Universidade do Minho
Escola de Engenharia

Daniel Eduardo Fernandes Martins

Characterization of genetic variants in 70 Portuguese individuals

Dissertação de Mestrado

Mestrado em Bioinformática

Trabalho efetuado sob a orientação de

Doutora Maria da Conceição Venâncio Egas

Mestre Hugo Jorge Calisto Froufe

Professor Doutor Rui Manuel Ribeiro Castro Mendes

Outubro de 2018

DECLARAÇÃO

Nome: Daniel Eduardo Fernandes Martins

Endereço eletrónico: daniel_5_martins@hotmail.com Telefone: 915828230

Bilhete de Identidade/Cartão do Cidadão: 14519869

Título da dissertação: Characterization of genetic variants in 70 Portuguese individuals.

Orientadores:

Doutora Maria da Conceição Venâncio Egas

Mestre Hugo Jorge Calisto Froufe

Professor Doutor Rui Manuel Ribeiro Castro Mendes

Ano de conclusão: 2018

Mestrado em Bioinformática

É AUTORIZADA A REPRODUÇÃO INTEGRAL DESTA DISSERTAÇÃO APENAS PARA EFEITOS DE INVESTIGAÇÃO, MEDIANTE DECLARAÇÃO ESCRITA DO INTERESSADO, QUE A TAL SE COMPROMETE.

Universidade do Minho, 31/10/2018

Assinatura: Daniel Eduardo Fernandes Martins

AGRADECIMENTOS

Antes de proceder à apresentação do trabalho realizado, aproveito este espaço para manifestar os devidos agradecimentos pela ajuda que se revelou essencial para a conclusão desta dissertação.

Começo por agradecer à Dra. Conceição Egas, pela oportunidade que me deu de desenvolver este trabalho com a equipa que coordena na Unidade de Sequenciação Avançada da Biocant. Agradeço igualmente a todos os membros dessa mesma equipa, pela forma como me receberam e acolheram no dia-a-dia da Unidade. Entre eles, agradeço principalmente ao Hugo, não só pela ajuda imprescindível que representou ao longo de todo o meu trabalho, como também por todo o tempo de convívio que tornou os dias de trabalho muito mais fáceis e no qual, sem dúvida, aprendi imenso.

Deixo também o meu agradecimento ao professor Rui Mendes pela sua disponibilidade enquanto orientador na Universidade do Minho e pela prontidão e celeridade da sua ajuda.

Apesar de não haver mais pessoas envolvidas diretamente no trabalho, este apenas pôde ser realizado graças ao apoio de muita gente a quem me tenho de dirigir nas seguintes linhas.

Assim, agradeço ainda à Beatriz, à Marta e ao Fernandes, pela próxima companhia que já se prolonga há cinco anos e por terem sido um apoio indispensável durante todo o mestrado, que estas linhas simbolizem as quinze páginas de agradecimentos que sempre disse que mereceriam da minha parte.

Devo também um grande agradecimento à minha secção do Hipermercado Pingo Doce de Aveiro, principalmente à Andreia e ao Fábio, com quem partilhei tantas horas de trabalho, companhia e risadas, e à Cristina, que fez sempre tudo o que pôde para me facilitar a gestão do meu tempo.

Fica ainda o agradecimento ao meu Clã, do Agrupamento de Escuteiros 970 da Palhaça, pela compreensão, pela flexibilidade, e acima de tudo, pela fraternidade que se manteve mesmo após longos períodos sem contacto.

Por último, os maiores e mais importantes agradecimentos, aos meus pais, ao meu irmão e à minha namorada, pelo suporte emocional que representam, pela paciência, por terem sido a minha principal motivação e por terem estado sempre disponíveis para me ouvir quando precisei.

A todos, o meu mais sincero “obrigado”.

RESUMO

A análise genómica das populações tem contribuído significativamente para o aumento do número de SNVs descritos em bases de dados. Estudos populacionais prévios têm contribuído com 18 a 57% novas variantes. A nova informação genética é particularmente relevante enquanto referência para propósitos clínicos. Iniciativas à escala global como o 1000 Genomes Project (1kG) incluem populações Ibéricas, contudo, nenhum indivíduo Português foi incluído no mesmo grupo. Tanto quanto se sabe, nenhum indivíduo Português foi incluído no projeto gnomAD, o maior conjunto de dados genómicos atual.

Acreditamos que uma coleção de informação genómica referente à população Portuguesa poderia trazer grandes benefícios ao diagnóstico molecular em pacientes Portugueses.

As alterações genéticas detetadas em 70 indivíduos Portugueses foram inseridas em uma base de dados não-relacional. A informação publicada pelos projetos 1kG e gnomAD para cada alteração incluída nas mesmas foi adicionada à referida base de dados. Frequências alélicas reportadas para sete populações incluídas na base de dados do gnomAD, cinco populações do 1kG e 5 subpopulações Europeias do mesmo projeto foram comparadas contra os valores calculados para os nossos dados. As diferenças das distribuições alélicas foram testadas com o Fisher's Exact test. Os p-values obtidos foram corrigidos de acordo com a sua False Discovery Rate (FDR).

Os exomas de indivíduos Portugueses analisados continham 224,155 alterações genéticas filtradas de acordo com critérios de qualidade definidos no presente estudo. Aproximadamente 16,4% das variantes não se encontravam descritas nas bases de dados dos projetos 1kG e gnomAD. Os resultados obtidos endossam evidências, previamente descritas na literatura, de uma correlação entre as diferenças genéticas das populações comparadas em relação à população Portuguesa e a distância geográfica das mesmas a Portugal. Diferenças significativas entre distribuições alélicas da população estudada e outras subpopulações Europeias foram encontradas para 7,284 alterações genéticas distribuídas por 2,571 genes. Os resultados obtidos sugerem a existência de marcadores genéticos populacionais e podem motivar futuros estudos com vista a detetar marcadores genéticos específicos da população Portuguesa.

O estudo apresentado representa uma contribuição significativa para, não só enriquecer iniciativas genómicas de grande escala, mas também para estabelecer uma referência auxiliar para análises genéticas a doentes Portugueses.

Este trabalho foi efetuado no âmbito do projeto In2Genome, ref. CENTRO-01-0247-FEDER-017800, apoiado pelo Programa Operacional Regional do Centro de Portugal (CENTRO 2020), ao abrigo do Acordo de Parceria Portugal 2020, através do Comité Regional Europeu Fundo de Desenvolvimento (FEDER).

Palavras-chave: Genómica, Alterações genéticas, Exomas, Distribuição alélica, População.

ABSTRACT

The in-depth study of the genomics of single populations has contributed significantly to the enlargement of known SNVs in databases. Each single population study has contributed with 18 to 57% of novel SNVs. The new genetic information is particularly relevant as a reference for clinical purposes. Global-scale initiatives as the 1000 Genomes Project (1kG) already include Iberian population; however, no Portuguese individuals were included in this cohort. Furthermore, to our knowledge, gnomAD, the most extensive genomic dataset, does not include Portuguese individuals either.

We believe that a Portuguese collection of genomic information would greatly benefit molecular diagnosis in Portuguese patients.

Variants detected in 70 Portuguese individuals were inserted in a MongoDB No-SQL Database. The 1kG and gnomAD information for each variant were uploaded to the same database. Allele frequencies for seven gnomAD populations, five 1kG populations, and five 1kG European subpopulations were compared to the values calculated for our data. Allele distribution differences were tested with Fisher's exact test. P-values were corrected for False Discovery Rate (FDR).

The exomes of the Portuguese individuals contained 224,155 variants filtered accordingly to defined quality criteria. Approximately 16.4% of the variants had not been previously reported by 1kG or gnomAD projects. The present work endorsed the evidence for a correlation between genetic and geographic distance previously reported in the literature. Finally, significative differences were found for the allele distribution between our population and the other 1kG European subpopulations in 7,284 variants distributed by 2,571 genes. Results suggest the existence of populational genetic markers and may prompt future studies for detection of Portuguese-specific genetic markers.

The present study is a significant contribution to enrich large-scale genomic initiatives and, to stand as a useful auxiliary reference for genetic analyses of Portuguese patients.

This work was supported by the In2Genome project CENTRO-01-0247-FEDER-017800, supported by Centro Portugal Regional Operational Programme (CENTRO 2020), under the Portugal 2020 Partnership Agreement, through the European Regional Development Fund (ERDF).

Keywords: Genomics, Variants, Exomes, Allele distribution, Population.

ÍNDICE

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Motivation

Nowadays, it is possible to re-sequence countless human exomes from all over the world aiding several working groups to connect genetic variants to clinical conditions. Nevertheless, large-scale genomic projects suggest that allele frequencies, which are highly relevant for clinical purposes, differ considerably across different populations [1], genetic diversity hinder specific conclusions regarding single populations from global-scale studies.

The populational genetic identity has significant relevance; once genes are responsible for biological traits, different populations may present different susceptibilities to certain conditions or display higher risk to develop a given disease depending on the carried genetic information. Said so, the disclosure of specific characteristics of a population might be very helpful to foresight and diagnose genetic diseases on its individuals. It may also result in a robust reference to find and develop new drugs since it can evidence population-specific resistances or adverse effects, for example, by identifying modified binding sites.

Although global-scale initiatives have already comprised Iberian population [2], no Portuguese individuals were included as all samples were obtained from Spanish individuals. Another study has denoted genetic proximity between Portuguese individuals and a Central Spanish sub-population but some differentiation to Eastern Spanish individuals [3] which disables a generalization of results obtained on Spanish samples to the Portuguese population. Moreover, this genetic similarity has not yet been characterized at the functional level hence it is not possible to assume similar disease risks for both populations.

Single-population level studies [1,4,5] have been performed and successfully provided a genetic reference for clinical purposes. Identifying a Portuguese genetic paradigm may be useful to constitute a reference for future studies.

1.2 Objectives

The main goal of the presented work is to identify Portuguese population-specific genetic characteristics and describe its outcomes. To accomplish this purpose, progressive objectives may be defined as follow.

1) Analyse previous initiatives, the methods and information needed to perform a similar study and understand the scientific context required to extract relevant conclusions.

2) Search for useful data that may be included on populational analyses, either from large-scale studies or from the previous initiatives analysed.
3) Organize genetic data in a scalable structure to aid analytical processing and enable the compilation of information from multiple sources.
4) Test genetic similarity between Portuguese individuals and other populations already described.
5) Search on Portuguese population for genetic alterations with significantly different frequencies in comparison with other populations.
6) Link Portuguese-characterising variants to its biologic effects and health consequences or benefits.

2. BIBLIOGRAPHIC REVISION

2.1 Human Diversity

The study of population-specific genetic variants requires the perception of how populations diverge on a genetic point of view, the biological fundament of genetic alteration and its consequences. Furthermore, these studies involve the methods to identify those alterations, the results obtained with that process, the way to store data and finally, the tools available to classify the alteration and predict its effects.

Among human beings, it is possible to notice differences between persons with the unaided eye; besides that, human physiology reveals some degree of differentiation across populations.[6] These differences are expressions of genetic diversity. Besides identical twins, no two human individuals share identical genetic information.[7] That genetic constitution of any human organism is called its genotype, which produces any observable trait constituting a phenotype.[8]

2.1.1 Biological context

Genetic data is encoded by a four-letter alphabet; adenosine (A), cytosine (C), guanine (G) or thymine (T) residues may be found on each position of the DNA sequence. Genetic terminology defines that position as *locus* (*loci* in plural), every *locus* is a template where an allele resides, and an allele is considered as the genetic information contained on a *locus*.[9]

Any information required by an organism to perform a biological process is stored on its DNA sequence. Since the conception of a new living being, every cell on the organism carry that information and transmits it to its "daughter cells" created by one of two division processes. Somatic cells divide by mitosis. On this event, DNA molecules are replicated into sister chromatids and each chromatid is segregated into separate nuclei, producing two new diploid cells. Meiosis division occurs in germline cells, the main difference to mitosis resides on the first round of cell division after DNA replication, where homologous chromosomes recombine. This first division separates recombined chromosomes and the latter segregates complementary chromatids to separate nuclei, producing four haploid cells in the end. Since each chromatid strand acts as a template for its complementary sequence, the result of the replication process is two DNA molecules almost equal to the original DNA chain. The differences to the original chain usually arise from unrepaired DNA damage, errors on replication or interference by mobile genetic elements. On each replication process, approximately one on every 10^9 nucleotides is altered, generating

a mutation.[10] Those may be point mutations, that is, small insertions or deletions, or substitutions of a single nucleotide. Inversions, translocations, deletions or insertions of regions of the chromosome with a variable size may result from chromosome recombination in meiosis.[11]

Mutation and *polymorphism* are two terms that could often be used afore this point, however, both are usually confused and incorrectly used. A mutation is intended as any permanent change in the nucleotide sequence, polymorphism is a genetic variation found on a population at an abundance meaningfully high to be caused by random events, nevertheless, on a genomic perspective, both "mutation" and "polymorphism" can be replaced by the term "variant".[12]

The most common variants are single-nucleotide polymorphisms (SNP), it accounts for about 90% of sequence alterations.[13] SNPs are *loci* at which different alleles may be found in a population in distinguishable abundances. Minor Allele Frequency (MAF) is an important measure on population genetics, MAF is the frequency value of the second most represented allele on a population [14] and may be used to infer how heterozygous is a population. To define a variant as a SNP, the MAF value threshold has been defined at 1%.[11,15] That is, two or more alleles shall be present on the population with a frequency of at least 1%.

SNPs arise from mutations [16] that occur in a germline cell of a common ancestor and may disperse across the population over time. A germline mutation is inherited by the offspring conceived by the mutated gamete then, that specific individual bears the variant as a somatic one and may transmit it to its progeny.[11] Over generations, a mutation may be transmitted to progressively more individuals until the sub-population of individuals presenting that specific mutation make up to 1% of the population.

One of the first endeavours to catalogue all single nucleotide polymorphism information in one platform was carried out by NCBI. dbSNP[17] was created as a repository for submitted SNP candidates, data is filtered, and a record is generated to be available in the database. Approximately a year after its creation, dbSNP accounted for 1.4 million submissions from 97 registered groups, data was referent to five species, among them, humans.[18]

Nowadays, it is the largest polymorphism database, counting over 900 million submissions for *Homo sapiens*.[19]

Henceforward, the term SNV (single nucleotide variant) will be used to refer to variants on a single *locus* independently of its population frequency, it must be distinguished from the SNP concept.

SNVs may either cause functional differences when they occur on coding or regulatory regions or be harmless if located elsewhere.[13]

The genetic information undergoes a process that ultimately results in a protein, those are then the effectors of biological processes, that flow of information is stated on the central dogma of molecular biology [20]. DNA encodes information that is transcribed to polynucleotide chains of RNA, which in turn are translated into proteins.

The translation process is based on codons, which are portions of a sequence with three nucleotides, each codon is read at a time by the ribosome to create and elongate the forming peptide chain. Each amino acid is added in agreement with a genetic code (Figure 1).[10] The genetic code consists of the correlation of the 64 possible combinations of the three nucleotides in the codon with the 20 natural amino acids (Figure 1). One of those codons, AUG, is responsible for the starting position of the nascent polypeptide, corresponding to the inclusion of a methionine. The three codons, UAA, UAG and UGA are represented on the genetic code as stop codons, they do not code for any amino acid and provide a signal for termination of the translation process.[21]

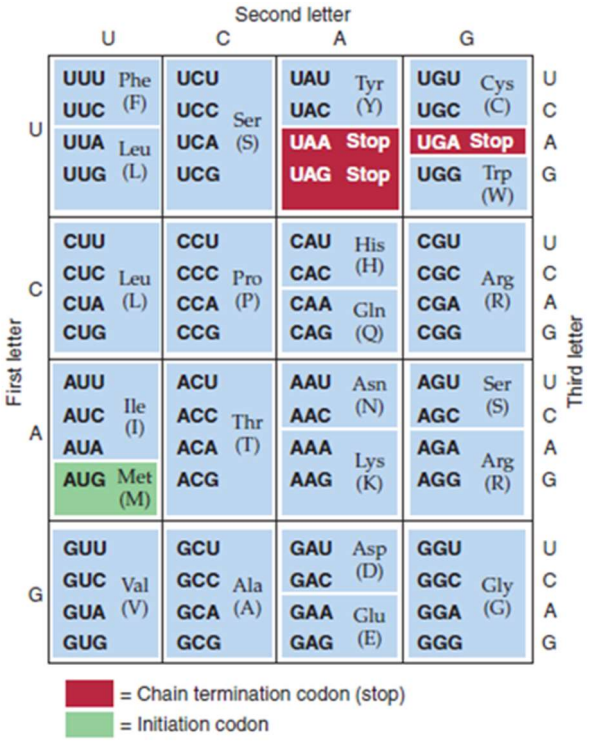


Figure 1. Codon table representing the genetic code. Adapted from iGenetics [8]

The genetic code is redundant; some codons may code for the same amino acid. When a SNV encodes for the same amino acid despite changing the codon, it is called synonymous (or silent) (Figure 2. a) and does not affect the protein. On the other hand, a nonsynonymous SNV (or missense) (Figure 2. b) do alter the coded amino acid and it may have diverse effects.[16]

A nonsynonymous SNV may constitute a neutral variant (Figure 2. c) if it replaces the amino acid with another that has similar chemical proteins, a loss-of-function variant, which eliminates the normal function

of a protein, a hypomorphic variant, if it reduces the normal function of a protein and a gain of function variant, which may increase the protein normal function. If a SNV creates a termination codon in the middle of a coding region, the variant is called nonsense (or stop gain) and results in a truncated protein (Figure 2. d). On the other hand, a variant that changes a stop codon to any other codon that translates into an amino acid is called a stop loss variant, the resulting protein may be longer and may lose its function or may be affected by structural alterations.[22]

Apart from SNVs, the most common polymorphisms are small insertions or deletions (INDELs) ranging from 1 to 10,000 base pairs.[23] INDELs may be either insertions and deletions of single base pairs, transposon insertions, INDELs of random DNA sequences or sequence expansions of small repeat units.[24]

INDELs may have many different effects on coding regions, insertions or deletions of any number of base pairs not divisible by three causes a frameshift. A frameshift variant changes the reading frame of a mRNA sequence downstream to the variant *locus*, so the translation process adds incorrect amino acids to the peptide sequence. That variant usually generates a non-functional protein.[8]

The modification of the reading frame may cause a bypass of the normal stop codon (stop loss variant) resulting on a protein longer than it should be or may generate a new stop codon resulting in a shortened polypeptide. On coding sequences, INDELs of a length divisible by three may also cause alterations on protein functions, either by deletion of important amino acids or length changes.[12]

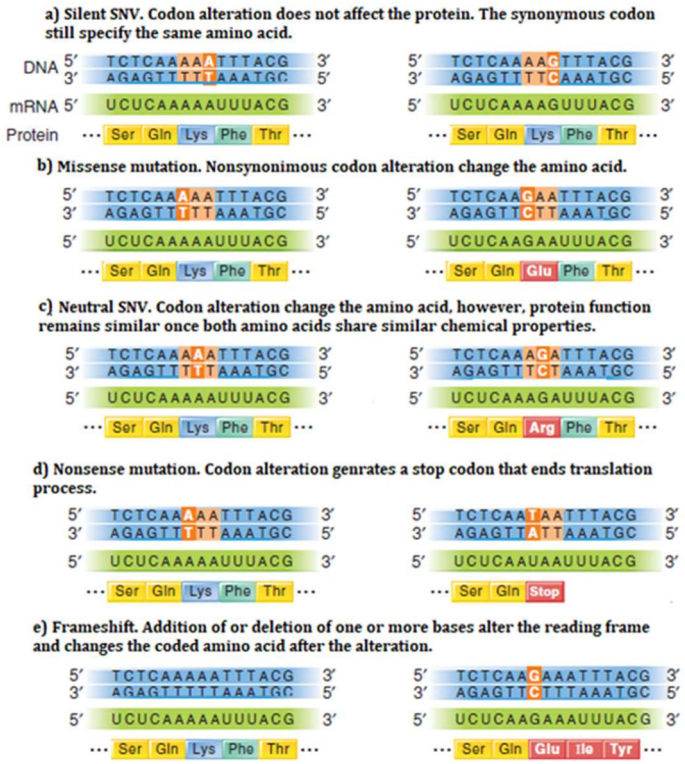


Figure 2. Types of base pair substitutions and frameshift examples. Adapted from iGenetics.[8]

At the end of the translation event, the DNA sequence information has been converted into a protein.[25] Reached this point, information cannot be transferred back from proteins [20], so the effects of genetic alterations are definitive, an altered protein cannot be corrected and may play a different role on organism physiology.

Those differences are responsible for the diverse characteristics (or traits) presented by the individuals of a population. [26] Alterations may also stand as the onset of diseases or health conditions. The term "susceptibility *locus*" is used to point out that an alteration on a specific *locus* may increase the risk but is neither necessary nor sufficient for disease expression. Variants on susceptibility *loci* may be taken as indicators of a greater propensity of an individual or population to develop any given disease or disorder with a genetical burden. [27–29] Susceptibility *loci* may also indirectly affect health conditions by being associated with eventual risk factors like blood lipids. [30]

2.1.2 Hardy-Weinberg Law and Genetic Drift

Understanding human evolution, requires knowledge and interpretation of the genetic profile of any given population.

The Mendelian principle of segregation states that any allele A has the same probability to be present on a gamete, so the offspring of two heterozygous progenitors has 25, 50 and 25% of chance to present respectively the A1A1, A1A2 and A2A2 genotypes. Based on that principle, the Hardy-Weinberg Law (HWL) is the groundwork for any evolution genetic study.

According to HWL, being, respectively, p and q the allele frequencies for A1 and A2, the genotypic frequencies appear at a proportion where $p^2 + 2pq + q^2 = 1$.[8]

However, the Hardy-Weinberg model builds on various assumptions, some of them concordant to the human biologic condition. The organism must be diploid and have sexual reproduction. The law is only applied to di-allelic genes ($p + q = 1$) and allele frequencies are assumed to be identical in males and females (only applies to genes on autosomal chromosomes). Finally, HWL does not take in account any selective pressure, random mating is assumed among individuals on very large population (assumedly infinite), migration and mutations events are ignored, and natural selection does not affect the allele frequencies under consideration.[7]

HWL provides a good approximation to real proportion values found on populations but, the influence of any of those ignored factors deviates frequency ratios from the predicted values. Besides that, the impossibility to comply with the assumption of infinite populations generates a genetic drift. That is, random changes in allele frequencies caused by a different number of descendants of individuals. On a

finite population, it leads to a frequency rise on some alleles over others across generations. Alleles of individuals who generate a bigger progeny will be perpetuated on the population over time. It leads to a dispersive pressure that vanishes genetic variation. The smaller the population, the bigger is the allele extinction rate.[9]

A gene pool is constituted by all unique alleles carried by the individuals of a population at a given moment. If the Hardy-Weinberg Equilibrium is not perturbed the individuals of the next generation can only inherit its progenitor alleles, therefore, the gene pool will be maintained, and the genetic diversity of that population preserved as it is.

However, HWL assumptions may be contravened and consequently this diversity may suffer alterations over time by evolution events as mutations, selective pressures or migrations and interaction between populations.[8]

2.1.3 Natural selection

It has been made clear that a mutation may alter various aspects of organism physiology. The consequence of an alteration ranges from prejudicial or fatal malfunctions to gainful traits regarding ability, health or any other aspect. On several cases, altered features affect fitness to a given environment, so, adaptation to the environment dictates whether a mutation is harmful, neutral or advantageous to the organism.

Greater suitability of some traits over others leads to the evolution of species [31] by natural selection. The organism traits which increase its chance of surviving or reproducing are perpetuated on the population by genetic drift, bearers of those traits naturally have more chances to generate a greater progeny which may carry the genetic information in charge to manifest the characteristic.[8,9]

2.1.4 Migration and Gene Flow

Associated with the factors approached at this point, there is an important phenomenon to the historical evolution of the human species, migration.

Migration has been a determinant aspect in settlement of genetic differences over geographically scattered populations.

Influence of new populations on a given location goes beyond culture or habits; migrations may introduce new alleles on the prevalent gene pool, thus altering allele frequencies and disturbing Hardy-Weinberg equilibrium.

The incorporation of new alleles into a population is called gene flow. The importance of gene flow on population genetics resides on the propagation of variants, a variant that arises on a given population allows to trace back population course over history, that variant may characterize descendant generations and allows to identify contacts with other populations on which the variant is found.[8]

Besides population interaction, migration is the basis of the founder effect, another major event on population genetics. The founder effect consists of the detachment of a small group of individuals (sub-population) from an established population. The detachment itself is pointed out as a bottleneck. It results in a noteworthy genetic drift; the new population experiences a significant loss in heterozygosity since it represents a "sample" of the original gene pool. Bottlenecks may also be the consequence of an event that drastically reduces population size, like natural disasters, disease outbreaks or population isolation on specific locations.[7]

Nowadays, population genetics studies had revealed that the modern-day patterns of genomic variation are the consequence of several major demographic events in the past.[32] Comparison of genomic data from ancient and modern populations provide a great amount of information, it may provide an approximation about the time of divergence of two given populations, unveil more complex genealogical relationships and facilitate determination of migration routes, founder effects and genetic admixture among various groups.[33]

2.1.5 Genetic ancestry of the Portuguese population

Since the first global level analysis of genetic variation, both human origin in Sub-Saharan Africa and the out-of-Africa model of population expansion had been mainly accepted.[33,34] Additionally, the correlation between geographic and genetic proximity has been strongly evidenced; the relationship extends to support the deduced migration routes and historic relations.[35,36]

Although the earliest fossil evidence of *Homo Sapiens* existence comes from East Africa – estimated to be 150 to 190 thousand years (kyr) old [37,38] – there is no consensus on pinpointing origin within the continent, there are evidence that supports either East African [39] or west to central origin [34].

The great genetic diversity among actual African sub-populations in comparison to the rest of the world also supports African human origin.[35] Several ancestral population clusters can be identified and a bigger divergence on Sub-Saharan populations is denoted.[40,41] This data is consistent with theoretical population genetics explained before. Subsets of an initial population may have migrated to distinct parts of the continent and evolved independently, the genetic differences would be more easily explained by the eventual small size of the founder groups.

From that diverse population, East African subsets of individuals began the population expansion to the Middle East instituting the first out-of-Africa event (Figure 3). The gene pool of the founder group undergone a major genetic drift. Then, all non-African human populations descended from that less diverse group.[34,36]

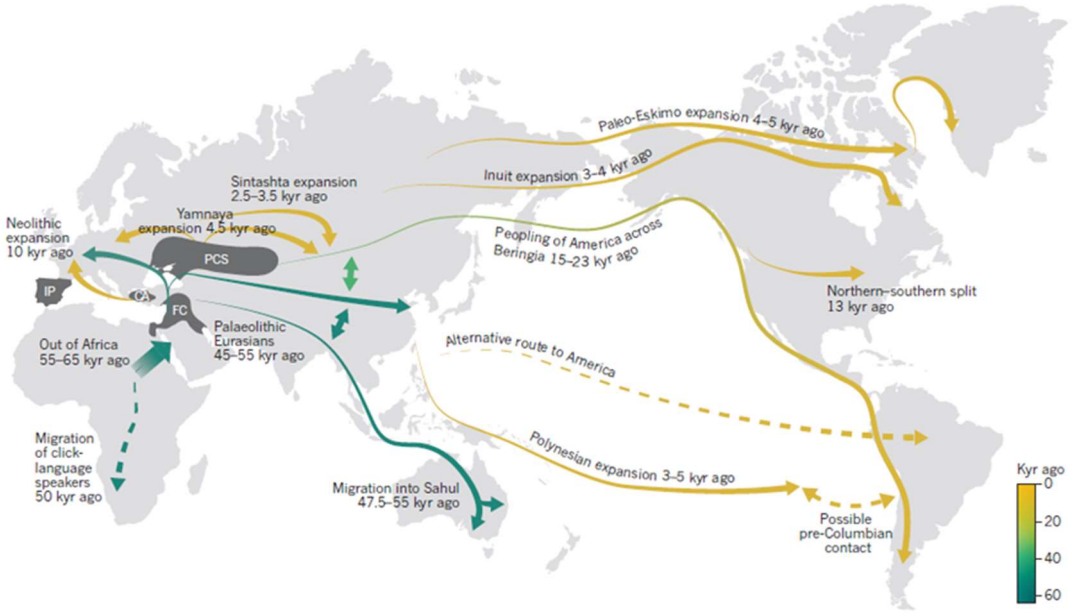


Figure 3. Major human migrations over the world. IP – Iberian Peninsula; CA – Central Anatolia, present-day Turkey; FC – Fertile Crescent, present-day Middle East; PCS – Pontic-Caspian Steppe. [33]

The earliest *Homo sapiens* evidence in the Middle East region is dated from over 100 thousand years ago (kya).[42] However, the population split that originated modern Eurasian population is estimated to have happened between 50 to 75 kya [43,44]. It is currently supposed that a single major migration group was on the origin of any other population worldwide, that is supported by the evidence of an equivalent effect on non-African populations of an encounter and admixture with Neanderthals shortly after humans leaving Africa.[45] All Non-African individuals analysed on this aspect until this moment bears roughly 2% Neanderthal ancestry.[46,47]

The arrival to Europe happened around 45 kya, which is the age of the earlier evidence of anatomically modern human presence.[48] However, the genetic contributions of these first inhabitants to present day Europeans is still under debate.[32]

Several individuals studied by Fu *et al.*[45] seem to share the same ancestry and do not display substantial differentiation evidence, it indicates the presence of a single population lineage in Europe from 37 kya until 14 kya. However, the contribution to modern-day Europeans comes from a subset of the original population. During the Last Glacial Maximum (LGM) – between 26 and 19 kya – northern Europe

was covered with Ice, it led to human migrations into southern refugia, causing a severe bottleneck that reduced the genomic diversity of later European populations. [49,50]

After the end of LGM, Europe was recolonized by a group named West European hunter-gatherers (WHG). Those individuals contribute ancestry to every European population and none to modern-day Middle Easterners. Indicating that this population never reached the Middle East.[46]

WHG individuals had been found over northern Iberia.[45,51,52] Nevertheless, the trace of that population in actual Europe is located mainly in northern Europe, individuals from that region bear up to 50% WHG ancestry.[46,53]

On early Neolithic, Early European farmers (EEF), a new population from Anatolia (modern day Turkey) expanded into Europe. There is evidences that indicates that this group arrived in Iberia roughly 7-6 kya[52], and influenced the gene pool of the local population.[32,54]

EEF presented 44% ancestry from a "basal Eurasian" population from Near East and spread over the continent founding a new lifestyle and setting themselves on given locations. [33] It was essential to the demographic growth verified by that time.[50] Nowadays, EEF genetic influence is more relevant in southern Europe.[46]

Finally, Ancient North Eurasians (ANE) the ancient population from which modern-day Siberians descended. Although it has been reported that by the Neolithic transition, EEF and WHG were the two main ancestral populations in Europe and there are no indications of the presence of ANE in central and western Europe on that period, today, its ancestry is found in nearly all Europeans at a percentage that reaches up to 20%.[46]

Both WHG populations who may not have migrated to northern Europe – Martiniano *et al.*[55] suggests a prolonged WHG interaction at European Atlantic littoral – and EEF populations size growth have a significant impact on the genetic composition of modern-day southwestern European populations. Both groups admixture [50] and natural increase in genetic variation of growing populations [56] led to a higher level of genetic diversity compared to most northern regions.[57]

A third influential factor arose more recently. In the Iberian Peninsula, North African invasions began in the 8th century of the common era and lasted until the 13th century. This occupation embodies a gene flow event that modified Iberian gene pool. This region has the biggest North African ancestry among European populations.[57,58]

On the present day, it is suggested that in Europe, even geographically distant individuals share a common ancestry from the past 3,000 years.[59] Nevertheless, a clear correspondence between genetic and geographic distances might be denoted, Iberia could be demarked from the rest of Europe by the

genetic characteristics of its population.[60] A continental-level study performed by Lao *et al.*[3] that includes Portuguese subpopulations denotes a close relation of that group to a Central Spanish sub-population, some differentiation to Eastern Spanish individuals and suggests some proximity with Italian individuals.

However, Iberian-Italian clusters relations may not be mostly caused by shared ancestry but rather by scarce similarities of both populations with other European regions. Both Portuguese and Spanish individuals show very low common ancestry with the rest of Europe but a high relation within them. The low rate of shared ancestry may be explained by the aforementioned North African occupation and possibly by geographic isolation due to the Pyrenees that may have difficulted the contact with neighbouring populations.[59]

2.1.6 Global-scale population studies

As it has been presented, the divergence between different populations results in a vast diversity of the global population. Data collection from sufficient independent populations enables large-scale analyses to deepen the comprehension of human population genetics. It triggered the development of various initiatives like the HapMap Project that aimed to characterize the variants by its frequencies and correlations between them. DNA samples of 270 individuals were obtained from 4 populations, including 90 samples from the Utah (United States of America) population with Northern and Western European ancestry.[61] HapMap has been discontinued by 2016.[62]

1000 Genome Project. On this regard, the 1000 Genome Project is one of the biggest endeavours to create a base for further genomic studies. It was launched in 2008 to create a public reference database for DNA polymorphism.

The main goal of the project is assumed to be the search on the "accessible genome" and characterize over 95% of genetic variants which presents an allele frequency of 1% or higher in each of five major population groups, which are the descendants from populations of ancient Europe, East Asia, South Asia, West Africa and the Americas. Clustered sampling was employed to ameliorate the detection capability of low frequency variants. That is, in a cluster of related populations, genetic drift may provide variants with a higher frequency than may be found on other populations that present an overall low frequency for that variant. Hence, that variant is easily detectable.[63] On the sampling process, not more than 100 unrelated individuals were sampled from related populations.[2]

On a first phase, the 1000 Genome Project performed whole-genome sequencing on low coverage, array-based genotyping and targeted sequencing of some coding regions for 1,092 individuals sampled from

14 populations. The design proved to be powerful and cost-effective on the task to search for almost every variant.[63]

For each variant, there was considered information about mapping quality, the quality of the reads and the distribution of variant calls in the population. Variant sites were ranked accordingly to results from machine-learning approaches using the quality information, this enabled to establish thresholds to ensure low False Discovery rates.

Additionally, genotype likelihoods were used to assess the evidence for each genotype found at bi-allelic sites, at every site in each sample it was inspected if there were 0, 1 or 2 copies of the variant.

On overall, the project discovered and genotyped 38 million SNVs, 1.4 million bi-allelic indels, and 14 thousand large deletions. The results were filtered and validated to avoid inconsistent or ambiguous data. This study estimates to have detected 1% frequency SNVs with a certainty of 99.3% over the genome and 99.8% on the exomic region. Moreover, the power to detect 0.1% frequency variants is nearly 70% on the genome and 90% across the exome. From the reported variants, 6% of the variants with frequencies over 5% were not known, 38% of variants in the range from 0.5 to 5% had never been described previously, as well as 87% variants with frequencies under 0.5%.

Among the results obtained by the Project, some findings may be highlighted, variants with a frequency above 10% are almost all found in all populations studied, at the other hand, 53% of variants at 0.5% or below were observed only in a single population. Allele frequency distributions show that African ancestry populations carry up to three times as many variants with low frequency (in the range of 0.5 to 5%) as the populations of European or East Asian ancestry, supporting ancestral bottlenecks in the origin of non-African populations.

Regarding the effect of the variant, at the most highly conserved coding sites, 85% of the nonsynonymous variants and 90% of nonsense and splice-disrupting variants are rare (frequency below 0.5%). Only 65% synonymous variants present that scarce frequency.[2]

By 2015, the project was completed, using the same approach as reported earlier, it reconstructed, in total, 2,504 individual genomes from 26 populations, finding 88 million variants distributed on 84.7 million SNVs, 3.6 million short indels and 60 thousand structural variations. It accounts for over 99% of the known SNPs with a frequency of 1% or above. On this phase, multi-allelic events were also analysed, expanding the restriction to bi-allelic polymorphisms that was previously imposed.

Almost ¾ variants reported are rare (frequency below 0.5%), more precisely, 64 million variants. Besides those, there are 12 million variants that present a frequency between 0.5 and 5% and approximately 8

million variants have a frequency above 5%. However, every single genome carries just 40 to 200 thousand rare variants, which represent 1 to 4% of the genome variants.

The final project report stated that a typical genome differs from the reference genome at a range between 4.1 and 5 million *loci*. It also supported the out-of-Africa model by finding that individuals with African ancestry carried the greatest number of variant sites.[64]

On present day, the 1,000 Genome Project contains 107 Iberian samples, all collected on Spanish territory from individuals who were born in Spain as well as their direct relatives on the last two generations.[65]

ExAC. The Exome Aggregation Consortium had a more ambitious initiative by calling variants from the exomes of 60,706 individuals.[66] Six ancestry groups were defined; AFR included African and African American individuals, AMR represent South America, EAS East Asian, SAS South Asian, FIN Finnish and NFE Non-Finnish individuals from Europe. The specific territories where the sampled individuals lived were not provided.

The study identified over 10 million variant candidates. After quality filters, a subset of approximately 7.4 million high-quality variants was defined. From that subset, over 317 thousand variants are indels. Those results correspond to one variant for every eight base pairs. Almost all of the high-quality variants, 99%, have a frequency below 1%, 54% are seen only once in the whole dataset and 72% were absent from the data sets of other projects.

ExAC has been built to serve as a support for medical genetic analysis and a scope to study the effect of different variants on human physiology. However, the dataset itself is among the most complete collections of exomic data.[67]

Although still available as a stand-alone browser, ExAC has recently been complemented with genomic data hence constituting gnomAD (Genome Aggregation Database) [68], the current dataset spans 123.136 exomes and 15,496 genomes from unrelated individuals.

2.1.7 Single-population genetic studies

Regarding diversity between populations, studies have been performed either to characterize the genetic proximity of isolated groups to other populations or to assess population predisposition to certain diseases or conditions based on populational genetics. Those searched for SNVs on susceptibility *loci* or any other variants that might affect health conditions of the individuals that are comprised of the population.

The comparison with 1,000 Genomes Project data is widely recurrent in recent studies, Zlobin *et al.*[69] examined 12 exomes from Yakuts, a secluded Siberian indigenous population, and searched for

similarities with populations described on the project. In this study, 746,396 variants were called. 56,949 of those were absent from variant databases. Finally, it succeeded to identify Yakuts as a genetically secluded population.

Einhorn *et al.*[70] also succeeded to find Ashkenazi Jewish specific variants. One hundred and twenty-eight individuals from that population were sampled, and the 1,000 Genome Projects populations from Europe, Africa, East Asia and South Asia were taken as a control group, one hundred and twenty-eight individuals were randomly selected from each population. The authors found 222,179 SNVs, of which 18.3% were not present in population genetics databases and 30.6% were present on only one individual.

Moreover, several studies addressed allele frequencies of specific variants to determine the propensity of populations to express a condition or the influence of a gene to a given clinical effect. Zhou *et al.*[71] integrated ExAC data from 56945 individuals and linkage information from the 1000 Genome Project to derive the frequencies of 176 alleles, representing 12 cytochrome P450 genes with the highest relevance for human drug metabolism. Their study establishes an overview for cytochrome p450 gene allele distributions on five major populations (Europeans, Africans, South Asians, East Asians, and admixed Americans). Slavin *et al.*[72] sampled 2,134 women with familial breast cancer. They identified 2,859 variants in 26 known or proposed breast cancer susceptibility genes, then, to detail the spectrum of susceptibility genes and its correlation with clinical outcomes and refine risk estimates for specific mutations, they analysed the results using the frequencies of 9,647 variants associated to those genes on the non-Finnish European group control from ExAC.

With a broader focus, the UK10K project was designed to characterize rare and low frequency variants from the United Kingdom population to assess the contributions of genetic variation to diverse biomedically relevant traits and diseases within the population.[4] A similar initiative took place in the Netherlands, there, 250 families, totalizing 769 individuals, were sampled.[5]

Dopazo *et al.*[1] collected samples from 267 healthy Spanish individuals and downloaded exomic data of 13 populations from the 1,000 Genomes Project in a VCF file. The selected populations represented European, Asian, American and African descendants. The final dataset totaled 1,359 individuals.

This study reported 170,888 variant positions. Almost one-third of the variants found were not described on public repositories. They were found, on average, approximately 19,000 variants per individual, 9,194 of those were nonsynonymous and 85 % were found in only one individual. The main goal of the study was to describe disease-related variants. Around 3,000 variants were present on disease databases. The authors then compared the MAFs of those variants for the 267 analysed individuals with the MAFs presented by the other populations included in the study, 193 variants were two times more frequent in

Spanish populations than in the 1,000 Genome Project populations, 69 of those variants still had a MAF 2-fold higher when concerning only European populations.

The study yielded useful conclusions, for example, it found significantly higher (between 4 and 18-fold) variant frequencies for various conditions like Marfan syndrome, Von Willebrand syndrome, Ellis-van Creveld syndrome, Wilson disease, cystinuria, Crohn's disease, or Charcot-Marie-Tooth disease among others. Additionally, the study identified drug binding sites which have a higher probability to be found affected on Spanish populations, such as the gene CYP11B2 from the Cytochrome P450 family, it is affected by a nonsynonymous alteration with 15-fold higher prevalence on the Spanish population [1]

On the sequence of this study, the Collaborative Spanish Variant Server (CSVS) was developed to store and supply variability information based on a bigger dataset with 1582 unrelated Spanish individuals. VCF files aggregated on the server are enriched with annotation data from NoSQL CellBase Database.[73]

Regarding what has been said until this point, the significant genetic similarity between Portuguese and Spanish populations is expected. However, Portuguese individuals samples have not been included in global scale endeavours, therefore, it is not possible to compare both populations to assess eventual differences, Dopazo *et al* results cannot be generalized to the Iberian Peninsula population, therefore, it is conceivable that different susceptibilities and risks may be found, once both populations may manifest different predisposition to health disorders or different responses to drugs or treatments, it is important to beacon the differences between both populations.

2.2 Technologic Context

So far, it has been presented why it is important to know the nucleotide order on the sequence, for that, the polynucleotide chain must be sequenced. It then allows identifying genetic variants [74] by mapping reads against a reference genome and identifying the differences between both sequences.[75]

Through the last 40 years, sequencing technologies had been developed and improved in order to perform, faster, cheaper and more accurate processes.[76]

On the 1970s, there were two predominant and influential protocols. Coulson and Sanger's "plus and minus" system relied on radiolabelled nucleotides and two different alternated reactions. The "minus" reaction, uses three different nucleotides to produce sequences until the missing nucleotide is required, in the "plus" reaction, that nucleotide is provided so all extensions will end with that base.[77] That method takes several polymerization reactions and neither the "plus" nor the "minus" reactions were completely accurate.[78]

2.2.1 First generation

The other protocol developed by Maxam and Gilbert on 1977 consists of sequencing by chemical degradation. To determine a nucleotide sequence, the DNA molecules were radiolabelled, cleaved and separated into four aliquots containing, respectively, fragments cleaved on adenine (A), adenine and guanine (A+G), cytosine (C) and cytosine and thymine (C+T) residues. On both protocols, fragments were separated by its length on polyacrylamide gels where the nucleotide order could be inferred (Figure. 4). [79,80]

Maxam and Gilbert method was the first widely used technique, standing as the initial first-generation DNA sequencing protocol. Nonetheless, on the year of 1977, Frederick Sanger and colleagues developed the chain-termination method, the first "sequencing-by-synthesis" process, meaning that it requires the direct action of DNA polymerase to produce an observable output. It would become one of the benchmarks of DNA sequencing processes (Figure 4). [78]

Sanger method improved accuracy, robustness and ease of the sequencing process by using dideoxynucleosides (ddNTP), chemical analogues of the DNA functional units that lack the hydroxyl group on the 3' end, thus impeding DNA sequence extension beyond that position. Those analogues were incorporated at random positions on the synthesis process rendering DNA strands of all possible lengths which were revealed by autoradiography.[79] In the following years, advances in fluorescence substituted radiolabelling with fluorometric detection, simplifying and enhancing sequencing protocols.[81]

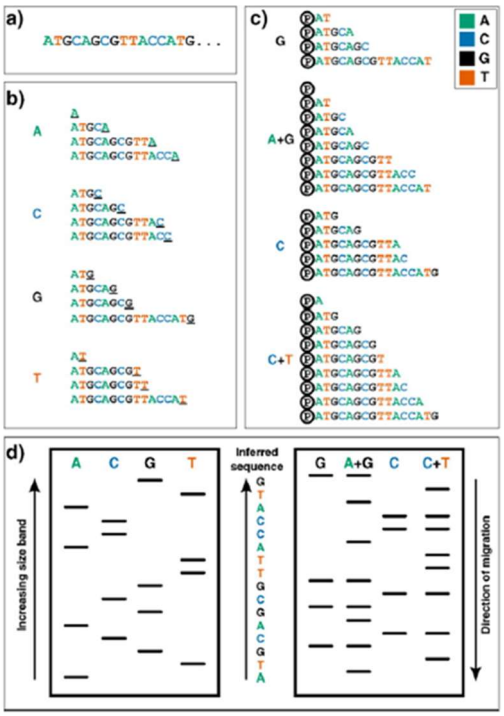


Figure 4. First Generation sequencing methods. (a) DNA molecule to be sequenced. (b) Sanger method. (c) Maxam-Gilbert method. (d) Fragments visualized via electrophoresis.[79]

Base detection was also improved with capillary-based electrophoresis.[82] It augmented the rate at which fragments could be separated by applying much higher electric fields that were used on standard electrophoresis. The progress verified on sequencing methods manifested itself on the development of automated sequencing machines. Nowadays, Sanger method is still used by some of those machines like ABI 3730xl DNA Analyzer, a Gold Standard on automated high-throughput sequencing. It produces reads up to 900 base pairs. A run yields a total of 96 kb (thousand bases) pairs in 3 hours.[81] The sequencing process cost lowered a lot, in 1985, for every single base, the read cost 10$. On the mid-2000s, 10$ rendered 10 kb pairs.[76]

2.2.2 Second generation

As Sanger method was still being improved, its limitations motivated the appearance of new alternatives. The development of emulsion PCR, an in vitro process that anchors DNA fragments to primer-coated beads, augmented the yield and accuracy of the sequencing process by using more than a hundred clonal copies of the DNA molecule per bead. Over the years, this parallelized process allowed to perform gradually larger amounts of reactions, increasing the sequencing throughput.[83]

Pyrosequencing arose in the early 1990s as the first real alternative to the Sanger method, a sequencing-by-synthesis technique that measures inorganic pyrophosphate synthesis with a luminescent chemical method. It consists of a series of enzymatic reactions following nucleotide incorporation events that ultimately emit light photons that are detected and measured by a sensor.[76,84]

The market reference to this technique is the Genome Sequencer FLX marketed by Roche 454 Life Science. It is capable of generating reads of 250 base pairs to a total of 80-120 Mb (million bases) in a 4-hour run. More recent instruments like FLX Titanium may yield 400 base pairs reads. As today, this technology is no longer commercialized.

On the late 1990s, the nowadays known as Illumina sequencing technology was developed. It relied on clonal bridge amplification, another method to amplify DNA that consists of attaching both ends of single-stranded DNA fragments to a solid surface covered with sequence adapters. That amplification process yields more than 40 million single molecule clusters, each one with a diameter of 1 μm and counting approximately 1000 copies of a single template.[76,81]

Sequencing itself is performed with fluorescently colour labeled dNTPs, one colour for each dNTP; all provided at a time. On each cycle, every cluster is scanned to identify the base incorporated by the colour emitted.[76,79,81]

The first instrument to apply this method, the Illumina GA genome analyser, generated 35-bp reads and sequenced a total of 1 Gbps (1 thousand million base pairs) on a 2-3 day run. This instrument sequenced each base by approximately 1% of the cost of Sanger sequencing.[76,81] There are more recent instruments that are capable of sequence small genomes in 4 hours and others may sequence 1.8 Tb (billion base pairs) in 3 days.[75]

Ion Torrent technology proceeds to an amplification step by emulsion PCR to scatter millions of copies of a DNA single-stranded chain over the surface of a bead. The method is based on measuring H^+ protons release into the solution during polymerization reactions. A pH sensor plays a fundamental role to detect those variations. Besides yielding the highest throughput (Ion Proton produces over 50 million reads with 200 bases length per run) [85], this process is much faster-completing runs within 2 to 4 hours.[86]

2.2.3 Third generation sequencing

The focus on improving speed, accuracy and reducing costs of sequencing large amounts of genetic information, ultimately led to the development of methods to sequence single molecules hence avoiding the amplification requirement.[79]

The most impacting single-molecule sequencing method in the present day is PacBio Real Time Sequencing-by-synthesis (SMRT). In SMRT, DNA polymerization occurs on nanometer-scale wells with holes that focuses light on approximately 20 zl (10^{-21} liters) of the volume of the well. A single DNA polymerase immobilised on that spot incorporates fluorescently labeled nucleotides thereby exposing the base-specific fluorophore. There are SMRT instruments distributed by PacBio that may produce reads with a length of 10 kb at a rate around 10 bases per second.[81,87] SMRT can generate reads with an average of more than 14 kb in length. Individual reads may reach 60 kb.[75]

Another approach to single-molecule sequencing consists of the use of nanopores. Nanopores are already used on detection and quantification of various biological molecules by trespassing a lipid bilayer through ion channels. The technique is based on that principle using 1.5 nanometres wide synthetic pores, a current is applied and the negatively charged DNA chain traverses the nanopore at a flow rate proportional to the size of the nucleotide. A major issue on nanopore technology is the low detection sensitivity. Once this method could surpass that problem, nanopore technology may disrupt DNA sequencing paradigm since it has the potential to produce very long reads (around 6 kb on average and a maximum length of more than 60 kb[75]) cheaper and faster than was previously possible.[79]

2.2.4 Comparison

A comparative analysis performed by Quail *et al.* in 2012 reported that Ion Torrent sequencing would cost approximately 1,000 US Dollars per Gb on Ion-318 chip model. That sequencing process would produce 98 to 99% accurate data.[88] Presently available Proton-I chips yield between 60 to 80 million 200 bp reads, reaching 10 Gb within a 4-hour run. This amount of data is equivalent to the sequence of two human exomes at a 50-times coverage.[86]

Quail *et al.* also compared SNV calling ability of Ion Torrent, PacBio and three Illumina sequencers. The rate of 82% correct calls for Ion Torrent was higher than for Illumina machines, which accomplished correct rates ranging 68 to 76%. PacBio identified 71% SNVs correctly, this low rate could be explained with the optimization of existing tools for short-read data instead of for long reads where errors are more prone to occur.[88]

2.3 Data Context

2.3.1 Data processing

It has been presented how the evolution of NGS sequencing technologies permit to generate the enormous amount of data that is daily obtained. However, this aroused issues with data storing and quality control. As an illustrative example, although currently, whole-genome sequencing is the most informative and complete genetic analysis method, exome sequencing is still advantageous due to lower cost, higher speed and greater ease of storage and analysis.[89]

Several steps are needed to process and extract genetic variant information from raw sequencing data (Figure 5). [90,91] After generating short reads from the sequencing process, the next step consists on quality control of those reads. Following this, the reads are aligned to a reference sequence and the alignment is post-processed to improve the quality of the final steps that fulfill the variant calling.

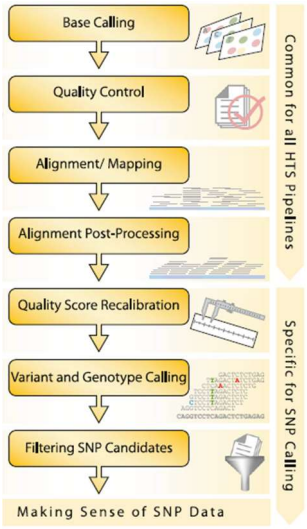


Figure 5. General pipeline for SNP calling. Adapted from "A beginners guide to SNP calling from high-throughput DNA-sequencing data", Altmann et al., 2012. [90]

Overall quality of the results must be assessed by universal classifiers, concordantly to this, base calling certainty is measured by statistical models that take in account factors such as signal intensities, number of sequencing cycles performed and distance to other sequencing clusters.[90] Error estimates are usually expressed as Phred-like quality scores which essentially relies on the expected error probability of the base call as is noted on the following formula.[92]

$$Q_{phr} = -10 \times \log_{10} P(error)$$

Currently, most platforms perform the entire process until this point, results are optimized removing redundant reads and contaminating primers, adapters or other artifacts. Manufacturer software verifies and compiles quality scores that are stored together with base calling data in FASTQ files. Those are widely accepted as the standard file for NGS raw data.[91,93]

2.3.2 Sequence Alignment

On the alignment step, the short reads contained in the FASTQ file are aligned to a reference sequence. This process was only possible after the completion of the Human Genome Project (HGP).

HGP was the first initiative taken to sequence the entire human genome. The project started in 1990 and in 2003 its conclusion was announced two years ahead of schedule. Project results presented 99% of the gene-containing region of human sequence with a 99.99% accuracy. Fifteen thousand full-length cDNA stretches have been identified and 3.7 million SNVs were mapped. To accomplish this endeavour, genome portions have been sequenced in several universities and research centres throughout the world.[94] HGP cost approximately 2.7 billion US Dollars. Afterward, sequencing costs progressively got lower allowing to generate larger amounts of data. [94]

The definition of a reference sequence established a major benefit to multiple scopes; biology, medicine and more specific subjects like proteomics, experienced a significant improvement supported by the enhancement of sequencing processes. As parallelization was enabled by a standard mapping against the reference sequence, second generation methods widespread was accompanied by higher throughputs and lower costs. [95,96] Reference sequence may be altered due to advances in the human genome sequence or more accurate determinations, the actual build is the Genome Reference Consortium Human Build 38 (GRCh38 or hg38). [62]

Notwithstanding this reference, it is known that there will be variations between sequences, so algorithms must be tolerant to imperfect matches toward finding optimal alignments.[91] Smith-Waterman score-based dynamic programming algorithm is implemented on alignment methods to provide at least one optimal local alignment.[97]

Two approaches have been used to speed up this process by compressing data, Burrows-Wheeler transform (BWT) rearranges a character string into runs of similar characters[98] and is faster and more efficient than hash-based aligners – which transforms strings into shorter fixed-length values or keys that are retrieved in a database – that by its time is more sensitive than BWT.[90]

The most popular BWT-implemented software are BWA[99], Bowtie[100] and its successor Bowtie2. As examples of hash-based software, there may be referred Novoalign[101], SHRiMP[102] and SHRiMP2[103].[104]

Besides those software, Ion Torrent technology has a specifically oriented set of utilities. Tmap[105] is designed to meet Ion Torrent data mapping challenges; it also applies BWT algorithms to build an index of the reference genome.

Subsequently to the alignment, the reads are stored in a Sequence Alignment Map (SAM) format file, besides the reads, also the mapped positions in relation to the reference sequence, the orientation of the read and the quality of the alignment are stored.[90] This file may then be compressed into a binary format (BAM) that reduces by 3 or 4 times the size of the file.[106] BAM is, nowadays, the de facto standard format for alignment files.[89]

BAM files are then used as the input for most processing tools, the most commonly used of them are SAMtools[106], Genome Analysis Toolkit (GATK)[107] and Picard[108]. Alignment post-processing process is useful to reduce even more the size of the file and avoid some errors that eventually could affect variant calling.[104] On this process, several tasks are performed, the reads are sorted in relation to their chromosomal positions, PCR-remaining artifacts (reads that start at the same position and have the same length) are removed, reads with more than one optimal alignment are also removed – because

it cannot be determined where it has been originated – and reads are realigned around small indels to avoid the detection of artificial SNVs that might be created by inaccurate alignments.[90]

Alternative formats have been developed to obtain more compressed files that still retain much of the original information. CRAM format[109] has been increasingly used in recent years, it may reduce the size of a BAM file by 38-55%.[110]

2.3.3 Variant Calling

The alignment of the reads provides a comparative perspective between sequences. It is powerful support for identifying rare and *de novo* variants and quantifies genotype expression levels.[111]

Variants are identified as nonreference alleles found on at least one sample. Then, variant calling may be defined as the search and identification of *locus* differing from a reference sequence. It should not be confused with genotyping which is the estimation of genotypes on each *locus*.[112]

By analysing reads from different samples concerning a genomic region, a probabilistic framework which facilitates variant calling is created. Although at the expense of speed and computing resources, a higher depth of coverage increases the possibility to find low-frequency alleles and guarantees a bigger certainty on the identification of alleles with low coverage in single samples. Besides, there is a reduction of the probability to call a random sequencing error as a variant.[91,104]

The biggest advantage in using a probabilistic framework is the possibility to integrate previous data on allele frequency and compare experimental results with these values. The baseline data may be found and derived from genomic and variation databases like ExAC [67], the 1,000 Genome Project [63] or dbSNP [18].[90]

The abovementioned GATK[107] and SAMtools[106] are two of the most used software to call variants. GATK includes two programs to perform the task, UnifiedGenotyper calls SNVs and indels assuming each variant *locus* as independent, HaplotypeCaller also detects SNVs and indels in addition to structural variations by performing local de novo assemblies of the aligned reads. SAMtools apply *mpileup* utility to scan every position of the genome and produce a BCF format file. It enables the calculation of the allele frequency for each position covered in the sample. Then, *bcftools* takes that file as input to call SNVs and indels.[91]

Among several variant calling tools, there is a third most used software to be mentioned, FreeBayes. This is a short polymorphism caller built on a Bayesian statistical framework, that is, it takes previous data and new evidence to infer variants. It can simultaneously detect SNVs, indels, multi-base mismatches, polyallelic sites and copy number variants.[113]

Like in the alignment step, Ion Torrent technology uses a variant caller designed and optimized to exploit its platforms data to evaluate variants. That customizable tool is called Torrent Variant Caller (TVC). [114] The standard output for the process is the Variant Call Format (VCF). Those files comprise the variants and their positions and it may have the genotype and haplotype information for each polymorphism.[115] Large VCF files may be compressed into Tabix format, which indexes the position sorted file to perform efficient querying of genome positions. In Tabix, the size is reduced by 3 to 5 times.[116] Another format, gVCF, store adjacent reference alleles as a block, in these files, start and end positions correspond to the first and the last position of an uninterrupted sequence of reference alleles.[117]

2.3.4 Comparison

The number of available tools to perform variant calling enables multiple comparisons between them. As it has been mentioned before, GATK and SAMtools are the most widely used and are present in almost every benchmarking studies.

Altmann *et al.* compared the results generated by both software and concluded that GATK provided five thousand additional SNV candidates compared to SAMtools.[90]

Concordantly, most performance analyses point the best general results to GATK, either over SAMtools[118] or both SAMtools and FreeBayes[119] among others.[111,120]

In some cases, the authors mention that GATK outperformance over other tools consists of a slight difference, Laurie *et al.*[121] concludes that GATK, SAMtools and FreeBayes are equivalent regarding SNV calling, the main difference between them consists on indel calling accuracy, on which GATK performance is substantially better than that of the other callers.

Once there Ion Torrent data is included to a benchmark, it is possible to compare those callers with TVC, Hwang *et al.*[122] results support the better suitableness from GATK to indel calling, but points SAMtools and FreeBayes as better SNV callers for Illumina platforms data. Regarding Ion Proton data, SAMtools outperformed any other caller. However, authors mention that they could not predicate TVC performance based on their results since one Ion Proton data set was tested, which also also had low exome coverage.

A comparative study performed by Zhang *et al.*[123] on the same year concluded that TVC, properly optimised to Ion Proton system, processed data better than GATK.

A recent study by Sandmann *et al.* [124] suggests not to use GATK to detect variants with low allele frequencies. They recommend FreeBayes and VarDict [125] to achieve more accurate results regarding those variants. Bao *et al.* [91] on a previous analysis also concluded that FreeBayes achieved best performances than GATK and SAMtools.

2.3.5 Annotation

The variant calling pipelines yield high-scale and complex sets of information, hence the management and interpretation in phenotypic context end up being a substantial challenge. Annotation tools have been developed to catalogue genetic variants to enhance data analysis and identify subsets of functionally important variants.

ANNOVAR. One of the first mainly used annotation tools was ANNOVAR [126]. ANNOVAR is a licensed software tool freely available for academic use. It needs to download pre-compiled gene annotation data sets and save them on local disk to annotate variants with respect to their functional consequences. These data sets are scanned to identify and report allele changes. It also performs genomic region-based annotations, compares variants to existing databases and evaluates subsets of variants not reported on them.

ANNOVAR can take VCF files to annotate, however, they must be converted since ANNOVAR requires a standard simple text-based input format. Each line represents one genetic variant and is composed of five mandatory columns which respectively represent chromosome, start position, end position, reference allele(s) and observed allele(s). Additional commentary columns can be supplied, in output files, these columns are printed out in identical form.

ANNOVAR can filter specific variants. Furthermore, it may automatically filter functionally important variants through a multi-step procedure that executes sequential annotations with several different parameters and generates a final output file containing the most likely phenotype causal variants and their corresponding candidate genes.[126]

VEP. Among a wide variety of prediction tool choices, there are useful options like Variant Effect Predictor (VEP)[127]. VEP is a software suite that performs annotation and analysis of most types of genomic variation in coding and noncoding regions of the genome using a wide range of reference data, including two of the most used scores, SIFT[128] and Polyphen-2[129].

Compilation of various prediction scores increases analysis reliability. Although the diversity of prediction methods, comparison studies have noted that none of the existing tools achieves repetitively satisfactory results. They report equivalent results of simpler approaches and more recent machine learning-based methods, however, the tools do not present constant performances, prediction accuracies greatly vary depending on the used dataset. These studies recommend considerable caution in interpreting the predictions generated, the available tools may not be accurate enough to give definitive conclusions.[130–132]

To improve analyses quality, there have been published guidelines to evaluate the pathogenicity of reported variants. Among other suggestions, Wallis *et al.* recommend using at least three programmes to obtain reliable results.[133]

GEMINI. Further initiatives arose from the studied evidence that damaging loss-of-function variants are frequently artifacts of data, annotation or analysis. [134,135] Moreover, the size and complexity of genome annotation datasets, its varying documentation, frequent updates and the storage on centralized repositories or individual laboratory websites constitute technical and methodological challenges to reliably identify and characterize genetic variation.

GEMINI [136] is an open-source framework built to stand as a flexible, reproducible, and scalable software for mining genome variation. It integrates genetic variation data in VCF format files with GRCh37-based genome annotations into a unified database framework that suppresses the need to develop complex analysis pipelines. This tool allows querying variants and genome annotations in a common SQL database which may be augmented with custom annotations. GEMINI developers preferred SQLite relational database engine over NoSQL approaches due to the SQL expressiveness on constructing data exploration queries and its intuitive syntax.

Annotated variants are also loaded as table rows. To facilitate individual samples comparison for observed variants, genotype information is stored for each sample as a compressed array in a single column for each variant row. This strategy enhances both query performance and scalability while still providing necessary access to individual sample genotype information. Additionally, data for each variant is complemented with calculated statistics and population metrics.[136]

dbNSFP. Integrative tools that include both variant annotation and its phenotypic effect predictions have also been developed. dbNSFP collects all possible nonsynonymous SNVs in the human genome as they are found on the annotation of the Consensus Coding Sequence (CCDS) Project.[137] CCDS set is built by consensus of genomic datasets from diverse public resources.

dbNSFP first version compiled over 75 million SNVs, their respective scoring scores from four prediction algorithms, SIFT[128], Polyphen2[129], LRT[138] and MutationTaster[139], and one conservation score from PhyloP [140].

Although CCDS was based on the human reference sequence built hg18 (GRCh36), the coordinates were converted to hg19 on the database building process. However, 561 SNVs were not successfully converted. Both coordinates are available on the database as well as both reference and alternative alleles, reference and alternative amino acid, gene name and ID, CCDS ID, reference codon, position on the codon, amino acid position on the protein and the referred algorithms scores and predictions.[141]

On the second version of dbNSFP, the database was rebuilt, the information was separated into two parts, one regarding variant annotation and the latter regarding gene annotation. CCDS annotation was replaced with hg19-based GENCODE 9 annotation[142] and two prediction scores were added, MutationAssessor[143] and FATHMM[144], as well as two conservation scores, GERP++[145] and SiPhy[146,147]. Allele frequencies were also added from the 1,000 Genome Project [2] and the Exome Sequencing Project[148]. The database then included over 89 million SNVs and their respective updated scores.[149]

The third and current version also presented some major updates, the database backbone has been rebuilt using GENCODE 22 annotation which is based on the latest human reference sequence version GRCh38. It now contains over 82 million SNVs, along with it, there is also distributed a database called dbscSNV [150]. It, compiles all potential human SNVs within splicing consensus regions and their deleteriousness predictions, accounting over 15 million additional SNVs.

This latest version adds various prediction scores, MetaSVM and MetaLR [151], CADD [152], VEST3 [153], PROVEAN [154], fitCons [155], FATHMM-MKL [156] and DANN [157] and the conservation score phastCons [158]. Every algorithm version has been updated. Finally, allele frequencies from UK10K cohorts [4] and ExAC [66] were also added. [159]

Various annotation tools have been developed relying on dbNSFP data.

Vanno. As an example of a web-based application, Vanno[160] is a freely available tool that generates an integrated database from multiple annotation sources, those sources are regularly updated, hence a batch script was created to be executed monthly to keep up-to-date information available.

The differentiating feature presented by this tool is the visual architecture based on Circos visualization tool [161] and interactive filters that update the results and regenerate the image in real time.

The tool is capable of processing any variant calling file only requiring to the user to select the correct targeted gene panel and the respective variant calling package from which the variant calling file is generated. Those files are converted into a standard format subsequently stored into a SQLite database. Variant annotation consists of compiling information from several tools, molecular consequences are obtained from ANNOVAR and function predictions from multiple tools are extracted from dbNSFP [141]. Finally, information on gene ontology terms, biological pathway, protein domain, protein structure, and interaction networks are also annotated by consulting available databases. The output summarizes genetic variants rendered in charts, tables, and Circos plots. The major limiting factor of this feature on large datasets is the time consumption of generating high-resolution plots. [160]

VarAFT. VarAFT [162] is a system freely available to non-commercial usage that annotates variants based either on GRCh37 or GRCh38 reference sequences and filters variant files besides enabling to compare several individuals.

It implements ANNOVAR as a basal annotation tool. As a useful feature, VarAFT provides a direct link to Integrative Genomics Viewers (IGV) visualization tool [163] to visualize any variations using a BAM file.

Instead of displaying standardized tables, VarAFT allows the user to select which columns are to be displayed. Among available information, VarAFT includes gene annotation data, function prediction scores and allele frequency information from diverse sources. VarAFT additionally allows the user to create a local database to filter and analyse given variants.

To examine data, VarAFT relies on a Java based interface with diverse filtering options and the possibility to customize analyses.

Final output presents a summarized table containing all variants, for each variant is possible to access a panel that provides full detailed information displayed on an organized interface. [162]

Highlander. Like VarAFT, Highlander [164] is an annotation tool that relies on a Java based interface. Yet, unlike VarAFT, Highlander is an open source software coupled to a local MySQL database which compiles available variant data and annotations and enables query-based filtering methods. Database information comes from various sources, among them, dbNSFP [141] provides functional predictions, prioritization scores and allele frequencies. Global statistics are then computed to enable variant discrimination through filtering queries. Complex queries are simplified by using shortcuts for certain standard criteria.

The software facilitates analysis of filtered data by allowing the user to sort, mask and highlight information besides accessing useful tools to visualize the alignment or explore variants on specific genes among other functionalities. Variant details may be accessed by selecting any given variant.

Highlander enables customized filtering functions that may be saved, edited and loaded at any moment. Conditional highlighting and implemented search tool are also meant to enhance result analyses.

Finally, BAM and VCF files for current analysis may be downloaded and the table content might as well be exported as Excel or TSV files. [164]

Finally, the amount of information required by these tools has motivated the development of alternative ways to store data.

CellBase. With the same purpose of joining information from various sources into only one database, CellBase is a NoSQL-based repository for genomic information of more than 20 species, for humans, annotations are based on the GRCh37 reference sequence.

It compiles protein annotation from UniProt and its functions from SIFT and Polyphen, allele frequencies provided by the 1,000 Genome Project, Exome Sequencing Project, ExAC, GnomAD, UK10K and GoNL and conservation scores obtained with PhastCons, PhyloP and GERP++ besides clinical information. Sequence effect prediction is calculated on the fly and described by Sequence Ontology (SO) terms.

Although it can be accessed by a Java application programming interface, CellBase has been designed to enable access by web services that can be queried by different programming languages like Python, R, Java and JavaScript or by user-developed methods. [73]

Considering the size of the database, the integration of the various scores and the allele frequencies derived from the major population genetics endeavours, and mainly for being built accordingly to the actual human reference sequence, GRCh38, dbNSFP is ideal to stand as a comparative reference for our project. It is already used by different tools to obtain annotation data. Allows to directly analyse the phenotypic outcomes of overrepresented alleles in the regarded population, beneficing our study even more. As it has been mentioned before, predictive approaches may not be entirely accurate, said that, the inclusion of diverse pathogenicity prediction scores is helpful in obtaining reliable conclusions.

2.4 Population analysis

To be able to compare genetic data between samples or populations, information must be organized. Different approaches may be taken, a VCF file may be used as groundwork for frequency analyses. However, basing genetic analyses solely on VCF files could hamper the linkage to more in-depth information and functional scores, nonetheless, the main defect on this approach would be the crescent complexity of analysis scripts with a bigger number of files. Therefore, data compilation in a database improves scalability and information retrieval through queries. Regarding it relational nature, a SQL-based approach is a more natural choice to provide a local structural support to comparative studies. As it has been presented, both Vanno and Highlander, compile all information in this kind of structure, on the other hand, CellBase uses a NoSQL repository to fetch annotation data yet this repository and variant call information are stored independently.

To store variant data in a SQL database, each variant must be stored on a table row, this because, as it has been presented, a variant call pipeline may detect millions of variants, SQL management systems compile a limited number of columns, for example, MySQL [165] and SQLite [166], two of the most used systems, have a table column count limit of 4,096 and 32,767 columns respectively.

The referred databases stores information row-by-row, the biggest issue in using these systems to query databases with, eventually, millions of rows is that each row must be scanned in totality, the more rows, the more is the time and resources consumed. For this kind of data, alternative systems stand as a ponderable choice. Column-store database management systems keep the same SQL syntax and the same methods found on row-based systems, the main difference consists of the way that information is accessed, instead of fetching the necessary information in each row, systems like MariaDB ColumnStore [167] and MonetDB [168] handle only the columns required to respond to a query, it is more suited for analytical workloads and particularly for datasets that compile diverse types of information which is the case of the addressed issue, as it may utilize genetic data, frequency values and diverse scores in the same analysis.

Once data has been compiled, it is possible to compute different measures for population genetic analyses.

Principal Component Analysis. PCAs have been widely used in populational studies. As before the advent of large-scale usage of SNP data, it is still used to summarize allele frequency data from diverse populations. It groups a set of observations by sets of values of linearly uncorrelated variables called principal components (PC). The first principal component of this set has the largest possible variance and the following components are increasingly more specific. A plot composed of the first 2 PCs is generally used to reproduce the geographic arrangement of sampled individuals. [169]

PCAs may be used to characterize differences among populations and individuals regarding its ancestry. The proportion of the variance explained by the first PC may indicate the presence of population substructure regardless of the influence of migrations or isolation between populations. [170]

Although PCA may be used for diverse types of analytical studies, there are tools specifically focused on a genetic application, LASER [171] is a program based on a PCA approach that estimates individual ancestry background by analysing its sequence reads.

3. METHODS

3.1 Samples

Samples of 71 Portuguese patients between the ages of 48 and 80 years old were sequenced previously to the present work by Genoinseq – Next Gen Sequencing unit at CNC - Centro de Neurociências e Biologia Celular da Universidade de Coimbra facilities, in the scope of the project DoIT (FCOMP-01-0202-FEDER-013853). Individuals were diagnosed with diabetes type 2 and had no congenital diseases. Raw sequencing data was also previously processed and aligned to the reference genome sequence GRCh37, the resulting BAM files stored internally stood as the initial working material for the present dissertation.
Of the 71 samples, "Exome 51" was excluded by parenthood with "Exome 41". All analyses, therefore, were based on 70 unrelated individuals.

3.2 Populations

Besides the 70 Portuguese samples, 1000 Genomes Project (1kG) data was also included in the present work. Genotypic information for the populations was downloaded from the 1000 Genomes Project web page [172]. Overall, 2,504 1kG samples were considered for frequency comparisons to world population. For European-level comparisons, only data from European populations from the 1000 Genome Project was included in the analysis: 107 Iberian samples from Spain (IBS), 91 British samples from England and Scotland (GBR), 99 Finnish from Finland (FIN), 99 Central-European residents of Utah, USA (CEU) and 107 Tuscan samples from Tuscany, Italy (TSI). The number of samples totalizes 503 European individuals. Besides frequency values, the genotypes provided in the VCF files corresponding to those samples were also used. According to the Coriell Institute collection information [65], all IBS samples were collected throughout the Spanish territory, the individuals were identified as born in Spain and having its entire two previous generations born in the same area.
GnomAD frequencies were also used for frequency comparisons. The used file, corresponded to the release 2.0.2 with gnomAD exomes, can be found on its website [68]. This dataset comprises 123,136 exomes, 55,860 of them are non-Finnish European samples. Corresponding allele frequency values were the ones used for the comparisons at the European level.

3.3 Database construction

3.3.1 Insertion of Portuguese variants

BAM files were used as input for TVC variant caller [114], this plugin was used to perform the variant calling procedure against the reference genome GRCh37 The results for each sequenced sample were stored in a different gVCF file, which store uninterrupted extensions of adjacent reference alleles as a block.

gVCF blocks were then decomposed by *gvcftools* break-blocks function [117], the output VCF, hence store the information for each position separately. Finally, *vcftools* [115] *merge* function was used to compile in a single VCF file all Portuguese variants from the 70 sequenced exomes.

Those variants were inserted into a MongoDB (version 2.4.14) [173] database by a script, developed to generate, for each variant, a database unit named *document*.

The script granted that the database would only include variants for which at least one Portuguese sample had a heterozygous or an alternative homozygous genotype. All unaltered positions were excluded of this work, otherwise, the database would be unmanageably large for the available resources.

The script also separated multiallelic variants (Figure 6. A to B) – various alternative alleles for the same position – to create distinct database documents for each variant. Reference homozygous genotypes were included in each decomposed document, altered genotypes were only attributed to the correspondent variant.

To normalize data from different sources, all variants were reduced to its minimal representation by this script through a function developed by the MacArthur Lab affiliate member, Eric Vallabh Minikel [174]. This reduction consists on the removal of nucleotides in common for both reference and alternative alleles at its extremities, therefore, the added alleles would start at the first altered nucleotide and end at the last altered nucleotide – which could be the same, if the subjacent variant was a SNV (Figure 6, B to C last case).

A				B				C		
POS	REF	ALT		POS	REF	ALT		POS	REF	ALT
1001	CTCC	CCC, C, CCCC	→	1001	CTCC	CCC	→	1001	CT	C
			→	1001	CTCC	C	→	1001	CTCC	C
			→	1001	CTCC	CCCC	→	1002	T	C

Figure 6. Schema depicting multiallelic variants decomposition – transition from A to B – and reduction to minimal representation – transition from B to C. POS stands for position, REF for reference allele and ALT for alternative allele.

After the insertion, MNVs were decomposed into multiple independent variants to avoid superposition cases with other variants (Figure 7). A script was developed to add each position and its corresponding alleles to the database along with all genotype information as a new document; the MNV document was then removed.

A					B			
POS	REF	ALT	SAMP		POS	REF	ALT	SAMP
1001	AG	CT	Ex1	→	1001	A	C	Ex1
				→	1002	G	T	Ex1
1002	G	T	Ex2, Ex3	→	1002	G	T	Ex2, Ex3

Figure 7. Schema depicting MNV decomposition. POS stands for position, REF for reference allele, ALT for alternative allele and SAMP for the samples for which the variant was called. Variant in position 1002 in A is not subjected to MNV decomposition procedure, nonetheless, the process generates a document containing the same variant with information for different samples.

Subsequently, this step generated some variants with the same general information as another variant detected for other samples. Genotypes called for different samples, were merged by a second script that joined all information of both documents and then removed the original ones. At this point, there were no repeated or superposed variants in our database. The merged file will be referenced as *VCF of Portuguese variants* (Figure 8, column Data Insertion).

3.3.2 1kG and gnomAD files processing

To reduce the size of the 1kG data files, *vcftools isec* function was used to intersect original files and the *VCF of Portuguese variants*, creating VCF files containing the information for positions shared by both VCFs. Along *vcf-isec* command, the options *-f* (*–force*), *-o* (*–one-file-only*) and *-n +2* (*–nfiles [+-=] <int>*) were provided. The first option forces the script to continue even if identifying a different number of columns in both files, otherwise, this condition would stop the process. For the present case, the number of columns corresponds to the number of samples. Since both populations include a different number of individuals, *-f* must be used.

The option *-o* orders to only print the entries from the left-most file (writing order) to the output document. In every case the left-most file was the 1kG VCF. Without this option, 1kG and PT data would be printed as separated entries for each variant found in common in both files.

Finally, *-n* establishes the number of files that must share the position, the *+2* modifier defines that it should be found on, at least, two files.

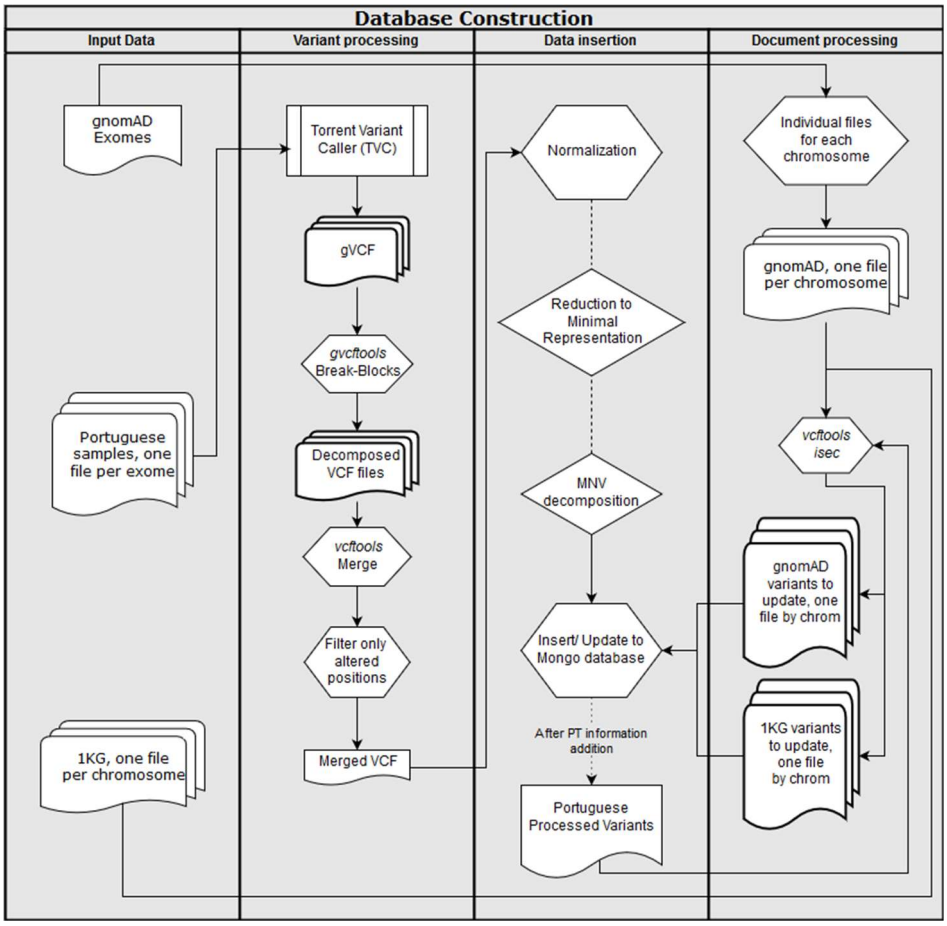


Figure 8. Schema depicting the construction of the database used for the present work.

gnomAD file information was extracted into a different smaller file for each chromosome using *vcftools* querying options. Subsequently, those files have undergone the same process as 1kG files to filter the variants which were not found in the Portuguese population (Figure 8, column Document processing). Finally, database documents were updated with the information contained in the obtained files.

3.3.3 Variant annotation

The abovementioned VCF containing all variants found in Portuguese samples and its genotypes was used as input for variant annotation with GEMINI [136]. This provided, for each variant, its respective gene name, an effect classification accordingly to Variant Effect Predictor (VEP), impact predictions by SIFT and Polyphen, clinical significance and disease name for the variants found in ClinVar and identified the exomic variants.

The information was recorded in a CSV file along with the variant chromosome position, reference and alternative allele and was updated to the respective database document.

3.3.4 Database structure

The database includes a separated collection of documents for each chromosome. Each document represents a sole distinct variant and its respective information (Figure 9), as variant position (Pos), reference allele (Ref), alternative allele (Alt) and all available genotypes (gt). Additionally, the alternative allele frequency for the Portuguese samples (AF) was calculated and included in the document as an independent field. (Figure 9, in orange)

The field PT_Counts (Figure 9, in blue) contains a dictionary that includes the values for the sum of reference (RefAll) and alternative alleles (AltAll), number of reference homozygous (Hom_Ref), heterozygous (Het) and alternative homozygous (Hom_Alt) samples and the total of samples (NSamps).

Two distinct fields include information referent to 1kG (Figure 9, in purple) and gnomAD (Figure 9, in green), respectively. From its files, for each variant in common to the Portuguese population, the Allele Frequency (AF) values for each population presented were updated together with the general AF value to the respective document. Additionally, Allele Count (AC) and Allele Number (AN) values for both European populations – NFE for gnomAD and EUR for 1kG – were updated to its respective field.

In relation to 1kG, alongside the provided information, the values for AC, AN and AF for each European subpopulation were calculated and updated to the database as a dictionary included in the 1kG field. Samples were associated to each subpopulation accordingly to the information provided by the sample list downloaded from the International Genome Sample Resource website[172].

For gnomAD, the Filter *flag* provided in its files for each variant is also included in its database field.

A field named *Annotation* consists of a dictionary of additional information for the variant. HWE p-value for the Portuguese samples was calculated and compared to a threshold of 0.05, the sub-field HWE_PT presents a Boolean *flag* that indicates whether the value is superior or inferior to the threshold, which corresponds, respectively to variants found at equilibrium (value: 1) or not (0).

Finally, information obtained from the annotation procedure was added to the Annotation field, a Boolean *flag* to distinguish exonic (1) and non-exonic variants (0), Gene name, SIFT and Polyphen predictions, a VEP field containing a dictionary that presents its impact and severity classification and a ClinVar dictionary field that contain both significance and the name of the disease associated to the variant.

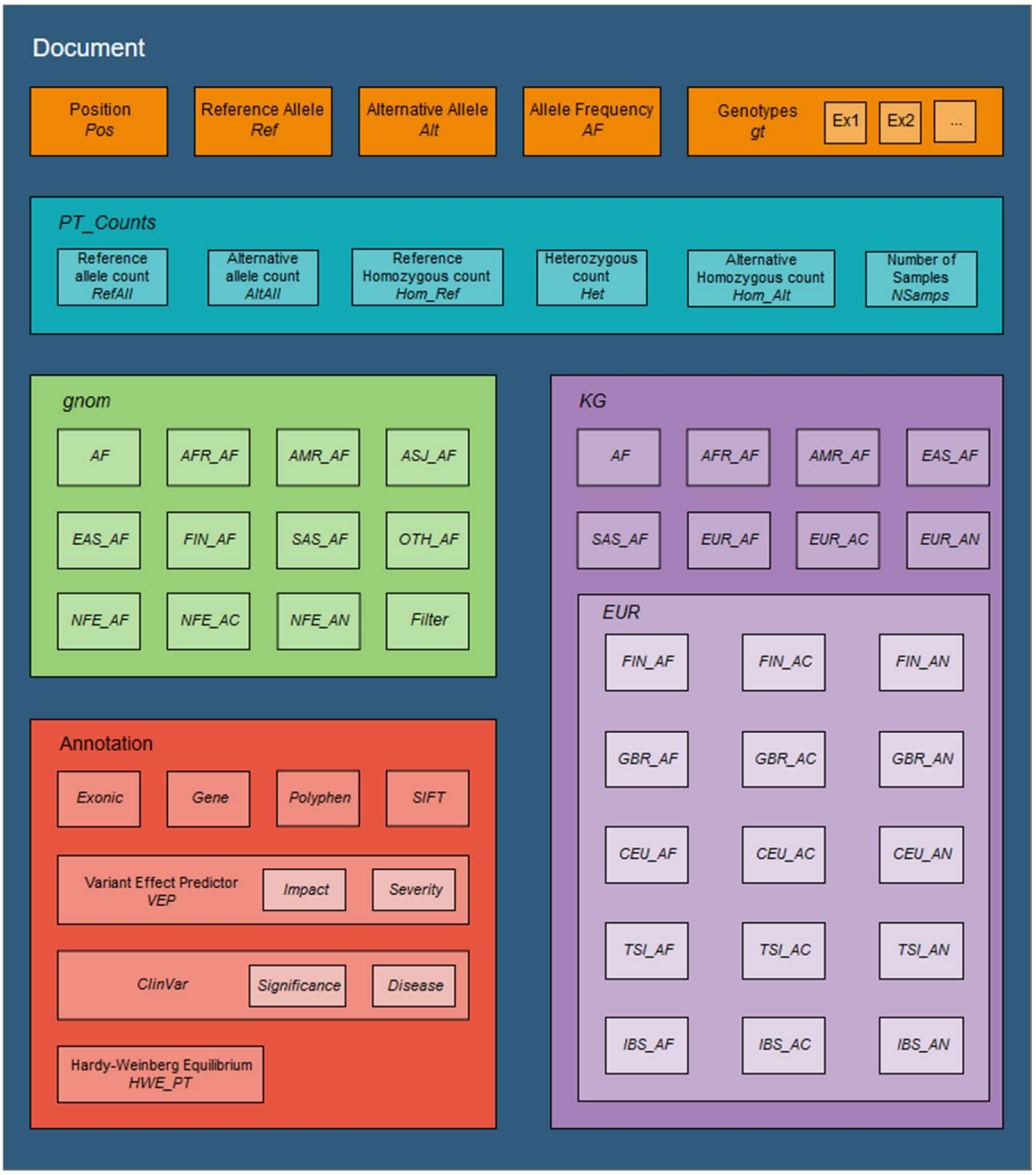


Figure 9. Representative schema of all the information that may be contained in a document. Field names are presented in italic. In gnom field: AFR – African/African American; AMR – Admixed American; ASJ – Ashkenazi Jewish; EAS – East Asian; FIN – Finnish; NFE – Non-Finnish Europeans; SAS – South Asian. In KG field: AFR – African; AMR – American; EAS – East Asian; EUR – European; SAS – South Asian. In EUR sub-field: FIN – Finnish; GBR – British; CEU – Central Europeans from Utah; TSI – Tuscans; IBS – Iberians.

The database does not present empty fields, if there is no information for a field, it is not generated. The non-existence of a field or a sub-field does not affect other information on the document.

Documents were indexed by position in ascending order through Mongo *index* function. Index enabled faster searches and information updates.

3.4 Population comparison

3.4.1 Allele Frequency scatterplots

From the created database, the allele frequencies values of each population were obtained for each SNV in HWE (HWE_PT: 1) that presents 60 or more Portuguese genotypes (*NSamps* >= 60). A distinct CSV document was generated for 1kG and gnomAD containing the AF values of its correspondent populations, in the case of 1kG, the allele frequencies of the European subpopulations were also included separetely. Each document was read in R environment and the *ggplot* package was used to generate scatterplots of AF values to compare Portuguese information to each population contained in each file. The graphics generated for each population in one or another project were presented in World Maps that linked a population to its location, *ggmap* package provided the map template and enabled to pinpoint each image to its respective location. The same approach has been applied for 1kG European subpopulations with a Europe map.

3.4.2 Principal Component Analysis

Another CSV file was generated from the presented database to contain all available genotypes for the 503 European individuals from 1kG and the 70 Portuguese samples for each variant that comply with a set of conditions. These conditions postulated that the variants might be autosomal SNVs, present HWE conditions in the Portuguese population, 60 or more Portuguese samples, an allele frequency above 0.1% for the European individuals and relatively to the latter, at least 453 reported genotypes (maximum of 50 missing values).

An *Adegenet* [175] genlight object was generated in R environment, providing the genotypes for each sample of the CSV file, a unique variant ID, and both chromosome and position of the variant. A population label was added to link each sample name to either 'PT' or its respective KG sub-population groups.

Finally, an *adegenet* glPCA object was generated by performing a Principal Components Analysis. It contained each Principal Component Eigenvalue and a matrix of scores containing a list of values by individual (rows) for each Principal Component axis (columns).

PCA scores were then converted into graphical representations reflecting the distribution of individual values for the first 4 Principal Components. The plots were generated using *ggplot2* package [176]. The package *scatterplot3d* [177] has been used to combine the distribution of the first three Principal Components in a single plot.

3.4.3 Fisher's exact test

The statistical significance of the differences between allele frequency values found for Portuguese and European populations of 1kG were measured with Fisher's exact test.

For that, alternative allele and total allele counts for each populational group compared were obtained from the constructed database.

Fisher's exact test of independence is preferentially applied to smaller sample sizes, yet it may be used to analyse contingency tables for any sample. Instead of relying on approximation approaches as other statistical tests like the Chi-squared test, Fisher's method calculates the exact deviation from a null hypothesis (same allele frequency in both populations).

In cases where at least one cell of the contingency table presents a value below 5, it is recommended to use an exact approach as Fisher's test. As it is, its choice over other statistical tests relies on the low allele count values, which were found on contingency tables in respect to many variants.

To ascertain the significance of the independence between the populations analysed, p-values were corrected for False Discovery Rate (FDR).

4. RESULTS AND DISCUSSION

4.1 Genomic position

The 70 sequenced samples contained 57,142,483 genomic positions. A variant was found in 272,207 of them (0.48%), 49.5% of these positions corresponded to exonic regions (Table 1).

Table 1. General characterization of the positions covered by the constructed database for the Portuguese population, the amount of exomic positions in each subset presented is calculated in relation to the total number of positions covered.

	Total	**Exomic**
Positions	272,207 (100%)	49.5%
Covered by all samples	49.5%	27.7%
One alternative allele (biallelic)	99.0%	49.1%
More than one alternative allele (multiallelic)	1.0%	0.4%

Less than half of the positions are covered in all the sequenced samples (Table 1). This finding raised a concern for future analysis as sample size influences AF values, that would be the base for several comparisons. The position coverage distribution (Figure 10) depicts a great difference between the amount of positions covered by 70 and by 60 or more individuals, suggesting a substantial lack of information for 1 to 10 samples.

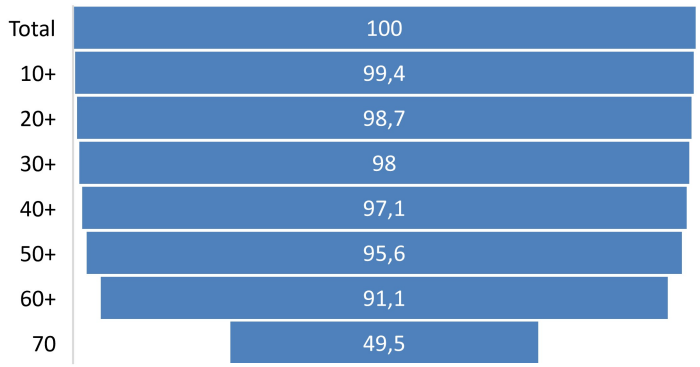


Figure 10. Count of positions by number of samples where are reported, in percentage. Number of samples were grouped in intervals of 10. Percentages (%) were calculated in relation to the total number of positions covered.

This result was confirmed by the coverage analysis for each sample. It was noticeable that there were 10 samples (1-8,166 and 168, Table 2) that presented considerably lower values than the remaining. Those samples had information for less than 90% of the positions, four of them for less than 80%, while the coverage average stood at 94.3%. In opposition, the 10 samples with higher values for this analysis covered more than 97.5% of the positions and 53 samples covered, at least, 95%.

Table 2. Position coverage by sample. The number of positions covered (Cov) for each exome (Ex) and the percentage (%) that it represents among the total number of positions where an identified variant is presented. The 10 samples with lowest coverages are presented in red, the 10 samples with highest coverages are presented in green.

Ex	Cov	%	Ex	Cov	%	Ex	Cov	%	Ex	Cov	%
1	242,355	89.03	19	260,130	95.56	37	262,265	96.35	130	256,083	94.08
2	241,827	88.84	20	261,175	95.95	38	261,889	96.21	132	256,319	94.16
3	225,778	82.94	21	262,910	96.58	39	261,523	96.08	134	266,737	97.99
4	211,528	77.71	22	263,081	96.65	40	263,104	96.66	135	265,828	97.66
5	243,766	89.55	23	262,831	96.56	41	261,646	96.12	136	266,398	97.87
6	236,178	86.76	24	260,246	95.61	42	262,760	96.53	138	265,757	97.63
7	234,847	86.28	25	262,762	96.53	43	262,601	96.47	140	266,630	97.95
8	215,166	79.04	26	260,484	95.69	44	261,521	96.07	142	265,979	97.71
9	262,429	96.41	27	259,228	95.23	45	261,604	96.10	149	266,881	98.04
10	262,889	96.58	28	262,709	96.51	46	260,800	95.81	151	264,158	97.04
11	260,658	95.76	29	262,415	96.40	47	262,984	96.61	155	266,131	97.77
12	263,161	96.68	30	258,201	94.85	48	262,266	96.35	158	266,316	97.84
13	261,955	96.23	31	261,785	96.17	49	261,372	96.02	161	264,948	97.33
14	261,240	95.97	32	263,642	96.85	50	259,613	95.37	164	263,839	96.93
15	260,673	95.76	33	265,819	97.65	114	252,398	92.72	166	196,958	72.36
16	263,121	96.66	34	264,682	97.24	116	252,316	92.69	168	196,363	72.14
17	262,191	96.32	35	261,214	95.96	123	252,847	92.89			
18	260,836	95.82	36	263,364	96.75	126	255,963	94.03			

For populational analyses, it was convenient to include positions that comprise information for as many samples as possible while not reducing the number of variants involved excessively. As it is, considering the last two results presented – 91.1% of the positions including 60 or more samples (Figure 10), and the 60 samples that present a coverage above 90% (Table 2) – the threshold of 60 exomes was recurrently used for diverse comparisons. This value does not represent a defined group of samples, each combination of 60 or more called genotypes for the same variant met this condition.

Finally, among all positions, 1% of them were multiallelic (Table 1), that is, they represent *loci* for which more than two alleles were found. This value is lower than previously reported, nonetheless, the cohort size presents a large influence on this proportion. For this case, this low percentage of multiallelic positions is concordant with our small number of samples. [178]

Regardless of that, the existence of multiple variants in common positions took the focus of the analysis from the positions to the variants.

4.2 Variants Characterization

4.2.1 Singletons

Considering that multiallelic positions contain multiple variants. There were 275,159 variants registered in the database, these variants have constituted the basis for all presented analysis – a detailed schema of the variants considered is presented in Attachment I. 90.3% (248,547) of the variants database documents included 60 or more samples. Subsequent comparisons were made among this subset.

39.9% of those variants (99,142) (Table 3) presented a sole alteration among all samples reported (singleton variants), that is, for these variants, all but one sample were homozygous for the reference allele. That value presented itself among the results obtained in other population-level projects, 31.6% for a Spanish initiative [1], 28.4% among Dutch [5] and 42.8% in a large-scale British study. [4].

Exomic regions comprised 50.1% of the variants (124,541). This group includes 53,860 singletons, which stand for 21.7% of the total. Overall, 94.7% of the singletons presented were heterozygous and 51.7% of the variants were exomic singletons.

Table 3. Characterization of singleton variants. Singleton variants are those for which a single sample presents an alteration. In each table section, its values are calculated in relation to a total indicated as 100% of the variants accounted.

	Total	**Exomic**
Variants	248,547 (100%)	50.1%
Percentage of singletons	39.9%	21.7%
Singletons	99,142 (100%)	54.3%
• Heterozygous	94.7%	51.7%
• Homozygous for alternative	5.3%	2.6%

For the present subset, all heterozygous singletons presented an allele frequency value below the threshold of 1% – the convention to define rare variants – henceforward, those variants were referred as low-frequency variants and presented AF values between 0.71% (1 alternative allele among 140 alleles) for variants that included information for all 70 samples and 0.83% (1/120) for variants that included information for 60 samples.

All homozygous singletons present 0 heterozygous genotypes. None of them was found at Hardy-Weinberg equilibrium, on the other hand, all heterozygous singletons met HWE conditions.

4.2.2 Hardy-Weinberg Equilibrium

Overall, 90.2% of the subset variants (224,155 variants) were at Hardy-Weinberg Equilibrium (HWE). HWE has been used as a measure to ascertain whether the genotypes detected for each variant reflected a conceivable allele distribution for a population. [7,8]

The alternative homozygous singletons may be taken as a representative example to endorse this approach. Those variants represented 36.5% of the non-HWE variants, they portray unusual cases once if a whole population does not include heterozygous individuals, homozygous genotypes for different alleles cannot be inherited. Said that, besides technology errors, the most viable explanation for these occurrences would be that the variant was more prevalent in some subsets of a population than in others. This would increase the probability of selecting, during the sampling process, an individual with that specific alteration, for that variant, the cohort would not be representative.

Two examples may be presented in accordance with this hypothesis. One of them relied on the 9.6% (855) of all non-HWE variants found in chromosome X. This value corresponded to 17.2% of the variants of that chromosome while the average for that proportion was 3.8% for each chromosome. That discrepancy may be explained by the fact that all males are hemizygous, hence, present an abnormal allele distribution in the population. These numbers draw attention to possible inaccuracies of some analyses when including variants with unbalanced amounts of alleles.

The presented hypothesis was also endorsed when regarding all SNVs, which are less prone than insertions or deletions to be wrongly called by Ion Torrent technology [88], thus enabling more reliable conclusions about the quality filters.

There were found 1,787 non-HWE autosomal SNVs that did not report any heterozygous occurrences, among them, 75.4% presented a sole homozygous genotype for the less abundant allele, that is, despite not representing a balanced population, it corroborated the assumption that homozygous genotypes were scarce on the actual population.

It is worthy of mention that, overall, 45.6% of those 1,787 variants were found neither on 1kG nor gnomAD exomes and may be eventual population-specific variants. Additionally, 10.7% presented an AF below 1% in both European populations of those projects and 20.5% presented those low values for one source and were not present in the other one. Those two scenarios included 31.2% of the variants in the analysis.
Nonetheless, these variants represent only 3.6% of the SNVs excluded. As most of the variants that did not pass the filters might be errors, excluding them in the populational analysis would not cause a substantial loss of effective information an might prevent result bias by erroneous values.
Henceforward, the minimal value of 60 genotypes and HWE conditions will be abbreviated as "quality filters", the term "filtered variants" will be used for the presented 224,155 variants that meet both conditions.

4.2.3 SNV, Insertions and deletions

The most notable effect of this quality filters was observed in the exclusion of insertions and deletions (Indels).
Although most of the variants were expected to be SNVs [8], the differences in relation to the number of indels were even more noticeable after applying the filters (Table 4).

Table 4. Number of variants by type of alteration. The presented percentages are calculated, for each row, in relation to the presented number of variants in the second column.

Variant Type	Total	Filtered	Filtered Exomic
SNV	253,336	215,511 (85.1%)	110,635 (43.7%)
Indel	21,823	8,644 (39.6%)	2,815 (12.9%)

The presented values denoted a tendency of the filters to exclude indels. This is concordant with the error propensity of the used technology in relation to that type of variants [179].
Only 39.6% of the indels passed the quality filters (Table 4), corresponding to 8,644 variants, which accounted for 3.9% of all the filtered variants. This value was lower than the one presented for large-scale exomic studies as ExAC [67], however, this proportion was even lower when taking into account only filtered variants in exomic regions, for which only 2.5% of the variants were indels. Besides constituting an additional indication of a higher error rate, this finding may be explained by the higher percentages of Loss-of-Function (LoF) Indels (Table 5). As these variants comprise more prejudicial effects, they tend to be rarer in the regions that code for proteins [134].

Table 5. Total count of filtered variants by impact according to GEMINI classification from Variant Effect Predictor (VEP). Impacts were grouped by its severity as classified by GEMINI (LOW, MED and HIGH). SNV and Indel counts were discriminated for each impact, the percentage of SNVs was calculated for the sum of variants by severity classification and depicts the proportion between SNVs and Indels. A more detailed table is available in Attachment II.

Impact	Total	SNVs	Indels	Severity	%SNVs
Synonymous variants	42,517	42,517	0	LOW	96.2%
Intron variants	99,273	93,960	5,313		
Other Low-impact variants	15,156	14,449	707		
Missense variants	54,949	54,949	0	MED	98.7%
Other Medium-impact variants	8,811	7,972	839		
Frameshift Variants	1,707	0	1,707	HIGH	48.2%
Stop loss variants	92	91	1		
Stop gain variants	835	821	14		
Other High-impact variants	815	752	63		
LoF variants	3,335	1,567	1,768		47.0%

The rareness of LoF variants was corroborated by the number of individual-specific alterations among them, 68.3% of the cases were found in a single occurrence among all reported genotypes.

4.3 Population Representativity

67.4% of the filtered variants (151,053) found in the Portuguese samples were also reported among the gnomAD exomes files. 73.9% (165,589) were reported by the 1kG project.

gnomAD contains information for a considerably larger number of samples than 1kG [64,68], however, the number of variants found in common with our data was lower. This might be explained by the usage of its exomic files. Those files include information for a considerably larger number of individuals than its genomic files (123,136 against 15,496), yet, less positions were screened for variants as exome sequencing spares a lesser extent than genome sequencing. [180]

By focusing the comparison in exomic variants, the percentages of variants not reported by each project were more concordant to the expected scenario. gnomAD presented both a lower number and percentage of unreported variants among the variants contained in exomic regions (21.7% for gnomAD against 37.3% against 1kG; Table 6).

The term “unknown” (Table 6) will be used when discussing variants that are not reported among the information extracted from both projects, assessments of a single project will be declared

Table 6. Number of variants reported by each genomic project. Percentages of unknown variants were calculated in comparison to the total number of variants for each column and project. Percentages of exomic variants were calculated against the values of the column "Total".

Variants	Total	Exomic	% Exomic
gnomAD	151,053	93,247	61.7%
▪ Unknown	73,102	20,203	27.6%
% Unknown	32.6%	17.8%	
1kG	165,589	82,853	50.0%
▪ Unknown	58,566	30,867	52.7%
% Unknown	26.1%	27.1%	

4.4 Unknown Variants

16.4% (36,732) of the filtered variants were unknown. This value was concordant to previous studies results, considering the number of individuals analysed (Table 7). The obtained value is higher than the results reported for 12 individuals of the Yakut population from North East Asia (7.6%) [69] and lower than the results presented for 128 Ashkenazi Jews (18.3%) [70], 267 Spanish (approximately 33%) [1], for 250 Dutch families (14.6%) [5] and 3781 British individuals (approximately 57%) [4]. A consistent relationship between the number of individuals included in each cohort and the percentage of novel variants detected among them was evidenced.

Table 7. Comparison of new variants reported by 5 populational endeavours against the results of the present work (highlighted in grey). The name for Ashkenazi Jews in the fourth column has been abbreviated to A. Jews. GoNL and UK10K represent, respectively, a Dutch and a British national-level projects. Populations are ordered by the number of samples involved in the respective study.

Population	Yakuts	PT	A. Jews	Spanish	GoNL	UK10K
Number of individuals	12	70	128	267	250 families	3,781
New variants	7.6%	16.4%	18.3%	33% (approx.)	38% (approx)	57% (approx.)

Among the reported unknown variants, 45.3% (16,624) are exomic, 9.8% (3,616) are indels and 75.0% (27,547) present an allele frequency below 1% (Table 8), which meets the expectation that most of them would be found at low frequencies. This value is much lower than the value reported by the UK10K project (99.9%) [4] and the Spanish population (85.6%) [1], on the other hand, it is concordant with the number of singletons among novel variants in GoNL (75.6%) [5]

Table 8. Distribution of unknown variants by allele frequency in Portuguese population.
A) Portuguese AF values grouped in intervals of 10%, the percentages presented correspond to the number of variants among all unknown variants that report 60 or more genotypes and present HWE conditions.

AF(%)	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Count	35,785	725	88	44	27	19	9	7	14	14
%	97.42	1.97	0.24	0.12	0.07	0.05	0.02	0.02	0.04	0.04

B) Portuguese AF values below 10% grouped in intervals of 1%. The percentages presented correspond to the number of variants in relation to the total number of unknown variants.

AF(%)	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10
Count	27,547	3,672	1,996	846	445	459	254	270	170	126
%	74.99	10.00	5.43	2.30	1.21	1.25	0.69	0.74	0.46	0.34

By comparing unknown low-frequency variants with the remaining variants, it was revealed that less severe classifications become more prevalent among the variants that present higher frequency values (Table 9). This might be explained by the lower selective pressure against those variants as they do not cause a phenotypic difference between individuals. [7,8]

A hypothesis to explain why those variants were not found among other populations consists in the appearance of that specific alteration among the Portuguese population throughout its history and the prevalence of the alteration in the succedent generations due to its low selective pressure. Said so, these variants may be eventual populational markers.

Table 9. Classification of unknown filtered variants accordingly to VEP classification for variant impact and severity (Sev.). Variants are divided into three groups, low-frequency variants, with an AF below 1%, variants with a frequency between 1% and 10% and variants with a frequency above 10%. This distinction shall reflect how variant impacts influences the probability to find more frequent alterations. The values in "%" columns are calculated in relation to the total number of variants found in each frequency interval.

Sev.	**Impact**	**AF < 1**	**%**	**AF 1-10**	**%**	**AF >= 10**	**%**
LOW	Intron Variants	13,278	48.20	4,397	53.37	549	57.97
	Synonymous variant	2,736	9.93	737	8.95	84	8.87
	Other low sev. variants	2,074	7.53	644	7.81	76	8.03
MED	Missense variant	7,156	25.98	1,852	22.48	171	18.06
	Other medium sev. variants	928	3.37	274	3.33	22	2.32
HIGH	Frameshift variant	904	3.28	242	2.94	32	3.38
	Other high sev. variants	471	1.71	92	1.11	13	1.38
	Loss-of-Function variants	1,362	4.94	330	4.01	42	4.44

On the other hand, and accordingly to the influence of the effect of the variant in its prevalence, variants classified by VEP as presenting a medium severity were progressively scarcer as AF ranges augmented. In the case of missense variants, SIFT [128] and Polyphen [129] provided predictions for its effects, however, there were discordances between both predictors (Table 10, yellow cells), 24.1% presented a damaging/deleterious effect in one of them and a benign/tolerated in the other. As it is, these results did not present general conclusive findings, yet, the percentage of Benign/Tolerated classifications (42.1%) suggest that a considerable amount of those variants may also be subject to low selective pressures and may constitute eventual populational markers.

Table 10. SIFT and Polyphen effect predictions of missense variants with Allele frequencies above 1%. Each document presents both predictions, the cell values represent the number of variants that share each combination of predictions. Benign/Tolerated predictions by both predictors are highlighted at green, damaging/deleterious at red and discordances at yellow.

		Polyphen			
		Benign	**Possibly_damaging**	**Probably_damaging**	**Unknown**
SIFT	**Tolerated**	33.1% (670)	5.8% (117)	3.0% (60)	0.1% (2)
	Tolerated low_confidence	9.0% (183)	0.9% (18)	0.4% (8)	0.3% (6)
	Deleterious low_confidence	4.2% (85)	1.7% (35)	2.2% (45)	0.1% (3)
	Deleterious	9.8% (198)	8.4% (169)	18.5% (374)	0
	None	0.9% (18)	0.4% (9)	0.6% (13)	0.5% (10)

Finally, regarding the variants classified by VEP as presenting a high severity, the higher percentage of frameshifts at larger frequency values would not be expected [181,182], this value might have been influenced by the reported error propensity associated to indel detection. [179] As it is, eventual analyses and conclusions regarding populational comparisons of this type of variants should be made cautiously. In a clinical point of view, it is worthy of notice that 7 of the 53 filtered unknown variants reported in ClinVar were included in the gene ADAMTSL2 on chromosome 9, that variants are associated by ClinVar to Geleophysic dysplasia. There were not known, by the time of writing, any populational comparisons concerning variants on this specific gene.

Although, only one of them constitutes a missense variant, this specific alteration will not cause a phenotypical manifestation by itself as the condition is inherited accordingly to an autosomal recessive pattern [183]. The alteration was found in a single heterozygous sample.

4.5 Low Frequency Variants

Among the 224,155 filtered variants, 41.9% (93,902) display a Portuguese AF below 1%. This value is considerably lower than the proportions of approximately 80% obtained for 3,781 genomes and 99% for 60,706 exomes respectively reported by the projects UK10K [4] and ExAC [67].

The value was also lower than the percentage of Minor Allele Frequency values below 0.5% presented by 1kG, 72.7% [64] and GoNL, 50% [5]. Allele frequencies for the present work do not represent MAFs, nonetheless, a great discrepancy of the values presented may still be noticed.

This confirms that the lower number of samples included in a cohort limits the capability to assess the rareness of its variants. As it was not possible to assess the rareness of the detected variants further the first decimal point, the following comparisons will not consider how rare is any variant in either gnomAD or 1kG. All values below the convention 1% [11,15] will be included in a single comparison group.

Low frequency variants were distributed in 29.3% (27,547) unknown variants, 41.1% (38,589) variants reported in both 1kG and gnomAD, 9.9% variants (9,339) only reported by 1kG and 19.6% (18,428) exclusively reported by gnomAD (Table 11).

Table 11. Distribution of the European population AF values reported by gnomAD and 1kG for the variants that present a Portuguese AF value below 1%. Only filtered variants were included in the comparison. As both projects include other population besides the European, the latter may present null frequencies for some variants, these variants are distinguished from the remaining variants. The red cells in the first row represent the variants that are absent from 1kG only, red cells in the first column represent the variants that are not found among gnomAD exomes. Green cells correspond to all variants reported by both projects, the bottom-right cells contain the sum of those values. White cells correspond to the sum of values for its respective row/column, percentages were calculated in relation to the number of variants included in the present comparison reported by the respective project.

		gnomAD				
		Not Found	0%	<1%	>=1%	**Total**
1kG	Not found	27,547	4,249	14,018	161	18,428
	0%	2,657	525	10,677	5	13,864 (28.9%)
	<1%	5,068	56	20,442	1,176	26,742 (55.8%)
	>=1%	1,614	2	982	4,724	7,322 (15.3%)
	Total	9,339	4,832 (8.5%)	46,119 (80.9%)	6,066 (10.6%)	38,589

Overall, 86.9% (57,692) of the known Portuguese low-frequency variants were concordantly reported with a low, or null, European allele frequency in the databases where they were found.

Among the 38,589 variants reported by both databases, 5.6% (2,165) presented an AF above 1% in only one of them, 2.5% (984) in the 1kG European population and 3.1% (1,181) in gnomAD Non-Finnish Europeans. Nonetheless, 94.5% (930) and 97.4% (1150) respectively of those cases present an AF value below 2% (not shown in Table 11).

The obtained values suggested that, although the presented data could not be analysed for its rareness, a differentiation between common and low-frequency variants might be established. Therefore, the analysis of the 12.2% of the variants present in both databases (4724) with an AF value above 1% assumed a greater interest.

Most of those variants – 79.0% – displayed an AF value below 3% in both compared populations (Figure 11), besides it, there was an observable diminution of the number of variants accounted as the AF intervals were considered for higher values. At this point, it was plausible to believe that some of these differences were effectively caused by a limited sampling [184] that increased the probability to produce deviations to the real populational scenario. This finding constitutes an indication that a larger number of samples might lead to a general approximation to the values found in the high-scale projects compared. This would provide a stronger certainty to indicate effective differences. Still, some abnormally discrepant values may be observed, 2% of the variants present an AF value above 10% for, at least, one of the European populations. 5.6% present values above 5% (Figure 11).

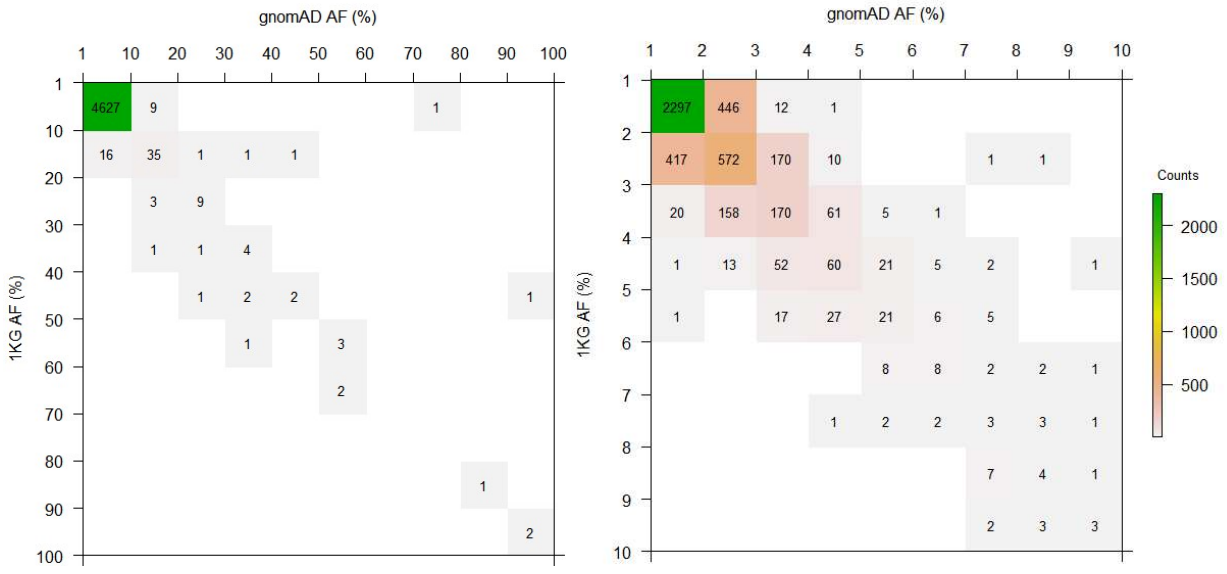


Figure 11. Distribution of variants AF values in both 1kG and gnomAD European populations. All variants used to construct this graphic present a Portuguese allele frequency below 1%. Left graphic represents the distribution of values by intervals of 10 percentual units. The second graph displays the number of variants with an AF value above 1% and below 10% for both projects grouped in intervals of 1%.

There could be expected more discrepancies between 1kG and gnomAD AF values since its respective European populations were used for the comparison. In the case of gnomAD, the Finnish population (FIN) constitutes an independent group in relation to Non-Finnish Europeans (NFE) [68], by its turn, 1kG include 99 Finnish individuals as a European subpopulation accounting approximately one fifth (19.7%) of the EUR population [65]. This could influence the comparison of AF values between both groups, nevertheless, it does not seem to cause atypical discrepancies. This might be intended as another benefit of larger sample sizes, while a subset may present a different frequency of a variable in relation to a larger group, yet, after combining both groups, that difference would only be noticeable if the difference was pronouncedly discrepant.

Table 12. Impact and severity (Sev.) classification provided by VEP for the variants that present frequency values below 1% for the Portuguese population and above 1% for both 1kG and gnomAD European populations. The four last columns present, respectively, a distribution for the entire set and for the subsets of variants that present AF values above 5, 10 and 25% for both European populations in comparison.

Sev.	Impact	>1%	> 5%	> 10%	> 25%
LOW	Intron Variants	1,593	100	39	13
	Synonymous variant	1,122	11	2	0
	Upstream gene variant	11	0	0	0
	Downstream gene variant	6	2	2	2
	Non-coding exon variant	20	3	3	3
	3-prime UTR variant	136	6	3	0
	5-prime UTR variant	98	7	4	1
	Stop retained variant	2	0	0	0
MED	Inframe deletion	17	4	3	0
	Inframe insertion	15	5	4	0
	Missense variant	1,432	22	5	2
	Splice region variant	221	8	4	1
HIGH	Frameshift variant	20	5	1	0
	Splice donor variant	8	1	1	0
	Splice acceptor variant	5	0	0	0
	Start loss	2	0	0	0
	Stop loss	2	0	0	0
	Stop gain	14	1	0	0

4.6 Population Comparison

AF values of Portuguese samples and other populations have been compared by scatterplots.

These graphics can be used as an indication of the genetic distance between two populations by the distribution of its points, narrower bulks of dots indicate a low variation of AF values in both populations. Few genetic differences in relation to European populations in gnomAD and 1kG were revealed (Figure 12). Interestingly, the plot of general AF values in respect to gnomAD is similar to the one obtained to its European population, this may be explained by the high percentage of NFE samples in relation to the total number of samples (45.4%) [185]. European samples in 1kG correspond to 20.1% of the total [172].

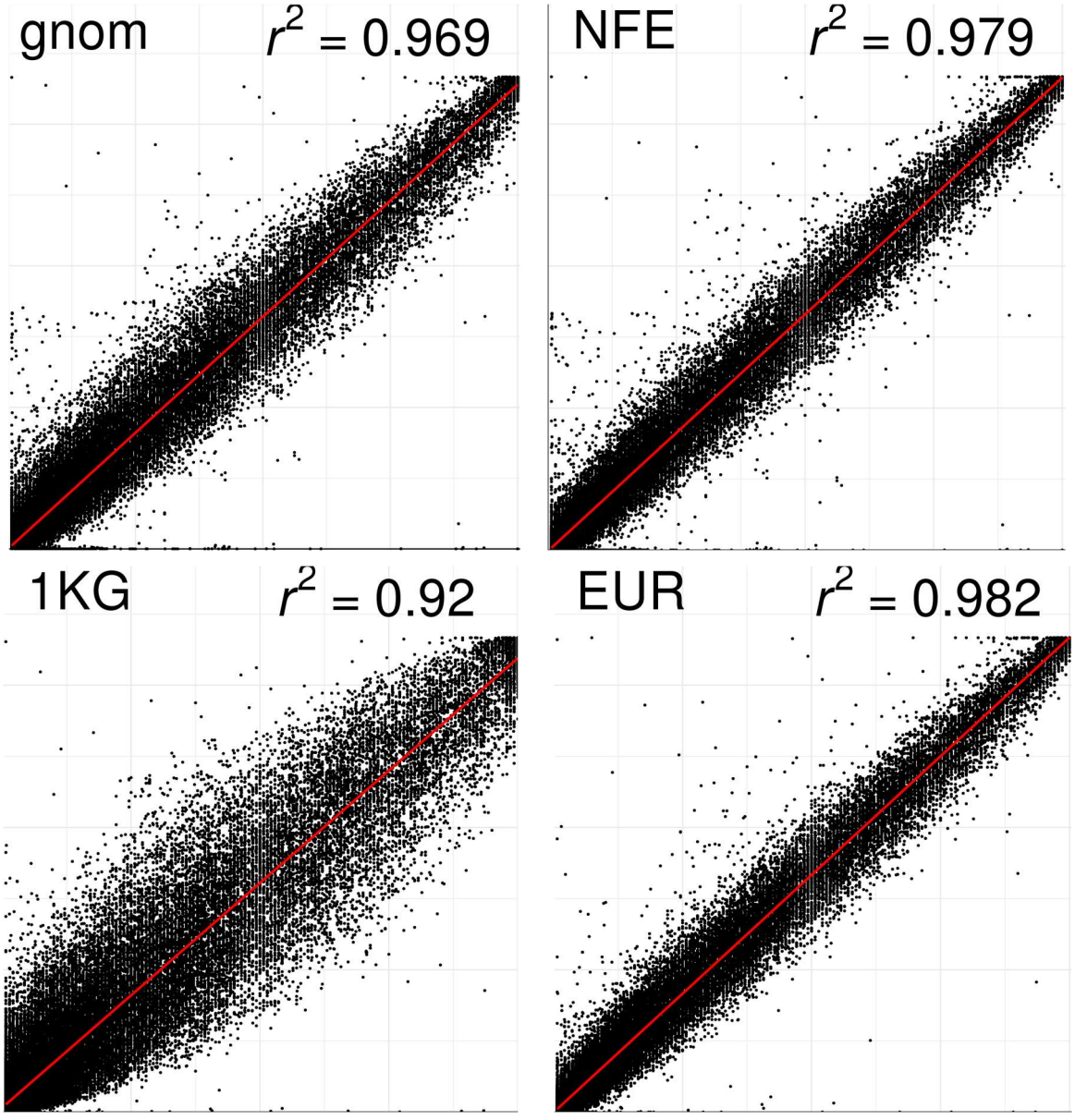


Figure 12. Comparison of Portuguese allele frequency values to the allele frequencies obtained from 1kG and gnomAD. Left-to-right and top-to-bottom, the presented scatterplots correspond, respectively to comparisons against gnomAD general population, gnomAD Non-Finnish European population, 1kG general population and 1kG European populations. Axis display AF values in decimal scale representation.

The same analysis have been performed for all populations in both projects (Figures 13 and 14), overall, the evidenced genetic differences support actual evolutionary models of world peopling [33].

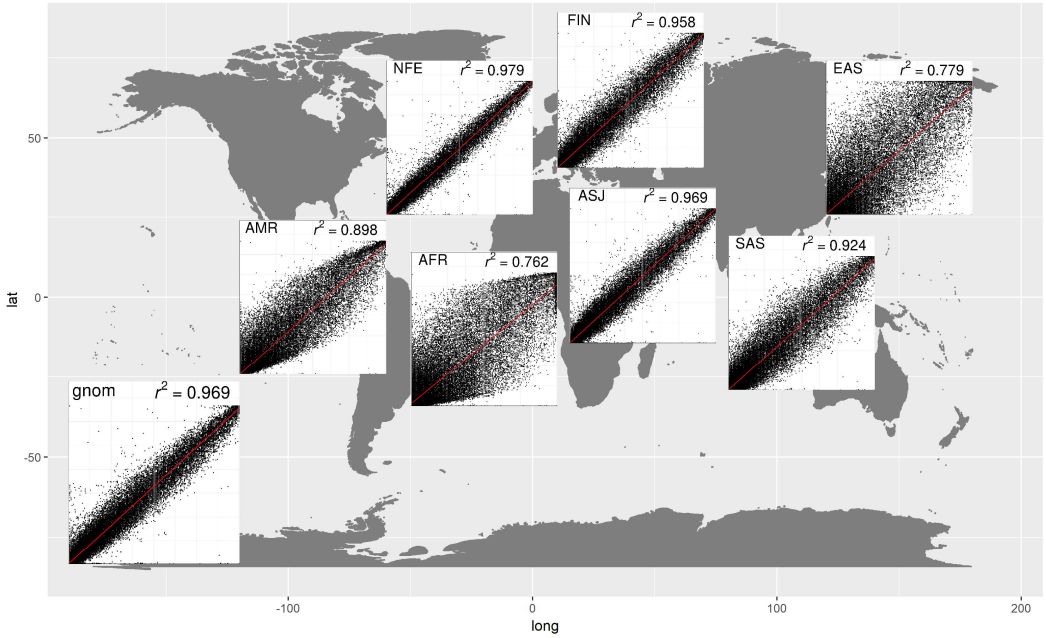


Figure 13. Allele Frequency scatterplots comparing Portuguese population against each gnomAD population. Plots are represented in its correspondent regions. The uppermost plot corresponds to Finnish (FIN) population, the bottommost to African/African Americans (AFR) and the other five plots represent, respectively, from left to right, Admixed Americans (AMR), Non-Finnish Europeans (NFE), Ashkenazi Jewish (ASJ), South Asians (SAS) and East Asians (EAS). The bottom-left scatterplot corresponds to general gnomAD population.

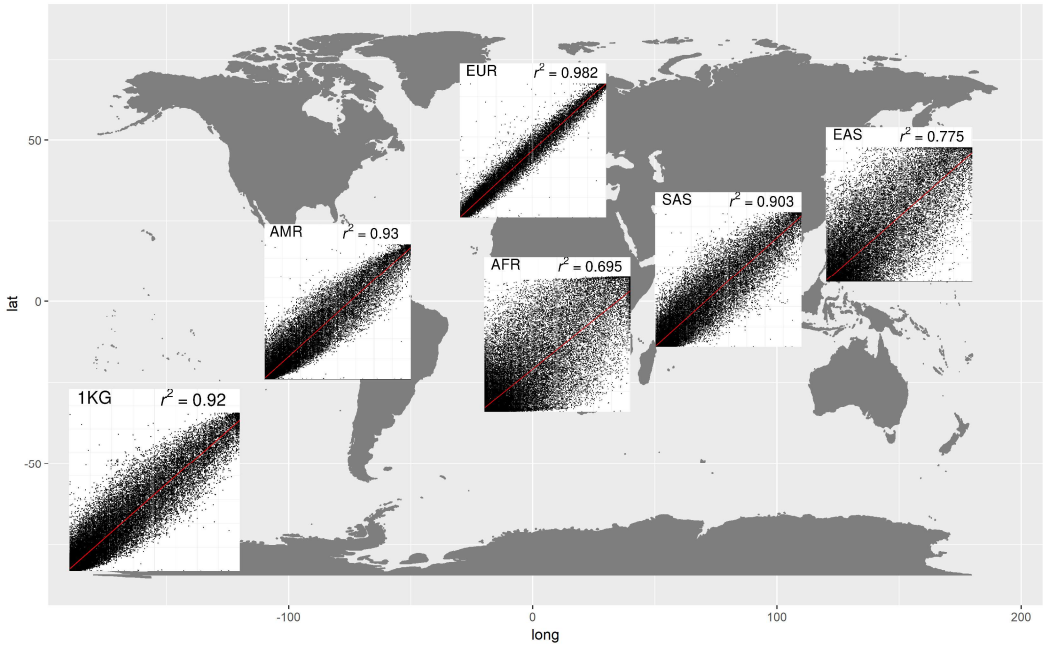


Figure 14. Allele Frequency scatterplots comparing Portuguese population against each 1kG population. Plots are represented in its correspondent regions. Plots represent, respectively, from left to right, Americans (AMR), Europeans (EUR), Africans (AFR), South Asians (SAS) and East Asians (EAS). The bottom-left scatterplot corresponds to general 1kG population.

As it was expected, European populations in both gnomAD (Figure 13) and 1kG (Figure 14), present the narrowest distribution of frequencies against Portuguese individuals. It could be suggested that Finnish are genetically more distant to Portuguese individuals in relation to the Non-Finnish Europeans (Figure 13), it is concordant with the suggestions that Northern Europeans and the rest of Europe descend from distinct ancient populations [46].

Both figures presented a great genetic distance to African population. The Out-of-Africa model [33,34] suggests that a single East African population migrated into the Middle East, constituting a founder population that represented only part of the African genetic variability. gnomAD map suggests a proximity to the Ashkenazi Jewish population, a representative group of that region from where migrated the first inhabitants of Southern Europe [32,46,54].

Additionally, both Asian populations present a proportion between genetic and geographical distance in both figures. Accordingly to suggested expansion models, Indigenous Americans would descend from ancient Siberians [33], however, colonization of the Americas by European civilizations caused a genetic approximation to European populations. Indications of a noticeable European ancestry in AMR 1kG populations have been reported [186]. This relation is reflected in both figures.

Once our database also includes AF values for the 1kG European subpopulations, the same procedure has been applied to establish European-level comparisons (Figure 15).

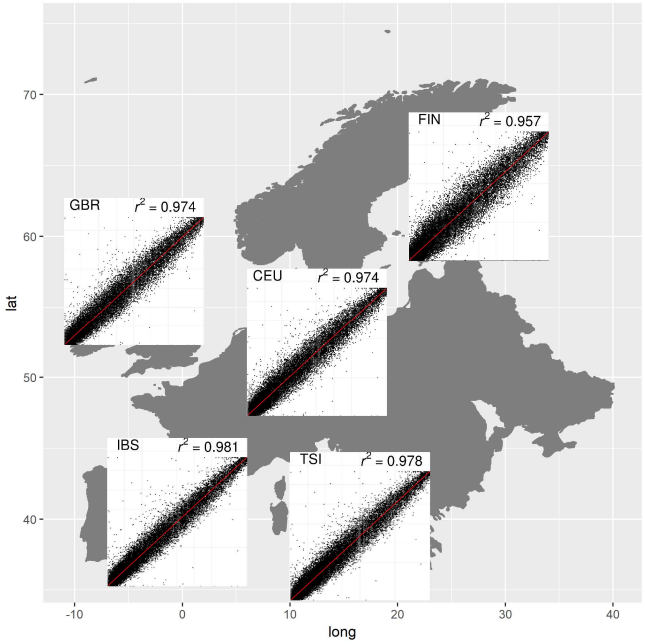


Figure 15. Allele Frequency scatterplots comparing Portuguese population against each European 1kG subpopulation. Plots are located above the regions that they represent. The uppermost plot corresponds to Finnish (FIN) population, the other four plots represent, respectively, from left to right and top to bottom, British from Great Britain (GBR), Central European from Utah (CEU), Iberians from Spain (IBS) and Tuscans from Italy (TSI). The bottom-right scatterplot correspond to the complete European population from 1kG.

A distinguishing genetic distance to Finnish individuals and a low genetic divergence among four of the five European 1kG subpopulations and our Portuguese samples were presented (Figure 15). Plots similarity suggested a genetic profile proximity of PT to either CEU, GBR, IBS and TSI populations. Nonetheless, specific differences may constitute a differentiation factor to some or all populations, as it is, the genetic profile of population individuals may be compared to denote population-related discrepancies.

4.7 Principal Component Analysis

A Principal Component analysis has been performed with Portuguese and European 1kG genotypes, subpopulations of the latter were discriminated. For this procedure, all the genotypes for the variants in HWE for the Portuguese population which are also reported by 1kG, presented an AF value for the European population above 0.1% and its document include, at least, 453 European – the total number of EUR individuals is 503 – and 60 PT genotypes were included. As in the case of the threshold of 60 Portuguese genotypes, the last condition has been stipulated to include an error margin for the number of genotypes detected.

A.

PC	EigenValue
1	50.22
2	23.13
3	19.90
4	19.23

B.

Figure 16. A) Eigenvalues by principal components. B) The 4 first scores are discriminated.

As it may be observed in the first obtained PCA (Figure 17), PC1 closely grouped Portuguese population with IBS and TSI populations. By its time, PC2 presented a slight differentiation in relation to IBS.
PC3 evidenced a detachment in opposing directions of Portuguese and Tuscan individuals from the remaining 1kG European populations (Figure 18). PC4 did not suggest any inter-populational difference as it clustered all populations together. By combining the three first components, a three-dimension

general view was created that suggested that the Portuguese individuals are most related to IBS population. (Figure 18.B)

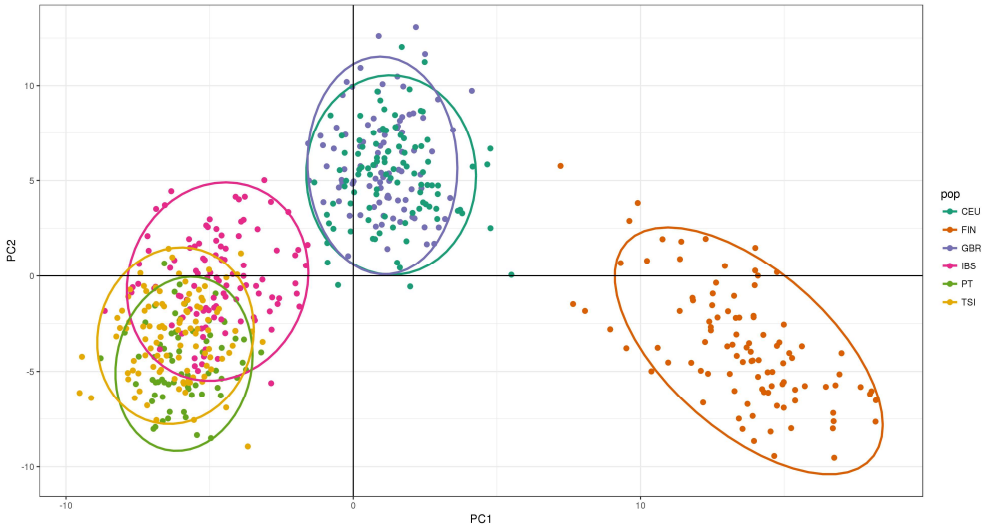
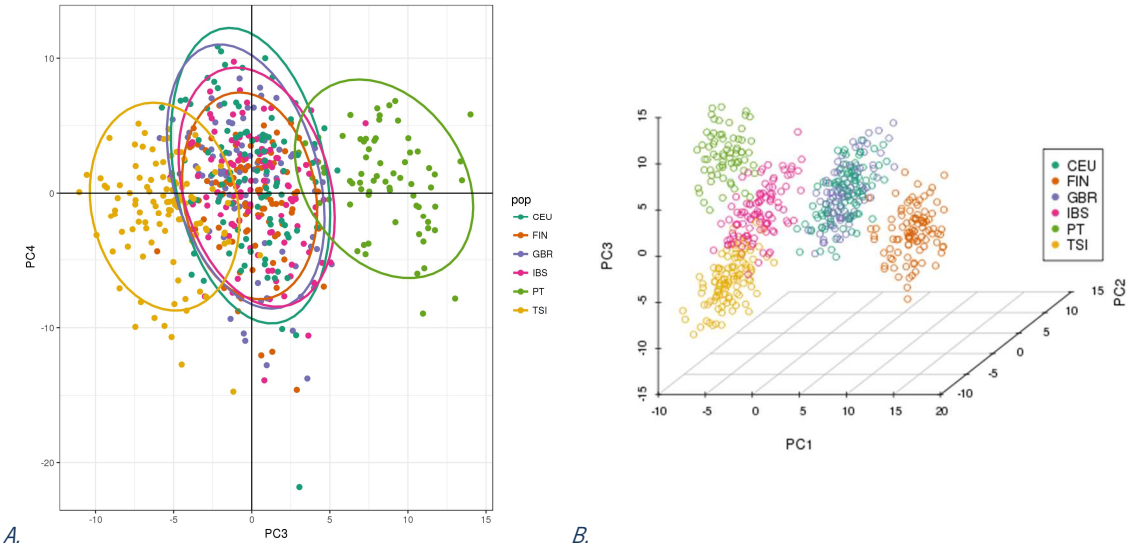


Figure 17. Scatterplot of the first two components. PC1 distinguish individuals horizontally, PC2 vertically.



A. *B.*

Figure 18. A) Scatterplot of the components 3 and 4. PC3 distinguish individuals horizontally, PC4 vertically. B) 3D scatterplot of the first three components, PC1 scores determines individuals position over the x-axis, PC2 on the y-axis and PC3 on the z-axis.

As the previous scatterplots revealed, Finnish population is the most distinct group among 1kG European subpopulations. The distance to the closest Non-Finnish groups that may be observed in Figure 18 suggest that its populationally characteristic variations are the most influential differences in PC1. This finding would be concordant to hypothesis that postulate the recolonization of Europe after the Last Glacial Maximum from southwestern refugia [32,45,46], as an apparent proportion is denoted between the genetic distance of the three clusters obtained from the Principal Component 1 and the geographic

distance of the European regions where each cluster may be situated – PT, IBS and TSI in South-western (SW) Europe. CEU and GBR in Central Europe and FIN in Northern Europe.

PC2 support a slight differentiation between SW and Central European groups, however, FIN individuals overlap SW populations in this component, suggesting that there may exist characteristic variants for Central European group. Finally, while PC3 depicts a detachment among SW group populations, the strong association of IBS to GBR, CEU and FIN populations does not permit to draw conclusions about regional-specific comparisons, nonetheless, the isolation of Portuguese individuals in relation to the other populations compared may signify the existence of potential genetic markers for Portuguese population.

4.8 Differences against Europeans in 1kG

4.8.1 Allele distribution differences

Using alternative allele and total allele counts for all 503 samples in the European population of 1kG it has been possible to ascertain the statistical significance of the difference between allele frequencies for this group and Portuguese population. Only filtered variants were included in this procedure.

Overall, 165,589 filtered variants were tested for difference against each population with Fisher's exact test, after adjusting p-values for False Discovery Rate (FDR). A significant difference was obtained to, at least, one population for 4.4% (7,284) of the tested variants. 8.5% of them (619 variants) displayed a significant difference to all the tested populations.

Analysing by population, there were vast differences to Finnish individuals in comparison to the other four populations tested (Table 13 and Figure 19), 73.0% of the differences found were exclusively significant for that group. This finding corroborate the previous results that suggest a distinctive genetic distance of Finnish individuals in relation to any other European population; this difference endorses the separation of both groups as it is done in gnomAD project.

It was also suggested that Finnish population differences conceal some eventual relations regarding the differences of Portuguese individuals to the other European populations (Figure 19 - red circle), a considerable number of variants present significantly different allele distributions in relation to Portuguese populations for GBR or CEU individuals but not for IBS or TSI, however, these relations are not reflected by the number of exclusively different variants (Table 13).

To asses these differences and be able to compare and evidence divergences among the closest populations, Finnish differences were excluded of the analysis. This filter would expectably avoid drawing

general European-level conclusions based on values that may be biased by the high number of exclusively different variants for Finnish.

Table 13. Number of variants with significantly different allele distributions in relation to the Portuguese samples. Successive analyses are presented. For each population comparison, the total number of Fisher test p-values below 0.05 is presented in the second column. The number of null hypothesis rejected after p-value correction for false discovery rate (FDR) in third column, the percentage of variants that these numbers represent among all variants tested (165589) is presented in the fourth column. Fifth and sixth columns contain, respectively, the number of variants that only present a significantly different allele distribution for a sole population and its percentage among the total results corrected for FDR (7284 variants).

Population	p-value < 0.05	Corrected for FDR	% of total	Exclusively different	% of total corrected
IBS	7,433	987	0.60%	57	0.78%
TSI	9,634	1,071	0.65%	137	1.88%
GBR	12,148	1,220	0.74%	140	1.92%
CEU	11,611	1,285	0.77%	172	2.36%
FIN	24,042	6,610	3.99%	5,320	73.0%

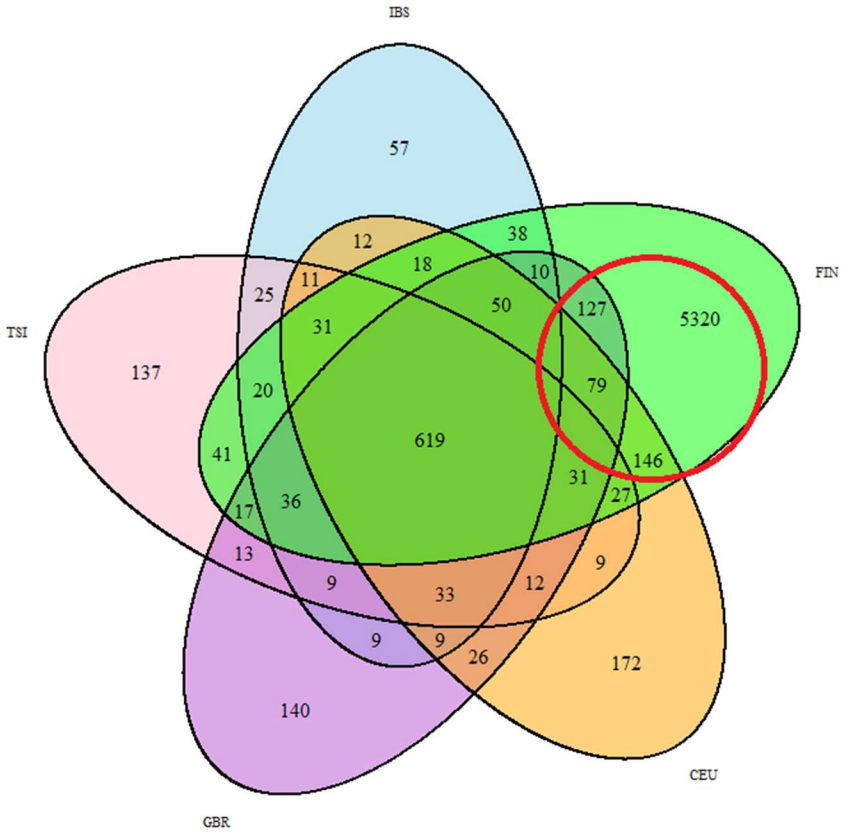


Figure 19. Venn Diagram for variants with significantly different allele distributions for each population after correction for FDR. Red circle highlight the larger bulk of variants that are not different for all populations, it include the variants are simultaneously different for FIN and GBR or CEU populations and do not present differences to IBS or TSI individuals.

27.0% (1,964) of the initially reported variants presented significant differences to, at least, one of the populations, 33.2% of them (652 variants) presented a significant difference to the four populations (Figure 20).

In the same way as the results obtained in previous large-scale comparisons, this approach denotes progressively larger amounts of exclusive differences that reflect a relationship between the genetic and geographic distance (Table 14). The most noticeable finding is that the expected higher proximity of Portuguese samples to IBS population is corroborated regardless of the approach. Additionally, results presented in Table 14 comply to previously reported relations suggested by Principal components 2 and 3, a larger genetic distance of Portuguese individuals to both GBR and CEU corroborates the distances between these groups displayed in both components. Differences between South-Western (Portuguese, IBS and TSI) populations are also concordant with the relations presented for those groups in PC3.

Table 14. Recalculation of the number of variants that only present a significantly different allele distribution for a sole population and its percentage among the results corrected for FDR for that population in a subset that do not account for Finnish differences to Portuguese individuals. Population location is presented for each group to reflect the geographical distance to Portugal, located in the South-westernmost extremity of Europe.

Population	Exclusively different variants	% of total corrected	Population Location
IBS	95	3.02%	South-western Europe; Bordering Portugal
TSI	178	5.66%	South of Europe; Mediterranean Coast of Italy
GBR	267	8.49%	North-western Europe; British islands
CEU	318	10.11%	USA; Northern and Western European ascendancy

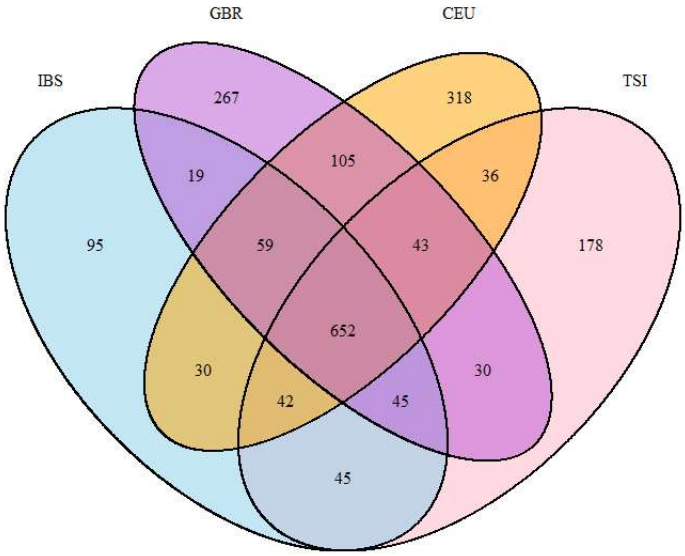


Figure 20. Reformulation of the Venn Diagram for variants with significantly different allele distributions for each population after excluding Finnish data.

4.8.2 European genetic differentiation

Since genetic differences has been reported throughout the present work at a scale-dependent variable extent, these differences may be analysed under diverse perspectives.

Availing this dataset and the obtained p-values, significative differences were found for the allele distribution between our population and the other 1kG European subpopulations in 7,284 variants distributed by 2,571 genes. Differences were grouped by gene and the number of differences in relation to each population revealed diverse patterns of gene variation across European populations (Table 15).

Table 15. Count of variants with significant differences to each population by gene. Results against FIN population were not included. Only the 21 genes with highest number of alterations reported in the results are presented. All those genes denoted differences for, at least, 8 variants among the four populations. Population-gene pairs that reported, by itself, 5 or more variants among these results were highlighted at red, on the other hand, population-gene pair that reported none or a single difference to Portuguese individuals was highlighted at green.

Gene	IBS	TSI	GBR	CEU	Total
MICA	8	7	11	18	18
SDHA	0	16	4	15	16
HLA-G	7	7	8	12	13
HERC2	1	1	11	10	13
ANKRD36	8	0	2	3	11
MYOM3	5	6	11	5	11
COL6A1	10	9	6	9	10
MUC5B	3	4	2	8	9
KRT38	3	3	9	6	9
ADH1C	1	1	6	9	9
HLA-B	1	1	8	9	9
KIR3DL1	8	6	7	8	8
GPATCH1	7	2	1	8	8
AHRR	6	1	7	6	8
C9orf84	2	2	8	7	8
RESP18	1	1	2	8	8
MRS2	1	1	8	1	8
DHX38	1	0	1	8	8
PIEZO2	0	8	6	6	8
LCT	0	8	7	8	8
PLK2	0	0	8	7	8

Among the 21 genes with the highest number of detected differences, some genes have been already reported in the literature as highly polymorphic, such as KIR3DL1, HLA-B and HLA-G [63,187–189], and genes that had been already used for genetic populational comparisons as SDHA [190] and ANKRD36 [191], none of the studies found for each case included Portuguese individuals. Additionally, studies for two other genes comprised Spanish individuals, Europeans samples from 1kG were included in a comparative analysis of allele frequencies for the gene AHRR [192] and the allelic diversity of the gene MICA has previously been analysed in a population of the region of Murcia, Spain [193].

Finally, LCT gene, from chromosome 2, a previously confirmed genetic marker [194,195] may be presented as a paradigmatic case of the power for this type of analysis. Significant differences were found for 8 variants from this gene. None of them presented those differences in relation to IBS populations, all 8 presented those differences in relation to TSI and CEU, and excepting one of them, all were significantly different in comparison to GBR (Table 16).

The most noticeable finding is the consistent pattern presented by the allele frequencies reported for each population – TSI < PT < IBS < GBR < CEU for all but one case, for which the order is reverted. Besides confirming an accordance with previous results.

Table 16. LCT gene variants that reported significantly different allele distributions in relation to, at least, one population. Allele frequencies for Portuguese individuals is highlighted in grey. Variant-population pairs for which significant corrected p-values were reported are highlighted in red for the cases where the correspondent allele frequency values are lower than the value calculated for Portuguese individuals and in green for the opposite case.

Position	Ref	Alt	PT AF	IBS AF	TSI AF	GBR AF	CEU AF
136546110	A	G	57.1%	65.0%	41.1%	75.8%	81.8%
136555659	T	C	51.5%	62.2%	36.9%	75.3%	80.8%
136558157	C	T	50.0%	57.5%	25.7%	73.1%	79.8%
136561557	G	A	71.4%	74.8%	50.5%	81.9%	87.4%
136569848	C	A	50.0%	57.5%	25.7%	73.1%	79.8%
136575199	G	T	50.0%	57.5%	25.7%	73.1%	79.8%
136590746	C	T	27.1%	21.5%	43.5%	15.9%	11.6%
136594158	G	A	49.3%	57.9%	25.7%	73.1%	79.8%

In this case, an Iberic relation is confirmed between Portuguese and IBS populations. Additionally, as the relations presented were found to be concordant with the genetic distribution of a well reported marker, this finding suggest a potentiality of the present dataset to find new genetic markers in further analyses.

4.9 Work relevance

The presented work was developed under the scope of the In2Genome project. This project aims, among other things to improve diagnosis of congenital diseases.

The combination of all reported genotypes and the information regarding location, gene and predictions provided by GEMINI enable a more direct and complete assessment of the variants.

As it is, the constructed database may constitute a relevant auxiliary reference for genetic analysis of Portuguese patients. The database is scalable and may include information for more samples sequenced in the future, it would benefit further analyses and would imporve the accuracy of the eventual diagnoses.

Additionally, the unknown variants reported constitute an increase on the genetic information available and may prompt future studies to constitute Portuguese-specific genetic markers.

5. CONCLUSION

The present work successfully created a database that enhanced diverse analyses and comparisons, the storage of the genotypes as accessible and manageable repositories was the principal advantage of this resource. Compilation of information from multiple sources enabled the assessment of the studied data under diverse perspectives. The 224,155 variants found on 70 Portuguese exomic samples constituted the basis for the present work. Among those variants, 16.4% constitute novel variants, a proportional value in relation to previous studies on different sample sizes. Comparisons to the same projects report that a lesser number of low frequency variants is found among the present work data, however, there was a concordance on the percentage of singletons found by the studies.

These findings and the low representativity for the present dataset in relation to the variants rareness allow to conclude that an eventual project containing a larger number of samples would approximate the obtained results to the scenario presented by large-scale endeavours.

At a populational point of view, the analysed data corroborate established hypothesis for historical migrations and its respective influence on evolution and genetic differentiation.

Notwithstanding, significative differences could be found despite the similarities presented between the Portuguese individuals and previously described 1kG European populations.

Among those differences, previously described genetic markers have been found and populational relations are plausible, thus, the present work may prompt future studies to confirm genetic differences and establish Portuguese-specific genetic markers.

Epitomizing, the present study represents a significant contribution to enrich large-scale genomic initiatives with complementary information and may stand as a useful auxiliary reference for genetic analyses of Portuguese patients.

BIBLIOGRAPHY

1. Dopazo J, Amadoz A, Bleda M, Garcia-Alonso L, Alemán A, García-García F, et al. 267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic Variation. Mol Biol Evol. 2016;33(5):1205–18.
2. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1 , 092 human genomes. Nature. 2013;491(7422):56–65.
3. Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, et al. Correlation between Genetic and Geographic Structure in Europe. Curr Biol. 2008;18(16):1241–8.
4. Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015;526(7571):82–9.
5. Francioli LC, Menelaou A, Pulit SL, Van Dijk F, Palamara PF, Elbers CC, et al. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet. 2014;46(8):818–25.
6. Pritchard JK, Pickrell JK, Coop G. The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. 2010;20(4):208–15.
7. Hartl DL, Clark AG. Principles of population genetics. Third. Sinauer Associates, Inc; 1997. 542 p.
8. Peter J. Russell. iGenetics. Vol. 53, Journal of Chemical Information and Modeling. 2013. 1689-1699 p.
9. Gillespie JH. Population Genetics. The Johns Hopkins University Press; 1998. 169 p.
10. Alberts B. Biologia Molecular da Célula. Fifth. S.A. AE, editor. 2010.
11. Karki R, Pandya D, Elston RC, Ferlini C. Defining “mutation” and “polymorphism” in the era of personal genomics. BMC Med Genomics. 2015;8(1):1–7.
12. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J. Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. 2015;17(5):405–24.
13. Collins F, Collins F, Brooks L, Brooks L, Chakravarti A, Chakravarti A. A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation. Genome Res. 1998;8(12):1229–31.
14. Kitts A, Phan L, Minghong W, Holmes JB. The database of short genetic variation (dbSNP). NCBI Handb [Internet]. 2013;(2nd). Available from: http://www.ncbi.nlm.nih.gov/books/NBK174586/
15. Brookes AJ. The essence of SNPs. Gene. 1999;234(2):177–86.
16. Shastry BS. SNP alleles in human disease and evolution. J Hum Genet. 2002;47(11):561–6.
17. Sherry ST, Ward M, Sirotkin K. dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. Genome Res. 1999;9(8):677–9.
18. Sherry ST. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–11.
19. NCBI NC for BI. dbSNP Short genetic variation [Internet]. [cited 2017 Dec 29]. Available from: https://www.ncbi.nlm.nih.gov/SNP/
20. Crick F. Central dogma of molecular biology. Nature. 1970;227(5258):561–3.
21. Crick FHC, Barnett FRSL, Brenner S, Watts-Tobin RJ. General nature of the genetic code for proteins. Nature. 1961;189:1227–32.
22. Rosenberg L, Rosenberg D. Human Genes and Genomes. First. Academic Press; 2012. 446 p.
23. Mullaney JM, Mills RE, Stephen Pittard W, Devine SE. Small insertions and deletions (INDELs) in human genomes. Hum Mol Genet. 2010;19(R2):131–6.
24. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. 2006;16(9):1182–90.
25. Griffiths AJF, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM. An Introduction to Genetic Analysis [Internet]. Seventh. W. H. Freeman and Company; 2000. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21766/
26. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segrè AV, Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, Heard-Costa NL, Rand HJ. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010;467(7317):832–8.
27. Phelan CM, Kuchenbaecker KB, Tyrer JP, Kar SP, Lawrenson K, Winham SJ, et al. Identification of twelve new susceptibility loci for different histotypes of epithelial ovarian cancer. Nat Genet. 2017;49(5):680–91.
28. van Heel DA, Fisher SA, Kirby A, Daly MJ, Rioux JD, Lewis CM, et al. Inflammatory bowel disease susceptibility loci defined by genome scan meta-analysis of 1952 affected relative pairs. Hum Mol Genet. 2004;13(7):763–70.
29. Voight BF, Scott LJ, Steinthorsdottir V, Andrew P, Aulchenko YS, Thorleifsson G, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet. 2010;42(7):579–89.

30. Teslovich TM, Musunuru K, Smith A V., Edmondson AC, Stylianou IM, Koseki M, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010;466(7307):707–13.
31. Darwin C. The Origin of Species. Pennsylvania State Univ. 2001;448 p.
32. Günther T, Jakobsson M. Genes mirror migrations and cultures in prehistoric Europe — a population genomic perspective. Curr Opin Genet Dev. 2016;41:115–23.
33. Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. Tracing the peopling of the world through genomics. Nature. 2017;541(7637):302–10.
34. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci. 2005;102(44):15942–7.
35. Rosenberg NA. Genetic Structure of Human Populations. Science (80-). 2002;298(5602):2381–5.
36. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538(7624):201–6.
37. White TD, Asfaw B, Degusta D, Gilbert H, Richards GD, Suwa G, et al. Pleistocene Homo sapiens from Middle Awash, Ethiopia. Middle Awash, Ethiop Nat. 2003;423(June):742–7.
38. McDougall I, Brown FH, Fleagle JG. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature. 2005;(433):733–6.
39. Ray N, Currat M, Berthier P, Excoffier L. Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. 2005;1161–7.
40. Tishkoff SA, Reed FA, Friedlaender FR, Ranciaro A, Froment A, Hirbo JB, et al. The Genetic Structure and History of Africans and African Americans. Science (80-). 2009;324(5930):1035–44.
41. Beltrame MH, Rubel MA, Tishkoff SA. Inferences of African evolutionary history from genomic data. Curr Opin Genet Dev. 2016;41:159–66.
42. Grün R, Stringer C, McDermott F, Nathan R, Porat N, Robertson S, et al. U-series and ESR analyses of bones and teeth relating to the human burials from Skhul. J Hum Evol. 2005;49(3):316–34.
43. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. Bayesian inference of ancient human demography from individual genome sequences. Nat Genet. 2011;43(10):1031–5.
44. Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, et al. Genomic analyses inform on migration events during the peopling of Eurasia. Nature. 2016;538(7624):238–42.
45. Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, et al. The genetic history of Ice Age Europe. Nature. 2016;534(7606):200–5.
46. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014;513(7518):409–13.
47. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, et al. The landscape of Neandertal ancestry in present-day humans. Nature. 2014;507(7492):354–7.
48. Benazzi S, Douka K, Fornai C, Bauer CC, Kullmer O, Svoboda J, et al. Early dispersal of modern humans in Europe and implications for Neanderthal behaviour. Nature. 2011;479(7374):525–8.
49. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475(7357):493–6.
50. Skoglund P, Malmström H, Omrak A, Raghavan M, Valdiosera C, Günther T, et al. Genomic Diversity and Admixture Foragers and Farmers. Science (80-). 2014;344(747):747–51.
51. Sánchez-Quinto F, Schroeder H, Ramirez O, Ávila-Arcos MC, Pybus M, Olalde I, et al. Genomic affinities of two 7,000-year-old Iberian hunter-gatherers. Curr Biol. 2012;22(16):1494–9.
52. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015;522(7555):207–11.
53. Skoglund P, Malmström H, Raghavan M, Storå J, Hall P, Willerslev E, et al. Origins and genetic legacy of neolithic farmers and hunter-gatherers in Europe. Science (80-). 2012;336(6080):466–9.
54. Omrak A, Günther T, Valdiosera C, Svensson EM, Malmström H, Kiesewetter H, et al. Genomic Evidence Establishes Anatolia as the Source of the European Neolithic Gene Pool. Curr Biol. 2016;26(2):270–5.
55. Martiniano R, Cassidy LM, Ó'Maoldúin R, McLaughlin R, Silva NM, Manco L, et al. The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. PLoS Genet. 2017;13(7):1–24.
56. Currat M, Excoffier L. The effect of the Neolithic expansion on European molecular diversity. Proc R Soc B Biol Sci. 2005;272(1564):679–88.
57. Botigue LR, Henn BM, Gravel S, Maples BK, Gignoux CR, Corona E, et al. Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. Proc Natl Acad Sci. 2013;110(29):11791–6.
58. Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, et al. The history of african gene flow into Southern Europeans, Levantines, and Jews. PLoS Genet. 2011;7(4).

59. Ralph P, Coop G. The Geography of Recent Genetic Ancestry across Europe. PLoS Biol. 2013;11(5).
60. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008;456(7218):98–101.
61. Consortium TIH. The International HapMap Project. Nature. 2003;426(6968):789–96.
62. NCBI NC for BI. No Title [Internet]. Available from: https://www.ncbi.nlm.nih.gov/
63. Altshuler D, Lander E, Ambrogio L. A map of human genome variation from population scale sequencing. Nature. 2010;476(7319):1061–73.
64. Auton A, Abecasis GR. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
65. Research CI for medical. Decoding the Genome [Internet]. Available from: https://www.coriell.org
66. Consortium EA. ExAC Browser Beta [Internet]. [cited 2017 Dec 29]. Available from: http://exac.broadinstitute.org
67. Lek M, Karczewski KJ, Minikel E V., Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91.
68. MacArthur D, Palotie A, Metspalu A, Remes A, Correa A, Franke A, et al. Genome Aggregation Database [Internet]. Available from: http://gnomad.broadinstitute.org/
69. Zlobin AS, Sharapov SS, Guryev VP, Bevova MR, Tsepilov YA, Sivtseva TM, et al. Population specific analysis of Yakut exomes. Dokl Biochem Biophys. 2017;474(1):213–6.
70. Einhorn Y, Weissglas-Volkov D, Carmi S, Ostrer H, Friedman E, Shomron N. Differential analysis of mutations in the Jewish population and their implications for diseases. Genet Res (Camb). 2017;99.
71. Zhou Y, Ingelman-Sundberg M, Lauschke VM. Worldwide Distribution of Cytochrome P450 Alleles: A Meta-analysis of Population-scale Sequencing Projects. Clin Pharmacol Ther. 2017;102(4):688–700.
72. Slavin TP, Maxwell KN, Lilyquist J, Vijai J, Neuhausen SL, Hart SN, et al. The contribution of pathogenic variants in breast cancer susceptibility genes to familial breast cancer risk. npj Breast Cancer. 2017;3(1):22.
73. Bleda M, Tarraga J, De Maria A, Salavert F, Garcia-Alonso L, Celma M, et al. CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources. Nucleic Acids Res. 2012;40(W1):609–14.
74. Warr A, Robert C, Hume D, Archibald A, Deeb N, Watson M. Exome Sequencing: Current and Future Perspectives. Genes|Genomes|Genetics. 2015;5(8):1543–50.
75. Reuter JA, Spacek D, Snyder MP. High-Throughput Sequencing Technologies. Mol Cell. 2016;58(4):586–97.
76. Pettersson E, Lundeberg J, Ahmadian A. Generations of sequencing technologies. Genomics. 2009;93(2):105–11.
77. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94(3):441–8.
78. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci. 1977;74(12):5463–7.
79. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016;107(1):1–8.
80. Maxam a M, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A. 1977;74(2):560–4.
81. Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. Genomics. 2008;92(5):255–64.
82. Swerdlow H, Gesteland R. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. Nucleic Acids Res. 1990;18(6):1415–9.
83. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. Nat Biotechnol. 2006;437(7057):376–80.
84. Nyrén P. Enzymatic method for continuous monitoring of DNA polymerase activity. Anal Biochem. 1987;167(2):235–8.
85. Hodkinson BP, Grice EA. Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. Adv Wound Care [Internet]. 2015;4(1):50–8. Available from: http://online.liebertpub.com/doi/abs/10.1089/wound.2014.0542
86. Buermans HPJ, den Dunnen JT. Next generation sequencing technology: Advances and applications. Biochim Biophys Acta - Mol Basis Dis. 2014;1842(10):1932–41.
87. Gupta PK. Single-molecule DNA sequencing technologies for future genomics research. Trends Biotechnol. 2008;26(11):602–11.
88. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencingplatforms: comparison of Ion Torrent, PacificBiosciences and Illumina MiSeq sequencers. BMC Genomics. 2012;13(1):1.
89. Lelieveld SH, Veltman JA, Gilissen C. Novel bioinformatic developments for exome sequencing. Hum Genet. 2016;135(6):603–14.
90. Altmann A, Weber P, Bader D, Preuß M, Binder EB, Müller-Myhsok B. A beginners guide to SNP calling from high-Throughput DNA-sequencing data. Hum Genet. 2012;131(10):1541–54.
91. Bao R, Huang L, Andrade J, Tan W, Kibbe W a, Jiang H, et al. Review of Current Methods, Applications, and Data

Management for the Bioinformatics Analysis of Whole Exome Sequencing. Lib Acad. 2014;13:67–82.
92. Ewing B, Green P. Base-Calling of Automated Sequencer Traces Using. Genome Res. 1998;(8):186–94.
93. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2009;38(6):1767–71.
94. NHGRI NHGRI. All About The Human Genome Project (HGP) [Internet]. [cited 2017 Nov 7]. Available from: https://genome.gov/
95. Hood L, Rowen L. The human genome project: Big science transforms biology and medicine. Genome Med. 2013;5(9):1.
96. Schendure J, Alden EL. The expanding scope of DNA sequencing. Nat Biotechnol. 2012;30(11):1084–94.
97. Smith TF, Waterman MS. Identification of common molecular subsequences. Mol Biol. 1981;147:195–7.
98. Burrows M, Wheeler D. A block-sorting lossless data compression algorithm. Algorithm, Data Compression. 1994;(124):18.
99. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
100. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3).
101. Bhd NTS. Innovative solutions for next-generation sequencing analysis [Internet]. 2014 [cited 2017 Nov 7]. Available from: http://www.novocraft.com/
102. Rumble SM, Lacroute P, Dalca A V., Fiume M, Sidow A, Brudno M. SHRiMP: Accurate mapping of short color-space reads. PLoS Comput Biol. 2009;5(5):1–11.
103. David M, Dzamba M, Lister D, Ilie L, Brudno M. SHRiMP2: Sensitive yet practical short read mapping. Bioinformatics. 2011;27(7):1011–2.
104. Mielczarek M, Szyda J. Review of alignment and SNP calling algorithms for next-generation sequencing data. J Appl Genet. 2016;57(1):71–9.
105. IonTorrent. Torrent Mapping Alignment Program [Internet]. 2014 [cited 2017 Nov 7]. Available from: https://github.com/iontorrent/TMAP
106. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
107. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data Aaron. Genome Res. 2010;(20):1297–303.
108. Institute B. Picard [Internet]. 2017 [cited 2017 Nov 7]. Available from: https://github.com/broadinstitute/picard
109. Fritz MH, Leinonen R, Cochrane G, Birney E. Efficient storage of high throughput DNA sequencing data using reference-based compression. 2011;734–40.
110. Bonfield JK. The Scramble conversion tool. Bioinformatics. 2014;30(19):2818–9.
111. Liu X, Han S, Wang Z, Gelernter J, Yang BZ. Variant Callers for Next-Generation Sequencing Data: A Comparison Study. PLoS One. 2013;8(9):1–11.
112. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12(6):443–51.
113. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012;1–9.
114. Domibel. IonTorrent-VariantCaller [Internet]. 2015 [cited 2017 Dec 19]. Available from: https://github.com/domibel/IonTorrent-VariantCaller
115. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
116. Li H. Tabix: Fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics. 2011;27(5):718–9.
117. Saunders C. gvcftools - utilities for gVCF files [Internet]. 2017. Available from: https://github.com/sequencing/gvcftools
118. Pirooznia M, Kramer M, Parla J, Goes FS, Potash JB, McCombie WR, et al. Validation and assessment of variant calling pipelines for next-generation sequencing. Hum Genomics. 2014;8:14.
119. Ni G, Strom TM, Pausch H, Reimer C, Preisinger R, Simianer H, et al. Comparison among three variant callers and assessment of the accuracy of imputation from SNP array data to whole-genome sequence level in chicken. BMC Genomics. 2015;16(1):1–12.
120. Cornish A, Guda C. A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. Biomed Res Int. 2015;2015.
121. Laurie S, Fernandez-Callejo M, Marco-Sola S, Trotta JR, Camps J, Chacón A, et al. From Wet-Lab to Variations: Concordance and Speed of Bioinformatics Pipelines for Whole Genome and Whole Exome Sequencing. Hum Mutat. 2016;37(12):1263–71.
122. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015;5(December):1–8.
123. Zhang G, Wang J, Yang J, Li W, Deng Y, Li J, et al. Comparison and evaluation of two exome capture kits and sequencing

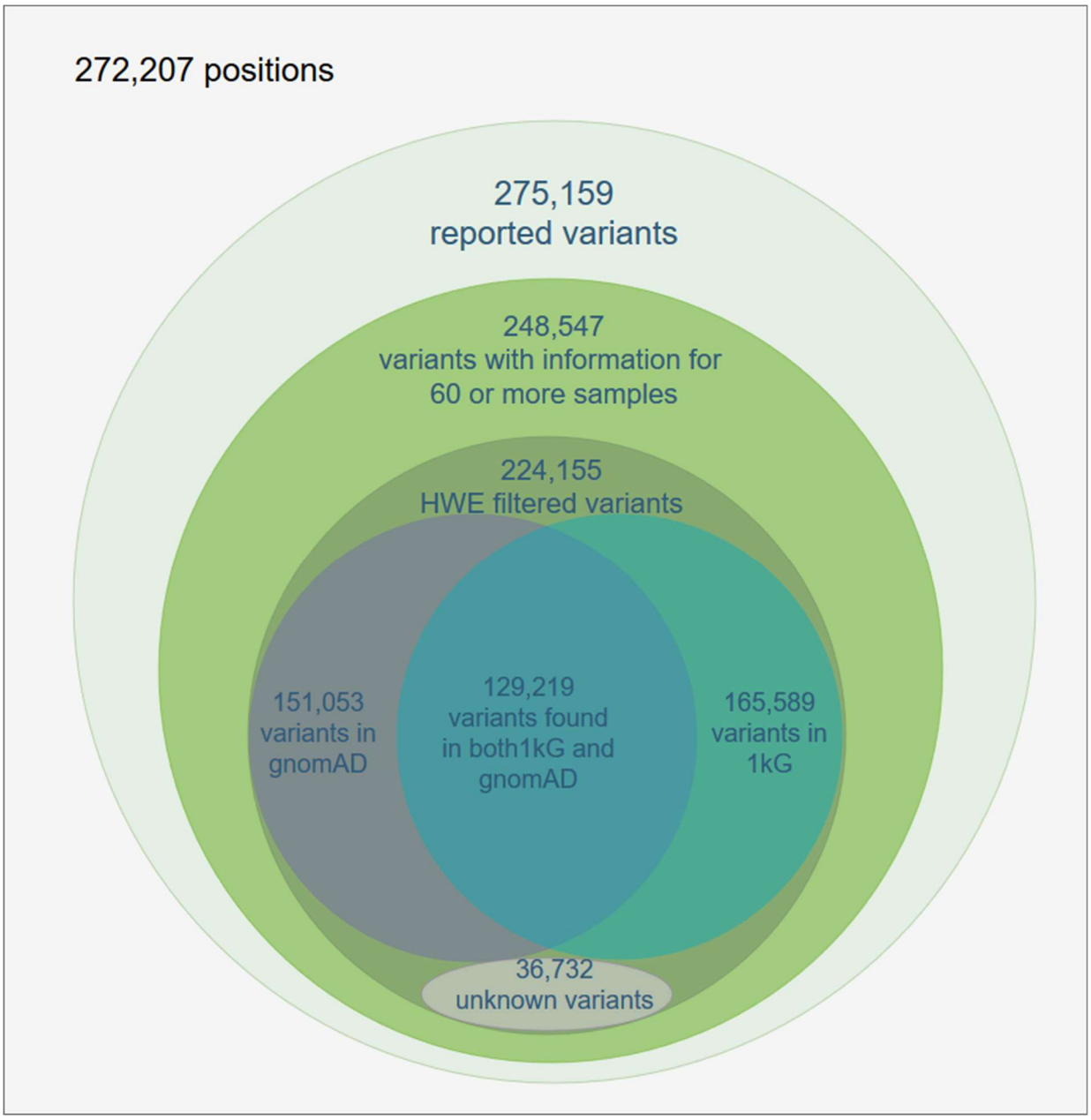
platforms for variant calling. BMC Genomics. 2015;16(1):1–9.
124. Sandmann S, De Graaf AO, Karimi M, Van Der Reijden BA, Hellström-Lindberg E, Jansen JH, et al. Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. Sci Rep. 2017;7:1–12.
125. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, Mcewen R, et al. VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;44(11).
126. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):1–7.
127. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1):1–14.
128. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073–82.
129. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9.
130. Grimm DG, Azencott CA, Aicheler F, Gieraths U, Macarthur DG, Samocha KE, et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat. 2015;36(5):513–23.
131. Schiemann AH, Stowell KM, Galley HF. Comparison of pathogenicity prediction tools on missense variants in RYR1 and CACNA1S associated with malignant hyperthermia. Br J Anaesth. 2016;117(1):124–8.
132. Mahmood K, Jung C, Philip G, Georgeson P, Chung J, Pope BJ, et al. Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. Hum Genomics. 2017;11(1):10.
133. Wallis Y, Payne S, Mcanulty C, Bodmer D, Sister-mans E, Robertson K, et al. Practice Guidelines for the Evaluation of Pathogenicity and the Reporting of Sequence Variants in Clinical Molecular Genetics. Acgs. 2013;(September):16.
134. MacArthur DG, Tyler-Smith C. Loss-of-function variants in the genomes of healthy humans. Hum Mol Genet. 2010;19(R2):125–30.
135. MacArthur D, Balasubramanian S, Frankish A. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. Science (80-). 2012;335(6070):1–14.
136. Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. PLoS Comput Biol. 2013;9(7).
137. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res. 2009;19(7):1316–23.
138. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Identif deleterious Mutat within three Hum genomes. 2009;19(9):1553–61.
139. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7(8):575–6.
140. Siepel A, Pollard KS, Haussler D. New Methods for Detecting Lineage-Specific Selection. Research in computational biology. 2006.
141. Liu X, Jian X, Boerwinkle E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat. 2011;32(8):894–9.
142. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F. GENCODE: The Reference Human Genome Annotation for The ENCODE Project. Genome Res. 2012;22:1760–74.
143. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: Application to cancer genomics. Nucleic Acids Res. 2011;39(17):37–43.
144. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. Hum Mutat. 2013;34(1):57–65.
145. Davydov E V., Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6(12).
146. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009;25(12):54–62.
147. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011;478(7370):476–82.
148. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493(7431):216–20.
149. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: A database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat. 2013;34(9).
150. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 2014;42(22):13534–44.
151. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction

methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24(8):2125–37.
152. Kircher M. A general framework for estimating the relative pathogenicity of human genetic variants. Nat g. 2014;46(3):310–5.
153. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. BMC Genomics. 2013;14(Suppl 3):S3.
154. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS One. 2012;7(10).
155. Gulko B, Hubisz MJ, Gronau I, Siepel A. Probabilities of Fitness Consequences for Point Mutations Across the Human Genome. Nat Genet. 2015;47(3):276–83.
156. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015;31(10):1536–43.
157. Quang D, Chen Y, Xie X. DANN: A deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31(5):761–3.
158. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate , insect , worm , and yeast genomes. Genome Res. 2005;(15):1034–50.
159. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat. 2016;37(3):235–41.
160. Huang PJ, Lee CC, Tan BCM, Yeh YM, Huang KY, Gan RC, et al. Vanno: A visualization-aided variant annotation tool. Hum Mutat. 2015;36(2):167–74.
161. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an Information Aesthetic for Comparative Genomics. Genome Res [Internet]. 2009;19(604):1639–45. Available from: http://circos.ca/intro/genomic_data/
162. University AM. VarAFT [Internet]. Available from: http://varaft.eu/
163. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief Bioinform. 2013;14(2):178–92.
164. Institute de D, Louvain U catholique de. Highlander - Variant analysis made easy [Internet]. Available from: https://sites.uclouvain.be/highlander/index.html
165. Axmark D, Widenius M. MySQL 5.7 reference manual [Internet]. California: Oracle.; 2015. Available from: https://dev.mysql.com/
166. Hipp DR, Kennedy D, Mistachkin J. SQLite [Internet]. SQLite Development Team; 2018. Available from: https://sqlite.org/index.html
167. Corporation M. MariaDB ColumnStore. Espoo, Finland; 2016.
168. Boncz PA. Monet ; a next-Generation DBMS Kernel for Query-intensive Applications [Internet]. University of Amsterdam; 2002. Available from: https://www.monetdb.org/Home
169. Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. Nat Genet. 2008;40(5):646–9.
170. Reich D, Price AL, Patterson N. Principal component analysis of genetic data. Nat Genet. 2008;40(5):491–2.
171. Wang C, Zhan X, Liang L, Abecasis GR, Lin X. Improved Ancestry Estimation for both Genotyping and Sequencing Data using Projection Procrustes Analysis and Genotype Imputation. Am J Hum Genet. 2015;96(6):926–37.
172. EMBL-EBI. IGSR: The International Genome Sample Resource [Internet]. 2018 [cited 2018 Sep 20]. Available from: http://www.internationalgenome.org/
173. Inc. M. MongoDB [Internet]. 2018. Available from: https://www.mongodb.com/
174. Eric Vallabh Minikel. Converting genetic variants to their minimal representation [Internet]. 2014 [cited 2018 Mar 5]. Available from: http://www.cureffi.org/2014/04/24/converting-genetic-variants-to-their-minimal-representation/
175. Jombart T, Kamvar ZN, Collins C, Lustrik R, Beugin M-P, Knaus BJ, et al. Exploratory Analysis of Genetic and Genomic Data [Internet]. 2018. Available from: https://github.com/thibautjombart/adegenet
176. Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, et al. Create Elegant Data Visualisations Using the Grammar of Graphics. 2018.
177. Ligges U. 3D Scatter Plot. 2018.
178. Campbell IM, Gambin T, Jhangiani S, Grove ML, Muzny DM, Shaw CA, et al. Multiallelic Positions in the Human Genome: Challenges for Genetic Analyses. 2016;37(3):231–4.
179. Song L, Huang W, Kang J, Huang Y, Ren H, Ding K. Comparison of error correction algorithms for Ion Torrent PGM data : application to hepatitis B virus. Sci Rep [Internet]. 2017;(July):1–11. Available from: http://dx.doi.org/10.1038/s41598-017-08139-y
180. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature [Internet]. 2009;461(7261):272–6. Available from: http://dx.doi.org/10.1038/nature08250
181. Hu J, Ng PC. Predicting the effects of frameshifting indels. Genome Biol [Internet]. 2012;13(2):R9. Available from:

http://genomebiology.com/2012/13/2/R9
182. Rausell A, Mohammadi P, Mclaren PJ, Bartha I. Analysis of Stop-Gain and Frameshift Variants in Human Innate Immunity Genes. 2014;10(7).
183. Marzin P, Cormier-Daire V. Geleophysic Dysplasia [Internet]. GeneReviews [Internet]. 2009 [cited 2018 Sep 14]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK11168/
184. Science NA of. Evaluating Human Genetic Diversity [Internet]. 1st ed. National Academies Press (US); 1997. Available from: https://www.ncbi.nlm.nih.gov/books/NBK100427/
185. Karczewski KJ, Francioli LC. The genome Aggregation Database (gnomAD) [Internet]. 2017 [cited 2018 Sep 5]. Available from: https://macarthurlab.org/2017/02/27/the-genome-aggregation-database-gnomad/
186. Gravel S, Zakharia F, Moreno-Estrada A, Byrnes JK, Muzzio M, Rodriguez-Flores JL, et al. Reconstructing Native American migrations from whole-genome and whole- Reconstructing Native American Migrations from Whole- Genome and Whole-Exome Data. 2013;9(12).
187. Fasano ME, Rendine S, Pasi A, Bontadini A, Cosentini E, Carcassi C, et al. The distribution of KIR-HLA functional blocks is different from North to South of Italy. 2014;168–73.
188. Manser AR. Human KIR repertoires : shaped by genetic diversity and evolution. 2015;267:178–96.
189. Sanchez-mazas A. HLA DNA Sequence Variation among Human Populations : Molecular Signatures of Demographic and Selective Events. 2011;6(2).
190. Baysal BE, Lawrence EC, Ferrell RE. Sequence variation in human succinate dehydrogenase genes : evidence for long-term balancing selection on SDHA. 2007;14:1–14.
191. Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, et al. Global diversity , population stratification , and selection of human copy number variation. 2015;349(6253):1–23.
192. Cavaco I, Hombhanje FW, Gil JP, Kaneko A. Frequency of the Functionally Relevant Aryl Hydrocarbon Receptor Repressor (AhRR) Pro185Ala SNP in Papua New Guinea. 2013;28(6):519–21.
193. Lucas D, Campillo JA, López-hernández R, Martínez-garcía P, López-sánchez M, Botella C, et al. Allelic diversity of MICA gene and MICA / HLA-B haplotypic variation in a population of the Murcia region in southeastern Spain. 2008;655–60.
194. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I. Identification of a variant associated with adult-type hypolactasia. 2002;30(february):233–7.
195. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, et al. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. 2004;(1973):1111–20.

ATTACHMENT I

AI – Schema of the subsets of variants analysed through the present work.



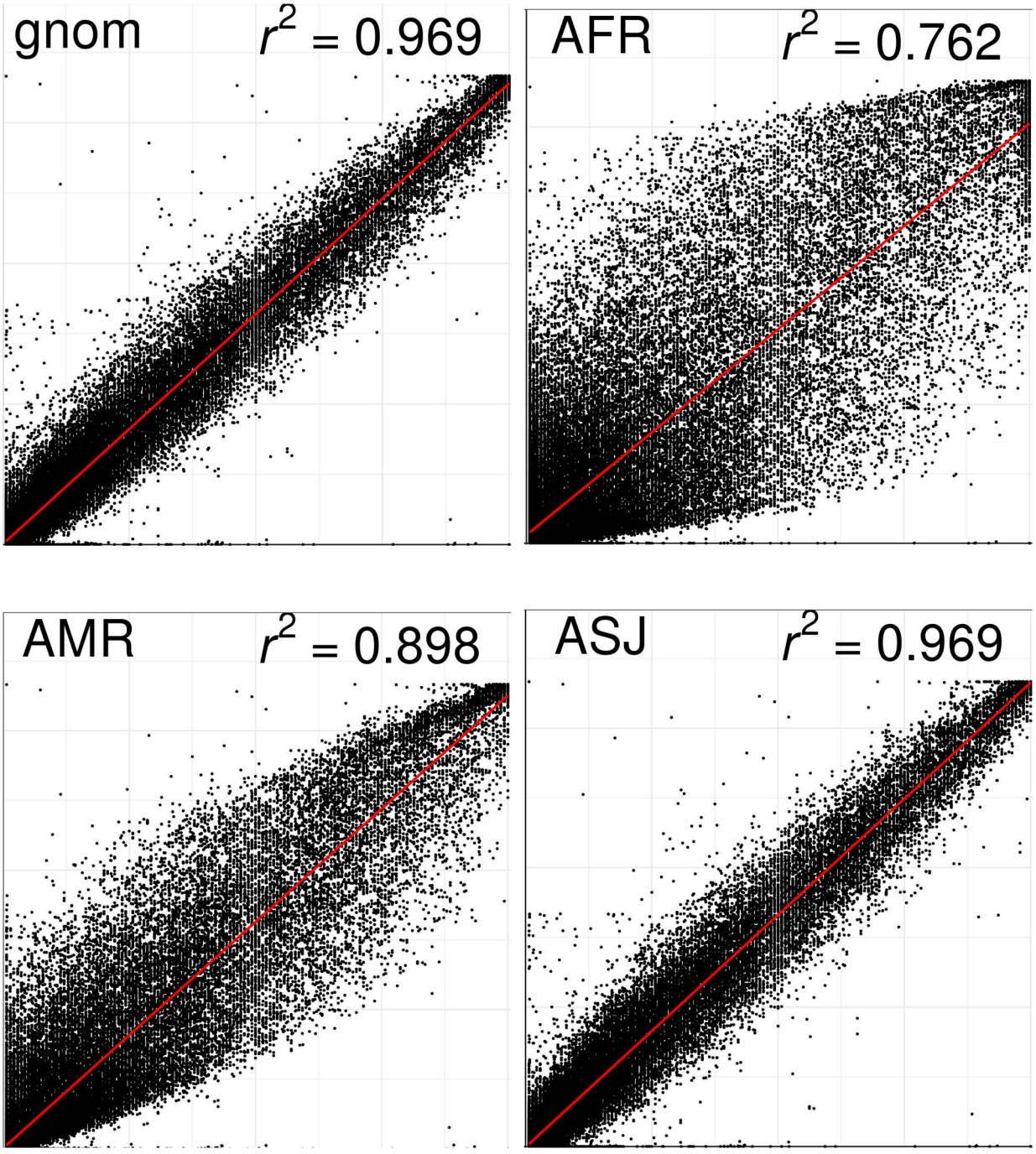
ATTACHMENT II

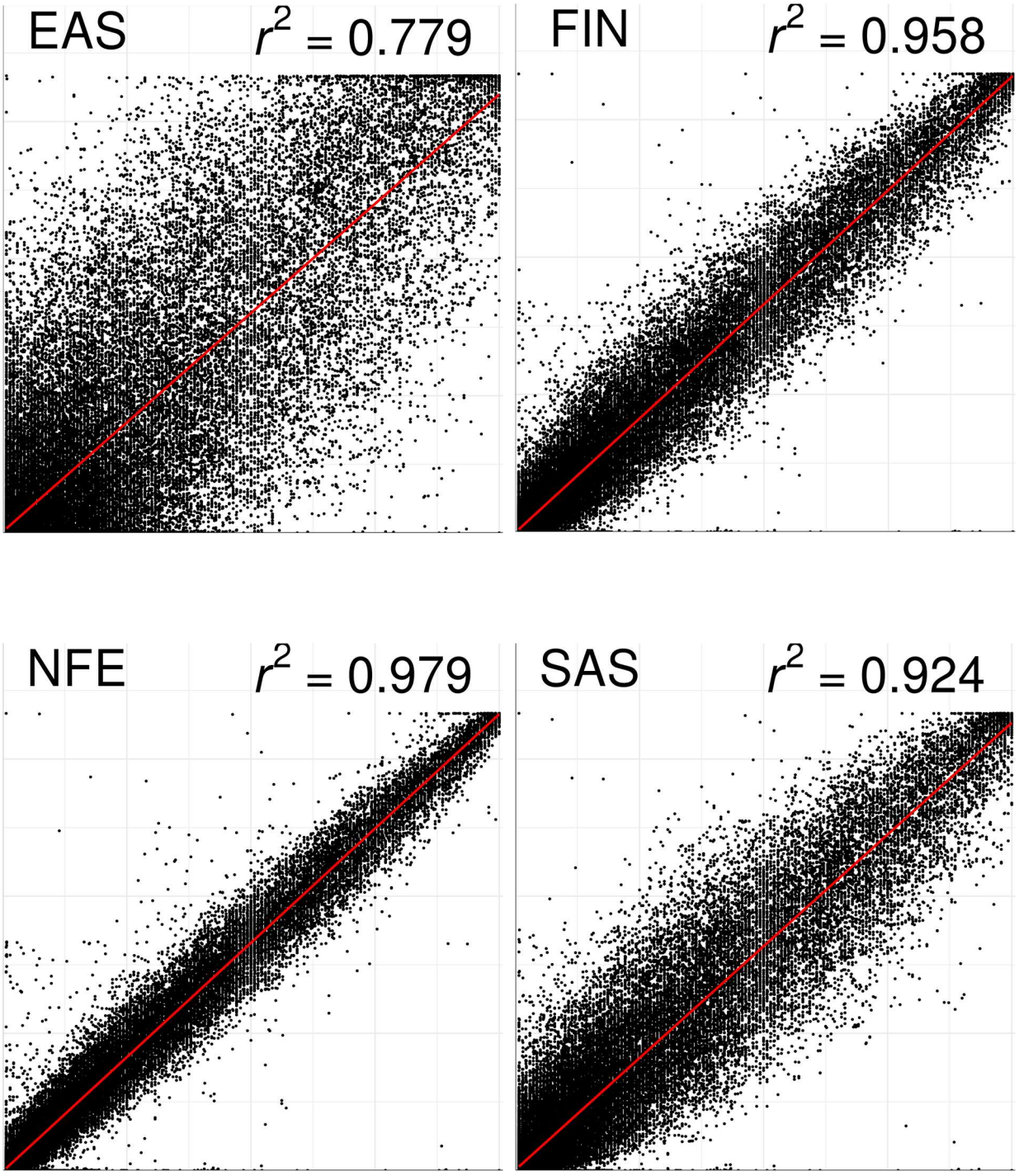
All - Total count of variants by impact. According to GEMINI impact classification.

Impact	Total	SNVs	Indels	Severity
Synonymous variants	42,517	42,517	0	
Intron variants	99,273	93,960	5,313	
3-prime UTR variants	7,244	6,877	367	
5-prime UTR variants	5,581	5,342	239	
Downstream Gene variants	443	424	19	LOW
Upstream Gene variants	1,030	985	45	
Intergenic variants	41	41	0	
Non-coding Transcript variants	780	743	37	
Stop Retained variants	37	37	0	
Missense variants	54,949	54,949	0	
Coding Sequence variants	4	1	3	
Inframe Deletions	366	0	366	MED
Inframe Insertions	116	0	116	
Protein Altering variants	2	0	2	
Splice Region variants	8,323	7,971	352	
Frameshift variants	1,707	0	1,707	
Stop Loss variants	92	91	1	
Stop Gain variants	835	821	14	HIGH
Splice Acceptor variants	364	327	37	
Splice Donor variants	328	306	22	
Start Loss variants	123	119	4	
Splicing variants	9,015	8,604	411	
LoF variants	3,335	1,567	1,768	

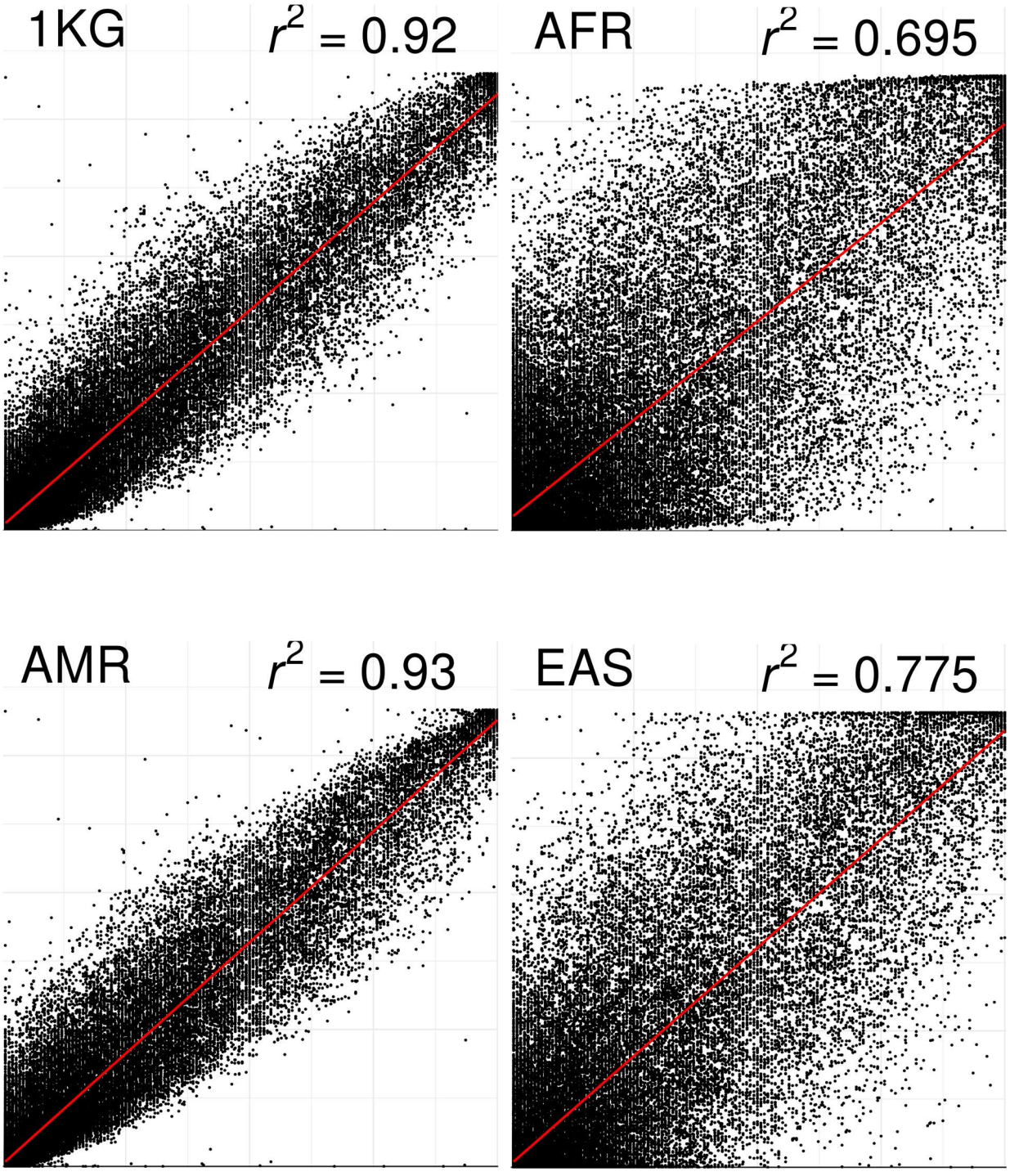
ATTACHMENT III

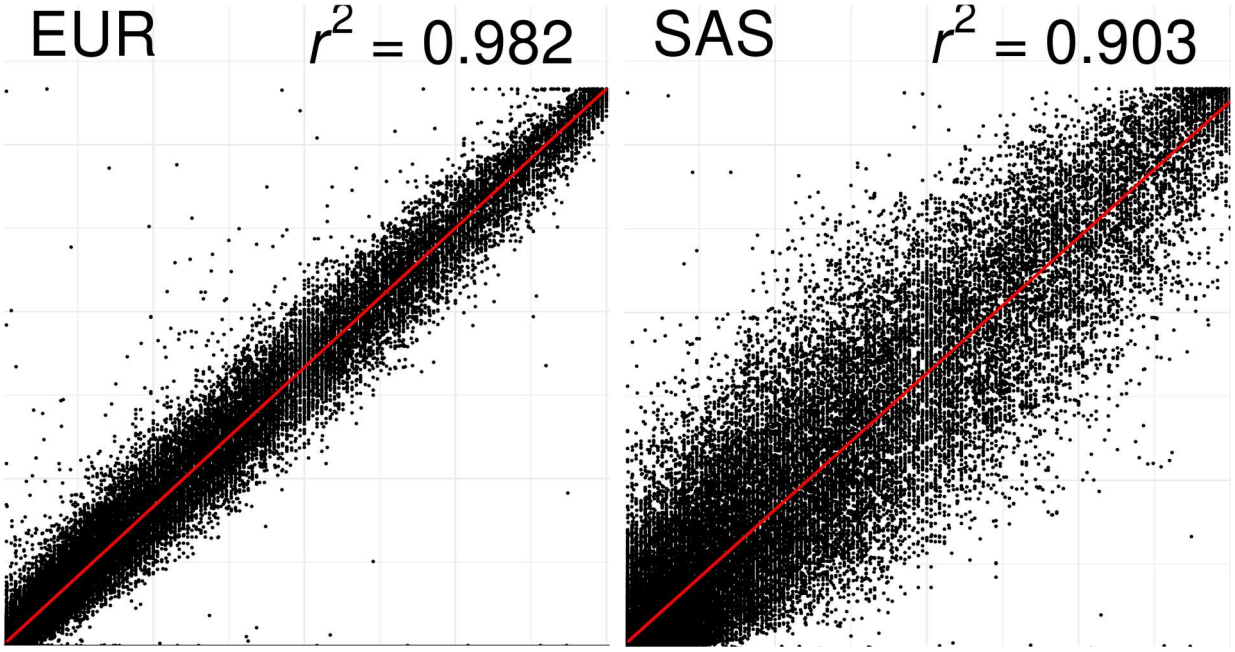
AIII.1 – Allele Frequency scatterplots comparing Portuguese population (x-axis) against each gnomAD population (y-axis). Axis display AF values in decimal scale representation. gnom – gnomAD general values; AFR – African/African Americans; AMR – Admixed Americans; ASJ – Ashkenazi Jewish; EAS – East Asians; FIN – Finnish; NFE – Non-Finnish Europeans; SAS – South Asians.



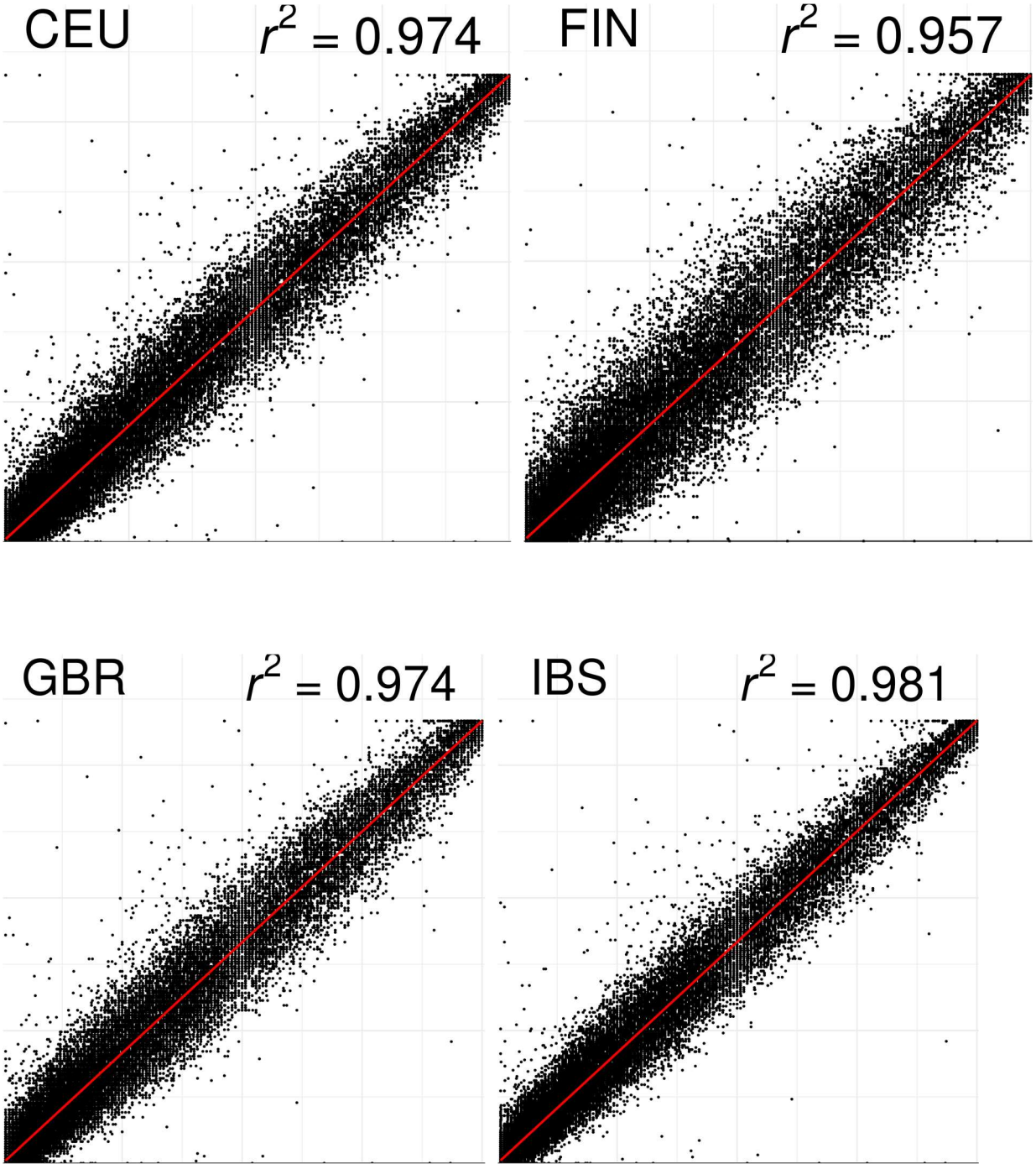

EAS
$r^2 = 0.779$
FIN
$r^2 = 0.958$
NFE
$r^2 = 0.979$
SAS
$r^2 = 0.924$

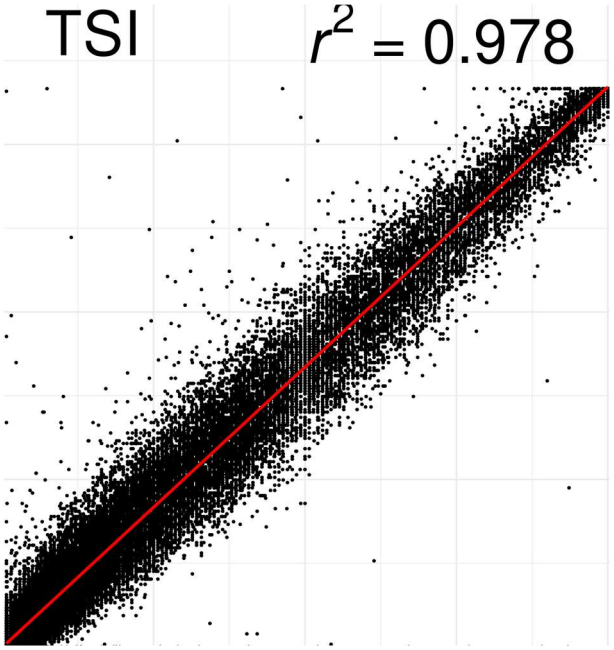
AIII.2 – Allele Frequency scatterplots comparing Portuguese population (x-axis) against each 1kG population (y-axis). Axis display AF values in decimal scale representation. KG – 1kG general values; AFR – Africans; AMR – Americans; EAS – East Asians; EUR – Europeans; SAS – South Asians.




EUR
r^2 = 0.982
SAS
r^2 = 0.903

AIII.3 – Allele Frequency scatterplots comparing Portuguese population (x-axis) against each 1kG European population (y-axis). Axis display AF values in decimal scale representation. CEU – Central Europeans from Utah; FIN – Finnish in Finland; GBR – British from England and Scotland; IBS – Iberians from Spain; TSI – Tuscans from Italy.




TSI
r^2 = 0.978