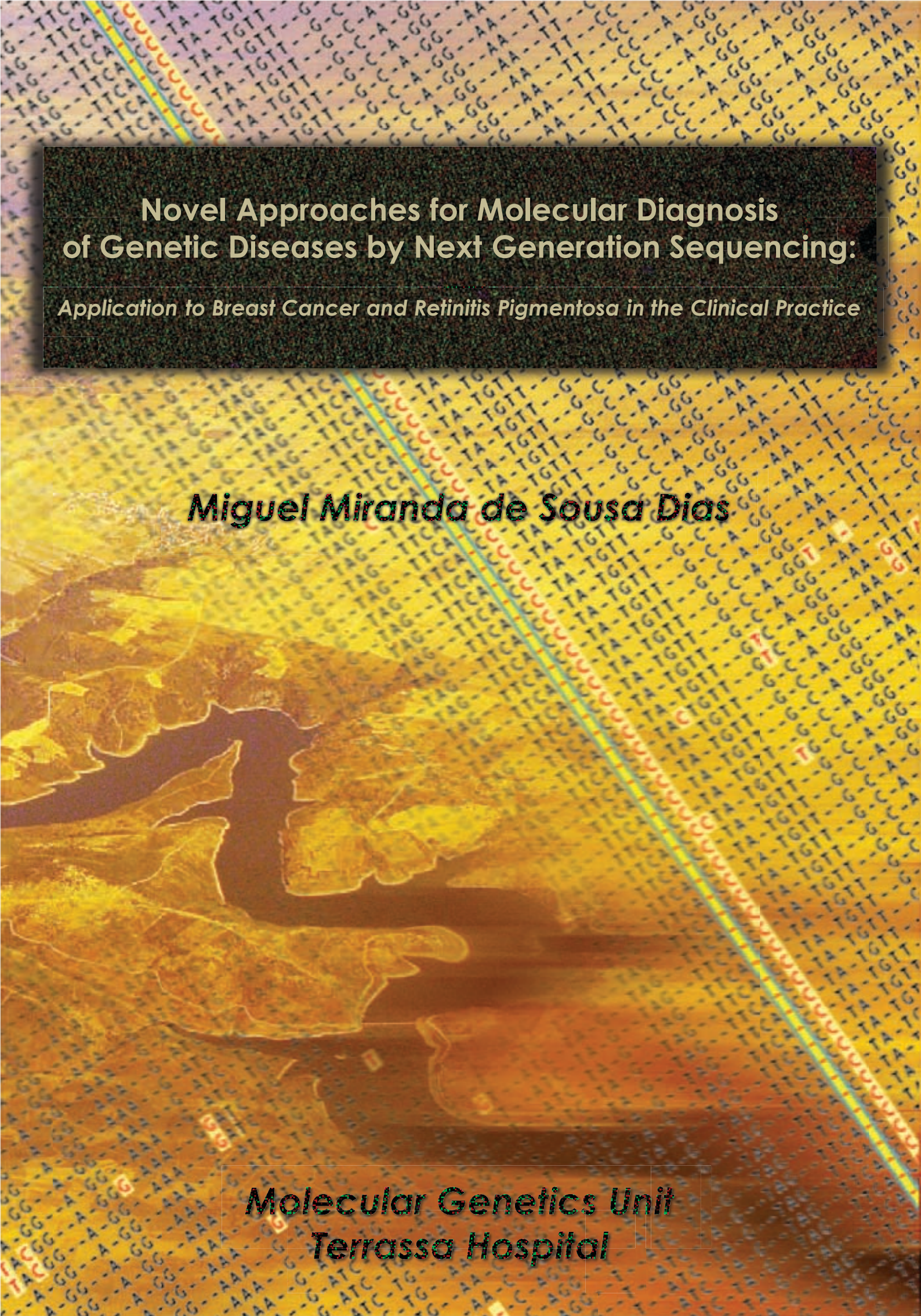
Novel Approaches for Molecular Diagnosis of Genetic Diseases by Next Generation Sequencing:

Application to Breast Cancer and Retinitis Pigmentosa in the Clinical Practice



Miguel Miranda de Sousa Dias

Molecular Genetics Unit
Terrassa Hospital

Work performed at the Molecular Genetics Unit of the Terrassa Hospital (Barcelona, Spain), and partially funded by the Carlos III Health Institute, the Fund for Health Research (FIS) from the Ministry of Health and Consumption of Spain and the Fundació Joan Costa Roma, of the Consorci Sanitari de Terrassa.

The cover is composed of three images. The background image is an aerial view of the Douro River, a river that connects Spain with Portugal. The second image corresponds to a DNA sequence alignment software representation and the third one, used as a background for the titles and writing, is an image of the PicoTiter Plate after a sequencing run. Cover design by Miguel Constantino Silva de Sousa Dias.

English edited by Richard May.

Novel Approaches for Molecular Diagnosis of Genetic Diseases by Next Generation Sequencing:

Application to Breast Cancer and Retinitis Pigmentosa in the Clinical Practice

A Thesis Presented to

The Department of Biochemistry and Molecular Biology

Universitat Autònoma de Barcelona

In Fulfilment of the Requirements for the Degree

Philosophiæ Doctor (PhD) in Biochemistry, Molecular Biology and Biomedicine

Director: Dr. José Miguel Carballo Villarino

Tutor: Dr. Jaume Farrés Vicén

by

Miguel Miranda de Sousa Dias

December 2013

El **Dr. José Miguel Carballo Villarino,** cap de la Unitat de Genètica Molecular de l'Hospital de Terrassa, i el **Dr. Jaume Farrés Vicén,** catedràtic del Departament de Bioquímica i Biologia Molecular de la Universitat Autònoma de Barcelona,

CERTIFIQUEN:

Que **Miguel Miranda de Sousa Dias** ha realitzat sota la seva direcció i supervisió el treball d'investigació que s'exposa en la memòria titulada **"*Novel Approaches for Molecular Diagnosis of Genetic Diseases by Next Generation Sequencing: Application to Breast Cancer and Retinitis Pigmentosa in the Clinical Practice*"** per optar al grau de Doctor en Bioquímica, Biologia Molecular i Biomedicina per la Universitat Autònoma de Barcelona.

Que aquest treball s'ha dut a terme a la Unitat de Genètica Molecular de l'Hospital de Terrassa.

I perquè consti, signem el present certificat

Dr. José Miguel Carballo Villarino
Director de la Tesi

Dr. Jaume Farrés Vicén
Tutor de la Tesi

Miguel Miranda de Sousa Dias
Autor de la Tesi

Terrassa, a 2 de Desembre de 2013

À minha mãezinha, ao meu grande pai e à maluca da minha mana,

cliché ou não, esta é para vocês!

"Nothing in ~~Biology~~ Genetics Makes Sense Except in the Light of Evolution"

Theodosius Dobzhansky

Acknowledgements

Even though I am the only author signing this thesis a great number of people have left their personal touch on it. For that reason not to mention them would be extremely unfair. My heart is Portuguese and because the next words come from my heart they too are in Portuguese.

Começo por quem fez tudo isto possível, o Doutor Miguel Carballo, chefe da Unidade de Genética do Hospital de Terrassa e meu director de tese. Obrigado por teres dado a este português de atitude "*un poco pardilla*" a oportunidade de aprender os (tantos) mistérios por trás da genética humana e de demonstrar que tem a capacidade para desenvolver todo um trabalho de investigação de qualidade contigo. À parte disso, te agradeço também o facto de teres sido um bom chefe de equipa, compreensível e sempre zelando pelos interesses de todos os membros do grupo de trabalho. Cientificamente sublime e com ideias e soluções a condizer, posso até afirmar que, dos nossos *brainstormings,* ideias potencialmente milionárias foram engendradas. Assim sendo, expresso também o meu sincero agradecimento pela orientação científica, pela revisão crítica desta dissertação e pelo teu contributo para o resultado final deste trabalho. Muitas vezes faltam os meios, mas Miguel e a sua varilha mágica são capazes de resolverem os mais variados obstáculos. Por essa razão também desejo manifestar a minha gratidão pela perseverança e determinação que permitiram criar as condições indispensáveis para a realização deste trabalho. Obrigado Miguel, pela honra de ser aprendiz do já chamado "*Famoso Doutor Miguel Carballo*".

Ao Professor Doutor Jaume Farrés, professor do departamento de Bioquímica e Biologia Molecular da Universidade Autónoma de Barcelona, apresento o meu sentido de agradecimento por ter aceitado ser meu tutor neste projecto e estender o seu apoio e disponibilidade, bem como pela revisão desta dissertação e as suas palavras de apreço generosamente dirigidas.

À Professora Doutora María Carmen Martínez, professora do departamento de Bioquímica e Biologia Molecular da Universidade Autónoma de Barcelona, agradeço o facto de ter aceite fazer parte da comissão de seguimento desta tese doutoral, como presidente.

Às minhas colegas do laboratório da Unidade de Genética Molecular, Alba Chacón (ainda que por pouco tempo), Beatriz Pascual, Begoña Mañé, Emma Borràs, Imma Hernan e Maria José Gamundi, que passados quase quatro anos – e depois de ter visto três novas vidas virem a este mundo – penso cá com os meus botões que este será o grupo de trabalho mais fértil onde alguma vez trabalharei. Brincadeira à parte, sem dúvida aprendi muito com cada uma de vocês nestas muitas horas partilhadas em laboratório e por esse motivo apresento o meu profundo agradecimento.

À cantina das dez da noite e todas as pessoas responsáveis por ela, apresento também o meu agradecimento pelos pitéus tardios no Hospital de Terrassa e por manterem este espécime de grande necessidade calórica sempre bem nutrido, mesmo naqueles dias – e foram bastantes – em que o trabalho ou a escrita desta tese foram A prioridade.

Ao *Instituto de Salud Carlos III*, ao *Fondo de Investigación Sanitaria* (FIS) e à *Fundació Joan Costa Roma*, gostaria de agradecer o apoio financeiro parcial dos projectos nos quais participei ao longo destes quatros anos de bolsa de investigação. Estendo também os meus agradecimentos ao Doutor Manel Balcells, o então director-gerente da *Fundació Joan Costa Roma*.

Ao Alex Perez, especialista de aplicações da *Roche Diagnostics*, pelo fornecimento do óleo de emulsificação necessário para levar a cabo inúmeras experiencias e também pela excelente receptividade que sempre demonstrou ao longo do presente trabalho.

Richard May, how could I not thank you in your mother language? My deepest thanks to you, for doing the English revision of this dissertation. Your advices were amusing and effective lectures improving both my written English and this thesis.

Lorena, zemra ime, diz-se que a paciência é uma grande virtude e tu sempre tiveste tanta comigo. Obrigado por acreditares em mim, pelo teu amor e confiança desde sempre demonstrados, que nos momentos de maior esmorecimento foram tão importantes para mim. Faleminderit shume jeta ime!

Por último, mas não menos importante agradeço a minha *Família*, nomeadamente os meus pais, irmã e os meus amigos chegados! Mãe, as despedidas doem e a saudade aperta, mas mesmo à distância consegues ser esta imensa força que me empurra para a frente. Pai, tenho muita sorte que a tua veia artística tocou esta dissertação mas é também pelo teu apoio incondicional que te agradeço. Mana, sempre disponível com opiniões válidas, agradeço-te por teres estado sempre disponível para o teu irmão predilecto e por me mostrares que afinal até não sou o "especial" da família ("Mom says I'm special!").

Contents

Abbreviations

A	Adenine	
C	Cytosine	
G	Guanine	
T	Thymine	
dNTP	deoxyriboNucleotide TriPhosphates	mM
DNA	Deoxyribonucleic Acid	ng. μl^{-1}
cDNA	complementary Deoxyribonucleic Acid	
dsDNA	double strand Deoxyribonucleic Acid	
RNA	Ribonucleic Acid	ng. μl^{-1}
cRNA	complementary Ribonucleic Acid	
mRNA	Messenger Ribonucleic Acid	
ssRNA	single strand Ribonucleic Acid	
bp	Base Pair	
Kb	1,000 base pairs	
Mb	1,000,000 base pairs	
Ta	Annealing Temparature	ºC
Tm	Melting Temperature	ºC
Taq	Thermophilic Bacterium Thermus aquaticus derived	U/µl

Tgo	Thermococcus Gorgonarius derived	U/µl
PCR	Polymerase Chain Reaction	
emPCR	Emulsion Polymerase Chain Reaction	
LR-PCR	Long Range Polymerase Chain Reaction	
LR-EF	LR-PCR allied with Enzymatic Fragmentation	
RT-PCR	Real Time Polymerase Chain Reaction	
ε	Extinction Coefficient	$\mu g.ml^{-1}.cm^{-1}$
$_{Em}$	Emission Wavelength	nm
$_{Ex}$	Excitation Wavelength	nm
EDTA	Ethylenediamine Tetraacetic Acid	g
TAE	Tris-Acetat-EDTA	mM
TE	Tris-EDTA	mM
Tris	Tris(Hydroxymethyl)Aminomethane	g
DMSO	Dimethyl sulfoxide	
aa	Amino Acid	
AAS	Amino Acid Substitution	
BLP	Bead-Linker-Primer complex	
BRCA	Breast Cancer	
CCD	Charge-Coupled Device camera	
CDS	Coding Sequence	
CGA	Candidate Gene Approach	
DGGE	Denaturing Gradient Gel Electrophoresis	
ERG	Electroretinography	
FA	Fatty acid group	
FRET	Fluorescence Resonance Energy Transfer	
x g	Gravitational Constant	
GS	Genome Sequencer	

IC	Interconnecting Cilium
IPG	Ion Personal Genome Machine
IS	Inner Segment
kDa	KiloDalton
MID	Molecular Identifier
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
nt	nucleotide
OS	Outter Segment
PGM	Ion Torrent Personal Genome Machine
PPi	inorganic PyroPhosphate
PR	Photoreceptor
PTP	Pico-Titer Plate
QC	Quality Control
RP	Retinitis Pigmentosa
RPE	Retinal Pigment Epithelium
SFF	Standard Flowgram Format
SMRT	Single Molecule Real-Time
TIRF	Internal Reflection Fluorescence
tSMS	true Single Molecule Sequencing
UTR	Untranslated Region
UTTD	Ultra Turrax Tube Drive
UV	Ultra-Violet
w/o	Water-in-Oil
WT	Wild-Type
ZMW	Zero-Mode Waveguides

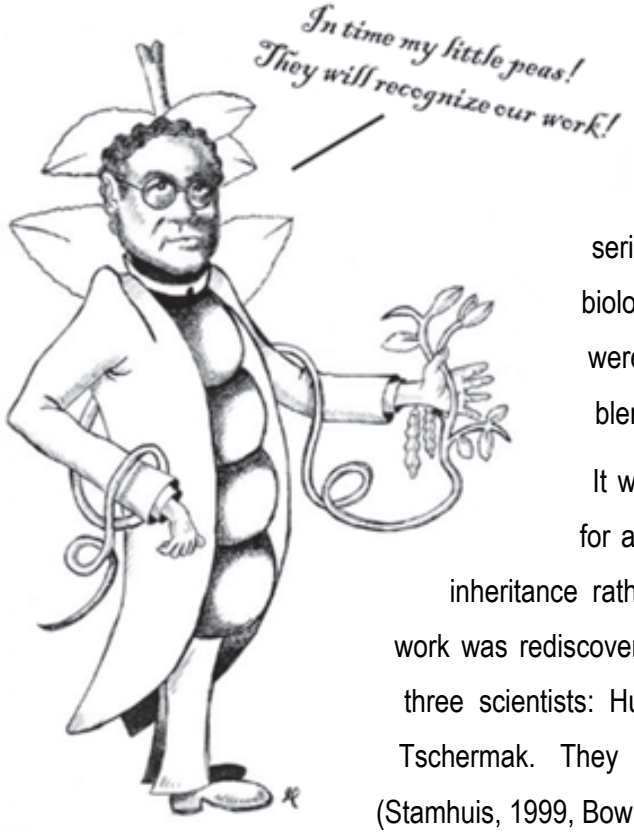
List of Tables

List of Figures

1. Introduction

- *The First Geneticist*

In the middle of the 1800s Gregor Mendel, an Austrian augustinian monk and scientist, was engaged in the experimental breeding of pea plants, and discovers that certain traits are inherited as units. We now call these units of inheritance 'genes' and for his discovery Mendel is now known as the father of modern genetics, which since monks aren't allowed to be the father of anything renders this identity somewhat ironic.



By 1866 Mendel published his work entitled "Versuche über Pflanzenhybriden" or "Experiments on Plant Hybridization" (Mendel, 1866), which wasn't taken very seriously by the scientific community; most biologists at the time held the idea that all traits were passed to the next generation through blending inheritance.

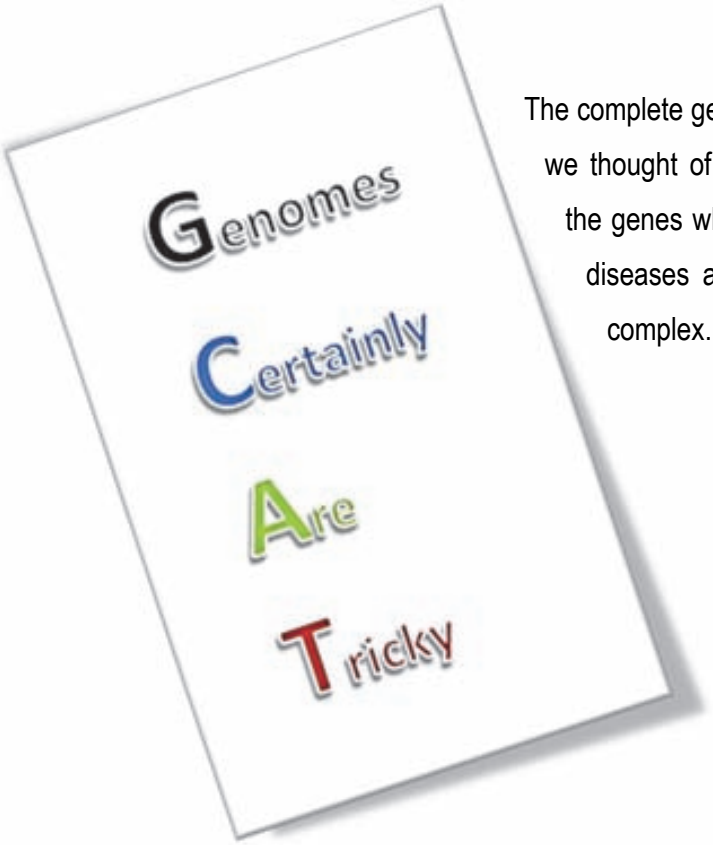
It was only by 1900, when scientists on a quest for a theory to successfully explain discontinuous inheritance rather than blending inheritance, that Mendel's work was rediscovered leading to its independent duplication by three scientists: Hugo de Vries, Carl Correns, and Erich von Tschermak. They all later acknowledged Mendel's priority (Stamhuis, 1999, Bowler, 2003).

Figure 1.1: Mendel's caricature, referring to his misunderstood findings (credits to Miguel-Constantino de Sousa Dias).

- *The Most Famous Biopolymer and its Code*

Like Gregor Mendel, we are built from around 50 trillion cells. Like him and his pea plants, our cells contain a set of instructions that are passed on to the next generation, the Genome. It is made entirely of DNA, which stands for deoxyribonucleic acid and is basically a biopolymer in which its repetitive units (or monomers) are nucleotides: the adenine, thymine, cytosine, and, guanine. Hence, DNA carries a code built from an alphabet of just four letters: A, T, C, and G (the nucleotide adenine, thymine, cytosine, and, guanine, respectively), but there are 3 billion of them in one cell of an average person; that makes it awfully complicated.

Ever since Mendel started breeding pea plants, staggering progress has been made in deciphering genomes, figuring out how the code is translated, how and when it is active, and what happens when it goes wrong. In 2003, with the Human Genome Project publishing the entire 3 billion letter code that makes an average person, many thought it would be the end of hereditary diseases, or genetic diseases. But genetics is not just about the code; it is about how that code is interpreted and how it makes different cells and different people.



The complete genome is rich with activity; even the parts we thought of as junk are actually instructions telling the genes what to do and when to do it. People and diseases are complex because our genomes are complex.

Figure 1.2: Genomes Certainty Are Tricky!

1.1. Genetic Disorders

DNA is a magnificent biopolymer that not only holds the code for who we are and where we come from, it also holds the key for what causes the diseases we might have. It can also show our predispositions for other diseases and cancers we do not yet have. This amazing macromolecule is in a constant state of replication and it's easy to imagine that sometimes something goes wrong. Just like trying to copy a 3 billion letter book by hand, you might miss a comma and this small mistake may or may not change the sentence's meaning. A mutation arises when a mistake occurs in the DNA's replication and this mistake changes an encoded message for an essential protein; this has the potential to originate a genetic disorder. A genetic disorder is then defined as an illness that is caused by abnormalities in our genes or chromosomes, which in turn are generated by inevitable errors in chromosome segregation at meiosis, DNA replication and DNA repair, and spontaneous chemical attack (Strachan and Read, 1999).

As we unravel the secrets of the human genome, we are learning that nearly all diseases have a genetic component. Some, including many cancers, are caused by a mutation in a gene or a group of genes in the cells of an individual. Such mutations occur randomly but can be enhanced by exposure to environmental mutagenic agents (e.g. exposure to any carcinogen). Other genetic disorders are hereditary - such as **retinitis pigmentosa** - where a mutated gene is passed down through a family and each generation of children can inherit the gene that causes the disease. Nevertheless, most genetic disorders are "multifactorial inheritance disorders," meaning they are caused by a combination of small variations in genes, often in concert with environmental factors. This is why most genetic disorders are quite rare and affect one person in every several thousands or millions. Indeed some types of recessive gene disorders confer an advantage in the heterozygous state in certain environments.

Through research on the human genome, we now know that many common diseases are usually caused by genetic alterations in the genes of an individual's cells. Some diseases, such as **breast and ovarian cancer**, also have rare hereditary forms. In these cases, gene variants that cause or strongly predispose a person to these cancers run in a family and significantly increase each member's risk of developing the disease. Accordingly, it is possible to divide genetic disorders into three groups:

1) Disorders that are caused by aberrations in our chromosomes like an excess or deficiency of the genes that are located on a specific chromosome or an entire chromosome extra copy, or structural changes within chromosomes themselves. Down

syndrome, for example, is caused by an extra copy of chromosome 21, but no individual gene on the chromosome is abnormal.

2) Complex multifactorial inheritance disorders are caused by a combination of small variations in multiple genes, often in concert with environmental factors. Heart disease, obesity and most cancers are examples of these disorders. Behaviours are multifactorial, complex traits involving multiple genes that are affected by a variety of other factors.

3) Monogenic or single gene disorders (also known as **mendelian disorders**) are caused by a mutation in a single gene. The mutation may be present on one or both alleles (one allele inherited from each parent). Sickle cell disease, cystic fibrosis, and Tay-Sachs disease are examples of single gene disorders.

1.1.1. Mendelian Inheritance Patterns

Genetic disorders are heritable thus passed down from the parents to sons. As noted above, other defects may be caused by spontaneous mutations or changes in the DNA. In these cases, these defects will only be heritable if they occurred in the germ line.

The simplest genetic characters are those whose presence or absence depends on the genotype at a single locus. That is not to say that the character itself is programmed by only one pair of genes - expression of any human character is likely to require a large number of genes and environmental factors. However, sometimes a particular genotype at one locus is both necessary and sufficient for the character to be expressed. Such characters are called mendelian or single gene. In humans over 10,000 mendelian characters are known. The essential starting point for acquiring information on any human mendelian character, whether pathological or non-pathological, is the OMIM (Online Mendelian Inheritance in Man) Internet database.

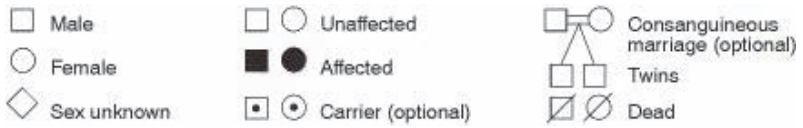


Figure 1.3: Symbols used for drawing pedigrees. Generations are usually labelled in Roman numerals, and individuals within each generation in Arabic numerals; III-7 the seventh person from the left in generation III. An arrow → can be used to indicate the index subject (Strachan and Read, 1999).

There are five basic mendelian pedigree patterns. Figure 1.3 shows the symbols used for drawing pedigrees. Mendelian characters may be determined by loci on an autosome or on the X or Y sex

chromosomes. Autosomal characters in both sexes and X-linked characters in females can be dominant or recessive. Nobody has two genetically different Y chromosomes (in the rare XYY males, the two Y chromosomes are duplicates). Thus there are five archetypal mendelian pedigree patterns (Strachan and Read, 1999).

- *Autosomal dominant inheritance*

Here, an affected person usually has at least one affected parent (in non-sporadic cases); it will affect either sex and it is transmitted by either sex; a child born from an affected and unaffected mating will have 50% chance of being affected as well. Of course this assumes that the affected parent is heterozygous, which is normally true for rare conditions (Figure 1.4A).

- *Autosomal recessive inheritance*

In this pattern, affected people are usually born to unaffected parents being that their parents are usually asymptomatic carriers. There is obviously an increased incidence of parental consanguinity and this pattern can affect either sex. Statistically speaking, after the birth of an affected child, each subsequent child has a 25% chance of being affected (Figure 1.4B).

- *X-linked recessive inheritance*

This mode of inheritance affects mainly males (but there are exceptions); affected males are usually born from unaffected parents, being that the mother is normally an asymptomatic carrier and may have affected male relatives; females may be also affected if the father is affected and the mother is a carrier, or occasionally as a result of non-random X-inactivation. Evidently there is no male-to-male transmission in the pedigree, although matings between an affected male and carrier female can give the appearance of male-to-male transmission (Figure 1.4C).

- *X-linked dominant inheritance*

This mode affects either sex, but more females than males; females are often more mildly and more variably affected than males due the random X-inactivation. The child of an affected female, regardless of its sex, will have a 50% chance of being also affected. In the case of an affected male, all daughters will be affected but none of his sons (Figure 1.4D).

- *Y-linked inheritance*

This inheritance pattern affects only males. Affected males always have an affected father (in non-sporadic cases); all sons of an affected man are also affected.

Undoubtedly, these basic patterns are subject to various complications. Some of which involve; common recessive conditions that can give a pseudo-dominant pedigree pattern; or failure of a dominant condition to manifest (incomplete penetrance). There are also conditions showing variable expression like imprinted genes, where their expression depends on parental origin. In addition, sporadic or new mutations often complicate pedigree interpretation, and what is more can lead to mosaicism.

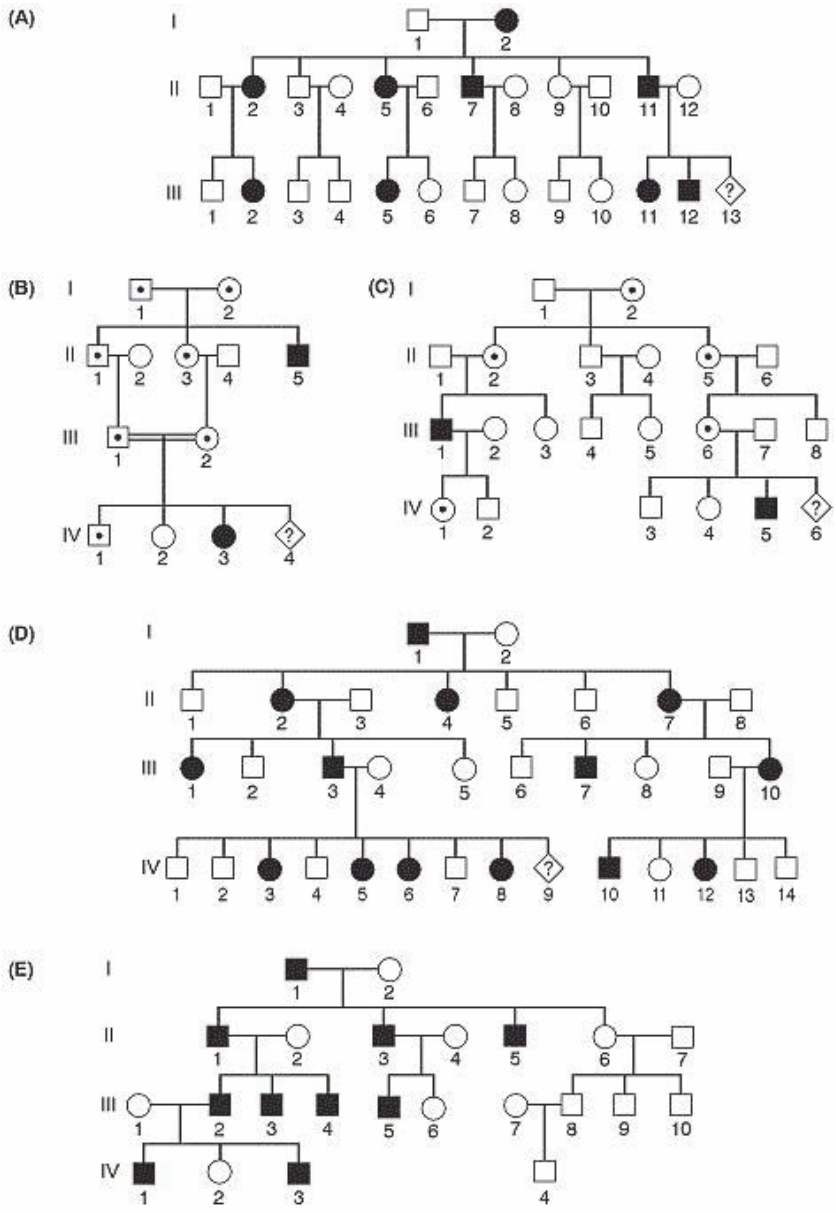


Figure 1.4: Basic mendelian pedigree patterns. **(A)** Autosomal dominant; **(B)** autosomal recessive; **(C)** X-linked recessive; **(D)** X-linked dominant; **(E)** Y-linked. The risk for the individuals marked with a query are **(A)** 1 in 2, **(B)** 1 in 4, **(C)** 1 in 2 males or 1 in 4 of all offspring, **(D)** negligibly low for males, 100% for females (Strachan and Read, 1999).

1.1.2. Risk of Breast and Ovarian Cancer Genes, *BRCA1* and *BRCA2*

More than 20,000 genes are encoded in the human genome. That doesn't imply that we know the purpose of all of these genes. We do know that over thirty of these genes are classified as tumour suppressors. The normal functions of these genes include repair of DNA, induction of programmed cell death (apoptosis), and prevention of abnormal cell division which prevents cancer formation. In contrast to proto-oncogenes, in tumour suppressors it is loss-of-function mutations that contribute to the progression of cancer. This means that tumour suppressor mutations tend to be recessive and thus both alleles must be mutated in order to allow an abnormal growth to proceed. It is perhaps not surprising that mutations in tumour suppressor genes are more likely than oncogenes to be inherited (Strachan and Read, 1999).

BRCA1 and *BRCA2* are large genes that encode tumour suppressor proteins. As said before, these proteins help repair damaged DNA and therefore play a role in ensuring the stability of the cell's genetic material. When either of these genes is mutated, or altered, such that its protein product is not made or does not function correctly, DNA damage may not be repaired properly. As a result, cells are more likely to develop additional genetic alterations that can lead to cancer.

Figure 1.5: *BRCA1* and *BRCA2* representation relative to their mutation frequency (extracted from http://www.cancer.gov).

Specific inherited mutations in *BRCA1* and *BRCA2* increase the risk of female breast and ovarian cancers (King *et al.*, 2003; Bordeleau *et al.*, 2010; Pruthi *et al.*, 2010), and they have been associated with increased risks of several additional types of cancer. Together, *BRCA1* and *BRCA2* mutations account for about 20-25% of hereditary breast cancers (Easton, 1999) and about 5 - 10% of all breast cancers (Campeau *et al.*, 2008). In addition, mutations in *BRCA1* and *BRCA2* account for around 15% of ovarian cancers overall (Pal *et al.*, 2005). Breast cancers

associated with *BRCA1* and *BRCA2* mutations tend to develop at younger ages than sporadic breast cancers.

Moreover, a harmful *BRCA1* or *BRCA2* mutation can be inherited from a person's mother or father. Each child of a parent who carries a mutation in one of these genes has a 50% chance of inheriting the mutation. The effects of mutations in *BRCA1* and *BRCA2* can be seen even when a person's second copy of the gene is normal.

1.1.2.1. Other cancers linked to mutations in BRCA1 and BRCA2 genes

Harmful *BRCA1* mutations may increase a woman's risk of developing fallopian tube cancer and peritoneal cancer (Brose *et al.*, 2002; Finch *et al.*, 2006). Men with *BRCA2* mutations, and to a lesser extent *BRCA1* mutations, are also at increased risk of breast cancer (Tai *et al.*, 2007). Men with harmful *BRCA1* or *BRCA2* mutations have a higher risk of prostate cancer (Levy-Lahad and Friedman, 2007). Men and women with *BRCA1* or *BRCA2* mutations may be at increased risk of pancreatic cancer (Ferrone *et al.*, 2009).

1.1.2.2. Patent implementation of BRCA1 and BRCA2

It was in 1990 when Mary-Claire King first announced the localization through linkage analysis of the *BRCA1* gene (Hall *et al.*, 1990). Its clinical importance was immediately acknowledged as being a gene associated with an increased risk for breast cancer. Four years later, in 1994, MYRIAD® Genetics (Salt Lake City, Utah, USA) was founded by scientists seeking breast cancer (BRCA) related genes. Later that year the *BRCA1* gene's entire sequence was isolated and its patent filed by MYRIAD® (US5747282). The very next year, MYRIAD® isolated the second BRCA related gene, the *BRCA2*, and filed its patent (US5867492) as well. In 1996, the first *BRCA1* and *BRCA2* analysis was introduced to the market.

Figure 1.6: Gene patent cartoon (credits to Cathy Wilcox).

The patent of these and other genes generated controversy especially when their licenses were used to apply abusive prices. At present, the entire genome can be sequenced for $5,826 (http://www.genome.gov) but the *BRCA1* and *BRCA2* analysis by MYRIAD® still comes with a price tag of over $6,000.

Although these patents would serve MYRIAD® for 20 years, in June of this year, in the 'Association for Molecular Pathology vs. Myriad Genetics (No. 12-398)', a case challenging the validity of gene patents in the United States, the US Supreme Court unanimously ruled that, "*A naturally occurring DNA segment is a product of nature and not patent eligible merely because it has been isolated.*" invalidating MYRIAD® patents on the *BRCA1* and *BRCA2* genes.

1.1.2.3. Traditional clinical screening of BRCA1 and BRCA2

Conventional protocols for mutation detection in the *BRCA1* and *BRCA2* genes involve PCR amplification of individual exons and Sanger sequencing of the products (Frank *et al.*, 1998; Walsh *et al.*, 2010). However, both *BRCA1* and *BRCA2* are very large genes, both encountering with more than 80,000 base pairs (bp) long sequences. As expected, screening for mutations in such large genes is rather complicated and even though there are well established clinical protocols for their genetic testing (Pruthi *et al.*, 2010), the task is nevertheless costly and time consuming as it results in over forty individual polymerase chain reactions (PCR) plus their purification and subsequent Sanger sequencing.

Additional strategies that employ massive sequencing technology have been reported to characterise variations in *BRCA1* and *BRCA2*. Walsh *et al.* uses a genomic DNA solution for the capture of both genes together with other genes involved in breast or ovarian cancer (Walsh *et al.*, 2010). Similarly, whole exome analysis involves DNA capture of all exonic sequences. However, these approaches have their limitations as they require large and expensive NGS platforms. Furthermore, DNA capture still generates libraries where a high percentage of the analysed sequences are not on the target.

Taking into account the above referred methods, this research work intends to develop a novel method to detect DNA genomic variants in large genes, such as *BRCA1* and *BRCA2*, which can be performed in normal genetics laboratories in hospitals.

1.1.3. Retinitis Pigmentosa

Retinitis Pigmentosa (RP; OMIM #268000) belongs to a group of human retinal dystrophies which are inherited ocular disorders characterised by a primary and progressive loss of photoreceptor cells leading to visual handicap. Monogenic retinal dystrophies are rare diseases (Ayuso and Millan, 2010) with RP being the most common form of this disease group. RP is characterised by primary degeneration of rod photoreceptors and it affects 1 in 3,000 to 5,000 people (Ammann *et al.*, 1965; Boughman *et al.*, 1980; Jay, 1982; Haim, 2002; Veltel *et al.*, 2008).

Typical symptoms include night blindness followed by visual fields decreased by the development of blind spots in the side (peripheral) vision. Over time, these blind spots merge leading to tunnel vision and eventually legal blindness or, in many cases, complete blindness (Hartong *et al.*, 2006).

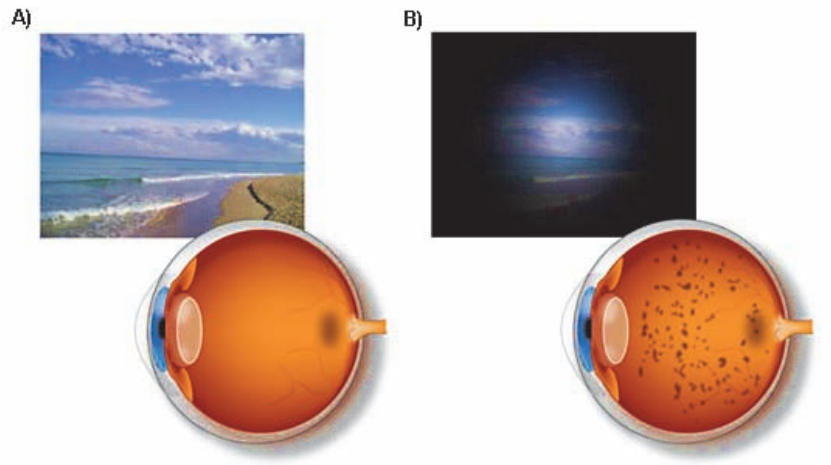


Figure 1.7: Landscape as seen by a healthy individual **(A)** and by a RP patient **(B)**. Below two cross sections of the eye are seen, corresponding to each of the cases. In the affected eye darker spots appear around the eyeball (edited from http://retinosis.umh.es/).

The diagnosis of retinitis pigmentosa relies mainly upon documentation of progressive loss in photoreceptor function by electroretinography (ERG) and visual field testing. Currently, there is no therapy that stops the progression of the disease or restores the vision. The therapeutic approach is restricted to slowing down the degenerative process by sunlight protection and vitamin-therapy, treating the complications (cataract and macular edema), and helping patients to cope with the social and psychological impact of blindness. Although at present the visual prognosis is poor, new therapeutic strategies are emerging from intensive research (gene therapy, neuroprotection, retinal prosthesis) (http://www.orpha.net/).

1.1.3.1. The Visual Cycle

In order to better comprehend how mutations in the genes involved in RP affect the activity of the proteins, a brief summary of the molecular bases behind vision is required. Hence, human vision begins when light enters the eye and is focused by the lens onto the photosensitive tissue at the back of the eye, the retina (Figure 1.8).

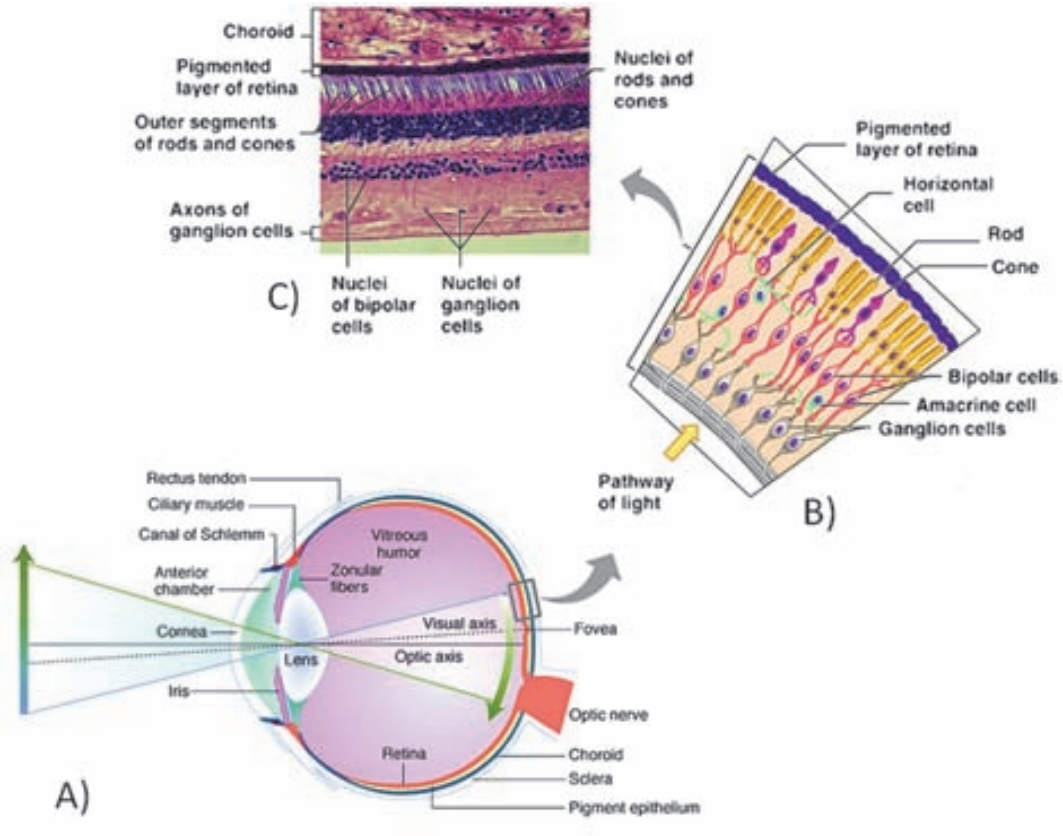


Figure 1.8: Basic anatomy of the human eye**(A)**; Cells composing the retina**(B)**; Haematoxylin-stained section of the human eye**(C)**.

Light-sensitive cells in the retina – the rod and cone photoreceptors – capture the incoming photons. The photoreceptors (PR) are polarised cells that consist of a synaptic region, a cell body, an inner segment (IS), and an outer segment (OS). The 'molecular machinery' involved in biosynthesis, energy metabolism, and membrane trafficking reside within the IS, which is connected to the OS *via* the so-called interconnecting cilium (IC). The OS is comprised of membranous discs surrounded by a plasma membrane. Over 90% of the total protein in the OS is rhodopsin, which resides both within the membrane of the discs and in the surrounding plasma membrane (Figure 1.10). Vision in vertebrates begins with the absorption of light by the rod photoreceptor 'visual pigment', rhodopsin (RHO), which consists of an apoprotein, opsin (a single 384 amino acid polypeptide chain), and a chromophore (11-cis retinaldehyde, derived from

vitamin A) attached to the opsin by a Schiff base bond. The absorption of light by rhodopsin causes isomerization of 11-cis retinal to all-trans retinal, which changes opsin conformation, thus initiating vision (Figure 1.9) (Ferrari *et al.*, 2011). All-trans retinal is released from rhodopsin, conjugated with the membrane lipid phosphatidylethanolamine and transported to the cytoplasm by ATP-binding cassette, subfamily A, member 4 (ABCA4). After modification to all-trans retinol by a retinol dehydrogenase (Figure 1.9, hydroxyl group shown as OH), it is transported to the retinal pigment epithelium (RPE), where it is esterified to a fatty acyl group (FA) by lecithin retinol acyltransferase (LRAT) to form all-trans retinyl ester. All-trans retinyl ester is subject to trans-isomerization to 11-cis retinal through the actions of two further enzymes (RPE65 and 11-cis retinol dehydrogenase). After transport back to the PR cell, 11-cis retinal binds rhodopsin, rendering it sensitive to light. Retinoid-binding proteins, such as interstitial retinol-binding protein (IRBP), cellular retinol-binding protein and cellular retinaldehyde-binding protein, are involved in the transport of the hydrophobic retinoids in an aqueous environment (Figure 1.9, show in purple) (Alan *et al.*, 2010).

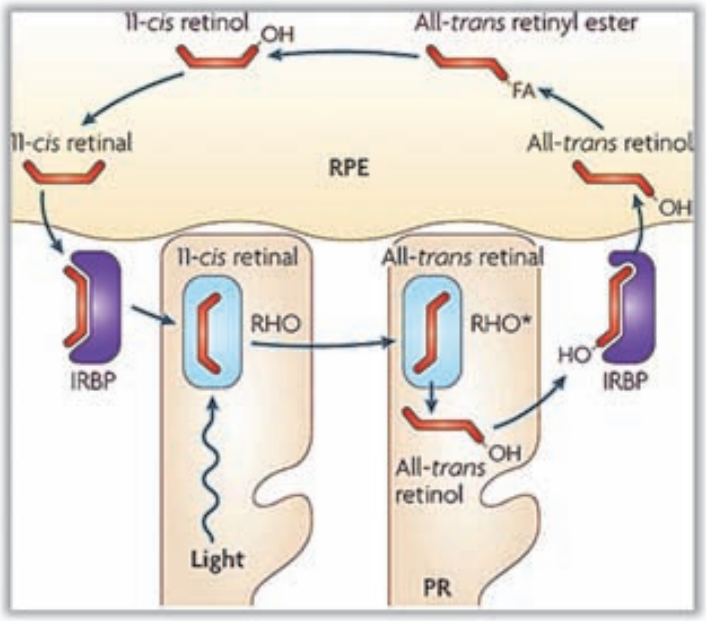


Figure 1.9: The visual cycle. Photoreceptor cells (PR); rhodopsin (RHO, RHO* when activated); retinal pigment epithelium (RPE); Fatty acyl group (FA); Interstitial retinol-binding protein (IRBP) (edited from Alan *et al.*, 2010).

1.1.3.2. Inheritance patterns of Retinitis Pigmentosa

The genetics of RP are complex as they are heterogeneous. Since the first report describing a linkage of an RP locus to a DNA marker on human chromosome X in 1984 by Bhattacharya *et al.*, over 50 genes have been associated with RP (http://sph.uth.edu/retnet/).

Non-syndromic or "simple" cases may be inherited as autosomal-recessive (ar), autosomal-dominant (ad), or X-linked traits (Ayuso *et al.*, 1995). For the Spanish population, a reliable estimate for the percentages of each inheritance pattern could be 15% for autosomal-dominant RP (adRP), 34% for autosomal-recessive RP (arRP), 7% for X-linked RP, and around 40% for syndromic RP (Ayuso and Millan, 2010) (Table 1.1). Also it is likely that more than 30% of all RP cases are sporadic and other roughly 5% could be early-onset and grouped as part of Leber congenital amaurosis (Ferrari *et al.*, 2011). In addition to the typical mendelian inheritance patterns other atypical and rather rare forms exist such as X-linked dominant, mitochondrial, and digenic (due to mutations in two different genes).

Table 1.1: Estimate of each RP inheritance pattern for the Spanish population (edited from Ayuso and Millan, 2010).

Non-syndromic RP (*n*)	adRP (%)	arRP (%)	xlRP (%)	Syndromic RP (%)
1,717	15	34	7	41 (3 unclassified)

While RP is a disease usually limited to the eye, there are syndromic forms involving multiple organs. Usher syndrome is the most common form of syndromic RP (where RP usually develops by the early teenage years) followed by Bardet-Biedl syndrome.

Because of the genetics behind RP are so complicated, a clear genotype-phenotype correlation is in most of the times not possible and the multiplicity of mutations makes it even harder. In fact, different mutations in the same gene may cause different diseases and the same mutation can exhibit intra- and inter-familial phenotypic variability (as for *BEST1* mutations) (Ferrari *et al.*, 2011). Then we have the vast majority of *RHO* mutations showing a classical autosomal dominant inheritance leading to RP and yet a few mutations show an autosomal recessive inheritance or lead to night blindness only (Berger *et al.*, 2010). Similarly, *RPGR*, the major X-linked recessive RP gene, is associated with mutations in male patients and no male-to-male transmission of the phenotype has ever been observed. However, families with dominant inheritance pattern and female carriers showing disease symptoms of variable degree have also been described; this could likely be due to the dominant nature of some of the mutations or non-random X-inactivation in the affected tissue (Berger *et al.*, 2010; Ayuso and Millan, 2010; Ferrari *et al.*, 2011). Lastly, mutations in some genes have been reported to be associated with incomplete penetrance and as a result making molecular diagnosis even more challenging (Ferrari *et al.*, 2011; Borràs and de Sousa Dias *et al.*, 2013).

1.1.3.3. Molecular genetics behind adRP

Many genes have been associated with adRP. To date mutations in over 20 different genes are reported to cause this disease (http://sph.uth.edu/retnet/). However, only four of these genes account with a relevant incidence; these genes are *RHO* (20-30%), *PRPH2* (5-10%), *PRPF31* (5-10%), and *RP1* (3-40%) (Fahim *et al.*, 2013) (Table 1.2; Figure 1.10).

Table 1.2: Genes associated with adRP (edited from Fahim *et al.*, 2013).

Gene	Estimated proportion of adRP attributed to mutations in this gene	Protein	OMIM
RHO	20%-30% (Daiger *et al.*, 2008)	Rhodopsin	180380, 613731
PRPF31	5%-10% (Daiger *et al.*, 2008)	U4/U6 small nuclear ribonucleoprotein Prp31	600138, 606419
PRPH2	5%-10% (Daiger *et al.*, 2008)	Peripherin-2	179605, 608133
RP1	3%-4% (Daiger *et al.*, 2008)	Oxygen-regulated protein 1	180100, 603937
IMPDH1	2%-3% (Daiger *et al.*, 2008)	Inosine-5'-monophosphate dehydrogenase 1	146690, 180105,
PRPF8	2%-3% (Daiger *et al.*, 2008)	Pre-mRNA-processing-splicing factor 8	600059, 607300
KLHL7	1%-2%	Kelch-like protein 7	611119, 612943
NR2E3	1%-2% (Daiger *et al.*, 2008)	Photoreceptor-specific nuclear receptor	604485, 611131
CRX	1% (Daiger *et al.*, 2008)	Cone-rod homeobox protein	120970, 602225,
PRPF3	1% (Daiger *et al.*, 2008)	U4/U6 small nuclear ribonucleoprotein Prp3	601414, 607301
TOPORS	1% (Bowne *et al.*, 2008)	E3 ubiquitin-protein ligase Topors	609507, 609923
CA4	Rare (Daiger *et al.*, 2008)	Carbonic anhydrase 4	600852, 114760
NRL	Rare (Daiger *et al.*, 2008)	Neural retina-specific leucine zipper protein	162080, 613750
ROM1	Rare (Daiger *et al.*, 2008)	Retinal outer segment membrane protein 1	180721
RP9	Rare (Daiger *et al.*, 2008)	Retinitis pigmentosa 9 protein	180104, 607331
RDH12	unknown	Retinol dehydrogenase 12	608830, 612712
SNRNP200	unknown	U5 small nuclear ribonucleoprotein 200 kDa helicase	601664, 610359
AIPL1	Rare (Sohocki *et al.*, 2000)	Aryl-hydrocarbon-interacting protein-like 1	604392
BEST1	Rare (Davidson *et al.*, 2009)	Bestrophin- 1	607854, 613194
PRPF6	Rare (Tanackovic *et al.*, 2011)	Pre-mRNA-processing factor 6	613979, 613983
RPE65	Rare (Bowne *et al.*, 2011)	Retinoid isomerohydrolase	180069, 613794
GUCA1B	4%-5% in Japan; rare in UK	Guanylyl cyclase-activating protein 2	602275, 613827
FSCN2	3% of Japanese with adRP; otherwise rare (Daiger *et al.*, 2008)	Fascin-2	607643, 607921
SEMA4A	3%-4% in Pakistan	Semaphorin-4A	607292, 610282

Data are compiled from the following standard references: gene symbol from HGNC; OMIM numbers from OMIM; protein name from UniProt.

- *Rhodopsin*

Rhodopsin (*RHO*) is the first component of the visual transduction pathway and is activated by absorption of light in the rod photoreceptor cells of the retina (Murray *et al.*, 2009). More than 100 *RHO* mutations have been reported; one, NM_000539.3: c.68C>A (NP_000530.1:p.Pro23His) associated with distinct sectorial disease, is found in approximately 12%-14% of Americans of European origin who have adRP (Sullivan *et al.*, 2006).

- *Peripherin 2*

The *Peripherin 2* gene (*PRPH2*), formerly known as the *retinal degeneration slow gene* (*RDS*), consists of 3 exons and encodes a 39-kDa integral membrane glycoprotein with 346 amino acids. The protein includes 4 transmembrane domains (M1-M4) and a large intradiscal domain (known as the D2 loop) and is located at the outer segment discs of the rod and cone photoreceptors. Mutations in this gene are associated with clinical phenotypes ranging from RP to macular degeneration to complex maculopathies.

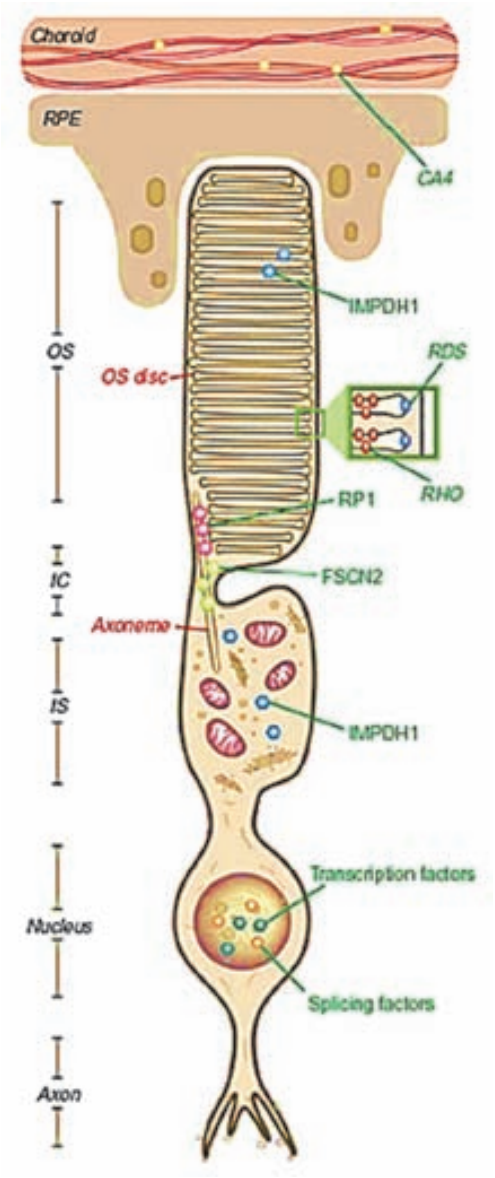


Figure 1.10: Rod cell in detail to the expression of several genes associated with adRP. The figure shows the distribution of different proteins throughout the photoreceptor, retinal pigment epithelium (RPE) and choroid. (OS) outer segment (IC) interconnecting cilium, (IS) inner segment (edited from Kennan *et al.*, 2005).

- *Pre-mRNA Processing Factor 31*

Pre-mRNA Processing Factor 31 (*PRPF31*) is a pre-mRNA splicing factor of 61kDa, which is integral to the U4/U6+U5 trimer. To date over 40 mutations have been located in different parts of the gene and comprise missense substitutions, splice-site mutations, deletions, and insertion. Mutation of *PRPF31*, previously thought to account for 5%-6% of adRP, is now known to account for 8% of adRP because 2.5% of adRP is caused by genomic rearrangements of this gene that are detected using deletion/duplication analysis rather than sequence analysis.

- *Retinitis Pigmentosa 1*

The Retinitis Pigmentosa 1 (RP1) gene encodes a 240-kD retinal photoreceptor-specific protein. RP1 is expressed prominently in the photoreceptor cells of the retina and is involved in the correct orientation and higher order stacking of outer segment discs. RP1 is a microtubule-associated protein forming part of the photoreceptor axoneme and thus plays an important role in photoreceptor function (Liu *et al.*, 2003, 2004). Of the RP1 known mutations, two account for half of adRP caused by an RP1 mutation: c.2029C>T (p.Arg677X) and c.2285_2289delTAAAT (p.Leu762Tyrfs*17); reference sequences NM_006269.1 NP_006260.1.

1.1.3.4. Current methods for molecular diagnosis of adRP

Due to the genetic heterogeneity of adRP, an accurate molecular diagnosis is challenging. Even with decades of improvement, the current diagnostics including Sanger sequencing and Asper Biotech (APEX) array still have many limitations (Ferrari *et al.*, 2011; Zaneveld *et al.*, 2013).

Since mutations in over 20 genes can cause adRP (Bowne *et al.*, 2011), Sanger sequencing, while highly accurate, requires that separate PCRs be performed for each region of interest which is costly and time-consuming when hundreds of PCR are required (de Sousa Dias *et al.*, 2012). This makes it impractical for testing such a large number of genes in a large numbers of patients.

The recently developed APEX array enables the analysis of 414 mutations in 16 genes: *CA4*, *FSCN2, IMPDH1, NRL, PRPF3, PRPF31, PRPF8, RDS, RHO, ROM1, RP1, RP9, CRX, TOPORS, KLHL7,* and *PNR* (http://www.asperbio.com/). However, it is designed to efficiently detect known mutations in known genes; with a similar array, a typical genetic diagnostic rate of less than 15 % was achieved (Avila-Fernandez *et al.* 2010).

In contrast, next generation sequencing (NGS) technology provides a new approach for molecular diagnosis of RP. Several recent studies reported the new NGS-based molecular diagnosis of RP (Simpson *et al.* 2011; Neveling *et al.* 2012; O'Sullivan *et al.* 2012; Shanks *et al.* 2012; Glockle *et al.* 2013; de Sousa Dias *et al.*, 2013; Borràs and de Sousa Dias *et al.*, 2013). Their results achieved a superior genetic diagnostic rate when compared to conventional methods. NGS also allows sequencing other candidate retinal disease genes in addition to RP disease genes without significantly increasing the cost. In addition, screening many patients in parallel is also possible again without significantly increasing the costs.

This research work targets the development of a novel, simple, and effective method for detecting DNA genomic variants in several genes associated with adRP which could then be incorporated into a molecular testing routine without the use of special equipment.

1.2. DNA Sequencing Technologies for Molecular Diagnosis

Historically speaking, certainly one of the most important events for DNA sequencing was the discovery of the DNA Double-helix structure in 1953 by two researchers, namely James Watson and Francis Crick (Watson and Crick, 1953).

Naturally, credits should also be given to Rosalind Franklin and Maurice Wilkins who produced the first X-ray crystallographic images that proved essential to solving the mystery of DNA's molecular structure. *"Photo 51"* was the finest on this lot (Figure1.11).

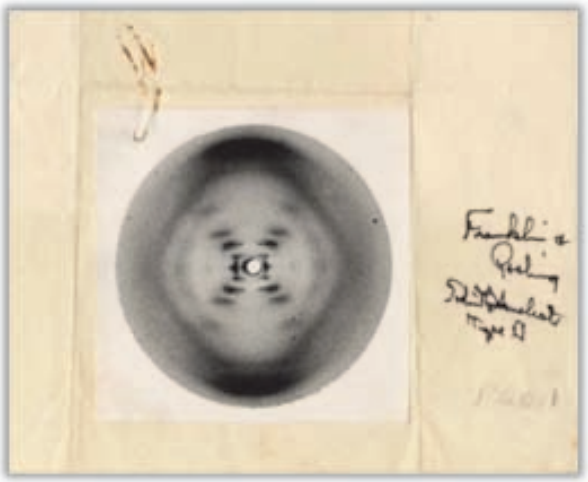


Figure 1.11: X-ray crystallographic image of DNA. *Photo 51*, taken by Rosalind E. Franklin and R.G. Gosling. Linus Pauling's holographic annotations are to the right of the photo. May 2, 1952.

DNA Sequencing entails several techniques and methods that are used to determine the sequence of the aforementioned nucleotide bases in a DNA molecule. The first DNA sequencing techniques were introduced in the 1970s and since then DNA sequencing has come a long way. The understanding of DNA sequences has become an integral part of basic molecular biology and has been applied to molecular diagnosis and biomedical research, allowing scientists to better understand genes and their role in the creation of the human body.

1.2.1. The First Generation

It was in 1975 when Sanger first introduced the concept of DNA sequencing method in his pioneering Croonian lecture (Sanger, 1975). Shortly thereafter he published a rapid method for determining sequences in DNA by primed synthesis with DNA polymerase (Sanger and Coulson, 1975). In the year of 1977, two breakthrough articles for DNA sequencing were published: the Frederick Sanger's enzymatic dideoxy DNA sequencing technique, later known as Sanger method (Figure 1.12), based on the chain-terminating dideoxynucleotide analogues (Sanger *et al.*, 1977) and the Allan Maxam and Walter Gilbert's chemical degradation DNA sequencing technique in which terminally labelled DNA fragments were chemically cleaved at specific bases and separated by gel electrophoresis (Maxam and Gilbert, 1977). These breakthrough findings

earned Sanger his second Nobel Prize in Chemistry in 1980, which he shared with Walter Gilbert and Paul Berg.

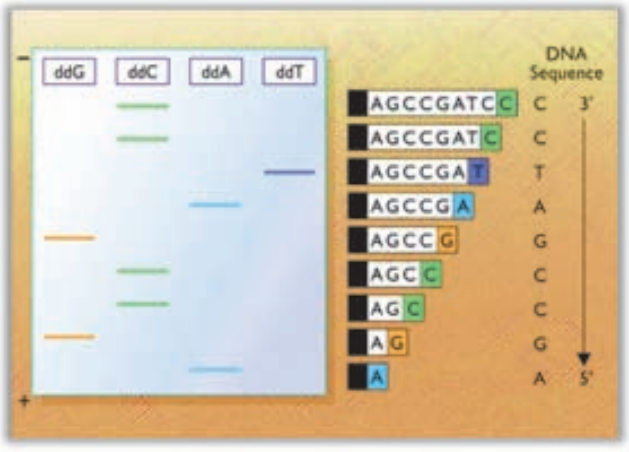


Figure 1.12: The Sanger (chain-termination) method for DNA sequencing.

These two noticeable laboratories (Sanger's and Maxam-Gilbert's) were then responsible for an alternative to the labelling of the primer by labelling the terminators instead. This is commonly known as 'dye terminator sequencing' and its major advantage is that the complete sequencing set can be done in a single reaction, rather than the four reactions required for the labelled-primer method. Owing to its greater practicality and speed, dye-terminator sequencing represented the basis for the introduction of the first automated DNA sequencers led by Caltech (Smith *et al.* 1986), which was subsequently commercialised by Applied Biosystems (ABI), the European Molecular Biology Laboratory (EMBL) (Ansorge *et al.* 1986, 1987), and Pharmacia-Amersham, later General Electric (GE) healthcare.

In 1996, ABI introduced the first commercial DNA sequencer, the ABI Prism 310, which utilised a slab gel electrophoresis. Two years later, the considerable labour of pouring slab gels was replaced with automated reloading of the capillaries with polymer matrix by the ABI Prism 3700 with its 96 capillaries.

This automated DNA sequencer was employed for the successful sequencing of the first human genome in 2003 after a 13-year, $2.7 billion world-wide effort of the human genome project consortium. It is important to recognise that all of this progress has been made using methods that are simply enhancements of the basic 'dideoxy' method introduced by Sanger in 1977 (Parrek *et al.*, 2011).

1.2.2. The Next Generation

The automated Sanger method is regarded as a 'first-generation' sequencing technology; newer methods are referred to as next-generation sequencing (NGS). These newer technologies constitute various strategies that rely on a combination of template preparation, sequencing and imaging, and genome alignment and assembly methods. The arrival of NGS technologies in the marketplace has changed the way we think about scientific approaches in basic, applied, and clinical research. In some respects, the potential of NGS is akin to the early days of PCR, with one's imagination being the primary limitation to its use. The major advance offered by NGS is the ability to produce an enormous volume of data cheaply - in some cases in excess of one billion short reads per instrument run (Metzker, 2010).

The GS 20 instrument, developed by 454 Life Sciences (http://www.454.com/), was the first commercially available NGS platform. It was introduced in 2005, just two years after the completion of the Human Genome Project.

1.2.2.1. Overview of the Second Generation Sequencing leading Platforms

The second generation of NGS platforms can generate about five hundred million bases of raw sequence (Roche) to billions of bases in a single run (Illumina, SOLiD). These methods rely on parallel, cyclic interrogation of sequences from spatially separated clonal amplicons (26 µm oil-aqueous emulsion bead [Roche: pyrosequencing chemistry]; clonal bridge [Illumina: sequencing by reversible dye terminators], 1 µm clonal bead [SOLiD: sequencing by sequential ligation of oligonucleotide probes]). These three (Roche, Illumina and SOLiD) are currently the commercially leading second generation NGS platforms (Figure 1.13) (Parrek *et al.*, 2011).

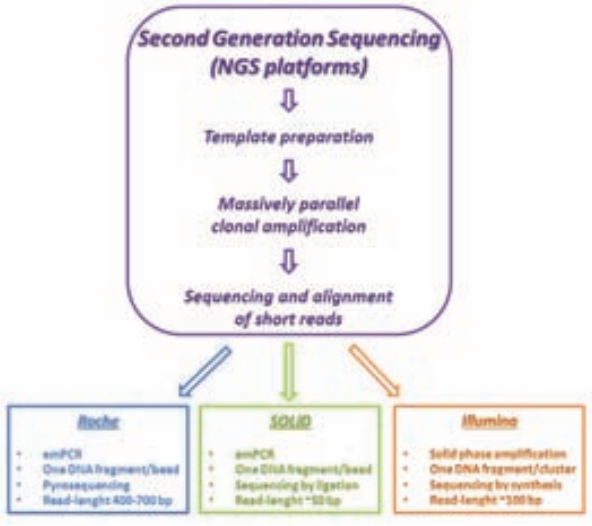


Figure 1.13: Advanced technological features of three leading second generation NGS platforms (edited from Pareek *et al.* 2011)

1.2.2.1.1. Roche GS FLX

Sharing the same technological principle as the above cited GS 20, the Roche GS FLX sequencing process consists of preparing an end-modified DNA fragment library, sample immobilization on streptavidin beads, and pyrosequencing:

Sample preparation of the GS FLX sequencing system begins with random fragmentation of genomic DNA into 300 - 800 base-pair (bp) fragments (Margulies *et al.*, 2005). After shearing, fragmented double-stranded DNA is repaired with an end-repair enzyme cocktail and adenine bases are added to the 3'-ends of fragments. Common adapters, named "A" and "B," are then nick-ligated to the fragments ends. The nicks present in the adapter-to-fragment junctions are filled in using a strand-displacing *Bst* DNA polymerase. Adapter "B" carries a biotin group, which facilitates the purification of homoadapted fragments (A/A or B/B). The biotin labelled sequencing library is captured on streptavidin beads. Fragments containing the biotin labelled B adapter are bound to the streptavidin beads while homozygous, nonbiotinylated A/A adapters are washed away. The immobilised fragments are denatured after which both strands of the B/B adapted fragments remain immobilised by the streptavidin–biotin bond and the single-strand template of the A/B fragments are freed and used in sequencing (Myllykangas *et al.*, 2012).

Clonal amplification of templates is done through emulsion Polymerase Chain Reaction (emPCR). The single-strand sequencing library is immobilised onto a specific DNA capture bead under conditions that favour one DNA molecule per bead (Figure 1.14). A one molecule per bead ratio is achieved by limiting dilutions. The bead-bound library is then amplified using a specific form of PCR. In emulsion PCR, parallel amplification of bead-captured library fragments takes place in a mixture of oil and water. Aqueous bubbles (or droplets) immersed in oil, form microscopic reaction entities for each individual capture bead. Hundreds of thousands of amplified DNA fragments can be immobilised on the surface of each bead (Figure 1.14).

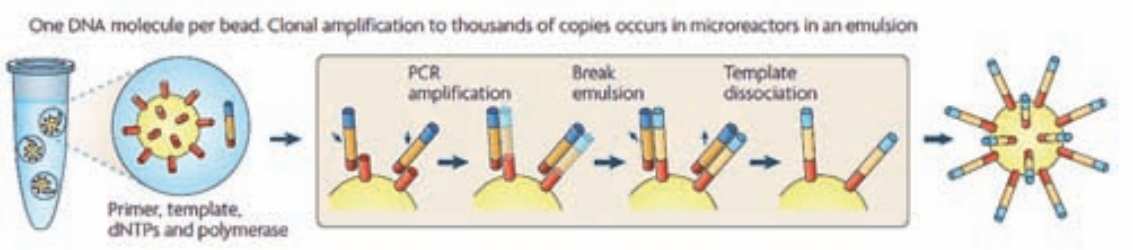
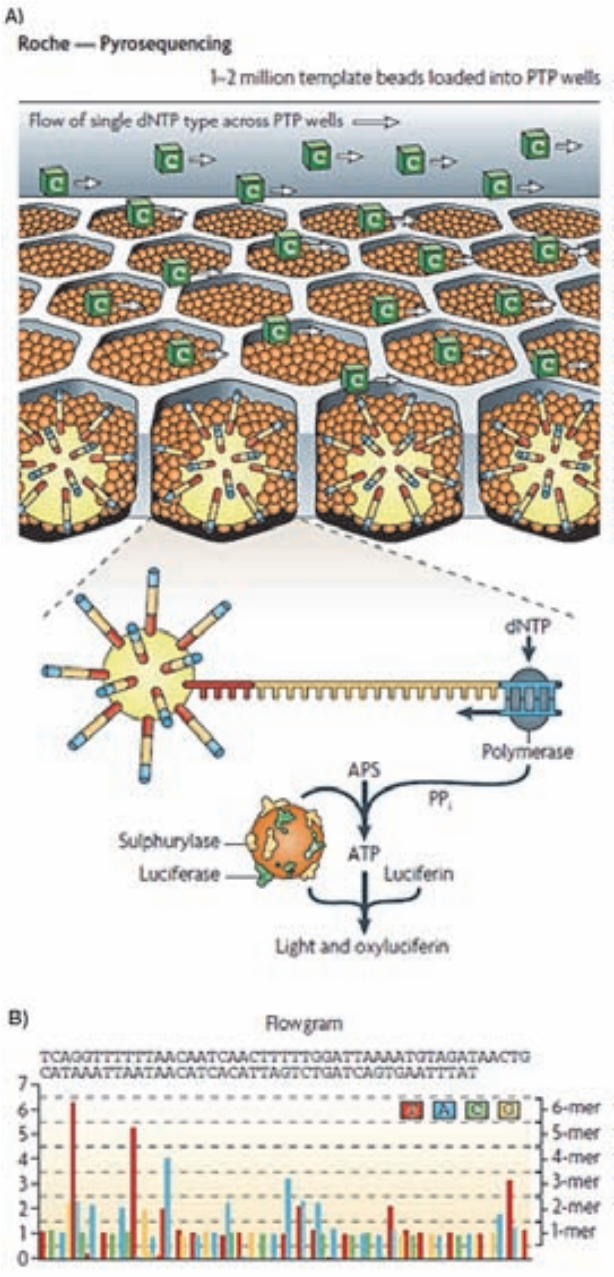


Figure 1.14: Template immobilization and amplification strategy by emulsion PCR (edited from Metzker, 2010).

In the GS FLX sequencing platform, beads covered with amplified DNA can be immobilised on a solid support (Figure1.15A). The GS FLX sequencing platform uses a "Picotiter plate," a solid

phase support containing over a million picoliter volume wells (Margulies *et al.*, 2005). The dimensions of the wells are such that only one bead is able to enter each position on the plate. Sequencing chemistry flows through the plate and insular sequencing reactions take place inside the wells (Myllykangas *et al.*, 2012).



Pyrosequencing is a non-electrophoretic, bioluminescence method that measures the release of inorganic pyrophosphate by proportionally converting it into visible light using a series of enzymatic reactions (Ronaghi *et al.*, 1996, 1998). The GS FLX sequencing reaction utilizes pyrosequencing (Figure 1.15B) to detect the base incorporation events during sequencing (Margulies *et al.* 2005). In pyrosequencing, Picotiter plates are flushed with nucleotides and the activity of DNA polymerase and the incorporation of a nucleotide lead to the release of a pyrophosphate. ATP sulfurylase and luciferase enzymes convert the pyrophosphate into a visible burst of light which is detected by a high-resolution charge-coupled device (CCD) imaging system. Each nucleotide species (dATP, dCTP, dGTP, and dTTP) is washed over the Picotiter plate and interrogated separately for each sequencing cycle (Myllykangas *et al.*, 2012).

Figure 1.15: **(A)** Pyrosequencing using Roche GS FLX titanium platform. Following loading of the DNA-amplified beads into individual PicoTiterPlate (PTP) wells, additional beads, coupled with sulphurylase and luciferase, are added. In this example, a single type of 2'-deoxyribonucleoside triphosphate (dNTP) — cytosine — is shown flowing across the PTP wells. The fibre-optic slide is mounted in a flow chamber, enabling the delivery of sequencing reagents to the bead-packed wells. The underside of the fibre-optic slide is directly attached to a high-resolution charge-coupled device (CCD) camera, which allows detection of the light generated from each PTP well undergoing the pyrosequencing reaction. **(B)** The light generated by the enzymatic cascade is recorded as a series of peaks called a flowgram. PPi, inorganic pyrophosphate (edited from Metzker, 2010).

Unlike other sequencing approaches that use modified nucleotides to terminate DNA synthesis, the pyrosequencing method manipulates DNA polymerase by the single addition of a dNTP in limiting amounts; upon incorporation of the complementary dNTP, DNA polymerase extends the primer and pauses. DNA synthesis is reinitiated following the addition of the next complementary dNTP in the dispensing cycle. The order and intensity of the light peaks are recorded as flowgrams (Figure 1.15B), which reveal the underlying DNA sequence (Metzker, 2009). However, multiple nucleotides can be incorporated to the extending DNA strand and accurate sequencing through homopolymer stretches represents a challenging technical issue for Roche GS FLX. For homopolymeric repeats of up to six nucleotides, the number of dNTPs added is directly proportional to the light signal. Insertions are the most common error type, followed by deletions.

1.2.2.1.2. Illumina Genome Analyser

The Illumina system (Bentley *et al.*, 2008) is based on immobilizing linear sequencing library fragments using solid support amplification. DNA sequencing is enabled using fluorescent reversible terminator nucleotides.

Sample preparation for the Illumina Genome Analyser involves adding specific adapter sequences to the ends of DNA molecules (Bentley, 2006; Bentley *et al.*, 2008). The production of a sequencing library also initiates with fragmentation of the DNA sample, which defines the molecular entry points for the sequencing reads. Then, an enzyme cocktail repairs the staggered ends, after which, adenines (A) are added to the 3-ends of the DNA fragments. A-tailed DNA is applied as a template to ligate double strand, partially complementary adapters to the DNA fragments. Adapted DNA library is size selected and amplified to improve the quality of sequence reads. Amplification introduces end-specific PCR primers that bring in the portion of the adapter required for sample processing on the Illumina system (Myllykangas *et al.*, 2012).

Solid-phase amplification is used to produce randomly distributed, clonally amplified clusters from fragment or mate-pair templates on a glass slide (Figure 1.16). High-density forward and reverse primers are covalently attached to the slide which creates an ultra-dense primer field. The sequencing library is immobilised on the surface of a flow cell (Figure 1.16). The immobilised primers on the flow cell surface have sequences that correspond to the DNA adapters present in the sequencing library. DNA molecules in the sequencing library hybridise to the immobilised primers and function as templates in strand extension reactions that generate immobilised copies of the original molecules. The primer functionalised flow cell surface serves as a support for amplification of the immobilised sequencing library by a process also known as “Bridge-PCR” (Figure 1.16). In the Illumina Bridge-PCR system, amplification is performed on a solid support

using immobilised primers and in isothermal conditions using reagent flush cycles of denaturation, annealing, extension, and wash. Bridge-PCR initiates by hybridisation of the immobilised sequencing library fragment and a primer to form a surface-supported molecular bridge structure. Arched molecule is a template for a DNA polymerase-based extension reaction. The resulting bridged double-strand DNA is freed using a denaturing reagent. Repeated reagent flush cycles generate groups of thousands of DNA molecules, also known as "clusters," on each flow cell lane. DNA clusters are finalised for sequencing by unbinding the complementary DNA strand to retain a single molecular species in each cluster, in a reaction called "linearization," followed by blocking the free 3'-ends of the clusters and hybridising a sequencing primer.

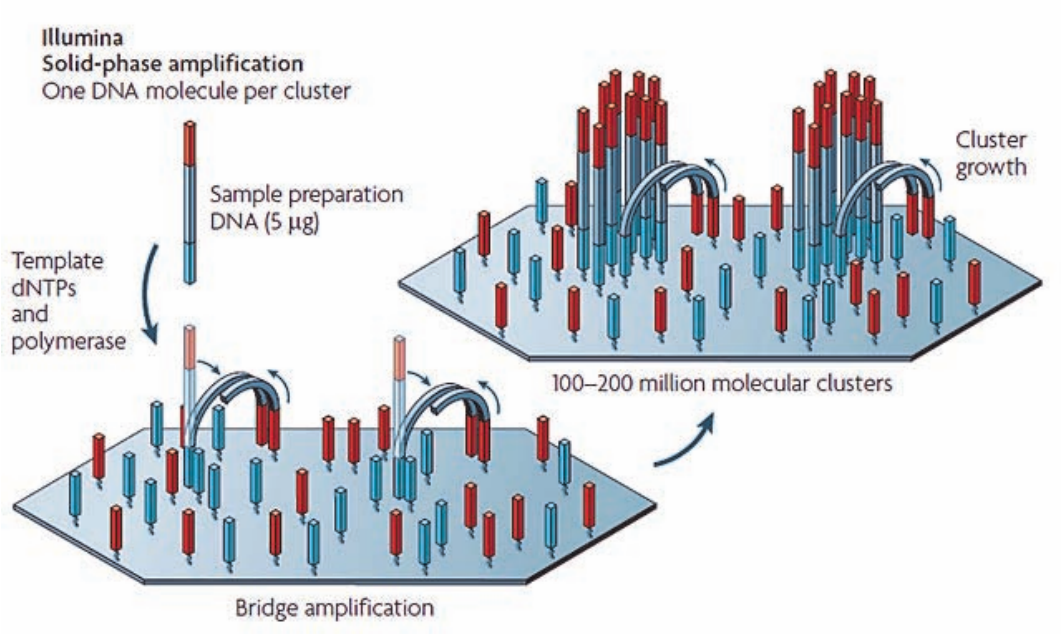


Figure 1.16: Template immobilization strategy in solid-phase amplification, which is composed of two basic steps: initial priming and extending of the single-stranded, single-molecule template, and bridge amplification of the immobilised template with immediately adjacent primers to form clusters (edited from Metzker, 2010).

Cyclic reversible termination is the method employed by Illumina for sequencing-by-synthesis. It uses reversible terminators in a cyclic method that comprises nucleotide incorporation, fluorescence imaging, and cleavage (Metzker, 2005). In the first step, a DNA polymerase, bound to the primed template, adds or incorporates just one fluorescently modified nucleotide, which represents the complement of the template base; termination of DNA synthesis after the addition of a single nucleotide is an important feature. Following incorporation, the remaining unincorporated nucleotides are washed away and imaging is performed to determine the identity of the incorporated nucleotide. This is followed by a cleavage step, which removes the terminating/inhibiting group and the fluorescent dye. Additional washing is performed before

starting the next incorporation step. Figure 1.17A depicts the four-colour cycle used by Illumina. The four colours are detected by total internal reflection fluorescence (TIRF) imaging using two lasers; the output of which is depicted in Figure 1.17B.The slide is partitioned into eight channels, which allows independent samples to be run simultaneously (Metzker, 2009). The synchronous extension of the sequencing strand by one nucleotide per cycle ensures that homopolymer stretches can be accurately sequenced. However, failure to incorporate a nucleotide during a sequencing cycle results in off-phasing effect – some molecules are lagging in extension and the generalised signal derived from the cluster deteriorates over cycles. Therefore, Illumina sequencing accuracy declines as the read length increases, which limits this technology to short sequence reads (Myllykangas *et al.*, 2012).

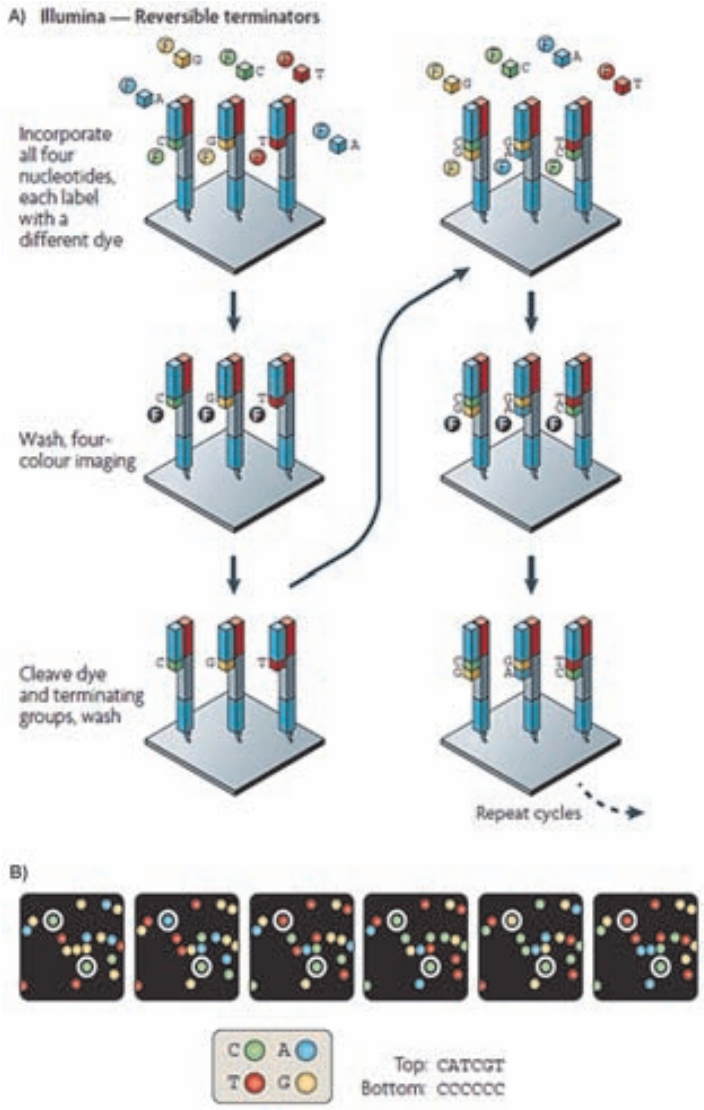


Figure 1.17: Four-colour cyclic reversible termination methods. **(A)** The four-colour cyclic reversible termination method uses Illumina's 3'-O-azidomethyl reversible terminator chemistry (Barnes *et al.*, 2002; Bentley *et al.*, 2008), using solid-phase-amplified template clusters. Following imaging, a cleavage step removes the fluorescent dyes and regenerates the 3'-OH group using the reducing agent tris(2-carboxyethyl)phosphine (Bentley *et al.*, 2008). **(B)** The four-colour images highlight the sequencing data from two clonally amplified templates (edited from Metzker, 2010).

1.2.2.1.3. ABI SOLiD Sequencer

The Applied Biosystems SOLiD sequencer (Valouev *et al.*, 2008; Smith *et al.*, 2010), is based on the Polonator technology (Shendure *et al.* 2005), an open source sequencer that utilizes

emulsion PCR to immobilise the DNA library onto a solid support and cyclic sequencing-by-ligation chemistry.

Sample preparation for SOLiD system again involves fragmentation of the DNA sample to an appropriate size range (400 – 850 bp), followed by end repair and ligation of "P1" and "P2" DNA adapters to the ends of the library fragments (Valouev *et al.* 2008). Emulsion PCR is applied to immobilise the sequencing library DNA onto "P1" coated paramagnetic beads. High-density, semi-ordered polony arrays are generated by functionalizing the 3'-ends of the templates and immobilizing the modified beads to a glass slide.

Sequencing by ligation is another cyclic method that differs from cyclic reversible termination in its use of DNA ligase (Tomkinson *et al.*, 2006) and either one-base-encoded probe or two-base-encoded probes. The ABI SOLiD platform method uses two-base-encoded probes (Valouev *et al.*, 2008), which has the primary advantage of improved accuracy in colour calling and SNV calling, the latter of which requires an adjacent valid colour change. Colour space is a unique feature of the SOLiD system. A sequencing primer is hybridised to the "P1" adapter in the immobilised beads; similar to the Roche GS FLX platform, the templates are amplified by emPCR (Figure 1.14). The SOLiD cycle of 1,2-probe hybridisation and ligation, imaging, and probe cleavage is repeated ten times to yield ten colour calls spaced in five-base intervals (Figure 1.18A). The extended primer is then stripped from the solid-phase-bound templates. A second ligation round is performed with an 'n – 1' primer, which resets the interrogation bases and the corresponding ten colour calls one position to the left. Ten ligation cycles ensue, followed by three more rounds of ligation cycles. Colour calls from the five ligation rounds are then ordered into a linear sequence (that is, the colour space) and aligned to a reference genome to decode the DNA sequence (Figure 1.18A). SOLiD uses two slides per run; each can be partitioned into four or eight regions called spots. The most common error type in SOLiD platforms is substitutions. Similar to the genome analysis of Illumina reads, SOLiD data have also revealed an underrepresentation of AT-rich and GC-rich regions (Harismendy *et al.*, 2009). Moreover, since the ligation-based method in the SOLiD system requires complex panel of labelled oligonucleotides and sequencing proceeds by off-set steps, the interpretation of the raw data requires a complicated algorithm (Valouev *et al.* 2008). However, the SOLiD system achieves a slightly better performance in terms of sequencing accuracy due to the redundant sequencing of each base twice by a dinucleotide detection core structure of the octamer sequencing oligonucleotides.

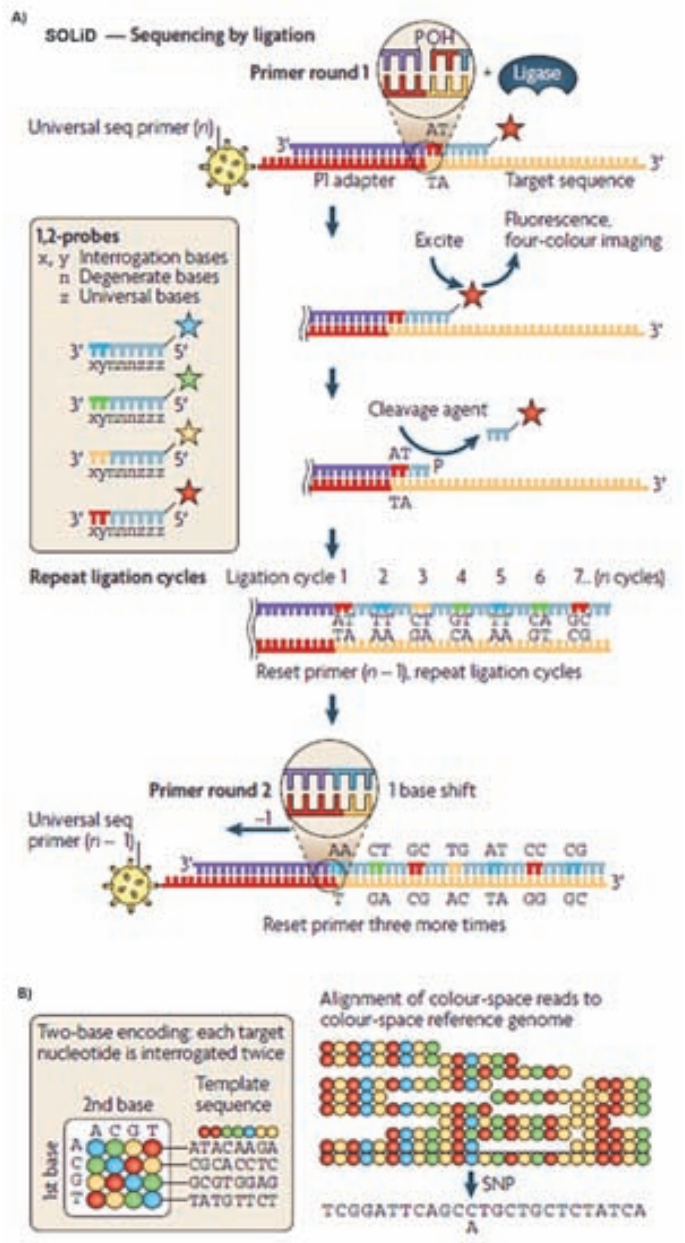


Figure 1.18: Sequencing by ligation. **(A)** four-colour sequencing by ligation method using SOLiD platform is shown. Upon the annealing of a universal primer, a library of 1,2-probes is added. Unlike polymerization, the ligation of a probe to the primer can be performed bi-directionally from either its 5'-PO4 or 3'-OH end. Appropriate conditions enable the selective hybridisation and ligation of probes to complementary positions. Following four-colour imaging, the ligated 1,2-probes are chemically cleaved with silver ions to generate a 5'-PO4 group. The SOLiD cycle is repeated nine more times. The extended primer is then stripped and four more ligation rounds are performed, each with ten ligation cycles. The 1,2-probes are designed to interrogate the first (x) and second (y) positions adjacent to the hybridised primer, such that the 16 dinucleotides are encoded by four dyes (coloured stars). The probes also contain inosine bases (z) to reduce the complexity of the 1,2-probe library and a phosphorothiolate linkage between the fifth and six nucleotides of the probe sequence, which is cleaved with silver ions (McKernan *et al.*, 2005). Other cleavable probe designs include RNA nucleotides (Macevicz, 1995; Mir *et al.*, 2009) and internucleosidic phosphoramidates (Macevicz, 1995), which are cleaved by ribonucleases and acid, respectively. **(B)** Two-base encoding scheme in which four dinucleotide sequences are associated with one colour (for example, AA, CC, GG and TT are coded with a blue dye). Each template base is interrogated twice and compiled into a string of colour-space data bits. The colour-space reads are aligned to a colour-space reference sequence to decode the DNA sequence (edited from Metzker, 2010).

1.2.2.2. Towards a third generation?

For the second generation sequencing platforms, or the first real massive parallel generation sequencing, the principle was based on the clonal amplification of DNA fragments, to make the light signal strong enough for reliable base detection by the CCD cameras. The issue here is the PCR amplification; while PCR revolutionised DNA analysis when first invented (Mullis et al., 1986), in some cases it may introduce base sequence errors or favour certain sequences over others, thus changing the relative frequency and abundance of the various DNA fragments which existed before amplification. To overcome this, the ultimate miniaturization into the nanoscale and the minimal use of biochemicals would be achievable if the sequence could be determined directly from a single DNA molecule without the need for PCR amplification and its potential for distortion of abundance levels. This sequencing from a single DNA molecule is now known as the "third generation of NGS technology" (Schadt et al. 2010). The concept of sequencing-by-synthesis without a prior amplification step or single molecule sequencing is currently pursued by a number of companies and is described below (Parrek et al., 2011).

1.2.2.2.1. Heliscope™ single molecule sequencer

The first commercial single-molecule DNA sequencing system was developed by Braslavsky *et al.* 2003 and licensed in 2007 by Helicos biosciences. The principle of the Heliscope sequencer relies on "true single molecule sequencing" (tSMS) technology. The tSMS technology begins with DNA library preparation through DNA shearing and the addition of an poly(A) tail to generated DNA fragments (Ozsolak *et al.* 2010), followed by hybridization of DNA fragments to the poly(T) oligonucleotides which are attached to the flow cell and simultaneously sequenced in parallel reactions. The sequencing cycle consists of DNA extension with one (out of four) fluorescently labelled nucleotides, followed by nucleotide detection with the Heliscope sequencer. The subsequent chemical cleavage of fluorophores allows the next cycle of DNA elongation to begin with another fluorescently labelled nucleotide, which enables the determination of the DNA sequence (Harris *et al.* 2008).

1.2.2.2.2. Single molecule real time (SMRT™) sequencer

The next revolutionary technology to hit the NGS commercial sector is most likely to be the single molecule real-time (SMRT) sequencing. Pacific Biosciences is currently leading this effort with its PacBio RS II (Eid *et al.*, 2009). The principle of the SMRT sequencer relies on single molecule real-time sequencing by a synthesis method provided by a sequencing chip which contains thousands of the so-called zero-mode waveguides (ZMWs) (Korlach *et al.* 2008b; Levene *et al.*

2003). For library preparation, genomic DNA is randomly fragmented and end-repaired. Then, 3'-adenine is added to the fragmented genomic DNA which facilitates ligation of an adapter with a T overhang. Single DNA oligonucleotide, which forms an intramolecular hairpin structure, is used as the adapter. The primed single-molecule templates are structurally linear molecules but the bubble adapters create topologically circular molecules. The sequencing reaction of a primed single-molecule template is performed by a single DNA polymerase molecule, which is attached to the bottom of each ZMW so that each DNA polymerase resides within the detection zone of ZMW (Figure 1.19A). This tethered Phi29 polymerase is a highly productive strand-displacing enzyme capable of performing rolling cycle amplification or RCA. During the sequencing reaction, the DNA fragment is elongated by the anchored DNA polymerase with dNTP's that are fluorescently labelled (each nucleotide is labelled with a fluorophore of different colour) at the terminal phosphate moiety (Korlach *et al.* 2008a).

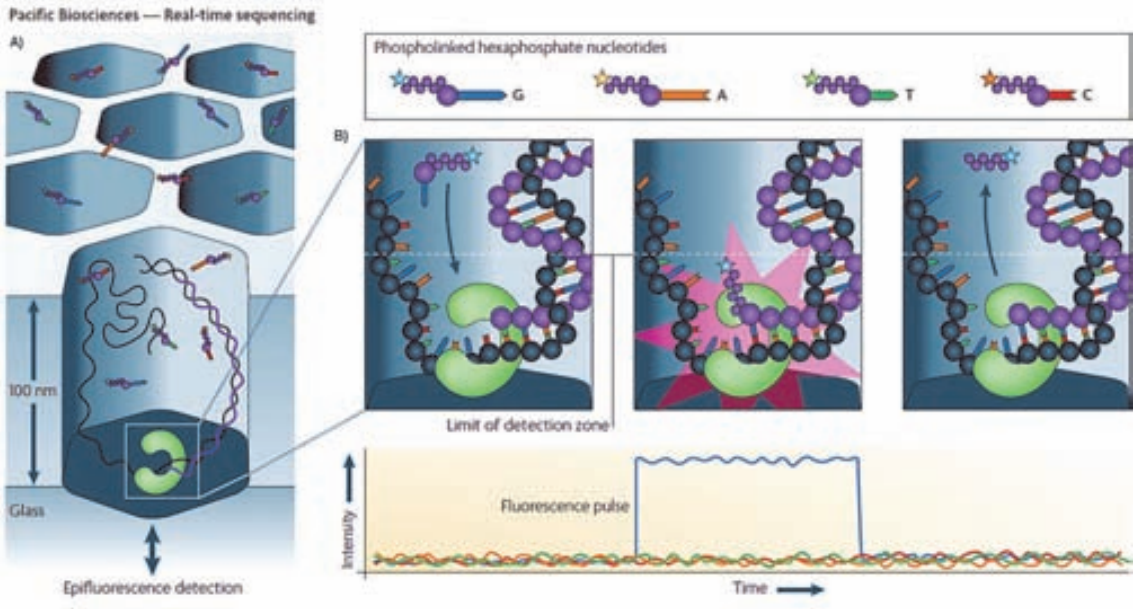


Figure 1.19: Real-time sequencing by Pacific Biosciences' four-colour real-time sequencing method. **(A)** The zero-mode waveguide (ZMW) design reduces the observation volume, therefore reducing the number of stray fluorescently labelled molecules that enter the detection layer for a given period. These ZMW detectors address the dilemma that DNA polymerases perform optimally when fluorescently labelled nucleotides are present in the micromolar concentration range, whereas most single-molecule detection methods perform optimally when fluorescent species are in the pico- to nanomolar concentration range (Metzker, 2009). **(B)** The residence time of phospholinked nucleotides in the active site is governed by the rate of catalysis and is usually on the millisecond scale. This corresponds to a recorded fluorescence pulse, because only the bound, dye-labelled nucleotide occupies the ZMW detection zone on this timescale. The released, dye-labelled pentaphosphate by-product quickly diffuses away, dropping the fluorescence signal to background levels. Translocation of the template marks the interphase period before binding and incorporation of the next incoming phospholinked nucleotide (edited from Metzker, 2010).

The DNA sequence is determined via CCD array on the basis of fluorescence nucleotide detection which is performed before nucleotide incorporation while the labelled dNTP forms a cognate association with the DNA template. The fluorescence pulse is stopped after phosphodiester bond formation, which causes the release of a fluorophore that diffuses out of ZMW (Figure 1.19B) (Levene *et al.* 2003; Eid *et al.* 2009). Thus, while the polymerase synthesizes a copy of the template strand, incorporation events of successive nucleotides are recorded in a movie-like format (Myllykangas *et al.*, 2012). Furthermore, the unique method of detecting nucleotide incorporation events in real time allows for the development of novel applications, such as the detection of methylated cytosines based on differential polymerase kinetics (Flusberg *et al.* 2010). Although the SMRT instrument PacBio RS II has only recently been made available to the market, the company claims that the SMRT analyser is capable of producing reads with average lengths of 4,200 to 8,500 bp, with the longest reads over 30,000 base pairs having a 99.99% accuracy (http://www.pacificbiosciences.com/).

1.2.2.2.3. Other emerging technologies

***Semiconductor Sequencing*:** Life Technology and Ion Torrent recently released the new Ion Personal Genome Machine (IPG), which represents an affordable and rapid benchtop system designed for small projects. The IPG system harbours an array of semiconductor chips capable of sensing minor changes in pH and detecting nucleotide incorporation events by the release of a hydrogen ion from natural nucleotides (Figure 1.20). The Ion Torrent system does not require any special enzymes or labelled nucleotides and takes advantage of the advances made in the semiconductor technology and component miniaturization (Myllykangas *et al.*, 2012).

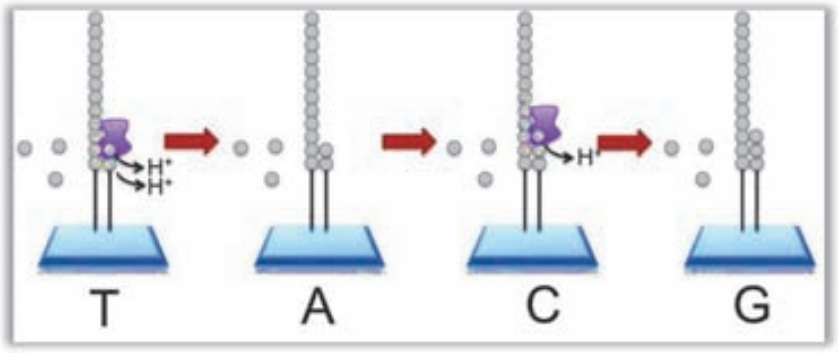


Figure 1.20: Semiconductor sequencing technology; Incorporation of deoxyribonucleotide (dNTP) into growing DNA strand by DNA polymerase. Hydrogen and pyrophosphate are released. (extracted from http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/ion-torre/ accessed in November 26, 2013).

Nanopore sequencing: DNA sequencing with the Nanopore instrument relies on the detection of an electrical signal created when nucleotides are passed through a nanopore; this is an α-hemolysin pore covalently attached to a cyclodextrin molecule – the binding site for nucleotides. The principle of this technique is based on the modulation of the ionic current through the pore as a DNA molecule traverses it, revealing characteristics and parameters (diameter, length and conformation) of the molecule (Figure 1.21). During the sequencing process, the ionic current that passes through the nanopore is blocked by the nucleotide, which in turn was previously cleaved by exonuclease from a DNA strand that interacts with cyclodextrin. The time period of the ionic current block is unique to each base and enables the DNA sequence to be determined (Astier *et al.* 2006; Rusk, 2009).

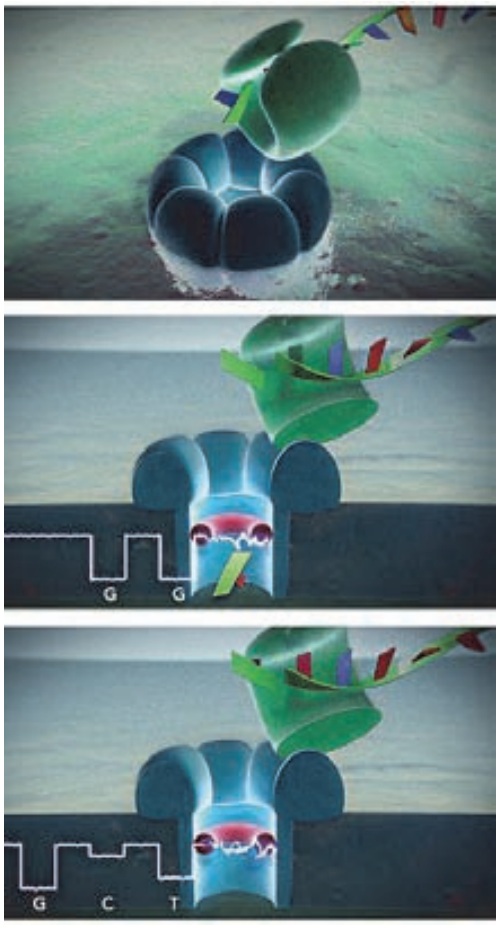


Figure 1.21: Nanopore DNA sequencing, as envisioned, will use an exonuclease (green) to cleave nucleotides from DNA. Each nucleotide will be directed into the nanopore (blue), where it interacts with a cyclodextrin "adapter" (red, donut-shaped) and interrupts a current to an extent that pinpoints its identity (such as G, C, and T) (Credits to: Oxford Nanopore Technologies).

1.2.2.2.4. Expected advances on the third generation

Second generation NGS technologies rely on PCR to grow clusters of a given DNA template, attaching the clusters of DNA templates to a solid surface that is subsequently imaged as the clusters are sequenced by synthesis in a phased approach. By contrast, the third generation NGS technologies interrogate single molecules of DNA in such a way that no synchronization (a limitation of second NGS) is required (Whiteford *et al.* 2009), thereby overcoming issues related to the biases introduced by PCR amplification and dephasing. Furthermore, third generation NGS technologies have the potential to more fully exploit the high catalytic rates and high processivity of DNA polymerase, or avoid any biology or chemistry altogether to radically increase read length

(from tens of bases to tens of thousands of bases per read) and time to result (from days to hours or minutes). Additionally, the third generation NGS technologies may offer the following advantages over second generation NGS technologies: i) higher throughput, ii) faster turnaround time (e.g., sequencing metazoan genomes at high fold coverage in minutes), iii) longer read lengths to enhance *de novo* assembly and enable direct detection of haplotypes and even whole chromosome phasing, iv) higher consensus accuracy to enable rare variant detection, v) small amounts of starting material (theoretically only a single molecule may be required for sequencing), and vi) low cost, where sequencing the human genome at high fold coverage for less than $1000 (Figure 1.22) is now a reasonable goal for the community (Table 1.23) (Pareek *et al.*, 2011).

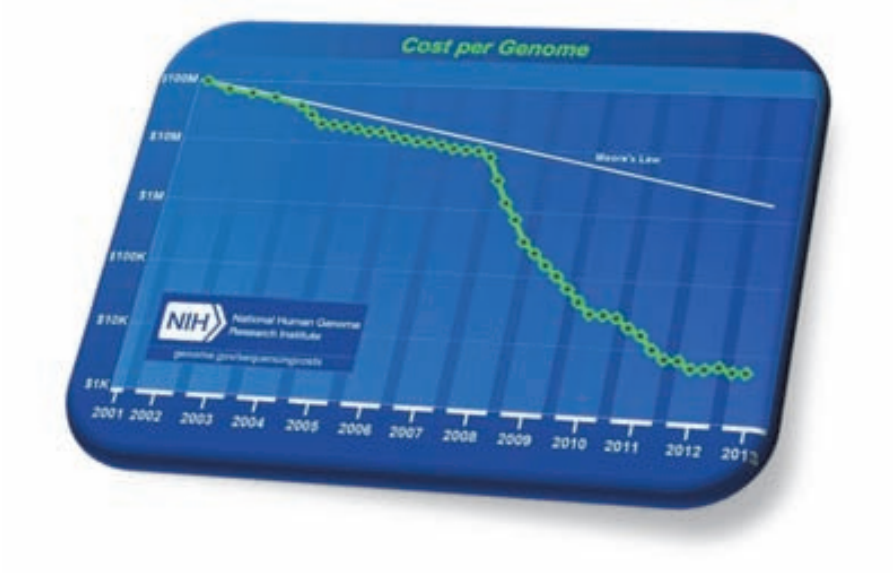


Figure 1.22: Cost per genome sequencing over the years in relation with the Moore's Law. (edited from http://www.genome.gov/)

Table 1.3: Comparison of second and third NGS platforms (edited from Pareek *et al.*, 2011).

	'Second Generation'			'Third Generation'	
Companies	Roche GS FLX	Illumina-Sollexa	Life Technologies	Helicos Biosciences	Pacific Biosciences
Platforms	GS FLX Titanium, GS Junior	HiSeq 2000, MiSeq Genome Analyzer IIX, Genome Analyzer IIE, iScanSQ	ABI SOLiD, SOLiD 4	HeliScope	SMRT, PacBio RS, PacBio RSII
Template preparation	Clonal-emPCR on bead surface	Clonal bridge enzymatic amplification on glass surface	Clonal-emPCR on bead surface	Single molecule detection	Single molecule detection
Sample requirements	1 µg for shotgun library, 5 µg for paired-end	<1 g for single or paired-end libraries	<2 µg for shotgun library, 5–20 µg for paired end	<2 µg, single end only	unknown
Detection method	Light emitted from secondary reactions initiated by release of pyrophosphate	Fluorescent emission from incorporated dye-labelled nucleotides	Fluorescent emission from ligated dye-labelled oligo-nucleotides	Real time detection of fluorescent dye in polymerase active site during incorporation	Real time detection of fluorescent dye in polymerase active site during incorporation
Length of library prep/ feature generation (days)	3–4	2	2–4.5	1	unknown
Method of feature generation	Bead-based/emulsion PCR	Isothermal *'bridge amplification'* on flow cell surface	Bead-based/emulsion PCR	Single molecule sequencing	Single molecule real time sequencing by synthesis
Paired ends/separation	3 kb (2 × 110 bp)	200 bp (2 × 36 bp)	3 kb (2 × 25 bp)	25–55 bp	unknown
Chemistry	Pyrosequencing	Reversible Dye Terminators	Oligonucleotide Probe Ligation	Reversible Dye Terminators	Phospho-linked Fluorescent Nucleotides
Bases/template	~400	~75 (35–100)	35–50	35	800–30,000
Templates run	1,000,000	40,000,000	85,000,000	unknown	unknown
Data production/day	400 MB/run/7.5 hr	3,000 MB/run/6.5 days	4,000 MB/run/6 days	8 days	0.02 days
Maximum samples	16 regions/plate	8 channels/flow cell	16 chambers/2 slides	unknown	unknown
Raw accuracy	99.5%	>98.5%	99.94%	>99%	>99%
Sequencing method	Pyrosequencing	Reversible dye terminators	Sequencing by ligation	One base-at-a-time	Sequencing by synthesis
Read lengths	400 bases	36 bases	35 bases	Longer than 1000	Longer than 1000
Sequencing run time	10 h	2-5 days	6 days	12	<1
Total Throughput bases/run	0.40–0.60 Gb, 0.035 Gb	3–6 Gb	10–20 Gb	28 GB	100 Gb per hour
Throughput/day (Gb)	~1	1.5	1.7–2	2.5	~1
Estimated system cost	$500,000	~$400,000	$525,000	Lower than second NGS	Lower than second NGS
Consumable cost per run	$5000	$3000	$4000	Lower than second NGS	Lower than second NGS
Cost per run (total direct)	$8439	$8950	$17,447	Lower than second NGS	Lower than second NGS
Cost per Mb	$84.39	$5.97	$5.81	Lower than second NGS	Lower than second NGS

1.2.3. Benchtop Next Generation Sequencers

Large NGS platforms have been used for massively parallel DNA sequencing. However, the cost and extremely large capacity of these platforms result in a loss of flexibility for the needs of many clinical genetics laboratories. Fortunately four different benchtop massive sequencing instruments are currently available, all roughly the size of a laser printer, with modest set-up and running costs that are feasible for routine molecular testing (Lonam *et al.*, 2012). In addition to the previously discussed IPG, its predecessor, the Ion Torrent Personal Genome Machine (PGM), was launched in early 2011 (Rothberg *et al.*, 2011) and has already seen its replacement born (the IPG); the MiSeq (Illumina) which was announced in January 2011 and is based on the existing Illumina sequencing-by-synthesis chemistry (Bentley *et al.*, 2008); the 454 GS Junior from Roche which was released in early 2010 and is a smaller, lower-throughput version of the Roche GS FLX platform, exploiting similar emulsion PCR and pyrosequencing approaches, but with lower set-up and running costs (Figure 1.23).

Figure 1.23: Available Bechtop Next Generation Sequencers. From left to right: PGM, GS Junior, MiSeq and IPG.

In this research work, different techniques were applied to carry out screening for disperse mutations in large genes (***BRCA1*** and ***BRCA2***) and for molecular testing of genetically heterogeneous diseases in which mutations in several genes may be involved (**autosomal dominant Retinitis Pigmentosa**) using the Roche GS Junior benchtop sequencing platform.

1.3. Introduction of DNA Massive Sequencing in the Clinical Practice

Molecular testing of genetic diseases is in increasing demand in routine clinical practice. For most recurrent genetic variants in genes responsible for disease, the methods for molecular testing are usually well established in clinical laboratories. For example, the expansion of nucleotide triplets that cause Fragile X Syndrome, Huntington disease, or various ataxias are directly screened for this expansion mutation in molecular testing (Fry and Usdin, 2006). However, the screening for disperse mutations in large genes or molecular testing of genetically heterogeneous diseases (in which mutations in several genes may be involved) is a more difficult task. *BRCA1* and *BRCA2*, which are associated with a high risk of breast cancer (King *et al.*, 2003), are large genes with a high number of exons and therefore involve a considerable number of individual PCRs and sequencing reactions to cover the coding and flanking sequences of both genes.

Although clinical protocols for genetic testing (Pruthi *et al.*, 2010) of ***BRCA1*** and ***BRCA2*** using PCR and Sanger sequencing are well established, the task is nevertheless costly and time consuming. On the other hand, in genetically heterogeneous diseases such as **autosomal dominant Retinitis Pigmentosa** (adRP), mutation screening may be required in more than 20 genes in order to establish the molecular cause of the disease (Bowne *et al.*, 2011). This requires traditional methods of PCR amplification of several hundreds of single amplicons and their corresponding individual sequencing by Sanger method. Next Generation Sequence (NGS) technology circumvents the limitation for single sequencing of amplicons allowing a massive sequencing of mixed DNA fragments.

Multiplex PCR is an attempt to approach this issue (Varley and Mitra, 2008). However, multiplex PCR with numerous primer pairs often results in interprimer interactions or an increase in mispriming events that prevent correct amplification (Fan *et al.*, 2006). Alternatively, DNA capture of targeted genomic coding and flanking sequences of several genes by hybridization with custom oligonucleotides followed by sequencing in large Next Generation Sequencing platforms has been used in molecular diagnostics (Simpson *et al.*, 2011; Walsh *et al.*, 2010). However, the cost and extremely large capacity of these platforms result in a loss of flexibility for the needs of many clinical genetic laboratories.

Nevertheless, NGS approaches are now the technical choice for molecular testing of large genes or simultaneous analysis for several genes causing a genetic disease. With the introduction of benchtop massive sequencing instruments like the GS Junior platform, such approaches are now possible in the clinical setting. Still, a critical step in NGS is the generation of DNA fragments eligible to be sequenced (libraries).

The research work hereby presented has taken advantages of the rapid progression of DNA sequencing technologies. Different methods for library preparation have been assayed, as Long-Range PCR, Multiplex PCR, or DNA capture by hybridization. Therefore, the challenge here was to introduce rational methods for molecular analysis and utilise them to study several patients in a relatively short time and thus satisfy an unmet clinical demand.

2. Objectives

This research work is committed to the following objectives:

1) To develop a method for molecular diagnosis based upon detection of genetic variants in large genes, such as *BRCA1* and *BRCA2* genes, by massive sequencing (also known as next generation sequencing or NGS) as alternative to traditional mutation screening and capillary sequencing (Sanger) methods;

2) To establish NGS technology in molecular diagnosis of complex heterogeneous monogenic diseases where several genes can individually be involved in their pathology, such as autosomal Retinitis Pigmentosa (adRP);

3) To investigate NGS technologies as innovative analysis process to detect genetic adRP-causing variants in new genes that are not yet associated with this disease on the Spanish population by two approaches: candidate genes screening and whole exome analysis.

3. Material and Methods

3.1. Patients and studied families

The study was conducted in patients with a family history of breast or ovarian cancer or with retinitis pigmentosa who had been admitted to the Terrassa Hospital. The study also included DNA samples from index cases of families with autosomal dominant retinitis pigmentosa who had been referred from the centres belonging to the EsRetNet network: Hospital de la Santa Creu i Sant Pau (Barcelona), Hospital Sant Joan de Déu (Barcelona), Fundación Jiménez Díaz (Madrid), Hospital Virgen del Rocío (Seville), Hospital Universitario La Fe (Valencia), and from other hospitals in Spain: Hospital Miguel Servet (Zaragoza) and Hospital de Basurto (Bilbao). Informed consent was obtained from all patients before the study, which was conducted in accordance with the Declaration of Helsinki and approved by the internal Clinical Research Ethics Committee (CEIC) of the Terrassa Hospital, Spain. To study the molecular genetic causes, DNA was extracted from peripheral blood of all patients. The mutation analysis was performed in the index case of each family. In cases in which a possible disease causing mutation was detected, a study of the cosegregation of the mutation with the disease in other family members was performed.

3.2. Nucleic acid extraction from peripheral blood samples

3.2.1. Genomic DNA

DNA isolation from peripheral blood lymphocytes collected in EDTA was performed automatically with the MagNA Pure Compact Instrument (Roche, Barcelona, Spain) in accordance with the

manufacturer's protocol and using MagNA Pure Compact Nucleic Acid Isolation Kit I (Roche). The initial sample volume was 400 µl and the final elution volume was 200 µl in TE buffer 1X.

3.2.2. Total RNA

Total RNA was extracted from 1.5 ml of fresh peripheral blood collected in EDTA using a QIAamp RNA blood kit following the manufacturer's instructions (Izasa, Barcelona, Spain). To prevent illegitimate splicing, blood samples were processed after venipuncture with a maximum delay of 4 hours (Ars *et al.* 2000, Wimmer *et al.* 2000). The final elution volume was 60 µl in RNAse free water.

3.3. Nucleic acid synthesis

3.3.1. Oligonucleotide design and synthesis

Oligonucleotide primers needed for polymerase chain reaction (see section 3.5) were designed by the Oligo 7.41 (Molecular Biology Insights, Cascade, CO, USA) program and delivered to StabVida (Oeiras, Portugal) for their synthesis and cartridge purification.

Fluorescence resonance energy transfer (FRET) probe pairs needed for real-time polymerase chain reaction (section 3.5.4) were synthesised and delivered by TIB MOLBIOL (Berlin, Germany). These probes are separated by a single nucleotide which allows a strong FRET signal to occur. The donor probe was labelled with fluorescein at its 3' end; the acceptor probe was labelled with LightCycler® Red 640 (LC Red 640) at its 5' end.

3.3.2. Reverse transcription of mRNA

Reverse transcription was performed using 500 ng of total RNA isolated and random hexamers with a First Strand cDNA Synthesis Kit for RT-PCR (AMV) (Roche) following the manufacturer instructions.

3.4. Nucleic acids quantification

Two methods were used to establish the concentration of a nucleic acids solution: the spectrophotometric quantification for concentrations above 10 ng.µl^{-1} and Ultra-Violet (UV) fluorescence in presence of a DNA dye for samples which are more dilute.

3.4.1. Spectrophotometric quantification

The method is based on the Beer-Lambert equation and operates on the fact that nucleic acids absorb in the ultraviolet range, at a wavelength of about 260 nm, due to their nitrogenous bases. At this wavelength, the average extinction coefficient (ε) for dsDNA is 0.020 $\mu g.ml^{-1}.cm^{-1}$ and for ssRNA it is 0.025 $\mu g.ml^{-1}.cm^{-1}$. This is the same as saying that, at 260 nm, a reading of 1.0 corresponds to 50 $\mu g.ml^{-1}$ for dsDNA and 40 $\mu g.ml^{-1}$ for ssRNA.

Measurements of absorbance were carried out using the Epoch™ Microplate Spectrophotometer combined with the Take3™ Multi-Volume Plate (Izasa) in accordance with the manufacturer instructions.

Nucleic acid samples are frequently contaminated with other molecules that absorb at 260 nm (proteins, organic compounds and others). The ratio of absorbance at 260 and 280 nm is used to evaluate the purity of the nucleic acid samples. This ratio is ~1.8 for pure DNA and ~2 for pure RNA.

3.4.2. Quantification using fluorescent dyes

3.4.2.1. Quant-iT™ PicoGreen®

Quant-iT™ PicoGreen® dsDNA Assay Kit (Life Technologies, Madrid, Spain) contains the Quant-iT™ PicoGreen® dsDNA reagent which is an ultra-sensitive fluorescent nucleic acid stain for quantitating double-stranded DNA (dsDNA) in solution.

The fluorescent intensity of the PicoGreen® dye was measured by fluorometry in a LightCycler® 480 instrument (Roche) producing the excitation ($_{Ex}$) wavelength of ~480 nm and recording at the emission ($_{Em}$) wavelength of ~520 nm (five points within 60 to 61°C). The dsDNA is then quantified by comparison of the sample fluorescence to the fluorescence of a set of standard reference samples that are included in every sample run.

3.4.2.2. QuantiFluor™

The QuantiFluor™ dsDNA System (Promega, Madrid, Spain) contains a fluorescent double-stranded DNA-binding dye ($504nm_{Ex}/531nm_{Em}$) that enables sensitive quantitation of small amounts of dsDNA. The QuantiFluor™ dsDNA System was used together with the QuantiFluor™-ST Handheld Fluorometer, selecting the blue optical channel (optimal excitation and emission wavelengths).

The DNA is then quantified by comparison of the sample fluorescence to the fluorescence of a set of standard reference samples that are included in every sample run.

3.5. Polymerase Chain Reaction

Polymerase Chain Reaction (PCR) (Mullis *et al.*, 1986) is a well-known and indispensable technique that is used in any molecular biology laboratory. Table 3.1 shows the general PCR reaction mix. For each PCR, pairs of primer sets were designed by the Oligo 7.41 (Molecular Biology Insights) program and synthesised as described in section 3.3.1.

Table 3.1: General PCR reaction mix.

Components*	Volume (µl)	Final Conc.
10X PCR buffer minus Mg	5 µl	1X
10 mM dNTP mixture	1 µl	0.2 mM each
50 mM $MgCl_2$	1 µl	1 mM
Primer Forward (25 µM) Primer Reverse (25 µM)	1 µl each	0.5 µM each
Taq DNA Polymerase (5 U/µl)	0.5 µl	2.5 units
Distilled water	to 50 µl	n/a
Template DNA	2 µl	n/a

*in order to inhibit secondary structures in the DNA template or the DNA primer, DMSO can be added to a max conc. of 10%.

This process takes place in an Analikt Jena FlexCycler thermo-cycler using a block assembly 96G (Ecogen, Barcelona, Spain). The general PCR temperature conditions on the thermo-cycler are presented below:

Table 3.2: General PCR program on the thermo-cycler.

	Temperature (°C)	Time (min)	Cycles
Initial Denaturation	95	2	1
Denaturation	95	0.5	
Annealing	40-65*	0.5	30-40
Extension	72	1	
Cooling	4	on hold	

*depends on the length and base sequence of the primers.

3.5.1. Long-Range PCR

Long-Range Polymerase Chain Reaction (LR-PCR) is a form of PCR that allows the efficient amplification of large genomic DNA fragments up to 30 kb long. This is possible using the Expand Long Template PCR system (Roche), which consists of a unique enzyme mix that contains a thermostable Taq DNA polymerase and Tgo DNA polymerase (Hopfner, K.P. *et al.* 1999), a thermostable DNA polymerase with proofreading activity. The following tables show the general LR-PCR reaction mix and its thermo-cycler program:

Table 3.3: General LR-PCR reaction mix.

Components	0.5 – 9 kb		9 – 12 kb	
	Volume (µl)	Final Conc.	Volume (µl)	Final Conc.
10X Long Expand buffer with $MgCl_2$	5 (***Buffer 1***)*	1X (***1.75 mM***)	5 (***Buffer 2/3***)*	1X (***2.75 mM***)
10 mM dNTP mixture	1.75	0.35 mM each	2.5	0.5 mM each
Primer Forward (25 µM) Primer Reverse (25 µM)	0.6 each	0.3 µM each	0.6 each	0.3 µM each
Long Expand Taq Polymerase (5 U/µl)	0.75	3.75 units	0.75	3.75 units
Distilled water	to 50	n/a	to 50	n/a
Template DNA	2	n/a	2	n/a

* Specific buffer used is annotated in the Table 3.5 and 3.6.

Table 3.4: General LR-PCR program on the thermo-cycler.

	Temperature (°C)	Time	Cycles
Initial Denaturation	92-94	2 min	1
Denaturation	92-94	10 s	10
Annealing	45-65[a]	30 s	
Extension	68	2 – 8 min[b]	
Denaturation	92-94	15 s	15-25
Annealing	45-65[a]	30 s	
Extension	68	2 – 8 min[b] + 20 s for each successive cycle[c]	
Final Elongation	68	7	
Cooling	4	on hold	

[a] Depends on the length and base sequence of the primers (Specific temperatures are presented in the Table 3.5 and 3.6). [b] Elongation time depends on fragment length: Use 2 min for up to 3 kb, 4 min for 6 kb, 8 min for 10 kb. [c] For example, cycle no. 11 is 20 s longer than cycle 10, cycle no. 12 is 40 s longer than cycle 10, cycle no. 13 is 60 s longer than cycle 10, and so on.

3.5.1.1. LR-PCR Amplification of BRCA1 and BRCA2 genes from gDNA

Genomic reference sequences of *BRCA1* and *BRCA2* genes were obtained from GenBank. LR-PCR fragments were prepared as reported previously. Again, all pairs of primer sets were designed by the Oligo 7.41 program, aiming for amplification of the complete coding exons and flanking splice junctions of each gene. For each library, eleven pairs of primers (five for *BRCA1* and six for *BRCA2*) were used to amplify the LR-PCR fragments. These fragments contain most of the genomic sequences of the analysed genes including the complete coding sequences (Table 3.5). The primers and LR-PCR specific conditions used to amplify each DNA fragment are shown in Table 3.5.

Table 3.5: Amplified LR-PCR Fragments of *BRCA1* and *BRCA2*.

Gene (NCBI Ref.)	Exons per Fragment	Primers (5' - 3')	Size (bp)	T. annealing (ºC)	Fragmentase Reaction Time (min)
BRCA1 (NG_005905.2)	A: 2-3	F: TGCCGGCAGGGATGTGCTTG R: TGCTTGCAGTTTGCTTTCACTGATGGA	9,436	60.8	45
	B: 5-7	F: GTTTAGGTTTTTGCTTATGCAGCATCCA R: TCAGGTACCCTGACCTTCTCTGAAC	3,081	61.8	35
	C: 8-10	F: GGAAAAGCACAGAACTGGCCAACA R: GTGGGTTGTAAAGGTCCCAAATGGT	4,196	60.9	25
	D: 11-13	F: GCCAGTTGGTTGATTTCCACCTCCA R: TGCCTTGGGTCCCTCTGACTGG	12,739	63.2	25
	E: 14-19	F: ACCCCCGACATGCAGAAGCTG R: GTGGTGCATTGATGGAAGGAAGCA	13,853	61.2	45
	F: 20-24	F: TGACGTGTCTGCTCCACTTCCA R: AGTGAGAGGAGCTCCCAGGGC	11,622	61	45
BRCA2 (NG_012772.3)	G: 2-9	F: CAGGCGGCGTTGGTCTCTAACTG R: AGGAGCAATCCTTCAATGGTGCC	15,240	62.1	25
	H: 10-13	F: ACAGGAGAAGGGGTGACTGACCG R: GGGGAAAGCATCTCTGTTTGCTCT	15,294	62.5	45
	I: 14-18	F: GCAAATGAGGGTCTGCAACAAAGGC R: TGAGAACAAGAGGGCAGCAAGC	9,043	62.8	45
	J: 19-24	F: TGGTGGCTGAATAACCTTGGGCA R: TGGCATGGGAACAATGTGGCTT	11,617	61.5	45
	K: 25-27	F: CCTGAGCTTTCGCCAAATTCAGCTAT R: TGCCCGATACACAAACGCTGAGG	4,382	63	35

3.5.1.2. LR-PCR Amplification of adRP associated genes from gDNA

Genomic reference sequences of twelve genes commonly associated with adRP were obtained from GenBank: *CA4, CRX, IMPDH1, NR2E3, RP9 (PAP1), PRPF3, PRPF8, PRPF31, PRPH2 (RDS), RHO, RP1,* and *TOPORS*. All pairs of primer sets were designed by the Oligo 7.41 (Molecular Biology Insights, Cascade, CO) program, aiming for amplification of the complete coding exons and flanking splice junctions of each gene. To obtain each library, 20 pairs of

primers were used in individual LR-PCR amplifications that rendered DNA fragments between 3 and 10 kb; these contained most of the genomic sequences of the study genes. The primers and LR-PCR specific conditions used to amplify each DNA fragment are shown in Table 3.6.

Table 3.6: Primers, annealing temperature, LR-PCR fragment size and fragmentase digestion times of 12 common genes associated with adRP.

Gene (NCBI Ref.)	Nº of exons per Fragment	Primers 5'- 3'	Size (bp)	T. annealing (ºC) Buffer (1-3)	Fragmentase reaction time (min)
IMPDH1 (NG_009194.1)	A: 6 exons	F: GGCCGCGCGGGTGTTTATGT R: GGAGTCAGGCTGGGGGTTG	5,858	65 (1*)	35
	B: 7 exons	F: CTCCCAAACTTCTCCCCACCAA R: CAGGCTGCAGCAAGAAAGACCT	5,952	59 (3)	45
RP9 (NG_012968.1)	5 exons	F: GCTCTTCTCAGGTGATTGCTTGTCT R: CAGCAGGGAGGACAACGATGAA	5,731	59 (1)	45
PRPF3 (NG_008245.1)	A: 3 exons	F: GTGTAGTATTGAGTCCTGTTT R: TCCACAATAAATACAGACCCATG	3,655	50 (1)	45
	B: 3 exons	F: GGTACTGTTTTCTTCCCCCTTC R: ATAATAAAGTACACAATAAATGCTA	3,153	50 (1)	45
	C: 5 exons	F: CCTGGGCAAGAAGAACGAAACTC R: GCTTTTAGGCCACAATGATTCTGC	6,596	65 (2)	35
	D: 4 exons	F: TTTTTACTTGCCACTCCATTCAGG R: GTGGGAGGGGGTTGGGAAATAGA	7,120	57 (1)	35
PRPF8 (NG_009118.1)	A: 11 exons	F: GATGCACTGTGTGGCGGGACTG R: AGTCATAGCTGCTCTCCAAATAGGTC	6,750	55 (1)	35
	B: 11 exons	F: CCCCACCTCTGCCCAGGTTTC R: GGACGAAGTGAAAGGGGTGTGAA	4,769	61-64 (1)	25
	C: 13 exons	F: GCCCAGCCTCATTGTGTTTTGC R: TGACCTCAAATGATCTACCCACCTT	7,130	62-64 (2)	35
	D: 6 exons	F: GAAGTGACCAAGGGGAAGGGATC R: CAAGCCATCAGGAGGTCAACAAC	3,398	59-61 (1)	45
PRPF31 (NG_009759.1)	A: 7 exons	F: CCTGTTGTCGTGGAGCCTGCAT R: ACGTCCCATGCCACCCTGTTC	6,659	63.5-65 (1)	45
	B: 6 exons	F: TTGCCCCGCTGTTACCTCTGTCTGT R: CCCTGCCAGAACCCGATCCTA	5,114	57 (1)	35
PRPH2 (NG_009176.1)	2 exons	F: GAGGCAGGGGTTGGGAGGAAGT R: CGGAGTTGGATGAGGGGGAGAT	6,475	62-63 (1)	35
NR2E3 (NG_009113.1)	8 exons	F: TGCAGGTGTGGCCAGTTGATC R: CTCATTCCTTCAGTGTCCCCAAA	7,832	60 (3)	45
TOPORS (NG_017050.1)	1 exon	F: TGCATTTTTGTTGAGACTAT R: AGACTGCAGTAGACGACATT	3,086	50-53 (1*)	25
RP1 (NG_009840.1)	3 exons	F: TATGGTGCTGTGATTCTGGAGA R: AGGTTATCTAGTCCGTATTCGTCAC	9,526	50 (3)	45
CRX (NG_008605.1)	3 exons	F: CCCCGAAGATCATGATGGCGTAT R: AAGGGGTGGTCTCTGTAAGCTGAAC	5,694	58.5 (1*)	35
CA4 (NG_012050.1)	7 exons	F: CACTTCACACCCTTCCTCTCTGCTG R: CTGGGAAGGCTAAGGACCGGAAG	4,215	58.5 (1*)	35
RHO (NG_009115.1)	5 exons	F: ATGAATGGCACAGAAGGCCCTAACT R: GTCTCCCATCCCCTACACCT	5,038	55-60 (1)	35

* 5% DMSO
Total bp amplified: 113751. Exons amplified: 116 (25914 bp)
Exons 1 of: *CA4*, *IMPDH1*, *RP9* (*PAP1)*, *PRPH2* and exon 2 of *TOPORS* are not included.

3.5.2. Multiplex PCR

Multiplex-PCR is powerful technique that enables the amplification of two of more fragments in one single PCR reaction tube. Multiplex-PCR experiments were performed using Qiagen Multiplex-PCR Kit (Izasa) and following the manufacturer guidelines as shown on tables 3.7 and 3.8.

Table 3.7: Multiplex-PCR reaction mix.

Components	Volume (µl)	Final Conc.
2X QIAGEN Multiplex PCR Master Mix*	25	1X
10X primer mix, 2 µM each primer (µl)†	5	0.2 µM†
Q-Solution, 5x	5	0.5X
RNase-free water	to 50 µl	n/a
Template DNA	2	≤1 µg DNA/50 µl

* Provides a final concentration of 3 mM MgCl2, † Exact mixes are shown in Table 3.17.

Table 3.8: Multiplex-PCR program on the thermo-cycler.

	Temperature (°C)	Time (min)	Cycles
Initial Activation Step	95	15	1
Denaturation	94	0.5	
Annealing	57-63*	1.5	30-45
Extension	72	1.5	
Final Extension	72	10	1
Cooling	4	on hold	

*depends on the length and base sequence of the primers.

3.5.3. Emulsion PCR

Emulsion PCR (emPCR) uses water-in-oil (w/o) emulsion to generate thousands of fine and robust water-phase compartments (droplets) in a bulk oil phase in which all PCR components can be encapsulated and primers and DNA molecules can be isolated. These small (15 to 20 µm diameter and 2 to 4 pl capacity) droplets are very stable, even at high temperatures, and will function well as micro-reactors for multiple PCRs. emPCR was performed by following these steps:

a) Prepare the reaction mixture as presented in this table and store it on ice:

Table 3.9: General reaction mixture for emPCR.

Components	Volume (µl)	Final Conc.
10X PCR buffer minus Mg	50 µl	1X
10 mM dNTP mixture	15 µl	0.3 mM each
50 mM MgCl2	15 µl	1.5 mM
Primer Mix (12 µM)*	41.7 µl each	1 µM each
BSA 100X	5 µl	1X
Taq DNA Polymerase (5 U/µl)	12 µl	6 units
Distilled water	to 500 µl	n/a
Template DNA†	20 µl	n/a

* Information about the exact primer mix can be found in section 3.9.2.3. † gDNA digested with fragmentase at 37 ºC for 15 min in accordance with the table 3.15.

b) Emulsion is performed in a Ultra Turrax Tube Drive (UTTD) (IKA® GmbH, Staufen, Germany), following these steps:

- Set the UTTD to 4000 rpm for 5 min.
- Pour 2 ml of the emulsion oil into a Turrax stirring tube.
- Add 1 ml of 1X Mock Mix to the Turrax stirring tube containing the emulsion oil.
- Place the stirring tube in the UTTD and start the UTTD to mix the emulsion (Both oil and Mock Mix are included in the GS 454 emPCR kit).

c) Transfer the entire mixture prepared in point a) into the Turrax stirring tube.

d) Set the UTTD to 2000 rpm for 5min.

e) Start the UTTD to make the final emulsion.

f) Using a Combitip, aliquot 100 µl of the final emulsion into one 96-well plate by slowly aspirating. Take care not to draw air.

g) Seal the 96-well plate with a sealing tape.

h) Place the sealed 96-well plate in a thermal cycle and used the following PCR cycle program:

Table 3.10: emPCR program on the thermo-cycler.

	Temperature (°C)	Time	Cycles
Initial Denaturation	95	4 min	1
Denaturation	95	30 s	15
Annealing	55	90 s +0.3 ºC for each successive cycle*	
Extension	72	1 min	
Denaturation	95	30 s	30
Annealing	65[a]	90 s	
Extension	7	1 min	
Cooling	4	on hold	

* For example, cycle no. 11 is 0.3 ºC higher than cycle 10, cycle no. 12 is 0.6 ºC higher than cycle 10, cycle no. 13 is 0.9 ºC higher than cycle 10, and so on.

After the reaction program is completed, the emPCR mixture is pooled to a 1.7 ml tube and 1 ml of water saturated di-ethyl-ether is added to each 0.5 ml of pooled emPCR mixture. This mixture is then vortexed shortly and centrifuged for two min. The upper phase is discarded and the process is repeated three times until a clear water phase is visible at the bottom of the tube. Aspirate the water phase taking care to not aspirate the upper phase. The water phase is where all amplicons are located and should be centrifuged for 10 min at 40 ºC in a speed-vacuum to eliminate any di-ethyl-ether contamination.

3.5.4. Real-Time PCR using Fluorescence Energy Transfer Probes

In order to screen the discovered mutations in RP patients and controls (or in the population), specific oligonucleotide primers and fluorescence resonance energy transfer (FRET) probe pairs targeting the variants were designed. Primers and probes (Table 3.11) were synthesised as described in section 3.3.1.

Table 3.11: Primer and FRET probes.

Gene	Type	5' – Sequence	Melting Temperature (ºC)
IMPDH1	Primer	F: GGTGTAGAGAGGACCCTCACG	57.7
		R: CTGGTACCTTTCTTGCTACGCT	57.3
	Probe	F: AGCCACCACCCGTTCAATC--FL	63.0
		R: LC640-TGGCGTCATCACCTGTGGGGCC--PH	71.3
PDE6G	Primer	F: CCCACAAGGGTTGGAAG	54.4
		R: GGGGGCTTGCTCTTGAA	56.1
	Probe	F: GGTGAGGCTGACGGAGACACC--FL	63.0
		R: LC640-CGGCAACCTTGGCTCCTGGACTCC--PH	71.3

* FL, donor dye Fluorescein; LC640, acceptor dye LightCycler Red 640; PH, Phosphate.

Real-time PCR (RT-PCR) amplification was performed using the LightCycler® 480 system and following the LightCycler FastStart DNA Master Hybridization Probes instruction manual (Roche). Data were acquired and analysed with the Melting Curve Genotyping software. The following tables show the general RT-PCR reaction mix used and the PCR temperature conditions on the thermo-cycler:

Table 3.12: RT-PCR reaction mix.

Components*	Volume (µl)	Final Conc.
25 mM $MgCl_2$	1.2 µl	1.5 mM
Primer Mix (25 µM)	0.4 µl	0.5 µM
Probes Mix (5 µM)	0.8 µl	0.2 µM
LightCycler FastStart DNA Master Hybridization Probes	2 µl	n/a
Distilled water	to 20 µl	n/a
Template DNA	2 µl	n/a

Table 3.13: RT-PCR program on LightCycler® 480 system.

Program	Cycles	Analysis mode	Target (°C)	Acquisition Mode	Hold	Ramp rate (ºC/s)	Acquisitions (per ºC)
Pre-incubation	1	None	95	None	10 min	4.40	-
Amplification	45	Quantification	95	None	5 s	4.40	-
			50	Single	5 s	2.20	-
			72	None	10 s	4.40	-
Melting	1	Melting Curves	95	None	1 min	4.40	-
			40	None	30 s	2.20	-
			85	Continuous	-	0.11	5
Cooling	1	None	40	None	30	2.20	-

3.6. DNA Electrophoresis

The gel electrophoresis method is used to visualise the DNA and RNA molecules. In our case, two forms of gel electrophoresis were used depending on the necessity of the analysis.

3.6.1. Agarose Gel

Most agarose gels used were between 1 and 3% concentration, having been dissolved in a solution containing 40 mM Tris, 20 mM acetic acid, and 1 mM EDTA (1X TAE). The exact

concentration of agarose used depends on the resolution needed. For example, a 0.7% agarose gel provides good resolution for fragments between 800 to 10,000 bp. On the other hand, for fragments ranging 200 to 300 bp a 1.5 % agarose gel should be used instead (Table 3.14).

Table 3.14: Resolution of Linear DNA on Agarose Gels.

Recommended Agarose (% p/v)	Optimum Resolution for Linear DNA (kb)
0.6	1.0 – 20
0.7	0.8 – 10
0.9	0.5 – 7
1.2	0.4 – 6
1.5	0.2 – 3
2.0	0.1 – 2

The separated DNA is then observed with stain (ethidium bromide) under UV light with a wavelength of about 210 nm to 285 nm.

3.6.2. Experion™ Automated Electrophoresis System

The Experion system employs LabChip microfluidic technology to automate electrophoresis and analysis by integrating separation, detection, and data analysis. It uses a much smaller sample volume (1 µl) than standard analysis methods such as agar and acrylamide gel electrophoresis. The Experion DNA 1K analysis kit (Bio-Rad, Barcelona, Spain) was used in accordance with the manufacturer instructions, which allows analysis of DNA fragments of 15 – 1,500 bp. This assay provides high sensitivity and excellent resolution (down to 5 bp) over a broad dynamic range**.** This method is used when a high resolution is needed.

3.7. PCR Product Purification

PCR leftovers like primers, mineral oil, salts, unincorporated nucleotides, and the thermostable DNA polymerase may inhibit further enzymatic reactions like sequencing and cloning of the PCR products. Accordingly, the PCR products were purified before any subsequent work.

3.7.1. Column Purification of PCR Products in Solution

Column purification was done using the High Pure PCR Product Purification Kit (Roche) following the manufacturer recommendations. In figure 3.1 the entire process overview is represented.

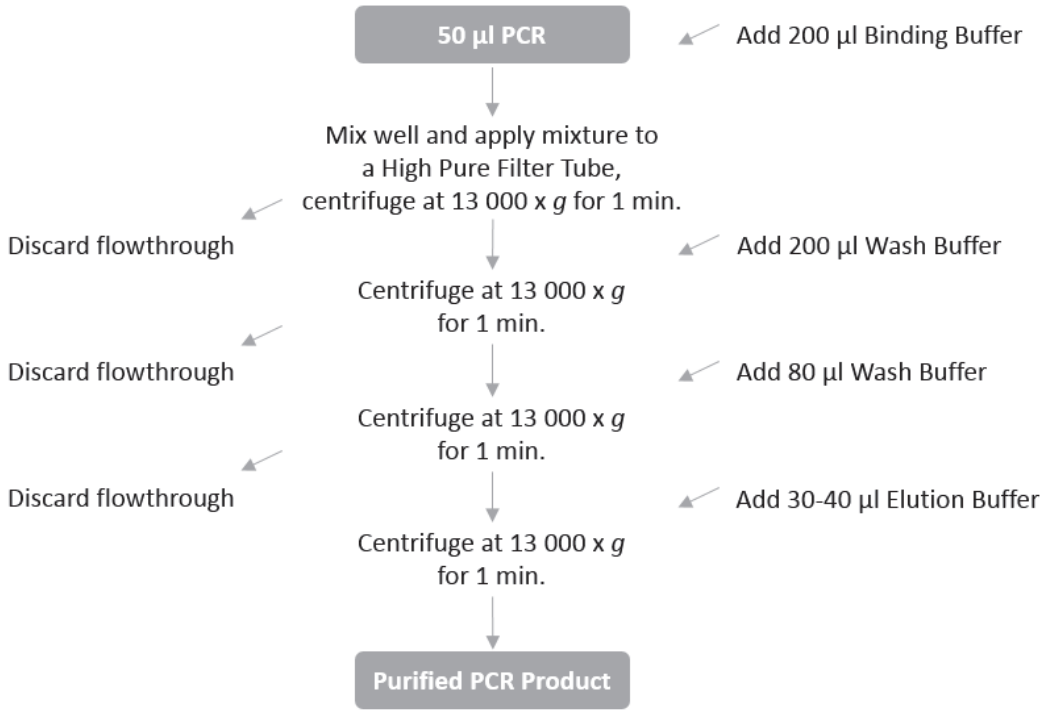


Figure 3.1: Column purification process overview.

3.7.2. Purification of PCR Products from Agarose Gel

To purify fragments within a very specific size or size interval, a PCR product can be purified from the agarose gel. This purification procedure was performed using a High Pure PCR Product Purification Kit (Roche) and following the manufacturer recommendations for purification of PCR Products from agarose gel.

3.7.3. AMPure® XP (small size removal)

This purification system utilises Agencourt's solid-phase paramagnetic bead technology for high-throughput purification of PCR amplicons. Agencourt® AMPure® XP (Beckman Coulter, Barcelona, Spain) utilises an optimised buffer to selectively bind PCR amplicons 100 bp and larger to paramagnetic beads. Excess primers, nucleotides, salts, and enzymes can be removed using a simple washing procedure. The resulting purified PCR product is essentially free of contaminants and small fragments like primer-dimers. This kind of purification is used primarily to purify sequencing libraries due to its effective removal of small sized fragments. Below is an overview of the entire process.

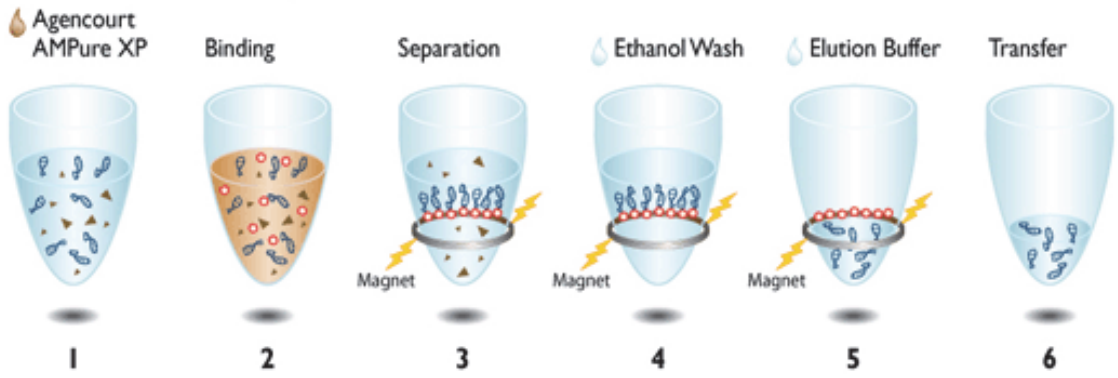


Figure 3.2: Agencourt® AMPure® XP process overview. (Image extracted from AMPure® XP user manual)

1) Addition of 1.8 ul AMPure® XP per 1.0 ul of PCR product.

2) Binding of PCR products to paramagnetic beads.

3) Separation of beads + PCR from contaminants.

4) Wash beads + PCR product 2x with 70% Ethanol to remove contaminants.

5) Elute purified PCR product from beads.

6) Transfer to new tube.

3.8. DNA Capillary sequencing (Sanger)

Amplified products of mutation positive or ambiguous samples were recovered from the plate, column purified with the High Pure PCR Product Purification Kit (Roche), and submitted to StabVida (Oeiras, Portugal) for direct sequencing on a 3730XL ABI DNA sequencer (Applied Biosystems, Foster City, CA, USA) using the Big Dye terminator V1.1 DNA sequencing kit and the subsequent PCR primers. The results were analysed using FinchTV V1.4.0 software (Geospiza, Seattle, WA, USA).

3.9. Next Generation Sequencing

Next Generation Sequencing (NGS) has revolutionised genomics and biology. This development has fuelled the demand for novel upstream techniques for optimal NGS library construction. Presented here are several novel library preparation methods developed specifically to fulfil the clinical demands of molecular diagnostics.

3.9.1. DNA sample quality control

One of the most critical aspects for an optimal NGS library preparation is the quality of the DNA to be sequenced. Hence, it is first necessary to evaluate the quality of the gDNA, its integrity, and its purity. With this in mind and before proceeding to the NGS library preparation itself, all gDNA samples were evaluated by agarose gel electrophoresis to check for sample degradation. The gDNA samples were also checked for contamination and quantified by absorbance measurement at 260 nm with an Epoch™ Microplate Spectrophotometer (Izasa) as described in section 3.4.1.

Moreover, when fragmented gDNA was needed for the NGS library preparation (NimbleGen 385K Array and SureSelect 60K and Whole-Exome Arrays), the fragment size interval and the average size of the fragmented gDNA were determined using the Experion™ Automated Electrophoresis (Bio-Rad).

3.9.2. NGS library preparation

3.9.2.1. Long-Range PCR allied with fragmentation

After the LR-PCR fragments were amplified (see section 3.5.1) they were individually purified with High Pure PCR Product Purification kit (Roche) and quantified using an Epoch Microplate spectrophotometer (Izasa).

A successful reading of the LR (3 kb – 15 kb) - PCR fragments required enzymatic digestion into 750 base pairs (bp) fragments (on average). The two distinct enzymatic fragmentation methods used are described below:

- *NEBNext dsDNA fragmentase (Fragmentase)*

The optimal Fragmentase digestion time for each DNA fragment, which rendered an average length of 750 bp, is annotated in the table 3.5 and table 3.6. The LR-PCR fragments obtained for each library are equimolar mixed into three distinct groups according to their reaction time (25, 35, and 45 min). Then they are digested with Fragmentase (Izasa) at 37 ºC to an average of 750

bp (Table 3.15) and purified with a High Pure PCR Product Purification kit (Roche Applied Science) as described in section 3.7.1.

Table 3.15: General NEBNext™ dsDNA Fragmentase reaction mix.

Components*	Volume (µl)	Final Conc.
DNA	variable	0.03 ng/µl
10X Fragmentase Reaction Buffer	5 µl	1X
100X BSA	0.5 µl	1X
Distilled water	variable	n/a
dsDNA Fragmentase	3 µl	n/a

After all LR-PCR fragments have been shorn, their concentration is measured in a microplate spectrophotometer (EPOCH, Izasa) and equimolar mixed. For each library preparation, 120 ng of each mix was end-repaired followed by a sequencer-specific adaptor ligation and small fragment removal by AMPure®, as reported in the Rapid Library (RL) Preparation Method: GS Junior Titanium Series (Roche) manual. Finally, the library is quantified with QuantiFluor as described in section 3.4.2.

- *Nextera™ fragmentation technology (Nextera)*

The LR-PCR fragments were equimolecular mixed and 50 ng of this DNA fragment mix was used in a Nextera (Ecogen) reaction in accordance with the manufacturer's instructions. Briefly, by incubation at 55 ºC for 5 minutes, 50 ng of pooled DNA is fragmented and tagged with 1 µl of the Nextera Enzyme Mix containing free transposon ends. The tagmented DNA is purified using a Zymo DNA Clean & Concentrator-5 Kit (Ecogen). The purified DNA is then amplified by a limited-cycle PCR with a four-primer reaction to add sequencer-specific-adaptors. Small fragment removal was then performed by AMPure®. Finally, the resulting product of this reaction is quantified using QuantiFluor™ as previously described.

3.9.2.2. Capture of gDNA by hybridization with target sequences

Three distinct DNA capture methods were employed for NGS library preparation. Each of these methods utilizes target sequences to capture the DNA regions of interest. These target sequences or baits can either be fixed to a solid phase (NimbleGen Array) or dispersed in solution (SureSelect 60K and Human All Exon Arrays).

3.9.2.2.1. DNA capture in solution by a custom 60K SureSelect Array

Here, a solution-based system utilizing ultra-long (120 bp) biotinylated cRNA baits is used to capture regions of interest and enrich them with the purpose of creating a NGS fragment library.

- *cRNA oligonucleotide bait custom-design*

A total of 23 genes commonly associated with adRP were selected for a 60K eArray (Table 3.16). These genes were converted to chromosomal and exon coordinates using the UCSC human genome assembly hg19 (http://genome.ucsc.edu/) and submitted to SureDesign (https://earray.chem.agilent.com/suredesign/) (Agilent Technologies, Barcelona, Spain) to generate a set of custom designed cRNA oligonucleotide baits to capture the coding and flanking regions.

- *Sequence capture and NGS library preparation*

Genomic DNA was first shorn in a Fragmentase assay. 500 ng of gDNA was digested with 1 µl of enzyme at 37 ºC for 15 min followed by purification with Agencourt® AMPure® XP. End-repair and adaptor ligation was carried out as previously described. The adapter-ligated library was then subjected to PCR amplification using the SureSelect 454 PCR primer mix (Agilent Technologies, Barcelona, Spain) following the manufacturer's instructions. 500 ng of the amplified adapter-ligated library was then captured by hybridization in solution with biotinylated custom-designed cRNA oligonucleotide baits following the SureSelect Target Enrichment System manual (Agilent Technologies). The target regions were selected using magnetic streptavidin beads, amplified as described in the manual. The library was then ready to be loaded onto the sequencer.

Table 3.16: List of the 23 candidate genes associated with adRP used for the library construction by gDNA target-capture.

Gene (NCBI Reference)	Location	Gene (NCBI Reference)	Location
CA4 (NG_012050.1)	17q23.2	*PRPF8* (NG_009118.1)	17p13.3
CRX (NG_008605.1)	19q13.3	*PRPH2* (NG_009176.1)	6p21.2
GUCA1B (NG_016216.1)	6p21.1	*RDH12* (NG_008321.1)	14q24.1
IMPDH1 (NG_009194.1)	7q31.1	*RHO* (NG_009115.1)	3q22.1
KLHL7 (NG_016983.1)	7p15.3	*ROM1* (NG_009845.1)	11q12.3
NR2E3 (NG_009113.1)	15q22.32	*RP1* (NG_009840.1)	8q11-q13
NRL (NG_011697.1)	14q11.2	*RP9* (NG_012968.1)	7p14.3
PROM1 (NG_011696.1)	4p15.32	*RPGR* (NG_009553.1)	Xp21.1
PRPF3 (NG_008245.1)	1q21.2	*SEMA4A* (NG_027683.1)	1q22
PRPF31 (NG_009759.1)	19q13.4	*SNRNP200* (NG_016973.1)	2q11.2
PRPF4 (NC_000009.11 116037914-116056079)	9q31-q33	*TOPORS* (NG_017050.1)	9p21.1
PRPF6 (NG_029719.1)	20q13.33		

3.9.2.2.2. DNA capture by a custom 385K NimbleGen Array

- *NimbleGen array custom-design*

A total of 448 selected genes (Appendix A.1) were converted to chromosomal and exon coordinates using the UCSC human genome assembly hg18 (http://genome.ucsc.edu/) and submitted to NimbleGen (Roche) to generate a 385K array to capture coding and flanking regions. For the array design, 385,000 unique probes 60-80 nucleotides in length across the coding and flanking sequences of 448 target genes were designed. Repetitive regions were not covered, overlapping target regions were merged into one, and regions were extended to 250 bp to increase capture efficiency. Uniqueness of probes was assessed with SSAHA (Sequence Search and Alignment by Hashing Algorithm) and an additional padding of 100 bases (offset) was

added to both sides of probes to obtain additional coverage. The final approved custom array covered 95% of the total 1,938,644 bases corresponding to 5,397 targets.

- *Sequence capture with NimbleGen array and NGS library preparation*

DNA capture was performed in accordance with NimbleGen protocols of sequence capture (NimbleGen Arrays User's Guide: Sequence Capture Array Delivery v3.2). The resulting library (500 ng) was hybridised to the custom 385K array with the use of the NimbleGen Sequence Capture Hybridization System of the NimbleGen System 4. The hybridised target DNA was washed and eluted with the use of a NimbleGen Wash and Elution Kit according to the manufacturer's instructions. The eluted sample was then amplified by ligation-mediated PCR with the use of primers complementary to the sequence of the adaptors and purified according to the instruction manual (NimbleGene v3.2). Finally, the average loci fold enrichment was estimated in each sequence capture sample by performing quantitative PCR on internal quality control (QC) loci (NimbleGen v3.2) before and after sequence capture enrichment and then calculating the relative changes in template concentration for those loci.

3.9.2.2.3. Whole exome capture

- *Whole exome capture and NGS library preparation*

Whole exome capture was performed in accordance with the SureSelect All Human Enrichment Target Exon (Agilent's) for 51 Mb protocols. The resulting library (500 ng) was hybridised to capture probes; unhybridised material was washed away and the captured fragments were amplified for ten PCR cycles followed by purification using AMPure XP beads. The quality of the enriched libraries was evaluated using the 2100 Bioanalyzer and a High-Sensitivity DNA-kit (Agilent). The adapter-ligated fragments were quantified by qPCR using the KAPA SYBR FAST library quantification kit for Illumina Genome Analyzer (KAPA Biosystems, Woburn, MA, USA). A 1 pM solution of the sequencing library was subjected to cluster generation for the HiSeq2000 following the manufacturer instructions.

3.9.2.3. Multiplex-PCR for NGS library prepapration

The construction of NGS-libraries by means of multiplex PCR consists of the amplification of over forty amplicons divided into 6-plex (Table 3.17). This was made possible by the support of thermodynamics-based programs for checking PCR primer specificity and compatibility. These primers included M13-tags in each 5'-end of each forward and reverse primer for *posteriori* sequencer adaptor insertion (see section 3.9.3). Multiplex-PCR reactions were performed using Qiagen Multiplex-PCR Kit (Qiagen, Barcelona, Spain) as previously described in section 3.5.2. Specific Plex conditions and primers used are presented in the table 3.17.

Table 3.17.A: Multiplex-Library Conditions (**Plex A to C**).

Plex group	T. annealing	Gene (refSeq)	Exon	Primers 5' - 3' included M13-tags*	Size (bp)	10X Primer Mix (ng/µl)
A	61 ºC	***RHO*** (NG_009115.1)	1	F: CGACGGCCAGTGCTCAGGCCTTCGCAGCAT R: CAGCTATGACCAGCCCGGGACTCTCCCAGA	516	4
			2	F: CGACGGCCAGTGGGGAGTGCACCCTCCTTA R: CAGCTATGACCACCCCTACCCTGAGTGGGC	392	2
			3	F: GACGGCCAGTTGTGAAGCCCCAGAAAGGGC R: CAGCTATGACCCTCCAGGGAGCAGGAGCCG	359	2
			4	F: CGACGGCCAGTTCACGGCTCTGAGGGTCCA R: CAGCTATGACCAGCTTGTCCTTGGCAGGCA	420	4
			5	F: GACGGCCAGTACGTGCCAGTTCCAAGCACA R: AGCTATGACCTGTGGCTGGGGGAAGGTGTA	260	4
		NR2E3 (NG_009113.1)	2	F: CGACGGCCAGTGTCACGCGTGGGTTCGTTCA R: CAGCTATGACCACCCCCTCACCCCTCCAGAA	345	2
B	61 ºC	***PRPF31*** (NG_009759.1)	1-2	F: CGACGGCCAGTAGGAGGGACTTTGTCGGGG R: CAGCTATGACCGATGGGGAGGGGCACAGAGT	470	2
			3	F: CGACGGCCAGTACATCAGCCTGTCCCTGGT R: CAGCTATGACCCTTGGGCTTAGGGGCAGGA	347	2
			4	F: CGACGGCCAGTCTGTATGCTGGTGCCCGTG R: CAGCTATGACCACTGAGCCCTCGTCCACTC	381	2
			5	F: CGACGGCCAGTGCCTTCCTGAGTTCCCGAGC R: CAGCTATGACCCCCAGCCTCCTGGATCTCCC	348	4
			7	F: CGACGGCCAGTTCGAGCCCCCAGGCAGATT R: CAGCTATGACCCTGGGCCAGATGGTGGGTG	356	2
			8	F: CGACGGCCAGTGGACCCCAGGTAGAGCCAG R: CAGCTATGACCTCTCCCTGCAGAGACACCC	283	2
			9-10	F: CGACGGCCAGTGCTCAGAGGAGGCCTGGGT R: CAGCTATGACCGGCTGGCTGTGGGGTTGAG	435	6
			11-12	F: CGACGGCCAGTTCGCTGAACTGCAGGGCG R: CAGCTATGACCTCTTCACAGGGGCAGAGGG	446	6
			13	F: CGACGGCCAGTAGACAGGGCAACTCCAGGG R: CAGCTATGACCCTGGGCAGTCCCAGCAATG	423	2
C	61 ºC	***KLHL7*** (NG_016983.1)	5	F: CGACGGCCAGTGCTTCATCTTGAATGTATACTTGG R: CAGCTATGACCTCCAAACATGAGTTTTAAGAAACC	342	2
		PRPF3 (NG_008245.1)	10	F: CGACGGCCAGTTGGGATTTTCAAGATAGGAGTTA R: CAGCTATGACCGAGAAAGGCATTGAAGATCAG	418	2
		PRPF8 (NG_009118.1)	42	F: CGACGGCCAGTATCGCCCTTTGCACTTGGG R: CAGCTATGACCCTGAATGTCAGCGGCCTGT	348	4
		PRPH2 (NG_009176.1)	1a	F: CGACGGCCAGTCTGGGCTCGTTAAGGTTTGG R: CAGCTATGACCGGCATACTTGGCTGGGTCC	365	2
			1b	F: CGACGGCCAGTCAACTCGCTGGCTGGGAAG R: CAGCTATGACCTGAGCCTCAGTGTCCCCAA	441	2
			2	F: CGACGGCCAGTAGGTTTCCAGAGGCAGGGG R: CAGCTATGACCACCCAAATGGGACCGGAGG	413	2
			3	F: CGACGGCCAGTGGTCCAGCTCCCAGCGATT R: CAGCTATGACCGATGGTGCCCTCCTTGGGA	404	2

*The colours **blue** and **orange** stand for the M13-tag sequences forward and reverse, corresponding.

Table 3.17.B: Multiplex-Library Conditions (**Plex D to F**).

Plex group	T. annealing	Gene (refSeq)	Exon	Primers 5' - 3' included M13-tags*	Size (bp)	10X Primer Mix (ng/µl)
D	61 ºC	***IMPDH1*** (NG_009194.1)	2-3	F: CGACGGCCAGTGAGCGGAGGAGAGGGAACA R: CAGCTATGACCTTCCCCTACAGACCCCAGC	379	2
			5	F: CGACGGCCAGTGTGCACGAGGTGGGAACTG R: CAGCTATGACCCATCTCTCCATGCCCTGCC	327	2
			6	F: CGACGGCCAGTGTTGCCAGTGGTCGCTTG R: CAGCTATGACCCCTGTGTGCCCTGGAGT	422	2
			7	F: CGACGGCCAGTCAGGGCACACAGGAACTGAC R: CAGCTATGACCACTGAGAGGAAGGACACGCA	311	2
			8	F: CGACGGCCAGTGGCCAGCCTGGACATCATCC R: CAGCTATGACCTCTGAGGCCCCAGCGTGA	355	2
			14	F: CGACGGCCAGTGTGCTGGGGATTGGGCAGG R: CAGCTATGACCGACTGGCTGCCATCTGGGG	342	2
E	61 ºC	***PRPF8*** (NG_009118.1)	41	F: CGACGGCCAGTTGGGCTCCTTGGGAGGAAG R: CAGCTATGACCCCCAAGTGCAAAGGGCGAT	394	2
		CRX (NG_008605.1)	1	F: CGACGGCCAGTGGAGAAGGAGGCAGGATTTGA R: CAGCTATGACCTGCCAAGAGAAACGACTGTACT	274	2
			2	F: CGACGGCCAGTTGGCAACCAGGATGGAATTCT R: CAGCTATGACCGGATGGTGGGGAGAGGGATTA	419	2
			3a	F: CGACGGCCAGTGGCCTCTTCCCCACTTACC R: CAGCTATGACCCAGGCAAAGGGGACTCTGA	334	2
			3b	F: CGACGGCCAGTGTGGCCACTGTGTCCATCTG R: CAGCTATGACCGAGGCCCGATGGAGAGAGATG	488	2
		NRL (NG_011697.1)	1a	F: CGACGGCCAGTGTGGCCTCCATGTGCTCCAGA R: CAGCTATGACCTGCAGGGTAGCCAGCCAGTA	325	2
			1b	F: CGACGGCCAGTGCCTCCTTCACCCACCTTCAG R: CAGCTATGACCCCTCTCTTGGGCAGTCCTCCTTC	327	2
		PRPF31 (NG_009759.1)	6	F: CGACGGCCAGTCAGGCGGGAGATCCAGGAGG R: CAGCTATGACCGGTGCCAAAGCCCCCATTCT	331	2
		RP1 (NG_009840.1)	4	F: CGACGGCCAGTTGCTCAGTGTGGTTTAACAAA R: CAGCTATGACCAGGTGCTCCTAAGCTTATTTT	446	1
F	61 ºC	***IMPDH1*** (NG_009194.1)	9	F: CGACGGCCAGTCTGGTGCCTGTGACCAGGG R: CAGCTATGACCCCCCAGGGCTCAGTCTGGT	308	6
			10	F: CGACGGCCAGTTCACCTAGTGGCTGACTGG R: CAGCTATGACCGGAGGGGCACAGGCTTAAT	455	2
			11	F: CGACGGCCAGTAGGCTCTCCCTCCTGCCTT R: CAGCTATGACCCATGCTCCCTGCCACCCAT	298	2
			12	F: CGACGGCCAGTCAGGCAGGGGCATCCCATC R: CAGCTATGACCGTCACCCCGGAGCCTACCA	292	2
			13	F: CGACGGCCAGTGCCCCGGAGTTGCTGTTGA R: CAGCTATGACCGCCAGCAGGGAGCCCATC	339	2
			15	F: CGACGGCCAGTGGGACCTTCCTGGGCGGTA R: CAGCTATGACCGGGCCACCAAGGGTGGAGA	331	2
			16	F: CGACGGCCAGTTGAGACTGGGGGTGGCTCC R: CAGCTATGACCGCCCCCGAAGAGAGGGTGA	350	2

*The colours **blue** and **orange** stand for the M13-tag sequences forward and reverse, corresponding.

3.9.2.4. Multiplex using functionalised beads associated with emPCR

This library construction method consists of a new primer isolation technique which uses an oligonucleotide functionalised with biotin at the 3'-end attached to a streptavidin-functionalised magnetic bead as a carrier of the PCR primers. emPCR is then used in combination with this technique in order to prepare libraries with a great number of amplicons. This method can be divided into three phases:

- Linker/Primer hybridisation.
- Attachment of the Linker/Primer (LP) complex to the magnetic bead (BLP).
- emPCR using the previously created BLP to accomplish multiplex.

- *Linker/Primer hybridisation.*

The Linker contains a short sequence (Table 3.18) that is complementary to the M13-tags of the primers presented in table 3.17 and, at the same time, it is functionalised with biotin at the 3'-end.

Table 3.18: Linkers sequences and Linker/Primer hybridisation scheme.

Linker-**Forward** ||||||| **Forward**-Primer*	Biotin-3'-ATA-**GCTGCCGGTCA**-5' ||||||||||| 5'-**CGACGGCCAGT**GCTCAGGCCTTCGCAGCAT-3'*
Linker-**Reverse** ||||||| **Reverse**-Primer*	Biotin-3'-ATA-**GTCGATACTGG**-5' ||||||||||| 5'-**CAGCTATGACC**AGCCCGGGACTCTCCCAGA-3'*

* Example of Primer Forward and Reverse for *KLHL7* gene taken from Table 3.17.

This sequence will serve as a connection between the primer and the streptavidin-functionalised magnetic bead. The following table shows the reaction mixture for the hybridisation of the linkers to the primers.

Table 3.19: Hybridisation reaction mix.

Components	Volume (µl)	Final Conc.
NaCl (5 M)	1 µl	0.1 M
Linker (**Forward** or **Reverse**) (100 µM) M13-tagged_Primer (**Forward** or **Reverse**) (100 µM)	12.5 µl each	25 µM each
Distilled water	to 50 µl	n/a

This mixture was taken to 90 ºC for about 1 minute then decreased to 80 ºC for another 1 minute. After that, the temperature was very slowly decreased (ideally using a water bath) to 37 ºC. This reaction was repeated individually for all primers presented in the table 3.17, always mixing

Primer Forward with **Linker – Forward** and **Primer Reverse** with **M13_Linker – Reverse**. The LP complexes were then ready for the next step. If not used immediately, such mixtures should be stored at -20 ºC.

- *Attachment of the Linker/Primer complex to the magnetic bead (BLP)*

Each LP complex forward was mixed with its corresponding LP complex reverse. This mix was then attached to the Dynabeads® MyOne™ Spreptavidin T1, which are streptavidin-functionalised magnetic beads, following the manufacturer instructions. An overview of the entire process is represented in figure 3.3.

- *emPCR using the previously created BLP to accomplish multiplex*

After the BLP complex has been created for each individual primer pair, all BLP complexes can be mixed, given that the primer pairs are now isolated from each other thru the magnetic bead. This BLP mix contained all 44 primer pairs from table 3.17; each primer pair attached to its corresponding bead. This mix was then used for emPCR, aiming for one BLP complex per droplet. This way, each droplet functioned as a microreactor containing just one primer pair. emPCR and amplicons extraction was performed as described in section 3.5.3.

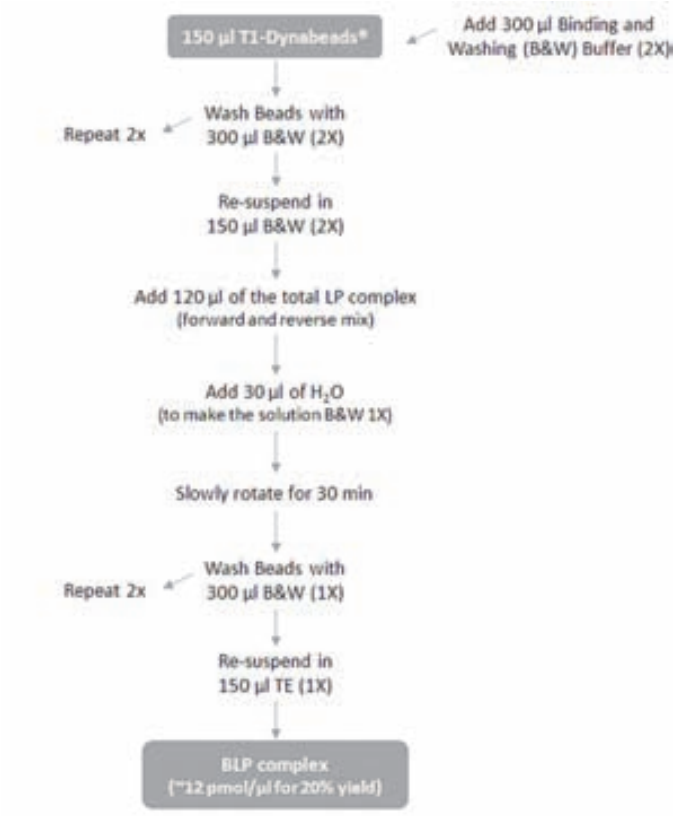


Figure 3.3: BLP process overview.

3.9.3. Pyrosequencing using the GS 454 Junior benchtop sequencer

- *Sample-specific DNA barcode with MIDs for parallel NGS using a GS Junior platform*

To simultaneously test several samples in a single sequencing run, each sample library was constructed with a specific adaptor. Each adaptor has a different sequence of ten nucleotides called the molecular identifier (MID) to distinguish each sample after NGS. For all library construction methods, except for Multiplex-PCR, fragmented DNA was mixed and 120 ng of each mix was end-repaired followed by MID-adaptor ligation and small fragment removal by AMPure® (Izasa), as reported in the Rapid Library (RL) Preparation Method: GS Junior Titanium Series (Roche) manual.

For libraries constructed through Multiplex-PCR the MID-adaptors are inserted in the DNA fragments through a second short-cycle PCR because this kind of library is based on amplicons instead of fragmented DNA.

The DNA libraries from each sample were then pooled so they could be clonally amplified through emulsion PCR and sequenced with the GS 454 Junior platform (Roche).

- *Clonal amplification of DNA libraries*

DNA libraries generated from the Fragmentase, sequence capture, and multiplex systems have different linker adaptors and need to be processed separately for sequencing in the GS Junior platform. Thus, emPCR amplification of a library sample performed by the Fragmentase and Sequence capture techniques was carried out according to the instructions of the emPCR Amplification Method Manual-Lib-L (Roche), while the emPCR amplification for DNA Multiplex libraries used the method described in the Lib-A manual.

- *Sequencing*

Once the DNA libraries possess the sequencing adaptors (or MID-adaptors) and have gone through the clonal amplification process, they are loaded into a PicoTiter sequencing plate. Preparation of the sequencing run was performed as described in the Sequencing Method Manual (GS Junior Titanium Series, Roche). DNA library sequencing was performed for 200 nucleotide cycles (approximately 500 bases) using a GS 454 *Junior* instrument (Roche) in accordance with the manufacturer's protocols.

3.9.4. NimbleGen Array System and Sequencing using SOLID platform

DNA sequencing using NimbleGen array was executed by Sistemas Genómicos, S.L. (Valencia, Spain). A total of 10 µg of target-enriched DNA from NimbleGen capture arrays was used to generate the libraries for SOLiD NGS in accordance with the protocols for sequencing with SOLiD v4 (Life Technologies). Internal controls were used to estimate the enrichment average of the capture DNA according to pre-formed libraries made by NimbleGene. The quality of these libraries was assessed with Qubit and the average size was determined using Bioanalyzer. Emulsion PCR to obtain microspheres for sequencing was also carried out in accordance with the protocols for SOLiD v4 sequencing. Thus, each library was subjected to a process of emulsion PCR for clonal amplification of the fragments followed by an enrichment process and chemical modification to allow loading into the reaction chamber. To facilitate this process the protocols and recommendations provided by Life Technologies for sequencing Solid v4 were followed. The quality and quantity of the microspheres obtained from each library was estimated taking into account the parameters obtained during the workflow analysis. The reactions for obtaining 50 nt sequences in Solid v4 were carried out afterward. The quality of the data obtained was estimated by the parameters provided by the SETS software (SOLiD Experimental Tracking System).

3.9.5. Whole exome Sequencing using the Illumina HiSeq2000© sequencer

Whole exome sequencing was executed by Sistemas Genómicos. Libraries underwent clonal amplification process (generation of clusters) and the reactions for obtaining sequences of 100 nt x 2 (Paired-end) were subsequently been carried out in the HiSeq2000 (Illumina, San Diego, CA, USA) following the manufacturer instructions. The quality of the obtained data was estimated using parameters provided by the computer software sequencing (see section 3.9.6.3).

3.9.6. Data Analysis

Each sequencer uses different a technology thus the resulting data and its analysis differ as well. Here it is explained how the data is assessed and used for each individual sequencer.

3.9.6.1. GS 454 Junior Data analysis

- *Obtaining raw sequences*

The GS Run Processor application performs the data processing of a sequencing Run to convert raw images into signal intensity values. This occurs in two steps:

1) *Image processing:* During a sequencing Run, the CCD camera on the GS Junior Instrument takes an image of the PicoTiterPlate Device for each nucleotide flow of the sequencing Run protocol. This image capturing step generates the image files (.pif) for all flows which can then be processed. The image analysis step finds raw wells (a well containing a DNA fragment that produced light due to base incorporations during the sequencing Run) across the entire PTP Device and implements algorithms to normalise the background.
2) *Signal processing:* Filters, corrects, and trims the raw flow signals to produce high quality sequence information, which is consolidated into a standard flowgram format (SFF) file.

- *Data Processing Pipelines*

The data processing steps are configured as a part of the sequencing Run using the Instrument Procedure Wizard as described in the GS Sequencer application manual (Roche). A data processing pipeline specifies the options required to perform data processing with respect to what should be processed (standard shotgun/Paired End or Amplicon libraries). The data processing pipeline options are shown in Figure 3.4.

- *Read Alignment*

For all experiments except the multiplex PCR assay we used Roche 454 GS Reference Mapper software (version 2.5p1) to assemble and compare the 454 sequencing reads to the gene reference sequences from GeneBank. After signal processing for Shotgun, reads were mapped to the reference sequence and “high confidence differences” (HCDiffs) were identified. The criteria for HCDiffs were defined by the GS Reference mapper as variants detected in at least

three non-duplicated high quality reads (both forward and reverse) and found in at least 10% of the total unique sequencing reads (i.e. non-duplicate, uniquely mapped reads that align at some location).

For the multiplex PCR assay analysis, the processed and quality-filtered reads were analysed with the GS Amplicon Variant Analyser (AVA). The amplicons (excluding adaptors and MIDs) were used as the reference to align amplicon reads; template-specific portions of the fusion primers were regarded as the forward and reverse primers. The known mutations of the selected samples were defined as substitutions relative to the reference sequence. Correspondence of samples and MID tags was specified and, as the same MID was present in both orientations, an "either" encoding multiplexer was used to demultiplex the reads.

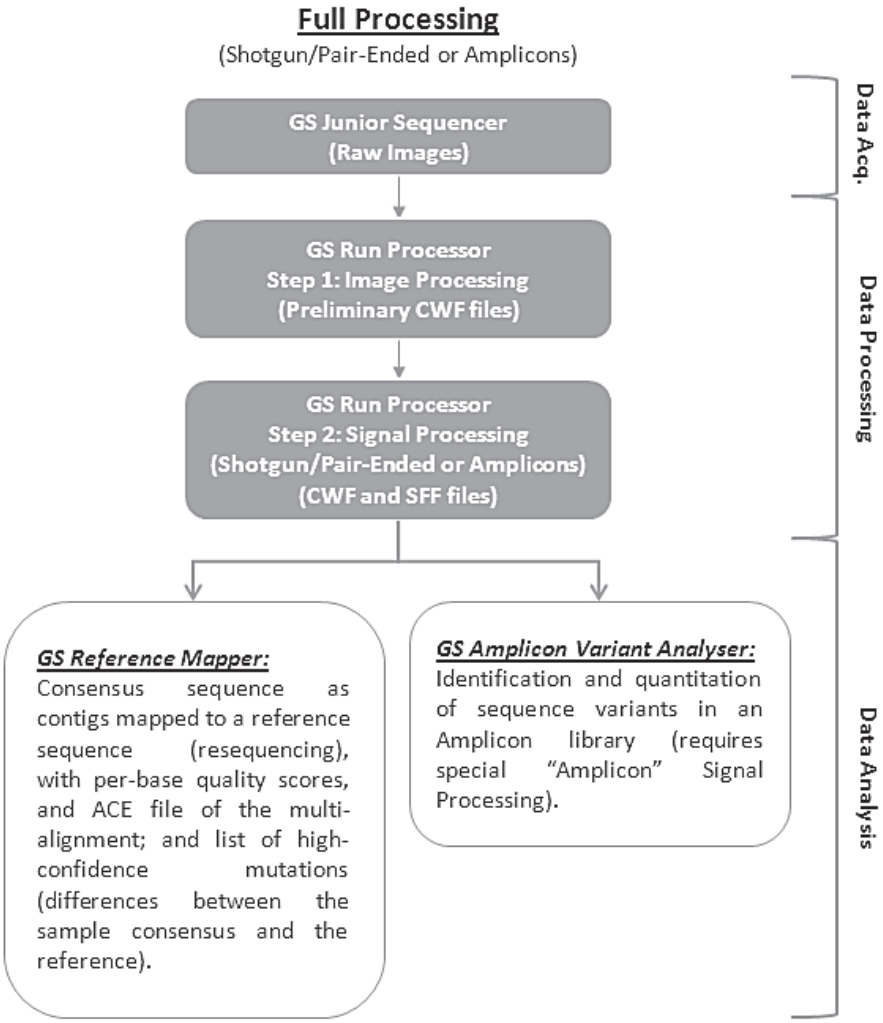


Figure 3.4: Data processing scheme in the GS *Junior* System. The blocks identify the various data acquisition, data processing, or data analysis applications and their outputs.

- *Variant Annotation*

All sequence variants were named according to the Human Genome Variation Society's recommended guidelines, using the A of the ATG translation initiation codon as nucleotide + 1. We classified each HCDiff as a single-nucleotide polymorphism (SNP) or a disease-causing mutation.

3.9.6.2. NimbleGen Array Data analysis

Data were analysed at three distinct levels:

1) *Primary* - obtaining raw sequences and their associated quality values.
2) *Secondary* - determination of variants (SNVs and indels).
3) *Tertiary* - results annotation.

- *Primary Analysis, obtaining raw sequences*

The images obtained in each ligation cycle were scanned with the software SOLiD Analysis Tools (SAT) (http://solidsoftwaretools.com/) to generate the colour-space sequences (colour codes) and the associated quality values. Then the generated raw data were normalised in order to calculate the platform standard quality control parameters.

- *Secondary analysis. Determination of variants*

The secondary analysis was divided into the following phases:

1) *Filtering:* The readings were filtered based on quality values associated with each of the generated bases. Readings with more than 25 bases and a quality value less than 9 were removed.
2) *Mapping:* The hg19 version of the human genome was used to perform this step. Up to five mismatches were allowed in the readings.
3) *Variants ID:* Prior to the identification of variants, redundant readings were eliminated in order to minimise false positives. Such readings would have been generated in the PCR amplification step which occurs during library preparation:
    - SNP Calling. This is the process by which a single nucleotide variation is identified. SNVs were called with Samtools (Li *et al.*, 2009) with the following filtering criteria: SNV quality of

20 (Phred-like score), genotype quality of 30 (Phred-like value), and a minimum coverage of 9X.

- Identification of small indels. For this step, a new alignment which permits holes was carried out with those previously unmapped reads from the phase 1 (*Filtering*). Indel calling was performed with the Small Indel tool, part of the Bioscope analysis suite. A unique filter of at least nine non-redundant reads was applied for indel detection.

The package v0.1.9 (Li *et al.*, 2009) was used to identify SNVs. The rest of the analysis was performed with Bioscope v1.2.1 (http://solidsoftwaretools.com/) and "in house" scripts written in Perl by Sistemas Genómicos.

- *Tertiary analysis. Results annotation*

The SNVs and indels identified in the secondary analysis were annotated using Application Programming Interfaces (APIs) of Ensembl V59 and several "in house" scripts written in Perl by Sistemas Genómicos. These variants then were classified attending to their position or effect on affected transcripts and accordingly with the version 59 of the Ensembl database, as follows:

- **Intergenic** – SNVs or indels in intergenic regions.
- **Regulatory region variations** – Variants in the regulatory regions.
- **Upstream** – Variants within the 5kb prior to the 5'-end of the transcript.
- **5prime UTR variations** – Variants located in the 5' untranslated regions.
- **Complex InDel** – Insertion or deletion in the intron/exon or UTR/coding region edge.
- **Intronic variations** – Variants identified in introns.
- **Synonymous** – Silent mutations. SNVs located in the coding sequence that do not cause a change in the amino acid.
- **Non-synonymous** – SNVs located in the coding sequence that cause a change in the amino acid.
- **Splice site** – Variants identified within 1-3 nt of an exon or within 3-8 nt of an intron.
- **Essential splice site** – Variants identified within the first two or the last two bases of an intron.
- **Frameshift variations** – Indel that causes a change in the reading frame.
- **Non_frameshift variations** - Indel that causes no change in the reading frame.
- **Stop gained** – SNVs that cause the appearance of a stop codon.
- **Stop lost** – SNVs that cause the loss of a stop codon.
- **3Prime UTR variations** – Variants located in the 3' untranslated regions.
- **Downstream** – Variants within the 5kb prior to the 3'-end of the transcript.

- **Partial codon** – Variants identified at the end of a transcript, in an incomplete codon, whose coding sequence has been cut-out and its end is unknown.
- **Within mature miRNA** – Variants located in a mature miRNA.
- **NMD transcript** – Variants located on a transcript which is degraded thru a cellular mechanism by nonsense-mediated decay (NMD).
- **Within non-coding gene** – SNVs or indels located in a non-coding gene.

At last, the SIFT program (Li *et al.* 2009) was used to determine the effect on the produced protein of the new variants whose transcript level effect had been previously classified as "**Non-synonymous**".

3.9.6.3. Whole Exome Data analysis

- *Raw data quality assessment*

Similar to Sanger sequencing, NGS platforms generate error probability values per base which are known as quality values. Quality values are provided in Phred-like scale (Ewing *et al.*, 1998a, 1998b) so that a value of 20 indicates a probability of 1 out of 100 of that base being wrong. The global study of the quality values provides information about the sequencing quality. Three different plots were generated for each sample's read group (R1 and R2, similar to Sanger's nomenclature "forward" and "reverse" to define reads obtained from the same DNA fragment):

1) Per base sequence quality – Quality base distribution across all read bases.
2) Per base sequence content – Nucleotide content across all bases.
3) Per sequence quality scores – Quality score distribution over all sequences.

- *Read alignment and target enrichment assessment*

Reads were aligned and compared with the human reference genome version GRCh37/hg19. Read alignment was performed using BWA and 'in-house' scripts. After read mapping, it is important to filter any sequences that can introduce major biases and/or noise in later steps. Low quality reads and sequences flagged as PCR duplicates were removed from the BAM formatted file obtained after read mapping. In addition, the overall sample coverage and the efficiency of the combination of the selected strategies (target enrichment system + NGS platform) were evaluated at this point. Three different parameters were calculated for this purpose:

1) Coverage distribution along targeted regions
2) Percentage of target bases covered at 1x, 10x, and 20x for each chromosome
3) Percentage of on-target reads against the total number of mapped reads.

Filtering processes were performed using Picard-tools (http://picard.sourceforge.net/) and SAMtools (Li *et al.*, 2009). Coverage metrics and evaluation of the target enrichment system were performed using custom scripts.

- *Variant calling*

Variant calling is the process in which variants are identified. Variant identification is performed using the information from read alignments. Thus mismatches found between the read and the reference genomes are more deeply studied to identify real variants. Variant calling was performed using a combination of two algorithms: VarScan (Koboldt *et al.*, 2009) and GATK (Flicek *et al.*, 2012). 'In-house' scripts were developed by Sistemas Genómicos to combine and filter variants. Identified variants were annotated using the Ensembl database (Flicek *et al.*, 2012). This database contains information from the most relevant human variation resources such as dbSNP, the HapMap project, the 1000Genomes project, COSMIC, and many others (http://www.ensembl.org/). Variants were classified according to their position or effect on affected transcripts as follows:

- **splice_donor_variant** – Splice variant that changes the 2 base region at the 5'-end of an intron.
- **splice_acceptor_variant** – Splice variant that changes the 2 base region at the 3'-end of an intron.
- **stop_gained** – A sequence variant whereby at least one base of a codon is changed resulting in a premature stop codon and leading to a shortened transcript.
- **frameshift_variant** – A sequence variant which causes a disruption of the translational reading frame because the number of nucleotides inserted or deleted is not a multiple of three.
- **stop_lost** – A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript.
- **initiator_codon_variant** – Codon variant that changes at least one base of the first codon of a transcript.
- **inframe_insertion** – An inframe non-synonymous variant that inserts bases into the coding sequence that are a multiple of three.
- **inframe_deletion** – An inframe non-synonymous variant that deletes bases from the coding sequence that are a multiple of three.

- **missense_variant** – A sequence variant that changes one or more bases, resulting in a different amino acid sequence but preserving the length of the transcript.
- **transcript_amplification** – A feature amplification of a region containing a transcript.
- **splice_region_variant** – A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron.
- **incomplete_terminal_codon_variant** – A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed.
- **synonymous_variant** – A variant where there is no resulting change to the encoded amino acid.
- **stop_retained_variant** – A sequence variant where at least one base in the terminator codon is changed but the terminator remains.
- **coding_sequence_variant** – A sequence variant that changes the coding sequence.
- **mature_miRNA_variant** – A transcript variant located with the sequence of the mature miRNA.
- **5_prime_UTR_variant** – A UTR variant of the 5' UTR.
- **3_prime_UTR_variant** – A UTR variant of the 3' UTR.
- **intron_variant** – A transcript variant occurring within an intron.
- **NMD_transcript_variant –** A variant in a transcript that is the target of NMD.
- **non_coding_exon_variant** – A sequence variant that changes non-coding exon sequence.
- **nc_transcript_variant** – A transcript variant of a non-coding RNA.
- **upstream_gene_variant** – A sequence variant located 5' of a gene.
- **downstream_gene_variant** – A sequence variant located 3' of a gene.
- **TFBS_ablation** – A feature ablation whereby the deleted region includes a transcription factor binding site.
- **TFBS_amplification** – A feature amplification of a region containing a transcription factor binding site.
- **TF_binding_site_variant** – A sequence variant located within a transcription factor binding site.
- **regulatory_region_variant** – A sequence variant located within a regulatory region.
- **regulatory_region_ablation** – A feature ablation whereby the deleted region includes a regulatory region.
- **regulatory_region_amplification** – A feature amplification of a region containing a regulatory region.
- **feature_elongation** – A sequence variant that causes the extension of a genomic feature with regard to the reference sequence.
- **feature_truncation** – A sequence variant that causes the reduction of a genomic feature with regard to the reference sequence.
- **intergenic_variant** – A sequence variant located in the intergenic region, i.e. between genes.

3.10. Prediction tools for mutation consequence

Next Generation Sequencing generates huge amounts of information thru an enormous quantity of sequence variations. The prediction tools serve to prioritise the discovered mutations, emphasising the mutations that have a negative impact on the coded protein. Hence, various *in silico* bioinformatic tools have been developed to predict the probable pathogenicity of missense variants. Here, three online prediction tools were put to use. These tools are described below:

3.10.1. Splice Site Prediction by Neural Network

In-silico splice site prediction tools can be used to predict the effect of a genetic variant on splicing. To do so, the Splice Site Prediction tool by Neural Network, from the Berkeley Drosophila Genome Project, was used. This tool can be accessed via the following webpage:

- http://www.fruitfly.org/seq_tools/splice.html

Here, the entire sequence of interest is loaded into the webpage, in FASTA format, taking care to use only single-letter nucleotides (A, C, G and T). After submitting the query, the results are divided into donor site and acceptor site predictions. To each predicted donor/acceptor site, a score is given which is presented as a probability (expressed as a figure between 0 and 1).

3.10.2. PolyPhen

PolyPhen (Polymorphism Phenotyping) is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations (Adzhubei *et al.*, 2010). This tool can be accessed via the following webpage:

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

Here, the entire protein sequence is loaded into the webpage, in FASTA format, then the amino acid change and position is indicated. After submitting the query, the results are presented as Damaging, Probably Damaging, Possibly Damaging, Benign, and Neutral, in descending order of severity.

3.10.3. MutPred

MutPred (Mutation Prediction) is an algorithm that predicts whether an amino acid substitution (AAS) will be disease-associated or neutral (Li *et al.* 2009). MutPred predicts the molecular cause of disease/deleterious AAS based upon the gain or loss of 14 different structural and functional

properties, for example, loss of a phosphorylation site. This tool can be accessed via the following webpage:

- http://mutpred.mutdb.org/

This tool requires a protein sequence in FASTA format, a list of amino acid substitutions, and an email address. After submitting the required information, the results are e-mailed and presented as a probability (expressed as a figure between 0 and 1) of an AAS to be deleterious/disease-associated. A missense mutation with a MutPred score > 0.5 could be considered as potentially 'harmful', while a MutPred score > 0.75 should be considered a high confidence 'harmful' prediction.

3.10.4. SIFT

SIFT (Sorting Intolerant From Tolerant) is an algorithm which predicts whether an AAS will affect protein function based on sequence homology and the physical properties of amino acids (Ng *et al.*, 2001, 2002, 2003, 2006; Kumar *et al.*, 2009;). SIFT prediction is based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences, collected through PSI-BLAST. This tool can be accessed at the following webpage:

- http://sift.jcvi.org/

It requires the ENSP (Ensembl Protein) numbers and any substitutions. The results are presented as a probability (expressed as a figure between 0 and 1) of an AAS to be Damaging or Tolerant. An AAS with a SIFT score of less than 0.05 is predicted to be deleterious (or damaging). One with a score greater than or equal to 0.05 is predicted to be tolerable.

4. Results

4.1. Detection of Genomic Variants in Large Genes by LR-PCR and NGS

- Context

For the analysis of large genes such as those associated with the risk of breast cancer and ovarian cancer (*BRCA1* and *BRCA2*) a new approach was evaluated. This approach used LR-PCR together with DNA enzymatic shearing for the NGS library preparation as explained in the section 3.9.2.1 of Material and Methods. Two experiments have been assayed to evaluate the appropriateness of this approach. The first experiment targeted the comparison between both shearing methods (Fragmentase/Nextera) while the second experiment sought to assess the number of samples that could be sequenced in parallel using this novel approach with the GS Junior platform.

The DNA samples were from Spanish patients with a family history of breast or ovarian cancer, who had been previously screened for mutations in *BRCA1* and *BRCA2* coding exons and flanking regions by direct sequencing. These samples had different types of breast cancer–causing mutations including point mutations, small deletions, and small insertions (Table 4.1).

Table 4.1: *BRCA* genes point mutations previously identified in each sample

DNA Sample	Gene	Primary mutation
1	*BRCA2*	c.8851G>A
2	*BRCA2*	c.3264dupT
3	*BRCA2*	c.8174_8185delGGTATGCTGTTinsTT
4	*BRCA1*	c.3359_3360delTT
5	*BRCA1*	c.5123C>A

These samples with these particular types of variants were chosen to evaluate the ability of the GS Junior sequencer to identify different types of sequence variation. Moreover, the samples carried several previously described single nucleotide polymorphisms (SNPs).

The *BRCA* genes were amplified thru eleven long range fragments as described in section 3.5.1.1 of Materials and Methods. The LR-PCR products were analysed by agarose gel electrophoresis (Figure 4.1) and used for library construction.

Figure 4.1: The PCR for *BRCA1* and *BRCA2*. Agarose gel electrophoresis,1%, of 11 long-range fragments fragments (A–K). MWM, molecular weight marker.

4.1.1. NGS libraries generated by LR-PCR with Fragmentase or Nextera Technology

4.1.1.1. LR-PCR Amplification of BRCA genes and library construction for NGS

For the analysis using the LR-PCR approach allied with two distinct DNA shearing methods (Nextera and Fragmentase), the DNA sample 2 (Table 4.1) was chosen. This sample presented the mutation c.3264dupT in the *BRCA2* gene and served as a control between the different shearing methods tested.

Both *BRCA1* and *BRCA2* genes have a large exon 11 that contains almost 50% of the coding sequencing (CDS) for *BRCA1* and 60% for *BRCA2*. This exon 11 can be amplified by LR-PCR in a single fragment: D for *BRCA1* and H for *BRCA2*. With 11 fragments (Table 3.5 from Material and Methods) of LR-PCR, all CDS of both genes were amplified. To generate long and accurate PCR fragments, a formulated blend of proofreading and high fidelity polymerase was used (as described in Materials and Methods). These DNA fragments were used to perform the NGS libraries by Fragmentase or the Nextera technology.

4.1.1.2. Fragmentase NGS Library

The NEBNext dsDNA fragmentase (Fragmentase) method uses two endonucleases: one randomly generates nicks on dsDNA and the other recognises the nicked site and cuts the

opposite DNA strand across from the nick, producing dsDNA breaks. Although these breaks are nearly random, differences in size and GC content of the target fragment may influence the fragmentation time.

- *Fragmentation kinetics study*

Due to the fact that fragmentation of DNA with Fragmentase is a time dependent reaction, a kinetic study for each LR-PCR fragment was performed prior to the library preparation. Briefly, Fragmentase reaction was carried out as described in the section 3.9.2.1 of the Materials and Methods chapter and aliquots of 10 µl were taken at different times. Agarose gel electrophoresis of each aliquot was used to analyse the range of lengths of the fragmented DNA. The time of Fragmentase digestion for each DNA fragment that rendered an average length of 750 bp was annotated (Table 3.5 from Material and Methods). Figure 4.2 shows the electrophoresis analysis of the fragments rendered in a Fragmentase kinetic reaction carried out on a LR-PCR fragment.

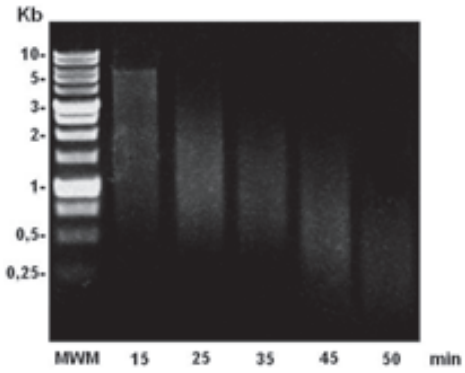


Figure 4.2: Fragmentase kinetics: MWM is a molecular weight marker: the numbers on the abscissa correspond to the minutes of digestion of a LR-PCR fragment with Fragmentase.

4.1.1.3. Nextera NGS Library

The Nextera (Caruccio, 2011) technology reaction involves an *in vitro* transposition reaction that catalyses the random nick of double-stranded breaks in the target DNA and covalently attaches the 3' end of the transferred transposon strand to the 5' end of the target DNA (Figure 4.3). In contrast with Fragmentase reaction, the Nextera reaction was assayed in the only condition indicated by the manufacturer as described in the section 3.9.21 of Materials and Methods.

Figure 4.3: Nextera technology to prepare RocheTitanium – compatible libraries. (**A)** A limited-cycle PCR with a four-primer reaction adds Roche Titanium– compatible adaptor sequences. The GS Junior Titanium Primers (**blue and red lines**), an optional MID (**circle**), and the transposon sequence (**grey line**) are shown. (**B)** Nucleotide sequences of the GS Junior Titanium Primers (P1and P2) are shown, together with 454-compatible adaptor sequences (Ad1 and Ad2), followed by the key (underlined), an optional MID, and the transposon sequence (**grey**) (edited from Hernan *et al.*, 2012).

4.1.1.4. Sequencing Data Analysis

In these two runs, one PicoTiter Plate contained just one library, one processed with Fragmentase and the other with Nextera. The sequenced libraries generated 99,097 high-quality reads (47.7% of the total raw reads) with Fragmentase and 82,334 high-quality reads (37.67% of the total raw reads) with Nextera in the GS Junior platform, in which 96,845 and 60,200 reads were successfully aligned with the references, respectively (Table 4.2). The average read length was 374 bases pairs for the Fragmentase run and 292 base pairs for the Nextera run. The resulting data generated from both Fragmentase and Nextera runs were analysed using GS Reference Mapper software version 2.5p1. The average depths obtained in CDS of BRCA1 and BRCA2 were 329X and 478X, respectively, when the Fragmentase library was analysed, and 108X and 299X, respectively, when the Nextera library was analysed.

Table 4.2: Sequence reads obtained by NGS in Fragmentase and Nextera runs.

Sample/Run	Nº of high-quality reads aligned to reference per sample	% of total high-quality reads
Fragmentase	96,845	97.73
Nextera	60,200	80.70

Comparing both (Fragmentase and Nextera) runs, in both methods some target sequence regions were seen to have a different depth representation (Figure 4.4). This shows that although the fragmentation of DNA by the recombinase of Nextera or the endonuclease mix formulated in Fragmentase are mostly random, a sequence effect may exist in the digestion mechanism of both systems.

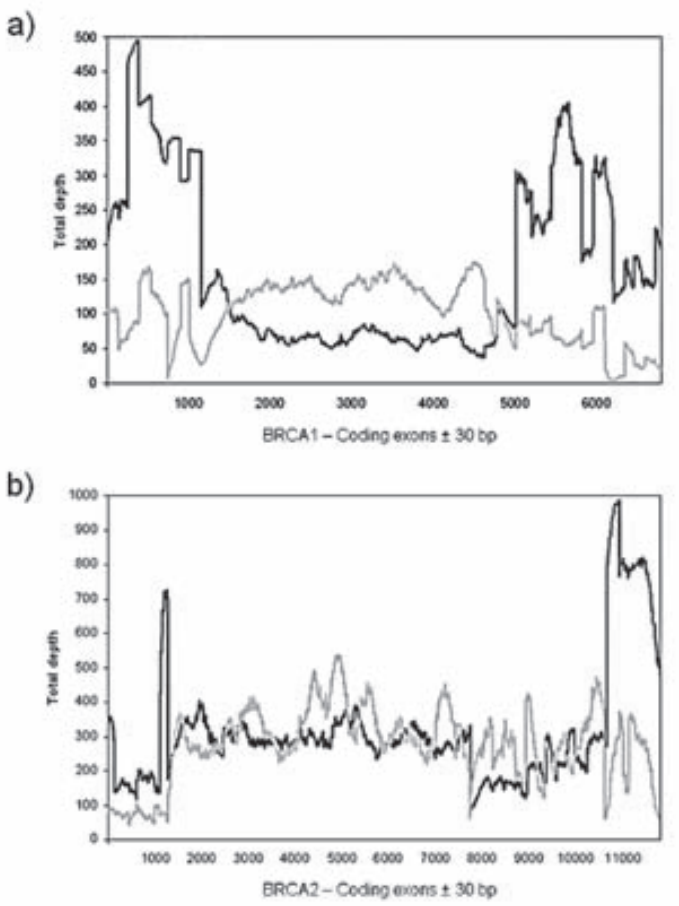


Figure 4.4: Depth profile of the amplified coding and the flanking regions (30bp) of the *BRCA* genes. Graphic representation of depth sequence obtained from Fragmentase (**dark grey**) and Nextera (**light grey**), runs for both a) *BRCA1* and b) *BRCA2* genes.

Graphic representation of the depth and coverage showed a similar profile on both runs in *BRCA1* and *BRCA2* sequencing runs, demonstrating the reproducibility of the method to prepare libraries for NGS.

An analysis of depth in intronic regions showed some sequences that were overrepresented, whereas others were under represented. An analysis of these sequences (Figure 4.5) illustrates that sequences of minor representation showed, in general, a significant content in homopolymeric A and T stretches. Moreover, these homopolymers were accurately sequenced up to seven nucleotides, but for longer sequences, inaccurate and inefficient sequencing was regularly observed. However, poor coverage depth in regions free of homopolymeric stretches were also found, suggesting that other unknown sequence effects or sequencing specificity could also be present on both Fragmentase and Nextera shearing mechanisms. On the other hand, overrepresented sequences showed no more regions of higher melting stability or content in guanine and cytosine nucleotides than average.

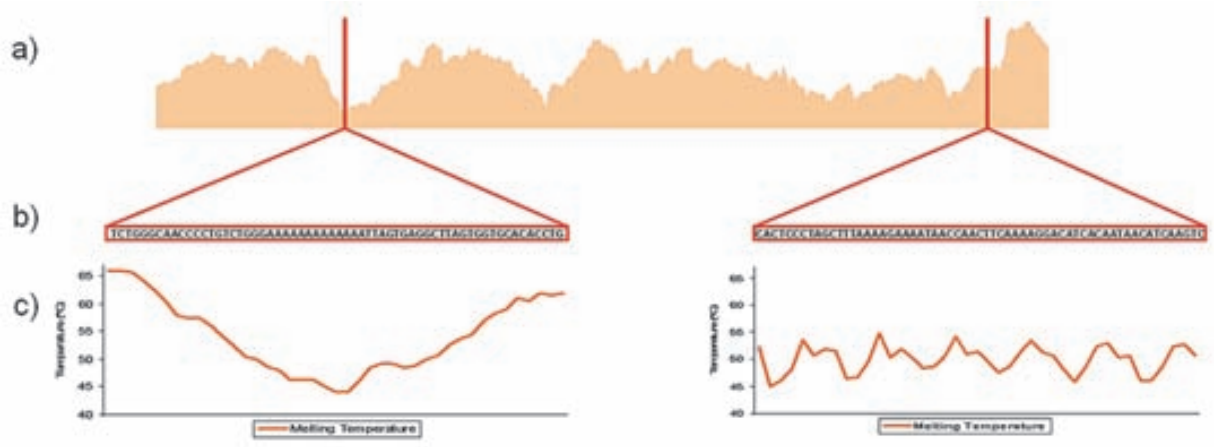


Figure 4.5: Depth, sequence, and melting profile analysis in *BRCA1*. Comparison of sequence and melting profile between a low and a high depth region. a) Depth profile of the entire *BRCA1* gene. b) Detail of the sequence analysed. c) Melting temperature profile of the specified sequence.

In both the Fragmentase and Nextera sequencing runs, the GS Junior platform detected the point mutation c.3264dupT associated with breast cancer found and the SNPs found with the Sanger method (Table 4.3).

Table 4.3: Point Mutation and SNPs identified in CDS of *BRCA* genes in Fragmentase and Nextera sequencing runs.

				Fragmentase		Nextera	
Point Mutation	Gene	Variant	Protein Change	% Variant	Total Depth	% Variant	Total Depth
	BRCA2	c.3264dupT	p.Gln1089fs	55.0	163	46.0	123
Single-Nucleotide Polymorphism	Gene	Variant	db SNP *	% Variant	Total Depth	% Variant	Total Depth
	BRCA1	c.2077G>A	4986850	56.7	161	40.2	132
		c.2082C>T	1799949	54.5	155	43.8	128
		c.2311T>C	16940	42.9	136	42.5	113
		c.2612C>T	799917	38.9	143	41.6	137
		c.3113A>G	16941	52.6	181	38.1	147
		c.3548A>G	16942	55.8	179	43.4	106
		c.4308T>C	1060915	45.6	168	40.6	69
		c.4837A>G	1799966	54.5	189	67.9	56
	BRCA2	c.3396A>G	1801406	99.4	166	96.1	129
		c.4558A>G	80358690	100.0	176	100.0	146
		c.6513G>C	206076	100.0	146	98.3	173
		c.7242A>G	1799955	98.9	73	99.0	104

* db SNP, single-nucleotide polymorphism database (http://www.ncbi.nlm.nih.gov/snp/)

The homozygous SNPs analysed always gave total variation values >96.1% (mean, 99.6% for Fragmentase and 98.4% for Nextera). The heterozygous SNPs showed values between a minimum of 38.9% and a maximum of 56.7% (mean, 50.2%) for Fragmentase and a minimum of 38.1% and a maximum of 67.9% (mean, 44.8%) for Nextera. The point mutation, which also worked as control between both runs, was successfully detected with a total variation of 55.0% and 46.0% for the Fragmentase and Nextera sequencing runs, respectively.

4.1.2. NGS library construction for parallel analysis of five samples by LR-PCR

4.1.2.1. LR-PCR Amplification of BRCA genes and NGS library construction

For the analysis of five samples in parallel, the DNA samples 1 to 5 (Table 4.1) were used to amplify the *BRCA1* and *BRCA2* genes, as described in the section 3.5.1.1 of Materials and Methods. *BRCA1* and *BRCA2* genes were amplified thru eleven fragments (Table 3.5 from Material and Methods) containing all coding and flanking sequences by LR-PCR. Primers and conditions were designed to allow the amplification of these genes in a simple robust PCR assay. The amplification of the eleven DNA fragments was performed in a PCR program with a gradient of annealing temperatures. Thus, by using a 96-well plate, it was possible to successfully amplify the 11 LR-PCR fragments in a single PCR run (Figure 4.6). By using Nextera technology, we obtained a library of *BRCA1* and *BRCA2* fragments from five different DNA samples ready to be sequenced in the 454 System.

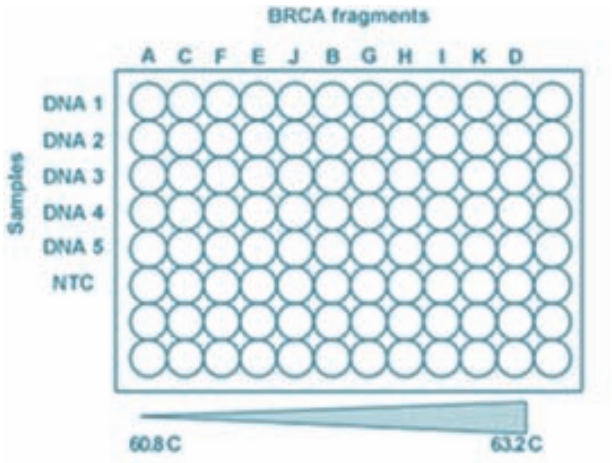


Figure 4.6: The PCR for *BRCA1* and *BRCA2*. PCR plate layout used to amplify the 11 long-range fragments (A–K) in a single PCR program with a gradient temperature annealing step (shown at the bottom of the diagram).The row and column schemes refer to DNA samples and amplified PCR fragments, respectively. NTC, nontemplate control (edited from Hernan *et al.*, 2012).

4.1.2.2. Barcode for Parallel NGS

The five DNA libraries, obtained in parallel from the samples 1 to 5 (Table 4.1), were performed by Nextera technology as previously described in section 3.9.3 of the Material and Methods. This was possible by adding a sequence of 10 nucleotides, called the molecular identifier (MID), in the upstream sequence of the transposon end in the Nextera adaptors (Figure 4.3) to distinguish each sample after NGS. In this study, the MIDs 1 to 5, corresponding to the DNA samples 1 to 5, of the 454Standard MID set (Roche), compatible with a shotgun library protocol, were used.

4.1.2.3. Sequencing Data analysis

The samples selected for parallel *BRCA* analysis by the GS Junior platform had different types of breast cancer–causing mutations, including point mutations, small deletions, and small insertions (Table 4.1). Samples with these particular types of variants were chosen to evaluate the ability of the GS Junior sequencer to identify different types of sequence variation. Moreover, the samples carried several previously described SNPs. The CDS and 30 bp (intronic flanking) of each exon of *BRCA1* and *BRCA2* was sequenced in the five selected samples by the standard Sanger method.

In this run, one PicoTiter Plate contained five different samples that were distinguished with MIDs. Variants detected by the GS Junior platform were then compared with those obtained by Sanger sequencing. The sequenced parallel library generated 80,250 high-quality reads (35.2% of the total raw reads) in the GS Junior platform, in which 66,351 reads were successfully aligned with the reference samples (Table 4.4). The average read length was 296 base pairs. The resulting data generated from the parallel run were analysed using the GS Reference Mapper software version 2.5p1 The average depths obtained in CDS of *BRCA1* and *BRCA2* were 75X and 90X, respectively, when five samples were analysed together.

Table 4.4: Sequence reads obtained by NGS in the parallel run.

Sample/MID	Nº of high-quality reads aligned to reference per sample	% of total high-quality reads
1	11,226	13.99
2	19,799	24.67
3	12,008	14.96
4	8,249	10.28
5	15,069	18.78
Total	66,351	82.68

Similar to all samples processed in parallel, analysis of total CDS depth demonstrated several regions in *BRCA1* with a low depth. These regions, comprising exons 10 (77 bp), 19 (41 bp), 20 (84 bp), 22 (74bp), and 24 (125 bp) also showed a significant low-depth coverage (<10X) (Figure 4.7).

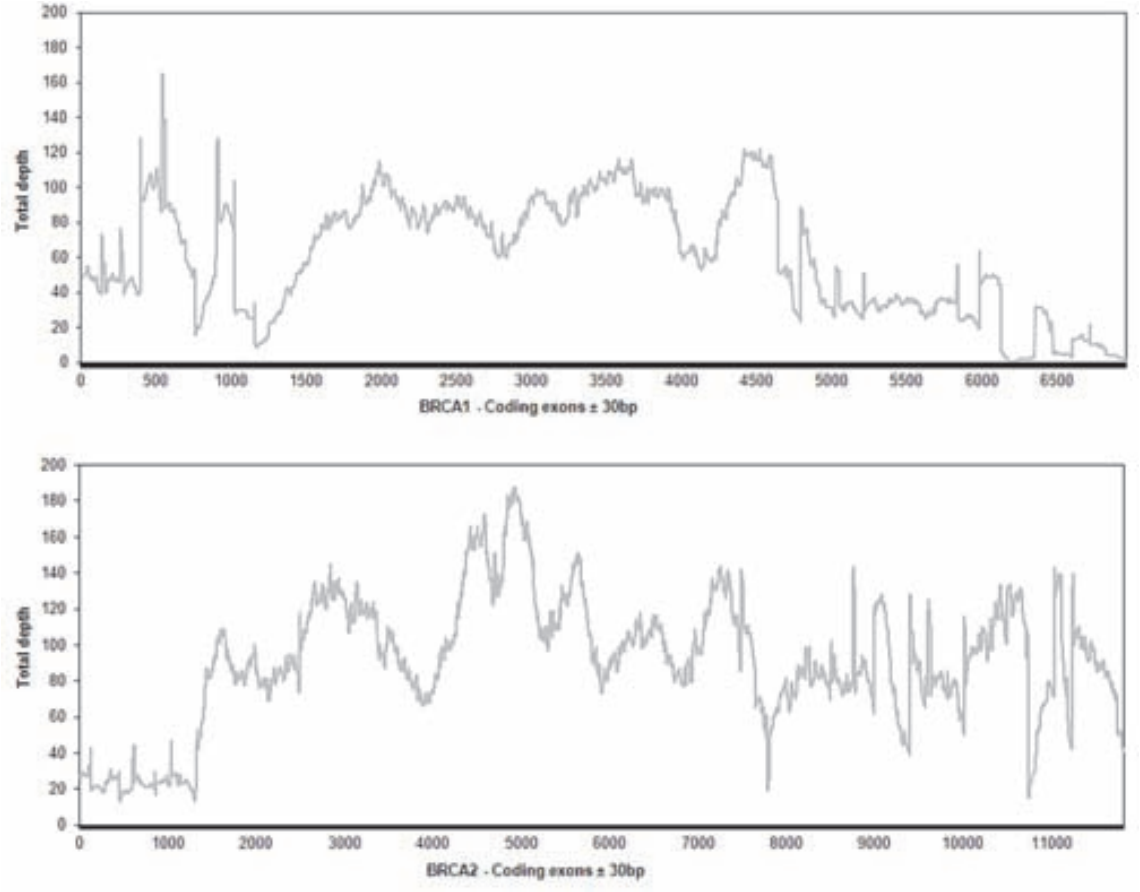


Figure 4.7: Average depth profile of the CDS and flanking regions (30bp) of the *BRCA* genes between the five samples.

As noted in the results of the single runs (see section 4.1.1.4), some under-expressed regions near homopolymer sequences were also observed. This lower representation may reflect selective discrimination by quality control of such sequence traits in NGS pyrosequencing platforms (Figure 4.5). However, every Sample/MID rendered sufficient sequence depth through almost all *BRCA1* and *BRCA2* coding and flanking sequences (Figure 4.7) with the exception of the previously commented regions of the *BRCA1* gene.

The point mutations associated with breast cancer found with the Sanger method were detected in the five samples (Table 4.5). These mutations were detected with a total variation between a minimum of 34.5% and a maximum of 54.0% (mean, 45.7%).

Table 4.5: Mutations Identified in *BRCA* Genes by the 454 Sequencer in the Parallel Run.

Sample/MID	Gene	Position	Reference base	Modified base	Protein Change	% Variant	Total depth
1	*BRCA2*	c.8851	G	A	p.Ala2951Thr	46.3	25
2	*BRCA2*	c.3264	T	TT	p.Gln1089fs	34.5	60
3	*BRCA2*	c.8174_8185	GGTATGCTGTTA	TT	p.Trp2725fs	50.8	21
4	*BRCA1*	c.3359_3360	TT	-*	p.Val1120fs	42.9	21
5	*BRCA1*	c.5123	C	A	p.Ala1708Glu	54.0	51

*Denotes the deletion of TT bases. PTP, PicoTiter Plate.

The SNPs found with the Sanger method were detected as well (Tables 4.6). The homozygous SNPs analysed always provided total variation values >91.7% (mean, 99.2%), whereas the heterozygous SNPs showed values between a minimum of 33.3% and a maximum of 69.2% (mean, 49.1%).

The results of the parallel run were also analysed with an external software program: CLC Genomics Workbench version 4.8 (CLCBio). The complete sequence analysis of *BRCA1* and *BRCA2* using this software showed a similar coverage profile to that obtained with GS Mapper software version 2.5p1 (Roche), but, in general, with a lower depth coverage; 99% of the validated variants were detected by both programs when default settings were used. However, CLC software detected more indels than Reference Mapper, although most of those were found in poly(A) regions and were probably false positives.

Table 4.6: SNPs identified in CDS of *BRCA*s in the parallel run for each sample.

Gene	Variant	db SNP *	MID 1		MID 2		MID 3		MID 4		MID 5	
			% Variant	Total Depth	% Variant	Total Depth	% Variant	Total Depth	% Variant	Total Depth	% Variant	Total Depth
BRCA1	c.591C>T	1799965	0.0	63	0.0	74	0.0	26	0.0	37	64.4	59
	c.2077G>A	4986850	44.4	36	34.4	64	48.8	41	0.0	41	40.4	47
	c.2082C>T	1799949	47.1	34	33.3	63	51.2	41	0.0	41	45.7	46
	c.2311T>C	16940	46.2	26	33.3	45	57.1	28	0.0	34	50.0	36
	c.2612C>T	799917	47.6	42	46.8	62	57.1	42	0.0	40	50.0	36
	c.3113A>G	16941	40.4	47	57.7	71	42.6	47	0.0	49	38.1	42
	c.3548A>G	16942	56.7	30	53.8	52	34.8	23	0.0	35	56.7	30
	c.4308T>C	1060915	53.3	21	40.6	32	68.8	22	0.0	38	60.0	25
	c.4837A>G	1799966	43.8	32	68.0	25	69.2	26	0.0	38	59.1	44
BRCA2	c.865A>C	766173	43.3	30	0.0	53	0.0	35	0.0	28	0.0	34
	c.1114A>C	144848	0.0	37	0.0	61	0.0	43	0.0	34	54.9	51
	c.1365A>G	1801439	48.4	31	0.0	62	0.0	39	0.0	36	0.0	43
	c.2229T>C	1801499	40.0	30	0.0	95	0.0	51	0.0	36	0.0	62
	c.2971A>G	1799944	43.2	37	0.0	65	0.0	40	0.0	29	0.0	48
	c.3396A>G	1801406	48.3	29	100.0	57	43.2	44	91.7	34	39.5	43
	c.4558A>G	80358690	100.0	33	100.0	78	100.0	37	100.0	33	100.0	39
	c.5744C>T	4987117	51.2	41	0.0	76	0.0	48	42.4	43	0.0	60
	c.6513G>C	206076	100.0	31	98.7	76	98.2	57	100.0	31	100.0	65
	c.7242A>G	1799955	53.1	32	100.0	47	62.5	24	100.0	43	0.0	35

* db SNP, single-nucleotide polymorphism database (http://www.ncbi.nlm.nih.gov/snp)

4.2. NGS Analysis of autosomal dominant Retinitis Pigmentosa

4.2.1. Detection of genetic by NGS in known genes causing adRP

- Context

In heterogeneous diseases such as autosomal dominant Retinitis Pigmentosa (adRP), where multiple candidate genes are involved, hundreds of individual exonic sequences should be analysed in order to achieve a successful molecular diagnosis. Thus, genetic variants in more than a dozen genes have been reported for adRP. The challenge here was to introduce rational methods of molecular analysis to study a few patients in a relatively short time and, thus, meet the demand for a more efficient clinical diagnostic.

Accordingly, several approaches to be used together with the GS Junior platform were devised and evaluated with the goal of developing simple and effective methods for detecting DNA genomic variations in several genes associated with adRP which can then be incorporated into a molecular testing routine protocol without the addition of any special equipment.

4.2.1.1. NGS libraries generated by LR-PCR allied with Fragmentase Technology

4.2.1.1.1. Selected cases for study

Two runs were performed to evaluate this approach, one run featuring a single library per sequencing plate (PicoTiter Plate or PTP) and another run in which multiple libraries were analysed in parallel.

- *Single NGS Library*

For the single sample run, a chimeric sample was generated that was composed of a mix of five long-range PCR (LR-PCR) fragments (amplified from five patients), each containing a previously characterised mutation causing adRP (Table 4.7), plus fifteen LR-PCR fragments from a control individual was generated.

Table 4.7: adRP point mutations previously identified present in the chimeric sample.

Gene	Primary mutation
CA4	c.700G>A
CRX	c.253-15G>A
PRPF31	c.769_770insA
PRPF8	c.6968_6988del21bp
PRPH2	c.641C>T

- *Parallel Barcoded NGS library*

For the parallel NGS run, a pool of four complete libraries was generated. Three of these four libraries were built using genomic DNA samples from three index patients with uncharacterised adRP.

The fourth library, which included the same chimeric sample used in the single sample run, contained the five previously characterised adRP-causing mutations (Table 4.7). These five mutations served as a positive control between the single and parallel NGS experiments.

The three index patients with uncharacterised adRP were identified with the MIDs 1 to 3 and the chimerical sample was identified with the MID 4; all from the 454Standard MID set (Roche), compatible with a shotgun library protocol, as described in section 3.9.3 of the Materials and Methods.

4.2.1.1.2. LR-PCR Amplification of 12 genes associated with adRP and NGS library construction

- *LR-PCR fragments Amplification*

The 12 genes associated with adRP, which contain more than 97% of the currently known adRP-causing mutations, were successfully amplified in 20 fragments (Figure 4.8), containing all coding and flanking sequences, with LR-PCR. The DNA fragments were used to perform the NGS libraries by Fragmentase technology.

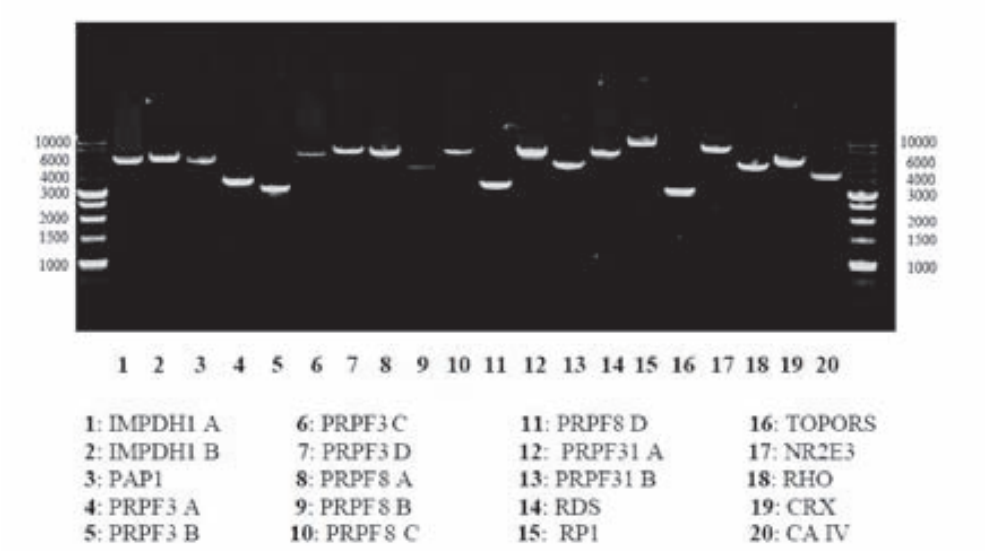


Figure 4.8: Electrophoresis of the LR-PCR fragments. 1.0% agarose gel electrophoresis of the long-range PCR fragments obtained using the primers and conditions shown in Table 3.6. Lanes at the left and right ends are molecular DNA markers.

- *Fragmentation kinetics study and library preparation*

Due to the heterogeneous size and sequence (GC content) of the LR-PCR fragments used, a kinetic study for each LR-PCR fragment was performed and the reaction time needed to obtain fragments to an average of 750 bp was annotated (Table 3.6 of the Material and Methods). An equimolar mix of the fragments with similar kinetics was made to facilitate simultaneous digestion in a single reaction as described in section 3.9.2.1. of the Material and Methods.

4.2.1.1.3. Sequencing Data analysis

The sequence analyses for both the individual and the parallel run were performed using GS Reference Mapper software. In the individual run, a total of 106,164 high-quality reads was obtained (72.4% of total raw reads), of which 104,874 of these filtered reads were matched with genomic sequences of the 12 genes (Table 4.8).

Table 4.8: Sequence reads obtained by NGS with the single chimerical run.

Nº of high-quality reads aligned to reference	% of total high-quality reads
104,874	98.78

In the parallel run, a total of 120,506 high-quality reads was obtained (61.4% of total raw reads), of which 112,984 of these filtered reads were matched with genomic sequences of the 12 genes (Table 4.9). The average sequences reading lengths were 331 and 377 bp for the individual and parallel run, respectively. In both the individual and parallel runs, around 50% (46.76% and 43.97%, respectively) of the reads that successfully aligned with the genomic reference were sequences from coding and flanking regions.

Table 4.9: Sequence reads obtained by NGS with the parallel run.

Sample/MID	Nº of high-quality reads aligned to reference per sample	% of total high-quality reads
1	16,132	13.39
2	27,476	22.80
3	20,556	17.06
Chimerical/4	48,820	40.51
Total	112,984	93.76

The coverage analysis of the complete coding and 30 bp flanking sequences contained in the LR-PCR fragments showed 100% of the bases covered with an average total depth >30X (Figure 4.9).

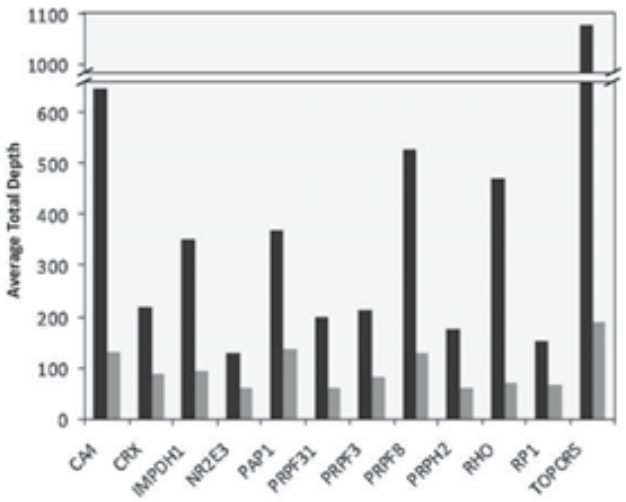


Figure 4.9: Average total depth for each of the 12 analysed genes for one sample/PTP (**Black**) and four samples/PTP (**Grey**).

The parallel sequence analysis of four libraries was also tested. In this assay, nearly 100% of the bases were covered for nearly all twelve study genes (Table 4.10). However, in this run, sample 2 presented a lower coverage (49%) of *NR2E3* gene, suggesting that, in the library preparation of this sample, the DNA fragment containing the *NR2E3* gene was underestimated.

Table 4.10: Parallel NGS of 12 adRP-associated genes of four samples.

Gene	Sample/MID 1		Sample/MID 2		Sample/MID 3		Chimerical/MID 4	
	Average total depth	% bases covered	Average total depth	% bases covered	Average total depth	% bases covered	Average total depth	% bases covered
CA4	50	97	85	100	76	100	316	100
CRX	42	100	98	100	49	100	167	100
IMPDH1	43	100	67	100	60	100	208	100
NR2E3	28	100	21	49	22	100	180	100
RP9	81	100	58	100	100	100	310	100
PRPF31	51	100	32	100	55	100	113	100
PRPF3	52	100	72	100	68	100	140	100
PRPF8	73	100	140	96	90	99	212	98
PRPH2	44	100	52	100	68	100	86	100
RHO	80	100	38	100	70	100	90	100
RP1	29	92	23	96	71	96	147	100
TOPORS	116	100	327	100	110	100	213	100

The analysis of sequence depth in the coding and flanking regions of the different genes showed different values (Figure 4.10). The differences in sequence depth may suggest unknown sequence effects of Fragmentase or that sequencing specificity could be present. However, graphic representation of the depth and coverage showed a similar gene profile for the one-sample and four-sample sequencing runs, demonstrating the reproducibility of the method for preparing libraries for NGS (Figure 4.10).

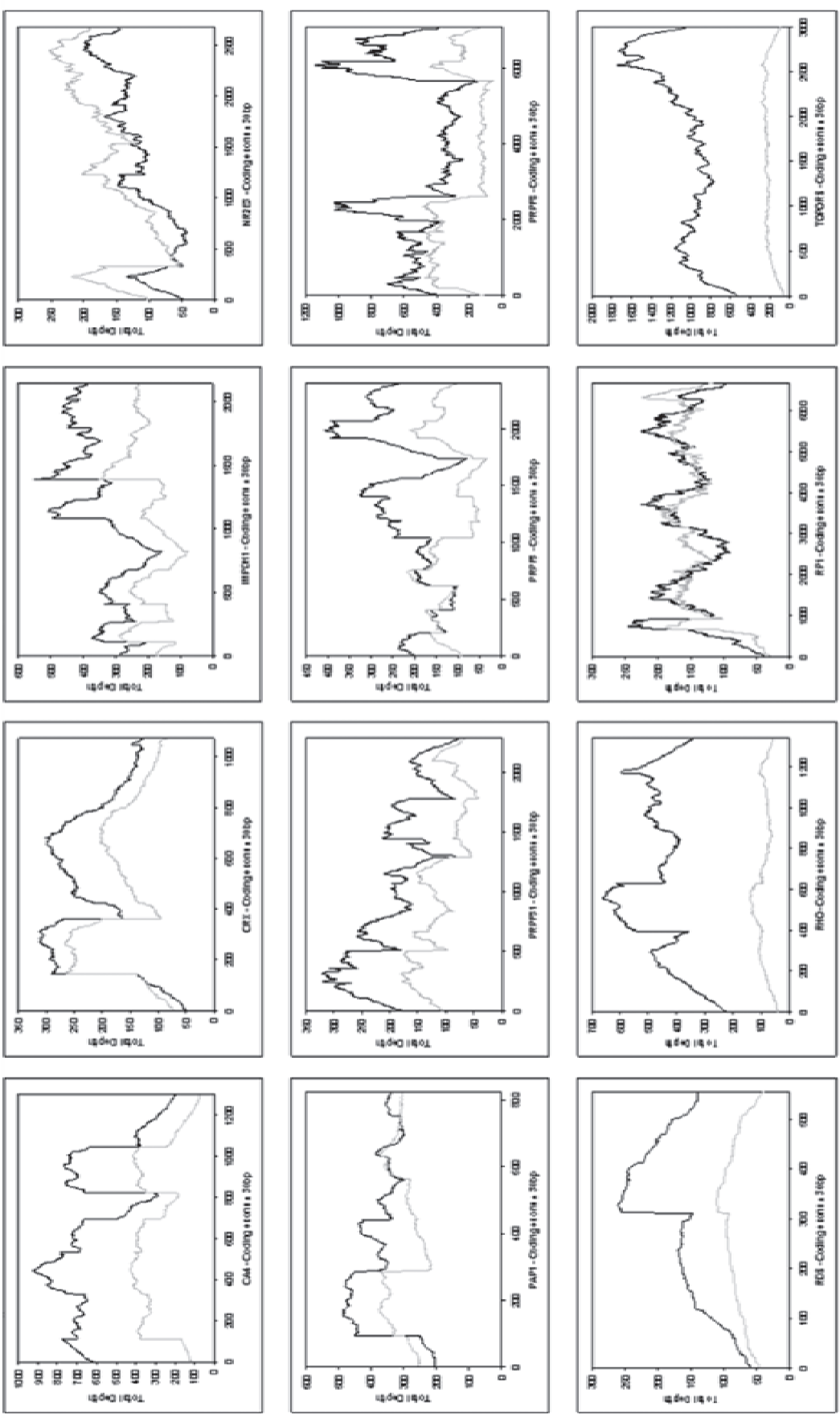


Figure 4.10: Depth profile of sequenced genes. The plot shows the depth coverage profile of the different fragments-amplified coding and intronic flanking regions (30bp) for single (**dark grey**) and multiple (**light grey**) sample runs.

Although inter (Figure 4.9) and intra (Figure 4.10) differences in the sequence depth representation of the 12 genes are shown, the coding sequences of the genes were 100% covered with a sequence depth >30X. Thus, the DNA fragmentation by enzymatic method proved effective. The five previously characterised mutations in *CA4*, *CRX*, *PRPF31*, *PRPF8*, and *PRPH2* presented in the chimeric libraries were also analysed (Table 4.11). The detection capacity of sequence variants was evaluated by analysing five previously characterised adRP-causing mutations, carried in the chimeric libraries. These mutations are heterozygous nucleotide substitutions, a deletion of one nucleotide, or small nucleotide deletions, and were detected with a sequence variation from 36% to 50% with a general sequence depth >48X for the individual run, and with a sequence variation from 34% to 67% with a general sequence depth >27X for the parallel run (Table 4.11). The sequences also revealed the polymorphism p.Arg310Lys in exon 3 of *PRPH2* (Sullivan *et al.*, 1999) in a heterozygous form and two polymorphisms in *RP1*, p.Ala1670Thr and p.Ser1691Pro (Brockman *et al.*, 2008), in a heterozygous and homozygosis form, respectively. These polymorphic variations were validated with Sanger sequencing.

In the four-sample parallel run in the GS Junior platform, the detected variants that could be considered negatives were always in regions containing a poor total sequence depth (less than 20X total depth) or with a total variation below 30%. In the total depth range below 20X, the variant validation as positive or negative may be uncertain because, despite some real heterozygous changes showing values approaching 50% variation, some variants with values around 30% that were validated with Sanger sequencing (Table 4.11) were found. Although the average sequence depth obtained decreased compared with the experiment analysing just one library per PTP, it was enough to detect the control genetic variations present in the chimeric library (Table 4.11). Additionally, sequencing the other three index patients revealed two novel mutations, one in Sample 1 and another in Sample 3. In Sample 3, it was possible to detect a deletion of the three nucleotides AAC at position g.312_314 in *RHO*, which causes the novel mutation p.Asn73del in the DNA sample of an index patient (Table 4.11). Additionally, in sequencing Sample 1, a deletion of the three nucleotides ATC at position g.7092_7094 of *PRPF31* was discovered, which causes the novel p.Ile109del mutation. The deletion of the trinucleotide ATC is located in the sixth nucleotide of exon 4 of *PRPF31*, within a putative splicing signal. A possible change in the splicing signal caused by this mutation was checked using two algorithms for splicing prediction (Splice-Site Finder [SSF] Human Splicing Finder version 2.4.1 and Berkeley Drosophila Genome Project (BDGP) NNSPLICE version 0.9). No change in splicing parameters was obtained.

Table 4.11: Mutation detection in CDS of the Single and Parallel NGS.

		Nº of Samples/PTP									
		1		4							
		Chimerical		Sample/MID 1		Sample/MID 2		Sample/MID 3		Chimerical/MID 4	
Gene	Protein Change	% Variant	Total Depth	% Variant	Total Depth	% Variant	Total Depth	% Variant	Total Depth	% Variant	Total Depth
CA4	p.Val234Ile	44	280	0*	-	0*	-	0*	-	48	222
CRX	None (c.253-15G>A)	50	99	0*	-	0*	-	0*	-	48	67
PRPF31	p.Lys257fsX277	47	76	0*	-	0*	-	0*	-	46	57
PRPF31	p.Ile109del	0*	-	47	25	0*	-	0*	-	0*	-
PRPF8	p.Val2325fsX2329	36	101	0*	-	0*	-	0*	-	34	44
PRPH2	p.Cys214Tyr	40	48	0*	-	0*	-	0*	-	67	27
PRPH2	p.Asp338Gly	0*	-	100	15	29	17	100	16	0*	-
PRPH2	p.Gln304Glu	0*	-	96	27	28	25	100	32	0*	-
PRPH2	p.Arg310Lys	52	31	100	25	92	24	96	28	48	40
RHO	p.Asn73del	0*	-	0*	-	0*	-	38	56	0*	-
RP1	p.Cys2033Tyr	0*	-	55	33	50	40	41	83	0*	-
RP1	p.Asn985Tyr	0*	-	60	20	53	17	46	55	0*	-
RP1	p.Ala1670Thr	43	122	0*	-	0*	-	30	54	71	119
RP1	p.Ser1691Pro	99	183	0*	-	0*	-	49	53	100	122

* In the % Variation column the value zero means no variation.

Both mutations were confirmed with Sanger sequencing and cosegregated in families with adRP (Figure 4.11). These mutations were not detected in other adRP cases or in controls that were screened.

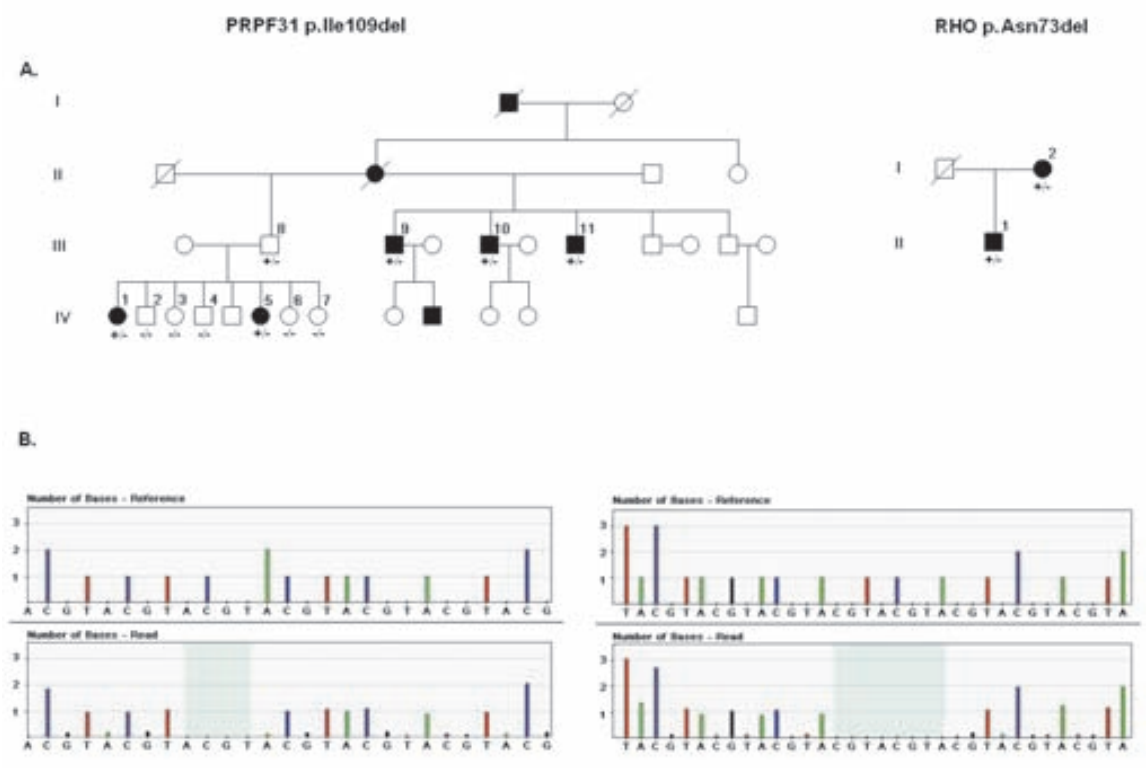


Figure 4.11: Family segregation of *RHO* (on the right) and *PRPF31* mutations (on the left). **(A)** Pedigrees of families carrying mutations in *RHO* (p.Asn73del) and *PRPF31* (p.Ile109del). **(B)** Pyrosequencing chromatogram of the mutations p.Asn73del and p.Ile109del. The top plots are idealised flowgrams for the selected reference sequences and the bottom plots are the aligned flowgrams for the selected reads. Each bar represents the signal intensity for each nucleotide and its height corresponds to the number of nucleotides. The deletion sequence is shown when comparing both flowgrams (shadowed region).

In these samples, four previously reported polymorphisms in *RP1* (Sullivan *et al.*, 1999) and three polymorphic variants in exon 3 of *PRPH2* (Jordan *et al.*, 1992) were also detected (Table 4.11). The sequencing highlights the previous finding that one *PRPH2* allele in a Spanish population contains the three polymorphic variations. All polymorphic variations were confirmed with Sanger sequencing.

The genomic variants found in intronic sequences were also analysed and 23 novel intronic variants were found. These variants were checked with Sanger sequencing and were shown to be positive. In addition, these variants were assayed for change in the splicing signal by using the two splicing prediction algorithms mentioned above. No changes produced by a single nucleotide or by indels in canonical splicing signals were found (Table 4.12).

Table 4.12: New mutation detection in the intronic regions.

	Sample/MID	Gene	Chromosome–Position	Reference Base	Modified Base	SNV	% Variation	Total Depth	Splicing Prediction
Single NGS	Chimenical	*PRPF3*	1-150314981	-	AC	NEW	75.0	88	
			1-150315755	-	TTGGTATACTAATATCTCTGCCTGACAG	NEW	89.5	133	Loss of one potential acceptor site
			1-150315759	-	ACA	NEW	92.9	140	
			1-150320958	-	T	NEW	44.4	36	
		PRPH2	6-42666875	CT	-	NEW	30.4	92	No change
		IMPDH1	7-128037226	-	T	NEW	36.4	220	No change
			7-128042852	G	-	NEW	100.0	196	
		RP1	8-55536890	C	T	NEW	53.6	84	No change
			6-55537000	-	A	NEW	42.5	195	
		NR2E3	15-72108622	-	T	unknown	41.9	116	No change
			15-72109600	C	A	unknown	52.7	93	
		CRX	19-48340889	TTG	A	NEW	44.6	92	No change
			19-48341016	-	AGAT	NEW	53.8	104	
		PRPF31	19-54622189	A	-	NEW	53.1	191	No change
Parallel NGS	1	*RHO*	3-129250305	C	T	NEW	41.4	58	No change
		PRPH2	6-42666506	-	A	NEW	100.0	39	No change
		PRPF8	17-1585810	G	A	NEW	43.2	37	No change
	2	*IMPDH1*	7-128040752	-	T	NEW	98.6	70	No change
			7-128042852	G	-	NEW	100.0	94	
		PRPF8	17-1562327	-	GTT	NEW	36.7	30	No change
		CRX	19-48338242	-	T	NEW	91.2	91	No change
			19-48340889	TTG	A	NEW	40.6	32	
			19-48341016	-	AGAT	NEW	60.0	35	
			19-48341513	-	ATGT	NEW	31.1	45	
	3	*RP9*	7-33140188	A	-	NEW	60.0	25	No change
		PRPF8	17-1558028	-	T	NEW	42.9	28	No change
	Chimenical/4	*PRPF3*	1-150314981	-	AC	NEW	72.2	64	Loss of one potential acceptor site
			1-150315755	-	TTGGTATACTAATATCTCTGCCTGACAG	NEW	96.7	97	
			1-150315759	-	ACA	NEW	98.2	150	
			1-150320958	-	T	NEW	45.1	50	
		PRPH2	6-42666875	CT	-	NEW	41.7	36	No change
		IMPDH1	7-128037226	-	T	NEW	39.0	242	No change
			7-128042852	G	-	NEW	100.0	203	
		RP1	8-55536890	C	T	NEW	50.7	160	No change
			6-55537000	-	A	NEW	41.0	61	
		NR2E3	15-72108622	-	T	unknown	31.4	70	No change
			15-72109600	C	A	unknown	47.3	112	
		CRX	19-48340889	TTG	A	NEW	50.9	57	No change
			19-48341016	-	AGAT	NEW	47.1	51	
		PRPF31	19-54622189	A	-	NEW	51.4	105	No change

4.2.1.2. NGS library construction by target-capture of 23 genes associated with adRP

Here, a method that uses a target DNA capture and NGS was used for mutation detection in 23 candidate genes associated with adRP; the capture and enrichment of coding and flanking sequences of the genes listed in Table 3.16 was carried out as described in the section 3.9.2.2.1 of Material and Methods. The DNA sample belonged to an adRP patient who had been studied by Sanger sequencing and who presented the mutation c.227C>T in the *NRL* gene (Nishiguch *et al.*, 2004) as well as other known SNPs. This point mutation and SNPs serve to evaluate the efficacy of this method.

4.2.1.2.1. Biotinylated cRNA baits design

The ultra-long – 120-mer – biotinylated cRNA baits – to capture regions of interest were designed and submitted as described in the section 3.9.2.2.1 of Material and Methods. The resulting design file, BaitTiling.bed, was loaded in the UCSC genome browser and the resulting baits observed (Figure 4.12).

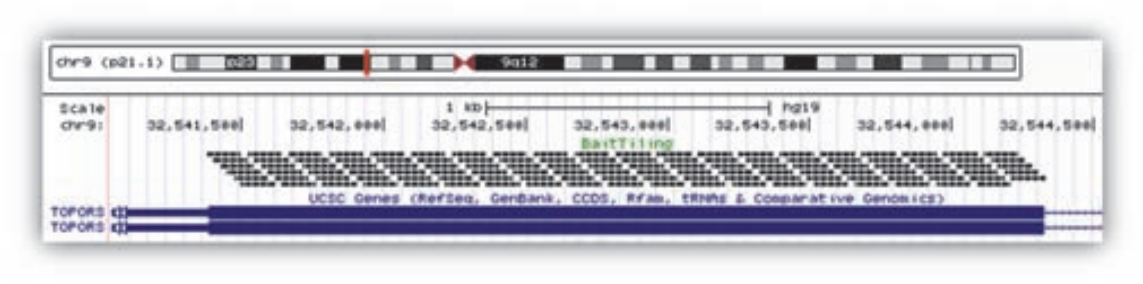


Figure 4.12: BaitTiling graphic representation of the cRNA baits (**Black**) for the TOPORS gene (**Blue**) on the UCSC Genome Browser webpage.

4.2.1.2.2. DNA capture in solution and NGS library quality control

Once the sequences of interest were captured by the cRNA baits and the GS Junior sequencer specific adaptors had been inserted as described in section 3.9.3 of Material and Methods, a quality control step was performed using Experion Automated Electrophoresis System as described in section 3.6.2 of Material and Methods. This step serves to evaluate if the size range of the library (Figure 4.13) is adequate to permit the GS Junior platform to read the sequences, approximately 750 bp.

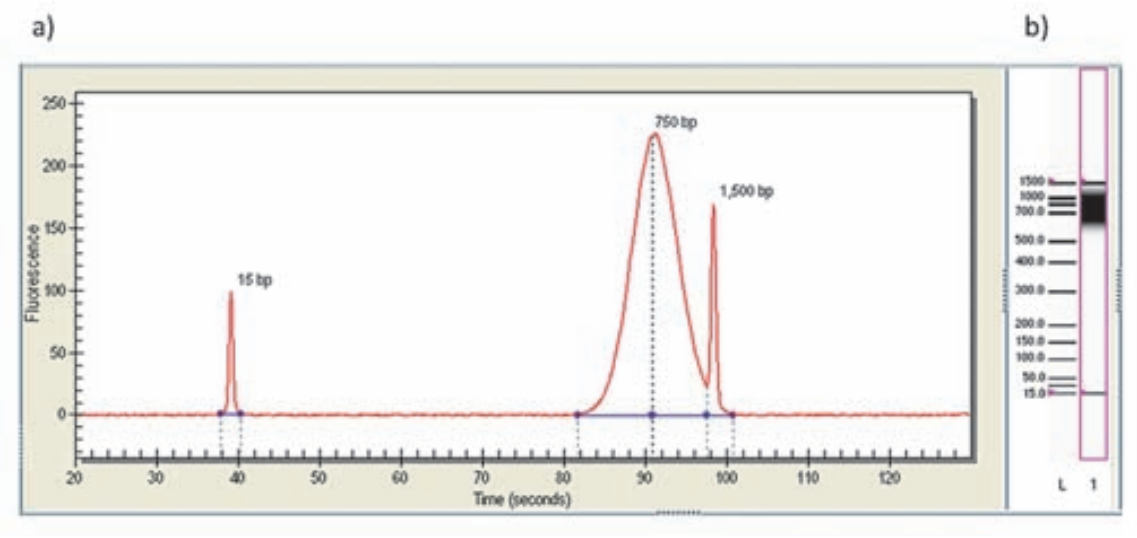


Figure 4.13: Experion Automated Electrophoresis of the NGS library performed with target-capture of DNA in solution. (**a)** Graphic representation of the Lower and Upper marker (15 bp and 1,500 bp, respectively) and the NGS library (peaking on the 750 bp size); (**b)** Gel electrophoresis like representation with "L" standing for molecular weight marker and "1" representing the NGS library.

4.2.1.2.3. Sequencing Data analysis

In this run, one PicoTiter Plate contained just one library, built by selected capture of DNA sequences from 23 genes associated with adRP. The sequenced library generated 96,737 high-quality reads (42.6% of the total raw reads) in the GS Junior platform, in which just 48,030 reads were successfully aligned with the references (Table 4.13). The average read length was 408 base pairs. The resulting data was analysed using GS Reference Mapper software version 2.5p1.

Table 4.13: Sequence reads obtained by NGS with target-capture of 23 genes associated with adRP.

Nº of high-quality reads aligned to reference	% of total high-quality reads
48,030	49.65

Nearly 50% of the captured sequences were outside the target genes (Table 4.13). This sequence background obtained in the target-capture approach may compromise the number of samples that can be run in parallel. Even so, all 23 genes were represented with average depth ranging from 33 to 1106. The average depth obtained in CDS for each of the 23 genes analysed is presented in Table 4.14.

Table 4.14: Average total depth for each of the 23 genes analysed by NGS.

Gene	Average total depth	Gene	Average total depth
CA4	55	*PRPF8*	206
CRX	73	*PRPH2*	124
GUCA1B	71	*RDH12*	111
IMPDH1	1106	*RHO*	77
KLHL7	85	*ROM1*	76
NR2E3	116	*RP1*	421
NRL	33	*RP9*	291
PROM1	76	*RPGR*	39
PRPF3	387	*SEMA4A*	101
PRPF31	69	*SNRNP200*	193
PRPF4	185	*TOPORS*	480
PRPF6	122		

In this sequencing run, the GS Junior platform detected the point mutation c.227C>T in the *NRL* gene, thought to be associated with adRP (Nishiguch *et al.*, 2004) and the SNPs found with the Sanger method (Table 4.15). The homozygous SNPs analysed always provided total variation values >94.4% (mean, 95.8%). The heterozygous SNPs showed values between a minimum of 44.8% and a maximum of 70.0% (mean, 67.1%). The point mutation was detected with a total variation of 40.0%.

Table 4.15: Point Mutation and SNPs identified in CDS of adRP genes.

	Gene	Variant	Protein Change	% Variant	Total Depth
Point Mutation	*NRL*	c.227C>T	p.Ala76Val	40.0	35
	Gene	Variant	db SNP *	% Variant	Total Depth
Single-Nucleotide Polymorphism	*PRPH2*	c.318T>C	7764439	66.0	106
	RP1	c.2953A>T	2293869	94.4	142
	RP1	c.6098G>A	61739567	97.1	140
	SEMA4A	c.1716C>T	12401573	70.0	60
	SNRNP200	c.5317C>T	772175	47.1	136
	SNRNP200	c.3550T>C	3171927	44.8	105
	TOPORS	c.2796T>C	12348918	50.0	94

* db SNP, single-nucleotide polymorphism database (http://www.ncbi.nlm.nih.gov/snp/)

While this approach allowed the enrichment of most gene regions of interest, a large variation in the coverage was observed. Figure 4.14 shows that the percentage of bases covered for most of the genes

analysed was 100%, but some regions of several genes presented a low read count (depth below 20X) or even uncovered regions (*PROM1, PRPF4, RP9* and *RPGR*).

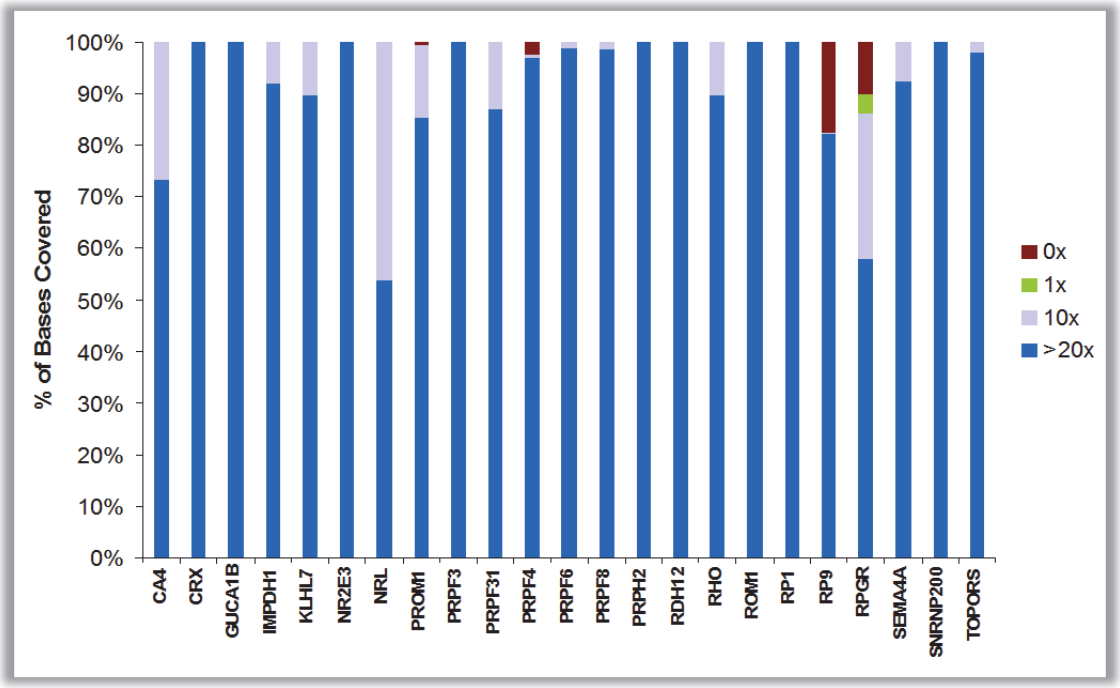


Figure 4.14: Target-capture coverage of 23 adRP candidate genes.

The fraction of targeted positions covered at 0x (**red**), 1x (**green**), 10x (**light blue**), and >20x (**dark blue**) are shown.

4.2.1.3. Parallel NGS library construction by Multiplex-PCR

Following the advances in multiplex-PCR, a mutation detection assay covering a gene panel of 11 candidate genes commonly associated with adRP was set-up through the construction of NGS-libraries containing over forty amplicons. These amplicons were obtained by multiplex-PCR of 6-plex as described in section 3.9.2.3 of the Material and Methods.

4.2.1.3.1. Selected cases for study

In this run, five chimerical samples (Chimerical 1 to 5) were prepared from DNA of several previously molecular diagnosed adRP patients. Each chimerical sample contained three point mutations distributed within its 6-plex (Table 4.16). These point mutations served to help evaluate the efficacy of this approach.

Table 4.16: Previously identified adRP point mutations present in the chimerical samples.

Chimerical Sample	Gene	Primary Mutation
1	*RHO*	c.119C>T
	PRPF31	c.735C>T
	IMPDH1	c.926G>C
2	*RHO*	c.644C>T
	IMPDH1	c.962C>T
	CRX	c.425A>G
3	*RHO*	c.1040C>T
	PRPF8	c.6968_6988del21bp
	RP1	c.2115delA
4	*PRFP31*	c.669_770insA
	PRPF3	c.1466C>A
	RP1	c.2038C>T
5	*RHO*	c.217_219delAAC
	PRPF31	c.328_330delATC
	PRPH2	c.641G>A

4.2.1.3.2. Barcode for Parallel NGS

Because libraries constructed through Multiplex-PCR are based on amplicons instead of fragmented DNA, the MID-adaptors are inserted in the DNA fragments through a second short-cycle PCR, as described in section 3.9.3 of Material and Methods. This is possible by including a short M13-sequence tag at the 5'-end of the specific primers (Figure 4.15). In this study, the MIDs 1 to 5, corresponding to the chimerical samples 1 to 5 (Table 4.16) of the 454Standard MID set (Roche) and compatible with the amplicon library protocol, were used.

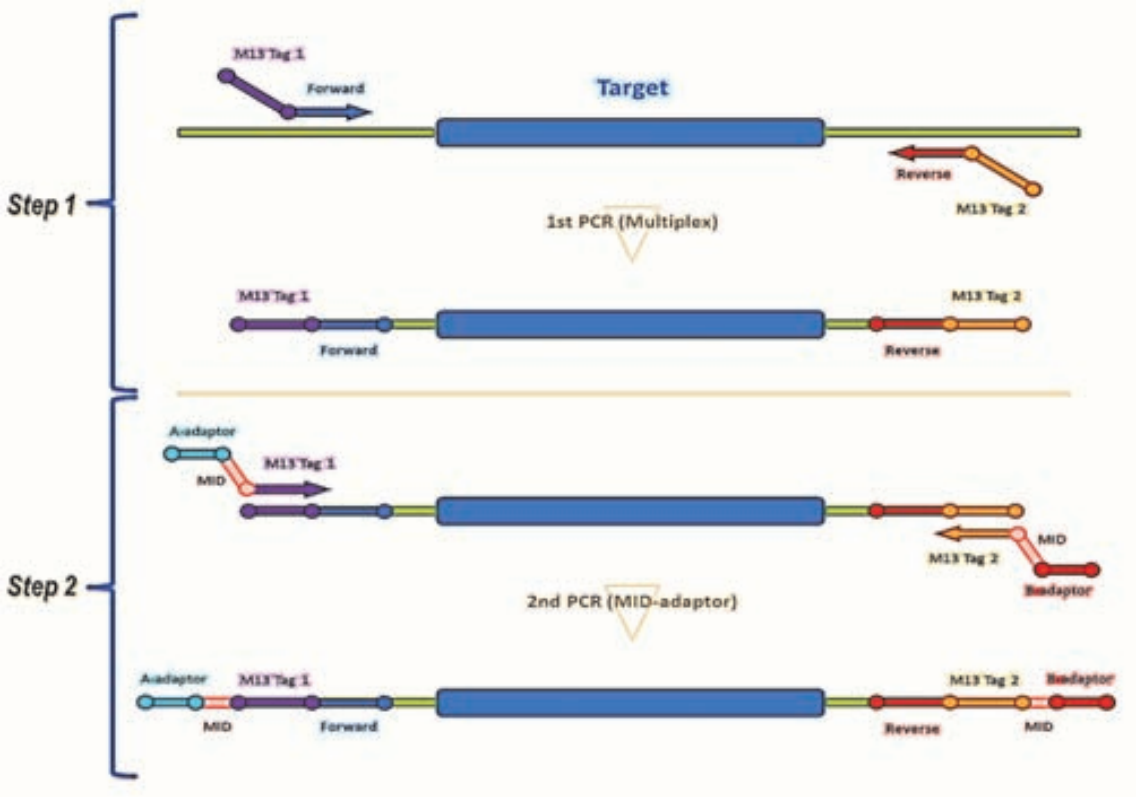


Figure 4.15: Two step amplification process over-view for Multiplex-PCR libraries.

Step 1) Targets are amplified. **Step 2)** MID-adaptors are inserted through a 2nd short cycle PCR.

4.2.1.3.3. Sequencing Data analysis

In this run, one PicoTiter Plate contained five different chimerical samples that were distinguished with MIDs. Variants detected by the GS Junior platform were then compared with those obtained by Sanger sequencing (Table 4.16). The sequenced parallel library generated 77,514 high-quality reads (28.3% of the total raw reads) in the GS Junior platform, in which 77,311 reads were successfully aligned with the reference (Table 4.17).

Table 4.17: Sequence reads obtained by NGS in the parallel run.

Chimerical/MID	Nº of high-quality reads aligned to reference per sample	% of total high-quality reads
1	13,943	17.98
2	15,160	19.55
3	15,447	19.92
4	14,619	18.86
5	18,142	23.40
Total	77,311	99.71

Although the average read length of the total high-quality reads was 327 base pairs, nearly 20% (42,511 raw reads) of total raw reads (the reads not yet qualified as high-quality or non-quality) were discarded for being too short or having too many poor incorporations or interruptions. Figure 4.16 illustrates a graphic representation of the total raw reads size distribution. Visible are a preponderance of reads sized 80, 100, and 150 bp, which could explain the low percentage of high-quality reads obtained (28.3%). Even so, sufficient high-quality reads were obtained to perform an acceptable analysis. (Table 4.18).

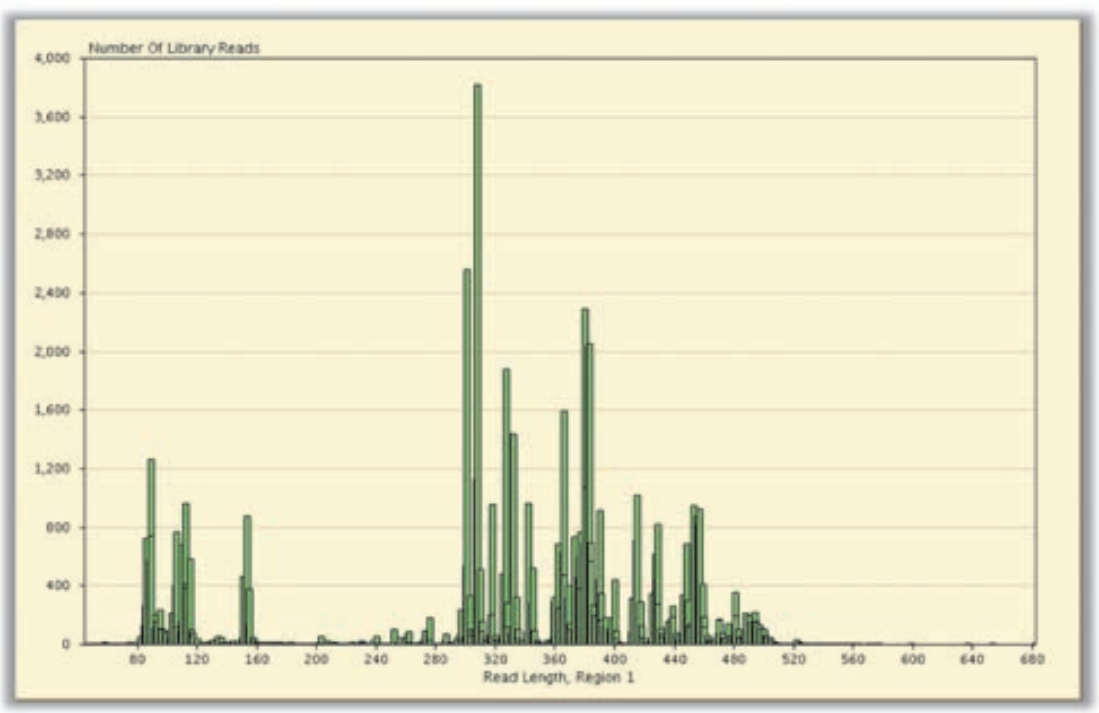


Figure 4.16: Graphic representation of the number of library reads per their read length.

The resulting data generated from the parallel run were analysed using the GS Amplicon Variant Analyser (AVA) software version 2.5p1. The average reads per amplicon obtained for each gene are presented in table 4.18. All amplicons from all genes obtained a high number of reads with the exception of the *RP1* gene, ranging between a minimum of 3 reads to maximum of 27 reads. This could indicate some unfortunate inhibition of the primers for the exon 4 of the *RP1* gene by other primers in the mix of Plex-E (see Table 3.17 of Material and Methods).

The point mutations associated with adRP found with the Sanger method were detected in the five chimerical samples (Table 4.19). These mutations were detected with a total variation between a minimum of 33.9% and a maximum of 60.0% (mean, 48.1%).

Table 4.18: Average number of reads per amplicon for each analysed gene.

Gene	Chimerical/MID 1	Chimerical/MID 2	Chimerical/MID 3	Chimerical/MID 4	Chimerical/MID 5
	Average Reads per Amplicon	Average Reads per Amplicon	Average Reads per Amplicon	Average Reads per Amplicon	Average Reads per Amplicon
RHO	107	153	161	139	188
NR2E3	768	795	852	829	816
PRPF3	634	802	772	715	857
PRPF8	389	334	338	365	403
PRPF31	279	270	293	284	291
PRPH2	250	251	238	234	319
KLHL7	439	248	255	198	343
IMPDH1	243	280	277	261	289
CRX	500	410	474	424	487
NRL	168	322	301	273	413
RP1	27	3	15	13	19
Total	346	352	361	339	402

All point mutations were successfully detected with a read count always above 30X, the only exception being the *RP1* gene-related point mutations for which just 15 reads were counted in the case of chimerical sample 3 and 13 reads in the case of chimerical sample 4. Even so, the primary mutations c.2115delA and c.2038C>T were effectively detected with a total variation of 40.0% and 53.9% in the chimerical samples 3 and 4, respectively.

Table 4.19: Point mutation detection in CDS for each of the samples in the multiplex NGS run.

Chimerical/MID	Gene	Protein Change	% Variant	Reads per Amplicon
1	*RHO*	p.Leu40Pro	60.0	30
	PRPF31	None (c.735C>T)	50.7	301
	IMPDH1	p.Arg309Pro	42.3	167
2	*RHO*	p.Pro215Leu	33.9	149
	IMPDH1	p.Ala321Val	56.4	126
	CRX	p.Try141Cys	49.6	140
3	*RHO*	p.Pro347Leu	51.2	84
	PRPF8	p.Val2325fsX2329	48.7	394
	RP1	p.Lys705fsX712	40.0	15
4	*PRFP31*	p.Lys257fsX277	35.0	314
	PRPF3	p.Ala489Asp	45.9	715
	RP1	p.Arg677X	53.9	13
5	*RHO*	p.Asn73del	58.1	74
	PRPF31	p.Ile109del	45.4	326
	PRPH2	p.Cys214Tyr	50.5	309

4.2.1.4. Bead-Linker-Primer complex combined with emPCR for multiplex NGS library generation

Here, the methodology of primer isolation by a particular carrier was employed. This carrier consists of an oligonucleotide functionalised with biotin at the 3'-end (Linker) which in turn is attached to a streptavidin-functionalized magnetic bead. This oligonucleotide possesses a complementary sequence that captures the specific primers and isolates them from the environment forming the Bead-Linker-Primer complex or BLP (see section 3.9.2.4 of Materials and Methods). Because the Linker has its 3'-end functionalised with biotin it will not further interfere with the PCR reaction and while its detachment from the magnetic bead is virtually impossible under the normal PCR conditions, the hybridisation with the primer is easily broken during the denaturation step of the PCR. This technique aimed for multiplexing by combining it with emPCR in order to prepare libraries where the number of amplicons produced in a single tube is virtually unlimited.

For starters and to check if this novel multiplexing technique worked, a small-in-scale multiplex experiment was performed. Briefly, nine distinct BLP, correspondent to the fragments of the Plex-E (Table 3.17 of Material and Methods) were mixed and used for emPCR followed by emulsion disruption and amplicon extraction, as described in section 3.5.3 of the Material and Methods. A small amount of the resultant amplicon mix was then diluted 1:2000 in TE 1X, and 1 µl of the diluted sample was used for a short cycle PCR using the nine correspondent individual primers (Table 3.17 of the Material and Methods), separately. The results showed that all nine amplicons were successfully amplified (Figure 4.17).

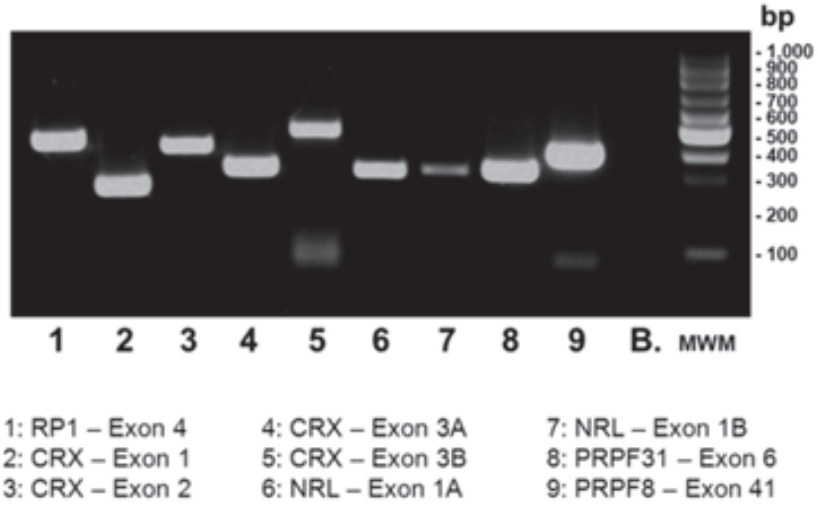


Figure 4.17: Agarose gel electrophoresis,1%, of the nine fragment multiplex done by BLP combined with emPCR; Control blank (B.); MWM, molecular weight marker.

4.2.1.4.1. BLP combined with emPCR base NGS library construction

In view of the previous small-in-scale multiplex experiment, a mix of BLPs (in a total of 45 distinct BLPs) was used for emPCR followed by emulsion disruption and amplicon extraction, for a complete NGS library. After, the sequencer-specific-adaptors were inserted as described in section 3.9.3 of Material and Methods. The resulting library was purified two times with AMPure®, in order to avoid small fragments as the ones visible in Figure 4.16, and analysed by agarose gel electrophoresis for quality control (Figure 4.18).

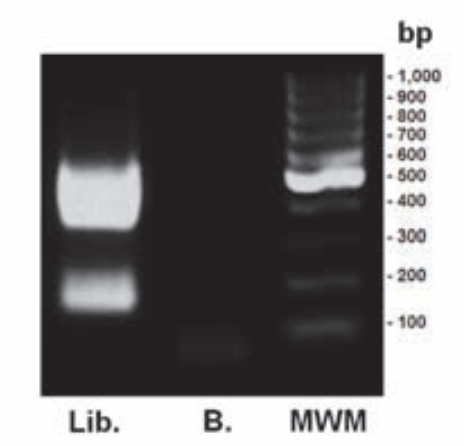


Figure 4.18: Agarose gel electrophoresis,1%, of the Multiplex Library done by BLP combined with emPCR (Lib.); Control blank (B.); MWM, molecular weight marker.

The agarose gel electrophoresis presented a large lane averaging 150 bp. This lane was not removed even after the double AMPure purification because it was larger than 100 bp. Additionally, a purification directly from the agarose gel could be done in order to remove this unwanted band but this would make the library unviable for the GS Junior platform due to an inhibition caused by the agarose gel purification procedure on the library clonal step (unreported data). All the same, the library was loaded into the GS Junior platform for sequencing as described in section 3.9.3 of the Material and Methods, though a negative interference of this large lane, averaging 150 bp, was expected.

4.2.1.4.2. Sequencing Data analysis

In this run, one PicoTiter Plate contained just one sample of an adRP characterized patient. The sequenced library generated just 14,611 high-quality reads (5.9% of the total raw reads), which is very low for a GS Junior platform single run; 8,998 of these high-quality reads were successfully aligned with the reference (61.6%). The average read length of the total high-quality reads was also lower, 155 bases pairs. These poor values are explained by nearly 50% (141,983 raw reads)

of the total raw reads (the reads not yet qualified as high-quality or non-quality) having been discarded as too short or having too many poor incorporations or interruptions.

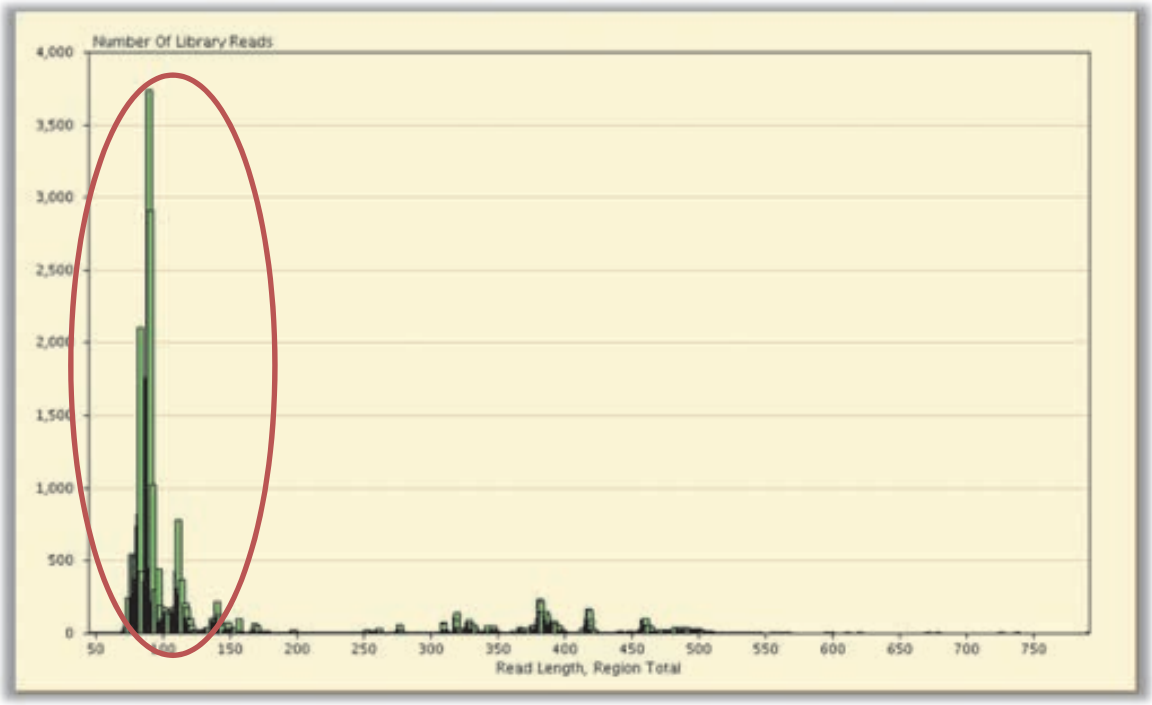


Figure 4.19: Graphic representation of the number of library reads per their read length. Marked in **red** colour are those reads considered to be short-quality reads.

Figure 4.19 show a graphic representation of the total raw reads size distribution. Comparable with the earlier observed multiplex-PCR approach results, it is visible an even greater amount of reads sizing 80, 100, and 150 bp (marked in colour in figure 4.19), which could explain the lower percentage of high-quality reads obtained (5.9%). Predictably, the analysis performed by the AVA software of these high-quality reads were not enough to obtain a good read per amplicon count for all amplicons. As a result, a total of 21 amplicons out of a library of 44 amplicons were analysed. Their reads per amplicon are presented in table 4.20.

Table 4.20: Number of reads per amplicon for BLP library.

Gene	Exon	Average Reads per Amplicon
CRX	1	126
IMPDH1	6	4
	7	28
	10	15
	11	14
	12	38
	13	91
	14	18
	15	35
	16	38
PRPF31	1-2	62
	3	14
	4	96
PRPF3	10	441
PRPF8	42	10
PRPH2	1A	8
	1B	27
	2	48
	3	94
NR2E3	2	211

Even though only 21 amplicons were represented, analysing these sequences with AVA software resulted in the detection of two SNPs in *PRPH2*: the p.Arg310Lys (rs425876) and the p.Asp338Gly (rs434102). No further variants were detected in the run.

To further investigate if indeed all amplicons were/were not present in the library or if the reason for such low read count was due to some other aspect of the sequence run protocol, a small trial was performed. Briefly and as previously described in this section, a small amount of the double AMPured library was diluted 1:2000 in TE 1X, and 1 µl of the diluted sample was used for a short cycle PCR using the individual primers (Table 3.17 of Material and Methods), separately. The results showed that 35 out of all 44 amplicons (Figure 4.20) were amplified in one single reaction using this novel multiplex technique (nearly 80% success); oddly just 21 of these 35 amplicons (60%) managed to reach the high-quality threshold (Table 4.20). It is also evident that their representation in the agarose gel is not linear; some fragments clearly present distinct concentration.

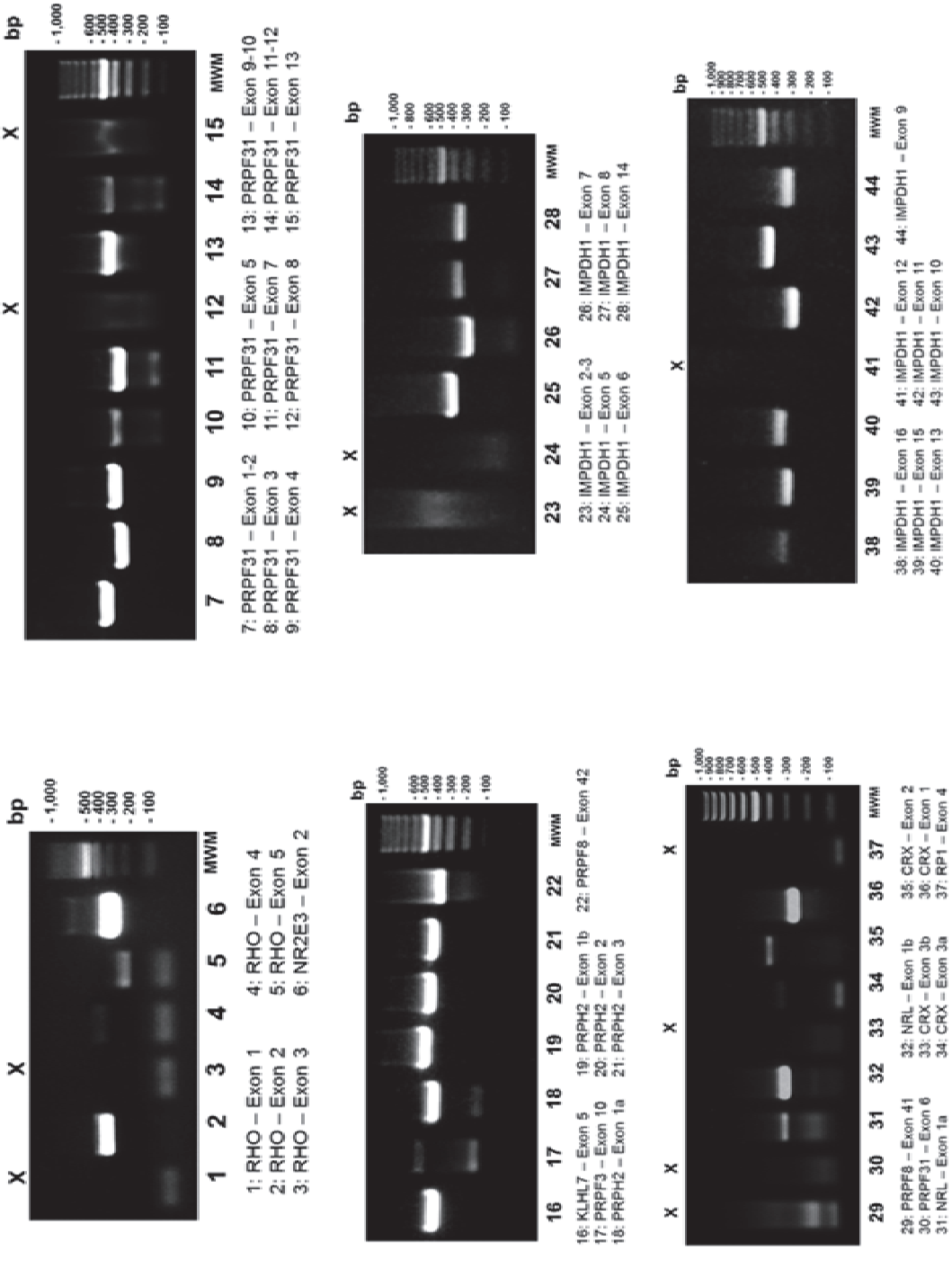


Figure 4.20: Agarose gel electrophoresis, 1%, of the Multiplex Library done with BLP using emPCR; individual fragments successfully amplified (1 - 44); the "**X**" marks failed amplification; MWM, molecular weight marker.

4.2.2. Detection of New Candidate Genes for adRP by Massive Sequencing

- Context

Approximately twenty genes are known to be associated with adRP (Sullivan *et al.*, 2006; Bowne *et al.*, 2011). Nevertheless, around 60% of adRP cases in the Spanish population remain molecularly-undiagnosed. With this in mind, the second part of this section aspires to the discovery of new gene defects associated with adRP. Hence, the approach of microarrays for DNA sequence capture and NGS for detecting pathogenic variants in 448 candidate genes in five index cases of adRP was explored.

During the 448 candidate genes analysis performed for 5 families, the costs for whole exome analysis had dropped significantly and it reached a point where it had become affordable. One of these five families was then chosen for whole exome analysis.

4.2.2.1. Candidate gene screening approach

Five index cases of Spanish adRP families who had previously been studied and clinically diagnosed with adRP were selected for this experiment. These cases and their family trees are represented in the Figures 4.23 to 4.26.

4.2.2.1.1. Selection of candidate genes

A sequence-capture experiment with a 385K array which has a sequence capacity of 5 Mb was designed. The array capacity allowed the inclusion of only a limited number of genes. Consequently, first a number of candidate genes that are possibly involved in adRP were selected. The genes selected for DNA capture (Appendix A.1) included autosomal genes that would be relevant in terms of specific expression or function in the retina. An initial set of candidate genes was retrieved from the NCBI Gene database (http://www.ncbi.nlm.nih.gov/gene/) by text search of adRP-related terms (retina, retinitis, eye, cone, rod, photoreceptor, macula, macular, ocular, blindness, visual perception, visual system, visual cycle, retbindin, Rho, Usher, opsin, and fovea) in the gene/protein name field and the summary text. In order to find novel disease-associated eye genes, those genes previously reported in the database eye disease genes (NEIBank website [http://neibank.nei.nih.gov/cgi-bin/eyeDiseaseGenes.cgi/]) associated with retinal dystrophies were also included in the candidate gene list. Most of the genes related with syndromic forms of RP are unlikely to be involved in the cause of disease in the index patients or affected relatives as these patients had no other apparent pathology apart from

dominant RP. Moreover, none of the known variants found in genes associated with syndromic RP have yet been associated with dominant RP. Consequently, although mutations in genes may show different phenotypes, most genes associated with syndromic RP in the candidate list of adRP genes were not included. The candidate genes list included 15 genes associated with adRP even if they had been partially or fully screened in the patients studied. Also included in this list were those genes that were previously associated with arRP for the reason that some mutations in these genes may be inherited in a dominant mode, as in the case of *NR2E3* (Bernal *et al.*, 2008; Coppieters *et al.*, 2007). Genes with major or specific expression in the retina included those showing a restricted eye pattern in the NCBI Unigene database (http://www.ncbi.nlm.nih.gov/unigene/) or being predominant in the NEI retina-specific library (http://neibank.nei.nih.gov/cgibin/showDataTablecgi?lib=NbLib0042/); and predicted targets of retina-specific transcription factors (Coppieters *et al.*, 2007) *CRX* (Freund *et al.*, 1997; Chen *et al.*, 2002), *NRL* (Bessant *et al.*, 1999) and *NR2E3* (Bowes *et al.*, 1989; Cheng *et al.*, 2004) Mutations in some genes that are involved in pre-mRNA splicing processing, although expressed ubiquitously, can cause adRP. Candidate genes involved in pre-mRNA splicing include those encoding the U4/U6; U5 tri-snRNP and associated Sm/LSm proteins (Liu *et al.*, 2006), and functional partners predicted by string (http://string-db.org/) of adRP-related *PRPF8*, *PRPF31*, *PRPF3*, *RP9* and *ASCC3L1*. In addition, genes that showed differential expression (Gamundi *et al.*, 2008) or splicing in previous arrays comparing wild-type and *PRPF8* mutant lymphocytes from RP patients were also included. Finally, genes associated with autosomal dominant macular dystrophy *PRPH2* (peripherin/*RDS*) (Farrar *et al.*, 1991; Kajiwara *et al.*, 1991; Kimura *et al.*, 1993; Wells *et al.*, 1993; Gamundi *et al.*, 2007) or cone + cone rod dystrophies like *GUCY2D*, *GUCA1*, *PDE6H,* or *RIM1* (Payne *et al.*, 1998; Gregory-Evans *et al.*, 2000; Johnson *et al.*, 2003; Piri *et al.*, 2005) were also included in the collection (Appendix A.1).

4.2.2.1.2. Library Construction

An average yield of 15 µg of captured DNA per sample was obtained. The average capture enrichment for quality control loci calculated using the NimbleGen internal controls run ranged from 229- to 1053-fold. These target-enriched samples were used for construction of the sequencing library (Table 4.21).

Table 4.21: Enrichment and average size of genomic DNA capture samples used in library construction for NGS.

	Sample Data		Library Data	
Sample	Amount of DNA (µg)	Average fold enrichment measured at QC loci	Average Size (bp)	Concentration (ng/µl)
RP-93	8.56	229	271	22.10
RP-95	9.18	297	264	14.38
RP-645	8.18	1053	270	9.93
RP-83	9.26	280	263	21.48
RP-65	8.14	356	273	21.65

4.2.2.1.3. Sequencing data and sequence variants

The specificity of the DNA capture was measured by the percentage of 'on target' or 'near target' mapped sequences. Thus, between 27% and 34% of total mapped sequences were on target while the near target specificity was 43 - 53% (considering a 100-bp interval on both sides) (Table 4.22).

Table 4.22: Sequence reads obtained by NGS of genomic DNA capture samples from the five adRP index cases.

Sample	Generated readings	Filtered readings	% Filtered readings	Mappable readings	% Mappable readings (generated)	% near target	% on target	% captured 20X
RP-93	59,725,546	47,145,390	78.94	31,433,758	52.63	43.88	27.80	94.64
RP-95	73,146,438	58,859,907	80.47	36,649,060	50.10	53.25	33.56	94.72
RP-645	38,064,457	31,820,656	83.60	21,679,940	56.96	50.20	32.17	93.29
RP-83	60,591,570	51,863,241	85.59	34,496,369	56.93	50.17	32.25	94.78
RP-65	39,151,911	32,138,326	82.09	23,910,338	61.07	47.13	30.69	93.51

The sensitivity of the DNA capture was measured by the percentage of bases contained in the array that were covered in the sequencing. The average capture percentage in all samples was 98 - 99% and 93 - 95% assuming 1X and 20X coverage, respectively (Figure 4.21).

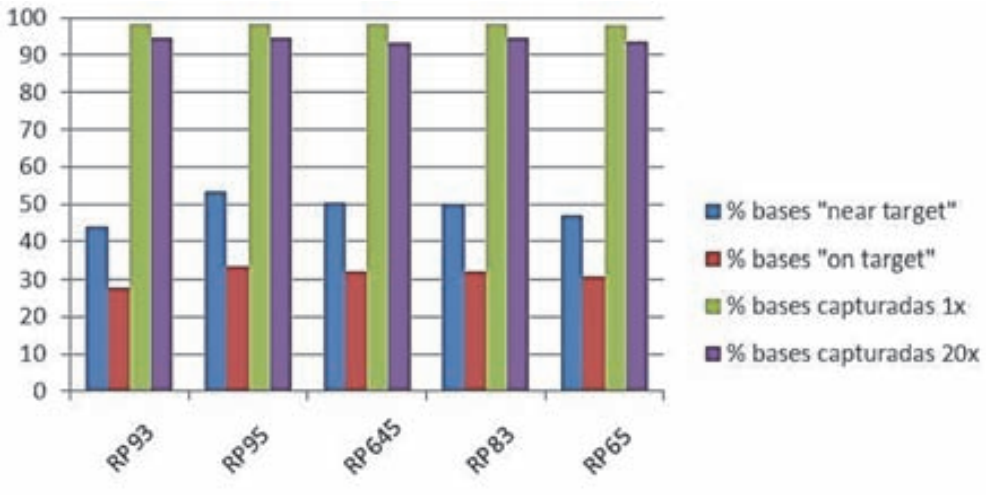


Figure 4.21: Graphic representation of the percentage of bases detected near and on target genes for each sample.

4.2.2.1.4. Sequencing Variants and Sanger Validation

The sequence variants detected were associated with their position in the gene transcript, annotated, and classified according to the nomenclature of the Ensembl database. Variants were classified into two groups: SNVs and Indels. The sequence variants (SNVs and small Indels) detected in all samples were compared with the variants reported in the database. A total of 4730 SNVs were obtained, 200 of which were unreported in the database. As for Indels, a total of 222 were found, being 124 of those unreported (Table 4.23).

Table 4.23: Sequence variants detected by NGS in samples from five adRP index cases.

Sample	Total SNVs	Non-described SNVs	Total indels	Non-described indels	Candidate genes (dominant model)	Sanger validated and cosegregating
RP-93	1028	51	71	31	11	11
RP-95	907	41	63	29	16	15
RP-645	932	38	67	20	9	9
RP-83	997	39	57	20	14	9
RP-65	866	31	64	24	6	6
Sum	4730	200	322	124	56	50

Figure 4.22 shows the workflow of the analysis and validation of the genomic variants found. The Ensembl v59 database was used to analyse the sequence variants found according to the functional consequence in the target transcript. As a candidate variant for causing adRP, the unreported SNVs that generate a non-synonymous change or affect a splicing site were selected first; in the same manner the Indels found in the coding or splicing regions of the targeted genes were also selected (Figure 4.22). All of the novel variants found in the analysis of the five index

adRP patients were annotated. A further analysis showed that 18 of these sequence variants (SNVs and indels) were previously annotated in a database but with a frequency <0.01 (Table 4.24), and are therefore considered relevant as well.

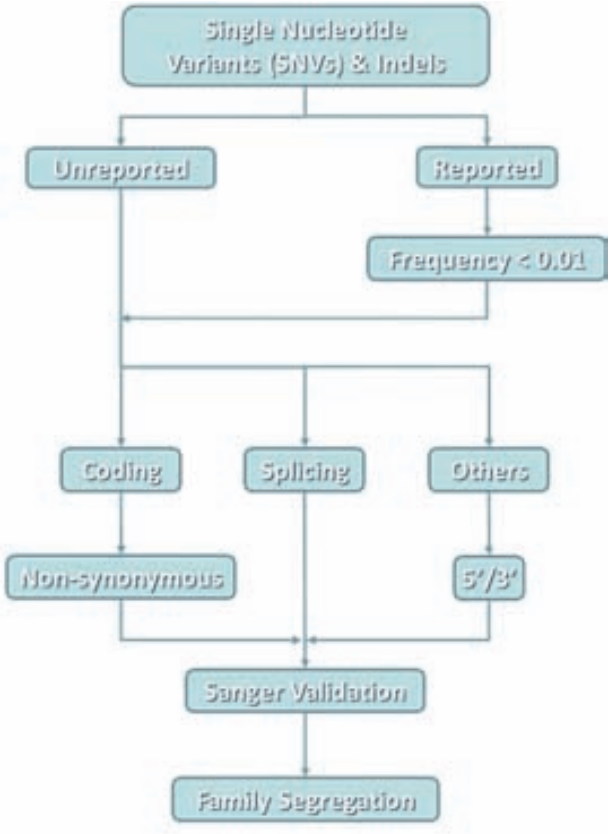


Figure 4.22: Workflow of the analysis of detected sequence variants.

To validate the 56 novel sequence variants found, direct Sanger sequencing of these variants was carried out. Six (10.7%) of these sequence variants were not confirmed by Sanger sequencing and were considered to be false positives. Moreover, 22 sequence variants that had been characterised in a previous survey by Sanger sequencing in some adRP genes of the index patient samples were all detected by NGS. These data point out the high specificity and sensitivity of this approach. The sequence variants found in each sample that were previously annotated in the Ensembl database were also analysed. Sequence variants were filtered and those causing a non-synonymous, splicing, or premature stop change for SNVs, and frame or in-frame change for Indels were selected. Those variants that were reported in the database with a frequency of 0.01 or lower (rare variants) were annotated, but none of them proved to be disease causing according to family segregation. Other sequence variants, reported in the database with a frequency >0.01, proved to be SNPs unrelated to adRP.

Table 4.24: Novel and rare variants (<0.01 frequency) detected in five adRP index cases.

Sample	Frequency	Gene	Nucleotide Variant	Protein Variant	SIFT*	MutPredΩ	PolyPhen
RP-93	novel	*NES*	c.1991T>A	p.Leu664*	n/a†	n/a†	Damaging
	novel	*C2orf71*#	c.2246A>C	p.Asp749Ala	0.17	0.091	Probably damaging
	<0.01	*EHBP1*	c.1210C>T	p.Pro404Ser	0.04	0.179	Probably damaging
	novel	*IMPG2*#	c.1300C>T	p.Pro434Ser	0.05	0.172	Probably damaging
	novel	*PROM1*#	c.55T>A	p.Ser19Ala	0.25	0.428	Benign
	novel	*GRK7*	c.338G>C	p.Cys113Ser	0.16	0.662	Probably damaging
	novel	*IMPDH1*	c.809T>G	p.Leu270Arg	0.20	0.761	Probably damaging
	novel	*MYO3A*	c.914T>C	p.Ile305Thr	0.00	0.432	Possibly damaging
	novel	*SPTBN5*	c.6751C>A	p.Leu2251Met	0.00	0.709	Benign
	novel	*WSB1*	c.331T>G	p.Trp111Gly	0.00	0.833	Probably damaging
	<0.01	*HSD17B14*	c.46G>T	p.Gly16Trp	0.00	0.814	Probably damaging
RP-95	novel	*SLC1A7*	c.355G>C	p.Ala119Pro	0.96	0.453	Benign
	novel	*C16orf92*	c.375C>T	p.His125His (splicing)	1.00	n/a†	n/a†
	n/a†	*C2orf71*	c.C9201_G9202insAGC	p.Ser1225_Glu1226insS	n/a†	n/a†	Damaging
	<0.01	*MYRIP*	c.971C>T	p.Pro324Leu	0.28	0.097	Benign
	novel	*PPEF2*	c.1670C>T	p.Ser557Leu	0.07	0.426	Possibly damaging
	novel	*FAM46A*	c.74G>A	p.Gly25Asp	0.10	0.433	Benign
	novel	*TULP4*	c.2425C>A	p.Pro809Thr	0.01	0.261	Possibly damaging
	<0.01	*ADAM9*#	c.280G>A	p.Val94Ile	0.49	0.608	Benign
	novel	*PDZD7*#	c.971G>A	p.Ser324Asn	0.07	0.199	Benign
	novel	*SF1*	c.898C>T	p.Pro300Ser	0.26	0.284	Benign
	novel	*SPTBN5*	c.7762T>C	p.Met2588Thr	0.45	0.576	Benign
	novel	*PDE6G*#	c.-14delC	-	n/a†	n/a†	Damaging
	novel	*GLTSCR2*	c.829C>T	p.Arg227Cys	0.03	0.528	Probably damaging
	novel	*CABP5*	c.122A>G	p.Asp41Gly	0.00	0.757	Probably damaging
	novel	*ZNF295*	c.2653_2655delGAG	p.Glu885del	n/a†	n/a†	Possibly damaging
	novel	*GRK7*ß	c.338G>C	p.Cys113Ser	0.16	0.662	Probably damaging
RP-645	novel	*SPTBN5*	c.5647C>T	p.Arg1883*	n/a†	n/a†	Damaging
	<0.01	*GLTSCR2*	c.59C>T	p.Ser20Phe	0.00	0.172	Possibly damaging
	novel	*CDH4*	c.206A>G	p.Gln69Arg	0.36	0.574	Benign
	n/a†	*SEPT8*	c.905G>A	p.Arg302His	0.12	0.507	Possibly damaging
	novel	*FUT8*	c.566G>A	p.Arg318Gln	0.44	0.413	Possibly damaging
	<0.01	*MAP4K3*	c.1228G>A	p.Ala410Thr	0.56	0.343	Benign
	novel	*ZNF764*	c.1072G>A	p.Val358M	0.04	0.488	Benign
	novel	*NRL*	c.287T>C	p.M96Thr	0.08	0.574	Benign
	<0.01	*PPEF2*	c.1441A>C	p.M481Leu	1.00	0.257	Benign
RP-83	<0.01	*INADL*	c.1007C>A	p.Pro336His	0.03	0.325	Probably damaging
	<0.01	*POU4F2*	c.417C>A	p.Asp139Glu	0.37	0.168	Possibly damaging
	novel	*CEP120*	c.2659G>A	p.Ala887Thr	0.44	0.106	Possibly damaging
	novel	*KCNV2*#	c.133G>A	p.Gly45Ser	0.63	0.228	Benign
	novel	*CACNA2D4*#	c.1646T>C	p.Leu549P	0.00	0.895	Probably damaging
	n/a†	*SARM1*	c.893C>A	p.Pro298Gln	0.09	0.657	Probably damaging
	<0.01	*PPEF2*	c.503C>T	p.Thr168Ile	0.07	0.416	Possibly damaging
			c.1441A>C	p.Met481Leu	1.00	0.257	Benign
	n/a†	*C2orf71*	c.C9201_G9202insAGC	p.Ser1225_Glu1226insS	n/a†	n/a†	Damaging
	novel	*PITPNC1*ß	c.28A>G	p.Ile10Val	0.09	0.641	Benign
	novel	*NPVF*ß	c.568G>A	p.Ala190Thr	0.10	0.106	Benign
	novel	*ABCC5*ß	c.1043T>C	p.Val348Ala	0.57	0.489	Benign
	novel	*ZNF764*ß	c.1014A>G	p.His305Arg	0.00	0.880	Probably damaging
	novel	*OPN4*ß	c.1209G>A	p.Thr403Ala	0.07	0.167	Probably damaging
RP-65	novel	*FRMPD1*	c.1729G>C	p.Gly577Arg	0.42	0.291	Possibly damaging
	novel	*GBP7*	c.349G>C	p.Ala117Pro	0.00	0.876	Possibly damaging
	<0.01	*TEAD4*	c.210C>G	p.Asp70Glu	0.23	0.887	Benign
	0,008	*NRL*	c.C9201_G9202insAGC	p.Ala76Val	0.06	0.612	Benign
	novel	*BEST4*	c.151C>T	p.Arg51Trp	0.03	0.945	Probably damaging
	<0.01	*SLC1A7*	c.190A>G	p.M64Val	0.09	0.488	Benign

* Considered "*Damaging*" if ≤0.05 and "*Tolerant*" if >0.05
Ω Probability of deleterious mutation
† n/a is not available
# Genes previously associated with recessive retinal dystrophies
ß Variant not confirmed by Sanger Sequencing

4.2.2.1.5. Family segregation of the selected genetic variants

The 50 confirmed variants were evaluated for their pathogenic potential by three prediction tools (Table 4.23). Some of these variants were predicted to cause adRP. Cosegregation of these 50 variants in probands' families were examined. A variant was considered as not directly causing adRP if it was absent in an RP patient in the family. Using this criterion, novel mutations in the *NRL*, *IMPDH1,* and *PDE6G* genes were each found in a family (Figures 4.23 to 4.26). The other variants did not segregate with RP in the families.

- *Family RP-65*

The p.Ala76Val mutation in *NRL* was detected in the index case of family 65 (Figure 4.23A). This variant, although not annotated in the database, had been previously reported as a mutation of uncertain cause of RP (Nishiguchi *et al.*, 2004). However, in family 65, one RP patient (IV-2, Figure 4.23A) does not carry the p.Ala76Val mutation while unaffected members of the family proved to be carriers of the mutation. Consequently, p.Ala76Val substitution is unlikely to be the cause of adRP in this particular family.

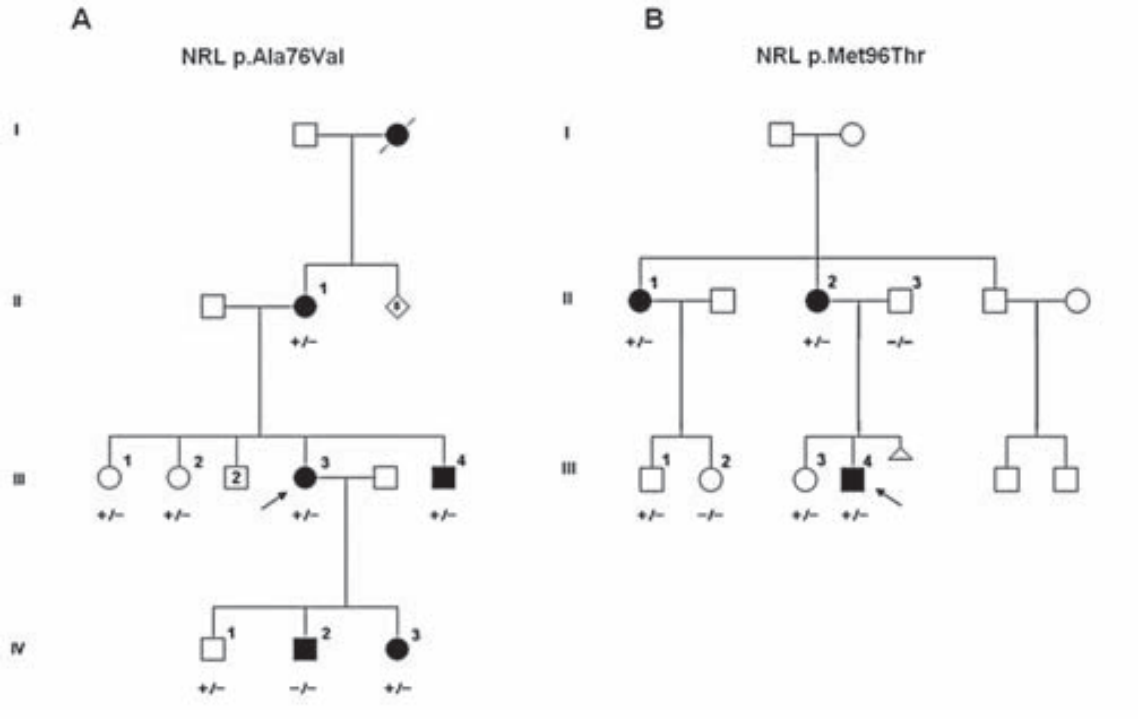


Figure 4.23: Family segregation of *NRL* sequence variants. (**A)** Family 65 shows an affected member (IV-2) who does not carry the p.Ala76Val mutation (-/-) while the unaffected members III-1 and III-2 are heterozygous carriers of this mutation.(+/-). (**B)** In family 645 we detected the heterozygous p.M96T mutation in *NRL* in all affected members (+/-) but also in the unaffected members III-1 and III-3. The arrow indicates the index patient.

- *Family RP-645*

This family was included in this study because an adRP mutation survey in a Spanish population detected the novel variant p.M96T in *NRL* in RP patients of this family. However, segregation of this variant showed two carriers of the mutation (Figure 4.23B) who remain asymptomatic thus far (Hernan et al, 2011), suggesting an incomplete penetrance for an NRL mutation. Thus, an alternative disease-causing mutation was investigated for this family. Still, none of the detected sequence variants (Table 4.23) were potentially adRP-causing in this family.

- *Family RP-93*

Family 93 had already been screened for mutations by DGGE (including *IMPDH1*) in a previous adRP survey without detecting any disease-causing mutation. However, in this study a novel variant, p.Leu270Arg in *IMPDH1* was detected and later confirmed by Sanger sequencing (Figure 4.24B). The variant is carried by all eight available patients and is absent in the unaffected members analysed (Figure 4.24A). This mutation was not detected in 100 adRP index patients or in 150 controls screened by real-time PCR using FRET probes (Figure 4.24C).

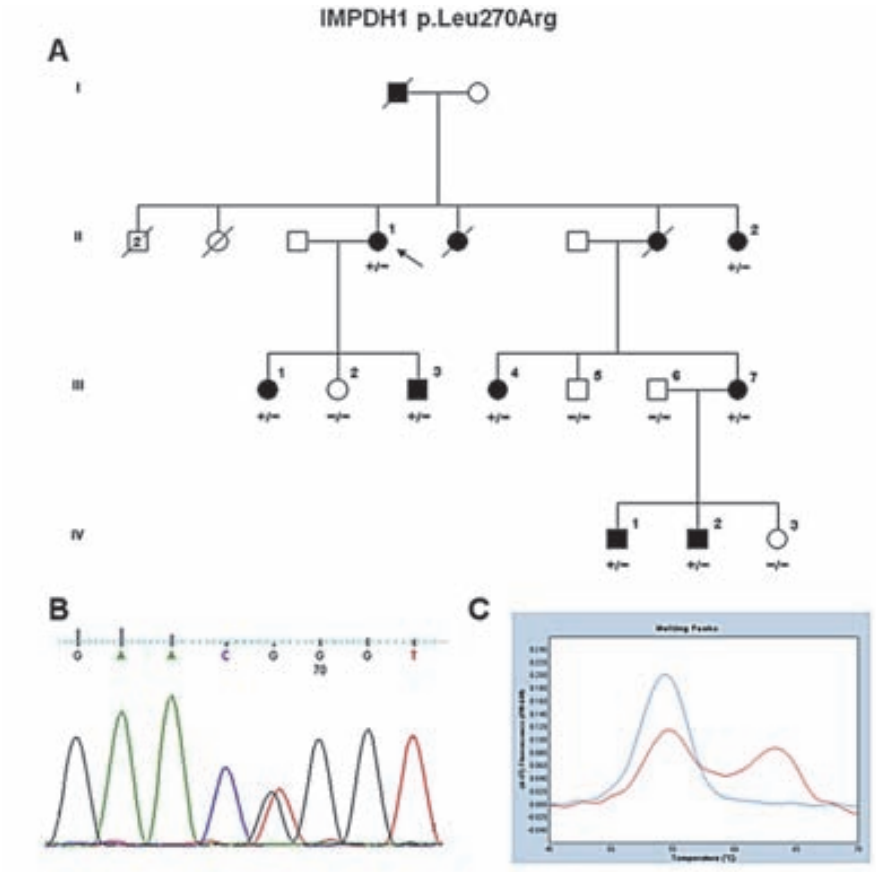


Figure 4.24: Family segregation of the p.L270R mutation in *IMPDH1*. (**A)** Family 93 pedigree; wild type (-/-), heterozygous (+/-) carrier of the mutation. (**B)** Sanger sequencing chromatogram of the mutation. (**C)** A FRET chromatogram obtained by real time PCR showing the wild type (**blue**) and heterozygous mutation (**red**). The arrow indicates the index patient.

- *Family RP-95*

Segregation of the novel sequence and rare variants detected in family 95 showed that only the c.-14delC variant in the *PDE6G* gene is carried by all the patients of the family, including an asymptomatic obligate carrier and two asymptomatic members (Figure 4.25A). This nucleotide deletion is located at position −14 of the *PDE6G* gene, in a promoter region conserved in primates. Here, real-time PCR with a pair of FRET probes was used to screen this variant in 150 controls, detecting 15.2% heterozygous and 1.6% homozygous individuals in our population (Figure 4.25C). To date, only one disease-causing mutation in *PDE6G* has been reported in an arRP family (Dvir *et al.*, 2010). This mutation, c.187+1G>T, in the conserved intron 3 donor splicing site in homozygous carriers was reported as causing RP in a large family in which the heterozygous carriers are unaffected.

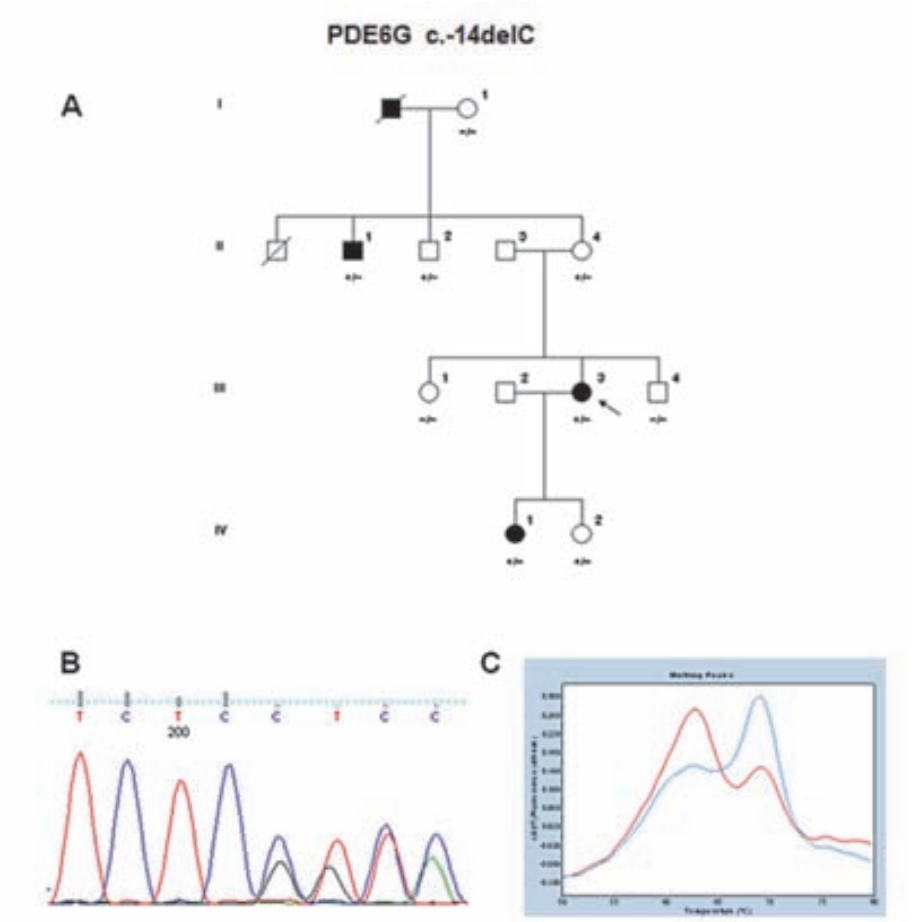


Figure 4.25: Family segregation of the mutation c.-14delC in *PDE6G*. **(A)** Family 95 pedigree; wild type (-/-), heterozygous (+/-) carrier of the mutation. The arrow indicates the index patient. **(B)** Sanger sequencing chromatogram of the mutation. (**C)** A FRET chromatogram obtained by real time PCR showing the wild type (**blue**) and heterozygous mutation (**red**). The arrow indicates the index patient.

- *Family RP-83*

Family 83 showed a complex trait with consanguinity in one branch of the family. Interestingly, a homozygous novel variant p.S1225_E1226insS in the *C2orf71* gene was detected in the index patient. Variants of this gene have been associated with arRP (Collin *et al.*, 2010; Nishimura *et al.*, 2010). However, segregation of the p.S1225_E1226insS variant in the family detected unaffected members who were homozygous and heterozygous carriers (Figure 4.26).

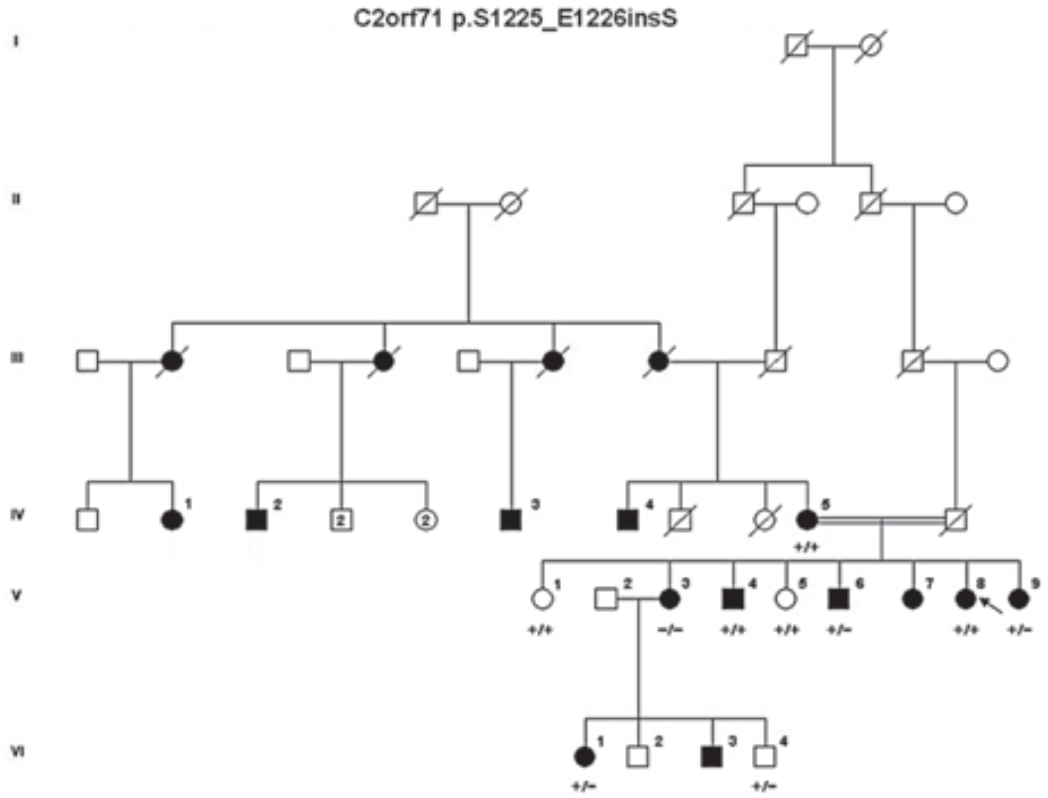


Figure 4.26: Family segregation of the mutation p.S1225_E1226insS in *C2orf71*. Family 83 pedigree; wild type (-/-), heterozygous (+/-) and homozygous (+/+) carriers of the mutation. The arrow indicates the index patient.

4.2.2.2. Whole exome comparative approach

One family – RP 65 – from the five previously studied families was chosen for whole exome analysis (Figure 4.23A). This family was selected for having a very solid clinical diagnosis for adRP and for presenting easy access to DNA samples from multiple members, both healthy and affected, spanning four generations which provides a very big statistical power when studying any candidate gene. Furthermore, all genes presently known to be associated with adRP had been previously analysed in this family in previous studies, inclusive of the 448 candidate genes screening approach.

Two affected members – RP-65_III-4 and RP-65_IV-2 – and one non-affected member – RP-65_IV-1 – were chosen for this study and their whole exomes were captured and sequenced (Figure 4.23A). Their results were then compared and it was anticipated that the majority of the SNPs would be filtered out (since they would have also appeared in the healthy subject and thus are not regarded as disease causing).

4.2.2.2.1. Raw reads quality and sequences assessment

Whole exome sequencing and analysis were executed by Sistemas Genómicos using the latest version of the Illumina HiSeq2000 sequencing platform and Agilent's SureSelect Target Enrichment System for 51Mb, as previously mentioned in the section 3.9.5 of Material and Methods.

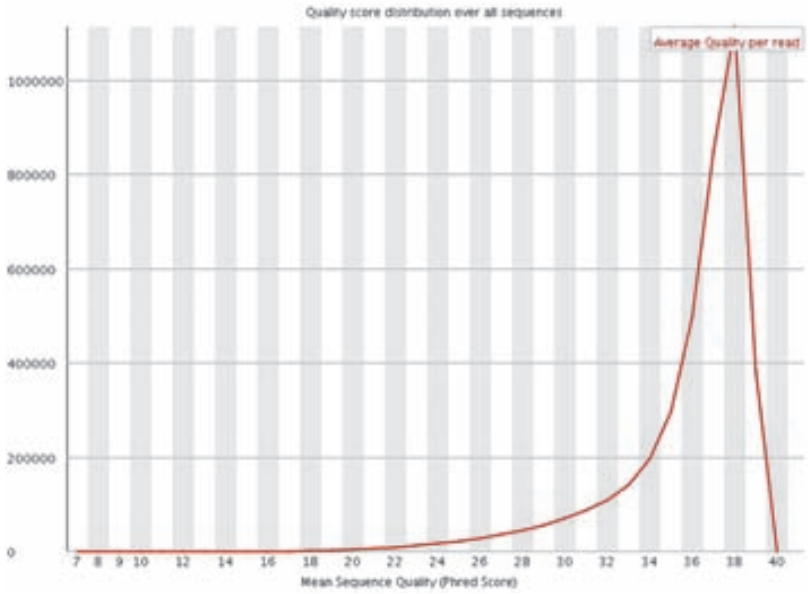


Figure 4.27: Quality score distribution over all sequences. The average quality read is displayed.

The global study of the quality values provided information about the sequencing quality (similar to Sanger's nomenclature). Figure 4.27 shows the average quality score obtained per sequence read. The percentage of mapped filtered samples was between a minimum of 88.73% and a maximum of 92.74% (Table 4.25).

Table 4.25: Sequence reads obtained by whole exome analysis of gDNA capture samples from the three family 65' members.

Sample	Total number of reads	% Filtered readings	Mapped readings
RP-65_III-4	64,879,370	92.74	60,169,128
RP-65_IV-1	66,164,556	91.76	60,712,597
RP-65_IV-2	191,535,462	88.73	169,949,415

4.2.2.2.2. Sequencing variants and Sanger Validation

Similar to the 385K NimbleGen array, the sequence variants detected were associated with their position in the gene transcript, annotated, and classified according to the nomenclature of the Ensembl database. Again, variants were classified into two groups: SNVs and Indels. The sequence variants (SNVs and small Indels) detected in all samples were compared with the variants reported in the database.

As expected in a whole exome analysis, a massive amount of SNVs and Indels were detected per sample (Table 4.26). Given that these samples were from the same family (Figure 4.23A), their results were compared and the first criteria used to filter out the majority of the SNPs was to classify as relevant only those variants present in all affected members, but at the same time not detected in the non-affected member. Figure 4.28 shows the workflow of the analysis and validation of the genomic variants found. The Ensembl v59 database was used to analyse the sequence variants found according to the functional consequence in the target transcript. Consequently, a total of 3,533 SNVs were obtained, 99 of which were unreported in the database. As for Indels, a total of 157 were found; 27 of those were unreported (Table 4.26).

Table 4.26: Sequence variants detected by Whole-exome analysis of genomic DNA capture samples from the three members of the RP-65 family.

Sample	Total Variants	SNVs		Indels	
		Total	Unreported	Total	Unreported
RP-65_III-4	58,537	49,747	3,514	4,177	1,099
RP-65_IV-1	58,099	49,423	3,481	4,117	1,078
RP-65_IV-2	55,691	50,560	1,653	2,703	775
Filtered Variants¥	3,816	3,533	99	157	27

¥ Variants present in all affected subjects but not in the non-affected subject.

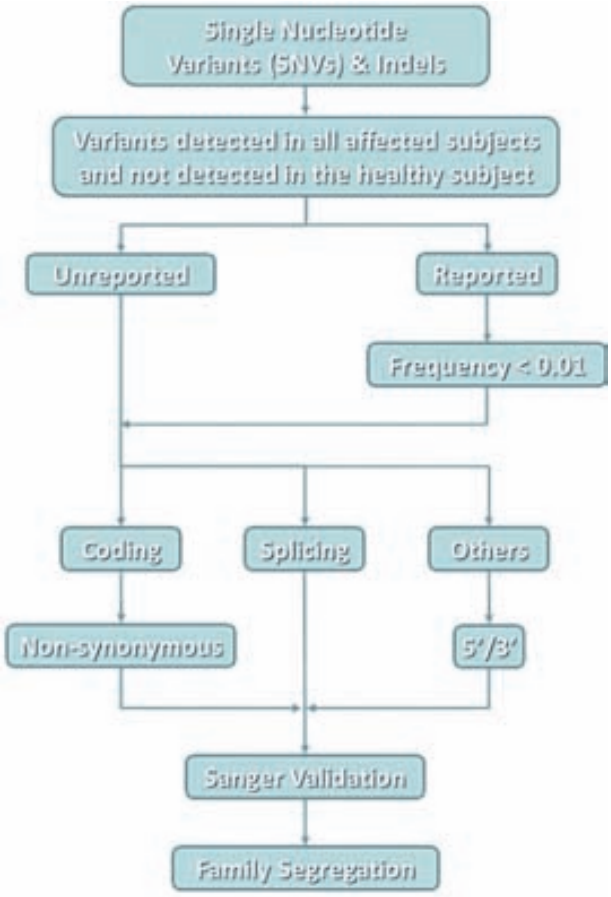


Figure 4.28: Workflow of the analysis of detected sequence variants.

The filtered variants (variants present in all affected subjects but not in the non-affected subject) were then subdivided in categories attending to their position or effect on affected protein. Low population frequency (<0.01) variants were considered relevant as well.

Table 4.27: Type of variants present in the filtered variants distributed by consequence.

Type of Variant	Filtered Variants¥	Unreported
Synonymous	650	15
Non-Synonymous	631	21
Intronic	1707	48
Splice-Site	126	4
Indels	157	27
Low population frequency	122	-

¥ Variants present in all affected subjects but not in the non-affected subject.

From the total of 126 remaining unreported variants, 99 SNVs, and 27 Indels (Table 4.25), 14 were selected for cosegregation studies. Likewise, from the total of 122 low population frequency variants (Table 4.27), 7 were selected for cosegregation studies. These variants were selected

firstly for having a the most deleterious effect on the transcript but also for being located in genes reported to express in the eye or expressing proteins with structural function (Table 4.28).

Table 4.28: Variants selected for cosegregating studies.

RefSeq_ID	Gene	Variant	Protein Change	Type of Variant	Frequency
NM_144663.1	*C11orf40*	c.599-600dupGT	p.Met201ValfsX211	Indel/Frameshift Coding	<0.01
NM_001837.3	*CCR3*	c.478G>A	p.Val160Met	Non-Synonymous	<0.01
NM_001852.3	*COL9A2*	c.150C>T	p.Pro51Leu	Non-Synonymous/Splice-Site	<0.01
NM_153002.2	*GPR156*	c.2426T>C	p.Leu809Ser	Non-Synonymous	<0.01
NM_015112.2	*MAST2*	c.3186G>A	p.Ser1062Ser	Synonymous/Splice-Site	<0.01
NM_001137604.2	*POLR1B*	c.1865C>G	p.Pro622Arg	Non-Synonymous	<0.01
NM_006583.2	*RRH*	c.900-4G>A	-	Intron/Splice-Site	<0.01
NM_022112.2	*ABCA7*	c.109G>A	p.Arg37His	Non-Synonymous	New
NM_001276713.1	*ANKDD1B*	c.567-571delCAGTG	p.Ala190fsX201	Indel/Frameshift Coding	New
NM_018186.2	*C1orf112*	c.171A>G	p.Gln57Gln	Synonymous/Splice-Site	New
NM_001102608.1	*COL6A6*	c.307G>A	p.Gly103Arg	Non-Synonymous	New
NM_000308.2	*CTSA*	c.459C>A	p.Asn153Lys	Non-Synonymous	New
NM_181453.3	*GCC2*	c.3793A>G	p.Ile1265Val	Non-Synonymous/Splice-Site	New
NM_002375.4	*MAP4*	c.4777C>T	p.Leu1593Phe	Non-Synonymous	New
NM_001127178.1	*PIGG*	c.661delC	p.Pro221GlnfsX222	Indel/Frameshift Coding	New
NM_014330.3	*PPP1R15A*	c.484A>T	p.Lys162X	Non-Synonymous/Stop-Gained	New
NM_175732.2	*PTPMT1*	c.361A>C	p.Thr121Pro	Non-Synonymous	New
NM_001031709.2	*RNLS*	c.871T>A	p.Ser291Thr	Non-Synonymous	New
NM_015540.3	*RPAP1*	c.1877G>A	p.Arg626His	Non-Synonymous	New
NM_022112.2	*TP53AIP1*	c.169G>A	p.Gly57Ser	Non-Synonymous	New
NM_003442.5	*ZNF143*	c.791G>A	p.Arg264Gln	Non-Synonymous	New

4.2.3. Molecular Diagnosis of Patients with adRP

All previously developed and studied approaches encompassed a healthcare mission as part of a translational development tasks concerning autosomal dominant Retinitis Pigmentosa. The inclusion of the previously described approaches in the clinical practice was investigated through the analysis of several patients with uncharacterised adRP, some of which were already commented before. A total of 32 adRP patients were analysed; the whole exome subjects were not considered in this count being that it is an ongoing study (Table 4.29).

Table 4.29: Number of patients analysed and their molecular diagnosis successful rate.

Used Approach¥	Nº of Patients Analysed	% of Molecular Diagnosed
LR+EF	9	33,3
MP	18	33,3
CGA	5	20,0
Total	32	31,3

¥ LR+EF, LR-PCR allied with Enzymatic Fragmentation; MP, Multiplex-PCR; CGA, Candidate Genes Array.

The results showed that, from a total of 32 adRP patients, 10 adRP patients were successfully molecular diagnosed (31.1%). A total of 9 disease-causing mutations were detected, 4 of which were unreported (Table 4.30). It is also observable that although LR+EF and MP approaches had similar molecular diagnosis rates (33.3%), LR+EF was more successful in detecting unreported disease-causing mutations; the MP approach primarily detected previously reported disease-causing mutations, with the exception of the new p.Val345Gly variant in the *RHO* gene.

Table 4.30: Variants causing adRP found in 32 patients with the different studied methods.

Gene	Variant	Protein Change	HGMD*	Used Method¥
*IMPDH1*ß	c.809T>C	p.Leu270Arg	New	CGA
PRPF31	c.328_330delATC	p.Ile109del	New	LR+EF
RHO	c.217_219delAAC	p.Asn73del	New	LR+EF
RHO	c.1034T>G	p.Val345Gly	New	MP
RP1	c.2029C>T	p.Arg677X	CM991103	MP/LR+EF
PRPH2	c.518A>T	p.Asp173Val	CM941209	MP
RHO	c.512C>T	p.Pro171Leu	CM910335	MP
PRPH2	c.584G>T	p.Arg195Leu	CM032999	MP
RHO	c.644C>T	p.Pro215Leu	CM003954	MP

* Human Gene Mutation Database

¥ LR+EF, LR-PCR allied with Enzymatic Fragmentation; MP, Multiplex-PCR; CGA, Candidate Genes Array.

ß Incomplete penetrance pattern

5. Discussion

In the clinical practice of sequencing candidate genes involved in a disease in individual patient samples, it is increasingly important to carry out molecular testing. Thus the use of massive DNA sequencing or next generation sequencing (NGS) technologies is a vital practice within any clinical genetic laboratory. Large next-generation platforms are indicated for this task. However, the cost and extremely large capacity of these platforms results in a loss of flexibility regarding the needs of many genetics laboratories where it is sometimes necessary to analyse samples from only one or just a few individuals in a reasonably short time. Thus, technologic firms participating in the NGS marketspace have introduced smaller NGS platforms adapted for clinical use. One such platform, the GS Junior, has successfully proven its potential for molecular diagnosis in molecular genetics laboratories.

5.1. Detection of genetic variants in *BRCA1* and *BRCA2* genes

- *Generation of NGS libraries for BRCA1 and BRCA2 genes analysis*

Here, *BRCA1* and *BRCA2* serve as model for other large genes where genetic testing is costly and time consuming. More specifically, the *BRCA1* and *BRCA2* genes were chosen for this study due to their social and health impact; mutations in these genes are responsible for approximately 7% of all cases of breast cancer and 11 - 15% of all ovarian cancer cases (Claus *et al.*, 1996; Risch *et al.*, 2001; Pal *et al.*, 2005). Consequently, there is a strong demand for screening for *BRCA1* and *BRCA2* genes in a clinical setting which represents a significant social and health care cost. Although clinical protocols for genetic testing (Pruthi *et al.*, 2010) of *BRCA1* and

BRCA2 are well established, the task is nevertheless costly and time consuming as it results in over 40 individual PCR reactions plus their purification and subsequent Sanger sequencing.

Taking advantage of NGS technology in a sequencer which has been appropriately scaled to clinical settings – the GS 454 Junior – the genomic variants in two relatively large genes, *BRCA1* and *BRCA2*, were analysed at the sequence level by means of LR-PCR associated with two distinct enzymatic DNA shearing methods: Fragmentase and Nextera. These analyses were then extended to five patients in parallel and their feasibility was demonstrated. Five parallel samples are clearly not cost-effective on large commonly available NGS platforms (e.g. Roche GS FLX Titanium). However, the results show that this type of analysis could be effective with an appropriately scaled benchtop sequencer, being that the run cost of a large NGS platform can be up to 10 times higher than with the GS Junior platform.

The approach for *BRCA1* and *BRCA2* analysis hereby tested uses a robust LR-PCR method to amplify most genomic regions of *BRCA1* and *BRCA2*, including all CDS and flanking regions, using just 22 primers in eleven PCR reactions. If necessary, it can be easily performed with seven DNA samples in parallel using one 96-well plate and a gradient PCR program (Figure 4.6).

For the preparation of DNA sequencing libraries, Nextera technology (Caruccio, 2011) was examined. This technology proved very effective for this assay and demanded no modification to the manufacturer's operating protocols except for the adaptors which must contain sequences compatible with the GS Junior platform. Fragmentase technology was also examined and was likewise proven effective for this assay. An enzymatic method to fragment the large amplicons was chosen for its superior reproducibility and improved kinetic control versus mechanical DNA fragmentation (recommended in the Roche GS Junior Library Preparation Manual). Moreover, no special device or laboratory installation is necessary.

A high reproducibility of intra-sequencing results was observed, noting a similar profile of depth and coverage in each sequenced sample even when different fragmentation methods were employed. Despite the fact that both Nextera and Fragmentase technology assures a near-to-random enzymatic action (recombinase) on DNA, a different representation of the sequences from LR-PCR fragments was observed. Thus, the sequences of the 5' or 3' ends are in general less represented. This compromises the depth sequences of exons 10, 19, and 20 of *BRCA1* located at these ends (at less than 75 bp). To avoid this poor depth, primers selected for LR-PCR should, in general, be located more than 100 bp from an exon. However, the poor depth obtained in exon 22 of *BRCA1* could, as in the case shown for the homopolymer stretch regions, be due to an enzymatic sequence effect of Nextera or due to a poor quality reading of these sequences by

the GS Junior Platform. In fact, the reads of homopolymers longer than seven nucleotides were inaccurate and these sequences may have been filtered by the GS Junior software.

Both LR-PCR and Fragmentase or Nextera technology could be performed easily in a typical molecular genetics laboratory (Figure 5.1). Nevertheless, the Nextera technology was acquired by Illumina, which made it less accessible and therefore this technology was not employed on further experiments.

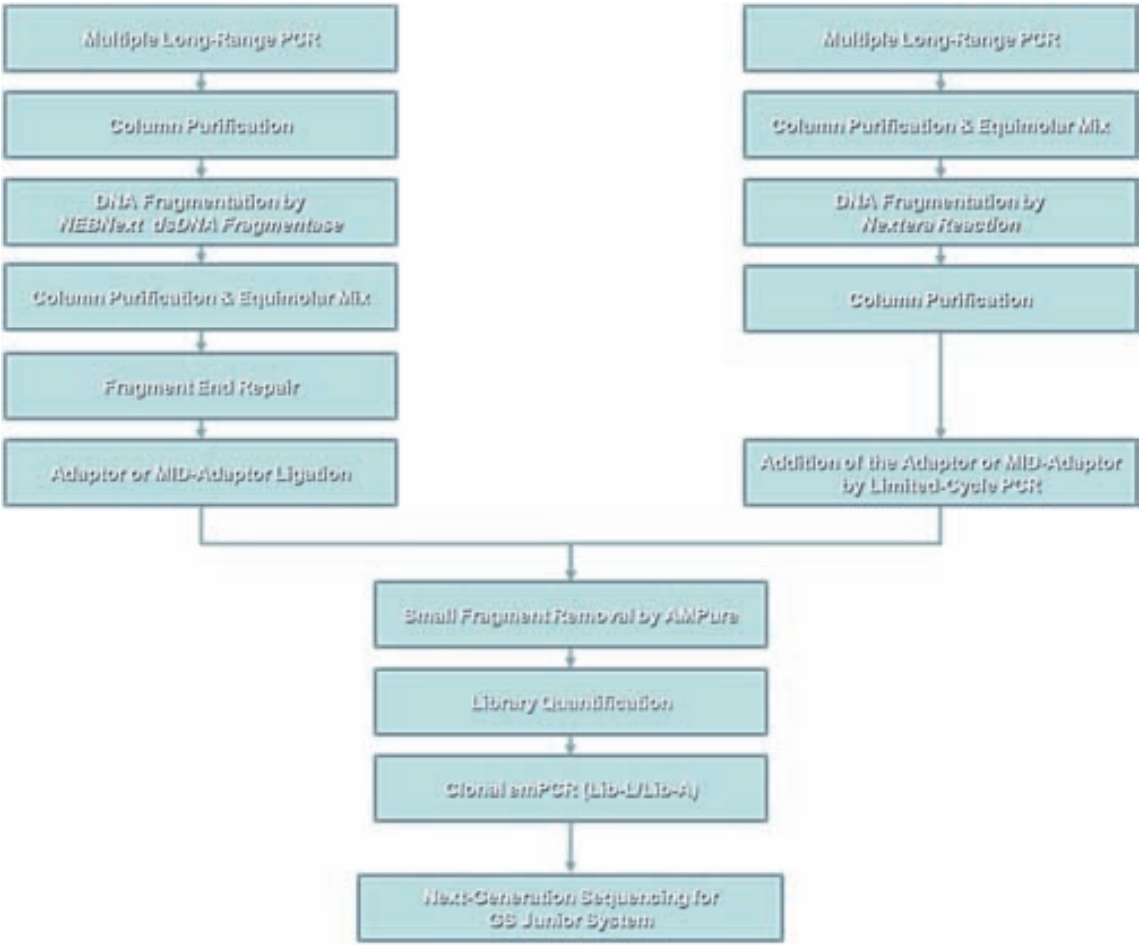


Figure 5.1: Fragmentase and Nextera used in NGS. Workflow comparing the Fragmentase and Nextera technologies.

Sequencing data were obtained and analysed using the Reference Mapper program of the 454 GS Junior. The resulting HCDiffs table showed all genetic variations previously detected by canonical PCR followed by Sanger sequencing, except the insertion of a T at position c.3264 in *BRCA2* gene. However, when this variation was checked manually in the flowgram, alignment, and AllDiffs table (the "All Differences" table contains all variants before passing the high confidence filters), the c.3264 mutation was present at a sequence depth of 60 with 34.5% total variation. Therefore, this false-negative in the HCDiffs table should be attributed to the software

features rather than a limitation in sequencing capacity. Accordingly, the commercially available software CLC Genomics Workbench v4.8 was also used to analyse the sequencing results from the SFF sample file from GS Junior platform. This software revealed all the variations previously identified including the T insertion at position c.3264. Therefore, its use together with the Reference Mapper program of GS Junior Platform is recommended.

Sequences with a total depth of less than 10X were filtered out and were not used in variants analysis. For diagnostic purposes only, variants with a statistical power of 99.9% were considered. This was achieved in a GS Junior platform by the employment of statistical analysis (De Leener *et al.*, 2011a) and fixed a cut-off of 38X with >25% variation as a discriminative threshold. Sanger sequencing was used to validate all of the variants detected in *BRCA1* and *BRCA2* which passed the cut-off. Notably, zero false-positives were identified in the coding and flanking sequences. However, while these genetic diagnostic methods are under development, the detected candidate mutations should always be validated by canonical Sanger sequencing before the results are reported to the patient (Mattocks *et al.*, 2010).

Using this approach, the greater number (60%) of sequence reads corresponded to intronic regions of *BRCA1* and *BRCA2*. These intronic sequences may constitute only background noise in a DNA capture or they may be absent in amplicon assay strategies. In spite of this, these intronic sequences were predominant in this approach and therefore imposed a limit on the number of samples that could be run in parallel. However, this limitation does not compromise the sequencing of CDS and the flanking regions. Moreover, certain intronic sequence variations may create aberrant splice sites that may generate mutant alleles (Chillon *et al.*, 1995; den Hollander *et al.*, 2006; Rio Frio *et al.*, 2009) and, for this reason, annotation of these sequences could be of interest in a future analysis.

Other strategies that employ NGS analysis have been reported to characterize variations in *BRCA1* and *BRCA2*. One strategy uses a genomic DNA solution for the capture of both genes together with other genes involved in breast or ovarian cancer (Walsh *et al.*, 2010). However, this approach has limitations as it requires large and expensive NGS platforms. Moreover, only about 50% of the total capture sequences are on the target. Similarly, whole exome analysis involves DNA capture of all exonic sequences, and although it appears cost-effective relative to the number of genes analysed, the reported coverage, concerning the *BRCA1* and *BRCA2* genes, is only 80%.

A different, previously reported approach involves the generation of more than one hundred amplicons that contain the coding and flanking sequences of *BRCA1* and *BRCA2*; indeed a

commercially available kit for *BRCA1* and *BRCA2* testing using a multiplex approach with NGS has been developed (Multiplicom N.V., Niel, Belgium). This is a multiplex open method for the GS Junior platform and it has been reported as very practical and cost-effective (De Leeneer *et al.*, 2011b). Undesirably, this commercially kit surely holds a significant amount of work and research behind it; the development of such commercial kits is expected only for highly incident diseases. Thus, while this commercially available kit holds excellent promise, the development of such commercial kits is expected only for highly incident diseases. On the other hand, the LR-PCR method allied with enzymatic fragmentation hereby proposed can be easily applied to other large gene associated diseases that have low incidence in the population and at readily assumable costs.

5.2. Detection of mutation in genes with known association to adRP

- *NGS libraries generated by LR-PCR allied with Fragmentase for adRP genetic testing*

Autosomal dominant retinitis pigmentosa is a good example of a monogenic disease with a clear heterogeneity that involves mutations in many genes. A method based on LR-PCR that avoids the use of large numbers of primers and PCR reactions and which results in a significant reduction in cost and time for PCR amplification of all coding and flanking sequences of 12 common genes associated with adRP was developed. Mutations in these genes account for more than 95% of the reported disease-causing mutations associated with adRP. In Western populations, most mutations causing adRP are found in *RHO*, with an incidence of 19% – 25% in a Spanish population (Ayuso and Millan *et al.*, 2010). In this population, the other known adRP genes account for between 0.7% and 10% (Ayuso and Millan *et al.*, 2010). Although in previous and recent NGS surveys for adRP no mutations in *TOPORS* or *RP9 (PAP1)* causing RP were detected, these genes were included as candidate genes in the study involving LR-PCR allied Fragmentase approach. Likewise, the complete sequence of *NR2E3* was included as well, which has previously been associated with recessive RP, being that some mutations in this gene had been reported to also be dominant (Gire *et al.*, 2007; Coppieters *et al.*, 2007; Escher *et al.*, 2009).

For the LR-PCR fragments design, the structure of each gene and the clustering of the coding sequences were considered. Accordingly, five exons (exon 1 of *CA4, IMPDH1, RP9,* and *PRPH2* and exon 2 of *TOPORS*) were omitted from the LR-PCR fragments because these exons possess a long 3'-intron. It is more cost-effective to analyse these exons for mutations separately with conventional techniques. Alternatively, these exons could be processed as individual

amplicons, added to the NGS library pool and be sequenced all together. For the similar reasons, the *NRL* gene, a common adRP-associated gene that accounts for 1% of the Spanish adRP population, was not included in this analysis because the genomic structure allows the gene to be analysed more easily with direct genomic Sanger sequencing.

All genetic control variants that had previously been characterized with conventional screening and Sanger sequencing (Table 4.7) were successfully detected (Table 4.11). These variants were identified with a sequence depth always above 30X. Initially, a cutoff at a minimum total sequence depth of 20X with a sequence variation >30% for a heterozygous variation was established (Brockman *et al.*, 2008). However, sequence variants with a total depth below 20X with a sequence variation between 30% and 65% were detected and later validated with capillary Sanger sequencing. Thus, and in accordance with a recent report (De Leeneer *et al.*, 2011a), a total depth of >10X values should be considered to increase the detection power to near 100%.

This approach detected missense variant and small deletions/insertions (indels), as well as zero false-positive calls in the coding and flanking sequences of any gene in any sample. The 23 novel intronic variants found (Table 4.12) were also validated as real variants with Sanger sequencing. However, sequence analysis of variants present in or near homopolymer stretches larger than six nucleotides proved unreliable, being areas of potential false positives and false negatives. This sequencing limitation has been reported when using pyrosequencing methodologies (Huse *et al.*, 2007; Gilles *et al.*, 2011). This too could be an issue with the alignment algorithm used by GS Reference Mapper Software, as previously noted in this section, due to the pyrosequencing inaccuracy reading homopolymers larger than six nucleotides; the sequence alignment in these regions is also affected. Nevertheless, it is advised that any apparent positive result continue to be validated with conventional Sanger sequencing (Walsh *et al.*, 2010; Hernan *et al.*, 2012; de Sousa Dias *et al.*, 2013).

Using LR-PCR amplification of genomic DNA, the highest number of sequences obtained corresponds to intronic regions of the adRP genes, similarly with *BRCAs* genes. These intronic sequences may constitute just background noise in DNA capturing or may be absent in amplicon assay strategies. Nevertheless, in this approach, these intronic sequences comprise the majority (53%) and thus limit the number of samples that can be run in parallel, though this does not compromise the sequencing of coding and flanking sequences. However, some intronic sequence variations may create aberrant splice sites that may generate mutant alleles (Rio Frio *et al.*, 2009); the annotation of these sequences could be an area of interest for future analysis (Houdayer *et al.*, 2008). In fact, all intronic variants were analysed *in-silico* with a splicing program

(Splice Site Prediction by Neural Network) to investigate an eventual putative change in the splicing. None of the intronic variants analysed demonstrated a significant change in splicing values.

This approach for screening for mutations in the common adRP genes demonstrated that using MIDs, at least four samples could be processed in parallel, proving an effective method for analysing multiple individual index patients with adRP in the same run, saving time and costs. Moreover, the limited number of novel putative disease-causing variants typically obtained must be cosegregated in the family resulting in a limited additional cost in this diagnostic approach. Recently, an effective targeted high-throughput DNA capture and sequencing method has been used to analyse 40 genes associated with RP in isolated cases of RP (Simpson *et al.*, 2011). However, this approach requires custom arrays and larger platforms for NGS which are better suited to a large survey of patients with unclassified (isolated) RP.

The validity of this method was demonstrated by its detection of two novel mutations, in two index patients of adRP, that had not been detected with commercially available arrays. These genetic variants correspond to two novel mutations that cause adRP. The mutation p.Asn73del in *RHO* was detected in one index patient and his mother; both diagnosed with RP. The Asn-73 residue is conserved among the four proteins: rhodopsin and red, blue, and green opsins (Raman *et al.*, 1999), suggesting that it plays an important structural or functional role. Moreover, *in vitro* studies of bovine *Rho* have demonstrated the critical role of Asn-73 in binding with arrestin opsins (Raman *et al.*, 1999). The other novel genetic variant was detected in *PRPF31*; an ATC deletion at position g.7092_7094, which corresponds to the sixth nucleotide of exon 4. Whether or not this deletion affects the acceptor signal splicing site of exon 4 remains to be investigated. Cosegregation of this mutation in the available members of the family showed one obligate asymptomatic carrier. However, incomplete penetrance for mutations in *PRPF31* that cause adRP (Rivolta *et al.*, 2006) has been reported previously.

The methodology reported here may be extended with new genes associated with adRP that could be processed together with these common genes or additionally used to create a second block of genes associated with adRP that could be analysed separately using this method. The LR-PCR allied with enzymatic shearing approach circumvents some limitations found in previous surveys of detection of adRP-causing mutations. Thus, all sequences of known candidate genes were analysed for mutations.

In most of the traditional routine analysis of adRP, some of these genes regions in which mutations that cause adRP have never been found, are usually left out from the analysis. This

may cause a bias of mutations in these genes (Sullivan *et al.*, 2006; Bowne *et al.*, 2011; Blanco-Kelly *et al.*, 2012; Millá *et al.*, 2002). Indeed, the novel variant p.Met1140Val in the *PRPF8* gene was discovered in one of our clinical analyses, in a region that is not typically analysed and thus would go undetected in most routine analyses; this mutation was later discarded as disease-causing by cosegregation studies. In this approach, not only were previously reported mutations detected but two novel mutations were also discovered whereas any novel mutation would go unnoticed in the array approach (APEX). The parallel analysis of four patient samples made this approach cost-effective. Moreover, the samples excluded for mutations by this approach comprise suitable candidates for seeking novel genes associated with adRP by a complete NGS exome analysis, for example. Furthermore, this approach could be used for mutation analysis in other heterogeneous monogenic diseases.

- *Target-capture of 23 genes associated with adRP*

Multiple genes can be involved in genetically heterogeneous diseases and molecular diagnosis in adRP requires mutation analysis of more than 20 genes. Such analysis by multiplex PCR requires hundreds of amplicons which is associated with a high cost and a time-consuming diagnostic (Bowne *et al.*, 2011). In an attempt to overcome these limitations, an alternative strategy that is typically deployed using large scale NGS platforms was assayed for the benchtop GS Junior platform. This assay consisted of target selection by hybridization capture of DNA fragments on RNA oligonucleotides in solution. While this approach allowed the enrichment of most gene regions of interest, a large variation in coverage was observed. As a result, some of the coding sequences from several genes (*PROM1, PRPF4, RP9, and RPGR*) were not captured or did not pass the established cut-off, and were thus insufficient for molecular diagnosis. In addition, other unwanted regions were captured and detected (approximately 50% of the total reads), which may limit the capacity to perform a parallel sample analysis in a benchtop NGS platform like that used. The incomplete coverage of target sequences and the inability of parallel sample analysis make this approach unsatisfactory in molecular diagnostics in a clinical genetics laboratory.

- *Parallel NGS library construction by Multiplex-PCR for adRP analysis*

After applying the approach of LR-PCR in conjunction with enzymatic fragmentation and following the advances in multiplex-PCR, a mutation detection assay was set-up. Comparing this multiplex-PCR methodology with the previously commented LR-PCR approach, it included less intronic bases compared with previous ones. It was therefore expected that this approach would allow for

an increased number of patients to be analysed in parallel. This assay covered a gene panel of 11 adRP candidate genes (which together also account for more than 95% of the reported disease-causing mutations) through the construction of NGS-libraries containing over forty amplicons. These amplicons were obtained by multiplex-PCR of 6-plex. Here, it was decided to not include the *RP9* (*PAP1)* and the *TOPORS* genes, which have been previously included in the LR-PCR and Fragmentase approach; as earlier commented, mutations in these genes are very much in the minority.

This methodology involved a vast primer-designed work, targeting the successful amplification of all amplicons at similar reaction (components and temperature) conditions and, obviously, primer compatibility. The amplicons size was taken into account as well, considering the GS Junior platform read length capacity. This was possible only through the support of thermodynamics-based programs (Qu *et al.*, 2009, 2012) for checking PCR primer specificity and compatibility.

To validate this assay, a parallel run of five samples (using 5 MIDs) was performed. Each MID contained 3 previously validated point mutations distributed within their 6-plex. All point mutations were successfully detected with the reads per amplicon count always above 30X.

Sequence analysis also showed that nearly 20% (42,511 raw reads) of total raw reads were discarded for being too short or having too many poor incorporations or interruptions during the sequencing process. These low quality sequences may have been originated due to primer interferences or interactions during the limited cycle PCR, used to introduce the sequencer-specific-adaptor for NGS analysis. This was further confirmed in agarose gel (Figure 4.18) where a large lane averaging 150 bp was detected. These unwanted sequences decrease the number of high-quality reads in the GS Junior platform and thus impose a limit on the number of patients that can be analysed in parallel. Nevertheless, all amplicons demonstrated a good read per amplicon count (Table 4.18), with the exception of the *RP1* gene, which demonstrated low read count in all five samples. This indicates that this specific amplicon is underestimated in its primer mix and therefore increasing its final concentration is advised. Despite that, all present *RP1* point mutations were successfully detected.

Additionally, and while not being a requirement of a clinical genetics laboratory, this approach could be employed to perform quick screening of the four most relevant genes (*RHO*, *PRPH2,* and *PRPF31*), in which nearly 45% of all disease causing mutations are found; this would be possible by performing multiplex-PCR using just Plex-A, B, and C (Table 3.17). This would allow the analysis of over a dozen patients in one single run in the GS Junior platform thus drastically decreasing the per-patient cost.

- *Bead-Linker-Primer complex combined with emPCR as a new multiplex system*

In an effort to improve the analyses of multiple genes, a new multiplex system has been successfully devised. Here, a new primer isolation technique, which uses an oligonucleotide functionalised with biotin at the 3'-end, attached to a streptavidin-functionalised magnetic bead as a carrier of the PCR primers (BLP), was developed and put to the test. The combination of this primer isolation technique with emulsion PCR (emPCR) was studied in order to ensure the amplification of dozens of different PCR fragments in just one tube thus simultaneously achieving a more capable multiplex and an optimization of the NGS library construction. Unfortunately, use of this technique resulted in too many low-quality reads generated by the GS Junior platform. This was due to the formation of unspecific fragments, averaging 150 bp, that interfered with the sequencing run (Figure 4.18 and 4.19). Furthermore, these unspecific fragments were not removed even after the double AMPure purification because they were larger than 100 bp. Also, direct agarose gel purification would make the library unviable for the GS Junior platform due to an inhibition caused by the agarose gel purification procedure on the library clonal step. Still, 21 of the intended 44 amplicon were detected by the AVA software.

A further investigation of the library generated by this method confirmed that 9 of the 44 amplicons failed to be amplified (Figure 4.20). However, this means that 35 fragments were successfully amplified in just one tube therefore a multiplex reaction with a success rate of approximately 80% was achieved. The detection of only 21 of the 35 amplified fragments by the GS Junior platform could be explained by the variety of reaction conditions needed to amplify all fragments and their kinetics, despite all efforts toward similarity. The M13-tag may also play a role in generating unwanted fragments which compete with the main reaction. These fragments, averaging 150 bp, are responsible for the high amount of low-quality reads on the GS Junior platform.

In theory, this technique would allow for the preparation of libraries where a virtually unlimited number of amplicons could be produced in a single tube. However, additional research and development would be required before its use in NGS library preparation for molecular diagnosis.

5.3. Detection of genetic variants in new genes responsible for adRP

- *Screening analysis of 448 adRP candidate genes*

Although this analysis was limited to 448 candidate genes, a considerable number (4730) of genetic variants were identified in the samples from the five patients analysed. After filtration of

the data, 56 novel sequence variants remained, six of which proved to be false positives after Sanger sequencing. Thus, in the five samples assayed, an average of ten novel sequence variants in RP candidate genes were obtained. The *in-silico* prediction of the pathogenic effect of these variants was performed with PolyPhen, MutPred, and SIFT algorithms, showing several potentially disease-causing variants per sample. These predictions were based on the putative loss of protein function. However, in a dominant disease (adRP) a pathogenic mechanism of gain-of-function or dominant-negative effect may occur rather than a haplo insufficient mechanism caused by a loss of function. Consequently, prediction of a pathogenic effect of genetic variants by PolyPhen, MutPred, and SIFT are limited in adRP. Moreover, segregation of variants in the families excluded most of these novel candidate variants as a cause of adRP. This clearly limits the use of this approach to large families with several RP-affected members and excludes an effective analysis for adRP in isolated cases.

However, in these analyses, three variants that could potentially be disease causing were discovered. In family 93, the novel p.Leu270Arg mutation in the *IMPDH1* gene cosegregated with the disease and it is likely to be the cause of RP in this family. The mutation p.Met96Thr in the *NRL* gene in family 645 was detected in this study was well. This *NRL* mutation had already been reported and could be the cause of adRP with incomplete penetrance in this family (Dvir *et al.*, 2010). Although none of the other sequence variants that were found segregated with RP in this family, the possibility cannot be ruled out that another variant outside the candidate genes analysed could be the cause of RP in this family. In the index case of family 65 the mutation p.Ala76Val in *NRL* was detected, which was previously considered to be possibly RP disease-causing (Hernan *et al.*, 2011). This mutation was not present in one RP affected member and present in an unaffected member of this family. Consequently, the variant p.Ala76Val seems not to be causing adRP in this family. In family 95 and family 83, variants were detected in *PDE6G* and *C2orf71* genes, respectively. In both cases the variant was present in all the affected members of the family. Mutations in those genes had been previously associated with arRP (Collin *et al.*, 2010; Nishimura *et al.*, 2010; Hahn *et al.*, 1994). However, the variants detected in *PDE6G* and *C2orf171* were found in homozygous unaffected carriers or in controls, suggesting that neither of these variants is disease-causing.

NGS analysis confirmed the large number of genetic variants in an individual. Our results show the large number of polymorphisms and rare variants in candidate genes that may be involved in retinal function. Although only two disease-causing genetic variants of adRP were found in our samples, our analysis revealed multiple variants that are predicted as potentially pathogenic. The

effect of these genetic variants on the structure and function of the retina is largely unknown. This existing variability in retinal function genes could explain the high heterogeneity in the clinical expression of RP, even in patients from the same family. Data obtained with NGS analysis could be a potential source for phenotype/genotype correlations in a novel multi-gene approach. Our results revealed the possible disease-causing mutation in two adRP families despite the high number of retinal candidate genes analysed, thus demonstrating the genetic complexity of adRP. As an alternative, exome sequencing in adRP samples was the method of choice to increase the chance of detecting a disease-causing genetic variant for one of the five families, the family 65.

- *Whole exome comparative analysis*

Following the preceding results, family 65 was chosen to perform whole exome analysis. Here, the methodology involved the comparison of two adRP affected members' whole exomes with one healthy member whole exome; one affected member was closely related (brother, same sex) to the healthy member while the other affected member was relatively distant (uncle) in the family tree concerning the same healthy member.

All obtained variants present in the two affected subjects but not detected in the non-affected subject were then selected. Using this method, 126 unreported variants (SNVs and Indels) were obtained; this is a significantly low number of variants having in mind the whole exome was analysed and that an average of 3867 unreported variants were obtained per individual whole exome analysis. Near twenty of these variants (selected by its strength as candidate gene) were analysed so far but none cosegregated in the family. Low population frequency (<0.01) variants were considered relevant as well.

In order to be effective this method is limited to big familial cases being that the results hereby presented demonstrate that at least two or more affected members are needed to filter most non-relevant variants and therefore make the analysis more direct. In fact, the rare variant c.900-4G>A, a splice-site variant in the *RRH* gene (a gene with specific expression in retina hat interacts with *RHO*) was, until a certain point, the prime suspect of disease-causing for this family but later excluded because it was not detected in one affected member (III-3, Figure 4.23A) which in turn had two affected sons (IV-2 and IV-3, Figure 4.23A). Improbably enough, this rare variant was present in the other parent. This variant would be initially excluded if the RP-65_III-3 would have been included in this study.

Given the dominant nature of adRP, variants may be very much in the minority or even unique in some particular families. Another consideration is its genetic complexity and heterogeneousness

and even though it is a monogenic disease, multiple gene interactions, incomplete penetrance, RNA aberrations, and big or relatively big deletions should be considered but would not be detected by this whole exome analysis.

Also, analysis of individuals is impractical for whole exome analysis. As demonstrated here, nearly 50,000 variants are found per sample.

With no healthy family member to compare, and even with the application of multiple *in-silico* mutation effect prediction tools and other available filters for these kinds of massive data analyses, thousands of novel variants would still need to be verified. Moreover, no family cosegregation would be possible and thus no confirmation could be obtained. Hence, the results obtained for a negative mutant cosegregation in relatively large families showed as a powerful tool to exclude potential disease-causing genetic variants. In contrast, a positive cosegregation of a variant in these families would strongly support its role in disease causing. According with these considerations it seems that the use of whole exome analysis for molecular diagnosis is limited to families with more than at least two patients that and when the disease causing-mutation was previously characterized in other patients (or families).

In addition to the above comments, it is necessary to consider the limitation of the whole exome capture approach: several regions of the exome are uncovered or poorly covered by whole exome analysis. On the other hand, new genes and new transcripts are being discovered at a rapid rate and it is not unlikely that many disease causing mutations remain undetected for these reasons.

5.4. General considerations of molecular diagnosis of genetic diseases by NGS

The development of massive sequencing technologies or NGS represented an important advance in DNA sequencing. Since the emergence of NGS in 2006, efforts were made to permit massive sequencing in a clinical setting. At its introduction, this technology was practical for only large centres that could support the substantial cost of a large NGS platform. Just four years later, the benchtop GS Junior was introduced as the first NGS platform scaled-down to meet clinical needs. It permitted the use of massive sequencing in molecular genetics laboratories albeit with undeveloped protocols that were both cost and time-intensive.

The approaches hereby developed for this platform have proven effective and have already been introduced to routine clinical practice allowing the reduction of the sequencing costs as well as

response time. Moreover, no special device or laboratory installation, other than the GS Junior platform itself, was necessary. Regrettably, the need to validate every genomic variant that is detected by these technologies is definitely a drawback. A future afford should be done by the scientific community towards the validation of the NGS platforms for molecular diagnosis uses.

On the other hand, for the molecular diagnosis of patients suffering from complex heterogeneous genetic diseases such as adRP, the use of NGS remains limited to large familial cases were probable disease-causing mutations can be cosegregated and subsequently studied on the population. When only one or two members of the same family are available, molecular diagnosis proves uncertain, difficult, or even impossible in the event of a novel disease-causing mutation (or one mutation located in a gene not yet associated with that disease). The significant genetic diversity from one individual compared to another, even if they are closely related, limits the detection of the probable genetic cause for their disease. Thus, in these cases new tools such as functional studies must be developed to achieve effective molecular diagnosis.

6. Conclusions

1) It is hereby demonstrated the value of next generation sequencing (NGS) in the detection of mutations in large genes associated with a genetic disease as well as heterogeneous diseases where multiple genes are involved. This technology represents an innovation in the field of DNA sequencing and it is remarkably valuable in biomedical research.

2) The introduction of NGS platforms scaled to the clinical setting has opened the doors to development of laboratory protocols for the routine use of molecular diagnosis in an effective way, significantly reducing both costs and response times.

3) Novel approaches using NGS were developed and studied resulting in the efficient detection of mutations in genes commonly associated with autosomal dominant Retinitis Pigmentosa (adRP), a genetically complex heterogeneous disease.

4) Among the studied approaches, Long-Range and multiplex-PCR techniques proved more efficient for mutation detection in the molecular diagnosis of adRP. With these methodologies, gene regions excluded from traditional screening methods were also analysed thus avoiding bias in the location of mutations in candidate genes.

5) The massive analysis of candidate genes, such as the 448 genes screening array of the five index cases from adRP affected families, resulted in an average of 10 candidate variants for causing adRP. Their cosegregation in the families represented an extensive labour of analysis resulting in the discovery of the new adRP-causing mutation p.Leu270Arg in the *IMPDH1* gene, successfully diagnosing family 93.

6) Whole exome analysis of adRP patients and their relatives revealed an enormous number of variants causing a real bottleneck for this approach in the task of segregation of the putative hundreds of variants. In contrast, the presumed decreasing costs of sequence capturing and NGS technologies makes it feasible in near future to sequence the whole exome of all affected and unaffected members of a family rather than just the index cases. Comparison of the data obtained should facilitate the characterization of novel disease-causing variants in large families.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248-9. doi: 10.1038/nmeth0410-248. PubMed PMID: 20354512; PubMed Central PMCID: PMC2855889.

Ammann F, Klein D, Franceschetti A. Genetic and epidemiological investigations on pigmentary degeneration of the retina and allied disorders in Switzerland. J Neurol Sci. 1965 Mar-Apr;2(2):183-96. PubMed PMID: 5878602.

Ansorge W, Sproat B, Stegemann J, Schwager C, Zenke M. Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. Nucleic Acids Res. 1987 Jun 11;15(11):4593-602. PubMed PMID: 3588303; PubMed Central PMCID: PMC340882.

Ansorge W, Sproat BS, Stegemann J, Schwager C. A non-radioactive automated method for DNA sequence determination. J Biochem Biophys Methods. 1986 Dec;13(6):315-23. PubMed PMID: 3559035.

Astier Y, Braha O, Bayley H. Toward single molecule DNA sequencing: direct identification of ribonucleoside and deoxyribonucleoside 5'-monophosphates by using an engineered protein nanopore equipped with a molecular adapter. J Am Chem Soc. 2006 Feb 8;128(5):1705-10. PubMed PMID: 16448145.

Ayuso C, Garcia-Sandoval B, Najera C, Valverde D, Carballo M, Antiñolo G. Retinitis pigmentosa in Spain. The Spanish Multicentric and Multidisciplinary Group for Research into Retinitis Pigmentosa. Clin Genet. 1995 Sep;48(3):120-2.

Ayuso C, Millan JM. Retinitis pigmentosa and allied conditions today: a paradigm of translational research. Genome Med. 2010 May 27;2(5):34. doi:10.1186/gm155. PubMed PMID: 20519033; PubMed Central PMCID: PMC2887078.

Barnes, C., Balasubramanian, S., Liu, X., Swerdlow, H. & Milton, J. Labelled nucleotides. US Patent 7,057,026 (2002)

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008 Nov 6;456(7218):53-9. doi: 10.1038/nature07517. PubMed PMID: 18987734; PubMed Central PMCID: PMC2581791.

Bentley DR. Whole-genome re-sequencing. Curr Opin Genet Dev. 2006 Dec;16(6):545-52. Epub 2006 Oct 18. Review. PubMed PMID: 17055251.

Berger W, Kloeckener-Gruissem B, Neidhardt J. The molecular basis of human retinal and vitreoretinal diseases. Prog Retin Eye Res. 2010 Sep;29(5):335-75. doi: 10.1016/j.preteyeres.2010.03.004. Epub 2010 Mar 31. Review. PubMed PMID: 20362068.

Bernal S, Solans T, Gamundi MJ, Hernan I, de Jorge L, Carballo M, Navarro R, Tizzano E, Ayuso C, Baiget M. Analysis of the involvement of the NR2E3 gene in

autosomal recessive retinal dystrophies. Clin Genet. 2008 Apr;73(4):360-6. doi: 10.1111/j.1399-0004.2008.00963.x. Epub 2008 Feb 20. PubMed PMID: 18294254.

Bessant DA, Payne AM, Mitton KP, Wang QL, Swain PK, Plant C, Bird AC, Zack DJ, Swaroop A, Bhattacharya SS. A mutation in NRL is associated with autosomal dominant retinitis pigmentosa. Nat Genet. 1999 Apr;21(4):355-6. PubMed PMID: 10192380.

Bhattacharya SS, Wright AF, Clayton JF, Price WH, Phillips CI, McKeown CM, Jay M, Bird AC, Pearson PL, Southern EM, *et al.* Close genetic linkage between X-linked retinitis pigmentosa and a restriction fragment length polymorphism identified by recombinant DNA probe L1.28. Nature. 1984 May 17-23;309(5965):253-5. PubMed PMID: 6325945.

Blanco-Kelly F, García-Hoyos M, Cortón M, Avila-Fernández A, Riveiro-Álvarez R, Giménez A, Hernan I, Carballo M, Ayuso C. Genotyping microarray: mutation screening in Spanish families with autosomal dominant retinitis pigmentosa. Mol Vis. 2012;18:1478-83. Epub 2012 Jun 5. PubMed PMID: 22736939; PubMed Central PMCID: PMC3380913.

Bordeleau L, Panchal S, Goodwin P. Prognosis of BRCA-associated breast cancer: a summary of evidence. Breast Cancer Res Treat. 2010 Jan;119(1):13-24. doi: 10.1007/s10549-009-0566-z. Review. PubMed PMID: 19789974.

Borràs E, de Sousa Dias M, Hernan I, Pascual B, Mañé B, Gamundi M, Delás B, Carballo M. Detection of novel genetic variation in autosomal dominant retinitis pigmentosa. Clin Genet. 2013 Nov;84(5):441-52. doi: 10.1111/cge.12151. Epub 2013 Apr 15. PubMed PMID: 23534816.

Boughman JA, Conneally PM, Nance WE. Population genetic studies of retinitis pigmentosa. Am J Hum Genet. 1980 Mar;32(2):223-35. PubMed PMID: 7386458; PubMed Central PMCID: PMC1686021.

Bowes C, Danciger M, Kozak CA, Farber DB. Isolation of a candidate cDNA for the gene causing retinal degeneration in the rd mouse. Proc Natl Acad Sci U S A. 1989 Dec;86(24):9722-6. Erratum in: Proc Natl Acad Sci U S A 1990 Feb;87(4):1625. PubMed PMID: 2481314; PubMed Central PMCID: PMC298573.

Bowler, Peter J. Evolution: The History of an Idea. Berkeley: University of California Press, 3rd revised edition, 2003. ISBN 0-520-23693-9.

Bowne SJ, Humphries MM, Sullivan LS, Kenna PF, Tam LC, Kiang AS, Campbell M, Weinstock GM, Koboldt DC, Ding L, *et al.* A dominant mutation in RPE65 identified by whole-exome sequencing causes retinitis pigmentosa with choroidal involvement. Eur J Hum Genet. 2011 Oct;19(10):1074-81. doi: 10.1038/ejhg.2011.86. Epub 2011 Jun 8. Erratum in: Eur J Hum Genet. 2011 Oct;19(10):1109. PubMed PMID: 21654732; PubMed Central PMCID: PMC3190249.

Bowne SJ, Sullivan LS, Gire AI, Birch DG, Hughbanks-Wheaton D, Heckenlively JR, Daiger SP. Mutations in the TOPORS gene cause 1% of autosomal dominant retinitis pigmentosa. Mol Vis. 2008 May 19;14:922-7. PubMed PMID: 18509552; PubMed Central PMCID: PMC2391085.

Bowne SJ, Sullivan LS, Koboldt DC, Ding L, Fulton R, Abbott RM, Sodergren EJ, Birch DG, Wheaton DH, Heckenlively JR, Liu Q, Pierce EA, Weinstock GM, Daiger SP. Identification of disease-causing mutations in autosomal dominant retinitis pigmentosa (adRP) using next-generation DNA sequencing. Invest Ophthalmol Vis Sci. 2011 Jan 25;52(1):494-503. doi:

10.1167/iovs.10-6180. PubMed PMID: 20861475; PubMed Central PMCID: PMC3053293.

Braslavsky I, Hebert B, Kartalov E, Quake SR. Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci U S A. 2003 Apr 1;100(7):3960-4. Epub 2003 Mar 21. PubMed PMID: 12651960; PubMed Central PMCID: PMC153030.

Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, Lee WL, Russ C, Lander ES, Nusbaum C, Jaffe DB. Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res. 2008 May;18(5):763-70. doi: 10.1101/gr.070227.107. Epub 2008 Jan 22. PubMed PMID: 18212088; PubMed Central PMCID: PMC2336812.

Brose MS, Rebbeck TR, Calzone KA, Stopfer JE, Nathanson KL, Weber BL. Cancer risk estimates for BRCA1 mutation carriers identified in a risk evaluation program. J Natl Cancer Inst. 2002 Sep 18;94(18):1365-72. PubMed PMID: 12237282.

Campeau PM, Foulkes WD, Tischkowitz MD. Hereditary breast cancer: new genetic developments, new therapeutic avenues. Hum Genet. 2008 Aug;124(1):31-42. doi: 10.1007/s00439-008-0529-1. Epub 2008 Jun 25. Review. PubMed PMID: 18575892.

Caruccio N. Preparation of next-generation sequencing libraries using Nextera™ technology: simultaneous DNA fragmentation and adaptor tagging by *in vitro* transposition. Methods Mol Biol. 2011;733:241-55. doi: 10.1007/978-1-61779-089-8_17. PubMed PMID: 21431775.

Chen S, Wang QL, Xu S, Liu I, Li LY, Wang Y, Zack DJ. Functional analysis of cone-rod homeobox (CRX) mutations associated with retinal dystrophy. Hum Mol Genet. 2002 Apr 15;11(8):873-84. PubMed PMID: 11971869.

Cheng H, Khanna H, Oh EC, Hicks D, Mitton KP, Swaroop A. Photoreceptor-specific nuclear receptor NR2E3 functions as a transcriptional activator in rod photoreceptors. Hum Mol Genet. 2004 Aug 1;13(15):1563-75. Epub 2004 Jun 9. PubMed PMID: 15190009.

Chillón M, Dörk T, Casals T, Giménez J, Fonknechten N, Will K, Ramos D, Nunes V, Estivill X. A novel donor splice site in intron 11 of the CFTR gene, created by mutation 1811+1.6kbA-->G, produces a new exon: high frequency in Spanish cystic fibrosis chromosomes and association with severe phenotype. Am J Hum Genet. 1995 Mar;56(3):623-9. PubMed PMID: 7534040; PubMed Central PMCID: PMC1801150.

Claus EB, Schildkraut JM, Thompson WD, Risch NJ. The genetic attributable risk of breast and ovarian cancer. Cancer. 1996 Jun 1;77(11):2318-24. PubMed PMID: 8635102.

Collin RW, Safieh C, Littink KW, Shalev SA, Garzozi HJ, Rizel L, Abbasi AH, Cremers FP, den Hollander AI, Klevering BJ, Ben-Yosef T. Mutations in C2ORF71 cause autosomal-recessive retinitis pigmentosa. Am J Hum Genet. 2010 May 14;86(5):783-8. doi: 10.1016/j.ajhg.2010.03.016. Epub 2010 Apr 15. PubMed PMID: 20398884; PubMed Central PMCID: PMC2869006.

Coppieters F, Leroy BP, Beysen D, Hellemans J, De Bosscher K, Haegeman G, Robberecht K, Wuyts W, Coucke PJ, De Baere E. Recurrent mutation in the first zinc finger of the orphan nuclear receptor NR2E3 causes autosomal dominant retinitis pigmentosa. Am J Hum Genet. 2007 Jul;81(1):147-57. Epub 2007 May

24. PubMed PMID: 17564971; PubMed Central PMCID: PMC1950922.

Daiger SP, Bowne SJ, Sullivan LS. Perspective on genes and mutations causing retinitis pigmentosa. Arch Ophthalmol. 2007 Feb;125(2):151-8. Review. PubMed PMID: 17296890; PubMed Central PMCID: PMC2580741.

Daiger SP, Sullivan LS, Gire AI, Birch DG, Heckenlively JR, Bowne SJ. Mutations in known genes account for 58% of autosomal dominant retinitis pigmentosa (adRP). Adv Exp Med Biol. 2008;613:203-9. doi: 10.1007/978-0-387-74904-4_23. PubMed PMID: 18188946; PubMed Central PMCID: PMC2582019.

Davidson AE, Millar ID, Urquhart JE, Burgess-Mullan R, Shweikh Y, Parry N, O'Sullivan J, Maher GJ, McKibbin M, Downes SM, Lotery AJ, Jacobson SG, Brown PD, Black GC, Manson FD. Missense mutations in a retinal pigment epithelium protein, bestrophin-1, cause retinitis pigmentosa. Am J Hum Genet. 2009 Nov;85(5):581-92. doi: 10.1016/j.ajhg.2009.09.015. Epub 2009 Oct 22. PubMed PMID: 19853238; PubMed Central PMCID: PMC2775838.

De Leeneer K, De Schrijver J, Clement L, Baetens M, Lefever S, De Keulenaer S, Van Criekinge W, Deforce D, Van Nieuwerburgh F, Bekaert S, Pattyn F, De Wilde B, Coucke P, Vandesompele J, Claes K, Hellemans J. Practical tools to implemente massive parallel pyrosequencing of PCR products in next generation molecular diagnostics. PLoS One. 2011a;6(9):e25531. doi: 10.1371/journal.pone.0025531. Epub 2011 Sep 30. PubMed PMID: 21980484; PubMed Central PMCID: PMC3184136.

De Leeneer K, Hellemans J, De Schrijver J, Baetens M, Poppe B, Van Criekinge W, De Paepe A, Coucke P, Claes K. Massive parallel amplicon sequencing of the breast cancer genes BRCA1 and BRCA2: opportunities, challenges, and limitations. Hum Mutat. 2011b Mar;32(3):335-44. doi: 10.1002/humu.21428. Epub 2011 Feb 8. PubMed PMID: 21305653.

de Sousa Dias M, Hernan I, Pascual B, Borràs E, Mañé B, Gamundi MJ, Carballo M. Detection of novel mutations that cause autosomal dominant retinitis pigmentosa in candidate genes by long-range PCR amplification and next-generation sequencing. Mol Vis. 2013;19:654-64. Epub 2013 Mar 21. PubMed PMID: 23559859; PubMed Central PMCID: PMC3611935.

den Hollander AI, Koenekoop RK, Yzer S, Lopez I, Arends ML, Voesenek KE, Zonneveld MN, Strom TM, Meitinger T, Brunner HG, Hoyng CB, van den Born LI, Rohrschneider K, Cremers FP. Mutations in the CEP290 (NPHP6) gene are a frequent cause of Leber congenital amaurosis. Am J Hum Genet. 2006 Sep;79(3):556-61. Epub 2006 Jul 11. PubMed PMID: 16909394; PubMed Central PMCID: PMC1559533.

Dvir L, Srour G, Abu-Ras R, Miller B, Shalev SA, Ben-Yosef T. Autosomal-recessive early-onset retinitis pigmentosa caused by a mutation in PDE6G, the gene encoding the gamma subunit of rod cGMP phosphodiesterase. Am J Hum Genet. 2010 Aug 13;87(2):258-64. doi: 10.1016/j.ajhg.2010.06.016. Epub 2010 Jul 22. PubMed PMID: 20655036; PubMed Central PMCID: PMC2917712.

Easton DF. How many more breast cancer predisposition genes are there? Breast Cancer Res. 1999;1(1):14-7. Epub 1999 Aug 23. PubMed PMID: 11250676; PubMed Central PMCID: PMC138504.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, *et al.* Real-time DNA sequencing from single polymerase molecules. Science. 2009 Jan 2;323(5910):133-8. doi: 10.1126/science.1162986. Epub 2008 Nov 20. PubMed PMID: 19023044.

Escher P, Gouras P, Roduit R, Tiab L, Bolay S, Delarive T, Chen S, Tsai CC, Hayashi M, Zernant J, Merriam JE, Mermod N, Allikmets R, Munier FL, Schorderet DF. Mutations in NR2E3 can cause dominant or recessive retinal degenerations in the same family. Hum Mutat. 2009 Mar;30(3):342-51. doi: 10.1002/humu.20858. PubMed PMID: 19006237; PubMed Central PMCID: PMC3658139.

Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998b Mar;8(3):186-94. PubMed PMID: 9521922.

Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 1998a Mar;8(3):175-85. PubMed PMID: 9521921.

Fahim AT, Daiger SP, Weleber RG. Retinitis Pigmentosa Overview. 2000 Aug 4 [Updated 2013 Mar 21]. In: Pagon RA, Adam MP, Bird TD, *et al.*, editors. GeneReviews™ [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2013. Available from: http://www.ncbi.nlm.nih.gov/books/NBK1417/

Fan JB, Chee MS, Gunderson KL. Highly parallel genomic assays. Nat Rev Genet. 2006 Aug;7(8):632-44. Review. PubMed PMID: 16847463.

Farrar GJ, Kenna P, Jordan SA, Kumar-Singh R, Humphries MM, Sharp EM, Sheils DM, Humphries P. A three-base-pair deletion in the peripherin-RDS gene in one form of retinitis pigmentosa. Nature. 1991 Dec 12;354(6353):478-80. PubMed PMID: 1749427.

Ferrari S, Di Iorio E, Barbaro V, Ponzin D, Sorrentino FS, Parmeggiani F. Retinitis pigmentosa: genes and disease mechanisms. Curr Genomics. 2011 Jun;12(4):238-49. doi: 10.2174/138920211795860107. PubMed PMID: 22131869; PubMed

Ferrone CR, Levine DA, Tang LH, Allen PJ, Jarnagin W, Brennan MF, Offit K, Robson ME. BRCA germline mutations in Jewish patients with pancreatic adenocarcinoma. J Clin Oncol. 2009 Jan 20;27(3):433-8. doi: 10.1200/JCO.2008.18.5546. Epub 2008 Dec 8. PubMed PMID: 19064968; PubMed Central PMCID: PMC3657622.

Finch A, Beiner M, Lubinski J, Lynch HT, Moller P, Rosen B, Murphy J, Ghadirian P, Friedman E, Foulkes WD, *et al.* Hereditary Ovarian Cancer Clinical Study Group. Salpingo-oophorectomy and the risk of ovarian, fallopian tube, and peritoneal cancers in women with a BRCA1 or BRCA2 Mutation. JAMA. 2006 Jul 12;296(2):185-92. PubMed PMID: 16835424.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, *et al.* Ensembl 2012. Nucleic Acids Res. 2012 Jan;40(Database issue):D84-90. doi: 10.1093/nar/gkr991. Epub 2011 Nov 15. PubMed PMID: 22086963; PubMed Central PMCID: PMC3245178.

Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods. 2010 Jun;7(6):461-5. doi: 10.1038/nmeth.1459. Epub 2010 May 9. PubMed PMID: 20453866; PubMed Central PMCID: PMC2879396.

Frank TS, Manley SA, Olopade OI, Cummings S, Garber JE, Bernhardt B, Antman K, Russo D, Wood ME, Mullineau L, *et al.* Sequence analysis of BRCA1 and BRCA2: correlation of mutations with family history and ovarian cancer risk. J Clin Oncol. 1998 Jul;16(7):2417-25. PubMed PMID: 9667259.

Freund CL, Gregory-Evans CY, Furukawa T, Papaioannou M, Looser J, Ploder L, Bellingham J, Ng D, Herbrick JA, Duncan A, Scherer SW, Tsui LC, Loutradis-Anagnostou A, Jacobson SG, Cepko CL, Bhattacharya SS, McInnes RR. Cone-rod dystrophy due to mutations in a novel photoreceptor-specific homeobox gene (CRX) essential for maintenance of the photoreceptor. Cell. 1997 Nov 14;91(4):543-53. PubMed PMID: 9390563.

Fry, M. and Usdin, K. Human Nucleotide Expansion Disorders. 1st ed. Vol. 19. Berlin: Springer, 2006. Print. Nucleic Acids and Molecular Biology.

Gamundi MJ, Hernan I, Muntanyola M, Maseras M, López-Romero P, Alvarez R, Dopazo A, Borrego S, Carballo M. Transcriptional expression of cis-acting and trans-acting splicing mutations cause autosomal dominant retinitis pigmentosa. Hum Mutat. 2008 Jun;29(6):869-78. doi: 10.1002/humu.20747. PubMed PMID: 18412284.

Gamundi MJ, Hernan I, Muntanyola M, Trujillo MJ, García-Sandoval B, Ayuso C, Baiget M, Carballo M. High prevalence of mutations in peripherin/RDS in autosomal dominant macular dystrophies in a Spanish population. Mol Vis. 2007 Jun 28;13:1031-7. PubMed PMID: 17653047; PubMed Central PMCID: PMC2776544.

Gilles A, Meglécz E, Pech N, Ferreira S, Malausa T, Martin JF. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics. 2011 May 19;12:245. doi: 10.1186/1471-2164-12-245. PubMed PMID: 21592414; PubMed Central PMCID: PMC3116506.

Gire AI, Sullivan LS, Bowne SJ, Birch DG, Hughbanks-Wheaton D, Heckenlively JR, Daiger SP. The Gly56Arg mutation in NR2E3 accounts for 1-2% of autosomal dominant retinitis pigmentosa. Mol Vis. 2007 Oct 17;13:1970-5. PubMed PMID: 17982421.

Glöckle N, Kohl S, Mohr J, Scheurenbrand T, Sprecher A, Weisschuh N, Bernd A, Rudolph G, Schubach M, Poloschek C, Zrenner E, Biskup S, Berger W, Wissinger B, Neidhardt J. Panel-based next generation sequencing as a reliable and eficiente technique to detect mutations in unselected patients with retinal dystrophies. Eur J Hum Genet. 2013 Apr 17. doi: 10.1038/ejhg.2013.72. [Epub ahead of print] PubMed PMID: 23591405.

Gregory-Evans K, Kelsell RE, Gregory-Evans CY, Downes SM, Fitzke FW, Holder GE, Simunovic M, Mollon JD, Taylor R, Hunt DM, Bird AC, Moore AT. Autosomal dominant cone-rod retinal dystrophy (CORD6) from heterozygous mutation of GUCY2D, which encodes retinal guanylate cyclase. Ophthalmology. 2000 Jan;107(1):55-61. PubMed PMID: 10647719.

Hahn LB, Berson EL, Dryja TP. Evaluation of the gene encoding the gamma subunit of rod phosphodiesterase in retinitis pigmentosa. Invest Ophthalmol Vis Sci. 1994 Mar;35(3):1077-82. PubMed PMID: 8125719.

Haim M. Epidemiology of retinitis pigmentosa in Denmark. Acta Ophthalmol Scand Suppl. 2002;(233):1-34. PubMed PMID: 11921605.

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC. Linkage of early-onset familial

breast cancer to chromosome 17q21. Science. 1990 Dec 21;250(4988):1684-9. PubMed PMID: 2270482.

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 2009;10(3):R32. doi: 10.1186/gb-2009-10-3-r32. Epub 2009 Mar 27. PubMed PMID: 19327155; PubMed Central PMCID: PMC2691003.

Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, *et al.* Single-molecule DNA sequencing of a viral genome. Science. 2008 Apr 4;320(5872):106-9. doi: 10.1126/science.1150427. PubMed PMID: 18388294.

Hartong DT, Berson EL, Dryja TP. Retinitis pigmentosa. Lancet. 2006 Nov 18;368(9549):1795-809. Review. PubMed PMID: 17113430.

Hernan I, Borràs E, de Sousa Dias M, Gamundi MJ, Mañé B, Llort G, Agúndez JA, Blanca M, Carballo M. Detection of genomic variations in BRCA1 and BRCA2 genes by long-range PCR and next-generation sequencing. J Mol Diagn. 2012 May-Jun;14(3):286-93. doi: 10.1016/j.jmoldx.2012.01.013. Epub 2012 Mar 16. PubMed PMID: 22426013.

Hernan I, Gamundi MJ, Borràs E, Maseras M, García-Sandoval B, Blanco-Kelly F, Ayuso C, Carballo M. Novel p.M96T variant of NRL and shRNA-based suppression and replacement of NRL mutants associated with autosomal dominant retinitis pigmentosa. Clin Genet. 2011 Nov;82(5):446-52. doi: 10.1111/j.1399-0004.2011.01796.x. Epub 2011 Nov 2. PubMed PMID: 21981118.

Hopfner KP, Eichinger A, Engh RA, Laue F, Ankenbauer W, Huber R, Angerer B. Crystal structure of a thermostable type B DNA polymerase from Thermococcus gorgonarius. Proc Natl Acad Sci U S A. 1999 Mar 30;96(7):3600-5. PubMed PMID: 10097083; PubMed Central PMCID: PMC22340.

Houdayer C, Dehainault C, Mattler C, Michaux D, Caux-Moncoutier V, Pagès-Berhouet S, d'Enghien CD, Laugé A, Castera L, Gauthier-Villars M, Stoppa-Lyonnet D. Evaluation of *in silico* splice tools for decision-making in molecular diagnosis. Hum Mutat. 2008 Jul;29(7):975-82. doi: 10.1002/humu.20765. PubMed PMID: 18449911.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 2007;8(7):R143. PubMed PMID: 17659080; PubMed Central PMCID: PMC2323236.

Jay M. On the heredity of retinitis pigmentosa. Br J Ophthalmol. 1982 Jul;66(7):405-16. PubMed PMID: 7093178; PubMed Central PMCID: PMC1039814.

Johnson S, Halford S, Morris AG, Patel RJ, Wilkie SE, Hardcastle AJ, Moore AT, Zhang K, Hunt DM. Genomic organisation and alternative splicing of human RIM1, a gene implicated in autosomal dominant cone-rod dystrophy (CORD7). Genomics. 2003 Mar;81(3):304-14. PubMed PMID: 12659814.

Jordan SA, Farrar GJ, Kenna P, Humphries P. Polymorphic variation within "conserved" sequences at the 3' end of the human RDS gene which results in amino acid substitutions. Hum Mutat. 1992;1(3):240-7. PubMed PMID: 1301931.

Kajiwara K, Hahn LB, Mukai S, Travis GH, Berson EL, Dryja TP. Mutations in the human retinal degeneration slow gene in autosomal dominant retinitis pigmentosa.

Nature. 1991 Dec 12;354(6353):480-3. PubMed PMID: 1684223.

Kennan A, Aherne A, Humphries P. Light in retinitis pigmentosa. Trends Genet. 2005 Feb;21(2):103-10. Review. PubMed PMID: 15661356.

King MC, Marks JH, Mandell JB; New York Breast Cancer Study Group. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. Science. 2003 Oct 24;302(5645):643-6. PubMed PMID: 14576434.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009 Sep 1;25(17):2283-5. doi: 10.1093/bioinformatics/btp373. Epub 2009 Jun 19. PubMed PMID: 19542151; PubMed Central PMCID: PMC2734323.

Korlach J, Bibillo A, Wegener J, Peluso P, Pham TT, Park I, Clark S, Otto GA, Turner SW. Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides. Nucleosides Nucleotides Nucleic Acids. 2008a Sep;27(9):1072-83. doi: 10.1080/15257770802260741. PubMed PMID: 18711669; PubMed Central PMCID: PMC2582155.

Korlach J, Marks PJ, Cicero RL, Gray JJ, Murphy DL, Roitman DB, Pham TT, Otto GA, Foquet M, Turner SW. Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. Proc Natl Acad Sci U S A. 2008 Jan 29;105(4):1176-81. doi: 10.1073/pnas.0710982105. Epub 2008b Jan 23. PubMed PMID: 18216253; PubMed Central PMCID: PMC2234111.

Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-81. doi: 10.1038/nprot.2009.86. Epub 2009 Jun 25. PubMed PMID: 19561590.

Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. Zero-mode waveguides for single-molecule analysis at high concentrations. Science. 2003 Jan 31;299(5607):682-6. PubMed PMID: 12560545.

Levy-Lahad E, Friedman E. Cancer risks among BRCA1 and BRCA2 mutation carriers. Br J Cancer. 2007 Jan 15;96(1):11-5. Review. PubMed PMID: 17213823; PubMed Central PMCID: PMC2360226.

Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009 Nov 1;25(21):2744-50. doi: 10.1093/bioinformatics/btp528. Epub 2009 Sep 3. PubMed PMID: 19734154; PubMed Central PMCID: PMC3140805.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8. PubMed PMID: 19505943; PubMed Central PMCID: PMC2723002.

Liu Q, Lyubarsky A, Skalet JH, Pugh EN Jr, Pierce EA. RP1 is required for the correct stacking of outer segment discs. Invest Ophthalmol Vis Sci. 2003 Oct;44(10):4171-83. PubMed PMID: 14507858; PubMed Central PMCID: PMC1904498.

Liu Q, Zuo J, Pierce EA. The retinitis pigmentosa 1 protein is a photoreceptor microtubule-associated protein. J Neurosci. 2004 Jul 21;24(29):6427-36. PubMed PMID: 15269252; PubMed Central PMCID: PMC1904502.

Liu S, Rauhut R, Vornlocher HP, Lührmann R. The network of protein-protein interactions within the human U4/U6.U5 tri-snRNP. RNA. 2006 Jul;12(7):1418-30. Epub 2006 May 24. PubMed PMID: 16723661; PubMed Central PMCID: PMC1484429.

Macevicz, S. C. DNA sequencing by parallel oligonucleotide extensions. US Patent 5,969,119 (1995).

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J,Braverman MS, Chen YJ, Chen Z, *et al.* Genome sequencing in microfabricated high-densitypicolitre reactors. Nature. 2005 Sep 15;437(7057):376-80. Epub 2005 Jul 31.Erratum in: Nature. 2006 May 4;441(7089):120. Ho, Chun He [corrected to Ho, Chun Heen]. PubMed PMID: 16056220; PubMed Central PMCID: PMC1464427.

Mattocks CJ, Morris MA, Matthijs G, Swinnen E, Corveleyn A, Dequeker E, Müller CR, Pratt V, Wallace A; EuroGentest Validation Group. A standardized framework for the validation and verification of clinical molecular genetic tests. Eur J Hum Genet. 2010 Dec;18(12):1276-88. doi: 10.1038/ejhg.2010.101. Epub 2010 Jul 28. PubMed PMID: 20664632; PubMed Central PMCID: PMC3002854.

Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A. 1977 Feb;74(2):560-4. PubMed PMID: 265521; PubMed Central PMCID: PMC392330.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297-303. doi: 10.1101/gr.107524.110. Epub 2010 Jul 19. PubMed PMID: 20644199; PubMed Central PMCID: PMC2928508.

McKernan, K., Blanchard, A., Kotler, L. & Costa, G. Reagents, methods, and libraries forbead-based sequencing. US Patent Application 11/345,979 (2005).

Mendel, G., 1866, Versuche über Pflanzen-Hybriden. Verh. Naturforsch. Ver. Brünn 4: 3–47 (in English in 1901, J. R. Hortic. Soc. 26: 1–32)

Metzker ML. Emerging technologies in DNA sequencing. Genome Res. 2005 Dec;15(12):1767-76. Review. PubMed PMID: 16339375.

Metzker ML. Sequencing in real time. Nat Biotechnol. 2009 Feb;27(2):150-1. doi: 10.1038/nbt0209-150. PubMed PMID: 19204695.

Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46. doi: 10.1038/nrg2626. Epub 2009 Dec 8. Review. PubMed PMID: 19997069.

Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 2012 May;30(5):434-9. doi: 10.1038/nbt.2198. Erratum in: Nat Biotechnol. 2012 Jun;30(6):562. PubMed PMID: 22522955.

Millá E, Maseras M, Martínez-Gimeno M, Gamundi MJ, Assaf H, Esmerado C, Carballo M; Grupo Multicéntrico Español de Retinosis Pigmentaria. [Genetic and molecular characterization of 148 patients with autosomal dominant retinitis pigmentosa

(ADRP)]. Arch Soc Esp Oftalmol. 2002 Sep;77(9):481-4. Spanish. PubMed PMID: 12221539.

Mir KU, Qi H, Salata O, Scozzafava G. Sequencing by Cyclic Ligation and Cleavage (CycLiC) directly on a microarray captured template. Nucleic Acids Res. 2009 Jan;37(1):e5. doi: 10.1093/nar/gkn906. Epub 2008 Nov 16. PubMed PMID: 19015154; PubMed Central PMCID: PMC2615607.

Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA *in vitro*: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol. 1986;51 Pt 1:263-73. PubMed PMID: 3472723.

Murray AR, Fliesler SJ, Al-Ubaidi MR. Rhodopsin: the functional significance of asn-linked glycosylation and other post-translational modifications. Ophthalmic Genet. 2009 Sep;30(3):109-20. doi: 10.1080/13816810902962405. Review. Erratum in: Ophthalmic Genet. 2010 Mar;31(1):52. PubMed PMID: 19941415; PubMed Central PMCID: PMC2881540.

Myllyjangas, S., J. Buenrostro, and PJ Hanlee. Chapter 2 - Overview of Sequencing Technology Platforms. Bioinformatics for High Throughput Sequencing. New York, NY: Springer, 2012. 11-25. doi.org/10.1007/978-1-4614-0782-9_2

Neveling K, Collin RW, Gilissen C, van Huet RA, Visser L, Kwint MP, Gijsen SJ, Zonneveld MN, Wieskamp N, de Ligt J, Siemiatkowska AM, *et al.* Next-generation genetic testing for retinitis pigmentosa. Hum Mutat. 2012 Jun;33(6):963-72. doi: 10.1002/humu.22045. Epub 2012 Mar 19. Erratum in: Hum Mutat. 2013 Aug;34(8):1181. PubMed PMID: 22334370; PubMed Central PMCID: PMC3490376.

Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. Genome Res. 2002 Mar;12(3):436-46. PubMed PMID: 11875032; PubMed Central PMCID: PMC155281.

Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001 May;11(5):863-74. PubMed PMID: 11337480; PubMed Central PMCID: PMC311071.

Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet. 2006;7:61-80. Review. PubMed PMID: 16824020.

Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;31(13):3812-4. PubMed PMID: 12824425; PubMed Central PMCID: PMC168916.

Nichols BE, Sheffield VC, Vandenburgh K, Drack AV, Kimura AE, Stone EM. Butterfly-shaped pigment dystrophy of the fovea caused by a point mutation in codon 167 of the RDS gene. Nat Genet. 1993 Mar;3(3):202-7. PubMed PMID: 8485574.

Nishiguchi KM, Friedman JS, Sandberg MA, Swaroop A, Berson EL, Dryja TP. Recessive NRL mutations in patients with clumped pigmentary retinal degeneration and relative preservation of blue cone function. Proc Natl Acad Sci U S A. 2004 Dec 21;101(51):17819-24. Epub 2004 Dec 9. PubMed PMID: 15591106; PubMed Central PMCID: PMC535407.

Nishimura DY, Baye LM, Perveen R, Searby CC, Avila-Fernandez A, Pereiro I, Ayuso C, Valverde D, Bishop PN, Manson FD, Urquhart J, Stone EM, Slusarski DC, Black GC, Sheffield VC. Discovery and functional analysis of a retinitis pigmentosa gene, C2ORF71. Am J Hum Genet. 2010 May 14;86(5):686-95. doi: 10.1016/j.ajhg.2010.03.005. Epub 2010 Apr 15. PubMed PMID: 20398886; PubMed Central PMCID: PMC2868997.

O'Sullivan J, Mullaney BG, Bhaskar SS, Dickerson JE, Hall G, O'Grady A, Webster A, Ramsden SC, Black GC. A paradigm shift in the delivery of services for diagnosis of inherited retinal disease. J Med Genet. 2012 May;49(5):322-6. doi: 10.1136/jmedgenet-2012-100847. PubMed PMID: 22581970.

Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B, Milos PM. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. Cell. 2010 Dec 10;143(6):1018-29. doi: 10.1016/j.cell.2010.11.020. PubMed PMID: 21145465; PubMed Central PMCID: PMC3022516.

Pal T, Permuth-Wey J, Betts JA, Krischer JP, Fiorica J, Arango H, LaPolla J, Hoffman M, Martino MA, Wakeley K, Wilbanks G, Nicosia S, Cantor A, Sutphen R. BRCA1 and BRCA2 mutations account for a large proportion of ovarian carcinoma cases. Cancer. 2005 Dec 15;104(12):2807-16. PubMed PMID: 16284991.

Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet. 2011 Nov;52(4):413-35. doi: 10.1007/s13353-011-0057-x. Epub 2011 Jun 23. Review. PubMed PMID: 21698376; PubMed Central PMCID: PMC3189340.

Payne AM, Downes SM, Bessant DA, Taylor R, Holder GE, Warren MJ, Bird AC, Bhattacharya SS. A mutation in guanylate cyclase activator 1A (GUCA1A) in na autosomal dominant cone dystrophy pedigree mapping to a new locus on chromosome 6p21.1. Hum Mol Genet. 1998 Feb;7(2):273-7. PubMed PMID: 9425234.

Piri N, Gao YQ, Danciger M, Mendoza E, Fishman GA, Farber DB. A substitution of G to C in the cone cGMP-phosphodiesterase gamma subunit gene found in a distinctive form of cone dystrophy. Ophthalmology. 2005 Jan;112(1):159-66. PubMed PMID: 15629837.

Pruthi S, Gostout BS, Lindor NM. Identification and Management of Women With BRCA Mutations or Hereditary Predisposition for Breast and Ovarian Cancer. Mayo Clin Proc. 2010 Dec;85(12):1111-20. doi: 10.4065/mcp.2010.0414. Review. PubMed PMID: 21123638; PubMed Central PMCID: PMC2996153.

Qu W, Shen Z, Zhao D, Yang Y, Zhang C. MFEprimer: multiple factor evaluation of the specificity of PCR primers. Bioinformatics. 2009 Jan 15;25(2):276-8. doi: 10.1093/bioinformatics/btn614. Epub 2008 Nov 27. PubMed PMID: 19038987.

Qu W, Zhou Y, Zhang Y, Lu Y, Wang X, Zhao D, Yang Y, Zhang C. MFEprimer-2.0: a fast thermodynamics-based program for checking PCR primer specificity. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W205-8. doi: 10.1093/nar/gks552. Epub 2012 Jun 11. PubMed PMID: 22689644; PubMed Central PMCID: PMC3394324.

Raman D, Osawa S, Weiss ER. Binding of arrestin to cytoplasmic loop mutants of bovine rhodopsin. Biochemistry. 1999 Apr 20;38(16):5117-23. PubMed PMID: 10213616.

Rio Frio T, McGee TL, Wade NM, Iseli C, Beckmann JS, Berson EL, Rivolta C. A single-base substitution within an intronic repetitive element causes dominant retinitis pigmentosa with reduced penetrance. Hum Mutat. 2009 Sep;30(9):1340-7. doi: 10.1002/humu.21071. PubMed PMID: 19618371; PubMed Central PMCID: PMC2753193.

Risch HA, McLaughlin JR, Cole DE, Rosen B, Bradley L, Kwan E, Jack E, Vesprini DJ, Kuperstein G, Abrahamson JL, Fan I, Wong B, Narod SA. Prevalence and penetrance of germline BRCA1 and

BRCA2 mutations in a population series of 649 women with ovarian cancer. Am J Hum Genet. 2001 Mar;68(3):700-10. Epub 2001 Feb 15. PubMed PMID: 11179017; PubMed Central PMCID: PMC1274482.

Rivolta C, McGee TL, Rio Frio T, Jensen RV, Berson EL, Dryja TP. Variation in retinitis pigmentosa-11 (PRPF31 or RP11) gene expression between symptomatic and asymptomatic patients with dominant RP11 mutations. Hum Mutat. 2006 Jul;27(7):644-53. PubMed PMID: 16708387.

Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P. Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem. 1996 Nov 1;242(1):84-9. PubMed PMID: 8923969.

Ronaghi M, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate. Science. 1998 Jul 17;281(5375):363, 365. PubMed PMID: 9705713.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, *et al.* An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011 Jul 20;475(7356):348-52. doi: 10.1038/nature10242. PubMed PMID: 21776081.

Rusk, N. Cheap Third-generation Sequencing. Nature Methods 6 (2009): 244-245

Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975 May 25;94(3):441-8. PubMed PMID: 1100841.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977 Dec;74(12):5463-7. PubMed PMID: 271968; PubMed Central PMCID: PMC431765.

Sanger F. The Croonian Lecture, 1975. Nucleotide sequences in DNA. Proc R Soc Lond B Biol Sci. 1975 Dec 2;191(1104):317-33. Review. PubMed PMID: 2920.

Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010 Oct 15;19(R2):R227-40. doi: 10.1093/hmg/ddq416. Epub 2010 Sep 21. Review. Erratum in: Hum Mol Genet. 2011 Feb 15;20(4):853. PubMed PMID: 20858600.

Shanks ME, Downes SM, Copley RR, Lise S, Broxholme J, Hudspith KA, Kwasniewska A, Davies WI, Hankins MW, Packham ER, *et al.* Next-generation sequencing (NGS) as a diagnostic tool for retinal degeneration reveals a much higher detection rate in early-onset disease. Eur J Hum Genet. 2013 Mar;21(3):274-80. doi: 10.1038/ejhg.2012.172. Epub 2012 Sep 12. PubMed PMID: 22968130; PubMed Central PMCID: PMC3573204.

Simpson DA, Clark GR, Alexander S, Silvestri G, Willoughby CE. Molecular diagnosis for heterogeneous genetic diseases with targeted high-throughput DNA sequencing applied to retinitis pigmentosa. J Med Genet. 2011 Mar;48(3):145-51. doi: 10.1136/jmg.2010.083568. Epub 2010 Dec 8. PubMed PMID: 21147909.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE. Fluorescence detection in automated DNA sequence analysis. Nature. 1986 Jun 12-18;321(6071):674-9. PubMed PMID: 3713851.

Sohocki MM, Bowne SJ, Sullivan LS, Blackshaw S, Cepko CL, Payne AM, Bhattacharya SS, Khaliq S, Qasim Mehdi S, Birch DG, Harrison WR, Elder FF, Heckenlively JR, Daiger SP. Mutations in a new photoreceptor-pineal gene on 17p cause Leber

congenital amaurosis. Nat Genet. 2000 Jan;24(1):79-83. PubMed PMID: 10615133; PubMed Central PMCID: PMC2581448.

Stamhuis IH, Meijer OG, Zevenhuizen EJ. Hugo de Vries on heredity, 1889-1903. Statistics, Mendelian laws, pangenes, mutations. Isis. 1999 Jun;90(2):238-67. PubMed PMID: 10439561.

Strachan T, Read AP. Human Molecular Genetics. 2nd edition. New York: Wiley-Liss; 1999. Chapter 16, Molecular pathology.

Strachan T, Read AP. Human Molecular Genetics. 2nd edition. New York: Wiley-Liss; 1999. Chapter 18, Cancer genetics.

Strachan T, Read AP. Human Molecular Genetics. 2nd edition. New York: Wiley-Liss; 1999. Chapter 3, Genes in pedigrees.

Sullivan LS, Bowne SJ, Birch DG, Hughbanks-Wheaton D, Heckenlively JR, Lewis RA, Garcia CA, Ruiz RS, Blanton SH, Northrup H, Gire AI, Seaman R, Duzkale H, Spellicy CJ, Zhu J, Shankar SP, Daiger SP. Prevalence of disease-causing mutations in families with autosomal dominant retinitis pigmentosa: a screen of known genes in 200 families. Invest Ophthalmol Vis Sci. 2006 Jul;47(7):3052-64. PubMed PMID: 16799052; PubMed Central PMCID: PMC2585061.

Sullivan LS, Heckenlively JR, Bowne SJ, Zuo J, Hide WA, Gal A, Denton M, Inglehearn CF, Blanton SH, Daiger SP. Mutations in a novel retina-specific gene cause autosomal dominant retinitis pigmentosa. Nat Genet. 1999 Jul;22(3):255-9. PubMed PMID: 10391212; PubMed Central PMCID: PMC2582380.

Tai YC, Domchek S, Parmigiani G, Chen S. Breast cancer risk among male BRCA1 and BRCA2 mutation carriers. J Natl Cancer Inst. 2007 Dec 5;99(23):1811-4. Epub 2007 Nov 27. PubMed PMID: 18042939; PubMed Central PMCID: PMC2267289.

Tanackovic G, Ransijn A, Ayuso C, Harper S, Berson EL, Rivolta C. A missense mutation in PRPF6 causes impairment of pre-mRNA splicing and autosomal-dominant retinitis pigmentosa. Am J Hum Genet. 2011 May 13;88(5):643-9. doi: 10.1016/j.ajhg.2011.04.008. Epub 2011 May 5. PubMed PMID: 21549338; PubMed Central PMCID: PMC3146730.

Tomkinson AE, Vijayakumar S, Pascal JM, Ellenberger T. DNA ligases: structure, reaction mechanism, and function. Chem Rev. 2006 Feb;106(2):687-99. Review. PubMed PMID: 16464020.

Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, Sidow A, Fire A, Johnson SM. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res. 2008 Jul;18(7):1051-63. doi: 10.1101/gr.076463.108. Epub 2008 May 13. PubMed PMID: 18477713; PubMed Central PMCID: PMC2493394.

Varley KE, Mitra RD. Nested Patch PCR enables highly multiplexed mutation discovery in candidate genes. Genome Res. 2008 Nov;18(11):1844-50. doi: 10.1101/gr.078204.108. Epub 2008 Oct 10. PubMed PMID: 18849522; PubMed Central PMCID: PMC2577855.

Veltel S, Gasper R, Eisenacher E, Wittinghofer A. The retinitis pigmentosa 2 gene product is a GTPase-activating protein for Arf-like 3. Nat Struct Mol Biol. 2008 Apr;15(4):373-80. doi: 10.1038/nsmb.1396. Epub 2008 Mar 23. PubMed PMID: 18376416.

Walsh T, Lee MK, Casadei S, Thornton AM, Stray SM, Pennil C, Nord AS, Mandell JB, Swisher EM, King MC. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. Proc Natl Acad Sci U S A. 2010 Jul 13;107(28):12629-33. doi: 10.1073/pnas.1007983107. Epub 2010 Jun 28. PubMed PMID: 20616022; PubMed Central PMCID: PMC2906584.

Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature. 1953 Apr 25;171(4356):737-8. PubMed PMID: 13054692.

Wells J, Wroblewski J, Keen J, Inglehearn C, Jubb C, Eckstein A, Jay M, Arden G, Bhattacharya S, Fitzke F, *et al.* Mutations in the human retinal degeneration slow (RDS) gene can cause either retinitis pigmentosa or macular dystrophy. Nat Genet. 1993 Mar;3(3):213-8. PubMed PMID: 8485576.

Whiteford N, Skelly T, Curtis C, Ritchie ME, Löhr A, Zaranek AW, Abnizova I, Brown C. Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics. 2009 Sep 1;25(17):2194-9. doi: 10.1093/bioinformatics/btp383. Epub 2009 Jun 23. PubMed PMID: 19549630; PubMed Central PMCID: PMC2734321.

Wimmer K, Eckart M, Rehder H, Fonatsch C. Illegitimate splicing of the NF1 gene in healthy individuals mimics mutation-induced splicing alterations in NF1 patients. Hum Genet. 2000 Mar;106(3):311-3. PubMed PMID: 10798360.

Wright AF, Chakarova CF, Abd El-Aziz MM, Bhattacharya SS. Photoreceptor degeneration: genetic and mechanistic dissection of a complex trait. Nat Ver Genet. 2010 Apr;11(4):273-84. doi: 10.1038/nrg2717. Review. PubMed PMID: 20212494.

Internet web sites

Cátedra Bidons Egara: Retinitis Pigmentaria. Available at: http://retinosis.umh.es/retinosis.html/
Accessed [December 1, 2013].

National Cancer Institute: Understanding Cancer Series. Available at: http://www.cancer.gov/cancertopics/understandingcancer/geneticbackground/AllPages/
Accessed [November 28, 2013].

Orphanet: Retinitis Pigmentosa. Available at: http://www.orpha.net/consor/cgi-bin/Disease_Search.php?lng=EN&data_id=659&Disease_Disease_Search_diseaseGroup=retinitis-pigmentosa&Disease_Disease_Search_diseaseType=Pat&Disease%28s%29/group%20of%20diseases=Retinitis-pigmentosa&title=Retinitis-pigmentosa&search=Disease_Search_Simple/
Accessed [November 29, 2013].

Retnet: Summaries of Genes and Loci Causing Retinal Diseases. Available at: https://sph.uth.edu/retnet/sum-dis.htm#B-diseases/
Accessed [November 29, 2013].

Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: http://www.genome.gov/sequencingcosts/
Accessed [November 28, 2013].

Frequently Consulted Online Databases

Ensemble (*!e*) Available at: http://www.ensembl.org/ Accessed [2010-2013]

National Center for Biotechnology Information (NCBI) Available at: http://www.ncbi.nlm.nih.gov/ Accessed [2010-2013].

Genome Browser (UCSC) Available at: http://www.hgmd.org/ Accessed [2010-2013].

Human Gene Mutation Database (HGMD) Available at: http://www.hgmd.org/ Accessed [2010-2013].

Retinal Information Network (RetNet) Available at: https://sph.uth.edu/retnet/ Accessed [2010-2013].

Rede Española de Retina (EsRetNet) Available at: http://www.esretnet.org/ Accessed [2010-2013].

Scientific contributions

i. Projects

Participation in public funded projects

- Research project FIS09/90754 “Assessment on the current and future of the molecular analysis of autosomal dominant retinal dystrophies: DNA Capture Technology and Next Generation Sequencing”, a project financed by the Fund for Health Research (FIS) from the Ministry of Health and Consumption of Spain.
- Research project FIS09/01271 “The genetics of inherited retinopathies: Characterization of new genes, mutant expression and pathogenic mechanisms associated with autosomal dominant retinitis pigmentosa” a project financed by the Fund for Health Research (FIS) from the Ministry of Health and Consumption of Spain.
- Research Project RD07/0064/2005 from the Spanish Research Network on Adverse Reactions to Allergens and Drugs (RIRAAF: Red de Investigacióńde Reacciones Adversas a Alérgenos y Fármacos) of the Carlos III Health Institute.

Participation in research contracts (private funded projects)

- Research agreement with ROCHE Applied Sciences for the development of new and more cost-efficient approaches for Next Generation Sequencing and gene panels design for several hereditary disorders.

ii. Publications

- **Miguel de Sousa Dias**, Imma Hernan, Beatriz Pascual, Emma Borràs, Begoña Mañé, Maria José Gamundi, Miguel Carballo. Detection of novel mutations that cause autosomal dominant retinitis pigmentosa in candidate genes by long-range PCR amplification and next-generation sequencing. Molecular vision (2013); 19:654-64. • 2.20 Impact Factor.

- Emma Borràs*, **Miguel de Sousa Dias***, Imma Hernan, Beatriz Pascual, Begoña Mañé, Maria José Gamundi, Barbara Delás, Miguel Carballo Detection of novel genetic variation in autosomal dominant retinitis pigmentosa. Clinical Genetics (2013); 84:441-52. • 3.13 Impact Factor. *These authors contributed equally and both should be considered as first authors.

- Elena Millá, Begoña Mañé; Susana Duch, Imma Hernan, Emma Borràs, Ester Planas **Miguel de Sousa Dias**, Miguel Carballo and María José Gamundi. Survey of familial glaucoma shows a high incidence of CYP1B1 mutations in non-consanguineous congenital forms in a Spanish population. Molecular vision (2013); 19:1707-22. • 2.20 Impact Factor.

- Imma Hernan*, Emma Borràs*, **Miguel de Sousa Dias**, María José Gamundi, Begoña Mañé, Gemma Llort, José A G Agúndez, Miguel Blanca, Miguel Carballo. Detection of genomic variations in BRCA1 and BRCA2 genes by long-range PCR and next-generation sequencing. The Journal of molecular diagnostics: JMD (2012); 14(3):286-93. • 3.48 Impact Factor. *These authors contributed equally and both should be considered as first authors.

- Emma Borràs, Emma Dotor, Angels Arcusa, Maria J Gamundi, Imma Hernan, **Miguel de Sousa Dias**, Begoña Mañé, José A G Agúndez, Miguel Blanca, Miguel Carballo. High-resolution melting analysis of the common c.1905+1G>A mutation causing dihydropyrimidine dehydrogenase deficiency and lethal 5-fluorouracil toxicity. Frontiers in genetics. (2012); 3:312.

- Emma Borràs, Ismael Jurado, Imma Hernan, María José Gamundi, **Miguel de Sousa Dias**, Isabel Martí, Begoña Mañé, Angels Arcusa, José A G Agúndez, Miguel Blanca, Miguel Carballo. Clinical pharmacogenomic testing of KRAS, BRAF and EGFR mutations by high resolution melting analysis and ultra-deep pyrosequencing. BMC Cancer (2011); 11:406. • 3.01 Impact Factor.

iii. Conferences

- **de Sousa Dias, M**; Hernan, I; Borràs, E; Gamundi, MJ; Pascual, B; Mañé, B; Carballo, M. Massive Sequencing of DNA in the Routine Clinical Practice: - in: XXVII Nacional Congress of the Spanish Association of Human Genetics, Madrid, Spain (2013)

Appendices

A. Detection of New Candidate Genes for adRP by Massive Sequencing

Appendix A.1: Selected 448 candidate genes for 385k array.

Gene	refSeq	Description
ABAT	NM_020686	4-aminobutyrate aminotransferase precursor
ABCA4	NM_000350	ATP-binding cassette, sub-family A member 4
ABCA7	NM_019112	ATP-binding cassette, sub-family A, member 7
ABCC5	NM_005688	ATP-binding cassette, sub-family C, member 5
ABCC6	NM_001171	ATP-binding cassette, sub-family C, member 6
ABLIM1	NM_002313	actin-binding LIM protein 1 isoform a
ADAM9	NM_003816	ADAM metallopeptidase domain 9 isoform 1
ADD1	NM_014189	adducin 1 (alpha) isoform b
ADFP	NM_001122	adipose differentiation-related protein
ADH4	NM_000670	alcohol dehydrogenase 4
AGPAT3	NM_020132	1-acylglycerol-3-phosphate O-acyltransferase 3
AIM2	NM_004833	absent in melanoma 2
ALDH1A2	NM_003888	aldehyde dehydrogenase 1A2 isoform 1
ALDH1A3	NM_000693	aldehyde dehydrogenase 1A3
ALDOC	NM_005165	fructose-bisphosphate aldolase C
AMY2B	NM_020978	amylase, pancreatic, alpha-2B precursor
AOC2	NM_009590	amine oxidase, copper containing 2 isoform b
ARSI	NM_001012301	arylsulfatase family, member I
ASCL1	NM_004316	achaete-scute complex homolog 1
ATF4	NM_001675	activating transcription factor 4
ATOH7	NM_145178	atonal homolog 7
ATP6V0C	NM_001694	ATPase, H+ transporting, lysosomal, V0 subunit
BEST1	NM_004183	bestrophin 1 isoform 1
BEST2	NM_017682	vitelliform macular dystrophy 2-like 1

BEST4	NM_153274	bestrophin 4
BHLHE22	NM_152414	basic helix-loop-helix domain containing, class
BHLHE23	NM_080606	basic helix-loop-helix domain containing, class
BHLHE41	NM_030762	basic helix-loop-helix domain containing, class
BSG	NM_001728	basigin isoform 1 precursor
BUD31	NM_003910	G10 protein
BZW2	NM_014038	basic leucine zipper and W2 domains 2
C16orf92	NM_001109659.1	chromosome 16 open reading frame 92
C17orf75	NM_022344	hypothetical protein LOC64149
C20orf39	NM_024893	hypothetical protein LOC79953
C21orf91	NM_001100420	early undifferentiated retina and lens isoform
C2orf71	NM_001029883	hypothetical protein LOC388939
C5orf41	NM_153607	luman-recruiting factor
CA14	NM_012113	carbonic anhydrase XIV precursor
CA4	NM_000717	carbonic anhydrase IV precursor
CABP1	NM_001033677	calcium binding protein 1 isoform 3
CABP2	NM_016366	calcium binding protein 2 isoform
CABP5	NM_019855	calcium binding protein 5
CACNA2D4	NM_172364	voltage-gated calcium channel alpha(2)delta-4
CAPRIN1	NM_005898	membrane component chromosome 11 surface marker
CAV1	NM_001753	caveolin 1
CAV2	NM_001233	caveolin 2 isoform a and b
CC2D2A	NM_001080522	coiled-coil and C2 domain containing 2A isoform
CD244	NM_016382.3	CD244 natural killer cell receptor 2B4
CD2BP2	NM_006110	
CDC40	NM_015891	cell division cycle 40 homolog
CDC5L	NM_001253	CDC5-like
CDH4	NM_001794	cadherin 4, type 1 preproprotein
CDKN2AIP	NM_017632	CDKN2A interacting protein
CDS1	NM_001263	CDP-diacylglycerol synthase 1
CDS2	NM_003818	phosphatidate cytidylyltransferase 2
CEP120	NM_153223	coiled-coil domain containing 100
CERKL	NM_201548	ceramide kinase-like isoform a
CETN1	NM_004066	centrin 1
CETN3	NM_004365	centrin 3
CHML	NM_001821	choroideremia-like Rab escort protein 2
CHMP1B	NM_020412	chromatin modifying protein 1B
CHRNA3	NM_000743	cholinergic receptor, nicotinic, alpha 3
CHST5	NM_024533	carbohydrate (N-acetylglucosamine 6-O)
CHST6	NM_021615	carbohydrate (N-acetylglucosamine 6-O)
CIRBP	NM_001280	cold inducible RNA binding protein
CKB	NM_001823	brain creatine kinase
CLDN5	NM_003277	claudin 5
CLSTN1	NM_001009566	calsyntenin 1 isoform 1
CLU	NM_001831	clusterin isoform 1
CLUL1	NM_199167	clusterin-like 1 (retinal)

CNGA1	NM_000087	cyclic nucleotide gated channel alpha 1 isoform
CNGB1	NM_001297	cyclic nucleotide gated channel beta 1 isoform
CNTF	NM_000614	ciliary neurotrophic factor
COBLL1	NM_014900	COBL-like 1
CORT	NM_001302	cortistatin preproprotein
CPLX4	NM_181654	complexin 4
CRABP1	NM_004378	cellular retinoic acid binding protein 1
CRABP2	NM_001878	cellular retinoic acid binding protein 2
CRB1	NM_201253	crumbs homolog 1 precursor
CREB5	NM_182898	cAMP responsive element binding protein 5
CRNKL1	NM_016652	crooked neck-like 1 protein
CRX	NM_000554	cone-rod homeobox protein
CRY1	NM_004075	cryptochrome 1 (photolyase-like)
CRY2	NM_021117	cryptochrome 2 (photolyase-like) isoform 1
CYGB	NM_134268	cytoglobin
D4S234E	NM_001040101	brain neuron cytoplasmic protein 1
DAB1	NM_021080	disabled homolog 1
DACH1	NM_080759	dachshund homolog 1 isoform a
DCT	NM_001129889	dopachrome tautomerase (dopachrome)
DDB1	NM_001923	damage-specific DNA binding protein 1
DDIT4	NM_019058	RTP801
DDX17	NM_001098504	DEAD box polypeptide 17 isoform 3
DDX21	NM_004728	DEAD (Asp-Glu-Ala-Asp) box polypeptide 21
DDX23	NM_004818	DEAD (Asp-Glu-Ala-Asp) box polypeptide 23
DFNB59	NM_001042702	deafness, autosomal recessive 59
DGKI	NM_004717	diacylglycerol kinase, iota
DHRS3	NM_004753	dehydrogenase/reductase (SDR family) member 3
DHRS7	NM_016029	dehydrogenase/reductase (SDR family) member 7
DHX30	NM_138615	DEAH (Asp-Glu-Ala-His) box polypeptide 30
DHX38	NM_014003	DEAH (Asp-Glu-Ala-His) box polypeptide 38
DHX9	NM_001357	DEAH (Asp-Glu-Ala-His) box polypeptide 9
DUSP1	NM_004417	dual specificity phosphatase 1
DYRK1A	NM_101395	dual-specificity tyrosine-(Y)-phosphorylation
EFEMP1	NM_001039349	EGF-containing fibulin-like extracellular matrix
EFTUD2	NM_004247	elongation factor Tu GTP binding domain
EHBP1	NM_015252	EH domain binding protein 1 isoform 1
EML2	NM_012155	echinoderm microtubule associated protein like
ENO2	NM_001975	enolase 2
EPB41L2	NM_001431	erythrocyte membrane protein band 4.1-like 2
ERO1LB	NM_019891	endoplasmic reticulum oxidoreductin 1-Lbeta
EVL	NM_016337	Enah/Vasp-like
EYA2	NM_005244	eyes absent 2 isoform a
EYA3	NM_001990	eyes absent 3
EYS	NM_198283	eyes shut homolog isoform 3
FAM46A	NM_017633	hypothetical protein LOC55603
FAM57B	NM_031478	hypothetical protein LOC83723

FIZ1	NM_032836	FLT3-interacting zinc finger 1
FLJ25404	NM_001109660	hypothetical protein LOC146378 isoform 2
FLT1	NM_002019	fms-related tyrosine kinase 1
FOXO3	NM_001455	forkhead box O3A
FRMPD1	NM_014907	FERM and PDZ domain containing 1
FRZB	NM_001463	frizzled-related protein
FSCN2	NM_001077182	fascin 2 isoform 1
FTH1	NM_002032	ferritin, heavy polypeptide 1
FUT8	NM_178154	fucosyltransferase 8 isoform a
FZD3	NM_017412	frizzled 3
GABRR2	NM_002043	gamma-aminobutyric acid (GABA) receptor, rho 2
GAPDH	NM_002046	glyceraldehyde-3-phosphate dehydrogenase
GBP7	NM_207398	guanylate binding protein 4-like
GJA10	NM_032602	gap junction protein, alpha 10
GJC1	NM_001080383	connexin 45
GLB1L	NM_024506	galactosidase, beta 1-like
GLTSCR2	NM_015710	glioma tumor suppressor candidate region gene 2
GNB1	NM_002074	guanine nucleotide-binding protein, beta-1
GNB2L1	NM_006098	guanine nucleotide binding protein (G protein),
GNB3	NM_002075	guanine nucleotide-binding protein, beta-3
GNB4	NM_021629	guanine nucleotide-binding protein, beta-4
GNB5	NM_016194	guanine nucleotide-binding protein, beta-5
GNG13	NM_016541	guanine nucleotide binding protein (G protein),
GNGT1	NM_021955	guanine nucleotide binding protein (G protein),
GNGT2	NM_031498	guanine nucleotide binding protein-gamma
GNL3	NM_206825	guanine nucleotide binding protein-like 3
GPD2	NM_000408	glycerol-3-phosphate dehydrogenase 2
GPM6A	NM_005277	glycoprotein M6A isoform 1
GPX3	NM_002084	glutathione peroxidase 3 precursor
GPX4	NM_001039847	glutathione peroxidase 4 isoform B precursor
GRIK2	NM_021956	glutamate receptor, ionotropic, kainate 2
GRK7	NM_139209	G-protein-coupled receptor kinase 7
GSTM1	NM_000561	glutathione S-transferase mu 1 isoform 1
GSTP1	NM_000852	glutathione transferase
GSTT1	NM_000853	glutathione S-transferase theta 1
GTF2H1	NM_005316	general transcription factor IIH, polypeptide 1,
GUCA1A	NM_000409	guanylate cyclase activator 1A (retina)
GUCA1B	NM_002098	guanylate cyclase activator 1B (retina)
GUCA1C	NM_005459	guanylate cyclase activator 1C
GUCY2D	NM_000180	guanylate cyclase 2D, membrane
HIF1A	NM_001530	hypoxia-inducible factor 1, alpha subunit
HK2	NM_000189	hexokinase 2
HMGB1	NM_002128	high-mobility group box 1
HMGN2	NM_005517	high-mobility group nucleosomal binding domain
HMOX1	NM_002133	heme oxygenase (decyclizing) 1
HNRNPK	NM_002140	heterogeneous nuclear ribonucleoprotein K

HPCA	NM_002143	hippocalcin
HPCAL1	NM_002149	hippocalcin-like 1
HSD17B11	NM_016245	dehydrogenase/reductase (SDR family) member 8
HSD17B14	NM_016246	dehydrogenase/reductase (SDR family) member 10
HSP90AA1	NM_001017963	heat shock protein 90kDa alpha (cytosolic),
HSPA1A	NM_005345	heat shock 70kDa protein 1A
IDH3B	NM_174856	isocitrate dehydrogenase 3, beta subunit isoform
IL11RA	NM_004512	interleukin 11 receptor, alpha isoform 1
IMMT	NM_006839	inner membrane protein, mitochondrial isoform 1
IMPDH1	NM_000883	inosine monophosphate dehydrogenase 1 isoform a
IMPG1	NM_001563	interphotoreceptor matrix proteoglycan 1
IMPG2	NM_016247	interphotoreceptor matrix proteoglycan 2
INADL	NM_176877	InaD-like
INO80B	NM_031288	high mobility group AT-hook 1-like 4
INTS10	NM_018142	integrator complex subunit 10
ISL1	NM_002202	islet-1
ISL2	NM_145805	ISL LIM homeobox 2
ITPRIPL2	NM_001034841	inositol 1,4,5-triphosphate receptor interacting
KCNB1	NM_004975	potassium voltage-gated channel, Shab-related
KCNJ14	NM_013348	potassium inwardly-rectifying channel J14
KCNV2	NM_133497	potassium channel, subfamily V, member 2
KIFC3	NM_005550	kinesin family member C3 isoform 1
KLF15	NM_014079	Kruppel-like factor 15
KLHL7	NM_018846	SBBI26 protein isoform 1
LAPTM4B	NM_018407	lysosomal associated transmembrane protein 4
LCN1	NM_002297	lipocalin 1 precursor
LENG8	NM_052925	leukocyte receptor cluster member 8
LHX1	NM_005568	LIM homeobox protein 1
LHX3	NM_014564	LIM homeobox protein 3 isoform b
LIF	NM_002309	leukemia inhibitory factor (cholinergic
LIN52	NM_001024674	lin-52 homolog
LRIT1	NM_015613	retina specific protein PAL
LRMP	NM_006152	lymphoid-restricted membrane protein
LRTM1	NM_020678	leucine-rich repeats and transmembrane domains
LSM2	NM_021177	LSM2 homolog, U6 small nuclear RNA associated
LSM3	NM_014463	Lsm3 protein
LSM4	NM_012321	U6 snRNA-associated Sm-like protein 4
LSM5	NM_012322	LSM5 homolog, U6 small nuclear RNA associated
LSM6	NM_007080	Sm protein F
LSM7	NM_016199	U6 snRNA-associated Sm-like protein LSm7
LSM8	NM_016200	U6 snRNA-associated Sm-like protein LSm8
LTB4R2	NM_019839	leukotriene B4 receptor 2
LZTFL1	NM_020347	leucine zipper transcription factor-like 1
MAB21L1	NM_005584	mab-21-like protein 1
MAB21L2	NM_006439	mab-21-like protein 2
MAP4K3	NM_003618	mitogen-activated protein kinase kinase kinase

MAP4K4	NM_145686	mitogen-activated protein kinase kinase kinase
MAPK10	NM_138980	mitogen-activated protein kinase 10 isoform 2
MAPK14	NM_139012	mitogen-activated protein kinase 14 isoform 2
MAPK8	NM_139049	mitogen-activated protein kinase 8 isoform JNK1
MBNL2	NM_144778	muscleblind-like 2 isoform 1
MBTPS1	NM_003791	membrane-bound transcription factor site-1
MDC1	NM_014641	mediator of DNA damage checkpoint 1
MDH1	NM_005917	cytosolic malate dehydrogenase
MEGF10	NM_032446	multiple EGF-like-domains 10
MERTK	NM_006343	MER receptor tyrosine kinase precursor
MFAP3L	NM_021647	microfibrillar-associated protein 3-like isoform
MSH5	NM_172165	mutS homolog 5 isoform b
MTFR1	NM_014637	mitochondrial fission regulator 1
MYH10	NM_005964	myosin, heavy polypeptide 10, non-muscle
MYL5	NM_002477	myosin regulatory light chain 5
MYL6	NM_021019	myosin, light chain 6, alkali, smooth muscle and
MYO3A	NM_017433	myosin IIIA
MYO9A	NM_006901	myosin IXA
MYRIP	NM_015460	myosin VIIA and Rab interacting protein
NAT1	NM_000662	N-acetyltransferase 1
NEDD8	NM_006156	neural precursor cell expressed, developmentally
NES	NM_006617	nestin
NEUROD4	NM_021191	neurogenic differentiation 4
NFKB1	NM_003998	nuclear factor kappa-B, subunit 1
NGB	NM_021257	neuroglobin
NHP2L1	NM_001003796	NHP2 non-histone chromosome protein 2-like 1
NOS1	NM_000620	nitric oxide synthase 1 (neuronal)
NOTCH1	NM_017617	notch1 preproprotein
NPVF	NM_022150	neuropeptide VF precursor
NR2E3	NM_016346	photoreceptor-specific nuclear receptor isoform
NRL	NM_006177	neural retina leucine zipper
NUB1	NM_016118	NEDD8 ultimate buster-1
NXNL1	NM_138454	nucleoredoxin-like 1
OAZ1	NM_004152	ornithine decarboxylase antizyme 1
OLFM3	NM_058170	olfactomedin 3
OPN3	NM_014322	opsin 3
OPN4	NM_001030015	opsin 4 isoform 1
OPN5	NM_181744	opsin 5 isoform 1
OPTC	NM_014359	opticin precursor
OSBP2	NM_030758	oxysterol binding protein 2 isoform a
OSBPL1A	NM_080597	oxysterol-binding protein-like 1A isoform B
OTX1	NM_014562	orthodenticle homeobox 1
PARP1	NM_001618	poly (ADP-ribose) polymerase family, member 1
PARP3	NM_005485	poly (ADP-ribose) polymerase family, member 3
PCBP2	NM_005016	poly(rC) binding protein 2 isoform a
PCBP4	NM_033008	poly(rC) binding protein 4 isoform c

PCDH21	NM_033100	protocadherin 21 precursor
PCMT1	NM_005389	protein-L-isoaspartate (D-aspartate)
PDC	NM_002597	phosducin isoform a
PDCL	NM_005388	phosducin-like
PDE6A	NM_000440	phosphodiesterase 6A, alpha subunit
PDE6B	NM_000283	rod cGMP-phosphodiesterase beta-subunit isoform
PDE6C	NM_006204	phosphodiesterase 6C, cGMP-specific, cone, alpha
PDE6D	NM_002601	phosphodiesterase 6D, cGMP-specific, rod, delta
PDE6G	NM_002602	phosphodiesterase 6G, cGMP-specific, rod, gamma
PDE6H	NM_006205	phosphodiesterase 6H, cGMP-specific, cone,
PDZD7	NM_024895	PDZ domain containing 7
PGAM1	NM_002629	phosphoglycerate mutase 1 (brain)
PHF17	NM_199320	PHD finger protein 17 long isoform
PITPNA	NM_006224	phosphatidylinositol transfer protein, alpha
PITPNC1	NM_012417	phosphatidylinositol transfer protein,
PITPNM1	NM_004910	phosphatidylinositol transfer protein,
PITPNM2	NM_020845	phosphatidylinositol transfer protein,
PITPNM3	NM_031220	PITPNM family member 3
PKM2	NM_002654	pyruvate kinase, muscle isoform M2
PLAU	NM_002658	urokinase plasminogen activator isoform 1
PLCB4	NM_000933	phospholipase C beta 4 isoform a
PLEKHB1	NM_021200	pleckstrin homology domain containing, family B
PLRG1	NM_002669	pleiotropic regulator 1 (PRL1 homolog,
PNOC	NM_006228	prepronociceptin
PNPLA2	NM_020376	patatin-like phospholipase domain containing 2
POU4F2	NM_004575	Brn3b POU domain transcription factor
POU6F2	NM_007252	POU domain, class 6, transcription factor 2
PPEF2	NM_006239	serine/threonine protein phosphatase with
PPIH	NM_006347	peptidylprolyl isomerase H
PPP1R2	NM_006241	protein phosphatase 1, regulatory subunit 2
PPP2R1A	NM_014225	alpha isoform of regulatory subunit A, protein
PRCD	NM_001077620	Homo sapiens cDNA FLJ43629 fis, clone SPLEN2029727.
PRDX1	NM_181696	peroxiredoxin 1
PRDX2	NM_005809	peroxiredoxin 2 isoform a
PRKAA2	NM_006252	AMP-activated protein kinase alpha 2 catalytic
PRKCI	NM_002740	protein kinase C, iota
PRKCZ	NM_002744	protein kinase C, zeta isoform 1
PROK1	NM_032414	prokineticin 1
PROM1	NM_006017	prominin 1
PROX1	NM_002763	prospero homeobox 1
PRPF3	NM_004698	PRP3 pre-mRNA processing factor 3 homolog
PRPF31	NM_015629	pre-mRNA processing factor 31 homolog
PRPF4	NM_004697	PRP4 pre-mRNA processing factor 4 homolog
PRPF6	NM_012469	PRP6 pre-mRNA processing factor 6 homolog
PRPF8	NM_006445	U5 snRNP-specific protein
PRPH2	NM_000322	peripherin 2

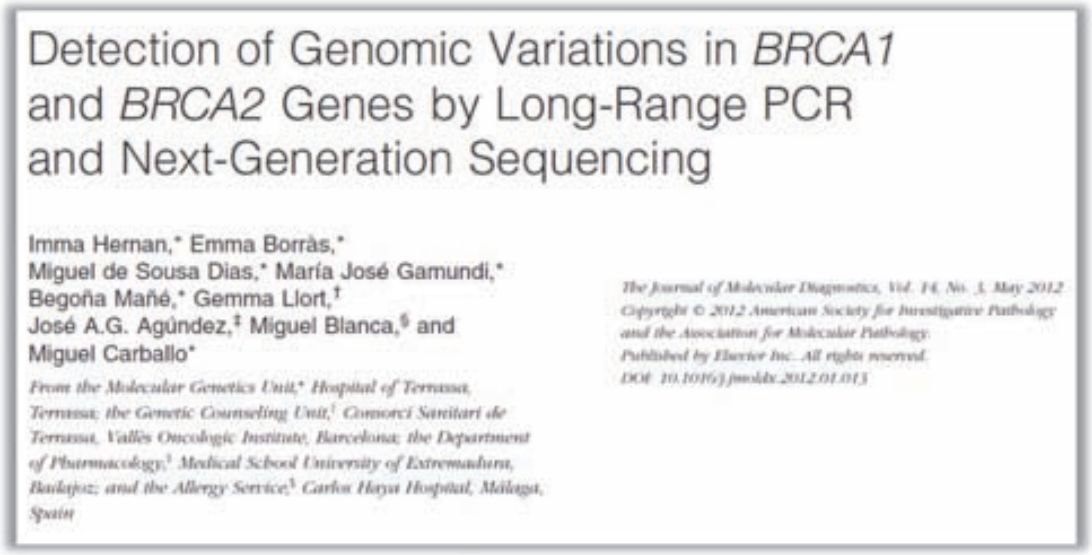
PTPN13	NM_080685	protein tyrosine phosphatase, non-receptor type
PTPRM	NM_001105244	protein tyrosine phosphatase, receptor type, M
PVALB	NM_002854	parvalbumin
RAB2B	NM_032846	RAB2B protein
RABGGTA	NM_182836	Rab geranylgeranyltransferase, alpha subunit
RABGGTB	NM_004582	Rab geranylgeranyltransferase, beta subunit
RAI14	NM_015577	retinoic acid induced 14
RANBP2	NM_006267	RAN binding protein 2
RASSF2	NM_170774	Ras association domain family 2
RAX2	NM_032753	retina and anterior neural fold homeobox 2
RBP3	NM_002900	retinol-binding protein 3 precursor
RCVRN	NM_002903	recoverin
RDH10	NM_172037	retinol dehydrogenase 10
RDH8	NM_015725	retinol dehydrogenase 8 (all-trans)
RGR	NM_002921	retinal G-protein coupled receptor isoform 1
RGS16	NM_002928	regulator of G-protein signalling 16
RHBDD3	NM_012265	rhomboid domain containing 3
RHO	NM_000539	rhodopsin
RIC8B	NM_018157	resistance to inhibitors of cholinesterase 8
RIMS1	NM_014989	regulating synaptic membrane exocytosis 1
RIMS2	NM_001100117	regulating synaptic membrane exocytosis 2
RING1	NM_002931	ring finger protein 1
RLBP1	NM_000326	retinaldehyde binding protein 1
ROM1	NM_000327	retinal outer segment membrane protein 1
RORB	NM_006914	RAR-related orphan receptor B
RP1	NM_006269	retinitis pigmentosa RP1 protein
RP1L1	NM_178857	retinitis pigmentosa 1-like 1
RP9	NM_203288	retinitis pigmentosa 9
RPL13A	NM_012423	ribosomal protein L13a
RPL3	NM_000967	ribosomal protein L3 isoform a
RPL4	NM_000968	ribosomal protein L4
RPL8	NM_000973	ribosomal protein L8
RPL9	NM_001024921	ribosomal protein L9
RPLP0	NM_053275	ribosomal protein P0
RPS12	NM_001016	ribosomal protein S12
RPS2	NM_002952	ribosomal protein S2
RPS3A	NM_001006	ribosomal protein S3a
RPS9	NM_001013	ribosomal protein S9
RPSA	NM_002295	ribosomal protein SA
RRAGD	NM_021244	Ras-related GTP binding D
RRH	NM_006583	peropsin
RTBDN	NM_031429	retbindin isoform 2
SARM1	NM_015077	sterile alpha and TIR motif containing 1
SART1	NM_005146	squamous cell carcinoma antigen recognized by T
SCAMP1	NM_004866	secretory carrier membrane protein 1
SCARB1	NM_005505	scavenger receptor class B, member 1 isoform 1

SCRT1	NM_031309	scratch
SEMA4A	NM_022367	semaphorin B
SEPT11	NM_018243	septin 11
SEPT5	NM_002688	septin 5
SEPT8	NM_001098811	septin 8 isoform a
SERPINB2	NM_002575	serine (or cysteine) proteinase inhibitor, clade
SERPINF1	NM_002615	serine (or cysteine) proteinase inhibitor, clade
SETD3	NM_032233	SET domain containing 3 isoform a
SF1	NM_004630	splicing factor 1 isoform 1
SF3A2	NM_007165	splicing factor 3a, subunit 2
SF3B1	NM_012433	splicing factor 3b, subunit 1 isoform 1
SFRP1	NM_003012	secreted frizzled-related protein 1
SFRP5	NM_003015	secreted frizzled-related protein 5
SFRS1	NM_006924	splicing factor, arginine/serine-rich 1 isoform
SLC12A5	NM_020708	solute carrier family 12 (potassium-chloride
SLC1A2	NM_004171	excitatory amino acid transporter 2
SLC1A7	NM_006671	solute carrier family 1 (glutamate transporter),
SLC24A1	NM_004727	solute carrier family 24
SLC24A2	NM_020344	solute carrier family 24
SLC25A27	NM_004277	solute carrier family 25, member 27
SLC25A29	NM_001039355	solute carrier family 25, member 29
SLC32A1	NM_080552	solute carrier family 32, member 1
SLC7A1	NM_003045	solute carrier family 7 (cationic amino acid
SLIT1	NM_003061	slit homolog 1
SMNDC1	NM_005871	survival motor neuron domain containing 1
SNCG	NM_003087	synuclein, gamma (breast cancer-specific protein)
SND1	NM_014390	staphylococcal nuclease domain containing 1
SNRNP200	NM_014014	activating signal cointegrator 1 complex subunit
SNRNP27	NM_006857	small nuclear ribonucleoprotein 27kDa
SNRNP40	NM_004814	WD repeat domain 57 (U5 snRNP specific)
SNRPB	NM_003091	small nuclear ribonucleoprotein polypeptide B/B'
SNRPD1	NM_006938	small nuclear ribonucleoprotein D1 polypeptide
SNRPD2	NM_004597	small nuclear ribonucleoprotein polypeptide D2
SNRPD3	NM_004175	small nuclear ribonucleoprotein polypeptide D3
SNRPE	NM_003094	small nuclear ribonucleoprotein polypeptide E
SNRPF	NM_003095	small nuclear ribonucleoprotein polypeptide F
SNRPG	NM_003096	small nuclear ribonucleoprotein polypeptide G
SNRPN	NM_022805	small nuclear ribonucleoprotein polypeptide N
SOLH	NM_005632	small optic lobes
SP4	NM_003112	Sp4 transcription factor
SPARC	NM_003118	secreted protein, acidic, cysteine-rich
SPC25	NM_020675	spindle pole body component 25
SPP1	NM_001040058	secreted phosphoprotein 1 isoform a
SPTBN5	NM_016642	spectrin, beta, non-erythrocytic 5
SSTR2	NM_001050	somatostatin receptor 2
STK35	NM_080836	serine/threonine kinase 35

SUPT16H	NM_007192	chromatin-specific transcription elongation
SV2B	NM_014848	synaptic vesicle protein 2B homolog
SYNJ2	NM_003898	synaptojanin 2
TAF11	NM_005643	TBP-associated factor 11
TBCC	NM_003192	beta-tubulin cofactor C
TBCD	NM_005993	beta-tubulin cofactor D
TBX2	NM_005994	T-box 2
TEAD4	NM_003213	TEA domain family member 4 isoform 3
THBS1	NM_003246	thrombospondin 1 precursor
THY1	NM_006288	Thy-1 cell surface antigen preproprotein
TMEM106C	NM_024056	transmembrane protein 106C isoform a
TNFSF18	NM_005092	tumor necrosis factor (ligand) superfamily,
TOP2B	NM_001068	DNA topoisomerase II, beta isozyme
TOPBP1	NM_007027	topoisomerase (DNA) II binding protein 1
TOPORS	NM_005802	topoisomerase I binding, arginine/serine-rich
TPD52	NM_001025252	tumor protein D52 isoform 1
TRAM2	NM_012288	translocation-associated membrane protein 2
TRPC3	NM_003305	transient receptor potential cation channel,
TSGA14	NM_018718	testis specific, 14
TSPAN3	NM_005724	transmembrane 4 superfamily member 8 isoform 1
TTYH2	NM_032646	tweety 2 isoform 1
TUBA1B	NM_006082	tubulin, alpha, ubiquitous
TUBB2A	NM_001069	tubulin, beta 2
TUBB2C	NM_006088	tubulin, beta, 2
TULP1	NM_003322	tubby like protein 1
TULP2	NM_003323	tubby like protein 2
TULP4	NM_020245	tubby like protein 4 isoform 1
TXNL4A	NM_006701	thioredoxin-like 4A
U2AF1	NM_001025204	U2 small nuclear RNA auxillary factor 1 isoform
UBA52	NM_001033930	ubiquitin and ribosomal protein L40 precursor
UBC	NM_021009	ubiquitin C
UBE2E3	NM_182678	ubiquitin-conjugating enzyme E2E 3
UBE2J1	NM_016021	ubiquitin-conjugating enzyme E2, J1
UBE4A	NM_004788	ubiquitination factor E4A
UNC119	NM_005148	unc119 (C.elegans) homolog isoform a
USH2A	NM_206933	usherin isoform B
USP39	NM_006590	ubiquitin specific protease 39
UTRN	NM_007124	utrophin
VAMP1	NM_014231	vesicle-associated membrane protein 1 isoform 1
VAX1	NM_001112704	ventral anterior homeobox 1 isoform a
VAX2	NM_012476	ventral anterior homeobox 2
VPS33A	NM_022916	vacuolar protein sorting 33A
VPS53	NM_001128159	vacuolar protein sorting 53 isoform 1
WBP4	NM_007187	WW domain-containing binding protein 4
WDR17	NM_170710	WD repeat domain 17 isoform 1
WDR5	NM_017588	WD repeat domain 5

WIF1	NM_007191	WNT inhibitory factor 1 precursor
WSB1	NM_015626	WD repeat and SOCS box-containing 1 isoform 1
YBX1	NM_004559	nuclease sensitive element binding protein 1
ZBTB17	NM_003443	zinc finger and BTB domain containing 17
ZFYVE9	NM_004799	zinc finger, FYVE domain containing 9 isoform 3
ZKSCAN1	NM_003439	zinc finger protein 36
ZNF295	NM_001098402	zinc finger protein 295 isoform L
ZNF385A	NM_015481	zinc finger protein 385A isoform c
ZNF764	NM_033410	zinc finger protein 764
ZSCAN18	NM_023926	zinc finger and SCAN domain containing 18

B. Publications derived from the research work of the presented thesis



Detection of Genomic Variations in *BRCA1* and *BRCA2* Genes by Long-Range PCR and Next-Generation Sequencing

Imma Hernan,* Emma Borràs,* Miguel de Sousa Dias,* María José Gamundi,* Begoña Mañé,* Gemma Llort,† José A.G. Agúndez,‡ Miguel Blanca,§ and Miguel Carballo*

From the Molecular Genetics Unit, Hospital of Terrassa, Terrassa; the Genetic Counseling Unit,† Consorci Sanitari de Terrassa, Vallès Oncologic Institute, Barcelona; the Department of Pharmacology,‡ Medical School University of Extremadura, Badajoz; and the Allergy Service,§ Carlos Haya Hospital, Málaga, Spain*

The Journal of Molecular Diagnostics, Vol. 14, No. 3, May 2012

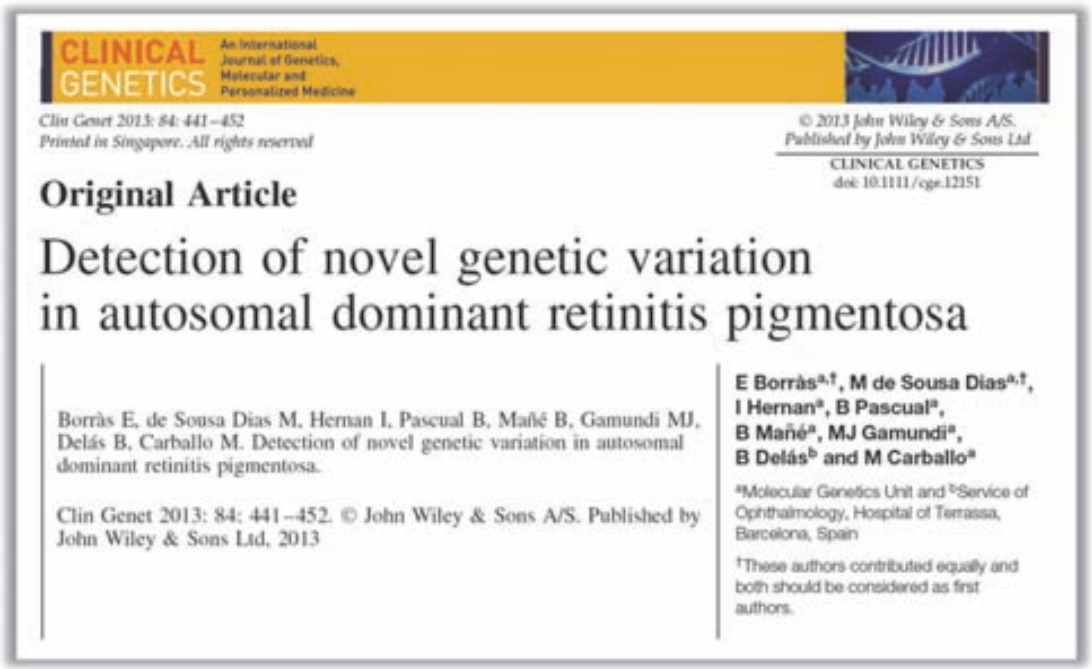
DOI: 10.1016/j.jmoldx.2012.01.013

Molecular Vision 2013; 19:654-664 <http://www.molvis.org/molvis/v19/654>
Received 20 December 2012 | Accepted 19 March 2013 | Published 21 March 2013

Detection of novel mutations that cause autosomal dominant retinitis pigmentosa in candidate genes by long-range PCR amplification and next-generation sequencing

Miguel de Sousa Dias, Imma Hernan, Beatriz Pascual, Emma Borràs, Begoña Mañé, María José Gamundi, Miguel Carballo

Molecular Genetics Unit, Hospital of Terrassa, Barcelona, Spain



CLINICAL GENETICS
An International Journal of Genetics, Molecular and Personalized Medicine

Clin Genet 2013: 84: 441–452
Printed in Singapore. All rights reserved

CLINICAL GENETICS
doi: 10.1111/cge.12151

Original Article

Detection of novel genetic variation in autosomal dominant retinitis pigmentosa

Borràs E, de Sousa Dias M, Hernan I, Pascual B, Mañé B, Gamundi MJ, Delás B, Carballo M. Detection of novel genetic variation in autosomal dominant retinitis pigmentosa.

Clin Genet 2013: 84: 441–452. © John Wiley & Sons A/S. Published by John Wiley & Sons Ltd, 2013

E Borràs[a,†], M de Sousa Dias[a,†], I Hernan[a], B Pascual[a], B Mañé[a], MJ Gamundi[a], B Delás[b] and M Carballo[a]

[a]Molecular Genetics Unit and [b]Service of Ophthalmology, Hospital of Terrassa, Barcelona, Spain

[†]These authors contributed equally and both should be considered as first authors.