

# IDENTIFICATION OF GENETIC MODIFIERS OF PENETRANCE IN HUMAN INHERITED DISEASE

Pau Puigdevall Costa

---

TESI DOCTORAL UPF / ANY 2018

DIRECTOR DE LA TESI

Dr. Robert Castelo Valdueza

Functional Genomics Group

Research Programme on Biomedical Informatics (GRIB)

DEPARTMENT OF EXPERIMENTAL AND HEALTH  
SCIENCES





A tot allò que som i podem ser encara.



## Acknowledgments

Arribats a aquest punt, un s'adona de la dificultat d'agrair a totes les persones que d'una forma directa o indirecta m'han ajudat a que aquesta tesi sigui una realitat. Estic convençut que sense vosaltres, o bé no hauria comès la imprudència de fer un doctorat o bé ho hauria deixat estar fa temps. Però per tossuderia heu-nos aquí! I durant aquests anys me n'ha fet falta i molta. Per això, ja demano per avançat disculpes si em deixo algú que hi hauria de ser.

En primer lloc, mil gràcies Robert per oferir-me l'oportunitat de fer el projecte de màster i aquest doctorat en un juny de fa ja més de 4 anys. Tinc un munt de coses a agrair-te: la passió docent, l'atreviment, la curiositat, el treballar fort, les formes, la comprensió i el temps invertit en mi en el dia a dia. Sempre he pensat que un anunci que es va emetre ja fa temps sobre la feina ben feta t'escau a la perfecció. Gràcies també pels recursos que m'has posat a disposició, en especial totes les oportunitats formatives en què he pogut assistir i totes les idees, suggeriments i correccions per aprendre i millorar.

Vull fer extensiu aquest agraïment als pares, un pilar bàsic per poder tirar endavant setmana rere setmana, especialment quan els dubtes apareixen. També pel xec en blanc de deixar-me escollir el meu camí, fos o no fos l'encertat. I als avis, pel seu suport incondicional i per cuidar-me amb tota l'estima.

Gràcies a les doctores de l'Hospital Clínic, l'Irene, la Cèlia, i també la Lucilla, per la dedicació en el projecte de la hipertensió pulmonar arterial. No em puc ni imaginar la d'hores que us heu passat repassant historials clínics, trucant a pacients, revisant informació i anant amunt i avall entre Clínic i PRBB. A més, heu sabut trobar finançament de sota les pedres per poder tirar endavant un projecte ambiciós i complex, i ens heu ajudat a fer-lo millor cada vegada que ens heu o us hem visitat.

I also want to thank Dan for providing us the genetic linkage “know-how” to face the study of incomplete penetrance. Your expertise has translated into great ideas that have enriched the project and have overcome the challenges that could easily arise without your support. And all of that in only one month. Hoping that your obligations as a dean do not stop you from enjoying your long bike trips!

A vosaltres també, Gerard i Rosa, per ajudar-me tant a canvi de res en els meus inicis en la bioinformàtica, per mantenir la pinya durant tots aquests anys i per fer-nos teràpia predoctoral de tant en tant. Que important és compartir èxits i fracassos, il·lusions i decepcions amb gent amb tan talent com vosaltres.

I què dir de la tribu del despatx 486? Estic encantat d'haver estat tan ben adoptat per vosaltres, tot i estar sempre en minoria. Estic segur que aquest bon rollo dins i fora del PRBB ens han fet millors a tots. Gràcies als companys de grup (Joan, Aina, Dani, Lijie, Eduardo), als veïns de Goedel (Emma, JC, Bea i Àngel), a tots aquells amb qui he

compartit una bona colla d'anys o mesos (Marina, Adrià, José Luis, Héctor, Babita, Carlota, Sabari, Sheena, Joan, Ivan, Joel, Gaurav, Bea, Audald) i especialment al trio calavera amb qui hi he estat de principi a fi (Will, Juanlu, Jorge). Espero que quan ja no sigui al PRBB, seguiu fent un equip de vòlei platja guanyador !

Igualment, vull agrair el suport imprescindible de la gent d'IT, l'Alfons i en Miguel, per la quantitat ingent de demandes ateses per programes, errors, instal·lacions i mil vicissituds més, entre elles l'apoderament temporal i no tan temporal del node "bigmem".

Als companys de pis, l'Alba i en Carles, per les millors tertúlies de bar que un pot tenir a casa, però sobretot per compartir amb mi les braves del Venezia, tot i saber que sempre hi sortiu perdent. També a l'Anna, per compartir a Barcelona el nostre König de Girona.

I als amics d'Olot i als companys d'equip de futbol. Uns especialistes en preguntar recurrentment si això del doctorat era una excusa per seguir "estudiant" en lloc de treballar. Sense dubte, heu estat uns excel·lents motivadors per no adormir-me ni un cap de setmana dels que he vingut.

I per acabar, gràcies també a tot el personal mèdic que m'ha tractat de forma exquisida, en especial les infermeres de l'ICO del Trueta.

Gràcies a tots, de tot cor.





## Abstract

Understanding the molecular basis of human inherited disease is a challenging task due to the complexity of mapping the genotype to the phenotype. The study of diseases running in families has shown that the inheritance of certain variants confers disease susceptibility. However, not all carriers of such pathogenic variants express the clinical symptoms, a phenomenon known as reduced penetrance. This thesis has investigated a potential mechanism of genetic modification for the penetrance of a missense *BMPR2* mutation in heritable pulmonary arterial hypertension (HPAH) by applying genetic linkage analysis to a large multiplex family. We have identified a candidate region for a modifier in the distal promoter region of the *FIGN* gene supported by lung-specific regulatory activity and risk factors associated with blood pressure. Taken together, these results suggest that common regulatory variants may have an important role in determining the penetrance of pathogenic coding variants. This thesis also provides a resource to integrate genome-wide position-specific scores routinely used in variant prioritization.



## Resum

Comprendre la base molecular de les malalties humanes hereditàries és un gran repte degut a la complexitat de les relacions entre genotip i fenotip. L'estudi de malalties de transmissió familiar ha demostrat que el fet d'heretar certes variants genètiques confereix una major susceptibilitat de patir aquestes malalties. No obstant això, no tots els portadors d'aquestes variants patogèniques expressen la simptomatologia clínica, un fenomen anomenat penetrància reduïda. Aquesta tesi s'ha centrat en investigar un possible mecanisme de modificació genètica de la penetrància d'una mutació del gen *BMPR2* en hipertensió arterial pulmonar hereditària. A tal efecte, hem utilitzat l'anàlisi per lligament genètic aplicat a una gran família multigeneracional amb múltiples portadors de la mutació, una fracció dels quals estan afectats per la malaltia. Hem identificat una regió candidata pel modificador en una regió promotora distant del gen *FIGN*, amb evidència d'activitat reguladora específica al pulmó i amb factors de risc associats a pressió sanguínia. Analitzats conjuntament, aquests resultats suggereixen que les variants reguladores comunes poden tenir un paper important en la penetrància de les variants patogèniques i codificants. Aquesta tesi també proporciona un recurs per integrar les puntuacions genòmiques específiques per posició, habitualment utilitzades per a la priorització de variants genètiques.



## **Preface**

Taking one-hour walk around my home town Olot, in the northeast of Catalonia, is enough for any curious mind to gain interest in natural science. This region is indeed an European hotspot for biodiversity, result of being a volcanic region and a transition territory for Pyrenean, Mediterranean and continental ecosystems. If you combine this with a particularly rainy weather and a mountainous territory, it is easy to understand the local devotion to meteorology and biology. And as a child, I was not an exception to such interest. I still keep the notebook where I used to write down the daily measurements from a precarious weather station. Later, as a teenager, I also became interested in robotics and with time, I organized a summer educational robotic course for kids that would become my first salary. Because of this, I probably end up studying biotechnology, which mixes a bit of the two worlds. Without the driving force of such curiosity, this thesis would not be a reality.

My first contact with human genetics was the reading of a newspaper article forecasting the future achievements that the completion of the Human Genome Project would entail. The article offered an overly optimistic view about the immediate endless applications to healthcare. A remote memory from this article came to my mind when after several years battling against a supposed Graves' disease, I was diagnosed with pituitary resistance to thyroid hormone, a rare genetic disease with mild consequences under proper treatment. By that time, my endocrinologist encouraged me to participate in a research project to identify the underlying genetic cause, but it was

suddenly cancelled due to the budget cuts on the healthcare system. Although this study is still pending, it undeniably raised my interest towards the understanding of the genomic architecture of diseases.

Learning bioinformatics was another step to feed such vital interest, as this discipline has revolutionized the way in which data is analysed and how genetic counselling is practiced. The advent of new genetic tools also shows that the current knowledge on human genetic disease is far from complete and many questions need still to be addressed. The big challenge ahead encouraged me to focus this thesis on reduced penetrance, a phenomenon that has been known for long, it has a major impact in human genetics, but it is still poorly understood. The project was ambitious and risky, but this is what basic science should be all about, pushing the boundaries of knowledge, even when you are unsure to go in the right direction. The findings reported in this thesis are a modest contribution for understanding reduced penetrance, as they have not yet been followed up by functional in vivo confirmatory assays. Still, in words of the Dutch painter Vincent Van Gogh “great things are done by a series of small things brought together”. And the different pieces are being collected right now.

Pau Puigdevall Costa

Barcelona, 1<sup>st</sup> October 2018.

## Table of contents

<i>Acknowledgments</i> .....	v
<i>Abstract</i> .....	ix
<i>Resum</i> .....	xi
<i>Preface</i> .....	xiii
<i>Table of contents</i> .....	xv
<i>List of figures (from Introduction)</i> .....	xix
<i>List of tables (from Introduction)</i> .....	xxi
<i>Abbreviations</i> .....	xxiii
<b>1. INTRODUCTION</b> .....	<b>1</b>
<i>1.1. Historical review of genetics and genomics</i> .....	1
1.1.1. Genetic disorders: from Ancient Greece to Mendel.....	1
1.1.2. Mendelian theory of inheritance.....	3
1.1.3. The dawn of molecular genetics .....	5
1.1.4. The genomics era.....	8
1.1.5. The Human Genome Project .....	10
1.1.6. The sequencing revolution.....	11
1.1.7. Genome editing .....	15
<i>1.2. Human genetic variation</i> .....	18
1.2.1. Origin and types of genetic variation .....	18
1.2.2. Building a catalogue of human genetic variation .....	21
<i>1.3. Functional genomics</i> .....	27
1.3.1. The Encyclopaedia of DNA elements (ENCODE) .....	27
1.3.2. The Genotype-Tissue Expression (GTEx) project .....	32
1.3.3. The NIH RoadMap Epigenomics Mapping Consortium .....	35
<i>1.4. Clinical interpretation of human genetic variation</i> .....	40
1.4.1. The variant annotation process .....	40
1.4.2. Variant localization and nomenclature .....	41
1.4.3. Position-based variant annotation.....	43
1.4.4. Other annotations at variant, gene and pathway level ..	45
1.4.5. Variant filtration and prioritization.....	50
<i>1.5. Human disease</i> .....	55

1.5.1. Disease definition and classification .....	55
1.5.2. Genetic disease .....	58
1.5.3. Inherited disease .....	60
1.5.3.1. Monogenic disease .....	61
1.5.3.2. Complex disease .....	66
1.5.3.3. New perspectives on the genetic architecture of disease.....	72
<i>1.6. Reduced penetrance in human disease</i> .....	76
1.6.1. Penetrance as a path from genotype to phenotype .....	76
1.6.2. The molecular basis of reduced penetrance.....	82
1.6.2.1. Mutation type.....	83
1.6.2.2. Modifier effects .....	85
1.6.2.3. Gene expression level.....	90
1.6.2.4. Allele dosage .....	91
1.6.2.5. Copy number variation .....	91
1.6.2.6. Sex .....	92
1.6.2.7. Age of onset.....	93
1.6.2.8. Epigenetic modification.....	94
<i>1.7. Genetic linkage analysis</i> .....	95
1.7.1. Genetic linkage and recombination .....	96
1.7.2. The fundamental aspects of linkage analysis .....	99
1.7.3. Types of genetic linkage analysis .....	106
1.7.3.1. Parametric linkage .....	106
1.7.3.2. Nonparametric linkage .....	108
1.7.4. Modelling reduced penetrance in parametric linkage.	109
<b>2. OBJECTIVES</b> .....	<b>115</b>
<b>3. RESULTS</b> .....	<b>117</b>
<i>3.1. Genetic linkage analysis of a large family identifies FIGN as     a candidate modulator of reduced penetrance in heritable     pulmonary arterial hypertension</i> .....	<i>117</i>
3.1.1. Abstract.....	118
3.1.2. Introduction .....	119
3.1.3. Methods .....	121
3.1.4. Results .....	130
3.1.5. Discussion.....	143
3.1.6. Acknowledgments .....	145
3.1.7. Contributors .....	145
3.1.8. Funding and competing interests.....	145



3.1.9. Ethics approval and data sharing statement.....	146
3.1.10. Web resources.....	146
3.2. <i>GenomicScores: seamless access to genomewide position-specific scores from R and Bioconductor</i> .....	147
3.2.1. Introduction .....	148
3.2.2. Results .....	149
3.2.3. Conclusions .....	155
3.2.4. Acknowledgments .....	156
3.2.5. Funding.....	156
<b>4. DISCUSSION.....</b>	<b>157</b>
4.1. <i>A pipeline to study reduced penetrance</i> .....	157
<b>5. CONCLUSIONS.....</b>	<b>171</b>
<b>6. ANNEX.....</b>	<b>173</b>
6.1. <i>Supplemental Material from Chapter 3.2</i> .....	173
6.1.1. Supplemental Methods .....	173
6.1.2. Supplemental Figures .....	175
6.1.3. Supplemental Tables.....	197
<b>7. REFERENCES .....</b>	<b>203</b>



## List of figures (from Introduction)

<b>Figure 1.</b> “Homunculus” representation (left) and hardcover of “De Generatione Animalum” from Aristotles (right). .....	3
<b>Figure 2.</b> Two-generation crosses of <i>Drosophila melanogaster</i> .....	6
<b>Figure 3.</b> The central dogma of molecular biology (Crick, 1956)...	7
<b>Figure 4.</b> Average annual sequencing cost per whole genome, in last two decades. ....	12
<b>Figure 5.</b> Transition of genomic research from the study of genome biology to effectiveness on clinics.....	14
<b>Figure 6.</b> CRISPR/Cas9 is part of the bacterial immune system against viruses. ....	17
<b>Figure 7.</b> Possible outcomes of DNA repair. An unrepaired mistake in DNA synthesis lead to point mutations. ....	19
<b>Figure 8.</b> Size distribution and novelty of variants discovered in the pilot phase of 1000 Human Genome Project.....	23
<b>Figure 9.</b> Fraction of novel variants in each allele frequency.....	24
<b>Figure 10.</b> Population structures of recent large-sequencing projects and novelty proportion from ExAC alleles.....	25
<b>Figure 11.</b> ENCODE functional and regulatory elements.....	29
<b>Figure 12.</b> Representation of the 44 tissues and cell lines included in the GTEx release v6p. ....	34
<b>Figure 13.</b> Illustration of the sampling protocol and the techniques used in the production of the reference epigenome from the RoadMap Epigenomics Project .....	37
<b>Figure 14.</b> Flowchart of a possible variant filtration and prioritization pipeline. ....	52
<b>Figure 15.</b> Genetic and environmental contributions to monogenic (A) and complex (B) disorders .....	60

<b>Figure 16.</b> Modes of inheritance in Mendelian disease. ....	63
<b>Figure 17.</b> The quantitative perspective of complex disease from phenotypic and genotypic viewpoints. ....	67
<b>Figure 18.</b> Heatmap and multi-generation pedigree showing the relationship among the age and frequency of variants with their effect size in disease. ....	72
<b>Figure 19.</b> Allelic spectrum of human inherited disease. ....	74
<b>Figure 20.</b> Large GWAS reveals different genomic architectures for different complex diseases and complex traits. ....	75
<b>Figure 21.</b> Complete and reduced penetrance in diseases with recessive and dominant mode of inheritance. ....	78
<b>Figure 22.</b> Clinical expressivity overlaps the trait distribution, being the disease only penetrant above a certain threshold of abnormal function. ....	81
<b>Figure 23.</b> Mechanisms of reduced penetrance. ....	84
<b>Figure 24.</b> The cis-regulation of a gene can modulate the penetrance of pathogenic mutations. ....	86
<b>Figure 25.</b> Models of modification for penetrance (a), dominance (b) and expressivity (c). ....	88
<b>Figure 26.</b> Age-dependent penetrance curve for Huntington's disease with a cohort of carriers with the 36-39 CAG triplet repeats in IT15 gene. ....	94
<b>Figure 27.</b> Model of single crossover in meiosis and computation of the recombination frequency ( $\theta$ ). ....	97
<b>Figure 28.</b> Conceptual core of IBD-based genetic linkage analysis. ....	99
<b>Figure 29.</b> Example of the application of the LOD scores method in a pedigree of two-generations. ....	103

## List of tables (from Introduction)

<b>Table 1.</b> Class and types of DNA mutation. ....	20
<b>Table 2.</b> Example of genome-wide position scores used in annotation. ....	47
<b>Table 3.</b> Heritability estimates of traits and disease-related phenotypes from twin studies .....	59
<b>Table 4.</b> Examples of the variability on rare disease prevalence... ..	64
<b>Table 5.</b> Summary of the main existing algorithms and their features.....	104
<b>Table 6.</b> Narrow-sense penetrance example. ....	110
<b>Table 7.</b> Wide-sense penetrance (recessive example).....	110
<b>Table 8.</b> Wide-sense penetrance (dominant example). ....	111
<b>Table 9.</b> Narrow-sense penetrance example with phenocopies... ..	112
<b>Table 10.</b> Age liability classes for the Charcot Marie-Tooth disease. ....	113



## Abbreviations

**CADD:** Combined Annotation Dependent Depletion scores

**ChIP-Seq:** Chromatin Immunoprecipitation Sequencing

**CRS:** Candidate Regulatory SNPs

**dbGAP:** NCBI Database of Genotypes and Phenotypes

**dbSNP:** NCBI database for short genetic variation

**dbVAR:** NCBI database for large scale genomic variants

**EGA:** European Genome-phenome Archive

**EFO:** Experimental Factor Ontology

**ENCODE:** The Encyclopedia of DNA Elements

**eQTLs:** Expression Quantitative Trait Loci

**ExAC:** Exome Aggregation Consortium

**gnomAD:** Genome Aggregation Database

**GTEx:** Genotype-Tissue Expression Project

**GWAS:** Genome-wide association study

**HGMD:** Human Gene Mutation Database

**(H)PAH:** (Heritable) Pulmonary Arterial Hypertension

**IBD:** Identity by descent

**IBS:** Identity by state

**LD:** Linkage Disequilibrium

**lincRNA:** Long intergenic non-coding RNA

**LoF:** Loss-of-function

**MAF:** Minor Allele Frequency

**M-CAP:** Mendelian Clinical Applicable Pathogenicity scores

**MOI:** Mode Of Inheritance

**mPAP:** mean Pulmonary Arterial Pressure

**NGS:** Next Generation Sequencing  
**NIH:** National Institutes of Health  
**NPL:** Non-parametric linkage  
**OMIM:** Online Mendelian Inheritance in Man database  
**PAR:** Pseudoautosomal Region  
**phastCons:** Phylogenetic Analysis with Space/Time models-  
Conservation scores  
**phyloP:** Phylogenetic p-values  
**REMC:** NIH RoadMap Epigenomics Mapping Consortium  
**SNP:** Single Nucleotide Polymorphism  
**TAD:** Topological Associating Domain  
**TF:** Transcription Factor  
**TOPmed:** Trans-Omics for Precision Medicine  
**WGS:** Whole Genome Sequencing







# 1. INTRODUCTION

## 1.1. Historical review of genetics and genomics

### 1.1.1. Genetic disorders: from Ancient Greece to Mendel

What are genetic disorders? How a genetic condition can be inherited? If a genetic disorder runs in my family, what are the chances that my offspring and I develop the condition? It is highly likely that the reader has heard these questions or personally formulated them to a doctor at least once in life. This need for responses regarding disease it is not exclusive from our time.

Genetic inheritance has been traditionally a main area of scientific and philosophical study, fascination and concern over centuries. The “likeliness” among related individuals had generated many theories in which science, pseudoscience and eventually, beliefs and religion had mixed to provide answers to insightful questions. Documented contributions on the subject are already present in ancient Greek literature starting from Pythagoras, who around 530 BC proposed the spermism theory. He suggested that all hereditary information was stored within the male body as a library where all the instructions to build a new organism are placed.

Almost two centuries after, Aristotle updated this view. He proposed that heredity information was transmitted in form of messages by observing systematically how traits were inherited within families. Aristotle was also the first to understand the complexity behind inheritance in *De Generatione animalum*, the so-called foundational

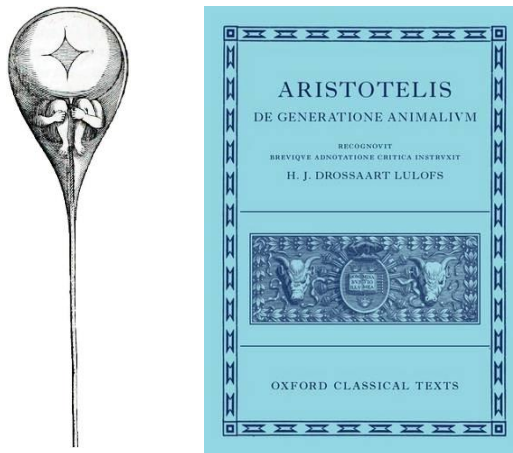
book for human genetics (Figure 1). For instance, in the passage below, he remarked that certain traits could keep hidden among family generations and manifest later in progeny.

*“And from deformed [parents] deformed [offspring] comes to be, just as lame come to be from lame and blind from blind, and in general they resemble often the features that are against nature, and have inborn signs, such as growths and scars. Some of such features have even been transmitted through three [generations]: for instance, someone who had a mark on his arm and his son was born without it, but his grandson had black in the same place, but in a blurred way. Such things happen rarely, and for the most part offspring with the body-parts intact come to be from mutilated parents, and nothing has been settled definitely about them. And they [the offspring] resemble the parents or their ancestors further away and sometimes nothing like that. And they can also transmit [features] through several generations: for instance, in Sicily a woman committed adultery with a man from Ethiopia – for the daughter did not become an Ethiopian, but her daughter [of the daughter] did.”*

*Generation of Animals (350 BC), Aristotle  
HA VII 6, 585b28-586a4*

Aristotle presumed the existence of encoded information, but he did not clarify the code identity, neither how it translated into function. In fact, these incipient observations would not be partially resolved after two millennia. Pre-Mendelian theories were essentially reformulations of Greek literature, whose discussion ended up in a scientific blind alley. Proof of that is the homunculus preformation

theory in the sixteenth century (Figure 1). It presupposed that the code and de-encryption were distributed all at once. This theory assumed that each male contained in his sperm the tiny body of his descendants following an infinite loop, being the development a mere consequence of its expansion. The preformation theory became hegemonic among medieval Christians as it fitted the beliefs of the fervent religious society of the moment.



**Figure 1.** “Homunculus” representation (left) and hardcover of “De Generatione Animalium” from Aristotles (right). Homunculus figure: Nicolas Hartsoeker in 1694. Hardcover of “De Generatione Animalium” from Aristotle: edited by H.J Drossart Lulofs (Oxford, 2005).

### **1.1.2. Mendelian theory of inheritance**

This period of scientific regression was surpassed thanks to notable scientific contributions during the nineteenth century. Among the most important findings, the Augustinian friar Gregor Mendel led several discoveries that supposed the first step towards the creation of genetics as a discipline. Contemporary to Mendel, extensive

experiments with plant hybridization were conducted by botanists such as Linnaeus. In these studies, heredity was studied indirectly using hybridizations and backcrosses, although no conceptual implications were formulated before Mendel. By the same time, Darwin published the *Origin of Species by Means of Natural Selection* in 1859, a theory of evolution that represented an unprecedented scientific revolution in biology. After that, Darwin struggled to propose a theory of heredity that supported the theory of evolution. He tried to explain how variance was generated and maintained over generations. However, the pangenesis theory that he proposed in the *Variation of Animals and Plants Under Domestication* (1867) was not consistent enough.

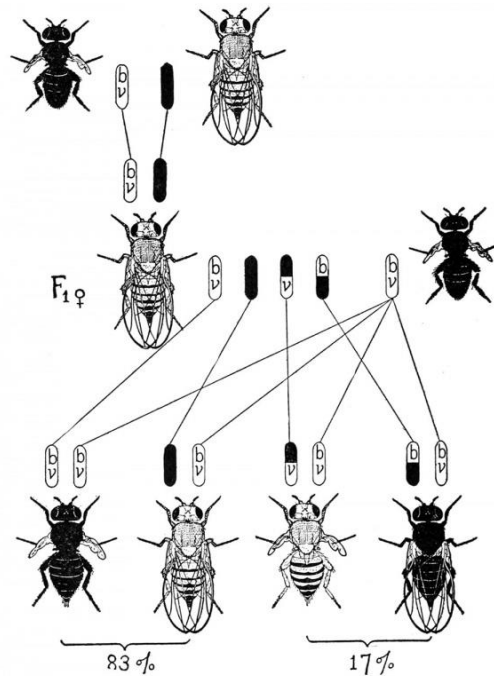
Nonetheless, Mendel had already proposed a theory that resolved this issue in 1865, but went completely unnoticed to the scientific community. Mendel identified discrete units of heredity passing between generations in his experiments on plant hybridization. These observations led to the formulation of the Mendelian laws of inheritance: the law of segregation of genes, the law of independent assortment and the law of dominance. Mendel theory was independently replicated thrice in 1900 by Hugo de Vries, Carl Correns and Erich von Tschermak-Seysenegg. Not much later, Bateson coined the term “genetics” for the first time to refer to the study of heredity and variation in 1905 (Mukherjee, 2016). In parallel to Mendel discoveries and with a completely different approach, Galton focused on the inheritance of quantitative traits, which was

extensively covered by the high impact book *Heredity Talent and Character*, published in 1865.

### **1.1.3. The dawn of molecular genetics**

After the presentation of the Mendel and Galton theories, the field of genetics experienced a big boost with successive achievements in the following years. This is the case of the germ-plasm theory formulated by Weismann in 1883, the chromosomal theory of heredity (Sutton 1903), the formulation of the Hardy-Weinberg law (Hardy 1908) and the observation of genetic linkage and crossing over on sex-linked traits of *Drosophila melanogaster* (Morgan 1910) (Figure 2). This finding was essential to understand why certain traits were transmitted together contradicting the Mendel law of independent assortment. This discovery is also at the conceptual core of the genetic linkage analysis technique, of major relevance within the scope of this thesis. The post-Mendel era ended with the first transformation experiments (Griffith 1928) and the identification of the DNA as the carrier of genetic information (Avery 1944).

In parallel, first findings regarding human traits based on Mendelian theories were described. For instance, a Mendelian mode of inheritance was identified in alkaptonuria, in different inborn errors of metabolism (Garrod 1902) and in the ABO blood groups discovered by Landsteiner in 1900. Both authors suggested that a familial component was behind the segregation of these traits.

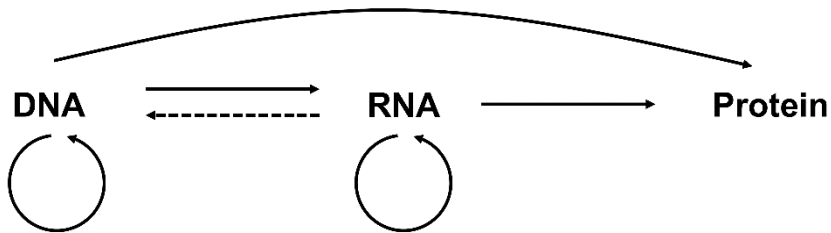


**Figure 2. Two-generation crosses of *Drosophila melanogaster*.** Two traits are linked (black body and vestigial; grey body and long, being the former the double recessive). F<sub>0</sub>: A black vestigial male is crossed with a grey long female. F<sub>1</sub>: Grey F<sub>1</sub> female is back-crossed with the black vestigial male. F<sub>2</sub>: Four different genotypes are observed in offspring, being the non-recombinant the 83% and the recombinant the 17%. From: Evolution and genetics. Morgan, 1925.

It would not be until fifty years later, though, that the elucidation of the DNA double helix structure would be published (Watson and Crick 1953) thanks to the support of the X-ray diffraction images provided by Rosalind Franklin. The observation of the DNA was a key event to build the hypothesis for DNA replication and transmission. During this period, not only replication, transcription and translation were described, but also the genetic code was disclosed (1961-1963) and the so-called “central dogma of molecular



biology” was proposed by Francis Crick in 1956. This dogma states that once the information is transferred from DNA and RNA into a protein, it cannot be transferred back to RNA or DNA (Figure 3).



**Figure 3. The central dogma of molecular biology (Crick, 1956)**

To achieve these advances, many microorganisms were used as model organisms. Simultaneously, new techniques in biochemistry and molecular genetics, like gel electrophoresis (Smithies 1955), were implemented. The new discoveries facilitated the observation of amino acid sequences, being the insulin the first one elucidated (Sanger 1952). Not only sequences were elucidated, but also certain mutations causing amino acids substitutions, deletions and frameshifts.

The emergence of this knowledge evolved jointly with the study of human inherited disease. Although medical genetics did not exist as a scientific field by that time, certain diseases such as hemoglobinopathies were studied from the genetics perspective. Sickle cell anemia was soon claimed to be a molecular disease (Pauling et al. 1949). During the 1950's, many metabolic diseases (methemoglobinemia, glycogen storage disease) were found to be

associated with genetic enzyme defects in structure, which in turn were caused by genetic mutations. With the advent of cytogenetics, the human diploid chromosome number was correctly fixed at 46 (Tjio and A. 1956). This technique started to be applied on prenatal diagnosis with the amniocentesis test. Not much later, the molecular basis of Down syndrome, a trisomy on chromosome 21, was unveiled by Lejeune in 1959, together with the Klinefelter (Jacobs and Strong 1959) and Turner (Ford 1959) syndromes. To recompile all the information being generated, a catalogue of all known Mendelian diseases was created by McKusick in 1966. Since then, it has been updated periodically with literature data, giving rise to the current database *Online Mendelian Inheritance in Man* (OMIM) that includes about 3,500 genes associated with monogenic disease.

#### **1.1.4. The genomics era**

The so-called genomics period is considered to start with the *HindIII* restriction enzyme finding (Smith and Welcox 1970), a milestone event for cloning techniques. Briefly after, the first recombinant DNA molecule was generated successfully (Jackson, Symons, and Berg 1972). Progressive advances on nucleic acid labelling and on Sanger sequencing techniques allowed the DNA sequencing of increasingly longer fragments. This happened jointly with the reverse transcriptase discovery (Baltimore 1970). Hence, only five years after sequencing the first gene, the bacteriophage MS2 coat protein (Jou et al. 1972), the entire genome of the bacteriophage  $\Phi$ -X174 was sequenced (Sanger et al. 1977). The whole process was further optimized with the incorporation of the polymerase chain reaction

(K.Mullis et al. 1986) that enabled an exponential amplification of target DNA, a limiting factor until the moment.

It was during this 70's decade that human genetics consolidated itself as a scientific field (McKusick 1975). At the beginning of the decade, many diseases were known to be genetic and the mode of inheritance unravelled, but the identity of the causal genetic elements was not established. New techniques already enabled the isolation and the sequencing of genes, as well the identification of DNA base pairs alterations contributing to disease. However, the chromosomal location of disease genes was still unknown. This bottleneck was partially solved by Botstein and Davis in 1978, who considered the linkage among DNA polymorphisms as a tool for gene mapping. This idea was especially suitable for genetic diseases segregating in families and led to the birth of genetic linkage analysis. This technique reduced enormously the gene search space as it provided specific chromosomal regions conferring disease susceptibility. In this sense, a proposal to build a genetic map of the human genome was raised (Botstein et al. 1980) as a valuable tool to systematically search for disease genes. Shortly after, a mapped region in chromosome 4 was related to Huntington's disease (Gusella et al. 1983). Despite this, the identification of the disease gene was not an automatic process as the detected regions had to be fragmented, isolated, cloned, sequenced, analysed and further validated in patients. For instance, the Huntington gene would not be identified until ten years later (Gusella et al. 1993), as it happened with cystic fibrosis (Riordan et al. 1989).

### **1.1.5. The Human Genome Project**

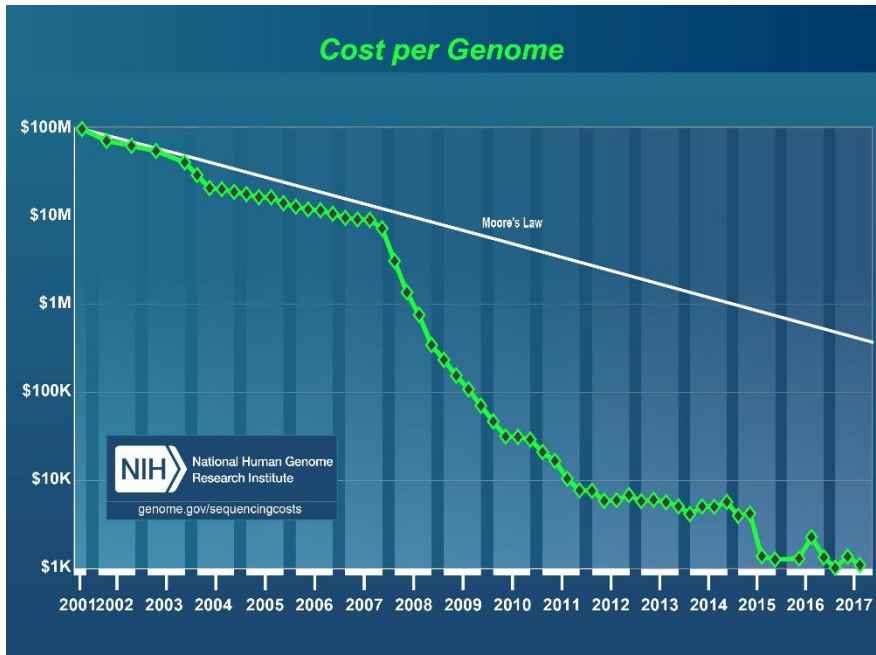
The slow and effortful task of isolating one by one every disease-causing gene raised the question of whether sequencing the whole genome would be a faster approach. It was not the unique reason. The one-gene one-disease approach was questioned as many forms of cancer and common human disorders could not be explained by a single-gene mutation, but a combination of them. Consequently, a scientific consensus appeared in favour of sequencing a reference for the whole genome. This reference genome was envisioned as a template for gene annotation, where the precise location of genes was known beforehand and to where mutated counterparts of affected patients could be compared with. The project was also expected to identify many new disease-causing genes that could be later used as druggable targets for monogenic disorders. This possibility was of major interest in a moment in which drug target discovery was strongly limiting (Penrod, Cowper-Sal-Lari, and Moore 2011).

Launched in 1989, the Human Genome Project was led by the National Institutes of Health (NIH) from United States and funded by a public international consortium. In 1992, the Institute for Genomic Research, a private foundation later converted to the Celera company, competed with the NIH to be the first to supply the sequence undertaking different sequencing strategies. After several years of a sequencing race, they both published a first draft at the same time (Lander et al. 2001) (Venter et al. 2001).

These drafts had a major impact among the scientific community since for the first time, the whole sequence of human DNA could be accessed. Nevertheless, the capacity of reading the encoded information did not necessarily entail the ability to interpret it.

### **1.1.6. The sequencing revolution**

After the final completion of the human genome sequence in 2004 (International Human Genome Sequencing Consortium 2004), genomic sequencing experienced an exponential growth. Only in the three forthcoming years, the whole sequences of many different species were achieved, including mouse (Waterston et al. 2002), rat (Gibbs et al. 2004), chicken (International Chicken Genome Sequencing Consortium 2004), dog (Lindblad-Toh et al. 2005), chimpanzee (Mikkelsen et al. 2005), sea urchin (Erica Sodergren, George M. Winstock 2006) and macaque (Richard A. Gibbs 2007). By the same time, the first personal human genomes were sequenced, underscoring the considerable reduction in time and sequencing costs fuelled by the massive parallel DNA sequencing technology (Metzker 2010). For instance, the first human genome cost around 1 billion \$ and took almost a decade to be fully completed. In 2017, the cost fell to 1,121 \$ according to the National Human Genome Research Institute (NHGRI), with further reductions still coming on the way. This reduction outperforms by several orders of magnitude the technological improvements anticipated by the Moore's law (Figure 4).



**Figure 4. Average annual sequencing cost per genome in the last two decades.**  
 Reprinted from: NHGRI, 2018.

This scenario represented a shift on research as the bottleneck moved from data production to the analysis capacity. This raised the need for creating platforms and databases to store the increasing amounts of DNA and RNA data regarding genotypes and phenotypes, such as dbGAP or EGA, as well to deposit first human genetic variation studies (Boyd and Silk 2007).

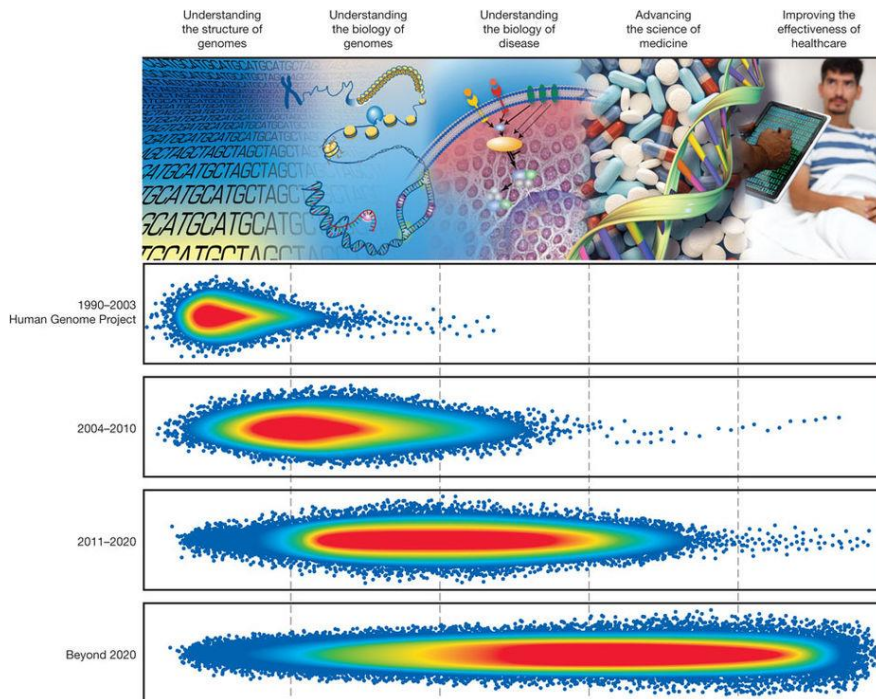
The appearance of high-throughput technologies has generated a sequencing revolution, by allowing the detection and sequencing of molecule fragments massively in parallel. In this breakthrough, two main techniques have been applied: genotyping microarrays and next-generation DNA sequencing techniques (NGS).

The genotyping microarrays permit the genome-wide identification of a large fraction of polymorphisms by taking advantage of DNA hybridization and fluorescence microscopy. The microarrays chips are designed to contain immobilized allele-specific oligonucleotide probes to which fragmented nucleotide sequences from an individual attach, being previously labelled with fluorescent dyes. The hybridization signal is then processed and analysed to determine which SNP genotype is present at each position. There are two major microarray manufacturers, Illumina and Affymetrix, which produces microarrays chips encompassing up to 4.3 and 10 million of variants, respectively. Genotyping microarrays have provided the input data for genetic linkage analysis and genome-wide association studies (GWASs).

The NGS techniques comprise the different deep DNA sequencing methods from Illumina, Roche 454, Ion Torrent and SOLiD. They are mostly based on the short-read sequencing strategy, in which millions of small DNA fragments are generated, amplified and then sequenced in parallel, multiple times and in an automated process. This strategy provides a supporting depth for almost any position of the human genome, including rare variants not observed before and absent from genotyping arrays. The sequenced fragments, known as reads, are mapped to the human reference genome using bioinformatic tools.

The advance in genomics has opened unprecedented possibilities for understanding the genetic contribution to health and disease. A

depiction of the genomic research achievements and future challenges (Figure 5) was summarized in five main steps reflecting the expected chronological transition from the genome biology comprehension to the clinical treatment (Green and Guyer 2011). In this transition, the first essential contributions belong to the genomic catalogues, which have provided valuable data on genetic variation, functional annotation of genomic elements, RNAs and proteins, among other “omics”.



**Figure 5. Transition of genomic research from the study of genome biology to effectiveness on clinics.** The density plots indicate the level of accomplishment. Reprinted from: Green et al., 2011.



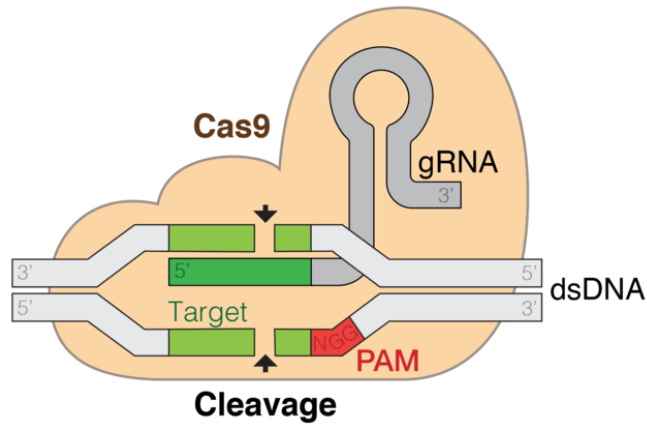
The outcome from these large catalogues and the improvement of the sequencing technology have provided a better knowledge on the biology of genomes. Advances on human genetics have also opened the doors to evolutionary studies, including the study of the genetic diversity of populations, historical migrations and estimations of the origin of modern humans in Africa about 200,000 to 400,000 years ago (Hublin et al. 2018).

### **1.1.7. Genome editing**

Nowadays, there is a pressing need to understand the genome in a level of detail that helps to develop more precise and personalized clinical treatment for disease. During previous decades, the approach was limited to a few monogenic disorders with extreme phenotypes. Genomics has enabled to move one step further by creating catalogues with millions of rare and common germline variants, whose clinical interpretation gives insightful clues on the underlying mechanism of genetic diseases. The most immediate benefit from these resources is the improvement in genetic counselling, which even in the absence of treatment represent a first step towards effectiveness on healthcare. Even so, many challenges stand ahead on diagnostics of rare and complex diseases. The room for improvement on medical genetics resides in the missing heritability of complex diseases, the still unknown genetic variation across human population and a better tracking of individual phenotypes. To disentangle that, a better comprehension of non-coding regions, gene regulation, gene-interaction networks, protein-protein interactions, triggers and environmental exposures are needed.

For many decades, genetic diseases did not have any viable drug treatment to alleviate or remove the symptoms, being still today a problem for many of them. In such cases, the only solution is to alter intentionally the human genome of progeny by negative eugenics, terminating pregnancy or using preimplantation genetic diagnosis to select healthy embryos. In the 90's decade, gene therapy was presented as a promising field and an alternative to previous drastic approaches. Initially, it comprised the usage of viruses to deliver genes in nonreproductive cells and projected the usage of germ-line cells in future. The biotechnological death of Jesse Gelsinger in 1999 halted all the clinical trials and running initiatives. It was not until fifteen years later than a successful gene therapy would be reported in the treatment for haemophilia (High et al. 2014) .

The most promising field currently under research is genome editing. It started by the discovery of a bacterial immune system that cuts viral DNA at specific sites (Figure 6). It is composed by two elements, RNA encoded in the bacterial genome that recognize viral DNA (CRISPR) and a bacterial protein (Cas9) that cuts the viral molecule after recognition (Jinek et al. 2012). The proper control of this system enables for the first time to modify any genome at precise locations in a clean way, opening gene therapy to treatment. For instance, a recent study using CRISPR/Cas9 in human embryos, corrected a pathogenic mutation in *MYBPC3* gene causing hypertrophic cardiomyopathy (Ma et al. 2017). The future success of this work, however, resides in the reduction of mosaicism and the undesired activity of this system.



**Figure 6. CRISPR/Cas9 is part of the bacterial immune system against viruses.** CRISPR is part of the bacterial genome, usually from previous viral infections, that recognize viral DNA. This recognition activates Cas9 protein to cut the viral DNA. Reprinted from: Wikipedia.

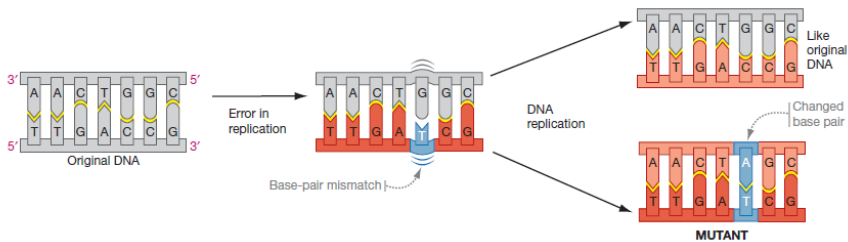
The genome editing relies mainly on the robustness of genetic counselling. This restricts any potential intervention to three principles: highly penetrant disease-causing mutations ( $\approx 100\%$ ), extraordinary suffering or short-term incompatibility with life and justifiable interventions (Mukherjee, 2016). The fast growth and optimization of the CRISPR system suggests that the edition of genomes will be a reality by next decade, if not before. It urges an ethical debate on society to set up the clinical limits of genome editing.

## **1.2. Human genetic variation**

### **1.2.1. Origin and types of genetic variation**

After publishing the *Origin of Species by Means of Natural Selection*, Darwin concentrated all his efforts on identifying the driving force that generated the genetic variation within species. The existence of this variation was indispensable to provide the substrate for natural selection to occur. Although Darwin exhaustively searched for an answer, he did not reach a plausible explanation during his lifetime.

Genetic variation is mainly generated by the inherent errors of polymerase enzymes in the process of DNA replication, which occurs at a rate of about 1 every 100,000 nucleotides. These changes are transmitted to the offspring when the DNA repair processes, mainly the mismatch repair and the proofreading system, cannot correct the errors in germ-line cells (Figure 7). These errors are mainly induced by the mispairing of wobbles and the mispairing of the non-tautomeric chemical forms of bases (Pray, 2008). Additionally, spontaneous lesions to DNA can take place due to the action of certain mutagens such as alkylating and intercalating agents or ultraviolet light (Griffiths, 1999).



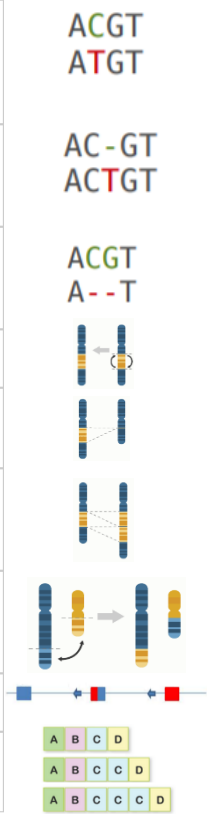
**Figure 7. Possible outcomes of DNA repair. An unrepaired mistake in DNA synthesis lead to point mutations.** Reprinted from: Griffiths, 1999.

Genetic variation is classified in two main groups according to the number of base pairs they affect. Thus, we distinguish point mutations and structural variants (Table 1). Point-mutations are classified in three subgroups: substitutions (also known as single nucleotide polymorphisms, SNPs), small insertions and small deletions, that summed account up to 99.9% of known human genetic variants (Auton et al. 2015).

Structural variants comprise gross alterations of DNA (>1kb) including chromosomal aberrations (inversions, deletions, duplications and translocations) and copy number variants. The latter group includes genetic rearrangements with locus tandem duplications, as well as big expansions of a trinucleotide repeat. The number of identified structural variants is remarkably small compared to SNPs. Still, they affect a bigger portion of the genome.

**Table 1. Class and types of DNA mutation.** Adapted from images in: Clancy, 2008, VCFtools poster, yourgenome.org and dbVAR.

Class of Mutation	Type of Mutation	Description
Point mutation	Substitution	One base is incorrectly added during replication and replaces the pair in the corresponding position on the complementary strand
	Insertion	One or more extra nucleotides are inserted into replicating DNA, often resulting in a frameshift
	Deletion	One or more nucleotides is "skipped" during replication or otherwise excised, often resulting in a frameshift
Chromosomal mutation	Inversion	One region of a chromosome is flipped and reinserted
	Deletion	A region of a chromosome is lost, resulting in the absence of all the genes in that area
	Duplication	A region of a chromosome is repeated, resulting in an increase in dosage from the genes in that region
	Translocation	A region from one chromosome is aberrantly attached to another chromosome
Copy number variation	Gene amplification	The number of tandem copies of a locus is increased
	Expanding trinucleotide repeat	The normal number of repeated trinucleotide sequences is expanded



In the narrow sense, mutations are the ultimate source of genetic variation. However, there are processes such as recombination, migration, inbreeding and assortative mating that shape the level and pattern of variation. For instance, homologous recombination increase variation by reshuffling the DNA from parental homologous chromosomes during meiosis. This generates a unique combination of variants for each individual. In the case of migrations, the introduction of genes from one population into another shifts the

original allele frequency. This also takes place with non-random mating, that occurs when individuals choose close relatives as mates (inbreeding) or because they resemble each other at a particular locus (assortative mating).

### **1.2.2. Building a catalogue of human genetic variation**

High-throughput DNA sequencing technologies have enabled the sequencing of genomes and exomes at an unprecedented pace. The data outcome from this technology, jointly with microarray chips, is the main source of information about human genetic variation currently available.

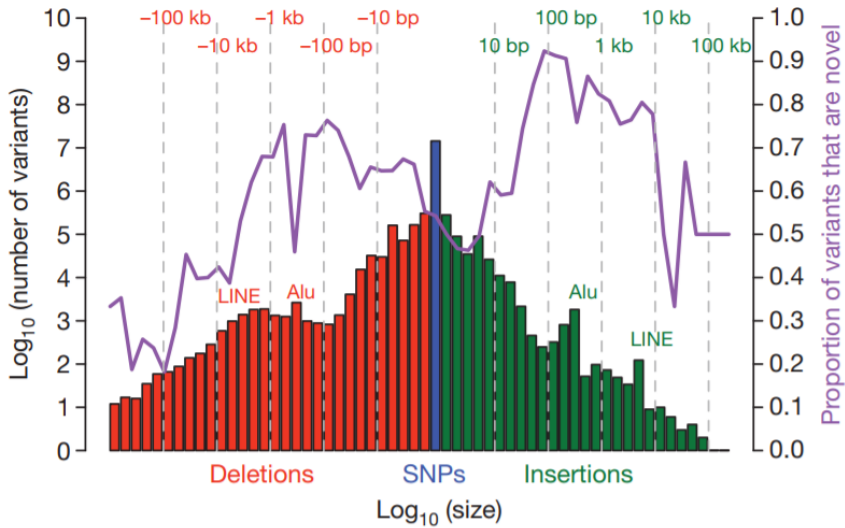
Shortly after the publication of the first human genome, the need for building a deep catalogue of human genetic variation was projected. It included a limited sample of individuals with the aim to be representative for population genetics. In this sense, the sampling strategy of the 1000 Human Genomes Project did not select individuals with a particular phenotype, but conversely, self-declared healthy individuals. Other important goals for the project were the identification of different sources of variation, the estimation of allele frequencies, the characterization of LD patterns and the phasing of haplotypes. Secondly, the catalogue was created to improve the human reference sequence, to better study the regions under selection and to support evolutionary studies on different populations.

Massive sequencing was used for the completion of the 1,000 Human Genome Project (Auton et al. 2015), which represented a

breakthrough on the knowledge of human genetics. It was the first catalogue that aimed to describe any kind of variation anywhere from the genome with a minor allele frequency as low as 1%, among 2,504 individuals from 26 populations. The project revealed 88 million variants phased onto high-quality haplotypes, being 84.7 of them SNPs, 3.6 small insertions-deletions and only 60,000 structural variants.

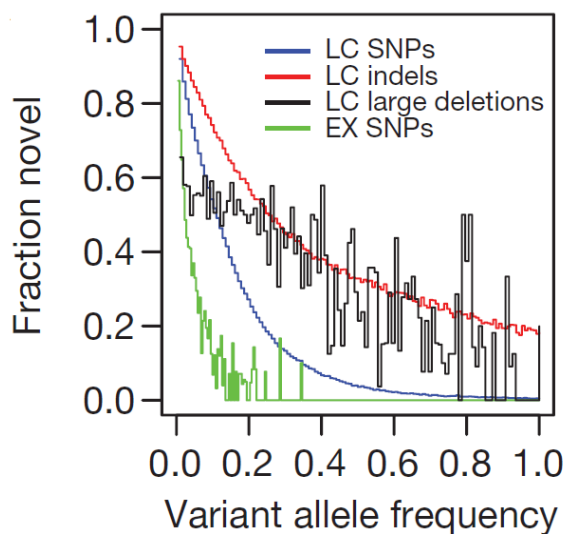
Although the number of novel SNPs in the project grew exponentially, the class with a higher proportion of novel variants corresponded to the insertions and deletions ranging from 10kb to 10bp (>70%), highlighting the room for better characterizing such variation type (Figure 8). Interestingly, the newly discovered variants differed among populations, variant types and allele frequencies (Altshuler et al. 2010). For instance, novel SNPs were mainly population-specific (84%), while already seen SNPs were shared across all populations (56%). In contrast to SNPs, most of the novel structural variants were generally observed in all populations, highlighting the lack of characterization of this type of variation. The pilot phase also included the result from an exome sequencing project with an increased depth of coverage and sample size that enabled the discovery of novel variants private to certain populations or even to specific individuals (Figure 9). This enrichment showed the need for even larger sequencing projects as much of the rare variation was predicted to remain unknown.





**Figure 8. Size distribution and novelty of variants discovered in the pilot phase of 1000 Human Genome Project.** In red, deletions respect to the human genome reference; in blue, SNPs; and in green, insertions or duplications. Purple line: Fraction of novel variants relative to existing databases (dbSNP, dbVAR, dbRIP, among others). Reprinted from: Altshuler, 2010.

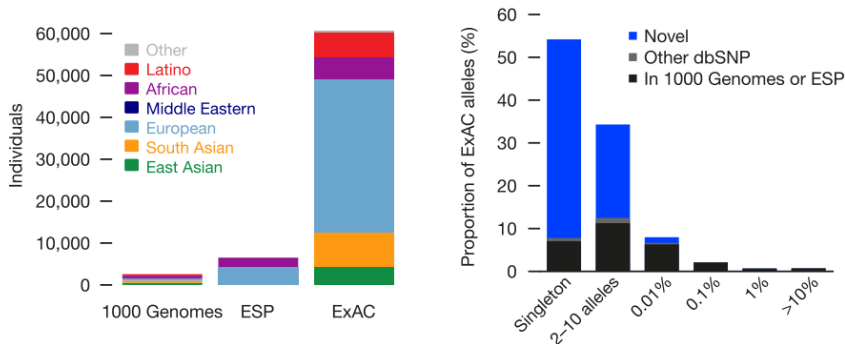
The very first effort that put together a large amount of sequencing data to produce a deep catalogue of variants associated with disease was the National Heart, Lung and Blood Institute Exome Sequencing Project (NHLB-ESP), which sequenced 6,500 exomes (<http://evs.gs.washington.edu/EVS/>). The goal of this project was to discover novel genes implicated on heart, lung and blood disorders.



**Figure 9. Fraction of novel variants in each allele frequency.** Novelty was determined in comparison to previously described variation in databases (dbSNP and dbVAR). LC: low coverage, EX: exon. Reprinted from: Altshuler, 2010.

The recent publication of the Exome Aggregation Consortium (ExAC), an analysis of protein-coding variation in 60,706 humans, and the Genome Aggregation Database (gnomAD), an analysis of 121,216 exomes and 15,136 genomes, has supposed a step further on the study of human genetic variation (Lek et al. 2016). Even more recently, the US National Institute of Health reported that the Trans-Omics for Precision Medicine (TOPMed) program achieved the sequencing of >130,000 whole genomes by July 2018. In the case of ExAC, the call set exceeded by nearly one order of magnitude previous exome databases. This increase was sufficient to study regions of the genome depleted for variation and to detect mutational recurrence in general population. Moreover, it provided better resolution on the analysis of very-low frequency variants globally

and for all populations (Figure 10). The analysis identified more than 7 million coding variants, including 300,000 indels. This observation represents in practice one variant every eight nucleotide bases of the exome.



**Figure 10. Population structures of recent large-sequencing projects and novelty proportion from ExAC alleles.** Left: ExAC is one order of magnitude larger in size than previous data sets for all studied populations. Right: Most of the variants are novel and rare, being singletons the most abundant class of human genetic variation. Reprinted from: Lek et al., 2016.

Again, most of the newly discovered variants (99%) were low-frequency variants (MAF<1%) and 54% of them were singletons, SNPs only observed in the allele of one individual (Figure 10). This indicates that even ExAC is not large enough to observe all coding non-lethal variation in human.

With the current world human population, 7,6 billion in 2018 (NU World Population Prospects, 2017), it is hypothesized that almost every position in the genome can mutate or have mutated in the past, as the rarest mutational rates are predicted to be in the order of  $10^{-9}$

(Samocha et al. 2014). Although this sequencing effort seems unattainable, mutational recurrence and high fractions of variation are already visible in coding-regions, suggesting that larger sample sizes, deep coverage and incremental improvements on the sequencing technology will provide an almost complete picture on missing variation.

Current databases of variation are experiencing an exponential growth of entries. The most paradigmatic example is the NCBI dbSNP database for short genetic variation. In less than three years, it has increased the number of variants from 62 to 660 million (build 151, March 2018). Similarly, the dbVAR database for human genomic structural variation (>50 bp) contains 5.2 million variant regions and 34.6 million variant calls by November, 2017. Two major conclusions are observed from this exponential increase. Firstly, there is a large abundance of rare and private SNPs in databases, which accounts for the largest part of novel variants. This is likely to continue in the forthcoming years as more individuals are sequenced. Secondly, there is a strong need to better characterize structural variation (SV), notably on the rare variation class. Only then, we will be able to explain if the lower discovery rate of SV in relation to SNP is certain or contrarily, shotgun techniques are inappropriate to detect rare SV. The nanopore technology (Jain et al. 2018) is called to be a sequencing breakthrough, as it will provide longer reads that will better cover the highly repetitive regions of the genome leading to an enhanced representation of the reference version.

## **1.3. Functional genomics**

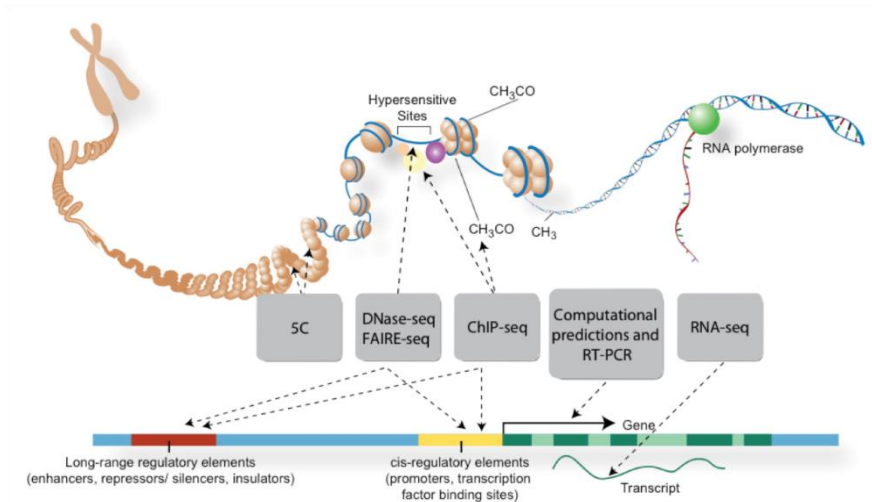
The clinical value of the human variation catalogues resides in the functional interpretation of the data. Although many protein-coding genes have been identified, only a small fraction of the reported variation falls within those regions. At the current pace of sequencing and variant discovery, it is currently unachievable to experimentally test the effect of each new variant to the phenotype. Instead, the challenge is to evaluate bioinformatically whether variation at specific genetic positions is neutral or exhibit a functional role (Mahmood et al. 2017). To that purpose, different projects have recently created large resources of functional genomics data, including transcriptomics and epigenomics, to have a better knowledge of regulatory regions, genome organization and expression variability in tissues. These resources are the result of applying many new techniques aimed to map RNA expression levels and identify the methylation patterns, the histone modifications and the chromatin accessibility of the genome. These elements also give insight on mechanisms frequently involved in human disease. Below, we describe three widely-used resources in functional genomics analysis.

### **1.3.1. The Encyclopaedia of DNA elements (ENCODE)**

The Encyclopaedia of DNA elements (ENCODE) (Dunham et al. 2012) aimed to characterize the biochemical activity at a genome-wide scale (Kellis et al. 2014). The results revealed pervasive activity on many regions of the genome, including non-coding and non-

conserved regions, mostly occurring in a specific cell type manner. For instance, from the total 2.9 million of open chromatin regions, only 3,700 are observed in all cell types (Maher 2012). This has raised the question on what can be considered functional fostering a debate between the genetic, the biochemical and the evolutionary approaches (Doolittle 2013). Still, the ENCODE major contribution resides in the identification of genomic elements showing molecular characteristics that may be used as a starting point for studying cellular processes and disease.

The ENCODE provides a catalogue of RNA transcripts, genomic regions bound by transcription factors, genomic regions occupied by nucleosomes, histone modifications and open chromatin regions. These signatures can describe how the genome packages, regulates and reads the information in different cell types (Maher 2012). The ENCODE project involved the usage of several techniques and approaches (Figure 11). The current last version (V4) provides genomics annotations at an integrative and at ground-level. Below, we shortly describe the nine different sources of ground-level annotation, directly obtained from the specific experimental data.



**Figure 11. ENCODE functional and regulatory elements.** Reprinted from: Encode Consortium.

**a) Open chromatin:** it provides a peak profile for the enzyme DNase I hypersensitive sites (DHS) (Song and Crawford 2010). By August 2018, DNA-seq data for 151 cell lines assays was available.

**b) Histone mark enrichment:** it provides a peak profile for the regions enriched on particular histone marks, including H3K4me3, H3K9ac, H3K27ac, among others. The experimental data is produced by the chromatin immunoprecipitation technique (ChIP-seq), which uses specific antibodies to bind to a protein, which in turn is bound to DNA. The capture of the protein-DNA crosslink allows the identification of the region where the DNA binds (Furey 2012). By August 2018, 414 cell lines were assayed for histone marks.

**c) Transcription factor (TF) binding:** ChIP-seq can be specifically addressed to identify the DNA regions where TF binds in living cells. The binding is usually presented as sequence motifs of 6 to 25 nucleotides positions. By August 2018, the ENCODE Factorbook resource hosted information up to 167 TFs.

**d) Gene expression:** it provides the expression levels of genes and non-coding transcripts annotated by GENCODE using RNA-seq (Wang, Gerstein, and Snyder 2009). The assay included information of total RNA-seq, small RNA-seq, polyA RNA-seq, siRNA-seq, microRNA-seq, among others.

**e) Promoter activity profiling:** the RNA Annotation and Mapping of Promoters for Analysis of Gene Expression (RAMPAGE) tool quantifies the gene expression and identifies the promoter locations (Batut and Gingeras 2013). By August 2018, only 28 RAMPAGE cell-lines assays were available.

**f) RNA binding protein occupancy:** it uses enhanced CLIP-seq data (UV crosslinking and immunoprecipitation of ribonucleoprotein complexes) to determine the binding sites where RNA-binding proteins (RBPs) interact (Van Nostrand et al. 2016). By August 2018, there was only data for two cell-lines encompassing 355 assays.

**g) DNA methylation:** it includes the genome-wide methylation state of CpG dinucleotides. By August 2018, 87 cell-line were assayed by



DNA methylation arrays and only 11 by whole-genome bisulfite sequencing.

**h) Three-dimensional chromatin interactions:** it informs about the genome-wide long-range chromatin interactions between regulatory regions such as promoters or distal enhancers mediated by specific target proteins. The chosen technique is the chromatin interaction analysis with paired-end tag sequencing (ChiA-PET), which uses an intermediate step of ChIP to enrich the number of interactions (Li et al. 2017). By August 2018, only 55 ChiA-PET cell-line assays were available.

**i) Topologically associating domain:** it provides information about the three-dimensional architecture of the genomes by displaying the topologically associated domains (TAD) and the cell-specific A and B compartments, which are respectively associated with open and closed chromatin (Fortin and Hansen 2015).

The jointly interpretation of the different signals (integrated-level annotation) is especially meaningful when the different patterns agree with the underlying biology. For instance, this happens when the open chromatin boundaries coincide with histone modifications or when the gene expression correlates with the binding of certain transcription factors at promoter regions. The outcome of the ENCODE has provided evidence for transcription, regulatory elements, chromatin structure and histone modifications, indicating

far more complexity than anticipated on gene regulation mechanisms.

Some critics have argued about the implications and the value of the ENCODE data. One of the most controversial ENCODE statements, that was refuted shortly after, hold that above the 80% of the genome have signatures of functionality. This percentage is unconceivable since it assumes that most of the genome (70%) can be functional, even being non-conserved, tolerating any type of mutation (Graur et al. 2013). This statement was based on an unappropriated definition of functionality, which confounded the pervasive activity expected from junk DNA with causality and presented as functional a lot of elements that just reflected biological activity (Eddy 2013). Moreover, the statement was based on unbalanced selection of sensitivity over specificity and it did not take into account the magnitude of the effect of the biochemical activity (Graur et al. 2013). Others have criticised the arbitrary choice of the cell lines and the transcription factors for analysis and the lack of negative controls in the assays. Despite these issues and the incompleteness of the data, ENCODE remains as one of the most widely used resources in functional genomics.

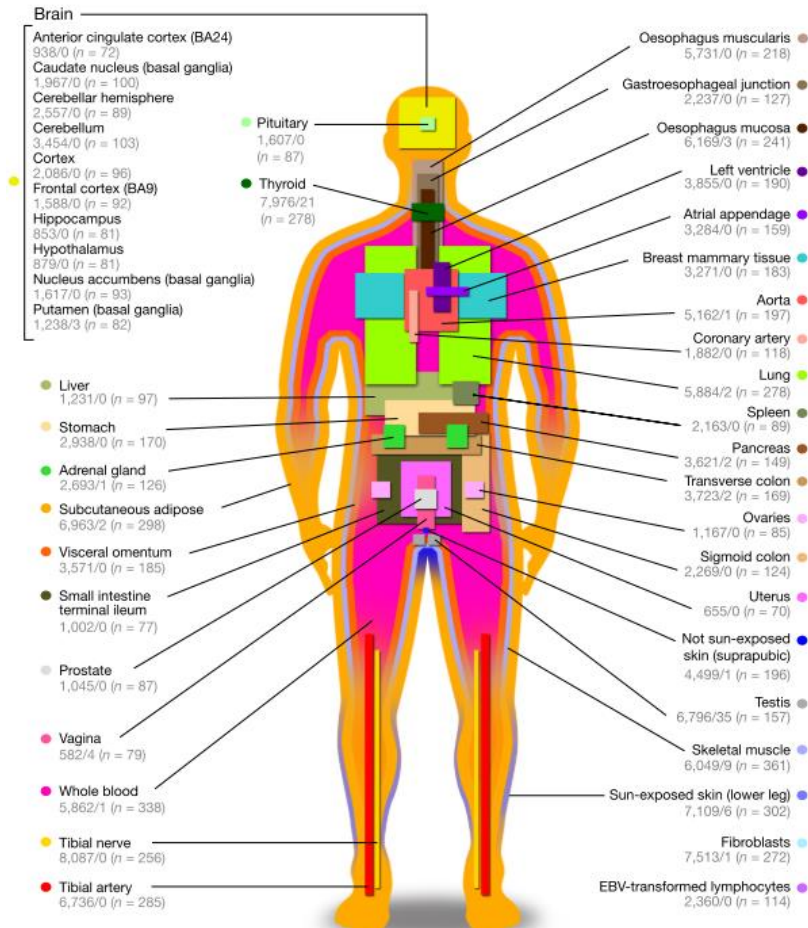
### **1.3.2. The Genotype-Tissue Expression (GTEx) project**

The GTEx project (Aguet et al. 2017) was created to answer a key question that emerged from the ENCODE data: How human variation could affect the activity of regulatory elements in the genome? The

answer was again oriented towards deciphering the molecular consequences of the genetic variation.

The GTEx project generated a catalogue of functional data that provides information on the gene regulatory mechanisms by correlating the observed genotypes with tissue-specific gene expression levels of the same individuals. This approach helps with the identification of genetic variants that are associated with expression changes and also determines the magnitude of such association. Variants that are highly associated with gene expression changes, are called expression quantitative trait loci (eQTLs). Depending on the distance from the gene, we distinguish cis-eQTLs (local action, within <1Mb from the gene transcription start site) from trans-eQTLs (distal) effects.

The study design involved the collection of multiple human tissues (Figure 12) from donors that were genotyped by a reference panel developed from the 1000 Genomes Project, and had their RNA sequenced. The GTEx is a pioneering project due to the protocols followed with data collection. A wide-scale sampling of histologically normal organs and tissues was obtained from donors with a healthy clinical history in the post-mortem interval. In the v7 release, data from 714 donors and 53 tissues was published, aiming to reach 1,000 individuals at the last phase. As for the data analysis, a linear model was applied correcting the expression associated with ancestry, age, genotyping platform and other unknown covariates.



**Figure 12. Representation of the 44 tissues and cell lines included in the GTEx release v6p.** The numbers in grey: cis-eQTLs/trans-eQTLs for each tissue. Between parenthesis the tissue sample size. Reprinted from: (Aguet et al. 2017)

The results from GTEx identified 341,316 cis-eQTLs from 31,403 genes and lincRNA in 48 tissues, affecting either most of tissues or only a small fraction of them. Identified trans-eQTLs were mainly tissue-specific and tended to overlap enhancer regions, although the

available data remained underpowered to detect most of them (Aguet et al. 2017). Both numbers are expected to grow, as the sample size from GTEx increases.

Apart from cis-eQTLs, the effect of local regulatory variation was quantified by the allele specific expression (ASE). This technique is applied to heterozygous variant sites to elucidate the difference on expression levels attributable to each allele. 88% of the testable genes in GTEx showed significant allelic imbalance, highlighting again the predominant effects of local regulation in gene expression. As for complex disease associations, half of the reported GWAS overlapped GTEx eQTLs. Interestingly, the phenotypic impact was higher in genes with cis-eQTLs shared across tissues. Those genes had substantially less loss-of-function mutations, indicating the action of purifying selection. A future step on GTEx prospects is the combination of gene expression data with molecular measurements from epigenomics and proteomics (Stranger et al. 2017).

The genotypes from GTEx individuals are not publicly available in order to preserve the donor identity and due to the nature of the consent agreement. They can only be downloaded from the dbGAP repository under authorized access.

### **1.3.3. The NIH RoadMap Epigenomics Mapping Consortium**

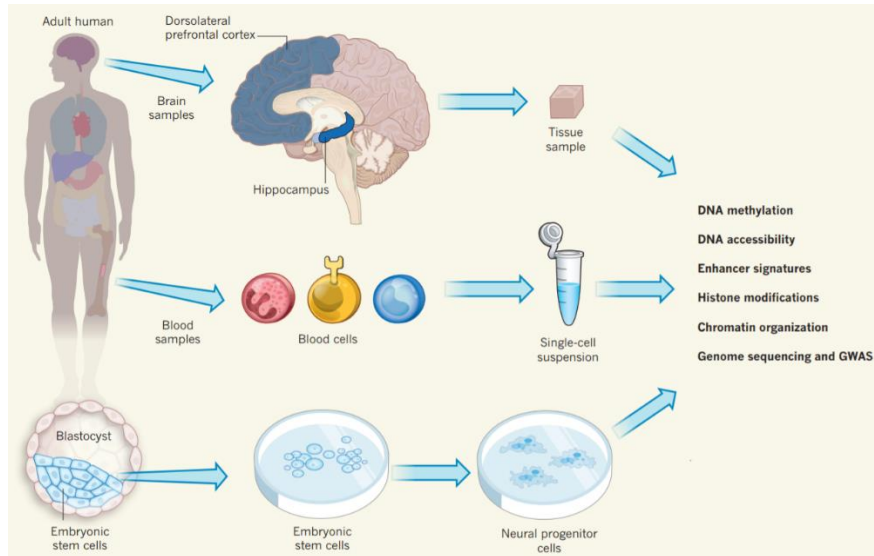
The Roadmap Epigenomics Mapping Consortium (REMC) is a public resource from the National Institutes of Health (NIH) of the

United States that contains high-throughput data on the functional elements that regulate gene expression in a cell. The epigenome provides information about the structural patterns in which DNA is methylated and how histones are modified. Also, it informs about the chromatin accessibility and the location of regulatory elements such as promoters and enhancers.

The project was created due to the lack of an integrative systematic analysis of the epigenomic landscape and to go further from existing annotations. Unlike the ENCODE project that catalogued cell-lines grown in culture, the REMC project focused on samples directly extracted from ex-vivo primary human tissues (Romanoski et al. 2015). Thus, the REMC aimed to build a high-resolution map of regulatory elements that could be used as a reference for understanding cellular circuitry, lineage specification and ultimately human disease (Figure 13).

Some of the techniques employed in the REMC were already used in ENCODE: ChiP, DNA digestion by DNaseI, bisulfite treatment or RNA profiling, as well as short-read sequencing. Two novel techniques were introduced: methylated DNA immunoprecipitation (MeDIP) and methylation sensitive restriction enzyme digestion (MRE). With MeDIP, the DNA is cut by sonication and immunoprecipitated with a specific antibody to measure the methylation enrichment of the fraction. It is combined with microarrays to be applied genome-wide (Mohn et al. 2009). As for

MRE, it uses restriction enzymes sensible to different DNA methylation states that generate fragments to be later studied.



**Figure 13. Illustration of the sampling protocol and the techniques used in the production of the reference epigenome from the RoadMap Epigenomics Project.** It includes embryonic and adult tissues, from healthy and diseased individuals. Reprinted from: *Roadmap Epigenomics Consortium et al., 2015*

The publication of the first human epigenomes led to several findings (Roadmap Epigenomics Consortium et al. 2015). The main highlights are discussed below:

- The structural organization of the genome is fundamental to understand its regulatory elements. The binding of transcription factors to DNA enhancers remodels the chromatin state. In specific tissues, the activated enhancers are enriched in sequences to which lineage-determining and

signal-dependent transcription factors bind. This reflects the importance of transcription factors in expression-specific stages to determine the cell development through changes on the gene expression (Tsankov et al. 2015).

- Chromatin remodelling is crucial for DNA accessibility and the binding of transcription factors. For instance, the histone marks correlate with different levels of DNA methylation and accessibility; and also predict the RNA expression. One of the best-known examples of histone modification is the one at histone H3 affecting the amino-acid residue lysine 27 (K27). The modification can either be an addition of an acetyl group (H3K27ac) or one trimethylation (H3K27me3). The former case correlates with major transcriptional activity, while the latter is linked to transcriptional repression (Ziller et al. 2015).
- Large regions of the chromosomes present domains of distinct epigenomic signatures that correlate with transcriptional activity. Also, the enhancers shared across tissues are enriched for common gene functions working as coordinated regulated modules.
- The comparison of the epigenomics signatures among healthy and diseased cells is essential to understand the drivers of human disease. Many GWAS associations located at non-coding regions have been studied in the context of specific cell-types. This has revealed that they mainly fall in regions



enriched in tissue-specific regulatory regions (Maurano et al. 2012).

- Long-ranging interactions through the action of distal enhancers implies large chromatin reorganization in space and time. This has also been observed during stem-cell differentiation (Dixon et al. 2015).
- Haplotype-specific differences on histone modification and chromatin architecture correlate with the allele specific expression across many tissues (Leung et al. 2015) . These differences could be associated with mutations that disrupt the transcription factor binding sites or long-range interactions.
- High-resolution microscopy will be used in the future to produce live imaging of the chromatin remodelling in space and time, with the assistance of systems biology.

The REMC resource provides an averaged epigenetic map from a population of cells extracted from a particular cell-line in a tissue, rather than a single cell map of that tissue. Although this restricts any stochastic variability estimation within the same cell-type, the objective of the project is to determine the epigenome of the entire collection of cell types in human body and also to monitor the ageing and environmental effects to predict its dynamic landscape. As a novel resource, the REMC also promotes the development of better protocols for producing the data, its fast dissemination and seeks the standardization of the analytical tools.

The ENCODE, GTEx and REMC projects, jointly with the last catalogue of human variation, have supposed a major breakthrough in the functional genomics field. The unprecedented number of putatively DNA regions involved in the regulation of gene expression and chromatin organization, provide a comprehensive picture of the information encoded in the genome. This is of major value for understanding the impact of human variation.

## **1.4. Clinical interpretation of human genetic variation**

### **1.4.1. The variant annotation process**

The identification of variants associated with a disease requires an accurate assessment of pathogenicity. The recent and extensive use of genome-wide sequencing techniques has supposed a dramatic shift in clinical genetics due to the exponential growth in variant discovery. The limitation is no longer the generation of the sequencing data, but its analysis and interpretation (Salgado, Bellgard, et al. 2016). Most of the newly discovered variants are rare in the population or even private to certain families, but not all of them show deleterious effects, complicating their classification among neutral, functional or pathogenic categories. This uncertainty opens an interpretative gap between the identification of variants and their clinical annotation (Cutting 2014) and it is likely to keep growing.

Variant annotation is the process that evaluates the functional significance of variants by associating different sources of information with them. The final goal is to predict the involvement of a variant in a trait or in a disease. The localization of variants at the genomic level is the key step in annotation, but many more features are incorporated. Examples of that are the measures of population allele frequency, sequence conservation, predictions of constraint, fitness and mutation tolerance, as well experimental evidence of their impact.

#### **1.4.2. Variant localization and nomenclature**

Predicting the impact of variants on DNA, mRNA and protein levels is complex. Depending on the variant position we mainly distinguish variants falling in coding and non-coding regions, being the former under strong purifying selection as they have greater impact on protein function and are more likely to lead to phenotype changes (Tennesen et al. 2012). Still, as mentioned before, a lot of regulatory regions, including 5' and 3'UTR, promoters, transcription factor binding sites and splicing positions in introns, can have a functional role by influencing the level, timing and tissue specificity of gene expression (Lonsdale et al. 2013) (Mohammadi et al. 2016).

Among coding SNPs, we distinguish between synonymous and non-synonymous variants with respect to the genetic code. Synonymous SNPs modify the DNA sequence, but they specify the same amino acid as in the original codon. Although they have often been misinterpreted as a neutral or a silent kind of variation, synonymous

variants can be in fact noisy or neutral depending on their ultimate effect. The effect of noisy synonymous variants can be observed with the creation and deletion of splicing sites, the mRNA folding, the microRNA binding or with the translation efficiency (Mueller et al. 2015) (Plotkin and Kudla 2011) .

On the other hand, non-synonymous SNPs change the original codon causing a different protein sequence. Depending on the downstream consequence from these changes, we distinguish missense and nonsense variants. A missense variant is a point mutation that changes the translation of the codon into a different amino acid, including the shift of a stop codon into another residue causing protein elongation. With a nonsense SNP, a termination codon is generated. Despite the protein change, not all non-synonymous SNPs show a deleterious effect (Tang and Thomas 2016) (Yngvadottir et al. 2008). For instance, loss of function (LoF) variants are clearly enriched on rare variants rather than missense ones (MacArthur et al. 2012).

Insertions and deletions (indels) in the coding region can lead to frameshift (FS) mutations if their size is not multiple of three. These FS variants typically lead to premature stop codon and putatively activate the nonsense-mediated RNA decay (Lin et al. 2017). However, they are not always processed, as indels that cluster towards the end of a protein can avoid the surveillance system. FS are also enriched on genes under relaxed selection, as those related to olfactory receptor activity (Hu and Ng 2012).

The consistent description of variant localization requires following a specific nomenclature. The Human Genome Variation Society (HGVS) nomenclature is recommended to unambiguously describe sequence variants (den Dunnen et al. 2016). Still, this nomenclature has limitations to describe consequences outside coding regions and should be regarded as a mere prediction (Salgado, Bellgard, et al. 2016), particularly when describing mutations at RNA level.

### **1.4.3. Position-based variant annotation**

One of the most essential steps in annotation is to map the genetic position of a variant respect to the coding regions of a reference genome. Although it may seem a straightforward step, it is not a perfectly resolved point and many tangential aspects gain importance when looked in detail. Classifying variants in relation to coding sequences is highly dependent on the knowledge we have about such regions. Current annotations are strongly biased to the transcript set choice. For instance, two widely-used transcripts sets such as REFSEQ (O’Leary et al. 2016) and ENSEMBL (Zerbino et al. 2018) notably differ among them since they respond to different building criteria (McCarthy et al. 2014). REFSEQ is a non-redundant curated resource that only includes genetic features with experimental evidence. On the contrary, ENSEMBL incorporates more exhaustive information from several sources. Thus, the transcript set only contains 65,648 transcripts in RefSeq – NCBI Homo Sapiens Annotation Release 109, while 261,276 are included in ENSEMBL release 93. The consequence of this imbalance is that a large number

of variants are only coding in ENSEMBL. In addition, both transcript sets are also subject to modifications, as periodical updates can cause substantial changes with time. Owing to these differences, it is recommended the usage of consensus coding sequences (CCDS), a dataset created to have an identical annotation of protein coding genes (Salgado, Bellgard, et al. 2016).

Another major component of the position-based variant annotations is the software tool used to retrieve or estimate the functional impact of variants. Such software tools are built following different algorithms and they report the results as a system of categories organized hierarchically. This ontology is not uniform among different software programs and, despite being equivalent under a fraction of the cases, this leads to certain disagreement between predicted annotations. The annotation category with the most deleterious consequences on genes is the loss-of-function (LoF) variant, which indicates gene product loss and a potential impact on phenotypes. This category encompasses frameshift indels, stop-gain and stop-loss variants, as well as some splicing mutations. Other categories with expected milder effects are the synonymous variant and the non-frameshift indels.

In McCarthy et al. 2014, they observed large differences on LoF variant annotations results, either when using different software tools, ANNOVAR (Wang, Li, and Hakonarson 2010) and VEP (McLaren et al. 2016), on the same transcript set (64% agreement); or when using different transcript sets, REFSEQ and ENSEMBL

(44%) with the same annotation tool. According to these results, the choice of the transcription set produced the major discordance on annotation. The choice was shown to be a compromise between minimizing the number of false positives (REFSEQ) or maximizing the detection of potential LoF variants (ENSEMBL). As for the disagreement on the annotation software tools, three major reasons were given. First, even when using the same transcript set, not all the predictions were based on the same transcript. Thus, while one tool reported only the most severe annotation for a variant, the other informed about all the possible annotations given the different transcripts of a gene for that same variant. Second, certain annotation categories such as splicing showed high discrepancies due to the different definitions considered for each tool in regard to splicing site positions. Third, the tools used different precedence rules that lead to different annotations, even when reporting exactly the same variant at the same transcript. All these differences highlight a large room for improvement and the need for standardization in the field.

#### **1.4.4. Other annotations at variant, gene and pathway level**

Beyond the position-based approach, many different types of annotations have been developed (Torkamani et al. 2011). The population characteristics, such as allele frequencies or LD patterns, have been extensively used as a surrogate for variant functionality (Zhu et al. 2011). This frequency is computed from large-scale sequencing projects, such as ExAC, gnomAD, 1000 Human Genome Project or TopMed, and should ideally match the population origin from the queried sample.

In the same line, the sequence conservation across species is used to detect regions under negative purifying selection (Siepel et al. 2005). This is measured using a model of neutral evolution in which sequences that are more conserved than expected are likely to have a functional role (Choi et al. 2012). Therefore, the higher is this conservation, the most chances than a variant shows deleterious effects on protein function, being not a definitive proof for pathogenicity.

Other annotation strategies make use of computational predictive models to evaluate the functionality of variants. These models include features such as fitness, deleteriousness, sequence homology, functional domains or physicochemical changes of amino acid residues. Many predictors have been developed to assess the pathogenicity of variants: SIFT (Sim et al. 2012), PolyPhen2 (Adzhubei et al. 2010), UMD-Predictor (Salgado, Desvignes, et al. 2016), among others. All these approaches can be summarized into genome-wide position scores that estimate the constraint or the impact of particular variants for different metrics, being highly useful for annotation. These scores are defined within a range of values that reflect a fitness gradient regarding the functional consequence of variants (Gulko et al. 2015). Variants that reduce the fitness of an organism are called deleterious and they are expected to be under the effect of purifying selection. Only under special circumstances and beyond the capacity of many predictors, a variant can become advantageous and increase the fitness. In the table below (Table 2) we describe a set of genome-wide examples scores.



**Table 2. Example of genome-wide position scores used in annotation.**

Score	Description	Range of values	Reference
<b>Minor Allele frequency (MAF)</b>	<b>ExAC:</b> 60,706 exomes <b>1000 Human Genome Project:</b> 2,504 genomes <b>GnomAD:</b> 15,496 genomes & 123,316 exomes <b>TopMed:</b> 62,784 genomes (package release)	From 0 (non-observed) to 0.5.	<b>ExAC:</b> (Lek et al. 2016) <b>1000 Human Genome Project:</b> (Altshuler et al. 2010) <b>GnomAD:</b> <a href="http://gnomad.broadinstitute.org/">http://gnomad.broadinstitute.org/</a> <b>TopMed:</b> <a href="https://nhlbiwgs.org/">https://nhlbiwgs.org/</a>
<b>phastCons</b>	Scores derived from a multiple alignment of the human genome to other vertebrate species.	From 0 (non-conserved) to 1 (conserved)	(Siepel et al. 2005)
<b>phyloP</b>	Phylogenetic p-values (phyloP) scores under a null hypothesis of neutral evolution. They are derived from a multiple alignment of the human genome to other vertebrate species.	Score>0, conservation Score>3: It measures slow evolution Score<0, acceleration Fast evolution	(Pollard et al. 2010)
<b>fitcons</b>	Fitness consequences of functional annotation. It integrates the data from several functional assays: ChIP-Seq data, DNase I peaks, chromatin states, normalized RNA-seq data as well as CDS annotation.	From 0 (non-selective pressure) to 1 (strong selective pressure indicating function).	(Gulko et al. 2015)
<b>CADD scores</b>	Combined Annotation Dependent Depletion (CADD) scores indicate deleteriousness of SNVs using scores from several sources: phastCons, GERP, phyloP, SIFT and PolyPhen.	PHRED scores: from 1 (neutral) to 99 (highly-deleterious). PHRED>10 = Top 10% most deleterious variants. PHRED>20 = Top 1% (threshold for pathogenicity)	(Kircher 2014)

<b>MCAP scores</b>	Mendelian Clinically Applicable Pathogenicity (MCAP) scores. It classifies the pathogenicity of rare missense variants. It integrates data from SIFT, PolyPhen-2 and CADD.	From 0 (benign) to 1 (pathogenic).  Threshold=0.025 to separate likely benign from pathogenic variants.	(Jagadeesh et al. 2016)
<b>Linsight</b>	It measures the probability of negative selection on noncoding sites (deleteriousness). The scores can be used to quantify the evolutionary constraint on regulatory sequences.	From 0 (non-deleterious) to 1 (deleterious).	(Huang, Gulko, and Siepel 2017)

Other annotations make use of the variants associated with clinical phenotypes encompassed in several databases such as OMIM (Amberger and Hamosh 2017), the GWAS catalog (MacArthur et al. 2017) or ClinVar (Landrum et al. 2018). Such an approach is limited by previous knowledge, false positives and it is uncompleted for many conditions.

Apart from the clinical phenotype, molecular measurements are also used to annotate variants. As explained in the previous chapter, the gene expression and protein levels, as well the epigenetics signatures are normally used as surrogates for functionality. A limitation of this method is that the tissue showing the disease expressivity, which is not always identifiable, must match the same tissue in molecular data.

The current annotations are limited when the disease risk implies several loci. Some risk assessments try to predict the collective effect

of variants making assumptions on how they interact, by modelling the variant effect sizes, frequencies and the disease incidence and prevalence (Torkamani et al. 2011). Similarly, pathway-based annotations are confounded by biochemical reactions full of compensatory mechanisms due to genetic epistasis (Jin et al. 2014). In other cases, when biological processes are not well known, co-expression and network inference techniques are used to predict potential effects on the phenotype (Wu et al. 2008). Other limitations include the difficulty of addressing context-specific perturbations such as environmental effects or the comorbidities reflecting a complex disease architecture.

The annotation process can also be challenging when upstream steps concerning sequencing data introduce variability. It is not negligible the effect on annotation of sequencing technologies platforms and the pre-processing pipeline steps involving genome mappers (Thankaswamy-Kosalai, Sen, and Nookaew 2017), the identification of technical duplicates and the assessment of variant quality (Regier et al. 2018). For instance, the identification of short tandem repeats or copy number variants is underpowered with short-read sequencing techniques. Also, the choice of the variant caller may strongly affect the annotation of some variants, as a fraction of them will be false positives, technological artefacts annotated as real variants, or true negatives, true variants discarded as artefacts.

In summary, rather than viewing annotation as a fully reproducible automatic step, we should be aware of the many decisions that can

affect its outcome. The simple choice of a particular transcript set makes a substantial difference having downstream implications. It is important to be aware of this and to perform annotation in a consistent, accurate and reproducible manner, especially when it concerns to research and clinical use.

#### **1.4.5. Variant filtration and prioritization**

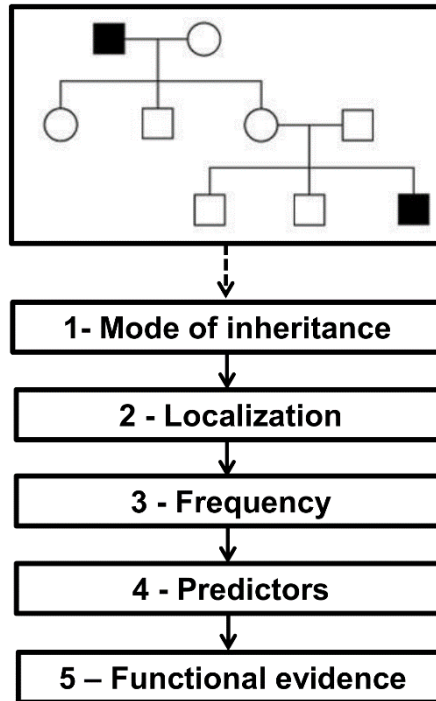
The variant annotation is the process of describing the nature and the effect of the DNA alteration produced by a variant (Eilbeck, Quinlan, and Yandell 2017), providing an assessment of their functionality. The variant filtration step aims to reduce the number of annotated variants that could be potentially associated with a trait, by using different filtering criteria such as the frequency variant threshold or the mode of inheritance (Dashti and Gamielidien 2017). When the filtering efforts reduce the number of likely functional variants to a few candidates, then it follows the variant prioritization step, a ranking process for those variants that are most likely to damage gene function and eventually trigger the disease. At this step, variants are evaluated using all the biomedical contextualizing information at hand.

Variant prioritization is mostly applied on a case-by-case approach, rather than with a fixed strategy of invariable steps. Here we describe a general variant filtration and prioritization scheme for families (Figure 14). One possible initial step consists of applying a particular mode of inheritance (MOI), which determines the model of Mendelian segregation that a variant must follow within an affected

family. The MOI discards all the variants that follow a different segregation model than the specified. A second step can be filtering variants by location using a transcript set. If this criterion fails to provide variants of interest, subsequent analysis with regulatory regions and synonymous variants are reconsidered. A third filtering step may consider an estimated threshold for minor allele frequency. This threshold usually excludes common variants, which cannot be pathogenic. The allelic frequency at which the threshold is fixed can be estimated by taking into account the disease prevalence and the mode of inheritance (Salgado, Bellgard, et al. 2016). A fourth step could use the predictors for fitness, deleteriousness, pathogenicity and conservation as extended evidence for variants. The researchers should be aware about the high discrepancies among predictors and understand the strengths and limitations of each approach. Finally, the last step of variant prioritization may require further exploration of the functional activity of the remaining candidates.

The results from filtering and prioritization are not a definitive evidence for pathogenicity. Other classes of genetic and experimental evidence are needed to implicate a variant to disease. In this sense, it is important to distinguish the terms “associated”, “damaging” and “pathogenic” to describe the filtered-in variants. A variant is associated with a disease when it is enriched in cases versus controls; and it is damaging when disrupts the coding protein, interferes with its function or diminishes the gene expression. Still, this does not imply evidence for causality. This term only applies to pathogenic variants with a mechanistically contribution to disease (MacArthur et

al. 2014).



**Figure 14. Flowchart of a possible variant filtration and prioritization pipeline.** Adapted from Salgado et al. 2016.

One of the most critical steps in clinical research is variant interpretation, which evaluates the potential causality of prioritized variants to disease phenotype, mainly based on expertise assessment and literature review (Eilbeck et al. 2017). The exponential growth of genetic data has also increased the complexity of variant interpretation, evidencing the need for consensus within the community. In order to assess objectively the pathogenicity evidence, it is strongly recommended to follow the most updated standards and

guidelines for variant interpretation from the American College of Medical Genetics (ACMG) (Richards et al. 2015). These guidelines provide a classification system for the clinical significance of variants relevant to Mendelian disease, which encompasses five categories: benign, likely benign, uncertain significance, likely pathogenic and pathogenic. The term likely establishes a 90% certainty threshold of a variant either being benign or pathogenic. This system of categories and classes follows a set of scoring rules based on objective and subjective criteria. For instance, in the pathogenic category, the evidence is ranked in four classes of evidence: supporting (i.e. variant present in a reputable source), moderate (i.e. variant absent in population databases), strong (i.e. functional studies show a deleterious effect) and very strong (null variant in a gene with previously reported LoF mechanism of disease). The classification system is consistent in all individuals, as pathogenicity assessment is invariable for each variant. These guidelines are limited by the heterogeneity and the rareness of disease, being not possible to provide quantitative evaluations for all variants. This fact explains the major abundance of variants of uncertain significance (VUS) in databases. The ACMG classification have been recently adopted by the ClinVar and OMIM resources, which in addition present other informative categories. For example, ClinVar incorporates the risk factor variants contributing to complex disease.

The guidelines for investigating the causality of sequence variants extend the recommendations in five areas: the study design, the implication at gene-level and at variant-level, the publications and

databases; and the clinical diagnosis (MacArthur et al. 2014). The most important considerations from this publication are discussed below.

The study design is highly conditioned to the expected genomic architecture of the disease (See Section 4) and the analytical methods must be adapted accordingly. To avoid reporting false-positives, large sample sizes are required. Still, one large family may be enough to discover a disease-causing variant by using the power of segregation.

As for the gene-level, several requirements are demanded. In any study, the first deleterious variants to investigate are those found in previously reported genes. In the case of new genes being reported, evidence must replicate in unrelated individuals. Also, it is required the development of a statistical framework to give insights on the gene burden. This approach requires the creation of a null model with whom to compare the expected number of loss-of-function mutations in a certain gene. As already demanded in the last step for variant prioritization, the usage of functional data is also essential to report gene contribution to disease. At variant-level, the statistical support for association and co-segregation are essential. However, the bioinformatics predictors on function and conservation cannot be considered evidence for pathogenicity, although being informative for deleteriousness. The implication of a variant in a pathogenic mechanism requires validation results through phenotype recapitulation and the wild-type rescue in an animal model.



The deposit of genotype and phenotype data from any published study is needed to guarantee the integrity, the reproducibility and the compliance with the scientific good practices. The technical conditions of the deposit must always comply with the ethical approval and the patient consent. As for the implications of pathogenic findings in the clinical setting, a widespread claim to proceed with caution is demanded, particularly when certain drug targets may be actionable for patients.

The implication of variants to disease is promoting the development of benchmarking tools to validate the clinical findings. Still, all these approaches rely on our understanding of disease: the definition, the classification and the genetic architecture.

## **1.5. Human disease**

### **1.5.1. Disease definition and classification**

What is a disease? There is not even consensus among physicians to that answer. Disease comes from the Old French “desaise” word, which literally means lack of ease. In general terms, a disease is an abnormal condition of an organism that impairs body functions. However, its definition is highly-context dependent, as it is created in relation to people (Scully 2004). Social and cultural components such as class, gender, ethnic group, but also economical or historical reasons can completely alter its meaning. Thus, disease is often used as a condition that causes pain, dysfunction, distress, social problems or death, including emotions. Even though, not all authors accept this

definition. According to the World Health Organization (WHO), disease and health are not antonyms, as the latter comprises a bigger scope. In the Preamble of the Constitution of WHO adopted in 1946, they considered health as “the state of complete physical, mental and social well-being and not merely the absence of disease or infirmity”. Remarkably, this definition has not been amended since then, but discussion on the term keeps taking place. A disease also underlies a relative mismatch between an organism and the environment, reinforcing the context-dependence of the definition (McKusick 1975). Other authors consider disease as a significant deviation from the normal phenotype (Coleman and Tsongalis, 2017). This deviation can be observed through systematic analysis of patient symptoms or signs measurements.

Classification of diseases is a system of categories to which abnormal function is assigned according to an established criterion. Many different criteria exist, being the nosology the discipline that addresses this question. Historically, diseases have been classified by clinical symptoms or by the organ they affect. The contemporary classification to define the syndromic phenotype depends mainly on observational skills and on laboratory tests. This procedure is limited by a lack of sensitivity and specificity, as the disease cannot be predicted before is manifested and additionally, it can be confounded in the presence of comorbidity (Loscalzo, Kohane, and Barabasi 2007). These limitations are linked to the insufficient information currently available on basal elements to fully understand disease

(Kim 2006). The aetiology and pathogenesis have been raised as alternative classifiers, although for most diseases they are unknown.

The most general classification based on aetiology divides diseases in two groups according to their communicability. Thus, diseases caused by infective agents such as bacteria, viruses or parasites are distinguished from the rest, a non-homogenous group including genetic, heritable and acquired, physiological and nutritional diseases. Another classification takes into account the molecular origin of diseases categorizing them among genetic or exogenous, which in a certain manner reflects the traditional genetics versus nurture dilemma. In this classification, exogenous diseases comprise infections, intoxications and nutritional problems, while genetics considers any change in germinal or somatic cell lines. Nonetheless, the international standard diagnostic classification in medicine is the *WHO International Statistical Classification of Diseases and Related Health Problems Tenth Revision (ICD-10)*, which translates diagnoses and other health problems from words to alphanumeric codes. The purpose of ICD is to build a systematic tool that records, interprets and compares morbidity and mortality data collected in different countries at different times.

The definition of disease is extremely important when studying the underlying genetic mechanism for diagnosis. A general definition, including a wide range of disease subtypes, is meaningless from this perspective, as different underlying causes encompasses genetic heterogeneity. For instance, blindness should never aggregate

information on cataracts, corneal defects or retinitis pigmentosa when being studied (Mitchell 2012).

### **1.5.2. Genetic disease**

A genetic disease is caused by one or more alterations in the DNA sequence affecting particular genes or due to chromosomal abnormalities. Following this broad-sense definition, all non-communicable diseases are indeed genetic, as DNA is directly or indirectly the ultimate cause triggering a disorder. Genetic diseases are not necessarily transmitted across generations in families, as they can be observed exclusively in one individual, in forms of *de-novo* or somatic mutations. When referring to the passing of genetic characteristics from one generation to another, we talk about heredity. It is not straightforward to determine if a certain genetic disease segregates within a family. Disease inheritance implies that related individuals are affected by the same condition due to the segregation of germ-line cell mutations. Consequently, an extended family history is usually required for genetic diagnosis.

Another term of main importance is disease heritability, which applies to the proportion of the disease expressivity that is explained by genetic factors and ranges from 0 (genes do not contribute to disease) to 1 (genes are the only reason for disease). To estimate this parameter, four different methods are used (Tenesa and Haley 2013), including the twin method (Table 3). This method compares the phenotypic resemblance among monozygotic and dizygotic twins, which respectively share 100% and 50% of their genome. Indirectly,

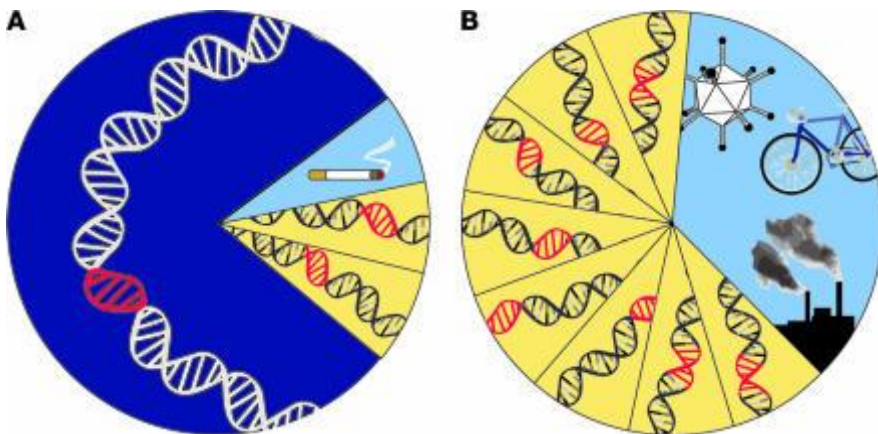
disease heritability provides a measure of the contribution of the environment to disease. Genetic diseases can be highly heritable and at the same time be poorly inherited. This means that they can be mostly driven by genes, but at the same time the exact combination of genes to cause the disease can be hardly transmitted. The heritability and inheritance of a disease are highly dependent on the disease itself and even under specific genomic architectures, different subtypes of the same disease can differ.

**Table 3. Heritability estimates of traits and disease-related phenotypes from twin studies.** Reprinted from: (Van Dongen et al. 2012)

Trait	Heritability	Number of twin pairs (or study type for multiple data sets)*
<i>Metabolic and cardiovascular</i>		
Diabetes, type 1	0.88	22,650
Diabetes, type 2	0.64	13,888
Coronary heart disease	M: 0.57; F: 0.38	10,483
Systolic blood pressure	0.42	1,617 <sup>h</sup>
Diastolic blood pressure	0.40	1,617 <sup>h</sup>
Markers for cardiovascular disease in blood		12,000 twins
High-density lipoprotein (HDL) level	0.66	
Low-density lipoprotein (LDL) level	0.53	
Triglyceride level	0.54	
Glucose level	0.53	
C-reactive protein (CRP) level	0.43	
<i>Brain and central nervous system disorders</i>		
Alzheimer's disease	0.48	662
Parkinson's disease	0.34	46,436 twins
Migraine	0.34–0.57 <sup>h</sup>	29,717
Multiple sclerosis	0.25–0.76 <sup>h</sup>	Review
Attention-deficit hyperactivity disorder	0.76	Review
Autism spectrum disorders	0.71	11,535 twins
Schizophrenia	0.81	Meta-analysis
Major depression	0.37	Meta-analysis
<i>Electroencephalography measures of brain activity</i>		
Alpha power	0.79	Meta-analysis
P300 amplitude	0.60	Meta-analysis
<i>Magnetic resonance imaging measures of brain structure</i>		
Total brain volume	0.66–0.97	Review
Frontal lobe volumes	0.90–0.95	Review
Hippocampal volumes	0.40–0.69	Review
<i>Skeletal features and disorders</i>		
Bone mineral density	0.60–0.80	Review
Osteoarthritis	0.40–0.70	Review
Rheumatoid arthritis	0.60	13,502

### 1.5.3. Inherited disease

Human inherited disease is classified according to the genetic contributors that cause the condition. Many examples of inherited disease are governed by individual genes (monogenic disease), while others are the result from polygenic contributions and environmental exposures (multifactorial or complex disease). Also, certain diseases have been linked to mitochondrial genes (Chinnery and Hudson 2013). The two former models have been traditionally used to classify diseases in a categorical way according to two main paradigms: a rare variant in a causal gene as the main contributor to Mendelian disease and common variants in many genes acting as risk factors ruling common disease (Figure 15).



**Figure 15. Genetic and environmental contributions to monogenic (A) and complex (B) disorders.** Reprinted from: (Manolio, Brooks, and Collins 2008)

### 1.5.3.1. Monogenic disease

Monogenic (also called *Mendelian*) disorders are the simplest category of human inherited disease. By definition, they are caused by the dysfunction of a single gene with large effects on disease outcome. They typically segregate within families following a Mendelian mode of inheritance (MOI), which can present different forms (Figure 16). Depending on the chromosome where the disease locus is located, we distinguish autosomal and X-linked disorders, and depending on the pattern of inheritance, they can be either dominant or recessive, having respectively one or two copies of the disease-causing mutation. In the autosomal recessive inheritance, we distinguish the homozygous and the heterozygous form. In the former case, the two recessive alleles in trans are the same and they are located at the same position of the locus. In the latter, also known as compound heterozygosity, the two alleles are located at the same locus, but at different positions causing the genetic disease in a double heterozygous state. This shows that two unrelated alleles in the same locus can be defective in combination. The disease-causing mutations can also appear *de-novo* as a result of a genetic alteration in the parental germ line cells, which may prompt the observation of affected individuals from unaffected parents.

The inheritance of a trait may also be caused by mutations in the mitochondrial genome. This non-mendelian inheritance follows an uniparental transmission, as the disease-causing mutations can only be inherited from the maternal line. This is explained by the fact that mitochondria are exclusively inherited from the oocytes.

It is estimated that above 10,000 human inherited diseases are monogenic (WHO) and show low or extremely low prevalence on population (Table 4). Consequently, they are also known as rare diseases. When considered collectively, they affect 1 in 50 individuals having a remarkable impact on morbidity, mortality and healthcare cost (Boycott et al. 2015).

The severity and the impact of monogenic disease on individual fitness are highly dependent on the deleterious effect of the DNA change. In monogenic disorders, the causal variant usually alters the protein-coding sequence or changes crucial regulatory sequences that affect protein structure, function or abundance. In those cases, the disease status is determined by the required gene dosage to achieve normal function (Rice and McLysaght 2017), the allele dominance and the disease penetrance (Cooper et al. 2013). For instance, certain diseases are triggered by the loss of function of one of the alleles, being insufficient the expression of the other functional copy to show the normal phenotype (haploinsufficiency). In other cases, the disrupted allele interferes with the wild-type allele impeding the normal function (dominant-negative effect).

The variants underlying monogenic disease that reduce the reproductive fitness of carriers are under strong purifying selection (Dudley et al. 2012). Even though, they can be maintained in population due to the existing mutation-selection balance (Reich and Lander 2001) as a result of mutational recurrence, heterozygosity tolerance and disruptive environmental changes.



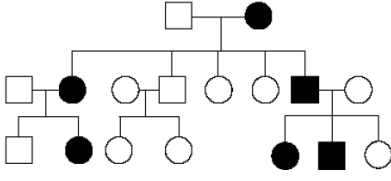
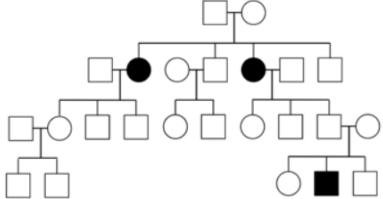
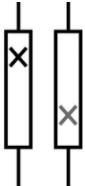
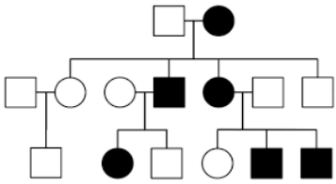
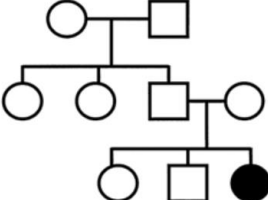
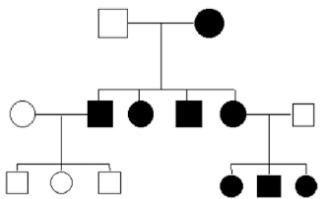
Mendelian inheritance	Pattern	Disease example
<i>Autosomal dominant</i>		Huntington's disease (Myers 2004)
<i>Autosomal recessive homozygous</i>		Cystic fibrosis (Cutting 2015)
<i>Autosomal recessive heterozygous</i>		Hemochromatosis (Rossi et al. 2000)
<i>X-linked</i>		Rett syndrome (Weaving et al. 2005)
<i>De novo</i>		Autism spectrum disorders (Ronemus et al. 2014)
<i>Mitochondrial</i>		Leber's hereditary optic neuropathy (Man, Turnbull, and Chinnery 2002)

Figure 16. Modes of inheritance in Mendelian disease.

Thus, in exceptional cases, the heterozygous carriers of these causal variants can have a selective advantage over wild-type homozygotes when they confer resistance towards lethal infectious disease, as observed in sickle cell anaemia with malaria, in cystic fibrosis with cholera or in Tay-Sachs disease with tuberculosis (Withrock et al. 2015). In this sense, the equilibrium shifts an allelic spectrum from removal or nearly removal to intermediate frequencies in population.

There are cases in which the best predictors for disease, given a certain genotype, are the age of onset (Huntington’s disease) or the diet (phenylketonuria).

**Table 4. Examples of the variability on rare disease prevalence.** Adapted from: Orphanet, 2017. (\* Data originally not-available)

<b>Mendelian disease</b>	<b>Estimated prevalence per 100,000 people</b>	<b>Number of surveyed cases / families</b>
Heritable pulmonary arterial hypertension	0.08	*
Cystic fibrosis	7.4	*
Huntington’s disease	2.7	*
Familial hyperthyroidism due to mutations in TSH receptor	*	28 families
Acute intermittent porphyria	0.54	*
Sickle cell anemia	22	*
Prata-Libéral-Gonçaves syndrome	*	2 cases

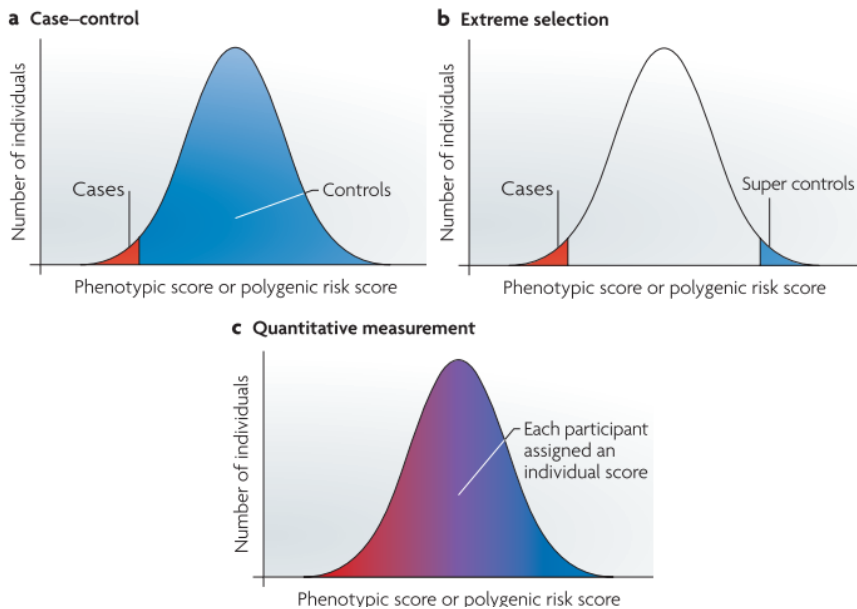
Monogenic diseases were the first to be studied in human genetics (Sections 1.1.3 and 1.1.4). The cases of study were mainly the affected families whose disease inheritance pattern could be traced back properly. The development of genetic linkage analysis as a statistical method (Section 1.6), the creation of recombination maps and the expansion of genotyping tools fostered the golden age of disease-causing gene discovery. Still, the completion of this discovery has not finished as the growth of rare variants is likely to continue, bringing new pathogenic variants into analysis. The existent knowledge makes monogenic disease the best candidate for future applications of gene editing techniques. Still, major limitations in genetic counselling need to be resolved to become potential future treatments (Section 5).

The worldwide research contributions on rare diseases are collected in different databases. OMIM is the reference database to consult information on monogenic disease (Amberger and Hamosh 2017). It currently hosts 5,098 single gene disorders and traits generated by 3,485 gene mutations (release: January 2018). The lack of correspondence between the number of disorders and mutations is given by the fact that gene mutations can cause more than one phenotype. Also, the same disease can be explained by different mutations in the same causal gene, being cystic fibrosis the paradigmatic example of that (Cutting 2015).

### **1.5.3.2. Complex disease**

There are many diseases that even having a genetic component, they are not likely to be explained by a single genetic cause. Complex or multifactorial disorders, such as hypertension, Alzheimer's disease, obesity or diabetes, are caused by the combination of multiple genes with lifestyle and environmental factors (Craig, 2008). They are normally observed as unrelated cases in population, but certain subtypes tend to segregate within families without showing patterns of Mendelian inheritance. In these cases, the risk to develop a certain condition usually increases with the proximity to an affected relative (Mitchell 2012). Nevertheless, this genetic liability is insufficient to diagnose the disease. The incidence of common diseases increases with aging, accounting also for an important fraction of mortality, morbidity and health care expenditure (Lupski et al. 2011).

The study of complex diseases has always been a matter for dispute among Mendelians and biometricians (Plomin, Haworth, and Davis 2009). Fisher resolved partially this debate in 1918 by suggesting that complex traits are caused by the involvement of multiple genes of small effect, which generate a continuous distribution per trait in population. According to this polygenic framework, the disease condition is just observed at the extremes of normally distributed traits (Figure 17). Hence, the major challenge is to use the right quantitative descriptor for disease and also to set up the proper threshold to differentiate affected cases from healthy controls.



**Figure 17. The quantitative perspective of complex disease from phenotypic and genotypic viewpoints.** This perspective regards complex disease as the extremes of a continuous distribution measured by a particular phenotypic score. a) GWAS classical case-control study. b) Extreme selection alternative strategy. c) Phenotype quantitative dimension study. Adapted from: (Plomin et al. 2009)

Complex diseases have been mainly studied from the common disease – common variant (CD-CV) hypothesis, that considers disease as the result of the contribution of many common variants (MAF>1%). By definition, common variants cannot show high penetrance on a deleterious trait as this would imply a large degree of affection in population (Bush and Moore 2012). Still, many complex diseases appear after the reproductive age and consequently, these variants are under very weak purifying selection, favouring their increase in population frequency (Spataro et al. 2017) (Wright et al. 2003). Another explanation for common variants to contribute

to complex disease is the balancing selection observed when the heterozygous state confers an advantage under certain circumstances (Pierre and Génin 2014). Moreover, changes in the direction of selection can trigger antagonist pleiotropy, in which harmful variants in old ages have shown beneficial at early ages (Rodríguez et al. 2017). Nevertheless, it exists an inverse correlation between the allele frequency of highly deleterious variants and the population prevalence. Consequently, the CD-CV hypothesis (Hemminki, Försti, and Bermejo 2008) only applies to variants of small effect (additive or regressive) that collectively impacts on interconnected biochemical pathways and non-linear networks that regulates homeostasis. Many combinatorial interactions within the cell, jointly with environmental exposures, is what finally determines the disease onset. As the number of possible variant combinations is enormously large, every individual is considered to be unique in that sense.

The genetic architecture that suggests the CD-CV hypothesis cannot be properly addressed by genetic linkage analysis due to the low power to detect variants of modest effect (Manolio et al. 2009). Instead, genome-wide association studies (GWASs) have been extensively used during the last decade.

GWASs compare the allele frequencies of common variants between a large set of cases versus controls. They are based on the idea that if a variant predisposes to disease, then it is expected to be enriched in cases (Visscher et al. 2017). GWASs mainly rely on the accumulation of recombination and natural genetic diversity in population to use

linkage disequilibrium (LD) blocks, SNPs that non-randomly correlate among them, to perform the association tests. The statistical power of a GWAS is affected by the sample, the variant effect size and the frequency of the disease-causing genetic variants. Statistically, one GWAS fits a model for each SNP to test the null hypothesis of no association with the phenotype. To reduce the number of false positives, GWASs are corrected for population stratification, familial relatedness and multiple testing (Pearson TA and Manolio TA 2008).

The association studies have been successful in identifying many loci contributing to complex diseases in cases such as type 2 diabetes (Fuchsberger et al. 2016), inflammatory bowel disease (De Lange et al. 2017) or blood pressure (Warren et al. 2017). In 2010, in response to the large growth of GWAS data, it was created the NHGRI-EBI *GWAS Catalog* (MacArthur et al. 2017), a curated resource with more than 3,500 published GWASs (September 2018).

Despite this advance, it is not straightforward to predict disease risk only from GWAS results. The identified variants in genotyping chips usually tag other causal SNPs in LD. In addition, many GWAS signals (>80%) fall in non-coding regions (Tak and Farnham 2015), suggesting that changes on gene expression through regulation rather than protein structure modifications are behind a large fraction of these findings (Lowe and Reddy 2015). Yet, the most important limitation on complex disease is the small capacity from GWAS findings to explain the observed phenotypic variability due to

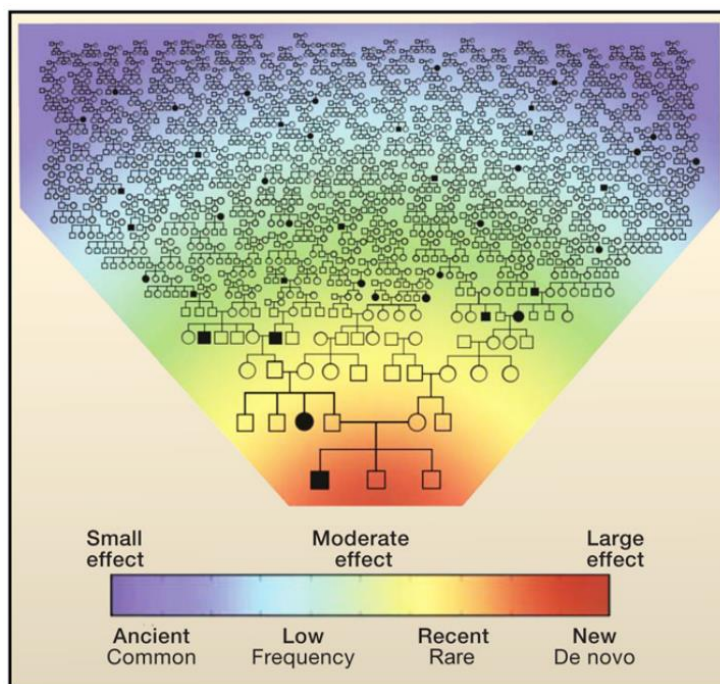
additive genetic factors, known as the missing heritability problem. This is mainly attributed to the large fraction of yet unknown common variants of small effect (Manolio et al. 2009).

Other explanations have argued that current heritability estimates are inflated as they were originally calculated from familial clustering without adjusting for the environment (Muñoz et al. 2016). Also, some authors suggest that structural variation must account for a large fraction of unexplained heritability, as recently observed in schizophrenia and autism (Stankiewicz and Lupski 2010). Epistasis could be another confounder for GWAS, as association studies cannot deal with the synergistic and antagonistic effects of genetic interactions. GWASs have discovered many new loci contributing to complex disease during last years. This is mean to continue, as future studies will better account for population diversity, will endow larger sample sizes and will properly address structural variation and environmental exposures (Murcray, Lewinger, and Gauderman 2009). Still, filling the gap of unexplained heritability will probably require the exploration of new hypothesis such as the one described as “common disease – rare variants” (CD-RV).

The CD-RV suggests that complex disease is caused by rare variants of large effect (Bomba, Walter, and Soranzo 2017), instead of common variants with tiny contribution. The genome of an individual includes many different types of mutational burden, but just a few may be responsible for the traits, mainly involving highly-penetrant variants that emerged in recent history in the family or clan, from



parents or de-novo (Lupski et al. 2011). Although many loci can confer subtle susceptibility risks to complex disease, medically actionable alleles are restricted to the rare variation according to the CD-RV hypothesis. In addition, the complexity of common disease would be explained by the high heterogeneity of these recent mutations observed in individuals (Figure 18). Under this scenario, GWAS would omit all this rare variation as a consequence of the microarray chip design. Even in the cases where causal rare variants were tagged by common variants in LD (Sun et al. 2011), their signal would get diluted in population studies. In other words, if the causal variants were meant to be rare or even private to certain individuals or families (Lupski et al. 2011), looking for shared susceptibility in groups that do not share such variants becomes worthless. Therefore, using data from whole genome sequencing would be a better approach for explaining more heritability and perform fine-mapping of causal variants (Wu et al. 2017), although the cost increases exponentially. Similarly, the concept of common disease has been put in quarantine (Maher 2008), as what we see as the same disease in unrelated population, might only be a collection of shared symptoms with different genetic backgrounds.



**Figure 18. Heatmap and multi-generation pedigree showing the relationship among the age and frequency of variants with their effect size in disease.** Recent mutations tend to show higher effect on the individual’s genetic risk as suggested by the CD-RV hypothesis. Image extracted from: (Lupski et al. 2011)

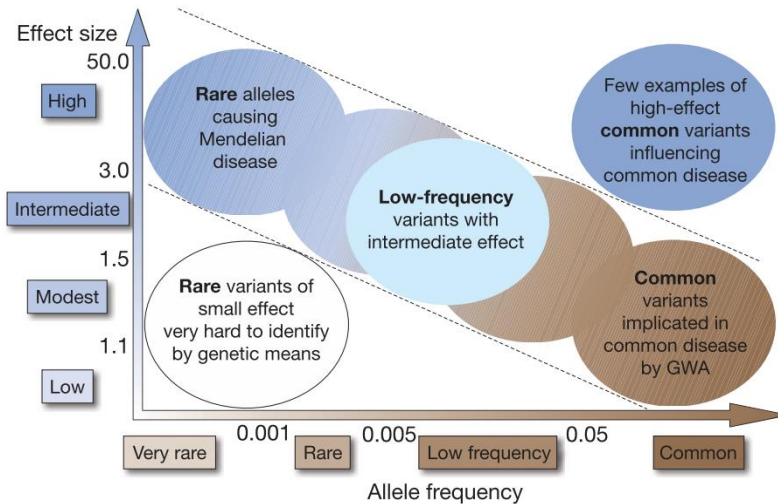
### 1.5.3.3. New perspectives on the genetic architecture of disease

Some authors have argued that the distinction among monogenic and complex disease is just a mere representation and that genetic diseases follow a continuum among those two extremes (Dipple and McCabe 2000). Regarding this architecture, all disease categories are explained by a spectrum of variants ranging from ancestral, recent or de-novo origin. This conceptual framework considers that the effect of a variant decreases as the number of variants needed to explain the

disease increases. It is accepted that the allelic architecture varies among traits depending on the number, type, effect size and the frequency of variants conferring susceptibility (Manolio et al. 2009). A continuum space among both categories would also provide insights on the missing heritability problem. Using the allele frequency as a surrogate for the effect size (Figure 19), we can distinguish several levels of variant rareness ranging from low-frequencies (0.5-5%) to extremely rare ( $10^{-7}$ ). This variant space can be operational for many variants of moderate penetrance effect, being sufficiently rare not to be present in chip arrays. This architecture would reduce significantly the number of variants needed to explain current estimates of heritability.

This variant spectrum would also be relevant to understand the contribution of modifier genes in digenic or oligogenic inheritance, in which they account for much of the observed variability. Most of the Mendelian diseases share genetic features of complex disease, reinforcing the idea of a continuum. Different phenomena such as clinical heterogeneity, variable expressivity, reduced penetrance, incomplete dominance or even codominance precludes the observation of an expected phenotype from a given genotype (Cooper et al. 2013). In such cases, the identification of modifiers would be strongly important to better understand the genetic architecture of rare disease. For instance, in widely studied cases of cystic fibrosis caused by the *CFTR* gene (Ivanov et al. 2018), there is a large variation on the clinical expressivity depending on the

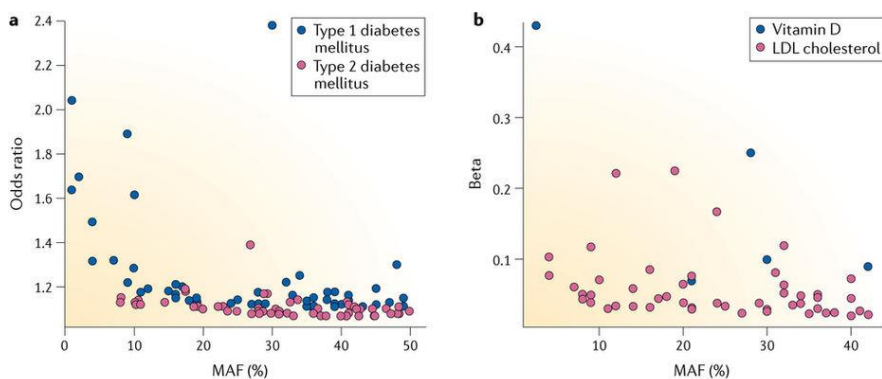
underlying pathogenic mutation. This also reveals different molecular mechanisms potentially driven by locus heterogeneity.



**Figure 19. Allelic spectrum of human inherited disease.** The two classical genetic categories of disease (Mendelian disease and complex disease) are represented, respectively, in the top-left and bottom-right corners. Linkage analysis and genome-wide association studies have been respectively the conventional tools to discover the contributing loci in each category. In between, there is an allelic spectrum of low-frequency variants with intermediate effect that must account for digenic and oligogenic inheritance, where most of allelic architectures would be located. Reprinted from: (Manolio et al. 2009)

Observing different genetic architectures is already possible in certain cases of common disease that have been extensively studied by GWAS (Timpson et al. 2017). These findings fit with this continuum transition hypothesis among Mendelian and complex diseases. For example, type I and type 2 diabetes mellitus differ in the number of loci, the allele frequency and the effect size (Figure 20a). Similarly, vitamin D shows an oligogenic structure (few loci

with large effects) and LDL cholesterol an omnigenic one (>50 loci with a broad distribution of effect sizes) (Figure 20b).



**Figure 20. Large GWAS reveals different genomic architectures for different complex diseases and complex traits.** (a) Genome-wide significant SNVs for type 1 and type 2 diabetes mellitus. Larger effects to disease risk (higher odds ratio) are observed in type 1, which also accumulates more rare variants (<0.1 minor allele frequency, MAF). (b) Genome-wide significant SNVs for the two biochemical traits, vitamin D and LDL cholesterol. Vitamin D is associated with few genetic variants, with relatively large effects. Reprinted from: (Timpson et al. 2017)

The next steps for unravelling the genetic architecture of disease are now focusing on the information provided by large sequencing efforts, jointly with different sources of functional data to better assess the impact of known loci. The main challenge is to integrate the disease risk information from common and rare disease to perform better genetic counselling in families and individuals.

## 1.6. Reduced penetrance in human disease

### 1.6.1. Penetrance as a path from genotype to phenotype

During the last decades, research efforts in clinical genomics have led to a notable growth of scientific literature about the genetic underpinnings of human disease. This growth has boosted the number of pathogenic variants associated with rare disease in databases such as OMIM, ClinVar or the Human Gene Mutation Database (HGMD) (Stenson et al. 2017). Although the pathogenicity of many of these variants is still considered valid for genetic counselling, an important fraction of them have been recently reclassified as harmless or benign thanks to the completion of the ExAC and gnomAD projects. These genetic studies have provided an unprecedented level of resolution for the allele frequency of human mutations highlighting the incongruity of annotating relatively frequent variants as lethal. This disagreement has revealed an apparent abundance of false positives in the aforementioned databases (Hayden 2016).

Penetrance is a concept that establishes the connection between genotypes and phenotypes or, in a narrow sense, between genotypes and diseases. The penetrance is measured as the proportion of individuals showing a disease phenotype divided by the number of carriers of the disease-causing mutation, also referred to as causal genotype. Statistically speaking, the penetrance is the conditional probability  $P(x|g)$  that an individual with a given genotype expresses the phenotype  $x$  (Ott, 1985). When all individuals carrying the causal genotype are affected, we talk about complete penetrance. On the

contrary, when only a fraction of carriers exhibits the disease symptoms, we refer to reduced or incomplete penetrance (Figure 21). One of the best-known examples of reduced penetrance is the phenylketonuria disease, which is caused by mutations in the *PAH* gene that inactivate the phenylalanine hydrolase enzyme. This metabolism dysfunction increases the concentration of phenylalanine to toxic levels in brain causing irreversible mental disability, which can be avoided with an appropriate diet, thereby reducing its penetrance (Blau, Van Spronsen, and Levy 2010). Reduced penetrance can also be observed because of germ line mosaicism, different age at disease onset or inaccurate classification of clinical symptoms (Hung et al. 2011). However, in the vast majority of disorders, the mechanisms ruling reduced penetrance remain unknown (Zlotogora 2003).

Although reduced penetrance is easily detected with disorders following an autosomal mode of inheritance, it can also occur with the recessive mode, in which higher allele frequencies are tolerated (Cooper et al. 2013). The mutations in the *GJA3* gene for congenital cataracts (Burdon et al. 2008) and in the *HFE* gene for hemochromatosis (Beutler 2003) are, respectively, examples for both modes of inheritance.

The phenomenon of reduced penetrance has been traditionally established as one of the main confounders in the analysis of simple and complex disease to the point of questioning the effectiveness of genetic counselling (Holtzman; and Marteau 2000). Nevertheless,

this phenomenon has remained underestimated during decades as only individuals and families affected by diseases have been considered for analysis.

Mode of inheritance		<i>Recessive</i>	<i>Recessive</i>	<i>Dominant</i>	<i>Dominant</i>
Penetrance		<i>Complete</i>	<i>Reduced</i>	<i>Complete</i>	<i>Reduced</i>
<b>Genotypes</b>	<i>+/+</i>				
	<i>m/+</i>				
	<i>m/m</i>				

**Figure 21. Complete and reduced penetrance in diseases with recessive and dominant mode of inheritance.** Genotypes: Wild-type (+) and mutated alleles (m). Phenotypes: Healthy (white) and affected individuals (black).

Most of the pathogenic mutations were incorporated in databases ignoring if the same genotypes were present in asymptomatic individuals. This lack of healthy controls generated an ascertainment bias in genetic studies. The large-scale genotyping and sequencing methods have partially resolved this issue by reversing genetics. It is now possible to assess the impact of a given genotype to a phenotype, instead of identifying a causal genotype given a disease (Xue et al. 2012). With this new paradigm, reduced penetrance has emerged as a widespread phenomenon in human inherited disease. In this sense,



the efforts to estimate its magnitude are essential to evaluate individual disease risks.

The classical approach to estimate the penetrance of a disease comprised all those cases reported in literature, often implying different mutations from the same causal gene. Nowadays, this strategy has shifted as penetrance is considered to be a function of the specific mutation or the genotype involved (Shawky 2014). Several studies have focused only on severe childhood disorders to estimate penetrance as the age of onset and the symptomatology are more tightly controlled. An study with 104 unrelated individuals encompassing only 448 paediatric diseases and 437 causal genes showed that on average each genome harbours 2.8 pathogenic mutations, ranging from 0 to 7 mutations (Bell et al. 2011). Another publication followed a strict protocol to identify individuals resilient to lethal mutations (Chen et al. 2016). In this effort, the considered panel for mendelian disorders (>6000, OMIM) was strongly reduced since diseases annotated with incomplete penetrance or unknown severity were filtered out, remaining only 584 rare disorders. This outlines that many reported Mendelian variants are substantially less penetrant than previously assumed (Flannick et al. 2013).

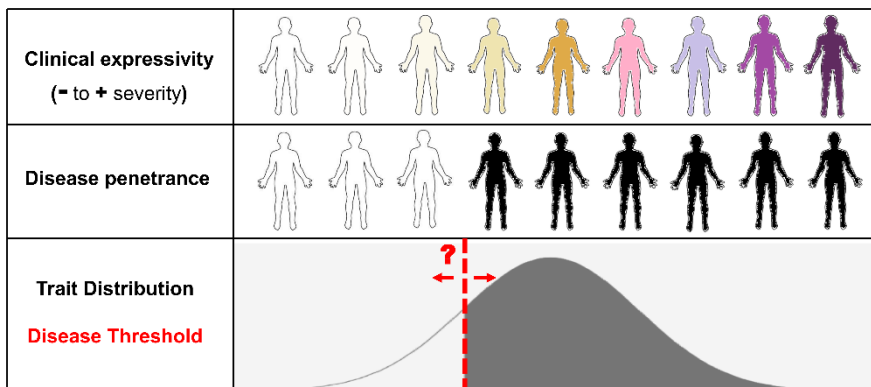
Other contributions have directly examined large cohorts to estimate the disease penetrance. For instance, with the estimation of penetrance for prion disease variants (Minikel et al. 2016), a basic principle was applied: fully penetrant causal genotypes cannot be more common in population than the prevalence of the disease they

cause. Following such approach, an excess of pathogenic variants in controls with the expectation of these variants being fully penetrant reveals the presence of either likely benign or low penetrance variants. The revaluation of the prion disease variants showed an spectrum of penetrance from  $<0.1$  to  $\sim 100\%$  (Minikel et al. 2016). These results may indicate that variants associated with reduced penetrance follow a risk continuum rather than a dichotomy of pathogenic versus benign, blurring the distinction between monogenic and complex disease (Sidransky 2006).

Another method to estimate penetrance is to directly calculate the mutational burden of reportedly pathogenic variants from control datasets. Only in the pilot phase of the 1000 Human Genome Project, between 0 and 8 highly damaging HGMD mutations (0-1 in homozygosity) were carried by each individual (Xue et al. 2012). At an more advanced phase, individuals were estimated to carry on average 24 to 30 ClinVar variants implicated in rare disease (Auton et al. 2015). This estimation was also performed using exome-sequencing in the ExAC project. 22,765 variants previously reported as pathogenic in ClinVar and HGMD were searched in the exomes of the 60,706 individuals. On average, 5.8 pathogenic variants were harboured per exome after filtering for false positives ( $>1\%$  allele frequency in at least one continental population), with 0.89 variants falling in known autosomal disease genes and 2 in autosomal recessive ones (Lek et al. 2016). These results implausibly suggest that most of the individuals should have a rare mendelian disorder, strengthening the idea that reduce penetrance is the norm, rather than

the exception in human genetics (Cooper et al. 2013).

The concept of penetrance is strongly imbricated with clinical expressivity, which is a different function of disease condition. The idea of penetrance is based on a binary output: the presence or not of a disease associated with a causal genotype. On the contrary, the clinical expressivity reflects the degree of variation of the disease phenotype. This is basically a range of signs and symptoms overlapping a trait distribution. A priori, all cases with a certain degree of expressivity are considered penetrant, but at low levels, it is difficult to determine if a variant is penetrant or not. This is usually represented as a trait distribution with a threshold for disease onset (Figure 22).



**Figure 22. Clinical expressivity overlaps the trait distribution, being the disease only penetrant above a certain threshold of abnormal function.**

The deleterious effect of pathogenic mutations is not always observed in all carrier individuals, and when it is manifested, it does not affect in the same way (Lobo 2008). Thus, patients with the same

pathogenic mutation can show differences on penetrance, as in retinoblastoma (J. William Harbour 2001), Huntington's disease (Quarrell et al. 2007), breast cancer (Gareth et al. 2008) or heritable pulmonary arterial hypertension (James White and Morrell 2012); or they can show large differences on the disease symptoms and severity, reflecting variable expressivity. This is the case of the *FBNI* mutations in Marfan syndrome (Arslan-Kirchner et al. 2010), which can range from mild symptoms to life-threatening complications. Many other diseases, such as neurofibromatosis (Sabbagh et al. 2009) or holoprosencephaly (Collins et al. 1993) show variable expressivity. The differences in penetrance and expressivity are caused by several factors which respond to the molecular basis of each disease. Both phenomena are closely related, and they are likely to share common mechanisms (Ahluwalia et al. 2009) (van Heyningen and Yeyati 2004) (Zlotogora 2003).

### **1.6.2. The molecular basis of reduced penetrance**

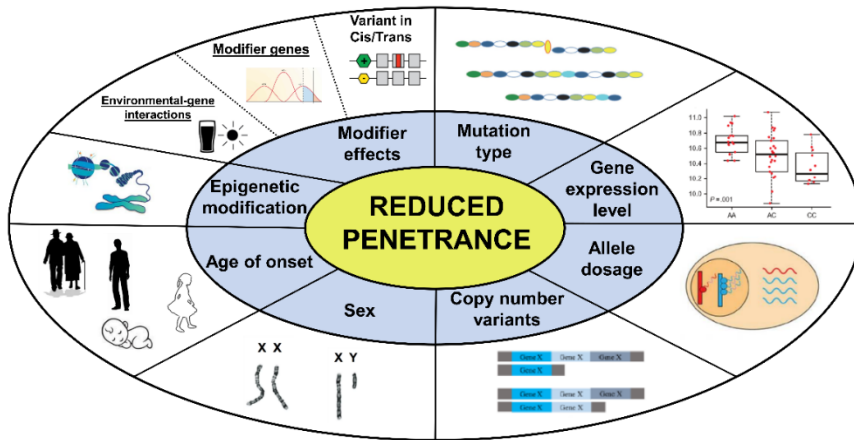
Understanding the reason why healthy individuals can tolerate a burden of pathogenic variants without suffering from disease is a challenge. Many of those variants have an impact on the biochemical function of genes, but this does not necessarily translate into an effect on health status (Xue et al. 2012). Additionally, the observation of protein-coding gene knockouts in population also provide evidence that the loss of certain genes is compatible with life (Lek et al. 2016), highlighting the complexity underlying the concept of penetrance.

Many disorders lack specific information on the molecular basis of

penetrance, implying a poor predictive power of genotypes for phenotypes, and therefore, impeding genetic counselling. Still, recent efforts have elucidated some of these mechanisms by explaining reduced penetrance as a combination of genetic, environmental and lifestyle factors. There are at least eight mechanisms proposed to modulate penetrance, including the type and molecular context of mutations, the patient characteristics such as age or sex; or the environmental conditions, such as diet or exposure to contaminants and pathogens. Below, we introduce the different mechanisms proposed in a review article about reduced penetrance (Cooper et al. 2013). Noteworthy, some of the distinguished categories overlap and could be interchangeably classified (Figure 23).

#### **1.6.2.1. Mutation type**

Many disorders are caused by pathogenic mutations identified in more than one gene, indicating the existence of locus heterogeneity. In those cases, the average penetrance of mutations differs among genes, as well within the same gene. For instance, the average penetrance for breast cancer at the age of 70 years is 65% for *BRCA1* gene mutations and 45% for *BRCA2* gene (Antoniou et al. 2003). Likewise, the mutation Arg1699Gln in *BRCA1* is associated with strongly reduced penetrance (24%) compared to the average penetrance in that same gene (Spurdle et al. 2012).



**Figure 23. Mechanisms of reduced penetrance.** Adapted from: Cooper et al., 2013

The impact of mutations on the protein-coding sequence is also important to determine its penetrance. Different consequences on disease manifestation have been observed from missense, truncating, splicing and even null mutations. An example of a disease influenced by the type of mutation is heritable pulmonary arterial hypertension (HPAH), discussed in the *Results* section. In this case, the missense mutations observed in *BMPR2* are linked to stronger severity and lower age of onset than the truncating variants in that same gene (Austin et al. 2009). The detrimental effect of missense variants is explained by the dominant-negative effect on the receptor structures, which does not occur with the truncated proteins. Also, the localization of the mutations within different domains of the protein structure is sometimes decisive on penetrance, as observed in prion disease (Minikel et al. 2016). Finally, the clinical penetrance of recessive disorders strongly depends on the nature of the mutation in the other haplotype copy, enhancing or rescuing the detrimental effect

when the alleles are different.

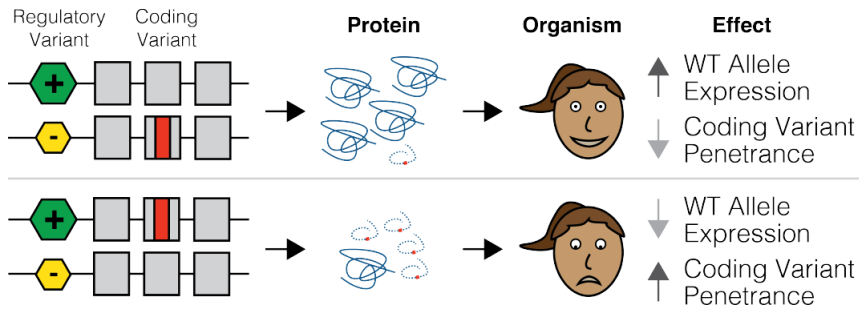
### **1.6.2.2. Modifier effects**

Many variants are insufficient to cause the disease alone, as they are conditionally pathogenic. This means they need to segregate with other genetic variants or occur under certain environments to manifest the disease. The modifier factors of penetrance can arise at the variant, gene or environmental level or as a combination of them.

#### **Variant modifiers in *cis* and *trans***

Regulatory variants can alter the penetrance of pathogenic mutations by altering the expression of the genes carrying such mutations. They can act either in a local or a distant way, both encompassing *cis*-acting and *trans*-acting factors (Rockman and Kruglyak 2006).

The regulatory variants acting in *cis* can exert a modulating effect on penetrance by changing the expression of the allele copy harbouring the pathogenic mutation. The allelic imbalance (Figure 24) that favours the expression of the dysfunctional allele over the wild-type one, reduces the dosage of functional gene product causing higher penetrance and higher disease risk (Lappalainen et al. 2011). For instance, an intronic enhancer SNP (rs2596623) in the thyroid hormone receptor  $\beta$  (*THRB*) gene stimulates over-expression of the pathogenic Arg338Trp pituitary-specific isoform of the receptor (TR  $\beta$ 2), leading to the phenotype of pituitary cell-specific resistance to thyroid hormone (Alberobello et al. 2011). This example shows that this mechanism is likely to manifest in a tissue-specific fashion.



**Figure 24. The cis-regulation of a gene can modulate the penetrance of pathogenic mutations.** Reprinted from: Castel et al. 2017.

The level of penetrance in cis-regulation depends on four factors: the functional importance and the dosage-sensitivity of a gene, the type of the pathogenic mutation impairing the same gene; and the effect size of the cis-regulation on expression. The relevance for this mechanism in human disease is demonstrated by the action of purifying selection on reducing the haplotype combinations of regulatory and coding variants associated with higher penetrance in general population (Castel et al. 2017). Likewise, the datasets with affected individuals are enriched in haplotypes associated with higher disease risk, lower functional dosage and lower fitness.

Other examples of regulatory variants acting in cis comprise intronic mutations modifying gene splicing. These variants can lead to aberrant splicing by creating cryptic acceptors and donors, by disrupting the wild-type existing ones or by breaking enhancer and silencer regions. Other regulatory regions such as 5'UTR, 3'UTR and non-coding regions are candidates to harbour regulatory variants with modifying effects on penetrance.



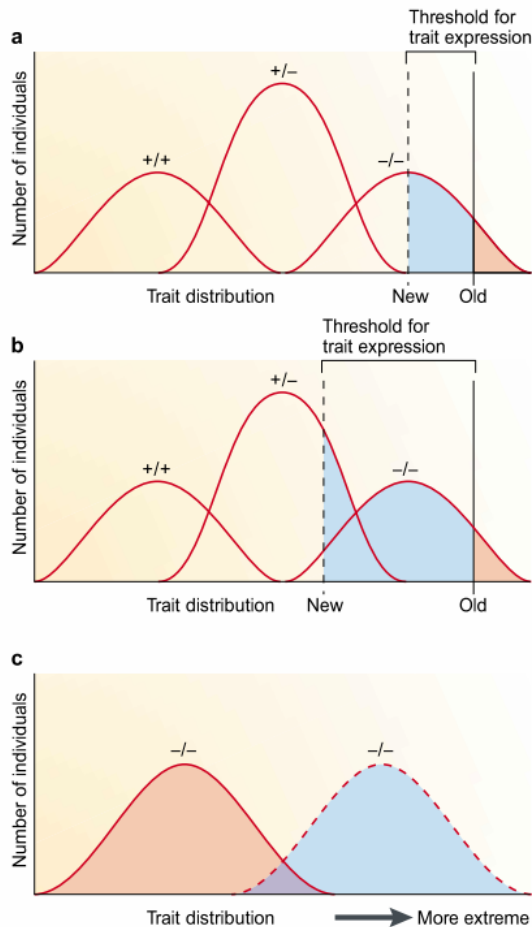
As for trans-acting variants, many examples support this mechanism despite being less characterized. For instance, two unlinked SNPs from the *TF* and *HFE* genes are shown to cause an increased risk for Alzheimer disease (Kauwe et al. 2010). More examples will be displayed in the gene modifier section.

### **Modifier genes**

Many disorders, once described as monogenic disease, are not genetically as simple as they were initially described (Dipple and McCabe 2000). This is partially explained by unlinked modifier genes (Steinberg and Sebastiani 2012) that influence the penetrance, the dominance and the expressivity of the majority of inherited diseases. The action of modifiers genes can also influence the risk and the age of onset for diseases with primarily responsible genes. This modification can act at any level influencing a trait, from transcription to intermediate phenotypes affecting both molecular and cellular levels.

Different models of modification have been proposed based on the idea that modifier genes may change the threshold for the trait or the disease expression (Nadeau 2001). Modifier genes can be protective when they reduce penetrance by moving this threshold to the right (Figure 25a) or confer susceptibility and higher penetrance by moving the threshold to the left. If this susceptibility contribution is high enough, it can even change the heterozygotes phenotype generating a dominance modification (Figure 25b). Likewise, modifiers can also shift the trait distribution of pathogenic carriers to

more (or less) extreme phenotypes modifying the clinical expressivity (Figure 25c).



**Figure 25. Models of modification for penetrance (a), dominance (b) and expressivity (c).** Reprinted from: (Nadeau 2001)

Many examples of unlinked modifier genes through digenic or oligogenic inheritance are reported in inherited disease. In digenic inheritance, mutations compromising two functional genes are required for explaining complete penetrance. This concept should not be confounded with mutation coinheritance, which takes place when

one of the two mutations can independently lead to the disease phenotype. The digenic interaction is observed in cases where different genes encode different subunits of the same multimeric protein, an oligomeric protein complex or two proteins that interact functionally with each other (Cooper et al. 2013). This inheritance can also be detected in functional pairs such as receptor/ligand or transcription factor/transcription factor binding site. In other scenarios without functional pairing, the digenic interaction can occur at the regulatory, biosynthetic or degradative level. Haemochromatosis is a disease example of digenic inheritance implying mutations at the *HFE* and *CYBRDI* genes (Constantine et al. 2009).

In oligogenic inheritance, multiple genes are required to be mutated to manifest the disease, blurring the frontiers between monogenic and complex disease. The gene mutations contributing to this inheritance have a net functional loss or gain effect on the phenotype onset. Consequently, the difficulty in understanding oligogenic inheritance resides in the interpretation of how these genes interact to modulate the clinical penetrance. The Bardet-Biedl syndrome follows an oligogenic inheritance, as a list of 17 genes are known to contribute to the penetrance and expressivity of the disease (Badano et al. 2006).

### **Gene-environment interaction**

Environmental factors can modify the penetrance and the expressivity of pathogenic mutations. These factors adopt many different forms, including the diet (Ramachandrappa and Farooqi

2011) or other activities such as smoking, alcohol intake, drugs or physical activity (Hunter 2005). Also, life episodes with exposition to certain pathogens or sun light, as well as living traumatic experiences or living at high altitude can respectively modulate the penetrance for infectious disease (Bach 2002), melanoma (Scherer and Kumar 2010), psychological disorders (Uher 2014) and oxygen-dependent traits (Astrom et al. 2003). Environmental factors can also interact differently depending on the sex (Hunter 2005). The most compelling evidence for environmental influence is found in cancer susceptibility, particularly in those cases of lung cancer where individuals that do not smoke have significantly less risk to develop the disease (Brennan, Hainaut, and Boffetta 2011). The known gene-environment interactions are based on suspected models of exposure, but much more cases of monogenic disease will be elucidated as the ability to identify and measure such interactions is improved (Aschard et al. 2012).

### **1.6.2.3. Gene expression level**

Gene expression levels show inter-individual differences because of the action of cis/trans-acting factors, the environmental interactions, and the presence of stochastic variability. As observed previously with variant modifiers, the differential allelic expression may boost either the effects of the deleterious or the wild-type allele in autosomal dominant inheritance, increasing or reducing the clinical penetrance. For instance, the penetrance for pulmonary arterial hypertension disease caused by *BMPR2* mutations partially depends on the level of expression of the wild-type *BMPR2* allele (Hamid et

al. 2009).

Some splicing mutations can generate truncated transcripts, but a fraction of wild-type production can be retained due to the binding affinity competition for the spliceosome. The different levels of gene expression in individuals can translate this phenomenon into different levels of penetrance. In the same line, alternative splicing may underly cases of reduced penetrance by a similar mechanism (Cogan et al. 2012).

#### **1.6.2.4. Allele dosage**

The penetrance of mutations depends on their genomic context, but also on the allele dosage. This explains why some mutations exhibit low penetrance in heterozygosity and much more severe expressivity and penetrance in homozygosity. For example, changes on the allele dosage due to mutations in the *PKDI* gene are associated with the initiation of the polycystic kidney disease, theoretically described as an autosomal dominant disorder (Rossetti et al. 2009).

#### **1.6.2.5. Copy number variation**

Structural variation is implicated in many disorders and encompasses the highest number of polymorphic genomic positions in the human genome (Sudmant et al. 2015). Copy number variants (CNVs) is a class of structural variation that contributes to disease susceptibility by gaining or losing genome regions ranging from 50 bp up to 3 Mb (MacDonald et al. 2014), including entire sequences of genes. Recently, CNVs are gaining importance in the study of the genetic

architecture of psychiatric (Bassett, Scherer, and Brzustowicz 2010) and development disorders (Marshall and Scherer 2012).

In this context, the estimates for penetrance are predicted from the CNV size, the genomic location and the context variants (Carvill and Mefford 2013). Still, they present high heterogeneity even when the same disease is considered. CNVs can modulate penetrance by altering the gene dosage or by interfering the immediate regulatory regions (Shawky 2014). In general, CNVs reduce the penetrance of a pathogenic mutation by providing in cis an extra wild-type copy of the causal gene. This genetic robustness is comparable to the functional redundancy and compensation dosage observed from homologous genes (Hsiao and Vitkup 2008). An example of CNV ameliorating effects is observed in spinal muscular atrophy, where multiple copies of the *SMN2* gene partially compensates the severity associated with the loss of the *SMN1* gene (Wirth et al. 2006).

#### **1.6.2.6. Sex**

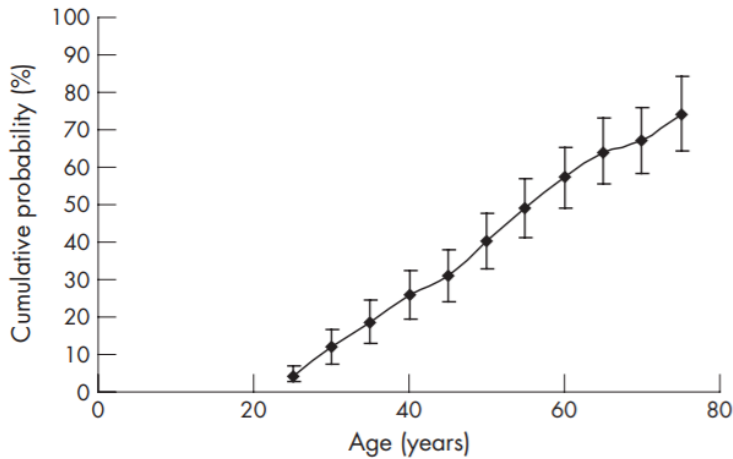
Certain disorders are more likely to manifest in a sex-specific manner due to the differential gene regulation between males and females, especially involving sex steroid-responsive genes (Dimas et al. 2012). For instance, oestrogens levels in females contribute to higher penetrance for hereditary pulmonary arterial hypertension (Austin et al. 2009). Another mechanism in which sex modulates penetrance is through genomic imprinting. With this phenomenon, an epigenetic modification silences the expression of either the maternal or the paternal allele, which essentially may correspond either to the mutant

or the wild-type copy. Depending on which copy is silenced, the risk for the disease is increased or decreased.

#### **1.6.2.7. Age of onset**

Many carriers with germline pathogenic mutations do not express disease at birth, but they can manifest it during lifetime, increasing the risk for carriers with the aging. The age of onset can vary considerably, not only among mutations causing the same disease, but also with mutations from the same gene or even when considering the same mutation. The best example to illustrate the relevance of the age of onset in the phenomenon of penetrance is Huntington's disease. In this disorder, most of the carriers only show the condition after midlife with an average onset at 40 years (Myers 2004), ranging from almost zero penetrance in youth to almost complete penetrance in old age.

The age-dependent effect on penetrance implies that an individual shifts its trait distribution towards an extreme phenotype during the lifetime. For certain disorders, this is supported by the different levels of expressivity observed among symptomatic carriers, asymptomatic carriers and healthy non-carriers individuals (Milanesi et al. 2013). The age of onset effect is usually represented with the cumulative penetrance curves, which reflects the probability of a carrier to be affected at a specific age (Figure 26).



**Figure 26. Age-dependent penetrance curve for Huntington's disease with a cohort of carriers with the 36-39 CAG triplet repeats in IT15 gene.** Reprinted from: (Quarrell et al. 2007)

### 1.6.2.8. Epigenetic modification

DNA methylation, histone modification and miRNA expression are different epigenetic mechanisms that could modulate penetrance by indirectly changing the gene expression (Wolffe and Matzke 1999). The role of epigenetics in disease was initially observed in monozygotic twin studies that showed discordant phenotypes (Gordon et al. 2012). Following this approach, DNA methylation has explained the observed reduced penetrance with the *IL4* gene in familial asthma (Soto-Ramírez et al. 2013) and other disorders affected by X-inactivation, like the *ABCD1* gene in familial adrenoleukodystrophy (Wang et al. 2013).



Reduced penetrance emerges from the complex relationship between the genotype and the phenotype. The prediction of the variant pathogenicity is a challenging task as reduced penetrance variants can easily be confounded with benign variants complicating genetic counselling. Still, reduced penetrance should not be regarded only as a problem. Further research on the mechanisms of reduced penetrance can provide insights into new therapeutic strategies if we are able to mimic the resilient effects of asymptomatic carriers on symptomatic ones.

## **1.7. Genetic linkage analysis**

Linkage analysis is a statistical genetics approach which was largely adopted at the dawn of clinical genetics to map loci involved in Mendelian disease. The development of genotyping microarrays increased the number of markers initially available and enhanced the statistical power of the technique. Although this technique has been progressively replaced by variant discovery and filtering approaches using short-read sequencing, it is emerging again as a powerful tool to detect variants responsible for disease using family-based strategies (Ott, Wang, and Leal 2015). Apart from being a low-cost option with genotyping microarrays, genetic linkage provides a model framework with statistical evidence of the variant contribution to disease susceptibility. Even more importantly, it can account for reduced penetrance, an overlooked phenomenon in variant filtering pipelines.

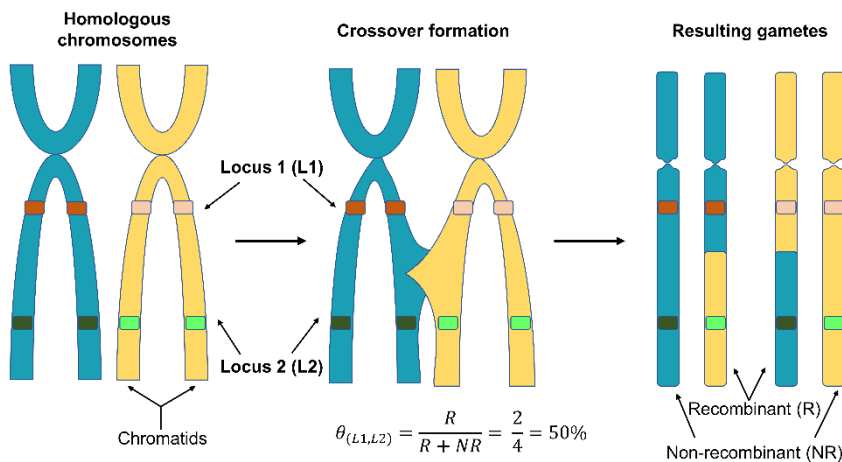
### 1.7.1. Genetic linkage and recombination

We refer to genetic linkage as the phenomenon by which loci located in a single chromosome are inherited together contradicting the Mendel's law of independent assortment. The genetic linkage is observed when the alleles from a certain haplotype are passed as intact groups through generations. For early geneticists, these groups were interpreted as the genetic equivalents of chromosomes, as different loci were observed to follow a linear order (Ott, 1999). However, this joint segregation is not observed for all the loci of a chromosome, as the phenomenon of recombination occurs.

In meiosis cell division, homologous chromosomes pair up and some physical contacts defined as chiasmata are formed between chromatids. These contacts allow the formation of the cross-bridged Holliday structure, whose resolution leads again to linear molecules, but with an exchange of genetic material between chromatids of parental chromosomes. The last meiosis step produces four sperm cells, corresponding to the two chromatids for each chromosome. In a model of a single crossover, two of these gametes are recombinant and two non-recombinants (Figure 27).

The probability of a recombination occurring between two loci is measured by the recombination fraction ( $\theta$ ). It reflects the proportion of recombinant haplotypes that a doubly heterozygous parent can produce. In absence of chromatid interference, which assumes that crossovers are random and independent among them, this recombination fraction ranges from a maximum of 50% for unlinked

loci to 0% for tightly linked ones. The maximum of 50% is obtained from considering two loci placed at the extremes of a chromosome with only one crossover being observed. In this hypothetical situation, only two gametes out of four are recombinant. The recombination fraction varies depending on the sex of the parental gametes and the region of the genome (Lenormand and Dutheil 2005), especially regarding the pseudoautosomal regions (PAR), where the sex chromosomes X and Y have homologous sequences that can recombine in meiosis.



**Figure 27. Model of single crossover in meiosis and computation of the recombination frequency ( $\theta$ ).**

Crossovers occur semi-randomly along a chromosome, as their distribution is usually biased towards recombination hotspots regions (Székvolgyi, Ohta, and Nicolas 2015). The presence of a crossover indicates that it is highly unlikely that another crossover is formed in the immediate vicinity. In the case of such small distances among two

loci, the recombination fraction is used as a measure for genetic distance in centimorgan (cM), being 1 cM equivalent to 1% of recombination fraction. This has allowed the creation of genetic maps encompassing the expected number of crossovers among the loci of a single chromosome (Botstein et al. 1980). These maps are based on the idea that two distant loci are more likely to recombine than those that are closer in a single chromosome.

The evolutionary meaning of recombination is hypothesized to be a source of variability for natural selection. In this line, the meiosis exchanges are shown to be a hallmark for mammals, contributing as a major determinant on haplotype diversity within populations (Dumont 2017).

Genetic linkage uses these two phenomena, recombination and linkage, to identify regions associated with monogenic disease. The linkage technique is based on the idea that a pathogenic mutation appears de-novo in an individual at an unknown location and segregates through a limited number of family generations. In this segregation, the recombination events statistically tend to occur more closely to the disease locus reducing progressively its extension up to 100 kb (Weiss and Clark 2002). Using genetic markers, it is possible to search the hypothetical region corresponding to the disease locus using the identical by descent (IBD) approach, in which two affected individuals share identical alleles from a common ancestor (Figure 28). This approach distinguishes genetic linkage from association studies, which are based on a much higher number

of generations and follows the identical by state (IBS) approach, in which two affected individuals have the same alleles, but not necessarily due to a common ancestor.

This gene mapping approach cannot be always applied straightforwardly, as many factors need to be arranged. For instance, not all the recombination events are informative for linkage and many different approaches, algorithms and implementations have been proposed so far. In the next subsections, we will briefly describe the most important aspects and the existing tools for genetic linkage.

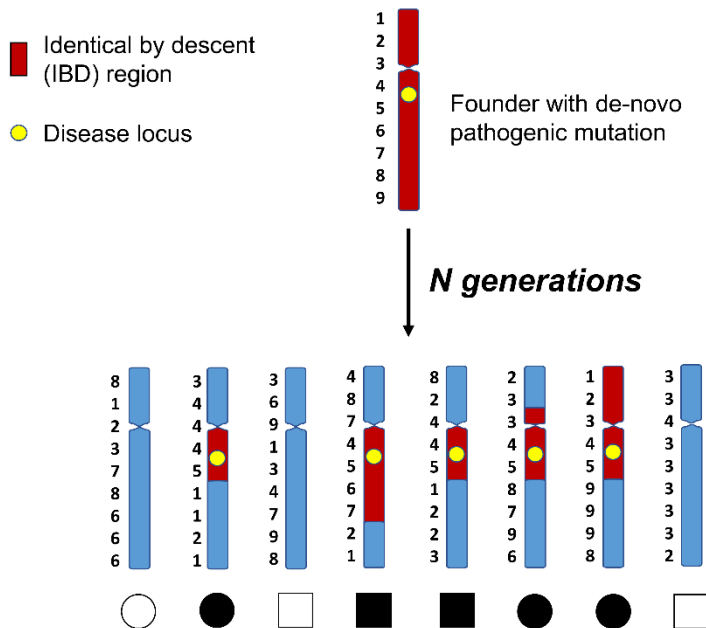


Figure 28. Conceptual core of IBD-based genetic linkage analysis.

### 1.7.2. The fundamental aspects of linkage analysis

Conceptually, linkage analysis study the joint transmission of a phenotypic character (i.e. disease) and genetic markers. To conduct a proper analysis, though, several requirements are needed.

One key aspect is the availability of genetic markers with a genome-wide uniform distribution. The informativeness of markers increases when they encompass multiple alleles, being variable number tandem repeat (VNTR) and short tandem repeats (STR) the most convenient markers for linkage. Still, the most used genetic markers are single nucleotide polymorphisms (SNPs), available in the order of millions at low cost in microarray chips. When adopting SNPs as markers, it is possible to include the population allele frequencies from human variations catalogues, such as the 1000 Human Genome Projects, to enhance linkage power detection. The linkage programs also require as input a genetic map, which contains the ordering and the genetic distances of markers in chromosomes.

Another basic component of linkage are the kinship relations encompassed in family pedigrees, which include founder and non-founder members, distinguished by the absence or the presence of parents in the pedigree. It is also common to attach the phenotypic information in pedigrees, which can adopt qualitative (affected – unaffected) or quantitative traits following a continuous distribution. This phenotyping step is challenging as symptoms for diagnoses are not always so clear due to different clinical expressivity. In most cases, however, the only solution is to use phenotypes that are expected to correlate with disease and presumably with the action of a gene. In any case, it is preferred to use quantitative trait linkage analysis when possible, as it encloses a definition closer to the underlying disease mechanism.

Before running a linkage analysis, it is necessary to perform different quality controls on genotyping data to minimize errors and inconsistencies. Using the family information, it is common to detect mendelian inconsistencies, but also pedigree problems such as swapped samples, adoption or false paternity (Neale, Neale, and Sullivan 2002).

Linkage analysis provides a genome-wide level of significance to determine the involvement of a variant in disease susceptibility by using the LOD score method proposed by Morton in 1950. This test evaluates the probability that a genetic marker cosegregates with the disease over the null hypothesis that it will occur by chance in the pedigree. In general, as closer the marker is from the disease allele, the stronger will be the cosegregation, and on the contrary, the further is located, the most chances that a recombination breaks this relationship.

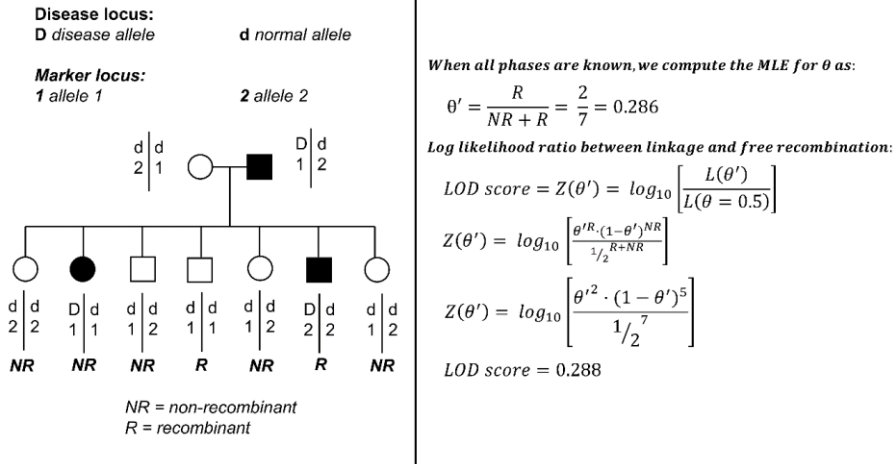
The LOD score is the log likelihood of obtaining the test data ( $X$ ) when the locus responsible for the disease maps at a given genetic distance to the marker ( $H_1$ , linkage hypothesis,  $\theta < 0.5$ ), divided by the likelihood of observing the same data if the marker and the disease are not linked ( $H_0$ , null hypothesis of free recombination,  $\theta = 0.5$ ).

$$LOD\ score = Z = \log_{10} \left( \frac{P(X|\theta < 0.5)}{P(X|\theta = 0.5)} \right)$$

To compute the LOD score, linkage programs use genetic distances that are translated into recombination fractions using map functions. As discussed above, for small distances and absence of interference, the appropriate map function establishes that genetic distances are equivalent to recombination fraction ( $x=\theta$ , Morgan's map function). Moreover, in those small intervals, recombination fractions are additive. However, when multiple crossovers occur, they are no longer additive and other map functions are required to correct for these differences, such as the Haldane or the Kosambi functions. In practice, the linkage programs look at different values of recombination fraction ( $\theta$ ) to maximize the LOD score in a step of maximum likelihood estimation (MLE). The value of recombination fraction that maximizes the LOD score ( $\theta'$ ) is then applied.

In Figure 29 we represent how the LOD score is calculated in a linkage analysis among a disease locus and a genetic marker with known phase. In the example, we assume that the penetrance is complete, there is no chromatid interference and the disease follows an autosomal dominant mode of inheritance.





**Figure 29. Example of the application of the LOD scores method in a pedigree of two-generations.**

In the example above, the LOD score is calculated analytically due to the simplicity of the example. In real situations, different factors such as reduced penetrance, population allele frequency, missing information, larger pedigree sizes and multiple generations make this task too complex to be computed in this way. To overcome this complexity, different algorithms encompassing different strategies have been implemented in the linkage programs to compute the LOD score (Table 5).

**Table 5. Summary of the main existing algorithms and their main features.**

Adapted from: Course on gene mapping: Linkage analysis by E.Sobel.

(\* Programs used in the Results section.)

Algorithm	Size restriction	Solution	Programs	Increase with the number of individuals	Increase with markers	Missing data effect
<b>Elson-Stewart</b>	~8 loci	Exact	Fastlink Linkage Mendel Vitesse	Linear	Exponential	Severe
<b>Lander-Green</b>	Aprox. 24 bits (complexity)	Exact	Allegro GeneHunter <i>Mendel*</i> <i>Merlin*</i>	Exponential	Linear	Modest
<b>Markov Chain – Monte Carlo</b>	>1000 individuals >1000 loci	Aprox.	Loki SimWalk2 <i>Morgan*</i>	Linear	Linear	Mild

LOD scores are interpreted from the perspective of an hypothesis test. Generally, a positive LOD score indicates evidence in favor of linkage and negative scores, evidence for free recombination. However, a more stringent criteria is used to determine if linkage really exists. The maximum of the LOD score is conventionally used as the measure for accepting or rejecting the null hypothesis of no-linkage, using a critical threshold of 3.3 LOD score units. This threshold is accepted to convey an adequate genomewide significance level that minimizes the number of false-positives when rejecting the null hypothesis of free recombination. Moreover, this

threshold is robust to a moderate practice of multiple testing. Still, for any positive result to be confirmed, it is a good practice to pursue other linkage tests for robustness, like selecting another set of markers or using other linkage programs if possible. Ideally, the positive linkage hit should be replicated in other families, although the low prevalence of certain diseases makes this suggestion difficult to be accomplished.

The results of a linkage analysis usually point to a locus or a region, rather than a specific variant. In fact, the regions showing significant linkage can expand more than 1 Mb and can contain more than one gene, so further filtering is always needed to identify pathogenic variants. Still, it unquestionably reduces the number of the candidate variants to follow up.

The power to detect linkage depends mainly on the magnitude of the variant contribution to the disease phenotype, but there are other factors that are also important. This includes the pedigree size and structure, and the marker informativeness for linkage, which is indirectly measured by their heterozygosity. In addition, the distance of markers to the disease locus and the genetic heterogeneity should be considered. Other factors such as genotyping errors, misdiagnoses in phenotyping or the usage of incorrect parameter models can compromise the power to detect linkage. The penetrance function and the phenocopies rates (discussed in Section 6.4) can dramatically modify the LOD score.

The genetic linkage technique presents limitations in studying complex traits, as it assumes the absence of epistasis and does not account for environmental interactions. On the other hand, it enables the identification of disease-causing genes by incorporating several parameters into a model of inheritance, including penetrance.

### **1.7.3. Types of genetic linkage analysis**

There are two main types of genetic linkage analysis: parametric and nonparametric linkage, also known as model-free linkage (Strauch et al. 2003). In linkage, the nonparametric tests are not equivalent to statistical distribution free tests. They simply refer to tests that do not need the specification of parameters for an hypothetical mode of inheritance.

#### **1.7.3.1. Parametric linkage**

Parametric linkage analysis relies on the specification of a list of parameters including the disease mode of inheritance (dominant or recessive), the disease gene frequency, which refers to the frequency of the sought allele causing the disease; the disease penetrance and the phenocopies rate. These parameters are chosen according to a hypothesis of how genotypes influence the phenotype. However, misspecifications on this model reduces the power for linkage detection. Two main parametric linkage tools can be used: two-point and multipoint linkage methods.

With two-point linkage analysis, the position of a disease-related locus is tested one marker at a time. Theoretically, this method is less powerful than multipoint linkage and it is widely used as a quick

evaluation of linkage. In spite of that, the two-point method is more flexible than multipoint as the recombination fraction is not constrained by neighbouring markers. For this reason, it is associated with a reduced computational cost and it does not present restrictions on pedigree size. Two-point linkage is highly sensitive, at the cost of generating a higher proportion of false positives. Consequently, two-point linkage results should be always followed up by multipoint linkage for robustness. Mendel (Lange et al. 2013) and Pseudomarker (Gertz et al. 2014) are the two-point linkage programs used in this thesis.

In multipoint linkage analysis, the position of a disease-related locus is tested along multiple markers, three or more, at a time. In this way, it addresses the allele transmission by considering short-range haplotypes. To that purpose, a correctly ordered genetic marker map is required. Multipoint linkage is more powerful and robust than two-point, but it is constrained by the pedigree size and the number of markers, easily reaching unfeasible computational scenarios. Two main algorithms are used to compute likelihoods in multipoint, the Elston-Stewart and the Lander-Green (Table 5). The first one scales exponentially with the number of markers and linearly with the number of individuals, while the latter performs in the opposite way. Linkage complexity is measured in bits and computed according to the number of founders ( $f$ ) and non-founders ( $n$ ) present in a pedigree, as follows:

$$\text{Complexity (bits)} = 2n - f$$

There is a maximum level of complexity that the Lander-Green algorithm can hold, also dependent on the software implementation. For instance, the algorithm is limited to 24 bits in Merlin (Abecasis et al. 2002). Above this threshold, a pruning step on the pedigree is required to be able to run the analysis. Merlin and Morgan (Wijsman, Rothstein, and Thompson 2006) are the two multipoint linkage programs that have been used in this thesis.

### **1.7.3.2. Nonparametric linkage**

The nonparametric linkage (NPL) does not require the specification of a model, since it is based on the concept of identity by descent (ibd). This method only uses information from affected individuals and produces robust results at the cost of losing information regarding unaffected individuals (Strauch et al. 2003). The parameter-free approach is useful when the mode of inheritance is unknown. Still, the “affected-only” strategy will be always limited in terms of statistical power in comparison to parametric linkage. Moreover, model-free linkage is usually equivalent to parametric linkage (Strauch et al. 2003), as shown by the statistically equivalence among certain NPL tests and parametric recessive models (Knapp, Seuchter, and Baur 1994).

Nonparametric linkage is build on descent graphs that represent the inheritance vectors among affected individuals. The method evaluates if there is more IBD sharing among affected individuals that it would be expected under the free recombination hypothesis. Several implementations from Whittermore and Halpern have been

widely used: NPL-pairs and NPL-all statistics. In the former case, it measures for specific positions, the degree of IBD allele sharing with pairs of affected, while in the latter, it measures the IBD allele sharing across all the affected individual at the same time. These implementations can be coupled with other methods to analyse IBD sharing in many small families (Linear Kong and Cox) or in a few large families (Exponential Kong and Cox). All of these implementations are available in the Merlin program.

The nonparametric tests present a major inconvenient regarding the high computational cost of large pedigrees. Also, its performance is time-consuming when it comes to the evaluation of significance.

#### **1.7.4. Modelling reduced penetrance in parametric linkage**

In many individuals, the phenotype does not exhibit the underlying genotype being necessary to account for reduced penetrance. Parametric linkage analysis is in fact the only genetic tool that can deal with reduced penetrance, as it is a necessary parameter of the linkage model. The phenomenon of reduced penetrance implies uncertainty in the disease onset, which may be the result of the action of gene-gene and/or gene-environment interactions.

In the models, penetrance is modelled as the phenotype risk given by some genotype, being usually expressed as a vector for each of the genotype components. Likewise, the penetrance can be interpreted in a narrow-sense, referring only to disease; or in a wide-sense, by considering all the possible phenotypes. The usage of the narrow or

the wide-sense definition depends on the linkage program. In the examples below (Table 6-8), we express how a fully penetrant disease with a disease allele  $D$  and a wild-type allele  $d$  would be expressed in a recessive and a dominant disorder:

**Table 6. Narrow-sense penetrance example.** Complete penetrance for each mode of inheritance (recessive and dominant) is displayed. Accordingly, for the recessive MOI, all the  $D/D$  homozygotes are affected, while in the dominant MOI, all the heterozygotes and all the  $D/D$  homozygotes are affected.

<b>d/d</b>	<b>D/d</b>	<b>D/D</b>	<b>MOI</b>
0	0	1	<b>recessive</b>
0	1	1	<b>dominant</b>

**Table 7. Wide-sense penetrance (recessive example).** The penetrance is specified for each of the phenotypes in a recessive MOI: for affected individuals is equivalent to Table 6 (row 1), for unaffected individuals is  $1-P(\text{affected})$  and for unknown individuals is 1, as the penetrance is the union of the affected and unaffected penetrance (Ott, 1999).

<b>d/d</b>	<b>D/d</b>	<b>D/D</b>	<b>Phenotype</b>
0	0	1	<b>affected</b>
1	1	0	<b>unaffected</b>
1	1	1	<b>unknown</b>



**Table 8. Wide-sense penetrance (dominant example).** The penetrance is specified for each of the phenotypes in a dominant MOI: for affected individuals is equivalent to Table 6 (row 2), for unaffected individuals is  $1-P(\text{affected})$  and for unknown individuals is 1, as the penetrance is the union of the affected and unaffected penetrance (Ott, 1999).

<b>d/d</b>	<b>D/d</b>	<b>D/D</b>	<b>Phenotype</b>
0	1	1	<b>affected</b>
1	0	0	<b>unaffected</b>
1	1	1	<b>unknown</b>

From the examples above, we can observe that the difference among the mode of inheritance in fully penetrant disease is basically the penetrance of heterozygotes. Additionally, when defining the penetrance for unknown individuals, it should be noted that the “unknown” phenotype is the union of the phenotypes “affected” and “unaffected”, so it is always equal to 1. In linkage analysis, it is common to perform a nonparametric test if the penetrance for a disease is unknown. This can be achieved by setting a small penetrance for affected individuals, which indirectly turns unaffected individuals into non-informative individuals for linkage.

Not only the individuals carrying the at-risk genotypes (D/d and D/D) can express the disease. There are individuals carrying protective genotypes that express the disease due to the action of genes at other loci or due to unknown environmental factors. These sporadic cases are known as phenocopies and they imply that the non-susceptible genotypes show a non-zero penetrance. Below, it is shown an

example (Table 9) with phenocopies (**bold**) and reduced penetrance (*italics*).

**Table 9. Narrow-sense penetrance example with phenocopies.** It is possible to observe a small fraction of phenocopies as 1% of individuals with protective genotypes (d/d homozygotes) are affected. Also, in both recessive and dominant MOI, reduced penetrance is observed, as penetrance do not reach 100% penetrance neither for the D/D homozygotes in a recessive MOI, nor for the heterozygotes and D/D homozygotes in a dominant MOI.

<b>d/d</b>	<b>D/d</b>	<b>D/D</b>	<b>MOI</b>
<b>0.01</b>	<b>0.01</b>	<i>0.83</i>	<b>recessive</b>
<b>0.01</b>	<i>0.83</i>	<i>0.83</i>	<b>dominant</b>

In linkage tests, it is common to account for phenocopies even when it is not evident that they actually exist. The inclusion of phenocopies maximizes the power to detect linkage when they exist, but they do not generate false positives when they do not exist (Strauch et al. 2003). The term phenocopy rate is widely-used to designate the penetrance of phenocopies for a certain disease, which refers to the non-zero penetrance of the protective genotypes (Ott, 1999). When phenocopies exist, it is possible to compute the penetrance ratio (g/f) by dividing the penetrance associated with genetic cases (g) and the penetrance associated with phenocopies (f). This penetrance ratio is analogous to the risk ratio in association studies. When the ratio is high, it is indicative that the disease phenotype can be easily discriminated by the underlying causal genotypes (Ott et al. 2015).

The consequence of reduced penetrance in linkage analysis is a loss of power that implies smaller values of LOD score. Moreover, when selecting smaller values of penetrance in a model, it is required to have larger sets of data to reach at least the same statistical power (Ott, 1999).

One of the issues to account for penetrance in parametric linkage is indeed to estimate its value, which can be done through segregation analysis, directly from pedigree data or by building penetrance curves to account for the age-of-onset.

When the penetrance changes according to a certain covariate, such as age or sex, it is common practice to build liability classes. For each class, a different penetrance vector is defined, and the likelihoods are computed with respect to those parameters (Table 10).

**Table 10. Age liability classes for the Charcot Marie-Tooth disease.** Adapted from: Ott, 1999. Only the penetrance for affected individuals is displayed (narrow-sense). The penetrance for the disease reaches 80% penetrance by age 80, being only 18% at age 40. Also, the penetrance for the phenocopies increase to 10% by age 80, while they are unobserved before 60.

<b>d/d</b>	<b>D/d</b>	<b>D/D</b>	<b>(age liability classes)</b>
0	0.02	0.02	<b>age&lt;20</b>
0	0.18	0.18	<b>age&lt;40</b>
0.05	0.55	0.55	<b>age&lt;60</b>
0.1	0.8	0.8	<b>age&lt;80</b>

In summary, linkage analysis is a technique that has been now used for more than two decades in genetics. Parametric linkage can account for the phenomenon of reduced penetrance and phenocopies. Both phenomena are of major importance in the identification of variants and modifiers associated with disorders, but they have been widely omitted in common variant filtering pipelines. Because of that, genetic linkage can be used before and after sequencing to account for them, especially regarding their growing importance in the genomic architecture of disease.

## 2. OBJECTIVES

- Develop a computational pipeline to search for genetic modifiers of reduced penetrance from microarray genotyping data.
- Apply the pipeline on a particular disease-case study of reduced penetrance: heritable pulmonary arterial hypertension (HPAH).
- Develop a resource to facilitate the integration of genome-wide specific scores in the context of genomic analysis.



### 3. RESULTS

#### 3.1. Genetic linkage analysis of a large family identifies *FIGN* as a candidate modulator of reduced penetrance in heritable pulmonary arterial hypertension

**Authors:** Pau Puigdevall<sup>1</sup>, Lucilla Piccari<sup>2,3</sup>, Isabel Blanco<sup>2,3</sup>, Joan A. Barberà<sup>2,3</sup>, Dan Geiger<sup>4</sup>, Celia Badenas<sup>5,6</sup>, Montserrat Milà<sup>5,6</sup>, Robert Castelo<sup>1,\*</sup>, Irene Madrigal<sup>5,6,\*</sup>

**Affiliations:** 1: Dept. of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain.

2: Hospital Clínic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain.

3: Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES), Instituto de Salud Carlos III, Madrid, Spain.

4: Dept. of Computer Science, Technion - Israel Institute of Technology, Haifa, Israel.

5: Dept. of Biochemistry and Molecular Genetics, Hospital Clínic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain.

6: Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Instituto de Salud Carlos III, Madrid, Spain.

**Under review at the *Journal of Medical Genetics***

**Supplementary Material:** Annex ([Section 6.1](#))

*\*To whom correspondence should be addressed*

Puigdevall P, Piccari L, Blanco I, Barberà JA, Geiger D, Badenas C, et al. [Genetic linkage analysis of a large family identifies FIGN as a candidate modulator of reduced penetrance in heritable pulmonary arterial hypertension.](#) J Med Genet. 2019 Jul;56(7):481–90. DOI: 10.1136/jmedgenet-2018-105669



### 3.2. GenomicScores: seamless access to genomewide position-specific scores from R and Bioconductor

**Authors:** Pau Puigdevall<sup>1</sup> and Robert Castelo<sup>1\*</sup>.

**Affiliations:** <sup>1</sup>Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona 08003, Spain

**Published in:** *Bioinformatics* (2018), 34(18):3208-3210.  
doi:10.1093/bioinformatics/bty311

**Full text URL:**

<https://academic.oup.com/bioinformatics/article-abstract/34/18/3208/4987140>

**Supplementary Material:** [Link to website](#)

**Abstract:**

Genomewide position-specific scores, such as those estimating conservation, constraint, fitness or mutation tolerance, are ubiquitous in current genome analyses. The diversity of sources and formats of these scores, as well as their size, increase the burden to use them. We present *GenomicScores*, a Bioconductor package that provides efficient storage and seamless access of genomewide position-specific scores from R, facilitating their use in genome analysis workflows.

*GenomicScores* is implemented in R and available at <https://bioconductor.org/packages/GenomicScores> under the open source ‘Artistic-2.0’ license.

*\*To whom correspondence should be addressed*

Puigdevall P, Castelo R. [GenomicScores: seamless access to genomewide position-specific scores from R and Bioconductor](#). *Bioinformatics*. 2018 Sep 15;34(18):3208–10. DOI: 10.1093/bioinformatics/bty311

## **4. DISCUSSION**

### **4.1. A pipeline to study reduced penetrance**

The practice of genetic diagnosis and counselling requires a full understanding and accountability of the genetic contribution to inherited disease. For this purpose, variant discovery and interpretation plays a fundamental role, yet current techniques often lead to uncertain results. This uncertainty mainly arises from biological phenomena such as genetic heterogeneity, pleiotropy, epistasis or composite phenotypes (Wright et al. 2018). One of the major impediments in mapping the genetic component of inherited disease is reduced penetrance, which hampers the genetic counselling of non-symptomatic patients carrying pathogenic mutations.

Although the phenomenon of reduced penetrance has been described for many decades, the observation of disease-causing variants at a large scale in healthy populations has revealed the need for better characterizing the mechanisms by which it manifests. There are no standardized procedures to study reduced penetrance, as it is likely to be dependent on the underlying mechanism of the pathogenic mutations. Recent efforts have tried to analyse the phenomenon from a population perspective, either by studying a cis-regulatory mechanism involving common variation (Castel et al. 2017) or by refining penetrance estimates of rare variants in large datasets (Wright et al. 2018).

In the latter case, higher penetrance estimates were observed for family-related carriers when compared to unrelated carriers in large cohorts (Wright et al. 2018). This observation suggests the existence of different disease mechanisms between individuals that are identical by descent (IBD) and identical by state (IBS). Such a higher degree of affection in related individuals could be explained by the cis-inheritance of the genetic modifier and its corresponding pathogenic mutation, which would be more unlikely to co-occur in population. Despite (Castel et al. 2017) shows a significant depletion of haplotype combinations leading to higher penetrance in the population, and conversely, an enrichment in disease cohorts, the penetrance differences between the detrimental and the neutral combinations remain small.

In this thesis, we have provided a strategy to identify genetic modifiers for the penetrance of disease-causing variants segregating in large families. We have run a model for genetic linkage analysis that assumes a genetic modification accounting for complete penetrance. The utility of our strategy relies in the compliance of several aspects that we proceed to discuss here below.

### **Aspects to consider before genetic linkage analysis**

Any study on reduced penetrance should begin with the re-evaluation of the pathogenicity of the variant under study. This can be achieved by following the guidelines for variant interpretation from the American College of Medical Genetics (ACMG) (Richards et al. 2015). In the PAH family case-study, the *BMP2* missense variant

(p.Arg491Gln) accomplishes the strongest criteria to be considered a pathogenic mutation: it is absent in population databases, it is a null variant located at a gene where loss of function is a known mechanism of disease; and other mutations in *BMP2* have been reported to be deleterious in mouse models (Frump et al. 2016).

Another necessary step is the estimation of the penetrance for the pathogenic variant. This can be achieved by aggregating all the individuals that carry this mutation in large population-based studies and from whom health records or systematic clinical phenotyping are available. In practice, if the variant is extremely rare in population, this estimation is restricted to one family, which should be large enough to minimize ascertainment biases. As for HPAH, the penetrance within the family (36.4%) approximately doubles the average *BMP2* penetrance (James White and Morrell 2012). Although the same variant has been reported in at least 6 independent PAH family cases (Deng et al. 2000) (Sankelo et al. 2005) (Pfarr et al. 2011) and also as a *de-novo* mutation in the idiopathic form of the disease (Machado et al. 2009) (Pfarr et al. 2013), we do not have penetrance data to be compared with. The penetrance differences between the p.Arg491Gln mutation in the reported family and other *BMP2* mutations can be informative of the underlying disease mechanism. For instance, we know that this mutation has a dominant-negative effect on the formation of the receptor, being more impairing than other mutations causing haploinsufficiency (Rudarakanchana et al. 2002). Other models also suggest that the action of non-sense mediated mRNA decay (NMD) may reduce the penetrance of

*BMPR2* variants (Frump et al. 2016). In general, rare variants located in genes showing a high probability of being loss-of-function intolerant (pLI) are likely to be more penetrant and pathogenic than those located in low pLI genes (Wright et al. 2018).

Accurate and up to date phenotyping data are fundamental for maximizing the statistical power to detect genetic modifiers of reduced penetrance. It is a common practice to use the affected and healthy qualitative traits in pedigrees. This classification is based on symptomatology and it may be biased for a fraction of cases, particularly when diagnosis is not straightforward. In general, it is better to adopt quantitative traits, as they correlate better with the underlying mechanism of disease and may account for the expressivity differences observed among the asymptomatic non-carriers, the asymptomatic carriers and the symptomatic carriers. Moreover, a quantitative trait may distinguish better those carrier individuals having a late onset of the disease. This late disease onset needs to be accounted in linkage using liability classes or other strategies, as it can reduce the ability to detect significant LOD scores. In addition to that, it is important to have a detailed understanding of the phenotype definition. Only for hypertension, three components of blood pressure can be reported (systolic, diastolic and pulse pressure), potentially involving different loci (Dubay et al. 1993) .

### **Aspects to consider during genetic linkage analysis**

After these initial steps, we applied genetic linkage analysis aiming

to account for complete penetrance when distinguishing affected carriers from unaffected ones. However, this technique requires a hypothesis for genetic modification and several considerations regarding the input family data and linkage parameters.

Traditionally, linkage analysis has lumped together families affected by the same disease, assuming it was monogenic and without considering if the underlying pathogenic variants were the same in each case. Using a SNP genotyping array as input data, this allelic heterogeneity does not affect linkage as the rare pathogenic variants are not included in the chips. In this manner, linkage detects only the relationship between loci, not alleles, through the common variants that indirectly tag the causal variant by linkage disequilibrium (Ott et al. 2015)

A completely different scenario encompasses the search for penetrance modifiers. It has been shown for many disorders, including HPAH, that several mechanisms can modulate the penetrance of the same disease leading to locus heterogeneity (West et al. 2008). This phenomenon usually involves the participation of common variation, which may underly different mechanisms depending on the identity of the primary mutation. In this context, only the families sharing the same pathogenic variant are likely to share the same modifier and consequently, they are the only ones that might be informative for linkage. In this thesis, we have only studied one family affected by HPAH, but its large size (5 generations, 22 carriers) is highly informative for searching a genetic modifier.

Parametric linkage analysis can take as input a model that assumes the participation of at least two genetic elements, knowing beforehand the condition of the carrier status. The mechanism may involve either a variant from a gene with independent and additive contribution to pathogenicity (digenic interaction), or a modifier that modulates the penetrance of the known pathogenic mutation (two-hit mechanism). In the former possibility, this variant is pathogenic and unlinked to the first one, while in the latter it is unlikely to be pathogenic and can exert its effect *in cis* or through downstream interaction with the molecular pathogenic mechanism.

The fact that at least one of the pathogenic mutations from the proposed mechanisms is known reveals that the phenocopies rate should be non-zero in a hypothetical digenic interaction. In this case, the identification of a second locus is usually limited by low penetrance, as this reduce the statistical power for detecting linkage of both loci, either by using single-locus analysis with phenocopies or by applying two-locus analysis (Strauch et al. 2003). For instance, in HPAH, we could not detect any locus, not even the known *BMPR2* one, when modelling for such digenic interaction.

Assuming a two-hit mechanism, we looked for a second variant that confers susceptibility and that could account for complete penetrance in combination with the pathogenic *BMPR2* mutation. In this context, a single-locus analysis is a better approach than a two-locus, as less parameters need to be estimated reducing misspecification errors. In a two-hit mechanism, the normal phenotyping is modified by



labelling asymptomatic non-carriers as individuals with unknown phenotype, asymptomatic carriers as healthy, and symptomatic carriers as affected. In this way, we force the model to look for the variants that distinguish healthy and affected carriers. If the genetic modifier is suspected to be under locus heterogeneity, which is highly likely when common variation is behind modulation, it is advisable to set a 1-2% of phenocopies rate to maximize the power of linkage detection (Strauch et al. 2000). The two-hit mechanism is by definition a susceptibility model, but it can be transformed into a protective model by switching to the parameters of the modifier alternative allele.

The search for penetrance modifiers can be performed via two-point or multipoint linkage analysis. Generally, the multipoint strategy is preferable since analysing multiple markers at a time is more powerful, except when the genotypes of the causal variant are available with the sequencing data (Ott et al. 2015). The performance of linkage analysis is enhanced by the addition of population allele frequencies, which is probably the more used type of genomic scores.

The value of our study resides in the identification of a candidate regulatory region that putatively modifies the penetrance of the pathogenic *BMP2* variant through a two-hit mechanism in HPAH. According to this linkage model, the mechanism is composed by two allele copies of a common regulatory variant from the *FIGN* distal promoter and the rare pathogenic mutation in *BMP2*, separated by 38 Mb in chromosome 2. This mechanism indicates a possible

downstream interaction of the two genes, rather than a mechanism involving *BMP2* and its promoter region. This observation would partially explain why several tests analysing the gene expression modification by promoter variants in *BMP2* do not correlate with the clinical manifestation of PAH (Song 2018).

### **Aspects to consider after genetic linkage analysis**

A limitation for the linkage results is that they provide a candidate region, usually encompassing several genes in the annotation, rather than a specific variant. In this thesis, we have characterized the candidate region using different functional genomics data, some of them with tissue-specific information, to narrow down the possible location of the modifier. Still, further efforts are required to decipher and validate the exact identity and mechanism of action of the genetic modifier.

### **Whole genome sequencing (WGS)**

The complete characterization of the linkage findings requires the sequencing of the candidate region to evaluate all the possible candidate variants within the region. As the region of interest is usually large (>1Mb), it is generally cheaper to sequence the whole genome instead. In that effort, it is of main importance to follow the best practices for variant pre-processing (Regier et al. 2018) and for variant calling (i.e. GATK caller †), to minimize the generation of technical errors.

---

† <https://software.broadinstitute.org/gatk/best-practices/>

Because the cost of WGS is still limiting for relatively large sample sizes, it is usually important to choose appropriately which individuals should undergo sequencing. In large families with reduced penetrance, this strategy should preferentially select carrier individuals, both affected and healthy. Again, the phenotype of unaffected carriers should be regarded with caution due to possible late onset.

Traditional filtering approaches applied to WGS data aiming to find rare pathogenic variants are inappropriate to identify common regulatory variation modulating penetrance. As an alternative, the WGS data can be analysed using two-point genetic linkage analysis, taking advantage of the family-based analysis, the LOD score statistical assessment and the prior knowledge on the inheritance model. Moreover, if a candidate region has been found previously, this strongly reduces the number of variants to follow up.

Allelic heterogeneity in WGS data has a negative effect in the analysis, since all the potentially causal variants are evaluated and this interferes with the different linkage signals. Still, there are methods such as the collapse haplotype pattern (CHP) that analyse them by aggregation. In the studies of reduced penetrance with WGS data, allelic heterogeneity does not compromise the linkage performance as only families carrying the same pathogenic mutation should be considered. On the other hand, locus heterogeneity will still be present, causing a fraction of carrier individuals to remain uninformative.

### **Using genomic scores for variant prioritization.**

The linkage results reduce again the number of variants that potentially map the modifier, but a list of candidates remain requiring further prioritization. In this context, genomic scores constitute a valuable resource to annotate and interpret variation by using information on conservation, constraint, fitness and mutation tolerance (Kircher 2014). These scores attempt to fill the gap between the data from high-throughput biochemical assays and the prediction of function by integrating statistical and machine-learning methods (Huang et al. 2017). Genomic scores may encode a single type of information, such as conservation or minor allele frequency, or combine multiple sources of information.

Using such data, however, may become difficult due to the heterogeneity of formats and the large size of the resources. In this thesis, we have developed the *GenomicScores* package (Puigdevall and Castelo 2018) to overcome such difficulties. This tool enables a common access point for the different existing resources and it also reduces the memory requirements of the data through lossy compression, while preserving its scientific integrity. Moreover, because it runs on top of R, it can be readily integrated into existing R workflows for the analysis of genetics and genomics data.

The compression proposed for genomic scores represents a simple and straightforward solution to store, manage and analyse big volumes of data associated to genetic variants. The advent of precision and personalized medicine in the next years will lead to a

further growth of genome sequencing and molecular profiling and, at the same time, the associated costs will be reduced. This will shift the limiting steps from data production to data storage and analysis, requiring further research on compression techniques for genomic data to meet the necessities of future clinical research.

### **Fine haplotype reconstruction**

In parallel with variant prioritization using genomic scores, the haplotype reconstruction is another important step to understand the mechanism of penetrance modulation. It consists in the determination of the parental origin of variants by placing them in the corresponding phase. The haplotype provides context information of the functional consequence of DNA variants and might be also helpful to identify the genetic modifiers when acting in *cis*. Haplotype phasing with computational methods can be obtained either from short-read sequencing or from the SNP genotyping arrays. There are two main methods for phasing: familial methods, that reconstruct haplotypes at a long range driven by the mendelian constraints imposed by the familial relatedness; and LD-based methods, which are useful for short-range phasing of unrelated individuals (Faux and Druet 2017).

In HPAH, we have used five different programs for phasing: Merlin (Abecasis et al. 2002), Beagle (Altshuler et al. 2010), Simwalk2 (Sobel and Lange 1996) and the duoHMM method from Shapeit2 (O'Connell et al. 2014). Still, these methods present inaccuracies due to limited reference panels or missing information in families. The new advent of long-read nanopore sequencing is expected to

substantially improve the haplotype phasing as it encompasses a low degree of uncertainty in reconstruction (Ammar et al. 2015).

### **Functional validation of genetic modifiers**

Results from computational pipelines on genotyping data still require further experimental confirmatory evidence for the contribution of genetic variants to disease. These computational tools are still limited by the inaccuracies and incompleteness of the existing data to interpret specific tissue biological contexts (Cox 2015). To overcome such limitations and to test the hypothesis of genetic modification, the prioritized variants should be validated by functional assays *in vivo*. These targeted assays evaluate the genotype-phenotype correlation by generating patient-derived models like cell cultures or animal models, and by rescuing the phenotype with the wild-type version of the gene (Rodenburg 2018). The generation of such models has been enhanced by the rapid development and simplicity of the CRISPR/Cas9 technology (Inui et al. 2014).

The utility of an animal model resides in how well the animal phenotype recapitulates the disease in the human phenotype. For that purpose, it is important to have a robust phenotyping system to interpret precisely the effects of genetic manipulation (Cox 2015). As for HPAH, *BMP2* defective mouse lines have been produced before (Ciuclan et al. 2011) (Frank et al. 2008) and a defined protocol for histological evaluation of pulmonary vascular remodelling is available (Frump et al. 2016).

An hypothetical functional in-vivo assay of the proposed two-hit mechanism for HPAH should ideally comprise three mouse types: a single mutant with the p.Arg491Gln *BMPR2* pathogenic mutation and two double mutants with the same *BMPR2* mutation with either one or two allele copies of the genetic modifier. If the putative genetic modifier was not homologous in the *FIGN* mouse promoter, an alternative would be the generation of two other mouse models. In this case, a genetic construct should overexpress or diminish *FIGN* expression to mimic the effects of the modifier to the pathogenic mutation. These models would reveal then whether the level of *FIGN* transcription modulates HPAH penetrance among *BMPR2* mutation carriers.

### **Open questions in reduced penetrance**

The study of disorders of reduced penetrance is complex because the mechanisms behind this phenomenon are diverse and mostly unknown. Its definition inherently depends on how precisely clinicians can diagnose a certain condition and how true is the assumption on the monogenic nature of the studied phenotype. Moreover, the concept of penetrance, although clinically relevant, it has not a biological meaning per se, as it is an outcome from the interaction of several biological phenomenon such as variable expressivity, composite phenotypes, pleiotropy and epistasis. This means that the observed estimates for penetrance are in general non-informative for the underlying biological mechanisms.

The pipeline discussed here represents an attempt to address a very

specific case of reduced penetrance. The proposed framework integrates old and new techniques to maximize the power to detect a genetic modifier. Still, many questions remain open due to the unknown genetic architecture of traits and diseases.

For instance, the idea of continuity among monogenic and complex disorders blurs the differences among genetic modifiers and risk factors. In our HPAH findings, we can observe a local enrichment of GWAS SNPs associated with blood pressure within the candidate region. This raises the question whether modifiers and risk factors inform about the same biological process in spite of using different tools.

Another critical issue is the phenomenon of locus heterogeneity concerning genetic modifiers. A myriad of common variants may influence the disease onset with the difficulty of detection and counselling that this entails. The problem is that the identity of such modifiers may not only change across families, but also across individuals from the same family, complicating even more the elucidation of reduced penetrance mechanisms.

Yet, studying reduced penetrance can help to elucidate the genotype-phenotype path and understand why there are carriers that develop a condition, while others remain healthy or show late onset. This information is expected not only to benefit the practice of genetic counselling, but also to reveal the mechanisms of disease providing new targets for potential future treatments.



## 5. CONCLUSIONS

1. We have developed a pipeline based on genetic linkage analysis to search for genetic modifiers of reduced penetrance accounting for the uncertainty of the age of onset and the locus heterogeneity.
2. We have identified a candidate region (2q24.2-q24.3) associated with PAH susceptibility among *BMP2* mutation carriers using the previous developed pipeline.
3. We have obtained evidence for the enrichment of the candidate region on risk factors associated with blood pressure and other cardiorespiratory traits.
4. We have obtained evidence for the impact of common variation on distal regulatory elements affecting the expression of *FIGN* gene in the same enriched region.
5. We have developed *GenomicScores*, a tool that enables a fast and efficient access to genomewide position-specific scores to be used in variant prioritization and other genetic analysis workflows.

## 6. ANNEX

### 6.1. Supplemental Material from Chapter 3.2

Puigdevall P, Castelo R. [GenomicScores: seamless access to genomewide position-specific scores from R and Bioconductor](#). *Bioinformatics*. 2018 Sep 15;34(18):3208–10. DOI: 10.1093/bioinformatics/bty311

## 7. REFERENCES

- Abecasis, Gonçalo R., Stacey S. Cherny, William O. Cookson, and Lon R. Cardon. 2002. “Merlin—rapid Analysis of Dense Genetic Maps Using Sparse Gene Flow Trees.” *Nature Genetics* 30(1):97–101.
- Adzhubei, Ivan A. et al. 2010. “A Method and Server for Predicting Damaging Missense Mutations.” *Nature Methods* 7(4):248–49.
- Aguet, François et al. 2017. “Genetic Effects on Gene Expression across Human Tissues.” *Nature* 550(7675):204–13.
- Ahluwalia, Jasmine K., Manoj Hariharan, Rhishikesh Bargaje, Beena Pillai, and Vani Brahmachari. 2009. “Incomplete Penetrance and Variable Expressivity: Is There a MicroRNA Connection?” *BioEssays* 31(9):981–92.
- Alberobello, Anna Teresa et al. 2011. “An Intronic SNP in the Thyroid Hormone Receptor  $\beta$  Gene Is Associated with Pituitary Cell-Specific over-Expression of a Mutant Thyroid Hormone Receptor B2 (R338W) in the Index Case of Pituitary-Selective Resistance to Thyroid Hormone.” *Journal of Translational Medicine* 9:144.
- Aldred, Micheala A. et al. 2010. “Somatic Chromosome Abnormalities in the Lungs of Patients with Pulmonary Arterial Hypertension.” *American Journal of Respiratory and Critical Care Medicine* 182(9):1153–60.
- Altshuler, David L. et al. 2010. “A Map of Human Genome Variation from Population-Scale Sequencing.” *Nature* 467(7319):1061–73.
- Amberger, Joanna S. and Ada Hamosh. 2017. “Searching Online Mendelian Inheritance in Man (OMIM): A Knowledgebase of Human Genes and Genetic Phenotypes.” *Current Protocols in Bioinformatics* 2017(June):1.2.1-1.2.12.
- Ammar, Ron, Tara A. Paton, Dax Torti, Adam Shlien, and Gary D. Bader. 2015. “Long Read Nanopore Sequencing for Detection of HLA and CYP2D6 Variants and Haplotypes.” *F1000Research* (0).
- Antoniou, A. et al. 2003. “Average Risks of Breast and Ovarian

- Cancer Associated with BRCA1 or BRCA2 Mutations Detected in Case Series Unselected for Family History: A Combined Analysis of 22 Studies.” *The American Journal of Human Genetics* 72(5):1117–30.
- Ardlie, K. G. et al. 2015. “The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans.” *Science* 348(6235):648–60.
- Arslan-Kirchner, Mine et al. 2010. “Clinical Utility Gene Card for: Marfan Syndrome Type 1 and Related Phenotypes [FBN1].” *European Journal of Human Genetics* 18(9):1070.
- Aschard, Hugues et al. 2012. “Inclusion of Gene-Gene and Gene-Environment Interactions Unlikely to Dramatically Improve Risk Prediction for Complex Diseases.” *American Journal of Human Genetics* 90(6):962–72.
- Astrom, Kristin, Joel E. Cohen, Joan E. Willett-Brozick, Christopher E. Aston, and Bora E. Baysal. 2003. “Altitude Is a Phenotypic Modifier in Hereditary Paraganglioma Type 1: Evidence for an Oxygen-Sensing Defect.” *Human Genetics* 113(3):228–37.
- Austin, E. D. et al. 2009. “Alterations in Oestrogen Metabolism: Implications for Higher Penetrance of Familial Pulmonary Arterial Hypertension in Females.” *European Respiratory Journal* 34(5):1093–99.
- Austin, E. D. and J. E. Loyd. 2014. “The Genetics of Pulmonary Arterial Hypertension.” *Circulation Research* 115(1):189–202.
- Austin, Eric D. et al. 2012. “Whole Exome Sequencing to Identify a Novel Gene (Caveolin-1) Associated with Human Pulmonary Arterial Hypertension.” *Circulation: Cardiovascular Genetics* 5(3):336–43.
- Auton, Adam et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526(7571):68–74.
- Avery, O. T. 1944. “Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation By a Desoxyribonucleic Acid Fraction Isolated From Pneumococcus Type Iii.” *Journal of Experimental Medicine* 79(2):137–58.

- Bach, Jean-François. 2002. "The Effect of Infections on Susceptibility to Autoimmune and Allergic Disease." *New England Journal of Medicine* 347(12):911–20.
- Badano, Jose L. et al. 2006. "Dissection of Epistasis in Oligogenic Bardet-Biedl Syndrome." *Nature* 439(7074):326–30.
- Baltimore, David. 1970. "Viral RNA-Dependent DNA Polymerase: RNA-Dependent DNA Polymerase in Virions of RNA Tumour Viruses." *Nature* 226(5252):1209–11.
- Bassett, a. S., S. W. Scherer, and L. M. Brzustowicz. 2010. "Copy Number Variations in Schizophrenia : Critical Review and New ..." *The American Journal of Psychiatry* 167(August):899–1010.
- Batut, Philippe and Thomas R. Gingeras. 2013. "RAMPAGE: Promoter Activity Profiling by Paired-End Sequencing of 5'-Complete cDNAs." *Current Protocols in Molecular Biology* (SUPPL.104):1–16.
- Bell, Callum J. et al. 2011. "Carrier Testing for Severe Childhood Recessive Diseases by Next-Generation Sequencing." *Sci Transl Med* 3(65):1–26.
- Beutler, E. 2003. "The HFE Cys282Tyr Mutation as a Necessary but Not Sufficient Cause of Clinical Hereditary Hemochromatosis." *Blood* 101(9):3347–50.
- Blau, Nenad, Francjan J. Van Spronsen, and Harvey L. Levy. 2010. "Phenylketonuria." *The Lancet* 376(9750):1417–27.
- Bomba, Lorenzo, Klaudia Walter, and Nicole Soranzo. 2017. "The Impact of Rare and Low-Frequency Genetic Variants in Common Disease." *Genome Biology* 18(1):1–17.
- Botstein, D., R. L. White, M. Skolnick, and R. W. Davis. 1980. "Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms." *American Journal of Human Genetics* 32(3):314–31.
- Boycott, Kym et al. 2015. "The Clinical Application of Genome-Wide Sequencing for Monogenic Diseases in Canada: Position Statement of the Canadian College of Medical Geneticists." *Journal of Medical Genetics* 52(7):431–37.
- Boyd, R. and Joan B. Silk. 2007. "Human Genetic Variation."

*Science* 318:1842–43.

- Brennan, Paul, Pierre Hainaut, and Paolo Boffetta. 2011. “Genetics of Lung-Cancer Susceptibility.” *The Lancet Oncology* 12(4):399–408.
- Burdon, K. P. et al. 2008. “A Novel Mutation in the Connexin 46 Gene Causes Autosomal Dominant Congenital Cataract with Incomplete Penetrance (Journal of Medical Genetics (2004) 41, (E106)).” *Journal of Medical Genetics* 45(4):256.
- Bush, William S. and Jason H. Moore. 2012. “Chapter 11: Genome-Wide Association Studies.” *PLoS Computational Biology* 8(12).
- Carvill, Gemma L. and Heather C. Mefford. 2013. “Microdeletion Syndromes.” *Current Opinion in Genetics and Development* 23(3):232–39.
- Castel, Stephane E. et al. 2017. “Modified Penetrance of Coding Variants by Cis-Regulatory Variation Shapes Human Traits.” *Nature Genetics* 190397.
- Chen, Rong et al. 2016. “Analysis of 589,306 Genomes Identifies Individuals Resilient to Severe Mendelian Childhood Diseases.” *Nature Biotechnology* 34(5):531–38.
- Chin, Kelly M. and Lewis J. Rubin. 2008. “Pulmonary Arterial Hypertension.” *Journal of the American College of Cardiology* 51(16):1527–38.
- Chinnery, Patrick Francis and Gavin Hudson. 2013. “Mitochondrial Genetics.” *British Medical Bulletin* 106(1):135–59.
- Choi, Yongwook, Gregory E. Sims, Sean Murphy, Jason R. Miller, and Agnes P. Chan. 2012. “Predicting the Functional Effect of Amino Acid Substitutions and Indels.” *PLoS ONE* 7(10).
- Ciuclan, Loredana et al. 2011. “A Novel Murine Model of Severe Pulmonary Arterial Hypertension.” *American Journal of Respiratory and Critical Care Medicine* 184(10):1171–82.
- Cogan, Joy et al. 2012. “Role of BMPR2 Alternative Splicing in Heritable Pulmonary Arterial Hypertension Penetrance.” *Circulation* 126(15):1907–16.
- Collins, A. L., P. W. Lunt, C. Garrett, and N. R. Dennis. 1993.

- “Holoprosencephaly: A Family Showing Dominant Inheritance and Variable Expression.” *Journal of Medical Genetics* 30(1):36–40.
- Constantine, Clare C. et al. 2009. “A Novel Association between a SNP in CYBRD1 and Serum Ferritin Levels in a Cohort Study of HFE Hereditary Haemochromatosis.” *British Journal of Haematology* 147(1):140–49.
- Cooper, David N., Michael Krawczak, Constantin Polychronakos, Chris Tyler-Smith, and Hildegard Kehrer-Sawatzki. 2013. “Where Genotype Is Not Predictive of Phenotype: Towards an Understanding of the Molecular Basis of Reduced Penetrance in Human Inherited Disease.” *Human Genetics* 132(10):1077–1130.
- Cox, Timothy C. 2015. “Utility and Limitations of Animal Models for the Functional Validation of Human Sequence Variants.” *Molecular Genetics & Genomic Medicine* 3(5):375–82.
- Cutting, Garry R. 2014. “Annotating DNA Variants Is the next Major Goal for Human Genetics.” *American Journal of Human Genetics* 94(1):5–10.
- Cutting, Garry R. 2015. “Cystic Fibrosis Genetics: From Molecular Understanding to Clinical Application.” *Nature Reviews Genetics* 16(1):45–56.
- Dashti, Mahjoubeh Jalali Sefid and Junaid Gamiieldien. 2017. “A Practical Guide to Filtering and Prioritizing Genetic Variants.” *BioTechniques* 62(1):18–30.
- Deng, Z. et al. 2000. “Familial Primary Pulmonary Hypertension (Gene PPH1) Is Caused by Mutations in the Bone Morphogenetic Protein Receptor-II Gene.” *American Journal of Human Genetics* 67(3):737–44.
- Devoto, Marcella et al. 2011. “Genome-Wide Linkage Analysis to Identify Genetic Modifiers of ALK Mutation Penetrance in Familial Neuroblastoma.” *Human Heredity* 71(2):135–39.
- Dimas, Antigone AS et al. 2012. “Sex-Biased Genetic Effects on Gene Regulation in Humans.” *Genome Research* 22(12):2368–75.
- Dipple, Katrina M. and Edward R. B. McCabe. 2000. “Modifier

- Genes Convert ‘simple’ Mendelian Disorders to Complex Traits.” *Molecular Genetics and Metabolism* 71(1–2):43–50.
- Dixon, Jesse R. et al. 2015. “Chromatin Architecture Reorganization during Stem Cell Differentiation.” *Nature* 518(7539):331–36.
- Van Dongen, Jenny, P. Eline Slagboom, Harmen H. M. Draisma, Nicholas G. Martin, and Dorret I. Boomsma. 2012. “The Continuing Value of Twin Studies in the Omics Era.” *Nature Reviews Genetics* 13(9):640–53.
- Doolittle, W. F. 2013. “Is Junk DNA Bunk? A Critique of ENCODE.” *Proceedings of the National Academy of Sciences* 110(14):5294–5300.
- Dresdale, David T., Martin Schultz, and Robert J. Michtom. 1951. “Primary Pulmonary Hypertension. I. Clinical and Hemodynamic Study.” *American Journal of Medicine* 686–705.
- Dubay, Christopher et al. 1993. “Genetic Determinants of Diastolic and Pulse Pressure Map to Different Loci in Lyon Hypertensive Rats.” *Nature Genetics* 3:354–57.
- Dudley, Joel T. et al. 2012. “Human Genomic Disease Variants: A Neutral Evolutionary Explanation.” *Genome Research* 22(8):1383–94.
- Dumont, Beth L. 2017. “Variation and Evolution of the Meiotic Requirement for Crossing over in Mammals.” *Genetics* 205(1):155–68.
- Dunham, Ian et al. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489(7414):57–74.
- den Dunnen, Johan T. et al. 2016. “HGVS Recommendations for the Description of Sequence Variants: 2016 Update.” *Human Mutation* 37(6):564–69.
- Eddy, Sean R. 2013. “The ENCODE Project: Missteps Overshadowing a Success.” *Current Biology* 23(7):R259–61.
- Eichstaedt, Christina A. et al. 2016. “EIF2AK4 Mutation as ‘Second Hit’ in Hereditary Pulmonary Arterial Hypertension.” *Respiratory Research* 17(1):141.



- Eilbeck, Karen, Aaron Quinlan, and Mark Yandell. 2017. "Settling the Score: Variant Prioritization and Mendelian Disease." *Nature Reviews Genetics* 18(10):599–612.
- Erica Sodergren, George M. Winstock, Er. 2006. "The Genome of the Sea Urchin." *Science* 314(February):941–52.
- Faux, Pierre and Tom Druet. 2017. "A Strategy to Improve Phasing of Whole-Genome Sequenced Individuals through Integration of Familial Information from Dense Genotype Panels." *Genetics Selection Evolution* 49(1):1–13.
- Flannick, Jason et al. 2013. "Assessing the Phenotypic Effects in the General Population of Rare Variants in Genes for a Dominant Mendelian Form of Diabetes." *Nature Genetics* 45(11):1380–87.
- Ford, C. E. 1959. "Turner 's Syndrome." *Lancet* i:711–13.
- Fortin, Jean Philippe and Kasper D. Hansen. 2015. "Reconstructing A/B Compartments as Revealed by Hi-C Using Long-Range Correlations in Epigenetic Data." *Genome Biology* 16(1):1–23.
- Frank, David B. et al. 2008. "Increased Susceptibility to Hypoxic Pulmonary Hypertension in *Bmpr2* Mutant Mice Is Associated with Endothelial Dysfunction in the Pulmonary Vasculature." *Am J Physiol Lung Cell Mol Physiol* 2372:L98–109.
- Frump, Andrea L., Arunima Datta, Sampa Ghose, James West, and Mark P. de Caestecker. 2016. "Genotype-Phenotype Effects of *Bmpr2* Mutations on Disease Severity in Mouse Models of Pulmonary Hypertension." *Pulmonary Circulation* 6(4):597–607.
- Fuchsberger, Christian et al. 2016. "The Genetic Architecture of Type 2 Diabetes." *Nature* 536(7614):41–47.
- Furey, Terrence S. 2012. "ChIP-Seq and beyond: New and Improved Methodologies to Detect and Characterize Protein-DNA Interactions." *Nature Reviews Genetics* 13(12):840–52.
- Galiè, Nazzareno et al. 2015. "2015 ESC/ERS Guidelines for the Diagnosis and Treatment of Pulmonary Hypertension." *European Respiratory Journal* 46(4):903–75.
- Gareth, D. Gareth et al. 2008. "Penetrance Estimates for BRCA1 and BRCA2 Based on Genetic Testing in a Clinical Cancer

- Genetics Service Setting: Risks of Breast/Ovarian Cancer Quoted Should Reflect the Cancer Burden in the Family.” *BMC Cancer* 8:1–9.
- Garrod, Archibald E. 1902. “The Incidence of Alkaptonuria : A Study in Chemical Individuality.” *The Lancet* 1:1616–20.
- Germain, Marine et al. 2013. “Genome-Wide Association Analysis Identifies a Susceptibility Locus for Pulmonary Arterial Hypertension.” *Nature Genetics* 45(5):518–21.
- Gertz, Edward et al. 2014. “PSEUDOMARKER 2.0: Efficient Computation of Likelihoods Using NOMAD.” *BMC Bioinformatics* 15(1):47.
- Gibbs, Richard A. et al. 2004. “Genome Sequence of the Brown Norway Rat Yields Insights into Mammalian Evolution.” *Nature* 428(6982):493–520.
- Girerd, Barbara et al. 2010. “Clinical Outcomes of Pulmonary Arterial Hypertension in Patients Carrying an ACVRL1 (ALK1) Mutation.” *American Journal of Respiratory and Critical Care Medicine* 181(8):851–61.
- Gordon, Lavinia et al. 2012. “Neonatal DNA Methylation Profile in Human Twins Is Specified by a Complex Interplay between Intrauterine Environmental and Genetic Factors, Subject to Tissue-Specific Influence.” *Genome Research* 22(8):1395–1406.
- Graur, Dan et al. 2013. “On the Immortality of Television Sets: ‘Function’ in the Human Genome According to the Evolution-Free Gospel of Encode.” *Genome Biology and Evolution* 5(3):578–90.
- Green, Eric D. and Mark S. Guyer. 2011. “Charting a Course for Genomic Medicine from Base Pairs to Bedside.” *Nature* 470(7333):204–13.
- Griffith, Fred. 1928. “The Significance of Pneumococcal Types.” *Journal of Hygiene* 27(2):113–59.
- Gulko, Brad, Melissa J. Hubisz, Ilan Gronau, and Adam Siepel. 2015. “A Method for Calculating Probabilities of Fitness Consequences for Point Mutations across the Human Genome.” *Nature Genetics* 47(3):276–83.

- Gusella, James F. et al. 1983. "A Polymorphic DNA Marker Genetically Linked to Huntington's Disease." *Nature* 306(5940):234–38.
- Gusella, James F., Marcy E. MacDonald, Christine M. Ambrose, and Mabel P. Duyao. 1993. "Molecular Genetics of Huntington's Disease." *Arch Neurol* 50:1157-.
- Hamid, Rizwan et al. 2009. "Penetrance of Pulmonary Arterial Hypertension Is Modulated by the Expression of Normal BMPR2 Allele."
- Hardy, G. H. 1908. "Mendelian Proportions in a Mixed Population." *Science* XXVIII:49–50.
- Hayden, Erika Check. 2016. "A Radical Revision of Human Genetics." *Nature* 538(7624):154–57.
- Hemminki, Kari, Asta Försti, and Justo Lorenzo Bermejo. 2008. "The 'Common Disease-Common Variant' Hypothesis and Familial Risks." *PLoS ONE* 3(6):1–6.
- van Heyningen, Veronica and Patricia L. Yeyati. 2004. "Mechanisms of Non-Mendelian Inheritance in Genetic Disease." *Human Molecular Genetics* 13(REV. ISS. 2):225–33.
- Higgins, M. et al. 1996. "NHLBI Family Heart Study: Objectives and Design." *Am. J. Epidemiol* 143(12):1219–28.
- High, K. H., A. Nathwani, T. Spencer, and D. Lillicrap. 2014. "Current Status of Haemophilia Gene Therapy." *Haemophilia* 20(S4):43–49.
- Holtzman, Neil A. and Theresa M. Marteau. 2000. "Will Genetics Revolutionize Medicine?" *The New England Journal of Medicine* 343(2):135–45.
- Hsiao, Tzu Lin and Dennis Vitkup. 2008. "Role of Duplicate Genes in Robustness against Deleterious Human Mutations." *PLoS Genetics* 4(3).
- Hu, Jing and Pauline C. Ng. 2012. "Predicting the Effects of Frameshifting Indels." *Genome Biology* 13(2).
- Huang, Yi Fei, Brad Gulko, and Adam Siepel. 2017. "Fast, Scalable Prediction of Deleterious Noncoding Variants from Functional

- and Population Genomic Data.” *Nature Genetics* 49(4):618–24.
- Huber, Wolfgang et al. 2015. “Orchestrating High-Throughput Genomic Analysis with Bioconductor.” *Nature Methods* 12(2):115–21.
- Hublin, Jean Jacques et al. 2018. “Author Correction: New Fossils from Jebel Irhoud, Morocco and the Pan-African Origin of Homo Sapiens (Nature (2017) DOI: 10.1038/Nature22336).” *Nature* 558(7711):E6.
- Humbert, Marc et al. 2006. “Pulmonary Arterial Hypertension in France: Results from a National Registry.” *American Journal of Respiratory and Critical Care Medicine* 173(9):1023–30.
- Hung, Chia Cheng et al. 2011. “Low Penetrance of Retinoblastoma for p.V654L Mutation of the RB1 Gene.” *BMC Medical Genetics* 12(1):76.
- Hunter, David J. 2005. “Gene-Environment Interactions in Human Diseases.” *Nature Reviews Genetics* 6(4):287–98.
- International Chicken Genome Sequencing Consortium. 2004. “Sequence and Comparative Analysis of the Chicken Genome Provide Unique Perspectives on Vertebrate Evolution.” *Nature* 432(December).
- International Human Genome Sequencing Consortium. 2004. “Finishing the Euchromatic Sequence of the Human Genome.” 431:931–45.
- Inui, Masafumi et al. 2014. “Rapid Generation of Mouse Models with Defined Point Mutations by the CRISPR/Cas9 System.” *Scientific Reports* 4:1–8.
- Ivanov, Maxim et al. 2018. “Targeted Sequencing Reveals Complex, Phenotype-Correlated Genotypes in Cystic Fibrosis.” *BMC Medical Genomics* 11(Suppl 1).
- J. William Harbour. 2001. “Molecular Basis of Low-Penetrance Retinoblastoma.” *Arch Ophthalmol* 119:1699–1704.
- Jackson, D. a, R. H. Symons, and P. Berg. 1972. “Biochemical Method for Inserting New Genetic Information into DNA of Simian Virus 40: Circular SV40 DNA Molecules Containing Lambda Phage Genes and the Galactose Operon of Escherichia

- Coli.” *Proceedings of the National Academy of Sciences of the United States of America* 69(10):2904–9.
- Jacobs, Patricia A. and J. A. Strong. 1959. “A Case of Human Intersexuality Having a Possible XXY Sex-Determining Mechanism.” *Nature* 183(4653):55–56.
- Jagadeesh, Karthik A. et al. 2016. “M-CAP Eliminates a Majority of Variants of Uncertain Significance in Clinical Exomes at High Sensitivity.” *Nature Genetics* 48(12):1581–86.
- Jain, Miten et al. 2018. “Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads.” *Nature Biotechnology* 36(4):338–45.
- James White, R. and Nicholas W. Morrell. 2012. “Understanding the Low Penetrance of Bone Morphogenetic Protein Receptor 2 Gene Mutations: Another Needle in the Haystack.” *Circulation* 126(15):1818–20.
- Jin, Lv et al. 2014. “Pathway-Based Analysis Tools for Complex Diseases: A Review.” *Genomics, Proteomics and Bioinformatics* 12(5):210–20.
- Jinek, Martin et al. 2012. “A Programmable Dual-RNA – Guided.” 337(August):816–22.
- Jou, W. Min, G. Haegeman, M. Ysebaert, and W. Fiers. 1972. “Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein.” *Nature* 237(5350):82–88.
- K.Mullis et al. 1986. “Specific Enzymatic Amplification of DNA in Vitro: The Polymerase Chain Reaction.” *Nature Communications* 51(0):263–73.
- Kato, Norihiro et al. 2011. “Meta-Analysis of Genome-Wide Association Studies Identifies Common Variants Associated with Blood Pressure Variation in East Asians.” *Nature Genetics* 43(6):531–38.
- Kato, Norihiro et al. 2015. “Trans-Ancestry Genome-Wide Association Study Identifies 12 Genetic Loci Influencing Blood Pressure and Implicates a Role for DNA Methylation.” *Nature Genetics* 47(11):1282–93.
- Kauwe, John S. K. et al. 2010. “Validating Predicted Biological Effects of Alzheimer’s Disease Associated SNPs Using CSF

- Biomarker Levels.” *Journal of Alzheimer’s Disease* 21(3):833–42.
- Kellis, M. et al. 2014. “Defining Functional DNA Elements in the Human Genome.” *Proceedings of the National Academy of Sciences* 111(17):6131–38.
- Kim, Jaegwon. 2006. “Emergence: Core Ideas and Issues.” *Synthese* 151(3):547–59.
- Kircher, Martin. 2014. “A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants.” *Nature G* 46(3):310–15.
- Knapp, M., S. A. Seuchter, and M. P. Baur. 1994. “Linkage Analysis in Nuclear Families: Relationship between Affected Sib-Pair Tests and Lod Score Analysis.” *Human Heredity* 44:44–51.
- Lander, E. S. et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409(6822):860–921.
- Landrum, Melissa J. et al. 2018. “ClinVar: Improving Access to Variant Interpretations and Supporting Evidence.” *Nucleic Acids Research* 46(D1):D1062–67.
- Lane, K. B. et al. 2000. “Heterozygous Germline Mutations in BMPR2, Encoding a TGF-Beta Receptor, Cause Familial Primary Pulmonary Hypertension.” *Nature Genetics* 26(september):81–84.
- De Lange, Katrina M. et al. 2017. “Genome-Wide Association Study Implicates Immune Activation of Multiple Integrin Genes in Inflammatory Bowel Disease.” *Nature Genetics* 49(2):256–61.
- Lange, Kenneth et al. 2013. “Mendel: The Swiss Army Knife of Genetic Analysis Programs.” *Bioinformatics* 29(12):1568–70.
- Lappalainen, Tuuli, Stephen B. Montgomery, Alexandra C. Nica, and Emmanouil T. Dermitzakis. 2011. “Epistatic Selection between Coding and Regulatory Variation in Human Evolution and Disease.” *American Journal of Human Genetics* 89(3):459–63.
- Lawrence, Michael et al. 2013. “Software for Computing and Annotating Genomic Ranges.” *PLoS Computational Biology*

9(8):1–11.

- Lek, Monkol et al. 2016. “Analysis of Protein-Coding Genetic Variation in 60,706 Humans.” *Nature* 536(7616):285–91.
- Lenormand, Thomas and Julien Dutheil. 2005. “Recombination Difference between Sexes: A Role for Haploid Selection.” *PLoS Biology* 3(3):0396–0403.
- Leung, Danny et al. 2015. “Integrative Analysis of Haplotype-Resolved Epigenomes across Human Tissues.” *Nature* 518(7539):350–54.
- Li, Xingwang et al. 2017. “Long-Read ChIA-PET for Base-Pair-Resolution Mapping of Haplotype-Specific Chromatin Interactions.” *Nature Protocols* 12(5):899–915.
- Lin, Maoxuan et al. 2017. “Effects of Short Indels on Protein Structure and Function in Human Genomes.” *Scientific Reports* 7(1):1–9.
- Lindblad-Toh, Kerstin et al. 2005. “Genome Sequence, Comparative Analysis and Haplotype Structure of the Domestic Dog.” *Nature* 438(7069):803–19.
- Lonsdale, John et al. 2013. “The Genotype-Tissue Expression (GTEx) Project.” *Nature Genetics* 45(6):580–85.
- Loscalzo, Joseph, Isaac Kohane, and Albert-Laszlo Barabasi. 2007. “Human Disease Classification in the Postgenomic Era: A Complex Systems Approach to Human Pathobiology.” *Molecular Systems Biology* 3(124).
- Lowe, William L. and Timothy E. Reddy. 2015. “Genomic Approaches for Understanding the Genetics of Complex Disease.” *Genome Res.* 25:1432–41.
- Lupski, James R., John W. Belmont, Eric Boerwinkle, and Richard A. Gibbs. 2011. “Clan Genomics and the Complex Architecture of Human Disease.” *Cell* 147(1):32–43.
- Ma, Hong et al. 2017. “Correction of a Pathogenic Gene Mutation in Human Embryos.” *Nature* 548(7668):413–19.
- Ma, Lijiang et al. 2013. “A Novel Channelopathy in Pulmonary Arterial Hypertension.” *New England Journal of Medicine* 369(4):351–61.

- MacArthur, D. G. et al. 2014. “Guidelines for Investigating Causality of Sequence Variants in Human Disease.” *Nature* 508(7497):469–76.
- MacArthur, Daniel G. et al. 2012. “A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes.” *Science* 335(6070):823–28.
- MacArthur, Jacqueline et al. 2017. “The New NHGRI-EBI Catalog of Published Genome-Wide Association Studies (GWAS Catalog).” *Nucleic Acids Research* 45(D1):D896–901.
- MacDonald, Jeffrey R., Robert Ziman, Ryan K. C. Yuen, Lars Feuk, and Stephen W. Scherer. 2014. “The Database of Genomic Variants: A Curated Collection of Structural Variation in the Human Genome.” *Nucleic Acids Research* 42(D1):986–92.
- Machado, Rajiv D. et al. 2009. “Genetics and Genomics of Pulmonary Arterial Hypertension.” *Journal of the American College of Cardiology* 54(1 SUPPL. 1):S32–42.
- Maher, Brendan. 2008. “Personal Genomes: The Case of the Missing Heritability.” *Nature* 456(7218):18–21.
- Maher, Brendan. 2012. “The Human Encyclopaedia.” *Nature* 489(7414):46–48.
- Mahmood, Khalid et al. 2017. “Variant Effect Prediction Tools Assessed Using Independent, Functional Assay-Based Datasets: Implications for Discovery and Diagnostics.” *Human Genomics* 11(1):10.
- Man, P. Y. W., D. M. Turnbull, and P. F. Chinnery. 2002. “Leber Hereditary Optic Neuropathy.” *J Med Genet* 39:162–69.
- Manolio, Teri A., Lisa D. Brooks, and Francis S. Collins. 2008. “A HapMap Harvest of Insights into the Genetics of Common Disease.” *Journal of Clinical Investigation* 118(5):1590–1605.
- Manolio, Teri a et al. 2009. “Finding the Missing Heritability of Complex Diseases.” *Nature* 461(7265):747–53.
- Marshall, Christian R. and Stephen W. Scherer. 2012. “Genomic Structural Variants.” 838:115–35.
- Maurano, Matthew T. et al. 2012. “Systematic Localization of



- Common Disease-Associated Variation in Regulatory DNA.” *Science* 337(6099):1190–95.
- McCarthy, Davis J. et al. 2014. “Choice of Transcripts and Software Has a Large Effect on Variant Annotation.” *Genome Medicine* 6(3).
- McKusick, VA. 1975. “The Growth and Development of Human Genetics as a Clinical Discipline.” *American Journal of Human Genetics* 261–73.
- McLaren, William et al. 2016. “The Ensembl Variant Effect Predictor.” *Genome Biology* 17(1):1–14.
- Metzker, Michael L. 2010. “Sequencing Technologies the next Generation.” *Nature Reviews Genetics* 11(1):31–46.
- Mikkelsen, Tarjei S. et al. 2005. “Initial Sequence of the Chimpanzee Genome and Comparison with the Human Genome.” *Nature* 437(7055):69–87.
- Milanesi, Elena et al. 2013. “Molecular Signature of Disease Onset in Granulin Mutation Carriers: A Gene Expression Analysis Study.” *Neurobiology of Aging* 34(7):1837–45.
- Miller, S. A., D. D. Dykes, and H. F. Polesky. 1988. “A Simple Salting out Procedure for Extracting DNA from Human Nucleated Cells.” *Nucleic Acids Research* 16(3):1215.
- Minikel, Eric Vallabh et al. 2016. “Quantifying Penetrance in a Dominant Disease Gene Using Large Population Control Cohorts.” *Science Translational Medicine* 8(322):322ra9-322ra9.
- Mitchell, Kevin J. 2012. “What Is Complex about Complex Disorders?” *Genome Biology* 13(1):1–11.
- Mohammadi, Pejman, Stephane E. Castel, Andrew A. Brown, and Tuuli Lappalainen. 2016. “Quantifying the Regulatory Effect Size of Cis -Acting Genetic Variation Using Allelic Fold Change.” *BioRxiv* 1–30.
- Mohn, Fabio, Michael Weber, Dirk Schübeler, and Tim-Christoph Roloff. 2009. “DNA Methylation.” *DNA Methylation: Methods and Protocols* 507:55–64.
- Morgan, T. H. 1910. “Sex Limited Inheritance in *Drosophila*.”

- Science* 23:120–22.
- Moyerbrailean, Gregory A. et al. 2016. “Which Genetics Variants in DNase-Seq Footprints Are More Likely to Alter Binding?” *PLoS Genetics* 12(2):1–27.
- Mueller, William F., Liza S. Z. Larsen, Angela Garibaldi, G. Wesley Hatfield, and Klemens J. Hertel. 2015. “The Silent Sway of Splicing by Synonymous Substitutions.” *Journal of Biological Chemistry* 290(46):27700–711.
- Mukherjee, Suranjana et al. 2012. “Human Fidgetin Is a Microtubule Severing Enzyme and Minus-End Depolymerase That Regulates Mitosis.” *Cell Cycle* 11(12):2359–66.
- Muñoz, María et al. 2016. “Evaluating the Contribution of Genetics and Familial Shared Environment to Common Disease Using the UK Biobank.” *Nature Genetics* 48(9):980–83.
- Murcray, Cassandra E., Juan Pablo Lewinger, and W. James Gauderman. 2009. “Gene-Environment Interaction in Genome-Wide Association Studies.” *American Journal of Epidemiology* 169(2):219–26.
- Myers, Richard H. 2004. “Huntington’s Disease Genetics.” *NeuroRx : The Journal of the American Society for Experimental NeuroTherapeutics* 1(2):255–62.
- Nadeau, Joseph H. 2001. “Modifier Genes in Mice and Humans.” *Nature Reviews Genetics* 2(3):165–74.
- Neale, M. C., B. M. Neale, and P. F. Sullivan. 2002. “Nonpaternity in Linkage Studies of Extremely Discordant Sib Pairs.” *Am J Hum Genet* 70(2):526–29.
- Newman, John H. et al. 2004. “Genetic Basis of Pulmonary Arterial Hypertension: Current Understanding and Future Directions.” *Journal of the American College of Cardiology* 43(12 SUPPL.):S33–39.
- Van Nostrand, Eric L. et al. 2016. “Robust Transcriptome-Wide Discovery of RNA-Binding Protein Binding Sites with Enhanced CLIP (ECLIP).” *Nature Methods* 13(6):508–14.
- O’Connell, Jared et al. 2014. “A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness.” *PLoS Genetics* 10(4).

- O’Leary, Nuala A. et al. 2016. “Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation.” *Nucleic Acids Research* 44(D1):D733–45.
- Obenchain, Valerie et al. 2014. “VariantAnnotation: A Bioconductor Package for Exploration and Annotation of Genetic Variants.” *Bioinformatics* 30(14):2076–78.
- Ott, Jurg, Jing Wang, and Suzanne M. Leal. 2015. “Genetic Linkage Analysis in the Age of Whole-Genome Sequencing.” *Nature Reviews Genetics* 16(5):275–84.
- Pauling, L., HA Itano, SJ Singer, and K. Wells. 1949. “Sickle Cell Anaemia, a Molecular Disease.” *Science* 110(2865):543–48.
- Pearson TA and Manolio TA. 2008. “How to Interpret a Genome-Wide Association Study.” *Jama* 299(11):1335–44.
- Penrod, Nadia M., Richard Cowper-Sal-Lari, and Jason H. Moore. 2011. “Systems Genetics for Drug Target Discovery.” *Trends in Pharmacological Sciences* 32(10):623–30.
- Pfarr, Nicole et al. 2011. “Hemodynamic and Clinical Onset in Patients with Hereditary Pulmonary Arterial Hypertension and BMPR2 Mutations.” *Respiratory Research* 12:1–10.
- Pfarr, Nicole et al. 2013. “Hemodynamic and Genetic Analysis in Children with Idiopathic, Heritable, and Congenital Heart Disease Associated Pulmonary Arterial Hypertension.” *Respiratory Research* 14(1):3.
- Pierre, Aude Saint and Emmanuelle Génin. 2014. “How Important Are Rare Variants in Common Disease?” *Briefings in Functional Genomics* 13(5):353–61.
- Plomin, Robert, Claire M. A. Haworth, and Oliver S. P. Davis. 2009. “Common Disorders Are Quantitative Traits.” *Nature Reviews Genetics* 10(12):872–78.
- Plotkin, Joshua B. and Grzegorz Kudla. 2011. “Synonymous but Not the Same: The Causes and Consequences of Codon Bias.” *Nature Reviews. Genetics* 12(1):32–42.
- Pollard, Katherine S., Melissa J. Hubisz, Kate R. Rosenbloom, and Adam Siepel. 2010. “Detection of Nonneutral Substitution Rates on Mammalian Phylogenies.” *Genome Research*

20(1):110–21.

- Portillo, Karina et al. 2010. “Study of the BMPR2 Gene in Patients with Pulmonary Arterial Hypertension.” *Archivos de Bronconeumología ((English Edition))* 46(3):129–34.
- Puigdevall, Pau and Robert Castelo. 2018. “GenomicScores: Seamless Access to Genomewide Position-Specific Scores from R and Bioconductor.” *Bioinformatics* (April):bty311-bty311.
- Purcell, Shaun et al. 2007. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.” *The American Journal of Human Genetics* 81(3):559–75.
- Quarrell, Oliver W. J. et al. 2007. “Reduced Penetrance Alleles for Huntington’s Disease: A Multi-Centre Direct Observational Study.” *Journal of Medical Genetics* 44(3):1–5.
- R Development Core Team. 2018. “R Internals.” *R Development Core Team*.
- Ramachandrappa, Shwetha and I. Sadaf Farooqi. 2011. “Genetic Approaches to Understanding Human Obesity.” *J Clin Invest* 121(6):2080–86.
- Regier, Allison A. et al. 2018. “Functional Equivalence of Genome Sequencing Analysis Pipelines Enables Harmonized Variant Calling across Human Genetics Projects.” *BioRxiv* 269316.
- Reich, David E. and Eric S. Lander. 2001. “On the Allelic Spectrum of Human Disease.” *Trends in Genetics* 17(9):502–10.
- Rice, Alan M. and Aoife McLysaght. 2017. “Dosage-Sensitive Genes in Evolution and Disease.” *BMC Biology* 15(1):1–10.
- Richard A. Gibbs, et al. 2007. “Evolutionary and Biomedical Insights from the Rhesus Macaque Genome.” *Science* 316(5822):222–34.
- Richards, Sue et al. 2015. “Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.” *Genetics in Medicine* 17(5):405–23.
- Riordan, John R. et al. 1989. “Identification the Cystic Fibrosis

- Gene : Cloning and Characterization of Complementary DNA.” *Science* 245:1066–73.
- Roadmap Epigenomics Consortium et al. 2015. “Integrative Analysis of 111 Reference Human Epigenomes.” *Nature* 518(7539):317–29.
- Rockman, Matthew V. and Leonid Kruglyak. 2006. “Genetics of Global Gene Expression.” *Nature Reviews Genetics* 7(11):862–72.
- Rodenburg, Richard J. 2018. “The Functional Genomics Laboratory: Functional Validation of Genetic Variants.” *Journal of Inherited Metabolic Disease* 41(3):297–307.
- Rodríguez, Juan Antonio et al. 2017. “Antagonistic Pleiotropy and Mutation Accumulation Influence Human Senescence and Disease.” *Nature Ecology and Evolution* 1(3):1–5.
- Romanoski, Casey E., Christopher K. Glass, Hendrik G. Stunnenberg, Laurence Wilson, and Genevieve Almouzni. 2015. “Epigenomics: Roadmap for Regulation.” *Nature* 518(7539):314–16.
- Ronemus, Michael, Ivan Iossifov, Dan Levy, and Michael Wigler. 2014. “The Role of de Novo Mutations in the Genetics of Autism Spectrum Disorders.” *Nature Reviews Genetics* 15(2):133–41.
- Rossetti, Sandro et al. 2009. “Incompletely Penetrant PKD1 Alleles Suggest a Role for Gene Dosage in Cyst Initiation in Polycystic Kidney Disease.” *Kidney International* 75(8):848–55.
- Rossi, Enrico et al. 2000. “Compound Heterozygous Hemochromatosis Genotype Predicts Increased Iron and Erythrocyte Indices in Women.” *Clinical Chemistry* 46(2):162–66.
- Rudarakanchana, N. et al. 2002. “Functional Analysis of Bone Morphogenetic Protein Type II Receptor Mutations Underlying Primary Pulmonary Hypertension.” *Hum.Mol.Genet.* 11(0964–6906 (Print)):1517–25.
- Sabbagh, Audrey et al. 2009. “Unravelling the Genetic Basis of Variable Clinical Expression in Neurofibromatosis 1.” *Human*

- Molecular Genetics* 18(15):2768–78.
- Salgado, David, Jean Pierre Desvignes, et al. 2016. “UMD-Predictor: A High-Throughput Sequencing Compliant System for Pathogenicity Prediction of Any Human CDNA Substitution.” *Human Mutation* 37(5):439–46.
- Salgado, David, Matthew I. Bellgard, Jean-Pierre Desvignes, and Christophe Bérout. 2016. “How to Identify Mutations among All Those Variations: Variant Annotation and Filtration in the Genome Sequencing Era.” *Human Mutation* 37(12):1272–82.
- Samocha, Kaitlin E. et al. 2014. “A Framework for the Interpretation of de Novo Mutation in Human Disease.” *Nature Genetics* 46(9):944–50.
- Sanger, F. et al. 1977. “Nucleotide Sequence of Bacteriophage Phi X174 DNA.” *Nature* 265(5596):687–95.
- Sanger, Frederick. 1952. “The Arrangement of Amino Acids in Proteins.” *Advances in Protein Chemistry* 7:1–67.
- Sankelo, Marja et al. 2005. “BMP2 Mutations Have Short Lifetime Expectancy in Primary Pulmonary Hypertension.” *Human Mutation* 26(2):119–24.
- Scherer, Dominique and Rajiv Kumar. 2010. “Genetics of Pigmentation in Skin Cancer - A Review.” *Mutation Research - Reviews in Mutation Research* 705(2):141–53.
- Schmitt, Anthony D. et al. 2016. “A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome.” *Cell Reports* 17(8):2042–59.
- Scully, Jackie Leach. 2004. “What Is a Disease?” *EMBO Reports* 5(7):650–53.
- Shawky, Rabah M. 2014. “Reduced Penetrance in Human Inherited Disease.” *Egyptian Journal of Medical Human Genetics* 15(2):103–11.
- Shintani, M., H. Yagi, T. Nakayama, T. Saji, and R. Matsuoka. 2009. “A New Nonsense Mutation of SMAD8 Associated with Pulmonary Arterial Hypertension.” *Journal of Medical Genetics* 46(5):331–37.
- Sidransky, E. 2006. “Heterozygosity for a Mendelian Disorder as a

- Risk Factor for Complex Disease.” *Clinical Genetics* 70(4):275–82.
- Siepel, Adam et al. 2005. “Evolutionarily Conserved Elements in Vertebrate , Insect , Worm , and Yeast Genomes.” *Genome Research* 15:1034–50.
- Silberstein, Mark et al. 2013. “A System for Exact and Approximate Genetic Linkage Analysis of SNP Data in Large Pedigrees.” *Bioinformatics* 29(2):197–205.
- Sim, Ngak-Leng et al. 2012. “SIFT Web Server: Predicting Effects of Amino Acid Substitutions on Proteins.” *Nucleic Acids Research* 40(W1):W452–57.
- Simino, Jeannette et al. 2014. “Gene-Age Interactions in Blood Pressure Regulation: A Large-Scale Investigation with the CHARGE, Global BPgen, and ICBP Consortia.” *American Journal of Human Genetics* 95(1):24–38.
- Smith, Hamilton O. and K. W. Welcox. 1970. “A Restriction Enzyme from Hemophilus Influenzae I.” *Journal of Molecular Biology* 51(2):379–91.
- Smithies, O. 1955. “Zone Electrophoresis in Starch Gels: Group Variations in the Serum Proteins of Normal Human Adults.” *Biochemical Journal* 61(4):629–41.
- Sobel, E. and K. Lange. 1996. “Descent Graphs in Pedigree Analysis: Applications to Haplotyping, Location Scores, and Marker-Sharing Statistics.” *American Journal of Human Genetics* 58(6):1323–37.
- Song, Jie. 2018. “A next Generation Sequencing Approach to Identify Mutations in Pulmonary Arterial Hypertension with a Functional Assessment of Bone Morphogenic Protein Receptor Type 2 Promoter Variants.”
- Song, Lingyun and Gregory E. Crawford. 2010. “DNase-Seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells.” *Cold Spring Harbor Protocols* 5(2):1–12.
- Soto-Ramírez, Nelís et al. 2013. “The Interaction of Genetic Variants and DNA Methylation of the Interleukin-4 Receptor Gene Increase the Risk of Asthma at Age 18 Years.” *Clinical*

*Epigenetics* 5(1):1.

- Spataro, Nino, Juan Antonio Rodríguez, Arcadi Navarro, and Elena Bosch. 2017. "Properties of Human Disease Genes and the Role of Genes Linked to Mendelian Disorders in Complex Disease Aetiology." *Human Molecular Genetics* 26(3):489–500.
- Spurdle, Amanda B. et al. 2012. "BRCA1 R1699Q Variant Displaying Ambiguous Functional Abrogation Confers Intermediate Breast And Ovarian Cancer Risk." *Journal of Medical Genetics* 49(8):525–32.
- Stankiewicz, Paweł and James R. Lupski. 2010. "Structural Variation in the Human Genome and Its Role in Disease." *Annual Review of Medicine* 61(1):437–55.
- Steinberg, Martin H. and Paola Sebastiani. 2012. "Genetic Modifiers of Sickle Cell Disease." *American Journal of Hematology* 87(8):795–803.
- Stenson, Peter D. et al. 2017. "The Human Gene Mutation Database: Towards a Comprehensive Repository of Inherited Mutation Data for Medical Research, Genetic Diagnosis and next-Generation Sequencing Studies." *Human Genetics* 136(6):665–77.
- Stranger, Barbara E. et al. 2017. "Enhancing GTEx by Bridging the Gaps between Genotype, Gene Expression, and Disease." *Nature Genetics* 49(12):1664–70.
- Strauch, K. et al. 2000. "Parametric and Nonparametric Multipoint Linkage Analysis with Imprinting and Two-Locus-Trait Models: Application to Mite Sensitization." *American Journal of Human Genetics* 66(6):1945–57.
- Strauch, Konstantin, Rolf Fimmers, Max P. Baur, and Thomas F. Wienker. 2003. "How to Model a Complex Trait: 1. General Considerations and Suggestions." *Human Heredity* 55(4):202–10.
- Sudmant, Peter H. et al. 2015. "An Integrated Map of Structural Variation in 2,504 Human Genomes." *Nature* 526(7571):75–81.
- Sun, Xiangqing, Junghyun Namkung, Xiaofeng Zhu, and Robert C.



- Elston. 2011. "Capability of Common SNPs to Tag Rare Variants." *BMC Proceedings* 5(Suppl 9):S88.
- Sutton, W. S. 1903. "The Chromosomes in Heredity." *Biological Bulletin* 231–51.
- Székvölgyi, Lóránt, Kunihiro Ohta, and Alain Nicolas. 2015. "Initiation of Meiotic Homologous Recombination: Flexibility, Impact of Histone Modifications, and Chromatin Remodeling." *Cold Spring Harbor Perspectives in Biology* 7:1–17.
- Tak, Yu Gyoung and Peggy J. Farnham. 2015. "Making Sense of GWAS: Using Epigenomics and Genome Engineering to Understand the Functional Relevance of SNPs in Non-Coding Regions of the Human Genome." *Epigenetics and Chromatin* 8(1):1–18.
- Tang, Haiming and Paul D. Thomas. 2016. "Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation." *Genetics* 203(2):635–47.
- Tenesa, Albert and Chris S. Haley. 2013. "The Heritability of Human Disease: Estimation, Uses and Abuses." *Nature Reviews Genetics* 14(2):139–49.
- Tennessen, Jacob A. et al. 2012. "Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes." *Science* 336(6090):64–69.
- Thankaswamy-Kosalai, Subazini, Partho Sen, and Intawat Nookaew. 2017. "Evaluation and Assessment of Read-Mapping by Multiple next-Generation Sequencing Aligners Based on Genome-Wide Characteristics." *Genomics* 109(3–4):186–91.
- Timpson, Nicholas J., Celia M. T. Greenwood, Nicole Soranzo, Daniel J. Lawson, and J. Brent Richards. 2017. "Genetic Architecture: The Shape of the Genetic Contribution to Human Traits and Disease." *Nature Reviews Genetics*.
- Tjio, J. H. and Levan A. 1956. "The Chromosome Number of Man." *Hereditas* 42(1-2):1–5.
- Torkamani, Ali, Ashley A. Scott-Van Zeeland, Eric J. Topol, and Nicholas J. Schork. 2011. "Annotating Individual Human

- Genomes.” *Genomics* 98(4):233–41.
- Tsankov, Alexander M. et al. 2015. “Transcription Factor Binding Dynamics during Human ES Cell Differentiation.” *Nature* 518(7539):344–49.
- Uher, Rudolf. 2014. “Gene-Environment Interactions in Severe Mental Illness.” *Frontiers in Psychiatry* 5(MAY):1–9.
- Venter, J. et al. 2001. “The Sequence of the Human Genome.” *Science* 291(5507):1304–51.
- Viales, Rebecca Rodríguez et al. 2015. “Mutation in BMPR2 Promoter: A ‘second Hit’ for Manifestation of Pulmonary Arterial Hypertension?” *PLoS ONE* 10(7):1–11.
- Visscher, Peter M. et al. 2017. “10 Years of GWAS Discovery: Biology, Function, and Translation.” *American Journal of Human Genetics* 101(1):5–22.
- Wain, Louise V. et al. 2017. “Novel Blood Pressure Locus and Gene Discovery Using Genome-Wide Association Study and Expression Data Sets from Blood and the Kidney.” *Hypertension* 70(3):e4–19.
- Wain, Louise V et al. 2011. “Genome-Wide Association Study Identifies Six New Loci Influencing Pulse Pressure and Mean Arterial Pressure.” *Nature Genetics* 43(10):1005–11.
- Wang, Dan et al. 2017. “Lower Circulating Folate Induced by a Fidgetin Intronic Variant Is Associated with Reduced Congenital Heart Disease Susceptibility.” *Circulation* 135(18):1733–48.
- Wang, Kai, Mingyao Li, and Hakon Hakonarson. 2010. “ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data.” *Nucleic Acids Research* 38(16):1–7.
- Wang, Yanli et al. 2017. “The 3D Genome Browser: A Web-Based Browser for Visualizing 3D Genome Organization and Long-Range Chromatin Interactions.” *BioRxiv*.
- Wang, Zhihong et al. 2013. “Familial Skewed X Chromosome Inactivation in Adrenoleukodystrophy Manifesting Heterozygotes from a Chinese Pedigree.” *PLoS ONE* 8(3).

- Wang, Zhong; Mark; Gerstein, and Michael Snyder. 2009. "RNA-Seq : A Revolutionary Tool for Transcriptomics." *Nature Reviews Genetics* 10:57–63.
- Ward, Lucas D. and Manolis Kellis. 2012. "HaploReg: A Resource for Exploring Chromatin States, Conservation, and Regulatory Motif Alterations within Sets of Genetically Linked Variants." *Nucleic Acids Research* 40(D1):930–34.
- Warren, Helen R. et al. 2017. "Genome-Wide Association Analysis Identifies Novel Blood Pressure Loci and Offers Biological Insights into Cardiovascular Risk." *Nature Genetics* 49(3):403–15.
- Waterston, Robert H. et al. 2002. "Initial Sequencing and Comparative Analysis of the Mouse Genome." *Nature* 420(6915):520–62.
- Watson, James and Francis Crick. 1953. "Molecular Structure of Nucleic Acids." *Nature*. 171(4356):737–38.
- Weaving, L. S., C. J. Ellaway, J. Gécz, and J. Christodoulou. 2005. "Rett Syndrome: Clinical Review and Genetic Update." *Journal of Medical Genetics* 42(1):1–7.
- Webb, Emily L., Gabrielle S. Sellick, and Richard S. Houlston. 2005. "SNPLINK: Multipoint Linkage Analysis of Densely Distributed SNP Data Incorporating Automated Linkage Disequilibrium Removal." *Bioinformatics* 21(13):3060–61.
- Weiss, Kenneth M. and Andrew G. Clark. 2002. "Linkage Disequilibrium and the Mapping of Complex Human Traits." *Trends in Genetics* 18(1):19–24.
- West, James et al. 2008. "Gene Expression in BMP2 Mutation Carriers with and without Evidence of Pulmonary Arterial Hypertension Suggests Pathways Relevant to Disease Penetrance." *BMC Medical Genomics* 1(1):45.
- White, Kevin, Joseph Loscalzo, and Stephen Y. Chan. 2012. "Holding Our Breath: The Emerging and Anticipated Roles of MicroRNA in Pulmonary Hypertension." *Pulm. Circ.* 2(3):278–90.
- Wijsman, Ellen M., Joseph H. Rothstein, and Elizabeth A. Thompson. 2006. "Multipoint Linkage Analysis with Many

- Multiallelic or Dense Diallelic Markers: Markov Chain-Monte Carlo Provides Practical Approaches for Genome Scans on General Pedigrees.” *American Journal of Human Genetics* 79(5):846–58.
- Wirth, Brunhilde et al. 2006. “Mildly Affected Patients with Spinal Muscular Atrophy Are Partially Protected by an Increased SMN2 Copy Number.” *Human Genetics* 119(4):422–28.
- Withrock, Isabelle C. et al. 2015. “Genetic Diseases Conferring Resistance to Infectious Diseases.” *Genes and Diseases* 2(3):247–54.
- Wolffe, Alan P. and Marjori A. Matzke. 1999. “Epigenetics: Regulation through Repression.” *Science* 286(5439):481–86.
- Wright, Alan, Brian Charlesworth, Igor Rudan, Andrew Carothers, and Harry Campbell. 2003. “A Polygenic Basis for Late-Onset Disease.” *Trends in Genetics* 19(2):97–106.
- Wright, Caroline F. et al. 2018. “Assessing the Pathogenicity, Penetrance and Expressivity of Putative Disease- Causing Variants in a Population Setting.” *BioRxiv*.
- Wu, Xuebing, Rui Jiang, Michael Q. Zhang, and Shao Li. 2008. “Network-Based Global Inference of Human Disease Genes.” *Molecular Systems Biology* 4(189).
- Wu, Yang, Zhili Zheng, Peter M. Visscher, and Jian Yang. 2017. “Quantifying the Mapping Precision of Genome-Wide Association Studies Using Whole-Genome Sequencing Data.” *Genome Biology* 18(1):1–10.
- Xue, Yali et al. 2012. “Deleterious- and Disease-Allele Prevalence in Healthy Individuals: Insights from Current Predictions, Mutation Databases, and Population-Scale Resequencing.” *American Journal of Human Genetics* 91(6):1022–32.
- Yngvadottir, Bryndis et al. 2008. “A Genome-Wide Survey of the Prevalence and Evolutionary Forces Acting on Human Nonsense SNPs.” *American Journal of Human Genetics* 84(2):224–34.
- Yucesoy, Berran et al. 2015. “Genome-Wide Association Study Identifies Novel Loci Associated With Diisocyanate-Induced Occupational Asthma.” *Toxicological Sciences : An Official*

*Journal of the Society of Toxicology* 146(1):192–201.

- Zender, Charles S. 2016. “Bit Grooming: Statistically Accurate Precision-Preserving Quantization with Compression, Evaluated in the NetCDF Operators (NCO, v4.4.8+).” *Geoscientific Model Development* 9(9):3199–3211.
- Zerbino, Daniel R. et al. 2018. “Ensembl 2018.” *Nucleic Acids Research* 46(D1):D754–61.
- Zhu, Qianqian et al. 2011. “A Genome-Wide Comparison of the Functional Properties of Rare and Common Genetic Variants in Humans.” *American Journal of Human Genetics* 88(4):458–68.
- Ziller, Michael J. et al. 2015. “Dissecting Neural Differentiation Regulatory Networks through Epigenetic Footprinting.” *Nature* 518(7539):355–59.
- Zlotogora, Joël. 2003. “Penetrance and Expressivity in the Molecular Age.” *Genetics in Medicine* 5(5):347–52.

**Education articles:**

- Craig, J. (2008) Complex diseases: Research and applications. *Nature Education* 1(1):184
- Lobo, I. (2008) Same genetic mutation, different genetic disease phenotype. *Nature Education* 1(1):64
- Pray, L. (2008) DNA Replication and Causes of Mutation. *Nature Education* 1(1):214

**Books:**

- Griffiths, A. J. F. (1999). *Modern genetic analysis*. New York: W.H. Freeman.
- Mukherjee, S. (2016). *The gene: An intimate history*. New York: Scribner.
- Coleman, W. B., & Tsongalis, G. J. (2018). *Molecular Pathology: The Molecular Basis of Human Disease*. Academic Press.

Ott, J. (1999). *Analysis of human genetic linkage*. Baltimore: Johns Hopkins University Press.

**Web resources:**

Morgan M. (2017). *AnnotationHub: Client to access AnnotationHub resources*. R package version 2.10.1. doi: 10.18129/B9.bioc.AnnotationHub.