Evolutionary analysis of the genome load of loss-of-function variants and their contribution to immunodeficiencies

Guillem de Valles Ibáñez

TESI DOCTORAL UPF / 2016

DIRECTORS DE LA TESI

Dr. Ferran Casals López (Genomics Core Facility)

Dr. Tomàs Marquès-Bonet (UPF)

Institut de Biologia Evolutiva (UPF-CSIC)

Als meus pares i al meu germà,
que m'han acompanyat fins aquí,
i a tots els que m'he trobat pel camí

Acknowledgments

Tinc molt que agrair i poca habilitat per fer-ho, però ho intentaré igualment:

Aquesta tesis ha estat possible gràcies, bàsicament, a tres persones. Agraeixo a en Javi per parlar-me del doctorat que estava fent i per animar-me a provar sort a l'IBE, i per tota la feina que ha fet i de la que jo m'he aprofitat per fer part d'aquesta tesis. I, fora de l'àmbit professional, per ser un bon amic i compartir birres, jocs i tonteria. Sobretot tonteria, s'ha de compensar tanta genialitat. Agraeixo a en Tomàs per acollir-me al seu grup i permetre'm fer el doctorat amb ell, per ensenyar-me i per compartir-me amb en Ferran. Em va impressionar molt la habilitat d'en Tomàs quan va venir a fer un seminari al meu màster, i ho he vist més vegades durant aquest doctorat. M'agradaria poder dir que n'he aprés alguna cosa, però segueixo fent presentacions com el cul. Malgrat tot, ja se on vull arribar. I agraeixo també a en Ferran, per ser el millor *jefe* que he tingut mai. Es un molt bon professional, i la seva relació propera i amigable fa que sigui molt fàcil treballar amb ell. Puc dir sense cap mena de dubte que ha sigut la persona que més m'ha ajudat durant aquests quatre anys de doctorat i qui més m'ha ensenyat sobre com treballar en ciència. M'ha tractat en tot moment com a una persona i fins i tot m'ha permès conservar el sentit de l'humor durant el doctorat, afegint-hi una bona cullerada del seu. Espero que no el perdis mai!

He d'agrair també a totes les persones que m'he trobat durant el doctorat, han fet que sigui una experiència que valgui la pena. M'agradaria començar fent el friki, citant a en Bilbo Bolsón i dir que no conec a la meitat de vosaltres ni la meitat del que voldria, i el que jo voldria es menys de la meitat que la meitat de vosaltres es mereix. Començant pel grup dels tomasinos -i afegits-, he de dir que es fan bons amics quan tens la boca plena (*pun intended*). Gracies a en Javi, la Irene, l'Ignasi, en Marcos, la Jess, en Marc (tant en Dabad con el de Manual), la Clàudia, en Xavi, la Raquel, en Tiago, la Irene Lobón, en Lukas, l'Aitor, la Meritxell, la Irune, en Manolo, en Ramón i en José per compartir dinars i temes de conversa eclèctics i, sovint, divertits. Durant la meva estada a l'IBE he passat per dues fases de cafès, ambdues molt productives i divertides, sobretot la segona. Gràcies als que les heu compartit amb mi. Quiero agradecer especialmente a Lukas por enseñar-me a mentir como una cabra en mi segunda fase cafetera, y por acompañarme en el birrabox cuando llega la hora de la gallina. También quiero agradecer a los que me han enseñado a venir, hacer un poco e irse.

L'Ignasi es mereix una menció especial, ja que hem compartit birrabox i llibres i ha sigut un molt bon company de doctorat i un bon amic. I a més em va deixar un motllo de pastissos en forma de pollón.

A parte de los tomasinos, hay muchos bioevos y demás fauna del PRBB con los que he compartido muy buenos momentos. Han sido pocos momentos, para mi gusto, pero espero volver a encontraros y ponerle remedio. Quiero remarcar especialmente al dúo del Dr. Caca y Mr. Juan, un hombre con una personalidad

tan grande y tan genial que tiene que repartirse en dos personas. Gracias a Juan y sus amigos (especialmente Mei y Tara), a Nino y su expresividad napolitana, a Marco y sus chupitos de tequila asesinos, a Ali y su radioclub de canciones, a Diego y sus 4 toneladas de molosidad, a la Marina i la seva companyia a l'Esperanzah, a Arturo y su saber vivir, a la Neus i el seu doctor de Harvard, a Iñigo y su Neus y sus pokemons, a Gio y su R, a María y sus tacos, a en Marc y les seves birres, a en Max i a la enveja que em foten els seus viatges, a en Txema i el seu hidromel, to Mayukh and his Indian birthday celebration, a Koldo y sus cafepintxos, a Fede y sus apariciones inesperadas, a la Judit i la seva ajuda en la burocracia kafkiana, a la Núria i a en Roger i les birres a Glasgow, a la Carla i a que em demani llibres frikis, als que ja paren per altres llocs (Urko, Ixa, Gabriel i Belén), als que ja no hi són i que vaig conèixer massa poc (Johaness i Margarita), i a tu que estás llegint això i ja sembla un anunci de coca-cola, moltes gràcies. Se que em deixo gent, però mentres escric això tinc unes ganes bojes d'imprimir la tesis i anar a fer una o varies cerveses.

A la Azu i l'Ana, que em van acollir molt bé quan anava a intentar ajudar-les amb el western blot, i no em van dir mai que feia més nosa que servei, encara que jo se quina es la veritat. També vull agrair a en Manel, l'Eva, l'Anna, la Laia, en Juantxo, la Mónica, en Pere, en Roger i en definitiva a tota la gent del Clínic, Sant Joan de Deu i Vall d'Hebron que ens han permés col·laborar amb ells i han acceptat l'ajuda d'un paio que té molt poca idea d'immunologia.

A mis amigos de la uni, que han hecho que estar lejos de casa merezca la pena. A Mon, mi hermano de otra madre, y su hidrocele. A Maxi y las clases de euskera y los kamarrierros pasados y los que nos quedan. A Arden y sus hostias a cámara lenta. A Esti por ser tan alta como buena gente. A Quechu y su Killa y sus oiiishs. A los muchos que ya no veo, y que echo de menos: Vio, Anna, Arturita, Myriam, Ra, Ángel, Mr. News, Toni, Cris, Ingmar, Manolo, les dues Marines, Itxi, Estitxu y mil más. A los que compartieron casa, juergas, y alguna que otra planta prohibida: Mifu, Janfri, Mauri, Narciso I il bello, Jorge, Alba, Xavi. Y a mi profesor de fisiología vegetal, Rafael Robaina, que fue quién me convenció definitivamente de que quería trabajar en genética.

Als meus amics de Puigcerdà, amb els que he compartit infància, adolescència, el que sigui que estem passant ara i el que sigui que vindrà. Menció especial a l'Uri i en Jambo, amb els que he compartit pis i als que he hostiat amb raquetes de squash. Sergim, Alejandrus, Helen, Lou, Jose, Busi, Nunu, Laia, Víctor, Edu, Llui, Pacmans, Mestres i un llarga llista de gent. No cal que posi gran cosa per aquí per que se que us aniré veient. Encara que no volgueu!

A Ju y a Uma y a Mad y a toda la gente que conocí en cierto oscuro foro. Que sepais que también os mereceis estar aquí. Por que han salido buenos amigos de ahí, y porqué las chorradas no se olvidan.

A Laura, que me enseñó a vivir en pareja y a querer a alguien complicado. Ha terminado pronto, pero han sido tres buenos años. Me dejas un legado de buenos gatos y un recuerdo de buenos ratos.

A la meva família, òbviament. M'han recolzat sempre i m'han donat amor incondicional. Espero correspondre'ls com es mereixen. Als meus pares que m'han ensenyat el valor de la lectura i com n'és de divertit aprendre coses. Al meu germà, que ha sigut el meu millor amic des de que va néixer, i que es un crack en moooooltsaspectes. A l'Alfons i la Pepa, que m'han demostrat que la família no es només genètica i han sigut uns segons pares per mi. Al Mochi i el Tino i la Titi i la Mina i l'Amaia i l'Aiona i l'Antonio i els bandarres, que també estan dintre d'aquesta categoria. A la Frida, que està com un llum. A la Lola i la Duda, que ja no estan. Al meu tiet Lauro, que tampoc està, que em va ensenyar moltes paraulotes i cançons guarres, i que pots estimar a algú amb una ideologia molt diferent. A la meva tieta Eli i a en Lluís. Als tiets Jordi i Jose Luis, i a la tieta Montse. Al meu cosinàs. Part de que m'agradi tant la ciència es cosa dels dos doctors Ibáñez. Al meu cosí Àlex. Als altres tiets i tietes amb els que ja no tinc contacte però que en guardo bons records.

També he d'agrair a la web sci-hub.cc per donarme accès a articles pels que hauria d'haver pagat una morterada.

A tothom que llegeixi aquesta tesis, espero que sàpiga perdonar els errors que conté i les tres pàgines d'agraïments. Gràcies per llegir-me.

Abstract

Human genomes have been found to harbor an unexpected number of ~100 loss-of-function (LoF) variants, with ~20 of them in an homozygous state, in most cases without a visible effect despite its potential truncation of proteins. This suggests that some of those variants should be neutral but also a fraction could be lethal alleles. In this work we study the implications of LoF variants in two different fields: in comparative genomics by exploring for the first time the mutational load of LoF variants segregating in 79 genomes belonging to six different great ape populations and its possible detrimental effects, and in medical genomics by its implication with other functional variants in 36 patients diagnosed with Common Variable Immunodeficiency, an heterogeneous disease with several genes implied in its etiology, using both monogenic and oligogenic models for this antibody deficiency.

Resum

Recentment s'ha descobert que els genomes humans contenen unes inesperades ~100 variants que causen pèrdua de funció (LoF), ~20 de les quals es troben en homozigosi, sense causar cap efecte visible malgrat el seu potencial per esguerrar una proteïna. Això suggereix que algunes d'aquestes variants han de ser neutres, però també que una fracció podrien ser al·lels letals. En aquesta tesis estudiem les implicacions de les LoF variants en dos camps diferents: en la genòmica comparativa explorant per primer cop la carrega mutacional de les variants LoF segregant en 79 genomes que pertanyen a sis poblacions diferents de grans simis i els seus possibles efectes deleteris, i en el camp de la genòmica mèdica per la seva implicació, junt amb altres tipus de variants, en 36 pacients diagnosticats amb Immunodeficiència Comú Variable, una malaltia heterogènia amb varis gens implicats en la seva etiologia, utilitzant models monogènics i poligènics per estudiar aquesta deficiència d'anticossos.

Preface

Since I have use of reason, I've been truly fascinated by living organisms. I remember collecting tadpoles and all kinds of insects that I found in my hometown when I was a child, completing a magazine series called *Bichos* and reading everything about animals that I could lay my hands on. The biology classes I took in high school boosted this passion and lead me to study a biology related career. And among the many wonders found in this field, I known none more amazing than genetics. It provides the final link between matter and the characteristics that we could observe with the naked eye on living organisms, and it has been an inexhaustible source of scientific discoveries, specially in recent times. The research in genetics allows us to understand the mechanisms in biology, and provides an unbiased answer to a many questions belonging, until recently, to the field of philosophy and metaphysics. Science works always in small steps (even groundbreaking discoveries need a previous knowledge, obtained in a careful and methodical way) but always keeps moving towards the ultimate goal of understanding the unknown. I intend with this thesis to collaborate, a little bit, with this adventure.

Index

1. INTRODUCTION

1.1 A brief history of DNA discovery

"Not one of your pertinent ancestors was squashed, devoured, drowned, starved, stranded, stuck fast, untimely wounded, or otherwise deflected from its life's quest of delivering a tiny charge of genetic material to the right partner at the right moment in order to perpetuate the only possible sequence of hereditary combinations that could result -- eventually, astoundingly, and all too briefly -- in you."

Bill Bryson, A Short History of Nearly Everything

In the very first chapter of his Origin of Species, Charles Darwin wonders about the causes of the variability observed inside a "variety or sub-variety" of plants or animals, attributed in his time to the conditions of the environment where the organisms were raised. But, remarkably, Darwin finishes the section pointing that the cause of the differences between the organisms may has a more intrinsic nature, writing, in a poetical way amidst the nineteenth-style verbiage, "[...] we clearly see that the nature of the conditions is of subordinate importance in comparison with the nature of the organism in determining each particular form of variation; perhaps of not more importance than the nature of the spark, by which a mass of combustible matter is ignited, has in determining the nature of the flames."(Darwin 1859)

After Darwin and Wallace's theory of natural selection struck the minds of their contemporaneous fellow scientists, it began a hunt for the particle of heredity on which selection acts upon. Darwin itself expressed his ideas on the matter with the “provisional theory of pangenesis”. He imagined that cells in the body where constantly creating microscopic hereditary particles -he called them gemmules- that travel to the reproductive organs carrying information about the acquired traits of their parents cells in that precise moment. The eventual blend of the gemmules with their opposite-sex counterparts, not necessarily in the same proportion, produced the next generation of organisms, with a mixture of the traits that each parent had in the moment that the gemmule was created(Darwin 1868; Schwartz 2008). Darwin's cousin, Francis Galton, infamous due to be the mind behind social Darwinism and eugenesia, proposed Darwin to test his Lamarckian hypothesis by a blood transfusion experiment in rabbits. He stated that, if cells where constantly shedding gemmules, they certainly must travel via bloodstream, and therefore an animal transfused with another breed's plasma should produce an offspring recapitulating the appearance of the original blood donor. The experiment failed(Galton 1870; Schwartz 2008) and gemmules were forsaken, but the hunt kept going on.

The focus of investigation moved to the nucleus of the cell, thanks to a bold guessing by Ernest Haeckel in, when he 1866 suggested that the factors responsible for heredity should be found in the

“inner nucleus”, without any experimental basis to support that claim(Haeckel 1866; Schwartz 2008). Scientists quickly realized that the fusion of the gametic nuclei was necessary for fertilization and that something strange was happening inside the nucleus. Before cell division the nuclear material seem to concentrate in a few rods, that, after a strange dance, migrate to the poles of the cells and disappeared again(Schneider 1873; Flemming 1882; Van Beneden 1883). Moreover, those rods were found to be variable in size, and also to differ in number between species and even between the two sexual forms of the same species(McClung 1899; McClung 1902; Sutton 1903). Males and females were found to had, in some cases, different number of chromosomes, and in other cases two unpaired chromosomes. Further research in this matter lead to the discovery of the chromosomal sex determination, greatly contributed by Thomas Hunt Morgan, who was its foremost detractor in its beginnings, and his student Nettie Stevens(Stevens 1905a; Stevens 1905b; Wilson 1905; Wilson 1906; Morgan 1908).

Well before sex determination was discovered, an Augustinian friar named Gregor Mendel wrote a book about hybridization in peas that contained one of the most revolutionary insights in the field of the biology(Mendel 1865), and that was pretty much ignored for the subsequent 35 years, until in 1900 was simultaneously and independently rediscovered by three scientists(De Vries 1900; Correns 1900; Tschermak 1900). Mendel's rediscovery was very controversial and sparkled a bitter debate that spanned one

decade(Schwartz 2008), until, again by Morgan and his students, the first gene maps in *Drosophila* chromosomes were produced(Sturtevant 1913), giving the reason to Mendel's party and closing the debate for good.

From sex determination to gene mapping, scientists have been working with the chromosomes without knowing how information was coded inside them, until Francis Crick and James Watson, working with the at the time unpublished X-ray model(Wilkins, Stokes, and Wilson 1953; Franklin and Gosling 1953) from Maurice Wilkins, Rosalind Franklin and Raymond Gosling, elucidated the three-dimensional structure of the DNA. The greatness of this discovery lies not in the mere description of how nucleotides place themselves along the DNA molecule, but in the implications that the nucleotide pairing has in the transmission of genetic material, which Crick and Watson resumed in their famous phrase “It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.”(Watson and Crick 1953)

The definition of a mutation has changed since then, and nowadays is viewed as any permanent modification in the genetic components of an organisms, to encompass in its definition the wide range of features that are known to affect the phenotype of an individual. Mutations during cell replications have great implications in the

field of biology, from evolution to its clinical applications, where are widely researched in genetic diseases and cancer.

1.2 Origins of life and mutations

"*In the beginning there was nothing, which exploded.*"
Terry Pratchett, Lords and Ladies

Mutations *per se* may be older than living organisms itself, as for sure before the first cell appeared there should have been a physical basis where genetic information has been stored, and that pre-cell molecule had to change or evolve trough mutations to develop in a future cell. The first clear evidence of living organism it's found in rocks dated to be 3,500 million years old(Schopf 1993; Schopf 2006), but some researchers claim to have found older traces of life, from 4,100 million years ago(Bell et al. 2015).

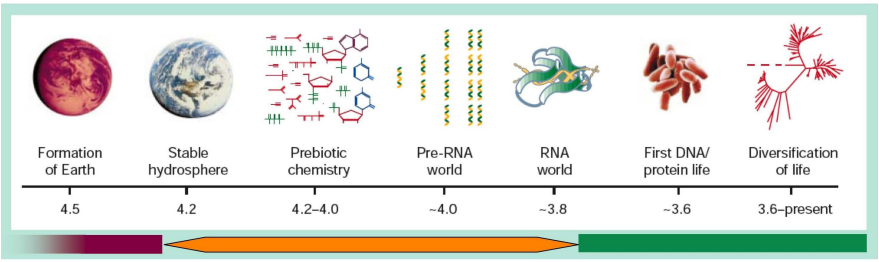


Figure 1: Timeline of early life events expressed in billions of years, taken from Bellini et al. 2012

What remains unclear are the mechanisms that have originated life, and many speculative science has been written on this subject. The building blocks of life are pretty common in the universe and they

can be found even in the dust debris that later may form solar systems or comets(Nuevo, Milam, and Sandford 2012), such as the recently famous 67P/Churyumov-Guerasimenko(Altwegg et al. 2016). The problem is that those chemicals need to produce even more complex molecules, work in an integrative and organized way to form a rudimentary metabolism, inside an enclosed lipidic membrane and under the management of some molecule that can store this information and replicate it. And it's not an easy problem to solve, as Crick recognized saying “An honest man, armed with all the knowledge available to us now, could only state that in some sense, the origin of life appears at the moment to be almost a miracle, so many are the conditions which would have had to have been satisfied to get it going.”(Crick 1981). Well before Crick wrote these words, and also well after then, many have taken the challenge to explain the origin of life. The most promising hypothesis place the event of life in the interface between hot water from geothermic activity and the colder water of the primitive ocean, and using RNA, not DNA, as the genetic material(Coveney et al. 2012; Bellini et al. 2012). Advocates of the RNA-world hypothesis argue that many of the core molecules essential for the cell's metabolism are in fact remnants of these early life forms(Coveney et al. 2012; Bellini et al. 2012; Cech 2010), and it's unquestionable that RNA is essential for modern cells. But many other hypothesis have been proposed, and the mystery of the origin of life might be unfathomable, because proving that some steps can create a living organism doesn't imply that it was the exact sequence

of events that created the original life forms(Serafino 2016). Moving outside the unstable quicksands of the origin of life, what we know for sure it's that DNA-based organisms eventually developed their complex DNA-RNA-proteins complex system and cells have used DNA as storage system for the rest of his history. All known living organisms use DNA as genetic material, but certain quasi-organisms, as virus and viroids (viroids are small strands of circular RNA without a protein coats and minute genomes of ~300-400 nucleotides that mostly infect plants(Flores et al. 2014)), use RNA and exist in a kind of limbo between inanimate matter and living cells, as a reminder that life is a category, a descriptive concept with entities below his scope. Some organelles seem to fall in that limbo too, because their existence as endosymbionts of other cells allowed them to reduce their genomes to a set of genes that is not enough to allow them to have a free-living existence. This pattern of genome reduction is also observed among parasites, specially intracellular parasites, which tend to evolve becoming more dependent of their host's cellular materials and therefore diminishing the need of producing it themselves. Recently, a study of prokaryotic genomes has identified 355 protein families that, traced back to the last universal common ancestor (LUCA) of all cells, suggest that it had a very small genome, lacking many genes that otherwise would have allowed a free existence outside those hydrothermal vents upon its products part of LUCA's metabolism relied(Weiss et al. 2016).

1.2.1. DNA damage and repair

Despite the greater stability that double stranded DNA shows in front of RNA molecules, it remains susceptible to be affected by a variety of natural means as well as by more aggressive agents from the media where it is found. Mutagenesis, aging and cancer are some of the outcomes of this limited stability, likewise some practical limits to restore the nucleotide sequence from ancient samples.

Due to complementary base pairing, the stability in the structure of DNA relies in the shape of the nucleobases and in the recognition by cellular proteins of the nucleotide sequence through the molecular electrostatic potential of the nucleophilic sites on nucleobases and base pairs(Liu and Wang 2015). The previous phrase implies that the replication machinery needs a "clean", undamaged strand of DNA in order to make a complementary strand. Mutations can arise spontaneously due to errors in replication or can appear when DNA is damaged and the repair machinery is not able to fix the damage, leading to abnormal replication and the subsequent incorporation of a mutation in the nucleotide sequence. This damage may be produced by both endogenous causes intrinsic of the cell or by environmental causes, being the former the most common ones. Table 1 shows the

estimated numbers of DNA damage induced by endogenous chemical reactions in mammalian cells.

DNA damages	Reported rate of occurrence
Oxidative	10,000-11,500 per cell per day in humans
	74,000 -100,000 per cell per day in rats
Specific oxidative damage products 8-hydroxyguanine, 8-hydroxydeoxyguanosine, 5-(hydroxymethyl) uracil	2,800 per cell per day in humans
	34,800 per cell per day in mice
Depurinations	2,000-13,920 per cell per day in humans
Depyrimidinations	696 per cell per day in humans
Single strand-breaks	55,200 per cell per day in humans
Double-strand breaks	10-50 per cell cycle in humans
O 6 -methylguanine	3,120 per cell per day in humans
Cytosine deamination	192 per cell per day in humans

Table 1: DNA damages due to endogenous causes in mammalian cells, adapted from Bernstein et al. 2013

There are many chemical reactions that can result in DNA damage, for example alkylation, oxidation, deamination, coordination, photo-addition or hydrolisis(Lindahl 1993; Lindahl and Wood 1999; Bernstein et al. 2013; De Bont and van Larebeke 2004; B. Liu et al. 2016). The principal causes of endogenous lesions are reactive oxygen species (ROS) product of normal oxygen metabolism and hydrolysis, but many other naturally occurring reactants can be found inside the cells and have a significant effect on the amount of lesions in DNA strands(Jackson and Loeb 2001). Estimates of the damage lesions in each cell per day vary greatly between different studies and upon the causes considered, but the order of magnitude estimated for each cause is more or less similar and in overall it's assumed that DNA of every cell can suffer more than 50,000 lesions each day(Lindahl and Wood 1999; Bernstein et al. 2013; Liu et al.

2016). Environmental damage can increase greatly those numbers but it's difficult to take into account all the possible extrinsic factors that can be affecting one cell and their quantitative effect. Nonetheless, many researchers tend to focus their studies in environmental mutagens like tobacco smoke, radiation or nutrition related chemicals due to their relationship with many kind of cancers.

Mutagens, either originated in the cells or outside them, are products that can react with DNA and form a DNA adduct, i.e. when there is a covalent bond between a DNA strand and other chemical, or capable to generate a DNA-DNA or a DNA-protein crosslink that doesn't allow normal replication. Other chemicals such those that can change epigenetic markers or induce a single or double strand-breaks are also considered mutagens, as well as radiation able to affect the helix structure. Among this kinds of damage double strand-breaks are specially detrimental, since they can impede replication and lead to cell death(Lindahl and Wood 1999; Liu et al. 2016). Transposable elements (TE) are other mutagenic source and may play an important role in genome evolution thanks to their capacity to activate or inactivate genes(Amariglio and Rechavi 1993). TE are repetitive nucleic acid fragments that propagate through the genome through a copy-paste mechanism, in the case of retrotransposons, or through a cut and paste translocation for the DNA transposons. Copies of TE constitute about 48% of the human genome, and this number could

be an underestimation because old transposition events may be masked thanks to the accumulation of mutations(Rayan, del Rosario, and Prabhakar 2016). Some TEs have a wide variety of functions in mammalian genomes and have been shown to shape the regulatory landscape of the genomes by modulating the expression of nearby genes, contributing to as much as a 20% of the human and mice transcription factor binding sites(Sundaram et al. 2014). They are specially important in primates since many of their regulatory elements are originated by transposition events and they have been extensively researched as a source of evolutionary innovation or due to their implications in human diseases(Hancks and Kazazian 2016). Although this thesis is focused on small mutations, TE contribution to the overall genome must be taken into account for other studies.

The sheer number of DNA damaged sites per cell gives us an idea of how efficient the repair mechanisms must be, although they are not error-free and sometimes the repair mechanism itself may induce mutations, as far as 5 kilobases from the region damaged(Chen and Furano 2015). The necessity of conservation of nucleotide sequence has brought evolution to develop a wide range of DNA repair mechanisms along all living organism, from bacteria to eukaryotes, whose genes show goods patterns of conservation through evolutionary distant organisms. Moreover, the majority of genes encoding repair proteins are essential in vertebrates and some of them have double roles in specific tasks involved in repair and

replication(Aze et al. 2013). Usually those mechanisms imply the excision of a short sequence surrounding the lesions and the replacement of that sequence with a copy coming from the complementary strand. Two usual repair pathways called base excision repair (BER) and nucleotide excision repair (NER) use this method to fix common base modifications introduced by endogenous sources and helix-distorting damage provoked mostly by environmental causes, respectively(Lindahl and Wood 1999; Aze et al. 2013). When damage is clustered, the attempt of repair can produce double strand-breaks and need to be avoided using specific polymerases that can skip the lesion, introducing sometimes incorrect bases. Single and double strand-breaks attract recombination events and therefore must be protected by specific means. An abundant nuclear protein called PARP1 works as antirecombinogenic factor protecting single strand-breaks lesions, while double strand-breaks have the potential to become sites of recombination by nonhomologous end-joining and attracts large numbers of proteins (including PARP proteins and homologous recombination factor RAD52) that not only protect the lesion but also signal the damage to cell-cycle proteins (the replisome) to act in consequence(Lindahl and Wood 1999; Aze et al. 2013). The repair by homologous recombination with another allele is usually achieved whit great fidelity, but non-homologous recombination, the main pathway used in mammals, may result in a change or a loss of genetic information.

1.3 Understanding mutations

"*Beneath the imposing building called heredity*
there has been a dingy basement called mutation"
H. J. Muller, Second International Congress on Eugenics

1.3.1 Genetic encoding

After the structure of DNA was unveiled, the scientific community got to understand how life is written, but not how is it read. By then, it was known that DNA was composed of 4 subunits (adenine, thymine, guanine and cytosine) while that proteins where made from 20 principal amino acids. The relationship between the nucleic acids and the amino acids was the missing link in molecular biology. In the early 40s it was clear that there were two different kinds of nucleic acids, one whose characteristic carbohydrate is d-2-desoxyribose (known at the time as animal nucleic acid or thymonucleic acid, nowadays called desoxyribonucleic acid or DNA) and other with d-ribose (ribonucleic acid or RNA, known before as plant nucleic acid or yeast nucleic acid)(Allen 1941). Those nucleic acids were known to differ also in terms of stability (RNA degrades faster than DNA) and in one of their bases (uracil instead of thymine in RNA), as well as their location: DNA it's located in the nuclei while RNA can be found also in the cytoplasm.

With all this data in mind, and after the publication of the DNA structure, a club of selected scientists tried to explain how proteins

are made. The RNA tie club was an idea from James Watson and George Gamow and it was composed of 20 members (one for each amino acid) plus 4 honorific members (relating to the four nucleotides). Among the club successes were the theoretical formulation of the non-overlapping codon by Sydney Brenner (designated as Val in the RNA tie club), the mathematical description of how three-nucleotide codons could code for the 20 aminoacids by George Gamow (Ala) and the adaptor hypothesis by Francis Crick (Tyr). Crick envisioned a set of adaptor molecules, at least one for each amino acid, that combined with an enzyme was able to confront each amino acid to his specific template through the nucleotide sequence hydrogen bonding surface(Crick 1955). Henceforth, those adaptor molecules should be nucleic acids, and indeed three years later a “soluble RNA” fitting those characteristics was found in rat livers(Hoagland et al. 1958). Nowadays those adaptor molecules are called transfer RNAs. This hypothesis was developed further in what is known as “the central dogma of molecular biology” to describe the flow of information within biological systems, where, in general, information from the DNA it's transcribed to RNA which in turn is used as template to synthesize proteins(Crick 1958; Crick 1970). In figure 2 the central dogma is schematized, with solid lines representing the general cases and dotted lines the special cases. Nonetheless, the central dogma isn't dogmatic and it's open to interpretations, a few examples of it are proteins that can induce changes in proteins through post-translational modifications, epigenetic changes

working in a different level regulating expression and prions able to mediate epigenetic inheritance of phenotypic traits in yeast(Koonin 2012). Crick himself recognized that he never meant that his central dogma was beyond doubt and that he might had not chosen the right word(Crick 1988).

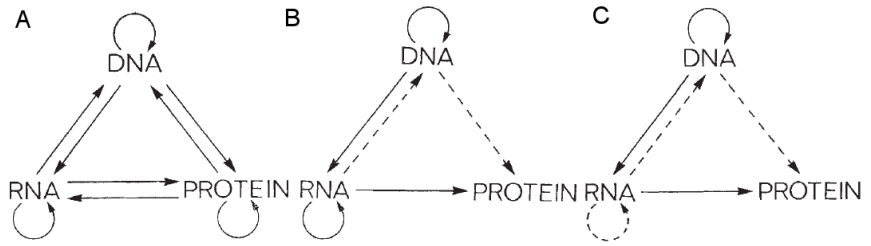


Figure 2: A) Representation of all the possible information flow B) How the central dogma was viewed in 1958 C) How the central dogma was viewed in 1970. Adapted from Crick, 1970.

At this point the code of life was ready to be cracked. There existed a theoretical basis, thanks to the RNA tie club efforts among other many scientists, pointing to a three-nucleotide, non-overlapping code. Three experiments performed between 1961 and 1964 shed light in the genetic code that orchestrate the synthesis of proteins. In 1961 Nirenberg and Matthaei were able to produce a polyphenylalanine protein from a pure uracil RNA chain in a free-cell system, thus deciphering the first codon-amino acid relationship: UUU encodes Phe(Nirenberg and Matthaei 1961; Matthaei et al. 1962). This major breakthrough was followed in the same year by a beautiful experiment done by Crick, Barnett, Benner and Watts-Tobin. Working with a bacteriophage from *E. coli* they prove that the code was in base three and therefore it should be

degenerate, *i.e.* that specific amino acids could be encoded by more than one codon (since 3 positions with 4 possible nucleotides in each one give 4^3 or 64 combinations, and there were only 20 commonly found amino acids). The experiment used a mutagen that generates a single base addition or deletion in the nucleotide sequence, putting the following sequence out of frame, coding for different amino acids and leaving the rest of the gene with a totally different sequence and thus unable to perform its task (in this case, allow the phage to grow in a certain strain of *E. coli*). A single mutation caused a loss of function of the gene, whereas two mutation restored the function with (usually) minimal consequences(Crick et al. 1961). The last experiment undergone by Leder and Niremberg was able to match each triplet of nucleotides to his specific amino acid thanks to a filtration method that allowed them to know the order of nucleotides in the codons(Nirenberg and Leder 1964). With that it was confirmed the degeneration of the code and were discovered whose codons encoded for a nonsense, also called stop. Figure 3 shows a circular representation of the relationship between codons and amino acids (https://kaiserscience.files.wordpress.com). Those three experiments provided directly the explanation of how exactly DNA encodes the amino acids, and indirectly how DNA can mutate, thus evolving or degenerating in a disease.

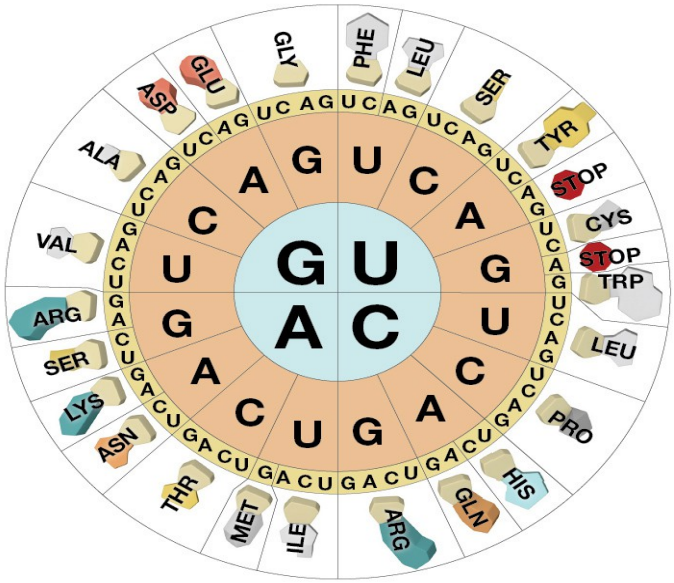


Figure 3: Circular genetic code, representing the nucleotide encoding of aminoacids. The innermost circle represents the first nucleotide, and moving to the perifery are the second and third nucleotides and the amino acid encoded for each triplet. From https://kaiserscience.files.wordpress.com.

1.3.2 Mutations in the coding DNA

We have seen that mutations can occur at different scales, ranging from polyploidies that imply the duplication of all the genome to simple substitutions of one base that can have no consequences at protein level. From now on I will refer to mutations in the sequence level, *i.e.* point mutations or single nucleotide polymorphisms (SNPs) or insertions and deletions (indels) of less than 50 bases. The indel length is an arbitrary value, taken to emphasize that are short insertion or deletions, in opposite as those that have the

potential to disrupt the entire sequence of a gene or more than one gene.

Nucleotide substitutions can be divided into transitions (changes between the purines A and G or between the pyrimidines C and T) and transversions (conversion between purines and pyrimidines). There are four possible transitions and eight possible transversions, and the direction of mutation is non-random: although under a random assumption transitions should account for 33% of the substitutions, they are found in 60-70% of the cases(Graur 2003), mainly because they are less likely to result in an amino acid substitution than transversions.

SNPs can be classified by their effect on the codon, where it could be a synonymous change when they encode for the same amino acid, non-synonymous when the mutation changes the amino acid encoded or nonsense where the encoding changes to a termination codon. Table 2 lists all the possibilities of nucleotide substitutions(Graur 2003).

Mutation	Number	Percentage
Total in all codons	549	100
Synonymous	134	25
Nonsynonymous	415	75
Missense	392	71
Nonsense	23	4
Total in first codon	183	100
Synonymous	8	4
Nonsynonymous	175	96
Missense	166	91
Nonsense	9	5
Total in second codon	183	100
Synonymous	0	0
Nonsynonymous	183	100
Missense	176	96
Nonsense	7	4
Total in third codon	183	100
Synonymous	126	69
Nonsynonymous	57	31
Missense	50	27
Nonsense	7	4

Table 2: Relative frequencies of different types of substitution mutations in a random protein coding sequence, taken from Graur 2003

Mutations can also be classified by their effect on the function of the protein, as silent mutations that have no effect at all or where they can produce a partial or total loss-of-function (LoF) or a gain-of-function (GoF); by their effect on fitness, being harmful or deleterious when they decrease the fitness, advantageous or beneficial when they produce an increase of the fitness, or neutral mutations when they have no effect in fitness. In this thesis, as well as the articles included in it, variants are classified by their effect in the protein and in the function, leading to three main categories: synonymous variants, non-synonymous variants and LoF variants, which are detailed in the section about types of mutations.

1.3.3 Gene and genome organization

Chromosomes are long stretches of DNA, packaged and wound around histones forming complexes called nucleosomes. Chromosomes usually differ in length and contain a high number of genes, usually in the order of thousands in each chromosomes. But biology, and therefore genomics too, it's the science of exceptions, and in nature some surprising ones can be found like the strange genome organization of the ciliates. For starters, ciliates have two distinct nucleus: a germline micronucleus which remains transciptionally inactive during vegetative growth and a macronucleus, which is product of the expression and editing of the genes in the micronucleus and has somatic functions. Some species of ciliates, for example *Oxytrica trifallax*, have the genome of their macronucleus fragmented in more than 16,000 nanochromosomes with an average length of 3.2 kilobases and usually encoding a single gene each(Swart et al. 2013). Going back to more "mainstream" genome organizations, organisms usually have their genomes divided into a few chromosomes, rarely more than a hundred.

When genetic code was understood and genes begun to be "read" with the developing of DNA sequencing methods(Wu and Taylor 1971; Maxam and Gilbert 1977; Sanger, Nicklen, and Coulson 1977), many laboratories found the striking fact that eukaryotic

genes where not coded as a single long chain of nucleotides but they have in between long stretches of sequence that is not translated as a protein, or even in sequences encoding tRNA and rRNA. This is rarely seen on prokaryotes (where mainly happens in regulatory genes) and single-celled eukaryotes, but more common in metazoans and specially in vertebrates(Koonin, Csuros, and Rogozin 2013), whose genome can have as much as 95% (as is the case for *Homo sapiens*) of their coding genes with this feature(Hubé and Francastel 2015). Gilbert called those untranslated sequences inside genes introns, and the expressed into proteins exons(Gilbert 1978). Introns range from few bases to sizes up to 1Mb, with an average of 5kb for coding genes and 7kb for non-coding genes and some of them can create overlapping units with other genes or even contain nested genes inside(Hubé and Francastel 2015). Introns and exons are defined by the dinucleotides GU at the 5' splice site and AG at the 3' splice site and are spliced in the nucleus by a large RNA-protein complex called the major spliceosome. Less than 1% of the genes have noncanonical splice sites that use other dinucleotide pairs to define intron/exon boundaries, and are spliced by the minor spliceosome outside the nucleus(Ng et al. 2004; König et al. 2007). This minor spliceosome his related with cell-proliferation and its decoupling from the nucleus appears to be an escape from mitotic downregulation(König et al. 2007).The splicing of genes into exons has important biological consequences, since alternative splicing is a significant source of protein diversity and mutations in splicing can modify drastically gene functionality.

Beyond the mere encoding of proteins, organisms' complexity makes necessary an elaborate organization of the genome. The ordeals that living systems have to face, even for the relatively simply prokaryotes, make mandatory to respond to certain stimuli and accomplish reproduction. Obviously, this can't be done only by a raw continuous translation of genes in to proteins, but it's necessary a certain spatial and temporal regulation of this translation. In organism with differentiated tissue this need is more apparent since they have different kind of cells with different proteins (at both quantitative and qualitative levels), growing or shrinking at different times, with the same genome directing everything. This can only be accomplished if there exists a fine tuning of the flow of information from DNA to proteins, to determine exactly when and in which amount a certain molecule is needed. Many of those organizing mechanisms can be found in non-coding sequences, ranging from short motifs in the sequence as enhancers, silencers and promotors to short and long non-coding RNA or enzymatic regulators. An important part of regulation of gene expression relies in epigenetic factors, *i.e.* functional changes in the genome that doesn't involve changes in the nucleotide sequence. Methylation patterns, specially on promoters(Jones and Takai 2001) but also in the gene sequence(Lister et al. 2009; Laurent et al. 2010) modulate gene expression and play important roles both in the development of the organism and in evolutionary scales(Hernando-Herraez et al. 2013). The era of the “omics”

studies is expanding our understanding of how the phenomena of gene regulation works and the different levels where it acts.

Furthermore, the majority of proteins don't work as separate entities since they need to collaborate between them to produce the final function in the cell, some of them forming protein complexes, interacting at determined times, or inducing the expression or repression of other proteins. More than 80% of the proteins need to interact in order to be functional(Keskin, Tuncbag, and Gursoy 2016). This adds a new layer of complexity beyond the mere encoding and expression, and therefore protein-protein interaction (PPI) networks should be taken into account for a better understanding of the mechanisms underlying cell function, specially when its defective. This implies that research can't take into account only the linear sequence affected by a mutation, but must also consider the functional domain where it lies, the impact in the protein structure and the possible interactions. PPI studies, either in a general view or in a more modular vision of the functional pathways with a subsets of protein from all the PPI, is a growing field both in evolutionary and clinical research. Nonetheless, by 2006 only the 10% of the possible interactions between proteins were mapped(Hart, Ramani, and Marcotte 2006), and current reviews of the matter are not able to found a consensus number of proteins and their interactions. In an extensive review by Keskin, Tuncbag and Gursoy (2016), the authors describe six databases for pathways and 18 databases of PPI with curated or predicted sources

from their data, with a number of proteins ranging from 11,836 to more than 5 millions (in STRING database, from more than 2,000 species of organisms(Szklarczyk et al. 2015)) and with 4,303 to more than 3.5 millions of interactions described(Keskin, Tuncbag, and Gursoy 2016). This amount of data makes necessary an extensive array of bioinformatic tools to analyze it, and several software applications have been designed to work with it, and specially to integrate and synthesize PPI information in a visual way, as Cytoscape(Shannon et al. 2003), the most widely used. The process to target specific PPI for its implication in diseases is challenging and relies extensively into a previous description of the specific PPI, although several laboratory methods are emerging to test and discover specific PPI interactions experimentally(Hayes et al. 2016). Predicted PPI interactions or even methods that rely in the physical contact between two proteins to infer a PPI have false positives and false negatives, since the spatial and temporal characteristics of a PPI have to be considered (a PPI could not be observed at a given moment where the two proteins are not in contact or the physical contact can also be product of random molecular movements rather than being related with a cell function) (Hayes et al. 2016). Nonetheless, NGS data of mutations in the coding sequence, combined with PPI and functional pathways data, specially when is extracted from curated databases, has proven to be a potent method to assign functional consequences to the mutations found in a genome(Porta-Pardo and Godzik 2014; Porta-Pardo et al. 2015), and allows to prioritize genes implied in certain pathways

specific for the disease, even if the mutated gene has not been related to the disease before(Quaynor et al. 2016), and also permits the discovery of novel defective pathways that can explain the phenotype(Aksentijevich 2015).

1.3.4 Mutation rates

DNA is damaged and repaired constantly, but sometimes this damage is not properly repaired, thus altering the original sequence through a mutation. The mutation may be heritable if it affects a germ line cell, while in other cases may have consequences for the phenotype of the organism, as the development of a cancer in proliferative cells or aging in other kinds of cells. Phenotypic mutation rates are larger than genotypic ones by one order of magnitude, as far as 20:1 estimated by Conrad et al. This proportion may be inflated due to the use of cultured cell lines that have undergone an abnormal number of replications and are under mutagenic conditions, but nonetheless it reflects a certain bias respect germ line mutation rates(Conrad et al. 2012). It appears that DNA replication has an accuracy far better than the transcription and translation machinery, and in general evolution doesn't seems to create a pressure enough to level phenotypic mutation rates to genotypic ones(Bürger, Willensdorfer, and Nowak 2006). As the consequences of somatic mutations are seen in the phenotype and are not inherited, this aspect of mutation is usually not taken in

account in evolutionary studies, but its contribution to cancer and other diseases makes somatic mutations an important factor to have in mind in clinical studies.

Mutation rates vary greatly between different organisms, although there is a strong relative correlation with the genome sizes. Before the implantation of the NGS technology, estimating mutation rates was a tricky business and empirical proof could only be obtained from small regions of the genome or organisms with small genomes. Initial estimates for mutation ratios where obtained from the hypothesis that Mendelian diseases arise from a balance between mutation and natural selection in the population, or from phylogenetic analysis that assumed that the rate at which changes accumulates on neutral evolving sequences over millions of years can be converted to the mutation rate per generation (*e.g.* the divergence between humans and chimpanzees in pseudogenes). Those early estimates are biased because they have to make assumptions over divergence times and population sizes, in the case of phylogeny-based approaches, or over phenotype-genotype relationships and mutational target sizes in the case of estimates from Mendelian diseases(Ségurel, Wyman, and Przeworski 2014). Nowadays whole genomes can be sequenced at affordable prizes and therefore mutations rates can be estimated using the sequencing of trios (parents and offspring). The most important difference in genotypic mutation rates occur across the genome sequence, where

in CpG sites in mammals can be tenfold higher respect the other sites(Hodgkinson and Eyre-Walker 2011).

There are also significant differences between chromosomes, mostly between sexual chromosomes an autosomes. Y chromosomes have rates that are at least 50% higher than autosomes, and in opposite X chromosomes mutations rates are 30% lower respect the autosomal chromosomes of great apes. This difference might be because point mutations are generated during replication, thus more cell divisions occur during spermatogenesis than oogenesis(Hodgkinson and Eyre-Walker 2011). The parental origin from the mutations affects all the genome, not only sexual chromosomes, where, although the ratio may vary greatly within and between families, as many as 92% of all *de novo* mutations have been found to be linked to the paternal germline(Conrad et al. 2012). Older fathers do not only produce gametes with more *de novo* mutations but also those mutations are more likely to have functional consequences, since there is an enrichment in exonic regions driven by CpG dinucleotides(Francioli et al. 2015). This age-fueled sex bias in mutation rates it's observed in many animal species, but is specially high in great apes, while animal species which produce a similar amounts of eggs and sperm experiment the same *de novo* mutation ratio in males and females(Sayres and Makova 2011).

If all the issues listed in the previous paragraph are taken in consideration, it's understandable that different estimates for the mutation rate arise according the genomic region considered. The lower estimates are obtained from studying the whole genome (around $1.1 \cdot 10^{-8}$ mutations per bp per generation) and the higher ones from phylogenetic studies (~$2.3 \cdot 10^{-8}$ mutations per bp per year) or from some particular cases. Figure 4 taken from an impressive review by Segurel, Wyman and Przeworski resumes various estimates from several studies(Ségurel, Wyman, and Przeworski 2014).

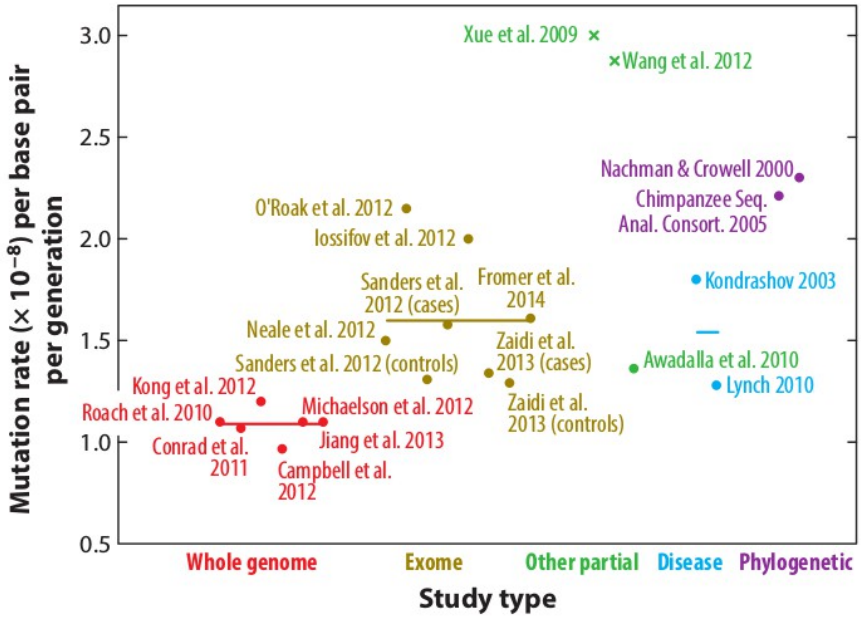


Figure 4: Different estimates of mutation rates according by the type of study, taken from Ségurel et al., 2014

1.3.5 Dominant and recessive mutations

Well before Mendel's time, breeders knew that certain traits could be hidden in an organism to appear again in his descendants. Mendel referred to this concept as dominance, where visible traits are dominant over recessive characters which are only manifest in some individuals(Mendel 1865). Mendel genius insight was the understanding that those characters come in pairs and that each one of the pairs is inherited from one of the parents, and he described that well before the concepts of gene or chromosomes were established. The pairing of characters in each individual is a direct consequence of the duplicity of genomic information, although not all living beings have two copies of the genome. Many unicellular organisms are haploids and others can have as many copies of their genome. Diving into ploidy strangeness, it is known that many algae and fungi, as well as some plants and animals, alternate generations between haploid and diploid organisms, sometimes producing entities with very different life styles(Nuismer and Otto 2004). Polyploidy is relatively common in plants and fungi, and in certain families of teleostei fishes and amphibians. A model proposed by Mable suggest that polyploidy is more common in organisms that produce a high number of gametes where random meiotic problems can be filtered, with assortive mating that can be modified by polyploidization and whose populations are exposed to environmental fluctuations during the breeding season that increase

frequencies of unreduced gametes(Mable 2004). Nonetheless, the focus of this thesis is on organisms with diploid genomes as humans or the great apes.

Most of the vertebrates, and almost all the mammals, are diploids, meaning that they have two copies of their genome. This allows a greater phenotypic flexibility than in haploid organisms, since at each position of the genome there could be two possible base pairs. Indeed, some studies point to this phenotypic flexibility as one of the driving forces towards polyploidy, putting as example host-parasite interactions. Most parasites are unicellular organism with haploid genomes, while their hosts are diploid. Parasites need to avoid their host's immune systems in order to survive, thus having an haploid genome that generates a narrow set of membrane antigens and elicitors potentially reduces the chances of being recognized. On the other hand, host defenses need a wide range of molecules to detect non-self cells in the organism, thus diploidy in hosts it's favored against haploidy since it implies more possibilities of variation(Nuismer and Otto 2004). There exists cases in humans of mixoploidy, meaning that some cell populations have 3 or 4 copies of the genome coexisting in the same individual. Mixoploidy in humans is a form of karyotipic disease since people afflicted have physical and mental anomalies(Järvelä et al. 1993; Edwards et al. 1994).

Which one of the two possibilities are used in the encoded protein

depends on many factors, being the most important the dominance of the allele. The dominance of an allele over another reflects that only the dominant allele has a visible contribution to the phenotype, so alleles can be dominant when the allele is expressed over the other, recessive when the allele it's not expressed in front of the other, co-dominant where both alleles are expressed or it can have an incomplete dominance, when the phenotype is intermediate between both alleles. To complicate things further, alleles can be epistatic, meaning that they can influence the expression of alleles from other loci. Study of Mendelian genetics it's important since there are many diseases that are influenced by a single loci. The molecular mechanisms that directs trait's dominance are related by the kind of mutation and the particularities of the gene, in special with his dosage dependency. Some genes need to have two functional copies in order to retain their normal function, thus if one of the copies of the genes loses his function due to a mutation, we would observe a dominant mutation and therefore the gene is haploinsufficient. In the opposite case, if the gene only needs one functional copy to remain functional, a single mutation in one of the two copies will be silent, so the mutation is recessive (it needs to affect both alleles to have a phenotypical effect) and the gene is haplosufficient. Haplosufficient genes can be inactivated either by homozygous mutations (the same mutation in both alleles) or by compound heterozygous mutations (two or more mutations in different alleles, affecting together the two copies of the gene). When the effect of the mutation on a single copy of the gene

recapitulates halfway the function of two healthy copies then it appears the phenomena of incomplete dominancy.

In the moment that a mutation appears in a population, it usually does it in an heterozygous state (unless it's in a sexual chromosome). Although some mechanisms can cause apparent homozygous mutations that aren't product of the inheritance from two parental sources, as uniparent isodisomy(Spence et al. 1988), gene conversion or the combination of a deletion and a heterozygous mutation, homozygous *de novo* mutations can also occur(Bafunno et al. 2013), but there are so few cases that they can be considered anecdotic. In general, homozygous mutations appears after the mutation has been fluctuating several generations in the gene pool, has reached a certain frequency in the population or in cases of inbreeding where there are large regions of the genome sharing a recent ancestry.

1.3.6 Number of genes in the genome

In 2001 the initial draft of the first human genome was published after 13 years of immense scientific effort, financed by both public(Lander et al. 2001) and private sources(Venter et al. 2001) (Celera Genomics joined the race with a big inversion in 1998, aiming to patent parts of the genome, but in 2000 the U.S. President Bill Clinton announced that the human genome would not

be patented and that it will be available for free to the scientific community). One of the first surprises surrounding the human genome where the number of genes. Pre-release estimates put that number in more than 100,000 protein genes and they varied greatly between the source of the estimation(Fields et al. 1994; Liang et al. 2000; Pertea and Salzberg 2010), but the more accurate estimates from genomic information lowered those estimates to 20,000 – 25,000 genes. The exact number of genes of a human genome it's still in debate and fluctuates according gene database policies, and recent studies suggests that the number of genes varies even between human genomes. A possible estimate (Figure 5) it's a number between 21,000 and 23,000(Pertea and Salzberg 2010). If we add concepts like ribosomal genes, non-coding genes or functional pseudogenes, the best answer to "What is the number of genes in the human genome?" might be an angry "First get straight what is a gene!". Post-ENCODE studies suggest to move from a protein-centric vision of the gene to a more functional definition, viewing a gene as a "union of genomic sequences encoding a coherent set of potentially overlapping functional products"(Gerstein et al. 2007).

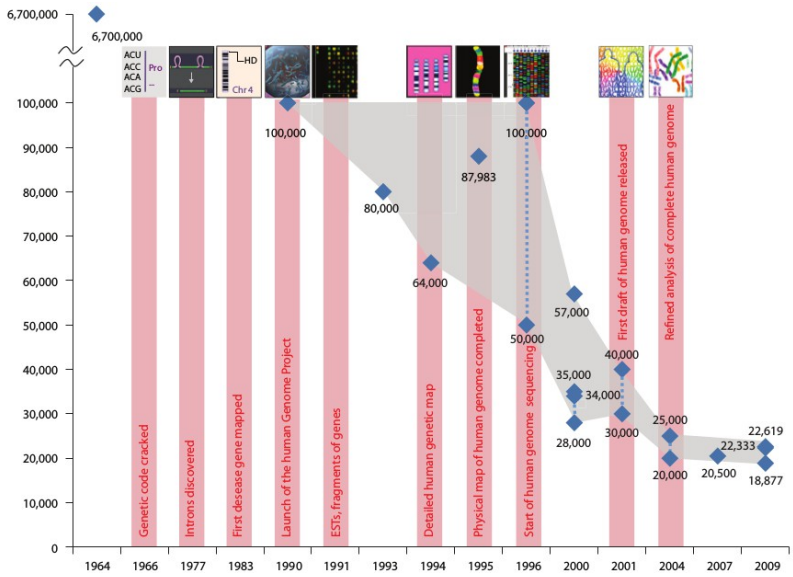


Figure 5: Trend of estimated number of human genes together with human genome-related milestones, from Pertea and Salzberg (2010)

The surprise was even bigger when the scientific community realized that less than 3% of the human genome's sequence encoded proteins(Lander et al. 2001; Elgar and Vavouri 2008). This proportion varies greatly between species, and specially between prokaryotes and eukaryotes. As little as 20% of the DNA of the prokaryotes is non-coding, whilst eukaryotes have much higher proportions of non-coding DNA. Interestingly, plants have huge variations in their non-coding genome sizes, from as low as 3% in the tiny carnivore plant *Utricularia gibba*(Ibarra-Laclette et al. 2013) to as much as 85% in the common corn *Zea mays*(Michael 2014). The case of *Utricularia gibba* is specially interesting because it harbors 28,500 genes in a genome of 82Mb, with a minimum of

its sequence dedicated to regulatory elements and non-coding DNA, but stills is perfectly functional(Ibarra-Laclette et al. 2013).

The non-coding DNA was initially debunked as “junk” but a growing body of evidence is pointing towards some functionality to many regions of the non-coding DNA. Aside from non-coding regulatory elements, pseudogenes and other regions that nowadays have been found to be functional, there are fragments of sequence originated from transposable elements, repetitive sequences and even mysterious ultraconserved regions that retains 100% homology between humans and rodents and that are under a stronger negative selection than protein-coding regions(Elgar and Vavouri 2008). Moreover, the recent ENCODE project has found that more than 80% of the genome is actively transcribed and it has sparkled an intense debate about where this transcription is related with functionality. Apparently, the genomic era is still a newborn and holds many surprises to be discovered.

Despite all that, investigation keep being biased towards the study of mutations in the coding regions. At the date as many as 85% of the mutations know to be related with diseases are found in the coding regions(Choi et al. 2009), but research in the field of regulatory elements is increasing exponentially. Nonetheless, there are still more than 3,000 Mendelian diseases with unknown genetic etiology(McKusick 2007; J. Kaiser 2010; Bamshad et al. 2011). Usually mutations are listed according their effect in the translated

protein, but other aspects should be considered like their frequency in the population or the evolutionary conservation of the position where they happen.

1.4 Types of mutations:

"*Quieres indentificarnos, tienes un problema.*"

La Polla, No Somos Nada

a) Synonymous variants:

The fact that the genetic code is degenerated allows that some substitutions in the nucleotide sequence doesn't imply an amino acid change, since the same amino acid can be encoded in different codons. It's estimated that a 20% of the nucleotide replacements caused by mutation are synonymous(Kimura 1968). Those mutations are thought to be evolutionary neutral and that is supported by the fact that in a typical human genome they account for, at least, as much as non-synonymous variants (10,000 to 12,000 synonymous variants per genome respect 10,000 to 11,000 non-synonymous variants(Abecasis et al. 2010)), but sometimes they are not exactly silent and show some evidences of selection(Chamary, Parmley, and Hurst 2006; Shabalina, Spiridonov, and Kashina 2013; Du et al. 2014). Some codons coding for the same amino acid are translated at different speeds or with different accuracies by the molecular translation machinery and this can have effects on the protein structure(Kimchi-Sarfaty et al. 2007; Tarrant and Von Der

Haar 2014; Jacobson and Clark 2016), and synonymous changes may also introduce changes in methylation patterns, in splicing or in the motif to be recognized by transcription factors, miRNAs or other regulative elements, or directly affect mRNA stability(Sauna and Kimchi-Sarfaty 2011). Despite that, in general synonymous substitutions are still regarded as neutral, at least in mammals. Motoo Kimura proposed in 1968 that many of the substitutions seen in a genome are neutral and that genetic drift is the main force controlling allelic frequencies(Kimura 1968). In order to be described as neutral, a mutation must have a fitness effect considerably smaller than the inverse of the effective population size. Henceforth, species with smaller population sizes, as the vast majority of mammals, have higher tolerance for slightly deleterious variants, which effectively act as neutral(Kimura 1968; Ohta and Gillespie 1996; Sauna and Kimchi-Sarfaty 2011). Those neutral or nearly neutral mutation have been understandably neglected in study of diseases, but recent research lines suggest that codon usage bias should be taken into account in the future(Sauna and Kimchi-Sarfaty 2011).

In evolutionary genetics synonymous changes are used to calculate the ratio of non-synonymous to synonymous mutations, which is informative of the selection acting upon a determined sequence due the neutrality attributed to this kind of mutations. This could be informative also of sequence constrains for mutation in some genes that are under purifying selection.

b) Non-synonymous variants:

Non-synonymous mutations, also called missense mutations, are sequence substitutions that have as a result a change in the amino acid encoded. Normally short insertions and deletions are not included in this category, but here I'm referring to mutations that change one or a few amino acids of a protein, independently if those changes are originated from one or few single nucleotide variations (SNVs) or from small insertions or deletions (indels) that don't cause a frameshift, *i.e.* indels of 3 or multiples of 3 bases that remove or introduce codons without altering the next ones in the sequence. Depending on the change the consequences in the final protein can be small and have no impact on the function of the protein or it can collapse in a total loss-of-function on the protein.

The main differences between amino acids are charge, hydrophobicity, size and their functional groups. Thus, when an amino acid is replaced in the protein, in general the impact of the change in the function of the protein is related with how different is the replacement in those aspects in respect to the original amino acid. Other important aspects are the functional domain of the protein affected by the change and the effects in the altered protein.

Two random human genomes usually differ between them in more than 10,000 non-synonymous SNV (nsSNV)(Ng et al. 2008; Kim et al. 2009; Lupski et al. 2010; Stitziel, Kiezun, and Sunyaev 2011),

and in the entire human population you can found one or more SNVs in at least 81% of the genes(Chakravarti 2001). Early estimates gave a number of nsSNPs much higher due to the overestimation in the number of genes, in the range of 24,000 to 40,000 nsSNPs per human genome(Cargill et al. 1999), and supposed that 2,000(Sunyaev et al. 2001) or as much as 9,500(Chasman and Adams 2001) of those nsSNPs in each individual could affect protein function being slightly deleterious, but lethality studies and the first prediction algorithms lowered those numbers in, at least, two orders of magnitude. Nowadays one of the biggest struggles that genomics is facing is not to found mutations, since NGS allows to do it relatively fast and easy, but to understand their effect on the protein. This is a huge problem since proteins are complex molecules, with a functionality dependent of their specific three-dimensional structures and the interaction with other proteins. The final proof of the impact of a specific mutation in the protein usually need long and costly experimental functional studies, but this approach is not possible when the study is dealing with great numbers of mutations. Therefore, an initial step where mutations are prioritized is needed. In order to predict the phenotypic impact of a non-synonymous mutations, several predictors have been developed. The most widely used are SIFT(Ng and Henikoff 2001) and Polyphen(Sunyaev et al. 1999). SIFT (Sorting Intolerant From Tolerant) is a sequence-homology based tool that relies in multiple alignment information to predict tolerated and deleterious substitutions for every portion of the query

sequence. Each sequence is queried against similar sequences that may share similar function to obtain the multiple alignment, and then calculates the probabilities for all the possible substitutions in each position of the alignment. Substitutions in uncharacterized proteins can only be evaluated when there are homologous sequences with known function(Ng and Henikoff 2001). Polyphen (from polymorphic phenotyping) uses a statistical and heuristic approach giving weights as a function of both sequence and alignment position to amino acid type occurrences(Sunyaev et al. 1999). It exploits several programs to find nsSNPs in known genes, determine the site of the substitution within the protein and perform a profile analysis of homologous sequences to test incompatibility with the possible spectrum of mutations. After that, the substitutions are mapped to the known three-dimensional structure of the protein and structural parameters are used to evaluate the effect of the substitution, as contacts with critical sites, ligands and other polypeptide chains(Ramensky, Bork, and Sunyaev 2002). Usually predictors rely in the structure of the protein, in the sequence or in the annotation of the substitution. Each method has his caveats, as the lack of knowledge for three-dimensional structures of many proteins or the use of homologs/orthologs in sequence methods that may have different functions. Currently amino acid substitution prediction methods are useful for predict the effect of rare variants but still fail at distinguish common variants involved in common diseases from normal variation(Castellana and Mazza 2013).

Besides prediction methods, sequence conservation can be a good estimator of the deleteriousness of a mutation. GERP (stands for Genomic Evolutionary Rate Profiling) is based in the comparison of orthologous genomic sequences between 29 mammals (33 in the most actual version (Davydov et al. 2010)) to characterize genomic regions under purifying selection. It assigns an score to each base pair in relation with the substitution deficit found, where higher scores are directly correlated with higher conserved bases(Cooper et al. 2005). Therefore, when a nsSNV affects a base with a high conservation score, is more likely that this change may have negative consequences.

c) Loss-of-function variants:

When the genetic code was discovered, it became evident that point mutations could change an encoded protein for another, or to produce a termination codon. Those codons tells the translation machinery where is the end of the protein, and therefore no more amino acids should be added from this point. It is evident that when a mutation introduces a stop codon prematurely in the sequence, the protein should be shorter and therefore it might loss some its functionality, in the case of LoF mutations with lower levels of activity or expression, called hypomorphic, or all of it, in the case of null LoF alleles. Furthermore, in eukaryotes, there's a strong

pressure against translating mRNAs that contains premature stop codons, to the point that evolution has created a nonsense-mediated decay (NMD) pathway that eliminate those transcripts. NMD is though to be necessary due to the deleterious effects that some shortened proteins may have. The process detects the aberrant transcripts after the first round of translation, when the exon-exon junction complexes (EJC) are removed from the mRNA. When a premature stop codon appears, the EJC downstream of the mutation remain attached when the ribosome releases the mRNA. This is the signal recognized by the NMD pathway and translation is interrupted. In vertebrates, when a stop codon is within 50 nucleotides of the last EJC translation proceeds as normal, otherwise usually NMD takes place. This implies that the majority of the homozygous stop gain mutations impacts on translation by not producing any amounts of protein and with the subsequent total loss-of-function of the gene in question(Baker and Parker 2004; Chang, Imam, and Wilkinson 2007). NMD impacts in the allele frequencies of the reads produced by NGS, producing a bias in the number of reads that map in the alternate allele to lower values in genes with NMD, as illustrated in figure 6(MacArthur et al. 2012a).

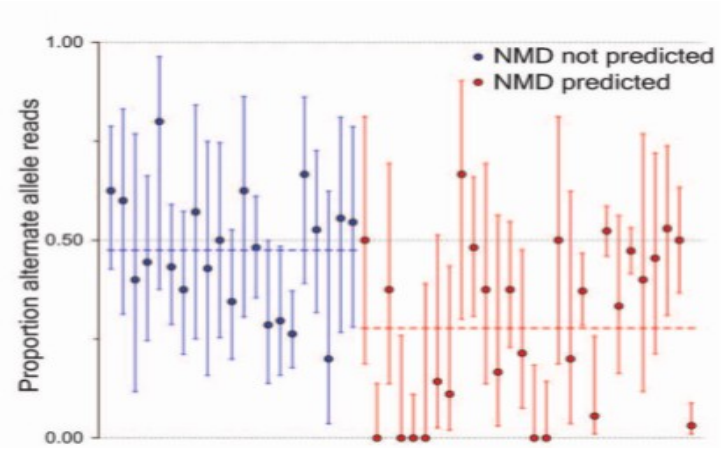


Figure 6: Allele-specific expression of nonsense variants obtained from RNA sequencig of 119 lymphocyte cell lines. The x axis distribution of the variants is arbitrary, the meaningful information is given by the red (NMD predicted) and blue (NMD escape) dotted lines, from MacArthur et al. (2012).

As mentioned before, one of the key experiments in the discovery of the genetic code used insertions and deletions to provoke a frame shift of the codons and therefore to produce a different protein, usually non-functional. Those frame-shifting indels are potentially deleterious, and some of them may produce a down regulation of the transcript by NMD, specially if the aberrant protein has a premature stop codon.

Besides stop gains and frameshifting indels, mutations can also produce the loss of an stop codon, adding amino acids to the protein until a new stop codon is found in the intronic sequence. Likewise, start codons can be lost and introduced erroneously in the sequence.

Finally, mutations in the splice sites have important consequences from translation, since they can remove exons or introduce intronic sequence in the translation. The most strong cause-effect relationship in LoF mutations is observed in stop gain mutations, frameshift indels and splice site disrupting mutations, and start gain and loss and stop loss are more difficult to asses since those mutations can be rescued by nearby features in the sequence.

Sometimes the change introduced in a protein by a mutation, regardless of it's nature, can produce a gain of function (GoF) in the protein. The distinction between LoF and GoF is not always clear, since both can cause a disease phenotype and many of the particular mechanisms of action of LoF remain unknown. GoF have been recently a hot topic as they can confer increased transmissibility or pathogenicity when affecting viral genes, as in the case of the highly pathogenic avian influenza virus (HPAI). Two laboratories have experimented with different strains of influenza A(H5N1) and have found GoF mutations that increase the airborne transmission of the virus between ferrets(Herfst et al. 2012; Imai et al. 2012), sparkling an intense debate in the grounds of biosecurity policy, which resulted in calling for a voluntary year long moratorium on this kind of research by the US government until measures to regulate the GoF research on potentially dangerous biological agents could be established(Kilianski, Nuzzo, and Modjarrad 2016). A positive outcome of this issue has been the increase of research in GoF mutations, specially in the field of immunology. For example,

heterozygous GoF mutations in the CARD11 gene have been found to be causative of B-cell expansion with nuclear factor kappa-B (NF-kB) and T-cell anergy (BENTA) disease, a primary immunodeficiency syndrome, as well as in the activated PI-3Kδ syndrome (APDS), caused by GoF mutations in the PI-3K subunits genes(Arjunaraja and Snow 2015). The research in those immunological diseases give an idea of how GoF mechanisms work, either through overactive signaling in certain pathways, hyperactivity of the encoded protein or neomorphisms (acquiring a new function rather increasing the normal function of the protein) (Arjunaraja and Snow 2015; Boisson, Quartier, and Casanova 2015).

The lack of knowledge for the precise molecular outcome and the fact that some LoF variants produce a GoF in the protein has lead to the most recent publications to use the category of protein truncating variants (PTVs) instead of LoF. Rivas et al. (2015) define PTVs as single nucleotide variants predicted to introduce a stop codon or to disrupt an splice site, small indels that disrupt the reading frame of the sequence or large deletions removing the entire coding sequence. The study has measured the allele-specific expression (ASE) of the reads generated by NGS in different tissues, in order to detect differences between the two haplotypes in an individual. They have found higher proportion of strong and moderate allelic imbalance between rare and singleton PTVs than in PTVs with higher frequencies, suggesting that rare PTVs are more

likely to trigger NMD. They also report a lack of evidence to support the idea of compensation dosage in heterozygous PTVs (that could maintain normal levels of expression by overexpressing the copy of the gene not affected by the PTV), which informs that biological function is maintained by homeostatic mechanisms at cellular level rather than by compensation of the expression(Rivas et al. 2015). Furthermore, some nsSNPs, although producing a change in only one nucleotide, can have profound effects at the function level, reducing or even producing a total loss of function of the protein. Those effects on functionality by nsSNPs vary widely between proteins and are dependent of the structure of the protein and the differences between the wild type amino acid and the mutated amino acid, and its characterization requires specific functional studies suited for the specific mutation and the specific protein, likewise it requires a previous knowledge of the protein function. Therefore, precise functional effects of every nsSNPs considered in exomic or genomic cannot be considered by the actual methodologies.

All LoF changes (resumed in figure 7) have the potential to be deleterious because they not only affect one amino acid, but they can change the reading frame of one or more transcripts. Therefore the characterization of LoF mutations has great importance at evolutionary levels and in clinical studies. Stop gain LoF variants have been implicated in at least 15% of the monogenic heritable diseases(Mort et al. 2008; Balasubramanian et al. 2011) and many

examples of disease phenotypes caused by other categories of LoF variants can be found in the literature. Nonetheless, a low proportion of LoF have been under positive selection and can have a positive outcome. For example, a recent study has found a link between a LoF variant and selenium resistance in *Arabidopsis thaliana*(Jiang et al. 2016), or the multiple LoF mutations involved in the resistance to the oxamniquine drug found in Brazilian parasite trematodes(Chevalier et al. 2016). Even some LoF variants had been historically relevant, as in the case of a stop gain mutation in the DMT3 gene on horses, which gave the ability to produce a comfortable pacing in the 'gaited' horses and therefore making them more suited for long rides(Wutke et al. 2016).

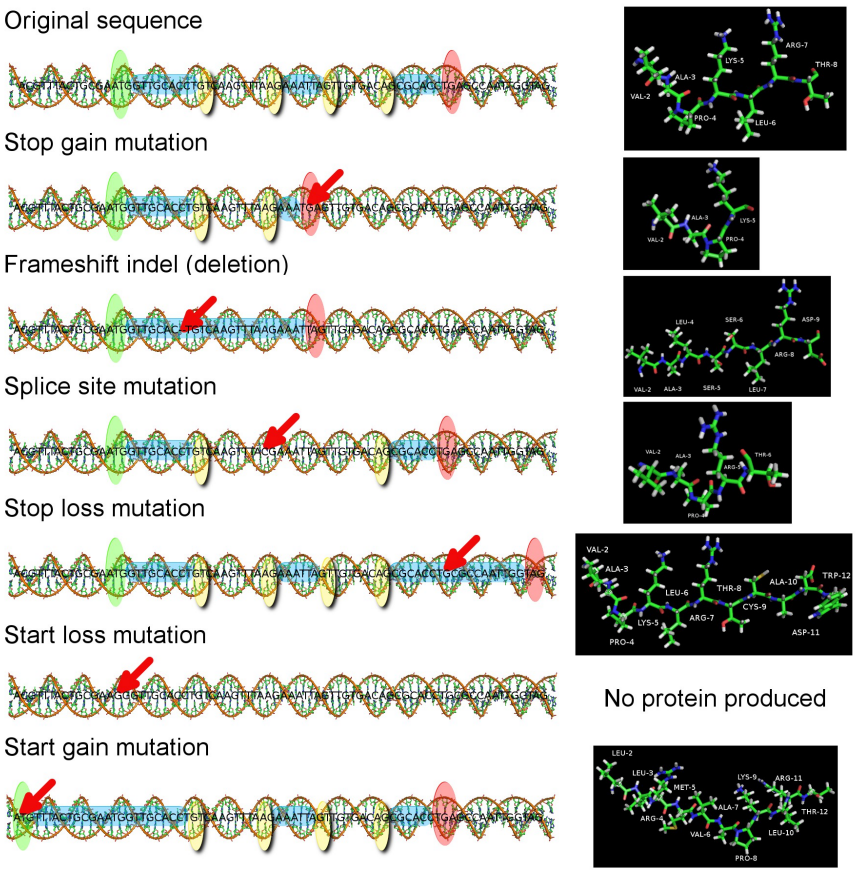


Figure 7: Effect of the different LoF in a protein sequence, with the protein produced (black background). The start codon is shadowed in green, the exonic sequence is shadowed in blue, the splice sites are shadowed in yellow and the stop codon is shadowed in red. The substitution producing each LoF is pointed by the red arrow.

Genes harboring LoF are strongly enriched in functional categories related with olfactory reception(MacArthur et al. 2012b; Kaiser et al. 2015) and taste receptors, specially for bitter and sour flavors(Fujikura 2015). This set of genes is also, unsurprisingly, depleted in categories involved in development, protein-binding and

transcriptional regulation. The accumulation of LoF in olfactory reception genes is not exclusive from humans, but is seen also in great apes, and even in domesticated bovines(Das et al. 2015). Moreover, LoF in high conserved regions have much lower frequencies, reinforcing the assumption of their potential deleteriousness.

It's logic to think that a loss-of-function in a protein, no matter how it happens, should produce abnormalities in the otherwise well-honed cellular machinery, but has the high number of LoF mutations illustrate, it's not always the case. A set of LoF-tolerant genes has emerged from the publication of MacArthur et al. (2012), which contains genes harboring homozygous LoF mutations in healthy humans. Those genes are less evolutionarily conserved and have a lower number of interactions than the genome average. They differ also from the norm in having a higher ratio of non-synonymous/synonymous mutations as well as a higher number of paralogs, indicating a possible functional redundancy in some of them, as well as an overall lower values in protein-protein interaction (PPI) networks, a higher distance from recessive genes in those networks and a lower tissue-specificity(MacArthur et al. 2012a). Figure 8 show the comparison between recessive disease genes, all the protein coding genes and LoF-tolerant genes in the aforementioned aspects.

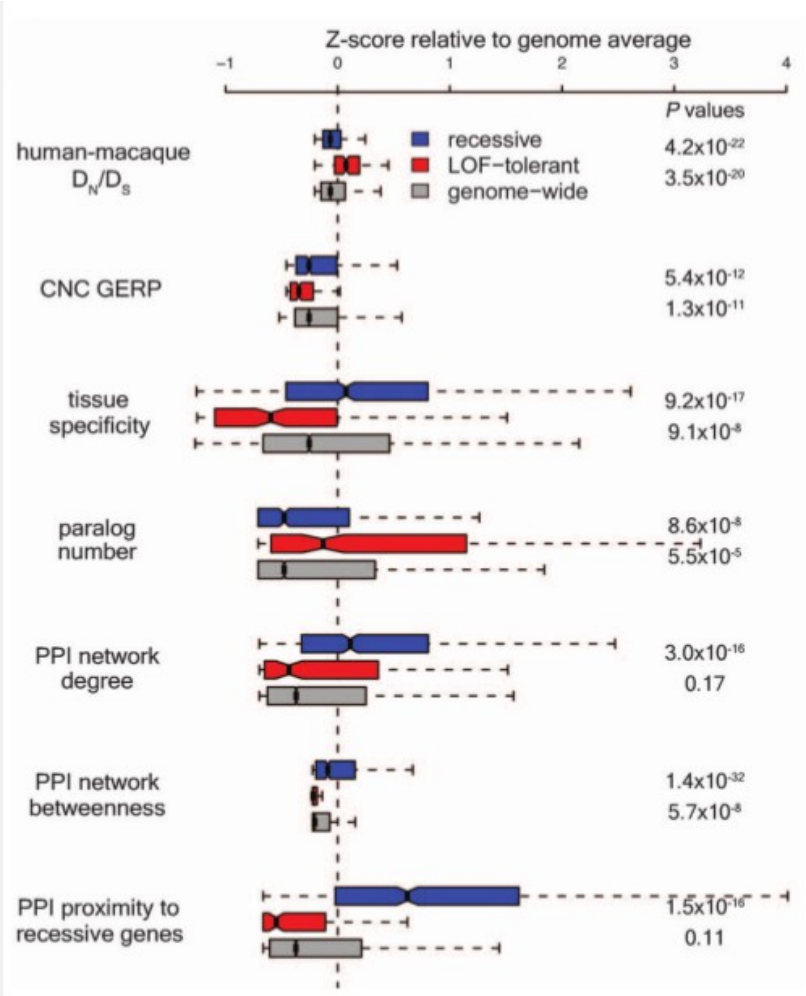


Figure 8: Distribution for a set of propierties in all protein-coding genes (gray), disease recessive genes (blue) and LoF-tolerant genes (red), from MacArthur et al. (2012)

1.5 Mutational load of LoF variants

"*And I have a tender spot in my heart for cripples and bastards and broken things.*"

George R. R. Martin, A Game of Thrones

Until the advent of NGS, the study of LoF was limited to punctual cases, usually focused on a specific gene or pathway in relationship with some disease. Due to the high potential of LoF variants to be deleterious, they were though to be rare both in the genome and in the population. But recent genome-wide studies have allowed quantitative characterizations of LoF in humans, unraveling an unexpected high number of LoF variants in humans. Those studies presented numbers ranging from ~600 in low-coverage samples(Abecasis et al. 2010) to ~800 LoF per genome in high-coverage samples, with an average of 165 homozygous LoF variants in each genome reported from high-coverage data(Pelak et al. 2010). LoF variant discovery by NGS methods is expected to be enriched in false positives respect other categories due to the imbalance between non-random genome wide natural selection, which acts strongly against putative deleterious variants as LoF changes, and random sequencing errors that are expected to be distributed uniformly across the genome. Moreover, the sequence at the start and end of the genes seems to be more tolerant to LoF mutations(Pelak et al. 2010; MacArthur et al. 2012b), possibly due to a small effect of the mutation in the overall function of the protein when it happens near the terminus or to a rescue by

transcriptional reinitiation at an alternative start codon when the LoF variant is near the start of the gene(MacArthur et al. 2012b).

MacArthur *et al.* applied strong filtering to both low-coverage and high-coverage genomic data, obtaining a more conservative and reliable estimate of ~100 LoF variants per genome, with ~20 of them in heterozygous state and therefore potentially inactivating the gene. This implies that genomes is more resilient to gene inactivating mutations than expected and that clinical analysis of genomic data should take into account LoF-tolerant genes in order to prioritize the selection of putative causing variants(MacArthur et al. 2012b). The majority of LoF found in human populations are common, with allelic frequencies higher than 5% (table 3), and tend to be found in a small set of genes tolerant to inactivation(Sleiman et al. 2014).

Variant type	Allele frequency (%)		
	<0.5	0.5–5	>5
Stop-gain	3.9–10	5.3–19	24–28
Stop-loss	1.0–1.2	1.0–1.9	2.1–2.8
Indel frameshift	1.0–1.3	11–24	60–66
Splice site donor	1.7–3.6	2.4–7.2	2.6–5.2
Splice site acceptor	1.5–2.9	1.5–4.0	2.1–4.6

Table 3: LoF allele counts in 1,092 genomes across three allele frequency bins, taken from Sleiman et al., 2014

NGS technologies are allowing to explore the genomes of many organisms (besides human genomes), but normally LoF are briefly covered. For example, in a study including 48 pig genomes there

are reported an average of ~30 nonsense mutations per genome, without providing any information about other LoF categories(Groenen et al. 2012). To the date I have found two studies that cover extensively LoF in animals. A study in genomes of four non-related cows from an specific breed, as well as in 288 genomes from the 1000 bull genome project to filter the LoF variants found in the 4 cows(Daetwyler et al. 2014), has found 2,145 LoF variants before filtering, with 714 in homozygosis in the four animals (probably breed-specific). After filtering, they have found 345 LoF variants for which none of the four cows nor the 288 bulls where homozygous, suggesting that those filtered LoF are truly deleterious variants. Roughly, those numbers gave ~536 LoF variants per genome before filtering and ~86 after filtering in cattle(Das et al. 2015). Previous studies in cattle, although not as thorough in the types of LoF studied, gave similar numbers (~40 heterozygous and ~4 homozygous LoFs including only stop gain introducing and splice site mutations)(Charlier et al. 2014). The second study analyzes LoF variants in the genomes of four rhesus macaques, but the publication is mainly focused on the differences found when using two different reference genomes of *M. mulatta* to map the reads. Nonetheless, they report ~390 LoF variants per genome, with ~44 of them in homozygosis in the newest and best genome(Cornish, Gibbs, and Norgren 2016).

1.6 Relationship between LoF mutations and disease

"*In examining disease, we gain wisdom about anatomy and physiology and biology. In examining the person with disease, we gain wisdom about life.*"

Oliver Sacks, The Man Who Mistook His Wife for a Hat

The potential deleteriousness of LoF mutations, either through heterozygous LoFs in haploinsufficient genes or through homozygous LoFs or compound heterozygous LoFs when the gene is haplosufficient, converts them in a perfect candidate for be causal variants in diseases. Molecular research in diseases it is widely accepted to be started with the identification of the molecular defect causative of sickle cell anemia. Vernon Ingram identified first the modified amino acid in the peptide(Ingram 1956) and later, altogether with Morgan, found the non-synonymous nucleotide substitution responsible in the HBB gene(Hunt and Ingram 1958). Since then the association between nucleotide changes and diseases has increased, specially when the prices of NGS dropped to affordable values by most of the scientific studies. The first variant discovered in a monogenic disease by exome sequencing was published in 2010(Ng et al. 2010), and nowadays more than a third (~3,000) of the Mendelian diseases known to be caused by a single gene have been associated to a genetic molecular defect(McKusick 2007; J. Kaiser 2010; Bamshad et al. 2011). LoF variants are very useful in this kind of studies since its impact on the protein usually more dramatic and harmful than non-synonymous variants and their cause-effect is usually easier to prove than in the case of non-

synonymous variants. Nonetheless, many examples of LoF variants have been found to be almost neutral or even benign or protective. Some examples are the AB0 and FUT2 polymorphisms found to be under balancing selection(Calafell et al. 2008; Ferrer-Admetlla et al. 2009; Casals et al. 2009), LoFs in the genes ACTN3 and CASP12 that increases athletic performance(Yang et al. 2003; MacArthur et al. 2007; MacArthur and North 2007) and survival to severe sepsis (respectively), or LoF mutations in the PCSK9 gene causing hypocholesterolemia and therefore protecting from coronary hearth disease(Cohen et al. 2006; Sleiman et al. 2014). Those examples warns against considering a direct causality when finding LoF variants in a clinical screening for a disease. Traditional methods for predicting mutation's deleteriousness have been focused on non-synonymous variants(Sunyaev et al. 1999; Ng and Henikoff 2001a; González-Pérez and López-Bigas 2011), and it exists a general tendency to assume that all LoF are highly deleterious, but those examples reinforce the idea that further functional validation is needed.

Phenomena like incomplete penetrance, where not all the individuals carrying the mutation present the disease phenotype, adds another layer of complication when assessing the deleteriousness of a LoF variant. It is specially difficult when the gene is thought to be haploinsufficient but some healthy LoF carriers can be found, as reviewed by Ropers & Wienker (2015) in haploinsufficient genes for intellectual disability(Ropers and

Wienker 2015). Moreover, some LoF variants may manifest in a disease phenotype when the carrier is in contact with some extrinsic substance, as in the case of LoF mutations in the FLG gene causing peanut allergy(Brough et al. 2014).

After all those remarks, we can say with extreme caution that, in general, the abolition of the protein function in certain genes is usually related with a disease. An extreme case are the genes essential for the development of the organisms where the phenotypical manifestation of a LoF implicates lethality. Lethal mutations are those where the organism dies before birth or shortly after that, or before it is able to reproduce. Selection acts strongly on those kind of mutations and therefore they are extremely rare. Indirect estimates, mainly from inbreeding, yield numbers of lethal alleles per individual in a wide range: from few lethal alleles (1.4(Bittles and Neel 1994) to 3-5(Morton, Crow, and Muller 1956) in humans or 3.14 in captive mammals(Ralls, Ballou, and Templeton 1988)) to higher numbers as 12.4 in overall for 30 wild inbreed mammalian or avian species(O'Grady et al. 2006), to 100 from estimates of the relationship between genome sizes and effective population sizes(Kondrashov 1995). All those numbers suffer a variety of faults, depending of the assumptions upon which the estimate rely, so hopefully direct methods could provide a real estimate. But to determine which LoF variants found in a genome are truly lethal is not a straightforward move. The function of many genes is not well understood or directly unknown, and therefore the

impact of a great number of LoF variants could not be inferred. Thanks to NGS a handful of genes have been implied in lethality in humans(Filges and Friedman 2015), but this kind of research in humans is (fortunately) constricted to few cases due to ethical reasons, so the assessment of lethality has to rely in animal models. A study including the exome of 96 animals from 10 different breeds of cows has found a close relationship between some of the LoF variants found in *Bos taurus* genomes and embryonic lethality, even in commonly found LoF variants(Charlier et al. 2014). Furthermore, a new program for phenotyping lethal embrionic mice (Deciphering the Mechanisms of Developmental Disorders, DMDD), still in course, aims to resolve part of this problems using homozygous knockout mice for all the genes. Initial results should be taken with caution, but Mohun et al. (2013) claim that a proportion as large to one third of the knockout genes tested to the date in mice are lethal when both copies are inactivated, with no phenotypic consequences in general when the causal mutation is found in heterozygosity(Mohun et al. 2013).

Other important aspect to consider is the frequency of the disorder and the genes implied in it. Most of rare disorders are product of mutations in one gene or even in a restricted region of a gene, but more common diseases may be due to defects in a few genes in oligogenic disorders or to a large set of genes in the case of multigenic disorders. Figure 8 depicts this differences in disorders and their mutational sizes(Gilissen et al. 2011). Usually those

common disorders have a wider spectrum of disease phenotypes which are due to the different genes causing it, and diagnostic complexity increases as well. It is expected that mutations causative of a rare disorder have low frequencies in the population or even to be novel, and that more common and genetically complex disorders are caused by mutations with higher frequency thresholds or even by a conjunction of interacting polymorphisms. This makes extremely difficult to prioritize variants as it's not possible to rely solely on allele frequencies and inheritance patterns for those common variants. Moreover, some of the complex diseases can be caused by either rare variants producing more extreme phenotypes or by more common variants producing mild phenotypes. LoF, through their potential complete gene inactivation, could provide a good marker to asses which genes in complex disease are more important and which ones could be considered more as modifiers or to have a smaller contribution to the disease.

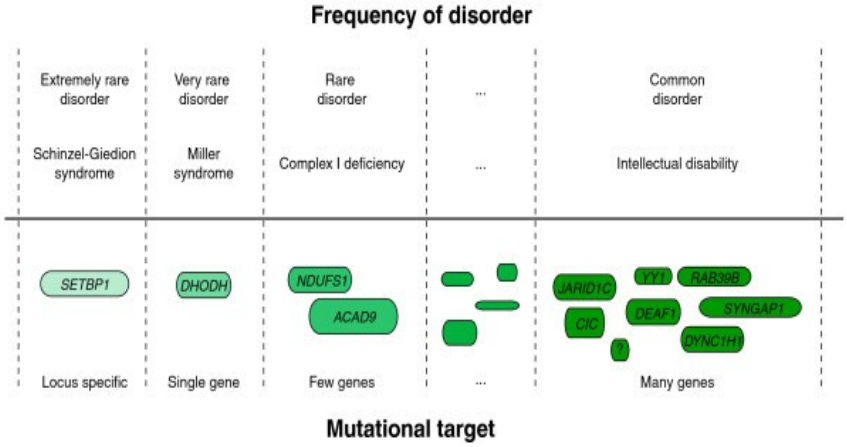


Figure 9: Representation of the frequency of a disease in relationship with the size of its mutational target, from Gilissen et al. (2012)

As disorders grow in frequency and complexity, they tend to have more genes implied in their etiology and therefore the analysis to discover causal variants in those genes increases in difficulty. The analysis of complex diseases is moving to analyze pathways rather than a few genes in order to tackle this problem, as traditional analysis exploring single genes only identifies a small amount of the susceptible genetic variants and its contribution to the understanding of complex diseases is limited. It exists a growing line of research demonstrating, through genome-wide association studies (GWAS), that genetic risk to complex diseases has is basis in large sets of genes with slightly effect factors through their interactions in a modular fashion, and that the dissection of the interaction between all the disease genes and their functionality is essential for their understanding(Freimer and Sabatti 2007; Cordell

2009; Moore and Williams 2009; Thomas 2010; Jin et al. 2014). Many studies show the power of pathway analysis using large-scale genetic datasets, as example for diabetes(Lee and Song 2016), Parkinson(Song and Lee 2013), autism(Wen, Alshikho, and Herbert 2016), schizophrenia(Crisafulli et al. 2015) or rheumatoid arthritis(Song, Bae, and Lee 2013).

1.7 LoF mutations in the context of Common Variable Immunodeficiency

> "Armour is part of a state of mind in which you admit the possibility of being hit."
>
> **Joe Abercrombie**, The Heroes

The human immune system is composed of two broad components that work together to fight infections: the innate immune system and the adaptive immune system. The innate immune systems precedes the adaptive systems and increases his response by allowing the host to differentiate self cells from pathogens, relying in a repertoire of receptors to target conserved microbial components or biochemical signatures that are telltales of infection. The recent discovery of receptor families instrumental in the first-line recognition of microbes and the regulation of the immune systems has empowered the recognition of human immunodeficiency phenotypes(Wong et al. 2013). Primary immunodeficiencies (PIDs) are a large and heterogeneous group of diseases caused by inborn defects in the innate immune system. Nowadays there are described

over 200 PIDs with an incidence rate of 1 in 10,000 births, of which 65% are related with antibody immunodeficiencies(Ebadi, Aghamohammadi, and Rezaei 2015). Classification of PIDs is complex and traditionally was based in the phenotypic description, but recently it has been expanded by the relationship between genotype and phenotype, which relies heavily on the molecular diagnostic of the gene implied, and henceforth it has been greatly improved thanks to the expanse of NGS technology. But research is revealing a confusing panorama were different mutations in the same gene can cause different phenotypes while defects in different genes can produce the same phenotype(Maggina and Gennery 2013). The PID's worldwide distribution is variable and the frequency of the subtypes is influenced by ethnicity and the rate of consanguinity. As survival of individuals affected by severe PIDs depends on the access to advanced medicine, the incidence in poor countries may be underestimated and is important to raise the public awareness on those matters(Ebadi, Aghamohammadi, and Rezaei 2015; Al-Herz et al. 2014).

Common variable immunodeficiency (CVID) is the most common primary immunodeficiency, with a prevalence of 1:10,000 to 1:50,000 in North America and Europe(Saikia and Gupta 2016). CVID is defined as an heterogeneous group of disorders characterized by antibody deficiency, hypogammaglobulinemia, recurrent bacterial infections and the inability to mount an antibody response to antigen, but it's diagnostic has changed since its first

molecular cause was discovered(Grimbacher et al. 2003). CVID was described as a late-onset agammaglobulinemia or hypogammaglobulinemia to differentiate it from X-linked agammaglobulinemia in children in the 1960s, but in the late 1970s was apparent that it can also have an early-onset and its variability in levels of serum immunoglobulins and B cell numbers from patient to patient, so the term common variable immunodeficiency was coined at that time(Saikia and Gupta 2016). Nowadays diagnosis is based in a set of criteria, resumed in table 4. The most important characteristics for its definition are the low levels of the serum immunoglobulins G and A (IgG and IgA) and the exclusion of other possible causes of hypogammaglobulinemia(Chapel and Cunningham-Rundles 2009; Saikia and Gupta 2016; Conley, Notarangelo, and Etzioni 1999).

At least one of the following: -Increased susceptibility to infection -Autoimmune manifestations -Granulomatous disease -Unexplained polyclonal lymphoproliferation -Affected family member with antibody deficiency
AND marked decrease of IgG and marked decrease of IgA with or without low IgM levels (measured at least twice; <2SD of the normal levels for their age)
AND at least one of the following: -Poor antibody response to vaccines (and/or absent isohemagglutinins) I.e., absence of protective levels despite vaccination where defined -Low switched memory B cells (<70 % of age-related normal value)
AND secondary causes of hypogammaglobulinemia have been excluded
AND diagnosis is established after the fourth year of life (but symptoms may be present before)
AND no evidence of profound T-cell deficiency, defined as two out of the following (y = year of life) -CD4 numbers/µl: 2-6y <300, 6-12y <250, >12 y < 200 -% Naïve CD4: 2-6y <25 %, 6-16y <20 %, >16y <10 % -T cell proliferation absent

Table 4: Diagnostic criteria from the European Society of Primary Immunodeficiencies (ESID), from Saikia and Gupta 2015.

CVID, as its own name appoints, is the most variable PID in age of onset (although usually two big groups with an early-onset and a late-onset are differentiated) , circulating B cell numbers, serum immunoglobulin levels and etiopathogenesis. It's usually sporadic an underlying molecular defects have been found in less than 20-25% of the cases studied(Bacchelli et al. 2007; Park et al. 2009; Rodríguez-Cortez et al. 2015; Saikia and Gupta 2016), and mostly in familial forms of the disease which are though to be an overall 10%(Li et al. 2016; Chapel and Cunningham-Rundles 2009), and under replacement immunoglobulin therapy most patients can endure its principal hallmark, although is frequent to find additional

complications other than hypogammaglobulinemia in CVID patients. Additional complications normally found in CVID patients include autoimmunity, viral infections, structural diseases as bronchiectasis, lymphoproliferation and cancers, specially lymphoma (figure 8)(Chapel et al. 2008).

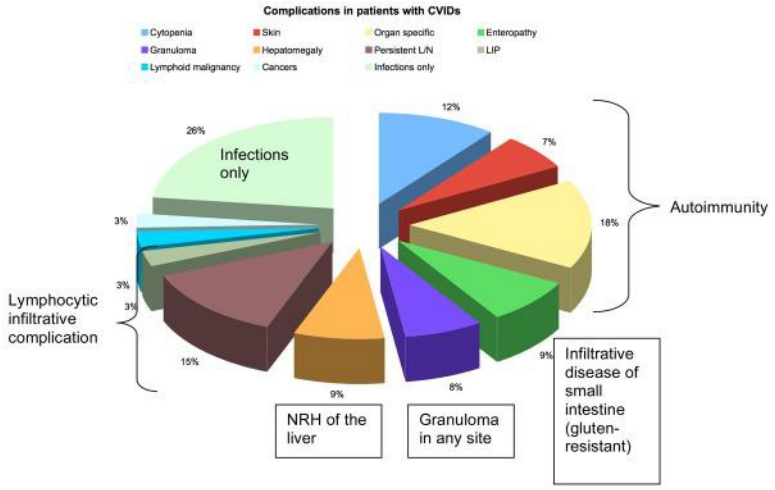


Figure 10: Types of complications in pacients with CVID and proportions of patients affected, from Chapel et al. (2009)

Altogether with those complications that CVID patients usually present, hypogammaglobulinemia and the subsequent risk of infections produce manifestations in a wide range of organs, which are the main causes of CVID morbidity, some of them depicted in figure figure 10.

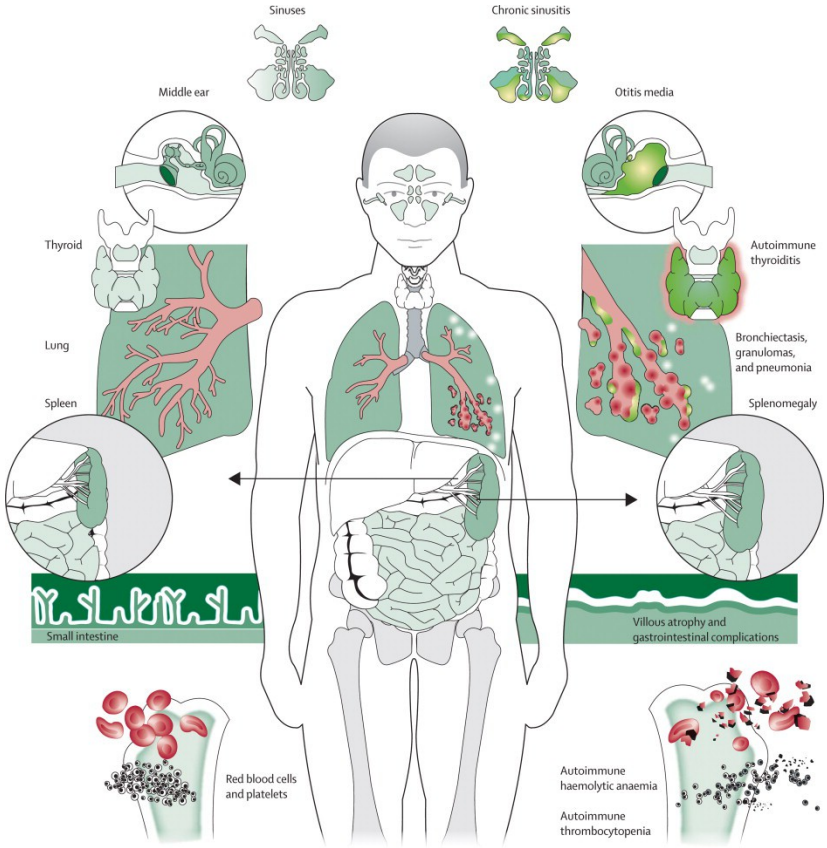


Figure 11: Representation of some healthy organs (left) in opposition to possible abnormalities found in organs from patients with CVID, from Park et al. (2008)

To the date (August 26th, 2016) there are listed 12 genes in the OMIM database implied in CVID (Table 5), but scientific literature can provide examples of many other genes that have been related with this disease or other diseases with similar phenotypes when defective.

Location	Phenotype	Phenotype MIM number	Gene/Locus	Gene/Locus MIM number
1q32.2	Immunodeficiency, common variable, 7	614699	CR2, C3DR, SLEB9, CVID7	120650
2q33.2	Immunodeficiency, common variable, 1	607594	ICOS, AILIM, CVID1	604558
4q24	Immunodeficiency, common variable, 12	616576	NFKB1, CVID12	164011
4q27	Immunodeficiency, common variable, 11	615767	IL21, CVID11	605384
4q31.3	Immunodeficiency, common variable, 8, with autoimmunity	614700	LRBA, LBA, CDC4L, CVID8	606453
7p12.2	Immunodeficiency, common variable, 13	616873	IKZF1, ZNFN1A1, IK1, LYF1, CVID13	603023
10q24.32	Immunodeficiency, common variable, 10	615577	NFKB2, LYT10, CVID10	164012
11p15.5	Immunodeficiency, common variable, 6	613496	CD81, TAPA1, CVID6	186845
11q12.2	Immunodeficiency, common variable, 5	613495	MS4A1, CD20, CVID5	112210
16p11.2	Immunodeficiency, common variable, 3	613493	CD19, CVID3	107265
17p11.2	Immunodeficiency, common variable, 2	240500	TNFRSF13B, TACI, CVID2	604907
22q13.2	Immunodeficiency, common variable, 4	613494	TNFRSF13C, BAFFR, CVID4	606269

Table 5: Genes related with CVID in the OMIM database

The wide phenotypical heterogeneity and the fact that diagnosis is made by exclusion makes that in some patients other hypogammaglobulinemia causes have to be ruled out, and in many cases, specially before the advent of NGS technologies, it was not considered by the clinicians due to the amount of time and effort that it requires. ESID lists several diseases that must be excluded before making the CVID diagnostic, which I have tried to relate with the genes implied (Table 6).

Disease	Genes
Ataxia-Telangiectasia	ATM
Severe combined immunodeficiency	ADA DCLRE1C IL2RG IL7R JAK3 MH2CTA PNP PTPRC RAG1 RAG2 RFX5 RFXANK RFXAP ZAP70
Hyper IgM syndrome with immunodeficiency	AICDA CD40 CD40LG UNG
Transcobalamin II deficiency and hypogammaglobulinemia	TCN2
X-linked agammaglobulinemia	BTK
Autosomal agammaglobulinemia	BLNK CD79A CD79B IGHM IGLL1 LRRC8 PIK3R1
X-linked lymphoproliferative syndrome	CD27 ITK SH2D1A XIAP

Table 6: Diseases provided by the ESID that must have to be excluded before CVID diagnosis and associated genes.

1.7.1 LoF mutations described as causal of CVID

The first description of the molecular defect inherent in a case of CVID was a homozygous large deletion encompassing two exons of the inducible T-cell co-stimulator (ICOS), found in 2003 in two families, with no suspected relationship between them(Grimbacher et al. 2003). The mutation was the same 1,815bp deletion in the two families and abolished the expression of ICOS in the surface of B-cells after stimulation by T-cells, producing a disease phenotype with low circulating B-cells count and hypogammaglubulinemia, but no other complications. The mutation segregated perfectly with the disease and it was likely to be originated through a homologous unequal recombination during meiotic recombination. Subsequent studies found more individuals affected in the same region and concluded that it must have its origin in a common founder(Salzer et al. 2004). This research opened the season to hunt genetic defects in CVID and since then many mutations have been related to the disease. Among them many LoF mutations have been described, mostly in the lipopolysaccharide-responsive, beige anchor protein, encoded by the LRBA gene(Lopez-Herrera et al. 2012), and in the tumor necrosis family factor superfamily, member 13B gene (TNFRSF13B)(Castigli et al. 2005; Salzer et al. 2005), which encodes the transmembrane activator and CAML interactor (TACI) protein. A few examples of types of highly deleterious LoF mutations in the OMIM CVID genes are listed in the table 7.

Gene	Types of LoF described	References
CD19	Frameshift indel	van Zelm *et al*., 2006; Vince *et al*., 2011
	Splice site	Kanegane *et al*., 2007
CR2	Splice site	Thiel *et al*., 2012
	Stop gain	Thiel *et al*., 2012
CD81	Splice site	van Zelm *et al*., 2010
TNFRSF13B	Stop gain	Salzer *et al*., 2005; Salzer *et al*., 2008
	Frameshift indel	Pan-Hammarstron *et al*., 2007; Salzer *et al*., 2008
	Splice site	Mohammadi *et al*., 2009
TNFRSF13C	Frameshift indel	Warnatz *et al*., 2009
LRBA	Stop gain	López-Herrera *et al*., 2012; Charbonnier *et al*., 2014; Lo *et al*., 2015
	Frameshift indel	Alangari *et al*., 2012; Charbonnier *et al*., 2014; Lo *et al*., 2015; Revel-Vilk *et al*., 2015; Seidel *et al*., 2015
	Splice site	Lo *et al*., 2015
	CNV	López-Herrera *et al*., 2012
NFKB2	Frameshift indel	Chen *et al*., 2013; Liu *et al*., 2014; Brue *et al*., 2014
NFKB1	Splice site	Fliegauf *et al*., 2015
	Frameshift indel	Fliegauf *et al*., 2015
MS4A1	Splice site	Kuijpers *et al*., 2010
ICOS	Frameshift indel	Takahashi *et al*., 2009; Robertson *et al*., 2015
	CNV	Grimbacher *et al*., 2003

Table 7: Types of LoF mutations found in OMIM CVID genes. To the date no LoF mutations have been found in the IL21 gene.

Since the first report of mutations in the LRBA gene causing CVID(Lopez-Herrera et al. 2012), many other laboratories have other variants in the same gene, in some cases with the typical phenotype of hypogammaglobulinemia and autoimmunity(Alangari et al. 2012; Charbonnier et al. 2015; Lo et al. 2015; Revel-Vilk et al. 2015; Seidel et al. 2015), and in few cases lacking the hypogammaglobulinemia(Burns et al. 2012; Serwas et al. 2015). LRBA seems to be a gene frequently defective in CVID, which has strong effects in the innate immune system, mainly through the

regulation of the expression of cytotoxic T-associated lymphocyte 4 gene (CTLA4). LRBA colocalizes with CTLA4 in endosomal vesicles, and its inactivation increases CTLA4 turnover, resulting in a reduction of the levels of CTLA4 proteins in the cells. Mutations in CTLA4 have been associated as well to immune dysregulation disorders with lymphoproliferation, but in the OMIM is not listed as forming part of the CVID genes, probably because B-cells of CTLA4-defective patients can produce normal levels of IgG and IgA *in vitro*. But *ex vivo* some subsets of their B-cells show an increased apoptosis that could explain the hypogammaglobulinemia found in some patients(Kuehn et al. 2014; Schubert et al. 2014). Nonetheless, the disease phenotype is very similar to CVID and therefore it should be taken into account as a possible disease causing gene, and it's haploinsufficiency makes it specially vulnerable to mutations.

In the other hand, the TNFRSF13B also has been in the focus on the genetics of the CVID, mostly to a non-synonymous variant, C104R, first thought to be causative and with a relative high-frequency in the population. Two independent and simultaneous studies found that individuals with genetic defects in TNFRSF13B had an expression of the gene comparable to controls, but were unable to induce IgG or IgA production, suggesting an impaired isotype switching(Castigli et al. 2005; Salzer et al. 2005). Further studies have found that it has incomplete penetrance since healthy individuals have been found with the variant(Salzer et al. 2009).

Nonetheless, this variant is related with CVID, although as a disease risk factor rather than a causative variant(Pan-Hammarström et al. 2007). The variant is found in 5% of the cases and has a strange behavior, protecting for autoimmunity in CVID when is found in homozygosity but increasing the risk of autoimmunity when is heterozygous(Salzer et al. 2009).

Many other genes have been related to CVID, but they haven't been so thoroughly researched in the CVID context as the mentioned above. Moreover, CVID doesn't even have an official list of genes which have to be screened. As it's diagnostic is mainly based on the phenotype and in the exclusion from other similar diseases, recent studies tend to focus in a more inclusive list of genes to compensate the possibility that some of the patients could be diagnosed with a more specific disease. For example, Maffuci et al. (2016) used an inclusive list of 269 PID genes in where they have screened mutations in 50 patients diagnosed with CVID using whole exome data, finding 17 probable mutations in 15 patients(Maffucci et al. 2016). Other studies have relied in the analysis of pathways and genes interacting with previously known CVID genes, which might be a good approach to identify novel genes in the disease(van Schouwenburg et al. 2015).

2.Objectives

In this thesis we aim to measure the load of LoF variants in the genomes and to understand their effect in human diseases, and particularly in the case of primary immunodeficiencies.

This thesis has two main objectives:

1) The determination of the load of LoF mutations in genomes of Great Apes. We want to provide an estimation of the number of LoF mutations segregating in the populations of these endangered species in order to compare it with the unexpected high number of LoF found in genomes of healthy humans, and to characterize the genes affected by those variants. We are going to use stringent filtering to pull apart errors produced in the sequencing and the mapping of the samples from deleterious variants which are more likely to have an effect in the organism.

2) To understand the impact of LoF variants in the genome, and to describe the molecular etiopathology behind the heterogeneity of the CVID. We intend to expand the knowledge of CVID disease mechanism through the analysis of of the genes that interact with known causative genes, as well as through the study of immunological pathways that could be related with the disease, relying in the comparison between patients and controls not diagnosed with CVID. In order to bypass the CVID heterogeneity, we will consider both the monogenic model and the oligogenic model in those pacients were we couldn't find a mendelian cause of the disease.

3.Results

de Valles-Ibáñez G, Hernandez-Rodriguez J, Prado-Martinez J, Luisi P, Marquès-Bonet T, Casals F. Genetic Load of Loss-of-Function Polymorphic Variants in Great Apes. Genome Biol Evol. 2016 Mar 26;8(3):871–7. DOI: 10.1093/gbe/evw040

Whole-exome sequencing of common variable immunodeficiency

Guillem de Valles-Ibáñez et al. In preparation.

Abstract

CVID is a frequent primary immunodeficiency characterized by hypogammaglobulinemias and poor response to vaccines with a diagnostic made by exclusion of other diseases that can cause similar phenotypes. We have analyzed whole-exome sequencing and copy number variants data of 36 patients diagnosed with early-onset CVID and 8 relatives. We have described causal genetic variants in LRBA, CTLA4, PIK3R1 and NFKB1, and the presence of two more compound heterozygotes in CR2 and PLCG2. Altogether it represents a monogenic origin for 13-19 % of the patients included in the study. Beyond the monogenic model for the disease, we have explored other models based in the presence of detrimental variants in a interacting proteins or the accumulation of functional variants in immunological pathways..

Introduction

Common variable immunodeficiency (CVID) is the most prevalent primary immunodeficiency with a prevalence from 1:10,000 to 1:50,000 in North America and Europe (Saikia and Gupta 2016). It's diagnosis criteria consists in low serum concentrations of IgG, IgA and/or IgM, recurrent bacterial infections and poor antibody response to vaccines, in addition to the exclusion of other known causes of hypogammaglobulinemia (Conley, Notarangelo, and Etzioni 1999; Chapel and Cunningham-Rundles 2009; Saikia and Gupta 2016). Phenotypes of the patients are highly heterogeneous due to different time onsets and to a high variety of related complications as autoimmune manifestations, lymphoproliferation, enteropathy and lymphoid malignancies, suggesting that CVID could be a common outcome from diverse autoimmune disregulations.

The clinical heterogeneity of CVID has hindered both the diagnostic and the identification of the underlying genetic defect of the disease, allowing a molecular characterization of the cause in less than 20% of the patients, and usually in familiar forms of the disease which constitute only a small fraction of the CVID cases (Bacchelli et al. 2007; Park et al. 2009; Rodríguez-Cortez et al. 2015; Saikia and Gupta 2016). Despite that, mutations in the genes CR2, LRBA, NFKB1, NFKB2, IL21, TNFRSF13B, TNFRSF13C, CD81, IKZF1, PRKCD, MS4A1 and CD19 are listed in the OMIM database as causative of the disease. In the literature specific

variants in those genes or in other not listed in the OMIM have been reported to confer susceptibility to the disease (NOD2, MSH5, TNFRSF13B, HLA) or to originate similar phenotypes to CVID (CTLA4, PLCG2, PIK3CD, PIK3R1), blurring even more the boundaries to define this disease. Furthermore, some of the mutations have an incomplete penetrance (Pan-Hammarström et al. 2007; Salzer et al. 2009) and many sporadic cases remain unexplained after deep genetic analyses, suggesting that in an important fraction of cases CVID does not follow a monogenic Mendelian model.

Recent studies using whole-exome sequencing to study CVID have reported that 15-30 % of patients have a monogenic origin(van Schouwenburg et al. 2015; Maffucci et al. 2016), with genetic variants both at candidate or new genes for CVID. In this work, we have analyzed the whole-exome sequence and copy number variants data for 36 CVID patients with early onset for the disease. Here, we estimate the proportion of patients with a monogenic origin, and propose a prioritized list of candidate genes for each case. For that, we consider the presence of rare genetic variants in an individual along with allele frequency, bioinformatic predictions of the phenotypical effect and evolutionary conservation rates for candidate gene prioritization. In addition, we expand the analysis to other oligo- or multigenic models for the disease, by considering the presence of mutations in interacting proteins or the accumulation of functional variants in immunological pathways (van Schouwenburg et al. 2015; Maffucci et al. 2016).

Materials and Methods

Individuals included in the study

This study includes 36 patients diagnosed with CVID, including sporadic and familiar cases, without any known immunodeficiency associated with Ig down-regulation, and completing the next criteria: from birth to 18 year-old at the age of diagnosis; lack of antibody production after immunization of antigen exposure in at least two assays; two years post-diagnosis to exclude lymphoid malignancy; IgG levels 2.5th centil for age and IgA or IgM low. CVID patients presenting one of these features have been excluded: Hyper IgM; Deficit in CD19+ or CD20+ B cell; ICOS or transmembrane activator and calcium-modulating cyclophilin ligand interactor (TACI) gene mutation already diagnosed; Complications as tumours associated, lymphomas or complications due to therapies (side effects of splenectomy, corticosteroid and immune suppressive therapies). In addition, parents and siblings have also been included in the study, when available. We used one set of controls with whole-exome sequences from 37 individuals from a Spanish cohort diagnosed with ASD (Codina-Solà et al. 2015).

Genetic analyses

DNA was extracted from blood samples. Genotyping was performed with the CytoScanHD array (Affymetrix) according to the manufacturer's protocol. The obtained cychp files were

analyzed with Chromosome Analysis Suite v.2.1.0.16 software and NetAffx na33 annotation version. For CNV detection and to prevent false positives, we considered alterations involving at least 25 markers and more than 150 Kb in length for gains, and 35 markers and more than 75 Kb for losses. For loss of heterozygosity (LOH) regions detection we considered alterations of at least 50 markers in more than 5 Mb. Exome capture was performed with the Agilent SureSelect XT enrichment system. DNA was sequenced in an Illumina HiSeq 2000 platform in a 2 × 75 paired-end cycles run. PCR duplicates were removed with Picard (http://www.picard.sourceforge.net). Sequence reads were mapped to the human reference genome (hg19) using GEM (Marco-Sola et al. 2012). Variant calling was performed using GATK(McKenna et al. 2010) and SNP annotation with SnpEff(Cingolani, Platts, et al. 2012) and SnpSift(Cingolani, Patel, et al. 2012). Candidate mutations were visually inspected with the Integrative Genomics Viewer (Thorvaldsdóttir, Robinson, and Mesirov 2013) and in some cases validated by Sanger sequencing.

Genetic data and statistical analyses

Only functional variants were considered, including non synonymous, stop gain and stop loss, mutations in splice donor or acceptor sites and frameshift insertions and deletions. In addition to standard filters for mapping and variant calling and annotation we also discarded indels clustering within 10 base pairs of another indel and for most of the analyses we excluded those variants present in 10 or more individuals from our study. We used allele frequencies

from The 1000 Genomes Project and the NHLBI and Exome Sequencing Project (ESP) to filter using allele frequencies ((Abecasis et al. 2010, http://evs.gs.washington.edu/EVS/). We used GERP (Cooper et al. 2005; Davydov et al. 2010) to asses for evolutionary conservation and Polyphen (Sunyaev et al. 1999) to predict the phenotypic impact of non-synonymous variants. We also have used predicted haploinsufficiency scores (Huang et al. 2010) to infer the possible model of the disease for the genes found in our analysis.

The PPI data was obtained from the Human Protein Reference Database (HPRD)(Keshava Prasad et al. 2009) considering the whole set of non-redundant interactions between two proteins and the genes in the pathways were extracted from the KEGG database(Kanehisa and Goto 2000; Kanehisa et al. 2015), considering those pathways that could be important for the disease and the pathways in the immune system category.

Functional validation

PBMC isolation: Peripheral blood mononuclear cells (PBMCs) from hUCB and healthy controls were isolated by Ficoll-Hipaque (Sigma-Aldrich, St. Louis, MO, USA) density gradient centrifugation of heparinized blood. Cells were cultured with complete medium [RPMI (Gibco, Grand Island, NY, USA) supplemented with 10% heat-inactivated fetal calf serum (FCS; Sigma-Aldrich, St. Louis, MO, USA), 1 μg/ml penicillin and 1 μg/ml streptomycin (Invitrogen, Grand Island, NY, USA)]. Viable

cells were counted using a hemocytometer in an inverted microscope.

Protein extraction and Western Blot: EBV immortalized cells were lysed with 1% NP-40 buffer. Protein concentration was normalized between control and patient. The products were analyzed by sodium dodecyl sulfate-polyacrylamide gel electrophoresis and western blotting. A nitrocellulose membrane was blocked with a 2% milk TBS, then incubated overnight with primary antibodies anti-LRBA (1:500, polyclonal, Abcam, United Kingdom) and anti-GAPDH (1:1000, polyclonal, Bio-Rad, United Kingdom) then the membrane was washed with TTBS and incubated for 1,5h with Goat Anti-Rabbit IgG H&L (HRP) (1:5000, Abcam). It was then developed with SuperSignal™ West Pico Chemiluminescent Substrate (Thermo Scientific, Waltham, MA, USA) following manufacturer's instructions.

Results

Sequencing and number of variants

We generated whole exome sequencing data for the 36 patients included in the study, as well as for 8 relatives, with an average coverage of 120X. Additionally, we also generated CNV data for all the samples except in one case where DNA was not available. Table S1 (Supplementary Information) shows the number of functional variants described in each sample, classified in different annotation categories: non-synonymous (or missense), stop-gain (or nonsense), start gain, splice site, and inframe and frameshift indels. Table S1 also contains the number of structural variants and loss of heterozigosity (LOH) regions detected in the genotypying analysis.

Known CVID mutations

The OMIM database includes known variants originating CVID in 13 genes: ICOS, TNFRSF13B, TNFRSF13C, CD19, CR2, MS4A1, CD81, IL21, LRBA, NFKB1, NFKB2, PRKCD and IKZF1. There is also evidence that defects in other genes (CTLA4, PLCG2) can cause a similar phenotype or modify the severity of the disease with co-morbidities (MSH5). These genes are thought to cause a CVID-like phenotype when are defective and are related mainly with T-cell and B-cell defects leading to a deficiency in antibody production. In total, we have found 96 nucleotide variants and 6 CNVs in the literature (Table S2) putatively related to CVID.

Table 1 shows the four CVID genetic variants that have been found in this study. Two of the reported variants are included in the TNFRSF13B gene, also called TACI, which is known to harbor functional mutations in 5-10% of the patients diagnosed with CVID (Martinez-Gallo et al. 2013). However, the existence of healthy controls with heterozygous mutations in this gene and the lack of a clear Mendelian pattern of inheritance in families has made some of the mutations at TNFRSF13B to be considered as risk factors (Pan-Hammarström et al. 2007; Salzer et al. 2009), which could be determinant only in case of homozygous patients (Salzer et al. 2005). Thus, TNFRSF13B would be considered a modifier gene rather than a casual gene in monogenic patients (Bogaert et al. 2016). The p.C104R variant is the most common TNFRSF13B functional mutation found in CVID patients (Bogaert et al. 2016). Three of the patients in this study present this mutation, in one case in homozygous state, being the second case found to the date (Koopmans et al. 2013). This mutation is significantly more frequent in CVID patients compared to controls ($P = 0.003$, Fisher's exact test). In the same gene we report nine samples with the protein change P251, although in this case the proportion is not significantly higher than in controls. In addition, a causal role for this variant can be discarded because of its high frequency in the population (14 % in the ExAc database, 11 % for the European population). On the other hand the p.P21R of the TNFRSF13C gene found in four patients, and in one case also in the healthy parent, shows a higher frequency compared to controls ($P = 0.003$, Fisher's

exact test). However, this variant (rs77874543) has also been found in non-CVID exomes in homozygosity, and has a population frequency higher to 5 %. Finally, we also detected two patients with the p.L85F change in the MSH5 gene (Sekine et al. 2007). The change was also present in the mother of one these patients, not diagnosed with CVID but with some of the clinical features described in the patient. Nonetheless this genetic variant has been found at lower frequencies in CVID patients compared to controls, and has a population frequency of 2 % or higher in some populations (7 % in Africans), which suggests that it has not a determinant role in CVID.

Molecular assessment for diagnosis of patients

Since CVID phenotypes are heterogeneous and in some cases similar to other syndromes, we also screened for variants at genes related to other hypogammaglobulinemias. We followed the diagnostic criteria of primary immunodeficiencies established by the European Society for Immunodeficiencies (ESID). We performed a search in the OMIM database to get a list of 32 genes related to other hypogammaglobulinemias. Table S3 shows the presence of rare variants (GMAF < 1%) in 23 genes for 21 of the patients. All the variants are present in heterozygosis and are non-synonmous, except a splicing variant at PIK3R1. This splicing variant has been reported to originate immunodeficiency thanks to its dominant gain of function effect on PI3K signaling (Deau et al. 2014) in agreement with its haploinsufficiency prediction value of

0.89 (Huang et al. 2010). Interestingly, two affected sisters harbor a missense variant in the same gene. This variant has not been previously reported and is located in a conserved nucleotide according to its GERP value, although it is not predicted to be damaging using SIFT and Polyphen.

For the rest of the variants, the fact that no compound heterozygote is detected (except for N216, but both ATM variants are also present in his healthy father N215), and the low values for haploinsufficiency prediction of less than 0.21 for all genes except for ITK (0.542) suggest that there is no direct relation between them and the disease in these patients. Polyphen and SIFT indexes are also in general low.

Loss-of-Function variants

LoF variants include stop gain and loss mutations, splice-site mutations and frameshift-indels, which are predicted to disrupt the protein and therefore likely to affect the phenotype and be related to the disease. In fact LoF variants only represent about 1% of the functional variants in a genome (Bamshad et al. 2011), but they account for approximately 20 % of the coding SNPs associated to disease (Mort et al. 2008). Whole-genome sequence analyses have showed that on average any human genome harbors about 100 LoF genetic variants, with 20 of them in homozygosity(MacArthur et al. 2012). However, the mere presence of a LoF variant, even in homozygosity, does not automatically indicate the phenotypic manifestation of a disease. Different levels of gene essentiality and

functional redundancy have been invoked to explain this seemingly excessive number of inactivating variants. Indeed, recent publications refers to these variants as protein-truncating variants (PTVs) and not LoF, since often their molecular and phenotypic effect (a supposed loss of function) has not been demonstrated(Rivas et al. 2015). Thus, in absence of functional evidence, filters based on allele frequency or evolutionary conservation (de Valles-Ibáñez et al. 2016) can be used to estimate the actual number of highly detrimental mutations in each individual and identifying candidate mutations and genes for a given disease. Gene features as tolerance to functional variants (Petrovski et al. 2013), expression levels or connectivity patterns (Huang et al. 2010) can also be used to prioritize candidate genes.

Table S1 shows the number of LoF variants identified in each individual of the study. The number of LoF variants ranges from 78 to 153, similar to what has been previously described(Bamshad et al. 2011; Mort et al. 2008; MacArthur et al. 2012). Applying different frequency thresholds substantially reduces the number of LoF variants per individual (Table 2). We established a permissive allele frequency threshold of 1 %, and first focused the analysis in the LoF variants described in candidate genes for CVID. To do that we have created a list of 97 candidate genes for CVID (Table S4), including genes in the OMIM database (http://omim.org), genes defined in Bogaert et al. (Bogaert et al. 2016), and others taken from the literature. Second, we also analyzed the presence of LoF variants in proteins interacting with

the proteins encoded by candidate genes. Finally, we also considered these LoF at other genes with very low frequency threshold (0.001). Table 2 and Table S5 show the number and list of LoF variants found in these three genes groups. A description including population frequencies and evolutionary conservation values for each variant from the three groups can be seen in Table S5 (Supplementary material).

Seven patients harbor a LoF variant at a frequency less than 1 % in CVID candidate genes (Table 3). Among them, L283 present a homozygous nonsense variant at the exon 4 of the LRBA gene (chr4:151392836G>A (hg19)). The stop codon change at LRBA (R2214*) is introduced at the beginning of the BEACH domain (IPR000409 in InterPro), a highly conserved domain with known crystal structure but unknown function(Gebauer et al. 2004). This mutation was validated by Sanger sequencing in the patient, and also detected in heterozygosis in both parents and three healthy siblings (Figure 1). Copy number and SNP analysis confirmed the existence of consanguinity in this patient. We estimated a consanguinity index of 0.058 compatible with descendants from third degree kinship marriages, based in the total of 174 Mb included in LOH regions (Sund et al. 2013) with ten LOH regions of more than 5 Mb. Functional studies were performed to comfirm the causalt role of this mutation (see below). In two more patients we have found a frameshift mutation in the gene NOD2, considered as a modifier of the disease.

For the remaining six patients presenting a low frequency heterozygous LoF variant in a CVID candidate gene (Table 3), one is located at the CTLA4 gene and three in NFKB1, two genes that have been reported to harbor heterozygous mutations originating CVID (Schubert et al. 2014; Kuehn et al. 2014; Fliegauf et al. 2015). In the case of the CTLA4 mutation, it consists of a frameshift deletion not described before in the databases. We performed Sanger sequencing of this mutation and detected that is a *de novo* mutation not present in the parents (Figure S1), and therefore a strong candidate to originate CVID. Regarding NFKB1, two patients share a start loss variant affecting one of the transcripts, although its frequency of 0.002 makes it unlikely to have a causal (monogenic) role in the disease. In contrast, a new splice-site mutation in NFKB1 is described in N234, being a good candidate to originate the disease. In addition, patient N227 presents a 13 MB heterozygous deletion (chr4: 94,135,868-107,295,574) not present in parents which includes NFKB1 gene among others. Finally, although the variants described at NOD2 and IL10RA are not present in any database, no CVID cases with heterozygous variants at these genes have been described, in agreement with their low haploinsufficieny values (0.119 and 0.173, respectively).

New genetic variants at candidate genes for CVID

We next explored the presence of functional variants, other than the candidate genetic variants described above, at candidate genes for CVID. In this case, we established a filter frequency with

a threshold of 1%. The final number of variants in each gene and individual is shown in Table S6, differentiating variants in candidate genes, variants in interacting proteins and in other genes (in the latter case with higher frequency thresholds). After selecting the variants with a GERP conservation score higher than 2 (Davydov et al. 2010), a Polyphen score higher than 0.5 and a frequency in the ExAC and GMAF databases below 1/1000, we have found functional variants in four patients that can explain the disease considering both the truncating potential of the variants and the predicted haploinsufficiency of the genes. We have found additional functional variants in some of our candidate genes, listed in Table 4. However, the variants shared between patients and healthy relatives cannot explain the disease *per se* without assuming a partial penetrance or a more complex model for the disease.

Compound heterozygotes

In absence of consanguinity, the standard WES sequencing approach to rare disease usually consist in including a few patients with the same syndrome, generating a list of compound heterozygote loci for each individual, and identifying the causal gene by the comparing these candidate gene list across patients (Ng et al. 2010). However, this approach can be of limited utility for diseases such as CVID, where the phenotypical heterogeneity of patients, as well as the already reported cases of several CVID causing genes, suggests diverse origins of the disease. Thus, once we produced a list of genes harboring compound heterozygotes in

each patient, we appied two different allele frequency thresholds of 0.01 and 0.001, in order to shorten the list of candidate genes for each patient. Table 5 shows the number of compound heterozygotes per patient, and gene names are reported in Table S8. The number of genes per patient can be reduced using additional filters based in evolutionary conservation or predicted phenotypic effect. We established a threshold of a GERP > 2 for the functional variants, since positions with values greater than 2 are considered to be conserved among mammals and therefore more to prone to be of functional importance (Davydov et al. 2010). On the functional effect, we used the Polyphen prediction (Adzhubei et al. 2010). Table S9 shows the number of times that a gene is found as a compound heterozygote in a patient, with GERP values higher than 2 or Polyphen predictions as possibly damaging. After applying those filters, only four genes are shared between more than one patient. Among them, gene SLC25A5 is found in 9 patients, but is a gene frequently found in NGS data(Fuentes Fajardo et al. 2012).

Interestingly, two candidate genes (CR2 and PLCG2) are also found as compound heterozygotes in patients N233 and N212, respectively. For the gene PLCG2 one of the variants is predicted to be damaging (Table 6). In contrast, both CR2 genetic variants show low Polyphen values, being therefore a less promising candidate to originate CVID. Finally, none of the remaining genes after filtering by conservation and predicted damage is included in the CVID candidate genes list.

Oligogenic disease

For the 31 patients without a clear candidate gene for a monogenic origin of the disease, we next considered an oligogenic model of inheritance. Features as variable penetrance and the phenotypical variability inside of families may suggest an oligogenic origin, where the disease is caused or modulated by a few genes (Robinson and Katsanis 2010). The prevalence of CVID would also fit to a model where the disease is produced by mutations in two or a few genes, between the very rare disorders originated by a single locus and common disease produced by the interaction of many genes (Gilissen et al. 2011). Among the oligogenic models, digenic disease is the simplest one. DIDA, a database of digenic diseases, included 44 diseases with 213 digenic combinations collected from the literature until June 2005 (Gazzo et al. 2015). This form of disease often refers to both situations with a primary locus or cases where two loci contribute to the disease with roughly the same importance (Schäffer 2013). Modifier genes, affecting the severity of the disease, can also be considered a type of digenic inherintace (Génin, Feingold, and Clerget-Darpoux 2008).

We have focused our study in patients with a previously known mutation in a CVID gene. The case of TNFRSF13B is especially important since its incomplete penetrance and its variants previously related to CVID with high frequency in the population may suggest a digenic model. We have searched in our patients with known variants in TNFRSF13B and in genes that interact with it.

Among them, the patient L297, besides having a rare homozygous non-synonymous variant (introducing the protein change C104R, commonly found in CVID patients, but not in homozygosis) in the gene TNFRSF13B, has another heterozygous non-synonymous mutation in a gene that interacts directly with TNFRSF13B (TNFRSF13C). This second mutation is more frequent than the first one, and is not predicted to be damaging with Polyphen (Table 7), but together with the C104R mutation in homozygosys has potential to be a candidate for the digenic model.

Accumulation of functional variants in immunological pathways

We have also tested if there is an accumulation of variants at the following KEGG pathways related to the immune function (Kanehisa and Goto 2000; Kanehisa et al. 2015): cytokine-cytokine receptor interaction, B-cell signaling, JAK-STAT signaling, NFKB signaling, TNF signaling, ras signaling, mismatch repair, PIK3-AKT signaling, primary immunodeficiency, T-cell signaling, mTOR signaling, and all the pathways related with the immune system (Hematopoietic cell lineage, antigen processing and presentation, chemokyne signaling, complement and coagulation cascades, cytosolic-DNA sensing, Fc epsilon RI signaling, Fc gamma R-mediated phagocytosis, intestinal immune network for IgA production, leukocyte transendothelial migration, natural killer cell mediated cytotoxicity, platelet activation RIG-I-like receptor signaling, NOD-like receptor signaling and toll-like receptor signaling). We first assessed a possible excess of functional variants

in these twenty-five pathways by comparing our patients to a set of controls (see Materials and Methods). To correct for differences in sequencing coverage between patients and controls, we estimated the ratios of functional to synonymous variants in each sample, and used a frequency threshold of 1 % . We corrected those values with the number of genes found in each pathway, since pathways with a higher number of genes could be overrepresented. We detected an excess of variants in four of the pathways in our samples respect the controls: NFKB signaling, T cell signaling, mTOR signaling, intestinal immune network for IgA production and chemokyne signaling, while remaining pathways are more similar (Figure 2). Nonetheless, many patients have ratios significantly higher than the mean in some of the pathways, reinforcing the heterogeneity of the CVID and suggesting a more personalized approach involving the relationship between the specific clinical history of each patient and the possible phenotypic outcomes of each pathway. For example, patient N233 has a higher ratio in the JAK-STAT signaling pathway and patient N295 in the TNF signaling pathway.

Functional validation

The homozygous stop gain mutation found in the LRBA gene of the patient L283 is a good candidate to be a disease-causative truncating mutation. We have tested if the variant causes the gene to undergo non-sense mediated decay or it produces an aberrant protein. The western blot gel electrophoresis separation (Figure 3) shows that the cells of the patient don’t produce any

detectable amounts of LRBA protein, thus validating the deleterious effect of the mutation.

Discussion

After its presentation as a valuable tool for identifying causal genes for rare disorders (Ng et al. 2010), whole-exome sequencing followed by the identification of compound heterozygotes for rare functional variants has become a standard approach in human genetic rare disease studies, except in case of consanguinity where homozygous variants inherited from both parents are the most plausible explanation. In this situations, exome sequencing can be assisted by homozygosity mapping, which increases the power to identify the causal variant even for studies with a single patient (Bolze et al. 2010; Walsh et al. 2010). In case of populations having experienced a demographic bottleneck with the consequent genetic diversity reduction, homozygous rare variants can play also an important role even in absence of consanguinity (Samuels et al. 2013). However, for diseases with higher frequencies, phenotypic variability or incomplete penetrancies as CVID the whole-exome sequencing approach is less promising, since by now only a reduced proportion of the cases have been attributed to a monogenic origin with Mendelian inheritance. We have first applied approaches to detect single genes at the origin of rare diseases, aiming to estimate the proportion of monogenic cases in CVID, and then different strategies to detect cases with an oligo or multigenic origin.

We have described five LoF variants (including one stop-gain and two splice-site nucleotide variants, one frameshift indel

and one large deletion) in the genes LRBA, CTLA4, PIK3R1 and NFKB1 that are most likely causative. We have performed functional tests validating the new stop gain mutation described at LRBA. Thus, a minimum of 14 % of the patients included in this study would have a monogenic origin for CVID. In addition, we provide different levels of evidences for the remaining 31 patients. For four more cases we propose a possible monogenic origin produced by a dominant mutation in PIK3R1 (two related patients) or compound heterozygotes at CR2 and PLCG2. Therefore, at most the proportion of monogenic cases detected would increase to 28 % of the patients.

In addition, we have also found one patient harboring a variant in the PIK3R1 gene that excludes the CVID diagnostic, and two other patients with a non-synonymous variant in the same gene that could explain the phenotype. This finding highlights the potentiality of genetic analysis to assess the diagnostic in heterogeneous diseases as CVID. We have also described other non-synonymous variants in genes related to diseases producing hypogammaglobulinemia and that therefore need to be ruled out in the diagnostic of CVID. Although we are not able to link them to the patients phenotype, future studies could provide the evidence necessary for some of those variants.

For the patients without a clear candidate gene for a monogenic model, we present a list of LoF variants and compound heterozygotes, prioritized according to different mutation and gene properties. Although the casual gene has been probably identified

for some of the patients in these lists, the phenotypic diversity of CVID patients hinders the identification of new genes because of overlapping between patients, and would therefore ultimately rely in functional analyses.

Finally, we have performed different approaches to detect CVID cases originated by genetic variants in two or several genes. The analyses including PPI and immunological pathways have expanded the group of possible causal genes. We propose that the combined action of genetic variants could disturb the fine tuning of the biological networks needed to attain a correct functionality. We have identified an excess of LoF and other functional variants in four immunological pathways, as well as particular individuals with an excess of mutations at a given pathways or at interacting genes putatively related to CVID. In particular the NFKB, T cell and mTOR signaling pathways seem to play an important role in CVID, from the comparison of patients in this study to healthy controls.

After the success of new sequencing technologies and in particular of whole-exome sequencing in unraveling the molecular mechanisms of many rare syndromes, rare diseases as CVID that do not completely fit to a Mendelian model represent a new challenge for medical genomics. In this manuscript we have proposed different approaches to the analysis of CVID from whole-exome sequencing data, and showed its power and limitations as a diagnostic tool for the study of these diseases. Beyond the identification of the casual gene in some patients, this kind of studies can also be of help to detect key pathways related to the

development of the disease, thus contributing to a better understanding of its etiology.

Acknowledgements

The authors thank funding to F.C. by grant SAF2012-35025 from the Ministerio de Economía y Competitividad (Spain) and FEDER and by Direcció General de Recerca, Generalitat de Catalunya (2014SGR-866) and the grant BES-2012-051794 .

This study makes use of data generated by the Medical Genome Project. A full list of the investigators who contributed to the generation of the data is available from http://www.medicalgenomeproject.com/en. Funding for the project was provided by the Spanish Ministry of Economy and Competitiveness, projects I+D+i 2008, Subprograma de actuaciones Científicas y Tecnológicas en Parques Científicos y Tecnológicos (ACTEPARQ 2009) and ERFD.

Bibliography

Abecasis, Gonçalo R, David Altshuler, Adam Auton, Lisa D Brooks, Richard M Durbin, Richard a Gibbs, Matt E Hurles, and Gil a McVean. 2010. "A Map of Human Genome Variation from Population-Scale Sequencing." *Nature* 467 (7319): 1061–73. doi:10.1038/nature09534.

Adzhubei, I A, S Schmidt, L Peshkin, V E Ramensky, A Gerasimova, and P Bork. 2010. "A Method and Server for Predicting Damaging Missense Mutations." JOUR. *Nat Methods* 7. doi:10.1038/nmeth0410-248.

Bacchelli, C, S Buckridge, a J Thrasher, and H B Gaspar. 2007. "Translational Mini-Review Series on Immunodeficiency: Molecular Defects in Common Variable Immunodeficiency." *Clinical and Experimental Immunology* 149 (3): 401–9. doi:10.1111/j.1365-2249.2007.03461.x.

Bamshad, Michael J, Sarah B Ng, Abigail W Bigham, Holly K Tabor, Mary J Emond, Deborah a Nickerson, and Jay Shendure. 2011. "Exome Sequencing as a Tool for Mendelian Disease Gene Discovery." *Nature Reviews. Genetics* 12 (11). Nature Publishing Group: 745–55. doi:10.1038/nrg3031.

Bogaert, Delfien J A, Melissa Dullaers, Bart N Lambrecht, Karim Y Vermaelen, Elfride De Baere, and Filomeen Haerynck. 2016. "Genes Associated with Common Variable Immunodeficiency: One Diagnosis to Rule Them All?" *Journal of Medical Genetics*, no. June: jmedgenet-2015-103690. doi:10.1136/jmedgenet-2015-103690.

Bolze, Alexandre, Minji Byun, David McDonald, Neil V Morgan, Avinash Abhyankar, Lakshmanane Premkumar, Anne Puel, et al. 2010. "Whole-Exome-Sequencing-Based Discovery of

Human FADD Deficiency." JOUR. *The American Journal of Human Genetics* 87 (6): 873–81. doi:http://dx.doi.org/10.1016/j.ajhg.2010.10.028.

Chapel, Helen, and Charlotte Cunningham-Rundles. 2009. "Update in Understanding Common Variable Immunodeficiency Disorders (CVIDs) and the Management of Patients with These Conditions." *British Journal of Haematology* 145 (6): 709–27. doi:10.1111/j.1365-2141.2009.07669.x.

Cingolani, Pablo, Viral M. Patel, Melissa Coon, Tung Nguyen, Susan J. Land, Douglas M. Ruden, and Xiangyi Lu. 2012. "Using Drosophila Melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift." *Frontiers in Genetics* 3 (MAR). doi:10.3389/fgene.2012.00035.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain W 1118; Iso-2; Iso-3." *Fly* 6 (2): 80–92. doi:10.4161/fly.19695.

Codina-Solà, Marta, Benjamín Rodríguez-Santiago, Aïda Homs, Javier Santoyo, Maria Rigau, Gemma Aznar-Laín, Miguel del Campo, et al. 2015. "Integrated Analysis of Whole-Exome Sequencing and Transcriptome Profiling in Males with Autism Spectrum Disorders." JOUR. *Molecular Autism* 6 (1): 1–16. doi:10.1186/s13229-015-0017-0.

Conley, Mary Ellen, Luigi D Notarangelo, and Amos Etzioni. 1999. "Diagnostic Criteria for Primary Immunodeficiencies." JOUR. *Clinical Immunology* 93 (3): 190–97. doi:http://dx.doi.org/10.1006/clim.1999.4799.

Cooper, Gregory M., Eric A. Stone, George Asimenos, Eric D. Green, Serafim Batzoglou, and Arend Sidow. 2005. "Distribution and Intensity of Constraint in Mammalian Genomic Sequence." *Genome Research* 15 (7): 901–13. doi:10.1101/gr.3577405.

Davydov, Eugene V., David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. 2010. "Identifying a High Fraction of the Human Genome to Be under Selective Constraint Using GERP++." *PLoS Computational Biology* 6. doi:10.1371/journal.pcbi.1001025.

Deau, Marie-Cèline, Lucie Heurtier, Pierre Frange, Felipe Suarez, Christine Bole-Feysot, Patrick Nitschke, Marina Cavazzana, et al. 2014. "A Human Immunodeficiency Caused by Mutations in the PIK3R1 Gene." JOUR. *The Journal of Clinical Investigation* 124 (9). The American Society for Clinical Investigation: 3923–28. doi:10.1172/JCI75746.

de Valles-Ibáñez, Guillem, Jessica Hernandez-Rodriguez, Javier Prado-Martinez, Pierre Luisi, Tomàs Marquès-Bonet, and Ferran Casals. 2016. "Genetic Load of Loss-of-Function Polymorphic Variants in Great Apes." *Genome Biology and Evolution* 8 (3): evw040. doi:10.1093/gbe/evw040.

Fliegauf, Manfred, Vanessa L. Bryant, Natalie Frede, Charlotte Slade, See Tarn Woon, Klaus Lehnert, Sandra Winzer, et al. 2015. "Haploinsufficiency of the NF-κB1 Subunit p50 in Common Variable Immunodeficiency." *American Journal of Human Genetics* 97 (3): 389–403. doi:10.1016/j.ajhg.2015.07.008.

Fuentes Fajardo, Karin V., David Adams, Christopher E. Mason, Murat Sincan, Cynthia Tifft, Camilo Toro, Cornelius F. Boerkoel, William Gahl, and Thomas Markello. 2012.

"Detecting False-Positive Signals in Exome Sequencing." *Human Mutation* 33: 609–13. doi:10.1002/humu.22033.

Gazzo, Andrea M, Dorien Daneels, Elisa Cilia, Maryse Bonduelle, Marc Abramowicz, Sonia Van Dooren, Guillaume Smits, and Tom Lenaerts. 2015. "DIDA: A Curated and Annotated Digenic Diseases Database." JOUR. *Nucleic Acids Research* , October. doi:10.1093/nar/gkv1068 .

Gebauer, Damara, Jiang Li, Gerwald Jogl, Yang Shen, David G Myszka, and Liang Tong. 2004. "Articles Crystal Structure of the PH - BEACH Domains of Human LRBA/BGL †" 43 (47).

Génin, Emmanuelle, Josué Feingold, and Françoise Clerget-Darpoux. 2008. "Identifying Modifier Genes of Monogenic Disease: Strategies and Difficulties." JOUR. *Human Genetics* 124 (4): 357–68. doi:10.1007/s00439-008-0560-2.

Gilissen, Christian, Alexander Hoischen, Han G Brunner, and Joris a Veltman. 2011. "Unlocking Mendelian Disease Using Exome Sequencing." *Genome Biology* 12 (9): 228. doi:10.1186/gb-2011-12-9-228.

Huang, Ni, Insuk Lee, Edward M. Marcotte, and Matthew E. Hurles. 2010. "Characterising and Predicting Haploinsufficiency in the Human Genome." *PLoS Genetics* 6 (10): 1–11. doi:10.1371/journal.pgen.1001154.

Kanehisa, Minoru, and Susumu Goto. 2000. "KEGG: Kyoto Encyclopedia of Genes and Genomes." JOUR. *Nucleic Acids Research* 28 (1). Oxford, UK: Oxford University Press: 27–30. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC102409/.

Kanehisa, Minoru, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2015. "KEGG as a Reference

Resource for Gene and Protein Annotation." JOUR. *Nucleic Acids Research* , October. doi:10.1093/nar/gkv1070 .

Keshava Prasad, T S, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, et al. 2009. "Human Protein Reference Database—2009 Update." JOUR. *Nucleic Acids Research* 37 (suppl 1): D767–72. doi:10.1093/nar/gkn892 .

Koopmans, Wikke, See Tarn Woon, Anna E S Brooks, P. Rod Dunbar, Peter Browett, and Rohan Ameratunga. 2013. "Clinical Variability of Family Members with the C104R Mutation in Transmembrane Activator and Calcium Modulator and Cyclophilin Ligand Interactor (TACI)." *Journal of Clinical Immunology* 33 (1): 68–73. doi:10.1007/s10875-012-9793-x.

Kuehn, Hye Sun, Weiming Ouyang, Bernice Lo, Elissa K Deenick, Julie E Niemela, Danielle T Avery, Jean-Nicolas Schickel, et al. 2014. "Immune Dysregulation in Human Subjects with Heterozygous Germline Mutations in CTLA4." *Science (New York, N.Y.)* 345 (6204): 1623–27. doi:10.1126/science.1255904.

MacArthur, Daniel G, Suganthi Balasubramanian, Adam Frankish, Ni Huang, James Morris, Klaudia Walter, Luke Jostins, et al. 2012. "A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes." *Science (New York, N.Y.)* 335 (6070): 823–28. doi:10.1126/science.1215040.

Maffucci, Patrick, Charles A Filion, Bertrand Boisson, Yuval Itan, Lei Shang, Jean-laurent Casanova, and Charlotte Cunningham-rundles. 2016. "Genetic Diagnosis Using Whole Exome Sequencing in Common Variable Immunodeficiency." *Frontiers in Immunology*. doi:10.3389/fimmu.2016.00220.

Marco-Sola, Santiago, Michael Sammeth, Roderic Guigo, and Paolo Ribeca. 2012. “The GEM Mapper: Fast, Accurate and Versatile Alignment by Filtration.” JOUR. *Nat Meth* 9 (12). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 1185–88. http://dx.doi.org/10.1038/nmeth.2221.

Martinez-Gallo, Monica, Lin Radigan, María Belén Almejún, Natalia Martínez-Pomar, Núria Matamoros, and Charlotte Cunningham-Rundles. 2013. “TACI Mutations and Impaired B-Cell Function in Subjects with CVID and Healthy Heterozygotes.” *Journal of Allergy and Clinical Immunology* 131 (2): 468–76. doi:10.1016/j.jaci.2012.10.029.

McKenna, A H M, E Banks, A Sivachenko, K Cibulskis, A Kernytsky, and K Garimella. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” JOUR. *Genome Res* 20. doi:10.1101/gr.107524.110.

Mort, Matthew, Dobril Ivanov, David N. Cooper, and Nadia A. Chuzhanova. 2008. “A Meta-Analysis of Nonsense Mutations Causing Human Genetic Disease.” *Human Mutation* 29 (8): 1037–47. doi:10.1002/humu.20763.

Ng, Sarah B, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, et al. 2010. “Exome Sequencing Identifies the Cause of a Mendelian Disorder.” JOUR. *Nat Genet* 42 (1). Nature Publishing Group: 30–35. http://dx.doi.org/10.1038/ng.499.

Pan-Hammarström, Qiang, Emanuela Castigli, Stephen Wilson, Lilit Garibyan, Rima Rachid, Francisco Bonilla, Lynda Schneider, Massimo Morra, John Curran, and Raif Geha. 2007. “Reexamining the Role of TACI Coding Variants in Common

Variable Immunodeficiency and Selective IgA Deficiency." *Nature Genetics* 39 (4): 430–31.

Park, Miguel A, James T Li, John B Hagan, Daniel E Maddox, and Roshini S Abraham. 2009. "Common Variable Immunodeficiency: A New Look at an Old Disease." JOUR. *The Lancet* 372 (August): 489–502. doi:http://dx.doi.org/10.1016/S0140-6736(08)61199-X.

Petrovski, Slavé, Quanli Wang, Erin L Heinzen, Andrew S Allen, and David B Goldstein. 2013. "Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes." *PLoS Genetics* 9 (8): e1003709. doi:10.1371/journal.pgen.1003709.

Rivas, Manuel A, Matti Pirinen, Donald F Conrad, Monkol Lek, Emily K Tsang, Konrad J Karczewski, Julian B Maller, et al. 2015. "Effect of Predicted Protein-Truncating Genetic Variants on the Human Transcriptome." JOUR. Edited by Ayellet V Young Segre Taylor R. Gelfand, Ellen T. Trowbridge, Casandra A. Ward, Lucas D. Kheradpour, Pouya Iriarte, Benjamin Meng, Yan Palmer, Cameron D. Esko, Tonu Winckler, Wendy Hirschhorn, Joel Kellis, Manolis Getz, Gad Shablin, Andrey A. Li, Gen Zhou, Yi-Hui Nobel,. *Science* 348 (6235): 666 LP-669. http://science.sciencemag.org/content/348/6235/666.abstract.

Robinson, Jon F, and Nicholas Katsanis. 2010. "Oligogenic Disease BT - Vogel and Motulsky's Human Genetics." CHAP. In , edited by Michael R Speicher, Arno G Motulsky, and Stylianos E Antonarakis, 243–62. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-37654-5_8.

Rodríguez-Cortez, Virginia C., Lucia del Pino-Molina, Javier Rodríguez-Ubreva, Laura Ciudad, David Gómez-Cabrero, Carlos Company, José M. Urquiza, et al. 2015. "Monozygotic

Twins Discordant for Common Variable Immunodeficiency Reveal Impaired DNA Demethylation during Naïve-to-Memory B-Cell Transition." *Nature Communications* 6: 7335. doi:10.1038/ncomms8335.

Saikia, Biman, and Sudhir Gupta. 2016. "Common Variable Immunodeficiency." JOUR. *The Indian Journal of Pediatrics* 83 (4): 338–44. doi:10.1007/s12098-016-2038-x.

Salzer, Ulrich, Chiara Bacchelli, Sylvie Buckridge, Qiang Pan-Hammarström, Stephanie Jennings, Vassilis Lougaris, Astrid Bergbreiter, et al. 2009. "Relevance of Biallelic versus Monoallelic TNFRSF13B Mutations in Distinguishing Disease-Causing from Risk-Increasing TNFRSF13B Variants in Antibody Deficiency Syndromes." *Blood* 113 (9): 1967–76. doi:10.1182/blood-2008-02-141937.

Salzer, Ulrich, H M Chapel, a D B Webster, Q Pan-Hammarström, A Schmitt-Graeff, M Schlesier, H H Peter, et al. 2005. "Mutations in TNFRSF13B Encoding TACI Are Associated with Common Variable Immunodeficiency in Humans." *Nature Genetics* 37 (8): 820–28. doi:10.1038/ng1600.

Samuels, Mark E, Jacek Majewski, Najmeh Alirezaie, Isabel Fernandez, Ferran Casals, Natalie Patey, Hélène Decaluwe, et al. 2013. "Exome Sequencing Identifies Mutations in the Gene TTC7A in French-Canadian Cases with Hereditary Multiple Intestinal Atresia." JOUR. *Journal of Medical Genetics* 50 (5): 324–29. doi:10.1136/jmedgenet-2012-101483 .

Schäffer, Alejandro A. 2013. "Digenic Inheritance in Medical Genetics." JOUR. *Journal of Medical Genetics* 50 (10): 641–52. doi:10.1136/jmedgenet-2013-101713 .

Schubert, Desirée, Claudia Bode, Rupert Kenefeck, Tie Zheng Hou, James B Wing, Alan Kennedy, Alla Bulashevska, et al. 2014. "Autosomal Dominant Immune Dysregulation Syndrome in Humans with CTLA4 Mutations." *Nature Medicine* 20 (12): 1410–16. doi:10.1038/nm.3746.

Sekine, Hideharu, Ricardo C Ferreira, Qiang Pan-Hammarström, Robert R Graham, Beth Ziemba, Sandra S de Vries, Jiabin Liu, et al. 2007. "Role for Msh5 in the Regulation of Ig Class Switch Recombination." *Proceedings of the National Academy of Sciences of the United States of America* 104 (17): 7193–98. doi:10.1073/pnas.0700815104.

Sund, Kristen Lipscomb, Sarah L Zimmerman, Cameron Thomas, Anna L Mitchell, Carlos E Prada, Lauren Grote, Liming Bao, Lisa J Martin, and Teresa A Smolarek. 2013. "Regions of Homozygosity Identified by SNP Microarray Analysis Aid in the Diagnosis of Autosomal Recessive Disease and Incidentally Detect Parental Blood Relationships." JOUR. *Genet Med* 15 (1). American College of Medical Genetics and Genomics: 70–78. http://dx.doi.org/10.1038/gim.2012.94.

Sunyaev, S R, F Eisenhaber, I V Rodchenkov, B Eisenhaber, V G Tumanyan, and E N Kuznetsov. 1999. "PSIC: Profile Extraction from Sequence Alignments with Position-Specific Counts of Independent Observations." *Protein Engineering* 12 (5): 387–94. doi:10.1093/protein/12.5.387.

Thorvaldsdottir, H, J T Robinson, and J P Mesirov. 2013. "Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration." JOUR. *Brief Bioinform* 14. doi:10.1093/bib/bbs017.

van Schouwenburg, Pauline A., Emma E. Davenport, Anne Kathrin Kienzler, Ishita Marwah, Benjamin Wright, Mary Lucas,

Tomas Malinauskas, et al. 2015. “Application of Whole Genome and RNA Sequencing to Investigate the Genomic Landscape of Common Variable Immunodeficiency Disorders.” *Clinical Immunology* 160 (2). Elsevier B.V.: 301–14. doi:10.1016/j.clim.2015.05.020.

Walsh, Tom, Hashem Shahin, Tal Elkan-Miller, Ming K Lee, Anne M Thornton, Wendy Roeb, Amal Abu Rayyan, et al. 2010. “Whole Exome Sequencing and Homozygosity Mapping Identify Mutation in the Cell Polarity Protein GPSM2 as the Cause of Nonsyndromic Hearing Loss DFNB82.” JOUR. *The American Journal of Human Genetics* 87 (1): 90–94. doi:http://dx.doi.org/10.1016/j.ajhg.2010.05.010.

Figure legends

Figure 1

Pedigree of the family of the patient L283 with the chromatogrames for each member of the family below, showing that the patient is homozygous and all their relatives heterozygous for the mutation.
II:1 - patient L283
II:2, II:3, II:4 – siblings of patient L283
I:1, I:2 – parents of patient L283

Figure 2

Boxplots representing the ratio of non-functional variants to synonymous variants, both with a frequency below 1%, in 25 immunological pathways, for the CVID patients without a clear molecular defect found (shaded in blue) and for the controls (shaded in red). Samples without synonymous variants below 1% in the genes of the pathway are not represented.
BCSP: B-cell signaling pathway, CCRI: cytokine-cytokine receptor interaction pathway, JSSP: JAK-STAT signaling pathway, NFKBSP: NFKB signaling pathway, PASP: PIK3-AKT signaling pathway, PIDP: primary immunodeficiency pathway, TCSP: T-cell signaling pathway, MRP: mismatch repair pathway, mTORSP: mTOR signaling pathway, TNFSP: TNF signaling pathway, RSP: ras signaling pathway, HCL: Hematopoietic cell lineage pathway, APP: antigen processing and presentation pathway, CCC: complement and coagulation cascades pathway, ChSP: chemokyne signaling pathway, CDNASP: cytosolic-DNA sensing pathway, FCERISP: Fc epsilon RI signaling pathway, FCGRIMP: Fc gamma RI-mediated phagocytosis pathway, IINIgA: intestinal immune network for IgA production pathway, LTM: leukocyte transendothelial migration pathway, NKCMC: natural killer cell mediated cytotoxicity pathway, NLRSP: NOD-like receptor signaling pathway, PAP: platelet activation pathway, RILRSP: RIG-I-like receptor signaling pathway, TRSP: toll-like receptor signaling pathway.

Figure 3

Western blot of from the patient L283 (right) and from a control (left). LRBA represents the band where LRBA protein should be found (none in the patient) and GAPDH is the loading control protein.

Tables from article in preparation

Gene CVID	Effect	Sample	Genotype
LRBA	STOP_GAINED(R2214*)	L283	1/1
MLH1	NON_SYNONYMOUS_CODING(R18L)	L283	0/1
IRF2BP2	CODON_CHANGE_PLUS_CODON_INSERTION(L93CM)	L287	0/1
PRKCD	NON_SYNONYMOUS_CODING(V276L)	L288	0/1
CLEC16A	NON_SYNONYMOUS_CODING(R305W)	L292	0/1
NOD2	FRAME_SHIFT	L292	0/1
CD37	NON_SYNONYMOUS_CODING(Y226D)	L293	0/1
DOCK8	NON_SYNONYMOUS_CODING(V759M)	L293	0/1
ANP32B	CODON_CHANGE_PLUS_CODON_DELETION(DE219E)	L294	0/1
CR2	NON_SYNONYMOUS_CODING(M989I)	N201	0/1
PIK3R1	SPLICE_SITE_DONOR	N202	0/1
DOCK8	NON_SYNONYMOUS_CODING(V759M)	N210	0/1
CTLA4	FRAME_SHIFT	N211	0/1
CD5	NON_SYNONYMOUS_CODING(S439F)	N212	0/1
IRF2BP2	CODON_CHANGE_PLUS_CODON_INSERTION(L93CM)	N213	0/1
CARD11	NON_SYNONYMOUS_CODING(R424W)	N214	0/1
IRF2BP2	CODON_CHANGE_PLUS_CODON_INSERTION(L93CM)	N216	0/1
CD84	NON_SYNONYMOUS_CODING(T78M)	N232	0/1
NOD2	FRAME_SHIFT	N233	0/1
NFKB1	SPLICE_SITE_DONOR	N234	0/1

Table 4: Functional variants found in the CVID candidate genes after filtering

polyphen	rs	gerp	esp5400_all	GMAF	Effect	gene	idsample	genotype
0.002	-	-3.25	-	-	NON_SYNONYMOUS_CODING(A293E)	CR2	N233	0/1
0.05	rs144572703	4.47	0.005763	0.0018	NON_SYNONYMOUS_CODING(V871L)	CR2	N233	0/1
0.598	rs187956469	5.18	0.002838	0.0032	NON_SYNONYMOUS_CODING(Y482H)	PLCG2	N212	0/1
0.005	rs75472618	-6.5	0.007067	0.0064	NON_SYNONYMOUS_CODING(N571S)	PLCG2	N212	0/1

Table 6: Compound heterozygous variants found in CVID candidate genes

rs	polyphen	gerp	esp5400_all	GMAF	Effect	gene	idsample	genotype
rs34557412	0.988	4.73	0.003997	0.0032	NON_SYNONYMOUS_CODING(C104R)	TNFRSF13B	L297	1/1
rs77874543	0.318	0.559	0.024034	0.0536	NON_SYNONYMOUS_CODING(P21R)	TNFRSF13C	L297	0/1

Table 7: Variants found in the interacting genes TNFRSF13B and TNFRSF13C in the same patient

Table 1.

Known CVID variants detected in CVID patients in this study.

Gene	cDNA	Aa change	Genotype[a] (reference)	hg19_pos	CVID (N=38)[b]	Controls (literature)	Controls (Autism, N=37)[b]	Controls (Spain, N=267)[b]
TNFRSF13B	c.752C>T	p.P251L	0/1 (Pan-Hammarstron et al., 2007)	17:16842991	9	yes	0	36 (3)
TNFRSF13B	c.310T>C	p.C104R	*/1 (Castigli et al., 2005)	17:16852187	3 (1)	yes	0	2
TNFRSF13C	c.62G>C	p.P21R	2*0/1 (Losi et al., 2005)	22:42322716	4	yes	0	16[c]
MSH5	c.253C>T	p.L85F	2*0/1 (Sekine et al., 2007)	6:31709045	2	yes	0	55 (2)

[a] 0/1 heterozygotes; 1/1 homozygotes; */1 heterozygotes and homozygotes; 2*0/1 compound heterozygotes.

[b] Homozygous individuals are shown in brackets.

[c] No data avialble for the 267 controls. Instead, we used data from 578 whole-exome sequences at the CIBERER Spanish Variant Server (csvs.babelomics.org).

Sample	CVID	PPI – CVID	ALL < 0.001
L283	1(1)	1(1)	61(5)
L287	0(0)	3(1)	52(5)
L288	0(0)	0(0)	49(3)
L289	0(0)	1(0)	67(4)
L290	0(0)	0(0)	52(4)
L291	0(0)	1(0)	59(1)
L292	1(0)	3(0)	52(2)
L294	0(0)	0(0)	53(3)
L295	0(0)	0(0)	48(5)
L296	0(0)	0(0)	53(3)
L297	1(0)	1(0)	43(1)
L298	0(0)	1(0)	35(0)
L299	0(0)	1(0)	55(1)
N201	0(0)	0(0)	42(1)
N202	1(0)	2(0)	46(4)
N203	0(0)	0(0)	57(3)
N204	0(0)	2(0)	50(2)
N205	0(0)	2(0)	51(1)
N206	0(0)	1(0)	63(1)
N207	1(0)	1(0)	59(3)
N208	0(0)	1(0)	59(4)
N209*	0(0)	2(0)	42(0)
N210	0(0)	3(0)	57(2)
N211	1(0)	1(0)	58(1)
N212	0(0)	0(0)	49(5)
N213	0(0)	2(0)	51(0)
N214	0(0)	0(0)	52(3)
N215*	0(0)	4(0)	96(2)
N216	0(0)	3(0)	78(10)
N223	0(0)	3(0)	58(1)
N224	0(0)	4(2)	47(4)
N225*	0(0)	1(1)	50(2)
N226*	0(0)	3(2)	63(5)
N227	0(0)	2(2)	64(3)
N228*	0(0)	3(2)	70(9)
N229	0(0)	2(0)	60(3)
N230*	0(0)	2(0)	59(4)
N231	0(0)	2(0)	74(4)
N232	0(0)	1(0)	63(5)
N233	1(0)	2(1)	62(4)
N234	2(0)	3(0)	62(6)
N235	0(0)	3(0)	66(4)
N237*	0(0)	5(0)	76(8)
N246*	0(0)	1(0)	55(4)

Table 2: Number of LoF per exome found after filtering, in three categories: variants in CVID genes with frequency below 0.01, variants in the PPI network of the CVID genes with frequency below 0.01%, and all the LoF variants in the exome with frequency below 0.001%

Table 3.

Genes with LoF homozygous or heterozygous variants in CVID candidate genes and interacting proteins.

Individual	CVID < 0.01	PPI < 0.01
L283	LRBA(hom)	
L287		C7orf64(het), PDGFRB(hom), RIPK4(het)
L289		HDAC1(het)
L291		GP6(het)
L292	NOD2(het)	SLA2(het), ZNF655(het)
L297	NFKB1(het)	
L298		MAPK8(het)
L299		FGFR3(het)
N202		FHOD1(het), PIK3R1(het)
N204		HP(het), PLSCR1(het)
N205		HNF1A (comp_het)
N206		RPA2(het)
N207	NFKB1(het)	
N208		EEF1G(het)
N209*		DERL3(het), HP(het)
N210		DERL3(het), HP(het), PDGFRB(het)
N211	CTLA4(het)	
N213		IBTK(het), PDGFRB(het)
N215*		CASP1(het), HCLS1(het), NCOR2(het), SPI1(het)
N216		CASP1(het), HCLS1(het) ,NCOR2(het)
N223		BCAP31(het), SLC6A8(het), TNFRSF12A(het)
N224		BCAP31(hom),CASP1,SLC6A8(hom),TNFRSF12A
N225*		TNFRSF12(hom)
N226*		BCAP31(hom), SLC6A8(het), TNFRSF12A(hom)
N227		BCAP31(het), SLC6A8(het)
N228*		BCAP31(hom), CSF3R(het), SLC6A8(hom)
N229		CR1(het), SPI1(het)
N230*		ITGB4(het), SPP1(het)
N231		PML(het), TNFRSF12A(het)
N232		TNFRSF12A(het)
N233	NOD2(het)	TNFRSF12A(het)
N234	IL10RA(het), NFKB1(het)	TNFRSF12A(het)
N235		C9(het), PIAS1(het), TRPV1(het)
N237*		C9(het), PRSS3(het), TNFRSF12A(het), TRPV1(het), XAF1(het)
N246*		PIAS1(het)

* healthy relatives.

Sample	Genes 1%	Genes 0.1%	Genes 1% filtered	Genes 0.1% filtered
L283	53	21	4	1
L287	26	11	2	0
L288	29	18	0	0
L289	39	19	3	0
L290	31	15	0	0
L291	27	18	2	1
L292	26	16	3	1
L294	31	20	1	1
L295	34	15	0	0
L296	25	13	0	0
L297	30	17	1	0
L298	28	7	1	0
L299	27	14	4	2
N201	33	23	3	1
N202	27	13	3	1
N203	33	13	2	0
N204	34	14	1	1
N205	28	15	0	0
N206	34	18	1	1
N207	29	17	0	0
N208	23	12	0	0
N209*	26	12	4	1
N210	33	18	1	0
N211	25	13	1	0
N212	25	11	1	1
N213	27	10	2	1
N214	19	6	1	0
N215*	84	29	3	1
N216	75	21	3	1
N223	35	19	1	0
N224	38	22	1	0
N225*	44	17	2	0
N226*	43	24	1	1
N227	37	25	1	0
N228*	46	35	3	1
N229	39	20	0	0
N230*	34	20	1	1
N231	35	20	3	1
N232	18	12	2	0
N233	35	19	0	0
N234	37	22	1	0
N235	42	22	0	0
N237*	45	17	3	0
N246*	41	23	0	0

Table 5: Number of genes harboring compound heterozygous mutations in each of the samples, in four categories determined by the frequency filter (below 1% or below 0.1%) and by conservation and damage prediction scores (GERP>2 and polyphen>0.5)

Figures from article in preparation

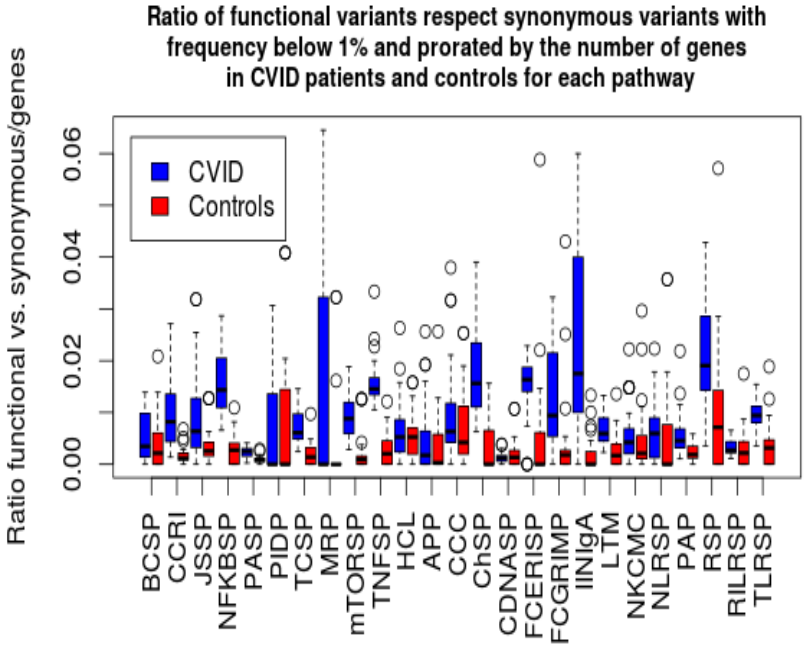
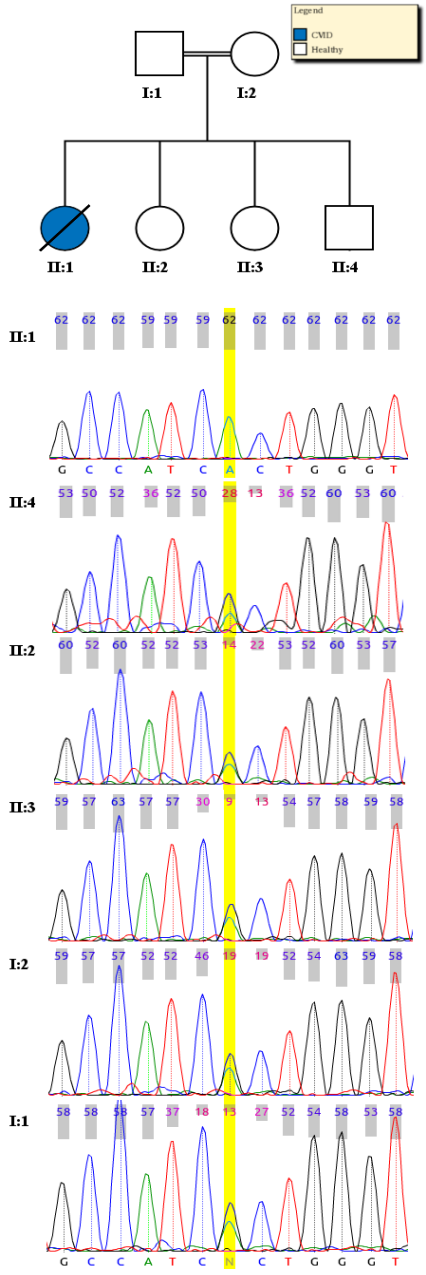


Figure 2: Ratio of functional to synonymous variants in 25 different pathways for CVID patients and controls, prorated by the number of genes.

Figure 1: Sanger sequencing of the patient N283 and her family, showing the segregation of the recessive mutation in the family



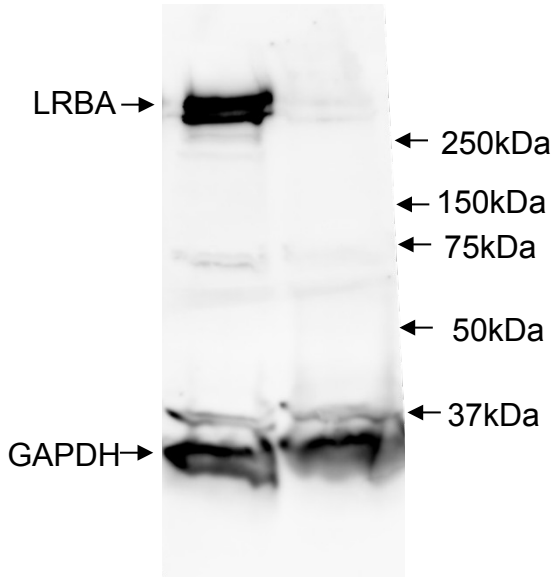


Figure 3: Western blot gel electroforesis for the control and the L283 with a homozygous stop mutation abolishing the expression of LRBA gene. GAPDH is the loading control protein.

4. Discussion

"*It's the questions we can't answer that teach us the most. They teach us how to think. If you give a man an answer, all he gains is a little fact. But give him a question and he'll look for his own answers*."
Patrick Rothfuss, The Wise Man's Fear

4.1 Polymorphic LoF variants in great apes

LoF variants have not been studied quantitatively until the omics revolution has made possible to sequence human genomes at an affordable price. The sequencing of thousands of genomes(Abecasis et al. 2010) and exomes(Lek et al. 2015) in the last years allowed to dispose of quantitative data and set a realistic threshold for the count of LoFs in a human genome, with evolutionary and medical implications. But NGS technology permits to go further and interrogate genomes others than those of humans or model organisms. The recent publication and analysis of several genomes of all the extant species of great apes(Prado-Martinez et al. 2013), as well as the draft genomes of chimpanzees(Chimpanzee Sequencing Consortium 2005), gorillas(Scally et al. 2012) and orangutans(Locke et al. 2011) is a huge step towards the diversification of comparative genomics that has allowed to resolve important issues, as the demographic history of the Hominidae family or its genetic diversity, and has encouraged as well related research in many facets, from the similar studies in other species to insights into ancient genomes. As our closest organisms, genetically speaking, great apes are of huge interest for the scientific

community and in special in the field of evolution, since it's thought that they hold the key to explain what makes us humans. Therefore it is natural, almost mandatory, to translate studies from humans to great apes. The study of potentially deleterious variants as are LoF has great relevance in those species, because the diminishing of their populations and their habitable space is putting great constrains in their genomes. The conservation efforts in those species makes necessary to know their mutational load, specially when variants as potentially deleterious as the LoF mutations are considered, specially when the use of captive populations, with inbreed individuals, will had to be use to recuperate their populations. The first chapter of this thesis aims to study the genetic load of polymorphic LoF variants in great apes and to present an initial comparative with humans.

Working with the set of genomes obtained for the Great Ape Genome Diversity Project (Prado-Martinez et al. 2013), we have provided the first quantitative study of polymorphic LoF mutations in great apes. The initial number of LoF obtained for each species ranged from two-fold to six-fold the number described in humans, increasing with the evolutive distance from *Homo sapiens*. Those high numbers are probably an overestimation due to several factors, being the more important ones the mapping and the annotation of the samples' reads against the human reference genome and the relative low quality of the NGS technologies, that introduce an important bias towards false positives that affects significantly more

LoF variants respect other kind o variants more abundant in the genome. Nonetheless they are, in general, lower than the first estimates obtained from genomic data from humans(Abecasis et al. 2010; Pelak et al. 2010). The differences in the number of LoF variants in the first human genomes obtained by NGS are mainly due to the use of samples with different coverage and to the software upon the mapping of the reads and the calling of the variants relied, which has been in a constant improving. In order to obtain an estimation more close to the reality and to prioritize those variants more likely prone to have deleterious effects, we have applied filters related with the position of the LoF inside the gene and the presence of another fixed LoF in the same gene, as well as those variants affecting high conserved positions and in genes known to accumulate LoF variants. We also have focused only in stop gain variants due to our high confidence in its annotation in primates respect the other variants, and in those variants with low frequency in the population as expected for harmful mutations with a real effect on the organism, in opposition to more frequent variants that are less visible to selection and can increase their frequencies by genetic drift. The final number obtained after those filters is relatively homogeneous across all the species and similar to the one found in humans (in the order of 8-10 missense mutations against the 27-37 reported by MacArthur et al. (2012)) , considering the limitations expected from the sample size in our study and the strong filtering that we have applied, specially the filter regarding the conservation of the position. In a recent attempt to screen LoF

variants in exomes of *Maccaca mulatta*, using the draft genome of this ape, the estimates obtained have many similarities with the work of this thesis can be appreciated(Cornish, Gibbs, and Norgren 2016). Although the *M. mulatta* study differs in many key aspects from the one presented in this thesis, as the different species of study and different caveats related to the reference genome, they report a similar number of stop gains (42-99) as ours (64-112) before filtering and they also have similar problems with the excessive numbers of other LoF variants, specially indels. A comparison between the LoF variants found in 4 individuals of *M. mulatta*, mapped against two different version of the *M.mulatta* genome, and the great ape genomes from our study (mapped against hg19 reference genome) can be found in table X. This reinforces the selection of stop gain introducing variants as good representatives of truly deleterious variants and adds weight to the use of the stringent filtering performed in our study, and adds weight to the necessary improvement of the reference genomes.

		Stop gain	Frameshift indels	Splice site	Stop loss	All
de Valles et al. (2016) Mapped against hg19	Homo sapiens	45.1(11.6)	78.5(16.7)	82.6(30.2)	7.6(3.2)	**214(61.8)**
	Pan paniscus	64.7(14)	171(58.8)	85.2(26.7)	11.1(7.1)	**332.2(106.6)**
	Pan troglodytes ellioti	86.8(24.2)	349(104.5)	118.5(34.2)	10.8(4.6)	**565.1(167.5)**
	Pan troglodytes schweinfurthii	76(16.6)	278.6(78.5)	100.5(24.6)	4.6(1)	**459.8(120.8)**
	Gorilla gorilla	112.7(31.9)	368.1(150.1)	136.9(53.1)	9.3(2)	**627(237.3)**
	Pongo abelii	101.4(21.4)	465.2(114.2)	124.6(34.8)	5(1.4)	**696.2(171.8)**
	Pongo pygmaeus	90.8(27.4)	408(98.6)	103.8(37.6)	8.6(4.6)	**611.2(168)**
Cornish *et al.*, 2016 4 *M. mulatta* mapped Against two different Reference genomes	rheMac2	291(63.25)	1206(705.75)	266(116.75)	177.5(107.5)	**1940,5(993.25)**
	MacaM7	76.5(5)	208(21)	96.75(17,25)	8.75(1)	**390(44.25)**

Table 8: Comparision between the LoF variants found in the genomes of great apes mapped against the human genome reference hg19, adapted from de Valles et al. (2016), and the LoF variants found in the genome of 4 M. mulatta mapped against two different references of the M. mulatta reference genome (rhesMac2 and MacaM7), adapted from Cornish et al. (2016)

The initial number of LoF mutations appears to be correlated with the effective population sizes, pointing to a neutral effect of a significant fraction of those LoFs in the fitness of the population, but when those numbers are divided by the number of synonymous changes the correlation with the effective population size is negative, probably a product of the increased efficiency of purging selection at high effective population sizes and supporting the idea that a significant fraction of those LoF have true detrimental effects and are not neutral variants. When only the stop gain variants are considered (still respect the synonymous variants), all correlation with the effective population size is lost. This suggests that the number of highly deleterious variants is not modified by the effective population sizes and that the selection component in their allelic frequency is very high compared to the effect of the

population size. A study by Xue *et al.* (2015) has found that mountain gorillas, whose populations have very low sizes, presents a great burden of genetic variation but a low charge of detrimental variants, due to their high consanguinity that increases the possibility of a LoF to be homozygous and therefore is more prone to be removed by selection. Moreover, in a recent analysis of homozygous LoF in isolated human populations(Kaiser et al. 2015), it has been found that, although those populations have more rare variants than expected, their homozygous LoF where rather disease risk alleles and that they don't produced a visible disease phenotype. Those publications gave force to our interpretation of the lack of correlation between highly deleterious variants and effective population sizes. It is obvious that low effective population sizes should increase the frequencies of the otherwise rare deleterious variants, but our study suggests that there is a limit for the numbers of those variants when they are highly deleterious and thus the population effect contribution is not significant against the purging selection. Previous studies of lethality have estimated, in general, low numbers of lethal alleles in the genomes, even when inbreed populations are the source of the estimates. The limit of highly deleterious variants respect the synonymous that we have reported and their independence to the effective population size suggests that what we are reporting is related with this genetic load of lethal alleles, and the genes harboring those variants could be good candidates to prioritize novel gene discovery in prenatal death studies or to found novel relationships between genes and essential

cell functions. Projects currently under development as the Human Knockout Project(J. Kaiser 2014) or the DMDD in mice(Mohun et al. 2013) could provide future confirmations about the lethality of those variants (or the lack of it in the variants that we have filtered out).

The distribution of LoF variants across the genes is weakly influenced by the presence of a fixed variant in the population and strongly influenced by the presence of a fixed LoF in at least one of the species considered, pointing to a non-essential gene whose function could be lost with little effect on the organism (at least in the great apes). In all the species almost one third of the polymorphic LoF variants are found in genes that harbor a fixed LoF in at least one of the populations, and as much as 60% of the variants are in genes that have a LoF variant in any of the populations considered. Those percentages are also high (~40%) when the genes with a LoF in its sequence are compared with those found in humans in the 1000 Genomes Project(Khurana et al. 2013). Nonetheless, the number of shared LoF variants across species is relatively low and mainly found in CpG sites, more likely to be byproduct of recurrent mutations that polymorphisms maintained in the population through large evolutionary times. This sharing of genes harboring LoF variants across great apes could provide a set of primate LoF-tolerant genes which has potential to be useful in the prioritization of variants in some clinical studies, as well as a focus to assess essentiality of some genes whose function is

nowadays still not well understood, and it could be interesting to compare them with the genes that will be found in the Human Knockout Project, since the differences and similarities could unravel previously unknown traits that separate humans from great apes or that bring us closer to them.

Due to the low sample sizes of our study and the many possible sources of error, the results of gene enrichment analysis yield pathways with probability values that made them not reliable. Otherwise those results could have been really interesting, but this work should be undergone in future studies with bigger sample sizes. The only pathways found to be enriched in LoF variants across great ape genomes from our study with a significant level where those related with olfactory reception, as reported in humans(MacArthur et al. 2012a; Kaiser et al. 2015). Genes encoding proteins with olfactory reception is the biggest family of genes in mammals with more than 1,500 genes in some families of mammals. Nonetheless, primates is one of the most microsmic (with olfactory identification deficit) lineages between the mammals, with less than 800 genes for olfactory reception, and many of them are pseudogenes or are in process of pseudogenitzation, specially in humans compared to the other great apes(Fleischer 2009). Curiously this enrichment in LoF variants in olfactory receptor genes has been also found in an animal defined as macrosmic (with an olfactory identification superior to the norm) like *Bos taurus*(Das et al. 2015), which has a sense of smell nearly

as good as the pig and likewise is expected to have lows amounts of olfactory receptor pseudogenitzation(Groenen et al. 2012). Pigs have a keen sense of smell, a proof of it could be the market price of truffle hogs, with values as high as 145,000$ (in Australian dollars, data extracted from http://www.breednet.com.au). The recent characterization of the *Sus scrofa* olfactory subgenome shows that this species has one of the largest olfactory receptor repertory, with many specific olfactory receptors for swine olfaction, and with the lowest proportion (14%) of olfactory receptor genes under pseudogenitzation(Nguyen et al. 2012)

In perspective, our study tries to provide a more reliable estimate of the number of truly deleterious variants in great ape genomes. In order to select the most deleterious variant, our approach relies in low frequency stop gain introducing LoF in conserved sites. We have limited power when assessing the frequency of the variants in our populations, since in some of them we had only 5 individual genomes and therefore the 1% frequency threshold may not be true for some of those variants when future studies include bigger sample sizes. Nonetheless, the use of the stop gain as deleterious variants, and specially the stringent filter of keeping only variants with GERP values higher than 4, allows us to estimate real numbers of LoF mutations with potential strong effects in the fitness of the individuals, although those variants will require a future functional validation.

Future studies in this direction should consider the possibility of include even more genomes of great apes or make insight in a single population, in order to have a sample size enough to get significance in gene enrichment analysis, as well as use the completed reference genomes for the respective reads and sequence the reads with technologies outputting greater read lengths and accuracy when they are available. There is also a need to improve gene models and annotations for the great apes to reduce the faulty calling of the variants in this kind of studies.

4.2 LoF variants in the common variable immunodeficiency

Finding the underlying molecular defect in PIDs is necessary not only to increase the knowledge of this wide group of diseases but also to provide specific treatments to the affected individuals. Early diagnosis his crucial and highly correlated with patients survival, and where traditional methods to quickly diagnose PIDs with high heterogeneity failed, new methods based on NGS technologies could provide a reliable and faster alternative(Raje et al. 2014). Research in CVID has been greatly improved since it introduced NGS methods. Initial studies were performed on few individuals or families, which provided limited power to find genes implied in the disease mechanism and only when it followed a Mendelian pattern

of inheritance. Nowadays studies of CVID are more inclusive, not only in the number patients but also in the genomic region interrogated. This has allowed the finding of several genes implied it the disease and the focus of the research is moving beyond the mere Mendelian model towards a more integrative perspective. Nonetheless, studies in general still only find the molecular cause in less than 20-25% of the familiar cases(Saikia and Gupta 2016; Bacchelli et al. 2007; Park et al. 2009; Rodríguez-Cortez et al. 2015), that are only a 10% of the patients of CVID(Chapel et al. 2008; Li et al. 2016). In overall, it is appointed that a clear molecular defect explaining the disease is found in roughly a 5% of the patients diagnosed with CVID(van Schouwenburg et al. 2015).

Our study includes 36 patients with both sporadic and familiar forms of CVID, using whole exome sequencing and genotypic data for all of the patients and with relatives of the patients in 6 cases, with a total of 8 exomes from individuals not diagnosed with CVID. This facilitates the screening of almost all the potential exonic variants and the analysis of structural variants, which has been useful to pinpoint specific SNPs in regions with low heterozygosity in the consanguineous samples as well as to detect CNVs in our samples. When our project was planned 4 years ago, there weren't any publications that used whole exome sequencing on a great number of samples to research in the CVID field. Since then, many reports have associated molecular defects to CVID using NGS, but usually in few patients or families, with the exception of the works

by van Schouwenbourg *et al.* (2015), Kuehn *et al.* (2016) and Maffuci *et al.* (2016). The first one has used WGS in 34 cases of CVID, mainly sporadic, plus RNA sequencing of samples enriched in B cells from 3 CVID cases and 3 controls, identifying many variants in known (TNFRSF13B, TNFRSF13C and LRBA) and novel genes (NRLP12) for the disease and in known and novel pathways of the disease(van Schouwenburg et al. 2015). Kuehn et al. (2016), in the other hand, have used WES and a comparative genomic array (a similar approach to our own) in 29 patients from 6 families, finding several mutations and two big deletions in the IKZF1 gene. After an extensive functional workout, they have established IKZF1 as a new CVID gene with a dominant Mendelian inheritance. Curiously, they forgot to mention in his paper how they have selected the patients in order to enclose a disease as heterogeneous as CVID into a single defective gene model(Kuehn et al. 2016). To the date, the most inclusive to my knowledge is the one published by Maffuci et al. (2016), containing WES data for 50 patients and finding 17 probable monoallelic or biallelic variants for a 30% (15) of their patients, in the genes NFKB1, STAT3, CTLA4, PIK3CD, IKZF1, LRBA and STXBP2, after a inital screening in 269 PID genes. 42 other damaging variants are also reported but thay don't consider it to be causative based on the inheritance model and the patetiens' phenotypes(Maffucci et al. 2016). Those three studies validate the power of NGS approaches in the study of CVID and provides further confirmation of the duality between Mendelian cases and a more complex model among the patients for this

disease. One possible criticism, applicable mainly to van Schoubert *et al.* and Maffuci *et al*. studies, as well as the majority of our causal genes, is the lack of functional validation that provides the definitive evidence to link each mutation to the disease phenotype, but the time and effort required made this horizon unattainable in this kinds of studies involving so many patients. Mutations in the TNFRSF13B gene have made evident that the assignation to causality to one variant in this disease should be complemented with functional data, because after its discovery some have proven to have incomplete penetrancy or even to be risk factors rather than causal variants(Pan-Hammarström et al. 2007).

One of CVID studies biggest handicaps is his difficult diagnostic, as it is defined mainly by exclusion. Other studies used an inclusive approach by screening several PID genes, and we have settled for an intermediate approach where we perform and initial screening for the genes involved in possible diseases that are causes of hypogammaglobulinemia that have to be ruled out to diagnose a patient with CVID. This screening has found one patient with a previously described(Deau et al. 2014) heterozygous mutation in the splice site of the dominant PIK3R1 gene. We have found also a novel non-synonymous mutation in the same gene for two sisters in our samples. The position has a high GERP conservation score of 5.41 but the deleteriousness predictors SIFT and Polyphen have values that suggest a more neutral role for the variant. Further validation is needed to assess causality of the variant in the affected

sisters' phenotypes. After this initial screening we have selected functional variants with frequencies below 1% in the 1000 genomes and ExAC databases in a list of 97 genes, including the most well known to be related with CVID, genes found in the literature to participate in a CVID-like phenotype and genes that have been suggested by the clinicians that participate in the study. Among those genes the easiest to evaluate are those containing variants already described as causative of CVID with a clear Mendelian model. We have found LoF mutations in the LRBA, CTLA4 and NFKB1 genes, including a large deletion affecting the NFKB1 gene. LRBA gene is one of the most frequent genes in CVID literature, with a recessive inheritance and several variants described. We have found a homozygous stop gain introducing mutation, in a region of low homozigosity confirmed by genotyping in a consanguineous female patient. The mutation has been validated by Sanger in the affected female as well as in his family, and has been found to produce a total lack of the protein in a western blot analysis, probably due to NMD. In the CTLA4 gene we report a heterozygous *de novo* frameshift mutation in a male patient, confirmed by Sanger. The CTLA4 gene is known to be haploinsufficient and the description of the variant permitted the clinician that is involved in that particular case to treat the patient with a specific drug that has improved visibly the condition of the patient(Alsina 2015). In the NFKB1 gene we have found a novel heterozygous variant in the splice site donor in one of our samples, as well as a big deletion including one copy of this gene, known to

be recessive. Two of our patients have one start loss mutation in the same gene, but that has a frequency relatively high (0.002), and therefore is unlikely to be causative or at least to have a monogenic contribution to the diseases. In overall, we have found the probable molecular cause of the disease phenotype in 5 of our 36 patients (13%) solely by analyzing the LoF variants. Through the analysis of compound heterozygous variants we can possibly add one more patient with compound heterozygous non-synonymous mutations in the gene PLCG2, besides the two sisters aforementioned. The confidence for the causality in the non-synonymous compound heterozygous model on the disease is far more tenuous than in the case of LoF mutations, so we cautiously report them without strongly assuming that they are the final molecular diagnostic for this patients. Further functional validations, outside the scope of this thesis, are needed for those variants. Besides the genes from our list to exclude the diagnostic of CVID or from our list of 97 candidate variants, we also have found LoF variants in some pathways related with the immune system, suggesting a connection with CVID that must be researched in future studies.

To infer which variants may be implied in CVID for the cases where we weren't able to find a variant in known monogenic genes, we have taken two distinct approaches: one similar to the one taken in the van Schouwenburg *et al*. study, searching for variants in genes that interact with one of the 97 candidate genes considered in our study, and other approach analyzing the functional variants

found in pathways important for the disease, extracted from the KEGG database. By finding functional variants that are in both a CVID gene and an interacting gene we can extend our study to a digenic model for the disease, with special power were those variants are affecting interacting functional domains of the protein. The pathway analysis also gave interesting results, specially when comparing the CVID patients with the controls. We have found a ratio of low frequency functional variants respect synonymous variants significantly higher in the pathways related with NFKB-signaling and in the T-cell signaling pathway consistent with previous reports(Giovannetti et al. 2007; Keller et al. 2016). Moreover, we have found differences inside our dataset, with some patients having an excess of variants in some of the pathways respect the overall found in all the patients. Although our knowledge about the implications of an excess of variants in those pathways in the disease phenotypes of the patients is limited, future studies could benefit from our findings. Furthermore, we have found 66 LoF variants in those pathways, some of them in genes implied in more than one pathway. Although the haploinsufficiency prediction would discard many of these as disease causative *per se*, it could be interesting to analyze the join effect of the LoF variants and other functional variants found in the same pathway.

In overall, our study has found several novel variants in genes related to the disease and has provided a possible relationship between many novel genes and the CVID, that will need further

research and functional validation to prove a direct cause-effect. It demonstrates the value of NGS in diagnostic of complex diseases, specially when there is a certain Mendelian components, and it stresses the need of new methodologies to tackle complex diseases. Our results suggests a possible path to follow, already suggested by others, in the using of protein-protein interacting networks and pathway analysis to understand the mechanisms involved in common and complex diseases.

Bibliografia

Abecasis, Gonçalo R, David Altshuler, Adam Auton, Lisa D Brooks, Richard M Durbin, Richard a Gibbs, Matt E Hurles, and Gil a McVean. 2010. "A Map of Human Genome Variation from Population-Scale Sequencing." *Nature* 467 (7319): 1061–73. doi:10.1038/nature09534.

Aksentijevich, Ivona. 2015. "Update on Genetics and Pathogenesis of Autoinflammatory Diseases: The Last 2 Years." JOUR. *Seminars in Immunopathology* 37 (4): 395–401. doi:10.1007/s00281-015-0478-4.

Al-Herz, W, H Aldhekri, M.-R. Barbouche, and N Rezaei. 2014. "Consanguinity and Primary Immunodeficiencies." JOUR. *Human Heredity* 77 (1–4): 138–43. http://www.karger.com/DOI/10.1159/000357710.

Alangari, Abdullah, Abdulrahman Alsultan, Nouran Adly, Michel J Massaad, Iram Shakir Kiani, Abdulrahman Aljebreen, Emad Raddaoui, et al. 2012. "LPS-Responsive Beige-like Anchor (LRBA) Gene Mutation in a Family with Inflammatory Bowel Disease and Combined Immunodeficiency." JOUR. *Journal of Allergy and Clinical Immunology* 130 (2): 481–488.e2. doi:http://dx.doi.org/10.1016/j.jaci.2012.05.043.

Alsina, Laia. 2015. Personal Communication.

Altwegg, Kathrin, Hans Balsiger, Akiva Bar-nun, Jean-jacques Berthelier, Andre Bieler, Peter Bochsler, Christelle Briois, et al. 2016. "Prebiotic Chemicals — Amino Acid and Phosphorus — in the Coma of Comet 67P / Churyumov-Gerasimenko," no. May: 1–6.

Amariglio, N, and G Rechavi. 1993. “Insertional Mutagenesis by Transposable Elements in the Mammalian Genome.” *Environmental and Molecular Mutagenesis* 21 (3): 212–18. doi:10.1002/em.2850210303.

Arjunaraja, Swadhinya, and Andrew L. Snow. 2015. “Gain-of-Function Mutations and Immunodeficiency: At a Loss for Proper Tuning of Lymphocyte Signaling.” *Current Opinion in Allergy and Clinical Immunology* 15 (6): 533–38. doi:10.1097/ACI.0000000000000217.

Aze, Antoine, Jin Chuan Zhou, Alessandro Costa, and Vincenzo Costanzo. 2013. “DNA Replication and Homologous Recombination Factors: Acting Together to Maintain Genome Stability.” *Chromosoma* 122 (5): 401–13. doi:10.1007/s00412-013-0411-3.

Bacchelli, C, S Buckridge, a J Thrasher, and H B Gaspar. 2007. “Translational Mini-Review Series on Immunodeficiency: Molecular Defects in Common Variable Immunodeficiency.” *Clinical and Experimental Immunology* 149 (3): 401–9. doi:10.1111/j.1365-2249.2007.03461.x.

Bafunno, Valeria, Chiara Divella, Francesco Sessa, Giovanni Luca Tiscia, Giuseppe Castellano, Loreto Gesualdo, Maurizio Margaglione, and Vincenzo Montinaro. 2013. “De Novo Homozygous Mutation of the C1 Inhibitor Gene in a Patient with Hereditary Angioedema.” *Journal of Allergy and Clinical Immunology* 132 (3). doi:10.1016/j.jaci.2013.04.006.

Baker, Kristian E., and Roy Parker. 2004. “Nonsense-Mediated mRNA Decay: Terminating Erroneous Gene Expression.” *Current Opinion in Cell Biology* 16 (3): 293–99. doi:10.1016/j.ceb.2004.03.003.

Balasubramanian, Suganthi, Lukas Habegger, Adam Frankish, Daniel G. MacArthur, Rachel Harte, Chris Tyler-Smith, Jennifer Harrow, and Mark Gerstein. 2011. "Gene Inactivation and Its Implications for Annotation in the Era of Personal Genomics." *Genes and Development* 25 (1): 1–10. doi:10.1101/gad.1968411.

Bamshad, Michael J, Sarah B Ng, Abigail W Bigham, Holly K Tabor, Mary J Emond, Deborah a Nickerson, and Jay Shendure. 2011. "Exome Sequencing as a Tool for Mendelian Disease Gene Discovery." *Nature Reviews. Genetics* 12 (11). Nature Publishing Group: 745–55. doi:10.1038/nrg3031.

Bell, E. A., P. Boehnke, T. M. Harrison, and W. L. Mao. 2015. "Potentially Biogenic Carbon Preserved in a 4.1 Billion-Year-Old Zircon." *Proceedings of the National Academy of Sciences of the United States of America* Early Edit (20): 1–4. doi:10.1073/pnas.1517557112.

Bellini, Tommaso, Marco Buscaglia, Andrea Soranno, and Giuliano Zanchetta. 2012. "Origin of Life Scenarios: Between Fantastic Luck and Marvelous Fine-Tuning," 113–40.

Bernstein, Carol, Anil R Prasad, Valentine Nfonsam, and Harris Bernstein. 2013. "DNA Damage , DNA Repair and Cancer." *New Research Directions in DNA Repair*, 413–66. doi:10.5772/53919.

Bittles, a H, and J V Neel. 1994. "The Costs of Human Inbreeding and Their Implications for Variations at the DNA Level." *Nature Genetics* 8 (2): 117–21. doi:10.1038/ng1094-117.

Boisson, Bertrand, Pierre Quartier, and Jean Laurent Casanova. 2015. "Immunological Loss-of-Function due to Genetic Gain-of-Function in Humans: Autosomal Dominance of the Third

Kind." *Current Opinion in Immunology*. doi:10.1016/j.coi.2015.01.005.

Brough, Helen A., Angela Simpson, Kerry Makinson, Jenny Hankinson, Sara Brown, Abdel Douiri, Danielle C M Belgrave, et al. 2014. "Peanut Allergy: Effect of Environmental Peanut Exposure in Children with Filaggrin Loss-of-Function Mutations." *Journal of Allergy and Clinical Immunology* 134 (4). Elsevier Ltd: 867–875.e1. doi:10.1016/j.jaci.2014.08.011.

Bürger, Reinhard, Martin Willensdorfer, and Martin A. Nowak. 2006. "Why Are Phenotypic Mutation Rates Much Higher than Genotypic Mutation Rates?" *Genetics* 172 (1): 197–206. doi:10.1534/genetics.105.046599.

Burns, Siobhan O, Helen L Zenner, Vincent Plagnol, James Curtis, Kin Mok, Michael Eisenhut, Dinakantha Kumararatne, Rainer Doffinger, Adrian J Thrasher, and Sergey Nejentsev. 2012. "LRBA Gene Deletion in a Patient Presenting with Autoimmunity without Hypogammaglobulinemia." JOUR. *Journal of Allergy and Clinical Immunology* 130 (6): 1428–32. doi:http://dx.doi.org/10.1016/j.jaci.2012.07.035.

Calafell, Francesc, Francis Roubinet, Anna Ram??rez-Soriano, Naruya Saitou, Jaume Bertranpetit, and Antoine Blancher. 2008. "Evolutionary Dynamics of the Human ABO Gene." *Human Genetics* 124 (2): 123–35. doi:10.1007/s00439-008-0530-8.

Cargill, M, D Altshuler, J Ireland, P Sklar, K Ardlie, N Patil, N Shaw, et al. 1999. "Characterization of Single-Nucleotide Polymorphisms in Coding Regions of Human Genes." *Nature Genetics* 22 (3): 231–38. doi:10.1038/10290.

Casals, Ferran, Anna Ferrer-Admetlla, Martin Sikora, Anna Ramírez-Soriano, Tomàs Marquès-Bonet, Stéphanie Despiau, Francis Roubinet, Francesc Calafell, Jaume Bertranpetit, and Antoine Blancher. 2009. "Human Pseudogenes of the ABO Family Show a Complex Evolutionary Dynamics and Loss of Function." *Glycobiology* 19 (6): 583–91. doi:10.1093/glycob/cwp017.

Castellana, Stefano, and Tommaso Mazza. 2013. "Congruency in the Prediction of Pathogenic Missense Mutations: State-of-the-Art Web-Based Tools." JOUR. *Briefings in Bioinformatics* , March. doi:10.1093/bib/bbt013.

Castigli, Emanuela, Stephen a Wilson, Lilit Garibyan, Rima Rachid, Francisco Bonilla, Lynda Schneider, and Raif S Geha. 2005. "TACI Is Mutant in Common Variable Immunodeficiency and IgA Deficiency." *Nature Genetics* 37 (8): 829–34. doi:10.1038/ng1601.

Cech, Thomas R. 2010. "The RNA World in Context." *RNA Worlds*, 9–13. doi:10.1101/cshperspect.a006742.

Chakravarti, Aravinda. 2001. "...to a Future of Genetic Medicine." *Nature* 409: 822–23. doi:10.1038/ncb2703.

Chamary, J V, Joanna L Parmley, and Laurence D Hurst. 2006. "Hearing Silence: Non-Neutral Evolution at Synonymous Sites in Mammals." *Nature Reviews Genetics* 7 (2): 98–108. doi:10.1038/nrg1770.

Chang, Yao-Fu F, J Saadi Imam, and Miles F Wilkinson. 2007. "The Nonsense-Mediated Decay RNA Surveillance Pathway." *Annual Review of Biochemistry* 76: 51–74. doi:10.1146/annurev.biochem.76.050106.093909.

Chapel, Helen, and Charlotte Cunningham-Rundles. 2009. “Update in Understanding Common Variable Immunodeficiency Disorders (CVIDs) and the Management of Patients with These Conditions.” *British Journal of Haematology* 145 (6): 709–27. doi:10.1111/j.1365-2141.2009.07669.x.

Chapel, Helen, Mary Lucas, Martin Lee, Janne Bjorkander, David Webster, Bodo Grimbacher, Claire Fieschi, Vojtech Thon, Mohammad R Abedi, and Lennart Hammarstrom. 2008. “Common Variable Immunodeficiency Disorders: Division into Distinct Clinical Phenotypes.” *Blood* 112 (2): 277–86. doi:10.1182/blood-2007-11-124545.

Charbonnier, Louis Marie, Erin Janssen, Janet Chou, Toshiro K. Ohsumi, Sevgi Keles, Joyce T. Hsu, Michel J. Massaad, et al. 2015. “Regulatory T-Cell Deficiency and Immune Dysregulation, Polyendocrinopathy, Enteropathy, X-Linked-like Disorder Caused by Loss-of-Function Mutations in LRBA.” *Journal of Allergy and Clinical Immunology* 135 (1). Elsevier Ltd: 217–27. doi:10.1016/j.jaci.2014.10.019.

Charlier, C, W Li, C Harland, M Littlejohn, F Creagh, M Keehan, T Druet, W Coppieters, R Spelman, and M Georges. 2014. “NGS-Based Reverse Genetic Screen Reveals Loss-of-Function Variants Compromising Fertility in Cattle.” CONF. In *Vancouver: 10th World Congress on Genetics Applied to Livestock Production*, 17–22.

Chasman, Daniel, and R.Mark Adams. 2001. “Predicting the Functional Consequences of Non-Synonymous Single Nucleotide Polymorphisms: Structure-Based Assessment of Amino Acid variation1.” JOUR. *Journal of Molecular Biology* 307 (2): 683–706. doi:http://dx.doi.org/10.1006/jmbi.2001.4510.

Chen, J, and A V Furano. 2015. “Breaking Bad: The Mutagenic Effect of DNA Repair.” *DNA Repair (Amst)* 32. Elsevier B.V.: 43–51. doi:10.1016/j.dnarep.2015.04.012.

Chevalier, Frédéric D, Winka Le Clec’h, Nina Eng, Anastasia R Rugel, Rafael Ramiro de Assis, Guilherme Oliveira, Stephen P Holloway, et al. 2016. “Independent Origins of Loss-of-Function Mutations Conferring Oxamniquine Resistance in a Brazilian Schistosome Population.” JOUR. *International Journal for Parasitology* 46 (7): 417–24. doi:http://dx.doi.org/10.1016/j.ijpara.2016.03.006.

Choi, Murim, Ute I Scholl, Weizhen Ji, Tiewen Liu, Irina R Tikhonova, Paul Zumbo, Ahmet Nayir, et al. 2009. “Genetic Diagnosis by Whole Exome Capture and Massively Parallel DNA Sequencing.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (45): 19096–101. doi:10.1073/pnas.0910672106.

Cohen, Jonathan C, Eric Boerwinkle, Thomas H Mosley, and Helen H Hobbs. 2006. “Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease.” JOUR. *New England Journal of Medicine* 354 (12). Massachusetts Medical Society: 1264–72. doi:10.1056/NEJMoa054013.

Conley, Mary Ellen, Luigi D Notarangelo, and Amos Etzioni. 1999. “Diagnostic Criteria for Primary Immunodeficiencies.” JOUR. *Clinical Immunology* 93 (3): 190–97. doi:http://dx.doi.org/10.1006/clim.1999.4799.

Conrad, Donald F, Jonathan E M Keebler, Mark a Depristo, Sarah J Lindsay, Ferran Cassals, Youssef Idaghdour, Chris L Hartl, Carlos Torroja, and V Kiran. 2012. “Europe PMC Funders Group Variation in Genome-Wide Mutation Rates within and

between Human Families" 43 (7): 712–14. doi:10.1038/ng.862.Variation.

Cooper, Gregory M, Eric A Stone, George Asimenos, Eric D Green, Serafim Batzoglou, and Arend Sidow. 2005. "Distribution and Intensity of Constraint in Mammalian Genomic Sequence." JOUR. *Genome Research* 15 (7): 901–13. doi:10.1101/gr.3577405.

Cordell, Heather J. 2009. "Detecting Gene-Gene Interactions That Underlie Human Diseases." *Nature Reviews. Genetics* 10 (6): 392–404. doi:10.1038/nrg2579.

Cornish, Adam S., Robert M. Gibbs, and Robert B. Norgren. 2016. "Exome Screening to Identify Loss-of-Function Mutations in the Rhesus Macaque for Development of Preclinical Models of Human Disease." *BMC Genomics* 17 (1). BMC Genomics: 170. doi:10.1186/s12864-016-2509-5.

Correns, Carl. 1900. "G. Mendels Regel Über Das Verhalten Der Nachkommenschaft Der Rassenbastarde." *Berichte Der Deutschen Botanischen Gesellschaft* 18: 158–68.

Coveney, Peter V., Jacob B. Swadling, Jonathan a. D. Wattis, and H. Christopher Greenwell. 2012. "Theory, Modelling and Simulation in Origins of Life Studies." *Chemical Society Reviews* 41 (16): 5430. doi:10.1039/c2cs35018a.

Crick, Francis. 1955. "On Degenerate Templates and the Adaptor Hypothesis: A Note for the RNA Tie Club." *Original Repository: Wellcome Library for the History and Understanding of Medicine.*, 18. http://archives.wellcome.ac.uk/.

———. 1958. "On Protein Synthesis." *The Symposia of the Society for Experimental Biology*, 138–66.

———. 1970. “Central Dogma of Molecular Biology.” *Nature* 227 (5258): 561–63. doi:10.1038/227561a0.

———. 1981. *Life Itself. Its Origin and Nature*. *Book*.

———. 1988. *What Mad Pursuit. A Personal View of Scientific Discovery.*

Crick, Francis, L Barnett, S Brenner, and R J Watts-Tobin. 1961. “General Nature of the Genetic Code for Proteins.” *Nature* 192: 1227–32. doi:10.1038/1921227a0.

Crisafulli, Concetta, Antonio Drago, Marco Calabrò, Edoardo Spina, and Alessandro Serretti. 2015. “A Molecular Pathway Analysis Informs the Genetic Background at Risk for Schizophrenia.” *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 59: 21–30. doi:10.1016/j.pnpbp.2014.12.009.

Daetwyler, Hans D, Aurelien Capitan, Hubert Pausch, Paul Stothard, Rianne van Binsbergen, Rasmus F Brondum, Xiaoping Liao, et al. 2014. “Whole-Genome Sequencing of 234 Bulls Facilitates Mapping of Monogenic and Complex Traits in Cattle.” JOUR. *Nat Genet* 46 (8). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 858–65. http://dx.doi.org/10.1038/ng.3034.

Darwin, Charles. 1859. *On the Origins of Species by Means of Natural Selection*. *London: Murray*. doi:10.1126/science.146.3640.51-b.

———. 1868. “The Variation of Animals and Plants under Domestication.” *Animals* 1: 1–411. doi:10.1017/CBO9780511709500.

Das, Ashutosh, Frank Panitz, Vivi Raundahl Gregersen, Christian Bendixen, and Lars-Erik Holm. 2015. "Deep Sequencing of Danish Holstein Dairy Cattle for Variant Detection and Insight into Potential Loss-of-Function Variants in Protein Coding Genes." *BMC Genomics* 16. BMC Genomics: 1043. doi:10.1186/s12864-015-2249-y.

Davydov, Eugene V., David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. 2010. "Identifying a High Fraction of the Human Genome to Be under Selective Constraint Using GERP++." *PLoS Computational Biology* 6. doi:10.1371/journal.pcbi.1001025.

De Bont, Rinne, and Nik van Larebeke. 2004. "Endogenous DNA Damage in Humans: A Review of Quantitative Data." *Mutagenesis* 19 (3): 169–85. doi:10.1093/mutage/geh025.

Deau, Marie-Céline, Lucie Heurtier, Pierre Frange, Felipe Suarez, Christine Bole-Feysot, Patrick Nitschke, Marina Cavazzana, et al. 2014. "A Human Immunodeficiency Caused by Mutations in the PIK3R1 Gene." JOUR. *The Journal of Clinical Investigation* 124 (9). The American Society for Clinical Investigation: 3923–28. doi:10.1172/JCI75746.

De Vries, Hugo. 1900. "Sur La Loi de Disjonction Des Hybrides." *Comptes Rendus de l'Academie Des Sciences* 130: 845–47.

Du, Jingjing, Sarah Z Dungan, Amir Sabouhanian, and Belinda S W Chang. 2014. "Selection on Synonymous Codons in Mammalian Rhodopsins: A Possible Role in Optimizing Translational Processes." *BMC Evolutionary Biology* 14 (1): 96. doi:10.1186/1471-2148-14-96.

Ebadi, Maryam, Asghar Aghamohammadi, and Nima Rezaei. 2015. "Primary Immunodeficiencies: A Decade of Shifting

Paradigms, the Current Status and the Emergence of Cutting-Edge Therapies and Diagnostics." JOUR. *Expert Review of Clinical Immunology* 11 (1). Taylor & Francis: 117–39. doi:10.1586/1744666X.2015.995096.

Edwards, Matthew J., Jonathan P. Park, Doris H. Wurster-Hill, and John M. Jr. Graham. 1994. "Mixoploidy in Humans: Two Surviving Cases of Diploid-Tetraploid Mixoploidy and Comparison With Diploid-Triploid Mixoploidy." *American Journal of Medical Genetics* 52: 324–30.

Elgar, Greg, and Tanya Vavouri. 2008. "Tuning in to the Signals: Noncoding Sequence Conservation in Vertebrate Genomes." *Trends in Genetics* 24 (7): 344–52. doi:10.1016/j.tig.2008.04.005.

Ferrer-Admetlla, Anna, Martin Sikora, Hafid Laayouni, Anna Esteve, Francis Roubinet, Antoine Blancher, Francesc Calafell, Jaume Bertranpetit, and Ferran Casals. 2009. "A Natural History of FUT2 Polymorphism in Humans." *Molecular Biology and Evolution* 26 (9): 1993–2003. doi:10.1093/molbev/msp108.

Fields, C, Md Adams, Owen White, and Jc Venter. 1994. "How Many Genes in the Human Genome?" *Nature Genetics* 7 (3): 345–46. doi:doi:10.1038/ng0794-345.

Filges, Isabel, and Jan M. Friedman. 2015. "Exome Sequencing for Gene Discovery in Lethal Fetal Disorders - Harnessing the Value of Extreme Phenotypes." *Prenatal Diagnosis* 35 (10): 1005–9. doi:10.1002/pd.4464.

Fleischer, Joerg. 2009. "Mammalian Olfactory Receptors." *Frontiers in Cellular Neuroscience* 3 (August): 9. doi:10.3389/neuro.03.009.2009.

Flemming, Walther. 1882. *Zellsubstanz, Kern Und Zelltheilung*.

Flores, Ricardo, Selma Gago-Zachert, Pedro Serra, Rafael Sanjuán, and Santiago F Elena. 2014. "Viroids: Survivors from the RNA World?" JOUR. *Annual Review of Microbiology* 68 (1). Annual Reviews: 395–414. doi:10.1146/annurev-micro-091313-103416.

Francioli, Laurent C, Paz P Polak, Amnon Koren, Androniki Menelaou, Sung Chun, Ivo Renkens, Cornelia M van Duijn, et al. 2015. "Genome-Wide Patterns and Properties of de Novo Mutations in Humans." *Nature Genetics* 47 (7). Nature Publishing Group: 822–26. doi:10.1038/ng.3292.

Franklin, Rosalind E, and R G Gosling. 1953. "Molecular Configuration in Sodium Thymonucleate." JOUR. *Nature* 171 (4356): 740–41. http://dx.doi.org/10.1038/171740a0.

Freimer, Nelson B., and Chiara Sabatti. 2007. "Variants in Common Diseases." *Nature* 445 (February): 828–30. doi:10.1098/rsif.2006.0197.

Fujikura, K. 2015. "Multiple Loss-of-Function Variants of Taste Receptors in Modern Humans." *Scientific Reports* 5. Nature Publishing Group: 12349. doi:10.1038/srep12349.

Galton, Francis. 1870. "Experiments in Pangenesis, by Breeding from Rabbits of a Pure Variety, into Whose Circulation Blood Taken from Other Varieties Had Previously Been Largely Transfused." *Proceedings of the Royal Society of London* 19: 393–410.

"Genetic Code Circular." n.d. https://kaiserscience.files.wordpress.com/2015/01/lut.jpg?w=960.

Gerstein, Mark B., Can Bruce, Joel S. Rozowsky, Deyou Zheng, Jiang Du, Jan O. Korbel, Olof Emanuelsson, Zhengdong D. Zhang, Sherman Weissman, and Michael Snyder. 2007. "What Is a Gene, Post-ENCODE? History and Updated Definition." *Genome Research* 17 (6): 669–81. doi:10.1101/gr.6339607.

Gilbert, W. 1978. "Why Genes in Pieces?" *Nature*. doi:10.1038/271501a0.

Gilissen, Christian, Alexander Hoischen, Han G Brunner, and Joris a Veltman. 2011. "Unlocking Mendelian Disease Using Exome Sequencing." *Genome Biology* 12 (9): 228. doi:10.1186/gb-2011-12-9-228.

Giovannetti, Antonello, Marina Pierdominici, Francesca Mazzetta, Marco Marziali, Cristina Renzi, Anna Maria Mileo, Marco De Felice, et al. 2007. "Unravelling the Complexity of T Cell Abnormalities in Common Variable Immunodeficiency." *Journal of Immunology (Baltimore, Md.□: 1950)* 178 (6): 3932–43. http://www.ncbi.nlm.nih.gov/pubmed/17339494.

González-Pérez, Abel, and Nuria López-Bigas. 2011. "Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel." *American Journal of Human Genetics* 88 (4): 440–49. doi:10.1016/j.ajhg.2011.03.004.

Graur, Dan. 2003. "Single-Base Mutation," no. c: 287–90.

Grimbacher, Bodo, Andreas Hutloff, Michael Schlesier, Erik Glocker, Klaus Warnatz, Ruth Drager, Hermann Eibel, et al. 2003. "Homozygous Loss of ICOS Is Associated with Adult-Onset Common Variable Immunodeficiency." JOUR. *Nat Immunol* 4 (3): 261–68. http://dx.doi.org/10.1038/ni902.

Grimbacher, Bodo, Andreas Hutloff, Michael Schlesier, Erik Glocker, Klaus Warnatz, Ruth Dräger, Hermann Eibel, et al. 2003. “Homozygous Loss of ICOS Is Associated with Adult-Onset Common Variable Immunodeficiency.” *Nature Immunology* 4 (3): 261–68. doi:10.1038/ni902.

Groenen, Martien A M, Alan L Archibald, Hirohide Uenishi, Christopher K Tuggle, Yasuhiro Takeuchi, Max F Rothschild, Claire Rogel-Gaillard, et al. 2012. “Analyses of Pig Genomes Provide Insight into Porcine Demography and Evolution.” *Nature* 491 (7424). Nature Publishing Group: 393–98. doi:10.1038/nature11622.

Haeckel, Ernst. 1866. *Generelle Morphologie Der Organismen*. Berlin.

Hancks, Dustin C, and Haig H Kazazian. 2016. “Roles for Retrotransposon Insertions in Human Disease.” JOUR. *Mobile DNA* 7 (1): 1–28. doi:10.1186/s13100-016-0065-9.

Hart, G Traver, Arun K Ramani, and Edward M Marcotte. 2006. “How Complete Are Current Yeast and Human Protein-Interaction Networks?” JOUR. *Genome Biology* 7 (11): 1–9. doi:10.1186/gb-2006-7-11-120.

Hayes, Sheri, Beatrice Malacrida, Maeve Kiely, and Patrick A. Kiely. 2016. “Studying Protein–protein Interactions: Progress, Pitfalls and Solutions.” JOUR. *Biochemical Society Transactions* 44 (4): 994 LP-1004. http://www.biochemsoctrans.org/content/44/4/994.abstract.

Herfst, S, E J Schrauwen, M Linster, S Chutinimitkul, E de Wit, V J Munster, E M Sorrell, et al. 2012. “Airborne Transmission of Influenza A/H5N1 Virus between Ferrets.” *Science* 336

(6088): 1534–41. doi:336/6088/1534 [pii]\r10.1126/science.1213362 [doi].

Hernando-Herraez, Irene, Javier Prado-Martinez, Paras Garg, Marcos Fernandez-Callejo, Holger Heyn, Christina Hvilsom, Arcadi Navarro, Manel Esteller, Andrew J. Sharp, and Tomas Marques-Bonet. 2013. "Dynamics of DNA Methylation in Recent Human and Great Ape Evolution." *PLoS Genetics* 9 (9). doi:10.1371/journal.pgen.1003763.

Hoagland, M. B., M. L. Stephenson, J. F. Scott, L. I. Hecht, and P. C. Zamenick. 1958. "A Soluble Ribonucleic Acid Intermediate in Protein Synthesis." *The Journal of Biological Chemistry* 231 (1): 241–57. doi:10.1126/science.1184725.

Hodgkinson, Alan, and Adam Eyre-Walker. 2011. "Variation in the Mutation Rate across Mammalian Genomes." *Nature Reviews Genetics* 12 (11). Nature Publishing Group: 756–66. doi:10.1038/nrg3098.

Hubé, Florent, and Claire Francastel. 2015. "Mammalian Introns: When the Junk Generates Molecular Diversity." *International Journal of Molecular Sciences* 16 (3): 4429–52. doi:10.3390/ijms16034429.

Hunt, J, and V Ingram. 1958. "Allelomorphism and the Chemical Differences of the Human Haemoglobins A, S and C." *Nature* 181 (4615): 1062–63.

Ibarra-Laclette, Enrique, Eric Lyons, Gustavo Hernández-Guzmán, Claudia Anahí Pérez-Torres, Lorenzo Carretero-Paulet, Tien-Hao Chang, Tianying Lan, et al. 2013. "Architecture and Evolution of a Minute Plant Genome." *Nature* 498 (7452): 94–98. doi:10.1038/nature12132.

Imai, Masaki, Tokiko Watanabe, Masato Hatta, Subash C. Das, Makoto Ozawa, Kyoko Shinya, Gongxun Zhong, et al. 2012. "Experimental Adaptation of an Influenza H5 HA Confers Respiratory Droplet Transmission to a Reassortant H5 HA/H1N1 Virus in Ferrets." *Nature* 486 (7403). Nature Publishing Group: 420–28. doi:10.1038/nature10831.

Ingram, Vernon. 1956. "A Specific Chemical Difference Between the Globins of Normal Human and Sickle-Cell Anaemia Haemoglobin." *Nature*, no. 178: 792–94.

Jackson, Aimee L., and Lawrence A. Loeb. 2001. "The Contribution of Endogenous Sources of DNA Damage to the Multiple Mutations in Cancer." *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* 477 (1–2): 7–21. doi:10.1016/S0027-5107(01)00091-4.

Jacobson, Giselle N, and Patricia L Clark. 2016. "Quality over Quantity: Optimizing Co-Translational Protein Folding with Non-'optimal' Synonymous Codons." *Current Opinion in Structural Biology* 38. Elsevier Ltd: 102–10. doi:10.1016/j.sbi.2016.06.002.

Järvelä, I E, M K Salo, P Santavuori, and R K Salonen. 1993. "46,XX/69,XXX Diploid-Triploid Mixoploidy." *Journal of Medical Genetics* 30 (June): 966–67.

Jiang, Li, Ziping Chen, Qiuchen Gao, Lingkun Ci, Shuqing Cao, Yi Han, and Weiyan Wang. 2016. "Loss-of-Function Mutations in the APX1 Gene Result in Enhanced Selenium Tolerance in Arabidopsis Thaliana." JOUR. *Plant, Cell & Environment*, January, n/a-n/a. doi:10.1111/pce.12762.

Jin, Lv, Xiao-Yu Zuo, Wei-Yang Su, Xiao-Lei Zhao, Man-Qiong Yuan, Li-Zhen Han, Xiang Zhao, Ye-Da Chen, and Shao-Qi

Rao. 2014. “Pathway-Based Analysis Tools for Complex Diseases: A Review.” *Genomics, Proteomics & Bioinformatics* 12 (5). Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China: 210–20. doi:10.1016/j.gpb.2014.10.002.

Jones, P A, and D Takai. 2001. “The Role of DNA Methylation in Mammalian Epigenetics.” *Science* 293 (0036–8075 (Print)): 1068–70. doi:10.1126/science.1063852.

Kaiser, Jocelyn. 2010. “Affordable ‘ Exomes ’ Fill Gaps in a Catalog of Rare Diseases.” *Science*, no. August: 903. doi:10.1126/science.330.6006.903.

———. 2014. “The Hunt for Missing Genes.” *Science* 344 (6185): 687–89. doi:10.1126/science.344.6185.687.

Kaiser, Vera B., Victoria Svinti, James G. Prendergast, You Ying Chau, Archie Campbell, Inga Patarcic, Inês Barroso, et al. 2015. “Homozygous Loss-of-Function Variants in European Cosmopolitan and Isolate Populations.” *Human Molecular Genetics* 24 (19): 5464–74. doi:10.1093/hmg/ddv272.

Keller, Baerbel, Zoltan Cseresnyes, Ina Stumpf, Claudia Wehr, Manfred Fliegauf, Alla Bulashevska, Susanne Usadel, et al. 2016. “Disturbed Canonical NF-kB Signaling in B Cells of CVID Patients.” *Journal of Allergy and Clinical Immunology* 0 (0). Elsevier Ltd. doi:10.1016/J.JACI.2016.04.043.

Keskin, Ozlem, Nurcan Tuncbag, and Attila Gursoy. 2016. “Predicting Protein–Protein Interactions from the Molecular to the Proteome Level.” JOUR. *Chemical Reviews* 116 (8). American Chemical Society: 4884–4909. doi:10.1021/acs.chemrev.5b00683.

Khurana, Ekta, Yao Fu, Vincenza Colonna, Xinmeng Jasmine Mu, Hyun Min Kang, Tuuli Lappalainen, Andrea Sboner, et al. 2013. “Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics.” *Science (New York, N.Y.)* 342: 1235587. doi:10.1126/science.1235587.

Kilianski, Andy, Jennifer B. Nuzzo, and Kayvon Modjarrad. 2016. “Gain-of-Function Research and the Relevance to Clinical Practice.” *Journal of Infectious Diseases* 213 (9): 1364–69. doi:10.1093/infdis/jiv473.

Kim, Jong-Il, Young Seok Ju, Hansoo Park, Sheehyun Kim, Seonwook Lee, Jae-Hyuk Yi, Joann Mudge, et al. 2009. “A Highly Annotated Whole-Genome Sequence of a Korean Individual.” *Nature* 460 (7258). Nature Publishing Group: 1011–15. doi:10.1038/nature08211.

Kimchi-Sarfaty, Chava, Jung Mi Oh, In-Wha Kim, Zuben E Sauna, Anna Maria Calcagno, Suresh V Ambudkar, and Michael M Gottesman. 2007. “A ‘silent’ polymorphism in the MDR1 Gene Changes Substrate Specificity.” *Science* 315 (January): 525–28. doi:10.1126/science.1135308.

Kimura, M. 1968. “Evolutionary Rate at the Molecular Level.” *Nature* 217: 624–26. doi:10.1038/217624a0.

Kimura, Motoo. 1968. “Genetic Variability Maintained in a Finite Population due to Mutational Production of Neutral and Nearly Neutral Isoalleles.” *Genetic Research* 11: 247–69.

Kondrashov, Alexey S. 1995. “Contamination of the Genome by Very Slightly Deleterious Mutations: Why Have We Not Died 100 Times Over?” *Journal of Theoretical Biology* 175 (4): 583–94. doi:10.1006/jtbi.1995.0167.

König, Harald, Nathalie Matter, Rüdiger Bader, Wilko Thiele, and Ferenc Müller. 2007. “Splicing Segregation: The Minor Spliceosome Acts Outside the Nucleus and Controls Cell Proliferation.” *Cell* 131 (4): 718–29. doi:10.1016/j.cell.2007.09.043.

Koonin, Eugene V., Miklos Csuros, and Igor B. Rogozin. 2013. “Whence Genes in Pieces: Reconstruction of the Exon-Intron Gene Structures of the Last Eukaryotic Common Ancestor and Other Ancestral Eukaryotes.” *Wiley Interdisciplinary Reviews: RNA* 4 (1): 93–105. doi:10.1002/wrna.1143.

Koonin, Eugene V. 2012. “Does the Central Dogma Still Stand?” *Biology Direct* 7: 27. doi:10.1186/1745-6150-7-27.

Kuehn, Hye Sun, Bertrand Boisson, Charlotte Cunningham-Rundles, Janine Reichenbach, Asbjørg Stray-Pedersen, Erwin W Gelfand, Patrick Maffucci, et al. 2016. “Loss of B Cells in Patients with Heterozygous Mutations in IKAROS.” *The New England Journal of Medicine* 374 (11): 1032–43. doi:10.1056/NEJMoa1512234.

Kuehn, Hye Sun, Weiming Ouyang, Bernice Lo, Elissa K Deenick, Julie E Niemela, Danielle T Avery, Jean-Nicolas Schickel, et al. 2014. “Immune Dysregulation in Human Subjects with Heterozygous Germline Mutations in CTLA4.” *Science (New York, N.Y.)* 345 (6204): 1623–27. doi:10.1126/science.1255904.

Lander, E S, a Heaford, a Sheridan, L M Linton, B Birren, a Subramanian, a Coulson, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409 (6822): 860–921. doi:10.1038/35057062.

Laurent, Louise, Eleanor Wong, Guoliang Li, Tien Huynh, Aristotelis Tsirigos, Chin Thing Ong, Hwee Meng Low, et al. 2010. “Dynamic Changes in the Human Methylome during Differentiation.” *Genome Research* 20 (3): 320–31. doi:10.1101/gr.101907.109.

Lee, Young Ho, and Gwan Gyu Song. 2016. “Genome-Wide Pathway Analysis for Diabetic Nephropathy in Type 1 Diabetes.” JOUR. *Endocrine Research* 41 (1). Taylor & Francis: 21–27. doi:10.3109/07435800.2015.1044011.

Lek, Monkol, Konrad Karczewski, Eric Minikel, Kaitlin Samocha, Eric Banks, Timothy Fennell, Anne O'Donnell-Luria, et al. 2015. “Analysis of Protein-Coding Genetic Variation in 60,706 Humans.” JOUR. *bioRxiv*, October. http://biorxiv.org/content/early/2015/10/30/030338.abstract.

Li, Jin, Zhi Wei, Yun R. Li, S. Melkorka Maggadottir, Xiao Chang, Akshatha Desai, and Hakon Hakonarson. 2016. “Understanding the Genetic and Epigenetic Basis of Common Variable Immunodeficiency Disorder through Omics Approaches.” *Biochimica et Biophysica Acta (BBA) - General Subjects*. Elsevier B.V. doi:10.1016/j.bbagen.2016.06.014.

Liang, F, I Holt, G Pertea, S Karamycheva, S L Salzberg, and J Quackenbush. 2000. “Gene Index Analysis of the Human Genome Estimates Approximately 120,000 Genes.” *Nature Genetics* 25 (2): 239–40. doi:10.1038/76126.

Lindahl, T. 1993. “Instability and Decay of the Primary Structure of DNA.” *Nature* 362: 709–15. doi:10.1038/362709a0.

Lindahl, T, and R D Wood. 1999. “Quality Control by DNA Repair.” *Science* 286 (5446): 1897–1905. doi:10.1126/science.286.5446.1897.

Lister, Ryan, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, et al. 2009. "Human DNA Methylomes at Base Resolution Show Widespread Epigenomic Differences." *Nature* 462 (7271). Nature Publishing Group: 315–22. doi:10.1038/nature08514.

Liu, Binyan, Qizhen Xue, Yong Tang, Jia Cao, F. Peter Guengerich, and Huidong Zhang. 2016. "Mechanisms of Mutagenesis: DNA Replication in the Presence of DNA Damage." *Mutation Research/Reviews in Mutation Research* 768. Elsevier B.V.: 53–67. doi:10.1016/j.mrrev.2016.03.006.

Liu, Shuo, and Yinsheng Wang. 2015. "Mass Spectrometry for the Assessment of the Occurrence and Biological Consequences of DNA Adducts." *Chem. Soc. Rev.* 44. Royal Society of Chemistry: 7829–54. doi:10.1039/C5CS00316D.

Lo, Bernice, Kejian Zhang, Wei Lu, Lixin Zheng, Qian Zhang, Chrysi Kanellopoulou, Yu Zhang, et al. 2015. "Patients with LRBA Deficiency Show CTLA4 Loss and Immune Dysregulation Responsive to Abatacept Therapy."

Locke, Devin P, LaDeana W Hillier, Wesley C Warren, Kim C Worley, Lynne V Nazareth, Donna M Muzny, Shiaw-Pyng Yang, et al. 2011. "Comparative and Demographic Analysis of Orang-Utan Genomes." *Nature* 469 (7331): 529–33. doi:10.1038/nature09687.

Lopez-Herrera, Gabriela, Giacomo Tampella, Qiang Pan-Hammarström, Peer Herholz, Claudia M Trujillo-Vargas, Kanchan Phadwal, Anna Katharina Simon, et al. 2012. "Deleterious Mutations in LRBA Are Associated with a Syndrome of Immune Deficiency and Autoimmunity." *American Journal of Human Genetics* 90 (6): 986–1001. doi:10.1016/j.ajhg.2012.04.015.

Lupski, James R, Jeffrey G Reid, Claudia Gonzaga-Jauregui, David Rio Deiros, David C Y Chen, Lynne Nazareth, Matthew Bainbridge, et al. 2010. "Whole-Genome Sequencing in a Patient with Charcot-Marie-Tooth Neuropathy." *The New England Journal of Medicine* 362 (13): 1181–91. doi:10.1056/NEJMoa0908094.

Mable, B. K. 2004. "'Why Polyploidy Is Rarer in Animals than in Plants': Myths and Mechanisms." *Biological Journal of the Linnean Society*, no. 82: 453–56.

MacArthur, D G, J T Seto, J M Raftery, K G Quinlan, G A Huttley, J W Hook, F A Lemckert, et al. 2007. "Loss of ACTN3 Gene Function Alters Mouse Muscle Metabolism and Shows Evidence of Positive Selection in Humans." *Nat Genet* 39 (10): 1261–65. doi:ng2122 [pii]\r10.1038/ng2122.

MacArthur, Daniel G, Suganthi Balasubramanian, Adam Frankish, Ni Huang, James Morris, Klaudia Walter, Luke Jostins, et al. 2012. "A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes." *Science (New York, N.Y.)* 335 (6070): 823–28. doi:10.1126/science.1215040.

MacArthur, Daniel G, and Kathryn N North. 2007. "ACTN3: A Genetic Influence on Muscle Function and Athletic Performance." *Exercise and Sport Sciences Reviews* 35 (1): 30–34. doi:10.1097/JES.0b013e31802d8874.

Maffucci, Patrick, Charles A Filion, Bertrand Boisson, Yuval Itan, Lei Shang, Jean-laurent Casanova, and Charlotte Cunningham-rundles. 2016. "Genetic Diagnosis Using Whole Exome Sequencing in Common Variable Immunodeficiency." *Frontiers in Immunology*. doi:10.3389/fimmu.2016.00220.

Maggina, Paraskevi, and Andrew R Gennery. 2013. “Classification of Primary Immunodeficiencies: Need for a Revised Approach?” JOUR. *Journal of Allergy and Clinical Immunology* 131 (2): 292–94. doi:http://dx.doi.org/10.1016/j.jaci.2012.10.008.

Matthaei, J H, O W Jones, R G Martin, and M W Nirenberg. 1962. “Characteristics and Composition of RNA Coding Units.” *Proceedings of the National Academy of Sciences of the United States of America* 48: 666–77. doi:10.1073/pnas.48.4.666.

Maxam, a M, and W Gilbert. 1977. “A New Method for Sequencing DNA.” *Proceedings of the National Academy of Sciences of the United States of America* 74 (2): 560–64. doi:10.1073/pnas.74.2.560.

McClung, Clarence. 1899. “A Peculiar Nuclear Element in the Male Reproductive Cells of Insects.” *Zoological Bulletin* 2: 187.

———. 1902. “The Accessory Chromosome, Sex-Determinant.” *Biological Bulletin* 3: 74–75.

McKusick, Victor a. 2007. “Mendelian Inheritance in Man and Its Online Version, OMIM.” *American Journal of Human Genetics* 80 (4): 588–604. doi:10.1086/514346.

Mendel, Gregor. 1865. “Experiments in Plant Hybridization.” *Journal of the Royal Horticultural Society* IV (1865): 3–47. http://www.esp.org/foundations/genetics/classical/gm-65.pdf.

Michael, Todd P. 2014. “Plant Genome Size Variation: Bloating and Purging DNA.” *Briefings in Functional Genomics and Proteomics* 13 (4): 308–17. doi:10.1093/bfgp/elu005.

Mohun, T, D J Adams, R Baldock, S Bhattacharya, A J Copp, M Hemberger, C Houart, et al. 2013. “Deciphering the Mechanisms of Developmental Disorders (DMDD): A New Programme for Phenotyping Embryonic Lethal Mice.” *Dis Model Mech* 6 (3): 562–66. doi:10.1242/dmm.011957.

Moore, Jason H., and Scott M. Williams. 2009. “Epistasis and Its Implications for Personal Genetics.” *American Journal of Human Genetics* 85 (3). The American Society of Human Genetics: 309–20. doi:10.1016/j.ajhg.2009.08.006.

Morgan, T H. 1908. “Sex-Determining Factors in Animals.” *Science* 25: 382– 384.

Mort, Matthew, Dobril Ivanov, David N. Cooper, and Nadia A. Chuzhanova. 2008. “A Meta-Analysis of Nonsense Mutations Causing Human Genetic Disease.” *Human Mutation* 29 (8): 1037–47. doi:10.1002/humu.20763.

Morton, N E, J F Crow, and H J Muller. 1956. “An Estimate of the Mutational Damage in Man From Data on Consanguineous Marriages.” *Proceedings of the National Academy of Sciences of the United States of America* 42 (11): 855–63. doi:10.1073/pnas.42.11.855.

Ng, Bernard, Fan Yang, David P. Huston, Yan Yan, Yu Yang, Zeyu Xiong, Leif E. Peterson, Hong Wang, and Xiao Feng Yang. 2004. “Increased Noncanonical Splicing of Autoantigen Transcripts Provides the Structural Basis for Expression of Untolerized Epitopes.” *Journal of Allergy and Clinical Immunology* 114 (6): 1463–70. doi:10.1016/j.jaci.2004.09.006.

Ng, Pauline C., and Steven Henikoff. 2001. “Predicting Deleterious Amino Acid Substitutions.” *Genome Research* 11 (5): 863–74. doi:10.1101/gr.176601.

Ng, Pauline C., Samuel Levy, Jiaqi Huang, Timothy B. Stockwell, Brian P. Walenz, Kelvin Li, Nelson Axelrod, Dana A. Busam, Robert L. Strausberg, and J. Craig Venter. 2008. "Genetic Variation in an Individual Human Exome." *PLoS Genetics* 4 (8). doi:10.1371/journal.pgen.1000160.

Ng, Sarah B, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, et al. 2010. "Exome Sequencing Identifies the Cause of a Mendelian Disorder." *Nature Genetics* 42 (1). Nature Publishing Group: 30–35. doi:10.1038/ng.499.

Nguyen, Dinh Truong, Kyooyeol Lee, Hojun Choi, Min-kyeung Choi, Minh Thong Le, Ning Song, Jin-Hoi Kim, et al. 2012. "The Complete Swine Olfactory Subgenome: Expansion of the Olfactory Gene Repertoire in the Pig Genome." JOUR. *BMC Genomics* 13 (1): 1–12. doi:10.1186/1471-2164-13-584.

Nirenberg, Marshall, and Philip Leder. 1964. "RNA Codewords and Protein Synthesis." *Science* 145 (3639): 1399–1407. doi:10.1073/pnas.53.5.1161.

Nirenberg, Marshall W., and J.H. H Matthaei. 1961. "The Dependence of Cell-Free Protein Synthesis in E. Coli upon Naturally Occurring or Synthetic Polyribonucleotides." *Proceedings of the National Academy of Sciences of the United States of America* 47 (10): 1588–1602. doi:10.1021/ja7114579.

Nuevo, Michel, Stefanie N Milam, and Scott a Sandford. 2012. "Nucleobases and Prebiotic Molecules in Organic Residues Produced from the Ultraviolet Photo-Irradiation of Pyrimidine in NH(3) and H(2)O+NH(3) Ices." *Astrobiology* 12 (4): 295–314. doi:10.1089/ast.2011.0726.

Nuismer, Scott L, and Sarah P Otto. 2004. “Host-Parasite Interactions and the Evolution of Ploidy.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (30): 11036–39. doi:10.1073/pnas.0403151101.

O’Grady, Julian J., Barry W. Brook, David H. Reed, Jonathan D. Ballou, David W. Tonkyn, and Richard Frankham. 2006. “Realistic Levels of Inbreeding Depression Strongly Affect Extinction Risk in Wild Populations.” *Biological Conservation* 133 (1): 42–51. doi:10.1016/j.biocon.2006.05.016.

Ohta, T, and Jh Gillespie. 1996. “Development of Neutral and Nearly Neutral Theories.” *Theoretical Population Biology* 49 (2): 128–42. doi:10.1006/tpbi.1996.0007.

Pan-Hammarström, Qiang, Emanuela Castigli, Stephen Wilson, Lilit Garibyan, Rima Rachid, Francisco Bonilla, Lynda Schneider, Massimo Morra, John Curran, and Raif Geha. 2007. “Reexamining the Role of TACI Coding Variants in Common Variable Immunodeficiency and Selective IgA Deficiency.” *Nature Genetics* 39 (4): 430–31.

Park, Miguel A, James T Li, John B Hagan, Daniel E Maddox, and Roshini S Abraham. 2009. “Common Variable Immunodeficiency: A New Look at an Old Disease.” JOUR. *The Lancet* 372 (August): 489–502. doi:http://dx.doi.org/10.1016/S0140-6736(08)61199-X.

Pelak, Kimberly, Kevin V. Shianna, Dongliang Ge, Jessica M. Maia, Mingfu Zhu, Jason P. Smith, Elizabeth T. Cirulli, et al. 2010. “The Characterization of Twenty Sequenced Human Genomes.” *PLoS Genetics* 6 (9). doi:10.1371/journal.pgen.1001111.

Pertea, Mihaela, and Steven L Salzberg. 2010. “Between a Chicken and a Grape: Estimating the Number of Human Genes.” *Genome Biology* 11 (5): 206. doi:10.1186/gb-2010-11-5-206.

Porta-Pardo, Eduard, Luz Garcia-Alonso, Thomas Hrabe, Joaquin Dopazo, and Adam Godzik. 2015. “A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces.” JOUR. *PLoS Comput Biol* 11 (10). Public Library of Science: e1004518. http://dx.doi.org/10.1371%2Fjournal.pcbi.1004518.

Porta-Pardo, Eduard, and Adam Godzik. 2014. “E-Driver: A Novel Method to Identify Protein Regions Driving Cancer.” JOUR. *Bioinformatics* , July. doi:10.1093/bioinformatics/btu499.

Prado-Martinez, Javier, Peter H Sudmant, Jeffrey M Kidd, Heng Li, Joanna L Kelley, Belen Lorente-Galdos, Krishna R Veeramah, et al. 2013. “Great Ape Genetic Diversity and Population History.” *Nature* 499 (7459): 471–75. doi:10.1038/nature12228.

Quaynor, Samuel D, Maggie E Bosley, Christina G Duckworth, Kelsey R Porter, Soo-Hyun Kim, Hyung-Goo Kim, Lynn P Chorich, et al. 2016. “Targeted next Generation Sequencing Approach Identifies Eighteen New Candidate Genes in Normosmic Hypogonadotropic Hypogonadism and Kallmann Syndrome.” JOUR. *Molecular and Cellular Endocrinology* 437 (December): 86–96. doi:http://dx.doi.org/10.1016/j.mce.2016.08.007.

Raje, Nikita, Sarah Soden, Douglas Swanson, Christina E Ciaccio, Stephen F Kingsmore, and Darrell L Dinwiddie. 2014. “Utility of next Generation Sequencing in Clinical Primary Immunodeficiencies.” *Current Allergy and Asthma Reports* 14 (10): 468. doi:10.1007/s11882-014-0468-y.

Ralls, Katherine, Jonathan D. Ballou, and Alan Templeton. 1988. "Estimates of Lethal Equivalents and the Cost of Inbreeding in Mammals." *Conservation Biology* 2 (2): 185–93. doi:10.1111/j.1523-1739.1988.tb00169.x.

Ramensky, Vasily, Peer Bork, and Shamil Sunyaev. 2002. "Human Non-Synonymous SNPs: Server and Survey." *Nucleic Acids Research* 30 (17): 3894–3900. doi:10.1093/nar/gkf493.

Rayan, Nirmala Arul, Ricardo C H del Rosario, and Shyam Prabhakar. 2016. "Massive Contribution of Transposable Elements to Mammalian Regulatory Sequences." JOUR. *Seminars in Cell & Developmental Biology* 57 (September): 51–56. doi:http://dx.doi.org/10.1016/j.semcdb.2016.05.004.

Revel-Vilk, Shoshana, Ute Fischer, B??rbel Keller, Schafiq Nabhani, Laura G??mez-D??az, Anne Rensing-Ehl, Michael Gombert, et al. 2015. "Autoimmune Lymphoproliferative Syndrome-like Disease in Patients with LRBA Mutation." *Clinical Immunology* 159 (1). Elsevier Inc.: 84–92. doi:10.1016/j.clim.2015.04.007.

Rivas, Manuel A, Matti Pirinen, Donald F Conrad, Monkol Lek, Emily K Tsang, Konrad J Karczewski, Julian B Maller, et al. 2015. "Effect of Predicted Protein-Truncating Genetic Variants on the Human Transcriptome." JOUR. Edited by Ayellet V Young Segre Taylor R. Gelfand, Ellen T. Trowbridge, Casandra A. Ward, Lucas D. Kheradpour, Pouya Iriarte, Benjamin Meng, Yan Palmer, Cameron D. Esko, Tonu Winckler, Wendy Hirschhorn, Joel Kellis, Manolis Getz, Gad Shablin, Andrey A. Li, Gen Zhou, Yi-Hui Nobel,. *Science* 348 (6235): 666 LP-669. http://science.sciencemag.org/content/348/6235/666.abstract.

Rodríguez-Cortez, Virginia C., Lucia del Pino-Molina, Javier Rodríguez-Ubreva, Laura Ciudad, David Gómez-Cabrero, Carlos Company, José M. Urquiza, et al. 2015. "Monozygotic Twins Discordant for Common Variable Immunodeficiency Reveal Impaired DNA Demethylation during Naïve-to-Memory B-Cell Transition." *Nature Communications* 6: 7335. doi:10.1038/ncomms8335.

Ropers, H. Hilger, and Thomas Wienker. 2015. "Penetrance of Pathogenic Mutations in Haploinsufficient Genes for Intellectual Disability and Related Disorders." *European Journal of Medical Genetics* 58 (12): 715–18. doi:10.1016/j.ejmg.2015.10.007.

Saikia, Biman, and Sudhir Gupta. 2016. "Common Variable Immunodeficiency." JOUR. *The Indian Journal of Pediatrics* 83 (4): 338–44. doi:10.1007/s12098-016-2038-x.

Salzer, Ulrich, Chiara Bacchelli, Sylvie Buckridge, Qiang Pan-Hammarström, Stephanie Jennings, Vassilis Lougaris, Astrid Bergbreiter, et al. 2009. "Relevance of Biallelic versus Monoallelic TNFRSF13B Mutations in Distinguishing Disease-Causing from Risk-Increasing TNFRSF13B Variants in Antibody Deficiency Syndromes." *Blood* 113 (9): 1967–76. doi:10.1182/blood-2008-02-141937.

Salzer, Ulrich, H M Chapel, a D B Webster, Q Pan-Hammarström, A Schmitt-Graeff, M Schlesier, H H Peter, et al. 2005. "Mutations in TNFRSF13B Encoding TACI Are Associated with Common Variable Immunodeficiency in Humans." *Nature Genetics* 37 (8): 820–28. doi:10.1038/ng1600.

Salzer, Ulrich, Andrea Maul-Pavicic, Charlotte Cunningham-Rundles, Simon Urschel, Bernd H Belohradsky, Jiri Litzman, Are Holm, et al. 2004. "ICOS Deficiency in Patients with

Common Variable Immunodeficiency." *Clinical Immunology (Orlando, Fla.)* 113 (3): 234–40. doi:10.1016/j.clim.2004.07.002.

Sanger, F, S Nicklen, and a R Coulson. 1977. "DNA Sequencing with Chain-Terminating Inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5463–67. doi:10.1073/pnas.74.12.5463.

Sauna, Zuben E, and Chava Kimchi-Sarfaty. 2011. "Understanding the Contribution of Synonymous Mutations to Human Disease." *Nature Reviews. Genetics* 12 (10). Nature Publishing Group: 683–91. doi:10.1038/nrg3051.

Scally, Aylwyn, Julien Y Dutheil, LaDeana W Hillier, Gregory E Jordan, Ian Goodhead, Javier Herrero, Asger Hobolth, et al. 2012. "Insights into Hominid Evolution from the Gorilla Genome Sequence." *Nature* 483 (7388): 169–75. doi:10.1038/nature10842.

Schneider, Friederich Anton. 1873. "Untersuchungen Über Plathelminthen." Bulletin. *Vierzehnter Bericht Der Oberhess*.

Schopf, J W. 1993. "Microfossils of the Early Archean Apex Chert: New Evidence of the Antiquity of Life." *Science (New York, N.Y.)* 260: 640–46. doi:10.1126/science.260.5108.640.

Schopf, J William. 2006. "Fossil Evidence of Archaean Life." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 361: 869–85. doi:10.1098/rstb.2006.1834.

Schubert, Desirée, Claudia Bode, Rupert Kenefeck, Tie Zheng Hou, James B Wing, Alan Kennedy, Alla Bulashevska, et al. 2014. "Autosomal Dominant Immune Dysregulation Syndrome in

Humans with CTLA4 Mutations." *Nature Medicine* 20 (12): 1410–16. doi:10.1038/nm.3746.

Schwartz, James. 2008. *In Pursuit of the Gene. From Darwin to DNA.* Harvard University Press.

Ségurel, Laure, Minyoung J. Wyman, and Molly Przeworski. 2014. "Determinants of Mutation Rate Variation in the Human Germline." *Annual Review of Genomics and Human Genetics* 15 (1): 47–70. doi:10.1146/annurev-genom-031714-125740.

Seidel, Markus G., Tatjana Hirschmugl, Laura Gamez-Diaz, Wolfgang Schwinger, Nina Serwas, Andrea Deutschmann, Gregor Gorkiewicz, et al. 2015. "Long-Term Remission after Allogeneic Hematopoietic Stem Cell Transplantation in LPS-Responsive Beige-like Anchor (LRBA) Deficiency." *Journal of Allergy and Clinical Immunology* 135 (5): 1384–1390e8. doi:10.1016/j.jaci.2014.10.048.

Sequencing, The Chimpanzee, and Analysis Consortium. 2005. "Initial Sequence of the Chimpanzee Genome and Comparison with the Human Genome." *Nature* 437 (7055): 69–87. doi:10.1038/nature04072.

Serafino, Loris. 2016. "Abiogenesis as a Theoretical Challenge: Some Reflections." *Journal of Theoretical Biology* 402: 18–20. doi:10.1016/j.jtbi.2016.04.033.

Serwas, Nina Kathrin, Aydan Kansu, Elisangela Santos-Valente, Zarife Kuloğlu, Arzu Demir, Aytaç Yaman, Laura Yaneth Gamez Diaz, et al. 2015. "Atypical Manifestation of LRBA Deficiency with Predominant IBD-like Phenotype." *Inflammatory Bowel Diseases* 21 (1): 40–47. doi:10.1097/MIB.0000000000000266.

Shabalina, Svetlana A., Nikolay A. Spiridonov, and Anna Kashina. 2013. “Sounds of Silence: Synonymous Nucleotides as a Key to Biological Regulation and Complexity.” *Nucleic Acids Research* 41 (4): 2073–94. doi:10.1093/nar/gks1205.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. “Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks.” *Genome Research*, no. Karp 2001: 2498–2504. doi:10.1101/gr.1239303.

Sleiman, Patrick, Jonathan Bradfield, Frank Mentch, Berta Almoguera, John Connolly, and Hakon Hakonarson. 2014. “Assessing the Functional Consequence of Loss of Function Variants Using Electronic Medical Record and Large-Scale Genomics Consortium Efforts.” *Frontiers in Genetics* 5 (April): 105. doi:10.3389/fgene.2014.00105.

Song, Gwan Gyu, Sang-Cheol Bae, and Young Ho Lee. 2013. “Pathway Analysis of Genome-Wide Association Studies on Rheumatoid Arthritis.” *Clinical and Experimental Rheumatology* 31 (4): 566–74.

Song, Gwan Gyu, and Young Ho Lee. 2013. “Pathway Analysis of Genome-Wide Association Studies for Parkinson’s Disease.” JOUR. *Molecular Biology Reports* 40 (3): 2599–2607. doi:10.1007/s11033-012-2346-9.

Spence, J E, R G Perciaccante, G M Greig, H F Willard, D H Ledbetter, J F Hejtmancik, M S Pollack, W E O’Brien, and A L Beaudet. 1988. “Uniparental Disomy as a Mechanism for Human Genetic Disease.” *American Journal of Human*

Genetics 42 (2): 217–26. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1715272&tool=pmcentrez&rendertype=abstract\nhttp://www.ncbi.nlm.nih.gov/pubmed/2893543\nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1715272.

Stevens, Nettie. 1905a. "A Study of the Germ Cells of Aphis Rosae and Aphis Oenotherae." *Journal for Experimental Zoology* 2: 313–33.

———. 1905b. "Studies in Spermatogenesis with Especial Reference to the Accessory Chromosome." *Carnegie Institution Publications*.

Stitziel, Nathan O, Adam Kiezun, and Shamil Sunyaev. 2011. "Computational and Statistical Approaches to Analyzing Variants Identified by Exome Sequencing." *Genome Biology* 12 (9): 227. doi:10.1186/gb-2011-12-9-227.

Sturtevant, Alfred. 1913. "The Linear Arrangement of Six Sex-Linked Factors in Drosophila, as Shown by Their Mode of Association." *Journal of Experimental Biology* 14: 45.

Sundaram, Vasavi, Yong Cheng, Zhihai Ma, Daofeng Li, Xiaoyun Xing, Peter Edge, Michael P Snyder, and Ting Wang. 2014. "Widespread Contribution of Transposable Elements to the Innovation of Gene Regulatory Networks." JOUR. *Genome Research* 24 (12). Cold Spring Harbor Laboratory Press: 1963–76. doi:10.1101/gr.168872.113.

Sunyaev, S R, F Eisenhaber, I V Rodchenkov, B Eisenhaber, V G Tumanyan, and E N Kuznetsov. 1999. "PSIC: Profile Extraction from Sequence Alignments with Position-Specific Counts of Independent Observations." *Protein Engineering* 12 (5): 387–94. doi:10.1093/protein/12.5.387.

Sunyaev, S, V Ramensky, I Koch, W Lathe 3rd, A S Kondrashov, and P Bork. 2001. “Prediction of Deleterious Human Alleles.” *Hum Mol Genet* 10 (6): 591–97. doi:10.1093/hmg/10.6.591.

Sutton, Walter. 1903. “The Chromosomes in Heredity.” *Biological Bulletin* 4: 231–51.

Swart, Estienne C., John R. Bracht, Vincent Magrini, Patrick Minx, Xiao Chen, Yi Zhou, Jaspreet S. Khurana, et al. 2013. “The Oxytricha Trifallax Macronuclear Genome: A Complex Eukaryotic Genome with 16,000 Tiny Chromosomes.” *PLoS Biology* 11 (1). doi:10.1371/journal.pbio.1001473.

Szklarczyk, Damian, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, et al. 2015. “STRING v10: Protein–protein Interaction Networks, Integrated over the Tree of Life.” JOUR. *Nucleic Acids Research* 43 (D1): D447–52. doi:10.1093/nar/gku1003.

Tarrant, Daniel, and Tobias Von Der Haar. 2014. “Synonymous Codons, Ribosome Speed, and Eukaryotic Gene Expression Regulation.” *Cellular and Molecular Life Sciences* 71 (21): 4195–4206. doi:10.1007/s00018-014-1684-2.

Thomas, Duncan. 2010. “Gene--Environment-Wide Association Studies: Emerging Approaches.” *Nature Reviews. Genetics* 11 (4): 259–72. doi:10.1038/nrg2764.

Tschermak, Erich. 1900. “Über Künstsliche Kreuzung Bei Pisum Sativum.” *Berichte Der Deutschen Botanischen Gesellschaft* 18: 232–39.

Van Beneden, Edouard. 1883. “Recherches Sur La Maturation de L’oeuf et La Fécondation.” *Archive de Biologie* 4: 265–640.

van Schouwenburg, Pauline A., Emma E. Davenport, Anne Kathrin Kienzler, Ishita Marwah, Benjamin Wright, Mary Lucas, Tomas Malinauskas, et al. 2015a. “Application of Whole Genome and RNA Sequencing to Investigate the Genomic Landscape of Common Variable Immunodeficiency Disorders.” *Clinical Immunology* 160 (2). Elsevier B.V.: 301–14. doi:10.1016/j.clim.2015.05.020.

van Schouwenburg, Pauline A, Emma E Davenport, Anne-Kathrin Kienzler, Ishita Marwah, Benjamin Wright, Mary Lucas, Tomas Malinauskas, et al. 2015b. “Application of Whole Genome and RNA Sequencing to Investigate the Genomic Landscape of Common Variable Immunodeficiency Disorders.” JOUR. *Clinical Immunology* 160 (2): 301–14. doi:http://dx.doi.org/10.1016/j.clim.2015.05.020.

Venter, J C, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, et al. 2001. “The Sequence of the Human Genome.” *Science* 291 (5507): 1304–51. doi:10.1126/science.1058040.

Watson, J. D.; Crick, Francis. 1953. “Molecular Structure of Nucleic Acids.” *Nature*. doi:10.1038/171737a0.

Weiss, Madeline C, Filipa L Sousa, Natalia Mrnjavac, Sinje Neukirchen, Mayo Roettger, Shijulal Nelson-Sathi, and William F Martin. 2016. “The Physiology and Habitat of the Last Universal Common Ancestor.” JOUR. *Nature Microbiology* 1 (July). Macmillan Publishers Limited: 16116. http://dx.doi.org/10.1038/nmicrobiol.2016.116.

Wen, Ya, Mohamad J Alshikho, and Martha R Herbert. 2016. “Pathway Network Analyses for Autism Reveal Multisystem Involvement, Major Overlaps with Other Diseases and Convergence upon MAPK and Calcium Signaling.” JOUR.

PLoS ONE 11 (4). Public Library of Science: e0153329. http://dx.doi.org/10.1371%2Fjournal.pone.0153329.

Wilkins, M H F, A R Stokes, and H R Wilson. 1953. "Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids." JOUR. *Nature* 171 (4356): 738–40. http://dx.doi.org/10.1038/171738a0.

Wilson, Edmund. 1905. "Studies on Chromosomes: 1. The Behavior of the Idiochromosomes in Hemiptera." *Journal for Experimental Zoology* 2: 371–405.

———. 1906. "Studies on Chromosomes: 3. The Sexual Differences of the Chromosome Groups in Hemiptera, with Some Considerations of the Determination and Inheritance of Sex with Six Figures." *Journal for Experimental Zoology* 3.

Wilson Sayres, Melissa A., and Kateryna D. Makova. 2011. "Genome Analyses Substantiate Male Mutation Bias in Many Species." *BioEssays* 33 (12): 938–45. doi:10.1002/bies.201100091.

Wong, Tiffany, Joanne Yeung, Kyla J Hildebrand, Anne K Junker, and Stuart E Turvey. 2013. "Human Primary Immunodeficiencies Causing Defects in Innate Immunity." *Current Opinion in Allergy and Clinical Immunology* 13 (6): 607–13. doi:10.1097/ACI.0000000000000010.

Worthington Allen, Frank. 1941. "The Biochemistry of the Nucleic Acids, Purines and Pyrimidines." *Annual Review of Biochemistry* 10: 221–45. doi:10.1146/annurev.bi.10.070141.001253.

Wu, R, and E Taylor. 1971. "Nucleotide Sequence Analysis of DNA." *Nature* 236: 198–201. doi:10.1038/10.1038/newbio236198a0.

Wutke, Saskia, Leif Andersson, Norbert Benecke, Edson Sandoval-Castellanos, Javier Gonzalez, Jón Hallsteinn Hallsson, Lembi Lõugas, et al. 2016. “The Origin of Ambling Horses.” *Current Biology* 26 (15): R697–99. doi:10.1016/j.cub.2016.07.001.

Yang, Nan, Daniel G MacArthur, Jason P Gulbin, Allan G Hahn, Alan H Beggs, Simon Easteal, and Kathryn North. 2003. “ACTN3 Genotype Is Associated with Human Elite Athletic Performance.” *American Journal of Human Genetics* 73 (3): 627–31. doi:10.1086/377590.