PhUSE 2017

Paper AS05

A real-life experience on the challenge called genetics

S. W. Fouchier, OCS Life Science, The Netherlands

ABSTRACT

Next-generation sequencing (NGS) technologies, such as whole-genome sequencing (WGS), whole-exome sequencing (WES), and targeted sequencing, are increasingly applied to medical study and practice to identify disease- and/or drug-associated genetic variants to advance personalised medicine.
The fast development of NGS technologies has made it increasingly easier to sequence a whole genome/exome against affordable costs. However, the current NGS platforms also generate huge amounts of raw data, which has moved molecular biology into the big data era.
Despite the large number of data analysis options, the lack of concordant results from different variant detection tools and the complexity of the human genome, challenges the detection of clinically relevant variants.

INTRODUCTION

Pharmacogenomics (PGx) investigates how the inter-individual differences of genomic components affect patient responses to disease and to treatment and has been widely recognized as the fundamental steps toward personalized medicine. While pharmacogenomics encompasses a more genome-wide association approach, incorporating genomics and epigenetics while dealing with the effects of multiple genes on drug response, **Pharmacogenetics (PGt),** a subset of PGx, focuses on single drug-gene interactions only, influenced by variations in the DNA sequence.
Fast development of next-generation sequencing (NGS) technologies not only has made it increasingly easier to sequence a whole genome/exome against affordable costs, it also unravelled a remarkable degree of genomic diversity of the human genome. with the discovery of a tremendous degree of individual-level variation with many single nucleotide variants (SNVs) that are “private” to each individual genome. In the absence of a reasonable understanding of how a SNV, e.q. the genotype, influence disease or response to treatment, e.q. the phenotype, the promise of personalized medicine may seem as a distant prospect.

Most so-called pharmacogenes encode drug-metabolizing enzymes, and each individual’s genotype for a particular gene can be categorized into phenotypes that describe the enzyme’s activity, ranging from ultra-rapid metabolizer to poor metabolizer. Other genes encode the enzymes that are the site of action where drugs exert their effects, patients with variations in these genes may be more sensitive or resistant to certain drugs than normal.
To date, the FDA has included pharmacogenomic information on the labels of more than 150 medications, but most of these do not translate the genetic test results into specific prescribing actions. Only ~7% of the ~1200 FDA approved drugs, represented by only 17 of ~19,000 human genes are considered clinically actionable for pharmacogenomics.[1] How come that the implementation of **PGt** to improve personalized drug safety and efficacy is rather limited in the current clinical setting?

THE BASICS OF GENETICS

For starters, the sequence of the human genome has been (almost) completely determined by DNA sequencing, however, it is far from fully understood.
The content of the human genome is commonly divided into coding and noncoding DNA sequences. Coding DNA, called exons, are defined as those sequences that can be transcribed into mRNA and translated into proteins during the human life cycle; these sequences occupy only a small fraction of the genome (<2%). Non-coding DNA, including introns, are made up of all of those sequences (ca. 98% of the genome) that are not used to encode proteins.
Most biological activities are carried out by proteins and thus these proteins are critical to the proper functioning of cells and organisms. Knowledge on how these proteins function in the human body is key. However, to date, of the ~19,000 predicted protein-coding genes in humans, only 2,937 genes have been discovered underlying a Mendelian disease, a disease

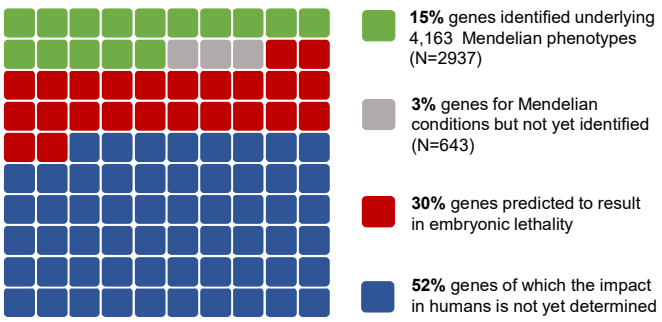


Figure 1. Relationship Human Protein-Coding Genes and Mendelian Phenotypes.

defined by patterns of inheritances, while the genes underlying ~50% (i.e., 3,152) of all known Mendelian phenotypes are still unknown, and many more Mendelian conditions have yet to be recognized (Figure 1).[1]
Additionally, in the past decade it has become more and more evident that non-coding regions, even located far away from a gene, can also have an substantial impact on protein expression and underlie Mendelian diseases.

DNA – RNA - PROTEIN

DNA sequencing reveals the order in which the four unique nucleotides; (**A**)denine, (**G**)uanine, (**C**)ytosine and (**T**)hymine lay within the genome. The bases on one strand of DNA form base pairs with a second strand of DNA to form the double helix, whereby A can only form a base pair with T and G can only form a base pair with C. The sequence of these nucleotides within the coding regions are coded in triplets (codons) which determines the sequence of 20 unique amino acids in a protein.
The instructions stored within the **DNA** are "read" in two steps, called transcription and translation (Figure 2). In transcription, a portion of the double-stranded DNA template gives rise to a single-stranded messenger RNA (**mRNA**) molecule, an exact copy of the DNA, with the only difference that the (**T**)hymine translates into (**U**)racil. This initial mRNA must then be processed before it becomes a mature mRNA that can direct the synthesis of protein. One of the steps in this processing is called RNA splicing, which involves the removal or "splicing out" of the intervening intronic sequences. The final **mature mRNA** thus consists of the exons, which are connected to one another through this splicing process. The processing of the mRNA molecule is then followed by a translation step, which ultimately results in the production of a **protein** molecule.

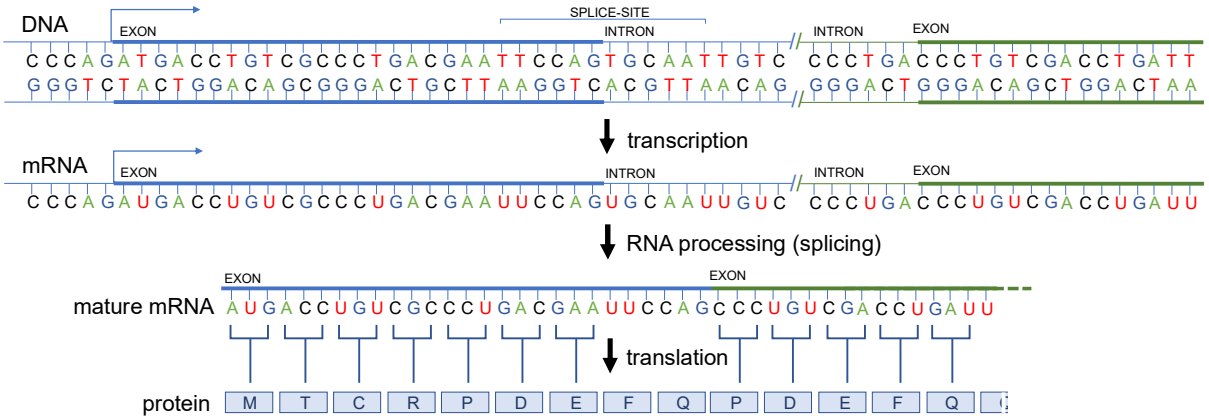


Figure 2. DNA transcription, RNA processing and translation results in the production of a protein.

SINGLE NUCLEOTIDE POLYMORPHISM (SNP) VERSUS MUTATION

A single nucleotide variation (SNV) is a variation in a single nucleotide, and can be considered to be a single nucleotide polymorphism (SNP) or a mutation. Although the definition is not black and white, but roughly we may say that a SNP is a common (found in at least 1% of the population), non-functional SNV often located in the non-coding regions of the genome, while a mutation is a rare disease-causing SNV. Different types of mutations do exist but their effect on protein functions may differ substantially (Figure 3).
A nonsense, frameshift or splicing mutation, however, often result in a severe phenotype. A point mutation that turns one codon into a stop codon is called a **nonsense** mutation, which always result in the early termination of protein translation. A **frameshift** mutation is a mutation caused by indels (insertions or deletions) of a number of nucleotides in a DNA sequence that is not divisible by three. Due to the triplet nature of gene expression by codons, the insertion or deletion can change the reading frame (the grouping of the codons), resulting in a completely different translation from the original. When the nucleotide change occurs within the region involving in the RNA processing, the splice-site, such mutations are referred to as **splicing** mutation. Depending on the location of the splice site, at the start or the end of an exon, such mutations could respectively result in translation of intronic regions or exon skipping. Translation of intronic regions also results in a completely different translation from the original, while exon skipping cause the production of a much shorter protein.
The effect of a missense or silent mutation, however, can be much more difficult to predict. A **missense** mutation is a point mutation in which a single nucleotide change results in a codon that codes for a different amino acid. It is a type of nonsynonymous substitution. When the single nucleotide change does not result in an amino acid change it is called a **silent** or synonymous mutation. Although the majority of silent mutations do not cause a phenotype and are often found in the general population with a frequency of >1%, and therefore may be considered a SNP, depending on the location in the DNA they do sometimes have a substantial effect on protein function and cause a phenotype. For missense mutations, it can work the other way around. Although most missense mutation cause a phenotype, depending on the location in the gene or which amino acid substation occur, the effect on protein function can be nihil. For instance, a pair change resulting in an amino change of a (V)aline into a (I)soleucine often

does not cause a phenotype, since the structure of these amino acids are very much alike and therefore such amino acid change does not have a substantial effect on protein function.

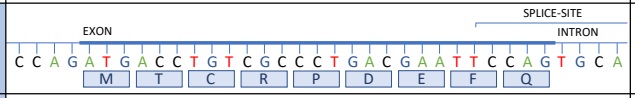
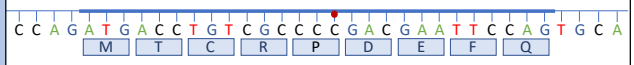
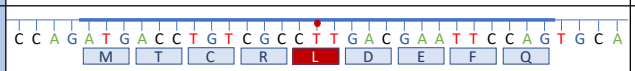
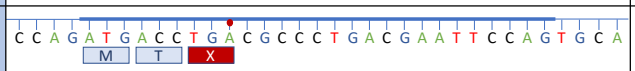
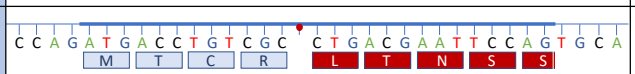
Type of mutation	Genotype	Effect	Phenotype
Consensus			
Silent		Base pair change NO amino acid change (synonymous)	none
Missense		Base pair change AND amino acid change (non-synonymous)	moderate
Nonsense		Base pair change, resulting in a stop codon	severe
Frame-shift (Indel)		Deletion of one base pair (indel), resulting in frame-shift	severe
Splice		Base pair change in splice-site, resulting in translation of intronic sequence	severe

Figure 3. Different types of mutations.

WHOLE EXOME SEQUENCING (WES)

Whole-exome sequencing (WES) is an application of the next-generation technology that determines the variations in and around all coding regions of known genes and thus focusing on only a very small proportion of the human genome. For novel-gene finding strategies, identifying the unknown gene underlying a Mendelian disease, a typical WES analysis pipeline consists of three parts, the whole exome capture and sequencing pipeline, the computational pipeline, and the interpretation pipeline (Figure 4).

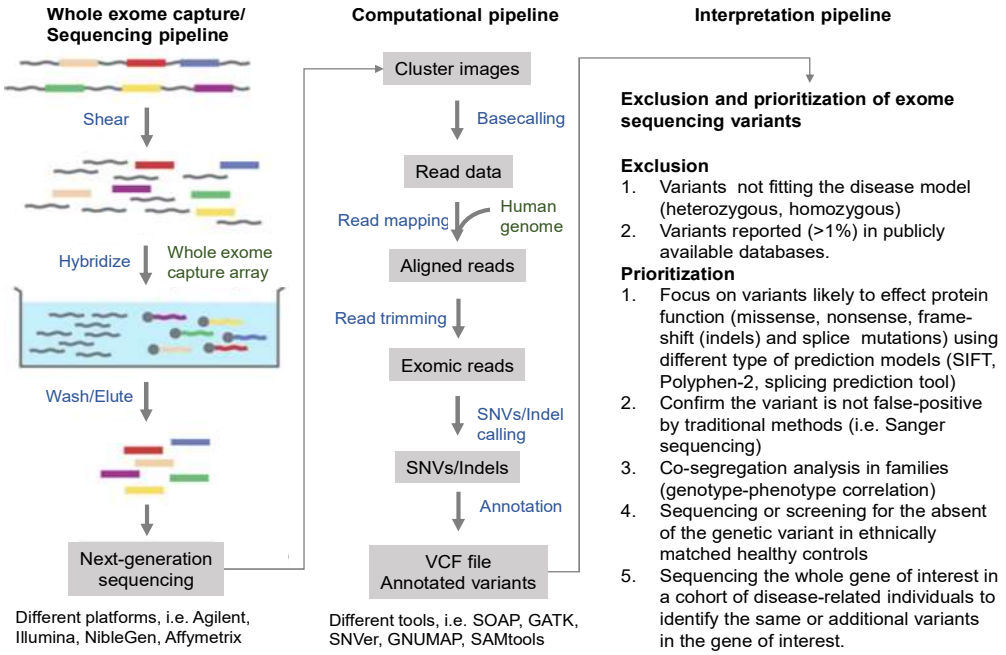


Figure 4. Scheme of a typical WES analysis pipeline.

For the whole exome capture and sequencing pipeline several different NGS platforms (i.e. Agilent, Illumina, NibleGen, Affymetrix) are available that selectively targeting and enrich genomic regions of interest before sequencing. After capturing these regions of interest the sequences can then be read by different machineries, the Roche 454 sequencer, the Illumina Genome Analyzer II and the Life technologies SOLiD & Ion Torrent, all of which have been used for exome sequencing.

After sequencing, there are several different **computational** tools (i.e. SOAP, GATK, SNVer, GNUMAP, SAMtools) to choose from to analyse the raw data by reading, cleaning and aligning the sequences to the genome.

The final step is to **interpretate** the many SNVs identified per sequenced exome. On average, WES identifies ~20,000 SNVs in individuals from European descent (Figure 5).[2] More than 95% of these SNVs are already known in human populations and reported in publicly available databases like dbSNP, the 1000 Genomes Project, ExAC Database and GoNL. **Exclusion** of these variants can greatly reduce the number of potential candidate SNVs by 90% to 95%. However, hundreds of novel rare and private potential disease-causing variants remain. It can be very time-consuming and expensive to investigate all these private SNVs in depth and therefore it is necessary to **prioritize** these SNVs based on the likelihood of being the disease-causing variant using prediction models like SIFT, Polyphen-2 and splice site predictions models. Since prioritizing largely depends on prediction and current knowledge there is always the risk that the disease-causing variant is filtered out in this step. Nevertheless, it makes sense to only focus on those SNVs that are likely to cause a phenotype (i.e. missense, nonsense, frameshift, deletions, insertions).

After confirmation by i.e. traditional Sanger sequencing that the variant is real and not falsely called, co-segregation analysis in families would be the next step to investigate whether a variant is likely to cause a phenotype whereby ideally each family member with the genotype. e.q. carrier of the variant, has the expected phenotype e.q. affected by the disease.

Variant type	Total variants	SD	Non-novel variants	SD	Novel variants	SD
synonymous	10,645	(± 286)	10,536	(± 288)	109	(± 16)
missense	9,511	(± 244)	9,319	(± 233)	192	(± 21)
nonsense	93	(± 6)	89	(± 6)	5	(± 2)
splice	34	(± 4)	32	(± 3)	2	(± 1)
total	20,283	(± 523)	19,976	(± 505)	307	(± 33)

Figure 5. Mean number of coding variants in one European individual

Additionally, the absence of the candidate variant in a cohort of healthy individuals or the present of the specific candidate variant or new potentially disease-causing variants in the same gene in a disease cohort might provide the proof you are ending up with the actual disease-causing variant.

THE FITFALLS OF WES

Given the existence of multiple platforms and multiple data-analysis pipelines for NGS, researchers and clinicians may be under the impression that these methods all work similarly to identify genetic variants from personal genomes. While the basic sample preparation protocols are similar among all available platforms, major differences however lie in the design, including selection of target genomic regions, sequence features and lengths of probes, and the exome capture mechanisms. The various sequence technologies therefore result in different error rates and generate various read-lengths which can pose challenges in comparing results from different platforms. This is extremely problematic due to the high sensitivity of these techniques resulting in high false positive and false negative rates which consequently results in the diversity of results.

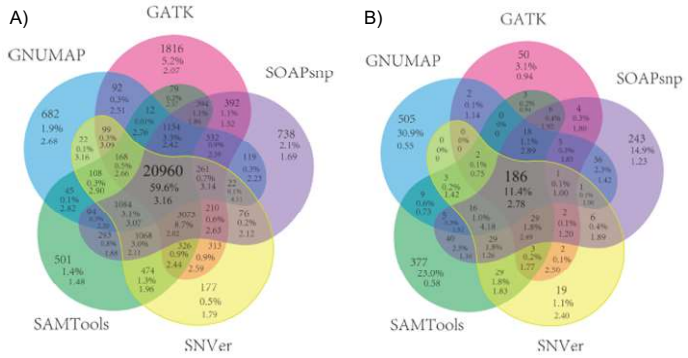


Figure 6. Mean SNVs concordance over 15 exomes between five alignment and variant-calling pipelines. A). Non-novel variants. B). Novel variants.

Analysis of the same DNA sample using different NGS platforms or using different computational pipelines still results in substantial different outcomes (Figure 6). This is especially true for novel SNVs showing an overall concordance of 11.4% compared to the overall concordance found for known SNVs (59.6%).[3]

Further development of genotype calling techniques and methods for validation are required in order to make NGS data usable and reliable. Additionally, no available NGS technique covers 100 % of the human genome. These inevitable gaps in coverage are a limitation that needs to be considered carefully when investigating the cause of a Mendelian disorder.

THE REAL-LIFE EXPERIENCE

Hypercholesterolemia is a disease characterized by elevated low-density lipoprotein cholesterol (LDL-C) levels and increased risk for coronary vascular disease. The inherited trait of this disease can be dominant in which the inheritance of one affected allele from father or mother results in an affected phenotype or can be recessive in which the inheritance of two affected alleles, inherited from father and mother, is needed to result in a phenotype. Within the recessive model both parents are only carriers of the affected allele, but do not express the phenotype (Figure 7). The chance to inherit a dominant familiar disease with one affected parent is 50%. The prevalence of a disease with a recessive trait is much higher compared to a dominant trait, since the chance that two individuals with the same mutation in one gene generates offspring together is very low, and the chance to inherit both affected alleles from father and mother is 25%.

For the hypercholesterolemic phenotype, autosomal dominant hypercholesterolemia (ADH) is a common disease with a prevalence of ≈1 in 200 individuals in most Western countries. While the prevalence of autosomal recessive hypercholesterolemia (ARH) is rare (≈1 in 300,000).[4]

Since the physical stigmata usually develop later in life, a molecular (genetic) diagnosis early in life is warranted when striving for maximum health benefit as recommended by the WHO.[5] Genetic screening of affected families is an efficient way of identifying subjects with ADH[6] and has contributed to reducing cardiovascular morbidity and mortality.[7,8]

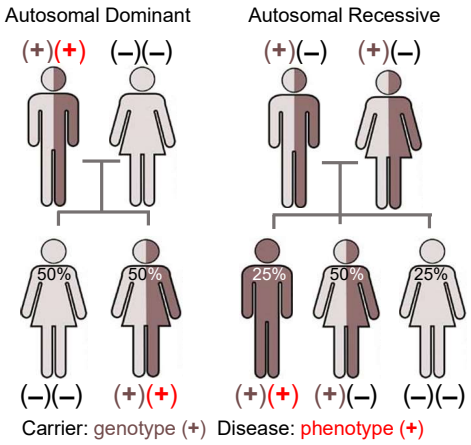


Figure 7. Monogenetic Mendelian traits of inheritance

THE GENOTYPE

While the disease hypercholesterolemia was already described in 1938 as being a genetic inherited disease, the molecular defect underlying this disease was not discovered until 1973. Goldstein and Brown, who received the Nobel price in 1985 for their discovery, found that severely affected individuals who were lacking the LDL receptor (LDLR) protein were unable to remove LDL-cholesterol particles from the circulation.[9] In 1990, a defect in a second gene coding for the apolipoprotein B (APOB) protein was discovered. This protein, as part of the LDL- cholesterol particle, is responsible for the binding of the LDL- cholesterol particle with the LDLR protein.[10] Mutations in specific regions of the *APOB* gene unables the LDLR to recognize the LDL-cholesterol particle. A third gene, discovered in 2003, coding for a protein called proprotein convertase subtilisin/kexin 9 (*PCSK9*), also showed to cause ADH.[11]

In the majority of molecular diagnosed patients, patients in which the genetic cause has been identified, *LDLR* mutations are most frequently found (~87.6%), while the occurrence of *APOB* mutations are less frequent (~12.3%) and *PCSK9* mutations are very rare (< 1%). The occurrence of mutations in different genes all resulting in a similar phenotype underline the genetic heterogeneity of this disease. More so, while in *APOB* and *PCSK9* only a dozen different mutations have been reported, over 800 different *LDLR* mutations have been reported to affect LDLR protein function ranging from mild to severe.[12,13] Additionally, mutations in these three genes only explain the phenotype in ~30% of all clinically diagnosed ADH patients.[13] Unravelling the molecular determinants in this large proportion of molecular unexplained ADH patients has been a major focus of cardiovascular research. However, identification of novel ADH genes is largely challenged by the genetic heterogeneity of this disease, as well as by the phenotypic heterogeneity.[14]

THE PHENOTYPE

Cholesterol levels in blood are sex and age dependent. To easily compare cholesterol levels between male and female of all ages, cholesterol levels are 'converted' into percentiles Subjects are considered as affected when cholesterol levels are above a cut-point of the 95th percentile, representing the 5% of the general population with the highest cholesterol levels. The discriminative ability of LDL-cholesterol levels to identify patients carrying a causal variant, however, is quite low and largely depends on the severity of the mutation. This became extreme clear when over 10 000 carriers of true pathogenic *APOB* and *LDLR* mutations and 20,000 relatives negative for the familial mutation, belonging to approximately thousand families collected via the cascade screening programme were evaluated on genotype-phenotype correlation.[15]

The Dutch nationwide genetic cascade screening programme for ADH was started in 1994. For this purpose, DNA samples from clinically suspected patients with ADH were analysed for the presence of an *LDLR* or *APOB* mutation. A patient was considered a proband for family screening when a pathogenic mutation was identified. Subsequently, first-degree relatives were offered DNA analysis for the presence of the specific ADH-causing familial mutation. The cascade screening was then further extended to identify distant relatives of the probands by using the inheritance pattern across the pedigree. Important to emphasise here is that only the probands were selected based on cholesterol levels, while relatives of these probands were identified based on being carrier of the familial mutation.

To understand why the discriminative ability of LDL-cholesterol levels to identify patients carrying a causal variant is quite low and largely depends on the severity of the mutation, the term **incomplete penetrance**, caused by non-penetrants and phenocopies, should be addressed here. As indicated before, for the ADH phenotype subjects are consider to be affected when LDL-cholesterol levels are above the 95th percentile. Additional in common novel-gene finding strategies a disease penetrance of 0.9 is expected, meaning that 90% of those with the mutation will develop the disease, while 10% will not. An individual affected by the genotype, but does not express the phenotype, is called a **non-penetrant**. The other way around, an individual not affected by the genotype, but does have the phenotype, due to other unknown causes resulting in the same phenotype, is called a **phenocopy**.

Applying the common ADH novel-gene finding strategies criteria on the thousand families collected via the cascade screening programme showed that these criteria were only applicable for mutations in the *LDLR* gene that cause a severe phenotype (i.e. nonsense, frameshift (indels) and splicing mutations). However, in de majority (~70%) of the families, the ADH phenotype was caused by a missense mutation, and 70% of all families affected by these *LDLR* missense mutation did not fulfil the currently accepted novel-gene finding criteria for ADH. This means that when these families were used to identify novel genes causing ADH, this novel gene would never have been discovered.

Thus, although ADH is considered to be a Mendelian disease (due to the influence of a single gene), the phenotype can clearly be influenced by additional factors, genetically or environmentally i.e. life-style, diet, emphasizing that the discovery of novel disease genes can be largely challenged by the genetic and phenotypic heterogeneity of a disease.

THE PRE-WES ERA

Since ADH is a common disease, a selection of severely affected individuals only, in an attempt to identify novel genes, can be easily made. However, as soon as a family becomes bigger and additional relatives are collected, the phenotypic heterogeneity comes in sight (Figure 8). Individuals with and LDL-cholesetrol >95th percentile are clearly affected, but individuals with LDL-cholesterol levels in the upper range (between 75th and 95th percentiles) of the LDL-cholesterol spectrum cannot simply be ignored. Nevertheless, focussing on the branches with the most severe phenotypes could increase the success-rate of the novel-gene finding attempt.

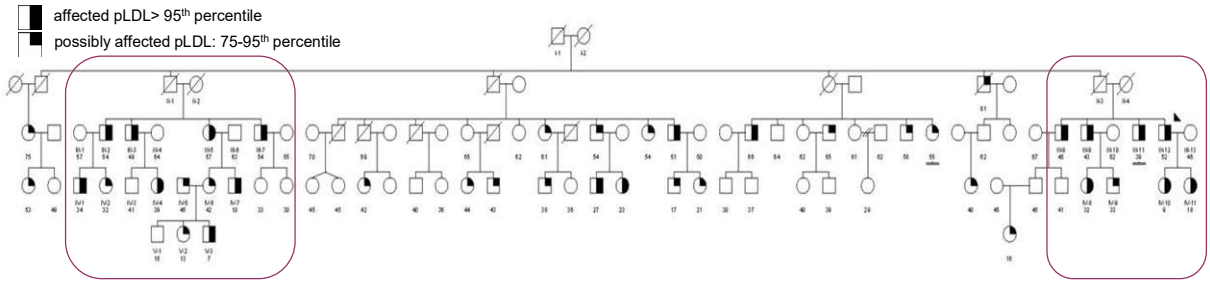


Figure 8. A large ADH family affected by an unknown gene.

The history of the family depicted here started in 1995 when the proband (represented by the triangle in Figure 8) was hospitalized due to his first signs of suffering from cardiovascular disease. Since the positive family history for hypercholesterolemia, including a brother who had suffered a myocardial infarction at the very young age of 35, blood samples of all first-degree relatives were collected to measure their cholesterol levels. In addition, the probands' DNA was screened for mutation in the known genes related to ADH. However, no *LDLR* mutation, nor an *APOB* mutation could be found. In the following years the family was further expanded, cholesterol levels were measured and DNA samples were collected.

In those days microsatellite markers were widely used as genetic markers in disease gene mapping and family linkage studies to identify the susceptibility loci or genes for monogenic and complex diseases. Although microsatellite markers are more informative than SNPs at the individual marker level, their number is far less than the several million SNPs in the human genome. With the introduction of SNP arrays, it became much easier and less expensive to identify 'smaller' regions of interest.

In 2009, linkage analysis using SNP arrays was performed on 15 relatives in the two branches with the strongest phenotype, in affected (>95th percentile) as well as unaffected (<50th percentile) relatives. This analysis revealed specific regions on chromosome 4, 11 and 13 shared by the affect individuals (Figure 9). Although these results narrowed down the regions of interest enormously, all three regions were still millions of base pair long, harbouring hundreds of different candidate genes.

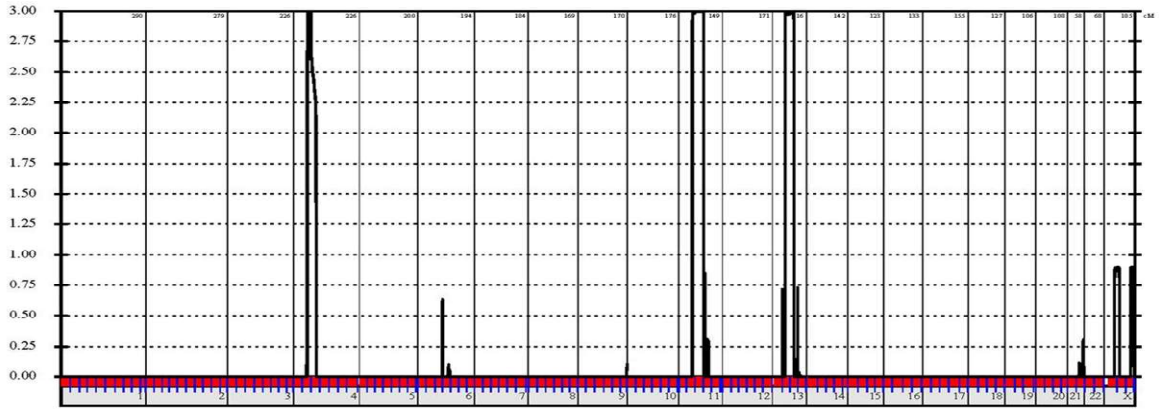


Figure 9. Multipoint-linkage analysis using SNP array data

INTO THE WES ERA

In 2011, WES was performed on the proband, his daughter and one most distant related relative sharing the genetic regions on Chromosome 4, 11 and 13. In all three individuals ~42,000 coding and non-coding SNVs were called. Since ADH is a dominant disease, a SNV was fitting the disease model if only one allele was affected and thus should be present in all three affected individuals in heterozygous form. Additionally, these SNVs should not be present with a high frequency in publicly available databases. These exclusion criteria resulted in the identification of 394 SNVs in 315 different genes (Figure 10). By selecting only SNVs located in the coding region and excluding silent mutations, since their effect on protein function is likely nihil, still dozens of candidate SNVs remain and follow-up all these SNVs in the additional prioritizing steps would be expensive and very time-consuming. Exclusion criteria

steps are preferred over prioritization step 1, because within this prioritization step the chance of filtering out the actual causal SNV is much higher. Extending the exclusion steps with the availability of an in-house database containing WES data of individuals with a known phenotype could help enormously to reduce the list of candidate SNVs. In this case 6 exomes of individuals with LDL-cholesterol levels <50th percentile was used. Additionally, the linkage analysis data pointed to a candidate SNVs that should be located on chromosome 4, 11 or 13, resulting in only a handful candidates. However, there was one problem.

Exclusion	# candidate variants		# candidate genes	
1. disease model (heterozygous)	5,953		2,822	
2. publicly databases (<1%)	394		315	
3. In-house database with phenotype		[167]		[143]
4. linkage analysis (Chr 4, 11, 13)		[33]		[27]
Prioritization				
1. Variants likely to effect protein function				
coding variant	149	[13]	131	[8]
exclusion of silent mutations	64	[3]	58	[3]
2. Confirm the variant is not false-positive by traditional methods (i.e. Sanger sequencing)				
3. Co-segregation analysis in families (genotype-phenotype correlation)				
4. Sequencing or screening for the absent of the genetic variant in ethnically matched healthy controls				
5. Sequencing the whole gene of interest in a cohort of disease-related individuals to identify the same or additional variants in the gene of interest				

Figure 10. Workflow for exclusion and prioritization of exome sequencing variants

The liver is the main organ responsible for the balance of cholesterol levels in the body. As a consequence, you would expect that a novel cholesterol-related gene would, at least in low quantities, being expressed in the liver. However, none of the candidate SNVs, even after evaluation of all SNVs identified by the exclusion steps only, made sense. The lack of knowledge on the genes in which the candidate SNVs were located hampered the identification of the actual causal variant. With lack of reasoning that a gene could potentially play a role in the processes related to the disease, following-up the candidate SNVs is a rather high-risk approach, because it makes you wonder whether the causal variant was actually called in the first place, and maybe the causal variant is located within the 98% non-coding region of the genome. In parallel to the research performed in the family described here, several other families were investigated with the same approach. In comparison to these other families it did seem that, although the candidate genes made less sense, at least we were most successful in tuning down the candidate SNVs, in this family. So, a year later, additional steps were undertaken. Co-segregation of one specific SNV (p.Glu97Asp) located in a gene called *STAP1*, showed that eleven carries had a total or LDL-cholesterol levels >95th percentile, while one clearly unaffected individual was not carrier of the variant. Additionally, four carriers had total or LDL-cholesterol levels in the upper range of the spectrum (between the 75th and 95th percentile). Including these individual as being “affected” resulted in a disease penetrance of 0.94, 15 out of 16 carriers (Figure 11A). Maybe even more important the *STAP1*: p.Glu97Asp was not found in a cohort of 500 healthy controls, nor were other coding *STAP1* SNVs found in this cohort.

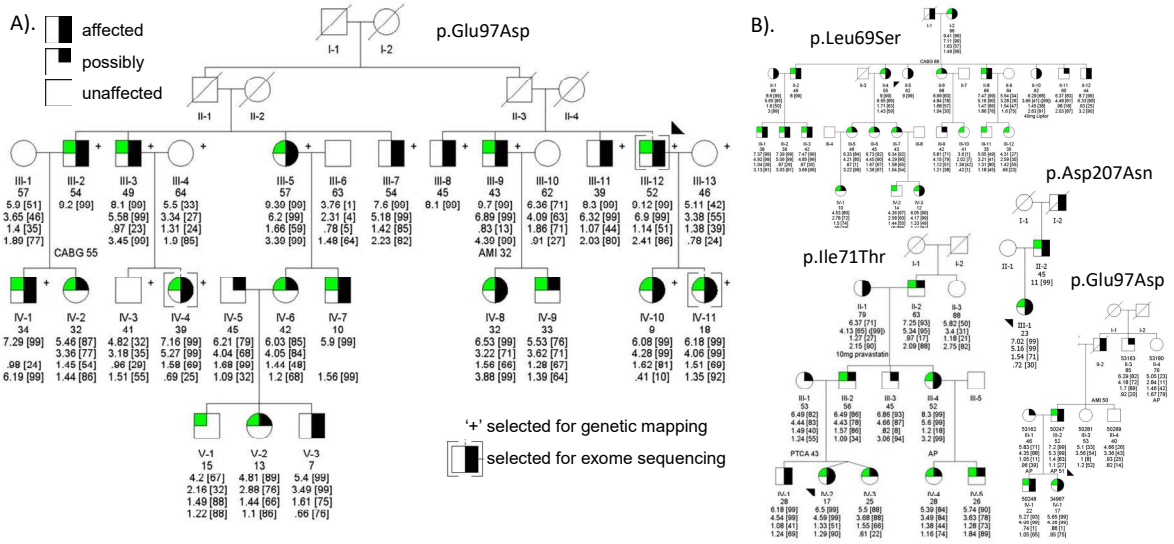


Figure 11. A). Sixteen *STAP1* mutation carriers (green square), 11 carriers with LDL-cholesterol >95th percentile, 4 carrier with LDL-cholesterol between 75th and 95thpercentile and 1 carrier with LDL-cholesterol below 75th percentile, reflecting a disease penetrance of 0.94. **B).** Four additional *STAP1* families.

The only remaining step was to find the *STAP1*: p.Glu97Asp or other putative protein affecting variants in a cohort of clinically diagnosed ADH patients. Over 500 ADH patients, in which no *LDLR*, *APOB* nor *PCSK9* mutations were found to explain the hypercholesterolemic phenotype, were screened by Sanger sequencing including the complete coding and exon-intronic boundaries of the *STAP1* gene. This resulted in the identification of one additional *STAP1*: p.Glu97Asp carrier and three additional novel missense *STAP1* mutations, which were not found, or only with very low frequency, in publicly available SNV databases (Figure 11B).

So after over a decade of research, walking through all novel-gene finding strategic steps possible within the field of genetics, the results were published in 2014[16]. Did we discover the fourth cholesterol gene causing ADH? Since *STAP1* is not expressed in the liver and the function of STAP1 is still largely unknown, building a hypothesis in how mutations in *STAP1* cause the ADH phenotype is mainly speculative. Did we convince the scientific world that we discovery the fourth ADH gene? I am not so sure, yet. It is now to others to support these findings or to proof that these findings were find just by chance.

CONCLUSION

In the past five years, WES increased our knowledge in the field of genetics enormously, but many hurdles are still to be taken, and many gaps are still needed to be filled, before this knowledge can lead us into a fully implemented personalized medicine era.
In the near future, whole genome sequencing (WGS) probably will overcome the fact that WES only covers a very small proportion of the human genome, but WGS is currently still too expensive for most laboratories. Additionally, the use of publicly available SNV databases is essential to filter out the majority of SNV that do not seem to be of any importance. For WGS, such databases are currently limited available. Also, genomic information even from a large number of individuals is of limited value, in the absence of deep phenotypic characterization.

Full discovery of the genes implicated in all Mendelian phenotypes will connect genes and their protein products to biological systems and clinical phenotypes, provide a pathway for developing better strategies to find genetic and environmental modifiers underlying common diseases, and enable understanding of the functional and phenotypic consequences of coding and non-coding SNVs. However, discovering the genetic basis of a Mendelian phenotype in families often is hampered by the large number of candidate variants that remain, no gene with causal variants or no compelling candidate variants that could be identified.
Missing even a single variant can mean the difference between discovering a disease-contributing mutation or not. One can minimize false positives by increasing stringency filters, but this automatically and correspondingly increases the false-negative rate. For this reason, using a single bioinformatics pipeline for discovering disease related variation is not always sufficient. A more comprehensive approach, all variants discovered by multiple variant-calling pipelines, when coupled with appropriate no-calling and quality filtering, could be included in downstream analyses, so as to not miss potentially disease-contributory variants.

Apart from all the technical hurdles and gaps related to WES and WGS, genetic and phenotypical heterogeneity are common phenomena, even for Mendelian diseases. This effect may be due to different mutations in the same gene resulting in similar, but not identical phenotypes or due to other genes influencing the disease phenotype resulting in multiple affected family members experiencing different levels of disease severity and outcomes. When there is uncertainty about the correct mode of inheritance for a phenotype the rate of gene discovery will be considerably lower than when the mode of inheritance is clear or easily predicted.
Even if a discovery is made, there is a considerable discordance between the number of discoveries and the number of publications reporting the discovery.[1] In part, this is reflected by the time that is required to identify and test additional families, but even more important, the time that is needed to elucidate the underlying molecular, developmental, and physiological mechanisms related to the novel gene. However, these efforts are essential to ensure the validity of each discovery.

REFERENCES

1. Chong JX, Buckingham KJ, Jhangiani SN, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet.* 2015;97(2):199-215.
2. Bamshad MJ, Ng SB, Bigham AW, et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet.* 2011;12(11):745-755.
3. O'Rawe J, Jiang T, Sun G, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.* 2013;5(3):28.
4. Sjouke B, Kusters DM, Kindt I, et al. Homozygous autosomal dominant hypercholesterolaemia in the Netherlands: prevalence, genotype-phenotype relationship, and clinical outcome. *Eur Heart J.* 2015;36(9):560-565.
5. World Health Organization. Human Genetics Programme: Familial hypercholesterolemia: report of a second WHO consultation. In:1999.
6. Umans-Eckenhausen MA, Defesche JC, Sijbrands EJ, Scheerder RL, Kastelein JJ. Review of first 5 years of screening for familial hypercholesterolaemia in the Netherlands. *Lancet.* 2001;357(9251):165-168.
7. Huijgen R, Kindt I, Verhoeven SB, et al. Two years after molecular diagnosis of familial hypercholesterolemia: majority on cholesterol-lowering treatment but a minority reaches treatment goal. *PLoS One.* 2010;5(2):e9220.
8. Versmissen J, Oosterveer DM, Yazdanpanah M, et al. Efficacy of statins in familial hypercholesterolaemia: a long term cohort study. *BMJ.* 2008;337:a2423.
9. Goldstein JL, Brown MS. Binding and degradation of low density lipoproteins by cultured human fibroblasts. Comparison of cells from a normal subject and from a patient with homozygous familial hypercholesterolemia. *J Biol Chem.* 1974;249(16):5153-5162.
10. Innerarity TL, Mahley RW, Weisgraber KH, et al. Familial defective apolipoprotein B-100: a mutation of apolipoprotein B that causes hypercholesterolemia. *J Lipid Res.* 1990;31(8):1337-1349.
11. Abifadel M, Varret M, Rabès JP, et al. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat Genet.* 2003;34(2):154-156.

12. Fouchier SW, Defesche JC, Umans-Eckenhausen MW, Kastelein JP. The molecular basis of familial hypercholesterolemia in The Netherlands. *Hum Genet.* 2001;109(6):602-615.
13. Fouchier SW, Kastelein JJ, Defesche JC. Update of the molecular basis of familial hypercholesterolemia in The Netherlands. *Hum Mutat.* 2005;26(6):550-556.
14. Varret M, Abifadel M, Rabès JP, Boileau C. Genetic heterogeneity of autosomal dominant hypercholesterolemia. *Clin Genet.* 2008;73(1):1-13.
15. Fouchier SW, Hutten BA, Defesche JC. Current novel-gene-finding strategy for autosomal-dominant hypercholesterolaemia needs refinement. *J Med Genet.* 2015;52(2):80-84.
16. Fouchier SW, Dallinga-Thie GM, Meijers JC, et al. Mutations in STAP1 are associated with autosomal dominant hypercholesterolemia. *Circ Res.* 2014;115(6):552-555.

CONTACT INFORMATION

Contact the author at:

Sigrid W. Fouchier
OCS Life Science
Ruwekampweg 2G
5222 AT 's-Hertogenbosch The Netherlands
Work Phone: +31 (0)73 523 6000
Fax:
Email: Sigrid.Fouchier@ocs-consulting.com
Web: