2013

Highly Parallel genetic approaches in Mendelian and complex disease

Paola Benaglio

UNIL | Université de Lausanne
Faculté de biologie et de médicine

Département de Génetique Médicale

Highly parallel genetic approaches in Mendelian and complex diseases

Thèse de doctorat ès Sciences de la Vie (PhD)

Présentée à la Faculté de Biologie et de Médicine de l'Université de Lausanne par

Paola Benaglio

Master de Biologie Moléculaire de l'Università degli Studi di Milano, Italie

Jury:

Dr. Luc PELLERIN, Président
Dr. Carlo RIVOLTA, Directeur de thèse
Dr. Colin WILLOUGHBY, Expert
Dr. Dror SHARON, Expert

Lausanne, 2013

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Ecole Doctorale

Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président	Monsieur	Prof. Luc **Pellerin**
Directeur de thèse	Monsieur	Dr Carlo **Rivolta**
Experts	Monsieur	Prof. Colin **Willoughby**
	Monsieur	Dr Dror **Sharon**

le Conseil de Faculté autorise l'impression de la thèse de

Madame Paola Benaglio

master de biologie moléculaire de l'Università degli Studi di Milano, Italie

intitulée

Highly parallel genetic approaches in Mendelian and complex diseases

Lausanne, le 15 novembre 2013

pour Le Doyen
de la Faculté de Biologie et de Médecine

Prof. Luc Pellerin

A mio papà

ABSTRACT

The recent advance in high-throughput sequencing and genotyping protocols allows rapid investigation of Mendelian and complex diseases on a scale not previously been possible. In my thesis research I took advantage of these modern techniques to study retinitis pigmentosa (RP), a rare inherited disease characterized by progressive loss of photoreceptors and leading to blindness; and hypertension, a common condition affecting 30% of the adult population.

Firstly, I compared the performance of different next generation sequencing (NGS) platforms in the sequencing of the RP-linked gene *PRPF31*. The gene contained a mutation in an intronic repetitive element, which presented difficulties for both classic sequencing methods and NGS. We showed that all NGS platforms are powerful tools to identify rare and common DNA variants, also in case of more complex sequences. Moreover, we evaluated the features of different NGS platforms that are important in re-sequencing projects.

The main focus of my thesis was then to investigate the involvement of pre-mRNA splicing factors in autosomal dominant RP (adRP). I screened 5 candidate genes in a large cohort of patients by using long-range PCR as enrichment step, followed by NGS. We tested two different approaches: in one, all target PCRs from all patients were pooled and sequenced as a single DNA library; in the other, PCRs from each patient were separated within the pool by DNA barcodes. The first solution was more cost-effective, while the second one allowed obtaining faster and more accurate results, but overall they both proved to be effective strategies for gene screenings in many samples. We could in fact identify novel missense mutations in the *SNRNP200* gene, encoding an essential RNA helicase for splicing catalysis. Interestingly, one of these mutations showed incomplete penetrance in one family with adRP. Thus, we started to study the possible molecular causes underlying phenotypic differences between asymptomatic and affected members of this family.

For the study of hypertension, I joined a European consortium to perform genome-wide association studies (GWAS). Thanks to the use of very informative genotyping arrays and of phenotipically well-characterized cohorts, we could identify a novel susceptibility locus for hypertension in the promoter region of the endothelial nitric oxide synthase gene (*NOS3*). Moreover, we have proven the direct causality of the associated SNP using three different methods: 1) targeted resequencing, 2) luciferase assay, and 3) population study.

RESUME

Le récent progrès dans le Séquençage à haut Débit et les protocoles de génotypage a permis une plus vaste et rapide étude des maladies mendéliennes et multifactorielles à une échelle encore jamais atteinte. Durant ma thèse de recherche, j'ai utilisé ces nouvelles techniques de séquençage afin d'étudier la rétinite pigmentaire (RP), une maladie héréditaire rare caractérisée par une perte progressive des photorécepteurs de l'œil qui entraine la cécité; et l'hypertension, une maladie commune touchant 30% de la population adulte.

Tout d'abord, j'ai effectué une comparaison des performances de différentes plateformes de séquençage NGS (*Next Generation Sequencing*) lors du séquençage de *PRPF31*, un gène lié à RP. Ce gène contenait une mutation dans un élément répétable intronique, qui présentait des difficultés de séquençage avec la méthode classique et les NGS. Nous avons montré que les plateformes de NGS analysées sont des outils très puissants pour identifier des variations de l'ADN rares ou communes et aussi dans le cas de séquences complexes. De plus, nous avons exploré les caractéristiques des différentes plateformes NGS qui sont importantes dans les projets de re-séquençage.

L'objectif principal de ma thèse a été ensuite d'examiner l'effet des facteurs d'épissage de pre-ARNm dans une forme autosomale dominante de RP (adRP). Un screening de 5 gènes candidats issus d'une large cohorte de patients a été effectué en utilisant la long-range PCR comme étape d'enrichissement, suivie par séquençage avec NGS. Nous avons testé deux approches différentes : dans la première, toutes les cibles PCRs de tous les patients ont été regroupées et séquencées comme une bibliothèque d'ADN unique; dans la seconde, les PCRs de chaque patient ont été séparées par code barres d'ADN. La première solution a été la plus économique, tandis que la seconde a permis d'obtenir des résultats plus rapides et précis. Dans l'ensemble, ces deux stratégies se sont démontrées efficaces pour le screening de gènes issus de divers échantillons. Nous avons pu identifier des nouvelles mutations faux-sens dans le gène *SNRNP200*, une hélicase ayant une fonction essentielle dans l'épissage. Il est intéressant de noter qu'une des ces mutations montre une pénétrance incomplète dans une famille atteinte d'adRP. Ainsi, nous avons commencé une étude sur les causes moléculaires entrainant des différences phénotypiques entre membres affectés et asymptomatiques de cette famille.

Lors de l'étude de l'hypertension, j'ai rejoint un consortium européen pour réaliser une étude d'association Pangénomique ou *genome-wide association study* Grâce à l'utilisation de tableaux de génotypage très informatifs et de cohortes extrêmement bien caractérisées au niveau phénotypique, un nouveau locus lié à l'hypertension a été identifié dans la région promotrice du gène endothélial *nitric oxide sinthase* (*NOS3*). Par ailleurs, nous avons prouvé la cause directe du SNP associé au moyen de trois méthodes différentes: i) en reséquençant la cible avec NGS, ii) avec des essais à la luciférase et iii) une étude de population.

ACKNOWLEDGEMENTS

I am sincerely grateful to my thesis director and supervisor Dr. Carlo Rivolta, who has given me the opportunity to join his group at the Department of Medical Genetics (DGM) and to participate in so many different and exciting projects. He has always supported me by giving me a lot of confidence and responsibility in every aspect of my research.

I thank very much the members of the jury Dr. Luc Pellerin, Dr. Dror Sharon and Dr. Colin Willoughby for having taken the time to read and evaluate my thesis and to meet with me for the examination (twice).

I want to express my special thanks to Dr. Eliot Berson, Dr. Christian Hamel, Dr. Carmen Ayuso and their collaborators for having provided the most precious material to conduce my research on: the samples from patients and their families affected with retinitis pigmentosa. Their commitment to improve the life conditions of these people serves as motivation for young researcher like me who want to continue to study inheritable disorders.

I had important collaborators from the University of Milan: Dr. Daniele Cusi, Dr. Cristina Barlassina, Dr. Fabio Macciardi and Dr. Erika Salvi, with whom I shared the eventful adventure of the HYPERGENES project.

From the University of Lausanne, I need to acknowledge the highly qualified support of Dr. Keith Harshman, for next generation sequencing and Dr. Zoltan Kutalik, for association studies.

At the DGM, I was surrounded by friendly people, who made my stay very pleasant. I first want to mention the secretary Suzanne, who has been always very kind and helpful, the new PhD students Beryl and Nicola, who helped me with the translation of my thesis summaries in French. Many colleagues also helped me with experimental work and scientific discussions and I would like to thank them all. In particular I want to thank Mrs Adriana Ransijn, who helped me with cell culture and other laboratory issues, Dr. Goranka Tanackovic, for her work on splicing factors, Drs. Koji Nishiguchi, Hanna Koskiniemi and Giulia Venturini for collaboration to an outstanding paper and nice colleagueship, Miss Giulia Ascari for assistance in sequencing, Mr. Luca Bartesaghi, for the continuous training and troubleshooting in molecular biology that he provided me, Dr. Alessandro di Gioia, for his

help in the laboratory, and Dr. Andrea Prunotto, for bioinformatic help (but mostly for coffee and cigarettes). Some of them also become very good friends of mine.
Speaking of that, Lausanne has been for me a place of great social entertainment, where I had good friends from the Italian Community, the Italian Group of Theater (The Pourquoi pas?), and other more international contexts…too many to be mentioned all. Except from one, my boyfriend Dr. Albrecht Lindner, who courageously supported me during these years.

Finally, my deepest thanks go to my family, and especially to my father, who has always and unconditionally encouraged and praised me during all my studies.

TABLE OF CONTENTS

Part I.

Next generation sequencing for the study of autosomal dominant retinitis pigmentosa

INTRODUCTION

1. THE MOLECULAR BASIS OF RETINITIS PIGMENTOSA

1.1 The Retina

1.1.1 Structural proprieties

The retina is a multi-layered sensory tissue that extends over a large portion of the posterior pole of the eye (**Fig. 1A**). Developing as an outpocketing from the neural tube, the retina is part of the central nervous system. It contains millions of photoreceptors that capture the rays of light and convert them into action potentials, which travel along the optic nerve to the visual centers of the brain where they are turned into images. Before the rays of light enter the photoreceptor layer, they must pass through layers of blood vessels, nerve fibers and four other types of neurons: bipolar cells, ganglion cells, horizontal cells, and amacrine cells (**Fig. 1B**). The most direct path for transmitting visual information to the brain is through a chain composed of three neurons: a photoreceptor, a bipolar cell, and a ganglion cell. Horizontal cells and amacrine cells mediate lateral interactions, which modify the main signal.

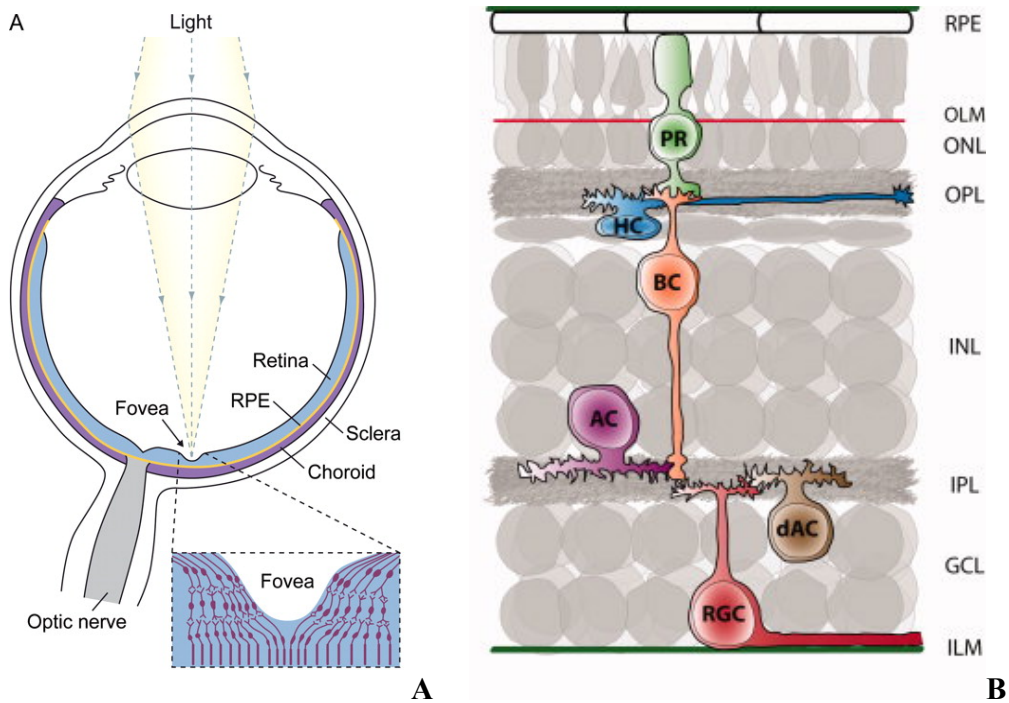


Figure 1. Anatomy of the vertebrate eye and structure of the retina. **A)** Anatomy of the eye. Image from [1]**. B)** The five neuronal cell types of the retina. Photoreceptors (PR), horizontal cells (HC), bipolar cells (BC), amacrine cells (AC) retinal ganglion cells (RGC) and displaced amacrine cells (dAC). RPE - retinal pigment epithelium, OLM - outer limiting membrane, ONL - outer nuclear layer, OPL - outer plexiform layer, INL - inner nuclear layer, IPL - inner plexiform layer, GCL - ganglion cell layer, ILM - inner limiting membrane. Image taken from [2].

The cell bodies and processes of these neurons are stacked into different, alternate layers, which are organized as follows, in the sense of the rays of light (**Fig. 1B**):

- *Inner limiting membrane* (ILM): is the inner surface of the retina, at the boundary with the vitreous body of the eye. It is composed of the conical feet of Müller glial cells and astrocytes.
- *Ganglion cell layer (GCL):* contains the nuclei of retinal ganglion cells, the axons of which form the optic nerve fibers, and some displaced amacrine cells.
- *Inner plexiform layer (IPL):* contains the synapses between the bipolar cell (BC) axons and the dendrites of the ganglion and amacrine cells (AC).
- *Inner nuclear layer (INL):* contains the cell bodies of the bipolar, amacrine, horizontal and Müller cells.
- *Outer plexiform layer (OPL)*: contains the synapses between photoreceptors axons and dendrites of horizontal and bipolar cells.
- *Outer nuclear layer (ONL):* contains the nuclei of photoreceptors and their *inner segments (IS)*, where metabolism, biosynthesis and endocytosis take place.
- *Photoreceptor layer:* contains the *outer segments*, i.e. the apical extensions of photoreceptors, composed of membranous disks containing light-sensitive photopigments and other proteins involved in the light transduction process. The outer segment communicates with the cell body trough the connecting cilium, essential for protein transport and structural integrity (**Fig. 2A**).
- *Retinal pigment epithelium* (RPE): a single layer of cells with tight junctions. They play an essential role in the turnover of the disks by phagocytosis and of the photopigment molecules after they have been exposed to light. Moreover the RPE contains melanin, which has a fundamental role in reducing light scattering in the back of the eye [3].

1.1.2 Functional aspects

There are two types of photoreceptors in the vertebrate retina: rods and cones. The first ones are specialized for vision at low light levels and the second for high visual acuity and the perception of colors in daytime lighting conditions. Three different pigments that absorb light of different wavelengths (red, green and blue) are packed in the disks of different types of cones, while there is only one kind of rod photopigment, called *rhodopsin*.

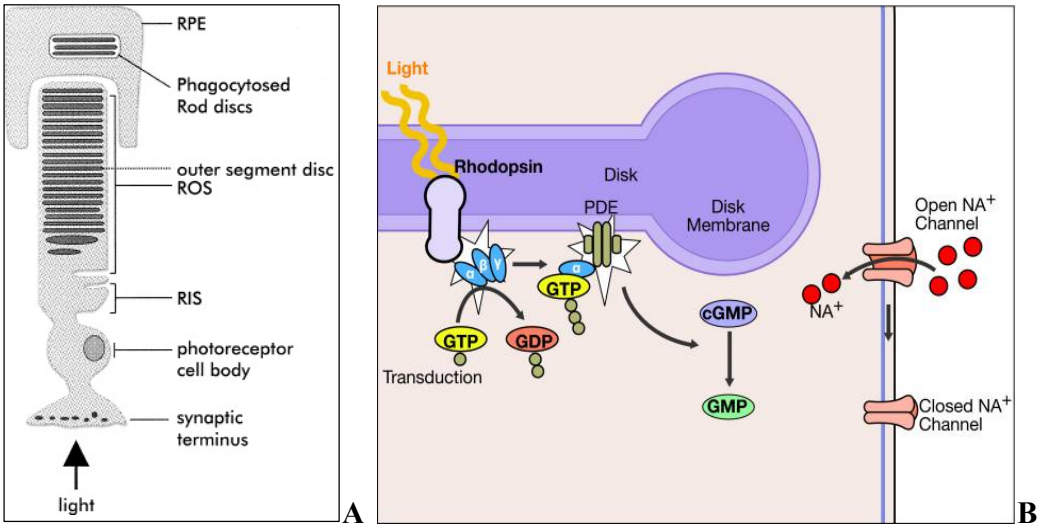


Figure 2. Schematic representations of **A)** a rod photoreceptor structure [4] and **B)** the phototransduction cascade [5]. RPE=retinal pigmented epithelium; ROS = rod outer segment; RIS = rod inner segment; PDE = phosphodiesterase.

Rods are the most abundant photoreceptors; in human retina there are approximately 120 million rods versus 6 million cones. Rods are also the most studied cell type of the retina, where mechanisms of phototransduction, visual cycle and electrophysiology have been elucidated. In primates, only the cones are densely packed in a region of the retina called *fovea centralis* (**Fig. 1A**), responsible for sharp central vision. There the cell bodies of the neurons are shifted to the side to reduce distortion of light paths, and each cone has a private line to the visual cortex. Visual acuity decreases toward the periphery of the retina where the cones become less dense and single bipolar-ganglion cell transmission lines serve many photoreceptors [3].

The cascade of events that transform light into an electric signal is called phototransduction (**Fig. 2B**) and starts with the absorption of a single photon, which induces the chromphore retinal (or vitamin A) to undergo a conformational change from 11-*cis*-retinal to all-*trans* retinal. Rhodopsin, which is bound to retinal, activates a G-protein (transducin), which amplifies the signal to hundreds of other G-proteins, which stimulate the cGMP phosphodiesterase (PDE) to hydrolyze cGMP into GMP. The decrease in cGMP concentration results in the closure of cGMP-gated cation channels, which in dark conditions pump sodium and calcium inside the cell. The hyperpolarization of the cells after a light stimulus causes a reduction in release of neurotransmitter glutamate at the synaptic terminal, and as a consequence of this, a signal is sent to the ganglion cells through bipolar cells [1].

The unique structural and functional specialization of the retina, and in particular of the photoreceptors, renders this tissue exceptionally susceptible to dysfunctions [6]. This vulnerability is enhanced by the continuous exposure to photons and free radicals and by the high metabolic energy demand, needed to sustain the phototransduction process. It is therefore not surprising that dystrophies of retinal photoreceptors cause most of adult blindness conditions in industrialized countries, among which retinitis pigmentosa represents the most prevalent of monogenic inheritable forms.

1.2 Retinitis Pigmentosa

1.2.1 Prevalence and classification

Retinitis pigmentosa (RP) is a heterogeneous group of hereditary conditions characterized by progressive retinal dystrophy with typically a major involvement of rod photoreceptors. It affects more than one million individuals worldwide, with a prevalence of one in 4000 [7]. It has typical Mendelian inheritance patterns: autosomal dominant (adRP, ~20% of cases), autosomal recessive (arRP, ~30%) or X-linked (~10%) [8]. The remaining fraction of RP patients are isolates cases, likely misrecognized recessive or *de novo*. Moreover, some cases of RP are due to digenic inheritance of mutations [9].

RP is usually a disease restricted to the eye (defined as non-syndromic RP), although there are several cases in which it can be part of more complex syndromes - 30 in total - and associates with non-ocular phenotypes (syndromic RP). The most common syndromic form of RP, which affects about 15% of all RP cases, is Usher syndrome, characterized by RP and hearing impairment [10]. The degree of deafness is variable and defines three types of Usher syndromes: it can be severe and present at birth (type I), moderate/mild (type II) or it can appear after childhood and progressively worsen during later years (type III). The second most common syndrome associated with RP (5% of cases) is Bardet-Biedl syndrome. The phenotype variably associates RP with obesity, developmental delay, polydactyly, hypogenitalism and structural renal abnormalities leading to renal failure and often transplantation.

1.2.2 Symptoms

Despite the name, retinitis pigmentosa does not involve an inflammation of the retina, as it was originally believed, but it is a result of progressive degeneration of photoreceptors. Conversely, the pigmented deposits observed in the *fundus* (the interior surface of the

posterior pole of the eye) are a hallmark for RP diagnosis. These pigmented granules, called bone-spicule deposits, are released by the RPE and accumulate in the neural retina, starting from the mid periphery. At later stages, the deposits are present all over the retina and the fundus appears depigmented (**Fig. 3**). Other characteristic consequences of retinal atrophy in RP are the attenuation of retinal vessels and cataracts. The visual symptoms of a typical RP patient reflect the primary degeneration of rods followed by degeneration of cones (often referred to as a rod-cone dystrophy). The first manifestation of RP is therefore nigh blindness in adolescence as consequence of early involvement of rods. This is followed by loss of peripheral visual field in young adulthood, the progressive constriction of the visual field leading to tunnel vision and the final loss of central vision by age of about 60, due to secondary, progressive degeneration of cones.

These clinical manifestations can be highly variable within patients, even belonging to the same family. The main features that distinguish the different types of RP are the age of onset, the rate of progression and the degree of rod or cone contribution to the disease. For example some patients experience visual loss in childhood while other have the first symptoms at mid adulthood. Some forms of RP are more similar to cone-rod dystrophies, because the cones degenerate simultaneously with the rods and this leads to a greater and earlier impairment of central vision.

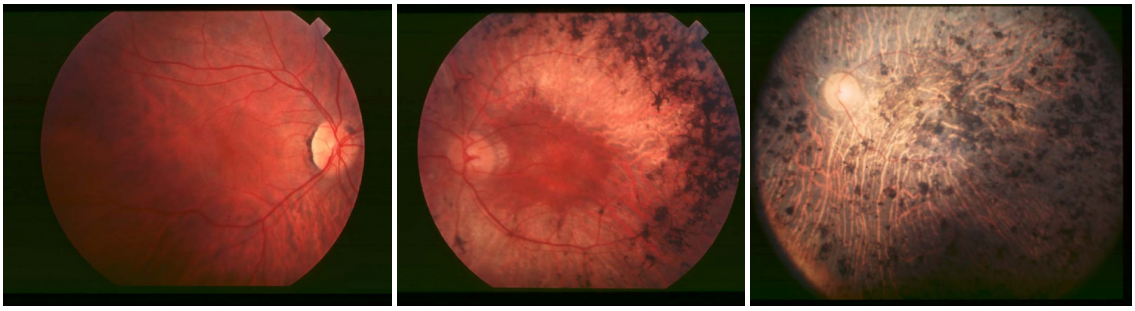


Figure 3. Fundus of patient with retinitis pigmentosa from left to right: at early stage (similar to a normal fundus), mid stage (retinal atrophy and bone spicule pigmented deposits at the periphery) and end stage (the whole retina is depigmented and deposits are present all over). From [10].

An objective tool to provide a reliable diagnosis of RP and to monitor the progression of the disease is the electroretinogram (ERG), which measures the electric response of the retina after stimulation with a short flash of light. The stimuli are applied after dark adaptation and consist of a flash of dim blue light to measure the function of the rods, a brighter white light (0.5 Hz), which stimulate both cones and rods, and flickering white flashes (30 Hz) to

stimulate cones only (**Fig. 4**). With the single white light stimuli, in patients with retinitis pigmentosa the amplitude of the hyperpolarizarion of the photoreceptors (a wave) and of the depolarizarion of the bipolar cells (b wave) is reduced; with flickering white flashes, the response is reduced and delayed (**Fig. 4**) [11].

Genetics plays an important role in the determination of different types of RP even though a correlation between a specific mutation and a phenotype is not always possible. In general, autosomal dominant forms are milder and appear at older ages, while autosomal recessive and X-linked forms are more severe and have early onsets [7, 10]. In dominant patients, the amplitude of the a and b waves is usually moderately reduced while in recessive and X-linked patients is severely reduced or undetectable (**Fig. 4**).

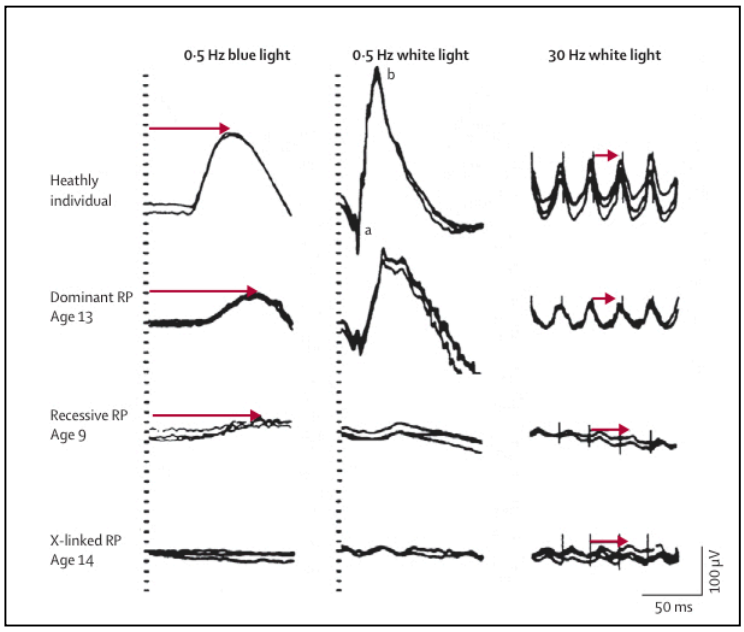


Figure 4. ERG responses from a normal subject and three patients with retinitis pigmentosa of autosomal dominant, recessive or X-linked inheritance. a=a wave, b=b wave. Arrows indicate the time from the stimuli (vertical bar) to the response. From [7].

1.2.3 Genetics

One of the most distinctive elements of RP is its genetic heterogeneity. Non-syndromic RP alone is caused by 56 different known genes, among which 23 genes cause adRP, 36 cause arRP and 3 cause X-linked RP. For many of these genes, several mutations have been reported, and their sum reaches the value of ~3,000 (**Table 1**) [12].

Function	Retina specificity	Symbol	Location	Protein	Type of retinitis pigmentosa	Reported pathogenic changes
Cell-cell interaction	-	***SEMA4A***	1q22	Semaphorin 4A	Autosomal dominant	3
Intermediary Metabolism	-	***CA4***	17q23.2	Carbonic anhydrase IV	Autosomal dominant	6
Intermediary Metabolism	-	***IMPDH1***	7q32.1	Inosine monophosphate dehydrogenase 1	Autosomal dominant	14
Phototransduction	+	***GUCA1B***	6p21.1	Guanylate cyclase activating protein 1B	Autosomal dominant	3
Pre-mRNA splicing	-	***PRPF3***	1q21.2	Human homolog of yeast pre-mRNA splicing factor 3	Autosomal dominant	3
Pre-mRNA splicing	-	***PRPF6***	20q13.33	Human homolog of yeast pre-mRNA splicing factor 6	Autosomal dominant	2
Pre-mRNA splicing	-	***PRPF8***	17p13.3	Human homolog of yeast pre-mRNA splicing factor C8	Autosomal dominant	21
Pre-mRNA splicing	-	***PRPF31***	19q13.42	Human homolog of yeast pre-mRNA splicing factor 31	Autosomal dominant	65
Pre-mRNA splicing	-	***RP9***	7p14.3	RP9 protein or PIM1-kinase associated protein 1	Autosomal dominant	2
Pre-mRNA splicing	-	***SNRNP200***	2q11.2	Small nuclear ribonucleoprotein 200 kDa (U5)	Autosomal dominant	7
Protein chaperones and degradation	-	***KLHL7***	7p15.3	Kelch-like 7 protein (*Drosophila*)	Autosomal dominant	3
Protein chaperones and degradation	-	***TOPORS***	9p21.1	Topoisomerase I binding arginine/serine rich protein	Autosomal dominant	8
Retinal Development	+	***CRX***	19q13.32	Cone-rod otx-like photoreceptor homeobox transcription factor	Autosomal dominant	51
Outer segment structure	+	***PRPH2/RDS***	6p21.1	Peripherin 2	Autosomal dominant; digenic with ROM1	123
Outer segment structure	+	***ROM1***	11q12.3	Retinal outer segment membrane protein 1	Autosomal dominant; digenic w/ PRPH2	11
Ciliary trafficking and structure	+	***RP1***	8q12.1	RP1 protein	Autosomal dominant; autosomal recessive	67
Ion channels	+	***BEST1***	11q12.3	Bestrophin 1	Autosomal dominant; autosomal recessive	232
Phototransduction	+	***RHO***	3q22.1	Rhodopsin	Autosomal dominant; autosomal recessive	161
Retinal Development	+	***NR2E3***	15q23	Nuclear receptor subfamily 2 group E3	Autosomal dominant; autosomal recessive	45
Retinal Development	+	***NRL***	14q11.2	Neural retina lucine zipper	Autosomal dominant; autosomal recessive	14
Visual cycle	+	***RDH12***	14q24.1	Retinol dehydrogenase 12	Autosomal dominant; autosomal recessive	66
Visual cycle	+	***RPE65***	1p31.2	Retinal pigment epithelium-specific 65 kDa protein	Autosomal dominant; autosomal recessive	134
Cell-cell interaction	+	***CRB1***	1q31.3	Crumbs homolog 1	Autosomal recessive	183
Ciliary trafficking and structure	+	***CLRN1***	3q25.1	Clarin-1	Autosomal recessive	23
Ciliary trafficking and structure	+	***FAM161A***	2p15	Family with sequence similarity 161 member A	Autosomal recessive	6
Ciliary trafficking and structure	+	***MAK***	6p24.2	Male germ-cell associated kinase	Autosomal recessive	9
Ciliary trafficking and structure	+/-	***TTC8***	14q32.11	Tetratricopeptide repeat domain 8	Autosomal recessive	14
Ciliary trafficking and structure	+	***TULP1***	6p21.31	Tubby-like protein 1	Autosomal recessive	31
Ciliary trafficking and structure	+	***USH2A***	1q41	Usherin	Autosomal recessive	392
Ciliary trafficking and structure	+	***C2ORF71***	2p23.2	Chromosome 2 open reading frame 71	Autosomal recessive	13
Ciliary trafficking and structure	-	***C8ORF37***	8q22.1	Chromosome 8 open reading frame 37	Autosomal recessive	4
Extracellular matrix	+	***EYS***	6q12	Eyes shut/spacemaker (*Drosophila*) homolog	Autosomal recessive	118
Extracellular matrix	+	***IMPG2***	3q12.3	Interphotoreceptor matrix proteoglycan 2	Autosomal recessive	10
Intermediary Metabolism	-	***IDH3B***	20p13	NAD(+)-specific isocitrate dehydrogenase 3 beta	Autosomal recessive	2
Ion channels	+	***CNGA1***	4p12	Rod cGMP-gated channel alpha subunit	Autosomal recessive	8
Ion channels	+	***CNGB1***	16q13	Rod cGMP-gated channel beta subunit	Autosomal recessive	6
Lipid metabolism	-	***CERKL***	2q31.3	Ceramide kinase-like protein	Autosomal recessive	8
Outer segment structure	+	***PROM1***	4p15.32	Prominin 1	Autosomal recessive	9
Phagocytosis	-	***MERTK***	2q13	c-mer protooncogene receptor tyrosine kinase	Autosomal recessive	27
Phototransduction	+	***PDE6A***	5q33.1	cGMP phosphodiesterase alpha subunit	Autosomal recessive	16
Phototransduction	+	***PDE6B***	4p16.3	Rod cGMP phosphodiesterase beta subunit	Autosomal recessive	39
Phototransduction	+	***PDE6G***	17q25.3	Phosphodiesterase 6G cGMP-specific rod gamma	Autosomal recessive	1
Phototransduction	+	***SAG***	2q37.1	Arrestin (s-antigen)	Autosomal recessive	11
Protein glycosylation	-	***DHDDS***	1p36.11	Dehydrodolichyl diphosphate synthetase	Autosomal recessive	1
Retinal Development	-	***ZNF513***	2p23.3	Zinc finger protein 513	Autosomal recessive	1
unknown	+	***PRCD***	17q25.1	Progressive rod-cone degeneration protein	Autosomal recessive	2
unknown	-	***SPATA7***	14q31.3	Spermatogenesis associated protein 7	Autosomal recessive	15
Visual cycle	+	***ABCA4***	1p22.1	ATP-binding cassette transporter—retinal	Autosomal recessive	680
Visual cycle	+	***LRAT***	4q32.1	Lecithin retinol acyltransferase	Autosomal recessive	10
Visual cycle	+	***RBP3***	10q11.22	Retinol binding protein 3, interstitial	Autosomal recessive	2
Visual cycle	+	***RGR***	10q23.1	RPE-retinal G protein-coupled receptor	Autosomal recessive	7
Visual cycle	+	***RLBP1***	15q26.1	Retinaldehyde-binding protein 1	Autosomal recessive	20
Ciliary trafficking and structure	-	***OFD1***	Xp22.2	Oral-facial-digital syndrome 1 protein	X-linked	127
Ciliary trafficking and structure	-	***RP2***	Xp11.23	Retinitis pigmentosa 2 (X-linked)	X-linked	76
Ciliary trafficking and structure	-	***RPGR***	Xp11.4	Retinitis pigmentosa GTPase regulator	X-linked	151

Table 1. List of genes linked to retinitis pigmentosa, as reported in the Retnet database (sph.uth.edu/retnet). Modified from [12] with classification modified from [6].

Some of these genes, such as *NRL*, *RP1* and *RHO* can cause both adRP and arRP, depending on the specific mutation involved. Mutations in the rhodopsin gene for example, although they cause the majority of adRP cases (**Fig. 5A**), can also account for a small percentage of arRP cases. Genes involved in RP are in some cases causing also syndromic forms of RP or other retinal degeneration. As an example, recessive mutations of the usherin gene *USH2A* can cause Usher syndrome but also underlie a large fraction of non-syndromic arRP cases.

Taken individually, the mutations in RP genes are very rare in the general population (minor allele frequency < 0.01) and they are usually highly penetrant, consistently with a mendelian monogenic disease [6]. However, low penetrance mutations and modifiers genes have been reported to play a role in RP and may even contribute to a large fraction of unresolved cases [13, 14]. The remarkable genetic and allelic variability of RP implies that each gene, with the exception of the most frequently mutated ones (*RPGR*, *RHO* and *USH2A*), contributes only to a small percentage of cases, and each allele - or particular mutation - even less. This has important consequences in the identifications of causal genes for diagnostics and also for the discovery of novel disease genes, which are predicted to underlie most of RP cases that still lack a molecular diagnosis.

The high number of recessive alleles, each one contributing to a small fraction of RP cases, implies also a high number of carriers in the healthy population. If we consider the totality of recessive RP mutations in all known genes, it results that at least one unaffected individual out of five carries a null RP mutation in heterozygous state. Nevertheless, RP is still a rare disease because the chance of having two mutations in the same RP gene is very low, and digenic inheritance is a rather infrequent mechanism [9, 15].

From the functional point of view, the classes of gene products implicated in of RP are also very heterogeneous. Many disease genes encode for proteins important for the specific functions and structures of rods photoreceptors, or can be also present in cones and RPE. These include proteins of the phototransduction cascade and signaling (*RHO, PDE, GUCA1B, CNGA1, CNGB1, SAG*), cytoskeleton proteins of the outer segment (*RDS, ROM1*), proteins of the connecting cilium (*C2orf71, C8orf37, CLRN1, FAM161A, MAK, OFD1, RP1, RP2, RPGR, TTC8, TULP1, USH2A)* and retina specific transcription factors (*NRL, NR2E3, CRX*) (**Table 1**). RP causing genes that are expressed in the RPE are involved in retinol metabolism and phagocytosis of the discs of the outer segment. Other RP genes do not localize specifically to the photoreceptors, but participate in more ubiquitous processes of all eukaryotic cells such as intermediate metabolism and pre-mRNA splicing. However, these processes are particularly

relevant for photoreceptors, which need intense activity of mRNA and protein synthesis for a massive production of proteins (rhodopsin, among all), therefore requiring high energy production and high rate and high fidelity of transcript processing.

Mutations in splicing factor genes have particular relevance in dominant RP, as they represent a large fraction of cases, second only to mutations in the rhodopsin gene (**Fig. 5A**). Moreover, among the functional categories of RP genes (as defined in **Table 1**), this is the largest one (**Fig. 5B**). It is therefore very likely that mutations in other genes belonging to this family may provoke RP. This is the assumption that drove the candidate gene screenings described in this research thesis.

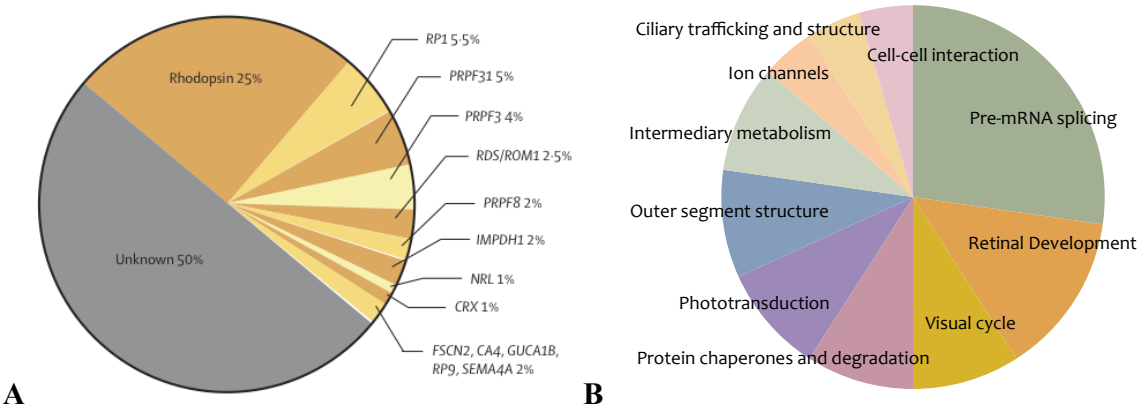


Figure 5. Genes causing autosomal dominant RP. **A)** Contribution of each gene to adRP cases [7]. **B)** Functional classification of genes causing adRP, according to [6].

2. SPLICING DEFECTS IN DOMINANT RETINITS PIGMENTOSA

2.1 The splicing machinery

2.1.1 Splicing catalysis

In the vast majority of eukaryotic genes, the protein-coding sequences, located in exons, are interrupted by non-coding regions, the introns, which are removed from the primary transcript (pre-mRNA) by the process of RNA splicing. On average, mammalian genes have 7-8 exons, of relatively short length (10-400 bp) and intervening introns that can be several kilobases long. The evolutionary meaning of such architecture can be linked to the role of exons in the creation of new genes by their duplication and diversification, each one representing a functional block [16]. In higher eukaryotes this allows obtaining different protein products staring from the same transcript, according to specific conditions or tissues. This process is called alternative splicing and concerns 90% of human intron-containing transcripts [17].

Splicing occurs in the nucleus, together with other modifications of the pre-mRNA: the capping at the 5' end and the polyadenylation at the 3' end. After this, the mature RNA is transported to the cytoplasm where can be translated. The specific removal of introns is ensured by *cis*-acting elements, i.e. consensus sequences and recognition sites, and trans-acting elements, i.e. the proteins composing the splicing machinery.

The consensus sequences necessary for intron removal are very short and conserved. They are located in the immediate surrounding of exon-inton boundaries and they define the splice sites (ss). The consensus sequence of the 5' ss, or donor site, is ag**GU**RAGU (exonic nucleotides are in lower case), while the 3' ss, or acceptor site is defined by the sequence Y**AG**. The GU and AG nucleotides are the most conserved elements among all introns. In metazoans, the other important recognition sequences in the introns are the stretch of pyrymidines that precedes the 3' ss and the so-called branch site (consensus YNCUR**A**C), 18-40 nucleotides upstream the 3' ss. In yeast, these consensus sequences are less degenerated and more conserved (**Fig. 6B**) [18].

The excision of the introns and joining of contiguous exons are achieved by two transesterification reactions (**Fig. 6A**). In the first step, a cut at the 5' ss separates the left exon from the right intron-exon molecule, whose new 5' end binds to the 2' –OH of the adenylated residue of the branch-site. Such structure is called *lariat*, because of its shape, and is formed by the nucleophilic attack of the 2-OH to the 5'ss, resulting in the atypical phosphodiesterase bond 5'-2' inside the intron. In the second step, the cleavage at the 3'ss releases the intron in

its lariat form and at the same time the two exons are joined together. This occurs through another nucleophilic attack by the 3'-OH of the 5' exon to the 3'ss.

RNA alone can catalyze the chemical reactions of splicing, since the transestereification reactions disrupt and create the same number of bounds with no energy consumption. This is well demonstrated by the existence of self-splicing RNAs [19]. However, the process needs to be fueled and coordinated by trans-acting elements, which are organized in a dynamic machinery [20]. They help *cis*-elements to come in close proximity for the reactions to take place and have a fundamental role in insuring accuracy, speed and regulation of intron removals. *Trans*-acting elements are both proteins and small RNA molecules, which form an elaborate network of dynamic interactions within a complex known as spliceosome.

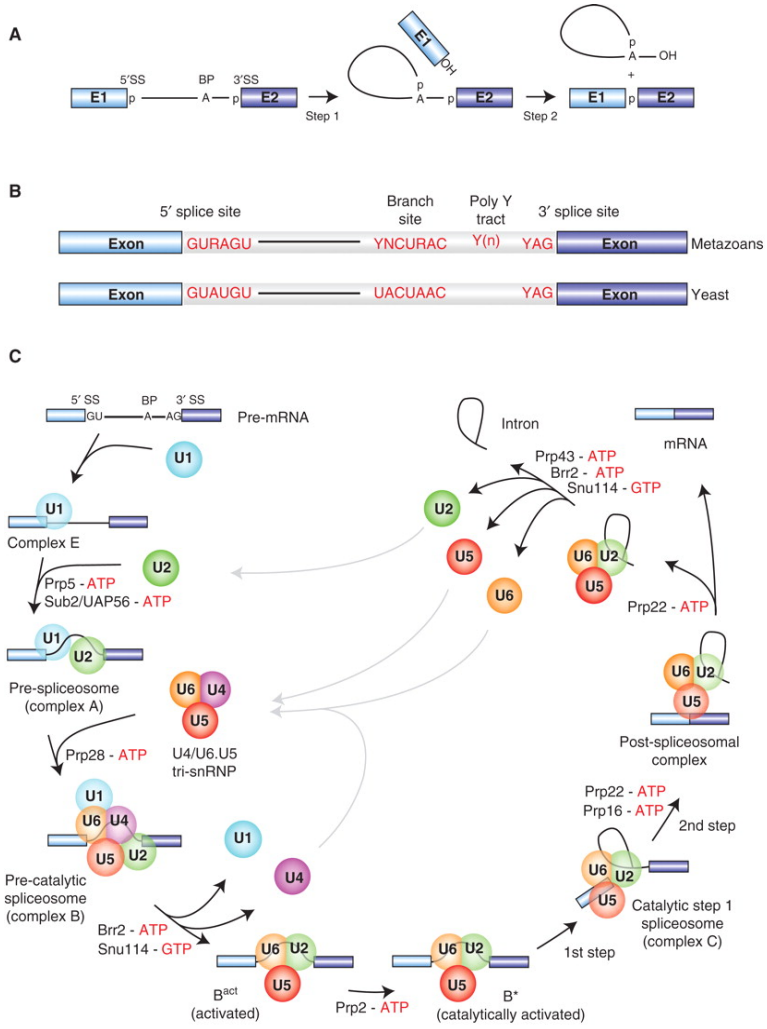


Figure 6. Pre-mRNA splicing of U2-type introns. **A)** Scheme of the two transesterification reactions necessary for intron removal. E= exon, A=branch point, p=phosphate groups of the splice sites (SS). **B)** Consensus sequences at the splice sites and branch site in metazoans and *S. cerevisiae*. Y=pyrimidine, R=purine. **C)** Steps of spliceosome assembly and disassembly during splicing. It is indicated when the DExH/D-box RNA ATPases/helicases are required for conformational rearrangements. Taken from [23].

2.1.2 Spliceosome assembly

The splice sites and branch sites sequences are recognized by the components of the spliceosome, which assembles on the pre-mRNA in an ordered and stepwise manner, before the two splicing reactions occur. The core components of the spliceosome are five uridine-rich small nuclear ribonucleoproteins (snRNPs) – U1, U2, U4, U5, and U6. Each one of these particles is composed of one small nuclear RNA (snRNA) molecule, and eight associated proteins (Sm proteins) that are common to all five snRNPs. A small fraction of introns found in vertebrates and invertebrates (only 800 in humans [21]) have different *cis* elements that need another type of machinery for their removal. This is the so-called minor spliceosome (or U12 dependent spliceosome), to differentiate it from the major spliceosome (or U2-dependent spliceosome), and contains different homologous snRNPs: U11, U12, U4atac, U6atac [22]). Additionally, each snRNP has a set of exclusively associated proteins, which are 70 in total, plus another 30 proteins that join at different stages. Once assembled, the spliceosome is a massive (12 MDa) complex, which resembles the ribosome for size and composition.

The first step of spliceosome assembly (**Fig. 6C**) is the recognition of the 5'ss by the U1 snRNP via the base paring of the consensus sequence within the splice site and the 5' hanging end of the U1 snRNA (complex E, for early pre-splicing). In the second stage, complex A is formed when U2 snRNP binds to the branch site, also by sequence complementarity between the two RNA elements. From this step on, the pre-mRNA is committed to splicing and the different rearrangements require ATP hydrolysis. Binding of the pre-assembled U4/U6.U5 (so-called tri-snRNP) complex converts complex A to complex B, which is activated (B^{act}) after conformational rearrangements that lead to the release of U1 and U4 from the complex. This event is due to the active disruption of the extensive base pairing engaged within the snRNAs of U4 and U6, allowing this latter to bind to the U2 snRNA instead, and bringing the two splice sites in juxtaposition. The catalytic active site is now created and transesterification reactions can occur (complex C). The spliced mRNA is released at first and after the disassembly of the U2/U6 and U5 complex, also the lariat is released, linearized and degraded.

In addition to the snRNPs, many proteins are involved in splicing processes: some of them have a direct role in splicing catalysis, but the majority of them is important for the conformational changes within snRNPs and regulation of spliceosome assembly. Some well-characterized proteins belong for example to the SR family, an important group of Arg-Ser rich splicing factors and regulators, important for the recognition of the splice site and formation of the early spliceosome complex [18].

Another class of proteins is composed of DExH/D-box RNA ATPases/helicases (in yeast: Prp5, Sub2/UAP56, Prp28, Brr2, Prp2, Prp16, Prp22 and Prp43), essential for timing and fidelity of the splicing process. They unwind short RNA helices in ATP-dependent manner and promote structural rearrangements of RNA and RNA-protein substrates [24].

Mutational and biochemical studies performed in the *S. cerevisiae* have identified and elucidated the function of the fundamental proteins for the correct functioning and dynamics of the splicing machinery, which are largely conserved in metazoans. Some of the genes encoding for these proteins have been found to be mutated in humans and to give rise to diseases.

2.1.2 Splicing and human diseases

According to the Human Genome Mutation Database (HGMD), 15% of reported mutations are predicted, and in some instances confirmed, to interfere with splicing of the affected transcript since they disrupt or create splicing consensus sequences [25]. In most cases *cis*-acting splicing mutations reduce or disrupt 5' or 3' ss strength and determine exon skipping or, less frequently, intron retention. Other mutations can activate cryptic splice sites that normally are not used. The final effect is often the disruption of the frame of the translation and the creation of a premature termination codon [21].

Conversely, trans-acting mutations affecting proteins that are part of the splicing machinery are much less frequent, probably because they undergo selection due to early lethality phenotype. Alterations of expression of several splicing factors have been frequently found in cancer, and recurrent somatic mutations affecting spliceosome components have been found in different kinds of leukemia [26]. Splicing factor mutations have been reported for a number of inheritable neurodegenerative diseases such as spinal muscular athophy (SMA), myotonic dystrophy (DM) and retinitis pigmentosa. In SMA, motoneurons are specifically affected due to mutations or deletions of the SMN1 gene, important for the assembly of snRNPs [27]. In DM, microsatellite expansions in untranslated regions of *DMPK* and *ZNF9* genes determine toxic gain of function of the mRNAs, which sequesters the MBNL splicing regulator, leading to splicing defects [28]. Finally, an emerging group of Mendelian pathologies caused by splicing factors is composed of craniofacial syndromic malformations like Nager syndrome [29], mandibulofacial dysostosis (MD) [30] and oesophageal atresia (OA) [31].

However, the most studied splicing factor-linked disease is autosomal dominant RP, for which mutations in genes of the U4/U6.U5 tri-snRNP complex account for 11% of all cases.

2.2 Pre-mRNA splicing factors and autosomal dominant retinitis pigmentosa (adRP)

2.2.1 AdRP-linked tri-snRNP components and mutations

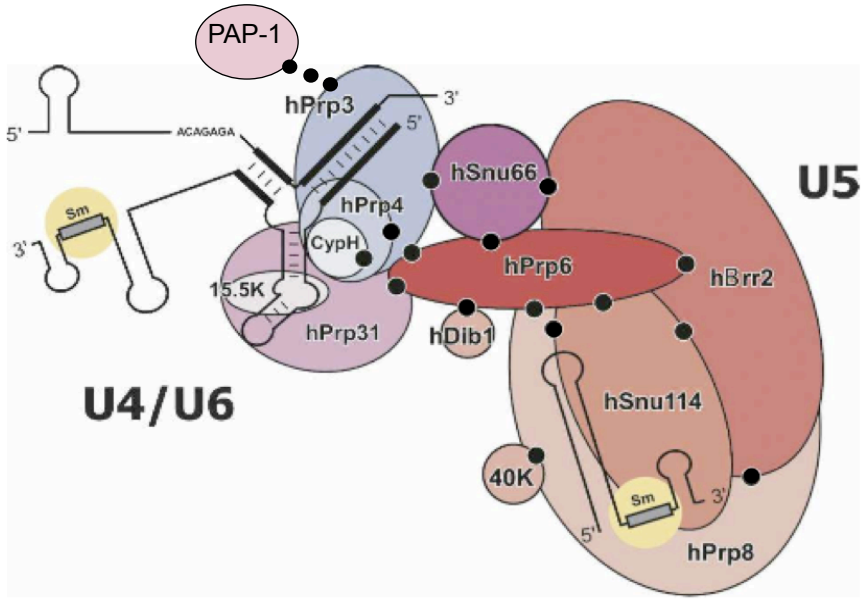


Figure 7. Scheme of protein-protein interactions within the human tri-snRNP. The dots indicate demonstrated interactions between proteins. Modified from [32].

So far, six splicing factors have been found mutated in adRP patients: *PRPF31* (RP11) [33], *PRPF8* (RP13) [34], *PRPF3* (RP18) [35], *PAP-1* (RP9) [36], *SNRNP200* (RP33) [37, 38] and *PRPF6* [39]. They are all components of the U4/U6.U5 tri-snRNP complex (**Fig. 7**), which is assembled before its recruitment on the complex A.

PRPF31 is the most frequently mutated RP gene among the splicing factor category, the second most common of all dominant cases [7, 40]. It was discovered by screening families with adRP linked to the RP11 locus on chromosome 19 [13, 33]. More than 60 mutations in several exons, including missense, frameshift, deletion, nonsense and splice site changes have been identified, but a recurrent mechanism of haploinsufficiency seems to be the main way of action of *PRPF31* mutations [41]. This is related to another particular feature of *PRPF31*-linked RP, which is the correlation of the phenotype with the expression level of the wild type protein [42] [43]. Within a same family, carriers of *PRPF31* mutations can either be asymptomatic lifelong or be severely affected at young age, because of a mechanism of incomplete penetrance [13]. *PRPF31* encodes a 61-kDa protein homologous to yeast Prp31p, a stable component of the U4/U6 di-snRNP. In yeast, it plays a key role in the recruitment of the

tri-snRNP on the pre-spliceosome [44] and in humans is required for the tri-snRNP assembly and spliceosome activity [45]. Silencing of *PRPF31* in HeLa cells was shown to inhibit tri-snRNP assembly, which accumulates in Cajal bodies and block pre-mRNA splicing, eventually leading to cell death [46]. The complete depletion of Prp31p in yeast [44], mouse [47] and zebrafish [48] is lethal.

The *PRPF8* gene was identified as causative for adRP after the discovery of the RP13 locus by linkage mapping [34, 49]. Mutations of the *PRPF8* gene determine a severe phenotype with early onset of night blindness and considerable reduction of the visual field. Although the phenotype is relatively uniform, it was lately reported that two families with adRP linked to two different *PRPF8* mutations showed variability in RP phenotype and incomplete penetrance, respectively [50]. The human PRPF8 protein, similarly to the yeast ortologous Prpf8p, is the core of the U5 snRNP: it interacts with all main pre-RNA splicing signals (5' and 3' ss as well as the branch site) and mediates interactions among the spliceosomal components, coordinating the rearrangements necessary for the formation of the catalytic centre. Moreover, it is required for the formation of the U4/U6.U5 tri-snRNP [51]. PRPF8 residues affected by RP mutations cluster in exon 42, encoding for the C terminal of the protein [52, 53]. This region was demonstrated to interact with an ATP-dependent helicase, Brr2p in yeast [54], hBrr2 in humans [55]. This 200-kDa enzyme is composed of two tandem helicase domains and is essential for the unwinding of the U4/U6 RNA duplex, which is the key step for the formation of the catalytic complex. Crystal structure of a hBrr2 fragment in complex with PRPF8 C-terminal domain demonstrated that PRPF8 extreme tail inserts into hBrr2's RNA-binding pocket, serving as an intermittent repressor of its ATPase activity and probably preventing premature unwinding of U4/U6 [56]. In the same study the authors also correlate the position of PRPF8 point mutations with molecular mechanisms involved in splicing impairment. Mutations located in the proximal part of the C- terminal domain were shown to inhibit U4/U6-U5 tri-snRNP formation, because of a reduced solubility or reduced affinity for Brr2 [54, 57]. Conversely, mutations located in the most extreme terminal do not interfere with U4/U6-U5 tri-snRNP formation but disrupt PRPF8 repression on hBrr2.

Very interestingly, hBrr2 has been recently implicated in retinitis pigmentosa. The *SNRNP200* locus, encoding for hBrr2, was discovered in a Chinese family, after linkage analysis pointing to the RP-33 locus on chromosome 2 [37, 38, 58]. RP33-linked family presented with variable RP phenotypes with relatively late onset of night blindness (16-18 years), and a full penetrant, autosomal dominant inheritance. The first two mutations found in

two Chinese families (S1087L and R1090L) locate to the ratchet helix of the first helicase domain and were shown to impair Brr2 mediated U4/U6 unwinding [38]. Other mutated residues are located in different regions of the first helicase domain [59] and may be important for the stabilization of domains folding [60].

PRPF3 mutations were identified in adRP families with linkage to the RP18 locus [35]. PRPF3 protein is specifically associated with the U4/U6 snRNP and interacts with another splicing factor, PRPF4. It is required for the assembly and activation of the spliceosome, although its specific role is not yet clear [61]. Overexpressed PRPF3 proteins carrying the hotspot mutation T494M in photoreceptor cells form large aggregates inside the nucleus that lead to apoptosis of photoreceptor cells, but not of epithelial or fibroblast cells [62]. The three missense mutations identified so far cluster to a conserved C-terminal domain, next to binding sites for other splicing factors. PAP-1 is one of these interactors, but its precise function is currently not known [63]. This gene was found to be involved in adRP as well, mapping to the RP9 locus [36, 64] and two mutations were found. Also in the case of RP9 mutations, variability in phenotype expressivity and incomplete penetrance were observed [65].

PRPF6 is the latest tri-snRNP associated gene found to be causative for adRP [39]. The protein is thought to have a role in protein scaffolding between the U5 snRNP and the U4/U5 di snRNP [66]. The mutation was identified by a candidate gene screening study on a cohort of American adRP patients and it was shown to cause accumulation of the mutated proteins within the Cajal bodies in the nucleus of patients' lymphoblasts and HeLa cells transfected with the mutated construct. A possible impairment in the tri-snRNP assembly or recycling was suggested [39].

Independently from the way of action of each mutation, the final effect of tri-snRNP associated mutations is the specific degeneration of rod photoreceptors.

2.2.3 Why the retina?

Considerable effort has been put in the elucidation of the mechanisms that would link mutations in the essential and ubiquitously expressed protein of the tri-snRNP complex to retina-specific phenotypes. Two main hypotheses have been addressed: one implies that RP mutations would affect splicing in the retina by disruption of specific protein-protein interaction or by affecting retina specific transcripts. A second one sees RP as a disease triggered by the lower threshold of tolerance for splicing defects in photoreceptor cells compared to other tissues.

Experimental approaches aimed at studying the effects of these mutations include the use of biochemistry techniques, in vitro assays for splicing efficiency assessment and animal models. Two knock-in mice for *PRPF3* and *PRPF8* missense mutations and a heterozygous knockout mouse mimicking haploinsufficient *PRPF31* mutations were generated, but did not provide conclusive mechanistic answer. Mice showed a certain degree of retinal degeneration, primarily affecting the RPE with a late onest, especially for *PRPF31* mouse model [47, 67].

Analysis of the eyes of a zebrafish model of *PRPF31* knockdown showed defects in photoreceptors morphology and visual processing, and a down-regulation of retina-specific mRNAs, supporting the hypothesis that reduced levels of tri-snRNP proteins could specifically affect the splicing of genes important for retinal functions [48]. However, a control transcriptomic analysis from a non-retina tissue was not provided. In other studies, which used as disease model lymphoblastoid cell lines from human patients with mutations in *PRPF3*, *31*, and *8*, aberrant splice products were detected in transcripts that do not relate with functions in the retina [68]. Also, in the same non-retina model, they observed impairments in spliceosome kinetics of assembly, protein and snRNPs composition, and in alternative splicing. These evidences suggest that general splicing defects are likely present in all tissues, but have deleterious effects only in the retina, as a consequence of its higher requirement for splicing with respect to the other tissues. This is supported by the fact that retina is the tissue in which a highest number of spliced genes is maximally expressed [68] including the *PRPF* genes themselves [69].

In agreement with this 'threshold' model, most of *PRPF* mutations seem to be hypomorphic, having as consequence a reduction in the quantity or in the activity of the proteins, rather than acting through dominant negative mechanisms. This is also supported by the many cases of incomplete penetrance - mainly associated to *PRPF31* - in which compensatory mechanisms for the function impairments have been observed [42].

Other non-verified hypotheses invoke unknown retina-specific functions or interactions of RP-linked splicing factors.

3. NEXT GENERATION SEQUENCING AS A NEW TOOL FOR GENE DISCOVERY IN RP

3.1 Next generation sequencing methods

3.1.1 Novelty and applications

The recent breakthrough in genetics since 2005 [70] has been the radical change of DNA sequencing methods. Next generation sequencing (NGS), ultra high throughput sequencing (UHTS) and massive parallel sequencing (MPS) are the terms designating such technologies. As their names suggest, the major novelty with respect to the Sanger sequencing - or dideoxy - method [71], which dominated the field since the 70s, is their extraordinary sequencing capacity and speed, currently associated with a 10,000 times lower per base cost [72]. As most paradigmatic example, the human genome can be sequenced now in about a few weeks of work, starting from a few micrograms DNA and a single sequencer machine, for less than 10 thousand dollars. Only 10 years ago, the sequencing of the human genome with traditional methods was achieved by a joint effort lasted 10 years and costed about 300 million dollars [73] [74] [75]. The possibilities offered by the huge throughput and low price of NGS are not limited to assemble genomes or to answer to "static" genetic questions, but inspired also a wide range of more "dynamic" applications, aimed at a better understanding of certain biological processes at the cell level. These latter applications include RNA-seq (sequencing of trasnscriptomes), ChIP-seq (sequencing of DNA fragment interacting with proteins), TRAP (purification of polysomal mRNA), Ribo-Seq (sequencing of ribosome-protected mRNA fragments), and many others [76].

Focusing back on human genetics questions, NGS transformed the approaches to comprehend the genetic basis of monogenic and complex diseases or cancer, with increasing implementation also in clinical diagnosis and in development of targeted therapies in the near future [77, 78]. NGS-based gene hunting has been particularly successful in identification of new genes associated to rare disorders and is predicted to solve most of Mendelian disorders in the next decade [79, 80]. Due to constant drop in costs, the sequencing of whole genomes (WGS) or whole known exonic sequences (WES) of an individual will probably replace all targeted approaches for gene identification (NGS based or not), such as sequencing of genes in large cohorts of patients or sequencing a linkage interval [81]. However, important ethical and practical issues still encourage many scientists to opt for targeted sequencing strategies. Among the practical issues, the bioinformatic analysis costs and knowledge represent one of

the main obstacles for whole-genome/exome approaches to become routine tools in research or diagnostic laboratories. Moreover, the analysis of a very restricted portion of the genome has the advantage to be more accurate, because each targeted region would be covered by a higher number of sequences than a whole exome or genome (the concept of *coverage* will be elucidated in the following section). Therefore, a series of DNA enrichment tools have been developed and commercialized to adapt the sequencing of small genomic portions to the incredible capacity of NGS instruments [82]. A review of these methods and examples of targeted approaches that I used in my study can be found in the publication section.

3.1.2 Chemistry

The "magic" of NGS relies on workflows leading to the sequencing of many templates in parallel. This is obtained through strategies based on cyclic-array sequencing, which is the sequencing of a dense array of DNA molecules by means of repetitive enzymatic reactions, and collection of data in the form of images [83]. First, DNA is randomly sheared into fragments of few hundred nucleotides, then fragments are ligated to universal adapters where primers for subsequent PCR amplification will anchor. The clonal amplification of these fragments is obtained using different approaches: emulsion PCR (used for instance by the 454-Roche and the SOLiD-Life Technologies systems) or bridge PCR (used by the Solexa-Illumina system). The first method is based on the amplification of single templates attached to beads in a water-in oil mixture, which are then distributed on a microarray of picoliters wells (**Fig. 8A**). The second method relies on PCR reactions occurring on a glass slide where primers are attached, generating clusters containing about 1000 copies of the same DNA fragment (**Fig. 8B**) [84].

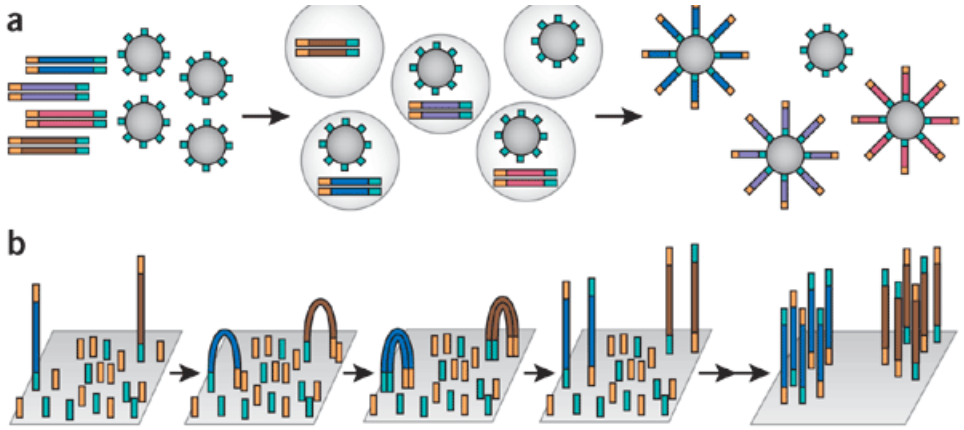


Figure 8. Schematic explanation of library amplification by the method of **A)** emulsion PCR and **B)** bridge PCR. From [84]

The cyclic reactions used for sequencing of such clusters of DNA clones employ different enzymatic processes. The 454-Roche instruments use pyrosequencing reactions, in which at each cycle, a single nucleotide species is added together with the substrate for the production of light when pyrophosphate is released, at wells where the incorporation of the base occurred (**Fig. 9A**) [70]. With the Solexa-Illumina instruments, a modified version of the Sanger method, using reversible-terminator, fluorescence-labeled dNTPs is implemented on a array setting (**Fig. 9B**) [85, 86]. As a last example, the SOLiD-Life technologies platforms employ yet another method, based on the ligation of fluorescently labeled octamers which anneal to the template with the 2 central bases, in multiple cycles of ligation and cleavage (**Fig. 9C**) [87].

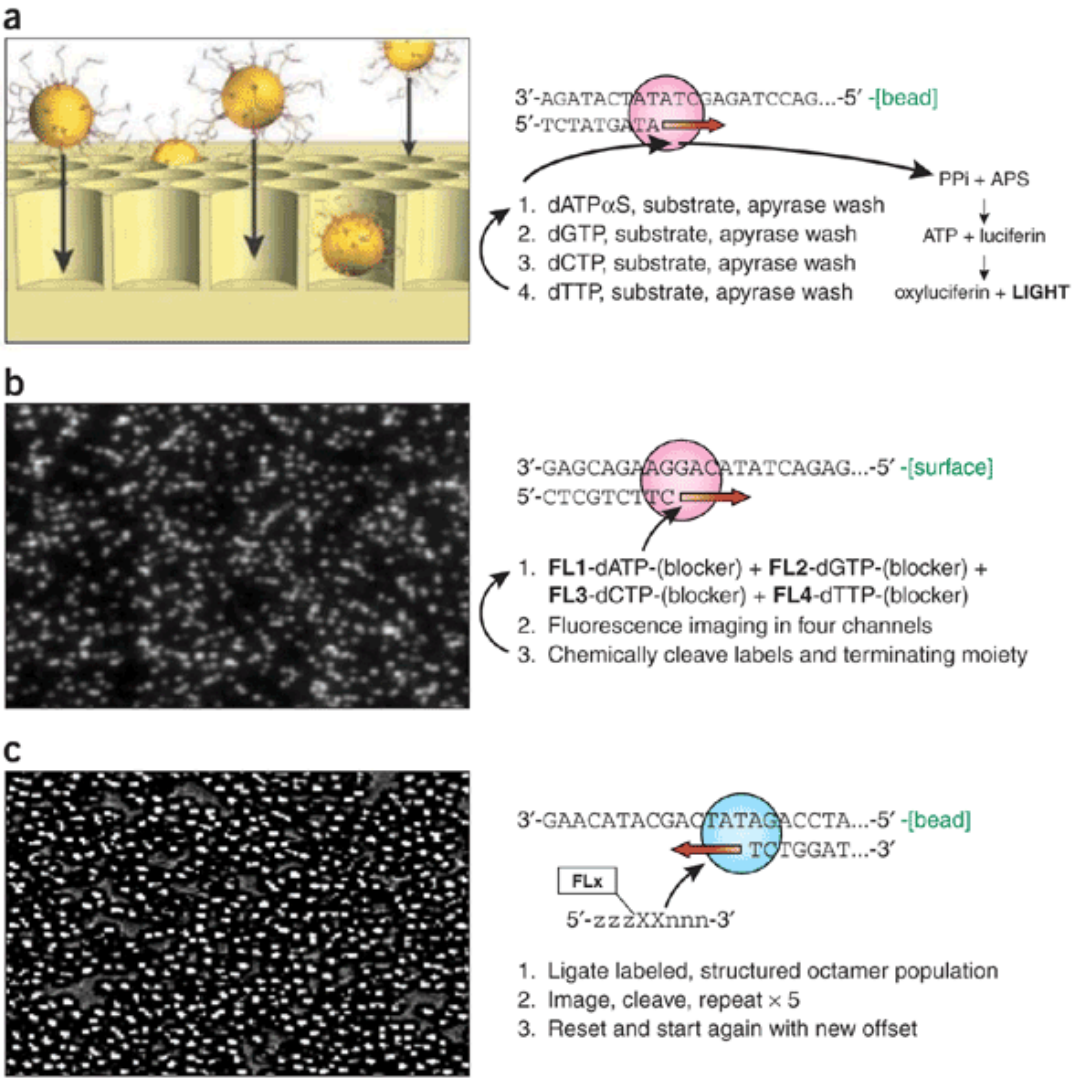


Figure 9 Examples of three different methods used in high throughput sequencing platforms: **A)** pyrosequencing, **B)** reversible terminators sequencing by synthesis, **C)** two-base encoding ligation method. From [84]

3.2 Definitions and interpretation of results

The analysis and interpretation of the data generated by NGS platforms require a certain familiarity with bioniformatic methods. This knowledge is rapidly diffusing also in laboratories and centers that are not specialized in computational biology, due to the progressive accessibility of NGS and to the development of bioniformatic pipelines with a graphic interface. The workflow for analysis of sequencing data generally follows the steps of alignment of the sequences and identification of differences between the sample and the reference, which in the context of medical genetics is important in order to identify pathogenic mutations. A summary of the terminology and the methods used in a NGS typical workflow for variation detection is listed below.

Reads. A list of sequence reads (or just *reads*) is the usable output of NGS sequencers, computed after the conversion of images into nucleotides. NGS reads are shorter than Sanger reads, since their length ranges between 50 and 400 nucleotides, depending on the platform used. The ambiguity that can derive from assembly of shorter reads is one of the weakest points of NGS with respect to Sanger sequencing. An improvement in the mapping of short reads is obtained with paired end reads, which, in contrast to single end reads, derive from both ends of a library of DNA fragments. The length of such fragment is known (generally 300-500 nucleotides) and helps to locate the paired sequences with respect to each other, as well as to detect insertion or deletions.

Accuracy. Each base listed in the output file - usually in a FASTQ format - is provided in association with a quality value, calculated upon parameters relevant to a particular sequencing chemistry. The accuracy of a base call is expressed with a Phred score, which indicates the probability that a base call is incorrect, with a logarithmic relation. For example a Phred of 30, which is the benchmark for quality in NGS, indicates a 1 in 1000 probability of incorrect base call, or, a 99.9% accuracy [88, 89].

Mapping. Alignment or mapping of the reads onto a genomic reference sequence is the first step in re-sequencing analyses. In absence of a reference genome for a given organism or gene, a *de novo* assembly of reads into a contig (i.e. a consensus DNA sequence) can be performed, but this is obviously more challenging than mapping each read to a known reference. Alignment algorithms can be generally classified as global or local. In the first case an optimal alignment between the reference and the full length of the read is searched, whereas in the second case highly similar sequences within a read are aligned to the reference even if some parts of it do not align. This is very useful for NGS in which base quality drops

at the ending of a read, and it is possible to automatically remove the non-matching errors at the end of the reads. Algorithms can be further categorized in gapped or un-gapped by whether they allow gaps in the sequences or not. Some commonly used open source programs for short reads alignments are: BWA, Bowtie, SOAP and MAQ. Others, commercially available, are Novoalign and CLC Genomics.

Coverage. A single read is not sufficient to determine accurately the sequence of the DNA of interest: it will contain an error at every given base (1000 if we consider a good Phred of 30, as mentioned before). To insure a reliable contig, and to allow confident identification of DNA variants, the target DNA must be sequenced at a higher depth or coverage, a safe threshold being at least 40 reads on average [90]. Different instruments have different throughputs, which adapt to a variety of applications requiring more or less coverage, depending on the size of the target DNA or on the number of samples to be sequenced. If more coverage is needed, multiple sequencing runs can be performed, if the target DNA is small, multiple samples can be sequenced in a same run.

Variant calling. After the alignment step, the sample DNA can be compared to the reference genome and differences can be identified. A major challenge is to distinguish between true variations and sequencing or mapping errors, which can be rather common especially in difficult sequence contexts like GC-rich or repetitive regions. An increase of coverage is usually sufficient to determine most of single nucleotide substitutions with high confidence, while it is not easy to correctly identify insertions and deletions (indels), especially if they are longer than a few nucleotides. Structural variants - including indels, copy number variations, inversions and translocations - present some difficulties for mapping algorithms, which must be able to integrate the presence of a big difference and still align it to the reference. Other parameters are taken into account for the identification of larger rearrangements and include significant alterations in the coverage or in the insert size in case of paired-end reads.

Filtering and annotation. Finally, the search of pathogenic variants is based on quality and functional criteria. Variants can be annotated with a number of elements such as predicted consequences at the protein level, presence in public polymorphism databases and position within the transcript. Annotation and filtering can be manual in case of short DNA sequences, but requires automatic procedures for analysis of large DNA sequences like exomes or genomes.

3.3 Implementation of NGS technology in RP genetic research and diagnostic

About half of cases diagnosed with RP have unsolved molecular diagnosis, since their screening for known RP genes has resulted negative [7]. It is likely that a large part of this estimate is composed by novel mutations in known genes. In fact, Sanger sequencing-based screenings of known genes is often restricted to the coding sequence or even only to mutational hotspots. Intronic mutations are therefore likely to be missed, although it has been shown that they can play a role in the pathogenesis of RP ([91, 92]. Mutation microarrays (arrayed primer extension or APEX- technologies) are useful to pre-screen patients for known mutations, but again novel base changes cannot be excluded [93]. Due to the lower costs and shorter time associated with NGS strategies, it is now possible to create sequencing panels of genes known to cause RP or other retinal degenerations. This is very useful in a clinical setting where many samples need to be processed in a short time and more importantly it can help to distinguish between different retinal degeneration types, whose phenotype is often overlapping [94-97].

The detection of novel disease genes is hampered by the genetic hetherogeneity of RP. Mutations in novel genes are in fact predicted to be very rare, although their number is hard to estimate [12]. A successful method that has led to the discovery of novel RP genes also in recent years has been linkage analysis or homozygosity mapping, very powerful in case of families with consanguinity history or originating from an isolated geographical region. However, the majority of RP cases are sporadic, therefore not suitable for linkage analysis. Genome-wide approaches like whole exome sequencing have already proven to be very effective methods to discover novel RP genes, also in absence of a linkage interval to guide the analysis [98]. On the other hand, whole genome sequencing is useful to detect changes lying outside commonly analyzed regions like intronic changes or structural variations. Such powerful methods, combined with novel approaches for analysis - for example by considering also non-monogenic or non-Mendelian mode of inheritance - will likely help to uncover the "missing heritability" of RP.

PUBLICATIONS AND MANUSCRIPTS

Project 1: **Comparison of NGS platforms in the detection of human DNA variants**

One of the first questions that we wanted to address about the use of NGS technologies was to determine which platform is the most suitable for analyzing human genes and variations. We specifically aimed at testing the three main platforms described in the introduction in a routine workflow of gene and variations analysis, by means of a user-friendly bioinformatic tool. The input for the project was given by the discovery of an RP dominant mutation with reduced penetrance, found within an intronic repetitive region of the *PRPF31* gene. This mutation was in fact missed by a first NGS experiment in our laboratory, because short reads could not uniquely align to repetitive sequences. Given this challenging peculiarity we considered it as a good benchmark for a comparative analysis. Furthermore, it gives an example of a mutation located in those non-coding regions that are normally excluded by the classical exon-PCR Sanger sequencing, but which should become more accessible thanks to NGS.

As template for sequencing, we used a pool of 4 long-range PCR products encompassing the entire gene and its vicinities, which represents the simplest and most unbiased strategy to target a limited region of the genome for NGS. The sequences obtained from the three different platforms were analyzed by a commercial software, which did not require specific competence in programming and allowed to treat the different outputs in a uniform way. For comparison purposes, we evaluated several technical aspects of the different platforms including the number of sequences produced, average coverage obtained, read accuracy, mapping accuracy, variant detection efficiency (SNPs, indels, mutation), and alignment to the repetitive element. We concluded that, at least for the specific region analyzed, all sequencing methods performed very well and that they could all be safely and efficiently used as a tool for the detection of targeted human DNA variations.

This study was published as "*Ultra high throughput sequencing in human DNA variation detection: a comparative study on the NDUFA3-PRPF31 region",* in PLoSONE journal on September 29, 2010.

Candidate's roles:

- Design and execution of the bioinformatic analyses.
- Writing of the manuscript.

OPEN ACCESS Freely available online

PLoS one

Ultra High Throughput Sequencing in Human DNA Variation Detection: A Comparative Study on the *NDUFA3-PRPF31* Region

Paola Benaglio, Carlo Rivolta*

Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland

Abstract

Background: Ultra high throughput sequencing (UHTS) technologies find an important application in targeted resequencing of candidate genes or of genomic intervals from genetic association studies. Despite the extraordinary power of these new methods, they are still rarely used in routine analysis of human genomic variants, in part because of the absence of specific standard procedures. The aim of this work is to provide human molecular geneticists with a tool to evaluate the best UHTS methodology for efficiently detecting DNA changes, from common SNPs to rare mutations.

Methodology/Principal Findings: We tested the three most widespread UHTS platforms (Roche/454 GS FLX Titanium, Illumina/Solexa Genome Analyzer II and Applied Biosystems/SOLiD System 3) on a well-studied region of the human genome containing many polymorphisms and a very rare heterozygous mutation located within an intronic repetitive DNA element. We identify the qualities and the limitations of each platform and describe some peculiarities of UHTS in resequencing projects.

Conclusions/Significance: When appropriate filtering and mapping procedures are applied UHTS technology can be safely and efficiently used as a tool for targeted human DNA variations detection. Unless particular and platform-dependent characteristics are needed for specific projects, the most relevant parameter to consider in mainstream human genome resequencing procedures is the cost per sequenced base-pair associated to each machine.

Citation: Benaglio P, Rivolta C (2010) Ultra High Throughput Sequencing in Human DNA Variation Detection: A Comparative Study on the *NDUFA3-PRPF31* Region. PLoS ONE 5(9): e13071. doi:10.1371/journal.pone.0013071

Editor: Kelvin Yuen Kwong Chan, The University of Hong Kong, China

Received June 24, 2010; **Accepted** September 2, 2010; **Published** September 29, 2010

Funding: This work was supported by the Swiss National Science Foundation (grant # 320030-121929) and the European Union (grant HEALTH-2007-201550). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: carlo.rivolta@unil.ch

Introduction

The recent commercialization of ultra high throughput sequencing (UHTS) technologies, initially applied to the *de novo* characterization of small genomes, is rapidly challenging the classical methods of human genetic research as well. The possibility of obtaining nucleotide sequences in the range of hundreds of millions base pairs from various types of DNA templates allows for example to extend mutational screenings to very large portions of the genome, an experimental strategy that would be too expensive and time consuming to perform with methods based on the Sanger procedure [1]. Thanks to UHTS, intronic and non-coding regions as well can theoretically be included in routine resequencing processes (i.e. the analysis of a DNA region for which a reference sequence is already known) of a particular candidate gene or linkage interval, with minimal additional costs and by a more complete approach with respect to classical exon-PCR and sequencing.

However, these "next-generation" technologies still have some limitations that must be taken into account. A well-recognized problem associated with the mapping of UHTS sequences is represented by the presence of repetitive elements or low-complexity stretches to which short UHTS reads cannot uniquely align [2,3]. To simplify assembly procedures of short sequencing reads, these DNA segments are therefore generally excluded, with the consequence of missing important disease-associated variants present in intronic or extra-genic areas.

Recently, we discovered a mutation (c.1347+654C>G) in one of these particular regions of the human genome associated with dominant retinitis pigmentosa, an hereditary blinding disease [4]. This single-base substitution is comprised in a repetitive element (variable number of tandem repeats, or VNTR) located within an intron of the *PRPF31* gene. As a proof of concept for UHTS to be used in routine human genetic screenings, we sequenced 31 kb of the human chromosome 19 encompassing the *PRPF31* region in a patient with this rare mutation as well as several common SNPs. For comparative purposes, we used the three currently most widespread UHTS platforms: Roche/454 GS FLX Titanium (Roche 454), Illumina/Solexa Genome Analyzer II (Illumina GA) and Applied Biosystems/SOLiD 3 (ABI SOLiD) instruments.

The Roche 454 technology is based on the clonal amplification of DNA fragments attached on individual beads in an emulsified PCR reaction. The beads are distributed on a 1.6 million wells substrate (PicoTiterPlate™) where pyrosequencing reactions

occur [5]. The most noticeable advantage of the Roche 454 platform is the large size of the reads produced (up to 500 nt), while Illumina GA and ABI SOLiD produce shorter reads (34 and 50 nt, at the time this research was performed). In the Illumina GA system the amplification step is achieved on the glass surface that covers the flow cell (bridge amplification) and the sequencing reactions are performed by using the "reversible terminator" chemistry [6]. ABI SOLiD is similar to Roche 454 in the amplification step (emulsified beads) but is unique for its ligase-dependent sequencing chemistry, based on multiple cycles of hybridization and ligation. The main advantage of ABI SOLiD is constituted by the possibility of reading each base twice by independent events, which provides internal error correction and enables higher accuracy, especially in SNP calling [7].

Materials and Methods

Ethics statement

This study was carried out in accordance with the tenets of the Declaration of Helsinki and was approved by the Institutional Review Boards of our University and of Harvard Medical School, where the blood was collected and the cell line derived. Written informed consent was obtained from the patient who participated in this study and donated her blood for research.

Sample preparation

We extracted DNA from a lymphoblastoid cell line derived from an affected individual carrying the *PRPF31* c.1347+654C>G mutation (cell line #13189) and amplified the 31-kb *NDUFA3-PRPF31* genomic region by 4 individually-amplified long-range PCR, designed as previously described [4] (Fig. 1). We specifically selected this region to have a well characterized reference sequence to compare our experimental results to. The following minor modifications were introduced to the original amplification protocol. Each PCR was performed in a final reaction volume of 10 µl, containing 1X GC Buffer I (TaKaRa, Otsu, Shiga, Japan), 0.4 mM dNTPs, 0.2 µM primers (each), 0.5 U of TaKaRa LA Taq (TaKaRa) and 100 ng of DNA. Such an amount of genomic template DNA allows virtually eliminating the possibility that errors introduced by the Taq polymerase are detected in subsequent sequencing procedures. Reactions were incubated at 94°C for 1 min, followed by 35 cycles of 98°C for 5 sec and 68°C for 17 min, and a final elongation step of 72°C for 10 min. After agarose gel analysis and quantification, the four PCR fragments were pooled together and processed for downstream applications.

Library preparation and sequencing

Preparation of DNA libraries was performed following the guidelines provided by the manufacturers of each platform and sequenced by using: 1/8 of a plate for the Roche 454 Genome Sequencer FLX, Titanium series, 1 lane of an Illumina Genome Analyzer version II, and 1 "quad" of an ABI SOLiD 3 instrument. The exclusion of reads with very low quality was performed automatically by the Roche 454 and Illumina GA sequencing instruments, while for ABI SOLiD this had to be carried out a posteriori with the ABI's csfasta_quality_filter.pl application, available from the SOLiD Software Development Community.

Alignment and analysis of reads

All analyses and statistics on quality-filtered reads were performed using the relevant tools of the software package CLC Genomics Workbench, version 3.7 (CLC bio, Denmark) as described below.

Trimming. In this process the parts of the reads with low quality scores were trimmed. The algorithm calculated base error probabilities based on their quality values, normalized to a PHRED scale. We set a cutoff value of 0.01, calculated as described in the software package manual, and discarded trimmed reads below 20 nt of length, independently from their residual score.

Assembly. The original reads as well as the trimmed sets were aligned to 31 kb of the corresponding reference sequence (NC_000019.8: 59,297,572-59,328,826). To ensure uniformity, we applied comparable settings to all platforms, considering the different read length of each platform, inclusive of the color-space option for the ABI SOLiD platform. Specifically, we used the local gapped alignment algorithm for all alignments, keeping the default parameters for mismatch and deletion costs. Reads that aligned to more than one position of the reference sequence were discarded.

For the intronic repetitive DNA fragment we also re-assembled the reads by using a *de novo* assembly procedure. The original reads were first aligned onto the Sanger-obtained sequence of the region by using the same parameters described above and by allowing random matches of reads with multiple mapping positions. Subsequently, we extracted the sequences that aligned to the region and used them for *de novo* assembly with the same parameters used for the reference assembly (no random matches).

Detection of variants. Variant detection was performed with the SNP and INDEL detection tools. The settings for calling a variant were described previously by Harismendy *et al.* [8]: if heterozygous, 20%–80% of the reads covering a particular nucleotide had to contain the alternative base with respect to the reference sequence; if homozygous, more than 80% of the reads had to contain the alternative allele. To test the limits of SNP detection, discovery by setting a minimum variant threshold of 10% was also performed. The minimum coverage allowed to call a SNP was of 15 reads for a given base. We applied the default restrictions on SNP calling: the average quality of the central base was set to 20 (PHRED score, corresponding to a base accuracy of 99%), the average quality

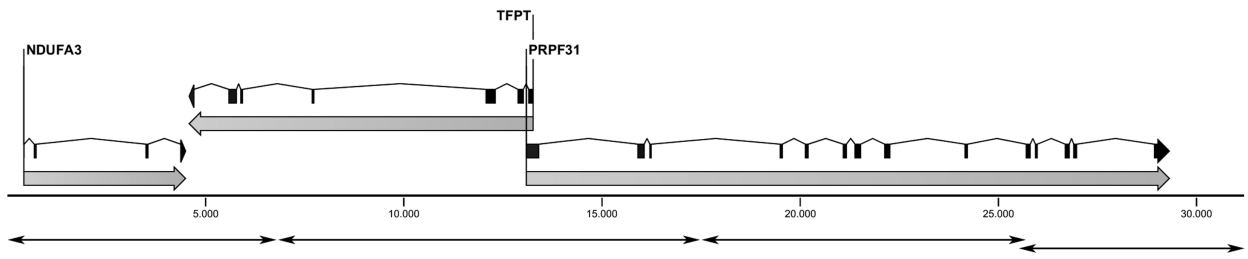


Figure 1. Schematic representation of the 31-kb genomic interval analyzed, containing the genes *NDUFA3*, *TFTP*, and *PRPF31*. Double-headed arrows indicate the position of the 4 amplicons used as template for UHTS.
doi:10.1371/journal.pone.0013071.g001

of the surrounding 10 bases was 20, and the maximum number of mismatches or indels accepted within an 11-nt window was 3. Low quality reads were removed from the calculation of SNP frequency and coverage and the un-aligned parts of a read counted as mismatches. We considered a variant as real if it was validated by Sanger or, in absence of Sanger validation, if it was found in at least one platform and previously annotated, or independently present in at least 2 platforms and not annotated. For the detection of indels we used the same criteria as for SNP detection.

Coverage simulation

We simulated different coverage depths by randomly sampling a subset of the reads from the *.fastq files exported after the trimming process. Each subset was sampled 3 times. Alignments to the reference sequence were performed as described above for the non-simulated sets of data. In order to have a balanced representation of the 4 amplicons, we calculated the average coverage at the level of the amplicons (and not of the entire region) and joined the amplicons with the same average coverage, considering them as a single (artificial) sampling event. For SNP detection we maintained the same parameters as for the full dataset, but with a minimum coverage of 5 and at least 2 reads carrying the variant allele, to compensate for the reduced coverage introduced by the simulations.

Sanger sequencing

Data from the Sanger sequencing of the *PRPF31* gene were available from previous analyses [4]. Additional SNPs located outside of the *PRPF31* gene were sequenced starting from the long-range PCRs used as UHTS templates or from short-range PCRs obtained using standard HotStartTaq DNA polymerase (Qiagen, Venlo, The Netherlands) protocols. PCR products were enzimatically purified using 1 µl ExoSAP-IT (USB, Cleveland, Ohio USA) for 10-µl reactions, according to the manufacturer's instructions. Sequencing reactions were performed by mixing 5 µl of purified PCR product, 0.75 µM of 20mer primers and 1 ul of BigDye Terminator v1.1 cycle sequencing kit (Applied Biosystems, Foster City, CA), and run on a ABI-3130XLS (Applied Biosystems).

Results

General considerations on the processing and analysis of the reads

All computer-based analyses were performed with a commercial, user-friendly software. This choice was taken in order to be as close as possible to the setup of the average laboratory performing routine genetic testing without the specific support of computer analysts. The use of a simple pipeline, compatible with outputs generated by different sequencing platforms, also allowed treating the data in a uniform manner, thus eliminating possible biases deriving from machine-specific software or algorithms.

Sequencing and trimming of the reads

For our analyses we used 1/8 of the total sequencing capabilities of each machine. The Roche 454 platform (1 sector of the 8-sector gasket) generated ~100,000 quality-passed reads with an average length of 318 nt, Illumina GA (1 lane) ~4.6 million reads of 34 nt, and ABI SOLiD (1 "quad") ~17,3 million reads of 50 nt, corresponding to a throughput of ~32 Mb, ~157 Mb and ~862 Mb, respectively (Table 1). We did not consider the option of using paired reads, since this technique would not provide any justified benefits to the analyses made on our standard resequencing project, given the absence of major genomic rearrangement or the necessity of creating a *de novo* assembly.

All raw sequences underwent quality filtering procedures consisting in the trimming of low quality nucleotides from the reads. After this procedure, 27.2% of the bases from the original throughput were discarded from the Roche 454 dataset, 12.4% from the Illumina GA dataset, and 38.0% from the ABI SOLiD dataset. However, despite the variable number of nucleotides that were rejected, for all platforms the large majority of the reads (>99.8%) were not eliminated, but simply shortened (Table 1, Fig. S1). This outcome changed when, during the trimming procedure, not only the quality of the reads, but also its length was considered. By imposing a minimal size of 20 nt, following the rationale that high-score reads of a few nucleotides are useless for practical resequencing applications, Roche 454 was left with >99.6%, Illumina GA with the 92.1%, and ABI SOLiD with the 78.7% of the original number of reads, corresponding to a loss of 27.2%, 15.8%, and 43.5%, respectively, in terms of nucleotides.

Alignment to the targeted interval

Trimmed sequences, as well as un-trimmed ones, were mapped to the reference sequence (ref_seq). Since the number and the length of trimmed reads was lower with respect to raw reads, the total amount of bases from trimmed sequences mapping to the ref_seq was also lower. However, trimmed reads mapped to the ref_seq in higher percents, as a consequence of their increased content in high-quality bases, with the effect of producing in principle more accurate consensus sequences (Table 2). These

Table 1. Sequence throughput obtained with the three UHTS platforms analyzed.

Sequencing technology	Reads	Count	Discarded reads	Average length of a read (nt)	Bases	Trimmed bases
Roche 454 (1/8)	Total (raw)	99,317		318	31,615,489	
	After trimming	99,317	0.00%	232	23,010,105	27.2%
	After trimming (>20 nt)	98,975	0.34%	232	23,005,448	27.2%
Illumina GA (1 lane)	Total (raw)	4,611,113		34	156,777,842	
	After trimming	4,610,388	0.02%	30	137,405,021	12.4%
	After trimming (>20 nt)	4,245,639	7.9%	31	132,052,133	15.8%
ABI SOLiD (1 quad)	Total (raw)	17,287,756		50	862,377,074	
	After trimming	17,287,610	0.00%	31	534,615,489	38.0%
	After trimming (>201 nt)	13,597,456	21.3%	36	487,053,627	43.5%

doi:10.1371/journal.pone.0013071.t001

Table 2. Features of reads mapping to the 31-kb reference sequence.

Sequencing technology		Count of Mapped Reads	Mapped reads	Average length of a mapped read (nt)	Total mapped bases	Mapped bases
Roche 454	Full length reads (99,3 K)	62,830	63%	372.85	23,426,369	74%
	Trimmed reads (98,9 K)	78,766	80%	263.45	20,751,152	90%
Illumina GA	Full length reads (4,6 M)	437,1967	95%	34.00	148,646,878	95%
	Trimmed reads (4,2 M)	418,3505	99%	31.14	130,291,345	99%
ABI SOLiD	Full length reads (17,3 M)	14,842,743	86%	49.92	740,988,589	86%
	Trimmed reads (13,6 M)	12,790,106	94%	36.00	460,470,548	95%

doi:10.1371/journal.pone.0013071.t002

observations were particularly relevant for Roche 454 and ABI SOLiD alignments, rather than Illumina GA, since the latter was less affected by the trimming process.

The selected interval was entirely covered using the three datasets, with the exception of 2 very small gaps originating from non-overlapping PCRs (Figs. 1 and 2), a 8-nucleotide gap (position 18,837-44 of the ref_seq) present in the assembled sequence from Illumina GA reads, and 3 small gaps in long homopolymeric stretches in the assembly of Roche 454's trimmed reads (positions 9135-40, 9313-17, 10723-36). The VNTR present in intron 13 of the *PRPF31* gene also presented platform-specific gaps, as detailed below.

Coverage varied depending on the specific LR-PCR product analyzed, because of uneven loading of the individual PCR products (Fig. 2). Similar to the effect of naturally-occurring copy number variants (CNVs) or large-scale deletions, coverage across the analyzed region displayed sudden changes, highlighting at the same time the boundaries between different LR-PCR products. Coverage also varied widely across platforms (Table S1), as a direct effect of the different throughput of the 3 sequencers. High coverage variation also occurred within the same PCR (coefficient of variation for local alignments of untrimmed reads: 0.46 Roche 454, 0.41 Illumina GA, 0.56 ABI SOLiD, Table S1), with a strong bias for the amplicons ends (Fig. 2), a well-known artifact of UHTS [9]. As expected, the average coverage for each amplicon decreased when trimmed sets were used, although it was still much higher than the one required for confident ascertainment of heterozygous genetic variations, estimated by others to be approximately in the range of 10- to 40-fold [6,7,10,11]. In downstream analyses, we kept saturating coverage values to ensure a reliable comparison across platforms and to avoid differences due to stochastic variations of single base coverage.

Read Accuracy

To evaluate the accuracy of a base call in each platform after the alignment procedure, we used the "conservation" score, generally used in relationship to alignments of sequences originating from different species. In a resequencing context and as defined by the software package used, this value indicates the percent of the most represented base across the reads covering the same nucleotide in a sequence. An alignment at a given position would have a conservation score of 100% if all the reads carry the same base. For sake of simplicity, to compare the three alignments we selected only one PCR fragment (amplicon #3, ~8 kb), brought to a simulated average coverage of ~250x by using sampled trimmed reads. This procedure also allowed evaluating reads that were already filtered by quality scores. The average conservation values were similar across the three platforms (99.38% for Roche 454, 99.56% for Illumina GA, and 99.72% for ABI SOLiD,). However, important differences appeared when values at each position were individually ascertained. In short-read assemblies almost all nucleotides had perfect conservation, with some outliers corresponding to heterozygous SNPs (around 50%). In the long-read assembly the number of outliers was higher, especially within the 80–100% range (Fig. 3). In this latter case, the less conserved positions of the manually-inspected bases were associated with homopolymers stretches and corresponded to either an incorrectly called base or, more frequently, to a gap.

SNP detection

For comparative analysis of SNP detection performances we considered neither the intronic VNTR containing the *PRPF31* pathogenic mutation, nor another VNTR in the *TFTP* gene, also present in this region.

The number of SNPs identified by setting an allelic threshold of 20% was very similar across all platforms (Table 3). Decreasing the detection threshold to 10% allowed identifying a few more real variants (confirmed by Sanger sequencing), but also 11 more false positives in the Roche 454 and 1 in the Illumina GA datasets, all in correspondence of homopolymeric traits (Table S2). No false positives were detected in ABI SOLiD sequences, even when the threshold was lowered to 10%. Performance in SNP detection was not significantly affected by the use of trimmed vs. raw reads, except for Roche 454 alignments, where the trimming process decreased the number of false positives. This was probably due to the reduction of the coverage below the minimal threshold needed to call a SNP, operated by the trimming procedure itself.

For some heterozygous SNPs, mostly located within the 2nd long range PCR fragment, the number of reads relative to one allele was substantially higher with respect to reads belonging to the other one. This effect was particularly visible for the short-read platforms, to a point that the experimental results did not allow a clear detection of the variant, or a clear ascertainment between homozygous and heterozygous SNPs (Table S3). Electropherograms from Sanger sequencing of the same PCR products used as sequencing template for UHTS revealed the same allelic imbalance for some of these SNPs (at positions 7661, 8337, 8564, 9081 of the ref_seq, Table S3). However, when using PCR products obtained by short-range PCR amplification as Sanger sequencing template, electropherograms showed clearly heterozygous peaks for these same SNPs. Taken together, these results may represent the effects of imbalanced amplification of the two alleles prior to sequencing [12], rather then a UHTS-specific or mapping effect.

In all three platforms, the algorithm interpreted the duplication of a CAAG next to an A stretch (dbSNP:5828571) as 2 SNP.

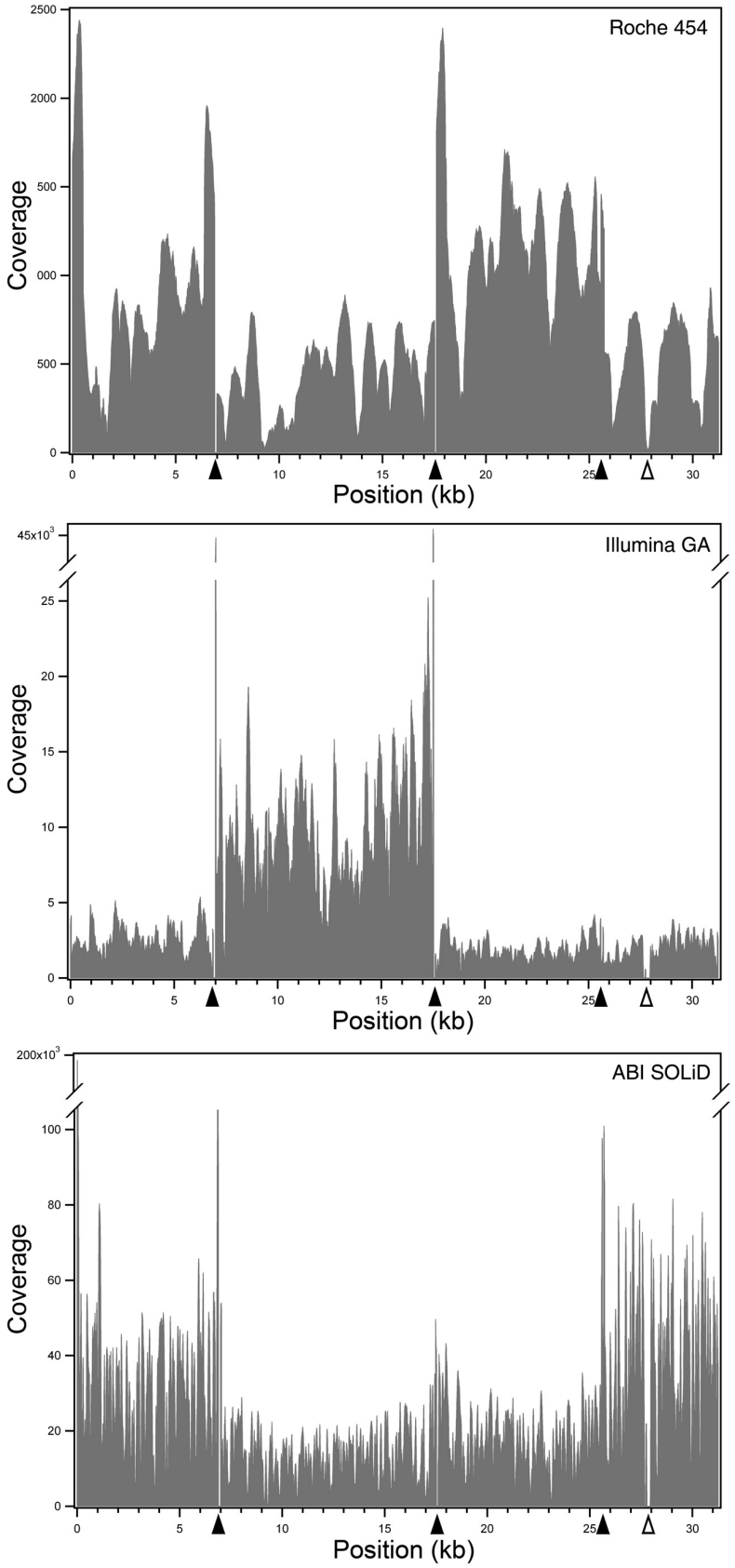

Roche 454
Coverage
Position (kb)
Illumina GA
45x10^3
Coverage
Position (kb)
ABI SOLiD
200x10^3
Coverage
Position (kb)

Figure 2. Coverage per bp of the analyzed region, by the assembly of untrimmed reads. Solid arrowheads indicate the boundaries of each long-range PCR amplicon, while open arrowheads show the position of the VNTR in intron 13 of *PRPF31*.
doi:10.1371/journal.pone.0013071.g002

SNP detection at simulated coverage depths

Coverage simulations were performed to ascertain the presence of features emerging from non-saturating conditions and to determine the minimum coverage required by each platform to detect the correct number of SNPs. We randomly sampled reads after the filtering and trimming procedure to obtain seven average depths (350, 250, 100, 50, 20, 15, and 10x). The average coverage of each fragment was proportional to the number of reads of a given length used in the assembly (data not shown), so that it was possible to calculate the number of reads to sample from each dataset in order to obtain the desired coverage depth.

For each simulated sequencing experiment, we counted the number of SNPs identified. We eliminated all variants detected having less than 5x coverage, allowing at least 2 high-quality reads carrying the variant, since these parameters were already ascertained to produce reliable calls [6]. We chose as "reference set" of detectable SNPs the list of variants reported in Table S3, with the exception of the two entries corresponding to the CAAG duplication (52 SNPs in total). SNP detection following mapping at simulated coverage depths showed some platform-specific differences in the number of variants detected as function of the average number of reads per base (Fig. 4). However, these differences quickly disappeared as soon as the threshold coverage value corresponding to ~50x was reached. After this limit there was little or no increase in the SNP discovery rate and the different samplings show nearly-similar results. Specifically, at 50x we detected 88% SNPs with Roche 454, and 95% SNPs with Illumina GA and ABI SOLiD, but at higher coverage all platforms reached a plateau score of ~95%.

False positive appeared in all three platforms. Regardless of the simulated average coverage, they were the outcome of random errors in sequencing that could not be corrected by additional reads covering the same position. At lower depths, this limitation was the obvious effect a reduced number of available reads. At higher depths, false positives invariantly showed to have local coverage that was at least 10 times lower than the average (simulated) one, likely because of mapping difficulties, and thus easily recognizable as false calls.

Insertions and deletions detection

The automated identification of small insertions and deletions (indels) is a difficult issue both for Sanger sequencing and UHTS technologies. One heterozygous cytosine deletion (dbSNP: 34064860) downstream of the *PRPF31* gene was found in alignments for the three platforms. For Illumina GA and ABI SOLiD this was the only indel detected, while for the Roche 454 we could identify 88 (Table S4) and 124 (not shown) additional deletions spanning one to four bp when trimmed and untrimmed

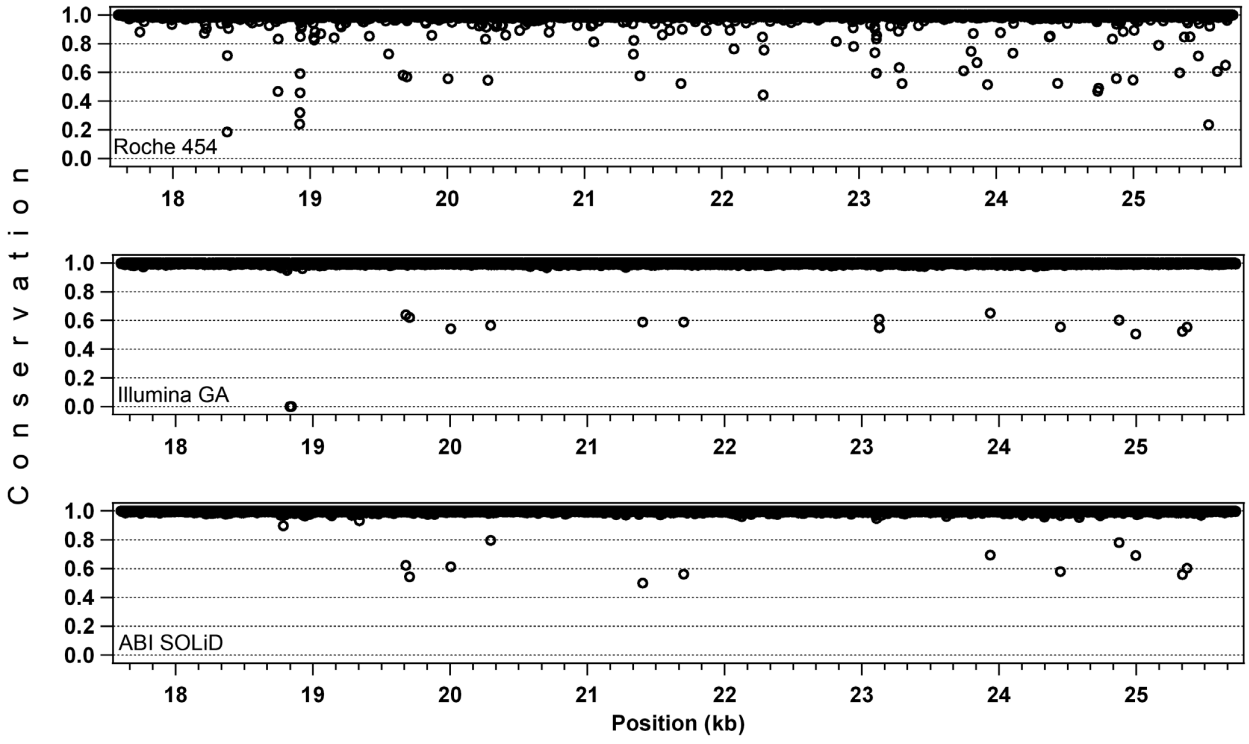


Figure 3. Read accuracy evaluation, on the 3rd long-range PCR fragment (~8 kb). Values on the Y axis (conservation) indicate the fraction of the most prevalent base at a given position, as detected from the reads covering that position. Conservation scores below 0.5 represent gaps of the assembly with respect to the reference sequence. In the Roche 454 assembly this occurs when the majority of the reads have an indel with respect to the ref_seq, while in the Illumina GA alignment the point at 0% conservation corresponds to a region with no coverage (position: 18,837-44).
doi:10.1371/journal.pone.0013071.g003

Table 3. Number of SNPs and false positive variants detected, after alignments of untrimmed reads.

Sequencing technology	Alignments	True variants >20%	False positives>20%	True variants >10%*	False positives>10%*
Roche 454	Full length reads	48	4	49	15
	Trimmed reads	46	1	47	6
Illumina GA	Full length reads	49	0	52	1
	Trimmed reads	49	0	52	1
ABI SOLiD	Full length reads	48	0	51	0
	Trimmed reads	50	0	51	0

*Values inclusive of the elements detected with a >20% threshold.
doi:10.1371/journal.pone.0013071.t003

reads, respectively, were used. All of them were found in correspondence of homopolymers stretches and were considered as false positives. Moreover, Sanger sequencing of the *PRPF1* gene did not reveal any of the deletions detected by Roche 454 in that interval.

No insertion was automatically found in any of the sequences generated by the three platforms, including the CAAG duplication, ascertained with Sanger sequencing and by manually checking the UHTS alignments (dbSNP:5828571).

Alignment to the repetitive region of *PRPF31* containing the c.1347+654C>G mutation

Repetitive regions represent more than 50% of the human genome [13]. These elements are generally masked in large-scale assembly processes to avoid non-specific alignment of the reads. To overcome this problem, which could have influenced the assessment of variant detection, reads that had multiple matches on the ref_seq were discarded from the analyses. This resulted in lowering the local coverage of low-complexity regions but did not create noise in variant detection. With respect to the VNTR, coverage patterns were not uniform and were platform-specific (Fig. 5A). Unlike reads from Illumina GA and ABI SOLiD, sequences generated by Roche 454 could cover the whole VNTR. Thanks to their longer range, they aligned to the non-repetitive (anchoring) flanking sequences and therefore represented the best option for sequencing this repetitive element. However, the reads deriving from the core repeats had multiple matches and were eliminated, thus resulting in the coverage dip in the corresponding region.

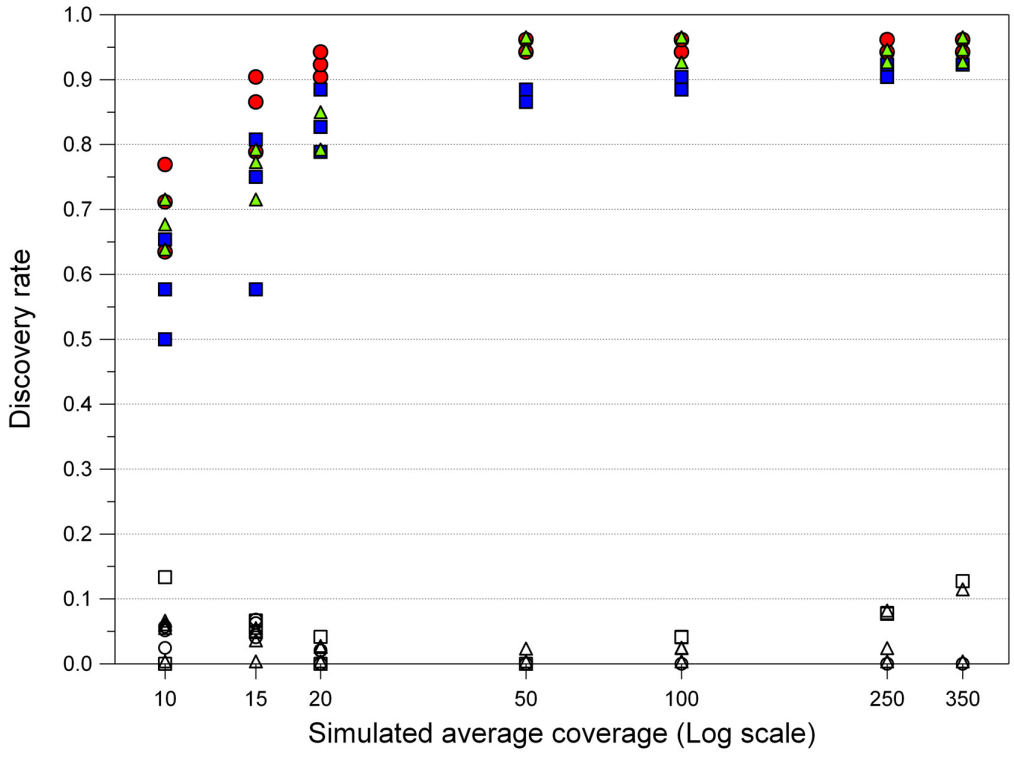


Figure 4. SNP discovery rate at simulated coverage depths. We tested seven average coverage depths, with three random samples for each point. SNPs and false positive hits are indicated by filled and open symbols, respectively. Squares, Roche 454; circles, Illumina GA; triangles, ABI SOLiD.
doi:10.1371/journal.pone.0013071.g004

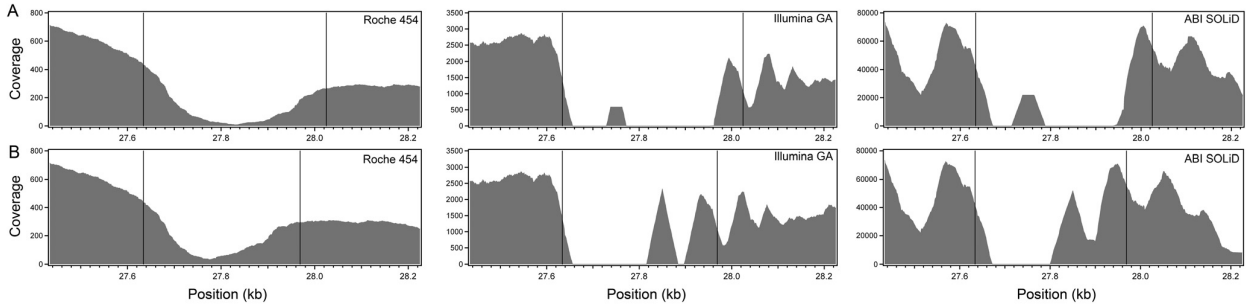


Figure 5. Coverage of the *PRPF31* intronic VNTR. Values shown are relative to the mapping of the original ref_seq entry containing 7 repeats (A) or of the experimentally-determined sequence containing 6 repeats (B). Untrimmed reads are represented here because they produced the best alignment to this repetitive element. Vertical lines show the boundaries of the VNTR.
doi:10.1371/journal.pone.0013071.g005

An additional element of complexity typical of repetitive elements such as VNTRs is that they are polymorphic. The individual analyzed here was homozygous for 6 VNTR repeats, while the ref_seq reported a VNTR carrying 7 elements. None of the three platforms analyzed could resolve the correct structure, which was disclosed only by Sanger sequencing. When UHTS reads were aligned onto the correct sequence, short reads assemblies still could not match to the central portion, although the sizes of the gaps were reduced and more sequence could be covered, as consequence of the increased number of uniquely-placed reads. On the contrary, long reads could precisely map the entire region (Fig. 5B). The same occurred also for another VNTR located in the *TFTP* gene: none of the three alignments could clearly detect two repeats present in the patient with respect to the four repeats reported in the ref_seq.

To bypass the limitation arising from forcing an alignment to a reference sequence, we tried also *de novo* assembly of the subset of reads matching the *PRPF31* VNTR (544 Roche 454's reads, 173,824 Illumina GA's reads, 773,064 ABI SOLiD's reads). A *contig* could be obtained only with Roche 454 reads but, as before, the number of the repeats did not correspond to the ones of the patient (one of them was missing).

The mutation associated with adRP in the patient was clearly detected by all three techniques with a frequency very close to 50% and a coverage similar to the rest of the fragment and regardless of the ref_seq used, likely thanks to its proximity with the 5′ anchoring non-repetitive region.

Discussion

To provide a proof of concept for routine genomic DNA resequencing by UHTS, specifically focused on the detection of disease-causing variants, we processed a 31-kb human genomic region with three next-generation sequencing platforms and analyzed the results with a commercial, user friendly software. In addition to several common SNPs and other typical variants of the human genome, this interval contained a rare mutation located in a particularly challenging region, thus representing an interesting benchmark for a comparative analysis.

The raw sequence throughputs obtained were consistent with the ones expected for the portion of the sequencing area used for each instrument, as specified by each manufacturer. For all platforms, the reads were minimally affected by the filtering (trimming) procedure, as only 0.2% or less of them were discarded. However, this result cannot be taken as a practical qualitative parameter, since reads of excellent quality but of very short length are basically useless in resequencing procedures. When a minimal length of 20 nt was included as a parameter in the filtering process, the three platforms began to reveal some differences. Roche 454 conserved basically all of the original reads, in virtue of its chemistry producing sequences much longer than 20 nt, while the other sequencers retained only 80–90% of them. It has to be noted, however, that this trimming procedure was heavily dependent on the strategies used by the single platforms to eliminate low quality and polyclonal sequences from the raw output and has only a relative value in terms of comparison across the different UHTS systems. For example, ABI SOLiD's low quality reads were not discarded *a priori* by the machine since this platform relies more on quality control steps (color space) during the mapping procedure than during the pre-filtering process.

Following mapping procedures, different platforms produced different coverage depths per base. This was simply the result of the initial different sequencing throughput typical of each platform, and not an issue related to the quality of the sequences or to the mapping procedure. Considering, however, that the same relative sequencing surface was used for all the machines (1/8 of the total sequencing area), mapping of Roche 454 raw reads produced an average coverage of ~770x/base, of Illumina GA reads ~4,000x/base, and of ABI SOLiD reads ~26,000x/base. The throughput of each machine is constantly increasing, following the technical development of the respective chemistries, making it difficult to provide updated comparisons relying on real data analyses. For example, the new released models from Illumina (HiSeq 2000) and ABI SOLiD (version 4) can reach a throughput of 100 Gb per run or more.

Mapping accuracy appeared to correlate with the quality of the individual reads, rather than with parameters related to the mapping procedure itself. Specifically, short-read platforms produced assemblies having higher accuracy than Roche 454, simply because this latter platform is prone to introduce errors (especially indels) when stretches of homopolymeric bases are present [10,14].

Once the contigs were obtained, we focused on the detection of the human variants contained in the targeted region (SNPs, small insertions and deletions, other polymorphisms), the principal aim being the simulated discovery of pathogenic mutations. SNP detection was overall comparable across the three platforms; however, some differences could be detected. In Roche 454's long-read alignments, false positives and negatives (undetected SNPs) could be again connected to the typical errors of the 454 technology, related to homopolymer effects. We observed that the use of quality-trimmed reads could reduce these false positive calls, but it also reduced the number of true variants automatically

detected. Nevertheless, when manually inspected, these variants could safely be identified. Similarly, in alignments from short-read platforms false negatives (one of which was in common between Illumina GA and ABI SOLiD) were due to the low frequency displayed by the "non ref_seq" allele, and they become detectable when the discovery threshold was lowered. In some particular instances, especially for Illumina GA data, the under-detection or the incorrect calling of SNPs as homozygous or heterozygous variants were not a consequence of UHTS errors, but could be explained by allelic unbalanced amplification. This phenomenon occurs when one of the two alleles is enriched during the PCR amplification of the template DNA, or perhaps during the amplification of the libraries, and results in a problem that is relevant also when high coverage depths are used [8]. Notably, in Roche 454 this unbalance was present but less pronounced, indicating probably an inferior sensitivity to this phenomenon. Another interesting observation is that some of the SNPs with low-limit frequency in short-read alignments were located in regions that presented similarities with other segments of the analyzed interval (Fig. S2). One hypothesis could be that some of the reads sequenced from a particular SNP were mistakenly aligned to other similar sequences and vice-versa, lowering the frequency of detection at the real position. Yet, in other regions of similarities SNPs were correctly identified, leading to the notion that errors in allelic calling due to sequence repeats may not represent an absolute rule, especially if the noise is reduced by eliminating reads displaying multiple matches. Taken together, these results indicate that, despite the fact that UHTS machines produce quantitative results, other causes may influence the detection of heterozygous variants when standard parameters are chosen in automated detection. However, in practical terms this issue should not represent a major concern, as the number of SNPs that were prone to this miscalling represented in our test only a small fraction of the total number of heterozygous SNPs.

Sensitivity in SNP detection with respect to the coverage increased from Roche 454 to ABI SOLiD and finally to Illumina GA. Since the differences detected were not too pronounced and SNP detection was heavily dependent on the regional sequence context, we can safely conclude that all platforms analyzed can be considered as having similar performances with respect to sensitivity at the same average coverage. Indeed, it is very hard to extrapolate the results from their specific sequence or random coverage contexts, as the mapping procedure (and the corresponding local coverage of a given SNP) was influenced by the complexity of the DNA to be sequenced and the number of reads available. At lower average coverage depths, the rate of discovery decreased sensitively and different random samplings gave different results because the number of poorly-covered regions was higher. As mentioned, in correspondence of false positive calls local coverage was low even when the average coverage depth was high, indicating a direct influence of the mapping procedure on automated identification of variants.

With respect to detection of small insertions and deletions, the most relevant observation relates to the identification of a large number of false positive deletions in homopolymers stretches obtained with Roche 454 alignments, as also noted by others in analyses of longer genomic intervals [8]. Considering the importance of indels in human hereditary diseases, our experiments indicate that Roche 454 sequences would require the use of specific downstream algorithms, able to systematically detect the presence of sequence-dependent false positives.

The c.1347+654C>G mutation in the 56-bp intronic VNTR of *PRPF31* was taken as a benchmark to assess whether "difficult" DNA variants could be detected by UHTS. Large-scale sequencing projects almost invariantly clash with the problem of mapping and carefully analyzing repetitive DNA elements [2,3]. Roche 454's long sequences (and presumably any newer UHTS chemistry or technology producing extended reads) represent without doubts the best tool for covering repetitive regions, at least for elements that do not exceed in size the average length of ~1.5 to 2 reads. Our results support this assumption, since the Roche 454 reads provided the most complete coverage of both the *PRPF31* and *TFPT* VNTRs analyzed. Nevertheless, it was not possible to precisely resolve the number of repeats composing these elements, neither by aligning them to a reference sequence, nor by *de novo* assembly.

Conversely, despite the presence of repeats, all three platforms tested could successfully detect the mutation associated with the disease in the patient's genome. This favorable outcome is probably due to the presence of the pathogenic variant within the first of the 6 elements composing the VNTR, thus allowing the "anchoring" of some reads to the non-repetitive DNA region in 5′ of this repeat. Although previous attempts to identify this mutation with an earlier version of the Illumina GA (the "GA I" platform) failed in such a task [4], this can be explained by the algorithms used for aligning Illumina reads, rather than by the improvements made by the Solexa technology. Specifically, all software used previously allowed random alignment of reads having multiple matches, thus creating noise in the detection of the variant in nearly-identical repeats.

Other rearrangements of the human genome characterized by a variable number of large and unique DNA copies (CNVs, large duplication and deletions, genetic amplification in cancer, etc.) are in general easily detected by UHTS. Because of the quantitative nature of the sequencing results, such rearrangements produce very noticeable variations of coverage when aligned to a ref_seq. For example, CNVs, sparse and non-repetitive elements spanning kilobases to megabases of DNA [15], are simply detected as sudden variations of the coverage by all UHTS platforms analyzed here [7,11,16,17].

The increase of read length in UHTS platforms, an issue on which manufacturers are putting constant efforts, will probably help reducing some of the current weaknesses of this technology and accelerate the transition from Sanger sequencing to UHTS. Illumina, for example, has increased the length of the reads from 35 nt to 100 nt in less than a year; ABI SOLiD, from 50 to 75 nt. However, if we exclude repeats-related concerns, our data seem to indicate that this ever changing dimension in UHTS systems should not have a major impact on DNA variants detection in resequencing efforts (since the reference sequence is known already), whereas the quality of the reads produced should. Hence, the data produced here can very likely be extrapolated to future longer reads from the same platforms, provided that the sequencing chemistry and procedures remain the same.

In our analysis we did not consider the costs of sequencing as a comparative parameter, although it obviously represents an important factor to be taken into account while designing a sequencing project. From our results, no striking qualitative difference appeared across the three platforms, when appropriate conditions in terms of reads and coverage depths were fulfilled. As a general rule then, the less expensive platform producing the needed amount of sequences for a given project would probably also be the most suitable one, unless platform-specific characteristics (e.g. long reads, usable throughput, etc.) are critical for the tests to be carried out or other endeavors with respect to genomic DNA resequencing (e.g. transcriptome sequencing) are performed.

In conclusion, in our work we show that identification of DNA variants in complex DNA sequences such as the human genome can be achieved by highly-parallel techniques, with investments in terms

of cost and time that represent a fraction of what is usually spent for conventional sequencing. Furthermore, our successful adoption of a user-friendly software and a straightforward analytical pipeline demonstrates that a strong bioinformatic background is not a compulsory requirement for investigators dealing with UHTS technology. In our example, we performed the analysis of a large genomic region from a single individual amplified by LR-PCR. However, the power of UHTS can be applied to sequence shorter DNA regions obtained by sequence capture or conventional PCR in multiple patients, i.e. to a procedure that is more similar to current routine setups in medical genetic laboratories. Although some limitations to this latter UHTS application still exist, the use of sample pooling [18,19] and individual DNA barcoding [20,21,22,23] is now facilitating the adoption of highly-parallel sequencers by conventional genetic labs. Taken together, our data indicate that the so-called "next-generation" sequencing, regardless of the platform used, can be efficiently and safely used by the current generation of human geneticists as well.

Supporting Information

Figure S1 Distributions of read lengths from the three platforms tested. The output generated from short-read platforms consists in reads having the same length: Illumina GA generated only reads of 34 nt and ABI SOLiD generated mostly reads of 50 nt, with only a small fraction of them (0.4%) having shorter lengths.
Found at: doi:10.1371/journal.pone.0013071.s001 (0.61 MB TIF)

Figure S2 Similarity plot of the region analyzed. The VNTRs within the *TPFT* and *PRPF31* sequences are indicated by arrows. Vertical lines (corresponding horizontal lines are omitted) indicate the position of SNPs rs35705606 and rs2668836 at coordinates 13,761 and 14,098, respectively, that were under-detected by short read platforms.
Found at: doi:10.1371/journal.pone.0013071.s002 (7.67 MB TIF)

Table S1 Coverage of individual amplicons.
Found at: doi:10.1371/journal.pone.0013071.s003 (0.05 MB DOC)

Table S2 Details on false positive results detected, after assembly of untrimmed reads.
Found at: doi:10.1371/journal.pone.0013071.s004 (0.06 MB DOC)

Table S3 SNPs detected after mapping of UHTS untrimmed reads. Black: SNPs detected by using the default threshold for heterozygozity (20%). Red: SNPs detected with a threshold between 10% and 20%. Blue: SNPs with a borderline limit definition of homo-heterozygosity. Green: variants corresponding to the CAAG insertion. Grey shadow: SNPs located in the 2nd long range PCR fragment, showing allelic imbalance. SNPs identified in the tandem repeats are not reported.
Found at: doi:10.1371/journal.pone.0013071.s005 (0.22 MB DOC)

Table S4 Deletions detected with Roche 454 (false positives), using trimmed reads. The deletion at position 30,672 was also found using the other 2 platforms, likely being the only real small deletion.
Found at: doi:10.1371/journal.pone.0013071.s006 (0.16 MB DOC)

Acknowledgments

We would like to acknowledge Drs. E. Brini, A. Felsani, and A. Guffanti, Genomnia srl, Milan, Italy, Dr. L. Farinelli, Fasteris SA, Plan-les-Ouates, Switzerland and Dr. M. Künzli, FGCZ, Zurich, Switzerland for help with UHTS experiments. We also thank Dr. G. Csárdi for Python programming.

Author Contributions

Conceived and designed the experiments: PB CR. Performed the experiments: PB. Analyzed the data: PB CR. Wrote the paper: PB CR.

References

1. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74: 5463–5467.
2. Pop M, Salzberg SL (2008) Bioinformatics challenges of new sequencing technology. Trends Genet 24: 142–149.
3. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11: 31–46.
4. Rio Frio T, McGee TL, Wade NM, Iseli C, Beckmann JS, et al. (2009) A single-base substitution within an intronic repetitive element causes dominant retinitis pigmentosa with reduced penetrance. Hum Mutat 30: 1340–1347.
5. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.
6. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53–59.
7. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, et al. (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res 19: 1527–1541.
8. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, et al. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10: R32.
9. Harismendy O, Frazer K (2009) Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology. Biotechniques 46: 229–231.
10. Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, et al. (2008) Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res 18: 1638–1642.
11. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452: 872–876.
12. Ikegawa S, Mabuchi A, Ogawa M, Ikeda T (2002) Allele-specific PCR amplification due to sequence identity between a PCR primer and an amplicon: is direct sequencing so reliable? Hum Genet 110: 606–608.
13. Richard GF, Kerrest A, Dujon B (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev 72: 686–727.
14. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8: R143.
15. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. Nature 444: 444–454.
16. Park H, Kim JI, Ju YS, Gokcumen O, Mills RE, et al. (2010) Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet 42: 400–405.
17. Massouras A, Hens K, Gubelmann C, Uplekar S, Decouttere F, et al. (2010) Primer-initiated sequence synthesis to detect and assemble structural variants. Nat Methods 7: 485–486.
18. Ingman M, Gyllensten U (2009) SNP frequency estimation using massively parallel sequencing of pooled DNA. European Journal of Human Genetics 17: 383–386.
19. Out AA, van Minderhout IJ, Goeman JJ, Ariyurek Y, Ossowski S, et al. (2009) Deep sequencing to reveal new variants in pooled DNA samples. Hum Mutat 30: 1703–1712.
20. Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, et al. (2008) Identification of genetic variants using bar-coded multiplexed sequencing. Nat Methods 5: 887–893.
21. Meyer M, Stenzel U, Myles S, Prufer K, Hofreiter M (2007) Targeted high-throughput sequencing of tagged nucleic acid samples. Nucleic Acids Res 35: e97.
22. Parameswaran P, Jalili R, Tao L, Shokralla S, Gharizadeh B, et al. (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. Nucleic Acids Res 35: e130.
23. Lennon NJ, Lintner RE, Anderson S, Alvarez P, Barry A, et al. (2010) A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. Genome Biol 11: R15.

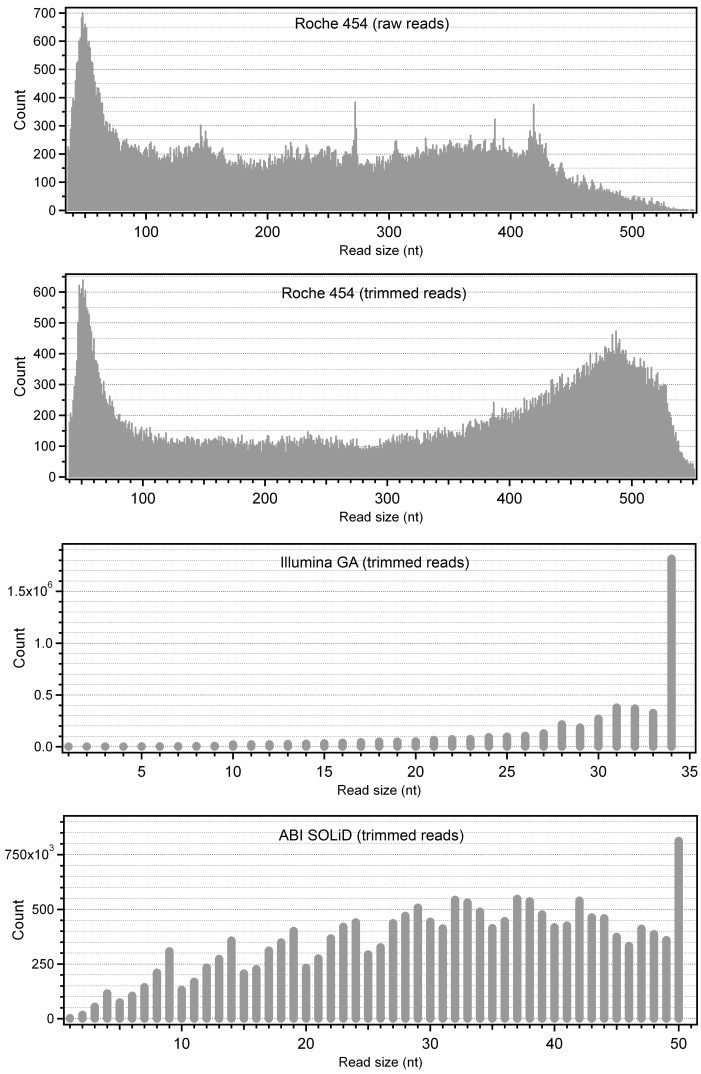


Figure S1.

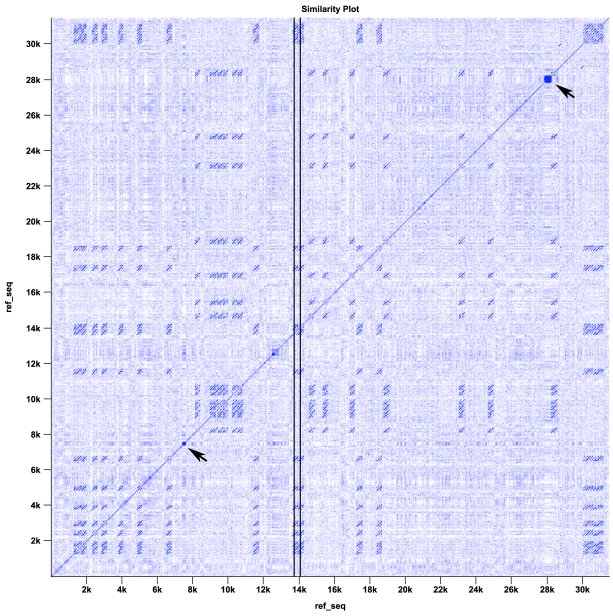


Figure S2.

Table S1.

		Assembly of raw reads				Assembly of trimmed reads			
Sequencing technology	**Amplicon**	**Median coverage**	**Mean coverage**	**Coefficient of variation***		**Median coverage**	**Mean coverage**	**Coefficient of variation***	
					Mean				Mean
Roche 454 (1/8)	1	821	916	0.56	0.46	727	786	0.56	0.48
	2	464	438	0.48		420	408	0.51	
	3	1172	1188	0.31		1039	1050	0.34	
	4	583	541	0.49		497	479	0.50	
Illumina GA (1 lane)	1	2391	2560	0.36	0.41	2082	2250	0.37	0.40
	2	9174	9823	0.46		7888	8635	0.48	
	3	1777	1894	0.38		1539	1663	0.40	
	4	2043	1955	0.44		1794	1776	0.37	
ABI SOLiD (1 quad)	1	32332	33266	0.61	0.56	19672	21074	0.69	0.64
	2	13569	13846	0.52		8202	8861	0.65	
	3	17368	18998	0.58		10614	11912	0.62	
	4	37036	37580	0.53		21700	22971	0.59	

*The coefficient of variation for each amplicon is calculated as the ratio of the standard deviation to the mean.

Table S2.

	Ref_seq pos.	**Alleles**	**Frequencies**	**Coverage**	**Count of 2nd allele**	**Gene**
Roche 454 at 20% threshold	17133	C/T	61.5/38.5	78	30	*PRPF31*
	23116	A/C	67.9/28.6	28	8	*PRPF31*
	29924	T/C	66.7/33.3	15	5	
	29925	T/C	78.3/21.7	23	5	
Roche 454 at 10% threshold	1008	A/G	82.1/17.9	28	5	*NDUFA3*
	2307	T/G	87.7/10.8	65	7	*NDUFA3*
	3758	T/G	81.2/15.6	32	5	*NDUFA3*
	3764	T/G	89.5/10.5	38	4	*NDUFA3*
	6610	T/A	80.6/16.1	31	5	*TFPT*
	7260	C/A	88.5/11.5	139	16	*TFPT*
	18939	A/G	80.0/17.8	45	8	*PRPF31*
	23117	A/C	82.4/17.6	17	3	*PRPF31*
	23126	A/C	84.2/15.8	19	3	*PRPF31*
	23320	T/C	87.0/13.0	23	3	*PRPF31*
	23321	T/G	83.3/16.7	18	3	*PRPF31*
Illumina GA at 10% threshold	9141	G/A	85.5/14.5	4558	661	*TFPT*

Table S3.

					ROCHE 454				ILLUMINA GA				ABI SOLiD			
GENE	Annotated SNPs	Validated by Sanger	Ref_seq position	Ref. allele	Alleles	Frequencies	Coverage	Count of 2nd allele	Alleles	Frequencies	Coverage	Count of 2nd allele	Alleles	Frequencies	Coverage	Count of 2nd allele
NDUFA3	rs254260		450	T	G	99.9	1897		G	100.0	1632		G	99.7	12814	
NDUFA3	rs254259		646	T	T/C	61.7/38.1	575	219	T/C	62.6/37.4	1919	718	T/C	50.8/49.1	8365	4110
NDUFA4	rs2118214*	yes	2286	G					G/T	71.4/28.5	2977	847	G/T	89.2/10.6	5450	575
NDUFA3			2476	G	G/A	59.4/40.4	636	257	G/A	56.4/43.4	3415	1483	G/A	56.9/43.0	21062	9066
NDUFA3	rs254257		3268	A	A/G	58.0/41.6	562	234	A/G	65.7/34.3	2643	906	A/G	55.1/44.8	22847	10240
NDUFA3			3389	C	C/T	61.2/38.8	619	240	C/T	61.4/38.5	1494	575	T/C	52.6/47.3	20754	9813
NDUFA3	rs45595133		3754	T	G/T	61.1/38.9	18	7	T/G	52.6/47.2	2243	1059	T/G	70.7/29.0	3615	1050
NDUFA3	rs11878334		3853	T	T/A	63.0/36.2	552	200	T/A	60.5/39.0	1650	643	T/A	60.8/38.9	16032	6235
NDUFA3. (CDS)	rs1061333		4411	C	C/T	52.2/47.8	813	389	C/T	58.9/41.1	1207	496	C/T	81.6/18.3	1706	312
TFPT			5224	G	G/T	62.5/37.1	774	287	G/T	63.5/36.0	1991	717	G/T	66.5/33.2	20108	6667
TFPT	rs56061812		5372	A	A/G	68.2/31.8	676	215	A/G	64.7/35.3	3208	1131	A/G	52.1/47.9	34549	16539
TFPT	rs2118213		7341	G	T	100.0	30		T	99.6	5319		T	99.3	6294	
TFPT	rs12609379	yes	7661	G	*A/G*	*80.8/18.8*	*240*	*45*	*A/G*	*81.4/18.4*	*7484*	*1378*	*A/G*	*86.9/13.0*	*9359*	*1221*
TFPT	rs254269		8128	G	A	99.1	323		A	99.3	6154		A	99.1	10488	
TFPT	rs60371156*	yes	8337	T					T/A	87.2/12.5	4049	505				
TFPT		yes	8564	C	G/C	70.1/29.9	616	184	*G/C*	*81.2/18.6*	*14914*	*2778*	G/C	76.7/23.2	9674	2245
TFPT	rs57911619	yes	9081	C	G/C	60.9/39.1	192	75	G/C	77.7/22.1	5641	1248	*G/C*	*82.8/16.9*	*3373*	*571*
PRPF31	rs4806711	yes	13432	A	G/A	57.3/42.7	503	215	G/A	70.8/29.1	4507	1313	G/A	70.0/30.0	12653	3790
PRPF31	rs35705606	yes	13761	A	A/G	74.3/25.7	113	29	A/G	84.4/15.5	4462	691	A/G	85.1/14.9	11367	1691
PRPF31	rs2668836	yes	14098	A	A/C	69.4/30.6	468	143	A/C	80.8/18.9	5588	1056	A/C	75.7/24.1	9431	2273
PRPF31	rs254277	yes	18785	G	A	99.6	231		A	98.4	1480		A	96.3	3577	
PRPF31		yes	19678	T	T/G	59.0/40.9	1116	456	T/G	60.1/39.5	2119	837	T/G	61.8/38.0	12695	4820
PRPF31	rs59977379	yes	19705	A	A/G	50.0/50.0	1051	525	A/G	58.9/41.1	2245	922	G/A	50.4/49.6	14174	7025
PRPF31		yes	20006	C	C/T	54.3/45.7	894	409	C/T	50.5/49.5	1846	913	C/T	50.3/49.7	9943	4938
PRPF31	rs56220912	yes	20296	C	C/G	57.0/42.9	994	426	G/C	52.4/47.3	1066	504	G/C	69.6/30.3	4504	1363
PRPF31	rs254275	yes	20527	C	G	99.5	409		G	100.0	1510		G	99.9	16010	
PRPF31	rs2303557	yes	21338	T	C	100.0	879		C	100.0	665		C	99.9	3080	
PRPF31. (CDS)	rs1058572	yes	21405	G	G/A	58.6/41.3	1139	470	G/A	58.7/41.3	993	410	A/G	55.5/44.4	8681	3858
PRPF31	rs56234781	yes	21703	G	A/G	53.1/46.8	1186	555	G/A	51.0/48.8	1435	700	G/A	63.6/36.4	5138	1870
PRPF31	rs33976447	yes	23127		C/A	86.2/13.8	65	9	C/A	52.7/47.1	831	391	C	99.3	539	
PRPF31		yes	23130		G/A	87.5/12.5	295	37	G/A	69.0/30.9	758	234	G	99.8	578	
PRPF31	rs254274	yes	23181	G	A	99.3	584		A	99.5	1315		A	99.0	6044	
PRPF31	rs254273	yes	23788	A	G	99.8	1235		G	99.6	1499		G	99.9	7961	
PRPF31	rs254272	yes	23938	T	T/C	53.3/46.7	1399	653	T/C	60.8/39.1	2092	819	T/C	64.2/34.2	16475	5631
PRPF31	rs10424816	yes	24449	C	C/A	58.0/41.9	776	325	C/A	50.4/49.5	1847	914	A/C	59.7/40.1	9779	3925
PRPF31		yes	24876	T	T/G	60.1/39.6	675	267	T/G	57.5/42.3	2132	901	G/T	76.0/23.4	9364	2187
PRPF31	rs254271	yes	24998	G	G/C	57.7/42.3	943	399	G/C	53.2/46.7	1582	739	G/C	65.5/34.2	12638	4324
PRPF31	rs10853869	yes	25338	G	G/A	56.1/43.8	1229	538	A/G	50.2/49.6	3089	1533	G/A	55.4/44.4	20382	9048
PRPF31		yes	25372	T	T/C	83.2/16.5	636	105	T/C	54.7/45.2	2239	1013	T/C	60.2/39.7	15970	6348
PRPF31	rs171703	yes	25619	T	C	100.0	542		C	100.0	604		C	98.4	6125	
PRPF31	rs34990810	yes	25871	C	T/C	58.5/41.5	388	161	T/C	54.4/45.5	794	361	T/C	70.2/29.7	6605	1964
PRPF31	rs10417221*	yes	26121	T					C	98.7	226		C	97.8	1963	
PRPF31	rs2668840	yes	26152	A	G	100.0	115		G	99.6	810		G	99.9	12033	
PRPF31	rs667324	yes	26177	G	A	98.6	143		A	99.6	764		A	99.7	10816	
PRPF31	rs2556367	yes	26332	G	G/A	52.7/47.3	245	116	A/G	51.7/48.3	1711	826	G/A	52.0/48.0	10107	4847
PRPF31	rs2576453	yes	26494	G	A	100.0	295		A	99.6	726		A	99.5	19730	
PRPF31	rs608608	yes	26593	C	G	99.7	298		G	99.6	687		G	99.6	14407	
PRPF31	rs655240	yes	26664	T	*C/T*	*79.7/20.3*	*138*	*28*	C	99.6	978		C	99.8	11411	
	rs12150988		29568	C	G/C	53.3/46.7	666	311	C/G	50.6/49.3	1901	938	C/G	54.4/45.4	15665	7107
	rs2668838		30444	C	T/C	71.6/28.4	95	27	T/C	51.7/48.1	1438	692	C/T	52.7/47.3	38101	18015
			30583	A	A/G	65.1/34.9	350	122	A/G	61.8/38.0	2567	976	A/G	61.8/38.1	32605	12422
	rs254248		30795	A	A/G	70.4/29.6	500	148	A/G	62.6/37.4	1769	661	A/G	54.5/45.4	20267	9200
	rs4806715		30871	A	C/A	58.0/41.3	460	190	A/C	62.2/37.7	2552	961	A/C	59.7/40.1	34714	13926
	rs2668837		30965	C	T/C	64.1/35.8	632	226	T/C	54.8/45.2	2192	990	T/C	52.0/48.0	22426	10757

*, within homopolymeric sequence.

Table S4.

Reference position	Reference allele	Allele variations	Frequencies	Coverage	Count of 2nd allele	Overlapping gene
728	G	G/-	64.4/35.6	407	145	*NDUFA3, CDS*
775	C	C/-	73.5/26.5	211	56	*NDUFA3*
782	C	C/-	65.1/34.9	195	68	*NDUFA3*
1001	A	A/-	56.1/37.8	180	68	*NDUFA3*
1211	G	G/-	51.4/47.5	183	87	*NDUFA3*
1217	G	G/-	76.4/23.6	165	39	*NDUFA3*
1238	T	-	42.0	143		*NDUFA3*
1238	TT	--	39.2	143		*NDUFA3*
1383	T	T/-	71.4/28.6	84	24	*NDUFA3*
1390	T	T/-	78.8/21.2	52	11	*NDUFA3*
1550	T	T/-	74.3/24.9	237	59	*NDUFA3*
2278	T	-/T	51.6/43.2	190	82	*NDUFA3*
2295	T	T/-	76.8/20.6	155	32	*NDUFA3*
3679	C	C/-	63.6/35.6	225	80	*NDUFA3*
3742	T	T/-	79.2/20.0	125	25	*NDUFA3*
3748	T	-/T	48.1/39.4	104	41	*NDUFA3*
3892	T	T/-	67.8/29.9	515	154	*NDUFA3*
4965	T	-/T	56.3/38.1	446	170	*TFPT*
5763	G	G/-	75.0/24.6	741	182	*TFPT, CDS*
6354	C	C/-	69.9/30.1	501	151	*TFPT*
6364	T	T/-	61.6/38.0	484	184	*TFPT*
6610	T	-/T	48.0/47.5	1222	580	*TFPT*
7261	AA	--	37.5	32		*TFPT*
7261	A	-/A	32.4/23.5	34	8	*TFPT*
7272	A	-	84.4	32		*TFPT*
7541	C	C/-	76.2/22.9	227	52	*TFPT*
7933	G	G/-	78.3/21.4	429	92	*TFPT*
8326	A	A/-	40.0/36.0	50	18	*TFPT*
8326	AA	AA/--	31.9/21.3	47	10	*TFPT*
8352	C	C/-	75.3/24.7	85	21	*TFPT*
9442	A	A/-	75.3/24.7	81	20	*TFPT*
9603	A	-/A	44.3/39.2	79	31	*TFPT*
9736	A	-/A	59.7/37.3	67	25	*TFPT*
9877	A	A/-	34.6/23.1	26	6	*TFPT*
9991	G	G/-	79.2/20.8	274	57	*TFPT*
10220	G	G/-	79.3/20.7	188	39	*TFPT*
10561	A	A/-	64.7/34.0	153	52	*TFPT*
10893	A	A/-	71.7/27.4	329	90	*TFPT*
11192	C	C/-	69.9/29.8	302	90	*TFPT*
11233	A	-/A	42.2/33.2	211	70	*TFPT*
11480	T	T/-	59.9/36.3	284	103	*TFPT*
12006	G	G/-	74.5/25.0	200	50	*TFPT*
12702	CTC	CTC/---	75.0/24.7	288	71	*TFPT*
13063	C	C/-	74.2/25.4	524	133	*TFPT*
13513	T	T/-	51.9/46.2	210	97	*PRPF31*
13519	T	-/T	62.0/34.0	200	68	*PRPF31*
14563	A	A/-	60.2/38.3	415	159	*PRPF31*
14716	A	A/-	48.9/39.4	94	37	*PRPF31*
14735	A	A/-	62.8/35.1	94	33	*PRPF31*
14858	G	G/-	68.8/30.5	311	95	*PRPF31*
15305	AAA	AAA/---	23.8/23.8	42	10	*PRPF31*
15474	A	A/-	77.4/22.0	164	36	*PRPF31*
16131	G	G/-	77.8/21.9	465	102	*PRPF31*
16370	T	-/T	37.1/24.3	70	17	*PRPF31*
16370	TT	--/TT	23.2/20.3	69	14	*PRPF31*
16599	T	T/-	54.3/41.0	427	175	*PRPF31*
17120	TTT	TTT/---	40.0/24.0	25	6	*PRPF31*
17289	T	T/-	71.8/25.1	529	133	*PRPF31*
18396	T	-/T	53.2/21.9	393	86	*PRPF31*
18767	A	-/A	46.2/43.6	117	51	*PRPF31*
18925	AAA	AAA/---	21.7/20.8	106	22	*PRPF31*
19573	G	G/-	70.6/28.9	948	274	*PRPF31*
21356	C	C/-	74.4/25.5	1036	264	*PRPF31*
22089	C	C/-	74.7/25.1	427	107	*PRPF31*
22302	C	-/C	55.7/43.0	619	266	*PRPF31*
22309	C	C/-	75.5/23.8	608	145	*PRPF31*
22961	A	A/-	73.2/25.2	523	132	*PRPF31*
23294	C	C/-	63.1/36.9	179	66	*PRPF31*
23314	T	-/T	50.0/40.9	176	72	*PRPF31*
23763	G	G/-	62.9/34.9	1135	396	*PRPF31*
23818	G	G/-	78.8/20.9	1313	275	*PRPF31*
23861	C	C/-	66.5/33.1	1109	367	*PRPF31*
24122	C	C/-	73.7/25.7	992	255	*PRPF31*
24741	A	-/A	57.1/40.1	352	141	*PRPF31*
24747	A	-/A	56.3/42.6	350	149	*PRPF31*
25184	G	G/-	78.6/21.3	1277	272	*PRPF31*
25474	G	G/-	70.5/29.4	797	234	*PRPF31*
25548	A	-/A	69.1/30.4	573	174	*PRPF31*
25613	G	G/-	58.5/40.5	1204	488	*PRPF31*
25671	C	C/-	63.8/35.9	1224	439	*PRPF31*
26120	CT	--	23.7	38		*PRPF31*
26121	T	-	60.5	38		*PRPF31*
26406	C	C/-	75.9/23.8	294	70	*PRPF31*
26848	G	G/-	70.5/29.2	359	105	*PRPF31*
28297	AAA	AAA/---	21.1/21.1	90	19	*PRPF31*
28297	AAAA	----	23.5	85		*PRPF31*
29326	T	T/-	76.2/22.0	395	87	*PRPF31, CDS*
30359	T	T/-	46.2/20.9	91	19	
30672	C	-/C	57.3/42.3	494	209	

Project 2: **Screening of the *SNRNP200* gene in a cohort of dominant RP patients**

At the time of the study, one of the major limitations of NGS, was the complexity of testing many individuals at the same time. Sequencing arrays allow in fact physical separation of only a few samples, and the enrichment systems for targeting particular genomic regions had limited scalability. A still prohibitive solution for multiplexing was the addition of nucleotide barcodes during the library preparation step, which is a limiting step in terms of time and cost when dealing with many samples.

We were confronted with this problem when we chose to test the *SNRNP200* gene as a candidate for dominant RP in a cohort of 96 patients. The strategy that we experimented with for this screening involved the pooling of long-range PCR products targeting the gene from each patient and their simultaneous sequencing via NGS (Roche 454). This approach avoided the expensive procedure of barcoding and required instead a validation step consisting of the Sanger sequencing of individual samples only for particular exons in which a novel change was detected. This method led us to the identification of novel RP mutations in the hBRR2 protein, and it proved to be more time and cost-effective than the classical method of sequencing single exons through the Sanger method. The results of this work were published as "*Next generation sequencing of pooled samples reveals new SNRNP200 mutations associated with retinitis pigmentosa*", in Human Mutation journal, on February 2011.

Candidate's roles:

- Preparation of samples for sequencing.
- Planning and execution of the sequence analyses.
- Downstream validation analyses by Sanger sequencing.
- Writing of the manuscript.

HUMAN MUTATION Mutation in Brief 32: E2246–E2258 (2011) Online

MUTATION IN BRIEF

HUMAN MUTATION

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

Next Generation Sequencing of Pooled Samples Reveals New *SNRNP200* Mutations Associated with Retinitis Pigmentosa

Paola Benaglio[1], Terri L. McGee[2], Leonardo P. Capelli[1,3], Shyana Harper[2], Eliot L. Berson[2], and Carlo Rivolta[1]

[1]Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland; [2]The Berman-Gund Laboratory for the Study of Retinal Degenerations, Harvard Medical School, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts, USA; [3]Department of Genetics and Evolutionary Biology, Institute of Biosciences, University of São Paulo, São Paulo, Brazil

*Correspondence to: Carlo Rivolta, Department of Medical Genetics, University of Lausanne, Rue du Bugnon 27, 1005 Lausanne, Switzerland, Phone: +41(21) 692-5451, Fax: +41(21) 692-5455, E-mail: carlo.rivolta@unil.ch

Contract grant sponsor: Swiss National Science Foundation, European Union, CAPES, and the Foundation Fighting Blindness; Contract grant number: 320030-121929, HEALTH-2007-201550, 3637/07-7

Communicated by Mark H. Paalman

ABSTRACT: The gene *SNRNP200* is composed of 45 exons and encodes a protein essential for pre-mRNA splicing, the 200 kDa helicase hBrr2. Two mutations in *SNRNP200* have recently been associated with autosomal dominant retinitis pigmentosa (adRP), a retinal degenerative disease, in two families from China. In this work we analyzed the entire 35-Kb *SNRNP200* genomic region in a cohort of 96 unrelated North American patients with adRP. To complete this large-scale sequencing project, we performed ultra high-throughput sequencing of pooled, untagged PCR products. We then validated the detected DNA changes by Sanger sequencing of individual samples from this cohort and from an additional one of 95 patients. One of the two previously known mutations (p.S1087L) was identified in 3 patients, while 4 new missense changes (p.R681C, p.R681H, p.V683L, p.Y689C) affecting highly conserved codons were identified in 6 unrelated individuals, indicating that the prevalence of *SNRNP200*-associated adRP is relatively high. We also took advantage of this research to evaluate the pool-and-sequence method, especially with respect to the generation of false positive and negative results. We conclude that, although this strategy can be adopted for rapid discovery of new disease-associated variants, it still requires extensive validation to be used in routine DNA screenings. ©2011 Wiley-Liss, Inc.

KEY WORDS: Next generation sequencing, retinitis pigmentosa, sample pooling, *SNRNP200*

INTRODUCTION

Retinitis pigmentosa (RP) is a group of hereditary retinal diseases characterized by the progressive degeneration of rod and cone photoreceptors. The disorder typically begins with night blindness in adolescence and proceeds

Received 3 November 2010; accepted revised manuscript 8 February 2011.

DOI: 10.1002/humu.21485

with gradual reduction of the peripheral visual field with eventual development of tunnel vision and, in some cases, virtual total blindness. Early detection of this condition has been achieved by measuring retinal function by electroretinographic (ERG) testing (Berson, 1993), which represents the most reliable diagnostic tool for RP at all ages. Vitamin A supplementation has been reported to slow the course of this condition (Berson et al., 1993). This disorder, which affects almost 1 in 4000 people worldwide, is genetically diverse and can be inherited as an autosomal-dominant, autosomal-recessive, X-linked trait, and in rare cases also as a non-Mendelian trait (Hartong et al., 2006; Rivolta et al., 2002).

By linkage mapping and candidate gene screening, more than 60 genes have been associated so far with non-syndromic RP (RetNet database, http://www.sph.uth.tmc.edu/retnet/); however, mutations in these genes account for only about half of all reported cases (Hartong et al., 2006; Sullivan et al., 2006). Discovery of new causative genes by a candidate-functional approach is hampered by the labor intensive and costly methods of sequencing target genes in large numbers of patients. New and efficient methods of screening are therefore necessary to aid in the discovery of the remaining fraction of RP genes. In this context, the development of strategies based on "next-generation," or ultra high-throughput DNA sequencing technologies, is starting to provide new tools to analyze panels of different genes in several patients and in a parallel fashion (Calvo et al., 2010; Daiger et al., 2010).

Some twenty causative genes for autosomal dominant forms of RP (adRP) have been identified so far, including several genes encoding pre-mRNA splicing factors: *PAP-1* (*RP9*) (Keen et al., 2002), *PRPF31* (*RP11*) (Vithana et al., 2001), *PRPF8* (*RP13*) (McKie et al., 2001), *PRPF3* (*RP18*) (Chakarova et al., 2002), and *SNRNP200* (*RP33*) (Li et al., 2010; Zhao et al., 2009).

Splicing is a ubiquitous process by which introns are removed from pre-mRNA to form mature mRNA. The enzymatic reactions take place in the spliceosome, a supermolecular complex containing five small nuclear ribonucleoproteins (snRNP) and ~200 other proteins (Jurica and Moore, 2003). The *SNRNP200* gene (or *ASCC3L1*; chromosome 2q11.2, MIM# 601664), encoding for the 200-kDa helicase hBrr2, is essential for the unwinding of the U4/U6 snRNP duplex, which is a key step in the catalytic activation of the spliceosome (Laggerbauer et al., 1998; Raghunathan and Guthrie, 1998). This protein is homologous to Brr2 from yeast and belongs to the DExD/H box protein family. It consists of two consecutive Hel308-like modules, each composed of a DExD/H box domain with ATPase activity and a Sec63 domain (Lauber et al., 1996; Pena et al., 2009; Zhang et al., 2009). Recently, two different mutations of hBrr2 have been found in two Chinese families with adRP that showed linkage to the *RP33* locus (Li et al., 2010; Zhao et al., 2009; Zhao et al., 2006). These mutations, p.R1090L and p.S1087L, were identified following screening of candidate genes within the *RP33* linkage interval and are both located in the first Sec63 domain. It has been shown that the corresponding mutations in yeast affect the helicase activity of Brr2 (Zhao et al., 2009). No other hBrr2 mutations have been identified and the prevalence of mutations in this gene among patients with adRP is yet unknown. No genetic analyses have been performed so far on large cohorts of patients or in families that were not pre-selected for segregation of the diseases with the *SNRNP200* (*RP33*) genomic region.

We present here the results of the screening of the *SNRNP200* gene (45 exons, 44 introns) in 96 patients from adRP families with unknown molecular genetic cause, mostly composed of Caucasian individuals. To reduce the time and costs required to screen such a large gene in several patients with classical techniques, we used ultra high-throughput sequencing technology on pooled samples from multiple patients (Ingman and Gyllensten, 2009; Out et al., 2009). The potential advantages and the limitations of this method are evaluated.

MATERIALS AND METHODS

Patients and controls

This study was carried out in accordance with the tenets of the Declaration of Helsinki and was approved by the Institutional Review Boards of the University of Lausanne and of Harvard Medical School and the Massachusetts Eye and Ear Infirmary, where the blood was collected and the patients were followed. Written informed consent was obtained from patients who participated in this study before they donated 10-30 ml of their blood for research.

In addition to a regular ophthalmologic examination, our evaluation included ERG testing, performed as previously described (Berson et al., 1993). Patients were characterized as autosomal dominant if their families

showed evidence of transmission of retinitis pigmentosa over two consecutive generations in at least one branch with or without evidence of reduced penetrance in other branches.

DNA from peripheral leukocytes was extracted from 191 unrelated patients with adRP. Ninety-six of these samples were used for screening with ultra high-throughput sequencing (UHTS), while the other 95 patients were analyzed only for those exons in which a mutation was confirmed after Sanger sequencing. Controls included 175 individuals with no history of retinal degeneration, and included 80 subjects with normal ERG. In instances where control genomic DNA was insufficient for direct genetic screening, it was amplified by using a whole-genome amplification kit, following manufacturer's instructions (REPLI-g Mini Kit, Qiagen, Venlo, The Netherlands).

UHTS and sequence analysis

The general workflow followed in this study is schematically illustrated in Figure 1. Specifically, the *SNRNP200* gene (chromosome 2, nt. 96,940,074 to 96,971,297 of GenBank entry NC_000002.11) was amplified in the initial set of 96 patients by 4 overlapping long-range (LR) PCRs of 9,009 bp, 10,474 bp, 12,145 bp and 5,594 bp in length, spanning in total an approximate 35-kb genomic region (nt. 96,937,041 to 96,972,001 of the same GenBank entry). Primers were those used in the work by Hinds *et al.* (Hinds et al., 2005), adapted in some instances to the region of interest (Supp. Table S1). PCR reactions were performed in a 10 µl final volume, including LA PCR Buffer II (TaKaRa, Otsu, Shiga, Japan), 4 mM of $MgCl_2$, 1µM of each primer, 0.4 mM of dNTPs, and 1 U of TaKaRa LA Taq (TaKaRa), with slight modification in the cycling conditions suggested by the supplier. For the quantification of PCR products, we loaded 1 µl of each reaction (96 x 4 LR-PCRs) on E-Gel 48 2% agarose gels (Invitrogen, Carlsbad, CA) and analyzed them by densitometry. We pooled equimolar PCR products, according to the measured intensity of the bands. To avoid over-representation of the overlapping regions after shotgun library preparation and sequencing, we chose to pool only fragments from non-overlapping LR-PCRs (fragment #1 pooled with fragment #3 and #2 with #4). Sequencing was performed with two runs of Roche 454 GS FLX Titanium, according to manufacturer's protocols and by using a gasket that separated the two pools. All sequence analyses were carried out with the software package CLC Genomics Workbench (CLC bio, Aarhus, Denmark). Sequence reads were first trimmed and filtered according to their quality score and length (quality limit value set to 0.001, defined in the software manual; minimum length of a read set to 25 nt) and then assembled onto the reference sequence. We used default local-gapped alignment, allowing reads to align if they have at least 98% identity for more than 98% of their length.

For detection of single-nucleotide substitutions, we applied the following restrictions on the sequence accuracy of the bases surrounding the variant to be called: within an 11-nt window, the average quality of the bases was set to 20 (PHRED score, corresponding to a base accuracy of 99%) and the maximum number of mismatches or indels accepted was 3 with respect to the reference sequence. We only considered calls having a minimal coverage of 1,000 reads, corresponding to at least 5 reads per allele per patient and to twice the threshold previously indicated for confident detection of variants (Ingman and Gyllensten, 2009). Finally, to be considered reliable DNA variants, all detected changes had to be present independently in the two technical replicates, represented by the two runs of sequencing, with at least a 0.5% frequency (corresponding roughly to 1 variant allele over 192 alleles).

Coordinates of detected DNA variants are given with respect to GenBank entry NM_014014.3, with +1 corresponding to the A of the ATG initiation codon.

Sanger sequencing and validation of mutations

To validate the changes detected by UHTS, we individually analyzed by the Sanger method (Sanger et al., 1977) the PCR products from each patient's DNA for 4 exons (16, 25, 37, and 38) containing putative mutations on either long- or short-range PCR templates (Supp. Table S1 and Supp. Methods). In addition, we sequenced exons 4 and 31 to further ascertain the precision of the variants called by the UHTS procedure.

Sequencing reactions were performed by mixing 5 µl of previously-purified PCR products (ExoSAP-IT, USB, Cleveland, OH), 0.75 µM of primers and 1 µl of BigDye Terminator v1.1 Cycle Sequencing kit (Applied Biosystems, Foster City, CA), and run on a ABI-3130XLS sequencer (Applied Biosystems).

To predict pathogenicity of amino acid substitutions we used both the PolyPhen (Ramensky et al., 2002) and MutPred (Li et al., 2009) algorithms. Possible mutations affecting splicing were tested with the NNSPLICE 0.9 program (Reese et al., 1997). Protein sequences were aligned by using tools form the CLC Genomics Workbench.

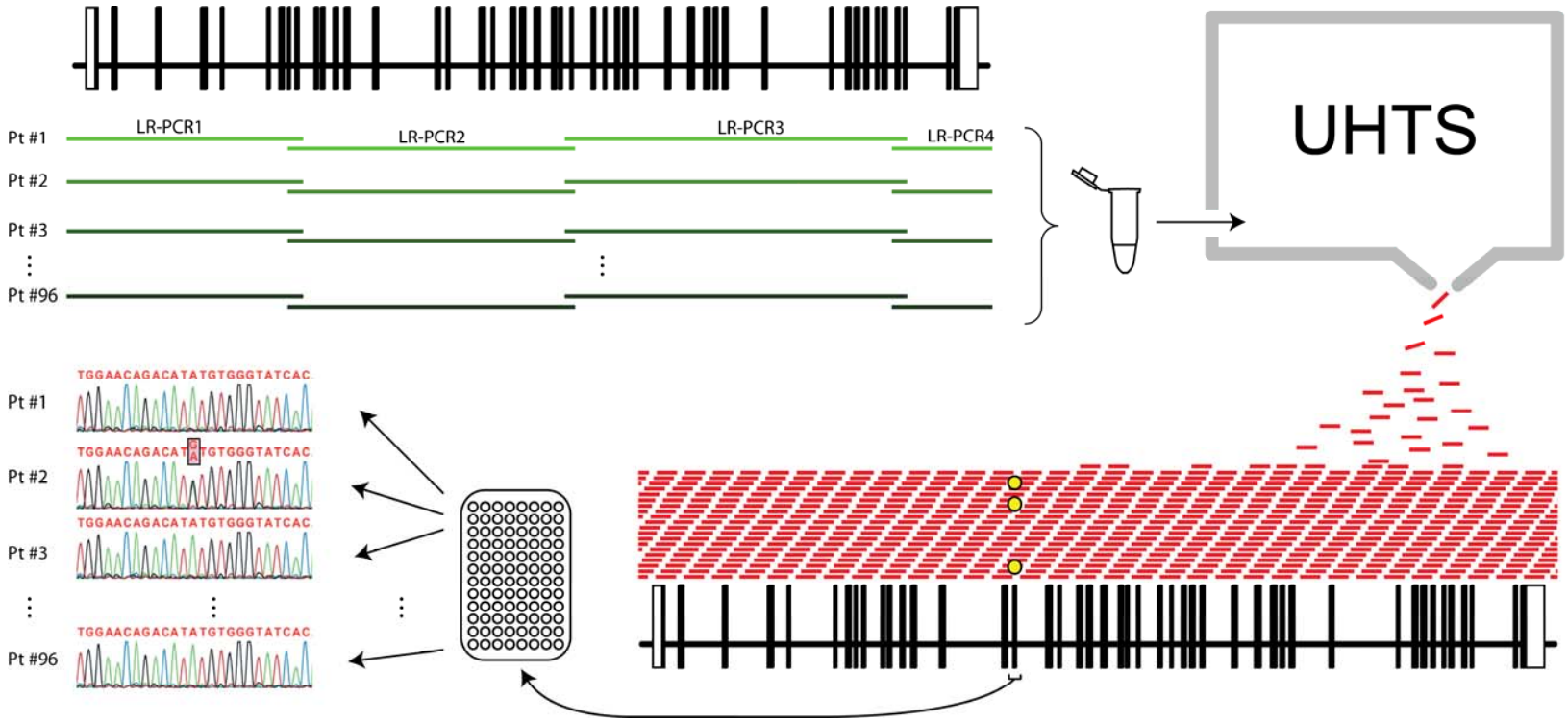


Figure 1. Schematic representation of the workflow used. Four long-range PCRs (LR-PCR, in shades of green) targeting the SNRNP200 gene (vertical boxes: exons; horizontal lines: introns) were performed on individual DNA samples from 96 patients, purified and pooled in equimolar quantities before UHTS. Following alignment to the reference sequence of the two million reads obtained (red bars), DNA variants were identified in the pool (yellow circles). To verify the presence of the detected variants, as well as to ascertain which patients carried them, relevant regions of *SNRNP200* were re-amplified by regular PCR in all 96 patients and sequenced individually by the Sanger procedure.

RESULTS

Sequencing

We sequenced the *SNRNP200* gene with the Roche 454 GS FLX Titanium instrument, as a pool of individually obtained LR-PCRs in 96 unrelated patients with adRP (4,320 exons, 4,224 introns, or ~3.5 Mb in total). We obtained in total ~2.3 million raw sequences of 314 nt in length on average. Following trimming and quality filtering procedures, 87% of them aligned to the reference sequence. The average base coverage obtained was about 7,500 fold, corresponding to ~40 sequences per single allele per patient in the pool, assuming an even representation of each sample. Ninety-six percent of the reference sequence was covered by at least 1,000 reads.

Ascertainment of DNA variants

We identified 79 DNA variants, including 33 annotated SNPs and 18 changes associated with homopolymeric stretches (i.e. AAAA…, CCCC…, etc.). Since these latter changes represent a well-known source of error for Roche 454 technology (Huse et al., 2007), they were immediately discarded from further analyses along with the identified known SNPs. Of the remainder variants, 21 were located within noncoding regions, 3 were predicted to produce isocoding changes, and 4 involved nonsynonymous changes. Putative isocoding changes were tested *in silico* for possible interference with the canonical splicing process, and none of them was predicted to be pathogenic. More specifically, the c.3315A>G (p.A1105) variant was in fact predicted to create a new donor site, but its associated likelihood score was not significant (0.43 out of 1.00).

To confirm the presence of DNA variants and identify the actual carriers among the patients' DNA composing the pool, we sequenced all exons carrying nonsynonymous variants as well as p.A1105 by the Sanger method in individual DNA samples. Whenever a change could be confirmed, the screening of that particular exon (and of its intron vicinities) was extended to the genomes of 95 additional unrelated adRP patients (Table 1).

In exon 16, the non-synonymous DNA change p.Y689C was confirmed by Sanger sequencing to be present heterozygously in one patient (Berman-Gund Laboratory patient ID: 001-107). Two missense variations, both affecting codon 681 (c.2041C>T and c.2042G>A, or p.R681C and p.R681H, respectively), were also identified by Sanger sequencing in two patients from the first cohort (IDs: 001-046, and 001-051). These variants were initially not detected by UHTS because they were present in the pool with frequency values that were below the 0.5% threshold and therefore can be considered as false negatives of the first method of screening. Sequencing of the second cohort allowed the identification of p.R681C in two additional unrelated patients (IDs: 001-061, and 001-303), as well as the detection of two new DNA changes, p.V683L and c.2160+42C>T in two patients (IDs: 001-485, and 001-346). None of these variants was present in 350 control chromosomes.

In exon 25, the change p.S1087L, detected by UHTS with a frequency of 1.4%, was present in two patients from the first cohort (IDs: 001-085, and 001-212) and 1 patient of the second cohort (ID: 001-367). The isocoding change p.A1105 was also confirmed to be present in two patients, one in each cohort (IDs: 001-090, and 001-060). Again, these DNA variants were absent in controls.

Sanger sequencing of the amplicons spanning exons 37 and 38 identified two false positives of the UHTS screening, p.F1717S and p.M1808V, both having a measured frequency corresponding exactly to the threshold value used in inclusion criteria. Alleles from SNPs rs772175 and rs78519182 were also confirmed to be present.

Cosegregation analyses

The p.S1087L mutation, found in 3 unrelated patients from our cohorts, was previously reported to be present in a family with adRP by Zhao *et al.* (Zhao et al., 2009).

The 4 new missense changes detected in exon 16, p.R681C, p.R681H, p.V683L, and p.Y689C involved highly conserved residues (Figure 2A and B) and were all predicted to be deleterious by *in silico* analyses. Family members were only available from two probands carrying p.Y689C and p.R681C. In these pedigrees, both changes were present heterozygously in patients and absent in unaffected members, following the classical pattern of inheritance of alleles causing a dominant disease with complete penetrance (Figure 2C).

The intronic change c.2160+42C>T, for which we also had other family members, did not co-segregate with the disease in the family and was therefore considered as non-pathogenic.

Table 1. SNRNP200 DNA variants in selected exons, detected by UHT and Sanger sequencing

DNA change*	Allele frequencies in the pool (%)	Coverage	Number of reads with allele	Putative amino acid change	Detected with UHTS	Patients in the 1st cohort (Sanger)	Patients in the 2nd cohort (Sanger)	Controls	Pathogenicity
Exon 16									
c.2041C>T	99.6/0.4	10,627	44	p.R681C	No (false negative)	1	2	0	Likely
c.2042G>A	99.9/0.1	10,656	10	p.R681H	No (false negative)	1	0	0	Likely
c.2047G>T				p.V683L	No	0	1	0	Undetermined
c.2066A>G	99.3/0.7	9,677	69	p.Y689C	Yes	1	0	0	Likely
c.2160+42C>T				Intronic	Yes	0	1	0	No
Exon 25									
c.3260C>T	98.6/1.4	20,258	290	p.S1087L	Yes	2	1	0	Confirmed
c.3315A>G	99.1/0.9	17,203	153	p.A1105	Yes	1	1	0	Undetermined
Exon 37									
c.5150T>C	99.5/0.5	12,064	65	p.F1717S	Yes (false positive)	0	ND	ND	No (not a real variant)
c.5317C>T (rs772175)	64.9/35.1	7,975	2,802	p.L1773	Yes	63 (alleles)	ND	ND	No
c.5324-31G>C (rs78519182)	98.0/2.0	7,553	149	Intronic	Yes	3	ND	ND	No
Exon 38									
c.5422A>G	99.5/0.5	9,409	51	p.M1808V	Yes (false positive)	0	ND	ND	No (not a real variant)

With the exception of c.5317C>T, all changes were detected in a heterozygous state. * Nucleotide numbering reflects cDNA numbering with +1 corresponding to the A of the ATG translational initiation codon in the reference sequence NM_014014.3. ND, not determined.

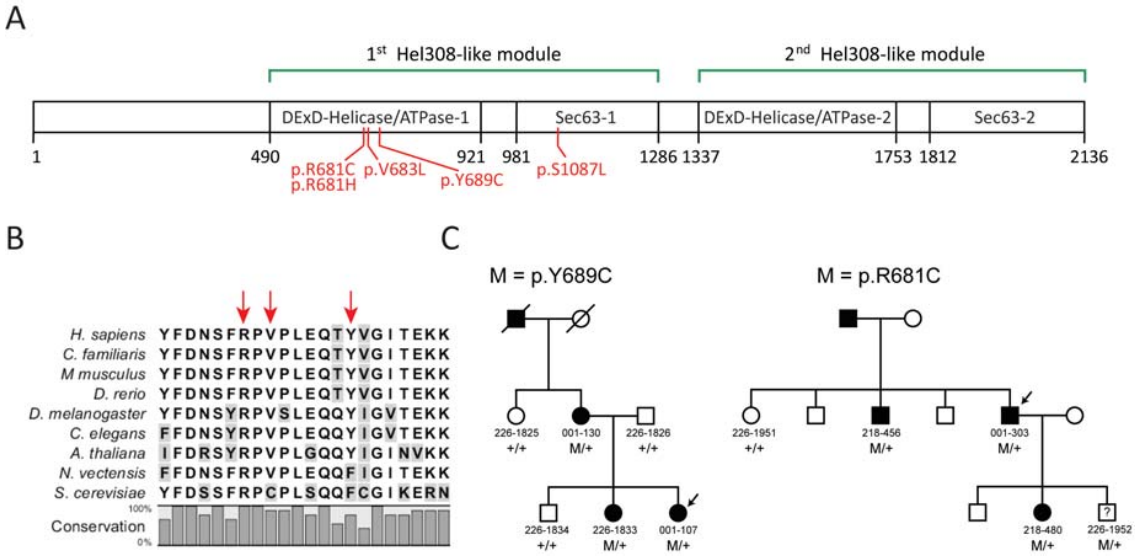


Figure 2. Structure and sequence of the hBrr2 protein and mutation cosegregation analysis. A) Functional domains of hBrr2. Positions of the mutations found in this screening are indicated in red. B) Alignment of Brr2 protein sequences from human, dog, mouse, fish, fly, worm, plant, sea anemone, and yeast. Non-conserved residues are shaded; arrows indicate the residues affected by DNA changes detected in exon 16. C) Pedigrees segregating the p.Y689C (Berman-Gund Laboratory family ID: 5632) and p.R681C (family ID: 0270) mutations (M) are shown. Black and white symbols represent clinically affected and unaffected members, respectively. The question mark indicates an individual for whom clinical examination was not possible. Arrows indicate probands analyzed in the UHTS screening.

Evaluation of variant detection specificity

To test the performance of the pooling method adopted here, we re-analyzed the sequence obtained by UHTS for exons 4, 16, 25, 31, 37, and 38 in the initial set of 96 samples. Specifically, we ascertained the number of variants detected by using different frequency thresholds (0.1, 0.2, … 1.0%) and compared them with the results obtained by individual Sanger sequencing of such samples. The number of false positives increased as the threshold of detection decreased, following a negative exponential trend (Figure 3).

DISCUSSION

The *SNRNP200* gene, encoding the splicing factor hBrr2, has been recently discovered by linkage analysis and molecular screening as a new autosomal dominant retinitis pigmentosa gene in two families from China. Two mutations were identified: p.S1087L and p.R1090L (Li et al., 2010; Zhao et al., 2009). Based on the evidence that hBrr2 is part of the same snRNP that includes PRPF31, PRPF3, and PRPF8, also involved in adRP, prior to the publication of these studies we screened this candidate sequence in a large cohort of unrelated patients from North America. Because of the elevated number of exons to be analyzed, we adopted previously published protocols consisting of the parallel sequencing of pooled and untagged DNA samples and evaluated them as potential methods for research on adRP, i.e. a rare disease with elevated genetic heterogeneity.

Three out of 191 patients from our screening carried p.S1087L (c.3260C>T), located in the first Sec63 domain of the protein. These patients were of Mexican, French Canadian, and English/Irish descent, indicating either that this mutation represents a relatively early event in human history, or that nucleotide c.3260 is a mutational hotspot. Haplotype studies on other ethnic groups and extended cohorts of patients are needed to verify which one of these hypotheses is the correct one.

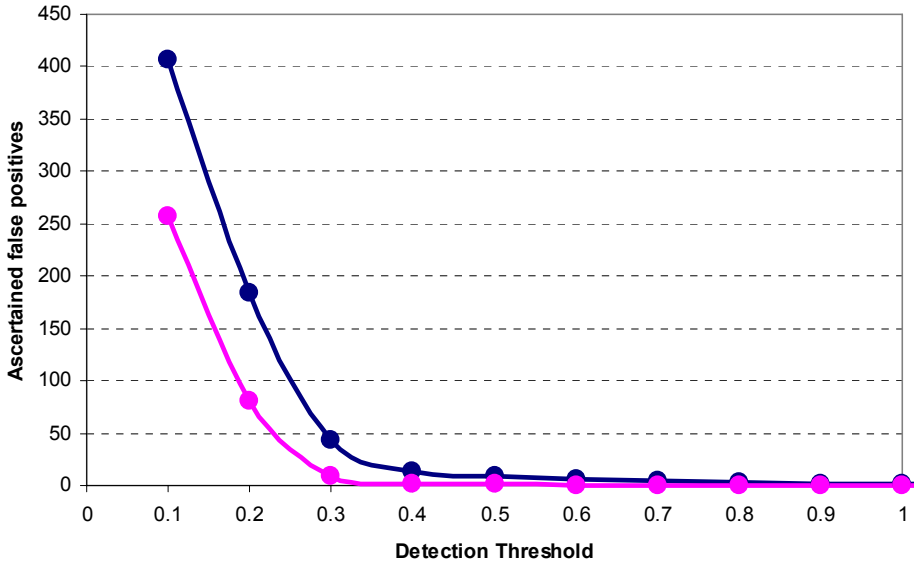


Figure 3. False positives identified as a function of different frequency thresholds. False positives were ascertained after Sanger sequencing of exonic DNA, covering in total ~3% of the entire *SNRNP200* region sequenced by UHTS. Blue line: number of false positive variants detected in the assembly of all the sequences obtained by two sequencing runs. Pink line: false positives detected in both of the independent technical replicates.

Importantly, we found 4 new missense variants in exon 16: p.R681C, p.R681H, p.V683L, and p.Y689C, which were present heterozygously in 6 patients and absent from 350 control chromosomes. Residues Arg681 and Tyr689 are phylogenetically very well conserved, and their replacement is predicted *in silico* to be damaging for the correct functioning of hBrr2. While both p.R681C and p.Y689C co-segregated with disease in the pedigrees analyzed, probands carrying p.R681H and p.V683L changes did not have other family members available for further genetic analyses. However, the p.R681H variation affects the very same conserved residue co-segregating with disease in the pedigree with p.R681C, strongly suggesting an association with RP. p.V683L was predicted to be possibly pathogenic by *in silico* analyses, involved a conserved amino acid, and was absent in the controls. In the absence of additional data (e.g. cosegregation) it is difficult to speculate at the present time whether it represents a rare benign variant or a true RP mutation.

All the newly detected changes fall in the Brr2 protein region containing the first DExD-helicase domain, which has been demonstrated to be essential for the U4/U6 unwinding function *in vivo* and *in vitro* and for cell survival in yeast (Kim and Rossi, 1999; Raghunathan and Guthrie, 1998). The first of the two consecutive Hel308-like modules, consisting of a DExD/H domain and a Sec63 domain, shows the highest level of conservation among species, reflecting its importance at the functional level (Zhang et al., 2009). It is therefore remarkable that all adRP mutations in hBrr2 so far identified are located in this first Hel308-like module. We hypothesize that these new mutations, similar to the ones already described, would impair hBrr2 helicase/ATPase activity, leading to defects in spliceosome catalysis.

Because of the high genetic heterogeneity displayed by RP, a very effective strategy for the identification of new disease genes consists in the screening of candidate genes in large cohorts of patients (Dryja, 1997). UHTS technologies (reviewed in (Metzker, 2010)) allow obtaining unprecedented amounts of DNA sequencing data, which makes them suitable for the screening of large genes. However, UHTS analysis of multiple samples is not a straightforward procedure, and unavoidably requires sample pooling to be economically sustainable. Current multiplexing procedures mostly rely on the addition of nucleotide barcodes to individual samples since the use of

physical separators does not grant sufficient parallelization (Craig et al., 2008; Lennon et al., 2010; Meyer et al., 2008). Detection of sequence variants in multiple samples can also be achieved through sequencing a pool of non-tagged DNA templates (for example PCR products covering the same gene) from different individuals and by ensuring an appropriate coverage in downstream UHTS procedures (Calvo et al., 2010; Ingman and Gyllensten, 2009; Out et al., 2009). This approach bypasses the expensive and laborious procedure of barcoding multiple libraries, and can theoretically lead to identification of rare variants, the frequency of which is as low as 0.5% with respect to the pool.

We followed this latter approach to analyze *SNRNP200* for mutations. In our screening we detected an unexpectedly high number of both false positive and false negative calls, which could be ascertained only by Sanger sequencing. False positive calling of mismatches is a necessary drawback, since in our study we were considering very low frequency variations that could also be caused by sequencing or alignment errors. We could reduce them by considering only the subset of variations detected in two independent sequencing runs, as demonstrated also by our simulation experiments using variable thresholds and as indicated by the reduction of the number of DNA changes associated to homopolymeric stretches (from more than 100 to 18, data not shown). However, false positive calls could not be completely eliminated. Using a higher threshold of detection could correct the problem but would also hide potentially true signals (false negatives), especially for variants that could be penalized by uneven pooling of different PCRs products and/or unbalanced allelic amplification during pre-sequencing procedures (Benaglio and Rivolta, 2010). In our specific case, we failed for example to identify two true changes, p.R681C and p.R681H, since they were present at a frequency that was below the theoretical limit of 1 variant allele in 96 samples (0.4% and 0.1% respectively). While failure to detect the first change could be attributed to statistical fluctuation, the second false negative call is more likely to depend on the underrepresentation of this allele in the pool, probably prior to sequencing. However, correcting the underdetection of true positive calls through the mere operation of decreasing the threshold would also result in an exponential increase of noise generated by false positive calls, making the fine tuning of this procedure a subtle and rather empirical process. A practicable possibility in this context could consist in pooling fewer samples and in raising the threshold of detection proportionally. In our case, for example, pooling 48 samples instead of 96 would have allowed detecting a single allelic variation in 1% of the sequences (instead of 0.5%), allowing therefore to increase the detection rate while keeping the noise under control.

The DNA screening strategy used in this work has proven to be extremely advantageous, especially if it is compared to the alternative option of individually sequencing all *SNRNP200* exons in the several dozen patients and controls examined. Specifically, the triage operated by UHTS of pooled samples allowed reducing the number of exons to be analyzed by an order of magnitude (from 45 to 4). However, in contrast to classical exon-PCR analyses by Sanger sequencing or to UHTS of single samples, the results obtained have a stochastic component that depends heavily on the settings used.

In conclusion, we identified new mutations in *SNRNP200* and confirmed that adRP associated to hBrr2 impairment is not limited to the Chinese population. Based on our data, the prevalence of *SNRNP200*-associated adRP seems to be rather high, since mutations were present in at least 4.2% (8 out of 191) of the screened patients. Considering that for 95 patients only 2 out of the 45 *SNRNP200* exons were analyzed and that the UHTS pooling technique used generated a number of false negatives, this value could potentially be even higher, making mutations in this gene one of the most frequent causes of adRP in Caucasians. Furthermore, we also tested the use of next-generation sequencing technology on pooled and untagged samples, highlighting the advantages and the limitations of this methodology for DNA analyses involving many patients. Based on our work, we are persuaded that candidate gene screening for RP and other genetic diseases will greatly benefit from the high-throughput revolution in the very near future, but this will probably follow the development of automated and inexpensive procedures for genetic barcoding or other solutions for sample multiplexing.

ACKNOWLEDGMENTS

We would like to acknowledge Ms. A. Title and Dr. M. Künzli, FGCZ, Zurich, Switzerland. Our work was supported by the Swiss National Science Foundation (grant # 320030-121929) and by the European Union (grant HEALTH-2007-201550).

This work was also supported by a Center Grant from the Foundation Fighting Blindness, Columbia, MD. (ELB).

Leonardo P. Capelli was sponsored by the CAPES program (Process 3637/07-7).

REFERENCES

Benaglio P, Rivolta C. 2010. Ultra high throughput sequencing in human DNA variation detection: a comparative study on the NDUFA3-PRPF31 region. PLoS ONE 5: e13071.

Berson EL. 1993. Retinitis pigmentosa. The Friedenwald Lecture. Invest Ophthalmol Vis Sci 34:1659-1676.

Berson EL, Rosner B, Sandberg MA, Hayes KC, Nicholson BW, Weigel-DiFranco C, Willett W. 1993. A randomized trial of vitamin A and vitamin E supplementation for retinitis pigmentosa. Arch Ophthalmol 111:761-772.

Calvo SE, Tucker EJ, Compton AG, Kirby DM, Crawford G, Burtt NP, Rivas M, Guiducci C, Bruno DL, Goldberger OA, Redman MC, Wiltshire E, Wilson CJ, Altshuler D, Gabriel SB, Daly MJ, Thorburn DR, Mootha VK. 2010. High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. Nat Genet 42:851-858.

Chakarova CF, Hims MM, Bolz H, Abu-Safieh L, Patel RJ, Papaioannou MG, Inglehearn CF, Keen TJ, Willis C, Moore AT, Rosenberg T, Webster AR, Bird AC, Gal A, Hunt D, Vithana EN, Bhattacharya SS. 2002. Mutations in HPRP3, a third member of pre-mRNA splicing factor genes, implicated in autosomal dominant retinitis pigmentosa. Hum Mol Genet 11:87-92.

Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA, Homer N, Huentelman MJ. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. Nat Methods 5:887-893.

Daiger SP, Sullivan LS, Bowne SJ, Birch DG, Heckenlively JR, Pierce EA, Weinstock GM. 2010. Targeted high-throughput DNA sequencing for gene discovery in retinitis pigmentosa. Adv Exp Med Biol 664:325-331.

Dryja TP. 1997. Gene-based approach to human gene-phenotype correlations. Proc Natl Acad Sci U S A 94:12117-12121.

Hartong DT, Berson EL, Dryja TP. 2006. Retinitis pigmentosa. Lancet 368:1795-1809.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. 2005. Whole-genome patterns of common DNA variation in three human populations. Science 307:1072-1079.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8:R143.

Ingman M, Gyllensten U. 2009. SNP frequency estimation using massively parallel sequencing of pooled DNA. European Journal of Human Genetics 17:383-386.

Jurica MS, Moore MJ. 2003. Pre-mRNA splicing: awash in a sea of proteins. Mol Cell 12:5-14.

Keen TJ, Hims MM, McKie AB, Moore AT, Doran RM, Mackey DA, Mansfield DC, Mueller RF, Bhattacharya SS, Bird AC, Markham AF, Inglehearn CF. 2002. Mutations in a protein target of the Pim-1 kinase associated with the RP9 form of autosomal dominant retinitis pigmentosa. Eur J Hum Genet 10:245-249.

Kim DH, Rossi JJ. 1999. The first ATPase domain of the yeast 246-kDa protein is required for in vivo unwinding of the U4/U6 duplex. RNA 5:959-971.

Laggerbauer B, Achsel T, Luhrmann R. 1998. The human U5-200kD DEXH-box protein unwinds U4/U6 RNA duplices in vitro. Proc Natl Acad Sci U S A 95:4188-4192.

Lauber J, Fabrizio P, Teigelkamp S, Lane WS, Hartmann E, Luhrmann R. 1996. The HeLa 200 kDa U5 snRNP-specific protein and its homologue in Saccharomyces cerevisiae are members of the DEXH-box protein family of putative RNA helicases. EMBO J 15:4001-4015.

Lennon NJ, Lintner RE, Anderson S, Alvarez P, Barry A, Brockman W, Daza R, Erlich RL, Giannoukos G, Green L, Hollinger A, Hoover CA, Jaffe DB, Juhn F, McCarthy D, Perrin D, Ponchner K, Powers TL, Rizzolo K, Robbins D, Ryan E, Russ C, Sparrow T, Stalker J, Steelman S, Weiand M, Zimmer A, Henn MR, Nusbaum C, Nicol R. 2010. A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. Genome Biol 11:R15.

Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. 2009. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25:2744-2750.

Li N, Mei H, MacDonald IM, Jiao X, Hejtmancik JF. 2010. Mutations in ASCC3L1 on 2q11.2 are associated with autosomal dominant retinitis pigmentosa in a Chinese family. Invest Ophthalmol Vis Sci 51:1036-1043.

McKie AB, McHale JC, Keen TJ, Tarttelin EE, Goliath R, van Lith-Verhoeven JJ, Greenberg J, Ramesar RS, Hoyng CB, Cremers FP, Mackey DA, Bhattacharya SS, Bird AC, Markham AF, Inglehearn CF. 2001. Mutations in the pre-mRNA splicing factor gene PRPC8 in autosomal dominant retinitis pigmentosa (RP13). Hum Mol Genet 10:1555-1562.

Metzker ML. 2010. Sequencing technologies - the next generation. Nat Rev Genet 11:31-46.

Meyer M, Stenzel U, Hofreiter M. 2008. Parallel tagged sequencing on the 454 platform. Nat Protoc 3:267-278.

Out AA, van Minderhout IJ, Goeman JJ, Ariyurek Y, Ossowski S, Schneeberger K, Weigel D, van Galen M, Taschner PE, Tops CM, Breuning MH, van Ommen GJ, den Dunnen JT, Devilee P, Hes FJ. 2009. Deep sequencing to reveal new variants in pooled DNA samples. Hum Mutat 30:1703-1712.

Pena V, Jovin SM, Fabrizio P, Orlowski J, Bujnicki JM, Luhrmann R, Wahl MC. 2009. Common design principles in the spliceosomal RNA helicase Brr2 and in the Hel308 DNA helicase. Mol Cell 35:454-466.

Raghunathan PL, Guthrie C. 1998. RNA unwinding in U4/U6 snRNPs requires ATP hydrolysis and the DEIH-box splicing factor Brr2. Curr Biol 8:847-855.

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30:3894-3900.

Reese MG, Eeckman FH, Kulp D, Haussler D. 1997. Improved splice site detection in Genie. J Comput Biol 4:311-323.

Rivolta C, Sharon D, DeAngelis MM, Dryja TP. 2002. Retinitis pigmentosa and allied diseases: numerous diseases, genes, and inheritance patterns. Hum Mol Genet 11:1219-1227.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463-5467.

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. 2006. Prevalence of disease-causing mutations in families with autosomal dominant retinitis pigmentosa: a screen of known genes in 200 families. Invest Ophthalmol Vis Sci 47:3052-3064.

Vithana EN, Abu-Safieh L, Allen MJ, Carey A, Papaioannou M, Chakarova C, Al-Maghtheh M, Ebenezer ND, Willis C, Moore AT, Bird AC, Hunt DM, Bhattacharya SS. 2001. A human homolog of yeast pre-mRNA splicing gene, PRP31, underlies autosomal dominant retinitis pigmentosa on chromosome 19q13.4 (RP11). Mol Cell 8:375-381.

Zhang L, Xu T, Maeder C, Bud LO, Shanks J, Nix J, Guthrie C, Pleiss JA, Zhao R. 2009. Structural evidence for consecutive Hel308-like modules in the spliceosomal ATPase Brr2. Nat Struct Mol Biol 16:731-739.

Zhao C, Bellur DL, Lu S, Zhao F, Grassi MA, Bowne SJ, Sullivan LS, Daiger SP, Chen LJ, Pang CP, Zhao K, Staley JP, Larsson C. 2009. Autosomal-dominant retinitis pigmentosa caused by a mutation in SNRNP200, a gene required for unwinding of U4/U6 snRNAs. Am J Hum Genet 85:617-627.

Zhao C, Lu S, Zhou X, Zhang X, Zhao K, Larsson C. 2006. A novel locus (RP33) for autosomal dominant retinitis pigmentosa mapping to chromosomal region 2cen-q12.1. Hum Genet 119:617-623.

SUPPORTING INFORMATION

Supp. Table S1. Sequences of primers used for long-range (LR), regular (SR) and sequencing (Seq) PCRs

Name	Sequence (5'-3')
LR_1.F	ACAGAGAAGACTTTGTAGCTGGGGAAGA
LR_1.R	CAGCCCTGAAATAAACTATATATGAAACAAGG
LR_2.F	TGGTTTGGTCATGAGACCAGTGACCTG
LR_2.R	ACACAAAACACAGTCATTAAAGGCAGACACTG
LR_3.F	GCTGCTCTTCATCTTTACCTCTAAGAA
LR_3.R	TAAACATGACAGTATCTGGTTTCTGCTATCAA
LR_4.F	CTGGTAGCTGGCTTGGTCAGGTGTCAACTCAC
LR_4.R	TGATGGGGAGGTGGCCTTCTGGAAGTATCAG
SR_ex16.F	GTTTTAGAAGGGCCTTTGGG
SR_ex16.R	TTTTAATTTCTGTCAATCTTCCCC
SR_ex25.F*	ACCGTGTGTAGAGTGGCTCA
SR_ex25.R*	TTCCCATCAGACCCTTGG
SR_ex37-38.F	GCGTATTGTCCACCAGTGATG
SR_ex37-38.R	TCCTCGATGCTGATGCACTT
Seq_ex4.F	TCCTTTAGTTGTGGCATCAGC
Seq_ex25.F	GCCGCAGCACTCTTCTAATTGT
Seq_ex31.R	TTTGGAATAGGGCAGCAGGTAG
Seq_ex37.F	TTAGGTCTCACACAGGGACCATG

* From Li et al, 2010.

Supp. Methods

Long-Range PCRs

Reaction mix

1x Buffer LA PCR™ Buffer II
4 mM MgCl2
1uM each primer
0.4 mM each dNTP
1 unit of LA Taq™ (TaKaRa)
10 ul final volume

Cycling conditions

LR 1
95˚C 5’(95˚C 30”, 67˚C 1’,68˚C 14’)x14, (95˚C 30”, 62˚C 1’,68˚C 14’)x16, 72˚C 10’

LR 2, 3, and 4
94˚C 1’(98˚C 5”,68˚C 15’)x30, 72˚C 10’

Regular PCRs

Exon 16

1x PCR Buffer
2 mM MgCl2
0.1 uM each primer
0.2 mM each dNTP
0.5 unit of HotStarTaq DNA Polymerase (Qiagen)
20 ul final volume

95˚C 15'(95˚C 30", 56˚C 30", 72˚C 1')x35, 72˚C 10'

Exon 25

1x PCR Buffer
1.5 mM MgCl2
0.1 uM each primer
0.2 mM each dNTP
1 unit of HotStarTaq DNA Polymerase
25 ul final volume

98˚C 8'(94˚C 30", 56˚C 30", 72˚C 1') x5, (94˚C 30", 54˚C 30", 72˚C 1')x5, (94˚C 30", 52˚C 30", 72˚C 1')x15, (94˚C 30", 50˚C 30", 72˚C 1')x15 72˚C 5'

Exons 37-38

1x PCR Buffer
0.5 mM MgCl2
0.1 uM each primer
0.2 mM each dNTP
0.5 unit of HotStarTaq DNA Polymerase
20 ul final volume

95˚C 15'(95˚C 30", 60˚C 30", 72˚C 1') x35, 72˚C 10'

Project 3 (Review)**: Methods for genetic screening of multiple samples using targeted NGS**

We were invited to write a paper for a chapter of a book titled “Genomics III: Methods, Techniques and Applications”, edited by iConcept Press. The editors were interested in an extended version of the paper "Ultra High Throughput Sequencing in Human DNA Variation Detection: A Comparative Study on the *NDUFA3-PRPF31* Region" (project 1), or other recent papers related to the book project. Since we were working on a new screening (that is described below, in project 4), and we had methodological materials from the previous one (project 2), we though it would be useful to perform a comparison of the two methods and discuss about their performances. This book chapter contains a detailed introduction in the form of a review of the methods used in NGS for targeted sequence enrichment and parallel analysis of multiple samples (multiplexing), with application in genetic screenings for Mendelian disorders. In the second part we reviewed the method already described in the paper of *SNRNP200* screening (project 2), and compared it with the new one used for project 4, consisting of a commercial method for the generation of multiple barcoded DNA libraries. The two strategies have been extensively dissected and discussed, providing useful recommendations for screenings projects. Additionally, two techniques for PCR product quantity normalization were evaluated.

The work has been peer reviewed by two anonymous experts, with the same process as for journal publications. The PDF of the accepted version is freely available on the website (www.iconceptpress.com) since October 2102, while the hardcopy of the book is in press.

Candidate’s roles:

- Design of the paper.
- Analysis of sequencing data.
- Writing of the manuscript.
- Corresponding author.

Strategies for Genetic Screening of Multiple Samples Using PCR-Based Targeted Sequence Enrichment

Paola Benaglio
Department of Medical Genetics
University of Lausanne, Switzerland

Carlo Rivolta
Department of Medical Genetics
University of Lausanne, Switzerland

1 Introduction

1.1 Next Generation Sequencing Technologies: an Overview

Biological research has been revolutionized by the introduction of dideoxy DNA sequencing, developed by Frederick Sanger et al. in the late '70s (Sanger et al., 1977). This technique, that essentially dominated the field of DNA analysis for the following 3 decades, was instrumental for the sequencing of the human genome in 2004 (International Human Genome Consortium, 2004) and still is very heavily used. The succeeding advent and rapid development of the so-called "next generation" or "ultra high throughput" sequencing (NGS or UHTS) technologies ushered in an era in which reading an organism's genome has almost become a routine practice. In addition to the sequencing of whole genomes, the development of different NGS methods and protocols has enabled a wide range of applications. The most used and best established ones are the sequencing of entire transcriptomes (RNA-seq), the sequencing of DNA from chromatin immunoprecipitation assays (ChIP-seq), DNA methylation profiling, and the analysis of genetic variations, especially in the field of medical genetics.

The distinctive feature of next generation sequencing is the possibility of producing very high number of sequences (or *reads*) in a fast and cost-effective manner. The most used platforms are currently commercialized by Roche 454 (GS FLX and GS Junior), Illumina (HiSeq, Genome Analyzer and MiSeq), and Life Technologies (SOLiD System and Ion Torrent sequencers). Each platform is characterized by a combination of different strategies for template preparation, amplification and sequencing, which lead in all cases to the parallelization underlying the drastic drop of the per-base cost of sequencing.

Template preparation is mostly based on a "shotgun cloning" approach (used to sequence long fragments of DNA) and includes the random shearing of the input DNA, usually through nebulization or sonication, followed by the clonal amplification of the fragments obtained. "Emulsion PCR", for example, is the method of library amplification used by Roche 454, SOLiD, and Ion Torrent technologies. Fragments of DNA, ligated to universal adaptors, are captured and amplified on individual beads in a water-in-oil mixture (Metzker, 2010). The enriched beads are then fixed on a glass surface (SOLiD) or deposited into PicoTiterPlate (PTP) wells

(Roche 454). In the Illumina platforms, the enrichment step occurs on a glass slide, where high-density primers are attached, and clonally amplified clusters are produced from the templates.

The sequencing reactions rely on different principles, based on either DNA polymerase or DNA ligase. Roche 454 uses pyrosequencing, in which the incorporation of complementary dNTPs results in the emission of photons (Margulies et al., 2005). The Illumina technology relies on a sequencing-by-synthesis approach, based on the cyclic incorporation of fluorescent nucleotides with reversible termination. Ion Torrent employs a similar approach, but uses non-modified dNTPs and a silicon chip that detects hydrogen ions released during each cycle of polymerization (Rothberg et al., 2011). In contrast to the previous methods, depending on the activity of DNA polymerase, SOLiD uses a ligase-based chemistry consisting of cycles of hybridization and ligation (McKernan et al., 2009).

The sequences produced by NGS are first collected as raw images and then processed to generate a readable output that has to be either aligned to a reference genomic sequence or assembled to form a "de novo" DNA sequence. In all cases, assembly, mapping, and analysis of sequences require the use of dedicated software.

Typically, NGS platforms produce reads of shorter length than those produced by Sanger sequencing, in the range of 50 to 400 nucleotides, depending on the platform. The throughput of these machines varies from several Megabases to hundreds of Gigabases per run and the time needed to produce such reads from a few hours to a few days. All of these features have to be taken into account and selected according to the desired output, in particular with respect to the desired *coverage*. The coverage is defined as the number of reads that interrogate a given DNA base and indicates how "deeply" a sample is sequenced. A rough way to estimate the average coverage of a sequencing experiment is to divide the total throughput (in base pairs) by the size of the DNA fragment that has to be sequenced. Sequencing redundancy, or high coverage, is necessary to reconstruct a correct sequence since NGS reads contain a higher proportion of errors than sequences obtained by the Sanger method and therefore every base has to be interrogated multiple times. In principle, the more accurate an instrument is, the less coverage is needed. Typical NGS errors are represented by an incorrect base call or small insertion and deletions. They can occur randomly or systematically in certain DNA regions that are more difficult to sequence, such as GC rich regions and homopolymeric stretches (i.e. TTTT..., AAAA..., etc) (Harismendy et al., 2009; Huse et al., 2007).

NGS companies are putting constant effort into improving the accuracy, the throughputs, and the flexibility of their systems. It is therefore rather difficult to give a contemporary picture of their technical features, due to the continuing evolution of the technology and specialization of these instruments.

1.2 Enrichment Strategies

The analysis and interpretation of genome-wide sequencing results is currently a step behind the technology that produces them. For certain applications it is more interesting -and economically more convenient- to concentrate the study only to a limited part of the genome. In medical genetics, for example, it is still a common practice to screen for genetic variants only a limited number of genes (for example a genomic interval associated to an inheritable disease) or all the coding sequences of a genome (the so-called exome) (Gilissen et al., 2011).

For this purpose, various strategies for enrichment of target DNA have been developed in recent years (Mamanova et al., 2010; Mertes et al., 2011). They can be divided mainly into PCR-, hybridization-, or circularization-based approaches. The choice of the enrichment strategy to be used depends on specific requirements of the project, and in particular on the target DNA size and the number of samples to be sequenced.

1.2.1 PCR-Based Enrichment Techniques

Polymerase chain reaction (PCR) is certainly the most reliable method for target enrichment, due to its high specificity, sensitivity and reproducibility. However, while this technique is very well suited to capillary electrophoresis (Sanger) sequencing, in which each amplicon is directly and separately analyzed, it is not completely adapted to NGS approaches. In fact, to exploit the full throughput of NGS and to perform an analysis that is economically viable, many samples and amplicons must be run at the same time. As it will be detailed in the next paragraph (1.3), the sequencing of multiple samples is limited by the long time and the relatively high costs required for library preparation. Multiple amplicons may be obtained via multiplex PCRs or pooling of single-plex PCRs. In both cases, the danger is an uneven representation of the amplicons forming the pool. This can occur from unequal PCR efficiency across the various amplicons or from unbalanced pooling of the fragments. Both of these events, if present, are in general difficult to correct. The manual production of PCRs is feasible for less than a hundred thousand bases of target DNA by using standard PCRs of few hundred bases, or by using long range PCR (LR-PCR) of approximately 10 Kbases. For larger target regions, the workflow involving primers design, optimization of robust PCRs, and generation and pooling of the products becomes less time- and cost- effective.

An automated solution for preparation of multiple PCRs (in this case standard PCRs) is provided by the RainStorm platform, produced by RainDance Technologies. In this system, emulsion-like PCR reactions occur as single-plexes in microdroplets, which are mechanically generated and assembled in a microfluidic system (Tewhey et al., 2009). Up to 20,000 primer pairs and corresponding number of reactions can be supported at the same time by this machine, which allows a relatively uniform enrichment of up to 10 Mb regions.

Some PCR-based approaches allow the simultaneous generation of short PCRs and library preparations by means of the incorporation of sequencing adaptors to the PCR primers, which must amplify fragments shorter than the sequencing reads. The advantage of this strategy is to avoid the cleaning, pooling, and shotgun library preparation required for longer DNA fragments (Mertes et al., 2011). The Fluidigm "Access Array System" employs this approach and allows, for example, the automatic preparation of 48 sample libraries using a microfluidic device that assembles and hosts 2,304 parallel separated PCR reactions.

Finally, certain platforms, such as Illumina Miseq and Life Technologies IonTorrent PGM, provide panels of ready-to-use highly multiplexed short PCR reactions to amplify target regions ranging from 1 Kb to 1 Mb, for up to 96 samples.

1.2.2 Hybrid Capture Enrichment Techniques

To select larger DNA regions (from 1 to 50 Mb of cumulative sequence), the hybrid capture method is preferred to the PCR method for its simplicity and rapidity, rather than for its specificity. The principle of sequence capture is based on the hybridization of a shotgun library to complementary probes of 60-150 nucleotides, designed to cover the target region. The hybridization reaction can occur in solution, or on a solid phase, where the probes are fixed on a microarray. The in-solution capture has the advantage of not requiring special equipment but a thermocycler, and of being more easily scalable. The main vendors of hybrid capture kits are NimbleGene (Roche) and Agilent; the first uses DNA probes, while the second longer RNA probes. The most used application of this technique is the enrichment of all the transcribed regions of the human genome, the so-called *whole exome capture*. One of the drawbacks of the hybrid capture method is the relatively high proportion of off-target and pseudogenes sequences, which have a negative effect on the coverage and the variant calling of the target region.

1.2.2 Circularization-based enrichment techniques

Another type of enrichment strategy is based on the use of custom molecular inversion probes (MIPs) or "gap-fill padlock", adapted from a SNP genotyping protocol (Akhras et al., 2007). MIPs are synthetic DNA oligonucleotides that contain a common linker sequence flanked by two single-stranded sequences designed to anneal to two nonconsecutive parts of a genomic target region. Such a region can be of up to 191 bp in length (Turner et al., 2009) or 500 bp when using longer padlock probes (LPPs) (Shen et al., 2011). Once hybridized, the gap between the two specific sequences is filled by the DNA polymerase and closed by a ligase reaction. These circular products are then amplified by PCR with primers annealing to the common linker. Since this latter sequence contains NGS adaptors as well, ready-to sequence templates that do not require further steps for library preparation are produced. A slightly different approach involves the use of "selector probes" which differ from the previous one because the genomic DNA is first digested by restriction enzymes and the resulting fragments circularize after the hybridization to the probes (Dahl et al., 2007). The advantages of these capture methods include high specificity and reproducibility, the characteristic of being library-free, and low input DNA (Turner et al., 2009). Disadvantages are represented by a poorer uniformity of the captured targets and high initial costs of the probes (Mamanova et al., 2010).

1.3 Multiplexing Strategies

The number of samples that can be simultaneously sequenced and resolved represents an important issue in NGS. While capillary electrophoresis sequencing offers the highest degree of scalability, with hundreds of samples and Kilobases sequenced per run, NGS platforms easily produce Gigabases of sequences, but typically distributed among few samples or only one. The reason for this is mainly physical: a capillary electrophoresis sequencer contains a few dozen capillaries that can be used simultaneously; NGS reactions occur on array-like surfaces with almost no separation to host different samples. The ideal application for NGS is therefore the sequencing of large amounts of DNA from an individual sample (e.g. a genome or an exome). In case of targeted resequencing, a smaller sequencing throughput is required and this is obtained either by using lower scale sequencers (the very new "benchtop sequencers" like Ion Torrent PGM, Illumina MiSeq and Roche 454 GS Junior) or by distributing the sequencing capacity of a big platform across many samples, to be sequenced at the same time. If such samples correspond to non-overlapping DNA sequences (e.g. each sample is constituted by an individual PCR product, representative of a unique genomic region), sample/sequence identification is performed a posteriori, when reads are assembled or mapped. Conversely, if each sample is constituted by a PCR product that originates from the DNA of a given individual and pairs of primers targeting the same DNA region for all samples, "multiplexing" procedures aimed at identifying individual samples become necessary. Such multiplexing is achieved during library preparation, via a step in which a sequence of DNA of 4-8 nt (the so-called *barcode*), unique to each sample, is ligated to all DNA fragments composing the library (Craig et al., 2008; Smith et al., 2010). Multiple libraries are then pooled in equal amounts and sequenced at once, along with their barcodes. The identification of samples occurs after the sequencing, thanks to the information contained in the genetic barcode tags.

Rather than technical, the real limitations of multiplexing are the high costs and the labor associated to sample preparation, which must be carried out separately for each sample in order to add individual nucleotide barcodes. As mentioned before, recent developments of PCR-based enrichment kits allow a greater automation in library preparation and a high level of multiplexing (up to 96 samples). However, this workflow is integrated for the moment only to low throughput sequencers (MiSeq, IonTorrent) or requires special equipment dedicated only to this process (Flugidim Access Array). Alternatively, other methods that do not require library preparation like MIPs can be appealing for processing many samples. Specifically, with MIPs barcodes can be

directly inserted in the primers that will amplify the captured sequences and the NGS adaptors (Akhras et al., 2007).

In case of recurrent genetic screenings performed on a same cohort of samples, a different strategy would be to initially tag the genomic DNA from different individuals with specific barcodes and NGS adaptor oligonucleotides, and then pool the barcoded fragments. Any downstream manipulation would be then performed on a single tube, which contains separable information of many individuals after the sequencing. This approach is available since the beginning of 2012 and is commercialized by PopulationGenetics, which also patented this workflow under the name of GenomePooling. In this approach, the regions of interest are extracted from the pool through specific primers and inverse simplex PCRs. Since they contain already both the individual barcode and sequences for NGS processing, they could be directly sequenced as a pool of samples (Casbon et al., 2011). If this technique demonstrates to have sensitivity and specificity comparable to other enrichment methods, it will represent a very powerful tool for genetic screenings via NGS.

Many scientists have tried to bypass the step of individual barcoding by using as a strategy the "anonymous" pooling (i.e. with no tagging nucleotides) of target DNA from different samples and the creation of a unique library for sequencing (Calvo et al., 2010; Lee et al., 2011; Otto et al., 2011; Out et al., 2009). Obviously, by this approach it is not possible to assign any direct relationship between reads and samples to which they belong, and further validations are required to track back the carrier of the variations via capillary sequencing or other methods. For some purposes, sample identification may not represent a primary necessity, for example if the aim of the project is to estimate the allele frequency of certain alleles in a population (Ingman & Gyllensten, 2009).

The mixed information contained in the results of such sequencing projects, in fact, must be interpreted based on the expected frequency of a single allele in a pool of chromosomes. For example, if 100 human samples have been pooled together, each autosomal allele will represent the 0.5% of the total sequence reads. In projects aiming at the identification of novel or rare variations (like disease causing mutations), the detection threshold of DNA changes must be therefore set at a frequency as low as 0.5%, depending on the number of individuals pooled together. For such experiments, it is important that the error rate of the sequencing platform does not exceed the expected frequency of one variation present in one individual. The risk is to produce many false positives, if the variant detection threshold is set too low, or false negatives, if this is set too high. This point will be elucidated in the following paragraphs through the description of a real example.

Two main different ways of obtaining pooled libraries are to group samples before or after target enrichment. For example, in case of PCR-based enrichment, template DNA can be quantified, pooled and amplified in a unique reaction. Alternatively, PCR fragments must be generated for each samples and pooled in a second time. The first approach is by far quicker and cheaper, but on the other hand the second one allows assessing the product of each reaction and produce a more balanced pool across all different samples (Otto et al., 2011; Out et al., 2009).

1.4 Genetic Screenings of Multiple Samples in Medical Genetics

In medical genetics, the discovery of genes causing Mendelian diseases has been classically achieved through linkage analysis of families or through the screening of candidate genes in large cohorts of patients. Linkage analysis and sequencing of the genomic region harboring the mutation are very powerful techniques if large families are available. In absence of large families to study, the candidate gene approach can be chosen. The hypothesis that a gene may cause the disease, based on its biological relevance and other known data, is tested through the screening of many patients with the same disease. Nowadays, with high throughput sequencing being progressively more affordable, the use of the new technologies simplifies and accelerates the discovery process.

Before NGS technologies emerged, candidate gene sequencing for detection of disease causing mutations was carried out through the Sanger dideoxy method, which is still the gold-standard method for molecular diagnosis in many hereditary diseases. However, the high costs and time required for sequencing entire genes in many patients by this method forced many laboratories to apply cheaper screening techniques such as single-strand conformation polymorphism (SSCP) and denaturing high-pressure liquid chromatography (DHPLC) prior to Sanger sequencing. NGS has the potential of substituting these procedures and offers a cost effective and accurate alternative to the Sanger method. The workflow of a NGS application is mostly defined by the enrichment and multiplexing strategies that are used, as described before. Successful examples include the use of commercial solutions such as the Raindance droplet-based multiplex PCR, or the Fluidigm microfluidic chip to test 86 known genes responsible for X-linked intellectual disability in 24 samples (Hu et al., 2009), or 3 known familial hypercholesterolemia genes in 144 samples (Hollants et al., 2012). Others have developed in-house methods to implement NGS in clinical diagnosis. For example, a pipeline based on a multiplex PCR enrichment step followed by a second PCR round to add sequencing adaptors was successfully applied to identify novel and known mutations in 3 genes responsible for the Marfan and Loeys-Dietz syndromes, in 87 patients (Baetens et al., 2011). The anonymous pooling strategy described before was also proven to be efficient in mutation discovery (Benaglio et al., 2011; Calvo et al., 2010; Otto et al., 2011; Out et al., 2009). In general, based on published data, it appears that the majority of NGS efforts aimed at analyzing multiple patients at once have been devoted to molecular diagnosis of known disease-genes, rather than to the discovery of new genes. Furthermore, it seems that a uniformed protocol for sequencing a small target region in many patients is not yet present.

Conversely, whole exome and genome sequencing are the most used and best-established strategies to discover new disease genes. Thanks to these approaches, discovery occurs through an unbiased analysis of the variants that are present in the entire genome or exome, in many cases helped by genotyping or sequencing information from family members. Whole exome or genome sequencing are more successful for discovering new genes associated with recessive conditions, since homozygous (or compound heterozygous) rare variations are less frequent in the genome with respect to heterozygous changes, and therefore are easier to identify and verify in terms of possible pathogenicity.

In the cases presented below we applied the candidate gene approach to identify new mutations in a cohort of patients affected with autosomal dominant retinitis pigmentosa (adRP), a diseases leading to progressive retinal blindness. RP may be caused by mutations in more than 100 genes, each of them responsible for a small fraction of the cases (Hartong et al., 2006). Diagnostic screening of known mutations are performed by using an arrayed primer extension chip by Asper Biotech or, more recently, by NGS of known genes after solid-phase capture arrays enrichment (Neveling et al., 2011; Simpson et al., 2010). Because of the high genetic heterogeneity displayed by RP, a very effective strategy for the identification of new disease genes, which are calculated to account for almost the half of RP patients, consists in the screening of candidate genes in large cohorts of patients (Dryja, 1997). Also, due to this genetic heterogeneity, the screening of single genes in a cohort of patients is expected to identify only a few individuals who are positive for a particular mutation.

We present two examples of single-gene screening in multiple patients (~100) using two different NGS strategies: the anonymous pooling approach and the tagged libraries pool approach. We chose long range PCR based methods, because of its high specificity, a crucial element to consider when many samples are processed together.

2 Anonymous Pooling Approach

The method presented here was applied to screen for heterozygous mutations an RP-associated gene (*SNRNP200*) and allowed us identifying new likely pathogenic DNA variants (Benaglio et al., 2011). Because of the elevated number of exons to be analyzed (45), we adopted a protocol consisting in the parallel sequencing of pooled and untagged DNA samples and evaluated it as a potential method for studying rare diseases with elevated genetic heterogeneity, such as RP. We sequenced this gene with the Roche 454 GS FLX Titanium instrument, by using as template a pool of individually-obtained LR-PCRs from 96 unrelated patients with adRP, accounting for a total of 4,320 exons, 4,224 introns, or ~3.5 Mb.

2.1 Experimental Methodology

2.1.1 Enrichment and Sequencing Method

The candidate gene of interest was amplified in a cohort of 96 patients by 4 overlapping LR-PCRs of 5 to 12 kbases in length, spanning in total approximately 35 contiguous kbases of the human genome. We quantified the resulting 384 PCR products by using pre-casted agarose gels and densitometry, before pooling them in equimolar amounts. Library preparation and sequencing was performed in agreement with the specific protocols for Roche 454 GS FLX Titanium. Two runs of such platform were performed.

2.1.2 Sequence Analysis

The analysis of the sequencing results was performed with the CLC Genomics Workbench software package (CLC bio, Denmark). We first polished the raw sequences by trimming the low-quality extremities of the reads and by eliminating the reads shorter than 25 nucleotides. Mapping was restricted to reads that could align to the reference sequence with at least 98% identity for more than 98% of their size. For reliable detection of single-nucleotide substitutions, we considered only high quality reads of the assembly, in highly-covered regions. Specifically, we set a minimum of 99% average base call accuracy (or 20 PHRED score) and allowed a maximum of 3 mismatches or insertions/deletions, calculated on a region of eleven nucleotides spanning the called variant. The variation detection frequency threshold was set to 0.5% over a minimum of 1,000 reads, which corresponded roughly to the identification of one heterozygous change in a pool of 96 samples (192 chromosomes), each allele being theoretically represented by at least 5 reads of relatively good quality. Finally, we used the information of the two independent sequencing runs as technical replicates and selected only DNA variants that were detected in both processes. Changes detected by UHTS were validated by sequencing individual PCR products from each patient's DNA by capillary electrophoresis, only for selected exons. Moreover, if a likely pathogenic change could be confirmed, we screened for that particular change an additional cohort of 95 unrelated individuals presenting with the same disease. Potential effects of amino acid substitutions were evaluated by the web-based software PolyPhen (Ramensky et al., 2002), while the involvement of isocoding DNA changes on gene splicing was tested by using NNSPLICE (Reese et al., 1997).

2.2 Results and Discussion

Each run of Roche 454 FLX produced roughly 1.2 million raw sequences of 314 nt in size on average. After quality trimming and filtering, 87% of the raw reads aligned to the 35-kb reference sequence, producing an average base coverage of about 3,750 fold, with a minimum coverage of 500 reads for the 96% of the targeted region. If we assume that each sample is equally represented in the pool, we obtain a ~20x average coverage

per single allele per patient, which is in the range of coverage recommended for confident DNA analysis (Bentley et al., 2008; McKernan et al., 2009).

The joint analysis of variant detection resulting from the two sequencing runs identified 79 DNA changes. We prioritized the analysis of candidate mutations according to the functional effect of such variations on the protein product of the gene. We therefore discarded all known polymorphisms (33), intronic and synonymous changes (24), and variations located within homopolymeric stretches (18), where pyrosequencing-based platforms are particularly prone to introduce errors (Huse et al., 2007) (Table 1). Validation of DNA changes by Sanger sequencing of individual DNA samples and the subsequent identification of the patient(s) carrying putative mutations was restricted to four non-synonymous substitutions, located in 4 different exons.

# Variants	Associated with homopolymers	Annotated SNPs	Intronic or non-coding	Synonymous	Nonsynonymous	Total
Merged	110	37	63	5	7	222
Intersection	18	33	20	4	4	79

Table 1: Number of single nucleotide variations identified in the pooled sequences with a frequency higher than 0.5%. Results obtained by either the assembly generated by merging the sequences of the two runs (Merged) or by retaining only those identified in both runs (Intersection - used in our experiment) are presented.

Of the four putative missense mutations detected in both sequencing runs, only two were confirmed by Sanger sequencing. They had an allele frequency detected via NGS of 0.7% and 1.4%, corresponding to one and two actual carriers, respectively. The other two missense variations identified by NGS with a frequency of 0.5% were false positives. Additionally, we identified two new missense variations through Sanger sequencing but not by NGS (false negatives). These variants, affecting the same codon, were initially not detected by NGS because they were present in the pool with frequency values that were below the 0.5% threshold (0.1% and 0.4%) and therefore were not included in the list of candidate variants.

These novel missense changes, were likely pathogenic mutations. Specifically, they involved highly conserved residues, were not detected in 350 control chromosomes, and were found in few additional unrelated patients after the screening of a second cohort. Moreover, for two of them the co-segregation of the DNA variant in the affected family members could be performed and was consistent with that of a pathogenic allele.

In addition to the clinical relevance of these findings, this study gave us the possibility of exploring the anonymous pooling method and to point out its advantages and limitations. One marked limitation of our screening was the high rate of false positive and false negative variation calls with respect to the true signals. The reason for this was mainly attributable to the low frequency that we set for calling variants, which inevitably brought to detection of mismatches due to sequencing or mapping errors.

To test the behavior of this method with respect to false positive discovery, we made a comparison between the number of variations detected by Sanger and NGS for 6 exons, covering in total ~3% of the entire gene. We were interested in ascertaining the number of false positives (i.e. variants detected by NGS but not by Sanger sequencing) at different thresholds of detection, ranging from a frequency of 0.1% to 2.0%. As predictable, false positives were detected in large amounts at low frequency thresholds. However, they were drastically reduced starting from a frequency threshold of 0.3% and virtually eliminated when such frequency was 1% or higher, according to an exponential curve (Figure 1).

Interestingly, we observed that the number of false positives was significantly lower when we counted only signals detected independently in both sequencing runs (i.e. the strategy that we adopted), with respect to merging the sequences obtained by the two runs. Similarly, we could also reduce by more than five folds the number of DNA changes that were present in homopolymeric stretches (Table 1), corresponding almost certainly to sequencing errors.

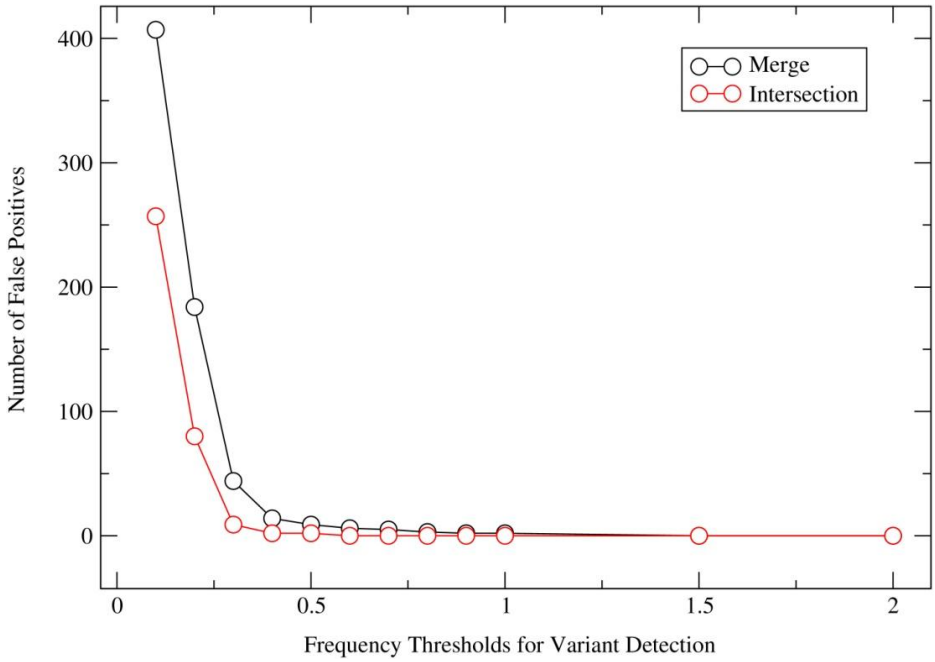


Figure 1: False positive variants as a function of different detection thresholds. Variations that were identified in the pooled sequencing experiments (merged dataset or intersection of the two sequencing runs) but not in the Sanger sequencing of the corresponding exons were considered as false positives. The comparison was performed over 6 exons of the gene.

Using higher thresholds of detection to correct the problem of false positives would also lead to the misidentification of potentially real variants, as we showed with our example. We could not in fact identify two true variations with an actual frequency of 0.4% and 0.1%, which were under the 0.5% limit that we chose for our analysis. While we can consider the first as a false negative (0.4%) because of stochastic deviation from the expected value, the second change (0.1%) was likely missed due to its underrepresentation with respect to the other alleles of the pool, rather than because of a sequencing error. This can happen for example when PCR products are pooled in an unbalanced quantity or when the two alleles from the same sample are differentially amplified due to the presence of a SNP near or inside the binding site of the PCR primers (Benaglio & Rivolta 2010; Ikegawa et al., 2002).

It seems therefore that the correct balance between low noise from false positive and sensitive detection of true variants is a fine process that cannot be predicted a priori. A feasible strategy to overcome this issue, also adopted in similar works (Calvo et al., 2010; Otto et al., 2011), is to pool a lower number of samples and

increase the detection frequency threshold consequently. For example, as it was showed by our simulation with different frequencies of detection, a marked improvement in specificity could be already observed at a frequency as low as 1%. This threshold could be used for variant detection of pools of 48 samples instead of 96 and theoretically allow the identification of 1 allele out of 96 with a lower noise due to sequencing errors.

To summarize, the DNA screening strategy presented here showed to be very efficient in finding new mutations, if compared to classical methods involving the individual sequencing and analysis of all the exons of a gene in hundreds of patients and controls. However, as a disadvantage with respect to classical exon-PCR analyses by Sanger sequencing or to UHTS of single samples, the results obtained with the pooled approach depend on stochastic variables that are difficult to control. The use of smaller pools of samples and more accurate sequencing platforms should almost certainly help increasing the efficiency of this method.

3 Tagged Libraries Pool Approach

Given the risk of missing potential mutations experienced in the anonymous pooling approach, we decided to test a safer, although more expensive technique involving the pool of tagged libraries. Tagged libraries should reduce the number of false positive and negative changes because the analysis is performed individually for each sample. A higher detection threshold for variant detection can therefore be used, and areas of potential errors can be more easily identified in those presenting low coverage. The target DNA in this screening consisted in two candidate adRP genes of 51 and 20 kb in size, amplified by 5 and 2 LR-PCRs, respectively. In order to achieve the parallelization required for library preparation of 95 samples, we chose a transposase-based method of fragmentation, described below. Moreover, we run a pilot test to evaluate the fragmentation protocol, the conditions for normalization of LR-PCRs, and the feasibility of merging untagged samples in single library preparations.

3.1 Experimental Methodology

3.1.1 Enrichment and Sequencing Method

Similar to the previous screening, long range PCR products of ~10 kb in length were individually obtained to target and amplify two candidate genes in a cohort of patients. The total number of exons and introns analyzed was 43 and 41, respectively. Prior to the screening of 95 samples, we conducted a pilot experiment on 18 libraries to test two methods of PCR product normalization. The first method, used for 8 samples, consisted in an approximate visual quantification of the 7 (5+2) long range PCRs on agarose gel and in an equimolar pooling that took into account their different PCR sizes. The PCR pools were subsequently purified and quantified. For 6 samples the purification and normalization of PCR products were performed with a commercial 96-well normalization plate (Invitrogen), which allows obtaining the same quantity of DNA for each product. Moreover, in the pilot test we wanted to assess the efficiency of variant detection when using the same barcode for multiple samples, in comparison with the results of individual sequencing. For this purpose, we included four libraries obtained by merging the pools of LR-PCRs from 2 or 4 samples, purified by the two methods just described (Figure 2).

For each sample (pool of LR-PCRs), a library was obtained by using a commercial transposase-based kit (Nextera), starting from 50 ng of input DNA. The protocol consisted in two thermocycler reactions. In the first reaction, an enzyme fragments and tags the DNA by means of appended transposon ends (“tagmentation”). After on-column purification of such product, a limited-cycle PCR is performed to add the barcoded adaptors, compatible with the sequencing platform. Eighteen different adaptors were used for the pilot study. After their

purification and quantification, the libraries were pooled together and run on one lane of Illumina GAII platform sequencer.

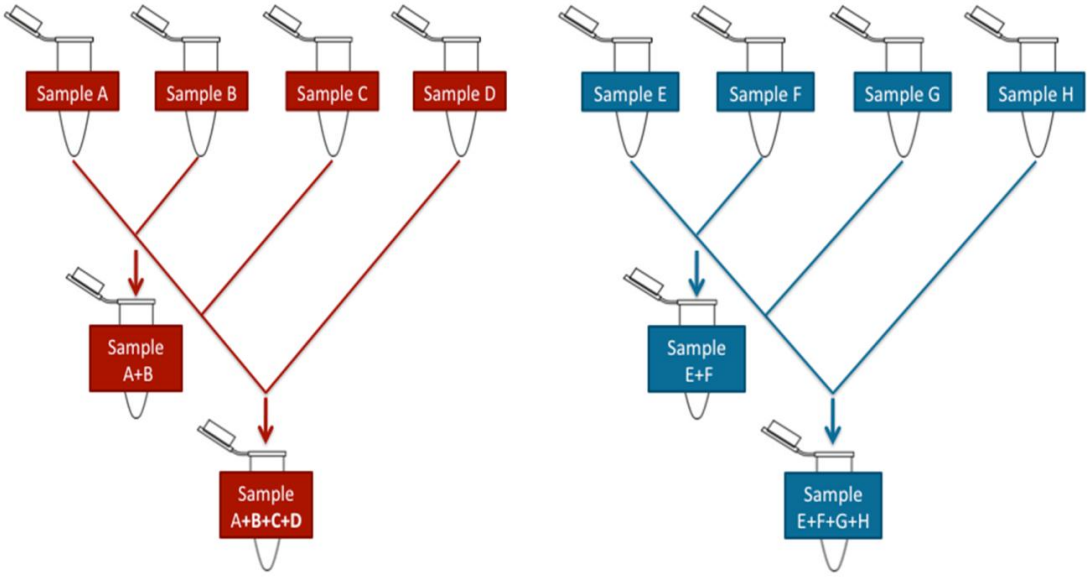


Figure 2: Design of the pilot study. LR-PCRs from samples in red were pooled by the "manual method" while the ones in blue by the "normalization method". We also merged two or four samples in order to test variant detection performance in non-tagged pools.

For the actual screening, we chose the first method of normalization ("manual method"), for reasons that will be detailed below. Libraries were again prepared for each sample via the commercial tagmentation protocol (Nextera), using 48 different barcoded adaptors optimized to allow the highest difference in sequence also in case of miscalled bases (Meyer & Kircher 2010). The 48 libraries were individualy purified and quantified, and then pooled together and sequenced. The same procedure was repeated for another set of 47 samples, to complete the sequencing of 95 different samples.

3.1.2 Sequence Analysis

The resulting sequences for each sample, separated according to the barcode that identified them (but which is not considered in the mapping), were aligned to the human genomic reference sequence of the two target genes (71 kb). We considered only reads having at least 95% identity for the 95% of their length. For confident variation calling (single nucleotide substitutions and small insertions and deletions), we set the threshold of detection to at least 20% of reads having a different base with respect to the reference sequence, and having a minimum coverage of 15 reads. In the experiments of pooled samples, the detection thresholds were set to 15% and 8% in the 2-sample and 4-sample pools, respectively.

3.2 Results and Discussion

The global metrics resulting from the pilot run, as well as the two discovery runs, are summarized in Table 2.

	Run #0 (Pilot)	Run #1	Run #2
Instrument	Illumina GAII	Illumina Hiseq	Illumina Hiseq
Number of multiplexed samples	18*	48	47
Number of Reads	22.7 M	143 M	149 M
Read length	77 nt	51 nt	101 nt
Average number of reads per sample	1.26 M	2.9 M	3.1 M
Target DNA size	71 kb	71 kb	71 kb
Total mapped reads	19 M (84%)	120 M (84%)	125 M (84%)
Average number of mapped reads per sample	1 M	2.5 M	2.6 M
Average base coverage per sample	1,140x	1,800x	3,900x
Percentage of target DNA covered at least 1000x	52%	80%	98%
Percentage of target DNA covered at least 500x	87%	95%	100%

Table 2: Summary of the results of the sequencing runs described in this section. (*) Eight samples were obtained by manual normalization of PCRs, 6 samples by commercial normalization, and 4 samples corresponded to the pools of untagged samples.

3.2.1 Pilot Run

The pilot run (one lane of Illumina GA II flow cell) produced a total of 22.7 million reads of 77 nt (1.7 Gbases) that could be assigned to a specific sample according to the index sequence. For each sample this resulted in 1.3 million raw on average, with a standard deviation of 380,000; i.e. there was on average 25% variation between the number of reads assigned to each sample. This indicated that the libraries were pooled in a relatively balanced proportion and there was no major over-representation of a sample with respect to another.

Due to the shorter length of the sequences produced by the Illumina with respect to the Roche 454 sequencer, we decided neither to trim nor to filter the reads before the alignment process. Indeed, we observed that using trimmed or raw reads gave similar results in terms of percentage of mapped reads, even by using stringent parameters. We also confirmed that reads that did not align to the reference sequence were automatically discarded from the mapping, thus avoiding creating noise in subsequent variation detection. Nineteen millions reads (84%) aligned to the reference sequence of the two genes analyzed, with one million reads per sample on average. The percentage of reads that are used in a mapping procedure gives an indication of the accuracy of the platform (depending on the stringency of the used parameters), but it also reflects the specificity of the enrichment procedure. PCR enrichment results in general in high sequence specificity, if primers are well designed and no off-target amplification occurs. Nevertheless, for long range PCR it is often difficult to optimize robust conditions to specifically amplify DNA of different qualities and therefore failure rate of amplification is relatively high. For all experiments performed, we checked each PCR product on agarose gels and selected for sequencing only those displaying good quality of bands. Even though this procedure ensures in principle to have a high level of specificity in the final sequencing results, for very few samples we still had mapping percentages that were significantly below the average. For example, one sample of the pilot run produced only 39% of the reads mapping on the target genes, while 54% of them mapped to other chromosomes. Such aspecific enrichment was likely due to the lower yield of PCR amplification for certain DNA samples of inferior quality, rather than to environmental genomic contamination, which nonetheless cannot be excluded.

The average coverage per base per sample was about 1,000x, a very high figure for analysis of single samples. We were therefore confident that pooling a higher number of samples for the real screening would guarantee a sufficient coverage for these analyses.

As mentioned, we also tested the efficiency of two different techniques for the normalization of PCR products before pooling (visual inspection vs. commercial normalization kit). The efficiency of these procedures was scored by using the coefficient of variation of the average coverages calculated for each long range PCR of one sample (i.e the standard deviation of the coverage of LR-PCRs divided by the mean of the average coverage of LR-PCRs). This value was on average 0.4 with no significant differences across the groups of samples normalized manually or by the commercial plates. In this latter system, the DNA is normalized according to the total amount of DNA, not to the number of molecules. Consequently, an equimolar normalization occurs only for DNA molecules of the same size, but not for long range PCR products of different sizes. When we then measured the coefficient of variation of the average coverage of a given long range PCR across different samples, we obtained indeed a slightly better level of normalization of products with the commercial respect to the manual method (mean coefficients of variation = 0.30 vs. 0.42). These concepts are more clearly represented in Figure 3, showing specifically that in the manual normalization method (panel A) the coverage of long range PCR products are almost randomly variable, despite the effort of visual quantification, while in the commercial normalization method (panel B) the coverage is more consistent across the same LR-PCRs but varies with an inverse proportion with respect to their size. Specifically, shorter fragments (such as fragment #3) tend to have a higher coverage than longer ones (e.g. fragments #2 and #4). From the same plots we can also observe peaks of higher coverage where contiguous LR-PCR fragments overlap and therefore artificially create DNA intervals having a higher number of copies.

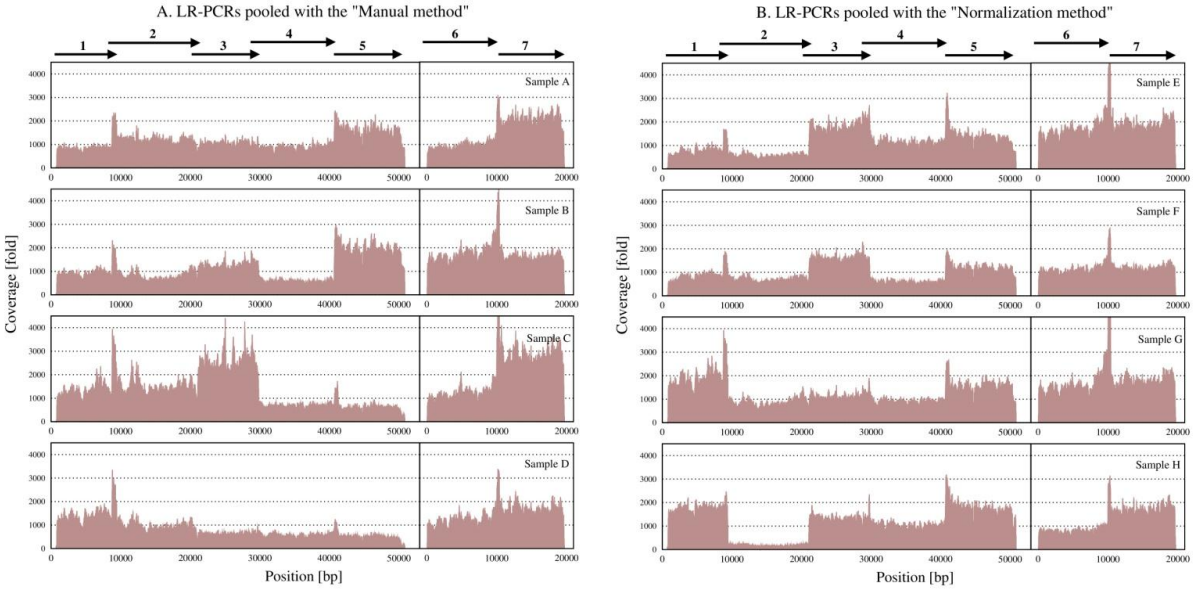


Figure 3: Coverage Plots. Base coverage values are represented for each point of the target regions that were amplified by the 7 LR-PCRs drawn at the top of the plots. The exact lengths of LR-PCRs are: 8.6, 12.2, 8.8, 11.6, 10.2, 10.3, and 9.7 kb. Panels on the left are examples form LR-PCRs that were pooled with the manual method, while those on the right show LR-PCRs normalized by the commercial kit.

Taken together, these observations indicate that there is no real advantage in using commercial normalization kits such as the one that we tested, especially if we consider that the coverage obtained with high throughput platforms is very high and allows a confident detection of variations also at depths that are inferior to the average. This said, for other applications such as detection of structural variants or copy number variations, in which the analysis of the local changes in coverage is important, it will be necessary to operate a reliable normalization step.

Another element that we wanted to test was variant calling sensitivity in pooled samples. The rationale was the same as the one used in the previous experiment, i.e. to save on library preparation reagents by merging different samples into a unique library. Unlike in the previous project, we tried to pool only two or four samples together, instead of 96 (we had previously concluded that the main limitations of the test was lack of accuracy due to the high number of pooled samples). We also used a different sequencing platform (Illumina instead of Roche 454). To perform this new test we sequenced four samples separately and as a pool of 2 or 4 samples, and then compared whether there were any differences in the variants detected by the two sequencing strategies. Four samples were purified by using the normalization kit and four others by manual pooling (Fig. 2). The results of such test are reported in Table 3, and examples of the output in Table 4. The minimum frequency of reads expected for one allelic change in a two-sample pool is 25% and in a four-sample pool is 12.5%. Therefore, we set compatible thresholds of detection (15% and 8%) to take into account normal experimental fluctuations as well. When we compared the results obtained from single nucleotide substitution analyses, we observed that globally the values at which polymorphic alleles were detected in the pools were very close to the expected frequency (Table 4, examples 1 and 2). Some exceptions were however present and resulted to be due to an uneven pooling or to an unbalanced amplification of specific long range PCRs. When such events are particularly intense, false negative results tend to occur, and the effect is larger when the number of samples increases (Table 4, examples 3). For example, in the manual pool experiment with four samples, we observed that false negatives were almost exclusive of one particular sample, which was likely under-represented in the pool with respect to the other three. Conversely, unbalanced allelic amplification, which occurs when the two alleles are differentially amplified, is recognizable while analyzing the frequency of detection in individual samples. For instance, a clear indicator for this event is a substantial deviation from the 50/50 proportion for heterozygous SNP alleles. False positive variations (Table 4, example 4) were also present with a frequency of detection immediately above the threshold, likely because of an effect of sub-optimal sequencing or mapping events.

We performed the same analysis for the detection of small insertions and deletions, using the same parameters. The rate of false positives for detection of these variant resulted to be significantly higher. These errors, which consisted in the wrong incorporation of a base during sequencing, were especially localized in long DNA stretches of the same nucleotide, similar to the phenomenon observed with the Roche 454 sequencer. It seems therefore that Illumina platforms are as well sensitive to this problem, at least when insertions or deletions are considered.

Based on all these observation, we concluded that the strategy of pooling samples might be an efficient technique; however, it does not guarantee perfect mutation detection, even when a few samples are pooled. Consequently, we chose to sequence individual samples as separate libraries for the screening of new candidate disease genes.

			Single Nucleotide Substitutions				Small Insertions/Deletions			
	Samples	Frequency of detection	# Detected Variants	# Expected Variants*	# False +	# False -	# Detected Variants	# Expected Variants*	# False +	# False -
Manual pooling	A	20%	91				12			
	B	20%	100				10			
	C	20%	70				12			
	D	20%	70				11			
	A+B	15%	112	115	1	4	28	16	12	0
	A+B+C+D	8%	151	148	12	9	44	22	22	0
Normalization pooling	E	20%	82				9			
	F	20%	96				9			
	G	20%	59				10			
	H	20%	112				18			
	E+F	15%	106	107	1	2	23	12	11	0
	E+F+G+H	8%	158	171	4	17	42	24	24	6

Table 3: Number of variants identified in individual samples and in pools of two or four of them, using appropriate detection frequency thresholds for single alleles. (*)Expected variants were calculated as the sum of unique variants found in the individual samples used for the pools.

Samples	Alleles	Alleles Frequencies	Read Count per Allele	Coverage	Alleles	Alleles Frequencies	Read Count Per Allele	Coverage
	Example 1				Example 2			
A	T/C	54.1/45.9	396/336	732	C/A	50.3/49.5	518/509	1029
B	T/C	51.1/48.9	393/376	769	C	99.9	1442	1444
C	T/C	50.9/49.1	556/536	1092	C	99.4	1108	1115
D	T/C	54.7/45.3	594/491	1085	C	99.9	999	1000
A+B	T/C	52.8/47.2	293/262	555	C/A	77.4/22.6	727/212	939
A+B+C+D	C/T	50.5/49.3	464/453	918	C/A	89.9/10.1	1118/126	1244
	Example 3 (False Negative)				Example 4 (False Positive)			
A	A	100	692	692	T	100	677	677
B	A	100	575	575	T/C	98.5/1.5	535/8	543
C	A	99.8	633	634	T/C	99.3/0.7	550/4	554
D	A/G	54.3/45.7	238/200	438	T/C	99.1/0.9	422/4	426
A+B	A	99.7	664	666	T/C	88.5/11.0	554/69	626
A+B+C+D	A/G	93.1/6.9	471/35	506	T/C	87.5/12.4	474/67	542

Table 4: Examples of SNP alleles detected in individual and pooled samples. The thresholds for variant detection are the same as in Table 3. In Example 1 all samples have the same heterozygous change, which was correctly detected with ~50% frequency in both pooled samples. In Example 2 one sample out of four carries a SNP that was detected at the corresponding frequency in the two-sample (~25%) and in the four-sample (~12.5%) pools. Example 3 shows a false negative: allele G of sample D was detected in the pooled sample at a lower frequency (6.9%, highlighted) than the detection threshold, set at 8%. Example 4 shows a false positive detection in the four-sample pool (highlighted), in which reads carrying likely a sequencing error were counted as carrying a variant, due to the lower frequency threshold set in the pooled sample.

3.2.2 Screening Runs

The screening was performed with two lanes of Illumina Hiseq flowcells, each one processing 48 different samples that were multiplexed with barcoded adaptors. As reported in Table 2, each run yielded in total about 150 million reads, almost 6 times higher than the ones obtained with previous version of the platform (GAII) for the pilot experiment. After mapping procedures, 84% of the reads aligned to the reference sequence of the two genes and produced an average coverage of about 2,000 and 4,000x for each sample of the first and the second run, respectively. This was largely due to the fact that in the first run reads were 51 nt long while in the second they were 101 nt long (Table 2). The coefficients of variation calculated over the average coverages of every LR-PCR were 0.46 and 0.38, respectively, indicating that PCRs were pooled with less than half a fold difference on average. If we consider the coefficient of variation of the average coverage for each sample, we obtain 0.23 and 0.16 for the first and second round, respectively, suggesting a good balance among samples.

We then performed variant detection on every sample to look for candidate mutations. When we excluded all variants reported in dbSNP, a database of human polymorphisms, we obtained a list of 139 unique variants. We also excluded 48 changes that were judged to be in stretches of bad alignment or low coverage by visual inspection of the mapping. Among the remainder 91 novel variants, we did not find any good candidate mutations, for the main reason that they were almost all intronic changes with no obvious effects on splicing of the gene transcripts, as predicted by the program NNSPLICE. The only novel nonsynonymous substitution was also predicted by PolyPhen to be non-pathogenic.

Although we could not identify any variants with possible implication in the disease that we were studying, we could get some insights on the sequencing methods that we selected for the screening of candidate disease genes. In particular, preparing many libraries in parallel through the “tagmentation” method was proven to be extremely fast (only a few hours needed to process 96 samples) and reliable. While at the time of this experiment reagents were still rather expensive (the cost was comparable to the one for normal library preparation via nebulization and size selection), their price is now starting to drop considerably, making screenings of few genes for many samples more affordable. This latter feature, together with the time effectiveness of such kits, renders the multiplexed barcode method a better choice with respect to the anonymous pooling screening. In fact, with a relatively limited amount of additional effort and, likely, of costs, the results produced by this new approach contain ready-to-use information for a complete panel of individuals, with minimal necessities of validation by additional methods. Moreover, if sufficient coverage and sequencing quality are ensured, false discovery rates should be in principle very low.

The biggest limitations that we could observe about this approach, but which are nonetheless present in the anonymous pooling approach as well, are mainly related to long range PCR robustness and scalability. For our purpose, which was based on the sequencing of a small region for many samples, LR-PCR was the best choice for several reasons. It (i) allowed very specific amplification of the target region, (ii) required relatively few reactions to cover up to 100 kbases, (iii) included non-coding region, and (iv) could be achieved by using standard laboratory procedures and equipment. On the other hand, long range PCRs are not as robust as short range PCRs and are very sensitive to the quality of the template. Optimization of robust PCRs may require a certain effort, and even when the right conditions are found it is still possible to observe failure in amplification of certain DNAs, which must be re-amplified or discarded from the screening. The time and effort necessary to obtain a complete panel of LR-PCRs for many samples to be processed at once can be indeed relatively long with respect for example to enrichment by hybridization. Additionally, PCRs must be individually checked, pooled in equimolar amounts, cleaned and quantified before starting the library preparation. Automated or semi-automated PCR-based enrichment methods, such as the ones outlined in the Introduction, could offer an interesting alternative to overcome these issues.

4 Conclusions

The two methods described in this chapter represent two different ways of addressing the scientific question of whether several patients affected by a specific disease carry pathogenic mutations in a given gene. Both approaches, the anonymous pooling and the tagged multiplexing, can lead to an answer. The first method is faster and cheaper, but provides positive answers only when mutations in a candidate gene have a relatively high frequency in the cohort that is analyzed. Furthermore, in case of negative results, it is not possible to rule out that the screening was poisoned by false negative events. Since the differences in costs and time required to perform the first vs. the second method are progressively diminishing, it may be worthier using tagged multiplexed libraries. Moreover, the rapid evolution of high throughput sequencing technologies and associated equipment is predicted to further facilitate the sequencing of many samples together.

Acknowledgements

We would like to thank Dr. Keith Harshman for precious technical support with NGS library preparation and sequencing, and Dr. Andrea Prunotto for helping in revising the manuscript. This research was supported in part by the Swiss National Science Foundation (Grants #320030-121929 and 310030_138346) and the Gebert Rüf Foundation (Rare Diseases - New Technologies grant).

References

Akhras, M. S., Unemo, M., Thiyagarajan, S., Nyren, P., Davis, R. W., Fire, A. Z. & Pourmand, N. (2007). Connector inversion probe technology: a powerful one-primer multiplex DNA amplification system for numerous scientific applications. *PLoS One*, 2(9), e915.

Baetens, M., Van Laer, L., De Leeneer, K., Hellemans, J., De Schrijver, J., Van De Voorde, H., Renard, M., Dietz, H., Lacro, R. V., Menten, B. and others (2011). Applying massive parallel sequencing to molecular diagnosis of Marfan and Loeys-Dietz syndromes. *Hum Mutat*

Benaglio, P., McGee, T. L., Capelli, L. P., Harper, S., Berson, E. L. & Rivolta, C. (2011). Next generation sequencing of pooled samples reveals new SNRNP200 mutations associated with retinitis pigmentosa. *Hum Mutat*, 32(6), E2246-58.

Benaglio, P. & Rivolta, C. (2010). Ultra high throughput sequencing in human DNA variation detection: a comparative study on the NDUFA3-PRPF31 region. *PLoS One*, 5(9),

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R. and others (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), 53-9.

Calvo, S. E., Tucker, E. J., Compton, A. G., Kirby, D. M., Crawford, G., Burtt, N. P., Rivas, M., Guiducci, C., Bruno, D. L., Goldberger, O. A. and others (2010). High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. *Nat Genet*, 42(10), 851-8.

Casbon, J. A., Osborne, R. J., Brenner, S. & Lichtenstein, C. P. (2011). A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res*, 39(12), e81.

Craig, D. W., Pearson, J. V., Szelinger, S., Sekar, A., Redman, M., Corneveaux, J. J., Pawlowski, T. L., Laub, T., Nunn, G., Stephan, D. A. and others (2008). Identification of genetic variants using bar-coded multiplexed sequencing. *Nat Methods*, 5(10), 887-93.

Dahl, F., Stenberg, J., Fredriksson, S., Welch, K., Zhang, M., Nilsson, M., Bicknell, D., Bodmer, W. F., Davis, R. W. & Ji, H. (2007). Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc Natl Acad Sci U S A*, 104(22), 9387-92.

Dryja, T. P. (1997). Gene-based approach to human gene-phenotype correlations. *Proc Natl Acad Sci U S A*, 94(22), 12117-21.

Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. (2011). Unlocking Mendelian disease using exome sequencing. *Genome Biol*, 12(9), 228.

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., Schork, N. J., Murray, S. S., Topol, E. J., Levy, S. and others (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol*, 10(3), R32.

Hartong, D. T., Berson, E. L. & Dryja, T. P. (2006). Retinitis pigmentosa. *Lancet*, 368(9549), 1795-809.

Hollants, S., Redeker, E. J. & Matthijs, G. (2012). Microfluidic amplification as a tool for massive parallel sequencing of the familial hypercholesterolemia genes. *Clin Chem*, 58(4), 717-24.

Hu, H., Wrogemann, K., Kalscheuer, V., Tzschach, A., Richard, H., Haas, S. A., Menzel, C., Bienek, M., Froyen, G., Raynaud, M. and others (2009). Mutation screening in 86 known X-linked mental retardation genes by droplet-based multiplex PCR and massive parallel sequencing. *Hugo J*, 3(1-4), 41-9.

Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. & Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol*, 8(7), R143.

Ikegawa, S., Mabuchi, A., Ogawa, M. & Ikeda, T. (2002). Allele-specific PCR amplification due to sequence identity between a PCR primer and an amplicon: is direct sequencing so reliable? *Hum Genet*, 110(6), 606-8.

Ingman, M. & Gyllensten, U. (2009). SNP frequency estimation using massively parallel sequencing of pooled DNA. *European Journal of Human Genetics*, 17(3), 383-386.

International Human Genaome Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), 931-45.

Lee, J. S., Choi, M., Yan, X., Lifton, R. P. & Zhao, H. (2011). On optimal pooling designs to identify rare variants through massive resequencing. *Genet Epidemiol*, 35(3), 139-47.

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J. & Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nat Methods*, 7(2), 111-8.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z. and others (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), 376-80.

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C. and others (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res*, 19(9), 1527-41.

Mertes, F., Elsharawy, A., Sauer, S., van Helvoort, J. M., van der Zaag, P. J., Franke, A., Nilsson, M., Lehrach, H. & Brookes, A. J. (2011). Targeted enrichment of genomic DNA regions for next-generation sequencing. *Brief Funct Genomics*, 10(6), 374-86.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1), 31-46.

Meyer, M. & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc*, 2010(6), pdb prot5448.

Neveling, K., Collin, R. W., Gilissen, C., van Huet, R. A., Visser, L., Kwint, M. P., Gijsen, S. J., Zonneveld, M. N., Wieskamp, N., de Ligt, J. and others (2011). Next-generation genetic testing for retinitis pigmentosa. *Hum Mutat*, 33(6), 963-72.

Otto, E. A., Ramaswami, G., Janssen, S., Chaki, M., Allen, S. J., Zhou, W., Airik, R., Hurd, T. W., Ghosh, A. K., Wolf, M. T. and others (2011). Mutation analysis of 18 nephronophthisis associated ciliopathy disease genes using a DNA pooling and next generation sequencing strategy. *J Med Genet*, 48(2), 105-16.

Out, A. A., van Minderhout, I. J., Goeman, J. J., Ariyurek, Y., Ossowski, S., Schneeberger, K., Weigel, D., van Galen, M., Taschner, P. E., Tops, C. M. and others (2009). Deep sequencing to reveal new variants in pooled DNA samples. *Hum Mutat*, 30(12), 1703-12.

Ramensky, V., Bork, P. & Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*, 30(17), 3894-900.

Reese, M. G., Eeckman, F. H., Kulp, D. & Haussler, D. (1997). Improved splice site detection in Genie. *J Comput Biol*, 4(3), 311-23.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M. and others (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356), 348-52.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12), 5463-7.

Shen, P., Wang, W., Krishnakumar, S., Palm, C., Chi, A. K., Enns, G. M., Davis, R. W., Speed, T. P., Mindrinos, M. N. & Scharfe, C. (2011). High-quality DNA sequence capture of 524 disease candidate genes. *Proc Natl Acad Sci U S A*, 108(16), 6549-54.

Simpson, D. A., Clark, G. R., Alexander, S., Silvestri, G. & Willoughby, C. E. (2010). Molecular diagnosis for heterogeneous genetic diseases with targeted high-throughput DNA sequencing applied to retinitis pigmentosa. *J Med Genet*, 48(3), 145-51.

Smith, A. M., Heisler, L. E., St Onge, R. P., Farias-Hesson, E., Wallace, I. M., Bodeau, J., Harris, A. N., Perry, K. M., Giaever, G., Pourmand, N. and others (2010). Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Res*, 38(13), e142.

Tewhey, R., Warner, J. B., Nakano, M., Libby, B., Medkova, M., David, P. H., Kotsopoulos, S. K., Samuels, M. L., Hutchison, J. B., Larson, J. W. and others (2009). Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat Biotechnol*, 27(11), 1025-31.

Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. (2009). Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat Methods*, 6(5), 315-6.

Project 4: **Screening of candidate splicing factors for mutations in adRP patients**

The results of the screening of *SNRNP200* confirmed the association of the members of the tri-snRNP complex with autosomal dominat RP, and that candidate gene screening is still an effective tool for gene hunting in RP, even in the exome/genome sequencing era. This fact encouraged us to perform a second screening of other splicing factor genes (*EFTUD2*, *PRPF4*, *NHP2L1* and *AAR2)* as novel candidate for adRP. The genes have been chosen based on their functional interaction with known genes of the tri-snRNP complex as documented in the literature. Since mutations in novel RP genes are expected to be very rare, a high number of patients with undiagnosed molecular cause is needed to add power to the screening. We expanded our adRP cohort used for the sequencing of *SNRNP200* with additional samples shared by our collaborators from France (Christian Hamel, M.D.) and Spain (Carmen Ayuso, M.D.). We reached a total of 200 samples for the discovery-screening phase and about 100 more for downstream validations. Following the methodological conclusions drawn in our previous publications, we adopted a similar NGS-based strategy using a pool of long-range PCRs as template, with the difference of using a high throughput commercial system for preparing barcoded libraries, by means of transposase reactions. This was used for the large genes, while for the genes composed of few exons we used capillary electrophoresis (Sanger) sequencing. The results did not reveal any plausible mutation linked to retinitis pigmentosa, although few confirmatory elements are missing. After their completion, the manuscript will be submitted to Molecular Vision or Ophthalmic Genetics, with minor modifications with respect to the following document.

Candidate's roles:

- Design of the project.
- Preparation of samples for sequencing: PCR enrichment and libraries preparation.
- Planning and execution of the sequence analyses.
- Downstream validation analyses by Sanger sequencing and other bioinformatic analyses.
- Writing of the manuscript.

Mutational screening of splicing factor genes as candidate cause for autosomal dominant retinitis pigmentosa

Paola Benaglio[1], Shyana Harper[2], Carmen Ayuso[3], Christian Hamel[4], Eliot L. Berson[2], Carlo Rivolta[1]

[1]Department of Medical Genetics, University of Lausanne, Lausanne; [2]The Berman-Gund Laboratory for the Study of Retinal Degenerations, Harvard Medical School, Massachusetts Eye and Ear Infirmary, Boston; [3]Servicio de Genética, IIS Fundación Jiménez Díaz, Madrid; [4]Service d'Ophtalmologie, Université et Hôpital de Montpellier, Montpellier.

ABSTRACT

Retinitis pigmentosa (RP) is a rare disease with heterogeneous genetic basis, characterized by the progressive degeneration of photoreceptors. Eleven percent of autosomal dominant RP (adRP) forms are caused by mutations in genes belonging to the tri-snRNP complex of the spliceosome. Although the exact mechanism by which splicing factor defects trigger specific photoreceptor death is not clear, their role in retinitis pigmentosa has been demonstrated by several genetic and biochemical studies. We sequenced four tri-snRNP splicing factor genes (*EFTUD2*, *PRPF4*, *NHP2L1* and *AAR2)* previously not associated with RP in 200 adRP patients of European origin, to test their possible role in the disease. We took advantage of both classic Sanger sequencing and next generation sequencing methods to obtain accurate and cost-effective sequence data. The result of the screening revealed a few novel missense changes, whose role in the disease was not possible to verify. This suggests that the connection between splicing factors and RP is far from being obvious, and that recent advances in genome-wide sequencing might offer more unbiased approaches than candidate screening procedures to discover the remaining fraction of genes causing RP.

INTRODUCTION

The most common form of hereditary blindness is retinitis pigmentosa (RP), affecting 1 in 4000 people worldwide. The disease typically begins with night blindness, due to early involvement of rod photoreceptors, and progresses with the reduction of visual field and eventual loss of central vision, as a result of later degeneration of cone photoreceptors [1]. Patients affected with RP display clinical heterogeneity with respect to age of onset, degree of severity, rate of progression and other secondary manifestations. These differences are partly explained by the different genes and mutations causing RP: to date, 56 genes have been

associated with non-syndromic RP, with about 3000 different mutations reported in total [2]. The inheritance mode is classically monogenic: dominant (~30-40%), recessive (~50-60%), x linked (5-15%), and a smaller fraction of non –Mendelian or complex inheritance [1]. Nevertheless, this genetic heterogeneity does not always determine differences in clinical manifestations, making it hard to assess clear genotype-phenotype correlation.

The functions of RP genes can be very diverse: some genes are specific for retinal functions such as phototransduction and retinal metabolism; some others have a more general function in cell development and maintenance [3]. A particular category that exemplifies the complexity of the molecular genetics of RP is represented by a number of very conserved and ubiquitously expressed pre-mRNA splicing factors. Splicing consists of enzymatic reactions leading to the removal of introns from pre-mRNA to form mature mRNA. A macromolecular complex, referred to as the spliceosome, ensures the fidelity and the correct timing or these reactions. The spliceosome is composed of pre-mRNA, five small nuclear ribonucleoproteins (snRNP), U1, U2, U4/U6 and U5, and ~200 other proteins [4]. Spliceosome assembly is a dynamic, stepwise process: U1 snRNP first recognizes the 5' splice site and U2 binds to the branch point; then U4/U6.U5 tri-snRNP complex is recruited and finally U1 and U4 are released, leading to catalytic activation [5].

To date, six splicing factors genes have been found to be mutated in adRP patients: *PRPF8* (RP13) [6], *PRPF31* (RP11) [7], *PRPF3* (RP18)[8], *PAP-1* (RP9) [9], *SNRNP200* (RP33) [10, 11] and *PRPF6* [12]. These discoveries have been achieved through linkage analysis and positional cloning for the first two genes discovered (*PRPF8* and *PRPF31*), followed by screening of other splicing factor genes in linkage intervals. All these genes have in common the high level of protein sequence conservation up to yeast, and their belonging to the U4/U6.U5 tri-snRNP complex. The growing evidences of a major role of these particular splicing factors suggested that also other partners of the complex could be meaningful candidate genes for adRP, which still includes about 50% of cases with unsolved genetic cause. The role of the *PRPF6* gene in adRP was found by such an approach, after screening its coding sequence in a cohort of 200 American patients [12].

Following this rationale, we screened four candidate genes (*EFTUD2*, *PRPF4*, *NHP2L1* and *AAR2)* in 200 adRP patients with unknown molecular diagnosis and previously analyzed for mutations in the most common adRP genes or hotspots. These genes were selected based on their physical and/or functional interaction with known RP-linked splicing factors. *EFTUD2* encodes for an essential GTPase hSnu114, homolog of *S. cerevisiae* Snu114p, which forms a

stable complex with the *SNRNP200* and *PRPF8* products (i.e. hBrr2 and PRPF8) [13]. hSnu114 regulates hBrr2 at the dissociation step of U4 from U6 snRNAs and is also necessary for spliceosome disassembly after splicing [14]. Aar2p (encoded by the *AAR2* gene) competes with hBrr2 in the binding of the C-terminal region of PRPF8 before the maturation of the U5 snRNP, supposedly regulating its assembly [15]. The 15.5-kDa protein (Snu13p in yeast), encoded by the *NHP2L1* gene, binds to the 5′-stem-loop of U4 snRNA probably playing a role in the late phase of the spliceosome assembly [16]. Finally, PRPF4 protein forms a complex with PRPF3 in the U4/U6 snRNP complex and its downregulation was found to induce photoreceptor defects in a Zebrafish model, similarly to PRPF31 [17, 18].

For the genetic screening we took advantage of a method that combines classical exon-PCR for the small genes, and long-range PCR followed by next generation sequencing for the large ones (**Table 1**). The latter approach provides a cost- and time- effective alternative to the Sanger method and well adapts to routine genetic screenings in large sets of samples.

METHODS

Samples and patients

The subjects analyzed in this cohort belong to three different collections of unrelated patients affected with autosomal dominant retinitis pigmentosa. One hundred ninety-one samples were collected and followed in the Massachusetts Eye and Ear Infirmary – Harvard Medical School; they are mostly American of European origin. Fourty-seven were collected in Spain (Servicio de Genética, IIS Fundación Jiménez Díaz, Madrid) and 96 were from France (Service d'Ophtalmologie, Université et Hôpital de Montpellier). DNA was extracted from peripheral leukocytes and quantified. For technical reasons, not all these samples could undergo the complete screening, but all of them were used in case of validation of putative mutations. Control DNA samples were obtained from 95 individuals with no history of retinal degeneration and 96 unrelated healthy individuals aged between 34 and 92, purchased from the Coriell institute for medical research. All subjects provided written, informed consent and the study was conducted in adherence with the Declaration of Helsinki.

Library preparation and next generation sequencing (NGS)

Genes *EFTUD2* and *PRPF4* were sequenced by a long-range PCR (LR-PCRs) enrichment and next generation sequencing approach, using Illumina instruments (San Diego, CA). Five

and two LR-PCRs were generated to amplify the entire 51 and 20 kb regions of each gene, respectively, for a total of 71 kb targeted region. LR-PCRs were obtained individually for each sample using TaKaRa LA Taq polymerase (Takara Bio, Shiga, Japan) with GC buffer and 1 μM of the primers reported in **Table S1**. The following cycling conditions were used: 94 °C for 1' followed by 30 cycles at 98 °C for 5`` and 68 °C for 15', and final extension of 72 °C for 10'. For each sample, the 7 LR-PCRs were pooled into a single tube, after estimation of their quantity by agarose 1% gel visualization. They were subsequently purified using DNA Clean and Concentrator columns (Zymo Research, Orange, CA). Only the DNA samples from the three cohorts that yielded 7 clear PCR bands underwent NGS.

Library preparation and sample barcoding were performed with the *tagmentation* method [19] using the Nextera DNA Sample Prep Kit (Epicentre, Madison, WI) and 48 barcodes adapted to Illumina platforms [20], following manufacturer instructions. Fourteen tagged samples were sequenced as a pool in one lane of the GAII instrument for testing purpose, after which two runs of Hiseq instrument (one lane) were used to sequence two different pools of 48 and 47 barcoded samples each. After the integration of the Nextera products by Illumina Company, we processed 91 additional samples using the Nextera XT DNA Sample Preparation (Illumina) protocol, reagents and barcodes, and sequenced them as a unique pool by one Miseq instrument run.

Analysis and variant calling of sequences from NGS

We mapped the reads obtained from NGS to the reference sequence of the genes (GRCh37.p10 assembly) by the CLC genomics Workbench package, v. 5.5 (CLC bio, Aarhus, Denmark). The parameters were in a way that a read could align only if it had at least a 90% identity for the 90% of its length. A more relaxed setting (80% identity over 70% of its length) was also tried. Single nucleotide variant calling and small insertion and deletion calling was achieved by imposing a minimum frequency of discordant bases of 20%, with minimum coverage of 5 nucleotides and an average base quality of 20 Phred. The analyses were performed as a batch for all individual samples (200) and the obtained variants were annotated with the hg19_snp137 track from the UCSC genome browser.

Functional predictions of variant identified

Missense changes were analyze by the online package PON-P, which integrates results of the most common prediction software including PolyPhen and SIFT [21]. The effect of intronic

changes was evaluated by the Shannon Human Splicing Pipeline implemented in the CLC software [22] and the NNSPLICE 0.9 algorithm [23].

Sanger sequencing and restriction analysis

The genes *NHP2L1* and *AAR2* were screened by Sanger sequencing of the coding exons. PCR reactions were obtained by the GoTaq polymerase (Promega, Madison, WI) standard protocol and 0.25 μM of the primers reported in **Table S2**. Reactions were purified from excess primers and nucleotides by ExoSAP-IT (Affymetrics, Santa Clara, CA) and subjected to sequencing reactions using Big Dye V1.1 Terminator Kit (Applied Biosystems, Foster City, CA) and an ABI automated DNA sequencer (Applied Biosystems). Sequences were analyzed using the CLC Genomics Workbench. The same procedure was applied for the validation of novel changes in specific exons identified by NGS, cosegregation analysis and screening of controls and additional patients. Primers used for these purposes are also listed in **Table S2**. Controls and patients were tested using restriction enzymes when a particular nucleotide change abolished or created a restriction site. In particular exon 5 of *PRPF4* was tested with *Msc*I, exon 8 of *EFTUD2* with *Hah*I and exon 1 of *NHP2L1* with *Mlu*I (New England Biolabs, Ipswich, MA).

RESULTS

NGS screening

A total of 200 unrelated individuals diagnosed with adRP were screened for the genes *EFTUD2* and *PRPF4*, in search for candidate mutations. Seventy-nine patients were from USA, 71 from France and 50 from Spain. The approach used to sequence these genes consisted of an enrichment step via LR-PCR, followed by multiplexed runs of NGS instruments. The analysis of the so-obtained reads involved firstly the alignment of such sequences to the reference genomic sequences of the targeted genes. Since different instruments were used, different samples had different coverage depths. Specifically, the samples sequenced using HiSeq had higher coverage with respect to the ones sequenced using MiSeq, due to the lower throughput and higher number of samples sequenced with the latter (**Table S3**). With the exception of a few samples, the targeted region was optimally covered for reliable variant calling. This consisted of the detection of both single nucleotide variations and small insertions and deletions by the CLC Genomics algorithm. After merging the results for each sample, we obtained a total figure of 1,195 variants identified, 591 of which are

known polymorphisms present at different frequencies in the analyzed cohort. By restricting the analysis to missense changes, we identified in total 6 variants, of which only 3 were novel (**Table 2**).

Three different patients carried three different variants in *PRPF4*, of which 2 had an *rs* number and one was a false positive due to low coverage, as ascertained by Sanger sequencing. Although the DNA residue c.559C (NM_004697.3 – exon 5) is tagged by the rs187531407 number, the ccc>tcc (p.Pro187Ser) change was found only in two not validated 1000Genomes reports. We could ascertain that one of these reports (low-coverage 1000Genomes) was a false positive, after having retrieved and sequenced the DNA sample that was identified to carry this variant. Moreover the base change in the patient of our cohort is different: it is a ccc>gcc change, which is translated into p.Pro187Ala. The residue is fairly conserved across species and, according to predictions with different tools, the likelihood of pathogenicity is uncertain (**Table 3**). We followed up this change by analyzing controls and available relatives. Public databases, as well as sequencing of in-house controls, did not reveal the presence of this change. Neither it was found in the remaining, unscreened patients from our cohorts. The affected sibling (226-1953) also carried the same change, but other members were not available to check further the segregation. Since the base change is located 5 nucleotides away from the exon-exon junction, we also checked if splicing of the exon could be affected. Bioinformatic prediction was negative and RT-PCR of patient's cDNA (data not shown) did not reveal missplicing events. The lack of further information, the possible presence of polymorphism in the same codon and a not striking prediction of pathogenicity prevented us to make significant conclusions.

Of the four missense changes found in the *EFTUD2* gene, two were novel and present in single individuals: p.Arg220Cys and p.Ile80Leu (**Table 2**). The first one was confirmed by Sanger sequencing and predicted to be pathogenic by a number of prediction tools (**Table 3**). The residue is in fact much conserved, from human to yeast. The change was absent in public variation databases (dbSNP, 1000Genomes, exome variant server, 42 unlerated control individuals from Complete Genomics). Moreover it was absent in 150 controls tested in-house and in the remaining patients from the cohort. Although we verified that patient's healthy sister and son did not carry this change, the unavailability of other family members prevented us for further investigation of the variant. The second change I80L was predicted to be a neutral change, based on conservation and strength of change and was not investigated further.

Since LR-PCR amplified both coding and non-coding regions of the target genes, we tested if any novel variant identified could affect sequences important for splicing signals also located in deep intronic regions. We analyzed 1,008 variants in both exons and introns of the *EFTUD2* and *PRPF4* genes with the Shannon Human Splicing Pipeline [22]. Only single nucleotide substitutions, but not insertions or deletion could be tested with this method. No change inactivated or reduced the strength of the natural splice sites. Four hundred thirty-four variants were predicted to modify in some way the sequence content information (defined with *R*) of cryptic splice sites. By filtering for variations that were not polymorphic and that resulted in the creation of a donor or acceptor splice site with greater strength than the natural splice site, only 5 variants were left (**Table S4**). Two of them were likely false positives because they were present only in one read out of 5-fold coverage. For the remaining three, other predictions were made using the NNSPLICE algorithm and did not agree with the one of the Shannon pipeline. In two cases the new cryptic splice sites were still weaker than the natural ones and in one case the already existing cryptic site was weakened by the change and not strengthen.

Finally a second run of variant calling was performed on alignments obtained with less stringent criteria, in order to exclude the possibility of false negatives, due to too rigid mapping parameters. Such an analysis increased by two times the number of known SNPs identified (999 vs. 476) and by three times the number of non-reported changes (3752 vs. 1195), indicating a gain in sensitivity, but also a loss in specificity (**Table S5**). In fact, when analyzing coding changes only, in addition to the variants found with the previous mappings, we obtained 8 false positives, all found in the same sample and localized in a stretch of wrongly aligned reads, as it was clear from inspection of the mapping.

Sanger sequencing screening

We selected two additional genes of the tri-snRNP complex to be screened by Sanger sequencing, in virtue of their small size. The gene *NHP2L1* consists of two coding exons and two alternative 5' UTR containing the start codon. The sequencing of the four exons in 320 patients did not reveal anything significant but a novel change introducing an ATG start codon in the 5' UTR, that was not found in the control population. This change was interesting because it creates an upstream open reading frame (uORF), whose effect would be to reduce the translation from the downstream, canonical ATG. However, this change did not

segregate with RP in the family. The sequence of AA2R gene was negative for novel variations.

DISCUSSION

The adRP-linked splicing proteins PRPF31, PRPF3, PRPF8, PRPF6, and hBrr2 are all components of the U4/U6.U5 tri-snRNP, suggesting that there is a common mechanism of pathogenesis in RP related to dysfunction of this complex. It has been shown that mutations in these proteins impair the assembly of tri-snRNP complex [24] [25] and/or affect catalytic activation of the spliceosome [26], leading to pre-mRNA splicing defects and eventually to cell death [27-29]. Mutations are thought to act through a haploinsufficiency mechanism because many of them either determine truncation and degradation of the protein and the transcript [30] or their instability and accumulation in Cajal bodies [27-29].

We wanted to investigate the hypothesis of the implication of other U4/U6.U5 tri-snRNP proteins in adRP by screening their DNA sequences in well-characterized cohorts of dominant patients who do not carry mutation in the most prevalent RP genes. We used an NGS-based approach that allowed a fast and parallel analysis of few candidate genes in a large set of patients, enabling in principle the identification of very rare mutations. The sequencing of the coding exons of the genes *NHP2L1* and *AAR2,* and of exons and introns of *EFTUD2* and *PRPF4* revealed a few variants that could have an effect at the protein level and that were absent from the general population. Only the p.Arg220Cys missense in the *EFTUD2* gene was predicted to be damaging but its putative pathogenicity could not be demonstrated. Moreover, during the course of this screening, the same gene has been linked by exome sequencing to another class of rare and sporadic congenital malformation syndromes, in particular to Mandibulofacial dysostosis with microcephaly (MIM 610536) [31, 32]. In these patients, mutations were de novo heterozygous missense, frameshift and null alleles. Although certain phenotypic variability was observed for *EFTUD2* mutations [33], it seems that they impact early developmental stages and lead to much more dramatic phenotypes than RP. However, it cannot be excluded that other mutations may have milder effects and trigger the same photoreceptor cell death pathway as for the RP-linked splicing factors. The negative results for this and the other genes screened in this study, and the evidence that mutations in components of the tri-snRNP have role in other diseases indicate that their link to adRP is not straightforward. The implication of splicing factors in RP likely involves very subtle mechanisms in which RP mutations determine gradual accumulation of mild splicing defects,

which are less tolerated by photoreceptor cells, because of their higher demand of splicing [29]. It is therefore rather difficult to predict which ones of such proteins would be implied in RP and a candidate gene screening strategy may result inefficient to identifications of new adRP genes. Otherwise, gene-based approaches have proven to be very effective in the discovery of novel genes involved in RP [12, 34], especially in the early phases of RP genetic research. Conversely, in recent years the rate of discovery of novel genes has slowed down with respect to the previous 20 years, because unknown RP genes are increasingly rare and fewer families suitable for linkage analysis are available [2]. With the introduction of next generation sequencing it is no longer an issue to obtain genome-wide information of patients affected with Mendelian diseases and identification of the molecular cause can be achieved by an unbiased prioritization of mutations and/or by gene-driven hypothesis (i.e. by looking only at a large set of known or candidate genes). Powerful techniques such as exome sequencing have proven to be extremely useful in identification on novel mutations and genes, also in RP. However discoveries via exome sequencing have been obtained when analyzing recessive conditions or dominant diseases with no genetic heterogeneity. For a dominant disease with elevated genetic heterogeneity, the computational analysis and validation elements necessary to find significant association with heterozygous changes become more important [35]. Therefore, in autosomal dominant RP, the screening of many samples for candidate genes can still be a feasible option, provided that 1) strong candidate (possible deriving from animal or cellular model) are tested and 2) cohorts highly enriched in novel genes are screened. For these applications as well, next generation sequencing can offer useful strategies to perform targeted re-sequencing in a fast and comprehensive manner, such as the one adopted for this study.

TABLES

Symbol	Protein	snRNP complex	# exons	Screening method	# screened patients
EFTUD2	Elongation factor Tu GTP binding domain containing 2- 116 kDa	U5	28	Nextera-NGS	200
PRPF4	PRP4 pre-mRNA processing factor 4 homolog – 60kDa	U4/U6	14	Nextera-NGS	200
NHP2L1	NHP2 non-histone chromosome protein 2-like 1- 15.5 kDa	U4/U6.U5	4	Sanger	320
AAR2	AAR2 splicing factor homolog	U5	4	Sanger	100

Table 1. Genes analyzed in this study and methods.

Gene	Sample count	Genomic Position	Coding region change	Amino acid change	DATABASE*	Allele counts (M/m)
EFTUD2	1	17:42953357	NM_004247.3:c.814A>G	NP_004238.3:p.Thr272Ala	rs150633454	541/497
EFTUD2	1	17:42956968	NM_004247.3:c.658C>T	NP_004238.3:p.Arg220Cys	-	779/649
EFTUD2	1	17:42963986	NM_004247.3:c.238A>C	NP_004238.3:p.Ile80Leu	-	292/292
PRPF4	3	9:116049532	NM_004697.4:c.233A>G	NP_004688.2:p.His78Arg	rs1138958	-
PRPF4	*1*	*9:116053294*	*NM_004697.4:c.481G>A*	*NP_004688.2:p.Glu161Lys*	-	*4/1 (False +)*
PRPF4	1	9:116053770	NM_004697.4:c.559C>G	NP_004688.2:p.Pro187Ala	rs187531407	1139/1363

Table 2. Variant output from NGS screening after filtering for aminoacidic changes. Genomic coordinates refer to assembly GRCh37.p10. Numbering of coding region starts at A of the ATG. *Human variation database search included dbSNP 137, 1000 Genome Project, Exome variant server, Complete Genomics control samples (52) and internal database of 500 exomes (CoLaus).

Gene	Origin	Change	Polyphen prediction based on alignment	PON-P prediction	Controls	Segregation
PRPF4	German	p.Pro187Ala	Probably damaging	Unclassified, Probability of pathogenicity: 0.27	0/189	Not conclusive
EFTUD2	American Indian-French Canadian/Irish	p.Arg220Cys	Probably damaging	Pathogenic, Probability of pathogenicity: 0.91	0/150	Not conclusive
EFTUD2	-	p.Ile80Leu	Benign	Neutral, Probability of pathogenicity: 0.02	NA	Not available
NHP2L1	Italian	chr22:42078408 NM_005008.2:c.-46G>A Creation of upstream out of frame ORF			0/150 (and absent from database too)	Negative

Table 3. Characterization of changes identified in the complete screening. Prediction of pathogenicity was made using public software of Polyphen [36] and PON-P [21]. Additional controls were performed.

SUPPLEMENTAL TABLES

PRIMER NAME	*SEQUENCE*
EFTUD2_LR1_for	CTCAGCCCTCCCCAGGCATTTAATACATAG
EFTUD2_LR1_rev	GCCAACCTCTTCCTTAAGAACACAAAACCC
EFTUD2_LR2_for	ATGGGGACAGGAAAGAAAGATGTCCCT
EFTUD2_LR2_rev	TCAAGGGGAAAATGTACACACCTGTCTCT
EFTUD2_LR3_for	CAATTGAGAGATGGAACACTTTGGCTAACCTT
EFTUD2_LR3_rev	TCTATTAGTAATTTCTGTGACCATGGGCATAG
EFTUD2_LR4_for	TACATGAGGGGAAACTATGCCCATGGTC
EFTUD2_LR4_rev	GACAAAGAATCTGGGAGAGACAGTCCCC
EFTUD2_LR5_for	ATTGTGCCATTGCTCTCCTTTTTGAGATGG
EFTUD2_LR5_rev	ATTAAGTTCTCCTGTTCTGGGCTCCACATC
PRPF4_LR6_for	TCCCTCATCTGTATCCTACTCTGCTGTTGT
PRPF4_LR6_rev	GTAATGAGGCTTGGAGAGGTGGACTCATTT
PRPF4_LR7_for	ACACAAAGCACTCTGTCACATTCTTGACAC
PRPF4_LR7_rev	CTTGAGCAAGTGTTTTTGGAGCCTGTTTTG

Table S1. Primers used for long-range PCR

PRIMER NAME	SEQUENCE
NHP2L1_Exon1-isoform1_for	CATGTGAAGGAGACATACTC
NHP2L1_Exon1-isoform1_rev	GTGGGACCACTCATCTATAG
NHP2L1_Exon1-isoform2_for	TCAGCTGGAGTGCTAGAGTG
NHP2L1_Exon1-isoform2_rev	AAACAGACCGTGCGGCAAAG
NHP2L1_Exon2_for	GCCATGGACTCAAATGGATG
NHP2L1_Exon2_rev	AGCCCATCAGCTTTAGCATC
NHP2L1_Exon3_for	AAGTGGTCAGATTCAGGACG
NHP2L1_Exon3_rev	AAGGATGAAGGATGGCAGAG
AAR2_Exon1_for	GCTGTGAAGTGAGTGTCCTTGCAT
AAR2_Exon1_rev	AACTGCGGAAGCCCCACTCC
AAR2_Exon2_for	TGCCAGGCTCTGCAGGAGTGA
AAR2_Exon2_rev	AGGGAGAGGGTGGTACAGAGAGA
AAR2_Exon3_for	AGTCCTGCCCTAGCATCTCA
AAR2_Exon3_rev	ATGGGTACAGCCTCTTTGGC
EFTUD2_exon8_for	TCAAGTTCTCTGGCTCCCAG
EFTUD2_exon8_rev	GCGAGGAGGAAAGGGGATAT
PRPF4_exon5-6_for	GAATTCACCTTCCTTCTGGG
PRPF4_exon5-6_rev	CAGAGGTGAGCCAAAGTAG

Table S2. Primers used for short range PCR and sequencing

Instrument	Multiplexed samples	Read Length	Number of reads	Average per sample reads	Average % mapping	Average per sample coverage
Illumina GA II 1 lane	14	77	1,304,633	1,008,000	0.77	2,383
Hiseq 2000 1 lane	48	51	143,417,293	2,987,860	0.85	1,876
Hiseq 2000 1 lane	47	101	148,691,083	3,163,640	0.84	3,826
Miseq full	91	200-250 (paired)	31,690,692	348,249	0.90	962
Total samples	200					

Table S3. Summary of metrics of NGS screening runs and per sample statistics.

Gene Name	*EFTUD2*	*EFTUD2*	*EFTUD2*	*EFTUD2*	*PRPF4*
Chromosome	17	17	17	17	9
Splice site coordinate	42939388	42954217	42963479	42938786	116041841
Input variant	T/A	C/T	G/C	C/G	T/G
Ri-initial	-13.32	-2.34	2.91	10.68	-2.09
Ri-final	5.3	12.37	6.78	13.05	1.28
ΔRi	18.62	14.71	3.87	2.37	3.38
Type	DONOR	ACCEPTOR	DONOR	ACCEPTOR	DONOR
Loc. of nearest nat. site	42940080	42953469	42963952	42937912	116041412
Ri of nearest nat. site	3.21	10.16	4.74	3.3	-1.45
Cryptic Ri relative to nat.	GREATER	GREATER	GREATER	GREATER	GREATER
Carriers	1	1	1	1	1
Variant coverage		1 out of 4			1 out of 4
Observations	not concordant with NNSPLICE	false positive	not concordant with NNSPLICE	not concordant with NNSPLICE	false positive

Table S4. Results of splice sites analysis using the Shannon Human Splicing Pipeline. Ri is a measure of the likelihood of a sequence to be used as acceptor or donor splice site. ΔRi is the difference between the R value of a genomic coordinate in presence and in absence of the nucleotide variation. Nat.site is the closest natural splice site.

	Used parameters	Low stringency parameters
Total variants	1195	3752
Total variants <1% frequency	591	1928
Tot SNPs	476	999
Total nonsynonimous changes	6	14
Tot nonsynonimous, non SNP change	3	11
False positives	1	8

Table S5. Total number of variants identified by two alignments with different mapping criteria.

REFERENCES

1. Hartong, D.T., E.L. Berson, and T.P. Dryja, *Retinitis pigmentosa.* Lancet, 2006. **368**(9549): p. 1795-809.
2. Daiger, S.P., L.S. Sullivan, and S.J. Bowne, *Genes and mutations causing retinitis pigmentosa.* Clin Genet, 2013. **84**(2): p. 132-41.
3. Berger, W., B. Kloeckener-Gruissem, and J. Neidhardt, *The molecular basis of human retinal and vitreoretinal diseases.* Prog Retin Eye Res, 2010. **29**(5): p. 335-75.
4. Jurica, M.S. and M.J. Moore, *Pre-mRNA splicing: awash in a sea of proteins.* Mol Cell, 2003. **12**(1): p. 5-14.
5. Staley, J.P. and C. Guthrie, *Mechanical devices of the spliceosome: motors, clocks, springs, and things.* Cell, 1998. **92**(3): p. 315-26.
6. McKie, A.B., et al., *Mutations in the pre-mRNA splicing factor gene PRPC8 in autosomal dominant retinitis pigmentosa (RP13).* Hum Mol Genet, 2001. **10**(15): p. 1555-62.
7. Vithana, E.N., et al., *A human homolog of yeast pre-mRNA splicing gene, PRP31, underlies autosomal dominant retinitis pigmentosa on chromosome 19q13.4 (RP11).* Mol Cell, 2001. **8**(2): p. 375-81.
8. Chakarova, C.F., et al., *Mutations in HPRP3, a third member of pre-mRNA splicing factor genes, implicated in autosomal dominant retinitis pigmentosa.* Hum Mol Genet, 2002. **11**(1): p. 87-92.
9. Keen, T.J., et al., *Mutations in a protein target of the Pim-1 kinase associated with the RP9 form of autosomal dominant retinitis pigmentosa.* Eur J Hum Genet, 2002. **10**(4): p. 245-9.
10. Zhao, C., et al., *Autosomal-dominant retinitis pigmentosa caused by a mutation in SNRNP200, a gene required for unwinding of U4/U6 snRNAs.* Am J Hum Genet, 2009. **85**(5): p. 617-27.
11. Li, N., et al., *Mutations in ASCC3L1 on 2q11.2 are associated with autosomal dominant retinitis pigmentosa in a Chinese family.* Invest Ophthalmol Vis Sci, 2010. **51**(2): p. 1036-43.
12. Tanackovic, G., et al., *A missense mutation in PRPF6 causes impairment of pre-mRNA splicing and autosomal-dominant retinitis pigmentosa.* Am J Hum Genet, 2011. **88**(5): p. 643-9.
13. Achsel, T., et al., *The human U5-220kD protein (hPrp8) forms a stable RNA-free complex with several U5-specific proteins, including an RNA unwindase, a homologue of ribosomal elongation factor EF-2, and a novel WD-40 protein.* Mol Cell Biol, 1998. **18**(11): p. 6756-66.
14. Liu, S., et al., *The network of protein-protein interactions within the human U4/U6.U5 tri-snRNP.* RNA, 2006. **12**(7): p. 1418-30.
15. Weber, G., et al., *Mechanism for Aar2p function as a U5 snRNP assembly factor.* Genes Dev, 2011. **25**(15): p. 1601-12.
16. Nottrott, S., et al., *Functional interaction of a novel 15.5kD [U4/U6.U5] tri-snRNP protein with the 5′ stem-loop of U4 snRNA.* EMBO J, 1999. **18**(21): p. 6119-33.
17. Lauber, J., et al., *The human U4/U6 snRNP contains 60 and 90kD proteins that are structurally homologous to the yeast splicing factors Prp4p and Prp3p.* RNA, 1997. **3**(8): p. 926-41.
18. Linder, B., et al., *Systemic splicing factor deficiency causes tissue-specific defects: a zebrafish model for retinitis pigmentosa.* Hum Mol Genet, 2011. **20**(2): p. 368-77.
19. Adey, A., et al., *Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition.* Genome Biol, 2010. **11**(12): p. R119.
20. Meyer, M. and M. Kircher, *Illumina sequencing library preparation for highly multiplexed target capture and sequencing.* Cold Spring Harb Protoc, 2010. **2010**(6): p. pdb prot5448.
21. Olatubosun, A., et al., *PON-P: integrated predictor for pathogenicity of missense variants.* Hum Mutat, 2012. **33**(8): p. 1166-74.
22. Shirley, B.C., et al., *Interpretation, stratification and evidence for sequence variants affecting mRNA splicing in complete human genome sequences.* Genomics Proteomics Bioinformatics, 2013. **11**(2): p. 77-85.
23. Reese, M.G., et al., *Improved splice site detection in Genie.* J Comput Biol, 1997. **4**(3): p. 311-23.
24. Makarova, O.V., et al., *Protein 61K, encoded by a gene (PRPF31) linked to autosomal dominant retinitis pigmentosa, is required for U4/U6*U5 tri-snRNP formation and pre-mRNA splicing.* EMBO J, 2002. **21**(5): p. 1148-57.
25. Huranova, M., et al., *A mutation linked to retinitis pigmentosa in HPRP31 causes protein instability and impairs its interactions with spliceosomal snRNPs.* Hum Mol Genet, 2009. **18**(11): p. 2014-23.

26. Mozaffari-Jovin, S., et al., *Inhibition of RNA Helicase Brr2 by the C-Terminal Tail of the Spliceosomal Protein Prp8.* Science, 2013.
27. Schaffert, N., et al., *RNAi knockdown of hPrp31 leads to an accumulation of U4/U6 di-snRNPs in Cajal bodies.* EMBO J, 2004. **23**(15): p. 3000-9.
28. Comitato, A., et al., *Mutations in splicing factor PRPF3, causing retinal degeneration, form detrimental aggregates in photoreceptor cells.* Hum Mol Genet, 2007. **16**(14): p. 1699-707.
29. Tanackovic, G., et al., *PRPF mutations are associated with generalized defects in spliceosome formation and pre-mRNA splicing in patients with retinitis pigmentosa.* Hum Mol Genet, 2011. **20**(11): p. 2116-30.
30. Rio Frio, T., et al., *Premature termination codons in PRPF31 cause retinitis pigmentosa via haploinsufficiency due to nonsense-mediated mRNA decay.* J Clin Invest, 2008. **118**(4): p. 1519-31.
31. Lines, M.A., et al., *Haploinsufficiency of a spliceosomal GTPase encoded by EFTUD2 causes mandibulofacial dysostosis with microcephaly.* Am J Hum Genet, 2012. **90**(2): p. 369-77.
32. Luquetti, D.V., et al., *"Mandibulofacial dysostosis with microcephaly" caused by EFTUD2 mutations: expanding the phenotype.* Am J Med Genet A, 2012. **161A**(1): p. 108-13.
33. Gordon, C.T., et al., *EFTUD2 haploinsufficiency leads to syndromic oesophageal atresia.* J Med Genet, 2012. **49**(12): p. 737-46.
34. Dryja, T.P., *Gene-based approach to human gene-phenotype correlations.* Proc Natl Acad Sci U S A, 1997. **94**(22): p. 12117-21.
35. Gilissen, C., et al., *Unlocking Mendelian disease using exome sequencing.* Genome Biol, 2011. **12**(9): p. 228.
36. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. **7**(4): p. 248-9.

Project 5. **Study of possible mechanisms of incomplete penetrance of *SNRNP200* mutations**

Incomplete penetrance is a phenomenon occurring most often in autosomal dominant diseases in which some of the members of a family segregating a mutation do not manifest the symptoms of the disease. Moreover, mutations can have variable expressivity determining differences in the presentation and in the severity of the symptoms. The causes for different penetrance and expressivity can be environmental, genetic or epigenetic. In retinitits pigmentosa there are very rare examples of incomplete penetrance, most of which are linked to mutations of the splicing factor gene *PRPF31* and determined by genetic modifiers influencing its expression. A few reports have indicated that also mutations of other RP-linked splicing factor genes can display this feature. Our collaborator Prof. Christian Hamel, MD, identified and examined a 4-generation French family with adRP and segregating the *SNRNP200* p.R681C mutation, which was first reported in three patients from our screening (project 2). The mutation is transmitted from generation 1 to 4. However, carriers from the 2nd and 3rd generation are asymptomatic. We were involved in this study to find possible explanations for incomplete penetrance observed in this family by studying lymphoblastoid cell lines derived from family members. We also had clinical information and cell lines from *SNRNP200* patients from our original paper, and included them in the analysis, to have better insight into *SNRNP200* mutations at both clinical and, possibly, molecular levels. The results are still preliminary and need further experimental work. I presented them here in the form of an article for coherence with the rest of the thesis. We plan to complete the analysis and to submit a manuscript for publication by the end of year 2013.

Candidate's role

- Patients' lymphoblastoid cell lines immortalization and culture
- Gene expression analysis by Real time PCR and Western Blot
- Writing of the manuscript (except for the clinical parts, provided by Prof. Hamel)

A Family with Autosomal Dominant Retinitis Pigmentosa Segregating *SNRNP200* Mutation shows incomplete penetrance and phenotypic variability

Paola Benaglio[1], Isabelle Meunier[2], Tremeur Guillaumie[2], Gaël Manes[2], Shyana Harper[3], Eliot L. Berson[3], Carlo Rivolta[1], Christian Hamel[2]

[1]Department of Medical Genetics, University of Lausanne, Lausanne; [2]Service d'Ophtalmologie, Université et Hôpital de Montpellier, Montpellier; [3]The Berman-Gund Laboratory for the Study of Retinal Degenerations, Harvard Medical School, Massachusetts Eye and Ear Infirmary, Boston.

ABSTRACT

Retinitis pigmentosa (RP) is a genetically heterogeneous monogenic disease leading to progressive degeneration of the photoreceptor layer of the retina. Mutations in splicing factor genes contribute to a large proportion of autosomal dominant RP (adRP) cases. The gene *SNRNP200* encodes a splicing factor protein of 200 kDa and has been linked to adRP by identification of several missense mutations. We identified for the first time a family segregating the *SNRNP200* p.R681C mutation showing incomplete penetrance and intra-familial variability with two severely affected members and one mildly affected, whose clinical phenotypes are reported in detail. The examination of other patients with *SNRNP200* mutations shows a picture of rather mild-variable phenotypes, consistent with previous literature. Molecular characterization of the mutation p.R681C in lymphoblastoid cell lines from patients reveals a decrease in *SNRNP200* protein compared to controls. Further investigation on changes in *SNRNP200* protein stability upon missense mutations might help to elucidate the mechanisms by which *SNRNP200* mutations may determine variable RP phenotypes.

INTRODUCTION

Retinitis pigmentosa (RP) is a heterogeneous group of inherited retinal dystrophies affecting more than one million people worldwide [1]. Typically patients present with difficult dark adaptation and night blindness in adolescence, and loss of mid-peripheral and far-peripheral vision in young adulthood. Central vision may be eventually lost as well. The main hallmarks of RP at *fundus* examination are the presence of melanin deposits resembling bone spicule

released from the RPE, retinal atrophy and retinal arterioles attenuation. Electroretinography (ERG) is used for objective diagnosis of RP and monitoring of the progression of the disease. Another tool used to analyze retinal thickness and integrity especially in the macula is optical coherence tomography (OCT). Patients display a wide variability in the manifestations of the disease, including age of onset, rate of progression and fundus appearance [2]. Genetic factors heavily influence the outcome of this pathology. RP is caused by mutations in more than 50 genes, which can be inherited most often as autosomal dominant, recessive or X-linked traits [3]. The genetic heterogeneity of RP also includes rather frequent cases of inter-individual variability, in which the same mutation may result in different expressivity of phenotype. Incomplete penetrance is far rarer, and is mostly observed in individuals carrying mutations in the second most common adRP gene, *PRPF31,* which encodes for a component of the pre-mRNA spliceosome [4]. In *PRPF31* families, incomplete penetrance is due to a differential mRNA expression between affected and asymptomatic carriers of mutations [5, 6]. Isolate reports of RP with incomplete penetrance have been described for other genes including *RP1* [7], *PAP-1* [8] and *PRPF8* [9], suggesting that the mechanisms of RP phenotype non-penetrance are multiple and possibly related to splicing factor genes.

The 200 kDa DExD/H-box RNA helicase hBrr2, encoded by the *SNRNP200* gene, is another core component of the spliceosome complex, which was first linked to adRP in two different Chinese families [10, 11]. They segregated two different missense changes (p.S1087L and p.R1090L) which are located in the N-terminal Sec63-like domain and were shown to impair Brr2 helicase activity in yeast [10]. Later, four additional missense mutations (p.Y689C, p.R681C, p.R681H [12] and p.Q885E [13]) were identified in the first and more conserved of the two Hel308-like helicase domains. Although their effect at the molecular level have not been yet tested, their pathogenicity is strongly supported by segregation in families, conservation of the residues, absence of the changes in control population, and crystal structure models [13, 14]. At the clinical level, *SNRNP200* mutations are usually associated with fully penetrant adRP with variable phenotypes and relatively late onsets [11, 13, 15].

MATERIALS AND METHODS

Clinical and genetic assessment

Patients R681C_1 to _5 were examined at the Centre of Reference for Genetic Sensory Diseases, in Montpellier, France, as described before [16]. Patients 001-303, 001-051, 001-061, 001-367 and their relatives were ascertained at the Berman-Gund Laboratory,

Massachusetts Eye and Ear Infirmary – Harvard Medical School, according to procedures that were previously published [17]. Molecular diagnosis of patients R681C_1 to _5 was performed using primers to amplify coding and intron flanking sequences of exons 16 and 25 of *SNRNP200*. The genetic analysis of patients 001-303, 001-051, 001-061, 001-367 and their relatives was performed in previous work [12]. All patients were enrolled in this project by signing written consent form. Our study was conducted in adherence with the tenets of the Declaration of Helsinki and approved by the Review Boards of our Institutions.

Lymphoblastoid cell lines

Lymphocytes from carriers of *SNRNP200* mutations were isolated from 20 ml peripheral blood samples using Ficoll (GE Healthcare, Munich, Germany) gradient centrifugation and were immortalized with Epstein-Barr virus. Control lymphoblastoid cells were obtained from the Coriell repository. Cells were grown in T25 flasks in RPMI-1640 medium and GlutaMax (Life Technologies, Carlsbad, CA), supplemented with 1% of penicillin/streptomycin and 10% fetal bovine serum (Life Technologies). Confluent cells (0.5-1 million cells/ml) were pelleted by centrifugation, washed with PBS buffer, and pelleted again for extraction of RNA or proteins.

RNA extraction and cDNA Synthesis

Ten million cells were lysed and homogenized with QIAshredder columns (Qiagen, Venlo, The Netherlands) and RNA was purified using RNeasy kit (Qiagen) and eluted in DHPC water. cDNA was synthesized from 2 μg of total RNA using PrimeScript RT-PCR Kit (Takara, Shiga, Japan) by priming with oligo dT and random hexamer oligonucleotides in a final volume of 20 microliters.

Real-Time PCR

The amplification of *SNRNP200* transcripts for quantitative Real-Time PCR in patients and controls was performed with Sybr Green PCR Master Mix (Life Technologies) and the following primers: CTTCTCGGAGTCTCTGCTGG (sense) and TGAAGCACGTCATAGATGGG (antisense), spanning exons 35 and 36 (NM_014014.4). The housekeeping gene *GAPDH* for relative quantification was amplified by TaqMan Universal PCR Master Mix and a premix of primers and a VIC/MGB probe obtained by Life Technologies. Reactions were assembled with 1% of cDNA (~20 ng retrotranscribed RNA) in

a final volume of 20 µl. Forty cycles at 95°C for 15 sec and 60°C for 1 min, after an initial denaturation of 10 minutes at 95 °C were performed in an ABI PRISM 7500 Sequence Detector. mRNA expression of *SNRNP200* was normalized with respect to *GAPDH*, and on the ΔCt averages using the ΔΔCt method.

Allelic imbalance assay with taqman probes

To determine the specific amplification from the two alleles harboring the wild type base and the c.2041C>T (p.R681C) mutation, we designed the following probes: VIC-CAACAGCTTCCGTCCAGTGC-MGB (wt), 6-FAM-CAACAGCTTCTGTCCAGTGC-MGB (mut) and primers: CAAGGGTCTCTTTTACTTTG (sense), TCTGGAAACGCTTGATAGC (antisense) using the AlleleID software (Premier Biosoft, Palo Alto, CA) and synthesized by Life Technologies. The mutation is located close to the junction of exons 15 and 16, therefore the amplification products derive from cDNA only. To obtain templates of known concentration for the standard curve, we sub-cloned a 200 bp fragment encompassing the mutation, amplified from a patient (heterozygous for the change) with the following primers: TGCCACCCTACCCAACTATGA (sense), TCCAGCATGTTCCATGATTTTT (antisense). The PCR product was inserted by blunt ligation into a pcDNA3 vector cut with EcoRV. Transformed competent cells were selected by colony-PCR and plasmids were purified and measured by Nanodrop (Thermo Scientific). We selected two clones for which Sanger sequencing confirmed the insertion of a single copy of the PCR, harboring either the wt or the mutated allele. Optimized reactions for allelic expression discrimination consisted of 100 nM of each probe, 300 nM of each primer, TaqMan Universal PCR Master Mix, 5% cDNA or 0.04 ng of standard plasmid, in a final volume of 20 µl. Cycling conditions were the same as for the Syber Green reactions. Each sample was tested by three replicates consisting of different RNA extractions. For the standard curve the wt and mutated plasmids were pooled at known ratios of 8:1, 4:1, 2:1, 1:1, 1:2, 1:4 and 1:8. To calculate the relative amount of the wt allele versus the mutated one, we substitute the ΔC_t ($C_{talleleC}$-$C_{talleleT}$) to the equation obtained from the standard curve: Log_2 [c.2041C/T]= -0.592*ΔC_t +0.0634 (R^2=0.986), as described in reference [18](**Supplemental Figure 1**).

Protein extraction and Western blot

Pelleted cells (~10 M) were lysed with 200-300 µl of lysis buffer (50mM Tris-HCl pH 7.6,

150 mM NaCl, 0.5% Triton and Protease Inhibitor Cocktail from Roche) in a rotating wheel at 4°C. After centrifugation at 10,000 rpm at 4°C, the supernatant was collected and quantified by the BCA Protein Assay (Thermo Scientific). Equal amounts of protein for each sample were loaded on a 6% acrylamide SDS page and transferred to nitrocellulose membrane. Proteins were revealed with HELIC2 (N-20) (Santa Cruz, sc-68563) and GAPDH (Ambion, 4300) or KIF3A (Abcam, ab11259) antibodies for normalization and revealed with LI-COR secondary fluorescent antibodies IRDye 800 anti goat and IRDye 680 anti mouse or rabbit, respectively. The bands were quantified using the program Image Studio (LI-COR Biosciences).

RESULTS

Clinical evaluation of adRP patients with *SNRNP200* mutations

By sequencing *SNRP200* exons 16 and 25 (hotspots for known mutations), we identified the p.R681C substitution segregating in family RP617 over four generations. Of the five carriers of the mutation, only two siblings and their great-grandmother presented symptoms of RP (**Figure 1**).

The elder of the two siblings (member #1 in the pedigree) was severely affected with first symptoms of night blindness appearing at age of 4 and attenuation of retinal vessels in fundus examination. Daylight activities and reading were normal. At the age of 7, her first visual field examination showed moderate narrowing of the isopter at 65-70° in temporal and 50-60° in nasal. At the age of 8, OCT showed a loss of photoreceptors outside the fovea. The full-field ERG was highly decreased. No rod responses at dim blue stimulation were found and responses at the maximum white stimulations were at about 25% of the normal value. The photopic responses were also decreased at 50% of the normal amplitudes and the 30-Hz flickers, specific for cones, were at 20% of the normal amplitude. At the age of 13, there were no more scotopic responses and the 30-Hz flickers dropped to 5% of the normal amplitude. At the time of last examination (age of 14), her visual acuity was normal in each eye with a correction for astigmatism. She had no cataract. On fundus examination, the foveal reflex was altered and the peripheral retina lacked shining reflexes. A few spots of retinal atrophy and pigment deposits were visible in temporal mid periphery and retinal vessels were attenuated (**Figure 2A-B**). The peripheral visual field was moderately reduced in periphery (75° to 90° in temporal, 45 to 50° in nasal) but there was a large annular scotoma in midperiphery. On OCT, the IS/OS line (the junction between the inner segment and the outer segment of

photoreceptors) was present in the fovea, but it disappeared beyond the center. Foveal thickness was slightly increased, with a few cysts of macular edema. At the autofluorescence test, there was a ring of autofluorescence typical of retinitis pigmentosa.

Her younger brother (member #2 in the pedigree) was mildly affected. The first visual field testing at 8 years old showed moderately decreased peripheral isopter at 70° in temporal and 50° in nasal. The full-field ERG showed an attenuated rod response with the dim blue stimulation at about 60% of the normal value. The same values were observed after three years. The photopic responses were moderately decreased at 70-80% of the normal amplitude and the 30-Hz flickers were in the same order. At the age of 9, the visual field was better at 85° temporal and 55° nasal. OCT showed that IS/OS line was present but it was attenuated in the periphery of the macula. At the time of last examination (age 12), his visual acuity was normal (20/20) in each eye with a correction for astigmatism. He had no cataract. The fundus showed slight depigmentation in the inferior retina but there were no spots of atrophy, no pigment deposits, no vessel attenuation and the macula was normal (**Figure 2C-D**). On OCT, there was no macular edema. The retinal autofluorescence was normal without macular ring of autofluorescence.

The patients' mother (member #3) and grandmother (member #4) were last seen at age of 28 and 49, respectively. They did not have any symptoms of RP. Both had normal visual acuity with correction for astigmatism, and normal ocular pressure. The full-field ERG was strictly normal for all scotopic and photopic stimulations. In the grandmother, only the 30-Hz flickers showed slightly decreased responses. The mother had a normal fundus (**Figure 3A-B**), while the grandmother had few small spots of atrophy in infero-temporal in each eye, but there were no pigment deposits, and the maculae, the retinal vessels as well as the optic disc were normal (**Figure 3C-D**). On slit lamp examination she had a cataract beginning in infero-temporal in the right eye.

The great-grandmother (member #5) was also a carrier of the p.R681C mutation and had a severe form of RP, with night blindness since infancy. At age of 18 years, she became aware of peripheral visual field defects and developed tunnel vision at the age of 70. She had cataract surgery at the age of 80. She also had an episode of acute glaucoma on the right eye at age 74. At the time of the examination (age 82), she could move only with the help of a stick and could not read. On examination, she had unoriented light perception of the right eye (because of acute glaucoma), and visual acuity was 1/20 on the left eye. The ocular pressure was normal at 13 mmHg in the right eye and at 18 mmHg on the left eye. The fundus showed

advanced retinitis pigmentosa with many pigment deposits and atrophy of the retina and choroid, mostly in pericentral regions (**Figure 2E-F**). The retinal vessels were very narrowed and the optic discs were waxy pale. There were small spots of foveal atrophy on both sides. A perifoveal ring of dark red-colored retina remained visible in both eyes. ERG was not performed because it would have been unresponsive.

The clinical details of this family as well as of other patients from France with *SNRNP200* mutations are reported in **Table 1**. The clinical examination details from patients originally identified in our previous paper to carry *SNRNP200* mutations, and used for expression studies are reported in **Table 2**.

Expression analysis

The molecular causes for incomplete penetrance are thought to derive from a different concentration of the disease gene product between affected and asymptomatic members, due to transcriptional regulation or modifier genes. We assessed by relative qPCR for possible expression differences between the members of family that are carrying the p.R681C mutation, as well as controls and unrelated patients carrying *SNRNP200* mutations p.R681C (patients 001-303 and 001-061), p.R681H (001-061) and p.S1087L (001-367). As shown in **Figure 4A**, the mRNA level in lymphoblastoid cell lines was similar for all individual tested. We then wanted to investigate a possible mechanism of allelic expression imbalance, under the hypothesis that a different level of expression from the wild type and the mutated alleles in affected and non affected carriers could underlie the different phenotypic outcome of the members of the family segregating p.R681C mutation. We calculated the expression ratios of Brr2 transcripts from the wild type allele versus the mutated allele by using allele specific Taqman probes. No significant difference between expressions of the two alleles was observed in asymptomatic, in affected members of the family or in unrelated patients (**Figure 4B**).

We then tested the hBrr2 protein level by Western Blot. Three protein extractions from lymphoblastoid cell lines grown in same conditions were performed. Although different protein preparations gave some variability in the results, we consistently observed lower amounts of Brr2 protein at the steady-state in carriers of the mutation p.R681C with respect to carriers of the other two mutations (p.R681H and p.S1087L) and, to a certain extent, to controls. In the group of control cell lines there was in fact certain variability in the protein amount (**Figure 5**). Surprisingly, in one of the three protein extraction replicates, the cell lines

from the two asymptomatic members of the family had very low level of protein with respect to the other tested cell lines (**Figure 6**).

DISCUSSION

We report here for the first time that retinitis pigmentosa due to heterozygous p.Arg681Cys in *SNRNP200* shows incomplete penetrance. The mutation segregates in a four-generation family with RP where mother and grandmother have no signs of retinitis pigmentosa, including normal ERG responses, while the great-grandmother her great granddaughter are severely affected. The great-grandson had moderate signs of retinal alteration and the decrease in ERG rod responses indicated a moderate loss of these photoreceptors. Yet, it is not certain whether these retinal alterations were progressing or not, in particular since there were no changes in ERG responses at a 3-year interval. In contrast, there were clearly signs of progressive disease in the great-grand daughter (ERG).

In general the phenotype of *SNRNP200* mutations reported so far and observed in the rest of the patients examined shows variability. The first description of a family with linkage to the RP33 locus, later identified to harbor the p.R1087L mutation, reports the diagnosis of dominant, fully penetrant RP, characterized by late onset of night blindness (16-18 years) and slow progression of the disease with variable phenotypes for both rods and cones vision [10, 15]. The second family that was identified, also from China, segregated the p.R1090L mutation. Patients had an earlier onset, between 7 and 11 years, with typical progressive narrowing of the visual field, while central vision was relatively preserved [11]. These two mutations locate in the first (N-terminal) of the two Sec63-like domains of the hBrr2 protein, a region that is important for the interaction with the RNA and other proteins. The aminoacid substitutions were shown to decrease binding to RNA and to reduce Brr2's helicase and ATPase activity in yeast [10, 14]. The last clinical report for *SNRNP200*-linked RP involves another Chinese family, segregating the p.Q885E mutation. Symptoms were more homogeneous within the patients of this family. Night blindness appeared between 10 to 15 years of age and was followed by a gradual decline in visual acuity and peripheral visual field loss at the age of 40. Also, glaucoma was diagnosed in two individuals, in their 40s [13].

Incomplete penetrance is defined when individuals with a given mutation do not always present the clinical manifestation of the disease. A similar, but yet distinct phenomenon is variable expressivity, where the type or the severity of a phenotype varies in individuals with the same mutations [19]. In the case of the family presented in this report, both incomplete

penetrance and variable expressivity are present. It is also possible to hypothesize that a common mechanism explains both phenomena and that incomplete penetrance is the "mildest" extreme of a spectrum of phenotypes [20]. Glöcke and collaborators have recently published the results of a genetic screening of known RP genes, which reported also that *SNRNP200* mutation p.S1087L was found in a patient and in the unaffected mother, indicating that incomplete penetrance can be a recurrent finding for different *SNRNP200* mutations [21]. It has to be remarked that, while for p.S1087L the authors provided a biochemical proof of the impairement of Brr2 activity in yeast, for p.R681C this evidence is missing. However, many genetic evidences support the causality of p.R681C for RP. It was found to cosegregate with RP in at least 5 families, it is absent in the general population, and it affects a very conserved residue. Moreover, Santos and coworkers provided recently a crystal structure of the two tandem helicase cassettes of human Brr2 protein where they highlighted the position of the residues that are mutated in RP. In particular, the cluster of residues R681, V683 and Y689, that we have found mutated in our first report, localize within a connecting domain in the first and more conserved helicase cassette, where they establish interdomain contacts, important for the stabilization of the folding [14]. In order to find clues to explain why we observed such divergent phenotypes in family RP612, we wanted to investigate whether there were any differences at the molecular level among carriers of the p.R681C mutation.

Despite incomplete penetrance and variable expressivity have been observed for several genetic disorders, their causes have been scarcely understood, and are likely to differ from case to case and influenced by environmental factors as well. The fact that these events have been found mostly for dominant diseases may indicate that they are related to the dose of the gene product. Therefore epigenetic, transcriptional or translational regulation of gene expression is thought to be involved [19]. In the case of *PRPF31*, for example, incomplete penetrance is due to differential mRNA expression that has genetic origins, between affected and asymptomatic members of a family. A *cis*-acting modifier gene increases the expression of the wild type protein, which in asymptomatic individuals compensates for the loss of the mutated allele, rescuing the hemizygous condition derived from *PRPF31* haploinsufficient mutations [5, 6, 22]. We first hypothesized that for *SNRNP200* mutations the mechanism could be similar. We tested this hypothesis with Real-Time PCR but no appreciable differences were found among affected and asymptomatic members of the family that was studied, in lymphoblastoid cell lines. However, to find significant correlations, mRNA

expression analyses should be performed on a higher number of family members, an experiment that was not possible in this specific case. Since *SNRNP200* mutations are missense and, unlike *PRPF31* mutations, are not predicted in principle to result in reduced protein concentrations, a more plausible mechanism for incomplete penetrance would be a difference in the relative amounts of the wild type over the mutated transcript between affected and asymptomatic relatives. This was observed for example for *BRCA1* and *BRCA2* mutations in familial breast cancer [18]. We could not see such a difference in our allelic imbalance expression assay. The sensitivity of this assay, constrained by the design of the probe, allowed to detect differences of more than 0.5-1 fold and not smaller, which would have been anyway of difficult interpretation in relation to the phenotype. We were then left to investigate possible differences in the total protein quantity. Although from our analysis it appears that hBrr2 protein has fluctuating concentrations in all cell lines that we analyzed, we could observe a certain reproducibility of the finding that p.R681C carriers have a decreased hBrr2 protein amount compared to other cell lines tested, including the ones from the carriers of p.R681H and p.S1087L mutations of the same gene. This may be in agreement with the structural role of the R681 residue in the conformational stability of the domain. The positive charge of arginine seems important to maintain this structural integrity and its disruption with a substitution by a non-polar aminoacid like cysteine, and maybe not by another positively charged aminoacid like histidine (as in the other mutation), might be detrimental for the stability of the whole protein. Since our analyses were performed at the steady-state, and may be confounded by a compensatory increase, we plan to confirm that the mutation causes indeed a reduction in protein stability by analyzing the protein after treatment with a translational inhibitor at different time points. The reason why in one instance it seems that asymptomatic carriers have even less protein than the affected relatives is difficult to interpret and further confirmatory experiments are needed to establish if there is a true difference or a stochastic finding, possibly due to the use of the particular antibody. Reduction or impairment in protein stability due to point mutations have been already indicated as a possible factor influencing the penetrance of mutations in different ocular diseases such as retinoblastoma (*RB* mutations) [23] primary congenital glaucoma (*CYP1B1* mutations) [24], and also in retinitis pigmentosa (*PRPF31* Ala216Pro mutation) [25]. The general hypothesis is that the protein derived from the hypomorphic allele would retain sufficient functionality to express the normal phenotype if concurrent factors are present. These factors can be of genetic nature, like compensatory variants in modifier genes or interacting protein partners or of

environmental origin, and their identification has proven to be challenging. By using novel whole genome sequencing approaches, it would be thinkable to perform genetic association analysis of non-penetrant p.R681C mutations, by focusing on sequence analysis of genes known to interact with hBrr2 or to play a role in retina development and pathogenesis. However, for this kind of study a larger number of non-penetrant p.R681C cases would be needed, and even in that case it is not sure that the cause for incomplete penetrance would be the same in all individuals.

In conclusion, our analyses so far exclude the possibility that penetrance of the p.R681C mutation is due to mechanisms of regulation at the RNA level, although only the steady state was investigated. An explanation at the level of protein stability can be hypothesized, for which confirmatory experiments are needed.

FIGURES AND TABLES

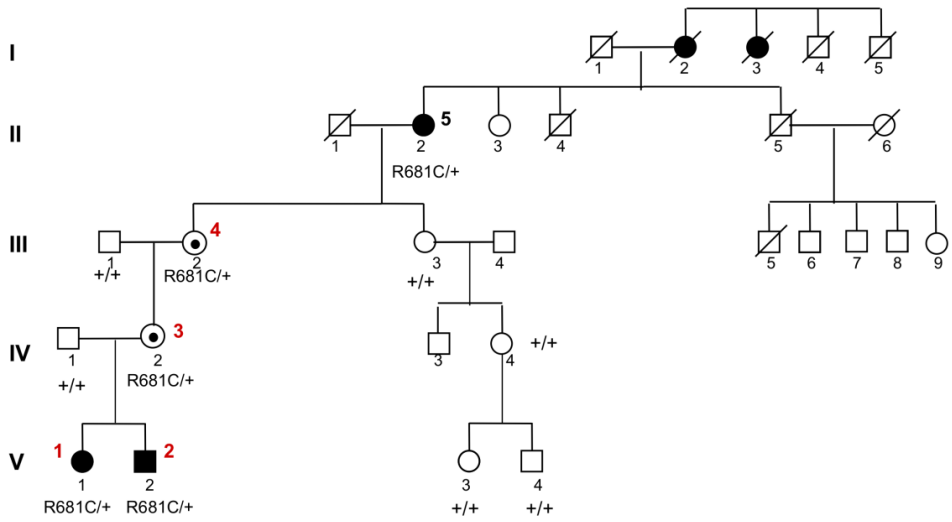


Figure 1. **Pedigree of family RP612.** The *SNRNP200* mutation p.R681C segregates in the branch of the family with RP. A red number designates the members of the family analyzed for RNA and hBrr2 protein.

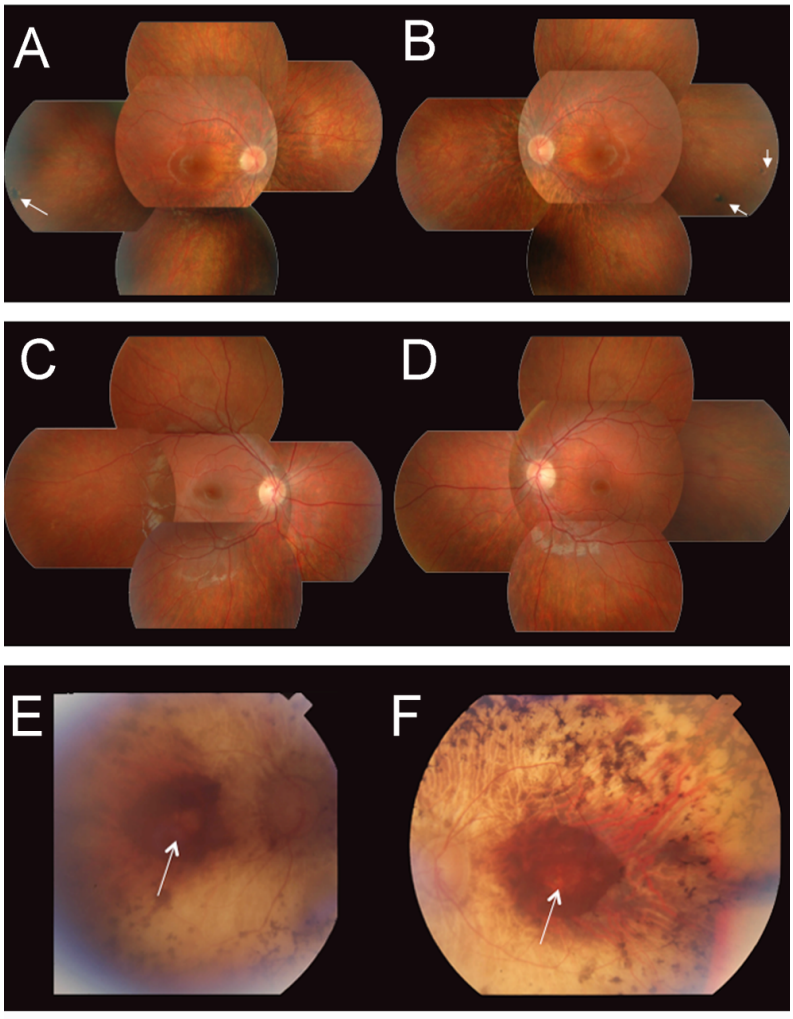


Figure 2. Fundus appearance of affected family members. A-B) Fundus photographs of right (A) and left (B) eyes of patient#1, age of 14, severely affected, showing attenuation of the retinal vessels and moderate atrophy of the peripheral retina. Pigment deposits in periphery are shown by white arrows. **C-D)** Fundus photographs of right (C) and left (D) eyes of patient#2, brother of patient 1, age of 12 and with less severe RP. A slight discoloration of the inferior retina is present. The retinal vessels are normal and there are no pigment deposits. **E-F)** Fundus photographs of right (E) and left (F) eyes of the great-grandmother, aged 82, showing atrophy of the retina and choroid (white color of the fundus), many pigment deposits, extreme narrowing of retinal vessels and waxy color of the optic discs. The atrophic foveas (white arrows) are surrounded by a dark red ring of retina.

Mutation	mgid	Family	Member	Age	Sex	Visual Acuity OD	Visual Acuity OS	Visual Field OD	Visual Field OS	DA	dim blue ERG OD	dim blue ERG OS	30HZ ERG OD	30HZ ERG OS	Lens OD	Lens OS	Macula OD	Macula OS	Periphery OD	Periphery OS
p.R681C	**p.R681C_1**	RP612	V:1	14	F	20/20	20/20	75°T, 45°N	75°T, 45°N	nd	ndr	ndr	5.0	5.0	-	-	-	-	+	+
p.R681C	**p.R681C_2**	RP612	V:2	12	M	20/20	20/20	80°T, 60°N	80°T, 60°N	nd	110 µV	110 µV	80 µV	80 µV	-	-	-	-	-	-
p.R681C	**p.R681C_3**	RP612	IV:2	28	F	20/20	20/20	nd	nd	nd	Nl	Nl	Nl	Nl	-	-	-	-	-	-
p.R681C	**p.R681C_4**	RP612	III:2	49	F	20/20	20/20	nd	nd	nd	Nl	Nl	85 µV	85 µV	+	-	-	-	-	-
p.R681C	p.R681C_5	RP612	II:2	82	F	LP	20/400	nd	nd	nd	nd	nd	nd	nd	+	+	+	+	+	+
p.R681C		RP845	II:2	71	M	HM	4/10	nd	nd	nd	nd	nd	nd	nd	+	IOL	+	+	+	+
p.R681C		RP845	III:7	31	F	20/20	20/200*	80°T, 40°N	80°T, 30°N	nd	ndr	ndr	ndr	ndr	-	-	-	+	+	+
p.S1087L		RP1767	IV:3	61	M	20/30	20/60	Tubular 8°	Tubular 7°	nd	ndr	ndr	ndr	ndr	IOL	IOL	+	+	+	+
p.S1087L		RP1767	V:1	35	M	20/20	20/20	60°T, 20°N	60°T, 20°N	nd	ndr	ndr	ndr	ndr	+	+	-	-	+	+

Table 1. Clinical summary of final visits of patients with SNRNP200 mutations associated with retinitis pigmentosa from Centre of Reference for Genetic Sensory Diseases, Montpellier. Visual Acuity: best corrected Snellen visual acuity; LP = light perception; HM = hand motion. **Visual Field**: Goldmann maximum extension in periphery of the V-4e white test light, T = temporal, N = nasal. **DA** (dark adaptation): final threshold in log units above normal after 45 minutes of dark adaptation. **ERG**: full field ERG amplitude in microvolts to scotopic dim blue stimulation, lower limit = 160 µV; 30HZ photopic white light (lower norm = 100 µV); ndr = no detected response. **Lens**: clear lens -; central posterior subcapsular cataract +, IOL = intraocular lens. **Macula**: within normal limits -; granular +. **Periphery**: bone spicule or clumped pigment in one or more quadrants: + present, - absent. nd = not done; Nl = normal; (*) macular hole.

Mutation	ID	Age	Sex	Visual AcuityOD	Visual AcuityOS	Visual Field OD	Visual Field OS	DA	0.5HZ ERGOD	0.5HZ ERGOS	30HZ ERG OD	30HZ ERG OS	Lens OD	Lens OS	Macula OD	Macula OS	Periphery OD	Periphery OS
p.R681C	**001-303***	8	M	20/20	20/20	NA	NA	1.5	45.0	45.0	17.0	17.0	-	-	-	-	-	-
p.R681C	218-480*	15	F	20/30	20/30	12197	6818	1.5	16.0	16.0	1.5	1.7	-	-	-	-	+	+
p.R681C	218-456*	30	M	20/70	20/40	2563	2517	2.5	2.4	3.5	1.7	1.4	+	+	+	+	+	+
p.R681C	001-046	25	F	20/50	20/50	3252	3523	NA	4.3	4.8	1.2	1.3	+	+	+	+	+	+
p.R681C	**001-061**	31	F	20/20	20/20	22252	22704	NA	51.9	79.6	35.8	58.0	-	-	+	+	+	+
p.R681H	**001-051**	40	M	20/30	20/30	2203	2765	NA	3.3	3.5	0.6	0.6	-	-	+	+	+	+
p.S1087L	001-085	22	F	20/20	20/25	21188	19437	NA	8.2	6.5	3.9	4.5	-	-	+	+	+	+
p.S1087L	**001-367**	32	F	20/200	20/50	1562	1193	3.0	2.5	4.2	0.3	0.9	+	+	+	+	+	+
p.S1087L	001-212	49	F	20/70	20/30	3708	4714	NA	5.9	5.4	5.0	5.1	+	+	+	+	+	+
p.T689C	001-107**	26	F	20/30	20/25	15675	16443	NA	6.8	5.7	3.8	4.4	-	-	+	+	+	+
p.T689C	001-130**	49	F	20/20	20/20	2881	2066	3.5	ND	ND	0.2	0.2	+	+	-	-	+	+

Table 2. Clinical summary of first visits of patients with SNRNP200 mutations associated with retinitits pigmentosa from Berman-Gund Laboratory, Boston. Visual Acuity: best corrected Snellen visual acuity. **Visual Field**: Goldmann maximum extension in periphery of the V-4e white test light (lower norm = 11.399 degrees squared). **DA** (dark adaptation): final threshold in log units above normal after 45 minutes of dark adaptation. **ERG**: full field ERG amplitude in microvolts to white light single 0.5HZ flash (lower norm = 350); 30HZ white light (lower norm = 50). **Lens**: clear lens -; central posterior subcapsular cataract +. **Macula**: within normal limits -; granular +. **Periphery**: bone spicule or clumped pigment in one or more quadrants: + present, - absent. NA means data not available. (*) Patients from family 0270, (**) Patients from family 5632.

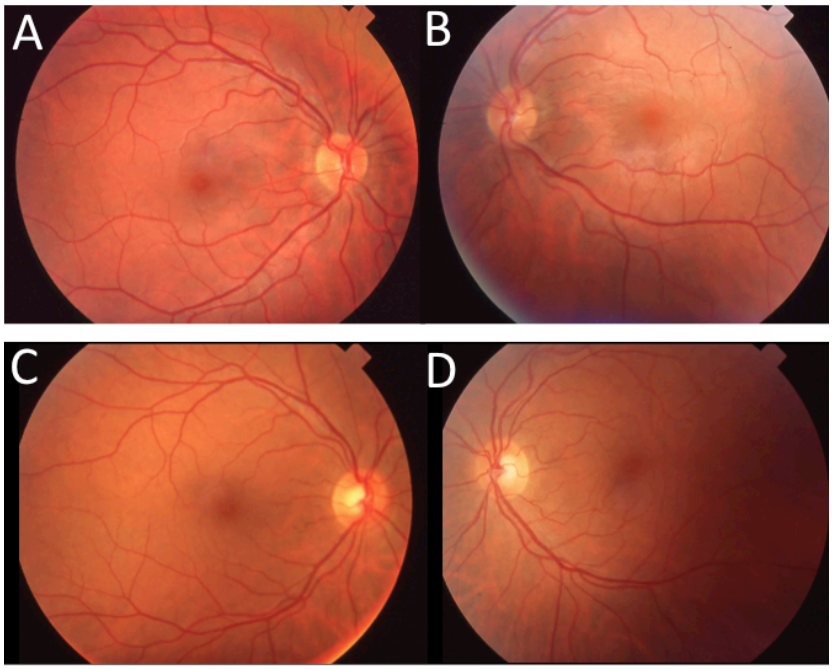


Figure 3. Fundus photographs of the unaffected carriers from the family. The right and left eyes of the mother, aged 28 **(A-B)** and the grandmother, aged 49 **(C-D)** show a normal posterior pole.

SAMPLES	Total transcript	Ratio wt/mut
p.R681C_1	1.016 ± 0.268	1.098 ± 0.035
p.R681C_2	1.024 ± 0.030	1.041 ± 0.102
p.R681C_3*	1.162 ± 0.134	1.036 ± 0.081
p.R681C_4*	0.926 ± 0.005	1.035 ± 0.055
p.R681C 01-303	1.066 ± 0.074	0.988 ± 0.107
p.R681C 01-061	1.095 ± 0.022	1.086 ± 0.128
p.R681H 01-051	0.838 ± 0.053	>4
p.S1087L 01-367	1.089 ± 0.067	>4
control 1	1.118 ± 0.092	>4
control 2	1.082 ± 0.192	>4
control 3	0.857 ± 0.082	>4
control 4	1.140 ± 0.019	>4

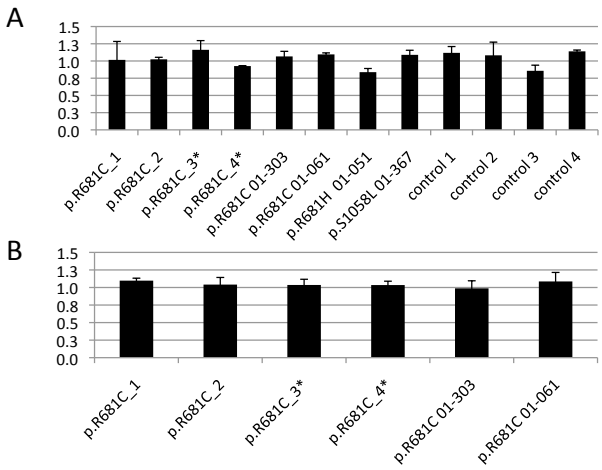


Figure 4. **Expression analysis in lymphoblastoid cell lines from patients with *SNRNP200* mutations**. **A)** Relative qPCR of *SNRNP200* transcript normalized on *GAPDH* expression and on the average expression. The values are an average of three experimets from different RNA extractions. Values are similar across patients, controls and asymptomatics. **B)** Taqman allelic imbalance assay from alleles c.2041T/C of *SNRNP200* in carriers of the mutation p.R681C. The relative abundance of the transcript containing the wt over the mutated allele was calculated using a standard curve of cloned cDNA fragments at known proportions, as described in the method and in the supplemental Figure S1. All the samples showed roughly a 1:1 proportion of the two variants. (*) asymptomatic individuals.

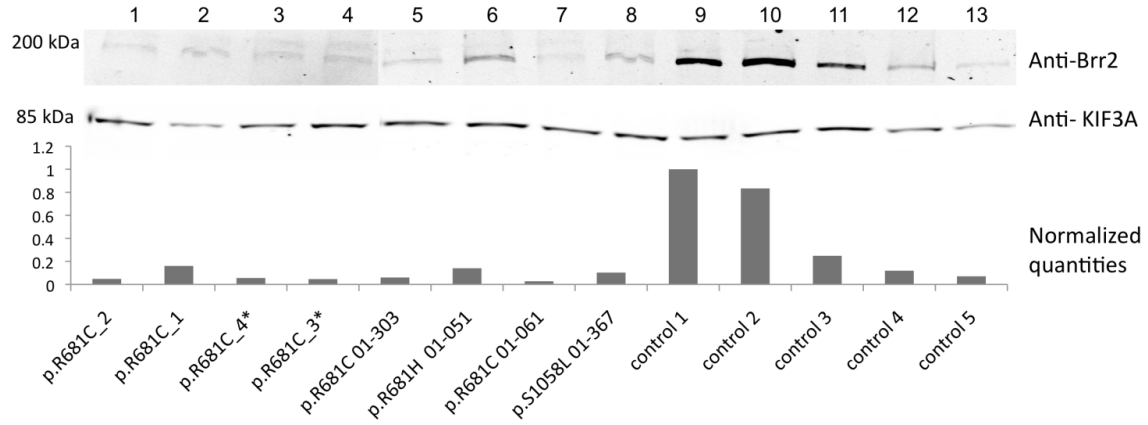


Figure 5. hBrr2 protein in lymphoblastoid cell lines from patients with *SNRNP200* mutations (I). Total protein extract form immortalized lymphoblasts were loaded in equal amount and revealed in western blot with Brr2 antibody (upper panel) and KIF3A antibody (middle panel). The relative densitometry quantification of the band is plotted in the panel at the bottom. (*) asymptomatic individuals.

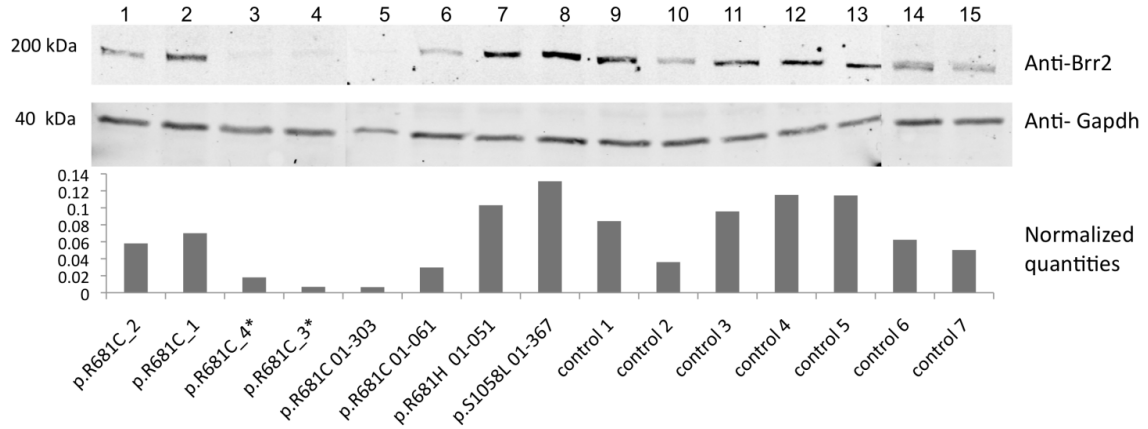
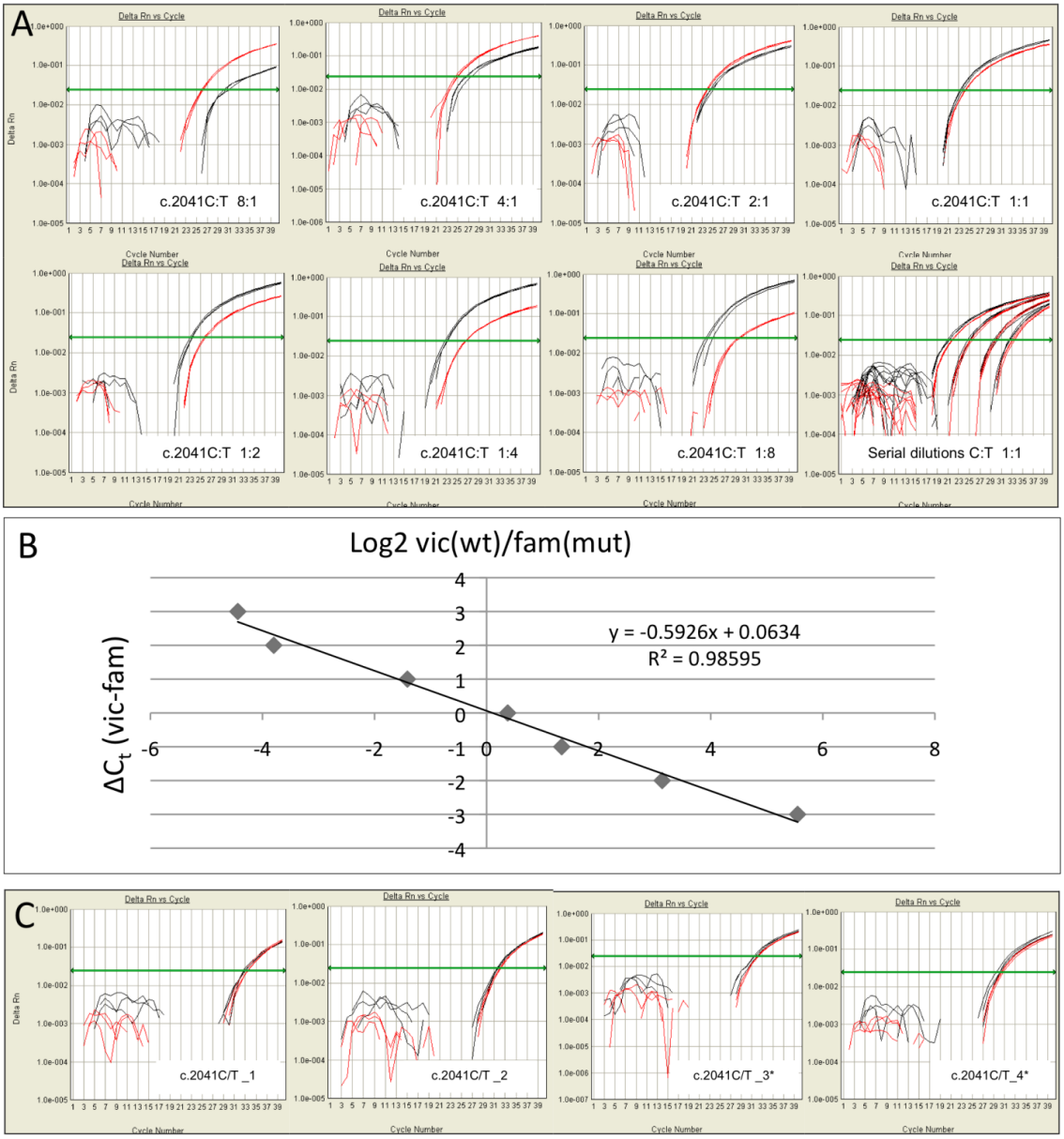


Figure 6. hBrr2 protein in lymphoblastoid cell lines from patients with *SNRNP200* mutations (II). Total protein extract form immortalized lymphoblasts were loaded in equal amount and revealed in western blot with Brr2 antibody (upper panel) and GAPDH antibody (middle panel). The relative densitometry quantification of the band is plotted in the panel at the bottom. Carriers of the p.R681C mutation (samples 1 to 6) had reduced protein amount compared to carriers of p.R681H and p.S1087L (samples 7 and 8). The two asymptomatic members (samples 3 and 4) of the family and patient 01-303 show very low protein levels. (*) asymptomatic individuals.

SUPPLEMENTAL FIGURES



Supplemental figure S1. Standard curve for allelic imbalance espression quantification. **A)** The panels from left to right show the amplification plots relative to the VIC Taqman probe, specifically recognizing the wt allele (c.2041C) in red, and the one from the FAM probe, recognizing the mutated allele (c.2041T) in black. From left to right plasmids containing cDNA from wt and mutated alleles are used at known ratios of 8:1, 4:1, 2:1, 1:1, 1;2, 1:4, 1;8. The last panel represents four 10-fold serial dilutions of the 1:1 plasmid mix. **B)** Standard curve obtained from the C_ts of panels A. The equation of the curve is displayed and used to calculate the relative expression of the two variants in the sample's cDNA. **C)** Amplification plots obtained from the cDNA of the affected (1-2) and unaffected (3*-4*) family members, showing almost overlapping curves from VIC and FAM probes, indicating a similar abundance of the two variants in all the individuals tested.

REFERENCES

1. Berson, E.L., *Long-term visual prognoses in patients with retinitis pigmentosa: the Ludwig von Sallmann lecture.* Exp Eye Res, 2007. **85**(1): p. 7-14.
2. Hamel, C., *Retinitis pigmentosa.* Orphanet J Rare Dis, 2006. **1**: p. 40.
3. Hartong, D.T., E.L. Berson, and T.P. Dryja, *Retinitis pigmentosa.* Lancet, 2006. **368**(9549): p. 1795-809.
4. Vithana, E.N., et al., *A human homolog of yeast pre-mRNA splicing gene, PRP31, underlies autosomal dominant retinitis pigmentosa on chromosome 19q13.4 (RP11).* Mol Cell, 2001. **8**(2): p. 375-81.
5. Vithana, E.N., et al., *Expression of PRPF31 mRNA in patients with autosomal dominant retinitis pigmentosa: a molecular clue for incomplete penetrance?* Invest Ophthalmol Vis Sci, 2003. **44**(10): p. 4204-9.
6. Rivolta, C., et al., *Variation in retinitis pigmentosa-11 (PRPF31 or RP11) gene expression between symptomatic and asymptomatic patients with dominant RP11 mutations.* Hum Mutat, 2006. **27**(7): p. 644-53.
7. Dietrich, K., et al., *A novel mutation of the RP1 gene (Lys778ter) associated with autosomal dominant retinitis pigmentosa.* Br J Ophthalmol, 2002. **86**(3): p. 328-32.
8. Kim, R.Y., et al., *Autosomal dominant retinitis pigmentosa mapping to chromosome 7p exhibits variable expression.* Br J Ophthalmol, 1995. **79**(1): p. 23-7.
9. Maubaret, C.G., et al., *Autosomal dominant retinitis pigmentosa with intrafamilial variability and incomplete penetrance in two families carrying mutations in PRPF8.* Invest Ophthalmol Vis Sci, 2011. **52**(13): p. 9304-9.
10. Zhao, C., et al., *Autosomal-dominant retinitis pigmentosa caused by a mutation in SNRNP200, a gene required for unwinding of U4/U6 snRNAs.* Am J Hum Genet, 2009. **85**(5): p. 617-27.
11. Li, N., et al., *Mutations in ASCC3L1 on 2q11.2 are associated with autosomal dominant retinitis pigmentosa in a Chinese family.* Invest Ophthalmol Vis Sci, 2010. **51**(2): p. 1036-43.
12. Benaglio, P., et al., *Next generation sequencing of pooled samples reveals new SNRNP200 mutations associated with retinitis pigmentosa.* Hum Mutat, 2011. **32**(6): p. E2246-58.
13. Liu, T., et al., *A novel missense SNRNP200 mutation associated with autosomal dominant retinitis pigmentosa in a Chinese family.* PLoS One, 2012. **7**(9): p. e45464.
14. Santos, K.F., et al., *Structural basis for functional cooperation between tandem helicase cassettes in Brr2-mediated remodeling of the spliceosome.* Proc Natl Acad Sci U S A, 2012. **109**(43): p. 17418-23.
15. Zhao, C., et al., *A novel locus (RP33) for autosomal dominant retinitis pigmentosa mapping to chromosomal region 2cen-q12.1.* Hum Genet, 2006. **119**(6): p. 617-23.
16. Bocquet, B., et al., *Relative frequencies of inherited retinal dystrophies and optic neuropathies in Southern France: assessment of 21-year data management.* Ophthalmic Epidemiol, 2013. **20**(1): p. 13-25.
17. Berson, E.L., *Retinitis pigmentosa. The Friedenwald Lecture.* Invest Ophthalmol Vis Sci, 1993. **34**(5): p. 1659-76.
18. Chen, X., et al., *Allelic imbalance in BRCA1 and BRCA2 gene expression is associated with an increased breast cancer risk.* Hum Mol Genet, 2008. **17**(9): p. 1336-48.
19. Ahluwalia, J.K., et al., *Incomplete penetrance and variable expressivity: is there a microRNA connection?* Bioessays, 2009. **31**(9): p. 981-92.
20. Zlotogora, J., *Penetrance and expressivity in the molecular age.* Genet Med, 2003. **5**(5): p. 347-52.
21. Glockle, N., et al., *Panel-based next generation sequencing as a reliable and efficient technique to detect mutations in unselected patients with retinal dystrophies.* Eur J Hum Genet, 2012.
22. Venturini, G., et al., *CNOT3 is a modifier of PRPF31 mutations in retinitis pigmentosa with incomplete penetrance.* PLoS Genet, 2012. **8**(11): p. e1003040.
23. Otterson, G.A., et al., *Temperature-sensitive RB mutations linked to incomplete penetrance of familial retinoblastoma in 12 families.* Am J Hum Genet, 1999. **65**(4): p. 1040-6.
24. Campos-Mollo, E., et al., *CYP1B1 mutations in Spanish patients with primary congenital glaucoma: phenotypic and functional variability.* Mol Vis, 2009. **15**: p. 417-31.
25. Huranova, M., et al., *A mutation linked to retinitis pigmentosa in HPRP31 causes protein instability and impairs its interactions with spliceosomal snRNPs.* Hum Mol Genet, 2009. **18**(11): p. 2014-23.

Part II.

Genome-wide association study of essential hypertension

As a parallel project during my PhD training I have collaborated with a European consortium named HYPERGENES (www.hypergenes.eu), which aimed at the identification of novel genetic loci associated with essential hypertension by case-control genome-wide association studies. In this section I enclose the two publications that I contributed to, after presenting a general introduction to the subject and the main findings of the study. In the first paper, published on *Hypertension* in Sepetember 2012, we discovered a novel locus associated with hypertension in the promoter of the *NOS3* gene and in the second one, to be published on *Hypertension* in November 2013, we proved the direct causality of the SNP in the susceptibility to hypertension.

Candidate's role:

- Management and quality control of DNA samples received from partner Universities for genotyping (paper #1 and #2).
- Array-based genotyping of a few thousands samples (paper #1 and #2).
- Luciferase assay to validate the effect of the risk allele: plasmid construction, assay optimization, experiment on HEK293T cells (paper #2).

INTRODUCTION

1. Hypertension and GWAS

Cardiovascular diseases (CVDs) such as heart failure and stroke are the major cause of death in Western countries [100] and hypertension causes about half of cardiovascular mortality [101]. Essential (or primary) hypertension (EH) is defined by a systolic blood pressure (SBP) of more than 140 mmHg or diastolic blood pressure (DBP) of more than 90 mmHg, in which causes of secondary hypertension such as renal failure and aldosternonism are not present. EH accounts for the 95% of cases of hypertension [102]. The values of blood pressure that define hypertension are a cutoff at the right end of the normal distribution present in the general population (**Fig. 1**). Hypertension is a complex trait resulting from the interactions among multiple genetic variants and environmental factors such as diet, smoking and exercise. Some forms of familial hypertension caused by mutations in single genes also exist, but these monogenic forms represent only a minority [103]. It was estimated by several studies that genetic factors contribute to 31-68% of blood pressure variance [104], but the identification of common variants with small effect is much more difficult than identification of rare variants with high penetrance and large effect on the phenotype. In the past 10 years the entire sequence of the human genome and a catalogue of its variants across different individuals and populations have been made available. This has been possible thanks to the effort of many groups participating to the International HapMap project [105] and the Human Genome Project [74]. As a consequence of this advancement and together with the development of high-throughput genotyping platforms, the genetic risk factors for many common diseases including hypertension have been identified [106].

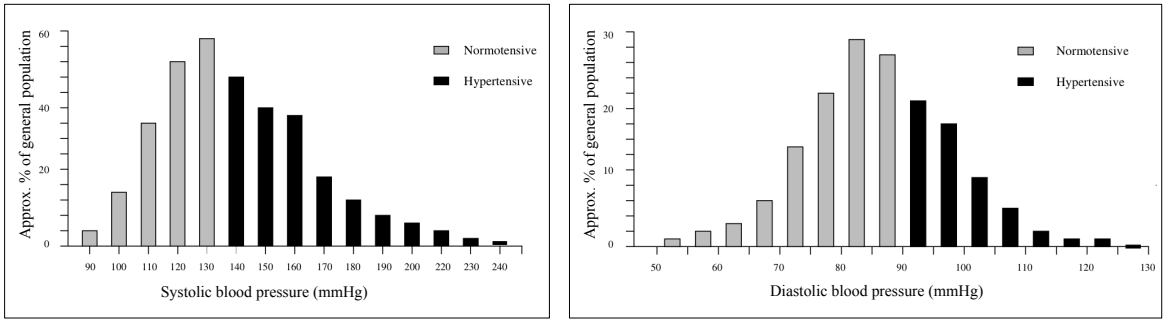


Figure 1. Normal distributions of SBP (left) and DBP (right) in the general population. Modified from [107].

Genome-wide association studies (GWAS) are used to find statistical association between common genetic variations and complex traits by comparing the genotypes of many individuals with different phenotypes. There are two types of GWAS: one is based on the

comparison of allele frequencies between individuals with a disease and matched control individuals (case-control study), while in the other type the phenotype is a quantitative trait that can be measured as a continuum within a population sample. High-density microarrays are used to genotype thousands of individuals for hundreds of thousands common SNPs spread over the whole genome. In fact, to reach the statistical power necessary to find genome-wide significant associations, and correct for the elevated number of independent tests, the sample size needs to be in the order of several thousand subjects [108]. The power of a GWAS increases with the number of individuals and SNPs tested. Some of the terms and the concepts frequently used in GWAS are explained in **Table 1**.

Term	Explanation
Bonferroni correction	Adjustment of the threshold of statistical significance in GWAS to limit the discovery of false positive results. Bonferroni correction divides the standard $P = 0.05$ by the number of statistical tests performed or, in GWAS, the number of variants being tested. Thus, for $P = 0.05$ and 10^6 variants tested, only associations of $P < 5 \times 10^{-8}$ are considered statistically significant.
Genetic risk score	A method to assess the combined effect of multiple variants that are individually associated with a disease. Scores are calculated by summing the number of alleles associated with increased disease risk. Top GWAS hits for a given condition are often used to identify the variants included in a genetic risk score.
Manhattan plot	A visual representation of GWAS results. Points graphed on a Manhattan plot represent *P values* for individual tests for association between a variant and the condition or trait of interest. The most important feature of the Manhattan plot is the presence of peaks formed by multiple points; this suggests a strong association at a particular locus.
Meta-analysis	A statistical method that combines results from multiple independent studies of the same condition. Meta-analyses of GWAS are used to increase sample size and boost statistical power to detect associations for variants of small effect.
Q-Q plot	A visual tool in which P values observed in GWAS are plotted against the expected P values derived from a χ^2 distribution. If the majority of observed P values deviate from expected P values, some artifact is likely inflating the significance of observed associations and must be adjusted in the GWAS regression model.
Statistical power	The probability of identifying statistically significant associations considering the desired significance threshold (i.e. $P = 0.05$), sample size, frequency of genetic variants in the general population, and frequency that specific variants are observed in cases. Well-powered GWAS require thousands of participants to detect associations between common variants with low effect size. Often, the power in GWAS is increased by meta-analysis of multiple GWAS that have studied the same phenotype.

Table 1. Glossary of main terms used in GWAS. Taken from [108].

Due to its high phenotypic heterogeneity, which is influenced by many factors including age, ethnicity, weight, diet, contractile state of the hearth and vascular tone, the genetic dissection of hypertension has progressed at a slower pace with respect to other common diseases. Nevertheless, in the last few years GWAS analyzing blood pressure as a continuous trait and by case-control design have identified about 40 loci robustly associated with it [103]. Some associated variants have been found near genes involved in important processes of regulation

of blood pressure including of the oxide-natriuretic signaling pathway (*NOS3, NPPA-NPPB, GUCY1A3-GUCY1B3*), vascular and endothelial function (*SH2B3, CSK, GNAS–EDN3, ATP2B1, LSP1/TNNT3, CACNB2*), renal electrolyte balance (*MTHFR, AGT, NPR3, UMOD)* and aldosterone synthesis (*CYP17A1)*. Some of these and other genes have been also found in Mendelian diseases and syndromes involving blood pressure [109]. Moreover, many of them were confirmed by meta-analyses and by large-scale association analysis of variants in candidate cardiovascular genes [110]. However, the evidence that many associated genes do not have a direct function in blood pressure regulation and that some of the most obvious candidate for hypertension have not been found near associated SNPs, indicate that less intuitive pathways of regulation constitute the genetic basis of hypertension.

A recognized limitation of the GWAS approach is that although statistically convincing associations have been identified for many conditions, in most of cases we lack the biological explanation for why a genomic interval associates with a complex trait [111]. The associated SNPs are often only proxies for variants lying in a same linkage disequilibrium (LD) block, among which the actual variants causing the observed phenotypic variability are hidden. The identification of causative variants affecting the phenotype is a complex task and involves the functional prediction of all possible SNPs in LD with the associated loci. In some cases the functional SNPs may directly affect the sequence of a protein by non-synonymous substitutions or truncating variants. However, it is largely agreed that the causative variants underlying GWAS signals have a regulatory function rather than coding. This hypothesis is supported also by the results of recent genome annotation efforts of the Encyclopedia of DNA Elements (ENCODE), which showed that GWAS SNPs were enriched in regions functionally labeled as enhancers, promoters, and transcription factor binding sites [112].

Another unforeseen limitation of GWAS is that although many genomic markers have been associated with a trait, they can explain only a small fraction of the variance of a quantitative trait or have low odd ratios for binary traits. As a consequence of this, most of the SNPs that have so far been found associated with common diseases have little predictive value and little clinical utility [113]. In particular, for hypertension, the aggregate common variants associated with blood pressure explain only the 3% of the phenotypic variance [103]. The concept of “missing heritability” refers to all genetic variation that was estimated to underlie common heritable diseases, and has not been uncovered yet. Genetic variants that are not usually captured with GWAS have been indicated as possible contributors to the missing heritability. They include rare variants, which are not in LD with the SNPs used for

genotyping, and structural variations such as copy number variants (CNVs- insertions and deletions) or copy neutral variations (inversions or translocations) [114]. Moreover, the lack of precise phenotypic information for a large number of individuals and of a measure of environmental effects are likely to contribute to our limit in understanding the molecular causes of complex diseases.

Current and future GWAS are aiming to overcome these drawbacks by different approaches. One of this is to analyze more accurate phenotypes through extensive clinical examination, or to evaluate more intermediate and molecular phenotypes, like gene expression in eQTLs association studies. Another one is to employ next generation sequencing to identify in an unbiased way rare variants, which may have a greater effect size. Additionally, targeted NGS can be applied for the identification of the actual causing variants among an associated genomic locus.

2. The "HYPERGENES" study

The design of the GWAS performed by the HYPERGENES consortium consisted of a case-control study organized in two stages. The strategy adopted to reduce the limitations explained above was to enroll extreme cases (DBP>90 mmHg and SPB>140 mmHg or under antihypertensive treatment before the age of 50) and controls (DBP <85 mmHg and SBP <135 at least until 55 years and never treated for hypertension) from extensively characterized cohorts followed up for many years in Europe. This determined a relatively small sample size, but had the aim to increase the discovery power by avoiding the effects of misclassifications biases. In the first phase of the study (discovery phase), 1865 cases and 1750 controls were analyzed for 1 million SNPs each. At the time of this study the 1M-SNPs chip from Illumina was the densest array used for GWAS and its use had a positive effect on statistical power. This was followed by a validation step where other 1385 cases and 1246 controls were genotyped for only 15 thousands SNPs including the top results of the discovery phase and variants in candidate genes for hypertension. Only one marker showed an association of genome-wide significance with hypertension and was the T allele (<10% MAF) of the *rs3918226* SNP in the promoter of the endothelial nitric oxide synthase gene (*NOS3*), a relevant protein for cardiovascular homeostasis and blood pressure regulation. The use of genotyping array with high number of markers played an important role in the discovery because this SNP was not included in arrays used in previous studies and, being in a region of low LD, was not possible to impute from genotypes of lesser dense arrays. Interestingly, the

polymorphism is located close to a predicted binding site for transcription factors important for *NOS3* expression, suggesting that the two alleles might result in a different transcription factor binding site (TFBS) strength and consequently to differential gene expression.

There are few cases in GWAS where the discovered associated variants have a direct effect on relevant gene products, and are thus likely to be the actual variants responsible for the association to the trait. The *NOS3* polymorphism *rs3918226* is one of these exceptions. In the follow up study in fact we could prove this causality using three different methods: i) targeted resequencing, ii) luciferase assay and iii) population study.

For the fine mapping of the *NOS3* locus, aiming at discovering other possible candidate causal variants, a 140-kb genomic area was re-sequenced by targeted high throughput sequencing in 44 hypertensive patients and 48 healthy controls from the HYPERGENES cohort. Two haplotypes including the risk allele were identified and one of it, about 27 kb long, was significantly more presents in patients than controls. However, the haplotype did not confer a higher association *P* value than rs3918226 alone. Five novel variants were located in the region of linkage disequilibrium upstream of rs3918226 but their annotation did not suggest any functional role. This indicated that rs3918226 was the closest marker responsible for the association signal. A direct evidence of the effect of the risk allele (T) on transcription with respect to the major one (C) was given by luciferase reporter assays in human cell lines (HeLa and HEK293T cells). The eNOS promoter containing variant T determined a significant reduction (20-40%) of transcription with respect to the C allele. Finally, a population study confirmed that the T allele present in homozygosis increased the risk of developing high blood pressure with age. Specifically, in a longitudinal study (7.6 years median of follow up) using blood pressure as a continuous phenotype in 2722 randomly recruited Europeans, systolic and diastolic blood pressures increased 5.9 and 4.8 mmHg more in TT homozygotes than in C allele carriers. Moreover, the hazard study on individuals that among the cohort developed hypertension (692) indicated that carriers of TT genotype had a two-fold risk of developing hypertension compared to the other genotypes.

3. eNOS and hypertension

Nitric oxide (NO) is a radical gas and in mammals is an important cellular signaling molecule involved in many processes. It is produced from L-arginine, oxygen and NADPH by three different nitric oxide synthase enzymes in different cells: nNOS (*NOS1*) in neurons, inducible iNOS (*NOS2*) in machrophages and eNOS (*NOS3)* in endothelial cells [115]. The role of NO

and eNOS in cardiovascular system is well known, as well as its implications in vascular disorders such as atherosclerosis and hypertension. NO biosynthesis is stimulated by various mechanical or humoral factors to induce vasodilatation [116] (**Fig. 2**). When diffusing to adjacent smooth muscle cells, NO stimulates the guanilate cyclase to form cyclic GMP, which will ultimately lead to relaxation of muscle cells and vasodilation. In addition, NO released from the endothelium to the lumen of the vessel has other functions, including the reduction of platelet aggregation and neutrophil adhesion to the endothelium, inhibition of vascular smooth muscle cell proliferation, and stimulation of angiogenesis. A decrease in nitric oxide synthesis produces an increase in blood pressure, as it was seen for example when introduding inhibitory L-arginine analogous in humans [117]. Moreover, it has been shown that in knockout mice for eNOS, acetylcholine-induced vascular relaxation is absent and mice are hypertensive [118]. Several genetic association studies have been addressing the role of genetic variants of eNOS in cardiovascular disease and other polymorphisms have been associated with a higher risk of hypertension [119]. Our work identified and characterized a novel susceptibility locus, consolidating the role of eNOS in cardiovascular risk in perspective of future therapeutic actions.

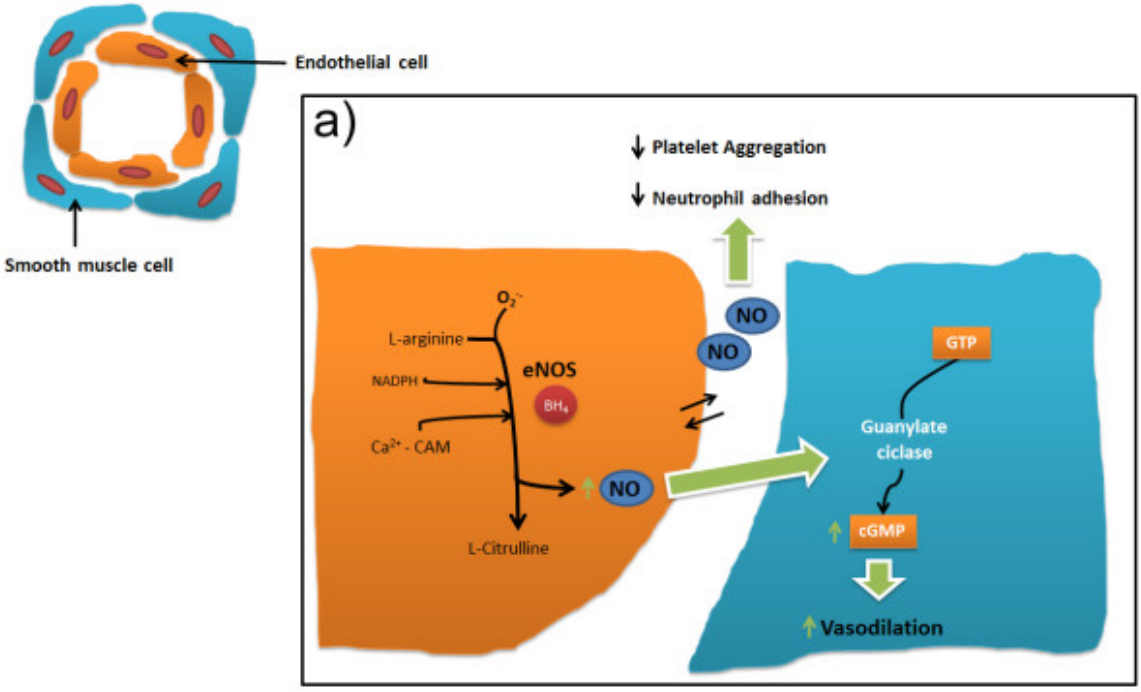


Figure 2. Scheme of nitric oxide pathway. Picture taken from [120].

PUBLICATIONS

Genomewide Association Study Using a High-Density Single Nucleotide Polymorphism Array and Case-Control Design Identifies a Novel Essential Hypertension Susceptibility Locus in the Promoter Region of Endothelial NO Synthase

Erika Salvi, Zoltán Kutalik, Nicola Glorioso, Paola Benaglio, Francesca Frau, Tatiana Kuznetsova, Hisatomi Arima, Clive Hoggart, Jean Tichet, Yury P. Nikitin, Costanza Conti, Jitka Seidlerova, Valérie Tikhonoff, Katarzyna Stolarz-Skrzypek, Toby Johnson, Nabila Devos, Laura Zagato, Simonetta Guarrera, Roberta Zaninello, Andrea Calabria, Benedetta Stancanelli, Chiara Troffa, Lutgarde Thijs, Federica Rizzi, Galina Simonova, Sara Lupoli, Giuseppe Argiolas, Daniele Braga, Maria C. D'Alessio, Maria F. Ortu, Fulvio Ricceri, Maurizio Mercurio, Patrick Descombes, Maurizio Marconi, John Chalmers, Stephen Harrap, Jan Filipovsky, Murielle Bochud, Licia Iacoviello, Justine Ellis, Alice V. Stanton, Maris Laan, Sandosh Padmanabhan, Anna F. Dominiczak, Nilesh J. Samani, Olle Melander, Xavier Jeunemaitre, Paolo Manunta, Amnon Shabo, Paolo Vineis, Francesco P. Cappuccio, Mark J. Caulfield, Giuseppe Matullo, Carlo Rivolta, Patricia B. Munroe, Cristina Barlassina, Jan A. Staessen, Jacques S. Beckmann, Daniele Cusi

Abstract—Essential hypertension is a multifactorial disorder and is the main risk factor for renal and cardiovascular complications. The research on the genetics of hypertension has been frustrated by the small predictive value of the discovered genetic variants. The HYPERGENES Project investigated associations between genetic variants and essential hypertension pursuing a 2-stage study by recruiting cases and controls from extensively characterized cohorts recruited over many years

Received September 15, 2011; first decision October 6, 2011; revision accepted November 21, 2011.

From the Department of Medicine, Surgery, and Dentistry (E.S., F.F., A.C., S.L., C.B., D.C.), Graduate School of Nephrology, University of Milano, Division of Nephrology, San Paolo Hospital, Milano, Italy; Filarete Foundation (E.S., F.F., A.C., S.L., C.B., D.C.), Genomic and Bioinformatics Unit, Milano, Italy; Department of Medical Genetics (Z.K., P.B., C.R., J.S.B.), University of Lausanne, Lausanne, Switzerland; Swiss Institute of Bioinformatics (Z.K.), Lausanne, Switzerland; Hypertension and Related Diseases Centre-Azienda Ospedaliero-Universitaria (AOU) (N.G., R.Z., C.T., G.A., M.F.O.), University of Sassari, Sassari, Italy; Studies Coordinating Centre (T.K., L.T., J.A.S.), Division of Hypertension and Cardiovascular Rehabilitation, Department of Cardiovascular Diseases, University of Leuven, Leuven, Belgium; George Institute for Global Health (H.A., J.C.), University of Sydney and the Royal Prince Alfred Hospital, Sydney, New South Wales, Australia; Department of Epidemiology and Biostatistics (C.H., P.V.), School of Public Health, Imperial College of London, London, United Kingdom; Institut inter Régional pour la Santé (J.T.), Tours, France; Institute of Internal Medicine (Y.P.N., G.S.), Siberian Branch of the Russian Academy of Medical Sciences, Novosibirsk, Russian; IMS (C.C., M.C.D.), Milano, Italy; 2nd Department of Internal Medicine (J.S.), Charles University, Medical Faculty, Pilsen, Czech Republic; Department of Clinical and Experimental Medicine (V.T.), University of Padova, Padova, Italy; First Department of Cardiology and Hypertension (K.S.-S.), Jagiellonian University Medical College, Krakow, Poland; Clinical Pharmacology and Genome Centre (T.J., M.J.C., P.B.M.), William Harvey Research Institute, Barts and London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom; Institut National de la Santé et de la Recherche Médicale (N.D., X.J.), UMRS-970, Paris Cardiovascular Research Center (PARCC), Paris France; Chair of Nephrology (L.Z., P.M.), Università Vita Salute San Raffaele, Nephrology, Dialysis and Hypertension Unit, San Raffaele Scientific Institute, Milan, Italy; Human Genetics Foundation (S.G., P.V.), Turin, Italy; Department of Medicine (B.S.), University of Catania, Catania, Italy; KOS Genetic (F.R., D.G., M.M.), Milano, Italy; Department of Genetics, Biology and Biochemistry (F.R., G.M.), University of Torino and Human Genetics Foundation, Torino, Italy; Genomics Platform (P.D.), National Center of Competence in Research (NCCR) "Frontiers in Genetics" University Medical Center (CMU) University of Geneva, Geneva, Switzerland; Center of Transfusion Medicine and Immunohematology (M.M.), Department of Regenerative Medicine, Foundation Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Ca' Granda Ospedale Maggiore Policlinico, Milano, Italy; Department of Physiology (S.H.), University of Melbourne, Melbourne, Victoria, Australia; 2nd Medical Department (J.F.), Cardiology and Angiology, 1st Medical Faculty, Charles University, Prague, Czech Republic; Institute of Social and Preventive Medicine (M.B.), Centre Hospitalier Universitaire Vaudois and University of Lausanne, Lausanne, Switzerland; Research Laboratories "John Paul II" Centre for High Technology Research and Education in Biomedical Sciences (L.I.), Catholic University, Campobasso, Italy; Murdoch Childrens Research Institute (J.E.), Department of Physiology, University of Melbourne, Melbourne, Victoria, Australia; Molecular and Cellular Therapeutics (A.V.S.), Royal College of Surgeons in Ireland, Dublin, Ireland; Institute of Molecular and Cell Biology (M.L.), University of Tartu, Tartu, Estonia; British Heart Foundation (BHF) Glasgow Cardiovascular Research Centre (A.F.D., S.P., N.J.S.), University of Glasgow, Glasgow, United Kingdom; Department of Cardiovascular Sciences (N.J.S.), University of Leicester, Leicester, United Kingdom; Hypertension and Cardiovascular Disease (O.M.), Department of Clinical Sciences, Lund University, Malmö, Sweden; Centre of Emergency Medicine (O.M.), Skåne University Hospital, Malmö, Sweden; University Paris Descartes (X.J.), Paris, France; Assistance Publique Hopitaux de Paris (APHP) (X.J.), Department of Genetics, Hopital Europeen Georges Pompidou, Paris, France; IBM Haifa Research Lab (A.S.), Haifa University Mount Carmel, Haifa, Israel; University of Warwick (F.P.C.), Warwick Medical School, Coventry, United Kingdom; Genetic Epidemiology Unit (J.A.S.), Department of Epidemiology, Maastricht University, Maastricht, The Netherlands; Service of Medical Genetics (J.S.B.), Centre Hospitalier Universitaire Vaudois, University Hospital, Lausanne, Switzerland.

This paper was sent to Friedrich C. Luft, associate editor, for review by expert referees, editorial decision, and final disposition.

Correspondence to Daniele Cusi, Department of Medicine, Surgery, and Dentistry, University of Milano, Division of Nephrology, San Paolo Hospital, Milano, Viale Ortles 22/4, 20139 Milano, Italy. E-mail daniele.cusi@unimi.it

Hypertension **is available at http://hyper.ahajournals.org** **DOI: 10.1161/HYPERTENSIONAHA.111.181990**

in different European regions. The discovery phase consisted of 1865 cases and 1750 controls genotyped with 1M Illumina array. Best hits were followed up in a validation panel of 1385 cases and 1246 controls that were genotyped with a custom array of 14 055 markers. We identified a new hypertension susceptibility locus (rs3918226) in the promoter region of the endothelial NO synthase gene (odds ratio: 1.54 [95% CI: 1.37–1.73]; combined $P=2.58 \cdot 10^{-13}$). A meta-analysis, using other in silico/de novo genotyping data for a total of 21 714 subjects, resulted in an overall odds ratio of 1.34 (95% CI: 1.25–1.44; $P=1.032 \cdot 10^{-14}$). The quantitative analysis on a population-based sample revealed an effect size of 1.91 (95% CI: 0.16–3.66) for systolic and 1.40 (95% CI: 0.25–2.55) for diastolic blood pressure. We identified in silico a potential binding site for ETS transcription factors directly next to *rs3918226,* suggesting a potential modulation of endothelial NO synthase expression. Biological evidence links endothelial NO synthase with hypertension, because it is a critical mediator of cardiovascular homeostasis and blood pressure control via vascular tone regulation. This finding supports the hypothesis that there may be a causal genetic variation at this locus. (***Hypertension*. 2012;59:248-255.) • Online Data Supplement**

Key Words: genetic epidemiology ■ risk factors ■ genetics association studies ■ NO ■ essential hypertension

Essential hypertension (EH) is a clinical condition affecting a large proportion (25% to 30%) of the adult population and is a major risk factor for cardiovascular and renal diseases.[1,2] It is a complex trait influenced by multiple susceptibility genes, environmental, and lifestyle factors and their interactions.[3] In the last years, huge efforts have been performed in recruiting and genotyping tens of thousands of individuals and meta-analyzing dozens of cross-sectional, population-based studies. In spite of this, the research on the genetics of EH has been frustrated by the small predictive value of the discovered genetic variants and by the fact that these variants explain a small proportion of the phenotypic variation.[4–13] EH is a late-onset disease and, therefore, the small discovered effect sizes could in part be because of the effect of misclassification, sample selection bias, and inappropriate phenotyping of cases and controls.[9,14,15] The selection of cases and controls may have important effects on the results, because misclassification bias can lead to loss of power. For common traits, such as EH, this bias can be remedied by defining more stringent selection criteria, by recruiting hypernormal controls and adopting a more stringent case definition.[14,15]

The HYPERGENES Project pursued a 2-stage study to investigate novel genetic determinants of EH. Cases and controls were recruited from extensively characterized cohorts over many years in different European regions using standardized clinical ascertainment. Particular care was devoted to control selection. A large proportion of the sample has been followed for 5 to 10 years after DNA collection, allowing for the exclusion of controls that developed hypertension at a later age, thereby defining the hypernormal controls.

Methods

Study Population

Cases and controls were recruited from extensively characterized cohorts using standardized clinical ascertainment, collected over many years in different European regions (balanced within North Europe, continental Italy, and Sardinia). The inclusion criteria are described in the Methods (S1) section of the online Data Supplement (available at http://hyper.ahajournals.org). To perform a genetic association with continuous blood pressure (BP) phenotypes, we considered 2 additional cohorts (FLEMENGHO-EPOGH, n=1514, and Wandsworth Heart & Stroke Study, n=306, see Methods [S2] of the online Data Supplement) that provided population-based data. Description of the different samples is reported in the Methods S2 section.

Genotyping and Imputation

Genotyping details are shown in Methods S3 through S6 of the online Data Supplement. Briefly, in the discovery phase, the samples were genotyped using the Illumina 1M-Duo array, and the imputation was performed with MACH[16] using as reference the 1000 Genomes haplotypes (release June 2010; Method S3). To validate and fine map the genes found associated with EH in discovery phase, an Illumina custom chip of 14 055 markers was created starting from the list of best-associated and of candidate single nucleotide polymorphisms (SNPs) based on a priori biological knowledge (Methods S4 and S5). For the replication stage, we used the in silico data of rs3918226 from Anglo-Scandinavian Cardiac Outcomes Trial/AIBIII/NBS, BRIGHT, EPIC Turin, HYPEST, and NORDIL/MDC studies (Methods S6).

Statistical Analysis

All of the quality controls and statistical analyses were performed in accordance with the protocols written by Anderson et al[17] and Clarke et al[18] (Methods S7 through S9). We tested each SNP for association with hypertension using a logistic regression under an additive model with adjustment for sex and for the first 10 principal components. Combined analysis for discovery, validation, and replication results was conducted using METAL.[19] The quantitative effect of rs3918226 on systolic BP and diastolic BP was tested on 2 additional population-based cohorts (Methods S2). Moreover, we tested for multiplicative interaction between rs3918226 and the most plausible interactive partners of the endothelial NO synthase (eNOS) gene, actin genes and heat shock protein (HSP) 90 genes (Methods S9). The quantitative effect of rs3918226 on systolic BP and diastolic BP has been tested on 2 additional population-based cohorts (FLEMENGHO-EPOGH and Wandsworth Heart & Stroke Study, see Methods S2). The recognition sequences for transcription factors in the eNOS region were searched using TRANSFAC[20,21] and the TFSEARCH database[22] (Methods S10).

Results

A classic 2-stage case-control strategy was used with a discovery phase of 1865 cases and 1750 controls (2294 males and 1321 females), all genotyped on the Illumina 1M Duo chip. The sample consisted of an ethnically diverse population (25.06% North Europeans, 38.70% Sardinians, and 36.24% continental Italy subjects). The discovery phase was followed by a validation phase of an additional 1385 cases and 1246 controls (1417 males and 1214 females). According to ethnicity, the validation sample was composed of 1262 North Europeans (47.97%), 788 Sardinians (29.95%), and 581 continental Italians (22.08%). Tables S1 and S2 (available in the online Data Supplement) show the demographic characteristics and baseline measures.

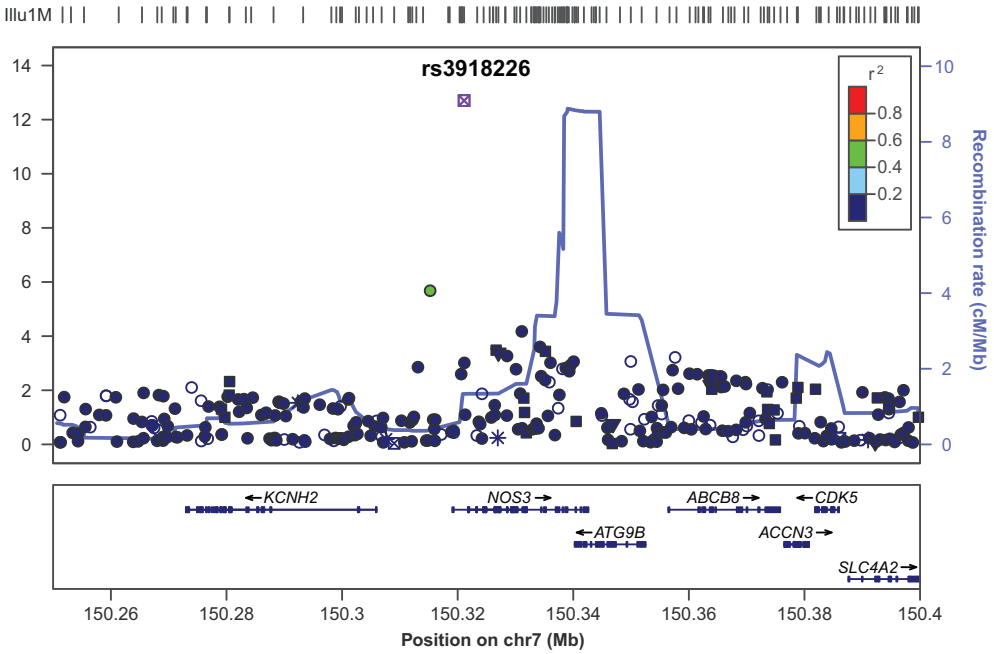


Figure 1. Local Manhattan plot for the NOS3 (endothelial NO synthase) region. Each circle represents a single nucleotide polymorphism (SNP), its *y* coordinate is the −log10 association *P* value for hypertension, and the *x* coordinate represents the physical position on the chromosome (on build 36, hg18). When replication data were available, the combined *P* value was used, otherwise the discovery *P* value was used. Circles are filled with colors according to the linkage disequilibrium (LD; r^2) between the given SNP and the lead SNP (rs3918266, violet square). Blue line indicates the recombination rate. The second best hit with *P* value 2.46E-6 in the discovery stage (named chr7:150 314 954 according to the 1000 Genome Project) was imputed based on the 1000 Genomes haplotypes (release June 2010), and its imputation quality was very high (r^2-hat=0.94). In validation stage, the imputation quality was very low (r^2-hat=0.17).

Principal component analysis of the genotype data were carried out to find the major axes of variation used as covariates to correct for population stratification.[23] The discovery samples in the principal component map showed 3 (roughly) equal-sized distinct clusters corresponding with the 3 main ethnic groups, as expected from the study design (Figure S1). All of the association analyses were adjusted for the ancestry principal components and sex by including them as covariates in the logistic regression model. In addition, genomic control correction was applied (because genomic inflation factor was 1.04). In the discovery phase, 90 SNPs (57% intragenic) with *P* value $<1 \cdot 10^{-4}$ were identified after genomic control (Figure S2 and Table S4). The most promising SNPs were genotyped in the validation samples using an Illumina Infinium Custom chip. The meta-analysis of the discovery and validation data revealed SNP rs3918226 to be associated with EH in whites, reaching a $P_{combined}$ of $2.58 \cdot 10^{-13}$ and odds ratio (OR) of 1.54 per T allele (95% CI: 1.37–1.73) under an additive model (Figure 1 and Table 1 and Figure S4). Estimated ORs in the discovery and validation

Table 1. Meta-Analysis Results for the Top SNPs in the HYPERGENES Study

Marker Name	Chr	Position	Effect/ Other Allele	Gene	OR Discovery	*P* Discovery	OR Validation	*P* Validation	OR Combined	CI Combined	Inverse Variance-Weighted *P* Combined	*Z* Score *P* Combined
rs3918226	7	150321109	T/C	NOS3	1.425	4.81E-06	1.71	2.55E-09	1.538	1.372–1.726	1.98E-13	2.58E-13
rs341408	15	58928982	G/A	RORA	0.786	1.74E-06	0.956	4.29E-01	0.856	0.79–0.92	3.98E-05	2.79E-05
rs4976593	5	167710021	G/A	WWC1	1.27	3.75E-06	1.045	4.60E-01	1.169	1.08–1.26	6.64E-05	5.29E-05
rs631208	16	9307225	G/A	RP11-473I1.1	0.798	8.09E-06	0.951	3.84E-01	0.862	0.80–0.93	8.89E-05	6.36E-05
rs7907270	10	78550949	G/A	KCNMA1	1.27	2.35E-06	0.989	8.53E-01	1.141	1.06–1.23	5.75E-04	4.25E-04
rs10519080	15	58925751	G/A	RORA	1.369	5.79E-06	0.979	7.95E-01	1.187	1.07–1.31	1.09E-03	8.49E-04
rs1406891	6	161107070	G/A	PLG	1.251	3.99E-06	0.949	3.50E-01	1.112	1.03–1.19	3.87E-03	2.97E-03
rs783182	6	161088538	G/A	PLG	0.797	2.95E-06	1.068	2.42E-01	0.902	0.84–0.97	5.31E-03	4.15E-03
rs1084656	6	161101282	C/A	PLG	1.243	6.67E-06	0.936	2.39E-01	1.103	1.03–1.18	7.66E-03	6.35E-03
rs783145	6	161072439	G/A	PLG	0.788	8.53E-07	1.102	8.45E-02	0.909	0.84–0.98	9.27E-03	6.85E-03
rs1247558	6	161110189	G/A	PLG	1.24	8.30E-06	0.932	2.14E-01	1.100	1.02–1.18	9.42E-03	7.93E-03

The table shows association results (OR and *P* values) for discovery and for validation samples and for the combined analysis (both inverse variance weighting and *z* score meta-analysis). *P* values and ORs with the associated 95% CIs have been calculated under an additive model using logistic regression adjusted for sex and principal components. To retrieve information about single nucleotide polymorphisms and their genomic context (the nearest gene) we used the hg18 (National Center for Biotechnology Information 36) assembly. OR indicates odds ratio; *P*, *P* values; CI, confidence interval; Chr, chromosome; SNP, single nucleotide polymorphism.

Table 2. In Silico Meta-Analysis Results for rs3918226 (T/C, Effect Allele/Other Allele)

Variable	Study	Sample Size	OR	SE	95% CI	*P*
HYPERGENES samples	HYPERGENES discovery	3596	1.43	0.11	1.224–1.657	4.81E-06
	HYPERGENES validation	2610	1.71	0.155	1.440–2.049	2.55E-09
	Combined analysis HYPERGENES	6206	1.54	0.038	1.372–1.726	2.58E-13
Replication samples	ASCOT/AIBIII/NBS	4049	1.06	0.092	0.895–1.256	4.97E-01
	BRIGHT	3641	1.39	0.126	1.168–1.663	2.32E-04
	EPIC Turin	2714	1.28	0.126	1.050–1.551	1.44E-02
	HYPEST	1204	1.13	0.236	0.754–1.705	5.45E-01
	NORDIL/MDC	3900	1.25	0.124	1.030–1.519	2.40E-02
	Combined Analysis of Replication Samples	15 508	1.23	0.056	1.125–1.344	6.50E-06
	Meta-analysis	21 714	1.34*	1.248–1.437†	1.032E-14‡	6.198E-16§

Top section shows association results (odds ratios, SEs, CIs, and *P* values) for discovery, validation, and combined analysis of the HYPERGENES samples. Middle section shows results for ASCOT/AIBIII/NBS, BRIGHT, Epic Turin, HYPEST, and NORDIL/MDC studies and combined analysis of replication in silico samples. Bottom section shows meta-analysis results for all of the samples using both the *z* score and inverse variance-weighted *P* value methods.

*Data are OR combined.

†Data are 95% CI combined.

‡Data are combined *P* (*z* score).

§Data are combined *P* (inverse variance weighted).

samples were consistent across the different white populations of the HYPERGENES sample (Figure S5).

The polymorphism rs3918226 maps to the promoter region of the eNOS gene (−665 C>T, NOS3).[24,25] The T-allele frequencies in the present study are 13.8% in cases and 8.9% in controls. SNP rs3918226 is monomorphic in the nonwhite HYPERGENES samples (Wandsworth Heart & Stroke Study cohort) and African and Asian HapMap samples. The second best hit chr7:150,314,954 (G/A SNP, minor allele frequency of A allele=3%) with *P* value $2.46 \cdot 10^{-6}$ and OR 2.25 was imputed based on the 1000 Genomes haplotypes (release June 2010); its imputation quality was very high (r2-hat=0.94). Unfortunately we could not replicate the observation in validation because of low imputation quality. An additional 7 SNPs within eNOS gene showed significant *P* values ($1 \cdot 10^{-3} < P < 1 \cdot 10^{-5}$): *rs2853792* (intronic, $P_{combined} = 7.76 \cdot 10^{-5}$), *rs1549758* (coding, $P_{combined} = 3.32 \cdot 10^{-4}$), *rs1800779* (intronic, $P_{combined} = 1.16 \cdot 10^{-3}$), *rs6951150* (intergenic, $P_{combined} = 1.64 \cdot 10^{-3}$), *rs743507* (intronic, $P_{combined} = 1.76 \cdot 10^{-3}$), *rs1800780* (intronic, $P_{combined} = 1.96 \cdot 10^{-3}$), and *rs1800783* (intronic, $P_{combined} = 2.89 \cdot 10^{-3}$; Figure 1).

Table 1 shows also other significant SNPs with *P* values between $1 \cdot 10^{-3}$ and $1 \cdot 10^{-5}$ mapping different genes as calcium-activated potassium channel subunit α-1 (*KCNMA1*), plasminogen (*PLG*), retinoid-related orphan receptor-α (*RORA*), and WW domain-containing protein 1 (*WWC1*). Moreover, the signals of SNPs presented previously in literature are in our study in the same direction as the original studies,[5,6,8] showing evidence of a marginally significant association in HYPERGENES (Table S5).

We meta-analyzed rs3918226 using in silico data from Anglo-Scandinavian Cardiac Outcomes Trial/AIBIII/NBS, BRIGHT, EPIC-Turin, HYPEST, and NORDIL/MDC samples (Methods S2 and S6), resulting in an overall OR of 1.34 per T allele (95% CI: 1.25–1.44; $P_{combined} = 1.032 \cdot 10^{-14}$; Table 2 and Figure 2) for a total of 21 714 subjects. Because case and control definitions differed between HYPERGENES and the in silico replication samples, the ORs are not directly comparable. In our study, the *P* value of heterogeneity calculated for HYPERGENES samples is 0.13. It decreased slightly but remained nonsignificant, as expected, when EPIC-Turin was also considered together in the meta-analysis

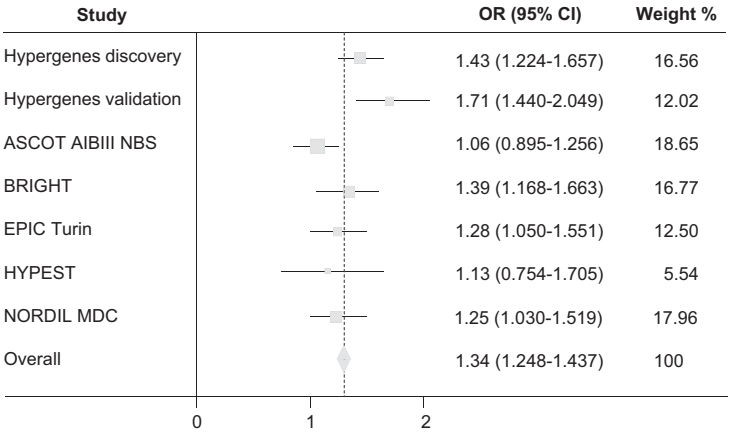


Figure 2. Forest plot of meta-analysis between HYPERGENES Discovery, HYPERGENES Validation, ASCOT/AIBIII/NBS, BRIGHT, EPIC Turin, HYPEST, and NORDIL/MDC studies. The squares and the horizontal lines correspond with the odds ratio (OR) and 95% CI of each study; the size of squares is proportional to weights (also shown as percentage); and the dotted line and the diamond represent the overall combined OR and 95% CI.

($P=0.092$), because the recruitment criteria for cases and controls were identical. Conversely, the heterogeneity increased significantly ($P=0.005$) when HYPERGENES samples were meta-analyzed with all of the other samples (Anglo-Scandinavian Cardiac Outcomes Trial/AIBIII/NBS, BRIGHT, HYPEST, and NORDIL/MDC).

Moreover, we tested for epistatic multiplicative interactions between *eNOS rs3918226* and all of the available polymorphisms in genes known to be involved in targeting and regulating the overall availability of eNOS at the cell membrane[26–28]: actin genes (ACTA1, ACTA2, ACTB, ACTG1, and ACTG2)[29,30] and HSP90 genes (HSP90AA1, HSP90AA2, and HSP90AB1).[26] Nominally significant interactions were observed between *rs3918226* and *rs13447427* ($P=1.34\cdot10^{-3}$) in actin-β gene (*ACTB*), *rs7503750* ($P=1.57\cdot10^{-3}$) in actin-γ1 (*ACTG1*), and *rs4922796* and *rs17309979* ($P=3.47\cdot10^{-3}$, $P=4.88\cdot10^{-3}$) in HSP-α2 (*HSP90AA2*; Table S6). When controlling for multiple testing, these interactions remained significant at a false discovery rate of 20%.

The quantitative analysis confirmed the qualitative observation. In fact, the β coefficient of the regression between systolic BP or diastolic BP with rs3918226 is, respectively, 1.91 (95% CI: 0.16–3.66) and 1.40 (95% CI: 0.25–2.55) per T allele. The coefficient is the effect size on BP in millimeter of mercury per coded allele based on an additive genetic model. The BP distribution according to rs3918226 genotype is shown in Table S7.

Because *rs3918226* maps to the promoter region of *eNOS*, we tested whether it may fall into a regulatory binding site. Using the PATCH algorithm of TRANSFAC database,[21] we characterized a putative binding site for transcription factors of the ETS family directly next to *rs3918226*. The ETS family members are present in endothelial cells and participated in activation of the eNOS promoter.[31] Using the TFSEARCH tool,[22] we confirmed this finding with a score of 87.3.

We also tested the degree of evolutionary conservation of rs3918226 locus in primates and placental mammals using the conservation track of the University of California, Santa Cruz genome browser. Figure S6 shows that the region in which rs3918226 lies is conserved from placental mammals to primates.

Discussion

EH is a complex clinical condition representing the main risk factor responsible for renal and cardiovascular complications. The HYPERGENES Project investigated undiscovered associations between genetic variants and EH pursuing a 2-stage study by recruiting cases and controls from extensively characterized cohorts recruited in different European regions.

We discovered rs3918226 in the promoter region of the eNOS gene to be significantly associated with hypertension (OR: 1.54 [95% CI: 1.37–1.73]; $P=2.58\cdot10^{-13}$). The result was confirmed by meta-analyzing in silico data for a total of 21714 subjects (OR: 1.34 [95% CI: 1.25–1.44]; $P=1.032\cdot10^{-14}$). We observed heterogeneity in the findings of meta-analysis ($P=0.005$ for Q test of heterogeneity) that could be attributed to both different sample sizes and recruitment criteria not directly comparable between HYPERGENES and the in silico replication samples (Figure 2).

The quantitative effect of rs3918226 was also estimated in continuous BP phenotypes, resulting in a β-coefficient of 1.91 for systolic BP and 1.40 for diastolic BP, despite the low P values of the regression probably because of the low sample size. This finding reinforces the observation on the qualitative phenotype.

We identified a potential transcription factor binding site for the ETS family domain directly next to rs3918226. The members of ETS family, as ETS-1 and ELF-1, are essential factors for the activation of the eNOS promoter.[31] This suggests that, by affecting transcription factor–binding affinity, *rs3918226* might modulate the transcription of the *eNOS gene*.

It is also worth noting that the region in which *rs3918226* lies is conserved from placental mammals to primates. We propose rs3918226 as a novel susceptibility SNP, because among the genomewide association studies so far published, this is the first that points to eNOS: the novelty of the rs39118226 finding is that the association between eNOS and hypertension has been found in whites using a genomewide association study approach.

The use of the Illumina 1M array and Human CVD BeadArray was crucial in detecting the association, because *rs3918226* is not present on other commercial arrays.[32] Other than being poorly covered by other genotyping platforms, the region has a relatively high recombination rate toward the coding region (Figure 1). This has resulted in low linkage disequilibrium with markers present on older platforms (eg, rsq-hat <0.2 for Affy500K platform). These facts largely limited the potential to replicate our finding using data from other genomewide association studies samples, almost all of which relied on older platforms.

Indeed, eNOS has been found inconsistently associated with hypertension with several underpowered candidate gene studies, many of which only focused on a few variants with relatively small numbers of cases and controls compared with the large sample sizes of genomewide association studies. Positive studies were substantially on Asian cohorts,[33–35] whereas the majority were negative in whites, as summarized in a recent meta-analysis.[36] The polymorphisms studied in our white sample G894T (rs1799983) and T-786C (rs2070744) did not reach genomewide significant association with hypertension. If looked with candidate gene threshold, the P value and the sample size of the present study by far outnumber all of the other published so far. rs1799983 was associated with EH with a P value of 2.63×10^{-3} (OR=1.038) and rs2070744 with a P value of 6.42×10^{-4} (OR=1.04), as shown in Table S8. To summarize, the ORs are clinically irrelevant. We underline the low linkage disequilibrium between rs3918226 and rs1799983 ($R^2=0.16$) and rs2070744 ($R^2=0.17$), suggesting that these 2 SNPs are independent from rs3918226 and do not have any additional effect on the phenotype.

There is considerable biological evidence linking eNOS with hypertension and hypertension-associated cardiovascular target organ damage.[37] eNOS, which catalyzes the synthesis of NO by vascular endothelium, is responsible for the vasodilator tone that is fundamental for the regulation of BP. Furthermore, eNOS is a critical mediator of cardiovascular homeostasis through regulation of blood vessels diameter and of the maintenance of an antiproliferative and antiapoptotic environment.

Because NO is highly active, it cannot be stored inside producing cells. Indeed, eNOS signaling capacity must be controlled, at least in part, by regulating its targeting from Golgi apparatus to plasma membrane by its compartmentalization within the plasma membrane and by its later internalization from the plasma membrane to the cytoplasm. eNOS is a dually acylated peripheral membrane protein that is targeted to endothelial plasmalemmal caveolae through an interaction with the caveolae structural protein, caveolin 1 (*Cav1*).[26,27] Cav1 inhibition of eNOS is lessened by calmodulin (*Calm*) causing dissociation of eNOS from caveolin. This regulatory mechanism is further altered by HSP90,[27] which binds to eNOS and facilitates displacement of Cav1 by Calm. Moreover, eNOS directly interacts with actin cytoskeleton.[29] Recently, Kondrikov et al[30] added that β-actin is associated with the eNOS oxygenase domain increasing eNOS activity and NO production. To explore such a pathway, we tested the interaction between the discovered eNOS SNP and its most plausible interactive partners. We observed nominally significant interactions between *rs3918226* and *rs13447427* in *ACTB*, *rs7503750* in *ACTG1*, and *rs4922796* and *rs17309979* in the *HSP90AA2* gene.

In conclusion, with a stringent case-control design and a population-based study, we identified a novel hypertension susceptibility locus in the promoter region of *eNOS* with a relatively high effect size. Our finding could provide new insights into the mechanism of vascular regulation and could help in better understanding the genetics of EH. Furthermore, we believe that this indication can be useful to guide fine mapping or sequencing efforts to single out causal variants.

Perspectives

Further investigations and high-throughput sequencing of region of interest will help to identify the real causal variant and to clarify the functional role of eNOS in EH.

Acknowledgments

The complete list is reported in supplemental material.

Sources of Funding

This work was supported by the HYPERGENES project (European Network for Genetic-Epidemiological Studies: building a method to dissect complex genetic traits, using essential hypertension as a disease model), grant HEALTH-2007-201550, funded by the European Union within the FP7. T.J. was supported by the Wellcome Trust (grant 093078/Z/10/Z). J.C. has received research grants from Servier International and from the National Health and Medical Research Council (Australia), administered through the University of Sydney, for the Perindopril Protection Against Recurrent Stroke Study and ADVANCE Trials and for the ADVANCE posttrial follow-up study. M.J.C. has received British Heart Foundation grant support for developing CVD bead array and KASPAr genotyping.

Disclosures

T.J. received honoraria for speaking about these studies at scientific meetings. M.J.C. was supported by British Heart Foundation grant for developing CVD BeadArray and KASPAr assay. J.C. received honoraria for speaking about the PROGRESS and ADVANCE Trials and for the ADVANCE-Post Trial follow-up study.

References

1. Kearney PM, Whelton M, Reynolds K, Muntner P, Whelton PK, He J. Global burden of hypertension: analysis of worldwide data. *Lancet*. 2005; 365:217–223.
2. Lawes CM, Vander Hoorn S, Rodgers A, for the International Society of Hypertension. Global burden of blood-pressure-related disease, 2001. *Lancet*. 2008;371:1513–1518.
3. Kunes J, Zicha J. The interaction of genetic and environmental factors in the etiology of hypertension. *Physiol Res*. 2009;58:S33–S41.
4. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447:661–678.
5. Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, Coin L, Najjar SS, Zhao JH, Heath SC, Eyheramendy S, Papadakis K, Voight BF, Scott LJ, Zhang F, Farrall M, Tanaka T, Wallace C, Chambers JC, Khaw KT, Nilsson P, van der Harst P, Polidoro S, Grobbee DE, Onland-Moret NC, Bots ML, Wain LV, Elliott KS, Teumer A, Luan J, Lucas G, Kuusisto J, Burton PR, Hadley D, McArdle WL; Wellcome Trust Case Control Consortium, Brown M, Dominiczak A, Newhouse SJ, Samani NJ, Webster J, Zeggini E, Beckmann JS, Bergmann S, Lim N, Song K, Vollenweider P, Waeber G, Waterworth DM, Yuan X, Groop L, Orho-Melander M, Allione A, Di Gregorio A, Guarrera S, Panico S, Ricceri F, Romanazzi V, Sacerdote C, Vineis P, Barroso I, Sandhu MS, Luben RN, Crawford GJ, Jousilahti P, Perola M, Boehnke M, Bonnycastle LL, Collins FS, Jackson AU, Mohlke KL, Stringham HM, Valle TT, Willer CJ, Bergman RN, Morken MA, Döring A, Gieger C, Illig T, Meitinger T, Org E, Pfeufer A, Wichmann HE, Kathiresan S, Marrugat J, O'Donnell CJ, Schwartz SM, Siscovick DS, Subirana I, Freimer NB, Hartikainen AL, McCarthy MI, O'Reilly PF, Peltonen L, Pouta A, de Jong PE, Snieder H, van Gilst WH, Clarke R, Goel A, Hamsten A, Peden JF, Seedorf U, Syvänen AC, Tognoni G, Lakatta EG, Sanna S, Scheet P, Schlessinger D, Scuteri A, Dörr M, Ernst F, Felix SB, Homuth G, Lorbeer R, Reffelmann T, Rettig R, Völker U, Galan P, Gut IG, Hercberg S, Lathrop GM, Zelenika D, Deloukas P, Soranzo N, Williams FM, Zhai G, Salomaa V, Laakso M, Elosua R, Forouhi NG, Völzke H, Uiterwaal CS, van der Schouw YT, Numans ME, Matullo G, Navis G, Berglund G, Bingham SA, Kooner JS, Connell JM, Bandinelli S, Ferrucci L, Watkins H, Spector TD, Tuomilehto J, Altshuler D, Strachan DP, Laan M, Meneton P, Wareham NJ, Uda M, Jarvelin MR, Mooser V, Melander O, Loos RJ, Elliott P, Abecasis GR, Caulfield M, Munroe PB. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet*. 2009;41:666–676.
6. Levy D, Ehret GB, Rice K, Verwoert GC, Launer LJ, Dehghan A, Glazer NL, Morrison AC, Johnson AD, Aspelund T, Aulchenko Y, Lumley T, Köttgen A, Vasan RS, Rivadeneira F, Eiriksdottir G, Guo X, Arking DE, Mitchell GF, Mattace-Raso FU, Smith AV, Taylor K, Scharpf RB, Hwang SJ, Sijbrands EJ, Bis J, Harris TB, Ganesh SK, O'Donnell CJ, Hofman A, Rotter JI, Coresh J, Benjamin EJ, Uitterlinden AG, Heiss G, Fox CS, Witteman JC, Boerwinkle E, Wang TJ, Gudnason V, Larson MG, Chakravarti A, Psaty BM, van Duijn CM. Genome-wide association study of blood pressure and hypertension. *Nat Genet*. 2009;41:677–687.
7. Munroe PB, Johnson T, Caulfield M. The genetic architecture of blood pressure variation. *Curr Cardiovasc Risk Rep*. 2009;3:418–425.
8. Padmanabhan S, Melander O, Johnson T, Di Blasio AM, Lee WK, Gentilini D, Hastie CE, Menni C, Monti MC, Delles C, Laing S, Corso B, Navis G, Kwakernaak AJ, van der Harst P, Bochud M, Maillard M, Burnier M, Hedner T, Kjeldsen S, Wahlstrand B, Sjögren M, Fava C, Montagnana M, Danese E, Torffvit O, Hedblad B, Snieder H, Connell JM, Brown M, Samani NJ, Farrall M, Cesana G, Mancia G, Signorini S, Grassi G, Eyheramendy S, Wichmann HE, Laan M, Strachan DP, Sever P, Shields DC, Stanton A, Vollenweider P, Teumer A, Völzke H, Rettig R, Newton-Cheh C, Arora P, Zhang F, Soranzo N, Spector TD, Lucas G, Kathiresan S, Siscovick DS, Luan J, Loos RJ, Wareham NJ, Penninx BW, Nolte IM, McBride M, Miller WH, Nicklin SA, Baker AH, Graham D, McDonald RA, Pell JP, Sattar N, Welsh P; Global BPgen Consortium, Munroe P, Caulfield MJ, Zanchetti A, Dominiczak AF. Genome-wide association study of blood pressure extremes identifies variant near UMOD associated with hypertension. *PLoS Genet*. 2010;6:e1001177.
9. Ehret GB. Genome-wide association studies: contribution of genomics to understanding blood pressure and essential hypertension. *Curr Hypertens Rep*. 2010;12:17–25.
10. Hong KW, Jin HS, Lim JE, Kim S, Go MJ, Oh B. Recapitulation of two genome-wide association studies on blood pressure and essential hypertension in the Korean population. *J Hum Genet*. 2010;55:336–341.
11. Fox ER, Young JH, Li Y, Dreisbach AW, Keating BJ, Musani SK, Liu K, Morrison AC, Ganesh S, Kutlar A, Ramachandran VS, Polak JF, Fabsitz RR, Dries DL, Farlow DN, Redline S, Adeyemo A, Hirschorn JN, Sun YV, Wyatt SB, Penman AD, Palmas W, Rotter JI, Townsend RR, Doumatey AP, Tayo BO, Mosley TH Jr, Lyon HN, Kang SJ, Rotimi CN,

Cooper RS, Franceschini N, Curb JD, Martin LW, Eaton CB, Kardia SL, Taylor HA, Caulfield MJ, Ehret GB, Johnson T, International Consortium for Blood Pressure Genome-wide Association Studies (ICBP-GWAS), Chakravarti A, Zhu X, Levy D, Munroe PB, Rice KM, Bochud M, Johnson AD, Chasman DI, Smith AV, Tobin MD, Verwoert GC, Hwang SJ, Pihur V, Vollenweider P, O'Reilly PF, Amin N, Bragg-Gresham JL, Teumer A, Glazer NL, Launer L, Zhao JH, Aulchenko Y, Heath S, Sõber S, Parsa A, Luan J, Arora P, Dehghan A, Zhang F, Lucas G, Hicks AA, Jackson AU, Peden JF, Tanaka T, Wild SH, Rudan I, Igl W, Milaneschi Y, Parker AN, Fava C, Chambers JC, Kumari M, Go MJ, van der Harst P, Kao WH, Sjögren M, Vinay DG, Alexander M, Tabara Y, Shaw-Hawkins S, Whincup PH, Liu Y, Shi G, Kuusisto J, Seielstad M, Sim X, Nguyen KD, Lehtimäki T, Matullo G, Wu Y, Gaunt TR, Onland-Moret NC, Cooper MN, Platou CG, Org E, Hardy R, Dahgam S, Palmen J, Vitart V, Braund PS, Kuznetsova T, Uiterwaal CS, Campbell H, Ludwig B, Tomaszewski M, Tzoulaki I, Palmer ND; CARDIoGRAM consortium; CKDGen consortium; KidneyGen consortium; EchoGen consortium; CHARGE-HF consortium, Aspelund T, Garcia M, Chang YP, O'Connell JR, Steinle NI, Grobbee DE, Arking DE, Hernandez D, Najjar S, McArdle WL, Hadley D, Brown MJ, Connell JM, Hingorani AD, Day IN, Lawlor DA, Beilby JP, Lawrence RW, Clarke R, Collins R, Hopewell JC, Ongen H, Bis JC, Kähönen M, Viikari J, Adair LS, Lee NR, Chen MH, Olden M, Pattaro C, Hoffman Bolton JA, Köttgen A, Bergmann S, Mooser V, Chaturvedi N, Frayling TM, Islam M, Jafar TH, Erdmann J, Kulkarni SR, Bornstein SR, Grässler J, Groop L, Voight BF, Kettunen J, Howard P, Taylor A, Guarrera S, Ricceri F, Emilsson V, Plump A, Barroso I, Khaw KT, Weder AB, Hunt SC, Bergman RN, Collins FS, Bonnycastle LL, Scott LJ, Stringham HM, Peltonen L, Perola M, Vartiainen E, Brand SM, Staessen JA, Wang TJ, Burton PR, Soler Artigas M, Dong Y, Snieder H, Wang X, Zhu H, Lohman KK, Rudock ME, Heckbert SR, Smith NL, Wiggins KL, Shriner D, Veldre G, Viigimaa M, Kinra S, Prabhakaran D, Tripathy V, Langefeld CD, Rosengren A, Thelle DS, Corsi AM, Singleton A, Forrester T, Hilton G, McKenzie CA, Salako T, Iwai N, Kita Y, Ogihara T, Ohkubo T, Okamura T, Ueshima H, Umemura S, Eyheramendy S, Meitinger T, Wichmann HE, Cho YS, Kim HL, Lee JY, Scott J, Sehmi JS, Zhang W, Hedblad B, Nilsson P, Smith GD, Wong A, Narisu N, Stančáková A, Raffel LJ, Yao J, Kathiresan S, O'Donnell C, Schwartz SM, Ikram MA, Longstreth WT Jr, Seshadri S, Shrine NR, Wain LV, Morken MA, Swift AJ, Laitinen J, Prokopenko I, Zitting P, Cooper JA, Humphries SE, Danesh J, Rasheed A, Goel A, Hamsten A, Watkins H, Bakker SJ, van Gilst WH, Janipalli C, Mani KR, Yajnik CS, Hofman A, Mattace-Raso FU, Oostra BA, Demirkan A, Isaacs A, Rivadeneira F, Lakatta EG, Orru M, Scuteri A, Ala-Korpela M, Kangas AJ, Lyytikäinen LP, Soininen P, Tukiainen T, Würz P, Ong RT, Dörr M, Kroemer HK, Völker U, Völzke H, Galan P, Hercberg S, Lathrop M, Zelenika D, Deloukas P, Mangino M, Spector TD, Zhai G, Meschia JF, Nalls MA, Sharma P, Terzic J, Kumar MJ, Denniff M, Zukowska-Szczechowska E, Wagenknecht LE, Fowkes FG, Charchar FJ, Schwarz PE, Hayward C, Guo X, Bots ML, Brand E, Samani N, Polasek O, Talmud PJ, Nyberg F, Kuh D, Laan M, Hveem K, Palmer LJ, van der Schouw YT, Casas JP, Mohlke KL, Vineis P, Raitakari O, Wong TY, Tai ES, Laakso M, Rao DC, Harris TB, Morris RW, Dominiczak AF, Kivimaki M, Marmot MG, Miki T, Saleheen D, Chandak GR, Coresh J, Navis G, Salomaa V, Han BG, Kooner JS, Melander O, Ridker PM, Bandinelli S, Gyllensten UB, Wright AF, Wilson JF, Ferrucci L, Farrall M, Tuomilehto J, Pramstaller PP, Elosua R, Soranzo N, Sijbrands EJ, Altshuler D, Loos RJ, Shuldiner AR, Gieger C, Meneton P, Uitterlinden AG, Wareham NJ, Gudnason V, Rettig R, Uda M, Strachan DP, Witteman JC, Hartikainen AL, Beckmann JS, Boerwinkle E, Boehnke M, Larson MG, Järvelin MR, Psaty BM, Abecasis GR, Elliott P, van Duijn CM, Newton-Cheh C. Association of genetic variation with systolic and diastolic blood pressure among African Americans: the Candidate Gene Association Resource Study. *Hum Mol Genet*. 2011;20:2273–2284.

12. Lettre G, Palmer CD, Young T, Ejebe KG, Allayee H. Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project. *PLoS Genet*. 2011;7:e1001300.
13. International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, Chasman DI, Smith AV, Tobin MD, Verwoert GC, Hwang SJ, Pihur V, Vollenweider P, O'Reilly PF, Amin N, Bragg-Gresham JL, Teumer A, Glazer NL, Launer L, Zhao JH, Aulchenko Y, Heath S, Sõber S, Parsa A, Luan J, Arora P, Dehghan A, Zhang F, Lucas G, Hicks AA, Jackson AU, Peden JF, Tanaka T, Wild SH, Rudan I, Igl W, Milaneschi Y, Parker AN, Fava C, Chambers JC, Fox ER, Kumari M, Go MJ, van der Harst P, Kao WH, Sjögren M, Vinay DG, Alexander M, Tabara Y, Shaw-Hawkins S, Whincup PH, Liu Y, Shi G, Kuusisto J, Tayo B, Seielstad M, Sim X, Nguyen KD, Lehtimäki T, Matullo G, Wu Y, Gaunt TR, Onland-Moret NC, Cooper MN, Platou CG, Org E, Hardy R, Dahgam S, Palmen J, Vitart V, Braund PS, Kuznetsova T, Uiterwaal CS, Adeyemo A, Palmas W, Campbell H, Ludwig B, Tomaszewski M, Tzoulaki I, Palmer ND, CARDIoGRAM consortium, CKDGen Consortium, KidneyGen Consortium, EchoGen consortium, CHARGE-HF consortium, Aspelund T, Garcia M, Chang YP, O'Connell JR, Steinle NI, Grobbee DE, Arking DE, Kardia SL, Morrison AC, Hernandez D, Najjar S, McArdle WL, Hadley D, Brown MJ, Connell JM, Hingorani AD, Day IN, Lawlor DA, Beilby JP, Lawrence RW, Clarke R, Hopewell JC, Ongen H, Dreisbach AW, Li Y, Young JH, Bis JC, Kähönen M, Viikari J, Adair LS, Lee NR, Chen MH, Olden M, Pattaro C, Bolton JA, Köttgen A, Bergmann S, Mooser V, Chaturvedi N, Frayling TM, Islam M, Jafar TH, Erdmann J, Kulkarni SR, Bornstein SR, Grässler J, Groop L, Voight BF, Kettunen J, Howard P, Taylor A, Guarrera S, Ricceri F, Emilsson V, Plump A, Barroso I, Khaw KT, Weder AB, Hunt SC, Sun YV, Bergman RN, Collins FS, Bonnycastle LL, Scott LJ, Stringham HM, Peltonen L, Perola M, Vartiainen E, Brand SM, Staessen JA, Wang TJ, Burton PR, Artigas MS, Dong Y, Snieder H, Wang X, Zhu H, Lohman KK, Rudock ME, Heckbert SR, Smith NL, Wiggins KL, Doumatey A, Shriner D, Veldre G, Viigimaa M, Kinra S, Prabhakaran D, Tripathy V, Langefeld CD, Rosengren A, Thelle DS, Corsi AM, Singleton A, Forrester T, Hilton G, McKenzie CA, Salako T, Iwai N, Kita Y, Ogihara T, Ohkubo T, Okamura T, Ueshima H, Umemura S, Eyheramendy S, Meitinger T, Wichmann HE, Cho YS, Kim HL, Lee JY, Scott J, Sehmi JS, Zhang W, Hedblad B, Nilsson P, Smith GD, Wong A, Narisu N, Stančáková A, Raffel LJ, Yao J, Kathiresan S, O'Donnell CJ, Schwartz SM, Ikram MA, Longstreth WT Jr, Mosley TH, Seshadri S, Shrine NR, Wain LV, Morken MA, Swift AJ, Laitinen J, Prokopenko I, Zitting P, Cooper JA, Humphries SE, Danesh J, Rasheed A, Goel A, Hamsten A, Watkins H, Bakker SJ, van Gilst WH, Janipalli CS, Mani KR, Yajnik CS, Hofman A, Mattace-Raso FU, Oostra BA, Demirkan A, Isaacs A, Rivadeneira F, Lakatta EG, Orru M, Scuteri A, Ala-Korpela M, Kangas AJ, Lyytikäinen LP, Soininen P, Tukiainen T, Würtz P, Ong RT, Dörr M, Kroemer HK, Völker U, Völzke H, Galan P, Hercberg S, Lathrop M, Zelenika D, Deloukas P, Mangino M, Spector TD, Zhai G, Meschia JF, Nalls MA, Sharma P, Terzic J, Kumar MV, Denniff M, Zukowska-Szczechowska E, Wagenknecht LE, Fowkes FG, Charchar FJ, Schwarz PE, Hayward C, Guo X, Rotimi C, Bots ML, Brand E, Samani NJ, Polasek O, Talmud PJ, Nyberg F, Kuh D, Laan M, Hveem K, Palmer LJ, van der Schouw YT, Casas JP, Mohlke KL, Vineis P, Raitakari O, Ganesh SK, Wong TY, Tai ES, Cooper RS, Laakso M, Rao DC, Harris TB, Morris RW, Dominiczak AF, Kivimaki M, Marmot MG, Miki T, Saleheen D, Chandak GR, Coresh J, Navis G, Salomaa V, Han BG, Zhu X, Kooner JS, Melander O, Ridker PM, Bandinelli S, Gyllensten UB, Wright AF, Wilson JF, Ferrucci L, Farrall M, Tuomilehto J, Pramstaller PP, Elosua R, Soranzo N, Sijbrands EJ, Altshuler D, Loos RJ, Shuldiner AR, Gieger C, Meneton P, Uitterlinden AG, Wareham NJ, Gudnason V, Rotter JI, Rettig R, Uda M, Strachan DP, Witteman JC, Hartikainen AL, Beckmann JS, Boerwinkle E, Vasan RS, Boehnke M, Larson MG, Järvelin MR, Psaty BM, Abecasis GR, Chakravarti A, Elliott P, van Duijn CM, Newton-Cheh C, Levy D, Caulfield MJ, Johnson T, Tang H, Knowles J, Hlatky M, Fortmann S, Assimes TL, Quertermous T, Go A, Iribarren C, Absher D, Risch N, Myers R, Sidney S, Ziegler A, Schillert A, Bickel C, Sinning C, Rupprecht HJ, Lackner K, Wild P, Schnabel R, Blankenberg S, Zeller T, Münzel T, Perret C, Cambien F, Tiret L, Nicaud V, Proust C, Dehghan A, Hofman A, Uitterlinden A, van Duijn C, Levy D, Whitteman J, Cupples LA, Demissie-Banjaw S, Ramachandran V, Smith A, Gudnason V, Boerwinkle E, Folsom A, Morrison A, Psaty BM, Chen IY, Rotter JI, Bis J, Volcik K, Rice K, Taylor KD, Marciante K, Smith N, Glazer N, Heckbert S, Harris T, Lumley T, Kong A, Thorleifsson G, Thorgeirsson G, Holm H, Gulcher JR, Stefansson K, Andersen K, Gretarsdottir S, Thorsteinsdottir U, Preuss M, Schreiber S, Meitinger T, König IR, Lieb W, Hengstenberg C, Schunkert H, Erdmann J, Fischer M, Grosshennig A, Medack A, Stark K, Linsel-Nitschke P, Bruse P, Aherrahrou Z, Peters A, Loley C, Willenborg C, Nahrstedt J, Freyer J, Gulde S, Doering A, Meisinger C, Wichmann HE, Klopp N, Illig T, Meinitzer A, Tomaschitz A, Halperin E, Dobnig H, Scharnagl H, Kleber M, Laaksonen R, Pilz S, Grammer TB, Stojakovic T, Renner W, März W, Böhm BO, Winkelmann BR, Winkler K, Hoffmann MM, O'Donnell CJ, Voight BF, Altshuler D, Siscovick DS, Musunuru K, Peltonen L, Barbalic M, Melander O, Elosua R, Kathiresan S, Schwartz SM, Salomaa V, Guiducci C, Burtt N, Gabriel SB, Stewart AF, Wells GA, Chen L, Jarinova O, Roberts R, McPherson R, Dandona S, Pichard AD,

Rader DJ, Devaney J, Lindsay JM, Kent KM, Qu L, Satler L, Burnett MS, Li M, Reilly MP, Wilensky R, Waksman R, Epstein S, Matthai W, Knouff CW, Waterworth DM, Hakonarson HH, Walker MC, Mooser V, Hall AS, Balmforth AJ, Wright BJ, Nelson C, Thompson JR, Samani NJ, Braund PS, Ball SG, Smith NL, Felix JF, Morrison AC, Demissie S, Glazer NL, Loehr LR, Cupples LA, Dehghan A, Lumley T, Rosamond WD, Lieb W, Rivadeneira F, Bis JC, Folsom AR, Benjamin E, Aulchenko YS, Haritunians T, Couper D, Murabito J, Wang YA, Stricker BH, Gottdiener JS, Chang PP, Wang TJ, Rice KM, Hofman A, Heckbert SR, Fox ER, O'Donnell CJ, Uitterlinden AG, Rotter JI, Willerson JT, Levy D, van Duijn CM, Psaty BM, Witteman JC, Boerwinkle E, Vasan RS, Köttgen A, Pattaro C, Böger CA, Fuchsberger C, Olden M, Glazer NL, Parsa A, Gao X, Yang Q, Smith AV, O'Connell JR, Li M, Schmidt H, Tanaka T, Isaacs A, Ketkar S, Hwang SJ, Johnson AD, Dehghan A, Teumer A, Paré G, Atkinson EJ, Zeller T, Lohman K, Cornelis MC, Probst-Hensch NM, Kronenberg F, Tönjes A, Hayward C, Aspelund T, Eiriksdottir G, Launer LJ, Harris TB, Rampersaud E, Mitchell BD, Arking DE, Boerwinkle E, Struchalin M, Cavalieri M, Singleton A, Giallauria F, Metter J, de Boer J, Haritunians T, Lumley T, Siscovick D, Psaty BM, Zillikens MC, Oostra BA, Feitosa M, Province M, de Andrade M, Turner ST, Schillert A, Ziegler A, Wild PS, Schnabel RB, Wilde S, Munzel TF, Leak TS, Illig T, Klopp N, Meisinger C, Wichmann HE, Koenig W, Zgaga L, Zemunik T, Kolcic I, Minelli C, Hu FB, Johansson A, Igl W, Zaboli G, Wild SH, Wright AF, Campbell H, Ellinghaus D, Schreiber S, Aulchenko YS, Felix JF, Rivadeneira F, Uitterlinden AG, Hofman A, Imboden M, Nitsch D, Brandstätter A, Kollerits B, Kedenko L, Mägi R, Stumvoll M, Kovacs P, Boban M, Campbell S, Endlich K, Völzke H, Kroemer HK, Nauck M, Völker U, Polasek O, Vitart V, Badola S, Parker AN, Ridker PM, Kardia SL, Blankenberg S, Liu Y, Curhan GC, Franke A, Rochat T, Paulweber B, Prokopenko I, Wang W, Gudnason V, Shuldiner AR, Coresh J, Schmidt R, Ferrucci L, Shlipak MG, van Duijn CM, Borecki I, Krämer BK, Rudan I, Gyllensten U, Wilson JF, Witteman JC, Pramstaller PP, Rettig R, Hastie N, Chasman DI, Kao WH, Heid IM, Fox CS, Vasan RS, Glazer NL, Felix JF, Lieb W, Wild PS, Felix SB, Watzinger N, Larson MG, Smith NL, Dehghan A, Grosshennig A, Schillert A, Teumer A, Schmidt R, Kathiresan S, Lumley T, Aulchenko YS, König IR, Zeller T, Homuth G, Struchalin M, Aragam J, Bis JC, Rivadeneira F, Erdmann J, Schnabel RB, Dörr M, Zweiker R, Lind L, Rodeheffer RJ, Greiser KH, Levy D, Haritunians T, Deckers JW, Stritzke J, Lackner KJ, Völker U, Ingelsson E, Kullo I, Haerting J, O'Donnell CJ, Heckbert SR, Stricker BH, Ziegler A, Reffelmann T, Redfield MM, Werdan K, Mitchell GF, Rice K, Arnett DK, Hofman A, Gottdiener JS, Uitterlinden AG, Meitinger T, Blettner M, Friedrich N, Wang TJ, Psaty BM, van Duijn CM, Wichmann HE, Munzel TF, Kroemer HK, Benjamin EJ, Rotter JI, Witteman JC, Schunkert H, Schmidt H, Völzke H, Blankenberg S, Chambers JC, Zhang W, Lord GM, van der Harst P, Lawlor DA, Sehmi JS, Gale DP, Wass MN, Ahmadi KR, Bakker SJ, Beckmann J, Bilo HJ, Bochud M, Brown MJ, Caulfield MJ, Connell JM, Cook HT, Cotlarciuc I, Davey Smith G, de Silva R, Deng G, Devuyst O, Dikkeschei LD, Dimkovic N, Dockrell M, Dominiczak A, Ebrahim S, Eggermann T, Farrall M, Ferrucci L, Floege J, Forouhi NG, Gansevoort RT, Han X, Hedblad B, Homan van der Heide JJ, Hepkema BG, Hernandez-Fuentes M, Hypponen E, Johnson T, de Jong PE, Kleefstra N, Lagou V, Lapsley M, Li Y, Loos RJ, Luan J, Luttropp K, Maréchal C, Melander O, Munroe PB, Nordfors L, Parsa A, Peltonen L, Penninx BW, Perucha E, Pouta A, Prokopenko I, Roderick PJ, Ruokonen A, Samani NJ, Sanna S, Schalling M, Schlessinger D, Schlieper G, Seelen MA, Shuldiner AR, Sjögren M, Smit JH, Snieder H, Soranzo N, Spector TD, Stenvinkel P, Sternberg MJ, Swaminathan R, Tanaka T, Ubink-Veltmaat LJ, Uda M, Vollenweider P, Wallace C, Waterworth D, Zerres K, Waeber G, Wareham NJ, Maxwell PH, McCarthy MI, Jarvelin MR, Mooser V, Abecasis GR, Lightstone L, Scott J, Navis G, Elliott P, Kooner JS. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*. 2011; 478:103–109.

14. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*. 2008;9:356–369.
15. Padmanabhan S, Melander O, Hastie C, Menni C, Delles C, Connell JM, Dominiczak AF. Hypertension and genome-wide association studies: combining high fidelity phenotyping and hypercontrols. *J Hypertens*. 2008;26:1275–1281.
16. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*. 2010;34:816–834.
17. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nat Protoc*. 2010;5:1564–1573.
18. Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT. Basic statistical analysis in genetic case-control studies. *Nat Protoc*. 2011;6:121–133.
19. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 2010;26:2190–2191.
20. Heinemeyer T, Wingender E, Reuter I, Hermjakob H, Kel AE, Kel OV, Ignatieva EV, Ananko EA, Podkolodnaya OA, Kolpakov FA, Podkolodny NL, Kolchanov NA. Databases on transcriptional regulation: TRANSFAC, TRRD, and COMPEL. *Nucleic Acids Res*. 1998;26:364–370.
21. Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Münch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*. 2003;31:374–378.
22. TFSEARCH: Searching Transcription Factor Binding Sites (ver 1.3). http://www.cbrc.jp/research/db/TFSEARCH.html. Accessed June 2011.
23. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD. Genes mirror geography within Europe. *Nature*. 2008;456:98–101.
24. Marsden PA, Heng HH, Scherer SW, Stewart RJ, Hall AV, Shi XM, Tsui LC, Schappert KT. Structure and chromosomal localization of the human constitutive endothelial nitric oxide synthase gene. *J Biol Chem*;. 1993; 268:17478–17488.
25. Zhang R, Min W, Sessa WC. Functional analysis of the human endothelial nitric oxide synthase promoter: Sp1 and GATA factors are necessary for basal transcription in endothelial cells. *J Biol Chem*. 1995;270: 15320–15326.
26. Oess S, Icking A, Fulton D, Govers R, Müller-Esterl W. Subcellular targeting and trafficking of nitric oxide synthases. *Biochem J*. 2006;396: 401–409.
27. Fleming I. Molecular mechanisms underlying the activation of eNOS. *Pflugers Arch*. 2010;459:793–806.
28. Moncada S, Higgs A. The L-arginine-nitric oxide pathway. *N Engl J Med*. 1993;329:2002–2012.
29. Su Y, Edwards-Bennett S, Bubb MR, Block ER. Regulation of endothelial nitric oxide synthase by the actin cytoskeleton. *Am J Physiol Cell Physiol*. 2003;284:C1542–C1549.
30. Kondrikov D, Fonseca FV, Elms S, Fulton D, Black SM, Block ER, Su Y. β-actin association with endothelial nitric-oxide synthase modulates nitric oxide and superoxide generation from the enzyme. *J Biol Chem*. 2010;285:4319–4327.
31. Karantzoulis-Fegaras F, Antoniou H, Lai SL, Kulkarni G, D'Abreo C, Wong GK, Miller TL, Chan Y, Atkins J, Wang Y, Marsden PA. Characterization of the human endothelial nitric-oxide synthase promoter. *J Biol Chem*. 1999;274:3076–3093.
32. Sõber S, Org E, Kepp K, Juhanson P, Eyheramendy S, Gieger C, Lichtner P, Klopp N, Veldre G, Viigimaa M, Döring A; Kooperative Gesundheitsforschung in der Region Augsburg Study, Putku M, Kelgo P; HYPertension in ESTonia Study, Shaw-Hawkins S, Howard P, Onipinla A, Dobson RJ, Newhouse SJ, Brown M, Dominiczak A, Connell J, Samani N, Farrall M; MRC British Genetics of Hypertension Study, Caulfield MJ, Munroe PB, Illig T, Wichmann HE, Meitinger T, Laan M. Targeting 160 candidate genes for blood pressure regulation with a genome-wide genotyping array. *PLoS ONE*. 2009;4:1–13.
33. Li J, Cun Y, Tang WR, Wang Y, Li SN, Ouyang HR, Wu YR, Yu HJ, Xiao CJ. Association of eNOS gene polymorphisms with essential hypertension in the Han population in southwestern China. *Genet Mol Res*. 2011;10:2202–2212.
34. Yan-Yan L. Endothelial nitric oxide synthase G894T gene polymorphism and essential hypertension in the Chinese population: a meta-analysis involving 11,248 subjects. *Intern Med*. 2011;50:2099–2106.
35. Men C, Tang K, Lin G, Li J, Zhan Y. ENOS-G894T polymorphism is a risk factor for essential hypertension in China. *Indian J Biochem Biophys*. 2011;48:154–157.
36. Niu W, Qi Y. An updated meta-analysis of endothelial nitric oxide synthase gene: three well-characterized polymorphisms with hypertension. *PLoS One*. 2011;6:e24266.
37. Huang PL, Huang Z, Mashimo H, Bloch KD, Moskowitz MA, Bevan JA, Fishman MC. Hypertension in mice lacking the gene for endothelial nitric oxide synthase. *Nature*. 1995;377:239–242.

Target Sequencing, Cell Experiments, and a Population Study Establish Endothelial Nitric Oxide Synthase (*eNOS*) Gene as Hypertension Susceptibility Gene

Erika Salvi,* Tatiana Kuznetsova,* Lutgarde Thijs,* Sara Lupoli, Katarzyna Stolarz-Skrzypek, Francesca D'Avila, Valerie Tikhonoff, Silvia De Astis, Matteo Barcella, Jitka Seidlerová, Paola Benaglio, Sofia Malyutina, Francesca Frau, Dinesh Velayutham, Roberta Benfante, Laura Zagato, Alexandra Title, Daniele Braga, Diana Marek, Kalina Kawecka-Jaszcz, Edoardo Casiglia, Jan Filipovský, Yuri Nikitin, Carlo Rivolta, Paolo Manunta, Jacques S. Beckmann, Cristina Barlassina,† Daniele Cusi,† Jan A. Staessen†

Abstract—A case–control study revealed association between hypertension and rs3918226 in the endothelial nitric oxide synthase (*eNOS*) gene promoter (minor/major allele, *T/C* allele). We aimed at substantiating these preliminary findings by target sequencing, cell experiments, and a population study. We sequenced the 140-kb genomic area encompassing the *eNOS* gene. In HeLa and HEK293T cells transfected with the *eNOS* promoter carrying either the *T* or the *C* allele, we quantified transcription by luciferase assay. In 2722 randomly recruited Europeans (53.0% women; mean age 40.1 years), we studied blood pressure change and incidence of hypertension in relation to rs3918226, using multivariable-adjusted models. Sequencing confirmed rs3918226, a binding site of E-twenty six transcription factors, as the single nucleotide polymorphism most closely associated with hypertension. In *T* compared with *C* transfected cells, *eNOS* promoter activity was from 20% to 40% ($P<0.01$) lower. In the population, systolic/diastolic blood pressure increased over 7.6 years (median) by 9.7/6.8 mm Hg in 28 *TT* homozygotes and by 3.8/1.9 mm Hg in 2694 *C* allele carriers ($P\leq0.0004$). The blood pressure rise was 5.9 mm Hg systolic (confidence interval [CI], 0.6–11.1; $P=0.028$) and 4.8 mm Hg diastolic (CI, 1.5–8.2; $P=0.0046$) greater in *TT* homozygotes, with no differences between the *CT* and *CC* genotypes ($P\geq0.90$). Among 2013 participants normotensive at baseline, 692 (34.4%) developed hypertension. The hazard ratio and attributable risk associated with *TT* homozygosity were 2.04 (CI, 1.24–3.37; $P=0.0054$) and 51.0%, respectively. In conclusion, rs3918226 in the *eNOS* promoter tags a hypertension susceptibility locus, *TT* homozygosity being associated with lesser transcription and higher risk of hypertension. **(*Hypertension*. 2013;62:844-852.)** • **Online Data Supplement**

Key Words: blood pressure ■ endothelial nitric oxide synthase gene ■ hypertension ■ population science ■ target sequencing ■ transfection

Hypertension is a chronic age-related disease influenced by a large number of genetic and environmental factors, lifestyle, and their interaction.[1] Hypertension affects an estimated 25% to 35% of the world's population and >60% of the elderly.[1–3] In 2001, hypertension caused 8 million deaths worldwide, representing 14% of global mortality.[4] High blood pressure (BP) is the main driver of ischemic heart disease and stroke[3,4] and substantially exceeds the contribution of the 2 other main modifiable risk factors, hypercholesterolemia and smoking, to the global burden of noncommunicable disease.[4] Several genome-wide association studies (GWAS) identified a number of single-nucleotide polymorphisms (SNPs), all with a small effect on BP.[5,6] In view of the impact of BP as a continuous risk factor, small genetic effects might entail substantial effects on morbidity and mortality.[1,3]

Using GWAS in a case–control design,[7] we recently identified rs3918226 as a new hypertension susceptibility

Received March 31, 2013; first decision May 10, 2013; revision accepted August 12, 2013.
From the European Working Party on *eNOS*.
The European Working Party on *eNOS* is an ad hoc collaboration between investigators involved in the Flemish Study on Environment, Genes, and Health Outcomes (FLEMENGHO), the European Project on Genes in Hypertension (EPOGH), and the HYPERGENES project.
*These authors are joint first authors.
†These authors are joint senior authors.
This paper was sent to Morris Brown, Guest editor, for review by expert referees, editorial decision, and final disposition.
The online-only Data Supplement is available with this article at http://hyper.ahajournals.org/lookup/suppl/doi:10.1161/HYPERTENSIONAHA.113.01428/-/DC1.
Correspondence to Jan A. Staessen, Studies Coordinating Centre, Research Unit Hypertension and Cardiovascular Epidemiology, KU Leuven Department of Cardiovascular Sciences, Campus Sint Rafaël, Kapucijnenvoer 35, Block D, Box 7001, BE-3000 Leuven, Belgium. E-mail jan.staessen@med.kuleuven.be

Hypertension **is available at http://hyper.ahajournals.org** **DOI: 10.1161/HYPERTENSIONAHA.113.01428**

locus. This locus lays in the promoter of the endothelial nitric oxide synthase (*eNOS*) gene, which encodes the enzyme that produces nitric oxide, a strong vasodilator with a key role in the regulation of systemic vascular resistance. GWAS usually points to genomic regions of interest in relation to a trait, but seldom directly identifies the causal or functional variant. In the present study, we aimed at consolidating the role of *eNOS* as a hypertension susceptibility gene by fine mapping the DNA sequence tagged by rs3918226, by studying the transcriptional functionality of the rs3918226 alleles in vitro, and by relating the change in BP over time to rs3918226 in a randomly recruited population sample.

Methods

Target Sequencing

From the HYPERGENES study,[7] we selected 44 hypertensive patients carrying ≥1 *T* allele and 48 healthy controls homozygous for the *C* allele. Analyses of the genetic data confirmed that all patients and controls were of continental Italian descent. We sequenced a 140-kb DNA region of chromosome,[7] which, in addition to *eNOS*, included *KCNH2* mapping upstream and 6 genes mapping downstream: *ATG9B*, *ABCB8*, *ACCN3*, *CDK5*, *SLC4A2*, and *FASTK*. Detailed information on the DNA sequencing methods and an accompanying glossary are available in the online-only Data Supplement. We sequenced indexed and multiplexed samples in a paired-end protocol implemented on an Illumina GAIIX platform (Illumina Inc, San Diego, CA).

Paired-end raw reads were checked for their quality using FastQC, version 0.10.0 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and PrinSeq, version 0.19.4 (http://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi). Reads were aligned to the human reference genome (hg19; UCSC assembly; February 2009) using BWA. SAMtools, Picard (http://picard.sourceforge.net), and Genome Analysis Toolkit (GATK) were used to handle the reads and for postalignment quality control checks. Multisamples variant call was performed by GATK, and quality filters were applied to variant call. Only variants with high SNP quality score were evaluated and annotated with Annovar software for type and impact on the gene product (http://www.openbioinformatics.org/annovar).

We imputed missing genotypes based on the HYPERGENES results with MiniMac, a low memory, computationally efficient implementation of the MaCH algorithm,[8,9] using as reference panel the 1000 Genome Database, released in June 2011. We tested imputed SNPs with high quality (mean R^2, 0.91; SD, ±0.11) for association with hypertension, using logistic regression as implemented in Mach2dat[8,9] with adjustments applied for sex and the first 10 significant principal components.

Luciferase Reporter Assays

We obtained the pGL2-eNOS promoter-luciferase plasmid, carrying the *C* allele of rs3918226, from Addgene (plasmid 19297; http://www.addgene.org) and constructed the *T* allele by site-directed mutagenesis (see Expanded Methods available in online-only Data Supplement). We transfected HeLa cells in 4 independent experiments and each construct was tested in triplicate. HEK293T cells were transfected in 3 independent experiments and each construct was tested in duplicate. We compared luciferase reporter activities between the *C* and *T* alleles by Student *t* test.

Population Study

Study Participants

Recruitment for the Flemish Study on Environment, Genes, and Health Outcomes (FLEMENGHO) started in 1985.[10,11] From August 1985 to November 1990, a random sample of the households living in a geographically defined area of Northern Belgium was investigated with the goal to recruit an equal number of participants in each of 6 strata by sex and age (20–39, 40–59, and ≥60 years). All household members aged ≥20 years were invited, provided that the quota of their sex–age group had not yet been satisfied. From June 1996 until January 2004, recruitment of families continued using the former participants (1985–1990) as index persons and also including teenagers. The participants were repeatedly followed up. In all study phases, we used the same standardized methods to measure BP and to administer questionnaires.[10,11] The European Project on Genes in Hypertension (EPOGH) recruited participants from 1999 to 2001.[11,12] The EPOGH investigators received training at the Studies Coordinating Centre in Leuven, Belgium, and applied the same protocol, questionnaires, and follow-up procedures, as used in FLEMENGHO. Questionnaires were translated from Dutch and English into Czech, Italian, Polish, and Russian and back-translated into Dutch and English to ensure that all questions kept the same meaning in all languages. The last follow-up examination took place from 2005 to 2008 in FLEMENGHO[11] and from 2006 to 2008 in EPOGH.[11] Both studies complied with the Helsinki Declaration for investigation of human subjects.[13] Each local institutional review board approved the study protocol. Participants gave written informed consent.

Measurements

At baseline and follow-up, experienced observers measured each participant's anthropometric characteristics and BP and administered the standardized questionnaire to collect information on medical history, smoking and drinking habits, and use of medications, including BP-lowering drugs, contraceptive pill intake, and hormonal replacement therapy. At each contact, BP was the average of 5 consecutive auscultatory readings in the sitting position (see Expanded Methods available in online-only Data Supplement). Digit and number preference was checked at 6-month intervals.[12] Hypertension was an untreated BP of ≥140 mm Hg systolic, or 90 mm Hg diastolic, or use of antihypertensive drugs. For adolescents (n=33), we used the thresholds specified by the European Society of Hypertension, which are stratified by sex, age, and height percentiles.[14] Body mass index was weight in kilograms divided by the square of height in meters.

Datasets for Analysis

From 3785 subjects who initially agreed to participate in FLEMENGHO (n=2593) and EPOGH (n=1192), 53 participants were excluded because the baseline BP measurements were missing, leaving 3732 subjects with a full set of required baseline measurements. Of these, 2981 subjects participated in ≥1 follow-up examination. We additionally excluded 259 participants from analysis because the BP measurements were missing (n=32) or their DNA was of bad quality (n=227). Thus, the blood pressure cohort used to study change in BP included 2722 participants (Figure 1). Changes in BP during follow-up were calculated as the last minus the baseline BP. The hypertension cohort used to study the incidence of hypertension encompassed 2013 participants, who were normotensive at baseline (Figure 1). We censored subjects from further analysis after occurrence of the first diagnosis of hypertension.

Statistical Analysis

For database management and statistical analysis, we used SAS software, version 9.3 (SAS Institute Inc, Cary, NC). Between-group comparisons of means, medians, proportions, and Kaplan–Meier survival functions relied on the standard normal z test or ANOVA, Kruskal–Wallis ANOVA or Fisher exact test, and the log-rank test, respectively. We applied McNemar test to evaluate changes over time in categorical variables. We computed 95% confidence intervals (CIs) of rates as $R \pm 1.96 \times \sqrt{(R/T)}$, where R and T are the rate and the denominator used to calculate the rate.

We studied the association between change in BP and the rs39188226 genotypes, using mixed models with indicator variables (0,1). Multivariable-adjusted models with BP changes as

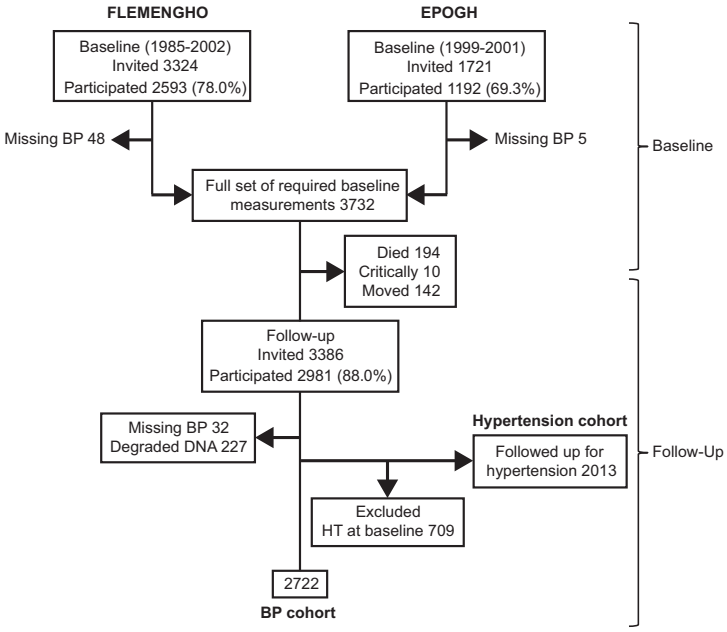


Figure 1. Flow diagram of participants. The blood pressure cohort and hypertension cohort refer to participants used to study changes in blood pressure and the incidence of hypertension over follow-up, respectively.

dependent variables accounted for cohort, sex, age, baseline BP, follow-up duration, baseline and follow-up body mass index, intake of female sex hormones or nonsteroidal antiinflammatory drugs at baseline and follow-up, and 3 indicator variables coding for antihypertensive drug intake (starting or stopping treatment between baseline and follow-up and remaining on treatment). A sensitivity analysis additionally accounted for family clusters modeled as a random effect.

To study the incidence of hypertension, we applied Cox regression adjusted for the same covariables as in the continuous analyses. We checked the proportional hazards assumption by the Kolmogorov supremum test. To account for family clusters, we used the PROC SURVIVAL procedure of the SUDAAN 10.01 software (Research Triangle Institute, NC). We computed the positive predictive value of *TT* homozygosity as $(R\times D)/([G/100]\times[R-1]+1)$, where R is the multivariable-adjusted hazard ratio, D is the incidence of hypertension in the whole population (34.5%), and G is the prevalence of *TT* homozygosity (1.09%).[15] The attributable risk is given by $([R-1]\times100)/R$ and the population-attributable risk by $([G/100]\times[R-1]\times100)/([G/100]\times[R-1]+1)$.[15]

Results

Target Sequencing

In 44 hypertensive patients and 48 healthy controls, we captured 91% (SD, 6) of the targeted genomic region with a 63-fold amplification above the genomic background. In each sample, the region of interest was covered on average 20 times with mapping and base quality scores of ≥20 and 17, respectively.

We identified 338 variants, of which 15 were nonsynonymous; 23 synonymous; 23 and 8 in the 3′ UTR and 5′ UTR, respectively; 198 intronic; and 71 intergenic. Among the 338 variants, 277 were already annotated in dbSNP135 (Table S1 in the online-only Data Supplement) and 61 were novel variants (Table S2). We had genotyped 76 of 277 annotated SNPs in the HYPERGENES GWAS, using the Illumina 1M array.[7] The genotype concordance rate between the 2 technologies was high (r^2=≈0.988). Of the 61 novel SNPs, 55 heterozygous variants were rare, only present in a single subject. Table S3 lists the 6 other novel variants. Of these, 3 were intronic in *KCNH2* (1 homozygous in 1 subject and 2 other heterozygous in 2 and 4 individuals). One variant was intergenic, mapping ≈3 kb from *KCNH2* and 9.5 kb from e*NOS* (3 heterozygotes). One was intronic in e*NOS* (2 heterozygotes). One was intronic in *ABCB8* (9 heterozygotes and 1 homozygote).

The haplotype analysis appears in page S5 and Figure S1. Five novel variants were located in a region of linkage disequilibrium upstream of rs3918226. Their annotation did not suggest any functional role (Table S3). We, therefore, did not consider them in further analyses. Variants imputed in the entire HYPERGENES study population with a probability value of $\leq10^{-3}$ appear in Table S4. rs3918226 remained the SNP most closely associated with hypertension.

Luciferase Reporter Assays

SNP rs3918226 is located in the promoter region of *eNOS*. Compared with the *C* allele, the risk-carrying *T* allele was associated with lower transcriptional activity of the *eNOS* gene ranging from ≈20% (P<0.0001) when tested in HeLa cells (Figure 2A) to ≈40% (P<0.01) in HEK293T cells (Figure 2B).

Population Study

Characteristics of the Participants

The Table lists the characteristics of the participants by cohort and rs3918226 genotype. The blood pressure cohort (n=2722) included 1442 (53.0%) women and 55 (2.0%) diabetic patients. All participants were white Europeans. Age averaged 40.7 years (range 19.5–83.5). At baseline, 709 (26.1%) participants had hypertension, of whom 322 (45.4%) were on antihypertensive drug treatment. The hypertension cohort consisted of 2013 participants, who were normotensive at baseline (Table). In 3 groups of unrelated participants (n=717), randomly selected from the blood pressure cohort,

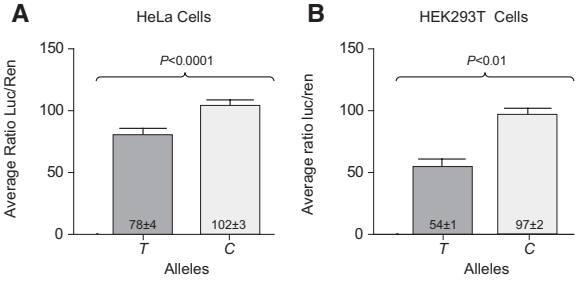


Figure 2. Luciferase activity in transfected cells. Luciferase activity was measured in HeLa cells (**A**) and HEK293T cells (**B**) transfected with constructs carrying the *C* or *T* allele. HeLa cells were transfected in 4 independent experiments and each construct was tested in triplicate. HEK293T cells were transfected in 3 independent experiments and each construct was tested in duplicate. Plus-minus values are mean±SE. Compared with the *C* allele, the risk-carrying *T* allele was associated with a ≈20% lower ($P<0.0001$) transcriptional activity of the *eNOS* gene in HeLa cells and with a ≈40% lower ($P<0.01$) transcriptional activity in HEK293T cells.

the frequencies of the rs3918226 genotypes did not deviate from Hardy–Weinberg equilibrium ($0.19<P<0.99$). The genotype and allele frequencies were equally distributed across cohorts ($P=0.67$; Table S6).

Cross-sectional Analyses of the Blood Pressure Cohort

At baseline (Table), systolic BP was similar across the rs3918226 genotypes ($P=0.42$), but *TT* homozygotes had higher ($P=0.030$) diastolic BP than *C* allele carriers. At follow-up (Table), systolic ($P=0.012$) and diastolic ($P=0.0002$) BPs were higher in *TT* homozygotes than in *C* allele carriers. The differences in baseline and follow-up systolic and diastolic BPs between the *CT* and *CC* genotypes were not statistically significant ($P\geq0.23$; Table). With adjustments applied for cohort, sex, age, body mass index, and antihypertensive drug intake, baseline ($P=0.035$) and follow-up ($P=0.0007$) diastolic BPs and follow-up systolic BP ($P=0.038$), but not baseline systolic BP ($P=0.48$), remained higher in *TT* homozygotes than in *C* allele carriers.

Blood Pressure Cohort

Follow-up data were available at 1, 2, or ≥3 occasions in 1269, 459, and 994 participants, respectively. Median follow-up was slightly longer ($P=0.053$) in *TT* homozygotes (10.4 years; interquartile range [IQR], 7.2–14.0) than in *C* allele carriers (7.6 years; IQR, 6.1–12.3; Table). In multivariable-adjusted analyses (see Methods) of the BP changes over follow-up (Figure 3), systolic BP increased 9.7 mm Hg (CI, 4.2–15.1; $P=0.0005$) in *TT* homozygotes and by 3.9 mm Hg (CI, 1.8–5.9; $P=0.0003$), 3.8 mm Hg (CI, 2.1–5.4; $P<0.0001$), and 3.8 mm Hg (CI, 2.1–5.4; $P<0.0001$) in *CT* heterozygotes, *CC* homozygotes, and *C* allele carriers, respectively. Diastolic BP increased by 6.8 mm Hg (CI, 3.3–10.3; $P<0.0001$) in *TT* homozygotes and by 2.0 mm Hg (CI, 0.6–3.3; $P=0.0036$), 1.9 mm Hg (CI, 0.9–3.0; $P=0.0004$), and 1.9 mm Hg (CI, 0.9–3.0; $P=0.0004$) in *CT* heterozygotes, *CC* homozygotes, and *C* allele carriers, respectively. Systolic BP increased 5.8 mm Hg (CI, 0.4–11.2; $P=0.034$) and 5.9 mm Hg (CI, 0.7–11.1; $P=0.028$) more in *TT* homozygotes than in *CT* heterozygotes and *CC* homozygotes. Similarly, diastolic BP increased 4.8 mm Hg (CI, 1.4–8.3; $P=0.0062$) and 4.8 mm Hg (CI, 1.5–8.2; $P=0.0046$) more in *TT* homozygotes than in *CT* and *CC* genotype carriers. Furthermore, the BP increases were 5.9 mm Hg systolic (CI, 0.6–11.1; $P=0.028$) and 4.8 mm Hg diastolic (CI, 1.5–8.2; $P=0.0046$) greater in *TT* homozygotes than in *C* allele carriers. In models additionally accounting for family clusters, the effect sizes were 5.3 mm Hg systolic (CI, –0.5 to 11.2; $P=0.071$) and 3.1 mm Hg diastolic (CI, –0.6 to 6.8; $P=0.097$). The interactions between the *TT* genotype and sex or age were not significant ($P\geq0.66$).

Hypertension Cohort

Follow-up data were available at 1, 2, or ≥3 contacts in 874, 351, and 788 participants, respectively. The median duration of follow-up (7.1 years; IQR, 5.5–10.4) was similar ($P=0.18$) among the rs3918226 genotypes (Table). In the entire cohort, 692 participants developed hypertension, of whom 216 (31.2%) were on antihypertensive drug treatment at the least follow-up visit. In 476 untreated patients (68.8%), the diagnosis of hypertension relied on thresholds being exceeded for systolic or diastolic BP or both in 171 (35.9%), 166 (34.9%), and 139 (29.2%) patients, respectively. The *TT* genotype conferred a higher risk of hypertension compared with the *CT* and *CC* genotypes ($P\leq0.011$) or *C* allele carriers ($P=0.003$), with no difference between the *C* allele–carrying genotypes ($P=0.55$; Figure 4). Expressed per 1000 person-years of follow-up, incidence rates of hypertension were 40.6 cases (CI, 37.6–43.6) in the entire hypertension cohort, 86.8 (CI, 44.3–129.3) in *TT* homozygotes, 43.7 (CI, 35.8–51.6) in *CT* heterozygotes, 39.4 (CI, 36.1–42.7) in *CC* homozygotes, and 40.1 (CI, 37.1–43.1) in *C* allele carriers. In multivariable-adjusted Cox regression (Figure 5), the hazard ratios associated with the *TT* homozygosity were 2.04 (CI, 1.23–3.37; $P=0.0056$), 2.06 (CI, 1.21–3.50; $P=0.0075$), and 2.04 (CI, 1.24–3.37; $P=0.0054$) compared with *CT* heterozygotes, *CC* homozygotes, and *C* allele carriers. In an analysis adjusted for family clusters, the hazard ratio expressing the risk of hypertension in *TT* homozygotes versus *C* allele carriers was 2.4 (CI, 1.20–3.46; $P=0.0082$). All Cox models complied with the proportional hazards assumption. The positive predictive values, attributable risk, and population-attributable risk associated with *TT* homozygosity were 69.3%, 51.0%, and 1.1%, respectively.

Discussion

The key finding of the present study was that rs3918226 in the *eNOS* promoter tags a hypertension susceptibility locus, *TT* homozygosity being associated with lesser transcription of the gene product and a 2-fold higher risk of hypertension. Our current findings confirm the previously reported HYPERGENES case–control study.[7] The discovery phase of this project[7] involved 1865 hypertensive patients and 1750 controls, who were genotyped with a Illumina 1M array. The validation study included 1385 cases and 1246 controls, who were genotyped with a 14-K Illumina Infinium custom array. HYPERGENES showed that rs3918226 in the *eNOS* gene promoter (–665 *C>T*) tags a hypertension susceptibility locus.[7] The odds ratio associated with the *T* allele was 1.54 (CI, 1.37–1.73; $P=2.58\times10^{-13}$). In a

Table. Characteristics of Participants by Cohort

	Blood Pressure Cohort			Hypertension Cohort		
Characteristic	*TT*	*CT*	*CC*	*TT*	*CT*	*CC*
Number of subjects (%)	28 (1.0)	411 (15.1)	2283 (83.9)	22 (1.1)	300 (14.9)	1691 (84.0)
Median follow-up, y (IQR)	10.4 (7.2–14.0)	8.4 (6.4–14.1)	7.4 (6.0–12.0)	7.2 (4.6–10.8)	7.2 (5.9–10.0)	7.0 (5.7–9.4)
Number (%) with characteristic						
FLEMENGHO	19 (67.9)	287 (69.8)	1612 (70.6)	17 (77.3)	221 (73.7)	1256 (74.3)
EPOGH	9 (32.1)	124 (30.2)	671 (29.4)	5 (22.7)	79 (26.3)	435 (25.7)
Women	10 (35.7)	214 (52.1)	1218 (53.4)	8 (36.4)	165 (55.0)	918 (54.3)
Baseline						
Hypertension	6 (21.4)	111 (27.0)	592 (25.9)	…	…	…
Antihypertensive treatment	2 (7.1)	51 (12.4)	269 (11.8)	…	…	…
Use of female sex hormones	2 (7.1)	47 (11.4)	208 (9.1)	2 (9.1)	42 (14.0)	182 (10.8)
Use of NSAID	3 (10.7)	53 (12.9)	326 (14.3)	3 (13.6)	39 (13.0)	232 (13.7)
Follow-up						
Hypertension	20 (71.4)†¶	191 (46.5)¶	973 (42.6)¶	16 (72.7)‡	118 (39.3)	558 (33.0)
Antihypertensive treatment	10 (35.7)‖	118 (28.7)¶	617 (27.0)¶	2 (9.1)	38 (12.7)	176 (10.4)
Use of female sex hormones	1 (3.6)	19 (4.6)¶	112 (4.9)¶	1 (4.6)	22 (7.3)‖	112 (6.6)§
Use of NSAID	3 (10.7)	56 (13.6)	334 (14.6)	3 (13.6)	28 (9.3)§	182 (10.8)‖
Mean (SD) characteristic						
Baseline						
Age, y	41.9 (16.0)	40.5 (14.9)	40.7 (15.1)	37.0 (14.3)	37.2 (13.8)	37.3 (14.2)
Body mass index, kg/m^2	25.8 (5.0)	25.7 (4.5)	25.4 (4.5)	25.4 (5.3)	24.7 (4.0)	24.4 (4.0)
Systolic pressure, mm Hg	128.6 (20.7)	126.1 (17.2)	125.3 (17.2)	121.5 (12.2)	118.8 (10.6)	118.4 (10.5)
Diastolic pressure, mm Hg	81.8 (9.4)*	77.8 (11.4)	77.2 (11.1)	80.2 (8.8)‡	73.8 (8.0)	73.3 (8.3)
Follow-up						
Age, y	52.8 (15.1)¶	51.0 (16.4)¶	50.1 (16.6)¶	45.4 (14.7)¶	46.2 (15.0)¶	45.6 (15.2)¶
Body mass index, kg/m^2	27.4 (5.6)‖	26.9 (5.0)¶	26.7 (7.0)¶	26.9 (4.9)¶	26.0 (4.7)¶	25.9 (7.4)¶
Systolic pressure, mm Hg	138.4 (18.1)*‖	130.6 (17.8)¶	129.5 (18.3)¶	136.7 (17.6)‡¶	126.1 (15.4)¶	125.1 (15.1)¶
Diastolic pressure, mm Hg	87.5 (10.2)‡‖	80.5 (11.2)¶	79.8 (10.5)¶	90.6 (11.5)‡¶	79.5 (10.9)¶	79.0 (10.3)¶

Blood pressure cohort and hypertension cohort refer to participants used to study the changes in blood pressure over follow-up and the incidence of hypertension, respectively. Blood pressure was the average of 5 consecutive readings at a single visit. FLEMENGHO indicates Flemish Study on Environment, Genes, and Health Outcomes, EPOGH, European Project of Genes in Hypertension; IQR, interquartile range; and NSAID, nonsteroidal antiinflammatory drugs. Significance of the difference with *TT* genotype: *P≤0.05; †P≤0.01; ‡P≤0.001; significance of the difference with baseline: §P≤0.05; ‖P≤0.01; ¶P≤0.001.

meta-analysis, using both in silico and de novo genotyping data in 21 714 subjects, the odds ratio was 1.34 (CI, 1.25–1.44; P=1.03×10^{-14}).[7] In the current study, using BP as a continuous phenotype in a randomly recruited European population sample followed up for 7.6 years (median), systolic and diastolic BPs increased 5.9 and 4.8 mm Hg more in *TT* homozygotes than in *C* allele carriers.

Fine mapping the *eNOS* genomic region in hypertensive patients and healthy controls and imputing approaches in the whole HYPERGENES cohort using 1000 Genome Database released in 2011 as reference further validated rs3918226 as the SNP most closely associated with hypertension. Of 61 novel variants discovered in the genomic area of interest, 55 were singletons with very low minor allele frequency (<0.5%), and 5 were located in a region of linkage disequilibrium and were not functional. The remaining newly discovered variant in the *ABCB8* gene located downstream of rs3918226 and was not genotyped in the HYPERGENES sample, because it could not have been tagged by rs3918226 in view of the high recombination rate in this genomic region.

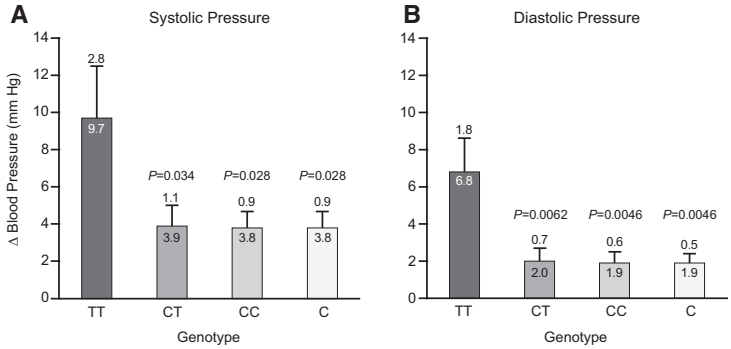


Figure 3. Association between blood pressure (BP) change (Δ) and rs3918226 genotype. The estimates were adjusted for cohort, sex, age, baseline BP, follow-up duration, baseline and follow-up body mass index, intake of female sex hormones or nonsteroidal antiinflammatory drugs at baseline and follow-up, and 3 indicator variables coding for antihypertensive drug intake (starting or stopping treatment between baseline and follow-up and remaining on treatment). Plotted values are mean±SE (values given) for systolic (**A**) and diastolic (**B**) BP. Probability values denote the significance of the difference with the *TT* genotype.

In our initial report,[7] we tested whether rs3918226 falls into a regulatory binding site. Using the PATCH algorithm of the TRANSFAC database[16] and the TFSEARCH software[17] (score 87.3[7]), we characterized a putative binding site for transcription factors of the ETS (E-twenty six) family only 1 nucleotide away from rs3918226. The members of the ETS family, ETS-1 and ELF-1, are present in endothelial cells and are essential for the activation of the *eNOS* promoter.[18] We then hypothesized that rs3918226 maps in an open chromatin region. Indeed, DNaseI and FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) experiments in HUVEC (Human Umbilical Vein Endothelial Cells) cells from ENCODE show significant signals around rs3918226 either for DNaseI (P=1.9e^{-10}) or FAIRE (P=5e^{-7}). Moreover, methylation and acetylation histone marks (H3K4ME1 and H3K27Ac) provide signals above the 98th percentile in the same region.

In our current study, we consolidated our previous results[7] in transfected HeLa and HEK293T cells showing that the *T* allele, which is the risk factor for hypertension, is associated with a significant reduction of *eNOS* transcription compared with the *C* allele. We hypothesize that this can impair endothelial NO production in vivo. Luizon et al[19] reported that the rs3918226 polymorphism does not affect plasma nitrite levels in 181 healthy self-reported blacks. However, Luizon et al's results are difficult to interpret, because rs3918226, according to HapMap data, is not polymorphic in blacks. We, therefore, presume a substantial admixture with whites in this black study population. The *T* allele frequency was only 0.04.[19] Thus, the low frequency of the risk-conferring *T* allele and the small sample size probably render a correct estimation of *T* allele effect on plasma nitrite levels impossible.[20]

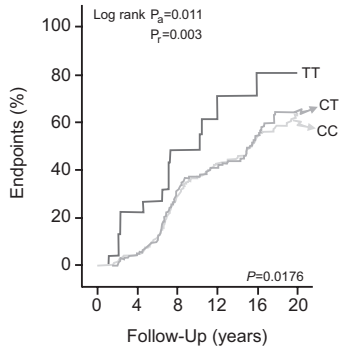


Figure 4. Incidence of hypertension according to rs3918226 genotype. p_a and p_r refer to the significance of the log-rank test according to an additive (*TT* vs *TC* vs *CC*) or a recessive model (*TT* vs *C* allele carriers).

In mammals, NO can be generated by 3 different isoforms of the enzyme NO synthase, referred to as neuronal nNOS (NOS1), inducible iNOS (NOS2) produced by macrophages, and endothelial *eNOS* (NOS3).[21] The human *eNOS* gene spans 21 kb with 26 exons on chromosome 7q35–q36. Blockade of NO synthesis with inhibitory L-arginine analogues leads to peripheral vasoconstriction and a rise in BP.[22–24] Genetically engineered mice with disrupted *eNOS* are hypertensive and have no endothelium-derived relaxant activity.[25] The physiologically most important determinants for the continuous generation of NO and thus the regulation of local blood flow are fluid shear stress and pulsatile stretch.[26] NO dilates all types of blood vessels by stimulating soluble guanyl cyclase and increasing the cGMP concentration in smooth muscle cells.[22] *eNOS* is not only a physiological vasodilator but also conveys vascular protection in several ways.[21,22] NO released toward the vascular lumen is a potent inhibitor of platelet aggregation and adhesion to the vascular wall and prevents the release of platelet-derived growth factors that stimulate smooth muscle proliferation. NO decreases the expression of chemoattractant protein MCP-1 (Monocyte chemoattractant protein) and of a number of surface adhesion molecules and inhibits leukocyte adhesion to vascular endothelium and leukocyte migration into the vascular wall. This offers protection against the early phases of atherosclerosis.[22] The decreased endothelial permeability, the reduced influx of lipoproteins into the vascular wall, and the inhibition of low-density lipoprotein oxidation contribute to the antiatherosclerotic properties of *eNOS*-derived NO. Finally, NO inhibits DNA synthesis and proliferation of vascular smooth muscle cells as well as smooth muscle cell migration, thereby protecting against the later stages of atherogenesis.[22]

Given the central role of *eNOS* in cardiovascular regulation, several previous studies addressed the association between hypertension or cardiovascular disease and genetic variation in *eNOS*. Niu and Qi[27] published a meta-analysis of 3 widely investigated polymorphisms, *G894T* (rs1799983) in exon 7, 4b/a in intron 4, and *T–786C* (rs2070744) in the promoter in relation to hypertension, published in English and Chinese. Overall comparison between allele *894T* and *894G* across all studies (cases/controls: 19 284/26 003) yielded an increased risk of hypertension, amounting to 16% overall, 32% in

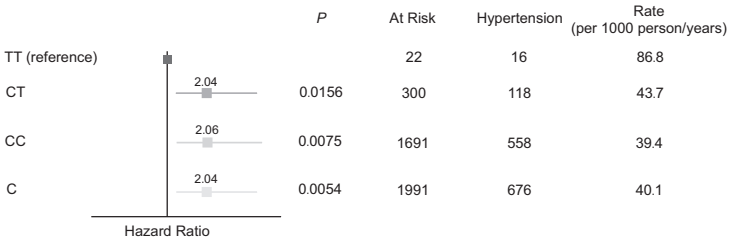


Figure 5. Hazard ratios for hypertension associated with *TT* homozygosity compared with the *CT* or *CC* genotypes or *C* allele carriers. The hazard ratios were adjusted as in the continuous analyses of the blood pressure changes (see Figure 3). Probability values denote the significance of the difference with the *TT* genotype.

Asians, and 40% in Chinese. The risk associated with the *4a* versus *4b* allele was 29% overall and 42% in Asians. For *T–786C*, ethnicity-stratified analyses suggested that in whites the risk of hypertension was 25% and 69% higher in carriers of the *–786C* allele and the *–786CC* genotype, respectively.[27] Moreover, the *T* allele of *T–786C* polymorphism is a predisposing factor to coronary spasm and reduces the *eNOS* promoter activity.[28]

The candidate gene and GWAS studies published so far on hypertension[6] identified 47 distinct genetic variants robustly associated with BP, but collectively these variants explained only a few percent of the heritability of BP. HYPERGENES was the first GWAS to identify rs3918226 (*C–685T*) as a hypertension susceptibility locus.[7] An international consortium applied a gene-centric assay in an independent discovery sample of 25 118 individuals that combined hypertensive case–control and general population samples and followed up 10 suggestive SNPs in a further 59 349 individuals.[6] An analysis of combined discovery and follow-up data identified rs3918226 (*T/C*: 0.08/0.92) as being significantly (2.2×10^{-9}) associated with diastolic BP. The effect size per *–690T* allele was +0.78 mm Hg (SE, 0.21; $P=9.5\times10^{-5}$).[6] The recent GWAS of systolic and diastolic BP performed by the International Consortium for Genome-Wide Association Studies (ICBP) used a multi-stage design in 200 000 individuals of European descent.[29] Overall, the *eNOS* region was poorly covered in this study. Genotyping had been performed in most of the cohorts with arrays older that the Illumina 1M, which do not include rs3918226. Moreover, imputation was done using the HapMap panel as reference that does not include rs3918226. In the ICBP data set,[29] a SNP mapping 779 bp from rs3918226, rs1800783 (position 150689397), shows a high D′ (1.000) but a low R-sq (0.141) with rs3918226. Because of the low R-sq, allele frequencies are different, and rs1800783 cannot be considered a proxy of rs3918226. In GenHAT,[30] the hazard ratio for the primary end point, fatal coronary heart disease, and nonfatal myocardial infarction in *T* allele versus *CC* genotype carriers was 1.12 (CI, 1.00–1.26; $P=0.048$). Conen et al[31] analyzed 3 SNPs in the *eNOS* gene (rs3918226, rs1800779, and rs1799983) in 18 436 white women enrolled in the Women's Health Study. The participants were all health professionals and normotensive at baseline. BP was self-reported. Over 9.8 years, 29.6% of the women developed hypertension. The hazard ratios for the *eNOS* polymorphisms were 1.01 (CI, 0.97–1.06), 1.06 (CI, 0.99–1.14), and 1.05 (CI, 1.01–1.09), respectively.[31] Progression of BP across 3 increasing categories was not associated with the *eNOS* polymorphisms, but follow-up for this soft end point was only 4 years. Seidlerová et al[32] reported a pilot study examining the association between arterial properties and the rs3918226 polymorphism in 101 untreated volunteers. Among 31 smokers, carriers of the mutated *T* allele (n=8) had a marginally higher aortic pulse wave velocity (10.0 versus 8.7 m/s; $P=0.051$) and a higher aortic augmentation index (172 versus 153%; $P=0.024$). Seidlerová et al[32] hypothesized that pending confirmation in a larger study genetic modulation of intermediate arterial phenotypes might lead to higher BP.

Taking into account our current findings and the literature, the *C* to *T* substitution at position –690 in the *eNOS* promoter strongly complies with the Bradford Hill criteria[33] as a cause of hypertension. The association is strong,[6,7,27] consistent across studies,[6,7,27] and specific for hypertension- or hypertension-related complications.[6,7,27,30] The current study established temporality and provided a possible mechanism adding to the plausibility. Some studies, but not ours, suggested a dose effect based on the number of *T* alleles.[6,27]

Perspectives

We demonstrated that *TT* homozygosity at the rs3918226 locus in the *eNOS* gene promoter enhances the age-related increase in BP and increases the risk of hypertension, probably by reducing the transcriptional activity of the *eNOS* gene. The implications of our current findings span both the prevention and treatment of hypertension and its associated cardiovascular complications. The prevalence of *TT* homozygosity is low, explaining why the population-attributable risk for hypertension is only 1.1%. However, the attributable risk in *TT* homozygotes is 51.0%. Combined with other genetic markers, the rs3918226 polymorphism might, therefore, contribute to the stratification of cardiovascular risk. Our findings also support pharmacological interference with the NO signaling pathway. In GenHAT,[30] amlodipine, compared with lisinopril, was more effective in the prevention of stroke in minor allele carriers (hazard ratios *CT*+*TT* versus *CC*: 0.49 versus 0.85; $P=0.04$). Overall, amlodipine reduced systolic BP 1.2 mm Hg more than lisinopril.[34] Stroke is the complication of hypertension that is most closely linked to the BP level.[1] Moreover, amlodipine enhances endothelial NO availability via stimulation of NO formation[35] and by prolonging the NO half-life through antioxidative properties.[35,36] Recently developed compounds act downstream in the NO signaling pathway by inhibition of cGMP-specific phosphodiesterase type 5[37] or by stimulation (haem-dependent) or activation (NO- and haem-independent) of soluble guanylate cyclase activity.[38] Further clinical research should establish whether *eNOS* might be a target for preventive or therapeutic intervention.

Acknowledgments

Dr Charles J. Lowenstein (University of Rochester Medical Center) provided the pGL2-eNOS promoter-luciferase plasmid. Dr Cédric Howald (University of Lausanne) and ENCODE researchers helped with the open chromatin analysis. Sandra Covens provided expert clerical assistance.

Sources of Funding

The European Union (grant FP7-HEALTH-2007-A-201550), HYPERGENES, and InterOmics (PB05 MIUR-CNR Italian Flagship Project) provided financial support for the genotyping and experimental studies. The European Union (grants IC15-CT98-0329-EPOGH, LSHM-CT-2006-037093-InGenious HyperCare, FP7-HEALTH-2007-A-201550- HYPERGENES, HEALTH-2011.2.4.2-2-EU-MASCARA, HEALTH-F7-305507 HOMAGE, and the European Research Council Advanced Researcher Grant-2011-294713-EPLORE) gave support to the Studies Coordinating Centre, Leuven, Belgium, and the FLEMENGHO and EPOGH population studies. The Fonds voor Wetenschappelijk Onderzoek Vlaanderen, Ministry of the Flemish Community, Brussels, Belgium (G.0734.09, G.0881.13 and G.0880.13), also supported the FLEMENGHO study.

Disclosures

None.

References

1. Staessen JA, Wang J, Bianchi G, Birkenhäger WH. Essential hypertension. *Lancet*. 2003;361:1629–1641.
2. Staessen JA, Kuznetsova T, Stolarz K. Hypertension prevalence and stroke mortality across populations. *JAMA*. 2003;289:2420–2422.
3. Prospective Studies Collaboration. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet*. 2002;360:1903–1913.
4. Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJ. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *Lancet*. 2006;367:1747–1757.
5. Newton-Cheh C, Johnson T, Gateva V, et al; Wellcome Trust Case Control Consortium. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet*. 2009;41:666–676.
6. Johnson T, Gaunt TR, Newhouse SJ, et al; Cardiogenics Consortium; Global BPgen Consortium. Blood pressure loci identified with a gene-centric array. *Am J Hum Genet*. 2011;89:688–700.
7. Salvi E, Kutalik Z, Glorioso N, et al. Genome-wide association study using a high-density SNP-array and case-control design identifies a novel hypertension susceptibility locus in the promoter region of eNOS. *Hypertension*. 2012;59:248–255.
8. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. *Annu Rev Genomics Hum Genet*. 2009;10:387–406.
9. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*. 2010;34:816–834.
10. Staessen JA, Wang JG, Brand E, Barlassina C, Birkenhäger WH, Herrmann SM, Fagard R, Tizzoni L, Bianchi G. Effects of three candidate genes on prevalence and incidence of hypertension in a Caucasian population. *J Hypertens*. 2001;19:1349–1358.
11. Stolarz-Skrzypek K, Kuznetsova T, Thijs L, Tikhonoff V, Seidlerová J, Richart T, Jin Y, Olszanecka A, Malyutina S, Casiglia E, Filipovský J, Kawecka-Jaszcz K, Nikitin Y, Staessen JA, on behalf of the European Project on Genes in Hypertension (EPOGH) Investigators. Fatal and nonfatal outcomes, incidence of hypertension and blood pressure changes in relation to urinary sodium excretion in White Europeans. *JAMA*. 2011;305:1777–1785.
12. Kuznetsova T, Staessen JA, Kawecka-Jaszcz K, Babeanu S, Casiglia E, Filipovsky J, Nachev C, Nikitin Y, Peleskã J, O'Brien E. Quality control of the blood pressure phenotype in the European Project on Genes in Hypertension. *Blood Press Monit*. 2002;7:215–224.
13. World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Bull World Health Organ*. 2001;79:373–374.
14. Lurbe E, Cifkova R, Cruickshank JK, et al; European Society of Hypertension. Management of high blood pressure in children and adolescents: recommendations of the European Society of Hypertension. *J Hypertens*. 2009;27:1719–1742.
15. Holtzman NA, Marteau TM. Will genetics revolutionize medicine? *N Engl J Med*. 2000;343:141–144.
16. Matys V, Fricke E, Geffers R, et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*. 2003;31:374–378.
17. Akiyama T. *TFSEARCH: Searching Transcription Factor Binding Sites*, ver 1.3 (accessed 16 December 2012).
18. Karantzoulis-Fegaras F, Antoniou H, Lai SL, Kulkarni G, D'Abreo C, Wong GK, Miller TL, Chan Y, Atkins J, Wang Y, Marsden PA. Characterization of the human endothelial nitric-oxide synthase promoter. *J Biol Chem*. 1999;274:3076–3093.
19. Luizon MR, Metzger IF, Lacchini R, Tanus-Santos JE. Endothelial nitric oxide synthase polymorphism rs3918226 associated with hypertension does not affect plasma nitrite levels in healthy subjects. *Hypertension*. 2012;59:e52; author reply e53.
20. Salvi E, Cusi D. Response to endothelial nitric oxide synthase polymorphism rs3918226 associated with hypertension does not affect plasma nitrite levels in healthy subjects. *Hypertension*. 2012;59:e53.
21. Förstermann U, Sessa WC. Nitric oxide synthases: regulation and function. *Eur Heart J*. 2012;33:829–837, 837a.
22. Li H, Förstermann U. Nitric oxide in the pathogenesis of vascular disease. *J Pathol*. 2000;190:244–254.
23. Haynes WG, Noon JP, Walker BR, Webb DJ. Inhibition of nitric oxide synthesis increases blood pressure in healthy humans. *J Hypertens*. 1993;11:1375–1380.
24. Rees DD, Palmer RM, Moncada S. Role of endothelium-derived nitric oxide in the regulation of blood pressure. *Proc Natl Acad Sci U S A*. 1989;86:3375–3378.
25. Huang PL, Huang Z, Mashimo H, Bloch KD, Moskowitz MA, Bevan JA, Fishman MC. Hypertension in mice lacking the gene for endothelial nitric oxide synthase. *Nature*. 1995;377:239–242.
26. Fleming I, Busse R. Molecular mechanisms involved in the regulation of the endothelial nitric oxide synthase. *Am J Physiol Regul Integr Comp Physiol*. 2003;284:R1–12.
27. Niu W, Qi Y. An updated meta-analysis of endothelial nitric oxide synthase gene: three well-characterized polymorphisms with hypertension. *PLoS One*. 2011;6:e24266.
28. Nakayama M, Yasue H, Yoshimura M, Shimasaki Y, Kugiyama K, Ogawa H, Motoyama T, Saito Y, Ogawa Y, Miyamoto Y, Nakao K. T-786–>C mutation in the 5'-flanking region of the endothelial nitric oxide synthase gene is associated with coronary spasm. *Circulation*. 1999;99:2864–2870.
29. The International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret GB, Munroe PB, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*. 2011;478:103–109.
30. Zhang X, Lynch AI, Davis BR, Ford CE, Boerwinkle E, Eckfeldt JH, Leiendecker-Foster C, Arnett DK. Pharmacogenetic associations of NOS3 variants with cardiovascular disease in patients with hypertension: the GenHAT Study. *PLoS ONE*. 2012;7:e34217. doi:10.1371/journal.pone.0034217.
31. Conen D, Glynn RJ, Buring JE, Ridker PM, Zee RY. Association of renin-angiotensin and endothelial nitric oxide synthase gene polymorphisms with blood pressure progression and incident hypertension: prospective cohort study. *J Hypertens*. 2008;26:1780–1786.
32. Seidlerová J, Filipovský J, Mayer O, Jr, Cífková R, Pešta M, Blatný R, Vanek J. Association between endothelial NO synthase polymorphism (rs3918226) and arterial properties. *Artery Res*. 2013;7:54–59.
33. Bradford-Hill A. The environment and disease: association or causation. *Proc Royal Soc Med*. 1965;58:295–300.
34. The ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic. The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). *JAMA*. 2003;288:2981–2997.

35. Berkels R, Taubert D, Bartels H, Breitenbach T, Klaus W, Roesen R. Amlodipine increases endothelial nitric oxide by dual mechanisms. *Pharmacology*. 2004;70:39–45.
36. Ganafa AA, Walton M, Eatman D, Abukhalaf IK, Bayorh MA. Amlodipine attenuates oxidative stress-induced hypertension. *Am J Hypertens*. 2004;17:743–748.
37. Reffelmann T, Kloner RA. Therapeutic potential of phosphodiesterase 5 inhibition for cardiovascular disease. *Circulation*. 2003;108:239–244.
38. Stasch JP, Pacher P, Evgenov OV. Soluble guanylate cyclase as an emerging therapeutic target in cardiopulmonary disease. *Circulation*. 2011;123:2263–2273.

Novelty and Significance

What Is New?

- In the general population, *TT* homozygosity at the rs3918226 locus (minor/major allele, *T/C* allele) in the *eNOS* gene promoter enhances the age-related increase in blood pressure and increases the risk of hypertension. Sequencing confirmed rs3918226, a binding site of E-twenty six transcription factors, as the SNP most closely associated with hypertension. In luciferase reporter assays, the risk-carrying *T* allele was associated with a 20% to 40% lower transcriptional activity than the *C* allele.

What Is Relevant?

- The prevalence of *TT* homozygosity is low, explaining why the population-attributable risk for hypertension is only 1.1%. However, the attributable risk in *TT* homozygotes is 51.0%. Combined with other genetic markers, the rs3918226 polymorphism might, therefore, contribute to the stratification of cardiovascular risk. Further clinical research should establish whether *eNOS* might be a target for preventive or therapeutic intervention.

Summary

rs3918226 in the *eNOS* promoter tags a hypertension susceptibility locus, *TT* homozygosity being associated with lesser transcription and higher risk of hypertension.

DISCUSSION

The years during which the experimental work for this thesis was made coincided with the most extraordinary technological change that occurred to medical genetics since decades. Next generation sequencing transformed all aspects of DNA investigations, starting from the sequencing of small genomes to the sequencing of human genomes and molecular diagnostics of inheritable disease. More specifically, I started working on this project during the early phases of implementation of NGS in molecular genetics research. In our first NGS-related work, we were interested in exploring the use of different NGS platforms (from Roche, Illumina and Life Technologies) for the sequencing of simple targets. A mutation in an intronic repetitive element of *PRPF31*, representing an obstacle for both classic sequencing methods and NGS, was chosen as a test case. We showed that all NGS platforms are powerful tools for molecular geneticist to identify rare and common DNA variants, even in case of more complex sequences. Moreover, we highlighted the features of different NGS platforms to be considered in re-sequencing projects. In particular, the identification of variants in repetitive elements was facilitated by the use of longer reads, such as the ones produced by Roche 454 Technology. Sequencing of repetitive regions has been a long lasting problem for finishing the sequence of genomes in many organisms, and has not been completely solved yet. In the human genome, for example, highly repetitive regions like centromeres, telomeres and Y chromosome are still scarcely covered and annotated [121, 122]. To correct the ambiguity of non-unique alignment of identical short reads on reference genomes, the use of paired-end reads and of specific bioinformatic tools have provided some help [123]. One of the most used strategies, especially for variant detection, is to simply ignore reads that map to multiple locations, with the awareness that only non-redundant sequences can be analyzed. In exome sequencing, for example, repetitive regions are not captured, limiting the efficacy of this methodology for comprehensive variant analysis. The homozygous 353 bp *Alu* inserion in the RP gene *MAK*, for example, has been identified by exome sequencing only due to a fortunate coincidence and because two different sequencing platforms were used [98]. Another recent example is the discovery of a single cytosine insertion in the gene *MUC1* causing the dominant rare disease medullary cystic kidney disease type 1 [124]. Despite linkage analysis clearly pointed to a defined locus, whole-exome (WES), sequence-capture

and whole-genome sequencing (WGS) of several patients failed to identify the causative mutation. Only by cloning, Sanger sequencing and *de novo* assembly of the specific gene it was possible to reconstruct the polymorphic repetitive element, a VNTR, where the mutation lied. Interestingly, this scenario is very similar to our original case of *PRPF31*, although the polymorphic number of repeats in *PRPF31* was smaller (6-7) than in *MUC1* (>30). The problem of repeats is still topical and should not be underestimated when analyzing whole genome or exome sequences, which often result in no candidate genes or mutation relevant for a particular disease.

Despite this and other limitations, WES or WGS are very powerful tools to discover new genes causing Mendelian diseases, including RP and especially for conditions that have recessive inheritance [81]. The constant plummeting of their costs allowed them to gradually substitute almost all other genetic approaches for research and diagnostic. However, for the study of dominant retinitis pigmentosa, we decided to use a "classic" candidate gene approach and not WES/WGS because: i) WES/WGS are less effective in the identification of dominant mutations in heterogeneous diseases, due to the high number of novel hetherozygous variants present at the genomic scale in each individual, ii) the cost was still too high to be applied to a elevated number of patients and their family members, iii) the candidate gene approach is an effective tool for discovering new genes in heterogeneous diseases such as RP [125]. We therefore performed NGS-based screenings of multiple samples to sequence candidate splicing factor genes in well-characterized cohorts of patients with adRP. From a methodological point of view, the main challenge was to adapt NGS procedures to multiple samples processing. For the first gene to screen, *SNRNP200*, as a cost-effective solution we sequenced all target PCRs from all the patients as a single library, thus avoiding the costs of many sequencing library preparations. This method required the development of downstream analysis and validation by Sanger sequencing, aimed at distinguishing true mutations from false positives, the latter being very frequent with the Roche 454 pipeline used in our analysis. Following the commercialization of cheaper and more scalable library preparation solutions (such as the "Nextera" kit), for screening of the second set of candidate splicing factor genes (*EFTUD2, PRPF4*) we sequenced PCRs from all samples by individual barcoded library preparations. This procedure allowed obtaining results that were less poisoned by false positive and negative discoveries. Overall, the approaches used in our screenings proved to be cost- and time-effective solutions for gene analysis in many samples. Both procedures allowed exploiting the full capacity of modern NGS sequencers and obtaining reliable results

for molecular analysis, especially when individual library preps were used. However, since the generation of many long-range PCRs is a rather tedious and low-throughput procedure, other enrichment methods such as micro-droplet PCR [126], microfluidic arrays-based PCRs [127] or molecular inversion probes [128] can be now used for faster and more scalable custom target selection.

Screening of these splicing factor genes confirmed *SNRNP200* as a novel RP gene, which was identified by other authors in two different families [38] [37]. Our study highlighted the effectiveness of screening large cohorts of patients with NGS to identify new RP genes and to enrich the mutational panel of disease genes, important also for diagnostics. Conversely, the genes analyzed in the second round of sequencing (*EFTUD2, PRPF4, NHP2L1* and *AAR2*) did not bear any pathogenic variant to be linked to RP, likely because these genes are not indeed causing RP or mutations are too rare to be identified by screening a small number of cohorts. WES and WGS are more expensive alternatives to candidate gene screenings and present the advantage to be hypothesis-free. However, applying these methods on large cohorts of patients has higher necessities than single candidate gene sequencing. For example, mutations in known genes have to be excluded in order to be cost-effective. Despite this precaution sounds obvious, in practical terms it is very hard to insure such pure enrichment because patients are usually collected and screened for known genes over a large period of time and with different methods. Custom NGS- based panels of genes for targeted enrichment are a good option for homogeneous and up-to-date screenings of known retinal degeneration genes, under the condition that the price per sample becomes significantly lower than WES itself [96, 97, 129]. Finally, for dominant RP, the analysis of families or trios may be critical to filter candidate variants from WES/WGS very large lists of DNA changes and to compensate for lack of commonly mutated genes in unrelated patients.

The hBrr2 protein, encoded by *SNRNP200*, is the core RNA helicase associated with the U5 snRNA. It is responsible for the disruption of U4/U6 base pairing necessary for spliceosomal activation and for the U2/U6 base pairing for spliceosomal disassembly and recycling. The dynamics of Brr2's functions are tightly and directly regulated by the PRPF8 C-terminal tail [56] and by the GTPase Snu114 (encoded by *EFTUD2* in humans) [130]. AAR2 is a small interactor of PRPF8 in the immature U5 snRNP and prevents PRPF8's binding to Brr2, which replaces AAR2 when the mature tri-snRNP is formed in the nucleus [131]. The genes *EFTUD2* and *AAR2* were chosen as candidate adRP genes because they are involved in the regulation of Brr2 activity and recruitment on the U5 snRNP, respectively. However, our

screening indicated that these genes are likely not involved in the disease. In particular, many heterozygous mutations of *EFTUD2*, the strongest candidate from our screening, have been linked to another set of congenital pathologies characterized by severe cranio-facial malformations [30]. This latter finding suggests that there might be different mechanisms underlying the development of these diseases and that we are still missing some key elements to link mutations of splicing factors to RP. Among all possible hypotheses, the one supported by the most numerous and convincing evidences indicates splicing defects as the triggering cause for degeneration of photoreceptors, which are unusually sensitive to splicing impairments. This theory implies a loss-of-function model, which is supported by several biochemical studies on tri-snRNPs associated mutations [35, 62, 68, 132] and by the haploinsufficiency of *PRPF31* mutations, which is also reflected in the many cases of incomplete penetrance [41-43]. Interestingly, we have indeed found a novel case of incomplete penetrance in one family segregating the *SNRNP200* p.R681C mutation. The affected members presented with typical symptoms of RP, but with variable severity, indicating variable expressivity as well. Two members of the family were healthy carriers of the mutation, and had no signs of RP, even at old ages. We started to investigate the causes of incomplete penetrance at the molecular level, by testing variations in expression of *SNRNP200*. We could only identify a possible reduction in total protein concentration at the steady-state in lymphoblastoid cell lines from carriers of p.R681C mutation. We plan to verify a possible mechanism of protein misfolding and degradation due to point mutations by measuring the protein levels in conditions where protein synthesis or proteasome machinery is blocked. Unless differences between asymptomatic and affected individuals are visible, this will not give clues on incomplete penetrance mechanism, but may confirm the haploinsufficiency model. Other factors - genetic or environmental - are likely to differentially influence hBrr2 expression or activity, and their identification could have a major impact on our understanding of the molecular basis of the disease, ultimately useful for clinical care.

Current animal models of haploinsufficient RP-linked splicing factors have not been conclusive to provide a mechanistic model for photoreceptor degeneration, either because the phenotype was not present (as in the case of mouse models [67]) or because proper controls have not been provided (in the case of the Zebrafish model [48]). Mammal cellular models include transfected human cell lines (such as HeLa and RPE-derived cells [132, 133]) or patients-derived lymphoblastoid cells [68, 134]. Although they have provided important

information, these cell models have been criticized because they are not relevant for modeling retina-specific degeneration in RP. An emerging tool to study the effect of mutations at the cellular level is induced pluripotent stem cells (iPSCs). Indeed, patient-derived iPSCs differentiated into photoreceptors or photoreceptor precursors have been generated for a series of mutations in RP genes such as *RP1*, *RP9*, *PRPH2, RHO, USH2A* and *MAK*, helping to study photoreceptor degeneration pathways and possible clinical interventions [98, 135-137]. The validation in these cell models, which recapitulate the disease phenotype in specific genetic background and in a relevant tissue, will undoubtedly help to provide final prove of pathogenic mechanisms of splicing factors as well as potential targeted therapeutic strategies.

The technological innovation in sequencing and genotyping technologies had a strong impact not only in the investigation of Mendelian disease but also in genetic research of complex diseases. The exceptional throughput and speed of sequencing allowed gaining deeper knowledge of human genomic variations and functional properties, useful in genome-wide association studies. For example, the 1000 Genomes Project has sequenced 2500 individuals from 25 worldwide populations, and catalogued common and rare variations. At the same time, the functional aspects of human genome were explored by the ENCODE Project [112], which performed extensive characterization of different cell types to annotate the human genome with regulatory regions such as enhancers, promoters and silencing regions, important for the functional characterization of GWAS signals. In our GWAS on hypertension, we identified a novel association with a regulatory SNP in the promoter region of *NOS3*. The risk allele determined a reduced transcription of endothelial nitric oxide synthase, and consequently a reduction of the vascular relaxing factor NO, predisposing to higher blood pressure. Furthermore, targeted NGS of the region surrounding the SNP in a panel of healthy and hypertensive individuals confirmed the rs3918226 polymorphism as the most strongly associated with hypertension.

In conclusion, massively parallel sequencing protocols allow the rapid and low-cost investigation of Mendelian and complex diseases on a scale not previously imaginable. We are already realizing that the technological progress that gave us access to unprecedented amounts of genomic information will crash against our still limited understanding of how specific variants cause or influence a disease. The necessity of valid models to test the impact of DNA changes that are constantly being identified by next generation genetic studies will be a key focus of future and current genetic research. This will ultimately open the way to the improvement of therapeutic strategies and personalized medicine.

BIBLIOGRAPHY

1. Sung, C.H. and J.Z. Chuang, *The cell biology of vision.* J Cell Biol, 2010. **190**(6): p. 953-63.
2. Randlett, O., C. Norden, and W.A. Harris, *The vertebrate retina: a model for neuronal polarization in vivo.* Dev Neurobiol, 2011. **71**(6): p. 567-83.
3. Kolb, H., *How the retina works.* American Scientist, 2003. **91**: p. 28-35.
4. van Soest, S., et al., *Retinitis pigmentosa: defined from a molecular point of view.* Surv Ophthalmol, 1999. **43**(4): p. 321-34.
5. Konieczka, K., et al., *Retinitis pigmentosa and ocular blood flow.* EPMA J, 2012. **3**(1): p. 17.
6. Wright, A.F., et al., *Photoreceptor degeneration: genetic and mechanistic dissection of a complex trait.* Nat Rev Genet, 2010. **11**(4): p. 273-84.
7. Hartong, D.T., E.L. Berson, and T.P. Dryja, *Retinitis pigmentosa.* Lancet, 2006. **368**(9549): p. 1795-809.
8. den Hollander, A.I., et al., *Lighting a candle in the dark: advances in genetics and gene therapy of recessive retinal dystrophies.* J Clin Invest, 2010. **120**(9): p. 3042-53.
9. Rivolta, C., et al., *Retinitis pigmentosa and allied diseases: numerous diseases, genes, and inheritance patterns.* Hum Mol Genet, 2002. **11**(10): p. 1219-27.
10. Hamel, C., *Retinitis pigmentosa.* Orphanet J Rare Dis, 2006. **1**: p. 40.
11. Berson, E.L., *Retinitis pigmentosa. The Friedenwald Lecture.* Invest Ophthalmol Vis Sci, 1993. **34**(5): p. 1659-76.
12. Daiger, S.P., L.S. Sullivan, and S.J. Bowne, *Genes and mutations causing retinitis pigmentosa.* Clin Genet, 2013. **84**(2): p. 132-41.
13. Al-Maghtheh, M., et al., *Evidence for a major retinitis pigmentosa locus on 19q13.4 (RP11) and association with a unique bimodal expressivity phenotype.* Am J Hum Genet, 1996. **59**(4): p. 864-71.
14. Venturini, G., et al., *CNOT3 is a modifier of PRPF31 mutations in retinitis pigmentosa with incomplete penetrance.* PLoS Genet, 2012. **8**(11): p. e1003040.
15. Nishiguchi, K.M. and C. Rivolta, *Genes associated with retinitis pigmentosa and allied diseases are frequently mutated in the general population.* PLoS One, 2012. **7**(7): p. e41902.
16. Sharp, P.A., *Split genes and RNA splicing.* Cell, 1994. **77**(6): p. 805-15.
17. Kalsotra, A. and T.A. Cooper, *Functional consequences of developmentally regulated alternative splicing.* Nat Rev Genet, 2011. **12**(10): p. 715-29.
18. Kramer, A., *The structure and function of proteins involved in mammalian pre-mRNA splicing.* Annu Rev Biochem, 1996. **65**: p. 367-409.
19. Kruger, K., et al., *Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena.* Cell, 1982. **31**(1): p. 147-57.
20. Staley, J.P. and C. Guthrie, *Mechanical devices of the spliceosome: motors, clocks, springs, and things.* Cell, 1998. **92**(3): p. 315-26.
21. Singh, R.K. and T.A. Cooper, *Pre-mRNA splicing in disease and therapeutics.* Trends Mol Med, 2012. **18**(8): p. 472-82.

22. Patel, A.A. and J.A. Steitz, *Splicing double: insights from the second spliceosome.* Nat Rev Mol Cell Biol, 2003. **4**(12): p. 960-70.
23. Will, C.L. and R. Luhrmann, *Spliceosome structure and function.* Cold Spring Harb Perspect Biol, 2011. **3**(7).
24. Cordin, O. and J.D. Beggs, *RNA helicases in splicing.* RNA Biol, 2012. **10**(1): p. 83-95.
25. Stenson, P.D., et al., *Human Gene Mutation Database (HGMD): 2003 update.* Hum Mutat, 2003. **21**(6): p. 577-81.
26. Makishima, H., et al., *Mutations in the spliceosome machinery, a novel and ubiquitous pathway in leukemogenesis.* Blood, 2012. **119**(14): p. 3203-10.
27. Hamilton, G. and T.H. Gillingwater, *Spinal muscular atrophy: going beyond the motor neuron.* Trends Mol Med, 2013. **19**(1): p. 40-50.
28. Cooper, T.A., L. Wan, and G. Dreyfuss, *RNA and disease.* Cell, 2009. **136**(4): p. 777-93.
29. Bernier, F.P., et al., *Haploinsufficiency of SF3B4, a component of the pre-mRNA spliceosomal complex, causes Nager syndrome.* Am J Hum Genet, 2012. **90**(5): p. 925-33.
30. Lines, M.A., et al., *Haploinsufficiency of a spliceosomal GTPase encoded by EFTUD2 causes mandibulofacial dysostosis with microcephaly.* Am J Hum Genet, 2012. **90**(2): p. 369-77.
31. Gordon, C.T., et al., *EFTUD2 haploinsufficiency leads to syndromic oesophageal atresia.* J Med Genet, 2012. **49**(12): p. 737-46.
32. Liu, S., et al., *The network of protein-protein interactions within the human U4/U6.U5 tri-snRNP.* RNA, 2006. **12**(7): p. 1418-30.
33. Vithana, E.N., et al., *A human homolog of yeast pre-mRNA splicing gene, PRP31, underlies autosomal dominant retinitis pigmentosa on chromosome 19q13.4 (RP11).* Mol Cell, 2001. **8**(2): p. 375-81.
34. McKie, A.B., et al., *Mutations in the pre-mRNA splicing factor gene PRPC8 in autosomal dominant retinitis pigmentosa (RP13).* Hum Mol Genet, 2001. **10**(15): p. 1555-62.
35. Chakarova, C.F., et al., *Mutations in HPRP3, a third member of pre-mRNA splicing factor genes, implicated in autosomal dominant retinitis pigmentosa.* Hum Mol Genet, 2002. **11**(1): p. 87-92.
36. Keen, T.J., et al., *Mutations in a protein target of the Pim-1 kinase associated with the RP9 form of autosomal dominant retinitis pigmentosa.* Eur J Hum Genet, 2002. **10**(4): p. 245-9.
37. Li, N., et al., *Mutations in ASCC3L1 on 2q11.2 are associated with autosomal dominant retinitis pigmentosa in a Chinese family.* Invest Ophthalmol Vis Sci, 2010. **51**(2): p. 1036-43.
38. Zhao, C., et al., *Autosomal-dominant retinitis pigmentosa caused by a mutation in SNRNP200, a gene required for unwinding of U4/U6 snRNAs.* Am J Hum Genet, 2009. **85**(5): p. 617-27.
39. Tanackovic, G., et al., *A missense mutation in PRPF6 causes impairment of pre-mRNA splicing and autosomal-dominant retinitis pigmentosa.* Am J Hum Genet, 2011. **88**(5): p. 643-9.
40. Vithana, E., et al., *RP11 is the second most common locus for dominant retinitis pigmentosa.* J Med Genet, 1998. **35**(2): p. 174-5.
41. Rio Frio, T., et al., *Premature termination codons in PRPF31 cause retinitis pigmentosa via haploinsufficiency due to nonsense-mediated mRNA decay.* J Clin Invest, 2008. **118**(4): p. 1519-31.

42. Vithana, E.N., et al., *Expression of PRPF31 mRNA in patients with autosomal dominant retinitis pigmentosa: a molecular clue for incomplete penetrance?* Invest Ophthalmol Vis Sci, 2003. **44**(10): p. 4204-9.
43. Rivolta, C., et al., *Variation in retinitis pigmentosa-11 (PRPF31 or RP11) gene expression between symptomatic and asymptomatic patients with dominant RP11 mutations.* Hum Mutat, 2006. **27**(7): p. 644-53.
44. Weidenhammer, E.M., M. Ruiz-Noriega, and J.L. Woolford, Jr., *Prp31p promotes the association of the U4/U6 x U5 tri-snRNP with prespliceosomes to form spliceosomes in Saccharomyces cerevisiae.* Mol Cell Biol, 1997. **17**(7): p. 3580-8.
45. Makarova, O.V., et al., *Protein 61K, encoded by a gene (PRPF31) linked to autosomal dominant retinitis pigmentosa, is required for U4/U6*U5 tri-snRNP formation and pre-mRNA splicing.* EMBO J, 2002. **21**(5): p. 1148-57.
46. Schaffert, N., et al., *RNAi knockdown of hPrp31 leads to an accumulation of U4/U6 di-snRNPs in Cajal bodies.* EMBO J, 2004. **23**(15): p. 3000-9.
47. Bujakowska, K., et al., *Study of gene-targeted mouse models of splicing factor gene Prpf31 implicated in human autosomal dominant retinitis pigmentosa (RP).* Invest Ophthalmol Vis Sci, 2009. **50**(12): p. 5927-33.
48. Linder, B., et al., *Systemic splicing factor deficiency causes tissue-specific defects: a zebrafish model for retinitis pigmentosa.* Hum Mol Genet, 2011. **20**(2): p. 368-77.
49. Greenberg, J., et al., *A new locus for autosomal dominant retinitis pigmentosa on the short arm of chromosome 17.* Hum Mol Genet, 1994. **3**(6): p. 915-8.
50. Maubaret, C.G., et al., *Autosomal dominant retinitis pigmentosa with intrafamilial variability and incomplete penetrance in two families carrying mutations in PRPF8.* Invest Ophthalmol Vis Sci, 2011. **52**(13): p. 9304-9.
51. Grainger, R.J. and J.D. Beggs, *Prp8 protein: at the heart of the spliceosome.* RNA, 2005. **11**(5): p. 533-57.
52. Pena, V., et al., *Common design principles in the spliceosomal RNA helicase Brr2 and in the Hel308 DNA helicase.* Mol Cell, 2009. **35**(4): p. 454-66.
53. Zhang, L., et al., *Structural evidence for consecutive Hel308-like modules in the spliceosomal ATPase Brr2.* Nat Struct Mol Biol, 2009. **16**(7): p. 731-9.
54. Maeder, C., A.K. Kutach, and C. Guthrie, *ATP-dependent unwinding of U4/U6 snRNAs by the Brr2 helicase requires the C terminus of Prp8.* Nat Struct Mol Biol, 2009. **16**(1): p. 42-8.
55. Lauber, J., et al., *The human U4/U6 snRNP contains 60 and 90kD proteins that are structurally homologous to the yeast splicing factors Prp4p and Prp3p.* RNA, 1997. **3**(8): p. 926-41.
56. Mozaffari-Jovin, S., et al., *Inhibition of RNA Helicase Brr2 by the C-Terminal Tail of the Spliceosomal Protein Prp8.* Science, 2013.
57. Boon, K.L., et al., *prp8 mutations that cause human retinitis pigmentosa lead to a U5 snRNP maturation defect in yeast.* Nat Struct Mol Biol, 2007. **14**(11): p. 1077-83.
58. Zhao, C., et al., *A novel locus (RP33) for autosomal dominant retinitis pigmentosa mapping to chromosomal region 2cen-q12.1.* Hum Genet, 2006. **119**(6): p. 617-23.
59. Benaglio, P., et al., *Next generation sequencing of pooled samples reveals new SNRNP200 mutations associated with retinitis pigmentosa.* Hum Mutat, 2011. **32**(6): p. E2246-58.
60. Santos, K.F., et al., *Structural basis for functional cooperation between tandem helicase cassettes in Brr2-mediated remodeling of the spliceosome.* Proc Natl Acad Sci U S A, 2012. **109**(43): p. 17418-23.
61. Gonzalez-Santos, J.M., et al., *Central region of the human splicing factor Hprp3p interacts with Hprp4p.* J Biol Chem, 2002. **277**(26): p. 23764-72.

62. Comitato, A., et al., *Mutations in splicing factor PRPF3, causing retinal degeneration, form detrimental aggregates in photoreceptor cells.* Hum Mol Genet, 2007. **16**(14): p. 1699-707.
63. Maita, H., et al., *Association of PAP-1 and Prp3p, the products of causative genes of dominant retinitis pigmentosa, in the tri-snRNP complex.* Exp Cell Res, 2005. **302**(1): p. 61-8.
64. Inglehearn, C.F., et al., *A new locus for autosomal dominant retinitis pigmentosa on chromosome 7p.* Nat Genet, 1993. **4**(1): p. 51-3.
65. Kim, R.Y., et al., *Autosomal dominant retinitis pigmentosa mapping to chromosome 7p exhibits variable expression.* Br J Ophthalmol, 1995. **79**(1): p. 23-7.
66. Makarov, E.M., et al., *The human homologue of the yeast splicing factor prp6p contains multiple TPR elements and is stably associated with the U5 snRNP via protein-protein interactions.* J Mol Biol, 2000. **298**(4): p. 567-75.
67. Graziotto, J.J., et al., *Three gene-targeted mouse models of RNA splicing factor RP show late-onset RPE and retinal degeneration.* Invest Ophthalmol Vis Sci, 2011. **52**(1): p. 190-8.
68. Tanackovic, G., et al., *PRPF mutations are associated with generalized defects in spliceosome formation and pre-mRNA splicing in patients with retinitis pigmentosa.* Hum Mol Genet, 2011. **20**(11): p. 2116-30.
69. Cao, H., et al., *Temporal and Tissue Specific Regulation of RP-Associated Splicing Factor Genes PRPF3, PRPF31 and PRPC8-Implications in the Pathogenesis of RP.* PLoS One, 2011. **6**(1): p. e15860.
70. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.
71. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
72. Wetterstrand, K.A., *DNA sequencing costs: data from the NHGRI large-scale genome sequencing program.* http://www.genome.gov/sequencingcosts/, 2012.
73. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
74. International Human Genome Sequencing Consortium., *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
75. Shendure, J., et al., *Advanced sequencing technologies: methods and goals.* Nat Rev Genet, 2004. **5**(5): p. 335-44.
76. Shendure, J. and E. Lieberman Aiden, *The expanding scope of DNA sequencing.* Nat Biotechnol, 2012. **30**(11): p. 1084-94.
77. Choi, M., et al., *Genetic diagnosis by whole exome capture and massively parallel DNA sequencing.* Proc Natl Acad Sci U S A, 2009. **106**(45): p. 19096-101.
78. Korf, B.R. and H.L. Rehm, *New approaches to molecular diagnosis.* JAMA, 2013. **309**(14): p. 1511-21.
79. Ng, S.B., et al., *Massively parallel sequencing and rare disease.* Hum Mol Genet, 2010. **19**(R2): p. R119-24.
80. Kuhlenbaumer, G., J. Hullmann, and S. Appenzeller, *Novel genomic techniques open new avenues in the analysis of monogenic disorders.* Hum Mutat, 2010. **32**(2): p. 144-51.
81. Gilissen, C., et al., *Unlocking Mendelian disease using exome sequencing.* Genome Biol, 2011. **12**(9): p. 228.
82. Mamanova, L., et al., *Target-enrichment strategies for next-generation sequencing.* Nat Methods, 2010. **7**(2): p. 111-8.

83. Mitra, R.D. and G.M. Church, *In situ localized amplification and contact replication of many individual DNA molecules.* Nucleic Acids Res, 1999. **27**(24): p. e34.
84. Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1135-45.
85. Fedurco, M., et al., *BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies.* Nucleic Acids Res, 2006. **34**(3): p. e22.
86. Turcatti, G., et al., *A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis.* Nucleic Acids Res, 2008. **36**(4): p. e25.
87. McKernan, K.J., et al., *Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding.* Genome Res, 2009. **19**(9): p. 1527-41.
88. Ewing, B. and P. Green, *Base-calling of automated sequencer traces using phred. II. Error probabilities.* Genome Res, 1998. **8**(3): p. 186-94.
89. Ewing, B., et al., *Base-calling of automated sequencer traces using phred. I. Accuracy assessment.* Genome Res, 1998. **8**(3): p. 175-85.
90. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.
91. Rio Frio, T., et al., *A single-base substitution within an intronic repetitive element causes dominant retinitis pigmentosa with reduced penetrance.* Hum Mutat, 2009. **30**(9): p. 1340-7.
92. Vache, C., et al., *Usher syndrome type 2 caused by activation of an USH2A pseudoexon: implications for diagnosis and therapy.* Hum Mutat, 2011. **33**(1): p. 104-8.
93. Avila-Fernandez, A., et al., *Mutation analysis of 272 Spanish families affected by autosomal recessive retinitis pigmentosa using a genotyping microarray.* Mol Vis, 2010. **16**: p. 2550-8.
94. Audo, I., et al., *Development and application of a next-generation-sequencing (NGS) approach to detect known and novel gene defects underlying retinal diseases.* Orphanet J Rare Dis. **7**: p. 8.
95. Daiger, S.P., et al., *Targeted high-throughput DNA sequencing for gene discovery in retinitis pigmentosa.* Adv Exp Med Biol, 2010. **664**: p. 325-31.
96. Neveling, K., et al., *Next-generation genetic testing for retinitis pigmentosa.* Hum Mutat, 2011. **33**(6): p. 963-72.
97. Simpson, D.A., et al., *Molecular diagnosis for heterogeneous genetic diseases with targeted high-throughput DNA sequencing applied to retinitis pigmentosa.* J Med Genet, 2010. **48**(3): p. 145-51.
98. Tucker, B.A., et al., *Exome sequencing and analysis of induced pluripotent stem cells identify the cilia-related gene male germ cell-associated kinase (MAK) as a cause of retinitis pigmentosa.* Proc Natl Acad Sci U S A, 2011. **108**(34): p. E569-76.
99. Zuchner, S., et al., *Whole-Exome Sequencing Links a Variant in DHDDS to Retinitis Pigmentosa.* Am J Hum Genet, 2011. **88**(2): p. 201-6.
100. O'Donnell, C.J. and E.G. Nabel, *Cardiovascular genomics, personalized medicine, and the National Heart, Lung, and Blood Institute: part I: the beginning of an era.* Circ Cardiovasc Genet, 2008. **1**(1): p. 51-7.
101. Kearney, P.M., et al., *Global burden of hypertension: analysis of worldwide data.* Lancet, 2005. **365**(9455): p. 217-23.
102. Carretero, O.A. and S. Oparil, *Essential hypertension. Part I: definition and etiology.* Circulation, 2000. **101**(3): p. 329-35.

103. Munroe, P.B., M.R. Barnes, and M.J. Caulfield, *Advances in blood pressure genomics.* Circ Res, 2013. **112**(10): p. 1365-79.
104. Hottenga, J.J., et al., *Heritability and stability of resting blood pressure.* Twin Res Hum Genet, 2005. **8**(5): p. 499-508.
105. Frazer, K.A., et al., *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**(7164): p. 851-61.
106. O'Donnell, C.J. and E.G. Nabel, *Genomics of cardiovascular disease.* N Engl J Med, 2011. **365**(22): p. 2098-109.
107. Pater, C., *The Blood Pressure "Uncertainty Range" - a pragmatic approach to overcome current diagnostic uncertainties (II).* Curr Control Trials Cardiovasc Med, 2005. **6**(1): p. 5.
108. Dube, J.B. and R.A. Hegele, *Genetics 100 for cardiologists: basics of genome-wide association studies.* Can J Cardiol, 2013. **29**(1): p. 10-7.
109. Padmanabhan, S., C. Newton-Cheh, and A.F. Dominiczak, *Genetic basis of blood pressure and hypertension.* Trends Genet, 2012. **28**(8): p. 397-408.
110. Johnson, T., et al., *Blood pressure loci identified with a gene-centric array.* Am J Hum Genet, 2011. **89**(6): p. 688-700.
111. Frazer, K.A., et al., *Human genetic variation and its contribution to complex traits.* Nat Rev Genet, 2009. **10**(4): p. 241-51.
112. Dunham, I., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
113. *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.
114. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.
115. Forstermann, U. and W.C. Sessa, *Nitric oxide synthases: regulation and function.* Eur Heart J, 2011. **33**(7): p. 829-37, 837a-837d.
116. Fleming, I. and R. Busse, *Molecular mechanisms involved in the regulation of the endothelial nitric oxide synthase.* Am J Physiol Regul Integr Comp Physiol, 2003. **284**(1): p. R1-12.
117. Haynes, W.G., et al., *Inhibition of nitric oxide synthesis increases blood pressure in healthy humans.* J Hypertens, 1993. **11**(12): p. 1375-80.
118. Huang, P.L., et al., *Hypertension in mice lacking the gene for endothelial nitric oxide synthase.* Nature, 1995. **377**(6546): p. 239-42.
119. Niu, W. and Y. Qi, *An updated meta-analysis of endothelial nitric oxide synthase gene: three well-characterized polymorphisms with hypertension.* PLoS One, 2011. **6**(9): p. e24266.
120. Puca, A.A., et al., *Endothelial nitric oxide synthase, vascular integrity and human exceptional longevity.* Immun Ageing, 2012. **9**(1): p. 26.
121. Genovese, G., et al., *Mapping the Human Reference Genome's Missing Sequence by Three-Way Admixture in Latino Genomes.* Am J Hum Genet, 2013.
122. Carvalho, A.B. and A.G. Clark, *Efficient identification of Y chromosome sequences in the human and Drosophila genomes.* Genome Res, 2013.
123. Treangen, T.J. and S.L. Salzberg, *Repetitive DNA and next-generation sequencing: computational challenges and solutions.* Nat Rev Genet, 2011. **13**(1): p. 36-46.
124. Kirby, A., et al., *Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing.* Nat Genet, 2013. **45**(3): p. 299-303.
125. Dryja, T.P., *Gene-based approach to human gene-phenotype correlations.* Proc Natl Acad Sci U S A, 1997. **94**(22): p. 12117-21.

126. Tewhey, R., et al., *Enrichment of sequencing targets from the human genome by solution hybridization.* Genome Biol, 2009. **10**(10): p. R116.
127. Halbritter, J., et al., *High-throughput mutation analysis in patients with a nephronophthisis-associated ciliopathy applying multiplexed barcoded array-based PCR amplification and next-generation sequencing.* J Med Genet, 2013. **49**(12): p. 756-67.
128. Akhras, M.S., et al., *Connector inversion probe technology: a powerful one-primer multiplex DNA amplification system for numerous scientific applications.* PLoS One, 2007. **2**(9): p. e915.
129. Corton, M., et al., *Exome sequencing of index patients with retinal dystrophies as a tool for molecular diagnosis.* PLoS One, 2013. **8**(6): p. e65574.
130. Small, E.C., et al., *The EF-G-like GTPase Snu114p regulates spliceosome dynamics mediated by Brr2p, a DExD/H box ATPase.* Mol Cell, 2006. **23**(3): p. 389-99.
131. Weber, G., et al., *Mechanism for Aar2p function as a U5 snRNP assembly factor.* Genes Dev, 2011. **25**(15): p. 1601-12.
132. Deery, E.C., et al., *Disease mechanism for retinitis pigmentosa (RP11) caused by mutations in the splicing factor gene PRPF31.* Hum Mol Genet, 2002. **11**(25): p. 3209-19.
133. Gonzalez-Santos, J.M., et al., *Mutation in the splicing factor Hprp3p linked to retinitis pigmentosa impairs interactions within the U4/U6 snRNP complex.* Hum Mol Genet, 2008. **17**(2): p. 225-39.
134. Ivings, L., et al., *Evaluation of splicing efficiency in lymphoblastoid cell lines from patients with splicing-factor retinitis pigmentosa.* Mol Vis, 2008. **14**: p. 2357-66.
135. Tucker, B.A., et al., *Patient-specific iPSC-derived photoreceptor precursor cells as a means to investigate retinitis pigmentosa.* Elife, 2013. **2**: p. e00824.
136. Jin, Z.B., et al., *Integration-free induced pluripotent stem cells derived from retinitis pigmentosa patient for disease modeling.* Stem Cells Transl Med, 2012. **1**(6): p. 503-9.
137. Jin, Z.B., et al., *Modeling retinal degeneration using patient-specific induced pluripotent stem cells.* PLoS One, 2011. **6**(2): p. e17084.

ABBREVIATIONS

AAR2P	A1-alpha2 repressin protein (yeast)
adRP	Autosomal Dominant Retinitis Pigmentosa
APEX	Arrayed Primer Extension
arRP	Autosomal Recessive Retinitis Pigmentosa
ATP	Adenosine Triphosphate
BCA	Bicinchonicic acid
BRR2	Bad Response to Refrigeration
cDNA	Complementary DNA
CEPH	Centre d'Etude du Polymorphisme Humain
cGMP	Cyclic Guanosine monophosphate
ChIP	Chromatine Immunoprecipitation
CNV	Copy Number Variant
CoLaus	Cohorte Lausannoise
CVD	Cardiovascular Disease
DBP	Diastolic Blood Pressure
DExH/D	ATP hydrolysis motif: Aspartic Acid, Glutammic Acid, Histidine
DNA	Deoxyribonucleic acid
dNTP	Desossinucleotide Triphosphate
EFTUD2	Elongation Factor Tu GTP-Binding Domain-Containing 2
EH	Essential Hypertension
eQTL	Espression Quantitative Trait Locus
ERG	Electroretinography
FAM	6-carboxyfluorescein
GAPDH	Glyceraldehyde-3-Phosphate Dehydrogenase
GC	Guanosine-Cytosine
GMP	Guanosine Monophosphate
GWAS	Genome Wide Association Study
HEK293T	Human Embryonic Kidney 293 cells
HGMD	Human Gene Mutation Database
INL	Innernuclear layer
IPL	Inner plexiform layer
IS	Inner Segment
KIF3A	Kinesin Family Member 3A
LD	Linkage Disequilibrium
LR-PCR	Long-Range PCR
MAF	Minor Allele Frequency
MAK	Male germ-cell associated kinase
MGB	Minor Groove Binder
MPS	Massively parallel Sequencing
mRNA	Messenger RNA
NA	Not Available

NGS	Next Generation Sequencing
NHP2L1	Non-Histone Chromosome Protein 2-Like 1 (S. Cerevisiae)
OCT	Optical Coherence Tomography
OD	Oculis Dexter
OS	Outer Segment
OS	Oculis Sinister
PCR	Polymerase Chain Reaction
PDE	Phosphodiesterase
PDF	Portable Document Format
pre-mRNA	Precursor Messenger RNA
PRPF	pre-mRNA processing factor
RNA	Ribonucleic acid
RP	Retinits Pigmentosa
RPE	Retinal Pigmented Epithelium
SBP	Sistolic Blood Pressure
SNP	Single Nucleotide Polymorphism
snRNA	Small Nuclear RNA
snRNP	Small Nuclear Ribonulcleic Particle
SNU114	116 KDa U5 Small Nuclear Ribonucleoprotein Component
ss	Splice site
TBS	Tris-buffered Saline
TFBS	Transcription Factor Binding Site
UHTS	Ultra High Throughput Sequencing
UTR	Untranslated Region
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
wt	Wild Type

SUMMARY OF PUBLICATIONS

Benaglio, P. & Rivolta, C. (2010). Ultra high throughput sequencing in human DNA variation detection: a comparative study on the NDUFA3-PRPF31 region. *PLoS One*, 5(9), e13071

Benaglio, P., McGee, T. L., Capelli, L. P., Harper, S., Berson, E. L. & Rivolta, C. (2011). Next generation sequencing of pooled samples reveals new SNRNP200 mutations associated with retinitis pigmentosa. *Hum Mutat*, 32(6), E2246-58

Valsesia, A., Rimoldi, D., Martinet, D., Ibberson, M., **Benaglio, P**., et al. (2011). Network-guided analysis of genes with altered somatic copy number and gene expression reveals pathways commonly perturbed in metastatic melanoma. *PLoS One*, 6(4), e18369

Salvi, E., Kutalik, Z., Glorioso, N., **Benaglio, P**., et al. (2012). Genomewide association study using a high-density single nucleotide polymorphism array and case-control design identifies a novel essential hypertension susceptibility locus in the promoter region of endothelial NO synthase. *Hypertension*, 59(2), 248-55

den Hoed, M., Eijgelsheim, M., Esko, T., […], **Benaglio, P.**, et al. (2013). Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nat Genet* doi:10.1038/ng.2610

Salvi, E., Kuznetsova, T., Thijs, L., Lupoli, S., Stolarz-Skrzypek, K., D'Avila, F., Tikhonoff, D., De Astis, S., Barcella, M., Seidlerová, J., **Benaglio, P**., et al. (2013). Target Sequencing, Cell Experiments and a Population Study Establish eNOS as Hypertension Susceptibility Gene. *Hypertension*, 62(5), 844-52

Nishiguchi, K.M., Tearle, R.G., Liu, Y., Miyake, N., **Benaglio, P**., et al. (2013). Whole genome sequencing in patients with retinitis pigmentosa reveals pathogenic DNA structural changes and NEK2 as a new disease gene. *Proc Natl Acad Sci U S A*, 110(40), 16139-44

PEER- REVIEWED BOOK CHAPTER:

Benaglio, P. & Rivolta, C. (2013). Strategies for genetic screening of multiple samples using PCR-based targeted sequence enrichment. *Genomics III - Methods, Techniques and Applications*. *iConcept Press*. ISBN: 978-1-922227-09-6