Introduction to Genetics and Its Value in Cardiovascular Disease Diagnosis: Basic Concepts and the Example of Familial Hypercholesterolemia

Fernando Civeira, José C. Rodríguez-Rey and Miguel Pocovi


cDepartamento de Bioquímica y Biología Molecular y Celular. Universidad de Zaragoza. I+CS. Zaragoza. Spain.

The main aim of this article is to provide clinicians with the basic knowledge needed to interpret data on genetic disorders that may be relevant to clinical decision-making. The human genome is likened to an encyclopaedia, “The encyclopaedia of life”, which is used to explain how genetic information is transmitted and to describe the causes and consequences of mutations, along with the main rules used for naming them. There follows a discussion of which mutations can contribute to the development of cardiovascular disease and how they do so. Finally, there is a description of how genetics has contributed to improvements in the diagnosis and treatment of familial hypercholesterolaemia, one of the conditions in which lipid-lowering therapy is efficacious and cost-effective. In addition, there is an explanation of how large-scale genetic assays, such as biochips, are used for detecting the principle mutations found in the genes responsible for familial hypercholesterolaemia.

Key words: Gene mutation. Genetics. Mutation nomenclature. Familial hypercholesterolemia.

THE HUMAN GENOME

When describing the human genome, whose sequence was first published in 2001, the most commonly used simile is that of an encyclopaedia. The “encyclopaedia of life” contains 44 volumes (two copies of each of the 22 chromosomes called autosomes) and two additional volumes, which are the sex chromosomes (XX for females and XY for males), for a total of 46 volumes.

The alphabet used in the human genome has only four letters, or bases: A adenine, G guanine (the purines), T thymine and C cytosine (the pyrimidines). The human genome (diploid genome) contains a total of 3 billion letters, and some of the letters are grouped into chapters (genes). As if it were a secret book, the chapters are not uniform. The zones with “meaning” (exons, which make up 1-2% of the entire encyclopaedia) are separated by

This study has been completed with the aid of the following projects: FIS.PI06/0365, PI06/1068, RTIC C06/01 (RECAVA), CIBERER and SA02005-07042.

Correspondence: Dr. M. Pocovi. Departamento de Biología Molecular y Celular. Facultad de Ciencias. Universidad de Zaragoza. P. Cerdana, 12. 50009 Zaragoza. Spain. E-mail: mpocovi@unizar.es


Introducción a la genética y su utilidad en el diagnóstico de las enfermedades cardiovasculares: conceptos básicos y el ejemplo de la hipercolesterolemia familiar

Nuestro objetivo es proporcionar al clínico los conocimientos básicos para la interpretación de datos referidos a enfermedades genéticas que le sirvan de ayuda a la hora de tomar decisiones clínicas. Se hace un símil del genoma humano con una enciclopedia, la «encyclopaedia de la vida», analizando como se transmite la información genética, las causas, las consecuencias y las principales reglas de nomenclatura de las mutaciones. Se discute a continuación qué mutaciones y cómo pueden contribuir a la aparición de la enfermedad cardiovascular. Por último, se describe la contribución de la genética a la mejora del diagnóstico y el tratamiento de la hipercolesterolemia familiar (HF), una de las entidades en que el tratamiento hipolipemiante es eficaz y coste-efectivo. Se justifica el empleo de técnicas de detección genética a gran escala como son los biochips que analizan las principales mutaciones de los genes que originan el HF.

zones with no meaning (introns) which make up about 25% of the total. Between chapters, there are large stretches (comprising 75% of the entire encyclopaedia) whose meaning is not known; these areas, or intergenic regions, are the most numerous.

The parts of the encyclopaedia’s chapters that have “meaning” (exons) are translated into another language made up of twenty different words (amino acids) that generate sentences (proteins). Words in this new language are the result of translating three-letter combinations from the original language (codons). Some three-letter combinations codify the same word, and for that reason the genetic code is said to be degenerate. There are letter combinations that indicate where phrases begin (ATG, the start codon) and three combinations (TGA, TAG and TAA) that mark where they end (stop codon).

At present, there is no complete index for this encyclopaedia; we do not know all of the genes, so we have only a partial index. Some chapters have not been located yet, and for that reason, they do not have a title. Like an encyclopaedia in which some chapter titles are missing, the genetic sequence contains regions that are very likely to be genes, but whose identity and function are not yet known. According to the Human Genome Project, our genome could contain some 30,000 genes that may trigger the synthesis of up to 100,000 different proteins.

TRANSMITTING GENETIC INFORMATION

Each chapter in our “encyclopaedia” is copied and transmitted from cell to cell and from generation to generation. When copies are made, errors (mutations) can occur in the DNA sequence; these include large and small insertions and deletions. These errors are transmitted in successive copies of the encyclopaedia (transmission to descendants). Although 99.9% of the human genome is identical in all individuals, there is a certain degree of variability resulting from the processes described above. Replacement of one letter, which we call a single nucleotide polymorphism (SNP), occurs approximately every 1,000 letters or base pairs (bp). Therefore, there are nearly 3 million SNPs in the entire genome that would explain most of the variations among individuals.

SNPs are the fundamental cause of the genetic variability that we observe among individuals. We can distinguish between various SNP types: random SNPs are those that are located in the “silent regions” of the genome and make up 90% of the total; gSNP (gene-SNP associations) that could influence gene control, total about one million, and cSNP (codon SNPs) located in coding regions and which often influence gene function. The SNP Consortium has recorded 1.7 million. Some of these changes (some 80,000 are recorded in the University of Cardiff’s database, http://www.hgmd.cf.ac.uk) have been linked with a disease, but the effect that most of them have is completely unknown and unpredictable.

TRANSCIBING GENETIC INFORMATION

Genomic deoxyribonucleic acid (DNA) has two sequence types: one carries information for protein synthesis, and the other, the internegic regions, do not lead to the synthesis of an active ribonucleic acid (RNA) or protein. The latter could carry out a wide variety of functions: structural functions in the chromosome, DNA packaging, organisation of chromatin within the nucleus, or even regulation of gene expression.

Genes in turn give rise to RNA transcription, a process involving the RNA polymerase enzyme. The synthesised RNA molecule may have its own functions (transfer [tRNA], ribosomal [rRNA] and microRNA), but in many cases, protein translation is necessary for it to be functional. As mentioned before, most structural genes in eukaryotic organisms are organised in coding regions that are interspersed with non-codifying regions called introns (fig. 1).

Transcription is the process by which an RNA molecule is synthesised based on a complementary DNA sequence. Gene transcription start and stop points are identified in the DNA template by a promoter sequence, which is located upstream from the gene (5’ end) and a terminator sequence downstream from the gene (3’ end). Promoter sequences contain conserved regions (called TATA, GC, CAAT boxes, etc.) to which proteins called
MOLECULAR MUTATIONS: CAUSES AND CONSEQUENCES

Interactions with chemical agents (chemical alterations of the DNA molecule as a result of hydrolysis, oxidation or methylation) or physical agents (ultraviolet light has the ability to form photoproducts, such as pyrimidine dimers between adjacent pyrimidine bases) can cause changes in DNA, but most of these changes occur spontaneously during cellular processes, such as those occurring during DNA replication or repair\(^5\). In many cases, these changes in DNA create base mispairing and give rise to point mutations. However, larger molecular changes may also occur, such as deletions and insertions of various sizes. Deletions and insertions occur when a fragment of DNA is eliminated or inserted in a specific location on the genome.

When a small deletion or insertion occurs in a coding region, the result is a mutated protein. The codon reading frame can be maintained in some cases (in-frame deletion or insertion), but the reading frame may also be modified (frame shift deletion or insertion), and in this case, the protein transcription factors bind. These proteins facilitate or prevent initiation of the transcription process by RNA polymerase II (fig. 2). In addition to these promoter sequences, other regulatory sequences located in the 5’ region of the promoter permit or facilitate the binding of molecules regulating gene transcription. Some genes also possess enhancer and silencer sequences that can be located in either the 5’ or 3’ region of the gene, or even in a region located within the gene in certain cases.

Once RNA has been synthesised, its non-codifying sequences (introns) are removed during the splicing process thanks to the presence of conserved sequences\(^5\). These steps lead to the formation of mRNA, which is made into protein inside ribosomes (fig. 3).

Fig. 1. In the transcription process, RNA polymerase II makes a copy of the RNA corresponding to the gene. Introns are removed from the immature RNA transcript by a process called splicing, resulting in a molecule of messenger RNA (mRNA) which is subsequently translated into a protein molecule in the ribosomes.

Fig. 2. Transcription factors (TF), whether directly or through coactivators (CoAct), stabilise the union of the preinitiation complex (PIC), which includes RNA polymerase II (POLII), to the gene promoter. Stabilisation causes RNA synthesis to begin.
"polymorphism" is used both when describing a change that does not cause illness and when indicating a change that is found in ≥ 1% of the population. In order to avoid confusion, it is best not to use the terms "mutation" and "polymorphism" (including SNP) and use instead terms such as "sequence variation", "alteration" or "allelic variant".7,8

Depending on the type of substituted nucleotide, point mutations are as follows: a) transition, which consists in replacing a pyrimidine for another or a purine for another, meaning a G-C pair would be replaced by an A-T pair or vice versa, and b) transversion, when a purine is replaced with a pyrimidine or vice versa, meaning an A-T pair would be replaced by a T-A or C-G pair.

When naming mutations, it is recommended that we use a series of standards created by the Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC), which are listed on their web page (www.hgvs.org/mutnomen/recs). We include a brief summary below:

- The most important rule is that variants must be described at the most basic level possible, that is, at DNA level. Descriptions should always be in relation to a reference sequence, either a genomic or a coding DNA reference sequence.
- For describing genes/proteins, only official HGNC symbols should be used (available at www.genenames.org/guidelines.html). When descriptions
at RNA or protein level are given in the text, on first mention, a format with the DNA first and then the protein or RNA in parentheses should be used. Example: c.48G>C (p.Trp16Cys).

– When several changes are described, a tabular listing should be provided with the findings, using separate columns for DNA, RNA and protein and indicating whether the changes were experimentally determined or theoretically deduced.

– A reference to a nucleotide must be preceded by “g” for genomic DNA, “c” for a coding DNA sequence (cDNA), “m” for mitochondrial DNA, “r” for a ribosomal RNA sequence, and “p” for a protein (all letters in lower case). To avoid possible confusion between the nucleotide number and the g or c etc., use a full stop. For example, g.576 refers to the nucleotide located in genomic DNA position 576.

**Description at DNA level**

When numbering nucleotides in the DNA reference sequence, the A in the start codon (ATG) is designated as nucleotide +1. There is no nucleotide 0. Nucleotides located upstream from +1 should be preceded by a minus sign: –1, –2, –3, etc. Nucleotides located downstream from +1 do not carry a sign.

It is best if the reference sequence used for DNA is the one described in the RefSeq database, available at www.ncbi.nlm.nih.gov/RefSeq/

For genomic reference sequences: a) nucleotide numbering is purely arbitrary and starts with 1 at the first nucleotide of the reference sequence database. No +, – or other signs are used. When the complete genomic sequence is not known, a coding DNA reference sequence should be used.

For describing specific changes:

– The symbol > indicates a substitution at DNA level: for example, c.75A>T.

– The symbol _ indicates a change affecting various residues which separates the first and last of the affected residues: for example, c.77_79delACT.

– A deletion is indicated with “del” : for example, c.77delA.

– A duplication is indicated with “dup”: for example, c.dupA.

– An insertion is indicated with “ins”: for example, c.75_76insG.

Replacements of nucleotides must begin with the nucleotide number followed by the change that occurs at that position. For example, 576G>C means that the nucleotide in position 576, which is G in the reference sequence, has been replaced by C.

When describing two sequence variations in the same individual:

– Two sequence changes in different alleles are shown between square brackets and separated by the + sign: for example, c.[76A>C]+[87delG].

– Two sequence variations in the same allele are shown between square brackets separated by a semicolon: for example, c.[76A>C; 83G>C].

– Two sequence variations with unknown alleles are shown between square brackets separated by (+): for example, c.[76A>C(+)]83G>C].

To designate the location of a variation in an intron, the nucleotide numbering for cDNA can be used. For example, c.1232+1G>C indicates that G is replaced by C in the first nucleotide of the intron corresponding to position 1232 in cDNA. When the complete genomic DNA sequence is not available for designate mutations occurring within introns, we add the prefix IVS followed by the intron number and a positive number, which corresponds to the nucleotide position beginning at the G of the donor splice site (GT) or a negative number, which corresponds to the nucleotide position beginning at the G of the splice acceptor site (AG). Therefore, for example, IVS4+1G>C refers to a substitution of G for C at nucleotide +1 of intron 4.

**Descriptions at protein level**

Descriptions of variations at protein level should reflect the changes observed at protein level and not try to incorporate any knowledge relative to changes at DNA-level.

The three-letter code must be the preferred code for describing amino acids (table I).

The description of reading-frameshift should not include the deletion from the site of the frame shift to the natural C-terminal end (stop codon) of the protein, so p.Arg97ProfsX23 and not p.Arg97_Pro109delfsX23. Similarly, for frame shifting insertions the inserted amino acid residues are not described; only the total length of the new shifted frame is given. For example, p.Glu5ValfsX5 and not p.Glu5Valins2fsX3.

For the nomenclature of mutations based on amino acids, the methionine start codon is numbered as +1.

The protein reference sequences should represent the primary translation product, not a processed protein, and thus include any signal peptide sequences.

Amino acids originating from changes introducing upstream translation initiation are numbered like nucleotides: for example, Ser-3, Thr-2.

Amino acids originating from changes resulting in translation of intronic sequences are numbered like nucleotides: for example, Phe5-3, Gln5-2, Cys4+1.
Amino acids originating from codon stop disappearance causing translation to extend are numbered Arg*1, Ser*2, etc.

“X” is used to designate a translation termination codon.

To describe a substitution, the wild type amino acid (the one with the highest frequency in the population) is named first and the mutant second: for example, Arg370Ser.

**CARDIOVASCULAR DISEASE AS A COMPLEX GENETIC DISORDER. IDENTIFICATION OF GENETIC VARIANTS THAT CAUSE IT**

In monogenic diseases, almost all bearers of an allele have the disease (this is known as a high degree of penetrance), which makes it easier to identify the genetic variant that causes it. Such is not the case for cardiovascular disease (CD). It is true that monogenic traits such as familial hypercholesterolaemia exist, (see following section), but in general, the disease does not follow a typical Mendelian pattern. Some authors point out that at least a hundred genes could be part of the process9, and their effect would depend not just on certain environmental conditions, but also on a characteristic pattern for common diseases.

The common disease-common variant hypothesis suggests that variants that cause common diseases have a high frequency in the general population10. Although this is the most widely-supported hypothesis, a significant body of scientists currently supports an alternative that also includes combinations of rare variants11. Both hypotheses, however, suggest that the isolated effect of each one of the variants must be quite small, whereas the combination of many pathogenic variants is what would increase the probability of developing the disease. Consequently, the population would be a continuum with the bearers of the most pathogenic variants at one end, suffering the disease even while maintaining a healthy lifestyle. The other end would include the bearers of healthy variants who are able to avoid developing the disease even if their lifestyle is far from being recommendable. Therefore, identification of variants that cause the disease could play an important predictive role in the future.

As stated previously, the genome contains millions of variants, out of which only a few hundred could be implicated in the development of CD. Large studies of the complete genome permit the simultaneous analysis of more than half a million genetic markers distributed throughout the genome in groups of individuals who have the disease and in control groups for comparison. Although these large-scale studies have revealed important associations12, analysing so many variants comes with large statistical limitations: only very strong associations (with values of \( p \approx 10^{-8} \)) are accepted as statistically significant in the analysis13.

A complementary strategy is the one for candidate genes in which there is a search for variants associated with certain genes. If, for example, it is known that cardiovascular risk increases in the presence of high cholesterol or high blood pressure, the genes that code proteins playing a well-known role in these processes would be an obvious topic of study. Logically, selecting which genes to study requires prior knowledge of the role they play and/or the aetiology of the disease.

**VARIANTS IN REGULATORY REGIONS AS IDEAL CANDIDATES**

Variants situated in coding regions normally cause serious effects and do not fit in with the current model of common diseases. For this reason, researchers’ attention is drawn toward the variants in the regulatory regions. There are two reasons why this type of variant is ideal for explaining common diseases. The regulation of gene expression is fundamentally carried out when transcription begins. This is when a multi-protein complex
The example of familial hypercholesterolaemia

FH is an autosomal co-dominant disease which is characterised by very high blood concentrations of low density lipoprotein cholesterol (LDL), increased risk of premature coronary heart disease and extravascular deposits of lipids such as tendon xanthomas and corneal arch in the first decades of life. FH is caused by mutations in the gene that codifies the cellular LDL receptor (LDLR). Its natural ligand is apolipoprotein B (apoB), which is the main protein for LDLs. FH is a frequent disease, given that in most countries FH prevalence is approximately 1/500 people in its heterozygous form; based on that number, we calculate that there are nearly 100,000 affected people in Spain.

FH is the most frequent (70%) and best-known member of the group of autosomal dominant hypercholesterolaemias (ADH) that are characterised by a family-related presentation of hypercholesterolaemia whose members show a clear bimodal pattern in LDL concentrations. Other less frequent causes (5%) of ADH include: a) mutations in the apoB gene, resulting in an ADH called familial defective apoB-100; b) PCSK9 gain-of-function mutations (< 1%) and protease involved in LDLR degradation, called HF3, and c) hyperlipoproteinaemia(a), which causes AHD in approximately 2-3% of cases. In approximately 30% of AHDs the genetic defect that caused them is unknown, although for some, an increase in intestinal sterol absorption could play an important pathogenic role. The lipid and cardiovascular phenotypes for AHDs that depend on LDLR and apoB are indistinguishable from each other, and more serious than other AHDs, which is why their clinical treatment is usually handled in a joint manner and called FH.

FH is a paradigmatic example of in vivo early onset atherosclerosis dependent on LDL, and one of the disorders for which lipid-lowering treatment is the most effective and cost-efficient. The prevalence of coronary heart disease in heterozygous FH subjects before statin drugs appeared was about 50% and 20% respectively in males and females younger than 50. Currently, all series show a significant drop in cardiovascular morbidity and mortality due to a much more extensive use of statins in these patients, whether as monotherapy or in combination with other lipid-lowering drugs such as ezetimibe.

Clinical diagnosis of FH

Since we do have a very effective treatment, early diagnosis of FH is a social health priority, as indicated by the WHO. The diagnosis of FH has traditionally been performed according to clinical criteria based on a family and personal history of early onset coronary heart disease, LDL concentrations in blood and the presence of cholesterol deposits in the corneal arch and in tendon xanthomas. However, clinical diagnosis has its problems: total cholesterol and LDL concentrations are not sensitive enough to identify affected individuals in the general population or even in a family with diagnosed FH. The overlap between affected and non-affected individuals is particularly significant in younger patients, of whom about 20% have LDL figures that could lead to a wrong diagnosis. Furthermore, corneal arch and xanthomas are very rare in heterozygotes in the first decades of life, so they are not helpful as diagnostic signs, particularly in younger patients. For these reasons, the clinical criteria used at the moment show little correlation with the genetic diagnosis. Recently, we assessed the
value of the three most frequently-used criteria for clinically diagnosing FH, and their sensitivity and specificity is about 70%\textsuperscript{19}. This is why we recommend the genetic diagnosis for those cases described in table 2.

**Genetic diagnosis of FH**

We know more than a thousand different LDLR mutations and at least three apoB mutations that cause FH, although the CGG/CAG mutation in codon 3500, by which glutamine replaces arginine (R3500Q) is the most frequent cause of familial defective apoB. This considerable molecular heterogeneity is constant in most populations, including Spain, where there are more than 250 known mutations. Defects are varied, ranging from point mutations or small deletions or insertions (< 20 bp) to substantial rearrangements that affect a large part of the LDLR. Only in certain populations with a high degree of genetic isolation are there founder effect mutations resulting in very few mutations that cause FH and a higher frequency of the disease than in other populations. This occurs in the French Canadian population, Lebanese Christians, the Finnish population, the Ashkenazi Jews or the Caucasian population in South Africa\textsuperscript{14,16}.

The confirmation of a functional mutation in the LDLR or apoB genes is preferred for FH because it provides a sure diagnosis. The usefulness of genetic diagnosis has been confirmed by many studies. It provides unequivocal identification of patients with a high level of cardiovascular risk at an early stage, facilitates genetic counselling, stratifies the prognosis more precisely according to the mutation type\textsuperscript{28}, aids in the search for affected family members, and makes indicating treatment easier for the doctor and complying easier for the patient\textsuperscript{29}. In our country, in many cases, it permits patients to purchase statin drugs at a reduced price.

Genetic analysis cannot currently be used across large segments of the population because of its cost, availability and complexity. Therefore, the principal recommendation offered by international guides to the clinical management of FH is to use genetic diagnosis in the following situations\textsuperscript{20}.

1. Populations in which a few mutations cause most of the cases of FH.
2. Populations in which most of the mutations are known and rapid genetic diagnosis tools are available.
3. Individuals with abnormal lipid phenotypes in families with a known FH-causing mutation.

**Tools for genetic diagnosis of FH**

The fact that at least three genes and no fewer than several hundred mutations can cause FH in most individuals obliges us to use large-scale genetic detection techniques. Essentially, there are two procedures: a) complete LDLR sequencing, and b) the use of DNA microarrays or biochips. The LDLR gene is a large one, so its sequencing and analysis are costly and labour-intensive. Microarrays are a technology consisting of a series of deposited DNA probes that include the mutations that cause FH in a certain population. Hybridising the DNA to be studied allows mutations to be identified quickly and with a sensitivity and specificity > 99%. Its principal drawback is that it requires a prior knowledge of the FH mutations in the population to be tested. Fortunately, a labour-intensive analysis of LDLR and apoB genes in FH has been carried out in Spain\textsuperscript{15}. By means of the analysis, a microarray has been optimised that contains 234 mutations for LDLR, the

---

**TABLE 2. Indications for genetic diagnosis in suspected FH*\**

<table>
<thead>
<tr>
<th>Patients with high cLDL and normal triglycerides</th>
<th>Personal or family history of tendon xanthomas</th>
</tr>
</thead>
<tbody>
<tr>
<td>If cLDL &gt; 160mg/dl</td>
<td></td>
</tr>
<tr>
<td>No personal or family history of tendon xanthomas</td>
<td>Age 18-30 if cLDL &gt; 220mg/dl</td>
</tr>
<tr>
<td>Age 30-39 if cLDL &gt; 225mg/dl</td>
<td>Age ≥ 40 if cLDL &gt; 235mg/dl</td>
</tr>
</tbody>
</table>

Patients with mixed hyperlipaemia (high total cholesterol and triglycerides > 200mg/dl)

Total cholesterol > 335mg/dl
or
Apolipoprotein B > 185mg/dl

\*cLDL: low density lipoprotein cholesterol; FH: familial hypercholesterolaemia.

*In families with clinically suspected FH due to bimodal cholesterol distribution in the family and an absence of secondary causes.
mutations for apoB, and the main mutations causing hypercholesterolaemia in PCSK9. This microarray, named Lipochip, is a first in the world for diagnosing a genetic disease. Its use has become widespread in Spain, and it is being developed in other countries with known genetic bases for FH such as the Netherlands, Italy, and the United States.

CONCLUSIONS

Comparing the human genome to an encyclopaedia is a way of understanding its structure and function. Genetic mutations must be named according to nomenclature rules accepted by the scientific community. Cardiovascular disease can be considered a complex genetic disease that results from both genetic and environmental factors. Familial hypercholesterolaemia is an autosomal codominant disease and a paradigmatic example of early atherosclerosis. Its diagnosis is a social health priority, since the cardiovascular risk to affected individuals can be reduced with effective treatments.

Conflict of interests

The authors declare they have no conflict of interests.

REFERENCES
