Mendelian Disease Gene Identification and Diagnosis Using Targeted Next Generation Sequencing

Daniel Trujillano Lidón

TESI DOCTORAL UPF / 2013

DIRECTOR DE LA TESI

Dr. Xavier Estivill Pallejà

Genomics and Disease Group
Bioinformatics and Genomics Program
Centre for Genomic Regulation
Universitat Pompeu Fabra (CRG-UPF)

A l'Alba

A mis padres

Acknowledgements

Primer de tot vull agrair al Xavier per haver-me acollit al laboratori i haver dirigit aquesta tesi. Ha estat un plaer poder treballar al teu costat aquests quatre anys. Admiro molt la passió què poses en tot el que fas i la teva capacitat per trobar sempre alguna cosa nova i interessant per a investigar. Sé què serà difícil tornar a trobar un altre “jefe” com tu...

Durant tot aquest temps he tingut la sort de creuar-me amb un munt de gent magnífica què ha anat passant pel laboratori. D’una manera o altra tots heu contribuït a aquesta tesi, però voldria agrair especialment l’ajuda d’en Justo, el Cristian, l’Elisa, la Marta, l’Anna, la Laia, la Geòrgia, la Rut i la Laura. Us trobaré molt a faltar a tots i totes.

També estic profundament agraït a tots els col·laboradors amb qui he tingut la oportunitat de treballar, com l’Stephan, el Jean-Jacques, la Tere, la Belén, l’Elisabet i, com no, el Tim. No vull tampoc oblidar-me de tota la gent de Seattle què em van rebre amb els braços oberts, en especial la Piper i en Kaelou.

Menció especial es mereixen el Rose, l’Ukta i el Cruz per sempre haver estat allà sense fer-se gaire pesats.

A mi madre, mi padre, mi hermana y mis abuelos por vuestro apoyo incondicional y por haber sido siempre una fuente de inspiración.

I finalment vull agrair a l’Alba per haver estat al meu costat tots aquests anys. Gràcies per haver-me suportat i cuidat tant.

Abstract

Next Generation Sequencing (NGS) technologies have emerged as a powerful tool for the discovery of causative mutations and novel Mendelian disease genes, and are rapidly impacting genetic diagnostics. NGS technologies can be used in combination with DNA enrichment methods to generate deep sequencing of target genome regions, such as the exome or known disease loci, delivering fast, inexpensive and detailed genetic information. This thesis describes the application of targeted NGS to identify a novel disease gene for familial hyperkalemic hypertension. In addition, it also explores the clinical translation of NGS technologies to the genetic diagnostics of a heterogeneous panel of Mendelian diseases, including cystic fibrosis, hyperphenylalaninemias and autosomal dominant polycystic kidney disease. The results of this thesis do not only ratify targeted NGS as a powerful tool for Mendelian disease gene discovery, but also show that it is ready to substitute traditional molecular methods in medical genetics.

Resum

Les tecnologies de seqüenciació de nova generació (NGS) han emergit com a una poderosa eina per al descobriment de mutacions causals i nous gens per a malalties Mendelianes, i estan tenint un ràpid impacte en l'àmbit del diagnòstic genètic. Les tecnologies de NGS es poden utilitzar en combinació amb mètodes d'enriquiment de l'ADN per a seqüenciar en profunditat regions genòmiques diana, com l'exoma o gens associats a malalties, entregant informació genètica d'una manera ràpida, barata i acurada. Aquesta tesi descriu l'aplicació de la NGS dirigida per a identificar un nou gen per a la hipertensió hipercalièmica familiar. També s'explora la traducció clínica de les tecnologies de NGS per a millorar el diagnòstic genètic d'un panell heterogeni de malalties Mendelianes, què inclou la fibrosi quística, hiperfenilalaninèmies i la malaltia renal poliquística autosòmica dominant. Els resultats d'aquesta tesi no només ratifiquen la NGS dirigida com a una potent eina per al descobriment de gens de malalties Mendelianes, sinó què també demostren que aquesta tecnologia està preparada per a substituir els mètodes moleculars tradicionals a l'àmbit de la genètica mèdica.

Preface

Next Generation Sequencing (NGS) coupled to genomic capture technologies represents an important milestone in genomics, revolutionizing the way geneticists screen for disease-causing mutations in Mendelian disorders. This has shifted the focus of molecular analysis in human diseases from Sanger sequencing and array-based methods to deep sequencing of linkage intervals, candidate regions or exomes at higher resolution and greater sensitivity than previously possible. The objective of this thesis was to examine the possibilities of targeted NGS for the identification of causative variants and novel disease genes to improve the genetic diagnostics of Mendelian disorders. The first chapter of this thesis gives a general introduction about genetic variability and disease. This section also contains a general description of both classic molecular methods and NGS technologies applied to the study of Mendelian diseases. The results chapter contains four articles describing the different studies and methodology followed in each of them. The first article describes the identification of a novel gene for familial hyperkalemic hypertension by exome sequencing. The other three report the application of targeted NGS to improve the genetic diagnostics of cystic fibrosis, hyperphenylalaninemias and autosomal dominant polycystic kidney disease, respectively. The following chapter provides a general discussion about the future implications of NGS technologies both at basic research and clinical levels. This thesis ends with a chapter that summarizes the main conclusions of the work presented here.

Contents

1. INTRODUCTION

1.1. Genetic variation

Genetic variation allows for the uniqueness of each individual, but is also the cause underlying several human diseases, mostly based on mutations that affect either the functionality of the genes or their regulatory elements. The discovery and study of human genetic variation has been traditionally driven by technological innovations. Following the completion of the human genome sequence,[1,2] research efforts at the population level, such as the SNP Consortium and HapMap projects, characterized population variation at approximately 3.5 million mostly-common single nucleotide polymorphisms (SNPs).[3]

Nowadays, the advent of next generation sequencing (NGS) has revolutionized the way in which we interrogate the genome. With the new sequencing technology platforms in action, the characterization of genetic variation has been extended into the 1000 Genomes Project (1000GP) in a subset of different ethnic groups,[4] and is being expanded to 2,500 samples representing populations from different geographic origins. In addition to this, other projects, such as the International Cancer Genomics Consortium (ICGC),[5] and other initiatives in several countries (UK, The Netherlands, Denmark, Iceland, and others) are producing genomic sequence from thousands of individuals across the world. This, in combination with other projects collecting dense genotype data from numerous disease cohorts, will provide a large collection

of different sources of genomic variation, which should help to gain knowledge on the genetic basis of human disease.

These recent large scale human population genetic studies have unveiled that genetic diversity between individuals is much larger than what was expected, both at the single nucleotide and at the genome structure levels. It has been estimated that even if all men and women were whole-genome sequenced, up to 60 unique mutations would be found per individual.[6] Fortunately, most of these variants have no medical consequences, especially those present in healthy individuals. However, their presence in databases can help to identify disease-related mutations.[7]

1.1.1. Single Nucleotide Variants

Watson-Crick DNA base pair changes or single nucleotide variants (SNVs) are the most prevalent class of genetic variation amongst individuals.[8] When at a given locus the nucleotide composition varies more than 1% in the general population, then the SNV is referred as single nucleotide polymorphism (SNP). It has been estimated that every two haploid human genomes would differ in 1 nucleotide per 1331 bp across the 3 billion bp of the human genome.[9]

1.1.2. Short Insertions and Deletions

A single or a sequence of nucleotides can also be deleted or inserted in the genome, giving rise to the second most common type of polymorphisms, collectively known as InDels.[10] A human genome

contains approximately one million of this type of genomic variation, although this figure might be an underestimation since InDels are more difficult to detect than SNVs.[11]

1.1.3. Structural Variants

Structural variants are genomic rearrangements of hundreds to millions of consecutive base pairs that include large insertions, deletions and duplications (also known as copy number variants or CNVs), inversions and complex combinations of different rearrangements. These changes can have an important impact on gene function and expression, and have been appreciated only more recently as a significant source for human genetic variation. In fact, structural variants have already been associated to human disease providing new insights on the genetic basis of phenotypic and disease-susceptibility differences between individuals.[12]

1.2. Classic methods for the detection of genetic variation and disease gene identification

Genetic variation is the driving force behind evolution. However, the phenotypic consequence of novel or inherited variation can provide a selective advantage or affect negatively the fitness of an individual, for example predisposing to disease. For this reason, it is a core activity of human genetics to identify the specific genetic defects that cause diseases, so that their diagnosis and treatment can be improved. So far, according to the Human Gene Mutation Database (HGMD; http://www.hgmd.org) more than 140,000

disease mutations, spread across >5,700 different nuclear genes, have been annotated, and these numbers keep increasing.

Following, the main methodological approaches used to detect genetic variation and disease gene identification are described. The main goal of all these approaches is to identify a certain sequence variation only found in patients and not in healthy individuals.

1.2.1. Candidate gene approach

Candidate genes are selected on the basis of previous evidence linking a given locus to the disease phenotype.[13] This hypothesis-based approach takes advantage of previous knowledge, including previously implicated genes, genes associated with similar diseases or on the biological activity of proteins relevant to the physiology of the disease. Then, the candidate genes are screened for causal mutations using Sanger sequencing or other methods specific for the analysis of genomic rearrangements. Although this approach has been successful in the identification of genes involved in monogenic and complex diseases,[14] it is clearly biased towards the current biological knowledge.

1.2.2. Positional cloning, linkage and homozygosity mapping

Positional cloning strategies aim at identifying in an unbiased and hypothesis-free manner genomic loci likely related to the disease. Once the candidate loci are defined, Sanger sequencing is applied to screen for causal variants. During the past decades, the development of highly reliable positional cloning strategies, including high-

throughput linkage analysis using DNA microarrays, along with the availability of the human genome sequence, has accelerated the search for the causative genes of diseases with Mendelian traits. Positional cloning starts with the ascertainment of large families inheriting the disease in a Mendelian fashion.

Linkage analysis uses highly polymorphic markers throughout the genome to identify chromosomal regions that segregate with disease susceptibility within families.[15] This strategy has led to the identification of genes underlying major human diseases, such as cystic fibrosis.[16-18] A limit of this approach is that it allows discovering genes that exert a major effect on susceptibility but it is less likely to be successful when several genetic determinants are involved with a small individual effect. Moreover, when the pedigree size is limited, it is difficult to narrow the candidate region by linkage analysis; hence, tremendous effort is still required to identify the causative genes.

Homozygosity mapping also benefits from the existence of dense SNP genotyping arrays, and it tests the assumption that a homozygous mutation in a recessive disease gene is identical by descent by segregating twice to the affected child from a common ancestor through both the maternal line and the paternal line. This short segment of homozygosity by descent can then be detected by multipoint homozygosity mapping and will harbor the sought-after disease gene. Until now, successful gene identification by homozygosity mapping has been mostly based on consanguineous families that have several affected individuals.[19]

1.2.3. Genome wide association studies

More recently, genome-wide association studies (GWAS) have been applied to discover novel disease loci for complex disorders and common traits. GWAS analyze genetic association by comparing SNP allele frequencies in affected individuals with those of controls. High-density genotyping arrays containing between 100,000 and 5 million SNPs, based on the existence of linkage disequilibrium blocks, allow screening variation in the human genome for any disease and trait with enough cases for adequate statistical power.[20] The unbiased approach of GWAS eliminates the disadvantages of the earlier association studies, which genotyped only few SNP from a candidate gene. Actually, GWAS using high-throughput approaches, have provided >2,600 associated common risk alleles, with convincing associations in >350 different complex diseases and common traits (Figure 1).[21]

Figure 1. Published Genome-Wide Associations for 17 trait categories ($p \leq 5x10^{-8}$). From Hindorff *et al.*, 2013.[22]

GWAS are more suited for the identification of variants associated to common disorders under the assumption of the "common-disease common-variant" hypothesis.[23] However, despite the success of GWAS in defining several hundred genes with disease associations, they fail to detect much of the genetic variation impacting on disease outcomes. This is likely due to the fact that GWAS cannot detect many sources of genetic variability such as complex copy number variants or low-frequency variants. In fact, low frequency (1%-5%), rare (<1%) and novel variants would not show up in either GWAS or linkage studies, and will only be detectable by efforts that explore the complete variability of the genome (Figure 2). Thus, it is clear that this approach can access only a small proportion of the genomic contribution to diseases and phenotypes.[24] Also, most of the GWAS findings do not yet translate into preventive measures or clinical testing, as most of the associations have modest effect size of odds ratio <1.5.

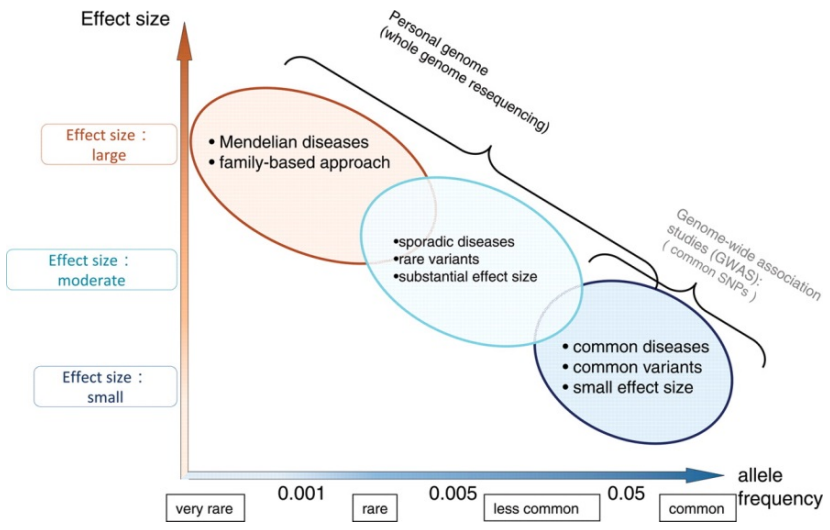


Figure 2. Identification of disease-related variations depending on risk allele frequency and genetic effect (odds ratio). From Tsuji *et al.*, 2010.[25]

1.3. Next generation sequencing

The identification of Mendelian disease genes has long been based on positional cloning approaches that often led to determine candidate regions spanning approximately 0.5-10 cM and containing around 300 genes. Thus, the identification required extensive Sanger sequencing of large numbers of genes, which is a costly and time-consuming process. Using these strategies, about half of the genes containing allelic variants responsible for monogenic disorders have been uncovered.[26] This has prompted the advent of NGS technologies, which come with the promise of overcoming the limitations of the previous genomic approaches, allowing the unbiased interrogation for all kind of causal mutations in broad candidate regions or even genome-wide.

During the last forty years, first-generation sequencing, also known as Sanger sequencing,[27] especially in its latest iteration (the automated sequencer by capillary electrophoresis ABI 3730XL), has been considered the "gold standard" because of its accuracy and ultimate resolution to identify sequence variants. It has been used in different historically significant large-scale sequencing projects for the discovery of important mutations and genomic structural variants. However, this technique is not easy to scale up allowing only up to 384 sequencing reactions in parallel in the most advanced capillary sequencers. It is also expensive, with an estimated cost of €400 per sequenced Mb.[28-30]

Thanks to the recently tremendous technology progress, current technologies utilizing NGS, based on shotgun approaches, are able to sequence in excess of one billion short reads in parallel per instrument run. Interestingly, NGS technologies are able detect all kinds of sequence variants in a single experiment, including SNVs, InDels and SVs, providing investigators access to a large spectrum of *de novo* and rare inherited variants (those with frequencies <1%) mutations, which are often omitted in standard genotyping panels.

Amongst the different NGS technologies recently developed, the most widely adopted have been the Roche's 454 GSFlex,[31,32] ABI's SOLiD,[33] and Illumina's GA and HiSeq.[34,35] The sequencers from these vendors show significant differences in many aspects, but their workflows are conceptually comparable, mainly involving template preparation, sequencing and imaging, and data analysis steps.[36]

Currently, Illumina's technology dominates the NGS market. This is the NGS technology used in this thesis and the one that will be explained here more in detail. The process starts by fragmenting the genomic DNA into small pieces (around 300 bp), which are then ligated to specific adapter sequences. Then, the DNA fragments are clonally amplified and clustered together (Figure 3) to serve as a template for the sequencing process, which consists of multiple alternating cycles of cyclic reversible termination (CRT) and imaging.[37]

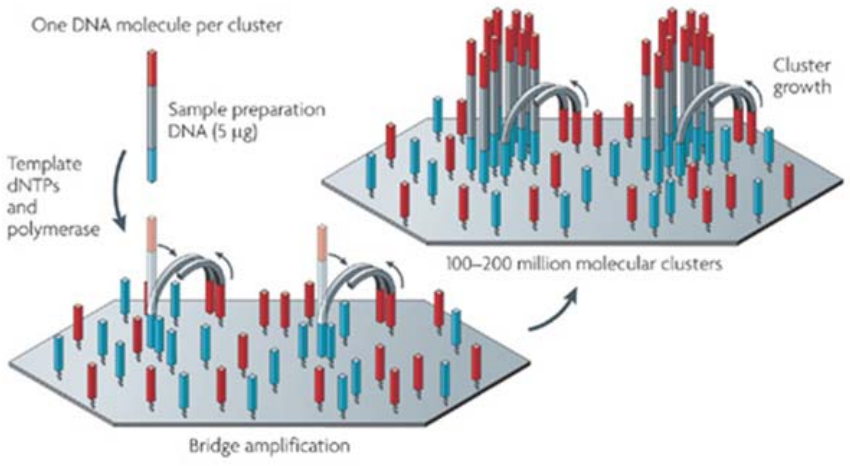


Figure 3. Clonal amplification. It is composed of two basic steps: initial priming and extending of the single-stranded, single-molecule template, and bridge amplification of the immobilized template with immediately adjacent primers to form clusters. Adapted from Metzker *et al.*, 2010.[36]

CRT uses four-color reversible terminators in a cyclic method that comprises nucleotide incorporation, fluorescence imaging and cleavage. More in detail, in every cycle just one fluorescently modified nucleotide is incorporated as the complement of the template base. Unincorporated nucleotides are washed before imaging. To determine the identity of the incorporated nucleotide at each cluster, the four colors are detected by total internal reflection fluorescent imaging. After imaging, the terminating group and the fluorescent dye of the incorporated nucleotide are removed during a cleavage step. Then a new CRT cycle begins to incorporate and read the following nucleotide (Figure 4). Illumina's technology allows sequencing runs ranging from 50 to 300 CRT cycles.[36]

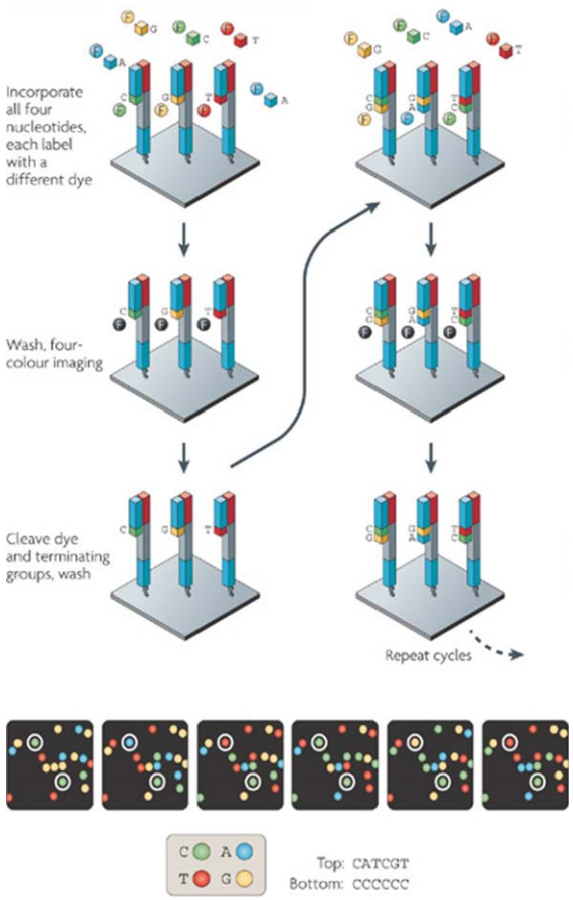


Figure 4. The four-color cyclic reversible termination (CRT) method.
Each cycle comprises nucleotide incorporation, fluorescence imaging and cleavage. The four-color images highlight the sequencing data from two clonally amplified templates. Adapted from Metzker *et al.*, 2010.[36]

Next generation sequencers can produce independent reads or reads from both sides of the DNA fragment (the insert) at an approximately known distance. In the latter case, two strategies can be followed. Mate pairs are generated when long DNA fragments of various kilobase (kb) in length are circularized and then randomly sheared before sequencing, allowing larger insert-size libraries. In contrast, paired-end reads are created by shearing the DNA into

short segments of approximately 300 bp which are then sequenced at both ends, providing a tighter insert size distribution.[38]

NGS technologies have shifted the scale of the sequencing runs from the order of kb of the first generation of capillary sequencers to the hundreds of gigabases (Gb) of the new breed of massive parallel sequencers (Figure 5). Also, the use of massively parallel technologies has reduced very rapidly the cost per sequenced base by a million-fold since 1990.[30,39]

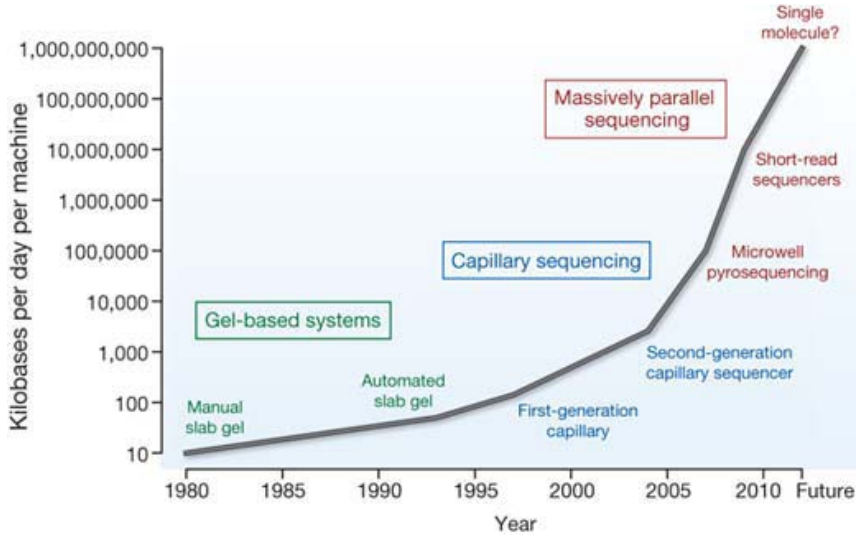


Figure 5. Increased throughput of next-generation sequencers. Improvements in the rate of DNA sequencing from slab gels to capillary sequencing and NGS technologies. From Stratton *et al.*, 2009.[40]

The first whole human genome sequenced using NGS technologies was that of James D. Watson.[41] Ever since, these technologies have permitted the whole-genome sequencing (WGS) of several individuals at much smaller costs than the sequencing of the first human genomes.[34,42-46] While the first human genome sequencing project supposed a total expense of the order of €2,000 million, the

cost of sequencing is dropping to the point where a whole human genome is now available for about a similar cost to that of a regular body imaging investigation. Moreover, the cost of WGS is becoming less expensive than many current genetic tests.

The broadest application of NGS is WGS, but it can also be used for many different reduced representation sequencing applications, such as resequencing of (subsets of) genomic DNA, ChIP-sequencing to identify binding sites of DNA-associated proteins,[47,48] and to profile the complete or selected transcriptome by sequencing cDNA derived from total RNA (RNAseq).[49] All these methods largely exceed their array-based precursor genomic technologies in terms of resolution.

1.3.1. Genomic enrichment

Despite the substantial cost reductions associated with NGS technologies in comparison with Sanger sequencing, WGS is still a prohibitively expensive endeavor for studies in which a large number of samples are required to achieve adequate statistical power or for routine clinical practice. Also, depending on the study it is not necessary to explore the entire genome of the subjects being investigated.

Thus, there is a growing demand for genome targeting methods to sequence specific subsets of interest of the genome, and several commercial alternatives are available. These protocols allow cost-effective capture of regions of interest in large numbers of samples in parallel. The capture target can be any genomic region, from a

single gene to linkage intervals or the complete list of exons of the genome, i.e. the exome (discussed below). The possibility of focusing only in to what is really interesting optimizes the cost and analysis required, especially considering that the function of much of the genome is still largely unknown. However, it is required an educated guess as to which regions or genes may be interesting. These methods produce a pool of desired molecules that are present in complex mixtures of irrelevant DNA sequences, and that are separated by the parallel nature of the sequencing technologies themselves.[50] Parallelized genome enrichment and sequencing of multiple samples also requires the incorporation of barcodes into the sequencing libraries to be able to trace back the sample source of each sequencing read. This approach allows for the sequencing of a smaller fraction of the genome across a much larger number of individuals.

Targeted NGS requires efficient methods for massive parallel enrichment of the templates to be sequenced. Targeted capture methods can be classified into those that rely on an enzymatic step to achieve specificity (i.e., molecular inversion probes, multiplex PCR), and those that rely purely on pullout by hybridization (into in-solution and on-array methods) to oligonucleotides complementary to the sequences of interest.[51]

1.3.1.1. Polymerase-mediated capture

Multiplexed PCR using several primer pairs in a single reaction can be used to generate multiple amplicons of the desired target

sequences.[52] However, this is a cumbersome process not feasible in terms of cost and labor input when large genomic regions are supposed to be sequenced. Alternatively, another PCR-mediated enrichment protocol based on Molecular Inversion Probes (MIPs) has been developed.[53,54] This technique consists in the performance of large numbers of individual amplification reactions using oligonucleotides that are synthesized on microarrays and subsequently cleaved off and amplified by PCR, to perform a padlock and molecular inversion reaction in solution where the probes are extended and circularized to copy, rather than directly capture, the targets to be sequenced. This is not a cost-effective approach given the significant investments in oligonucleotides, enzymes, and infrastructure required. The major problems associated with this technique are an unbalanced representation of the targets and poor reproducibility.

1.3.1.2. Solid-phase hybridization

Solid-phase hybridization makes use of a solid support, such as microarrays, where probes complementary to the regions of interest are affixed. The fragments of the randomly sheared genomic DNA that match the synthetic oligonucleotides of the surface of the microarray hybridize and are retained for the posterior sequencing, whereas the resting untargeted genomic fragments are washed away.[55-57]

1.3.1.3. Liquid-phase hybridization

Liquid-phase hybridization is a modification of the solid-phase method, where the probes are not attached to a solid-phase, but instead are in solution and biotinylated. In solution hybrid-selection capture takes advantage of the economy of oligodeoxynucleotide synthesis on an array and the favorable kinetics of hybridization in solution to cheaply and effectively target multiple regions in the genome. It consists in "fishing" the genomic targets out of a pool of genomic DNA fragments. Following hybridization, the biotinylated baits and their bound complementary genomic fragments are pulled-down using streptavidin-coated magnetic beads and a magnet (Figure 6). This makes this method a highly flexible, scalable, and efficient approach since it does not need any other specialist equipment than a magnet and standard tools for manipulating liquids.[58] In addition, respect to microarray-based hybrid-selection, solution capture is less expensive and requires less DNA. The major drawbacks of in-solution hybrid-selection are that is more time-consuming than its array-based counterparts, mainly due the need for relatively long hybridization times compared to solution phase capture.

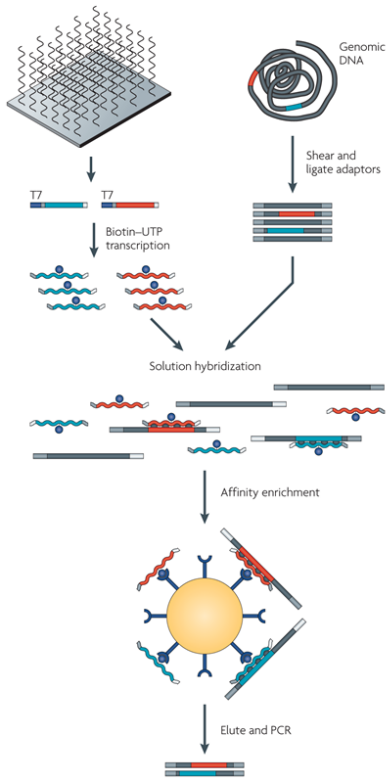


Figure 6. Liquid-phase hybridization. Solution-phase methods use target-specific oligonucleotides synthetized in arrays, which are cleaved as a pool of probes from the support and eluted into a single tube. The oligonucleotides are modified with a T7 promoter sequence, which allows the incorporation of biotin-labeled uridine-5′-triphosphate (UTP) into the probe sequence. Then the baits are mixed with a size selected pond library of fragments modified with sequencing adaptors. Hybridized fragments are then captured to streptavidin beads and eluted for sequencing. Adapted from Metzker *et al.*, 2010.[36]

1.3.2. Exome sequencing

Only 1% of the human genome corresponds to protein-coding genes, which represent around 50 Mb split across ~200,000 exons of ~21,000 genes. The function of the rest of the genome remains largely unknown. The regions with coding potential are generally referred as the "exome" or total exon complement. Noteworthy, most of the human inherited disease-related mutations identified so

far are located into this small coding portion. Actually, up to 85% of all disease-causing mutations in Mendelian disorders are within coding exons. Thus, it seems convenient for disease-gene discovery projects to concentrate sequencing efforts on the exome to avoid the additional cost and complexity of WGS.

As a more cost-conscious alternative to WGS, the extension of genomic enrichment to the human exome allows the unbiased study of the complete set of protein-coding regions of the genome without the need of having to choose a subset of genes for interrogation, allowing larger numbers of samples than are currently practical with WGS. Exome sequencing provides a complete perspective of coding genetic variation to a degree that has never before been possible, overshadowing traditional methods for the study of common, rare and novel genetic variation, such as SNP arrays or single locus resequencing studies.

Exome sequencing is a clear example of pragmatism applied to science, since sequencing only 1% of the genome represents a good compromise between cost and scientific ambition. However, it must be noted that that one percent costs more than one percent of a whole genome to obtain. The available methods for the capture of the exome (commercialized by Agilent Technologies [SureSelect] and Roche-NimbleGen [SeqCap/SeqCap EZ]) use array hybridization or liquid hybridization to biotinylated probes (previously discussed) that target all exonic and flanking sequences and may also include probes to target non-coding regions of interest, such as micro RNAs (miRNAs).

When exome sequencing is applied to the study of Mendelian diseases it is done under the assumption that this group of diseases are caused by rare genetic variants with complete or very high penetrance. Thus, the downstream analyses are focused on the identification of very rare or novel loss-of-function mutations that introduce truncations to the encoded protein, i.e. nonsense and non-synonymous variants, splice acceptor and donor site mutations and coding InDels, anticipating that synonymous variants are far less likely to be pathogenic.

Exome sequencing has achieved ground-breaking success in identifying genes associated with Mendelian diseases, as demonstrated by a slew of recent publications. As a proof of concept, Ng *et al.* reported the first study targeting the exome of patients suffering of a Mendelian disease in which they identified variants in the known disease-causative gene. Concretely, the sequencing of the exomes of four unrelated individuals with Freeman-Sheldon syndrome, a rare dominantly inherited disorder, demonstrated that the causative gene (*MYH3*) for a monogenic disorder (FSS) can be identified directly by exome sequencing of a small number of unrelated affected individuals.[59] Remarkably, *MYH3* was the only candidate gene after the application of multiple filters, including the requirement of identifying the variants in each sample studied. However, as mentioned above, this was just a proof-of-concept experiment, since *MYH3* was already known to underlie FSS. Thus, the actual success of this experiment was the demonstration of the power of exome sequencing for the

identification of disease-causing mutations in genes underlying Mendelian disorders.

Later on, the same group successfully applied exome sequencing for the first time in the identification of a novel disease gene. Ng *et al.* sequenced the exomes of four individuals from three independent kindred suffering of Miller syndrome, a rare Mendelian disorder (autosomal recessive) of unknown etiology, and identified a novel causal mutation in *DHODH*, which encodes for a protein involved in pyrimidine biosynthesis, as the origin of the disease.[60]

Meanwhile, another research group used exome sequencing to make an unanticipated genetic diagnosis. Choi *et al.* applied exome sequencing for the genetic diagnosis of a five-month-old Turkish boy with a mysterious genetic illness. This boy was initially diagnosed by the doctors to suffer of a kidney disease (Bartter syndrome), but the genetic cause could not be identified. However, the sequencing of his exome identified a homozygous missense substitution in *SLC26A3* (a known congenital chloride diarrhea gene).[61] Thus, the boy was finally diagnosed congenital chloride diarrhea, a diagnosis that was confirmed by follow-up clinical evaluation by the doctors, validating the use of exome sequencing as a clinical tool for the study of patients with undiagnosed genetic illnesses.

These studies demonstrated for the first time the potential of exome sequencing, even with reduced sample numbers, in combination with appropriate bioinformatics filtering against public variant

databases to exclude benign and unrelated variants, as an efficient strategy for the identification of single candidate genes for unsolved Mendelian disorders. Ever since after this, other reports have described the successful use of this strategy to several diseases and phenotypes, such as the Mabry syndrome,[62] Fowler syndrome,[63] severe brain malformations[64], and also in autosomal dominant disorders, such as Kabuki syndrome,[65] spinocerebellar ataxias,[66] familial amyotrophic lateral sclerosis (ALS),[67] or X chromosome-linked disorders, such as non-syndromic intellectual disability.[68]

Exome sequencing has also been applied for the identification of variation between populations. As an example, Yi *et al.* used exome sequencing to identify changes in allele frequency between high altitude populations (Tibetans), and low altitude populations (Han Chinese and Danes). Exome sequencing of 50 residents of the Tibetan Plateau, a region situated 4,000 meters over the sea, helped to the identification of several genetic variants associated to extreme altitude and low oxygen concentration adaptations. The most significant variant corresponds to a SNP at the Per-Arnt-Sim (PAS) domain protein 1 (*EPAS1*), a transcription factor involved in response to hypoxia. This variant also associates with high erythrocyte levels, reinforcing the potential involvement of *EPAS1* in adaptation to hypoxia.[69]

1.3.3. Detection of genetic variants

Due to the sheer magnitude of the genomic information produced by the current NGS equipment (several hundred Gb can now be

generated in just one sequencing run), the experimental bottleneck has shifted from data acquisition towards its correct storage and processing. As an example, while the sequence of the human genome is 3 Gb, the full collection of files in use for one whole human genome may reach various terabytes (TB), including intensity files, SAM (Sequence Alignment Map format), BAM (binary version of SAM), and other files with coordinates, variations etc.

Efficient methods to align millions of short-read sequences to the human genome (matching the short reads to a preexisting reference genome)[70-73] and the calling of variants (determination of the best guess for the genotype, or other sequence feature, at each aligned position) have been developed,[74-76] allowing to access most of the reference genome and to align de novo sequences that are missing in the reference genome sequence. Since DNA sequence variants may involve from single nucleotides up to several kb, specific algorithms have been developed for single base substitutions, insertions/deletions and structural variations.[77-79] All steps of the bioinformatics analysis have huge computational demands with which ordinary computers cannot cope.[80,81]

Once the variant calling process of a given sample or project is finished, the next task is the annotation of each detected sequence variant (Figure 7). During this process, information regarding the alignment of the variant to a specific base position in a gene, the *in silico* assessment of the variant's potential to disrupt gene function ("pathogenicity"),[82-85] and the presence of the variant in databases

such as dbSNP, are gathered and recorded. Several annotation tools are available, such as the Genome Analysis Toolkit (GATK),[75] SeattleSeq (www://gvs.gs.washington.edu/), or ANNOVAR,[86] among many others.

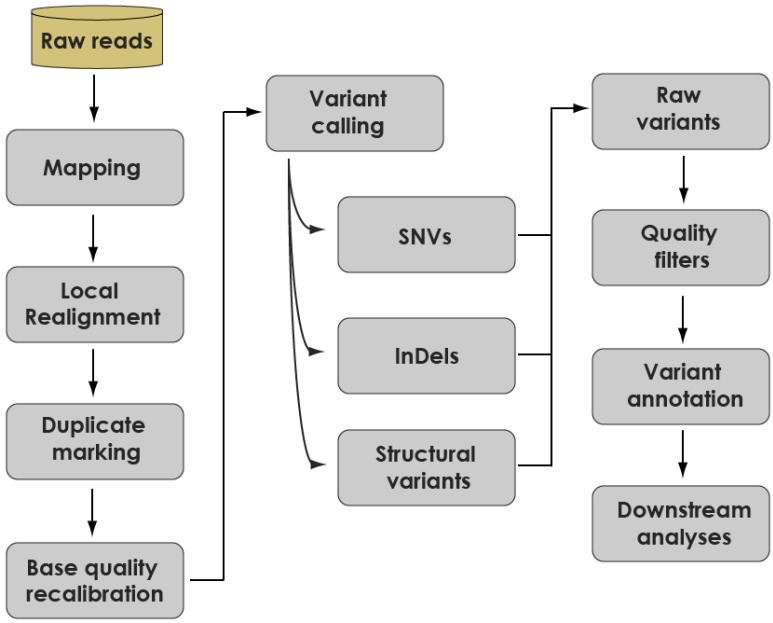


Figure 7. Overview of the bioinformatics analyses for NGS data. From the mapping of the raw sequencing reads to the annotation of the detected variants.

The sensitivity to detect sequence variants is a key parameter in the process of mutation identification. However, sensitivity is conditioned by three main factors: the coverage, the sequencing quality, and the read mapping quality. Although NGS technologies are highly accurate, with average rates of error of less than 0.1–0.2%, it is necessary to consider the consequences of these errors during the calling of variants in poorly covered loci.[87] Other sources of spurious variant calls are biases during library preparation[88] or

amplification[89], difficulty making genotype calls at the end of short reads[74], and platform-specific mechanistic problems[90,91]. Also, false positive calls can arise from misalignment of the sequencing reads to the reference sequence.[92]

1.3.4. Disease gene/variant identification strategies for NGS

As previously described, NGS technologies produce sheer numbers of genotype calls on the order of 10^4 per exome, 10^5 for the combined exomes of a small family, and 10^6 per genome.[93] Thus, after data acquisition and variant calling the main challenge in the downstream analysis of NGS data is to "winnow" the list of variants to be able to differentiate known and potential novel disease-causing mutations (the "wheat") from both technical artifacts and benign genetic variation (the "chaff").

Depending on the capture design and the depth of coverage, an average exome contains between five to ten thousand variant calls representing either non-synonymous substitutions in protein coding sequences, small InDels, or alterations of the canonical splice-site dinucleotides (NS/SS/I), being between 100 and 200 homozygous protein truncating or stop loss variants.[94] Thus, the mere identification of an apparently causative variant cannot be taken as a proof that it is relevant to the disease being investigated, and additional variant filtering and functional analyses are required to assign causativeness.[26]

When NGS are applied to Mendelian disorders, the filtration strategy is designed to highlight rare or *de novo* (filtering out

common variants from dbSNP, the HapMap and the 1000GP), high penetrance protein-modifying mutations responsible for a large phenotypic effect, as well as all variants previously associated with the disease. The inclusion of classical genetic mapping information (linkage[95]/homozigosity[96] mapping) to exclude irrelevant genomic regions prior to the application of other computational filters is important to maximize the success of Mendelian disease gene identification by NGS. This filtering strategy, which has been successfully applied in several studies, substantially reduces the list of candidate variants making feasible their individual confirmation prior expression and functional testing (Figure 8).

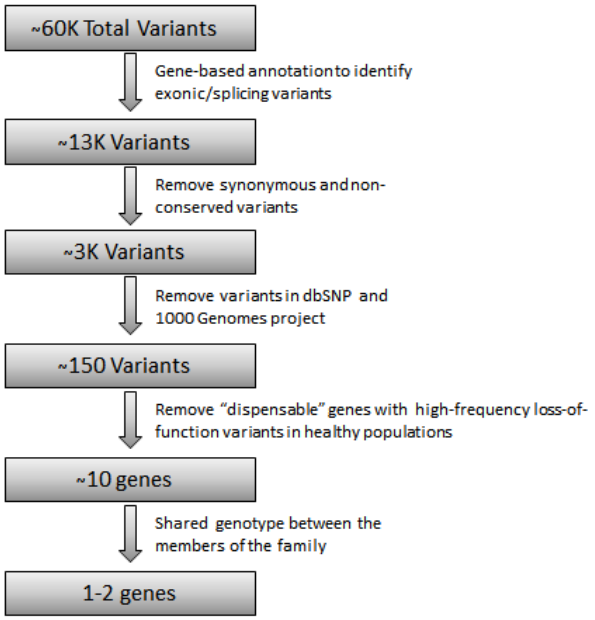


Figure 8. Variant prioritization process for Mendelian disorders using exome sequencing.

Another parameter that has to be taken in to account during variant/gene prioritization is the inheritance pattern, since for

autosomal dominant disorders each gene must show at least one potentially causative variant per individual, whereas in autosomal recessive disorders, candidate genes must have either homozygous or compound heterozygous mutations.[26]

All in all, the variant filtering strategy must be flexible enough to allow adjustment of all analytic parameters. But even more importantly, those performing the analysis must understand the rationale, procedures, and assumptions inherent in each step.[93]

2. OBJECTIVES

The extraordinary progress on genome sequencing technologies has produced one of the major scientific breakthroughs in the last years. NGS, combined with targeted enrichment and robust bioinformatics analyses, provides a rapid, cost-effective approach for identifying causative mutations and novel Mendelian disease genes.

The aim of this thesis was to fully leverage the potential of targeted NGS to identify the genetic causes of Mendelian disorders.

This aim can be subdivided into:

1. Identification of a novel disease gene for familial hyperkalemic hypertension by exome sequencing
2. Assessment of the amenability of targeted NGS for clinical diagnostics of cystic fibrosis, hyperphenylalaninemias, and autosomal dominant polycystic kidney disease.

3. RESULTS

3.1. Identification of a new gene for familial hyperkalemic hypertension

In this study we screened by exome sequencing two families affected by a dominant Mendelian form of arterial hypertension. The initial list of tens of thousands of variants was filtered to highlight functional *de novo* variants within a previously defined linkage region on human chromosome 5. In doing so, we identified mutations in *KLHL3* as a cause of familial hyperkalemic hypertension. Additional functional studies demonstrated the key role of *KLHL3* in regulating ion homeostasis in the distal nephron and blood pressure.

The results of this study led to the following publication:

KLHL3 mutations cause familial hyperkalemic hypertension by impairing ion transport in the distal nephron.

Louis-Dit-Picard H*, Barc J*, Trujillano D*, Miserey-Lenkei S, Bouatia-Naji N, Pylypenko O, Beaurain G, Bonnefond A, Sand O, Simian C, Vidal-Petiot E, Soukaseum C, Mandet C, Broux F, Chabre O, Delahousse M, Esnault V, Fiquet B, Houillier P, Bagnis CI, Koenig J, Konrad M, Landais P, Mourani C, Niaudet P, Probst V, Thauvin C, Unwin RJ, Soroka SD, Ehret G, Ossowski S, Caulfield M; International Consortium for Blood Pressure (ICBP), Bruneval P, Estivill X, Froguel P, Hadchouel J, Schott JJ, Jeunemaitre X.

*equal contribution

Nature Genetics 2012, 44: 456-60, S1-3

Louis-Dit-Picard H, Barc J, Trujillano D, Miserey-Lenkei S, Bouatia-Naji N, Pylypenko O et al. KLHL3 mutations cause familial hyperkalemic hypertension by impairing ion transport in the distal nephron. Nat Genet. 2012 Mar 11; 44(4): 456-60, S1-3. DOI: 10.1038/ng.2218

Corrigendum (May 2012)

Supplementary information

3.2. Improving genetic diagnostics of cystic fibrosis and CFTR-related disorders

Cystic fibrosis is one of the most common, life-threatening, autosomal recessive genetic disorders, with a carrier frequency in the Caucasian population of around one in 25 people. Although *CFTR* is one of the most extensively studied human disease genes, its high allelic heterogeneity makes the molecular diagnostics of cystic fibrosis and *CFTR*-related disorders challenging. This study describes the application of combined target enrichment of the *CFTR* gene, NGS and sophisticated bioinformatics analysis to develop an assay able to completely screen *CFTR* in cystic fibrosis and *CFTR*-related disorders patients and carriers. This approach was able to identify all types of mutations, including large and complex rearrangements, and polymorphic variants, achieving a diagnostic rate of 99%.

The results of this study led to the following publication:

Next generation diagnostics of cystic fibrosis and CFTR-related disorders by targeted multiplex high-coverage resequencing of CFTR.

Trujillano D, Ramos MD, González J, Tornador C, Sotillo F, Escaramis G, Ossowski S, Armengol L, Casals T & Estivill X.

Journal of Medical Genetics 2013, 50: 455-62.

Trujillano D, Ramos MD, González J, Tornador C, Sotillo F, Escaramis G et al. Next generation diagnostics of cystic fibrosis and CFTR-related disorders by targeted multiplex high-coverage resequencing of CFTR. J Med Genet. 2013 Jul; 50(7): 455-62. DOI:10.1136/jmedgen

Supplementary Data

3.3. Differential genetic diagnostics of hyperphenylalaninemias

A second example of the application of targeted NGS to the genetic diagnosis of Mendelian disorders is reported here. In this case we developed an assay for phenylketonuria and tetrahydrobiopterin deficient hyperphenylalaninemia. Hyperphenylalaninemias can be caused by mutations in four different genes (*PAH*, *GCH1*, *PTS*, and *QDPR*), complicating the mutation identification process. Interestingly in this assay, we optimized the different steps of the diagnostics process to be able to deliver results as soon as five days after receiving the DNA samples.

The results of this study led to the following publication:

Accurate molecular diagnosis of phenylketonuria and tetrahydrobiopterin-deficient hyperphenylalaninemias using high-throughput targeted sequencing.

Trujillano D*, Pérez B*, González J, Tornador C, Navarrete R, Escaramis G, Ossowski S, Armengol L, Cornejo V, Desviat LR, Ugarte M & Estivill X.

*equal contribution

European Journal of Human Genetics. 2013 August 14. [Epub ahead of print]

Trujillano D, Perez B, González J, Tornador C, Navarrete R, Escaramis G et al. Accurate molecular diagnosis of phenylketonuria and tetrahydrobiopterin-deficient hyperphenylalaninemias using high-throughput targeted sequencing. Eur J Hum Genet. 2014 Apr; 22(4): 528-34. DOI: 10.1038/ejhg.2013.175

Supplementary info

3.4. Diagnosis of autosomal dominant polycystic kidney disease using targeted sequencing

Yet another clinical application of NGS, in this case for the genetic diagnostics of autosomal dominant polycystic kidney disease (ADPKD), which is caused by mutations in *PKD1* and *PKD2*. To date, genetic diagnosis of ADPKD by conventional techniques requires cumbersome long-range polymerase chain reaction (LR-PCR) of the repeated region of *PKD1* followed by nested PCRs. For the first time, we successfully applied in-solution hybridization to capture *PKD1* gene, a complex gene, which is duplicated six times resulting in 6 pseudogens which share 98% sequence identity with the genuine gene. Our sophisticated bioinformatics analysis was able to detect single nucleotide variants as well as to characterize small insertions/deletions and large structural variants, even in the repeated region of *PKD1*.

The results of this study led to the following submitted manuscript:

Diagnosis of autosomal dominant polycystic kidney disease by efficient PKD1 and PKD2 multiplex high-throughput targeted sequencing.

Trujillano D, Bullich G, Ossowski S, Ballarín J, Torra R, Estivill X & Ars E.

Human Mutation [Submitted]

Trujillano D, Bullich G, Ossowski S, Ballarín J, Torra R, Estivill X et al. Diagnosis of autosomal dominant polycystic kidney disease using efficient PKD1 and PKD2 targeted next-generation sequencing. Mol Genet Genomic Med. 2014 Sep; 2(5): 412-21. DOI: 10.1002/mgg3.82

Diagnosis of autosomal dominant polycystic kidney disease by efficient *PKD1* and *PKD2* multiplex high-throughput targeted sequencing

Daniel Trujillano[1,2,3,4], Gemma Bullich[5,6], Stephan Ossowski[7,2], José Ballarín[6], Roser Torra[6], Xavier Estivill[1,2,3,4*] and Elisabet Ars[5,6*]

[1]Genomics and Disease Group, Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Barcelona, Catalonia, Spain;

[2]Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain;

[3]Hospital del Mar Medical Research Institute (IMIM), Barcelona, Catalonia, Spain;

[4]CIBER in Epidemiology and Public Health (CIBERESP), Barcelona, Catalonia, Spain;

[5]Molecular Biology Laboratory, Fundació Puigvert, Instituto de Investigaciones Biomédicas Sant Pau (IIB-Sant Pau), Universitat Autònoma de Barcelona, REDinREN, Instituto de Investigación Carlos III, Barcelona, Catalonia, Spain;

[6]Nephrology Department, Fundació Puigvert, Instituto de Investigaciones Biomédicas Sant Pau (IIB-Sant Pau), Universitat Autònoma de Barcelona, REDinREN, Instituto de Investigación Carlos III, Barcelona, Catalonia, Spain;

[7]Genomic and Epigenomic Variation in Disease Group, Centre for Genomic Regulation (CRG), Barcelona, Catalonia, Spain;

[*]X. Estivill and E. Ars contributed equally to this work.

Correspondence to:

X. Estivill, Genomics and Disease Group, Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Doctor Aiguader 88, Barcelona, Catalonia 08003, Spain; xavier.estivill@crg.cat

E. Ars, Molecular Biology Laboratory, Fundació Puigvert, Cartagena 340-350, Barcelona, Catalonia 08025, Spain; ears@fundacio.puigvert.es

Running title: ADPKD diagnosis using targeted sequencing

For Peer Review

ABSTRACT

Genetic diagnostics of autosomal dominant polycystic kidney disease relies on linkage analysis and mutation screening of *PKD1* and *PKD2*, which is complicated by extensive allelic heterogeneity and especially by the presence of six highly homologous sequences of *PKD1* (exons 1-33). We have implemented a much more efficient strategy based on multiplex targeted resequencing of *PKD1* and *PKD2*, which avoids the laborious long-range PCRs followed by nested PCRs in the repeated part of *PKD1*. We have validated this approach in a cohort of 36 samples with previously known *PKD1* and *PKD2* mutations and five control samples. Pooled barcoded DNA libraries were enriched with a custom NimbleGen SeqCap EZ Choice library and sequenced with a HiSeq2000 sequencer. The combination of several robust bioinformatics tools allowed us to detect 35 out of 36 known definitely, highly likely, and likely pathogenic mutations. Then we used the same capture assay in a discovery cohort of 12 uncharacterized patients using a MiSeq sequencer. Our study is a proof-of-principle that targeted resequencing of the two genes involved in the most common cystic kidney disease, even in the complex duplicated regions of *PKD1*, can be successfully applied to routine genetic diagnostics of autosomal dominant polycystic kidney disease.

KEYWORDS

ADPKD, Molecular diagnostics, Genetic counseling, Targeted resequencing

For Peer Review

INTRODUCTION

Autosomal dominant polycystic kidney disease (ADPKD; OMIM ID: 173900) is the most common inherited cystic kidney disease, with an incidence of 1 in 400 to 1000 (Dalgaard and Norby, 1989; Iglesias, et al., 1983). ADPKD is caused by mutations in *PKD1* (16p13.3) in approximately 85% of the cases (The European Polycystic Kidney Disease Consortium, 1994), and in *PKD2* (4q21) in the remaining 15% (Mochizuki, et al., 1996). ADPKD is characterized by the development and progressive enlargement of cysts in the kidneys and other organs, eventually leading to end-stage renal disease (ESRD). The ADPKD phenotype displays a significant variability that is greatly influenced by the affected gene. Thus, *PKD1* patients have a median age at ESRD of 58 years compared to 79 years for *PKD2* mutated patients (Cornec-Le Gall, et al., 2013).

Diagnosis of ADPKD is mainly performed by renal imaging such as ultrasonography, computed tomography, or magnetic nuclear resonance. For patients with positive family history, specific ultrasonographic criteria regarding the number of cysts at an individual's age have been developed for the diagnosis of ADPKD (Pei, et al., 2009). However, these criteria are not very effective at diagnosing ADPKD in young individuals, in those with mild disease caused by mutations in *PKD2* and in *de novo* cases. Therefore, molecular diagnostics is necessary in several situations: 1) when a definite diagnosis is required in young individuals, such as a potential living related donor in an affected family with equivocal imaging data; 2) in patients with a negative family history of ADPKD, because of potential phenotypic overlap with several other kidney cystic diseases; 3) in families affected by early-onset polycystic kidney disease, since in this cases hypomorphic alleles and/or oligogenic inheritance can be involved (Bergmann, et al., 2011; Harris and Hopp, 2013; Rossetti, et al., 2009); and 4) in

For Peer Review

patients requesting genetic counseling, especially in couples wishing a preimplantation genetic diagnosis (Harris and Rossetti, 2010).

Approximately 70% of the 5' *PKD1* gene (exons 1–33) is duplicated six times on chromosome 16p within six pseudogenes (*PKD1P1-P6*), which share a 97.7% sequence identity with the genuine gene (Bogdanova, et al., 2001; Rossetti, et al., 2012). This, together with a high GC content, the presence of many missense variants, the absence of mutation hotspots, and the high allelic heterogeneity of ADPKD, makes the molecular diagnostics of ADPKD challenging. Most mutations are private variants, with a total of 929 pathogenic *PKD1* and 167 pathogenic *PKD2* mutations reported to date (June 2013, ADPKD Database [PKDB], http://pkdb.mayo.edu). Thus, genetic diagnosis by conventional techniques of a new ADPKD family requires long-range polymerase chain reaction (LR-PCR) of the repeated region of *PKD1* followed by nested PCRs (Rossetti, et al., 2002), combined with Sanger sequencing of all 46 *PKD1* and 15 *PKD2* exons. When pathogenic mutations are not identified by Sanger sequencing, multiplex ligation-dependent probe amplification (MLPA) analysis is also performed to identify potential deletions.

A recent study evaluated next generation sequencing (NGS) for the diagnostics of ADPKD, using a complex and laborious enrichment of *PKD1* and *PKD2* by pooling LR-PCR amplicons, which is the main obstacle for a high-throughput implementation in routine medical genetics. The authors assumed that genome capture strategies would not be suitable for the genetic screening of the duplicated regions of *PKD1* (Rossetti, et al., 2012). Therefore, there is a demand for more simple and cost-effective molecular approaches that could be used for routine diagnosis, especially now with the coming specific therapies that will require differential genetic diagnosis (Torres and Harris, 2006).

John Wiley & Sons, Inc.

For Peer Review

To address these challenges, we have developed and validated an assay that couples genome partitioning and NGS, to comprehensively explore in one-step the genetic complexity of *PKD1* and *PKD2*, as an alternative to cumbersome conventional genetic testing methods. We performed in-solution hybrid capture to enrich the complete genomic sequence of the *PKD1* and *PKD2* genes in a heterogeneous panel of ADPKD patients and control samples. We assessed the performance of this assay in a validation cohort of 36 patients and 5 controls, and then used it in a discovery cohort of 12 samples with unknown mutations, allowing accurate test reporting 5 days after receiving the DNA samples.

For Peer Review

MATERIALS AND METHODS

Subjects

High-quality genomic DNA from 53 unrelated samples was obtained from peripheral blood lymphocytes, using standard protocols. The validation cohort included 36 ADPKD patients and five control samples that had previously undergone conventional genetic diagnosis by Sanger sequencing of all *PKD1* and *PKD2* exons and, in cases with no definitive mutation, MLPA was also applied. The discovery cohort consisted of 12 ADPKD consecutive samples received for genetic diagnosis for which no mutations were known. ADPKD diagnosis was based on standard clinical and imaging criteria. All samples were anonymized in order to ensure the protection of their identity and the list of confirmed mutations was not provided to the investigators performing the bioinformatics mutation analysis until the end of the variant prioritization process. In the discovery cohort informed consent was obtained for all patients. The study was approved by the institutional review boards of each participating hospital and complies with the guidelines of the Declaration of Helsinki.

Capture and multiplexed resequencing of the *PKD1* and *PKD2* genes

To carry out DNA capture, we designed an elaborated custom NimbleGen SeqCap EZ Choice Library (Roche, Inc., Madison, WI, USA) to target the complete genomic sequence of the *PKD1* and *PKD2* genes, and 1 kb of genomic sequence flanking at the 5′ and 3′ ends of each gene, accounting for 121,322 bp. Our capture design also included probes to target the entire genomic region (plus 1 kb at each end) or all exons, splice sites and the immediately adjacent intronic sequences (plus 100 bp at each end) of 125 additional genes related to kidney diseases, for a total of 2.1 Mb of captured DNA after removal of repetitive sequences. DNA baits were selected using the most stringent settings for probe design (uniqueness tested by Sequence Search and Alignment by

John Wiley & Sons, Inc.

For Peer Review

Hashing Algorithm [SSAHA]) (Ning, et al., 2001). In order to overcome the limitations of in-solution hybridization for the capture of the duplicated *PKD1* regions, we altered the parameters for probe design of this specific region to allow probes to have up to 10 close matches in the genome. No probe redundancy was allowed in the final capture design for the rest of target regions. The BED file of captured regions is available on request to the authors.

Libraries were prepared with the TruSeq DNA Sample Preparation Kits (Illumina, Inc., San Diego, CA, USA). Genomic capture from pooled libraries was carried out using NimbleGen SeqCap EZ Library (Roche, Inc.) following User's Guide v3.0 instructions, as previously described (Trujillano, et al., 2013). The libraries of the samples of the validation cohort and the five control samples were prepared and sequenced together with 7 samples of other kidney diseases enriched using the same capture design and enrichment protocol in two pools of 24 samples, for a total of 48 samples multiplexed in two HiSeq 2000 (Illumina, Inc.) lanes to generate 2x100 bp paired-end reads. The 12 samples of the discovery cohort were enriched in a single capture reaction and were sequenced in a Miseq (Illumina, Inc.) run to generate 2x250 bp paired-end reads.

Bioinformatics analysis and mutation identification and classification

The resulting fastq files were analyzed with an in-house developed pipeline previously described (Trujillano, et al., 2013). All the bioinformatics tools used in this study were run using standard parameters unless stated otherwise. Briefly, reads were aligned to the human reference genome hg19 using the Burrows-Wheeler Aligner (bwa aln) version 0.5.9 (Li and Durbin, 2009), allowing for maximally six mismatches and one gap of up to 20 bp. After re-alignment around potential insertions/deletions and SNP clusters, base-quality recalibration and duplication marking using the GATK pipeline (McKenna, et al., 2010), and picard-tools (http://picard.sourceforge.net), the resulting alignments

John Wiley & Sons, Inc.

For Peer Review

were used as input for three different variant prediction tools, namely GATK Unified Genotyper (DePristo, et al., 2011), samtools mpileup (Li, et al., 2009) and SHORE (http://1001genomes.org). Large InDels and SVs were identified using Pindel (Ye, et al., 2009), Conifer (Krumm, et al., 2012), and PeSV-Fisher (Escaramis, et al., 2013). Functional annotation of variants was performed using Annovar (Wang, et al., 2010).

In order to identify pathogenic mutations that could cause ADPKD, we applied the following cascade of filtering steps (Walsh, et al., 2010):

1. We required all candidate variants on both sequenced DNA strands and to account for ≥15% of total reads at that site.
2. Common polymorphisms (≥5% in the general population) were discarded by comparison with dbSNP 132, the 1000G, the Exome Variant Server (http://evs.gs.washington.edu), and an in-house exome variant database to filter out both common benign variants and recurrent artifact variant calls. However, since these databases contain known disease-associated mutations, all detected variants were compared to gene-specific mutation databases (The Human Gene Mutation Database [HGMD], www.hgmd.cf.ac.uk and ADPKD Database [PKDB], http://pkdb.mayo.edu).
3. Mutations that could give rise to premature protein truncating mutations, that is, stop mutations, exonic deletions/insertions and large genomic rearrangements were classified as definitely pathogenic.
4. Missense and non-canonical splicing variants were considered *a priori* Unclassified Sequence Variants (UCV) and their potential pathogenicity was evaluated using an *in silico* scoring system developed for *PKD1* and *PKD2* genes as previously described (Rossetti, et al., 2007). This scoring system takes into consideration a number of *in silico* predictors (Grantham, 1974; Rossetti, et al., 2007; Tavtigian, et al., 2006) and

For Peer Review

population data. We scored each of these factors, the sum of which resulted in an overall Variant Score (VS). The UCV were classified into 4 groups (Rossetti, et al., 2007): highly likely pathogenic (VS≥11); likely pathogenic (5≤VS≤10), indeterminate (0≤VS≤4), and highly likely neutral (VS ≤1).

We considered to be pathogenic mutations those sequence variants predicted to result in a truncated protein (classified as definitely pathogenic) and those not found in healthy controls, that segregated with the disease in families and expected to severely alter the protein sequence using *in silico* predictors (classified as highly likely pathogenic and likely pathogenic variants).

Validation of newly identified single nucleotide variants and large deletion

Validation of variants of the discovery cohort was performed by LR-PCRs of the repeated region of *PKD1* followed by nested PCRs (exons 1-33) and by conventional PCRs for the non-duplicated *PKD1* exons using conditions described previously (Rossetti, et al., 2002) as well as for *PKD2* exons (Hayashi, et al., 1997), combined with direct sequencing. The *PKD1* deletion was confirmed by MLPA using the Salsa MLPA kit P351-B1/P352-B1 (MRC-Holland, Amsterdam, Netherlands).

For Peer Review

RESULTS

PKD1 and *PKD2* enrichment

We designed an elaborated pool of specific and unspecific oligonucleotides to capture the duplicated region of *PKD1* and the rest of *PKD1* and *PKD2*, including all exons, introns, and 1 kb of 5′ and 3′ flanking genomic regions. After removal of repetitive sequences, 81.21% of the targeted bases could be covered with capture baits for a final targeted region of 98,524 bp divided in 99 individual regions, with lengths ranging from 65 to 6,493 bp (average of 995 bp) (Table 1). Noteworthy, 100% of all coding sequences, i.e. the complete 46 and 15 exons of *PKD1* and *PKD2*, respectively, were covered by capture baits. The target regions that precluded bait tilling correspond only to intronic and intergenic sequences. Our capture design also included probes to target 125 additional genes related to kidney diseases, for a total of 2.1 Mb of captured DNA. However, this study focuses only in *PKD1* and *PKD2*.

Sequencing statistics

In the validation cohort, it was achieved an evenly distributed mean depth of coverage for the targeted genes of 331X and 481X for *PKD1* and *PKD2*, respectively, on average across samples. The percentage of targeted bases that were covered by at least 5 reads (the minimum that we require for variant calling) was of 96.78% for *PKD1* and 99.43% for *PKD2* (Table 2). Regarding the coverage specific to the exonic regions, we achieved a sequencing depth of 289X for the 46 exons of PKD1 and 453X for the 15 exons of *PKD2*, on average across samples (Table 3). Ninety-five percent of the coding basepairs of *PKD1* and 94% of *PKD2* were covered by more than 20 reads, which is enough for an accurate detection of known and novel mutations. Only exons 1 and 42 of *PKD1* and exon 1 of *PKD2* were not captured and sequenced at an adequate read depth (Figure 1).

John Wiley & Sons, Inc.

For Peer Review

Due to the lower throughput of the MiSeq sequencer, the average coverage achieved in the discovery cohort was of 81X and 174X for *PKD1* and *PKD2*, respectively, across the 12 samples (Table 2). For a comprehensive summary of the obtained sequencing results, see also Supplementary Tables 1 and 2.

Identification of *PKD1* and *PKD2* mutations

The selection of the samples for the validation cohort was done with the idea to include as many different types of *PKD1* and *PKD2* mutations as possible to simulate a real-world ADPKD diagnostics scenario, including single nucleotide variants (SNVs), short insertions/deletions (InDels) and large structural variants (SVs), so that we could test the effectiveness of our assay for all these types of genetic variation. To assess the sensitivity of our assay to detect pathogenic mutations, we blindly inspected all mapped sequence reads from the 36 ADPKD patients with previously defined mutations in *PKD1* and *PKD2*, and five control samples. Then we applied our variant prioritization strategy to identify definitely, highly likely and likely pathogenic mutations and detected 35 previously known different pathogenic mutations (30 in *PKD1* and 5 in *PKD2*) in their correct heterozygous state across the 36 ADPKD patients included in the validation cohort (Table 4). Noteworthy, a previously unknown *PKD1* deletion was identified in patient 03-106-P6, located *in cis* with the highly pathogenic missense variant c.4645C>T (p.Arg1549Trp), with predicted breakpoints in chr16:2154344-2186386 (Figure 2A). This deletion was confirmed by MLPA analysis showing a deletion starting at 5'UTR until exon 22. We also detected a previously known large *PKD2* deletion in patient 11-571-P2 (Figure 2B) with predicted breakpoints in chr4:88952828-89050618.

For Peer Review

The discovery cohort consisted of 12 consecutive samples received for ADPKD diagnosis, for which no mutations were previously known. We detected pathogenic mutations in 10 out of 12 patients of the discovery cohort carrying a total of 11 different pathogenic mutations (10 in *PKD1* and one in *PKD2*). All variants were confirmed by Sanger sequencing (Table 5). Interestingly, we identified one patient (12-444) harboring one definitively pathogenic mutation in *PKD2* and one highly likely pathogenic mutation in *PKD1*, presenting a more severe phenotype compared to the rest of the family.

Sensitivity, specificity and accuracy of the assay

This study led to the identification of 35 out of 36 previously known mutations of the validation cohort, including SNVs, InDels and a large deletion in their correct heterozygous state. Manual inspection of the missing c.2_3insT p.Met1IlefsX113 variant in sample 03-393-P3 revealed that it is localized in a region of *PKD1*'s exon 1 which had not been captured and, thus, no NGS data was available. Also, we reached a diagnostic rate of 10 out of 12 patients in the discovery cohort in which we detected 11 pathogenic ADPKD mutations. The two samples that were not successfully characterized by our NGS assay were then screened by conventional Sanger sequencing and the two causal mutations were identified. Then, by manually inspecting the mapping files of the NGS reads, we realized that we had lost p.Val2768Met (sample 13-102) and p.Arg4021fs (sample 07-335) because their locations were in poorly covered areas of *PKD1* and the algorithms discarded them as potential false positives calls.

We included 5 control samples to determine the clinical specificity of our assay, since these individuals had no personal or family history of ADPKD. No spurious pathogenic mutations were found in these samples. These samples had been previously genotyped

John Wiley & Sons, Inc.

For Peer Review

with a HumanOmni 2.5-8 BeadChip (Illumina, Inc.) and were also used to determine the analytic sensitivity of our assay to detect heterozygous and homozygous SNVs. Genotype data were available for a total of 12,108 sites in aggregate across the five control samples within the whole 2.1 Mb of captured regions, being 80 and 269 genotyped sites within the targeted regions of *PKD1* and *PKD2*, respectively. Our assay correctly identified 3,344/3,511 homozygous and heterozygous SNVs in the 2.1 Mb captured, for a sensitivity of 95.2%, demonstrating high sensitivity of calling all variants across each captured region. Sensitivity was of 100% both for *PKD1* (20/20) and *PKD2* (103/103). We next assessed analytic specificity by analyzing 8,597 known non-variant (reference sequence) sites in the 5 genotyped control samples. Our assay correctly identified 8,593/8,597 sites as non-variant from the reference genome for an analytic specificity of 99.9%. Analytic specificity was 100% both for *PKD1* (60/60) and *PKD2* (166/166). We calculated the positive predictive value (PPV) as [number of true positives]/[number of true positives + number of false positives], to assess the performance of our assay as diagnostic method. PPV was 99.9%, 100% and 100% for all captured regions, *PKD1* and *PKD2*, respectively (Supplementary Table 3).

For Peer Review

DISCUSSION

The purpose of this study was to establish targeted resequencing by in-solution hybridization as a routine method for the molecular diagnostics of ADPKD. We have identified 35 of 36 previously known mutations in the validation cohort, in addition to a previously unknown large *PKD1* deletion, with zero spurious calls in the control samples. Noteworthy, most of these mutations were located within the segmentally duplicated regions of the *PKD1* gene. In the discovery cohort, we reached a diagnostic rate of 10 out of 12 patients, allowing test reporting five days after receiving the DNA samples.

Recently, targeted resequencing by NGS has been used in the identification of mutations in ADPKD. Rossetti et al. (2012) did not apply a capture protocol for *PKD1* and *PKD2* enrichment since they speculated that the duplicated genomic regions of *PKD1* would lead to concurrent capture of the six *PKD1* pseudogenes. Instead, these authors used a strategy of pooling equimolar LR-PCR amplicons and multiplexing barcoded libraries. Their approach showed a high sensitivity, specificity and accuracy, but it is a very laborious task more amenable to characterize large ADPKD populations than for routine genetic diagnosis. Moreover, their approach did not allow detecting large genomic rearrangements. Here, we do not only demonstrate that genome enrichment by in-solution hybridization using an elaborated probe design is an accurate strategy for mutation identification in the duplicated regions and the rest of *PKD1* and *PKD2*, but also that this strategy is ready to substitute LR-PCR-based methods in the routine genetic diagnostics of ADPKD to detect all sorts of sequence variants, including SVs. Only minor modifications are required in the design of the probes to fix the issues with the capture and sequencing of exons 1 and 42 of *PKD1* and exon 1 of *PKD2*.

For Peer Review

When we conceived this study we assumed that it would be extremely difficult to design capture probes specific to the duplicated region of the genuine *PKD1*, i.e. there would always be residual enrichment of the six pseudogenes. So, instead of excluding this region from our assay we decided to include in our capture library unspecific probes (with up to 10 close matches in the genome) to the duplicated region of *PKD1*. From our point of view the critical point of the assay was not the presence of sequencing reads coming from both the genuine *PKD1* and its pseudogenes. Instead, the major challenge was to map these reads coming from duplicated regions unambiguously to the genuine *PKD1* or to the six pseudogenes. We performed the mapping of the sequencing reads using standard mapping parameters, but instead of restricting it to the targeted region we allowed mapping to the whole genome. Moreover, the length of the sequencing reads produced in this study (2x100 bp and 2x250 bp in the validation and discovery cohorts, respectively) and the insert sizes in the DNA libraries (300 bp approximately) allowed us to unambiguously map a big proportion of the sequencing reads to *PKD1* or to its pseudogenes. However, as evidenced by the lower sequence coverage achieved for *PKD1* when compared to the rest of the captured regions, the proportion of unambiguously mapped sequencing reads in this region was significantly lower. This means that part of the sequences generated for *PKD1* and its pseudogenes were discarded during the mapping steps of the bioinformatics pipeline, since they could not be attributed to solely one genomic region. Fortunately, the high depth of coverage achieved neutralized this problem allowing confident variant calling across almost all captured regions of *PKD1* (Table 2).

The main drawbacks of conventional genetic diagnostics of ADPKD are the absence of mutation hot-spots in *PKD1* and *PKD2,* the lack of analysis of intronic and regulatory regions and, of course, the complexity to screen the duplicated region of *PKD1*,

For Peer Review

resulting in a costly and time-consuming diagnosis protocol that may even take two months to complete diagnosis. NGS technologies overcome these limitations offering great sensitivity and specificity at the necessary throughput for the detection of all sorts of mutations, including atypical, rare and deep intronic variants, which is especially important for inherited disorders for which no mutation hot-spots are known.

We estimate that with our NGS-based assay a 60% of cost savings per sample could be achieved, and the whole diagnostics process could be a minimum of 5 times faster than the conventional techniques currently used for the genetic diagnostics of ADPKD. In addition, our strategy offers an almost complete definition of the captured genes, without the need for stepwise testing anymore and having to choose which gene to sequence first, and capable to detect large genomic rearrangements and deep intronic variants. For the discovery cohort, in which the MiSeq (Illumina, Inc.) was used, the complete process of library preparation, sequence enrichment, NGS, and bioinformatics analysis was completed in five days after reception of the DNA samples. In the validation cohort we detected 97% of the known mutations, while in the small discovery cohort we reached a definitive diagnosis in 10 out of 12 patients (83%). Although the size of our cohort is modest, these results are very encouraging since these numbers represent a diagnostic rate comparable to data recently obtained by Sanger sequencing (Audrezet, et al., 2012; Cornec-Le Gall, et al., 2013) and the 63% obtained in the previous *PKD1*/*PKD2* NGS study (Rossetti, et al., 2012). We suspect that the lower mutation detection rate in the discovery cohort with respect to the validation cohort may be explained by the lower depth of coverage yielded by the MiSeq (Illumina, Inc.). However, we plan in the future to produce a capture design specific for *PKD1*/*PKD2* that will not be limited by carrying-over the additional 125 genes that we have included in our capture design prototype. This would significantly reduce the total captured DNA

For Peer Review

per sample, allowing multiplexing more samples per MiSeq (Illumina, Inc.) run, achieving depths of coverage comparable to those obtained for the validation cohort.

In conclusion, we illustrate here the first study successfully using in-solution hybridization coupled to NGS to detect ADPKD pathogenic mutations, both in the duplicated regions of *PKD1* and the rest of *PKD1* and *PKD2* genes. Our approach is cost and time-saving, and meets the sensitivity and specificity criteria required for genetic diagnostics, being ready to substitute classic molecular tools in routine genetic diagnostics of ADPKD.

For Peer Review

AKNOWLEDGEMENTS

We thank the subjects and referring physicians who participated in this study. We thank Justo Gonzalez, Sheila Santin, Laia Ejarque, Estefania Eugui and Patrícia Ruiz for technical support, Cristian Tornador and Georgia Escaramis for bioinformatics assistance, and IIB Sant Pau-Fundació Puigvert Biobank for kindly providing some of the ADPKD samples, ReTBioH (Spanish Biobank Network) RD09/76/00064. This project was funded by the Spanish Plan Nacional SAF2008-00357 (NOVADIS); the Instituto de Salud Carlos III (FIS/FEDER PI11/00733); the European Commission 7th Framework Program, Project N. 261123 (GEUVADIS), and Project N. 262055 (ESGI); the Spanish Healthy Ministry (FIS 12/01523), REDINREN (Spanish Renal Network for Research 16/06, RETICS, Instituto de Investigación Carlos III), RD09/76/00064; the Catalan Government (AGAUR 2009/SGR-1116); and the Fundación Renal Iñigo Álvarez de Toledo in Spain. D.T. is a PhD student supported by the Spanish Ministry of Economy and Competiveness. R.T. is supported by Intensification Programm of Research Activity ISCIII/Generalitat de Catalunya (programm I3SN).

For Peer Review

DISCLOSURES

All the authors declare no competing interests.

AUTHOR CONTRIBUTIONS

This study was conceived and designed by D.T., R.T., X.E. and E.A. Selection of samples was performed by G.B., R.T. and E.A. NGS libraries were prepared by G.B. The bioinformatics pipeline and the NGS analysis were performed by D.T. and S.O. Sanger confirmation of mutations was performed by G.B. The manuscript was written by D.T., G.B., X.E. and E.A.

REFERENCES

The European Polycystic Kidney Disease Consortium. 1994. The polycystic kidney disease 1 gene encodes a 14 kb transcript and lies within a duplicated region on chromosome 16. Cell 77(6):881-94.

Audrezet MP, Cornec-Le Gall E, Chen JM, Redon S, Quere I, Creff J, Benech C, Maestri S, Le Meur Y, Ferec C. 2012. Autosomal dominant polycystic kidney disease: comprehensive mutation analysis of PKD1 and PKD2 in 700 unrelated patients. Hum Mutat 33(8):1239-50.

Bergmann C, von Bothmer J, Ortiz Bruchle N, Venghaus A, Frank V, Fehrenbach H, Hampel T, Pape L, Buske A, Jonsson J, et al. 2011. Mutations in multiple PKD genes may explain early and severe polycystic kidney disease. J Am Soc Nephrol 22(11):2047-56.

Bogdanova N, Markoff A, Gerke V, McCluskey M, Horst J, Dworniczak B. 2001. Homologues to the first gene for autosomal dominant polycystic kidney disease are pseudogenes. Genomics 74(3):333-41.

Cornec-Le Gall E, Audrezet MP, Chen JM, Hourmant M, Morin MP, Perrichot R, Charasse C, Whebe B, Renaudineau E, Jousset P, et al. 2013. Type of PKD1 Mutation Influences Renal Outcome in ADPKD. J Am Soc Nephrol 24(6):1006-13.

Dalgaard OZ, Norby S. 1989. Autosomal dominant polycystic kidney disease in the 1980's. Clin Genet 36(5):320-5.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation

For Peer Review

discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491-8.

Escaramis G, Tornador C, Bassaganyas L, Rabionet R, Tubio JM, Martinez-Fundichely A, Caceres M, Gut M, Ossowski S, Estivill X. 2013. PeSV-Fisher: identification of somatic and non-somatic structural variants using next generation sequencing data. PLoS One 8(5):e63377.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185(4154):862-4.

Harris PC, Hopp K. 2013. The Mutation, a Key Determinant of Phenotype in ADPKD. J Am Soc Nephrol 24(6):868-70.

Harris PC, Rossetti S. 2010. Molecular diagnostics for autosomal dominant polycystic kidney disease. Nat Rev Nephrol 6(4):197-206.

Hayashi T, Mochizuki T, Reynolds DM, Wu G, Cai Y, Somlo S. 1997. Characterization of the exon structure of the polycystic kidney disease 2 gene (PKD2). Genomics 44(1):131-6.

Iglesias CG, Torres VE, Offord KP, Holley KE, Beard CM, Kurland LT. 1983. Epidemiology of adult polycystic kidney disease, Olmsted County, Minnesota: 1935-1980. Am J Kidney Dis 2(6):630-9.

Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, Quinlan AR, Nickerson DA, Eichler EE. 2012. Copy number variation detection and genotyping from exome sequence data. Genome Res 22(8):1525-32.

For Peer Review

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754-60.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078-9.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297-303.

Mochizuki T, Wu G, Hayashi T, Xenophontos SL, Veldhuisen B, Saris JJ, Reynolds DM, Cai Y, Gabow PA, Pierides A, et al. 1996. PKD2, a gene for polycystic kidney disease that encodes an integral membrane protein. Science 272(5266):1339-42.

Ning Z, Cox AJ, Mullikin JC. 2001. SSAHA: a fast search method for large DNA databases. Genome Res 11(10):1725-9.

Pei Y, Obaji J, Dupuis A, Paterson AD, Magistroni R, Dicks E, Parfrey P, Cramer B, Coto E, Torra R, et al. 2009. Unified criteria for ultrasonographic diagnosis of ADPKD. J Am Soc Nephrol 20(1):205-12.

Rossetti S, Consugar MB, Chapman AB, Torres VE, Guay-Woodford LM, Grantham JJ, Bennett WM, Meyers CM, Walker DL, Bae K, et al. 2007. Comprehensive molecular diagnostics in autosomal dominant polycystic kidney disease. J Am Soc Nephrol 18(7):2143-60.

For Peer Review

Rossetti S, Chauveau D, Walker D, Saggar-Malik A, Winearls CG, Torres VE, Harris PC. 2002. A complete mutation screen of the ADPKD genes by DHPLC. Kidney Int 61(5):1588-99.

Rossetti S, Hopp K, Sikkink RA, Sundsbak JL, Lee YK, Kubly V, Eckloff BW, Ward CJ, Winearls CG, Torres VE, et al. 2012. Identification of gene mutations in autosomal dominant polycystic kidney disease through targeted resequencing. J Am Soc Nephrol 23(5):915-33.

Rossetti S, Kubly VJ, Consugar MB, Hopp K, Roy S, Horsley SW, Chauveau D, Rees L, Barratt TM, van't Hoff WG, et al. 2009. Incompletely penetrant PKD1 alleles suggest a role for gene dosage in cyst initiation in polycystic kidney disease. Kidney Int 75(8):848-55.

Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, de Silva D, Zharkikh A, Thomas A. 2006. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J Med Genet 43(4):295-305.

Torres VE, Harris PC. 2006. Mechanisms of Disease: autosomal dominant and recessive polycystic kidney diseases. Nat Clin Pract Nephrol 2(1):40-55; quiz 55.

Trujillano D, Ramos MD, Gonzalez J, Tornador C, Sotillo F, Escaramis G, Ossowski S, Armengol L, Casals T, Estivill X. 2013. Next generation diagnostics of cystic fibrosis and CFTR-related disorders by targeted multiplex high-coverage resequencing of CFTR. J Med Genet 50(7):455-62.

Walsh T, Lee MK, Casadei S, Thornton AM, Stray SM, Pennil C, Nord AS, Mandell JB, Swisher EM, King MC. 2010. Detection of inherited mutations for breast and

For Peer Review

ovarian cancer using genomic capture and massively parallel sequencing. Proc Natl Acad Sci U S A 107(28):12629-33.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38(16):e164.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25(21):2865-71.

For Peer Review

FIGURE LEGENDS

Figure 1. Representation of the average depth of coverage of *PKD1* (A) and *PKD2* (B) in the validation cohort. Red lines and the numbers underneath represent the exons of the genes. Green lines represent the regions tilled by capture baits.

Figure 2. Detection of large deletions in the *PKD1* and *PKD2* genes by normalised depth of coverage analysis. Representation of the SVD-ZRPKM values calculated by Conifer for the 36 samples and 5 controls of the validation cohort. Yellow peaks indicate the two large deletions identified in this study. A) Sample 03-106-P6 *PKD1* deletion with breakpoints in chr16:2154344-2186386. B) Sample 11-571-P2 *PKD2/ABCG2* deletion with breakpoints in chr4:88952828-89050618.

For Peer Review

Table 1. Target regions for capture of *PKD1* and *PKD2*

Gene	Targeted region		Covered by baits		
	Coordinates	**Size (bp)**	**Size (bp)**	**%**	**Individual regions**
PKD1	chr16:2137710-2186899	49189	41086	83,53	21
PKD2	chr4:88927798-88999931	72133	57438	79,63	78
All	-	121322	98524	81,21	99

Table 2. Average sequencing quality control and coverage statistics of *PKD1* and *PKD2* in the validation and discovery cohorts

Cohort	Validation		Discovery	
	Average	SD	Average	SD
QC-passed reads	14452006,67	2252761,13	1303016,25	293339,48
mapped	14328976,12	2236282,41	1002567,63	269009,02
properly paired	14140971,70	2203337,48	780154,25	265250,15
ALL Mean coverage (X)	487,23	80,86	167,85	27,81
% ALL target bases covered = 0X	0,08	0,03	0,31	0,04
% ALL target bases covered >= 1X	99,92	0,03	99,84	0,02
% ALL target bases covered >= 5X	99,82	0,09	99,55	0,06
% ALL target bases covered >= 10X	99,74	0,13	99,28	0,13
% ALL target bases covered >= 20X	99,59	0,22	98,46	0,44
% ALL target bases covered >= 50X	98,88	0,80	91,25	3,83
% ALL target bases covered >= 100X	96,42	2,75	65,34	7,82
PKD1 Mean coverage (X)	331,14	89,20	80,60	13,60
% PKD1 target bases covered = 0X	1,98	0,35	3,70	0,39
% PKD1 target bases covered >= 1X	98,02	0,35	98,15	0,20
% PKD1 target bases covered >= 5X	96,78	0,68	95,86	0,62
% PKD1 target bases covered >= 10X	96,28	1,13	92,97	1,19
% PKD1 target bases covered >= 20X	95,54	1,98	86,69	2,59
% PKD1 target bases covered >= 50X	92,40	4,13	65,03	4,76
% PKD1 target bases covered >= 100X	84,78	6,84	52,01	1,65
PKD2 Mean coverage (X)	480,73	87,98	174,22	28,72
% PKD2 target bases covered = 0X	0,36	0,11	0,90	0,20
% PKD2 target bases covered >= 1X	99,64	0,11	99,55	0,10
% PKD2 target bases covered >= 5X	99,43	0,15	99,25	0,04
% PKD2 target bases covered >= 10X	99,35	0,17	99,17	0,06
% PKD2 target bases covered >= 20X	99,19	0,20	98,75	0,31
% PKD2 target bases covered >= 50X	98,68	0,37	92,74	3,46
% PKD2 target bases covered >= 100X	97,10	2,22	67,51	8,70

Table 3. Average coverage statistics of *PKD1* and *PKD2* exons in the validation cohort

Gene	Exon #	Average Coverage	% bp = 0X	% bp >= 1X	% bp >= 5X	% bp >= 10X	% bp >= 20X	% bp >= 50X	% bp >= 100X
PKD1	All	288,97	3,22	96,78	96,20	95,14	94,64	92,05	84,89
	1	0,37	88,18	11,82	1,89	0	0	0	0
	2	141,94	0	100	100	100	100	100	100
	3	146,97	0	100	100	100	100	100	100
	4	137,44	0	100	100	100	100	100	100
	5	114,21	0	100	100	100	100	77,50	46,50
	6	186,27	0	100	100	100	100	100	81,42
	7	48,67	0	100	100	100	100	30	0
	8	145,50	0	100	100	100	100	100	87,83
	9	99,66	0	100	100	100	100	89,68	53,17
	10	124,93	0	100	100	100	100	100	100
	11	155,15	0	100	100	100	100	100	84,64
	12	319,85	0	100	100	100	100	100	97,71
	13	619,92	0	100	100	100	100	100	100
	14	300,52	0	100	100	100	100	100	100
	15	385,25	0	100	100	100	100	100	100
	16	209,83	0	100	100	100	100	100	100
	17	243,94	0	100	100	100	100	100	100
	18	78,10	0	100	100	100	100	100	26,52
	19	92,43	0	100	100	100	100	100	0,94
	20	291,07	0	100	100	100	100	100	100
	21	201,80	0	100	100	100	100	100	100
	22	112,97	0	100	100	100	100	100	68,06
	23	228,21	0	100	100	100	100	100	100
	24	356,27	0	100	100	100	100	100	100
	25	301,58	0	100	100	100	100	100	100
	26	283,75	0	100	100	100	100	100	100
	27	491,08	0	100	100	100	100	100	100
	28	258,90	0	100	100	100	100	100	100
	29	238,94	0	100	100	100	100	100	100
	30	346,30	0	100	100	100	100	100	100
	31	421,09	0	100	100	100	100	100	100
	32	336,12	0	100	100	100	100	100	100
	33	331,53	0	100	100	100	100	100	100
	34	413,05	0	100	100	100	100	100	100
	35	306,75	0	100	100	100	100	100	100
	36	178,56	0	100	100	100	100	100	100
	37	460,48	0	100	100	100	100	100	100
	38	610,01	0	100	100	100	100	100	100
	39	378,85	0	100	100	100	100	100	100
	40	512,62	0	100	100	100	100	100	100
	41	171,52	0	100	100	100	100	100	100
	42	4,59	46,55	53,45	33,33	20,69	0	0	0
	43	36,26	0	100	98,28	57,24	45,17	29,31	9,31
	44	407,56	0	100	100	100	100	100	100
	45	395,70	0	100	100	100	100	100	100
	46	449,00	0	100	100	100	100	100	100
PKD2	All	452,51	3,76	96,24	95,65	95,33	93,62	90,74	86,59
	1	32,62	28,05	71,95	67,55	65,20	52,42	30,98	0
	2	401,73	0	100	100	100	100	100	100
	3	696,56	0	100	100	100	100	100	100
	4	689,57	0	100	100	100	100	100	100
	5	700,07	0	100	100	100	100	100	100
	6	622,80	0	100	100	100	100	100	100
	7	291,75	0	100	100	100	100	100	100
	8	578,01	0	100	100	100	100	100	100
	9	501,30	0	100	100	100	100	100	100
	10	367,60	0	100	100	100	100	100	100
	11	469,00	0	100	100	100	100	100	100
	12	795,21	0	100	100	100	100	100	100
	13	947,86	0	100	100	100	100	100	100
	14	755,93	0	100	100	100	100	100	100
	15	427,93	0	100	100	100	100	100	100

Table 4. ADPKD (*PKD1* and *PKD2*) mutations identified in the 36 samples of the validation cohort

Sample	Gene	Exon/Intron	Duplicated Region	cDNA change	Protein change	PKDB	# patients	Classification	Ref counts	Var Counts	Var %
12-331-P1	PKD1	EX05	Yes	c.566C>G	p.Ser189X	Present	1	Definitely Pathogenic	188	62	0,25
12-382-P1	PKD1	EX05	Yes	c.736_737del	p.Ser246Profs*14	Absent	0	Definitely Pathogenic	19	14	0,42
04-016-P6	PKD1	EX09	Yes	c.1831C>T	p.Arg611Trp	Present	1	Likely Pathogenic	25	13	0,34
12-235-P1	PKD1	EX11	Yes	c.2329C>T	p.Gln777*	Absent	0	Definitely Pathogenic	59	37	0,39
12-366-P1	PKD1	EX11	Yes	c.2478delC	p.Ile827Serfs*71	Absent	0	Definitely Pathogenic	114	83	0,42
03-106-P6	PKD1	EX15	Yes	c.4645C>T	p.Arg1549Trp	Absent	0	Highly Likely Pathogenic	132	125	0,49
	PKD1	EX1_EX22	Yes	c.1-?_8161+?del	p.Met1fs	Absent	0	Definitely Pathogenic	-	-	-
02-010-P6	PKD1	EX15	Yes	c.6583_6589del7	p.Cys2195fs*14	Present	1	Definitely Pathogenic	74	79	0,52
10-326-P3	PKD1	EX15	Yes	c.6778_6780delATT	p.Ile2260del	Present	1	Highly Likely Pathogenic	190	148	0,44
11-220-P2	PKD1	EX15	Yes	c.6221delA	p.Asn2074fs*42	Absent	0	Definitely Pathogenic	123	113	0,48
11-247-P7	PKD1	EX15	Yes	c.6384C>A	p.Asn2128Lys	Absent	0	Highly Likely Pathogenic	181	131	0,42
11-517-P1	PKD1	EX15	Yes	c.6736C>T	p.Gln2246*	Present	1	Definitely Pathogenic	185	124	0,40
11-525-P2	PKD1	EX15	Yes	c.6827T>C	p.Leu2276Pro	Absent	0	Highly Likely Pathogenic	280	245	0,47
12-010-P1	PKD1	EX15	Yes	c.4888C>T	p.Gln1630*	Present	1	Definitely Pathogenic	136	114	0,46
12-161-P1	PKD1	EX15	Yes	c.6586C>T	p.Gln2196*	Present	1	Definitely Pathogenic	86	50	0,37
10-388-P3	PKD1	IVS22	Yes	c.8161+1G>C	-	Absent	0	Definitely Pathogenic	21	25	0,54
10-463-P3	PKD1	EX23	Yes	c.8311G>A	p.Glu2771Lys	Present	18	Highly Likely Pathogenic	58	63	0,52
11-468-P1	PKD1	EX23	Yes	c.8251C>T	p.Gln2751*	Absent	0	Definitely Pathogenic	92	92	0,50
12-363-P1	PKD1	EX23	Yes	c.8285delT	p.Ile2762Thrfs*13	Absent	0	Definitely Pathogenic	120	89	0,43
11-457-P2	PKD1	EX24	Yes	c.8858A>G	p.Asn2953Ser	Absent	0	Highly Likely Pathogenic	236	191	0,45
11-287-P2	PKD1	EX26	Yes	c.9240_9241delAT	p.Ala3082fs*95	Present	3	Definitely Pathogenic	164	80	0,33
10-193-P3	PKD1	EX27	Yes	c.9412G>A	p.Val3138Met	Present	2	Likely Pathogenic	208	188	0,47
11-595-P2	PKD1	EX27	Yes	c.9455_9456insC	p.Arg3152fs*27	Absent	0	Definitely Pathogenic	283	184	0,39
07-172-P5	PKD1	EX29	Yes	c.9889G>A	p.Val3297Met	Absent	0	Likely Pathogenic	130	114	0,47
09-403-P3	PKD1	IVS31	Yes	c.10170+25_+45del19	p.Gln3390fs	Present	2	Highly Likely Pathogenic	96	26	0,21
10-182-P3	PKD1	EX37	-	c.11017-10C>A	p.Arg3672fs	Present	7	Highly Likely Pathogenic	109	83	0,43
12-144-P1	PKD1	EX37	-	c.10847C>A	p.Ser3616*	Absent	0	Definitely Pathogenic	224	177	0,44
10-353-P3	PKD1	EX40	-	c.11359_11360del	p.Pro3788fs*26	Absent	0	Definitely Pathogenic	292	203	0,41
11-256-P2	PKD1	EX41	-	c.11471G>T	p.Gly3824Val	Absent	0	Likely Pathogenic	85	63	0,43
11-168-P8	PKD1	IVS43	-	c.12004-2A>G	-	Absent	0	Definitely Pathogenic	143	113	0,44
09-446-P3	PKD1	EX44	-	c.12031C>T	p.Gln4011*	Present	4	Definitely Pathogenic	106	85	0,45
11-133-P8	PKD2	EX01	-	c.224delC	p.Pro75fs*42	Absent	0	Definitely Pathogenic	41	25	0,38
11-008-P3	PKD2	EX02	-	c.637C>T	p.Arg213*	Absent	0	Definitely Pathogenic	198	172	0,46
11-170-P2	PKD2	EX04	-	c.965G>A	p.Arg322Gln	Present	4	Highly Likely Pathogenic	396	330	0,45
12-149-P1	PKD2	EX10	-	c.2050_2053del4	p.Tyr684Leufs*2	Present	1	Definitely Pathogenic	212	181	0,46
11-571-P2	PKD2	EX3-EX15	-	c.709-?_2907+?del	p.Leu237_Val968del	Absent	0	Definitely Pathogenic	-	-	-
09-393-P3	-	-	-	-	-	-		-	-	-	-

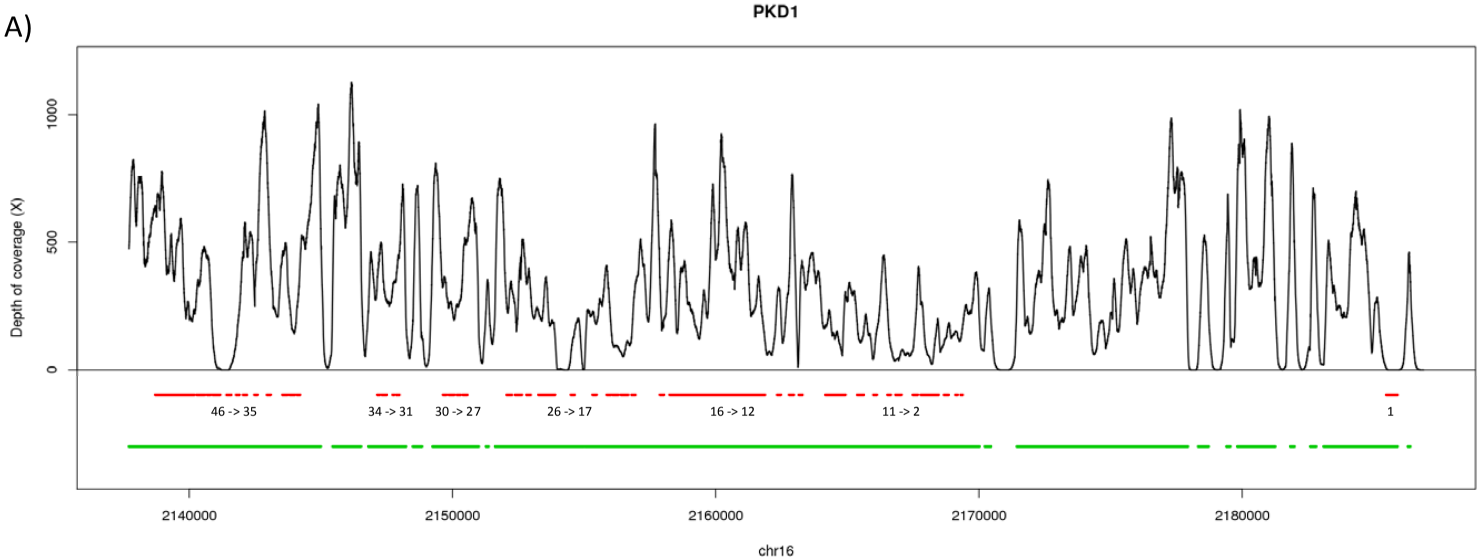
patients in previous studies

Table 5. ADPKD mutations in *PKD1* and *PKD2* identified in the 12 samples of the discovery cohort

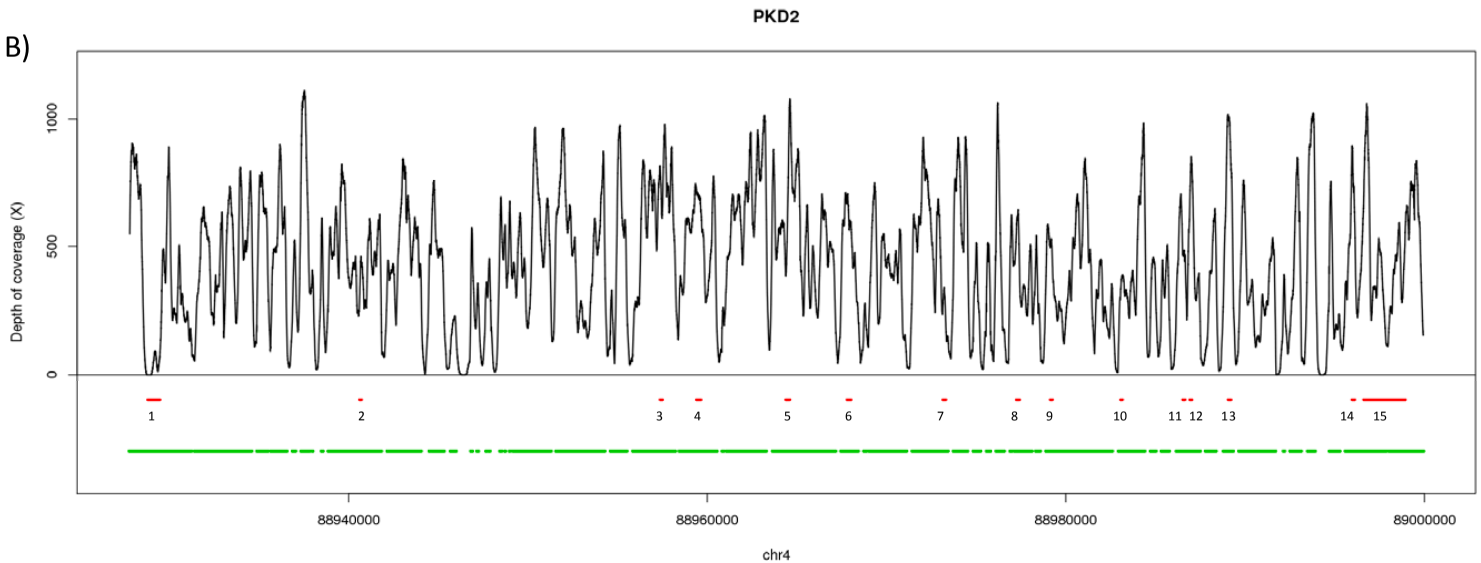
Sample	Gene	Exon/Intron	Duplicated Region	cDNA change	Protein change	PKDB	# patients	Classification	Ref counts	Var Counts	Var %
06-056	PKD1	EX03	Yes	c.348_352delTTTAA	p.Asn116fs*2	Present	1	Definitely Pathogenic	28	20	0,42
06-122	PKD1	EX17	Yes	c.7204 C>T	p.Arg2402*	Present	2	Definitely Pathogenic	24	12	0,33
07-032	PKD1	EX23	Yes	c.8421_8422insC	p.Ile2808fs	Absent	0	Definitely Pathogenic	22	18	0,45
11-444	PKD1	EX22	Yes	c.8041 C>T	p.Arg2681Cys	Absent	0	highly likely	38	20	0,34
12-444	PKD2	EX06	-	c.1532_1533insAT	p.Asp511Glufs*15	Absent	0	Definitely Pathogenic	156	70	0,31
	PKD1	EX37	-	c.10921 C>T	p.Arg3642Cys	Absent	0	highly likely	40	52	0,57
12-505	PKD1	EX15	Yes	c.50174_5015delAG	p.Arg1672Glyfs*98	Present	28	Definitely Pathogenic	118	88	0,43
13-199	PKD1	EX16	Yes	c.7039delC	p.Arg2347fs	Ausent	0	definitely pathogenic	34	32	0,48
12-628	PKD1	EX11	Yes	c.2180T>C	p.Leu727Pro	Ausent	0	highly likely	20	8	0,29
08-258	PKD1	EX21	Yes	c.7925 C>T	p.Arg2639*	Present	5	Definitely Pathogenic	28	14	0,33
10-484	PKD1	EX44	-	c.12010 C>T	p.Gln4004*	Present	4	Definitely Pathogenic	26	38	0,59
13-102	-	-	-	-	-	-	-	-	-	-	-
07-335	-	-	-	-	-	-	-	-	-	-	-

patients in previous studies

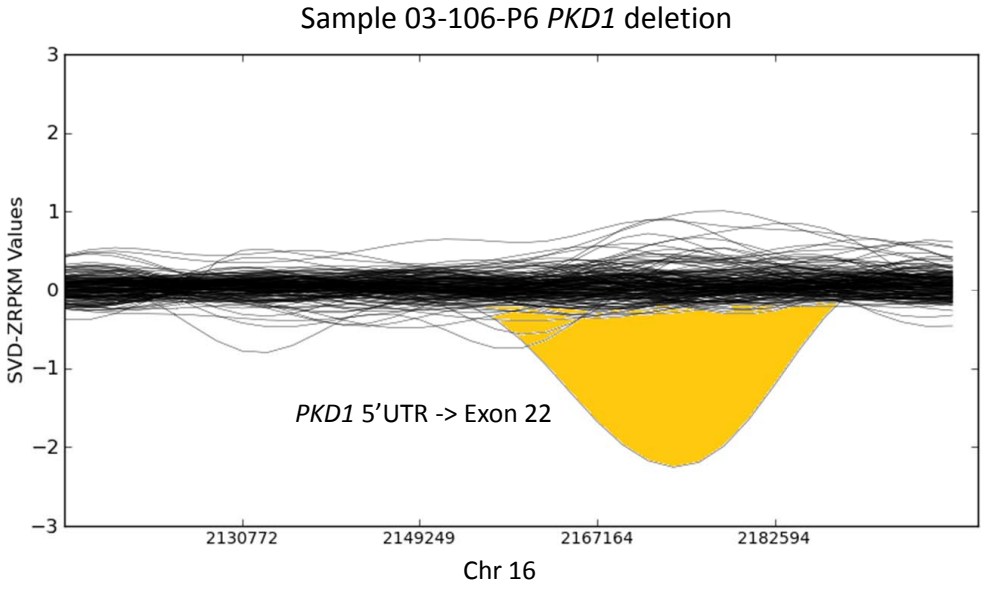
A)


PKD1
Depth of coverage (X)
0
500
1000
46 -> 35
34 -> 31
30 -> 27
26 -> 17
16 -> 12
11 -> 2
1
2140000
2150000
2160000
2170000
2180000
chr16

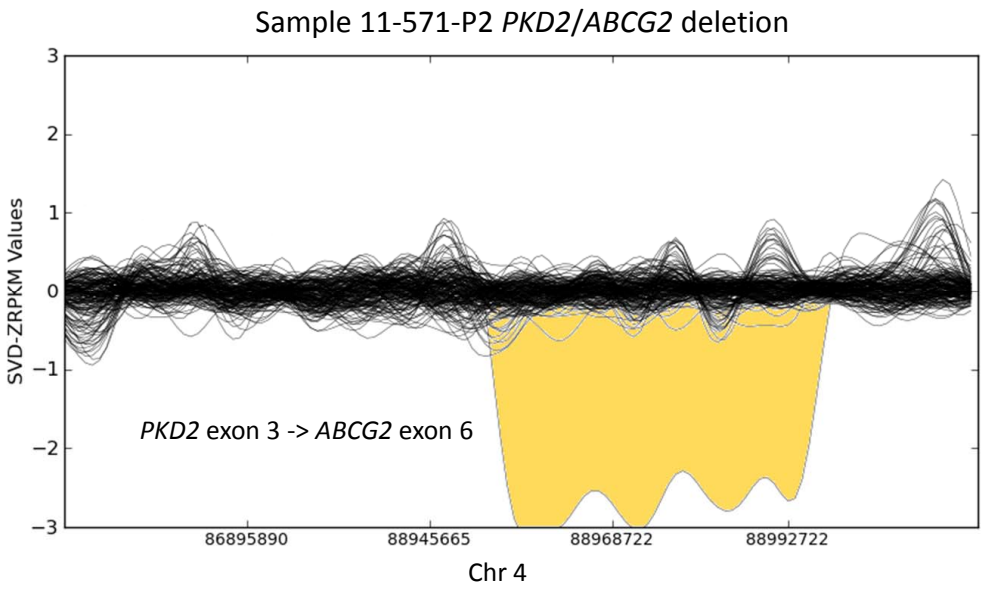
B)


PKD2
Depth of coverage (X)
0
500
1000
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
88940000
88960000
88980000
89000000
chr4

A)


Sample 03-106-P6 *PKD1* deletion
SVD-ZRPKM Values
3
2
1
0
−1
−2
−3
PKD1 5'UTR -> Exon 22
2130772
2149249
2167164
2182594
Chr 16

B)


Sample 11-571-P2 *PKD2*/*ABCG2* deletion
SVD-ZRPKM Values
3
2
1
0
−1
−2
−3
PKD2 exon 3 -> *ABCG2* exon 6
86895890
88945665
88968722
88992722
Chr 4

Supplementary Table 1. By sample sequencing of quality control and coverage statistics in the validation cohort

Sample	02-010-P6	03-106-P6	04-016-P6	07-172-P5	09-393-P3	09-403-P3	09-446-P3	10-182-P3	10-193-P3
QC-passed reads	21293178	14036390	12118708	13221808	14462320	13060902	14006548	12168782	15140412
mapped	21010161	13876408	11971701	13140256	14333995	12955439	13887613	12060599	15014787
paired in sequencing	21293178	14036390	12118708	13221808	14462320	13060902	14006548	12168782	15140412
read1	10646589	7018195	6059354	6610904	7231160	6530451	7003274	6084391	7570206
read2	10646589	7018195	6059354	6610904	7231160	6530451	7003274	6084391	7570206
properly paired	20604794	13643978	11813984	12962986	14108796	12740350	13677226	11861448	14795934
with itself and mate mapped	20957147	13846819	11946431	13124793	14310312	12936716	13866243	12041336	14992993
singletons	53014	29589	25270	15463	23683	18723	21370	19263	21794
with mate mapped to a different chr	94149	47481	37893	31415	121398	120616	102161	110222	110447
with mate mapped to a different chr (mapQ>=5)	69312	32998	26432	23266	107063	108135	88901	97585	96727
ALL Mean coverage (X)	688,56	449,54	398,59	442,37	503,35	459,49	485,71	421,25	523,95
% ALL target bases covered = 0X	0,09	0,09	0,07	0,09	0,11	0,10	0,10	0,11	0,09
% ALL target bases covered >= 1X	99,91	99,91	99,93	99,91	99,89	99,90	99,90	99,89	99,91
% ALL target bases covered >= 5X	99,70	99,83	99,82	99,83	99,80	99,79	99,81	99,77	99,81
% ALL target bases covered >= 10X	99,55	99,75	99,75	99,76	99,73	99,70	99,73	99,68	99,73
% ALL target bases covered >= 20X	99,35	99,57	99,58	99,58	99,58	99,55	99,60	99,52	99,60
% ALL target bases covered >= 50X	98,81	98,86	98,78	98,59	99,09	98,98	99,06	98,87	99,17
% ALL target bases covered >= 100X	97,41	96,44	96,30	94,56	97,92	97,59	97,61	96,88	98,04
PKD1 Mean coverage (X)	405,96	305,37	215,65	347,30	247,27	200,47	251,66	216,28	287,55
% PKD1 target bases covered = 0X	1,88	2,06	1,67	2,14	2,41	2,55	2,30	2,21	1,80
% PKD1 target bases covered >= 1X	98,12	97,94	98,33	97,86	97,59	97,45	97,70	97,79	98,20
% PKD1 target bases covered >= 5X	96,35	96,97	96,95	96,63	96,40	96,46	96,39	96,32	97,11
% PKD1 target bases covered >= 10X	94,96	96,37	96,55	96,33	96,09	96,06	96,17	95,83	96,27
% PKD1 target bases covered >= 20X	93,39	95,93	95,53	95,90	95,20	94,80	95,44	95,21	95,67
% PKD1 target bases covered >= 50X	89,34	93,05	89,58	93,35	90,32	88,08	89,98	88,22	92,17
% PKD1 target bases covered >= 100X	80,95	85,88	77,15	87,90	80,80	72,78	81,19	76,07	84,39
PKD2 Mean coverage (X)	710,70	434,75	400,98	440,23	521,94	478,37	522,50	453,31	533,95
% PKD2 target bases covered = 0X	0,34	0,41	0,30	0,46	0,44	0,17	0,42	0,52	0,47
% PKD2 target bases covered >= 1X	99,66	99,59	99,70	99,54	99,56	99,83	99,58	99,48	99,53
% PKD2 target bases covered >= 5X	98,93	99,49	99,54	99,46	99,35	99,37	99,43	99,32	99,37
% PKD2 target bases covered >= 10X	98,83	99,38	99,46	99,34	99,25	99,23	99,28	99,20	99,24
% PKD2 target bases covered >= 20X	98,75	99,19	99,21	99,16	99,04	99,01	99,08	98,96	99,06
% PKD2 target bases covered >= 50X	98,57	98,65	98,61	98,63	98,57	98,47	98,58	98,51	98,61
% PKD2 target bases covered >= 100X	98,05	96,76	96,93	96,57	98,08	97,96	98,06	97,79	98,11

10-326-P3	10-353-P3	10-388-P3	10-463-P3	11-008-P3	11-133-P8	11-168-P8	11-170-P2	11-220-P2	11-247-P7	11-256-P2	11-287-P2	11-457-P2
14585572	16602278	17485864	15893862	15302228	13615744	14239120	18607890	16291488	13873150	13845138	16555634	17067640
14449866	16441775	17332601	15731240	15161636	13425148	14039159	18514877	16217368	13685164	13783789	16470226	16996588
14585572	16602278	17485864	15893862	15302228	13615744	14239120	18607890	16291488	13873150	13845138	16555634	17067640
7292786	8301139	8742932	7946931	7651114	6807872	7119560	9303945	8145744	6936575	6922569	8277817	8533820
7292786	8301139	8742932	7946931	7651114	6807872	7119560	9303945	8145744	6936575	6922569	8277817	8533820
14233116	16153062	17076432	15455802	14946488	13139540	13710636	18305326	16051756	13492394	13675800	16318854	16810846
14423531	16412120	17303340	15698100	15136778	13393103	14004095	18498012	16203776	13653722	13772560	16454993	16983654
26335	29655	29261	33140	24858	32045	35064	16865	13592	31442	11229	15233	12934
119991	181038	148430	156428	117510	94687	98017	37918	28182	44270	22722	32889	32736
105134	163151	130639	138373	102580	75660	79025	28510	20783	30064	16785	24559	25221
506,12	573,88	610,18	548,79	523,77	380,55	401,10	608,47	543,60	453,83	463,39	539,71	568,01
0,07	0,07	0,07	0,07	0,10	0,09	0,09	0,07	0,06	0,09	0,06	0,06	0,07
99,93	99,93	99,93	99,93	99,90	99,91	99,91	99,93	99,94	99,91	99,94	99,94	99,93
99,82	99,85	99,86	99,82	99,82	99,79	99,81	99,87	99,85	99,80	99,85	99,85	99,84
99,75	99,79	99,81	99,76	99,74	99,70	99,72	99,82	99,80	99,73	99,78	99,80	99,78
99,60	99,68	99,69	99,61	99,62	99,47	99,52	99,71	99,67	99,52	99,66	99,65	99,65
99,12	99,24	99,31	99,16	99,14	98,47	98,63	99,19	99,18	98,61	99,00	99,00	99,06
97,85	97,85	98,18	98,06	97,80	94,83	95,39	97,22	97,73	94,89	96,52	96,44	96,95
280,20	352,97	355,63	272,55	308,04	264,13	270,10	504,89	352,64	301,71	320,95	433,01	436,07
1,72	1,80	1,67	2,05	2,40	1,66	2,24	1,57	1,72	1,98	1,80	1,51	1,87
98,28	98,20	98,33	97,95	97,60	98,34	97,76	98,43	98,28	98,02	98,20	98,49	98,13
96,58	96,90	97,36	96,64	96,78	96,64	96,40	97,36	97,36	96,48	97,16	97,41	97,29
96,25	96,58	96,78	96,23	96,35	96,19	96,16	96,92	96,95	96,21	96,48	96,96	96,68
95,65	96,03	96,17	95,35	95,83	95,69	95,69	96,38	96,11	95,70	96,07	96,35	96,39
91,08	93,61	93,73	90,66	93,16	91,44	91,43	95,45	93,97	93,04	93,32	95,26	95,13
83,24	87,00	86,97	82,11	86,08	82,73	83,13	92,05	86,65	83,95	86,21	90,54	90,37
513,64	602,14	634,34	570,76	526,49	365,61	391,12	591,33	529,96	455,95	472,29	525,37	553,62
0,33	0,32	0,35	0,23	0,23	0,43	0,47	0,41	0,05	0,47	0,24	0,42	0,42
99,67	99,68	99,65	99,77	99,77	99,57	99,53	99,59	99,95	99,53	99,76	99,58	99,58
99,54	99,53	99,52	99,42	99,36	99,43	99,47	99,50	99,52	99,33	99,53	99,50	99,48
99,48	99,44	99,50	99,29	99,26	99,31	99,36	99,46	99,44	99,25	99,48	99,46	99,42
99,38	99,35	99,32	99,09	99,09	99,09	99,14	99,39	99,20	99,11	99,36	99,38	99,27
98,78	98,81	98,85	98,71	98,58	98,47	98,54	99,05	98,79	98,52	98,70	98,92	98,92
98,00	98,17	98,14	98,22	97,92	95,30	95,87	98,14	97,82	96,87	97,55	97,74	97,94

11-468-P1	11-517-P1	11-525-P2	11-571-P2	11-595-P2	12-010-P1	12-144-P1	12-149-P1	12-161-P1	12-235-P1	12-331-P1	12-363-P1	12-366-P1
15189086	14465126	14294536	13152764	13454982	14653842	14365658	15559044	14096696	12236468	17056084	13428952	16833346
15059865	14353269	14234959	13096441	13395749	14531574	14251175	15436649	13992406	12139500	16927353	13324615	16726055
15189086	14465126	14294536	13152764	13454982	14653842	14365658	15559044	14096696	12236468	17056084	13428952	16833346
7594543	7232563	7147268	6576382	6727491	7326921	7182829	7779522	7048348	6118234	8528042	6714476	8416673
7594543	7232563	7147268	6576382	6727491	7326921	7182829	7779522	7048348	6118234	8528042	6714476	8416673
14893230	14212728	14142016	13004914	13308502	14411940	14138746	15300952	13888464	12040588	16737764	13172240	16552434
15038600	14334868	14223748	13086076	13384570	14511029	14232354	15415093	13975039	12123387	16907071	13308957	16709600
21265	18401	11211	10365	11179	20545	18821	21556	17367	16113	20282	15658	16455
39720	34004	31956	32140	31308	48647	41272	44773	39721	36225	49361	37571	48940
27775	24081	25759	26262	25450	37404	30557	32848	29677	27468	36800	28353	37709
526,67	520,90	479,81	439,38	448,37	509,43	501,90	540,98	496,27	430,95	600,22	484,62	605,50
0,09	0,08	0,06	0,06	0,07	0,08	0,07	0,07	0,09	0,09	0,07	0,07	0,08
99,91	99,92	99,94	99,94	99,93	99,92	99,93	99,93	99,91	99,91	99,93	99,93	99,92
99,85	99,86	99,85	99,86	99,85	99,84	99,87	99,87	99,84	99,82	99,87	99,87	99,86
99,81	99,81	99,80	99,80	99,79	99,79	99,81	99,82	99,79	99,77	99,83	99,82	99,81
99,69	99,68	99,69	99,66	99,64	99,68	99,68	99,72	99,69	99,65	99,71	99,72	99,74
99,14	99,28	99,14	98,98	98,95	99,21	99,24	99,28	99,29	99,15	99,16	99,18	99,39
96,94	97,83	97,20	96,43	96,03	97,36	97,70	97,59	97,89	97,06	96,45	97,19	98,21
408,08	340,41	355,05	323,83	323,52	368,33	372,32	414,74	344,53	330,35	619,12	372,10	517,96
2,40	1,97	1,69	1,61	1,95	1,99	1,77	2,02	2,75	2,42	1,90	1,90	2,03
97,60	98,03	98,31	98,39	98,05	98,01	98,23	97,98	97,25	97,58	98,10	98,10	97,97
97,12	97,38	97,06	97,06	96,75	96,80	97,35	97,20	96,46	96,29	97,31	97,37	97,40
96,95	96,74	96,70	96,86	96,46	96,37	96,94	96,59	96,33	96,04	96,96	96,84	96,61
96,39	96,16	96,29	96,13	96,00	96,06	96,39	96,22	96,00	95,68	96,37	96,09	96,32
94,86	94,12	94,64	94,31	93,76	94,47	94,31	95,23	94,39	94,08	95,68	94,42	95,58
89,62	86,27	88,19	86,69	86,65	88,44	89,19	89,67	87,78	86,81	94,03	88,34	92,88
520,01	541,05	451,99	286,33	444,58	510,86	475,83	534,19	472,89	404,55	545,76	452,26	554,99
0,41	0,14	0,35	0,35	0,42	0,43	0,39	0,42	0,41	0,49	0,46	0,32	0,40
99,59	99,86	99,65	99,65	99,58	99,57	99,61	99,58	99,59	99,51	99,54	99,68	99,60
99,46	99,46	99,51	99,52	99,45	99,46	99,52	99,52	99,44	99,37	99,49	99,53	99,50
99,42	99,41	99,44	99,47	99,38	99,37	99,47	99,47	99,38	99,30	99,48	99,49	99,47
99,27	99,25	99,39	99,31	99,22	99,25	99,35	99,36	99,22	99,18	99,39	99,40	99,44
98,88	98,92	98,92	97,48	98,66	98,99	99,01	99,07	98,92	98,83	99,09	99,06	99,16
97,68	98,22	97,36	89,29	97,53	98,07	98,04	98,37	97,95	97,45	98,36	97,91	98,61

12-382-P1	C-462-P4	C-586-P5	C-606-P8	C-616-P7	C-624-P6
12099566	12856374	12455250	14531168	13753210	6584983
12014816	12715977	12377395	14325234	13582529	6506999
12099566	12856374	12455250	14531168	13753210	6584983
6049783	6428187	6227625	7265584	6876605	3292492
6049783	6428187	6227625	7265584	6876605	3292491
11864234	12518678	12189256	13986206	13419388	6416009
12000565	12693283	12364383	14290403	13554306	6493941
14251	22694	13012	34831	28223	13058
32675	62835	27955	86399	38698	22478
24406	49138	20265	66145	25575	15829
437,23	418,79	417,25	404,81	451,81	210,87
0,08	0,07	0,06	0,07	0,07	0,22
99,92	99,93	99,94	99,93	99,93	99,78
99,81	99,83	99,84	99,82	99,83	99,29
99,76	99,75	99,75	99,71	99,76	98,94
99,67	99,57	99,55	99,48	99,57	98,27
99,12	98,77	98,42	98,27	98,57	94,08
96,75	95,92	93,66	93,48	94,56	80,63
347,74	286,75	319,62	300,89	353,05	119,69
1,78	1,62	1,52	1,62	1,68	2,85
98,22	98,38	98,48	98,38	98,32	97,15
96,53	96,98	96,98	96,59	96,69	93,09
96,24	96,40	96,64	96,38	96,55	89,45
95,89	95,86	95,96	95,66	96,21	83,36
93,94	92,31	93,78	92,76	94,19	69,33
87,19	84,17	86,33	85,02	87,41	50,79
401,25	404,88	409,08	402,13	445,23	215,64
0,50	0,35	0,13	0,30	0,45	0,48
99,50	99,65	99,87	99,70	99,55	99,52
99,37	99,54	99,45	99,37	99,50	98,76
99,29	99,43	99,38	99,27	99,48	98,59
99,16	99,19	99,26	99,14	99,22	98,32
98,84	98,48	98,69	98,46	98,68	97,08
97,53	96,25	95,73	95,56	96,73	86,56

Supplementary Table 2. By sample sequencing quality control and coverage statistics in the discovery cohort in sequencing of in PKD1 a

Sample	Average	Desvest	06-056	06-122	07-032	07-335	08-258	10-484	11-444	12-444	12-505	12-628	13-102	13-199
QC-passed reads	1444939,83	209921,69	1539934	1922710	1635464	1397528	1466252	1392912	1475046	1213858	1497954	1439158	1240018	1118444
duplicates	0,00	0,00	0	0	0	0	0	0	0	0	0	0	0	0
mapped	1139763,42	179226,45	1216361	1525006	1310916	1103131	1104980	1112500	1176885	934458	1202940	1176094	974870	839020
paired in sequencing	1444939,83	209921,69	1539934	1922710	1635464	1397528	1466252	1392912	1475046	1213858	1497954	1439158	1240018	1118444
read1	722469,92	104960,85	769967	961355	817732	698764	733126	696456	737523	606929	748977	719579	620009	559222
read2	722469,92	104960,85	769967	961355	817732	698764	733126	696456	737523	606929	748977	719579	620009	559222
properly paired	916518,83	158150,27	979398	1231384	1072404	890450	832856	906052	959096	739024	986356	983250	783644	634312
with itself and mate mapped	941485,67	162528,02	1006328	1265754	1102158	914554	857388	929926	984160	759766	1011264	1011154	803460	651916
singletons	198277,75	29119,39	210033	259252	208758	188577	247592	182574	192725	174692	191676	164940	171410	187104
with mate mapped to a different chr	280,33	58,63	282	344	410	288	262	264	278	196	282	306	262	190
with mate mapped to a different chr (mapQ>=5	193,58	47,74	191	242	296	196	182	181	202	138	203	210	179	103
ALL Mean coverage (X)	167,85	27,81	179,14	225,36	197,06	163,45	159,65	163,99	172,39	136,60	177,69	177,04	141,19	120,62
% ALL target bases covered = 0X	0,31	0,04	0,29	0,25	0,27	0,31	0,33	0,29	0,37	0,38	0,31	0,32	0,31	0,34
% ALL target bases covered >= 1X	99,84	0,02	99,85	99,87	99,86	99,84	99,84	99,86	99,82	99,81	99,84	99,84	99,85	99,83
% ALL target bases covered >= 5X	99,55	0,06	99,58	99,67	99,59	99,56	99,52	99,56	99,55	99,49	99,57	99,55	99,49	99,46
% ALL target bases covered >= 10X	99,28	0,13	99,38	99,51	99,38	99,32	99,24	99,30	99,29	99,14	99,34	99,29	99,13	99,05
% ALL target bases covered >= 20X	98,46	0,44	98,78	99,09	98,83	98,58	98,42	98,55	98,53	98,08	98,71	98,58	97,67	97,66
% ALL target bases covered >= 50X	91,25	3,83	93,80	96,36	94,91	92,20	91,56	92,15	90,95	87,69	93,32	92,80	86,36	82,88
% ALL target bases covered >= 100X	65,34	7,82	68,67	80,61	74,26	63,50	62,44	63,67	67,68	55,56	68,33	68,42	58,49	52,39
PKD1 Mean coverage (X)	80,60	13,60	82,56	112,37	87,07	78,88	66,32	78,80	88,91	70,86	89,52	81,34	70,66	59,95
% PKD1 target bases covered = 0X	3,70	0,39	4,18	3,44	3,62	3,85	3,21	2,85	3,75	4,17	3,89	4,02	3,62	3,85
% PKD1 target bases covered >= 1X	98,15	0,20	97,91	98,28	98,19	98,08	98,39	98,57	98,13	97,91	98,05	97,99	98,19	98,08
% PKD1 target bases covered >= 5X	95,86	0,62	95,99	97,24	96,50	95,66	95,16	95,81	96,34	95,36	96,09	95,20	95,67	95,30
% PKD1 target bases covered >= 10X	92,97	1,19	93,70	95,23	93,78	92,56	91,76	93,24	93,60	91,85	94,13	92,49	92,34	91,01
% PKD1 target bases covered >= 20X	86,69	2,59	87,44	91,34	88,42	86,87	83,75	87,15	88,40	84,25	88,74	87,47	84,12	82,30
% PKD1 target bases covered >= 50X	65,03	4,76	65,98	76,05	68,32	65,03	59,54	64,44	67,84	61,77	66,82	65,15	61,65	57,77
% PKD1 target bases covered >= 100X	52,01	1,65	51,80	56,48	52,20	51,55	50,90	51,47	52,93	50,81	53,14	51,61	51,21	50,08
PKD2 Mean coverage (X)	174,22	28,72	186,64	231,65	205,36	165,80	165,67	169,40	184,44	137,93	180,81	184,26	154,88	123,76
% PKD2 target bases covered = 0X	0,90	0,20	0,67	0,80	0,99	0,60	0,90	0,87	1,13	0,92	0,76	0,87	1,35	0,93
% PKD2 target bases covered >= 1X	99,55	0,10	99,67	99,60	99,50	99,70	99,55	99,57	99,43	99,54	99,62	99,56	99,33	99,54
% PKD2 target bases covered >= 5X	99,25	0,04	99,27	99,32	99,29	99,25	99,25	99,23	99,26	99,21	99,24	99,25	99,19	99,20
% PKD2 target bases covered >= 10X	99,17	0,06	99,22	99,24	99,22	99,21	99,12	99,18	99,18	99,11	99,17	99,20	99,09	99,04
% PKD2 target bases covered >= 20X	98,75	0,31	98,94	99,15	99,00	98,93	98,78	98,86	98,84	98,34	98,74	98,88	98,47	98,06
% PKD2 target bases covered >= 50X	92,74	3,46	94,71	97,19	96,41	92,45	92,39	92,86	94,96	88,44	94,22	93,61	90,97	84,67
% PKD2 target bases covered >= 100X	67,51	8,70	71,92	82,97	77,19	65,09	65,90	65,80	70,79	55,37	70,38	71,42	61,74	51,55

Supplementary Table 3. NGS of *PKD1* and *PKD2* vs genotyping calls in the five control samples

Sample	ALL						*PKD1*						*PKD2*					
	Sum	**462**	**586**	**606**	**616**	**624**	**Sum**	**462**	**586**	**606**	**616**	**624**	**Sum**	**462**	**586**	**606**	**616**	**624**
Total genotyped positions	12108	2419	2422	2422	2423	2422	80	16	16	16	16	16	269	54	53	54	54	54
Total genotyped SNPs	3511	666	706	680	762	697	20	4	6	5	0	5	103	14	24	14	27	24
Total NGS SNPs	3348	634	674	650	726	664	20	4	6	5	0	5	103	14	24	14	27	24
NGS TP	3344	633	673	649	725	664	20	4	6	5	0	5	103	14	24	14	27	24
NGS FP	4	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
NGS TN	8593	1752	1715	1741	1660	1725	60	12	10	11	16	11	166	40	29	40	27	30
NGS FN	167	33	33	31	37	33	0	0	0	0	0	0	0	0	0	0	0	0
NGS PPV	0,999	0,998	0,999	0,998	0,999	1	1	1	1	1	-	1	1	1	1	1	1	1
NGS Sensitivity	0,952	0,950	0,953	0,954	0,951	0,953	1	1	1	1	-	0	1	1	1	1	1	1
Genotyping Het	2109	412	414	375	459	449	20	4	6	5	-	5	64	5	19	14	11	15
Het TP	2048	404	400	364	442	438	20	4	6	5	-	5	64	5	19	14	11	15
Het TP (correct allele)	2032	404	395	359	437	437	15	4	6	5	-	0	64	5	19	14	11	15
Het TP (wrong allele)	16	0	5	5	5	1	0	0	0	0	-	0	0	0	0	0	0	0
Het FP	3	1	1	0	1	0	0	0	0	0	-	0	0	0	0	0	0	0
Het FN	61	8	14	11	17	11	0	0	0	0	-	0	0	0	0	0	0	0
Het PPV	0,999	0,998	0,998	1	0,998	1	1	1	1	1	-	1	1	1	1	1	1	1
Het Sensitivity	0,971	0,981	0,966	0,971	0,963	0,976	1	1	1	1	-	1	1	1	1	1	1	1
Genotyping Hom	1402	254	292	305	303	248	-	-	-	-	-	-	39	9	5	-	16	9
Hom TP	1296	229	273	285	283	226	-	-	-	-	-	-	39	9	5	-	16	9
Hom TP (correct allele)	1281	225	270	283	280	223	-	-	-	-	-	-	39	9	5	-	16	9
Hom TP (wrong allele)	15	4	3	2	3	3	-	-	-	-	-	-	0	0	0	-	0	0
Hom FP	0	0	0	0	0	0	-	-	-	-	-	-	0	0	0	-	0	0
Hom FN	106	25	19	20	20	22	-	-	-	-	-	-	0	0	0	-	0	0
Hom PPV	1	1	1	1	1	1	-	-	-	-	-	-	1	1	1	-	1	1
Hom Sensitivity	0,924	0,902	0,935	0,934	0,934	0,911	-	-	-	-	-	-	1	1	1	-	1	1

4. DISCUSSION

In a recent deliberately provocative publication, it is wondered whether genomics is ready for a plateau in sequencing costs.[97] During the second semester of last year we observed for the first time since the National Human Genome Research Institute (NHGRI) began to record the cost of sequencing, that the cost of sequencing a human genome increased a 12% (an increase of $717). Although the following months were accompanied by modest decreases, if we compare the cost of sequencing a human genome between April 2012 ($5.901) and April 2013 ($5.826), which is the latest record available (Figure 9), we observe that during the last year it has only decreased a 1.3% ($75).[98]

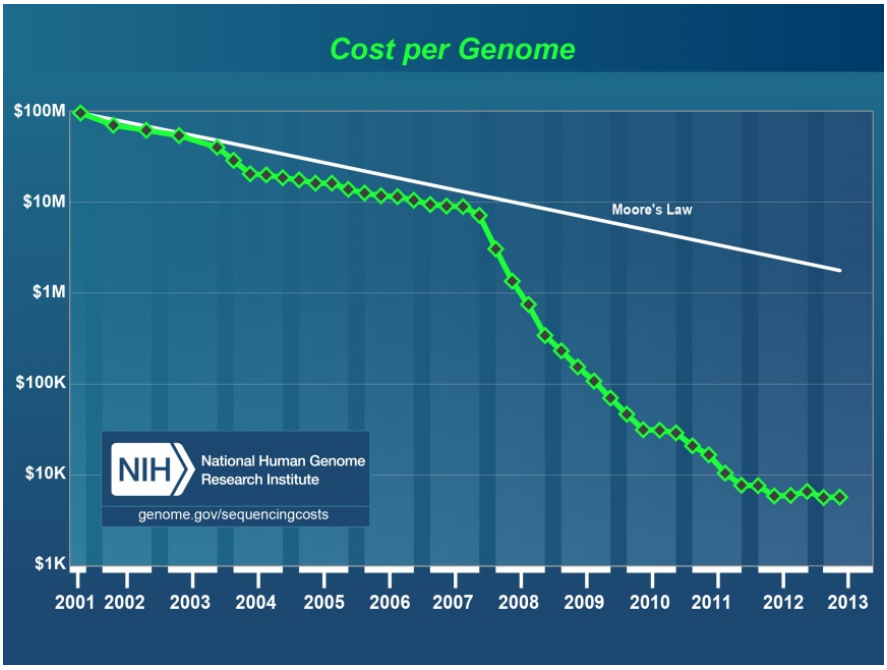


Figure 9. Evolution of the cost of sequencing a human genome. From Wetterstrand, 2013.[98]

These numbers challenge a trend that has been maintained during years beating Moore's law, and indicates that in the medium term we should not expect sequencing to be significantly cheaper until the next technological revolution arrives. This can have profound consequences in the genomics field, which has heavily relied during the last years in the fact that each sequencing run was cheaper than the last. Most of the recent big genomic achievements have been based in brute force experiments, made possible by the rapid technological advances. The present change of cycle would require more ingenious and elegant ideas to keep publishing interesting findings after the genomics boom of the last decade. So, it is time to fully exploit the potential of the current sequencing technologies, which seem that will be static for a while. Moreover, it is also time to tone down the rhetoric around the advent of $10 human genomes and to start working seriously on the clinical translation of our research based on the technology that we currently have, no what we expect to have tomorrow.[97] Thus, the main goal of this thesis was to prove the feasibility of targeted NGS in routine clinical diagnostics for different clinical entities.

Current NGS technologies have the potential to address the mismatch between promises and achievements in medical genetics, still present more than ten years since the human genome project was drafted. Shortly before the completion of the first human genome, Francis Collins, one of the leaders of the Human Genome Project predicted that by 2010 the genetic causes of most Mendelian diseases would have been unveiled and therapies would be available for most of them, that disease gene associations for most of the

common disorders would have been established, and that personalized preventive medicine would be a reality.[99] Although some of his claims have come true,[100] most of the post-genomic era promises are yet to be accomplished.

Genomics as an enhanced approach to healthcare has the potential to transform the quality of life worldwide, allowing the widespread implementation of more tailored medical care based on individual risk. It seems quite likely that whole human genome sequencing (and eventually, proteome, metabolome, microbiome, etc.) would be a routine component of everyone health record available to both patients and physicians for predictive and preventive healthcare purposes. This is poised to have a transforming effect in clinical practice, including diagnosis and decision-making for appropriate therapeutic procedures. The personal genome era is expected to be realized in the near future, and the genetics community and society in general needs to prepare for this new era.[25]

4.1. Identification of novel Mendelian disease genes by NGS

During the last decades positional cloning strategies have led to the discovery of several new insights in human genetics. However, when this approach is used for rare Mendelian disorders the results often are not conclusive due to the lack of complete pedigrees, the unavailability of large family collections and marked locus heterogeneity. Thus, in small and non-consanguineous families, neither linkage analysis nor homozygosity mapping are likely to

succeed in identifying the responsible gene of a given rare Mendelian disease.[101] As a consequence of these limitations, the underlying genetic cause of more than three thousand disorders of Mendelian inheritance still remain to be determined (http://www.ncbi.nlm.nih.gov/Omim/mimstats.html).

In that respect, advances in NGS technologies, especially exome sequencing, represent an important milestone in genomics, providing an effective alternative for the discovery of candidate genes and mutations that underlie Mendelian disorders, that have been resistant to conventional approaches, thanks to an unprecedented ability to identify rare variants. NGS technologies have been a much awaited step forward from linkage mapping for Mendelian disease gene discovery, since it has enabled mapping of genes for monogenic traits in families with small pedigrees and even in as few samples as two unrelated individuals.[63] Also, NGS technologies do not have the typical microarray-based methods restrictions, such as single nucleotide mismatches and melting temperature issues, and provide a more comprehensive and complete picture of human variability. In fact, NGS enables genotyping of all possible variant sites compared to array based genotyping methods, which use a set of predetermined common variants.

The first proof-of-principle report published back in 2009 using exome sequencing to identify Mendelian disease genes supposed a revolution in human genetics, which has led to the discovery of the genetic basis of several Mendelian disorders using similar

approaches. We can expect that during the coming years, most of the yet unresolved Mendelian disorders will be genetically characterized using exome sequencing, but at some point the transition towards more unbiased WGS will occur.

In this thesis it is reported a novel gene for familial hyperkalemic hypertension discovered using exome sequencing in combination with linkage analysis and functional studies. Familial hyperkalemic hypertension is a rare autosomal dominant form of arterial hypertension, associated to metabolic abnormalities involving hyperkalemia, metabolic acidosis and hyperchloremia.[102] Back in 2001, two genes were identified responsible for this disease, *WNK1* and *WNK4*, which encode two members of the with-no-lysine (WNK) serine threonine kinases family.[103] Functional studies demonstrated the regulatory functions of WNK1 and WNK4 on several ion transporters, especially on the Na^+-Cl^- cotransporter (NCC), which shows increased expression and activity in affected individuals.[104] However, mutations found in the *WNK1* and *WNK4* genes explain only a minority of the cases described. In the study included in this thesis, the search for a novel gene was done on a shared linkage interval on chromosome 5 in two affected families who did not have mutations in *WNK1* and *WNK4* genes.

Using exome sequencing and bioinformatics analysis one gene, *KLHL3*, was found in both families to be affected by *in silico* predicted pathogenic missense mutations. The involvement of *KLHL3* in the disease was quickly verified by mutation-pathology cosegregation in the initial two families and the identification of 14

additional missense mutations in 17 additional unrelated affected subjects. In 13 cases, the mutation was present in heterozygous state, in agreement with the expected dominant transmission. In four cases from consanguineous families, the mutation was present in homozygous state, corresponding to an apparently recessive form. The strong expression of *KLHL3* in the nephron and its cellular colocalization with NCC in the distal convoluted tubule strengthened the arguments of causality of this gene in the disease.

At the same time, an American team identified independently by exome sequencing the same gene, *KLHL3,* as responsible for the disease.[105] They reported 24 different mutations, mostly missense, but also nonsense, small deletions and splicing mutations. Sixteen mutations were heterozygous and eight homozygous or compound heterozygous, thus explaining the sometimes dominant or recessive presentation of the disease. In addition, the American team showed the presence of causal mutations in *Cullin3* in other 17 affected individuals. These mutations, most of them *de novo,* were detected in the more severe forms of the disease.

KLHL3 is a BTB-BACK-Kelch family protein that serves as a substrate adapter in Cullin3 E3 ubiquitin ligase complexes, which direct target proteins to degradation by the proteasome after ubiquitination.[106] Recently, some groups reported that an interaction of KLHL3 with CULLIN3 and WNK4 induced WNK4 ubiquitination and reduced the WNK4 protein levels, while a reduction in the interaction between KLHL3 and WNK4 attenuated the ubiquitination of WNK4, resulting in an increased level of the

WNK4 protein and leading to an elevated activation of the WNK-OSR1/SPAK-NCC signal cascade that cause familial hyperkalemic hypertension.[107-109] All together, our initial genetic findings have led to remarkable progress in understanding the molecular mechanisms regulating blood pressure.

Nowadays the sequencing of non-trivial numbers of complete human genomes is a reality, but as demonstrated in this thesis it is often more convenient to focus on informative regions of the genome, such as the exome. Targeted sequencing deepens information content and minimizes reagent costs of NGS technologies, allowing more samples to be analyzed in parallel given a set amount of sequencing capacity. Exome sequencing represents the most cost-effective alternative to WGS for the discovery of highly penetrant rare variants because it supposes a drastic reduction in the sequencing required. In fact, respect to WGS, exome sequencing requires about 20-fold less (~5%) sequencing to achieve the same depth of coverage, which is translated into considerably less raw sequence and lower costs.[110] Despite the inherent costs of genomic capture in addition to sequencing, according to the list prices, the all-in cost of exome sequencing is roughly 10- to 20-fold less than for the whole genome.

If the target genomic region is smaller than the exome, as in the case of diagnosis assays for known disease loci, cost reduction is even greater. This has a direct impact on the power to detect causative variants in projects where the number of samples required

to get meaningful statistical power makes WGS prohibitively costly.[50] Also, exome sequencing requires less onerous analyses than WGS, and the number of variants detected is up to two orders of magnitude lower as a consequence of only retrieving variants affecting the coding regions of the genome. This reduces data fatigue and simplifies the analyses for the identification of disease-causative variants.

A key point when designing an exome sequencing study is to choose the right amount of sequencing, i.e. the coverage depth, in order to achieve a good sensitivity to detect sequence variants. Considering the unavoidable enrichment uniformity differences between the different regions of the genome captured using hybridization-based approaches, it has been proposed that the safest strategy is to intend an average coverage ≥20X. This average coverage depth ensures a good sensitivity able to detect ≥95% of homozygous and heterozygous variants.[61] This guarantees sufficient allele sampling, as well as prevents sequencing errors from appearing to be actual variants. Uniformity also affects exome sequencing efficiency. Since not all the targeted bases are read at the same rate, different genomic region are covered at very different depths. This is not a problem for those regions that are oversampled and have high coverage depths, but supposes an increment in the amount of sequencing needed to collect the missing or low-covered sequences to bring the underrepresented sequences up.

In addition to the technical limitations inherent to hybrid capture, such as selection bias and uneven capture efficiency, the main

limitation of the targeted resequencing approach is the impossibility to efficiently capture and sequence the repetitive and low-complexity, and GC-rich genomic sequences that are refractory to enrichment. However, the constant optimization of the capture and NGS chemistries will gradually close the capture gaps (mainly due to uniqueness constraints, homopolymer runs, ambiguous bases or other factors that are known to cause issues in either oligonucleotide synthesis or hybridization), and reduce enrichment variability between samples and targets.

Exome sequencing has proven its reliability for the identification genetic variability underlying relatively simple, single-gene disorders. However, the step from rare monogenic and simple Mendelian disorders to more-complex multigenic disorders is going to be a challenging move. Exome sequencing studies done so far have to be considered as a starting point in the effort to apply these technologies to multigenic diseases. The extent of heterogeneity associated with common complex disorders will have to be mitigated with larger sample sizes and more sophisticated weighting of non-synonymous variants by predicted functional impact.[110]

Ultimately, targeted sequencing will continue to be a cost-effective approach as long as the cost of genomic capture does not dominate[50]. Therefore, within the next years targeted sequencing will be used alongside of WGS for different research and clinical applications.

4.2. Diagnostics of Mendelian disorders using targeted NGS

It has been suggested that the impact that NGS technologies will have on clinical genetics during the upcoming years will be comparable to the introduction of X-rays to medicine many decades ago.[111] After the tremendous impact of NGS technologies to the discovery of disease-causing genes during the last four years, we are now witnessing the introduction of these technologies for diagnostic applications. The aim is to rapidly revolutionize the field of genetic diagnostics, making it much more cost- and time-effective, but also advance in accuracy. However, since the first published report on massive parallel sequencing back in 2005,[31] few examples have been published on the use of NGS for routine genetic diagnostics.[112-119]

In this thesis it is reported the successful application of targeted NGS for the diagnostics of three different common Mendelian disorders, namely cystic fibrosis, hyperphenylalaninemias and autosomal dominant polycystic kidney disease. Here, it is demonstrated that NGS technologies, in combination with robust bioinformatics tools, can simultaneously detect all sorts of sequence variants, including complex genomic rearrangements, in multiple samples in just one experiment even in complex duplicated genomic regions. Actually, the assays reported here have been optimized to offer conclusive diagnostics five days after receiving the DNA samples, at a fraction of the cost of their traditional counterpart methods. Since the genetic screening of mutations for these three

diseases is already an integral part of routine clinical practice, the increased speed and efficiency offered by targeted NGS may enable the widespread application of the tests reported here, which should be easily adopted by routine molecular diagnostics laboratories.

Currently, our ability to discover genetic variation in a patient genome is running far ahead of our ability to interpret that variation. The success of NGS for medical genetics hinges on the accuracy in distinguishing causal from benign alleles, which is the key challenge for interpreting DNA sequence data for diagnostics. Over the last three decades, PCR amplification of target regions followed by Sanger sequencing has been the gold standard for the identification of clinically relevant mutations in the terms of routine diagnostics. It offers great accuracy, at the expense of being laborious and costly, especially when it comes to the analysis of disorders of heterogeneous etiology for which multiple targets might be tested in a stepwise fashion. Such disorders may require extensive screening of several genes, using different molecular approaches for every type of sequence variant being tested.

However, this rather costly, stepwise, and time-consuming technology will be gradually replaced by NGS technologies, which offer higher throughput and scalability and, as a corollary, have reduced costs per sequenced nucleotide and shorter turnaround time. In fact, interrogating a short list of candidate genes by Sanger sequencing will rapidly lead to spend thousands of euros. Given the current cost of targeted NGS of small genomic regions, it is inciting

to use NGS-based approaches to screen these genes for diagnostics purposes.

Although WGS is becoming part of the clinical practice for some specific medical problems,[120] until it can offer at an affordable price the sensitivity required for routine diagnostics purposes, where depth is preferred instead of width in terms of sequence coverage, targeted NGS will be the preferred method in diagnostic laboratories, since it exploits the full potential of the NGS devices to process several samples and loci in parallel. Targeted NGS by genome partitioning and capture methods offers significant advantages both in terms of cost and effectiveness for clinical diagnostics applications, overcoming the limitations of WGS and exome sequencing, which mainly are the cost and additional bioinformatics burden for the first, and the uneven distribution of depth of coverage of the later. Other advantages of targeted NGS vs. WGS and exome sequencing are scalability (several samples can be multiplexed both at the capture and sequencing steps), cost, and clinical validity (analysis restricted to genomic regions of known clinical significance, which prevents findings of uncertain relevance in other loci of the genome and that might raise ethical and legal concerns).

The transition over the next years of NGS technologies from basic research to the routine detection of mutations in genetic loci with well documented diagnostic value will take advantage not only of the new benchtop NGS platforms, such as the Illumina's Miseq, which can be much more easily incorporated in the daily clinical

practice, but also of automated workflows and simplified bioinformatics analyses able to generate medical report-like outputs adapted to clinical laboratories. However, the correct interpretation, storage, and dissemination of the large amount of the datasets generated remain a major challenge on the path of NGS to medical applications.[121] These challenges could be addressed with extensive exchange of data, information and knowledge between medical scientists, sequencing centers, bioinformatics networks and industry. Some genomic centers working in biomedicine have developed collaborative initiatives aiming at bringing everyone together to harmonize genomic medical research, set up standards in medical sequencing and review the current diagnostic standards according to the new insights gained from genomic and phenotypic data integration.

An example of such initiatives is GEUVADIS (Genetic European Variation in Disease) Consortium (www.GEUVADIS.org), which aims at harmonizing medical sequencing efforts across Europe by developing guidelines for the biological and medical interpretation of sequence data for monogenic (and complex) disorders, as well as to set up standards for the ethics of phenotype prediction from sequence variation.

4.3. Concluding remarks

Genomics is making faster progress than any other area of biomedical research. Especially, the advances in the field of NGS development and applications, make this an exciting time for the

study of how genetic variation affects health and disease. The ultimate game changer in clinical genetics will be the routine sequencing of individual genomes, but until this becomes feasible, targeted approaches are the more convenient interim solution. The standardization and further development of the methods used in this thesis will provide powerful and cost-effective techniques for the identification of causative variants of heritable disorders caused by known and unknown genes. The results presented in this thesis give evidence of the reliability and clinically accepted performance of NGS technologies, and represent another step towards the translation of NGS technologies and bioinformatics to medical genetics and diagnostic applications.

5. CONCLUSIONS

Exome sequencing, in combination with linkage analysis and functional studies, identified *KLHL3* as a major gene for familial hyperkalemic hypertension.

- *KLHL3* is coexpressed with the Na^+ Cl^- cotransporter (NCC) and downregulates NCC expression at the cell surface.
- *KLHL3* has a key role in the complex signaling pathway that regulates ion homeostasis in the distal nephron and blood pressure.

Targeted resequencing is an accurate and cost-effective tool for the diagnostics of monogenic and heterogeneous Mendelian disorders, able to detect all sorts of sequence variants and ready to substitute traditional molecular approaches.

- It provides great accuracy for the screening of *CFTR* in cystic fibrosis and *CFTR*-related disorders patients and carriers.
- It allows an efficient differential genetic diagnostics of phenylketonuria and tetrahydrobiopterin deficient hyperphenylalaninemia.
- It permits a comprehensive screening *PKD1* and *PKD2* in autosomal dominant polycystic kidney disease patients, even in the complex repeated region of *PKD1*.

6. BIBLIOGRAPHY

1. Lander ES, Linton LM, Birren B *et al*: Initial sequencing and analysis of the human genome. *Nature* 2001; 409: 860-921.

2. Venter JC, Adams MD, Myers EW *et al*: The sequence of the human genome. *Science* 2001; 291: 1304-1351.

3. Frazer KA, Ballinger DG, Cox DR *et al*: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; 449: 851-861.

4. Abecasis GR, Auton A, Brooks LD *et al*: An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; 491: 56-65.

5. Hudson TJ, Anderson W, Artez A *et al*: International network of cancer genome projects. *Nature* 2010; 464: 993-998.

6. de Ligt J, Veltman JA, Vissers LE: Point mutations as a source of de novo genetic disease. *Curr Opin Genet Dev* 2013; 23: 257-263.

7. MacArthur DG, Balasubramanian S, Frankish A *et al*: A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 2012; 335: 823-828.

8. Frazer KA, Murray SS, Schork NJ, Topol EJ: Human genetic variation and its contribution to complex traits. *Nat Rev Genet* 2009; 10: 241-251.

9. Kruglyak L, Nickerson DA: Variation is the spice of life. *Nat Genet* 2001; 27: 234-236.

10. Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA: Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum Mol Genet* 2005; 14: 59-69.

11. Harismendy O, Ng PC, Strausberg RL *et al*: Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* 2009; 10: R32.

12. Lupski JR: Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* 1998; 14: 417-422.

13. Tabor HK, Risch NJ, Myers RM: Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet* 2002; 3: 391-397.

14. Arts HH, Doherty D, van Beersum SE *et al*: Mutations in the gene encoding the basal body protein RPGRIP1L, a nephrocystin-4 interactor, cause Joubert syndrome. *Nat Genet* 2007; 39: 882-888.

15. Dawn Teare M, Barrett JH: Genetic linkage studies. *Lancet* 2005; 366: 1036-1044.

16. Kerem B, Rommens JM, Buchanan JA *et al*: Identification of the cystic fibrosis gene: genetic analysis. *Science* 1989; 245: 1073-1080.

17. Riordan JR, Rommens JM, Kerem B *et al*: Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* 1989; 245: 1066-1073.

18. Rommens JM, Iannuzzi MC, Kerem B *et al*: Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* 1989; 245: 1059-1065.

19. Wang S, Haynes C, Barany F, Ott J: Genome-wide autozygosity mapping in human populations. *Genet Epidemiol* 2009; 33: 172-180.

20. Hardy J, Singleton A: Genomewide association studies and human disease. *N Engl J Med* 2009; 360: 1759-1768.

21. Hindorff LA, Sethupathy P, Junkins HA *et al*: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 2009; 106: 9362-9367.

22. Hindorff LA, MacArthur J, Morales J, Junkins HA, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed September 2013.

23. Reich DE, Lander ES: On the allelic spectrum of human disease. *Trends Genet* 2001; 17: 502-510.

24. Manolio TA, Collins FS, Cox NJ *et al*: Finding the missing heritability of complex diseases. *Nature* 2009; 461: 747-753.

25. Tsuji S: Genetics of neurodegenerative diseases: insights from high-throughput resequencing. *Hum Mol Genet* 2010; 19: R65-70.

26. Robinson PN, Krawitz P, Mundlos S: Strategies for exome and genome sequence data analysis in disease-gene discovery projects. *Clin Genet* 2011; 80: 127-132.

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

28. Tucker T, Marra M, Friedman JM: Massively parallel sequencing: the next big thing in genetic medicine. *Am J Hum Genet* 2009; 85: 142-154.

29. Metzker ML: Emerging technologies in DNA sequencing. *Genome Res* 2005; 15: 1767-1776.

30. Shendure J, Ji H: Next-generation DNA sequencing. *Nat Biotechnol* 2008; 26: 1135-1145.

31. Margulies M, Egholm M, Altman WE *et al*: Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005; 437: 376-380.

32. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B: Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci U S A* 2003; 100: 8817-8822.

33. Shendure J, Porreca GJ, Reppas NB *et al*: Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 2005; 309: 1728-1732.

34. Bentley DR, Balasubramanian S, Swerdlow HP *et al*: Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008; 456: 53-59.

35. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G: BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res* 2006; 34: e22.

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

37. Mardis ER: Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 2008; 9: 387-402.

38. Fullwood MJ, Wei CL, Liu ET, Ruan Y: Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res* 2009; 19: 521-532.

39. Pettersson E, Lundeberg J, Ahmadian A: Generations of sequencing technologies. *Genomics* 2009; 93: 105-111.

40. Stratton MR, Campbell PJ, Futreal PA: The cancer genome. *Nature* 2009; 458: 719-724.

41. Wheeler DA, Srinivasan M, Egholm M *et al*: The complete genome of an individual by massively parallel DNA sequencing. *Nature* 2008; 452: 872-876.

42. Wang J, Wang W, Li R *et al*: The diploid genome sequence of an Asian individual. *Nature* 2008; 456: 60-65.

43. Kim JI, Ju YS, Park H *et al*: A highly annotated whole-genome sequence of a Korean individual. *Nature* 2009; 460: 1011-1015.

44. Ahn SM, Kim TH, Lee S *et al*: The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* 2009; 19: 1622-1629.

45. McKernan KJ, Peckham HE, Costa GL *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: 1527-1541.

46. Pushkarev D, Neff NF, Quake SR: Single-molecule sequencing of an individual human genome. *Nat Biotechnol* 2009; 27: 847-850.

47. Johnson DS, Mortazavi A, Myers RM, Wold B: Genome-wide mapping of in vivo protein-DNA interactions. *Science* 2007; 316: 1497-1502.

48. Bhinge AA, Kim J, Euskirchen GM, Snyder M, Iyer VR: Mapping the chromosomal targets of STAT1 by Sequence Tag Analysis of Genomic Enrichment (STAGE). *Genome Res* 2007; 17: 910-916.

49. Morozova O, Hirst M, Marra MA: Applications of new sequencing technologies for transcriptome analysis. *Annu Rev Genomics Hum Genet* 2009; 10: 135-151.

50. Fisher S, Barry A, Abreu J *et al*: A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol* 2011; 12: R1.

51. Igartua C, Turner EH, Ng SB *et al*: Targeted enrichment of specific regions in the human genome by array hybridization. *Curr Protoc Hum Genet* 2010; Chapter 18: Unit 18 13.

52. Meuzelaar LS, Lancaster O, Pasche JP, Kopal G, Brookes AJ: MegaPlex PCR: a strategy for multiplex amplification. *Nat Methods* 2007; 4: 835-837.

53. Fredriksson S, Baner J, Dahl F *et al*: Multiplex amplification of all coding sequences within 10 cancer genes by Gene-Collector. *Nucleic Acids Res* 2007; 35: e47.

54. Dahl F, Gullberg M, Stenberg J, Landegren U, Nilsson M: Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res* 2005; 33: e71.

55. Albert TJ, Molla MN, Muzny DM *et al*: Direct selection of human genomic loci by microarray hybridization. *Nat Methods* 2007; 4: 903-905.

56. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME: Microarray-based genomic selection for high-throughput resequencing. *Nat Methods* 2007; 4: 907-909.

57. Hodges E, Rooks M, Xuan Z *et al*: Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc* 2009; 4: 960-974.

58. Gnirke A, Melnikov A, Maguire J *et al*: Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 2009; 27: 182-189.

59. Ng SB, Buckingham KJ, Lee C *et al*: Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 2010; 42: 30-35.

60. Mochizuki T, Wu G, Hayashi T *et al*: PKD2, a gene for polycystic kidney disease that encodes an integral membrane protein. *Science* 1996; 272: 1339-1342.

61. Choi M, Scholl UI, Ji W *et al*: Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* 2009; 106: 19096-19101.

62. Krawitz PM, Schweiger MR, Rodelsperger C *et al*: Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nat Genet* 2010; 42: 827-829.

63. Lalonde E, Albrecht S, Ha KC *et al*: Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. *Hum Mutat* 2010; 31: 918-923.

64. Bilguvar K, Ozturk AK, Louvi A *et al*: Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* 2010; 467(7312):207-10

65. Ng SB, Bigham AW, Buckingham KJ *et al*: Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* 2010; 42: 790-793.

66. Wang JL, Yang X, Xia K *et al*: TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. *Brain* 2010; 133: 3510-3518.

67. Johnson JO, Mandrioli J, Benatar M *et al*: Exome sequencing reveals VCP mutations as a cause of familial ALS. *Neuron* 2010; 68: 857-864.

68. Shoubridge C, Tarpey PS, Abidi F *et al*: Mutations in the guanine nucleotide exchange factor gene IQSEC2 cause nonsyndromic intellectual disability. *Nat Genet* 2010; 42: 486-488.

69. Yi X, Liang Y, Huerta-Sanchez E *et al*: Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 2010; 329: 75-78.

70. Lin Y, Li J, Shen H, Zhang L, Papasian CJ, Deng HW: Comparative studies of de novo assembly tools for next-

generation sequencing technologies. *Bioinformatics* 2011; 27: 2031-2037.

71. Miller JR, Koren S, Sutton G: Assembly algorithms for next-generation sequencing data. *Genomics* 2010; 95: 315-327.

72. Schatz MC, Delcher AL, Salzberg SL: Assembly of large genomes using second-generation sequencing. *Genome Res* 2010; 20: 1165-1173.

73. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; 25: 1754-1760.

74. Ledergerber C, Dessimoz C: Base-calling for next-generation sequencing platforms. *Brief Bioinform* 2011; 12: 489-497.

75. McKenna A, Hanna M, Banks E *et al*: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; 20: 1297-1303.

76. Li H, Handsaker B, Wysoker A *et al*: The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; 25: 2078-2079.

77. Escaramis G, Tornador C, Bassaganyas L *et al*: PeSV-Fisher: identification of somatic and non-somatic structural variants using next generation sequencing data. *PLoS One* 2013; 8: e63377.

78. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z: Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 2009; 25: 2865-2871.

79. Krumm N, Sudmant PH, Ko A *et al*: Copy number variation detection and genotyping from exome sequence data. *Genome Res* 2012; 22: 1525-1532.

80. Stein LD: The case for cloud computing in genome informatics. *Genome Biol* 2010; 11: 207.

81. Richter BG, Sexton DP: Managing and analyzing next-generation sequence data. *PLoS Comput Biol* 2009; 5: e1000369.

82. Kumar P, Henikoff S, Ng PC: Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009; 4: 1073-1081.

83. Adzhubei IA, Schmidt S, Peshkin L *et al*: A method and server for predicting damaging missense mutations. *Nat Methods* 2010; 7: 248-249.

84. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A: Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 2005; 15: 901-913.

85. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D: MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010; 7: 575-576.

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

87. Fuentes Fajardo KV, Adams D, Mason CE *et al*: Detecting false-positive signals in exome sequencing. *Hum Mutat* 2012; 33: 609-613.

88. Teer JK, Bonnycastle LL, Chines PS *et al*: Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Res* 2010; 20: 1420-1431.

89. Aird D, Ross MG, Chen WS *et al*: Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 2011; 12: R18.

90. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 2008; 36: e105.

91. Meacham F, Boffelli D, Dhahbi J, Martin DI, Singer M, Pachter L: Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* 2011; 12: 451.

92. Church DM, Schneider VA, Graves T *et al*: Modernizing reference genome assemblies. *PLoS Biol* 2011; 9: e1001091.

93. Adams DR, Sincan M, Fuentes Fajardo K *et al*: Analysis of DNA sequence variants detected by high-throughput sequencing. *Hum Mutat* 2012; 33: 599-608.

94. Pelak K, Shianna KV, Ge D *et al*: The characterization of twenty sequenced human genomes. *PLoS Genet* 2010; 6: e1001111.

95. Rehman AU, Morell RJ, Belyantseva IA *et al*: Targeted capture and next-generation sequencing identifies C9orf75, encoding taperin, as the mutated gene in nonsyndromic deafness DFNB79. *Am J Hum Genet* 2010; 86: 378-388.

96. Walsh T, Shahin H, Elkan-Miller T *et al*: Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. *Am J Hum Genet* 2010; 87: 90-94.

97. Hall N: After the gold rush. *Genome Biol* 2013; 14: 115.

98. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program (GSP). Available at: www.genome.gov/sequencingcosts. Accessed September 2013.

99. Collins FS, McKusick VA: Implications of the Human Genome Project for medical science. *JAMA* 2001; 285: 540-544.

100. Collins F: Has the revolution arrived? *Nature* 2010; 464: 674-675.

101. Pierce SB, Walsh T, Chisholm KM *et al*: Mutations in the DBP-Deficiency Protein HSD17B4 Cause Ovarian Dysgenesis, Hearing Loss, and Ataxia of Perrault Syndrome. *Am J Hum Genet* 2010; 87: 282-288.

102. Gordon RD: Syndrome of hypertension and hyperkalemia with normal glomerular filtration rate. *Hypertension* 1986; 8: 93-102.

103. Wilson FH, Disse-Nicodeme S, Choate KA *et al*: Human hypertension caused by mutations in WNK kinases. *Science* 2001; 293: 1107-1112.

104. Hadchouel J, Jeunemaitre X: Life and death of the distal nephron: WNK4 and NCC as major players. *Cell Metab* 2006; 4: 335-337.

105. Boyden LM, Choi M, Choate KA *et al*: Mutations in kelch-like 3 and cullin 3 cause hypertension and electrolyte abnormalities. *Nature* 2012; 482: 98-102.

106. Pintard L, Willems A, Peter M: Cullin-based ubiquitin ligases: Cul3-BTB complexes join the family. *EMBO J* 2004; 23: 1681-1687.

107. Wakabayashi M, Mori T, Isobe K *et al*: Impaired KLHL3-mediated ubiquitination of WNK4 causes human hypertension. *Cell Rep* 2013; 3: 858-868.

108. Ohta A, Schumacher FR, Mehellou Y *et al*: The CUL3-KLHL3 E3 ligase complex mutated in Gordon's hypertension syndrome interacts with and ubiquitylates WNK isoforms: disease-causing mutations in KLHL3 and WNK4 disrupt interaction. *Biochem J* 2013; 451: 111-122.

109. Shibata S, Zhang J, Puthumana J, Stone KL, Lifton RP: Kelch-like 3 and Cullin 3 regulate electrolyte homeostasis via ubiquitination and degradation of WNK4. *Proc Natl Acad Sci U S A* 2013; 110: 7838-7843.

110. Ng SB, Turner EH, Robertson PD *et al*: Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 2009; 461: 272-276.

111. Hennekam RC, Biesecker LG: Next-generation sequencing demands next-generation phenotyping. *Hum Mutat* 2012; 33: 884-886.

112. Bell CJ, Dinwiddie DL, Miller NA *et al*: Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med* 2011; 3: 65ra64.

113. Pritchard CC, Smith C, Salipante SJ *et al*: ColoSeq provides comprehensive lynch and polyposis syndrome mutational analysis using massively parallel sequencing. *J Mol Diagn* 2012; 14: 357-366.

114. De Keulenaer S, Hellemans J, Lefever S *et al*: Molecular diagnostics for congenital hearing loss including 15 deafness genes using a next generation sequencing platform. *BMC Med Genomics* 2012; 5: 17.

115. Redin C, Le Gras S, Mhamdi O *et al*: Targeted high-throughput sequencing for diagnosis of genetically heterogeneous diseases: efficient mutation detection in Bardet-Biedl and Alstrom Syndromes. *J Med Genet* 2012; 49(8):502-12

116. Tucker EJ, Mimaki M, Compton AG, McKenzie M, Ryan MT, Thorburn DR: Next-generation sequencing in molecular diagnosis: NUBPL mutations highlight the challenges of variant detection and interpretation. *Hum Mutat* 2012; 33: 411-418.

117. Baetens M, Van Laer L, De Leeneer K *et al*: Applying massive parallel sequencing to molecular diagnosis of Marfan and Loeys-Dietz syndromes. *Hum Mutat* 2011; 32(9):1053-62

118. Morgan JE, Carr IM, Sheridan E *et al*: Genetic diagnosis of familial breast cancer using clonal sequencing. *Hum Mutat* 2010; 31: 484-491.

119. Walsh T, Lee MK, Casadei S *et al*: Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. *Proc Natl Acad Sci U S A* 2010; 107: 12629-12633.

120. Saunders CJ, Miller NA, Soden SE *et al*: Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci Transl Med* 2012; 4: 154ra135.

121. Pop M, Salzberg SL: Bioinformatics challenges of new sequencing technology. *Trends Genet* 2008; 24: 142-149.

7. ANNEX

List of publications of this thesis

Trujillano D, Bullich G, Ossowski S, Ballarín J, Torra R, Estivill X & Ars E.
Diagnosis of autosomal dominant polycystic kidney disease by efficient PKD1 and PKD2 multiplex high-throughput targeted sequencing. *Human Mutation*, [Submitted]

Trujillano D*, Pérez B*, González J, Tornador C, Navarrete R, Escaramis G, Ossowski S, Armengol L, Cornejo V, Desviat LR, Ugarte M & Estivill X.
Accurate molecular diagnosis of phenylketonuria and tetrahydrobiopterin-deficient hyperphenylalaninemias using high-throughput targeted sequencing. *European Journal of Human Genetics*, 2013 Aug 14. doi: 10.1038/ejhg.2013.175. [Epub ahead of print]

Trujillano D, Ramos MD, González J, Tornador C, Sotillo F, Escaramis G, Ossowski S, Armengol L, Casals T, Estivill X.
Next generation diagnostics of cystic fibrosis and CFTR-related disorders by targeted multiplex high-coverage resequencing of CFTR. *Journal of Medical Genetics*, 2013 Jul;50(7):455-62.

Louis-Dit-Picard H*, Barc J*, Trujillano D*, Miserey-Lenkei S, Bouatia-Naji N, Pylypenko O, Beaurain G, Bonnefond A, Sand O, Simian C, Vidal-Petiot E, Soukaseum C, Mandet C, Broux F, Chabre O, Delahousse M, Esnault V, Fiquet B, Houillier P, Bagnis CI, Koenig J, Konrad M, Landais P, Mourani C, Niaudet P, Probst V, Thauvin C, Unwin RJ, Soroka SD, Ehret G, Ossowski S, Caulfield M; International Consortium for Blood Pressure (ICBP), Bruneval P, Estivill X, Froguel P, Hadchouel J, Schott JJ, Jeunemaitre X.
KLHL3 mutations cause familial hyperkalemic hypertension by impairing ion transport in the distal nephron. *Nature Genetics*, 2012 Mar 11;44(4):456-60, S1-3.

*Shared first authorship

Additional publications during the thesis period

Ramos M*, Trujillano D*, Olivar R, Sotillo F, Ossowski S, Manzanares J, Costa J, Gartner S, Oliva C, Quintana E, Gonzalez M, Vazquez C, Estivill X, Casals T.
Extensive sequence analysis of CFTR, SCNN1A, SCNN1B, SCNN1G and SERPINA1 suggests an oligogenic basis for cystic fibrosis-like phenotypes. *Clinical Genetics*, 2013 Jul 9. doi: 10.1111/cge.12234. [Epub ahead of print]

Caley DP, Pink RC, Trujillano D, Carter DRF.
Long noncoding RNAs, chromatin, and development. *The Scientific World Journal*, 2010 Jan 8;10:90-102.

*Shared first authorship

Communications to scientific meetings related to this thesis

Next-Generation Sequencing: From the Lab to the Clinic
Trujillano D.

Invited Speaker, *"Integrated Biology – Agilent Usermeeting"*, Barcelona, Spain (2013)

Invited Speaker, *"Business Innovation Day – Centre for Omic Sciences"*, Reus, Spain (2013)

Invited Speaker, *"Advances in Genome Sciences – Illumina Seminar Series"*, Madrid, Spain (2012)

CFTR Gene Direct Sequencing Versus Scanning Techniques. Improving the Sensitivity to Identify Cystic Fibrosis Mutations
Trujillano D, Ramos MD, Vazquez A, Estivill X, Manzanares J, Casals T.

Poster, *"35th European Cystic Fibrosis Conference"*, Dublin, Ireland (2012)

Additional communications to scientific meetings during the thesis period

Genetic Clues of Healthy Ageing Identified by Sequencing the Supercentenarian's Genomes
Trujillano D, Badarinarayan N, Rosentiel P, Bayés M, Ossowski S, Tornador C, Escaramís G, Flachsbart F, Nebel A, Lathrop M, Franke A, Gratacòs M, Schreiver S, Gut I, Estivill X.

Poster, *"American Society of Human Genetics 62nd Annual Meeting"*, San Francisco, CA, USA (2012)

Poster, *"UniSR-CRG-Dimet Joint PhD retreat"*, Presezzo, Italy (2012).

Poster, *"The Biology of Genomes"*, Cold Spring Harbor Laboratory, NY, USA (2012)

Common and Rare Variants in Obsessive Compulsive Disorder Identified by Exome and Targeted Resequencing
Trujillano D, Ossowski S, Tornador C, Alonso P, Gratacòs M, Estivill X.

Poster, *"American Society of Human Genetics 62nd Annual Meeting"*, San Francisco, CA, USA (2012)
Poster, *"The Biology of Genomes"*, Cold Spring Harbor Laboratory, NY, USA (2012)

Application of next generation sequencing technologies to find new variants in Progressive Cardiac Conduction Defect
Trujillano D, Daumy X, Kyndt F, Ossowski S, Le Marec H, Estivill X, Probst V, Redon R, Schott JJ.

Poster, *"Printemps de la Cardiologie"*, Bourdeaux, France (2012)

Exome Sequencing in Familial Cases of Mitral Valve Prolapse
Trujillano D, Ossowski S, Estivill X, Schott JJ.

Oral Communication, *"Leducq Mitral Meeting"*, Johns Hopkins Medical Institute, Baltimore, USA (2011).

Targeted Resequencing Applied to the Study of Obsessive-Compulsive Disorder
Trujillano D, Ossowski S, Tornador C, Alonso P, Gratacòs M, Estivill X.

Oral Communication, *"V CRG PhD Student Symposium"*, Barcelona, Spain (2011)

A Novel, Diet-induced, murine model of Obsessive Compulsive Disorder
Trujillano, D, McDonal, D, Gratacòs, M, Dierssen, M, Estivill, X.

Poster, *"XIXth World Congress on Psychiatric Genetics"*, Washington DC, USA (2011)

Whole-Exome Sequencing of 40 Obsessive-Compulsive Disorder Patients
Trujillano D, Ossowski S, Tornador C, Alonso P, Gratacòs M, Estivill X.

Poster, *"12th International Congress of Human Genetics"*, Montreal, Canada (2011)
Poster, *"XIXth World Congress on Psychiatric Genetics"*, Washington DC, USA (2011)
Poster, *"12th International Meeting on Human Genome Variation and Complex Genome Analysis"*, Berkeley, USA (2011)

Poster, *"The Genomics of Common Diseases 2011"*, Wellcome Trust Genome Campus, Cambridge, UK (2011)

Targeted Resequencing of Susceptibility Genomic Regions for Psoriasis

Trujillano D, Riveira E, Gratacòs M, Estivill X.

Poster, *"European Human Genetics Conference"*, Amsterdam, The Netherlands (2011)

Oral Communication, *"7th Workshop on Biomedical Genomics and Proteomics"*, Barcelona, Spain (2010)