCR to preferentially amplify only the gene copy of *STRC* and *OTOA* and reduce as much as possible (ideally to zero) their pseudogenes amplification came from two papers [11, 12]. In particular, the starting point was the Mandelker's protocol, which was initially reproduced (with the exception of the enzyme used) to analyse *STRC*. However, although the 20kb lPCR seemed to be easily achievable (see fig. 4.1), a simple check for the co-amplification of *pSTRC*, through a pseudogene specific nPCR, suggested this was not the case. Of course, given the implications that such a finding could have had, this nPCR was Sanger sequenced to confirm it was truly pseudogene-specific (see fig. 4.2). In fact, it seemed that not only there was an equal amplification of both *STRC* and *pSTRC* during lPCR, but also this *pSTRC* co-amplification could not have been entirely due to residual gDNA amplification, as another PCR, with primer amplifying regions outside the lPCR, revealed (see fig. 4.3).

1.2 Confounding variability

Therefore, it was thought that something in the experimental conditions should have been changed, but most of the subsequent lPCR attempts, with either varying annealing-extension temperature or buffer, miserably failed. Later on, it was eventually found that a major and minor determinant for the success of the lPCR amplification were respectively the thermocycler used and the specific well position chosen to load tubes in (see fig. 4.4 and Discussion). In fact, the lPCR failed on Veriti Thermal Cycler (ThermoFischer) but succeeded in both Agilent SureCycler 8800 and Mastercycler nexus (eppendorf). Therefore, from this point on, only the SureCycler 8800 was used for lPCR. Moreover, the *STRC* lPCR annealing-extension temperature was found to be too high to get consistent results over several replicates, thus it was lowered from 68°C to 66°C (see fig. 4.5).

1. STRC/PSTRC AND OTOA/POTOA VARIANTS SEQUENCING

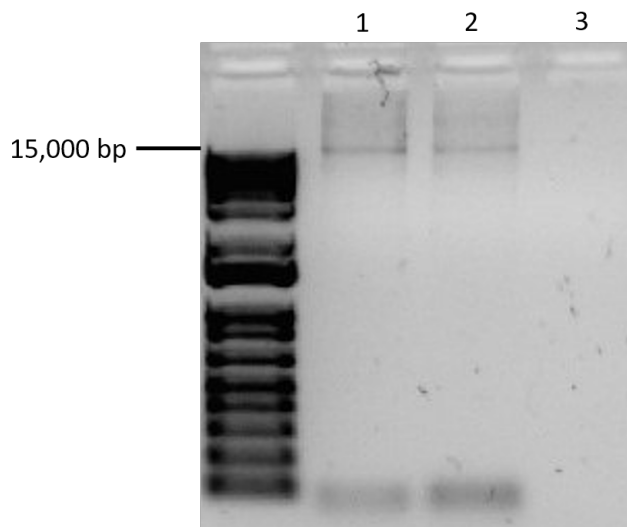


Figure 4.1: Agarose gel of *STRC* IPCR using Mandelker’s approach using two DNA from healthy donors (1. and 2.). The expected length of the amplified product is 20,343. 3. negative PCR control.

```

pSTRCex25      CTGCATCTGCAGTGCTCTGAGGAACAACCTGGAGTTTCTGGCCCACCTCTTTGTA CTGCT
RefSeq_pSTRC   CTGCATCTGCAGTGCTCTGAGGAACAACCTGGAGTTTCTGGCCCACCTCTTTGTA CTGCT
RefSeq_STRC    CTGCATCTCCAGTGCTCTGAGGAACAACCTGGAGTTTCTGGCCCACCTACTTGTACTGCCT
*****
    
```

Figure 4.2: Extract from the sequence alignment showing that *pSTRC* ex25 is indeed specific only for pSTRC amplification.

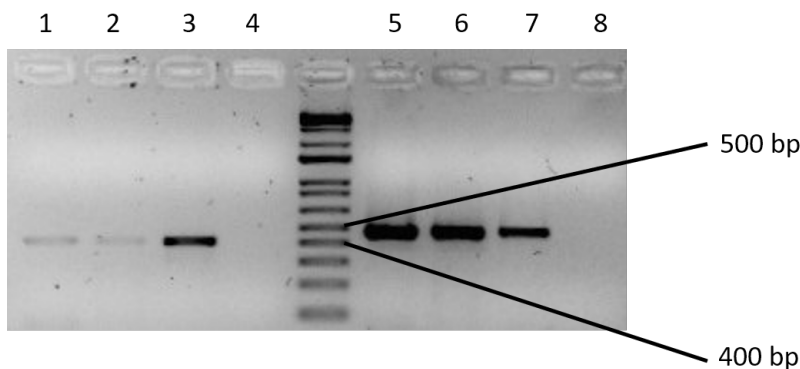


Figure 4.3: Agarose gel showing in 1. and 2. *SPG7* ex6 nPCRs (399 bp) from Mandelker’s STRC IPCRs performed using two DNA from healthy donors. *SPG7* is a gene outside the region amplified in the IPCR therefore is a marker of gDNA template residual. 5. and 6. are *STRC* ex25 nPCRs (483 bp) from Mandelker’s *STRC* IPCRs performed using two DNA from healthy donors. They both show increased amplification compared to genomic template (7.). 4. and 8. are negative PCR controls.

1. STRC/PSTRC AND OTOA/POTOA VARIANTS SEQUENCING

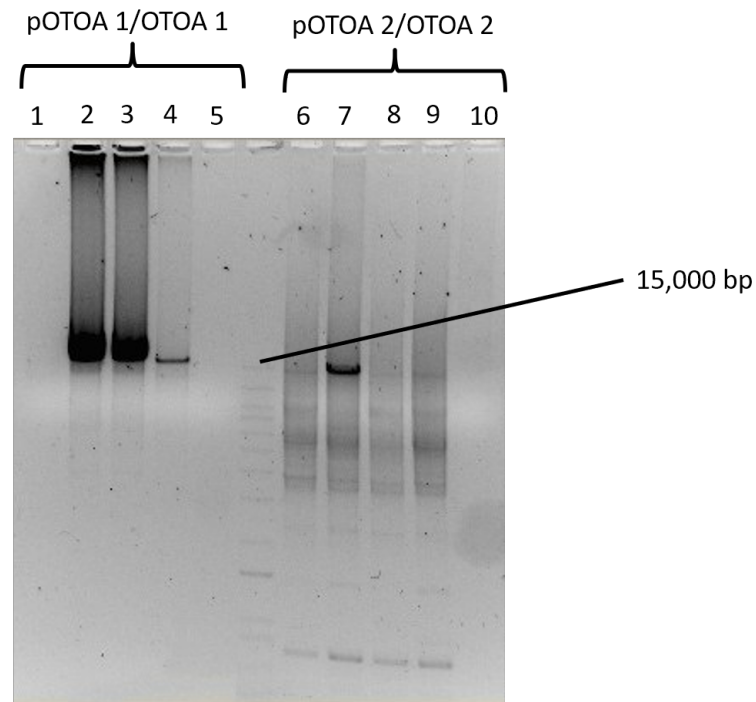


Figure 4.4: Agarose gel showing how the use of few amount of DMSO improved both *OTOA* IPCRs. DMSO concentrations were as follow (1.-10.): 0.8%, 1.2%, 1.6%, 2%, 2%, 0.8%, 1.2%, 1.6%, 2% and 2%. Annealing and extension temperatures were 64°C in all samples. 5. and 10. are negative PCR controls. This image is also a perfect example of how results varied by changing well position. In particular, in 1. the PCR completely failed.

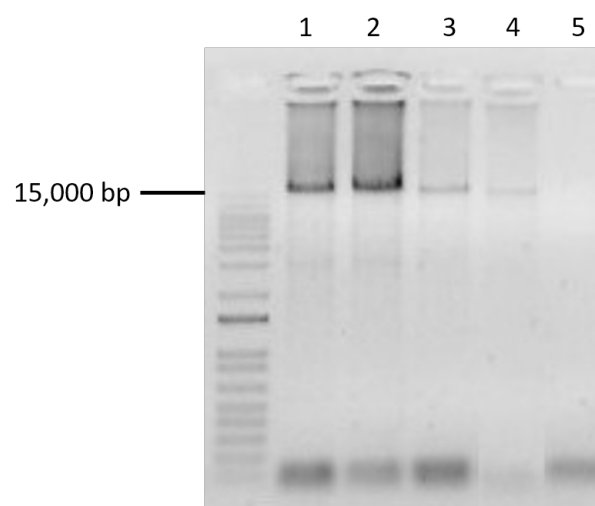


Figure 4.5: Agarose gel showing how *STRC* IPCR amplification changed lowering the annealing and extension temperature from 68°C to 65°C: 1. 65°C, 2. 66°C, 3. 67°C and 4. 68°C. 5. negative PCR control. Clearly, 2. (66°C) gave the best yield of IPCR product.

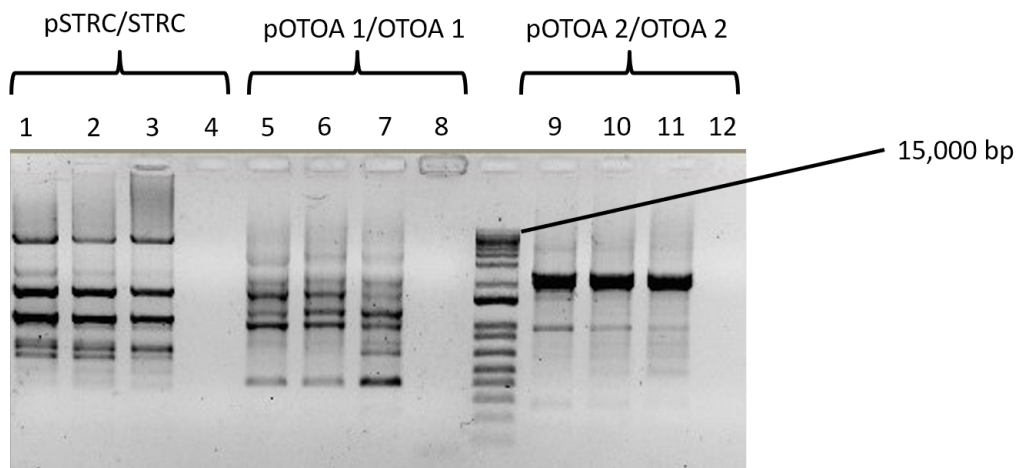


Figure 4.6: Agarose gel showing results from IPCR of *pSTRC/STRC* (1.-4.), *pOTOA 1/OTOA 1* and *pOTOA 2/OTOA 2* using short (approx. 20 bp) primers. Annealing and extension temperatures were as follow (1.-12.): 53°C, 54°C, 55°C, 53°C, 49°C, 50°C, 51°C, 51°C, 48°C, 49°C, 50°C and 50°C. In each IPCR, the expected products (20,323 bp for *STRC*, 15,764 and 12,359 bp for *OTOA 1* and *OTOA 2*, respectively) were not observed. 4., 8. and 12. were negative PCR controls.

1.3 First and second IPCR primers redesign

Meanwhile, owing to the aforementioned findings about Mandelker’s approach, forward and reverse *STRC* IPCR primers were redesigned and *OTOA* ones were designed de novo (see table 3.2). However, this first attempt, with shorter primers expected to hybridize more specifically to the respective genes, turned out to be instead really naïve and results were worse than expected (see fig. 4.6). Fortunately, the complete absence of correct IPCR products (which should have been 20,323 bp long for *STRC* and 15,764 or 12,359 bp long for *OTOA 1* and *OTOA 2*, respectively; see table 3.1) was strongly and rapidly convincing to change approach and not underestimate the task difficulty.

Therefore, a second attempt was made and primers for *OTOA* amplification were re-designed following the criteria presented in section 1.3 PCR primer design; on the other hand, *STRC* primers used by Mandelker et al. were used once again. This time, both IPCRs of *OTOA* gave better results, although with some non-specific products (see fig. 4.7), at the annealing and extension temperature of 64°C using buffer 1. In particular, these non-specific bands were greatly reduced, at least for *OTOA 1*, with the use of 1.6% DMSO (see fig. 4.4). On the contrary, the effect of DMSO for *OTOA 2* IPCR was somehow controversial, but the same amount was still used from then on. It is important to note that in this study the terms “*OTOA 1*” and “*OTOA 2*” are not used to indicate distinct genes, but rather to address the two LR PCRs in which the 32kb-long homologous region shared by *OTOA* and *pOTOA* was divided.

1.4 Searching for a rationale

Even though at this point some progresses had been made and the IPCR products of the expected lengths were obtained, the core of the problem was still

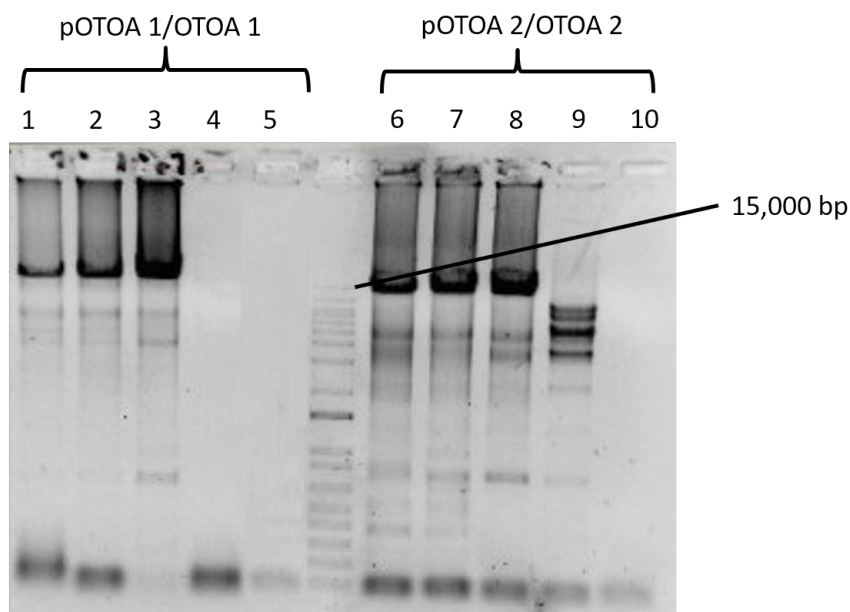


Figure 4.7: Longer IPCR primers improved the specificity of *OTOA* amplification: 1.-5. *pOTOA 1/OTOA 1* and 6.-10. *pOTOA 2/OTOA 2*. Annealing and extension temperatures were as follow (1.-10.): 62°C, 63°C, 64°C, 65°C, 62°C, 62°C, 63°C, 64°C, 65°C and 62°C. 5. and 10. are negative PCR controls.

intact, with both *STRC* and *OTOA* IPCRs not being able to selectively amplify the gene copy. For this reason, a complicated series of scalar dilutions was tried in vain before performing the nPCRs, in an attempt to exploit and maximize possible slight differences in amplification (based on gel images) between each gene and the corresponding pseudogene. Ideally, those differences should have been able then to mask the pseudogene signal in Sanger sequencing. However, this approach turned out to be totally useless since, at any degree of dilution, gene and pseudogene were still equally amplified (see fig. 4.8).

Another tested hypothesis was that if some residual activity of the polymerases blend used for IPCR had been retained at the end of the IPCR itself and up to the beginning of subsequent nPCR, the specificity of this nPCR could have been seriously compromised (for a detailed explanation see Discussion). Unfortunately, this was not the case as no significant polymerase activity was observed after 1 hr from the end of IPCR (see fig. 4.9).

1.5 Tackling homology: third and final IPCR primer redesign

3'-phosphorothioate and non-mismatch-ending primers were respectively used for *STRC* and *OTOA 1/OTOA 2*. Since they were essentially at the same genomic position of previous primers (apart those for *OTOA 2* amplification), the pattern of bands observed in agarose gel was pretty much unchanged. Nevertheless, those new non-mismatch-ending primers reduced drastically the amplification of lower-weight non-specific products for *OTOA 2*. However, two simple nPCRs with non-specific primers (see table 3.2) designed over each

1. STRC/PSTRC AND OTOA/POTOA VARIANTS SEQUENCING

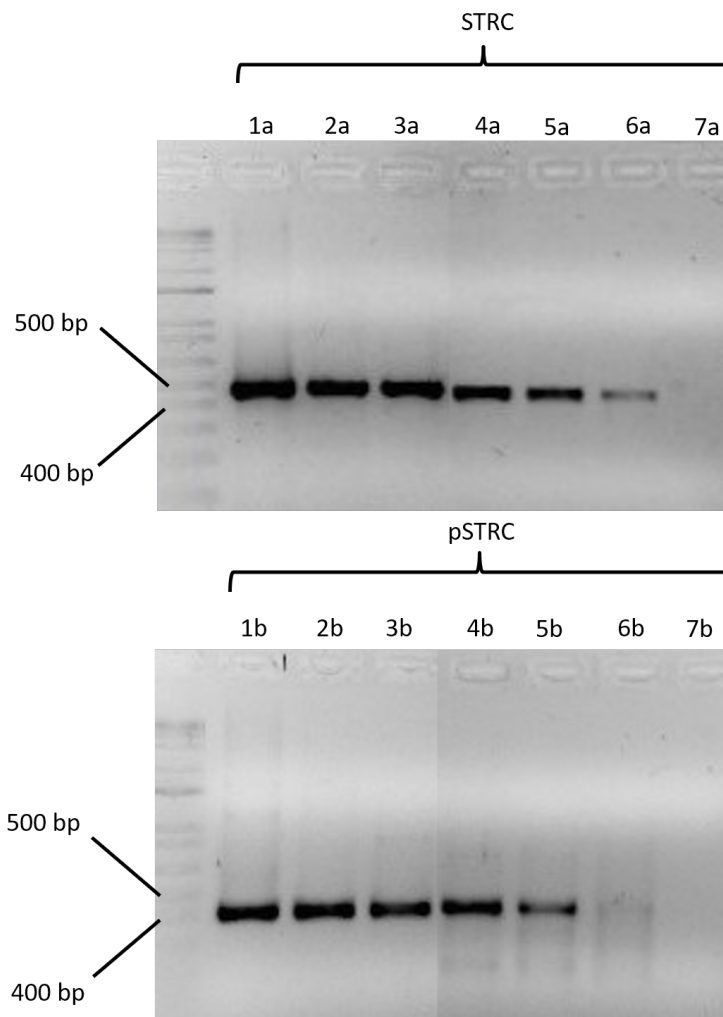


Figure 4.8: Agarose gel showing *STRC* ex25 (1-3.) and *pSTRC* ex25 (1-3.) nPCRs performed on Mandelker's IPCR. 4., 5., and 6. show the amount of nPCR amplification due to residual gDNA that can be expected respectively in 1., 2. and 3. Each lane from a. to b. can be directly compared and shows approximately equal amplification of both templates. 7. is the negative PCR control.

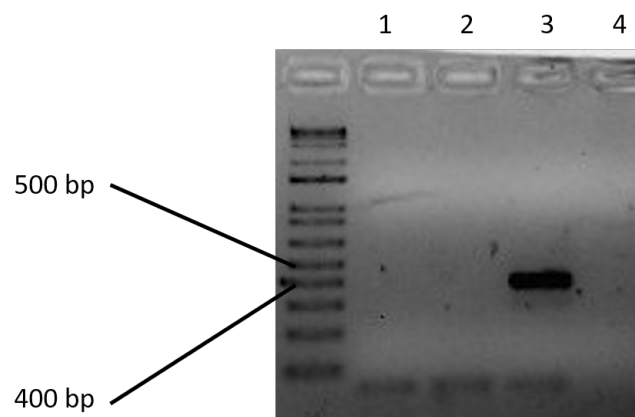


Figure 4.9: Agarose gel showing the absence of nPCR products due to residual IPCR polymerase activity after 1hr of storing samples at -20°C (1.) or 4°C (2.). 3. Adding IPCR polymerase activity to sample stored at -20°C for 1hr resulted in nPCR success. 4. negative PCR control.

1. STRC/PSTRC AND OTOA/POTOA VARIANTS SEQUENCING

```

pOTOA_1/OTOA_1_ex21 ATAAGATCCCCAGCTMTGACCCATGCCTGGTGAGTGTTTYCAGGGTATCTGAGCCAYTG
RefSeq_pOTOA        ATAAGATCCCCAGCTCTGACCCATGCCTGgtgagtgtttccagggatctgagccactg
RefSeq_OTOA         ATAAGATCCCCAGCTATGACCCATGCCTGgtgagtgtttccagggatctgagccattg
*****

pOTOA_1/OTOA_1_ex21 YTGRCATAGTAATTAATGTTTTGGGCAGGGTCCCTRACATCAAGAGGCCTCCTTATGCAG
RefSeq_pOTOA        ttggcatagtaattaatgTTTTGGGCAGGGTCCCTAACAATCAAGAGGCCTCCTTATGCAG
RefSeq_OTOA         ctgacatagtaattaatgTTTTGGGCAGGGTCCCTGACATCAAGAGGCCTCCTTATGCAG
** *****

pOTOA_2/OTOA_2_ex28 TGGTGAGTGGCCTGAGCRCATCGTCTGTGTTGCCCAAGCAGCTGGCCAACRTGTGTAG
RefSeq_pOTOA        TGGTGAGTGGCCTGAGCACATCGTCTGTGTTGCCCAAGCAGCTGGCCAACGTGTGTAG
RefSeq_OTOA         TGGTGAGTGGCCTGAGCGCATCGTCTGTGTTGCCCAAGCAGCTGGCCAACATGTGTAG
*****

pOTOA_2/OTOA_2_ex28 AGACAGGATGCTCCAGATGGTGGGACACCSTTCCCTGGATCCAGACCCTCATCTAGGGCA
RefSeq_pOTOA        AGACAGGATGCTCCAGATGGTGGGACACCGTTCCCTGGATCCAGACCCTCATCTAGGGCA
RefSeq_OTOA         AGACAGGATGCTCCAGATGGTGGGACACCSTTCCCTGGATCCAGACCCTCATCTAGGGCA
*****

```

Figure 4.10: Alignments for *pOTOA 1/OTOA 1* ex21 and *pOTOA 2/OTOA 2* ex28 non-specific nPCR products over long-range amplicons from unmodified IPCR primers.

OTOA long amplicon and amplifying, respectively, exon 21 (*pOTOA 1/OTOA 1*) and 28 (*pOTOA 2/OTOA 2*), revealed that the IPCR was still not able to preferentially amplify only the gene. In fact, as clearly shown in fig. 4.10, for each divergent base expected, these nPCRs had always both the gene- and pseudogene-specific one.

Surprisingly, this was not the case for the 3'-phosphorothioate primers used for *STRC* IPCR. In fact, using a slight different principle than for *OTOA*, two nPCRs, with (this time) specific primers for either *STRC* or *pSTRC*, showed that the two templates were no more equally amplified during IPCR. This could be clearly seen in fig. 4.11 as opposed to previous attempts (see fig. 4.8). However, the effect of 3'-phosphorothioate primers was not all-or-nothing as ideally desirable, but indeed allowed for some *pSTRC* amplification to go on, though in a reduced way. Therefore, there still was an important question to be answered: whether or not this biased amplification was already sufficient to obtain non contaminated Sanger sequences of *STRC*. This was easily cleared up by direct sequencing of non-specific nPCRs products amplifying *STRC/pSTRC* intron 18 compared to previous results (see fig. 4.12).

1.6 IPCR time and cost optimization

Meantime, further optimization of IPCR was carried on with the aim of reducing the time and cost of the technique as much as possible. In fact, the original Mandelker's protocol had taken almost 8 hr to be completed, while reducing by 8 cycles or 22% the IPCR allowed to get the same accurate results in 6 hr, thus reducing TAT. Moreover, this had also two positive effects. The former was to have drastically lowered the probability of chimeric products formation during IPCR, which is greater during the plateau phase, i.e. when amplicons can eventually compete with primers [28, 29]. The latter was to have further reduced the amount of non-specific products, thus improving gel and sequencing results.

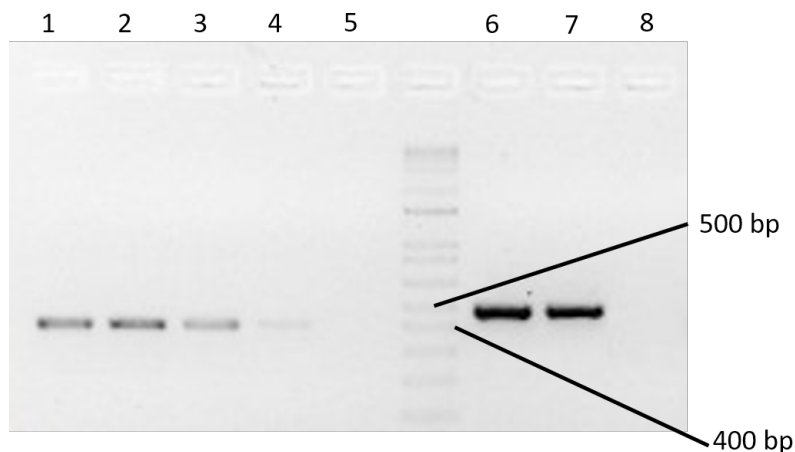


Figure 4.11: Agarose gel showing nPCRs of *pSTRC* ex 25 (1.-2.) and *STRC* ex 25 (6.-8.) performed on 3'-phosphothioate *STRC* IPCR. 3. and 4. show the amount of *pSTRC* ex 25 amplification due to residual gDNA that can be expected respectively in 1. and 2. . 5. and 8. are negative PCR controls.

On the other hand, in order to reduce costs, it was checked how many DNA polymerases blend enzymatic units were really necessary. It was found that 2 U were sufficient to get optimal results, achieving a 33% save in contrast to Mandelker's protocol. Finally, also the DNA template quantity was optimized to 100 ng.

1.7 Validating NGS variants: before and after

Having obtained the preferential amplification of the gene (both *STRC* and *OTOA*) using 3'-phosphorothioate IPCR primers, each variant identified by NGS could be confirmed or excluded performing the amplification of the target region with nPCR followed by Sanger sequencing as previously explained. The results obtained following both the approach described by Mandelker et al. and the one developed in the present study are shown in the following tables (4.1 and 4.2).

In most cases, the chromatograms obtained following the IPCR protocol described by Mandelker et al. were the sum of the chromatograms obtained after the IPCR protocol described in the present study for the selective amplification of, respectively, *STRC* gene and its pseudogene (see fig. 4.13 and 4.14).

This could not be seen for *OTOA*, since specific 3'-phosphorothioate primers of its pseudogene were not bought. Anyway, the nPCR and Sanger sequencing of specific regions containing different nucleotides in the gene compared to the pseudogene, allowed to clearly demonstrate the selective amplification of the gene for most IPCR (see table 3.2 and fig. 4.15 and 4.16).

Finally, it is remarkable to note also that, although some nPCR performed invariably well in SureCycler 8800 and Veriti Thermal Cycler, most of *STRC*-related and *OTOA*-related nPCRs were successful only in the Veriti Thermal Cycler or SureCycler 8800, respectively.

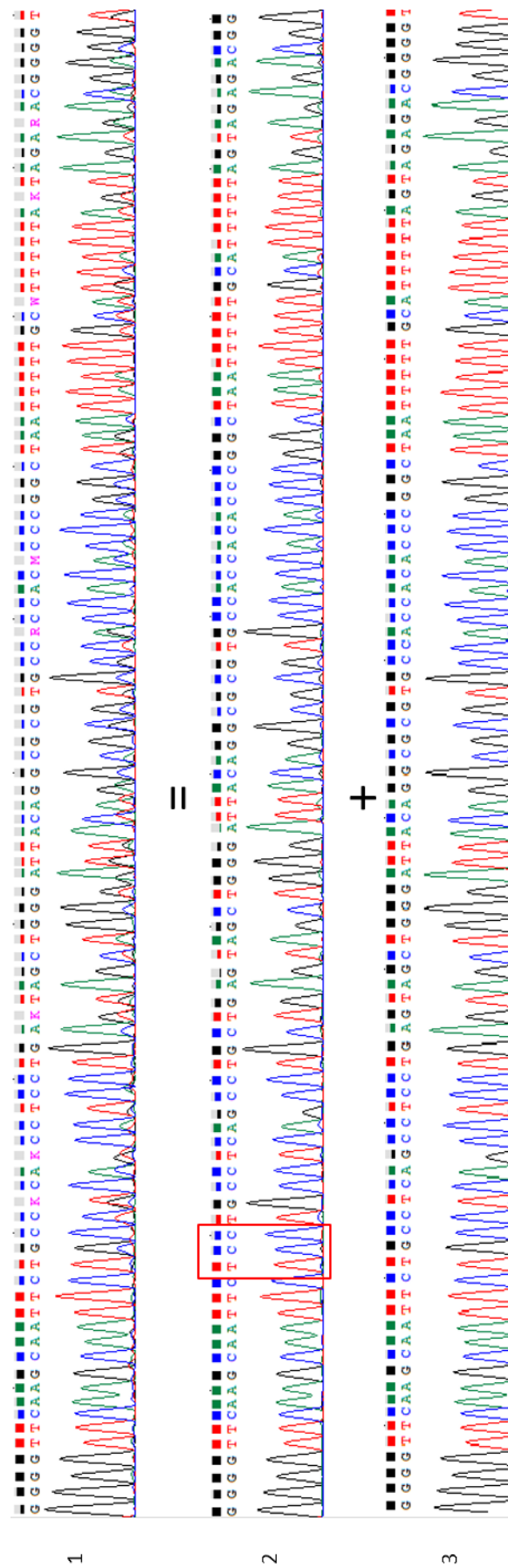


Figure 4.12: Sanger chromatograms showing a region from the *STRC/pSTRC* 18 intron. 1., 2. and 3. are the result of a nPCR performed respectively on Mandelker's IPCR, 3'-phosphorothioate *STRC* IPCR and 3'-phosphorothioate *pSTRC* IPCR. The red rectangle highlights a 3nt *STRC* specific insertion.

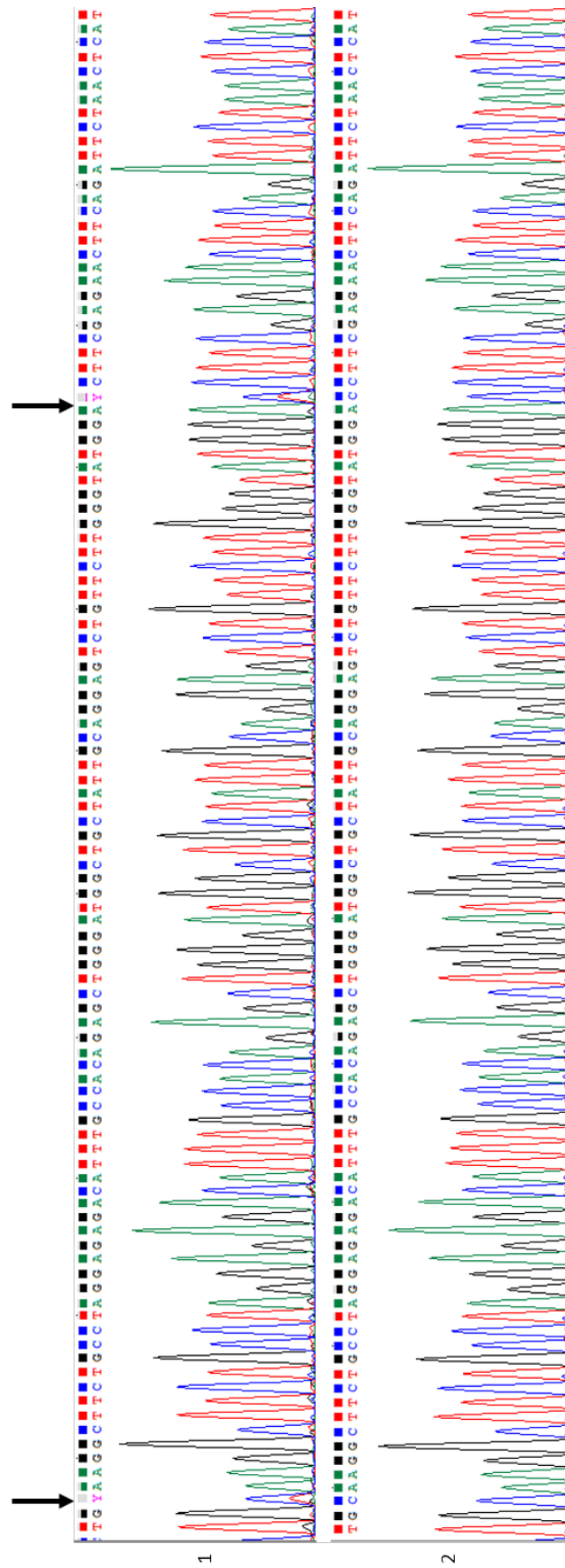


Figure 4.13: Sanger chromatograms showing a region from the *STRC/pSTRC* ex 20. 1. and 2. are the results of a nPCR performed on 3'-phosphorothioate *pSTRC* IPCR and 3'-phosphorothioate *STRC* IPCR. The arrows highlight were 1. shows co-amplification of *STRC*. This is an example of IPCR failed quality control.

1. STRC/PSTRC AND OTOA/POTOA VARIANTS SEQUENCING



Figure 4.14: Sanger chromatograms showing a region from the *STRC/pSTRC* ex 7. 1., 2. and 3. are the result of a nPCR performed respectively on Mandelker's IP-CR, 3'-phosphorothioate *STRC* IP-CR and 3'-phosphorothioate *pSTRC* IP-CR. The arrows highlight the NM_153700.2: c.2356delC *STRC* specific variant in heterozygosis.

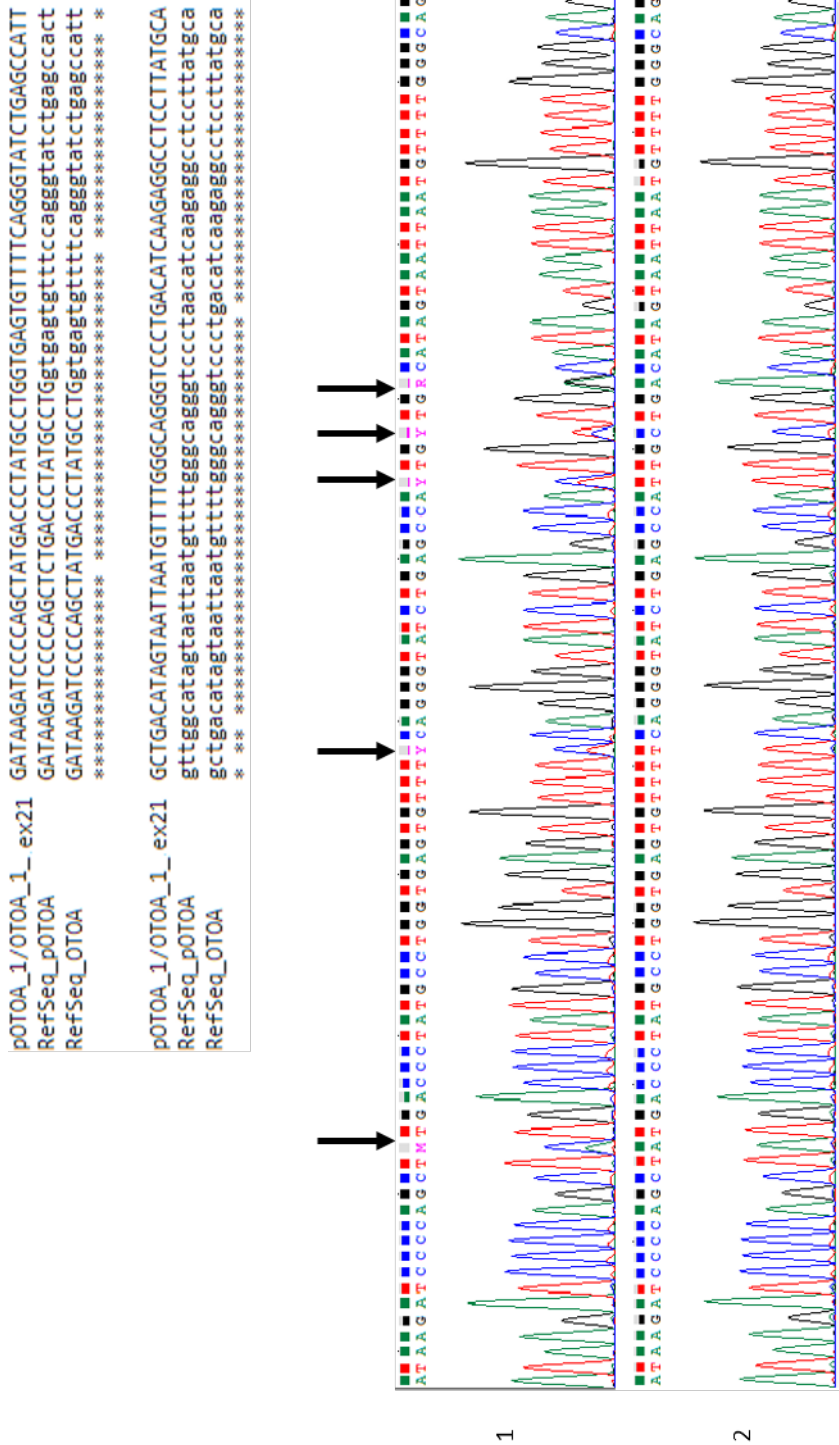


Figure 4.15: Sanger sequences from *pOTOA 1/OTOA 1* ex21. The specificity of the nPCR changed depending on which type of IPCR primers were used: unmodified ones (1.) or 3'-phosphorothioate ones (2.). Arrows show gene-pseudogene divergent bases. Alignment sequences refer to 2. .

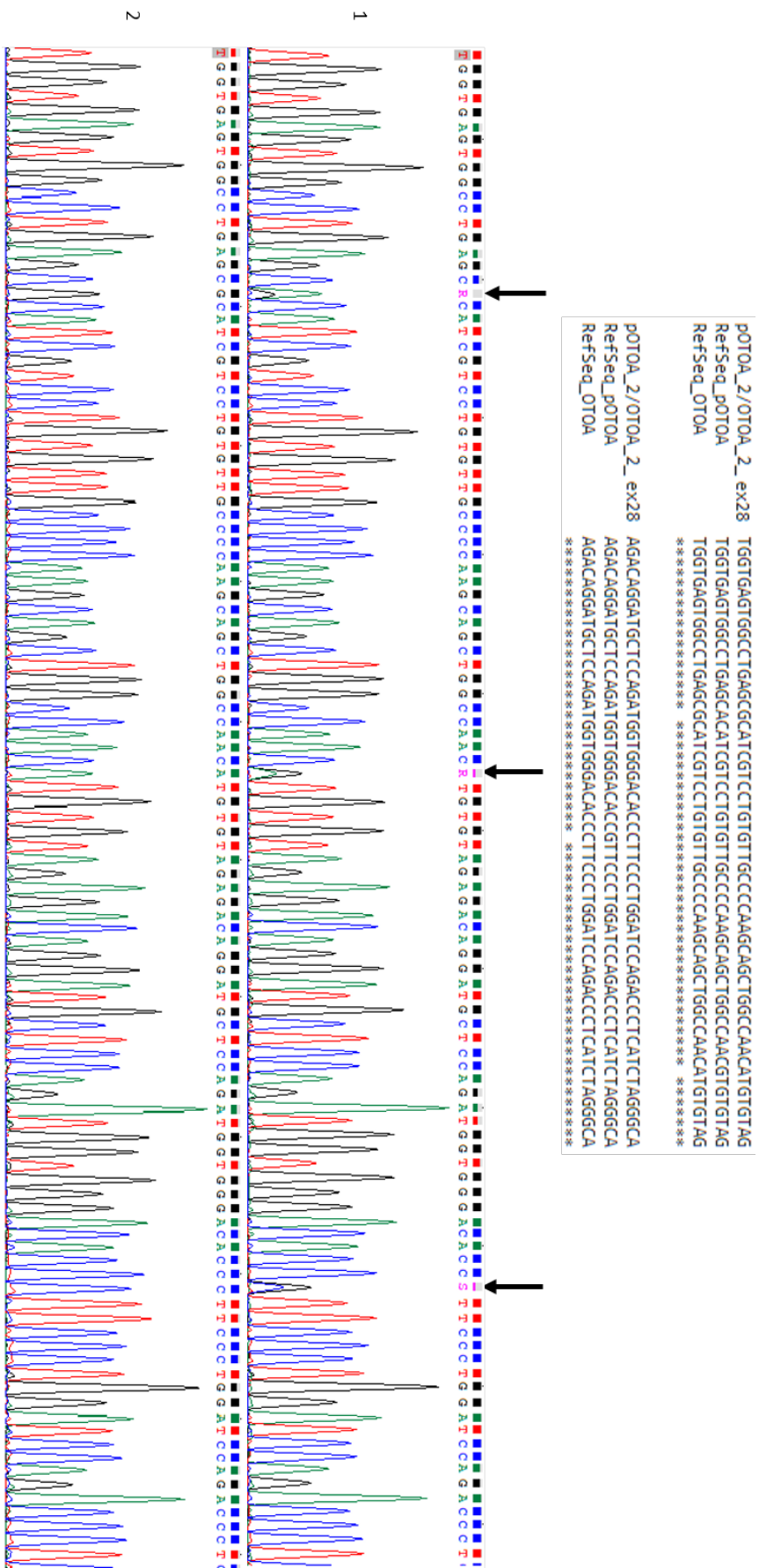


Figure 4.16: Sanger sequences from *pOTOA 2/OTOA 2* ex28. The specificity of the nPCR changed depending on which type of PCR primers were used: unmodified ones (1.) or 3'-phosphorothioate ones (2.). Arrows show gene-pseudogene divergent bases. Alignment sequences refer to 2. .

2 CNVs assessment

2.1 A statistical model for CNV detection

After the sequencing of all *STRC/pSTRC* variants, a simple statistical analysis for CNV detection (particularly deletions) was conducted across all their respective NGS data as described previously (see section in Materials and methods 2.1 A statistical model for CNV detection). All samples were thus classified into three groups based on the variant being predicted as heterozygous, hemizygous (i.e. in trans with a deletion) or homozygous/double heterozygous (see table 4.3). The same three variants (NM_153700.2: c.1631A>G, c.2914C>T and c.3947C>T) out of 16 were categorized as emizygous by different approaches tested, while the number of heterozygotes and homozygotes/double heterozygotes were changing based on the test sensitivity for lower or higher than expected mean allele frequencies. Moreover, in each test, some samples were not categorized at all; either because they were outside confidence intervals or in between overlapping maximum difference intervals. However, given the low incidence of duplications and the decreased sensitivity and specificity that this approach would have had, this statistical model can be used only for discovering deletions.

2.2 SureCall pair analysis

In order to prove additional evidence to the previous predictions about CNVs, all patients analysed for *STRC/pSTRC* were also checked with the SureCall pair analysis tool and results are presented in the following fig. 4.17. Interestingly, two of the three deletions (NM_153700.2: c.1631A>G and c.2914C>T) found with the statistical model were actually confirmed, but the software identified also another hemizygous variant (NM_153700.2: c.2356delC). For these positive samples, the analysis was repeated in parents (see fig. 4.17), further confirming these conclusions. Therefore, in the end, six patients (three probands + one parent each) were classified as carriers of a multiexonic (possible genic) heterozygous deletion in *STRC*. On the contrary, all the other variants (including the NM_153700.2: c.3947C>T variant previously classified as hemizygous) were not predicted to show any CNVs (both deletions and duplications).

However, because in all samples at least two or three different exons were reported as being duplicated (red) or deleted (blue), it is clear that some precautions should be taken to cope with this sensitivity like the concordance of signals (duplication or deletion) across the gene exons, the number of marked exons (the more the better), their distribution and score (the darker the better). Nevertheless, it should be noted as there are some exons that are specifically marked only in real deleted cases (see fig. 4.17).

2.3 Trios analysis

When the previously described method (with IPCR, nPCR and Sanger sequencing) resulted in variants being classified as (apparently) homozygous in

2. CNVS ASSESSMENT

Table 4.3: Allele frequencies calculated from NGS coverages for 16 variants found in *STRC/pSTRC*. Predicted genotypes from each statistical model tested are shown as green (homozygosis/double heterozygosis), orange (heterozygosis), yellow (hemizygosis) or grey (not classified). Show in red there is a mispredicted variant (c.2356delC) seen as hemizygous from SureCall pair analysis tool. All variants are referred to NM_153700.2.

Patient ID	variant	STRC allele freq.	pSTRC allele freq.	mean allele freq.	95% conf. int.	99% conf. int.	max diff. from mean	SureCall pair analysis
13554	c.1631A>G	0,316	0,319	0,318				mutiexonic deletion
14352	c.1108C>G	0,237	0,283	0,260				wt
15115	c.1027C>T	0,244	0,203	0,223				wt
15320	c.2172T>C	0,206	0,234	0,221				wt
15564	c.3988G>A	0,485	0,440	0,467				wt
15861	c.1873C>T	0,201	0,207	0,204				wt
16020	c.2914C>T	0,334	0,320	0,327				mutiexonic deletion
16284	c.1065C>T	0,164	0,165	0,165				wt
16356	c.2779C>T	0,172	0,206	0,189				wt
16428	c.2356delC	0,255	0,231	0,242				mutiexonic deletion
17326	c.3947C>T	0,435	0,201	0,366				wt
17990	c.2356delC	0,248	0,243	0,245				wt
18403	c.4779G>A	0,239	0,265	0,251				wt
18654	c.2914C>T	0,242	0,277	0,261				wt
19101	c.1873C>T	0,248	0,248	0,248				wt
19101	c.5252C>A	0,232	0,251	0,241				wt

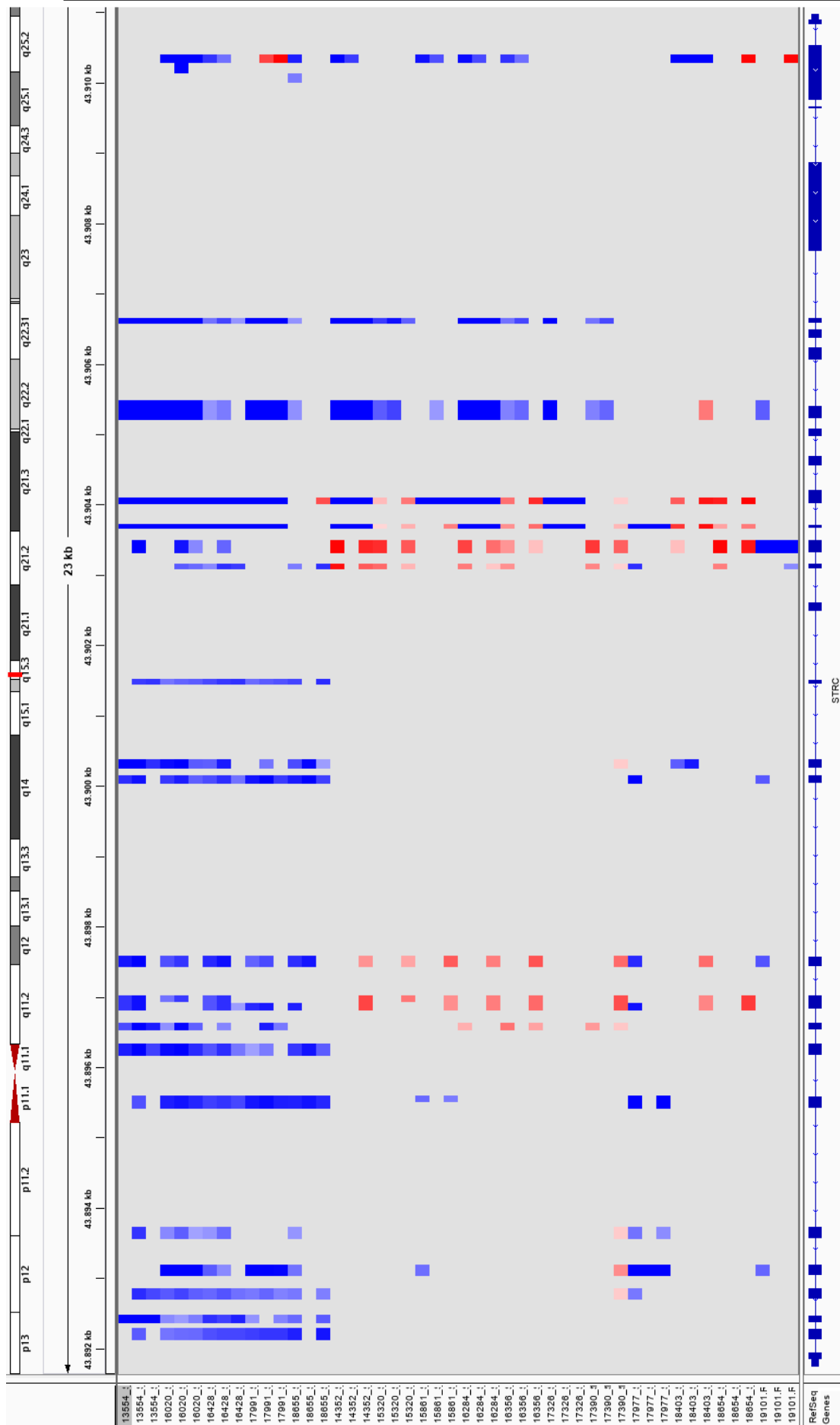


Figure 4.17: Screenshot from SureCall pair analysis tool showing the predicted CNVs in triplicates in the *STRC* locus, based on NGS coverages. Patients that were considered hemizygous for the variant are aligned at the top. Blue exons mean duplications while red exons mean deletions.

3. HYBRID MINIGENE ASSAYS

Table 4.4: List of prediction results for each variant from different online software. As it can be seen, *COL1A1* c.1515G>A variant was the only one that had been predicted by all software to cause a possible splicing alteration.

Gene and exon	Refseq and variant	Human Splicing Finder	NetGene2	NNSplice
OTOG ex 48	NM_001277269.1: c.7926C>T p.(=)	no effect predicted	slightly (2%) higher score for the donor splice site	no effect predicted
MYO15A ex 14-15	NM_016239.3: c.4779+9G>A	new ESS	no effect predicted	no effect predicted
COL1A1 ex 21-23	NM_000088.3: c.1515G>A p.(=)	alteration of ESE and activation of exonic cryptic acceptor splice site	71% lower score for the donor splice site (= loss)	loss of the donor splice site
COL2A1 ex 24-26	NM_001844.5: c.1734+3A>G	new ESS	loss of the donor splice site	20% lower score for the donor splice site
COL11A2 ex 19-21	NM_080680.2: c.1819-5T>C	alteration of ESE	slightly (2%) higher score for the acceptor splice site	no effect predicted

STRC or *OTOA* and SureCall pair analysis showed a possible CNV (involving at least the exon with the mutation in trans), it was possible to indirectly confirm these findings repeating all the analyses in the proband's parents (like in the aforementioned example with SureCall). If, for instance, those results were compatible with a possible deletion (i.e. one parent is apparently wt and the other is heterozygote for the same variant of the proband), one could then conclude with high accuracy that the proband is a true hemizygote rather than a homozygote. Actually, this was the case for all *STRC* suspected cases analysed (see table 4.1).

3 Hybrid minigene assays

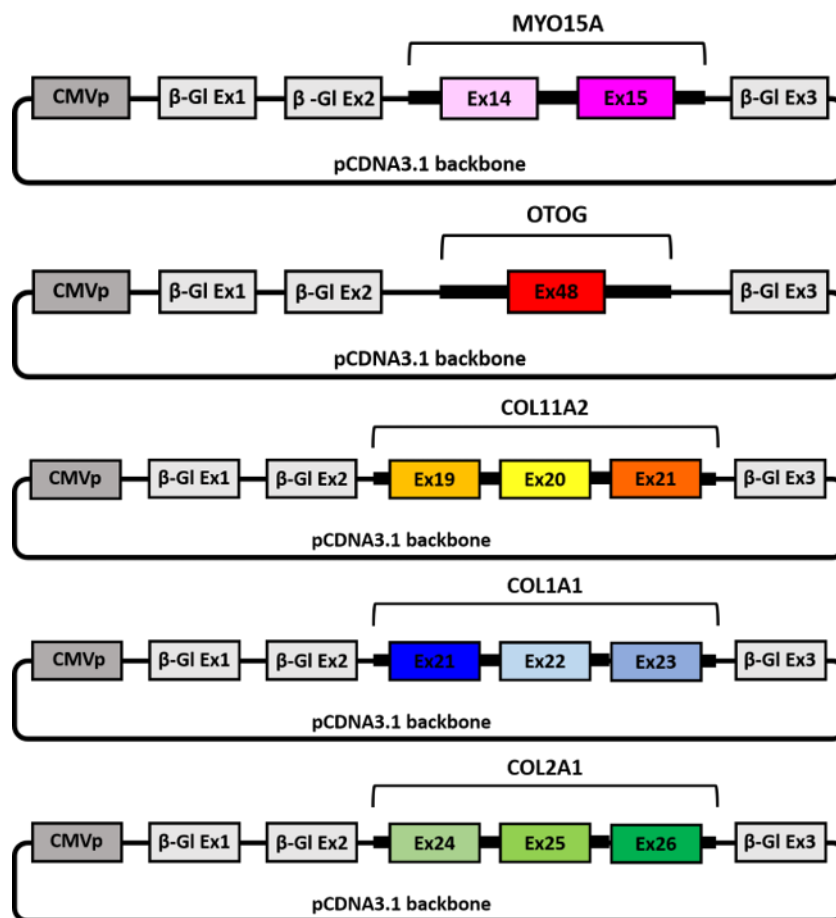
3.1 Bioinformatic analyses

Bioinformatic analyses were conducted to predict the effect on splicing, if any, of the tested variants. Software used includes Human Splicing Finder, NetGene2 and NNSplice; their characteristics were presented in the materials and methods section. Results were then summarized in table 4.4.

3.2 Hybrid minigene constructs

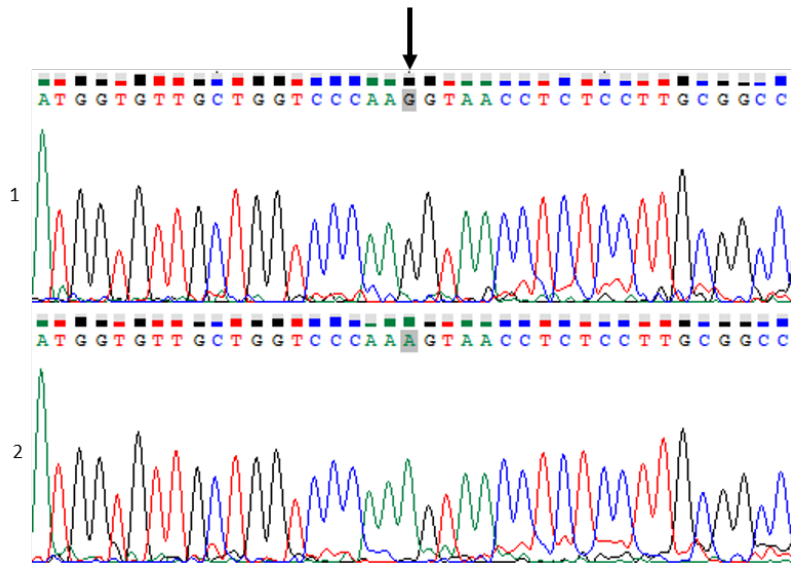
Depending on the particular primer pair used (see table 3.3), inserts contained at least the nearest exon for each variant and its flanking 150-200 intronic nucleotides. In some cases, to preserve intron-exon boundaries, this approach resulted in the inclusion of more exons than only the nearest one. Inserts were cloned into pcDNA3.1 hygrob β -globin vectors forming hybrid minigenes. A schematic representation of all hybrid minigene constructs can be seen in fig. 4.18.

Figure 4.18: Schematic representation of all different hybrid minigene constructs. The length of each insert is reported in table 3.3. The figure is not in scale.



3. HYBRID MINIGENE ASSAYS

Figure 4.19: Sanger sequencing of wt (1) and mutated (2) *COL1A1* hybrid minigene constructs. The point mutation of the c.1515G>A variant is highlighted by the arrow.



3.3 *COL1A1* hybrid minigene mutagenesis

The mutagenesis of *COL1A1* hybrid minigene was required to introduce the variant under study (NM_000088.3: c.1515G>A p.(=)), since the patient's DNA was not available. The primers used are listed in table 3.3 and the mutagenesis was confirmed by Sanger sequencing (see fig. 4.19). In the other cases this step was not needed because all DNA samples were heterozygous for the studied variants, thus providing both the mutant and control alleles.

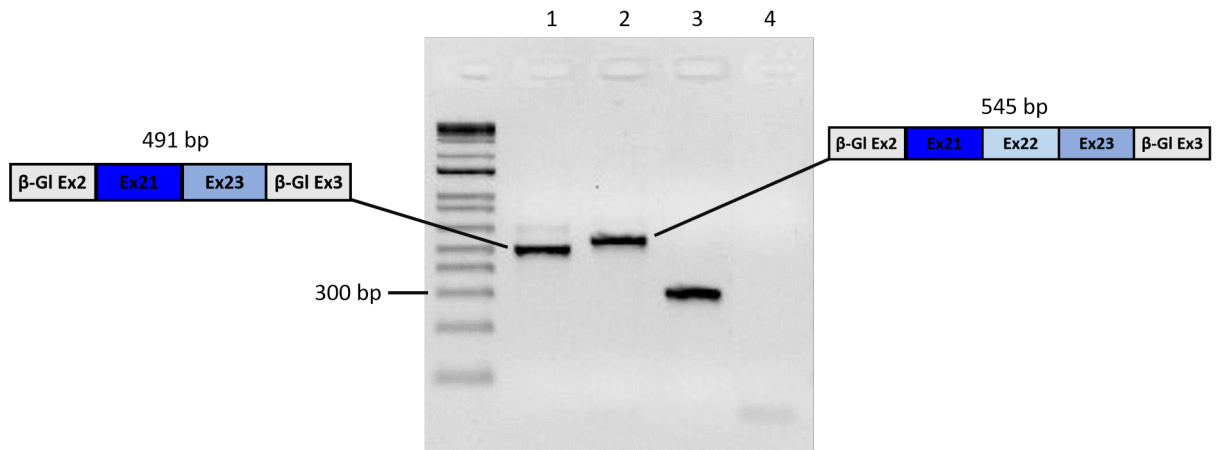
3.4 cDNA analysis

HEK293 cells were transfected with pcDNA3.1 hygro β -globin empty vector and both wt and mutated hybrid minigene constructs. RNA was extracted 24h after the transfection and it was retrotranscribed to cDNA. The subpopulation of cDNA fragments of interest was amplified using betaglobin specific primers (β -globin ex 2-3, see table 3.3). As expected, no evidence of expression of the endogenous betaglobin gene was observed (see fig. 5.2, [26]). Finally, the amplification products were run on 2% agarose gel and the specific results are presented below.

3.5 *COL1A1* NM_000088.3: c.1515G>A p.(=)

The *COL1A1* c.1515G>A variant was predicted to potentially alter an ESE and activate an exonic cryptic acceptor site by Human Splicing Finder and indeed the other two software did find a potential disruption of a donor splice site. Given the size of insert and the relative position of primers over 2 and 3 β -globin exons, a correctly spliced transcript would have been 545 bp long. As it can be seen in the figure 4.20, both the wt and the mutated hybrid minigenes

Figure 4.20: Agarose gel of the resulting *COL1A1* cDNA amplification. 1. mutated hybrid minigene; 2. wt hybrid minigene; 3. empty vector (amplifying only exons 2 and 3 of β -globin) and 4. negative PCR control. As it can be seen, the c.1515G>A variant tested alters the splicing pattern.



gave rise to two transcripts (one of them is really faint) of the approximate size of 500-650 bp, clearly distinguishable from the empty vector.

The size of the bands obtained by wt and mutated hybrid minigenes clearly differed, thus suggesting a splicing alteration due to the tested variant. Sanger sequencing of the lower bands confirmed this hypothesis and revealed that there was a skipping of *COL1A1* exon 22 (54 bp) in the mutated minigene compared to the 545 bp correctly spliced transcript in the wt one.

Faint bands were also sequenced, but they were contaminated by the lower ones, so no conclusion could be done. However, it seems reasonable to speculate that they would have been the same transcript if it was taken into account the 54 bp expected exon skip discussed above. As such, this is supported from two facts: the presence of the faint band also in the wt minigene and the lower size of that faint band in mutated minigene.

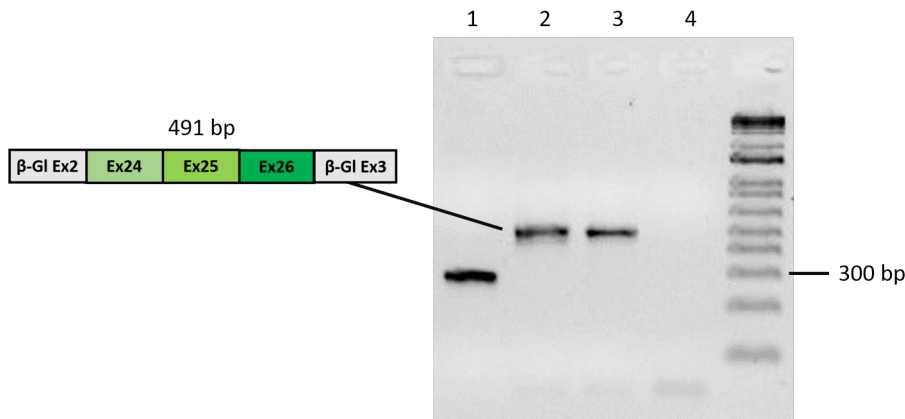
In conclusion, these data suggest a pathogenic role for this variant, although further study (for example transfection of osteosarcoma cells) may be desirable in order to accumulate more evidence.

3.6 *COL2A1* NM_001844.5: c.1734+3A>G

The *COL2A1* c.1734+3A>G variant was predicted to potentially create a new ESS site by Human Splicing Finder. Given the size of insert and the relative position of primers over 2 and 3 β -globin exons, a correctly spliced transcript would have been 491 bp long. As it can be seen in the figure 4.21, both the wt and the mutated hybrid minigene gave rise to only one transcript of the approximate size of 500 bp, clearly distinguishable from the empty vector. Sanger sequencing of the band revealed that it was indeed the same 491 bp correctly spliced transcript. Therefore, these data suggest the absence of a pathogenic role for this variant, although further, more specific study are needed.

3. HYBRID MINIGENE ASSAYS

Figure 4.21: Agarose gel of the resulting *COL2A1* cDNA amplification. 1. empty vector (amplifying only exons 2 and 3 of β -globin); 2. mutated hybrid minigene; 3. wt hybrid minigene and 4. negative PCR control. As it can be seen, the c.1734+3A>G variant tested does not alter the splicing pattern.



3.7 *COL11A2* NM_080680.2: c.1819-5T>C

The *COL11A2* c.1819-5T>C variant was predicted to potentially alter an ESE site by Human Splicing Finder, although the other two software did predict slight to any effect. Given the size of insert and the relative position of primers over 2 and 3 β -globin exons, a correctly spliced transcript would have been 437 bp long, while a potential alternative spliced transcript for any of the three *COL11A2* exons would have resulted in a 45-54 bp shortening. As it can be seen in the figure 4.22, both the wt and the mutated hybrid minigene gave rise to two transcripts of the approximate size of 400 and 450 bp, clearly distinguishable from the empty vector.

Sanger sequencing of the upper bands revealed that they were indeed the 437 bp correctly spliced transcript. In contrast, the lower bands were actually a 383 bp alternative transcript which skipped *COL11A2* exon 21. Even though this transcript may be seen as the effect of the mutation, this hypothesis is completely ruled out by the presence of the same transcript in wt minigene. Therefore, these data do not support a pathogenic role for this variant.

3.8 *MYO15A* NM_016239.3: c.4779+9G>A

The *MYO15A* c.4779+9G>A variant was predicted to potentially create a new ESS site by Human Splicing Finder, although the other two software did not predict any effect. Given the size of insert and the relative position of primers over 2 and 3 β -globin exons, a correctly spliced transcript would have been 467 bp long, while a potential alternative spliced transcript for any of the two *MYO15A* exons would have resulted in a 59-124 bp shortening. As it can be seen in the figure 4.23, both the wt and the mutated hybrid minigene gave rise to three transcripts of the approximate size of 300, 350 and 450 bp.

The longer transcripts were clearly distinguishable from the empty vector, but the shortest one seemed to be at the same position of this latter. In fact, Sanger sequencing of the that last band confirmed this suspect, since

Figure 4.22: Agarose gel of the resulting *COL11A2* cDNA amplification. 1. empty vector (amplifying only exons 2 and 3 of β -globin); 2. mutated hybrid minigene; 3. wt hybrid minigene and 4. negative PCR control. As it can be seen, the c.1819-5T>C variant tested does not alter the splicing pattern.

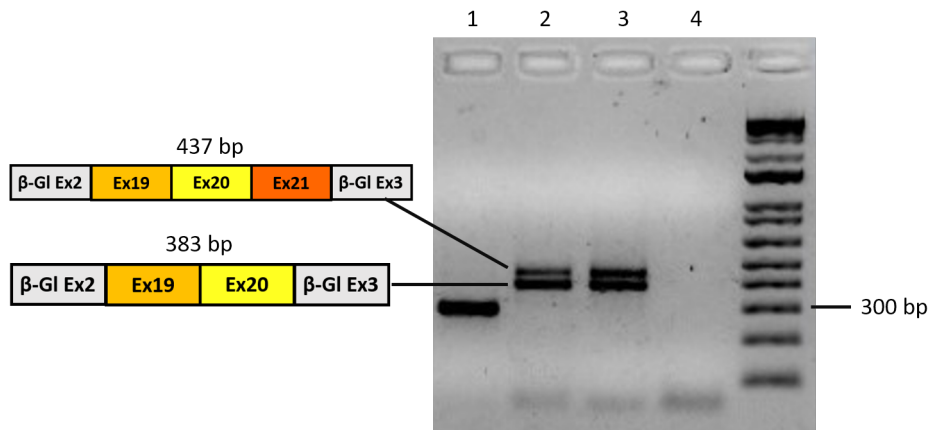
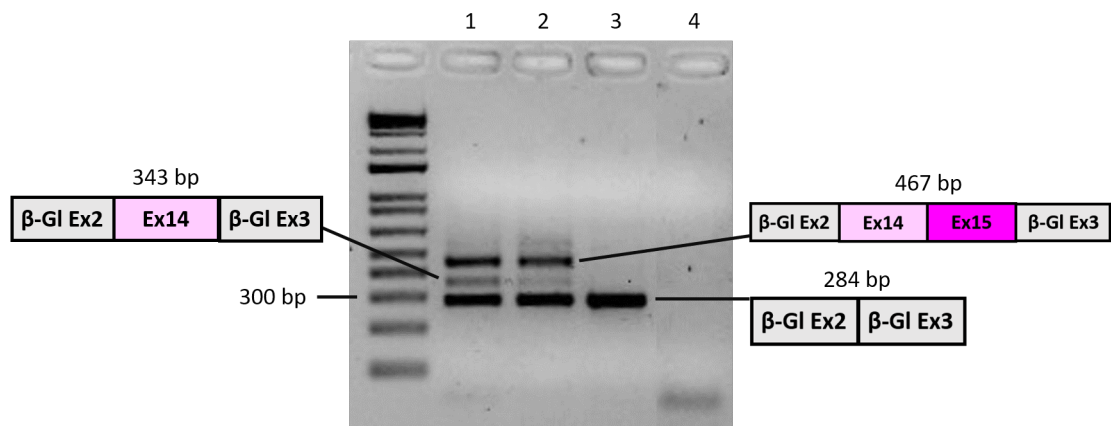
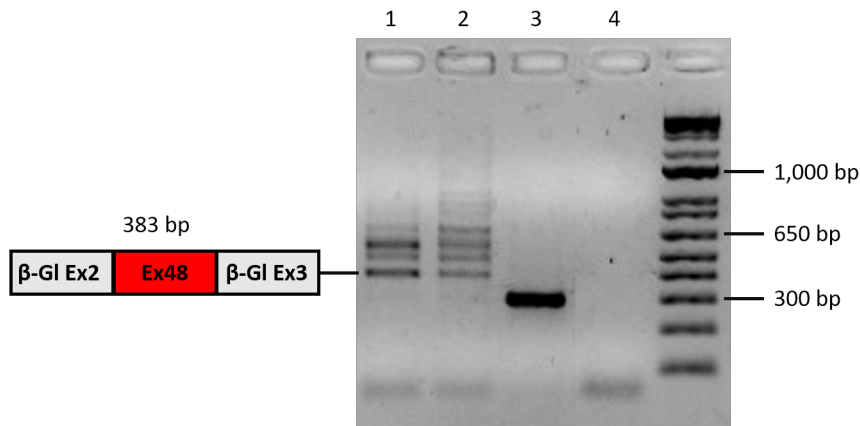


Figure 4.23: Agarose gel of the resulting *MYO15A* cDNA amplification. 1. mutated hybrid minigene; 2. wt hybrid minigene; 3. empty vector (amplifying only exons 2 and 3 of β -globin) and 4. negative PCR control. As it can be seen, the c.4779+9G>A variant tested does not alter the splicing pattern.



3. HYBRID MINIGENE ASSAYS

Figure 4.24: Agarose gel of the resulting *OTOG* cDNA amplification. 1. mutated hybrid minigene; 2. wt hybrid minigene; 3. empty vector (amplifying only exons 2 and 3 of β -globin) and 4. negative PCR control. As it can be seen, the c.7926C>T variant tested does not alter the splicing pattern, although not all transcripts were successfully sequenced.



the transcript included only betaglobin exons. Moreover, while sequencing of the upper bands revealed that they were indeed the 467 bp correctly spliced transcript, the middle ones came out as a 343 bp alternative transcript with skipping of *MYO15A* exon 15. Therefore, these data suggest, as a whole, the absence of a pathogenic role for this variant, although further study may really be desirable, as the shortest band could probably be a clear sign of insert misrecognition from HEK293 cells.

3.9 *OTOG* NM_001277269.1: c.7926C>T p.(=)

The *OTOG* c.7926C>T variant was one of a kind, since it was predicted to practically have no effect by all the software. However, algorithms can have flaws and an in vitro assay to confirm or not the prediction is always recommended. Given the size of insert and the relative position of primers over 2 and 3 β -globin exons, a correctly spliced transcript would have been 383 bp long. As it can be seen in the figure 4.24, both the wt and the mutated hybrid minigene gave rise to many transcripts comprised in between almost 400 and 1000 bp.

Although it was not expected to obtain all these transcripts because the insert contained only exon 48 of *OTOG*, from a simple visual check it seemed that all of them were in common between the wt and mutated hybrid minigene. Nevertheless, none of the bands were aligned with empty vector, at least excluding the possibility of exon skipping. Furthermore, this time no attempt was carried on in cutting gel bands and sequencing them all, since not only they were very close to each other, but also in previous minigenes contamination from less close band was indeed obtained. The only exception was the shortest band, which was more clearly separated and thus was sequenced. It came up to be the 383 bp correctly spliced transcript.

Therefore, all other transcripts should have retained some intronic regions and this was confirmed with a more detailed bioinformatic analysis with Hu-

Figure 4.25: Capture of Human Splicing Finder prediction results for the *OTOG* c.7926C>T variant. As it can be seen, after the end of exon 48, there is a hotspot (circled in red) of both donor and acceptor splice sites that are above the threshold (in yellow).

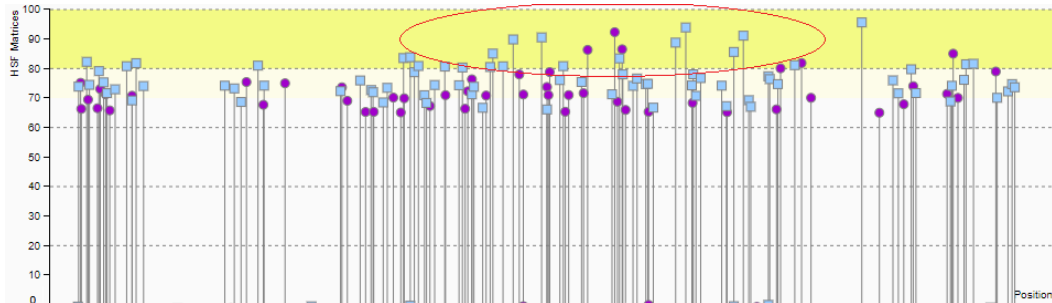
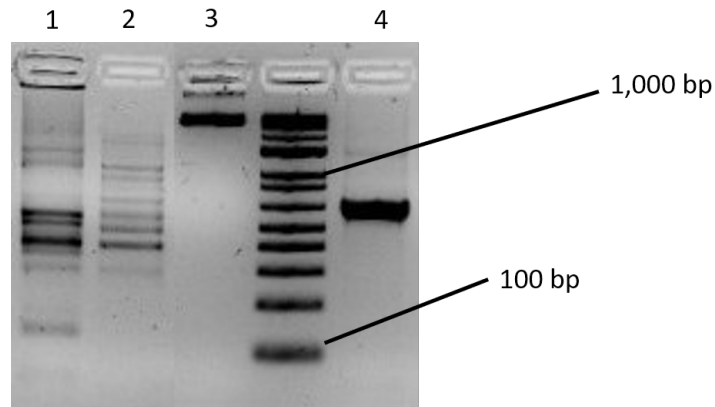


Figure 4.26: Agarose gel of the resulting HindIII digestion after *OTOG* cDNA amplification. 1. digested mutated hybrid minigene; 2. undigested mutated hybrid minigene; 3. pcDNA3.1 hygro vector (Invitrogen) and 4. digested pcDNA3.1 hygro vector for control. Although the pattern of bands is complex, it is clear that at least one transcript has been digested by HindIII both because 1. lacks some bands at around 1,000 bp and it has one low-weight band at around 100 bp. Therefore the tested hypothesis has been proved.



man Splicing Finder. In fact, it was predicted that some donor and acceptor splice sites should have resulted in transcripts including the HindIII restriction site reconstituted with the ligation of the insert within the vector (see fig. 4.25). Although the result may seem a bit complex, from the presence of a much lower band in the digested cDNA it is clear that the testing hypothesis has received some evidence (see fig. 4.26). The same pattern of bands was obtained both for wt and mutated minigene (data not shown).

In conclusion, these data suggest the absence of a pathogenic role for this variant, although further more specific studies may really be desirable, as the number of bands obtained could probably be a clear sign of intron misrecognition leading to illegitimate splicing by HEK293 cells.

Chapter 5

Discussion

1 *STRC/pSTRC* and *OTOA/pOTOA* variants sequencing

Hearing loss is a complex and heterogeneous disease with more than 100 genes involved. Therefore, NGS-based approaches are currently the most powerful way to tackle this complexity, although in some cases they may fail. This happens, for example, when large regions with high homology are sequenced together: in these cases, the short reads generated by standard NGS methodologies cannot be unambiguously aligned to the reference gene sequence. In other words, massive parallel sequencing may identify a variant, but in most cases it does not allow to understand where the variant maps (e.g. in the gene or the pseudogene) and if it is present in a heterozygous or homozygous state.

Therefore the best way to solve this problem is to couple the high-throughput capacity of NGS with long reads (possibly even longer than a long-range PCR). Currently, two methods from Pacific Biosciences and Oxford Nanopore exist that try to achieve a lot of high quality reads (for this reason they are called third-generation sequencing) over 160kb in length or even more. However, these techniques, although promising, are still expensive. Therefore, other approaches specifically designed to address this challenge need to be integrated in the standard analysis workflow. In this study, an efficient IPCR based approach was developed and tested.

However, before discussing other details about it, it may be interesting to analyse two other methods that, at least in theory, could have been successful, as well.

The first one, which indeed is pretty elegant, is to exploit the differential presence of a restriction site (hopefully more than one) between gene and pseudogene to cut gDNA and make sure that only one IPCR product (i.e. gene) is obtained. When the restriction site is located on the pseudogene, things start complicating a bit, because, although the concept remains valid, the digestion should be performed after the IPCR and there is the possibility that the size of digested and undigested fragments do not differ enough to be clearly separated through agarose gel electrophoresis. In conclusion, although the use of restriction enzymes is feasible, it is time consuming and expensive.

A second method is to exploit any existing difference in DNA methylation

between gene and pseudogene, even though both of them are not expressed in white blood cells, to design more specific IPCR primers after conversion of cytosines to uracils with bisulfite. However, it is necessary to know the DNA methylation pattern of each locus in advance and this could require a lot of work. Moreover, it is necessary that the bisulfite conversion step proceeds at completion both in the gene and in the pseudogene, before designing any primers. In fact, unless all cytosines are always converted into uracils, results will be hardly reproducible, not only because the PCR itself could fail (e.g. if primers hybridized not well on templates), but also because the alleged specificity of a PCR could change if by chance the same primers could hybridize on the homologous template.

Obviously this list can go on, but now it is a bit more clear why it was decided to use 3'-phosphorothioate primers which at least were simple, not toxic and not so expensive. Interestingly, what is really amazing about this technique is the universality of its potential applications that may include as well the sequencing of extremely long and repeated regions like centromeres. In fact, this technique could really be considered the scaled-up version of an allele specific PCR (ASPCR, [30]) that specifically amplifies up to 20kb of DNA requiring only two nucleotides that differ between the homologous regions. However, it is really important to remember that performing a IPCR is not like performing a standard PCR. Even though the two techniques are strongly related, it is fundamental to understand deeply all the molecular processes that might be involved in order to benefit from the advantages of both tests.

The success of this study was actually made possible because it became clear that the principles of ASPCR could not have been rigidly applied to this problem unless any particular expedient was taken. Of course, this not only came at the cost of failing several times, trying to modify all possible parameters (annealing and extension temperature, DMSO concentration, scalar dilution, buffers, primers . . .), but also required to disentangle external variables as well. Particularly challenging, in this sense, was the discovery that the IPCR of both *STRC* and *OTOA* could be obtained only by using two of the three different thermocyclers available in the laboratory. However, once external and internal variables were understood and specifically addressed, some achievements were made.

The first of them was the confirmation of a reported SNP (chr15:43998186 G>T) in intron 18 of *pSTRC* that was seen 4 times out of 4 samples analyzed. As reported in the gnomAD database (see fig. 5.1) this is a really common SNP with very low heterozygosity (data obtained only by whole genome sequencing). In fact, the allele frequency is >90% in many populations, including non-Finnish Europeans. Nevertheless, the finding of this SNP was important not only because it occurs in one of the divergent bases between *STRC* and *pSTRC*, but also because it may potentially confound nPCR results. In fact, the polymorphisms is towards the corresponding *STRC* specific base. As a matter of fact, the first time it was observed during *pSTRC* intron 18 sequencing, near other pseudogene specific and unique bases and before knowing it was actually a SNP, it warned of a possible phenomenon which was not still considered at that time, the PCR product chimera formation. PCR chimeras



Population Frequencies 				
Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency 
▶ East Asian	1548	1548	774	1.000
▶ African	7514	7622	3719	0.9858
▶ Latino	818	838	399	0.9761
▶ Ashkenazi Jewish	270	284	128	0.9507
▶ European (non-Finnish)	14135	15180	6603	0.9312
▶ Other	978	1064	451	0.9192
▶ European (Finnish)	3060	3444	1362	0.8885
▶ South Asian	0	0	0	0.000
Male	15823	16696	7537	0.9477
Female	12500	13284	5899	0.9410
Total	28323	29980	13436	0.9447

Figure 5.1: Allele frequencies from only whole genome sequencing projects for the chr15:43998186 G>T SNP in intron 18 of *pSTRC*.

are amplicons that originated from at least two different templates due to incomplete primer elongation. In simplistic terms, it is like as those incomplete PCR products act themselves as primers in the following PCR cycles, but having changed template first. As such, this SNP was eventually useful to start taking into account effective countermeasures like reducing the number of IPCR cycles for both *STRC* and *OTOA*.

Interestingly, nPCRs designed to be a specificity control for *STRC* and *OTOA* IPCRs suggested non-specific amplification in four cases (see tables 4.1 and 4.2). In fact, those regions analyzed by the nPCRs had some nucleotides that were different between gene and pseudogene and therefore allowed the discrimination of the two templates by Sanger sequencing. However, the fact that the results of all the other IPCRs were confirmed to be specific supported that the cause was not in the functioning principle of 3'-phosphorothioate primers, rather in the biological and technical features of each case. Some possible causes are discussed below.

In two of the four cases, nPCRs amplified exon 20 in both *STRC* and *pSTRC*. However, when the nPCR was conducted over the IPCR product made with 3'-phosphorothioate primers for *STRC*, it had hardly any pseudogene specific nucleotides, while when the IPCR was made with 3'-phosphorothioate primers for *pSTRC*, the nPCR was indeed a clear example of non-specific amplification with both gene and pseudogene specific bases being sequenced (see fig. 4.13). This could be explained by the fact that even with use of 3'-phosphorothioate specific primers, the amplification by IPCR is not 100% specific and there is also some amplification of the other template. While in most cases, such co-amplification was not sufficient to be detected when the specificity was tested, in these ones it might be higher due to some *STRC* specific bases in 3' of forward and reverse primers.

The other 2 cases regarded *OTOA* and were a bit more complicated to evaluate because it was not possible to specifically genotype *pOTOA*, since 3'-phosphorothioate primers were not available in the laboratory. Nevertheless, both long-range templates showed pseudogene contamination through a non-specific nPCR able to discriminate gene and pseudogene. Of course, it was clear that those results were not justifiable with the previous hypothesis because those nPCRs were the standard controls for all *OTOA* IPCRs and therefore their non-specificity had been deeply checked on healthy controls. If the previous hypothesis had been true in these cases, also other *OTOA* IPCRs should have shown this fact. For this reason, two other scenarios were suspected. The first one was the presence of at least one rare variant or SNP at the 3' end of forward or reverse annealing site for pseudogene IPCR. In fact, if this would be the case, the gene IPCR could lead also to pseudogene amplification much more than actually expected. Moreover, it is also interesting to test if a single SNP would be sufficient per se or rather two were required. But unfortunately, it was not possible to test these hypotheses in vitro because of lack of time. However, it should be noted that SNPs can occur also in the IPCR gene specific primers, not only in the pseudogene ones.

Another interesting hypothesis that could particularly fit one of the two cases nPCRs was the non-allelic homologous recombination (NAHR). In fact, this nPCR analyzed a *OTOA/pOTOA* recurrent variant (NM_144672.3: c.3281 C>T) which was seen four times in our laboratory: in two samples it was associated to *OTOA*, in one sample it was not possible to conduct the IPCR and in the other one, the case of interest, it was not clear if it would be associated with either *OTOA* or *pOTOA*. However, there is evidence that NAHR is a rather common mechanism in genes with high homologous genomic regions [31, 32]. Therefore, NAHR could explain not only the nPCR failure, but also could represent a molecular mechanism for the association of this apparently gene-specific variant to at least some regions of *pOTOA*. In fact, if NAHR had occurred between a mutated gene and a pseudogene in internal regions (i.e. IPCR primer sites had not recombined), not only some pseudogene regions could have been translocated to the gene, but also the variant could have been in cis with some pseudogene regions. Nevertheless, because of the real technical challenges that testing this hypothesis would have had, I did not elaborate any possible approaches, though a future solution could come from the aforementioned long read NGS platforms.

Anyway, another part of this work was composed by the simple statistical test created to assess CNVs (particularly deletions). It is now worth to discuss its limitations. In fact, the major structural drawback this system had is its evident multicollinearity, i.e. the usage of the same dataset to infer more than one type of information. When this happens, not only the conclusions drawn are somewhat biased and interdependent from one another, but also the possible effects of sampling error are magnified. In fact, this might be the case for this study, since, for ease, the dataset came from the same deaf patients analyzed by NGS and then sequenced for *STRC/pSTRC*.

Also, another limit is the implicit assumption of a negative correlation between the different variances made when translating the first population

distribution with mean 50%. This actually, may not be true since less frequent alleles (e.g. one mutated allele and 3 wt one, 2 of which are coming from pseudogene) are more exposed to sampling error during first amplification cycles than frequent ones are. In fact this effect was particularly strong in heterozygous + wt chromosomal assets with some cases falling outside the lower extreme of the 99% confidence interval. A partial solution was to consider the maximum deviation from the mean in percentage as the variance for each reference population, but at the cost of decreasing sensitivity for heterozygous cases with higher than expected mean allele frequencies. Finally, the comparison of results with the SureCall pair analysis tool could not be considered as a way to measure sensitivity and specificity for the statistical method, since both of them are, in the end, prediction software, and therefore prone to errors.

Nonetheless, this statistical system has helped the laboratory in formulating a diagnosis for a family with non-syndromic hereditary hearing loss due to a STRC missense mutation in exon 4 (NM_153700.2: c.1631A>G p.Tyr544Cys) and a possibly complete deletion (confirmed also with SureCall and trio analysis) of the gene. As such, this system not only could provide precise genotyping information for SNVs but also give, at least, some indications for possible multiexonic deletions.

In conclusion, it is important to remember that this study is not an actual demonstration that the Mandelker's approach for tackling genes with high homology is not effective, but rather it strongly suggests that it could be improved, for example by controlling the proofreading activity of IPCR polymerases blend. In addition, further work may be advocated to fill actual important gaps like the implementation of other bioinformatic tools to analyse NGS data and the development of a method to discover NAHR.

2 Hybrid minigene assays

In this study, five hybrid minigene assays were performed: four of them analysed genes associated with hearing loss (*COL2A1*, *COL11A2*, *OTOG*, *MYO15A*) and one associated with osteogenesis imperfecta (*COL1A1*). The approach used aimed to assess the potential effects on splicing of variants identified in patients at exon-intron boundaries through NGS targeted exon panels. Variants affecting the canonical ± 1 and 2 canonical splice sites were excluded. The vector used for the assays was the pcDNA3.1 hygrob-globin, previously generated in our laboratory and already exploited for similar functional mutation studies [26, 33].

The human β -globin gene was chosen for the construction of this vector because it is structurally simple, consisting in only three exons. Moreover, this gene is not constitutively expressed in HEK293 cells (see fig. 5.2), which are an easy system to work with, thus avoiding misinterpretation of results. However, hybrid minigenes are not perfect systems and have flaws which strongly depend on various factors such as: the tissue specificity of splicing regulation, the respective origin species of both insert and scaffold gene, the scaffold-insert intron specific boundaries (i.e. the vector cloning site) and the designed primer

2. HYBRID MINIGENE ASSAYS

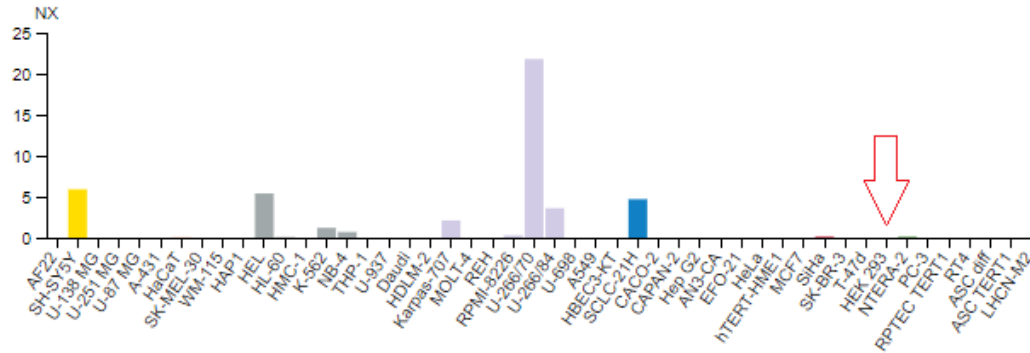


Figure 5.2: Capture from The Protein Atlas website that shows in which cell line the human β -globin gene is expressed constitutively. The red arrows highlight the absence of expression in HEK293 cells.

pair for insert generation.

For instance, it may not seem surprising that the amount of illegitimate splicing (which is a technical artefact) is expected to be higher in a human hear-related gene transfected into HepG2 than into human hair cells, since HepG2 cells are human liver cancer cells. Therefore, when working with hybrid minigenes, it is always a good practice to transfect the nearest cell line, in terms of gene expression, to the one where the tested gene could manifest molecular, biochemical or physiological alterations, if mutated.

Nevertheless, most laboratories cannot always afford to invest time and money needed to assure they have the best experimental setup and so it is convenient to them to perform human hybrid minigene assays in an easy cell line like HeLa or HEK293 (where they may had established some expertise), and then, whenever possible, confirm positive results in a more indicated cell line. However, this approach, though feasible and largely applied, might result in a serious loss of results interpretability. To tackle this and other aforementioned drawbacks, hybrid minigene assays must always include, along the gene mutated copy, one with the wild-type allele, in order to rule out splicing alterations due to the variant or to the hybrid minigene construct itself.

Among the five variants tested, only one is suggested to have a pathological role: *COL1A1* NM_000088.3: c.1515G>A p.(=). This variant was found in heterozygosis in a foetus died of osteogenesis imperfecta, a disease caused in 85% of cases by mutations in *COL1A1* and *COL1A2* [34]. These genes encode, respectively, for pro alpha 1 and 2 chains of collagen type I and are arranged in a 2:1 ratio. Since the variant is at the very last nucleotide of the 22nd exon, it could potentially alter the exon definition process and thus splicing. This was found indeed to be true, as the variant causes skipping of exon 22, a variation reported as being lethal [35]. In fact, even though the Gly-Xaa-Yaa triplet pattern is not altered by the 54 bp shortening of mRNA, the chain alignment can be shifted, affecting proper folding of the trimeric collagen structure. In other words, this variant is a perfect example of a dominant negative mutation, but unfortunately it was not possible to directly replicate the assay on osteosarcoma cells which are definitely more indicated in this case.

On the other hand, the fact that the results regarding the other four variants, all affecting hear-related genes, did not support their pathogenicity is not evidence of the inappropriateness of the chosen cell line for this kind of genes, but rather may reflect random effects. To actually prove such a hypothesis, it would be useful to test in HEK293 cells a lot of similar hybrid minigenes that had been previously demonstrated to have a pathogenic role.

Therefore, the only thing one could conclude from these results is that the high degree of illegitimate splicing, from one (*COL11A2*) to more than 5 gel bands (*OTOG*), may suggest both to redesign primers used to generate the insert and to assess the same hybrid minigene in other cell lines. Interestingly, it was found that the degree of illegitimate splicing did not correlate with the number of cloned exons, as the *OTOG* hybrid minigene included only one, but rather with the vector-insert specific intron boundaries. In particular, the *MYO15A* hybrid minigene showed a band corresponding to the fully spliced transcript, containing only β -globin exons. Such a condition is commonly conceived to be a major indicator of insert misrecognition or hybrid minigene failure. A possible explanation could be the loss of important splicing regulation elements (i.e. ISS and ESE) during the first amplification step when gene-specific primers were designed.

Another aspect which might be misleading is the fact that almost always the same gel band had different intensities across wt and mutated hybrid minigenes. In fact, it is not correct to conclude, from such an experimental setup, that the differential intensity and thus the different quantity of transcript is due to the variant itself, because, even when this is the case, there are multiple confounding factors not taken into account, such as the two steps (retrotranscription and amplification) required to get cDNA from RNA and the preferential amplification of shorter fragments.

For instance, it may seem that the gel band in the middle of the three bands of *MYO15A* (see fig. 4.23) is enhanced in the mutated hybrid minigene with respect to the wt condition. However, a leading factor to this arrangement could be the faster reaching of plateau phase during cDNA amplification of wt hybrid minigene than the mutated one. Moreover, if by chance, during the first cDNA amplification cycles or RNA retrotranscription, there was an unequal sampling of the three different transcripts, this initial bias could easily generate the aforementioned asymmetry. In conclusion, this happens because all the steps performed from the cDNA amplification up to the agarose gel are not meant to draw these kinds of conclusions and other approaches (e.g. rtPCR) are instead much more indicated.

Finally, it is worth to speculate a bit on bioinformatic variant prediction results. Generally speaking, they performed well, although it seems that Human Splicing Finder is more sensitive than both NetGene2 and NNSplice. In fact, in most cases it predicted some kind of splicing alteration, particularly through splicing regulatory elements, but unfortunately those hear-related hybrid minigene assays did not prove evidence of that. However, on the other hand, there was a case where NNSplice did not even see an actual acceptor splice site in *COL2A1* wt insert (data not shown). Interestingly, the only case where all algorithms agreed was also in the only variant with a clear effect on

splicing (*COL1A1* c.1515G>A), thus confirming that one criteria required to support pathogenicity is that multiple software predict deleterious effects on splicing. Finally, it is not clear how the different scores for each prediction (compared to wt controls) might be used to determine a threshold within a chosen confidence interval, leaving the management of false positives totally random.

Nevertheless, splicing prediction software still came useful when it was tried to address complex tasks like forecasting which intronic regions were retained in the *OTOG* hybrid minigene (see fig. 4.25 and 4.26). Actually, Human Splicing Finder found a 300 bp hotspot for donor and acceptor splice sites downstream the inserted exon. Fortunately, this region turned out to comprehend the insert-vector ligation site, where a HindIII restriction site was present. Therefore, an easy enzymatic step allowed to test whether or not the forecast was accurate or, in other words, if at least one transcript retained the intronic insert-vector ligation site. However, this enzymatic test could not exclude the possibility of DNA contamination during RNA extraction from HEK293 cells, which of course could have happened in other minigenes as well, although for *OTOG* the suspects (i.e. a high molecular faint smear) were higher. But in any case, this proves once again the usefulness of having accurate and updated bioinformatic tools.

References

- [1] Koeppen BM and Stanton BA. *Berne and Levy Physiology*. Elsevier., 2018 (cit. on p. 1).
- [2] Litovsky R. “Development of the auditory system.” *Handbook of clinical neurology*. (2015) (cit. on p. 1).
- [3] Shearer AE, Hildebrand MS, and Smith RJH. “Adam MP, Ardinger HH, Pagon RA, et al., editors. GeneReviews®.” University of Washington, Seattle., 1999. Chap. Hereditary Hearing Loss and Deafness Overview. (Cit. on pp. 1, 2).
- [4] Korver AM, Smith RJ, Van Camp G, et al. “Congenital hearing loss.” *Nat Rev Dis Primers*. (2017) (cit. on p. 2).
- [5] Toriello HV and Smith SD. *Hereditary hearing loss and its syndromes*. Oxford University Press., 2013 (cit. on p. 2).
- [6] Mamanova L, Coffey AJ, Scott CE, et al. “Target-enrichment strategies for next-generation sequencing.” *Nature methods*. (2010) (cit. on p. 3).
- [7] Strachan T and Read A. *Human molecular genetics*. Garland Science., 2011 (cit. on pp. 4, 24).
- [8] Zwaenepoel I, Mustapha M, Leibovici M, et al. “Otoancorin, an inner ear protein restricted to the interface between the apical surface of sensory epithelia and their overlying acellular gels, is defective in autosomal recessive deafness DFNB22.” *Proceedings of the national academy of sciences*. (2002) (cit. on p. 7).
- [9] Lukashkina AN, Legana PK, Weddella TD, et al. “A mouse model for human deafness DFNB22 reveals that hearing impairment is due to a loss of inner hair cell stimulation.” *Proceedings of the national academy of sciences*. (2012) (cit. on p. 7).
- [10] Cartagena-Rivera AX, Le Gal S, Richards K, et al. “Cochlear outer hair cell horizontal top connectors mediate mature stereocilia bundle mechanics.” *Science advances*. (2019) (cit. on pp. 7, 8).
- [11] Mandelker D, Amr SS, Pugh T, et al. “Comprehensive Diagnostic Testing for Stereocilin. An approach for analyzing medically important genes with high homology.” *The journal of molecular diagnostics*. (2014) (cit. on pp. 9, 24, 25, 39).

REFERENCES

- [12] Vona B, Hofrichter MAH, Neuner C, et al. “DFNB16 is a frequent cause of congenital hearing impairment: implementation of STRC mutation analysis in routine diagnostics.” *Clinical genetics*. (2015) (cit. on pp. 9, 25, 39).
- [13] Wang Z and Burge CB. “Splicing regulation: From a parts list of regulatory elements to an integrated splicing code.” *RNA*. (2008) (cit. on pp. 10, 12, 14, 15).
- [14] Will CL and Lührmann R. “Spliceosome structure and function.” *Cold Spring Harb perspectives in biology*. (2011) (cit. on pp. 11, 12).
- [15] Matlin AJ, Clark F, and Smith CWJ. “Understanding alternative splicing: towards a cellular code.” *Nature reviews*. (2005) (cit. on p. 14).
- [16] Cartegni L, Chew SL., and Krainer AR. “Listening to silence and understanding nonsense: exonic mutations that affect splicing.” *Nature reviews genetics* (2002) (cit. on pp. 14, 17).
- [17] Liu HX, Cartegni L, Zhang MQ, et al. “A mechanism for exon skipping caused by nonsense or missense mutation in BRCA1 and other genes.” *Nature genetics*. (1998) (cit. on p. 14).
- [18] Black DL. “Mechanisms of alternative pre-messenger RNA splicing.” *Annual reviews in biochemistry* (2003) (cit. on pp. 14, 15).
- [19] Hui J, Hung LH, Heiner M, et al. “Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing.” *The EMBO journal* (2005) (cit. on p. 15).
- [20] Chasin LA. *Searching for Splicing Motifs. Alternative Splicing in the Postgenomic Era*. Landes Bioscience, 2007 (cit. on p. 15).
- [21] Lewandowska MA. “The missing puzzle piece: splicing mutations.” *International Journal of Clinical and Experimental Pathology*. (2013) (cit. on p. 17).
- [22] Abramowicz A and Gos M. “Splicing mutations in human genetic disorders: examples, detection and confirmation.” *Journal of Applied Genetics*. (2018) (cit. on p. 17).
- [23] Sharma N, Sosnay PR, Ramalho AS, et al. “Experimental assessment of splicing variants using expression minigenes and comparison with in silico predictions.” *Human mutation*. (2014) (cit. on pp. 18, 19).
- [24] Jian X, Boerwinckle E, and Liu X. “In silico tools for splicing defect prediction – A survey from viewpoint of end-users.” *Genetics in medicine*. (2014) (cit. on pp. 18, 19).
- [25] Richards S, Nazneen A, Bale S, et al. “Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.” *Genetics in medicine*. (2015) (cit. on p. 29).

-
- [26] Forzan M, Salviati L, Pertegato V, et al. "Is CFTR 621+3 A>G a cystic fibrosis causing mutation?" *Journal of human genetics* (2010) (cit. on pp. 30, 33, 58, 69).
- [27] Chomczynski P and Sacchi N. "Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction." *Analytical biochemistry*. (1987) (cit. on p. 35).
- [28] Smythab RP, Schlubdm TE, Grimm A, et al. "Reducing chimera formation during PCR amplification to ensure accurate genotyping." *Gene*. (2010) (cit. on p. 45).
- [29] Haas BJ, Gevers D, Earl AM, et al. "Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons." *Genome research*. (2011) (cit. on p. 45).
- [30] Ruano G and Kidd KK. "Direct haplotyping of chromosomal segments from multiple heterozygotes via allele-specific PCR amplification." *Nucleic acid research* (1989) (cit. on p. 66).
- [31] Bunge S, Rathmann M, and Steglich C. "Homologous nonallelic recombinations between the iduronate-sulfatase gene and pseudogene cause various intragenic deletions and inversions in patients with mucopolysaccharidosis type II." *European Journal of Human Genetics*. (1998) (cit. on p. 68).
- [32] Tanooka H, Ootsuyama A, and Sasaki H. "Homologous recombination between p53 and its pseudogene in a radiation-induced mouse tumor." *Cancer research*. (1998) (cit. on p. 68).
- [33] Cassina M, Cerqua C, Rossi S, et al. "A synonymous splicing mutation in the SF3B4 gene segregates in a family with highly variable Nager syndrome." *European Journal of Human Genetics*. (2017) (cit. on p. 69).
- [34] Marini JC, Forlino A, Bächinger HP, et al. "Osteogenesis imperfecta." *Nature review disease primers*. (2017) (cit. on p. 70).
- [35] Steiner RD, Adsit J, and Basel D. "Adam MP, Ardinger HH, Pagon RA, et. al, editors. GeneReviews®." University of Washington, Seattle., 2005. Chap. COL1A1/2-Related Osteogenesis Imperfecta. (Cit. on p. 70).

