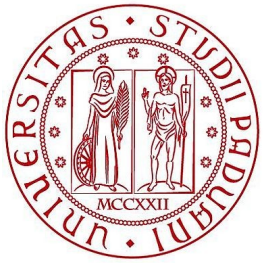

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Head Office: Università degli Studi di Padova

Department of Biology

Ph.D. COURSE IN: BIOSCIENCES
CURRICULUM: GENETICS, GENOMICS AND BIOINFORMATICS
SERIES XXXI

From exome to whole genome sequencing: mining for inconsistencies and functional elements in coding and non-coding regions

Coordinator: Prof. Ildikò Szabò
Supervisor: Prof. Giorgio Valle

Ph.D. student: Margherita Ferrarini

Contents

Abbreviations

CNV	Copy Number Variation
GRC	Genome Reference Consortium
GWAS	Genome-Wide Association Studies
HF	High Frequency
HGP	Human Genome Project
MAiR	Minor Allele in Reference
NGS	Next Generation Sequencing
SNP	Single Nucleotide Polymorphisms
SNV	Single Nucleotide Variant
TS	Targeted sequencing
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

Abstract

Over the last two decades the advancement in DNA sequencing technologies has enormously increased the amount of sequencing data available to researchers and geneticists. This has been accompanied by the development of tools for sequencing data analysis, including the human reference genome, that is undoubtedly an indispensable resource. It is known that the reference genome does not always represent the real consensus sequence of the human population, due to the inclusion of rare alleles and sequencing errors. Moreover, genomic duplications are often misassembled and, as a result, they may be found in the reference genome as a collapsed consensus, thus generating false variants. In this work I performed a thorough search for conflicting information between the human reference genome (GRCh37 and GRCh38) and some of the most popular human genetic resources such as the 1000 Genomes Project, to disclose minor alleles and to mine genetic inconsistencies. To search for unreported genomic duplications, I performed a genome wide screening for unbalanced heterozygosity. I found that inaccuracies and errors are much higher than expected. Minor alleles occurring with a frequency <10% are found on average every ~7,000 bases and include many rare variants that are never found elsewhere, producing high numbers of false positives as well as possible false negatives. The systematic screening for unbalanced heterozygosity revealed ~86,000 variants that are likely the result of unreported genomic duplications, involving functionally relevant genes such as *MAP2K3* and *KCNJ12*. My findings may help the ongoing quest to obtain a highly accurate human genome reference sequence. Moreover, the results presented in this thesis will be useful to human geneticists in the process of filtering and selecting causative variants.

The advancement in DNA sequencing technologies also accounts for the increasing usage of Whole Genome Sequencing approaches both in the research and clinical fields, thus revealing that the large majority of disease-associated SNPs are located in non-coding regions of the human genome. However, the functional interpretation of non-coding variants is still challenging. Part of my

work also addressed this problem, aiming to develop a method for non-coding variant prioritization. The method, presented in the last chapter of this thesis, is based on a comparative genomics approach for the identification of functional constraints in primate orthologous genes. The first steps of my approach have proved to be powerful in identifying orthologous genes, but further work is necessary to optimize the multiple sequence alignment step and the identification of conserved domains.

Chapter 1

Introduction

Over the last two decades the advancement in DNA sequencing technologies has enormously increased the amount of sequencing data available to researchers and geneticists. Handling and interpreting these data still remain the major challenge in the human genetics field. This challenge has driven the entire research carried out during my PhD.

My PhD project mainly addressed the problem of improving the quality of the human reference genome. The ideal reference genome should represent the consensus sequence of the human population[1]. In its standard format it consists in a linear haploid DNA sequence and serves as the foundation for sequencing analyses, by providing a substrate for read alignment. Since its first draft release in 2001[2], the reference sequence of the human genome underwent several updates and improvements and, even now, continuous efforts aim to ameliorate it.
Starting from the study of recurrent variants in both exome and genome sequencing data, I contributed to identify some inconsistencies in the reference sequence, both at base-pair and assembly level, not yet reported in literature. These findings could be useful to make the reference more accurate and, no less important, to help researchers and geneticists to avoid wrong inferences in sequencing data analyses due to the incompleteness of the reference genome.

In the first sections of this thesis, I will describe how the introduction of Next Generation Sequencing (NGS) technologies led to the reduction of DNA sequencing costs and, as a result, to an increasing use of high throughput sequencing in both the research and clinical context. This has been accompanied by the development of tools for sequencing data analysis, including the human reference genome that is undoubtedly an indispensable resource. The development of the reference sequence of the human genome and its limitations are extensively described in the following paragraphs.

The following chapters illustrate my project outline and the methods that I used to identify errors in the reference genome, together with the results that I obtained. These results have been also described in a manuscript entitled *'Data mining of recurrent variants reveals inconsistencies in the human reference genome'* that has been submitted for publication. The draft of the manuscript is enclosed at the end of the thesis.

Part of my PhD project also addressed the problem of interpreting the functional effect of nucleotide variants in non-coding regions of the human genome. As extensively discussed below, the sequencing of the whole genome is more and more widespread. Nevertheless, due to difficulties in determining the functional relevance of mutations outside protein coding genes, non-coding variants remain largely understudied.
In particular, my study aims to develop a method for non-coding variant prioritization starting from the identification of functional constraints in long non-coding RNAs. For this purpose I used a comparative genomics approach by aligning orthologous sequences found in phylogenetically related organisms and looking for regions that are conserved across species.
Although the approach is still under development and requires further work, preliminary results seem to be very promising. For this reason this part of my research will be described as future perspectives in the last chapter of this thesis.

1. 1 The Next Generation Sequencing era

1.1.1 The cost of DNA sequencing

Accurately determining the cost of the first human genome sequence is not simple. It was reported that the initial draft published in 2001 by the Human Genome Project (HGP) international consortium[2] required ~300 million dollars, with a further cost of ~150 million dollars to obtain the complete human genome sequence published in 2003[3].
Over the last two decades advances in the field of genomics have led to a remarkable reduction in DNA sequencing costs (Figures 1 and 2). The costs associated with the DNA sequencing are periodically tracked by the National

Human Genome Research Institute (NHGRI)[4]. In this evaluation only 'production' costs are included (costs of labor, sequencing reagents and instruments and data processing), while 'non-production' activities are not considered (quality control, technology development or data analysis). To highlight the impressive reduction in DNA sequencing costs, data are compared to the hypothetical line reflecting the Moore's Law, according to which the compute power doubles every two years.

A deep separation between DNA sequencing costs trend and Moore's law line occurred in 2008: data from 2001 through 2007 represent the costs of sequencing based on Sanger chemistries and capillary instruments (first generation sequencing), followed by the introduction of Next Generation Sequencing (NGS) technologies accountable for the cost drop. In 2005 454 Life Sciences released the Genome Sequencer 20 (GS20), the first next generation DNA sequencing machine

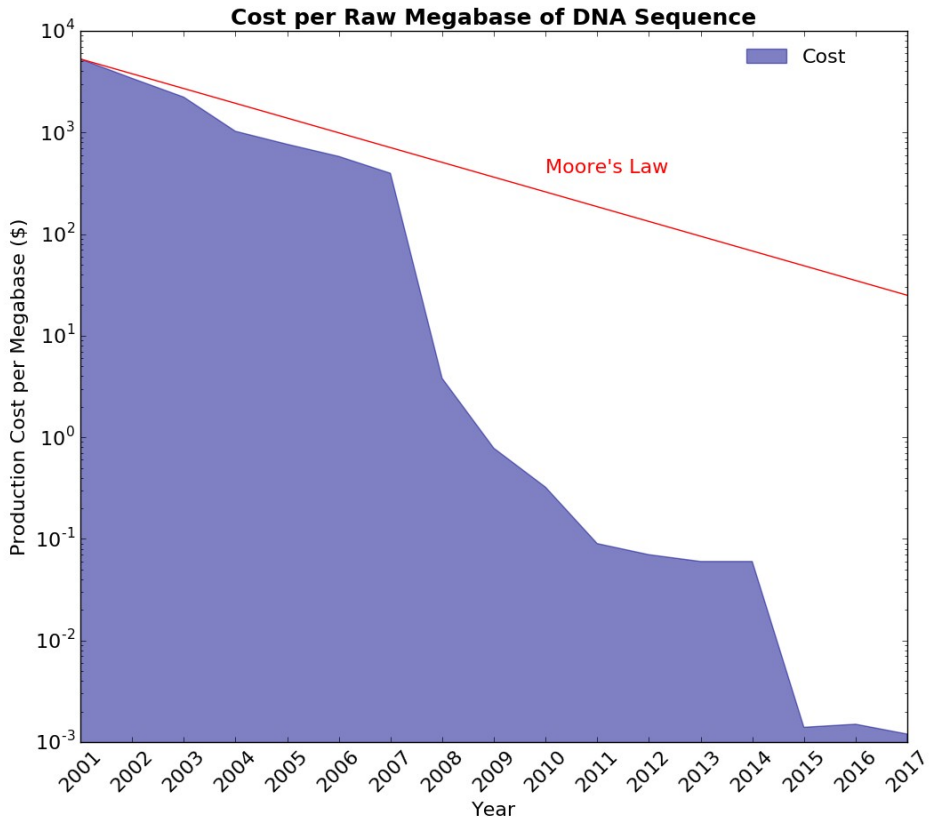

Figure 1. Cost per megabase of DNA sequence. The cost per megabase deeply decreased in 2008 with the introduction of NGS technologies in DNA sequencing. Data of production cost were calculated by National Human Genome Research Institute[4]. When multiple data were provided for the same year, corresponding to different months, only the last one was considered. The red line represents the Moore's Law. Note the logarithmic scale on the y-axis.

on the market, in 2006 Solexa launched the Genome Analyzer instrument and in 2007 Applied Biosystem announced the SOLiD™ system. As consequence of the use of these new technologies in the DNA sequencing, the price for a whole human genome dropped rapidly since 2008. In July 2017, the last available data at the NHGRI website, the estimated cost was 1,121 dollars. To figure out where we are right now, with the introduction of the NovaSeq™ series Illumina promised the sequencing of the human genome for 100 dollars[5].

Simultaneously to the decrease of whole genome sequencing cost, its usage has grown: the number of citations for 'Whole Genome Sequencing' in Scopus increased from 74 in 2001 to 4,688 in 2017, with a rapid growth in the last few years (Figure 2)[6]. This data reflects the fact that WGS is becoming the leading strategy routinely used not only in the research context but also in the clinical one[7].

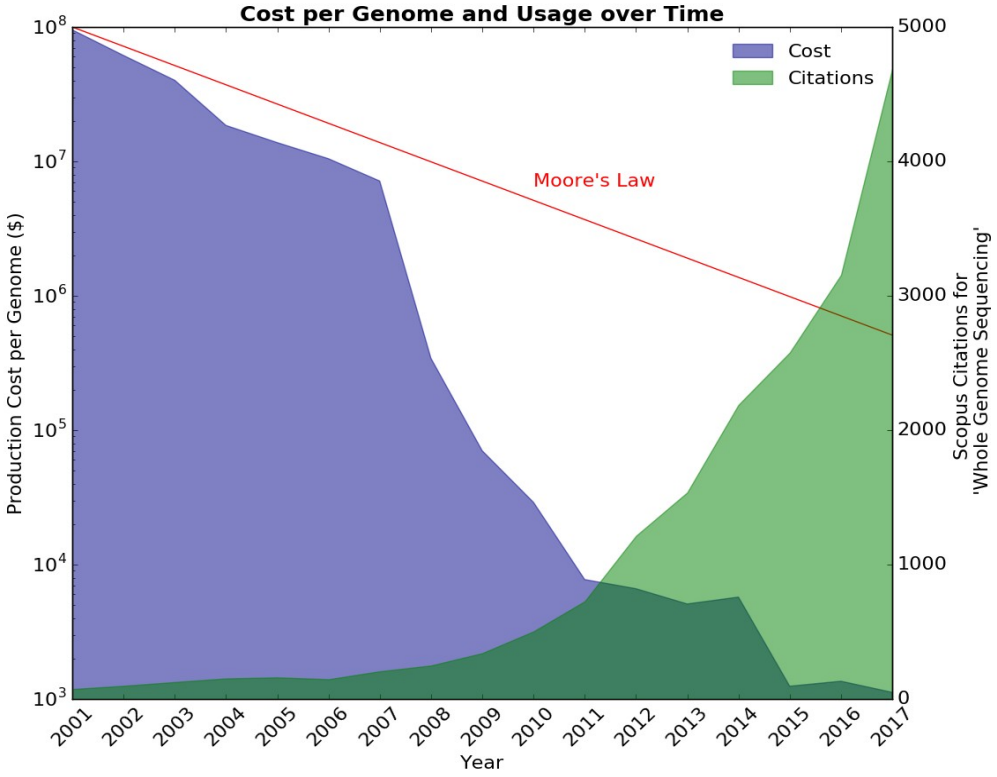

Figure 2. Production cost and usage of whole genome sequencing over time. The decrease of the cost of genome sequencing (blue) is shown together with the increase of the number of articles containing the phrase 'whole genome sequencing' (green). Data of production cost were calculated by National Human Genome Research Institute[4]. When multiple data were provided for the same year, corresponding to different months, only the last one was considered. The number of citations derives from Scopus[6]. The red line represents the Moore's Law. Note the logarithmic scale on the y-axis. Figure adapted from Katsonis *et al.*[8].

1.1.2 Next Generation Sequencing in diagnostics

As a result of the remarkable reduction in DNA sequencing costs, NGS has been widely introduced in the diagnosis field. NGS encompasses three different approaches: i) targeted sequencing (TS) to analyze a subset of genes or regions of the genome; ii) whole exome sequencing (WES) to obtain the sequence of protein coding regions (exons); iii) whole genome sequencing (WGS) to determine the complete sequence of the entire genome.

For many years TS has been used as the gold standard method for the molecular diagnosis of genetic diseases with a good knowledge of the associated genes, as in the case of inherited cardiomyopathies for which many target panels have been developed since 2007[8–10]. Advantages of customize targeted gene panels include the possibility to focus on the most relevant genes associated to the disorder, a higher coverage in the interesting regions compared to WES and WGS, a faster and cheaper sequencing and minimal chance of incidental findings. However, this approach requires a good *a priori* knowledge of the disease and limits the possibility to discover novel unsuspected disease genes. In this respect gene panels could be view as an inexpensive and rapid first-tier test, followed by WES or WGS in case of negative results[11].

WES was used in genetic diagnosis for the first time in 2009, when patients suspected to have Bartter syndrome were tested for a homozygous missense mutation at the known congenital chloride diarrhea locus[12]. The authors were able to capture approximately 95% of the targeted coding sequences with high sensitivity and specificity. The estimated sensitivity to detect heterozygous variants was 81%, 90% and 95% at mean coverage of 20x, 30x and 40x respectively, while the specificity reaches 99.9% at mean coverage of 30x. These data supported the clinical utility of WES for the first time. Moreover the authors highlighted WES strength in new disease genes discovery: compared to TS, WES not only covers exons of genes already associated to the disease, but also allows to identify novel causative genes in diseases with yet unknown molecular basis. However WES has two important limitations: the risk to insufficiently cover coding exons, especially those GC-rich[11,13], and the impossibility to identify non-coding pathogenic variants[12].

These limits are overcome by WGS that provides the most continuous coverage as the capturing is no longer necessary and WGS is much less sensitive to GC content[14–17]. Moreover WGS has the advantage to detect variants in non-coding regions and to improve copy number variations (CNVs) detection[11,16]. For these reasons several recent papers reported WGS as more powerful than WES[11,16].

Finally, the cost of sequencing must be taken into account: WGS cost is directly related to the cost per megabase of DNA which has decreased very much faster than the cost of any capture kit. This advantage, together with the high diagnostic yield, explains why several papers suggested the use of WGS as the first-tier test in diagnosis[18–20].

1.1.3 Prioritization of genetic variants

In the NGS era the major challenge researchers and geneticists face has moved from obtaining DNA sequences to interpreting the enormous amount of generated data. Even now the ability to sequence DNA exceeds the ability to analyze it.

Sequencing platforms provide the DNA sequence in the form of sequencing reads, generally collected in a FASTQ* format file. Sequencing reads are then aligned to the human reference genome, a database of all human DNA sequence that should ideally represent the entire human population[21]. The alignment is performed by appropriate alignment tools, for example the commonly used Burrows-Wheeler Aligner (BWA)[22]. The resulting alignment file is called Sequence Alignment Map (SAM**) file - BAM and CRAM for the corresponding binary and compressed files, respectively. The following steps are to identify those sites in which the sequenced DNA differs from the reference genome and assign a genotype to the subject. These steps are carried out by using a variant caller, such as the Genome Analysis Toolkit (GATK) developed by the Broad Institute[23], that provides a list of identified variants and corresponding genotypes in a Variant Call Format (VCF***) file.

*FASTQ format: It is a text-based format for storing base call and quality information for sequencing reads. Each entry in a FASTQ file consists of four lines: sequence identifier, sequence, quality score identifier line (consisting only of a +), quality score.

**SAM (Sequence Alignment Map) format: It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. Header lines start with '@'. Each alignment line has 11 mandatory fields for essential alignment information, such as mapping position, and variable number of optional fields for flexible or aligner specific information.

***VCF (Variant Call Format) format: It is a text file containing meta-information lines (included after the '##' string), a header line (included after the '#' string) and data lines, each containing information about a position in the genome.

Sequencing the exome of an individual, an average of 48,785 variants is identified, while sequencing the entire genome the average number of variants is 4,491,760 (data shown in the Results and Discussion). With this wealth of information and with more candidate variants to evaluate than ever before, researchers' efforts have been directed to understand which variations drive disease or contribute to phenotypic traits. Especially in the clinical field, disease associated genetic variants must be separated from the broader background of rare, potentially functional, but not pathogenic variants present in all individuals. This can be done by applying a sequential series of filters until the candidate mutation list is sufficient short for *in vitro* and or *in vivo* functional studies. This procedure is called 'prioritization of variants'.

Thorough guidelines on prioritization strategy can be found in literature[24–26] and several filter-based tools have been implemented, for example the recent Queryor platform developed by the CRIBI Center of the University of Padua[27].

Researchers' ability of prioritizing variants is good mainly for non-synonymous variations. In this case the severity of a sequence alteration on protein function can be predicted *in silico* by using multiple computational tools based on the conserved sequence of protein coding genes and amino acid changes. Examples of these tools are MutationAssessor[28], SIFT[29] and PolyPhen-2[30].

More difficult is to understand the functional effect of variants in non-coding regions, regulatory regions or splice sites. It is well known that these variants play an important role in determining human traits and complex diseases[31–34]. Several aspects can be investigated to reveal the role of non-coding variants, for example chromatin interactions and gene expression. Moreover different tools allow their prioritization, such as Genome-Wide Annotation of VAriants (GWAVA)[35] and Combined Annotation-Dependent Depletion (CADD)[36]. However the functional interpretation of non-coding variants remains a challenging and demanding task[34] and, especially in the case of non-coding regions, the classification of a genetic variant into deleterious or neutral, although very convenient, may be too simplistic.

A further level of complexity in analyzing NGS data derives from the awareness that no human reference genome - the fundamental necessity for all resequencing

test - is fully complete and correct at the moment[1,37]. As discussed in the following sections, erroneous bases and missing sequences still affect the reference genome and it has been demonstrated that these inconsistencies can be the source of misunderstandings in variant prioritization and interpretation[38,39]. The work begun in the 1990s to create the most complete and correct human reference genome is not over yet.

1.2 The human reference genome

1.2.1 From the first draft to the GRCh38 release

In February 2001 the Human Genome Project (HGP) international consortium announced the publication of the first draft of the human genome. The draft was produced with a clone-based approach and collapsing sequences from over 50 individuals into a single consensus haplotype representation of each chromosome. It was three billion base pairs long and covered more than 90% of the human genome. The announcement paper appeared on 15 February in the journal *Nature*[2]. The following day another draft sequence was published also in the journal *Science* by Celera Genomics[40].

The HGP required 2.7 billion dollars and the collaboration of 20 groups from United States, United Kingdom, Japan, France, Germany and China and it took 13 years to complete with the publication of the full sequence in April 2003.

Since then the reference sequence of the human genome has undergone several updates and improvements (Table 1). The aim was and still remains to obtain a 'human pan-genome', defined as 'the nonredundant collection of all human DNA sequence present in the entire human population'[21]. In this respect a good reference genome should represent the entire human genetic variability as much as possible. The last reference assembly, called GRCh38 and published in December 2013, is now the reference genome most able to satisfy this necessity.

Release name	Date of release	UCSC version	Total sequence length
NCBI Built 33	April 2003	hg15	3,104,781,186
NCBI Build 34	July 2003	hg16	3,091,959,510
NCBI Build 35	May 2004	hg17	3,091,649,889
NCBI Build 36	March 2006	hg18	3,104,054,490
GRCh37	February 2009	hg19	3,137,144,693
GRCh38	December 2013	hg38	3,209,286,105

Table 1: Human reference genome releases since 2003. Data from NCBI website[41].

1.2.2 The GRCh37 and GRCh38 assembly model

The first assembly models allowed a simple linear genome sequences representation, with low diversity in sequence and structure. It was thought that the genome assembly should be represented by a single 'Golden Path', that is a non-redundant chromosome sequence that would fully represent the sequence at all loci[42]. The following and progressive identification of large-scale structural variations and regions with complex allelic diversity unveiled the limit of this model for those genomic regions that required more than one sequence path.

In 2007 the Genome Reference Consortium (GRC), consisting of The Genome Institute at Washington University, The Wellcome Trust Sanger Institute, The European Bioinformatics Institute and The National Center for Biotechnology Information, formalized a new assembly model (Figure 3)[43]. This new model was used for the first time with the GRCh37 release and then with the GRCh38 release. The main advance was the introduction of ‘alternative sequence paths’ in regions with complex sequence and structural variations. According to this model: i) assembly is constructed by one or more assembly units; ii) sequences for the non-redundant haploid assembly are contained in the primary assembly unit, that includes also unlocalized scaffolds (known chromosome but unknown location or order) and unplaced scaffolds (unknown chromosome); iii) alternate loci and patches are placed in separate assembly units.

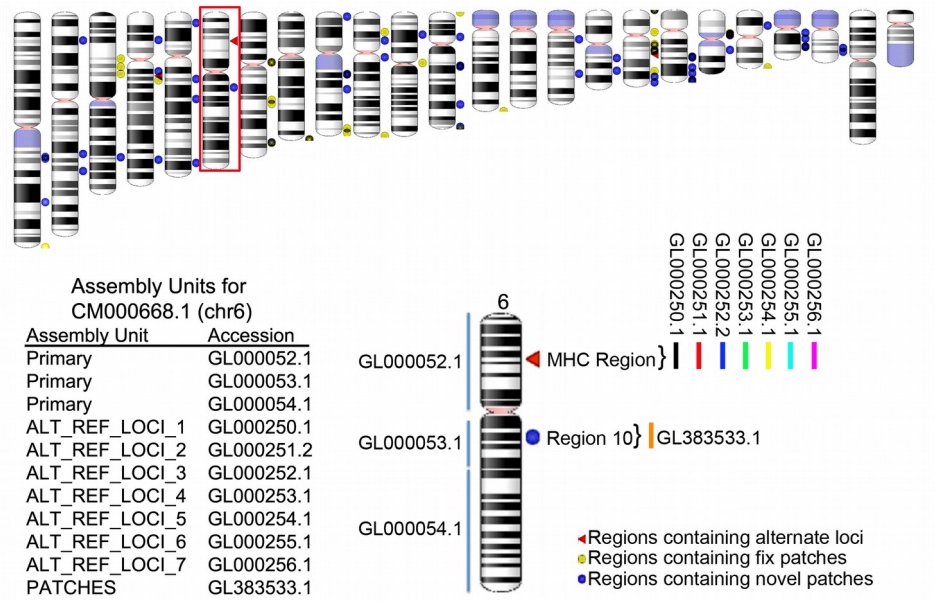

Figure 3. The GRCh37 and GRCh38 assembly model. The figure shows an ideogram representation of the human genome with a blow-up on chromosome 6. The primary assembly unit contains the chromosome sequence, unplaced scaffolds and unlocalized scaffolds. The separate assembly units contain alternate loci and patches. The highly variant MHC region is represented by 7 alternate scaffolds placed in different assembly units, as they are different representations of the same sequence. All patches are placed in a unique assembly unit. Red triangles indicate alternate loci, yellow circles represent fix patches and blue circles represent novel patches. Reprinted from Church *et al.*[43].

As a consequence, the new assembly model is neither haploid nor diploid; instead it includes alternate loci scaffolds providing an alternate representation of highly variable regions and divergent haplotypes and necessary for representing structurally complex loci. Patches allow the continuous correction of errors (fix patches) and addition of alternate loci (novel patches) in the assembly without changing the chromosome sequences or coordinate system. The introduction of patches is considered a 'minor' assembly update and in the next 'major' assembly release fix patches will be introduced as sequence corrections, while novel patches will be moved to a proper assembly unit.

The new assembly model allows to maintain the linear chromosome representation and to add improvements and corrections without frequently change the coordinate system. Even more importantly, with the possibility to introduce complexity and heterogeneity to the assembly, the model satisfy the need to make the human genome reference a pan-human genome, rather than the representation of single individuals or population groups. For example, the Major Histocompatibility

Complex (MHC) region, known to have a high degree of allelic complexity, is represented by 8 different paths in GRCh38[42].

1.2.3 From GRCh37 to GRCh38

The GRCh38 was released to the International Nucleotide Sequence Database Collaboration (INSDC) on December 2013 (GCA_000001405.15) by the GRC. The consortium was aware of the efforts required to move to a new assembly, but at the same time it stated that updating the coordinate system had become essential[1]. In fact 13 patches releases for GRCh37 were made available on public databases in the period from 2009 to 2013 (GCA_000001405.2 – GCA_00000 1405.14), but softwares and file formats were unable to handle the complexity introduced and the use of the new information was limited.

GRCh38 was assembled from the DNA of multiple donors and the gold standard Sanger sequencing was used to produce longer reads and more accurate sequences than high throughput short read sequencing.

Compared to the previous GRCh37.p13 release (GCA_000001405.14, June 2013), the total sequence length of GRCh38 decreases, as well as the total gap length, even if the gap number increases (Table 2). The increase in gap count is mainly due to the replacement of the single centromere gap on all GRCh37 chromosomes. Also the number of regions with alternate loci or patches increases. Lastly, a 26.9.0% increment in exome size is present in GRCh38[37].

Major improvements in the new release include:

I. correction of 8248 erroneous bases, 35 of which were annotated as ClinVar variants in GRCh37;
II. addition of missing copies of segmental duplications with emphasis on paralogous sequences;
III. introduction of centromere sequence replacing the 3 Mb centromeric gap on all GRCh37 chromosomes with modeled centromeres derived from a database of centromeric sequences;
IV. introduction of 261 alternative scaffolds (ALT) to represent diverse haplotypes in 178 chromosomal regions;
V. reduction of sequence gaps.

Genome assembly GenBank accession Submission date	Total sequence length	Gap number	Total gap length	Regions with alternate loci or patches
GRCh37.p13 GCA_000001405.14 June 2013	3,234,834,689	271	243,146,473	182
GRCh38 GCA_000001405.15 December 2013	3,209,286,105	349	159,970,007	207
GRCh38.p12 GCA_000001405.27 December 2017	3,257,319,537	349	161,368,351	317

Table 2. Comparison between assemblies. Data from the NCBI website[44–46].

Thanks to these improvements, GRCh38 is now the reference assembly most able to represent the extent of structural variation and population genomic diversity. However it is still not a perfect representation of the human reference genome[37]. The Genome Reference Consortium declared that none of the recently published individual human *de novo* assemblies yet overcome the quality of GRCh38, even if some sequences are still missing[1].

Since the first release of GRCh38 on December 2013 to the submission of the latest GRCh38.p12 release (GCA_000001405.27, December 2017), 12 patches updates have been submitted. The total sequence length increases, as well as the number of regions with alternate loci or patches (Table 2).

1.2.4 Missing sequences in the human reference genome

As just mentioned, none currently available reference genome can be considered entirely complete and correct. One of the main problems are missing sequences. In the last decade several studies have addressed the problem of missing sequences in the human reference genome and have significantly contributed to the continuous update and amelioration of the reference itself.

In 2010 Li *et al.* estimated that the NCBI Build 36.3 lacked about 20-40 Mb of novel sequences[21]. This and other preliminary studies[47–49] identified novel sequences absent in the reference genome analyzing only few individuals. More recent studies overcome the analysis of individual genomes and discovered novel sequences common in the human populations.

In 2013 Genovese and colleagues proposed a new approach for localizing human genome sequences that had not been included or mapped in GRCh37[50]. The approach was based on patterns of sequence variation that have been created by the admixture of human population. It allowed to successful localize 70 scaffolds spanning 4 Mb pairs of unplaced euchromatic sequences. Even more important, they highlighted the presence of 8 new cryptic segmental duplications (or paralogs) of known genomic sequences; these duplications are missing in the reference genome as they have been considered as the same sequence of their known paralogs. Utilizing the same admixture mapping approach and adding new genome data, some months later the authors published a similar study in which they described the localization of 569 scaffolds containing almost 20 Mb of sequences unlocalized or missing from GRCh37[51]. Only 38 of these scaffolds were shared with the previous work.

In the same year a list of gene fragments missing in GRCh37 was provided by Chen *et al.*[52]. They compared the NCBI human reference genome build 37.2 with the Celera genome[40] and the genome assembly from the Craig Venter Institute, called HuRef genome[53]. They reported that none of the compared human genome assemblies was fully complete and estimated that 3.78 Mb from Celera and 2.37 Mb from HuRef were either missed from their homologous chromosomes on the NCBI 37.2 assembly or partially or completely absent from it.

In 2014 a study performed by Liu *et al.* identified 309 missing common sequences (micSeqs) with a length of at least 100 bp and present in at least 1% of the human population, but absent in GRCh37[54]. They reported also that on average each individual had 50 micSeqs comprised of 5 kb or more sequences that were absent in the reference genome. The comparison with GRCh38 revealed that sequences with similarity higher than 95% were detected in the latest reference genome for only 43 of 309 micSeqs. The authors suggested the remaining micSeqs as candidate for integration in the following release.

With the single-molecule sequencing of a haploid human genome, in 2015 Chaisson *et al.* resolved 50 gaps and extended the boundaries of others 40 gaps, adding respectively 398 kb and 721 kb of novel sequence to GRCh37[55].

1.2.5 'Decoy' and 'sponge' databases to compensate for missing sequences

In 2014 Li performed the first study about the incompleteness of the reference genome as source of mistaken inference in NGS data analyses[38]. The author evaluated the number of heterozygous variants in a haploid genome using three different assemblies: GRCh37 and GRCh38 primary assemblies and hs37d5. The latter contains extra 35.4 Mb sequences missing from the GRCh37 primary assembly and called 'decoy' sequences, as they are supposed to attract many mismapped reads. The number of heterozygous Single Nucleotide Polymorphisms (SNPs) and insertions/deletions (INDELs) called with GRCh37 was double in comparison to hs37d5. This demonstrated that the lack or the under-representation of sequences in the commonly used GRCh37 reference account for reads misalignments and false positive variants identification. GRCh38 further resolved a fraction of heterozygotes called from hs37d5, but it retained some heterozygotes called from GRCh37 but not from hs37d5. Li was not able to clarify the source of these false heterozygous variants and in general concluded that hs37d5 and GRCh38 are more complete than GRCh37.

One year later Miga *et al.* focused on the problem of the under-representation of repeat-rich sequences in both GRCh37 and GRCh38 in mapping and peak-calling steps of Chromatin Immunoprecipitation Sequencing (ChIP-Seq) pipelines[39]. Reads deriving from these missing sequences are forced to map to a small number of homologous regions, resulting in inappropriate alignments and high read-depth signals. To address this problem, they constructed mapping targets, defined as the 'sponge' sequence database, that represent roughly 8.2% of the HuRef genome generally omitted from the reference assembly. The sponge database provides a larger representation of sequences (128,636 fasta sequences, 201 Mb) compared to the previously published decoy genome (4,715 fasta sequences, 35.4 Mb). The integration of this wider database in standard mapping and peak-calling protocols lead to a 10-fold reduction in bases aligned to the so called 'blacklisted regions', *i.e.* a collection of signal artifact regions in the human genome[56].

Despite the more accurate representation of segmental duplications and alternate loci in the latest release of the human reference genome[1], the problem of missing sequences concerns also GRCh38 and in December 2014 the Decoy version 1 for

GRCh38 (hs38d1) with a total length of 5.7 Mb was submitted to NCBI[57]. The inclusion of these sequences in the read alignment process allows a better read mapping of highly repetitive sequences that are difficult to align[58].

1.2.6 Segmentally duplicated genes not represented in the reference genome

As mentioned above, a source of missing sequences in the human reference genome is the presence of cryptic segmental duplications of known genomic sequences[50]. The problem of copy number-variable genes incorrectly classified as diploid in the reference genome was studied in 2010 by Sudmant *et al.*[59]. By analyzing 159 human genomes from the 1000 Genome Project, they discovered 173 segmentally duplicated regions present in the majority of genomes with a copy number greater than that of the reference genome. Among the 44 'hidden' duplicated gene families, they cited *ANKRD* (about six missing copies), *NBPF* (more than nine missing copies) and *NPIP* (about five missing copies) gene families. In 2015 always Sudmant and colleagues sequenced the genome of 236 individuals from 125 different human populations and identified 2,026 loci (corresponding to 6.2 Mbp) of fixed-copy 2 in all human genomes but absent from the reference genome[60].

It is well known that duplicated genes can gradually accumulate mutations over time, becoming non-functionalized, sub-functionalized or neo-functionalized[61]. However, they may retain a high degree of sequence homology, especially if they duplicated recently. These paralogous differences are known to contribute to the false positive variants calls in NGS analysis[62]. A database of nucleotide variants in duplicated gene loci was published in 2011[62,63], but it was based on the NCBI build 36.3 of the reference genome and it has not been updated.

Given the difficulty in distinguishing variants between duplicated genes, in 2013 Nuttle *et al.* proposed a sequencing-based method for genotyping duplicated genes using molecular inversion probes (MIPs), short oligonucleotides designed to target unique paralogous sequences variants[64].

However, the identification of variants in duplicated genes is still a challenging task and these regions are often excluded from NGS data analysis as considered low-confidence regions[65,66].

1.2.7 GRCh38 in NGS data analysis

In 2017 the Genome Reference Consortium published a paper to describe the updates introduced with the release of the latest human reference genome and the improvements provided by GRCh38 compared to GRCh37 on genomic analysis[1]. They mapped genome reads on the GRCh37 and GRCh38 primary assemblies and to the GRCh38 full assembly, which consists of the primary assembly plus alternate loci and patches[42]. They reported that 64.32% of the unmapped reads in GRCh37 mapped to the GRCh38 primary assembly, mainly on new sequences added at GRCh37 gaps. These data confirmed GRCh38 as a more robust mapping target. They reported also that 23.71% of the unmapped reads in the GRCh38 primary assembly mapped to the GRCh38 full assembly, thanks to the more accurate representation for population variation provided by alternate loci. Furthermore, the read mapping was investigated considering only the 2.6 Gb of unchanged reference sequence: 4.19% of reads mapped uniquely but imperfectly to a not modified region of GRCh37 mapped to a different location in GRCh38. In many cases this new location corresponded to GRCh38 centromeres. In conclusion, Schneider *et al.* recommended to use GRCh38 as substrate for genomic analyses.

In the same period Guo *et al.* analyzed 30 exomes using both GRCh37 and GRCh38 and quantified the difference in using the two releases[37]. First of all, all 30 samples showed an improved mapping rate with GRCh38, in particular in exome regions. Moreover, the number of both Single Nucleotide Variants (SNVs) and INDELs identified with GRCh38 decreased; also fewer structural variants were identified using GRCh38.

The higher percentage of reads mapped to GRCh38 compared to GRCh37 was confirmed also by the realignment of 1000 Genomes Project reads to GRCh38 performed by Zheng-Bradley *et al.* in 2017[58]. In this study they used a complete version of GRCh38 that includes the primary assembly, the mitochondrial genome, unlocalized contigs (known chromosome but unknown location or order), unplaced contigs (unknown chromosome), the Epstein-Barr virus (EBV) sequence, alternative contigs, decoy sequences and more than 500 HLA sequences. To

manage a such complex reference, they performed the alignment with a new BWA version (v. 0.7.12) able to handle alternate contigs.

Although the documented improvements in using GRCh38, researchers have been slow to switch to the latest human reference genome and GRCh37 is still largely being used. It was calculated that in 2016 the total number of GRCh38 BAM submissions to the NCBI Sequence Read Archive (SRA) represented only one third of the total number of GRCh37 BAM submissions; furthermore in the period October 2013 - December 2016 the total amount of CRAM submissions to the European Nucleotide Archive (ENA) consisted of 39% GRCh38 and 60% GRCh37[1]. Several reasons explain the hesitation of researchers to switch to GRCh38, first and foremost the delay in updating analysis pipelines and tools previously developed on GRCh37 reference. For example only with the Torrent Suite™ Software v5.2.1 (17 October 2016) (Thermo Fisher Scientific), GRCh38 was introduced as optional reference to be used[67].

1.2.8 Towards the graph of human variation

In the last years a new idea of the assembly model is spreading in the scientific community: the graph-based assembly, with edges representing all variation found within the source sequences (Figure 4). To better describe human diversity, a genome graph that compactly includes an ensemble of possible sequences would be more appropriate than a single reference genome[68].

One of the first attempts for creating a graph-based genome was made in 2015 by Dilthey *et al.*[69]; they proposed a population reference graph (PRG) to represent known genetic variation combining multiple reference sequences and catalogues of variation.

A very recent work carried out by a task team of the Global Alliance for Genomics and Health (GA4GH) reported different methods for graph construction and demonstrate that, in comparison to GRCh38, genome graphs improve the fractions of reads that map uniquely and perfectly[70].

The task is surely not trivial and researchers have explored the different possibilities for solving some aspects, for example how to define the coordinate systems[71] and the handling of hundreds of different sequences that overlap each

genomic location. With this premises, further efforts will be necessary not only to define the new model of the reference genome, but also to develop databases and tools able to support it. With this scope, the first graph genome toolkit have been recently developed by Rakocevic and colleagues[72]. The toolkit includes a Graph Genome Aligner, which maps sample reads to the Graph Genome Reference taking into account many alternate haplotypes for each locus, and a Graph-Genome-Assisted Variant Caller. The authors claimed that their pipeline improves read mapping sensitivity and improves SNP recall by around 0.5% over the coupled use of BWA and GATK.

At present, the graph-based assembly seems to be the most promising alternative to represent the human genome reference.

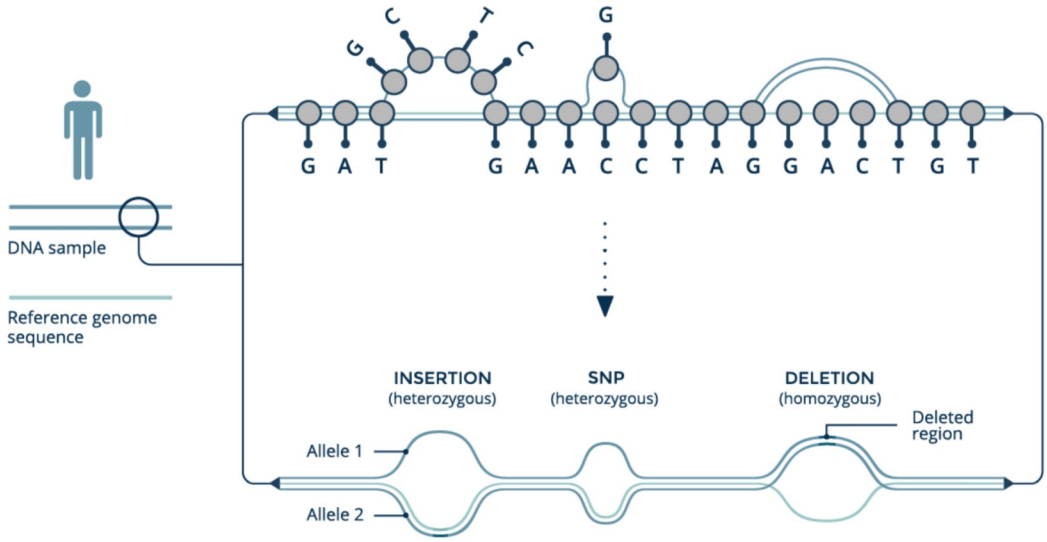

Figure 4: Schematic representation of a graph genome reference. The graph backbone is the linear reference assembly, while edges are additional variants. Reprinted from the SevenBridges website[73].

Chapter 2

Project outline

The previous introductory paragraphs describe the state of the art of the human genome reference, the route taken to date and the way forward. In this perspective, my PhD project aimed to contribute to the identification of further inconsistencies in the reference genome and to determine the impact of these inaccuracies on exome and genome sequencing analyses.

The work originated from the analysis of a heterogeneous dataset of 222 exomes produced at the CRIBI center of the University of Padua using the Ion Proton technology (Thermo Fisher Scientific). The 222 samples derived from patients included into different medical studies and from healthy controls. Surprisingly, using the GRCh37 reference genome in the alignment and variant calling steps, I found that some variants were unexpectedly frequent and, given the heterogeneity of the projects involved, they were supposed to be not correlated with any pathology. I defined variants found with an allelic frequency higher than 50% as high frequency (HF) variants. It should be considered that the reference genome should ideally contain the most common alleles in the population and, as a result, variants with allelic frequencies above 50% should not be theoretically caught. Interestingly, when I performed the alignment and the variant calling with GRCh38, the latest release of the human reference genome, I found that the large majority of HF variants were again identified.

First of all, I excluded the possibility that HF variants might be instrument-specific sequencing errors. For this purpose I performed some technology assessments using independent datasets obtained with Illumina and SOLiD sequencers. As will be further detailed, I found that exomes obtained by different technologies exhibit a largely overlapping set of these HF variants, indicating that the problem was not due to artefacts of a particular chemistry or sequencing platform.

It was clear that the nature of HF variants should be found elsewhere. Thus, I proposed a wide range of analyses to discover inconsistencies in the human reference genome, both in terms of assembly and nucleotide sequence. In fact, as

extensively discussed above, an incomplete or not entirely correct reference genome could cause the identification of false positive variants and, as a results, errors in the interpretation of exome data[38]. Importantly, this applies both to GRCh37 and GRCh38[38].

The literature reports that several thousand positions of the reference genome do not carry the major allele of the population[27,74]. In these positions, defined as Minor Alleles in Reference (MAiRs), variant callers identify an alternative allele that indeed represent the most common one, thus increasing the number of false positives. To identify HF variants in the Ion Proton dataset falling in MAiR positions, I analyzed their allelic frequencies in three different variant databases. A genomic position was marked as MAiR if the reference allele frequency was lower than any alternative allele frequency in all three databases. Using GRCh37 I found 18,839 HF variants mapped on MAiRs. I further checked whether they have been corrected in GRCh38 and found that this occurred only for 1,808 HF variants, while the remaining 17,031 were unchanged.

MAiRs undoubtedly provide a very easy and satisfactory explanation of the identification of HF variants. However, even after removing what is reported as common variant in the database with appropriate tools[27], many shared variants were still remaining. Therefore, the presence of some HF variants cannot be explained only by MAiRs.

A closer look at some of the HF variants revealed that assembly errors in the reference genome may also be involved in the problem. In particular, I observed that the sequence coverage was consistently higher than the average in some specific regions of all individuals. A possible explanation is that there may be genomic duplications that are reported as single regions in the reference, which could be the source of false variant calling[62]. This can be experimentally verified because any ‘collapsed’ repeated sequence in the reference genome would be the target for reads derived from two or more real genomic regions, resulting in a disproportion between frequency, heterozygosity and homozygosity of alleles. Indeed, any difference between two repeats would be seen in all individuals as a heterozygous variant mapping on the ‘collapsed’ reference. Figure 5 provides a schematic representation of a tandem duplication that in the reference is collapsed

into a single region.

To prove my hypothesis I performed a statistical test to compare the observed and the expected heterozygous genotype frequency of each variant. Overall, in the Ion Proton exome target I found 45 gene presenting variants with unbalanced heterozygosity in GRCh37. I decided to investigate whether or not the reference genome of these 45 'unbalanced' genes was modified in the GRCh38 release. I selected reads previously mapped on the unbalanced genes and I re-mapped them to GRCh38.

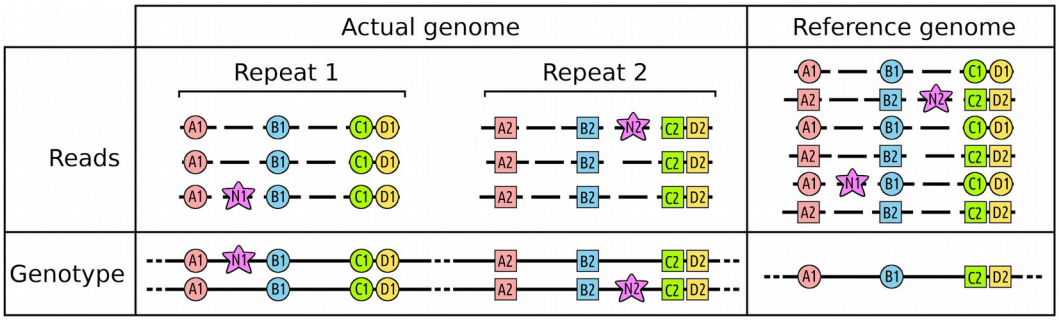

Figure 5. Hypothetical genomic region with a tandem repeat. The hypothetical tandem repeat is almost identical with the exception of four positions: A, B, C and D. This condition may be ancestral and shared by the entire population, repeat 1 having A1, B1, C1, D1 and repeat 2 having A2, B2, C2, D2. Two new variants are also shown as N1 and N2. Sometimes this kind of repeat may be misassembled in the reference genome, being reported as a single collapsed sequence, as shown in the bottom frame on the right. As a result, the four loci A, B, C and D will show a heterozygous genotype in all the individuals and the consequent variant call in all the loci, which is incompatible with the genetics.

This analysis demonstrated that for only 15 genes a duplicated region have been reported in the primary assembly of GRCh38. As a result, these amended genes lost their unbalanced variants in GRCh38. However, the remaining 30 genes remained unchanged or only partially corrected in GRCh38.

Being aware of the importance of these findings, I decided to move my analyses towards a new and wider direction and I extended the analysis to whole genome sequencing data available at the 1000 Genomes Project website[75].

First of all, I found that the surprisingly high number of HF variants observed in exomes was also confirmed in whole genomes. Then I evaluated the total number of HF variants found in GRCh37 that have been amended in GRCh38 and I observed that this occurred for only the 3% of HF variants. These findings clearly

suggested that a further deep revision of base pair level errors is necessary to make the reference genome the accurate representation of the most common DNA sequence in the population.

In addition to base pair level errors, I detected several exomic regions hiding duplications not reported in GRCh37. In search of all the possible unreported duplicated regions in the entire genome, I performed a statistical test for the unbalanced heterozygosity on the 1000 Genomes Project data. Many regions with a strong unbalanced heterozygosity were detected. All these unbalanced regions might conceal a duplication and require to be carefully revised.

Results described so far highlighted two important aspects. Firstly, all resequencing analyses should take into account that false positive variants could originate from the reference used. Secondly, although the improvements of GRCh38, some reference driven problems are still detectable. Exome and genome data analyses allowed me to accurately identify some of these problems, for example unreported gene duplications and genomic positions that do not represent the most frequent alleles. I believe that these findings should encourage the Genome Reference Consortium to update and correct the human reference genome. Moreover, providing a repertoire of the recurrent miscalls, my work will help geneticists in analysing exome data, facilitating the process of variant prioritization.

Chapter 3

Materials and Methods

3.1 Exome Datasets

In this study, three different exome datasets were used. The main dataset was composed by 222 exomes enriched with the Ion AmpliSeq Exome panel and sequenced with the Ion Proton system (Thermo Fisher Scientific) at the CRIBI facility at the University of Padua[76]. These samples came from a wide range of projects including cohorts of individuals, trios and individual patients (Table 3). The second dataset included 22 exomes enriched with the Illumina TruSeq Exome panel and sequenced with the Illumina NextSeq 500 platform at CRIBI[76]. The third dataset referred to the study published by de Ligt *et al.*[77] on 300 exomes enriched with SOLiD-optimized target enrichment and sequenced with SOLiD 4 System (Life Technologies), belonging to 100 trios composed of patients with unexplained severe intellectual disability and their unaffected parents[78].

Project	Exomes number
1	47
2	45
3	29
4	22
5	18
6	17
7	10
8	9
9	9
10	6
11	5
12	2
13	2
14	1
Total exomes	222

Table 3. Number of Ion Proton exomes for each project.

3.2 Alignment and variant calling

Unless otherwise specified, in this work I took as reference genome the GRCh37 primary assembly, that is about 3.1 Gb long, containing one single consensus base per position. However, full assemblies are also available, which consist of the primary assembly plus alternate loci and patches[42]. In this respect, the GRCh37 and GRCh38 full assemblies contain respectively 31 and 46 Gb.

3.2.1 Ion Proton dataset

Exomes were sequenced to reach a final mean coverage of 80x and a target uniformity higher than 90%. Reads were aligned against the GRCh37, as recommended by the manufacturers of the enrichment kits. Alignment and variant calling were carried out according to the Torrent Suite 5.0 exome analysis pipeline. Briefly, alignment was performed with tmap (v. 5.0 with the following parameters: -J 25 --end-repair 15 --do-repeat-clip stage1 map4) and variant calling was performed with the Torrent Variant Caller (v. 5.0) with germline high stringency parameters, as supplied by the producer. Variants were merged into a unique file using CombineVariants of Genome Analysis Toolkit (GATK v. 3.6) and then normalized applying the method proposed by Tan and colleagues[79] in order to eliminate different representations of the same variant. Figure 6 recapitulates the main steps of the Ion Proton exome pipeline from the alignment to the variants collection in the VCF file.

Variant annotation, based on GRCh37.82 version of Ensembl transcripts, was performed using an in-house software.

In 2014 Life Technologies provided a new smaller exome BED file, the Ion AmpliSeq Exome Hi-Q Effective Regions, without actually changing the AmpliSeq Exome panel. In this file poor performing regions are masked during the variant calling step. According to the manufacturer, the usage of this file should guarantee a higher confidence variant calling. To remove possible discrepancies in our dataset caused by the usage of different BED files, only variants covered by the new BED file were considered in this study. The complete list of these variants is available at https://github.com/margheritaferrarini/PhD-Thesis-Exome-Data.

The 222 Ion Proton exomes were also aligned on GRCh38.p10, downloaded from Ensembl. Alignment and variant calling were performed according to the Torrent Suite 5.0 exome analysis pipeline. The target regions (release 2014) were migrated from GRCh37 to GRCh38 coordinates using CrossMap (v 0.2.5)[80]. Variants from all samples were merged and processed as described above.

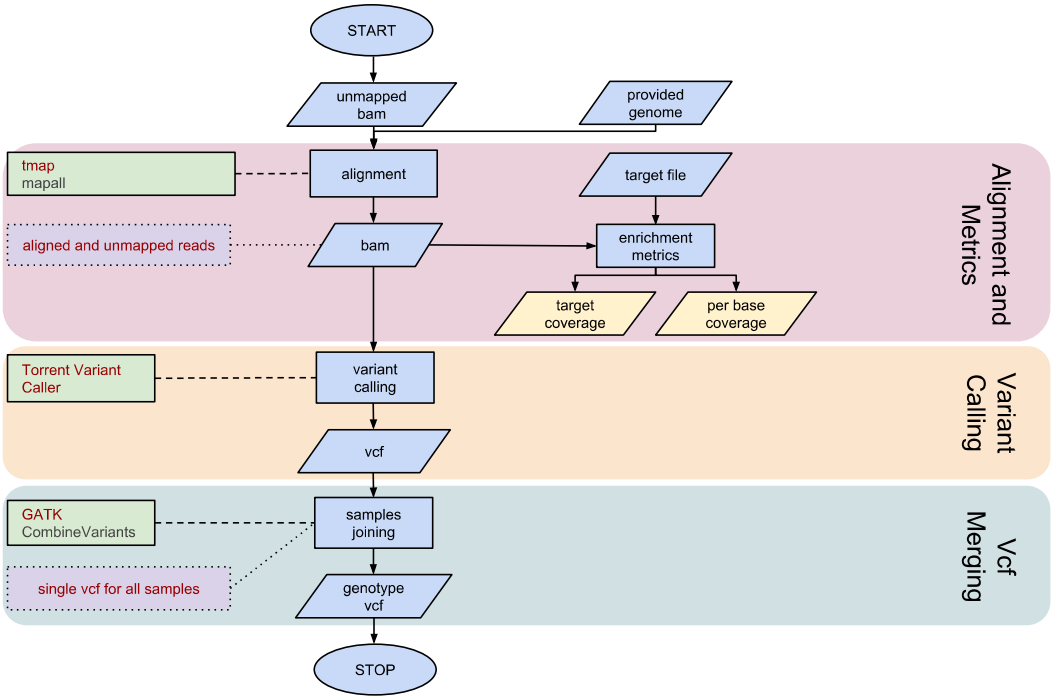

Figure 6. Ion Proton exome pipeline.

3.2.2 Illumina dataset

Each sample was sequenced with 75 bp paired-end reads by Illumina NextSeq 500 to a final average coverage of 103x. Reads were aligned against the GRCh37 primary assembly. Alignment and variant calling were performed according to the recommendations of the GATK Best Practices[81]. Briefly, reads were aligned using BWA mem (v. 0.7.12) with default parameters. The resulting BAM files were further processed by Picard MarkDuplicatesWithMateCigar (Picard v. 1.55) and GATK BaseRecalibrator (GATK v. 3.6). Variant calling was performed using GATK HaplotypeCaller (GATK v. 3.6) with default parameters. Single VCF files were then combined with the GATK JointGenotype (GATK v. 3.6). The collected variants were firstly filtered using GATK VariantRecalibrator (GATK v. 3.6) and then normalized as previously described. The complete list of these variants is available at https://github.com/margheritaferrarini/PhD-Thesis-Exome-Data.

Figure 7 recapitulates the main steps of the Illumina exome pipeline from FASTQ files preprocessing to variants filtering.

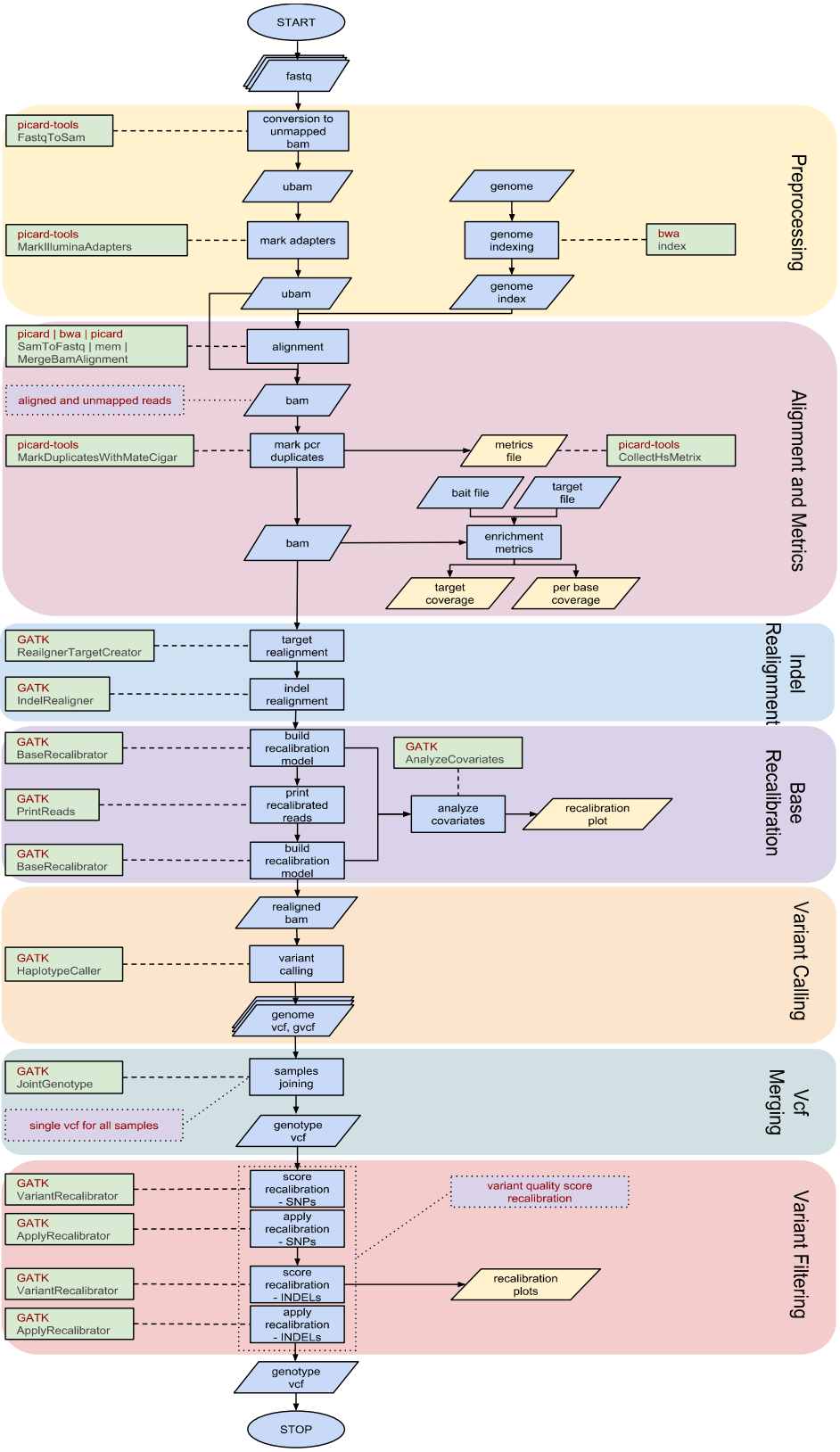

Figure 7. Illumina exome pipeline.

3.2.3 SOLiD dataset

VCF files of de Ligt *et al.*[77] were downloaded from The European Genome-phenome Archive[78]. Variant normalization was performed as indicated above.

3.3 Identification of exome variants mapped on MAiRs

Minor Allele in Reference (MAiR) are those positions of the human reference genome with an allele that is not the most frequent in the population[27]. To identify variants in the Ion Proton dataset falling in these positions their allelic frequencies were analyzed in 3 different variant databases: i) the Single Nucleotide Polymorphism database (dbSNP)[82,83] version 144, modified to recover old variants excluded from this release but present in the online version; ii) the NHLBI Exome Sequencing Project (ESP) database version ESP6500SI-V2[84]; iii) the Exome Aggregation Consortium (ExAC) database version 0.3.1[85,86]. When different populations frequencies were present, only the total one was considered. A genomic position was marked as MAiR if the reference allele frequency was lower than any alternative allele frequency in all three databases.

3.4 Impact of MAiR positions at the protein level

Variants in GRCh37 MAiR positions confirmed in GRCh38 were annotated using both SnpEff (v. 4.2)[87] and VEP (v. 84)[88], employing respectively UCSC and RefSeq transcripts. These two different annotations were chosen to avoid transcript-dependent biases. Missense variants were selected from the two annotated VCF files and analyzed with an in-house Python script (available at https://github.com/margheritaferrarini/PhD-Thesis-Scripts, file MAiR_Uniprot.py) to allocate them to one of the following three classes: i) match to the manually reviewed human protein sequence of SwissProt, ii) match to the Human Polymorphisms and Disease Mutations release 2017_05 of UniProt, iii) neither of the above.

3.5 Statistical test on heterozygous genotype frequencies

The Ion Proton dataset was searched for variants with unbalanced heterozygous genotype frequency. Each variant was tested with a one-tailed binomial test

between observed and expected heterozygous genotype frequencies. Observed frequencies were calculated as the number of times that genotype occurred divided by the total number of exomes (222 exomes). Expected frequencies were computed with the Hardy-Weinberg formula in which the heterozygous genotype frequency can be calculated as $x=2pq$, where q is the alternative allele frequency, *i.e.* the number of times that specific allele was found divided by the total allele number (444 alleles), and p is the reference frequency calculated as $1-q$. Resulting probabilities were corrected for false discovery rate using the Benjamini-Hochberg procedure[89]. Variants were considered significantly unbalanced if the corrected probability was lower than 0.01. This analysis was performed only for biallelic variants, defined as loci that have two observed alleles: the reference and one alternative allele.

The same test was performed also on whole genome data, in particular on the GRCh37 dataset of Phase3 1000 Genomes Project. In this case observed frequencies were computed as the number of heterozygous genotypes divided by the total number of genomes and the alternative allele frequency (q in the $x=2pq$ formula) was reported in the VCF file. A WIG* format file was then obtained as follow: for each chromosome the percentage of resulting unbalanced variants on the total number of biallelic variants was calculated in non-overlapping 10 kb sized windows. Values range from 0.0 to 100, with 0.0 indicating the absence of unbalanced variants in the given window; NaN values indicate that any biallelic variant was found in the given window.

These analyses were performed with several in-house developed Python scripts (available at https://github.com/margheritaferrarini/PhD-Thesis-Scripts, files unbalanced_heterozigosity_EXOMES.py, unbalanced_heterozigosity_1000G.py, unbalanced_heterozigosity_gnomAD.py).

*WIG (wiggle) format: The WIG format is designed for display of dense continuous data such as probability scores. A WIG file consists of one of more blocks, each containing a declaration line followed by lines defining data elements. There are two main formatting options: fixedStep and variableStep. VariableStep format is designed for data with irregular intervals between data points and is the more commonly used format. It begins with a declaration line, followed by two columns containing chromosome positions and data values. FixedStep format is designed for data with regular intervals between data points and is the more compact of the two wiggle formats. It begins with a declaration line, followed by a single column of data values.

3.6 Confirmation of unbalanced variants of *MAP2K3*

VCF files containing variants in chromosome 17 were downloaded from two different databases: Genome Aggregation Database (gnomAD) version 2.0.1[86] and 1000 Genomes Project database Phase1 release[90] and Phase3 release[91]. Using the one-tailed binomial test described above, variants with a heterozygous genotype frequency significantly higher than the expected were selected and subsequently compared with variants identified in *MAP2K3* in the Ion Proton dataset.

3.7 Frequency of bases updated in GRCh38

VCF file with the 8,248 bases updated in GRCh38 by Schneider *et al.*[1] was downloaded from Genome Research Supplemental Material[92]. Frequencies of variants in these positions were downloaded from 1000 Genomes Project database Phase1 release[90], the same used by Schneider *et al.*, and plotted with an in-house developed Python script (available at https://github.com/margheritaferrarini/PhD-Thesis-Scripts, file schneider_plot.py).

3.8 HF variants in 1000 Genomes on GRCh37 and GRCh38

1000 Genomes Project VCF files Phase3 for both GRCh37 and GRCh38 were collected respectively from 1000 Genomes ftp website[91] and Ensembl ftp website[93]. Variants reported with frequencies higher than 50% were marked as high frequency (HF). A HF variant was considered amended if it was called against GRCh37, but not in GRCh38. These analyses were performed using an in-house pipeline (available at https://github.com/margheritaferrarini/PhD-Thesis-Scripts, file 1000G_phase3_comparison.job).

3.9 Analysis of the physical coverage in mate pair whole genome data

Whole genome sequencing mate pair data were downloaded from the Genome In A Bottle project[94]. Samples were parents of an Ashkenazi Trio[95] and a Chinese trio[96]. For details on libraries preparation and sequencing refer to the work of Zook and colleagues[97]. Reads were aligned against the GRCh37 and GRCh38 primary assemblies with BWA mem (v. 0.7.12 with default parameters). An in-house script (available at https://github.com/margheritaferrarini/PhD-Thesis-Scripts, file

local_tracks.py) was used to produce a physical coverage profile in *MAP2K3* and *KCNJ2* regions on chromosome 17.

3.10 Identification of conserved domains in lncRNAs

The method for prioritizing variants in lncRNAs is based on the identification of conserved functional domains by a comparative genomics approach. The pipeline structure consists of three different steps: i) identification of orthologous genes of the human lncRNA in the genomes of 28 primates; ii) multiple alignment of orthologous sequences; iii) identification of conserved domains in the human lncRNA. Each of these steps is detailed below.

3.10.1 Identification of orthologous genes in the genomes of 28 primates

This first step to identify orthologous genes was developed with a Python3 script (available at https://github.com/margheritaferrarini/PhD-Thesis-Scripts, file orthologous_genes_identification.py). It requires four different inputs: i) a FASTA* format file with the human lncRNA sequence; ii) a BED** format file with the human lncRNA genomic coordinates in GRCh38; iii) the GRCh38 human reference genome; iv) a list of files with primate genomes, all of them in FASTA format and downloaded from the NCBI website[98]. The 28 primate organisms included in the analysis are: *Aotus nancymaae*, *Callithrix jacchus*, *Carlito syrichta, Cebus capucinus*, *Cercocebus atys*, *Chlorocebus sabaeus*, *Colobus angolensis*, *Daubentonia madagascariensis*, *Eulemur flavifrons*, *Eulemur macaco*, *Gorilla gorilla*, *Macaca fascicularis*, *Macaca mulatta*, *Macaca nemestrina*, *Mandrillus leucophaeus*, *Microcebus murinus*, *Nasalis larvatus*, *Nomascus leucogenys*, *Otolemur garnettii*, *Pan paniscus*, *Pan troglodytes*, *Papio anubis*, *Piliocolobus tephrosceles*, *Pongo abelii*, *Propithecus coquereli*, *Rhinopithecus bieti*, *Rhinopithecus roxellana*, *Saimiri boliviensis boliviensis*.

First of all, a BLAST alignment is performed between the human lncRNA

*FASTA format: A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The definition line (defline) is distinguished from the sequence data by a greater-than (">") symbol at the beginning. It is recommended that all lines of text be shorter than 80 characters in length.

**BED (Browser Extensible Data) format: A BED file is a tab-delimited text file that defines a feature track. BED lines have three required fields and nine additional optional fields. The first three required BED fields are the name of the chromosome or scaffold, the starting position of the feature in the chromosome or scaffold and the ending position of the feature in the chromosome or scaffold. The optional fields report additional information.

sequence and the genome of the first primate. BLAST default parameters are maintained at this stage, including the soft masking option, as suggested by Moreno-Hagelsieb and Latimer[99]. With this option, BLAST uses the database mask only during the initial word-finding phase.

The BLAST match with the highest *bit-score* is considered the best hit. The *bit-score* is defined as the required size of a sequence database in which the current match could be found just by chance. Thus, the higher the *bit-score*, the more significant the match is. If two matches have the same *bit-score*, the *E-value* is considered. The *E-value* is the number of expected hits of similar score that could be found just by chance. Thus, the lower the *E-value*, the more significant the match is. If two matches have also the same *E-value*, more than one best hit is present.

A further BLAST alignment is performed between the identified best hit sequence and the GRCh38 human reference genome. BLAST default parameters are maintained again. Start and end positions of this second BLAST alignment are then compared with genomic coordinates of the human lncRNA: if start and end positions correspond to lncRNA coordinates or are included between them, the best hit match resulting from the first BLAST alignment is considered the orthologous gene of the human lncRNA. Since the orthologous gene and the human lncRNA, each in a different genome, find each other as the best scoring match in the other genome, they are defined 'reciprocal best hits' (RBHs).

Both BLAST alignments are then repeated for each of the remaining 27 primates. At the end of this first step, an output FASTA format file is produced with the human lncRNA sequence followed by the 28 orthologous sequences. If one of the 28 primate genomes lacks the orthologous sequence, the organism will be excluded from the following steps.

3.10.2 Multiple alignment of the orthologous sequences

In this second step of my pipeline, the FASTA file with the human lncRNA plus the orthologous genes is used to perform a multiple sequence alignment. For this purpose, the T-Coffee package developed by Notredame *et al.*[100] was chosen because, according to several benchmarks, it is on overall much more accurate

than the most widely used ClustalW[101,102]. At each alignment step of the progressive alignment, T-Coffee considers information from all sequences, not just those being aligned at that stage. The increase in accuracy makes T-Coffee slower than ClustalW (about N times for N Sequences)[102]. Since the number of sequences to align is very low (only 29 sequences) and since the program is by default parallelized, meaning that it can use multiple cores when running on a cluster, T-Coffee slowness does not affect the compute time of the pipeline. Moreover, T-Coffee was preferred rather than other multiple sequence aligners, such as MUSCLE[103] and MAFFT[104], because their usage is recommended when the number of sequences to align is very high.
The Linux/Unix T-Coffee version 11.00.8cbe486 is used as follows:

```
t_coffee -seq sequences.fa -mode regular -output fasta_aln
score_html -n_core=12
```

With this command line three different output files are obtained: i) the alignment file in FASTA format (.fasta_aln); ii) the alignment file in html format (.html); iii) the guide tree in Newick* format (.dnd). This last file can be visualized in R with the Analyses of Phylogenetics and Evolution (ape) package (version 5.1). The guide tree is not a phylogenetic tree, it is used in the alignment process for clustering the sequences.

3.10.3 Identification of conserved blocks in the human lncRNA

The last step defines a set of conserved blocks starting from the multiple sequence alignment by using a program called Gblocks, developed by Castresana in 2000[105]. Gblocks is able to eliminate poorly aligned positions and divergent regions of a DNA alignment. The conserved blocks selected by Gblocks satisfy the lack of large segments of contiguous non-conserved positions, the lack or low density of gap positions and the high conservation of flanking positions. Several parameters can be modified to make the selection of blocks more or less stringent.

*Newick format: In mathematics, Newick tree format (or Newick notation or New Hampshire tree format) is a way of representing graph-theoretical trees with edge lengths using parentheses and commas.

The Linux Gblocks version 0.91b is used as follows (see the Gblocks online documentation[106] for a detailed explanation of each parameter):

```
Gblocks alignment.fasta_aln -t=d -b1=X -b2=Y -b3=3 -b4=3 -b5=a
-s=y -p=t -p=y -v=10000 -n=n -u=n
```

In this case *b1* (threshold for the definition of conserved positions) and *b2* (threshold for the definition of flank conserved positions) are equal to X and Y, since they should be set every time on the basis of the number of previously identified orthologous sequences.

With this command line two different output files are obtained: i) the alignment file with the selected blocks in FASTA format (.fasta_aln-gb); ii) the HTML file with colored conserved positions (.fasta_aln-gb.htm).

Chapter 4

Results and Discussion

4.1 Recurrent variants in the Ion Proton exome dataset

The work presented in this thesis originated from the analysis of a heterogeneous dataset of 222 exomes produced at the CRIBI center using the Ampliseq chemistry and the Ion Proton technology, as detailed in the Materials and Methods. The strength of this dataset came from the fact that the samples derived from patients included into different medical studies and from healthy controls. The overall analysis of the data led to the identification of 264,303 variants called against the GRCh37 reference genome, including 239,255 SNPs and 25,048 small INDELs (14,075 deletions and 10,973 insertions). Among the total variants, 245,088 were detected as biallelic, whereas 19,215 were multiallelic.

Surprisingly, I found that 9,313 variants were present in more than 90% of the individuals and, among these, 2,349 variants were shared by the 100% of the individuals. Given the heterogeneity of the projects involved, these variants were supposed to be not correlated with any pathology. I immediately realized that such a high number of recurrent variants was unexpected, considering that the average number of variants in each sample was 48,785. This finding was even more surprising since a considerable number of recurrent variants was reported with a low frequency in databases and in some cases they were not reported at all as known variants, making difficult the process of recognizing them as false positives. Interestingly, when I used the GRCh38 reference genome in the mapping and variant calling steps, I found that 8,132 out of 9,313 remained uncorrected.

The above values refer to the presence of variants in a diploid genome. In terms of allelic frequency, I found that 3,898 variants had an allelic frequency equal or greater than 90%; of these, 841 scored 100% allelic frequency, being homozygous in all the samples. I defined variants with an allelic frequency higher than 50% as high frequency (HF) variants. A total number of 18,043 HF variants were identified.

It should be considered that the reference genome should ideally contain the most common alleles in the population and, as a result, variants with allelic frequencies above 50% should not be theoretically caught. Fluctuations due to subsampling and/or ethnicity are certainly possible, but cannot explain this high number of HF variants. These findings are not completely unexpected because Minor Alleles in Reference (MAiRs) are a known problem[27,74]; however the large number of their occurrences was notable.

4.2 Comparison with Illumina and SOLiD exome datasets

First of all I verified whether the recurrent variants found in more than 90% of the exomes could result from Ion Proton specific errors. For this purpose I analyzed the exomes of two independent datasets produced with Illumina and SOLiD technologies, using their respective enrichment, sequencing and analysis pipelines, as detailed in the Materials and Methods. In Table 4 it can be seen that the large majority of variants that occurred in more than 90% of the Ion Proton exomes was confirmed also with the Illumina and SOLiD platforms.

Unfortunately, the exomic target regions captured with the three different technologies were not precisely overlapping; thus, of the 9,313 Ion Proton variants, only 6,085 fell in regions covered by the Illumina target and 7,046 in regions covered by the SOLiD target (*'on target'* variants in Table 4).

Ion Proton	**Illumina control dataset**			**SOLiD control dataset**		
	on target		**off target**	**on target**		**off target**
9313	6085		3228	7046		2267
	confirmed	**not confirmed**		**confirmed**	**not confirmed**	
	6008	77		5733	1313	

Table 4. Recurrent variants in Ion Proton exomes and their sharing in Illumina and SOLiD datasets. The 9,313 variants found in more than 90% of the Ion Proton exomes were analyzed to verify whether they were also present in at least 50% of the exomes obtained with Illumina and SOLiD technologies. As the exomic target regions captured with the three technologies did not precisely overlap, *'confirmed'* and *'not confirmed'* refer to variants falling in target regions shared between Ion Proton and Illumina or between Ion Proton and SOLiD (*'on target'* variants). However, also the number of Ion Proton variants outside Illumina and SOLiD target regions is reported (*'off target'* variants). A large percentage of variants that shared the exomic target was confirmed: 6,008 of the 6,085 on target variants were confirmed by Illumina (99%) and 5,733 of the 7,046 on target variants were confirmed by SOLiD (81%).

Variants were considered *'confirmed'* if they were respectively present in at least 50% of Illumina or SOLiD control datasets. As shown in Table 4, the very large majority (99%) of the on target variants were confirmed by Illumina, while 81% were confirmed by SOLiD. Unconfirmed variants could either be false positives of the Ion Proton or false negatives of the Illumina and SOLiD. Only 41 variants that were localized in common target regions were not confirmed in both Illumina and SOLiD and therefore these could be Ion Proton specific systematic errors.

4.3 European and total population allele frequencies

All the exomes of the study belonged to European people. I wondered if variants in the Ion Proton dataset could present a higher alternative allele frequency in the European population in respect to the general population (only frequencies from the ExAC database were considered). In fact, the high number of shared variants in our samples could be explained as european-specific polymorphisms. The plot in Figure 8 shows the almost perfect correlation between the frequencies in the two populations, indicating that there is no evidence of a possible bias due to ethnical origin of the samples.

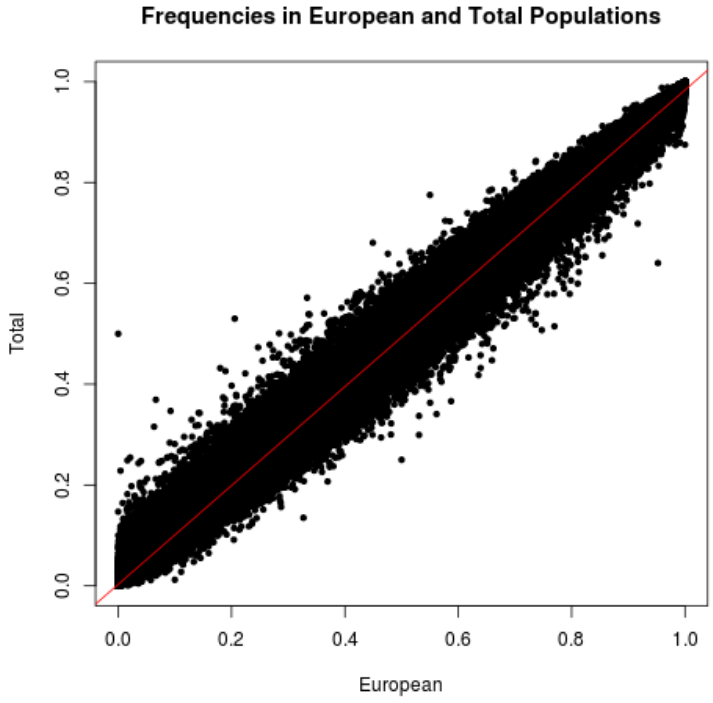

Figure 8. Correlation of allele frequencies between European and Total populations. Frequencies derived from ExAC database. Red line shows the high correlation between the two datasets.

4.4 GRCh38 variants comparison

Alignment and variant calling of the Ion Proton dataset were performed also using GRCh38 as genome reference. I identified 255,124 variants, a smaller number compared to the previous genome release. This was somehow expected as data published by Guo and colleagues[37] claimed a lower number of SNPs due to the improvements introduced in the last release of the human genome reference. The number of variants shared between GRCh37 and GRCh38 was 242,259 (91.66% of the GRCh37 variants dataset).

4.5 Identification of exome variants mapped on MAiRs

As mentioned above, the literature reports that several thousand positions of the reference genome do not carry the major allele of the population[27,74]. In these positions, defined as Minor Alleles in Reference (MAiRs), variant callers identify an alternative allele that indeed represents the most common one, thus increasing the number of false positives.

To identify HF variants in the Ion Proton dataset falling in MAiR positions, I analyzed their allelic frequencies in three different variant databases (dbSNP, ExAC and ESP databases, see Materials and Methods). A genomic position was marked as MAiR if the reference allele frequency was lower than any alternative allele frequency in all three databases. Using GRCh37 I found 18,839 HF variants mapped on MAiRs. I further checked whether they have been corrected in GRCh38 and I found that this occurred only for 1,808 HF variants, while the remaining 17,031 were unchanged.

I also examined whether any of the 17,031 HF variants retained in GRCh38 matched their corresponding protein in the UniProt variation database. I annotated the variants both with SnpEff[87] and VEP[88], employing respectively UCSC and RefSeq transcripts. I detected a comparable number of missense variants in the two databases: 3,814 with RefSeq and 3,761 with UCSC. I found that ~100 of them were included in the reference protein primary sequence, indicating that the alternative allele found at genome level represented the most common amino acid at protein level. Moreover, ~2,800 of missense variants were known as minor protein variants. Finally, ~880 alleles found with high frequency in exomes did not

show any known counterpart at the protein level. These data confirm a previous finding by Barbitoff *et al.*[74] and indicate that some general revision of the reference is required also at the protein level.

4.6 Mining for incongruities

MAiRs undoubtedly provide a very easy and satisfactory explanation of the identification of HF variants. However, the observation that the sequence coverage was consistently higher than the average in some specific regions of all individuals revealed that assembly errors in the reference genome may also be involved in the problem. I hypothesized the presence of gene and region duplications not already annotated in the reference genome as one possible cause for a misleading variant calling in the target regions. Since these duplicated regions can be enriched and sequenced together with the original target gene, the corresponding reads will align to an improper position, causing the identification of variants not really present in the gene. Consequently, a heterozygous genotype should be expected for these variants, with the reference allele deriving from the original target gene and the alternative allele from the duplicated region. Indeed, as shown in Figure 5 of the Introduction, any difference between two repeats in the genome would be seen in all individuals as a heterozygous variant mapping on the 'collapsed' reference.

To verify my hypothesis, I performed a statistical test to compare the observed and the expected heterozygous genotype frequency of each variant. According to the Hardy-Weinberg equation I should expect that $p^2+2pq+q^2=1$, where q is the alternative allele frequency and p is the reference frequency calculated as $1-q$. This null hypothesis was verified with a one-tailed binomial test, corrected for false discovery as discussed in the Materials and Methods. Variants were considered significantly unbalanced if their corrected *p-values* were lower than 0.01.

4.7 Exome target regions with unbalanced heterozygosity

Overall, in the Ion Proton exome target I found 753 variants identified with GRCh37 presenting an unbalanced heterozygosity. In particular I found 145 target regions (amplicons) containing more than one unbalanced variant, for a total of 560 variants spanning over 45 genes. In the process of investigating these regions,

I observed that two different groups of aligned reads were always distinguishable: reads having all the selected unbalanced variants and reads having none of them. This observation suggested a possible different genomic origin of the two pools of reads, even if they aligned on the same region using GRCh37. This was in agreement with the hypothesis of duplicated regions not present in the reference genome used for the analysis. Therefore I wanted to investigate whether or not the reference genome of these 45 'unbalanced' genes was modified in the GRCh38 release.

As explain in the Material and Methods, unless otherwise specified, in this work I took as reference genome the GRCh37 primary assembly, that is about 3.1 Gb long, containing one single consensus base per position. Full assemblies are also available, which consist of the primary assembly plus alternate loci and patches[42]. In this respect, the GRCh37 and GRCh38 full assemblies contain respectively 31 and 46 Gb. To better understand the progress of the current human reference genome, I selected reads previously mapped on the 45 unbalanced genes and I re-mapped them on GRCh37 and GRCh38 full assemblies using BLAST (with an identity percentage cutoff set to 90%). The alignment to the GRCh37 full assembly was useful to check if duplicated regions had been introduced already in the GRCh37 release in the form of alternate loci or patches. Instead the alignment to the GRCh38 full assembly helped to understand if these duplications had been subsequently inserted in the GRCh38 primary assembly or remained in the form of alternate loci or patches.

Assuming that the highest identity percentage indicated the real genomic origin of reads, BLAST results showed several possible scenarios and led to the classification of the 45 unbalanced genes in 5 different classes (Table 5).

I classified 11 genes as 'unchanged' in GRCh38 (Table 5, column 5) since both classes of reads with unbalanced variants and reads without variants aligned only to the target gene, thus indicating that neither duplicated regions nor alternative loci had been reported in the latest release of the reference genome. These genes still presented the same heterozygosity problem as their unbalanced variants were identified also with GRCh38.

Of the remaining genes, only 15 were classified as 'fully amended' in GRCh38

(Table 5, column 1), *i.e.* duplicated in the chromosomal primary sequence. For these genes, reads with unbalanced variants aligned to a different position of the same chromosome where the target gene localized or to a different chromosome, whereas reads with none variant aligned to the target gene. These fully amended genes lost their unbalanced heterozygosity in GRCh38. Among them I found *PRIM2*. This was expected since it had been previously reported as a paralog gene misassembled in GRCh37[50], which was fully amended in GRCh38[1]. *PRIM2* paralog contains only exons 6-14 of the original transcripts[50], that actually correspond to the exons covered by the enriched target regions with unbalanced variants in GRCh37. Since my screening process placed *PRIM2* as an amended gene in GRCh38, it could be considered a 'positive control' that validates my criteria for classifying genes.

5 genes were only partially duplicated and they still had some regions with unbalanced heterozygosity (Table 5, column 2). An interesting case is the *MAP2K3* gene, extensively discussed in the following paragraph.

For other 4 genes reads with unbalanced variants aligned on unplaced scaffolds, so the location of the duplication on chromosome is unknown (Table 5, column 3).

Finally, 10 genes were not duplicated, but reads with variants aligned to alternative loci in the full assembly (Table 5, column 4). According to the Assembly Terminology of the Genome Reference Consortium, an alternate locus is 'a sequence that provides an alternate representation of a locus found in a largely haploid assembly'[107]. For example the *KIR2DL3* gene, coding the killer cell immunoglobulin like receptor, is known to be highly polymorphic in the population[108,109]; many alternate loci for this gene were introduced in the full assembly and as a result they 'trapped' reads with unbalanced variants. However, it should be noticed that highly polymorphic alleles should not produce unbalanced heterozygosity when aligned on the primary reference genome. For the majority of variants in the genes reported in column 4, I found a heterozygous genotype in all the 222 individuals. This was unexpected, as highly polymorphic loci should lead to a mixture of homozygous and heterozygous genotypes. Therefore, some of these genes should be revised as they could be duplications.

1 **Fully amended**	2 **Partially amended**	3 **Unplaced scaffold**	4 **Alternative loci**	5 **Unchanged**
BCLAF1	*FRG2B*	*CTBP2*	*CES1*	*ALG1L2*
CCDC144NL	*FRG2C*	*FAM104B*	*HLA-DQA2*	*ANKRD36*
FRG1	*KCNJ12*	*MLL3*	*HNRNPCL1*	*FAM131C*
HYDIN	*KRT6B*	*NBPF1*	*KIR2DL3*	*FAM194B*
KRTAP4-11	*MAP2K3*		*KIR2DS4*	*GPRIN2*
*LOC653486**			*KRTAP9-2*	*OR1D5*
NBPF10			*MUC20*	*PCDH11X*
NOTCH2NL			*OR9G1*	*PDPR*
OR4C3			*PRSS3*	*PER3*
OR4C45			*TNXB*	*TPTE*
OR4M2				*ZDHHC11*
PDE4DIP				
*PPYR1**				
PRIM2				
SEC22B				

Table 5. Genes with unbalanced heterozygosity in GRCh37 and their status in GRCh38. Column 1: Fully amended genes that have been duplicated within chromosomes in GRCh38 and as a result lost the variants with unbalanced heterozygosity. Column 2: Partially amended genes that are still showing unbalanced variants in some of the exons. Column 3: Genes whose duplication was found on extra chromosomal scaffolds in the primary assembly and as a result lost the variants with unbalanced heterozygosity. Column 4: Genes that have not been duplicated, but reported as different alternative loci in the full assembly. Column 5: Unchanged heterozygosity in GRCh38. **LOC653486* and *PPYR1* have changed name in GRCh38 respectively to *SCGB1C1* and *NPY4R*. More details are given in the text.

4.8 *MAP2K3* as an example of partially amended gene

The *MAP2K3* (MAP Kinase Kinase 3) gene, also known as *MKK3*, encodes the mitogen-activated protein kinase kinase 3. This protein participates in the MAP kinase-mediated signaling cascade and has a well known role in tumor invasion and progression[110,111]. The gene maps on chromosome 17 and includes 12 exons.

Globally, in the Ion Proton dataset I identified 54 unbalanced variants localized in *MAP2K3*. First of all I verified whether these variants were found with an unbalanced heterozygous genotype frequency also in public databases. In particular I selected three genomes databases (gnomAD, 1000 Genomes Project Phase1 release and 1000 Genomes Project Phase3 release, see the Materials and Methods) and I performed the previously described statistical test to check if the 54 variants were reported with an unbalanced heterozygosity. Results are summarized in Table 6.

Chromosome	Position	Reference Allele	Alternative Allele	GnomAD	1000 Genomes Project Phase1	1000 Genomes Project Phase3
17	21201719	T	C	unbalanced	unbalanced	not in database
17	21202056	G	A	not in database	not in database	not in database
17	21202063	G	C	not in database	not in database	not in database
17	21202067	A	G	not in database	not in database	not in database
17	21202078	G	A	not in database	not in database	not in database
17	21202102	C	T	not in database	not in database	balanced
17	21202123	A	G	not in database	not in database	balanced
17	21202191	C	A	unbalanced	unbalanced	balanced
17	21202237	G	C	unbalanced	unbalanced	balanced
17	21202272	C	G	unbalanced	unbalanced	balanced
17	21203893	T	C	unbalanced	unbalanced	not in database
17	21203907	T	C	unbalanced	unbalanced	not in database
17	21203934	G	A	unbalanced	unbalanced	not in database
17	21203941	G	A	unbalanced	unbalanced	not in database
17	21203949	C	T	unbalanced	unbalanced	not in database
17	21203998	G	A	unbalanced	unbalanced	not in database
17	21204153	C	T	unbalanced	unbalanced	not in database
17	21204187	G	T	multiallelic	unbalanced	not in database
17	21204192	C	T	unbalanced	unbalanced	not in database
17	21204210	C	T	unbalanced	unbalanced	not in database
17	21204257	G	A	unbalanced	not in database	not in database
17	21204266	T	C	unbalanced	unbalanced	balanced
17	21204308	G	T	multiallelic	not in database	not in database
17	21204315	T	C	unbalanced	not in database	not in database
17	21204316	G	A	not in database	not in database	not in database
17	21204318	A	G	multiallelic	not in database	not in database
17	21205460	C	T	unbalanced	unbalanced	not in database
17	21207844	C	T	unbalanced	unbalanced	not in database
17	21208413	C	T	unbalanced	unbalanced	not in database
17	21208449	G	T	unbalanced	unbalanced	not in database
17	21208456	A	G	unbalanced	unbalanced	not in database
17	21215483	C	T	unbalanced	unbalanced	balanced
17	21215537	C	A	unbalanced	unbalanced	balanced
17	21215552	C	T	unbalanced	unbalanced	not in database
17	21215557	G	A	unbalanced	unbalanced	not in database
17	21215637	G	A	unbalanced	unbalanced	not in database
17	21215643	A	G	unbalanced	unbalanced	not in database
17	21215682	G	A	not in database	not in database	not in database
17	21215700	T	G	not in database	not in database	not in database

Chromosome	Position	Reference Allele	Alternative Allele	GnomAD	1000 Genomes Project Phase1	1000 Genomes Project Phase3
17	21216661	C	T	not in database	not in database	not in database
17	21216664	A	G	not in database	not in database	not in database
17	21216686	C	A	not in database	not in database	not in database
17	21216710	T	C	not in database	not in database	not in database
17	21216758	G	GCTTC	unbalanced	not in database	not in database
17	21216788	C	T	unbalanced	unbalanced	not in database
17	21216846	G	C	unbalanced	unbalanced	not in database
17	21217397	A	G	not in database	not in database	balanced
17	21217400	G	A	not in database	not in database	balanced
17	21217411	T	C	unbalanced	unbalanced	balanced
17	21217513	G	A	unbalanced	unbalanced	not in database
17	21217547	T	C	unbalanced	unbalanced	balanced
17	21217554	C	T	unbalanced	unbalanced	not in database
17	21217566	C	T	unbalanced	unbalanced	balanced
17	21217586	G	T	unbalanced	unbalanced	not in database

Table 6. List of unbalanced variants in *MAP2K3* compared with gnomAD, 1000 Genomes Project Phase1 release and 1000 Genomes Project Phase3 release. Among the 54 variants with an unbalanced heterozygosity in *MAP2K3*, 36 variants were unbalanced in gnomAD, 34 in 1000 Genomes Project Phase1 release and none in 1000 Genomes Project Phase3 release. Unbalanced: *p-value*<0.01; balanced: *p-value*>=0.01; not in database: variant not reported in the VCF file; multiallelic: variant with multiple alternative alleles.

The majority of variants were confirmed to have an unbalanced heterozygosity in gnomAD and 1000 Genomes Project Phase1 release, where GRCh37 was used as reference genome. The not confirmed variants were absent in the databases or present with multiple alternative alleles (these variants were not included in the statistical test). Differently, in 1000 Genomes Project Phase3 release variants were collected using the hs37d5 genome reference, which corresponds to the GRCh37 primary assembly integrated with rCRS mitochondrial sequence, Human herpesvirus 4 type 1 and the concatenated decoy sequences. As reported by Li *et al.*[38], the integration in standard pipelines of decoy sequences allows the resolution of false heterozygous calls. In fact, the majority of the 54 unbalanced variants in *MAP2K3* were not present in 1000 Genomes Project Phase3 release, while those reported in the database had a balanced heterozygosity. As a result, none of *MAP2K3* variants were unbalanced in 1000 Genomes Project Phase3 release.

Then I focused on 8 enriched regions of *MAP2K3* carrying more than one unbalanced variant: the first three regions match exons 3, 4 and 5, while the last five regions match exons 9, 10, 11, 12 (see Table 7). Reads from these regions were realigned against the GRCh37 and GRCh38 full assemblies.

Results in Table 7 show that exons 3, 4 and 5 behaved very differently from exons 9, 10, 11 and 12. In the former, reads aligned only to the target gene independently of the reference used. In the latter, the amelioration of the reference genome led to different results: using the GRCh37 full assembly reads carrying all variants aligned to a fix patch, called HG987_PATCH (NCBI Reference Sequence: NW_003315950.2), while using the GRCh38 full assembly they aligned to a new region added in the chromosome 17; on the other side, reads with none variant aligned to the target gene using both the references. These results indicate that a sequence very similar to the last portion of *MAP2K3* was included in the HG987_PATCH added in the GRCh37 full assembly. This sequence is indeed absent in the GRCh37 primary assembly. This patch was then inserted in the GRCh38 release and its coordinates correspond to the new region in the chromosome 17 where reads aligned. Therefore I can conclude that *MAP2K3* has been partially amended in GRCh38, with the insertion of a duplication of the last part of the gene, including exons 9 to 12, while it remains with its original unbalanced heterozygosity at the beginning of the gene, as seen in exons 3 to 5. In fact, a BLAST search confirmed that the duplication spans only exons from 8 to 12. However, results suggested the presence of a duplication also for exons 3, 4 and 5. In fact, I saw two groups of reads, one carrying all the variants and the other any of them, but they both aligned only to the target gene.

Importantly, of the 51 unbalanced variants localized in the 8 analyzed exons of *MAP2K3*, 26 variants were confirmed using GRCh38 for variant calling. Of these, 25 were localized in exons 3, 4 and 5 and only one variant was localized in exon 12 - it should be pointed out that this variant was identified in only one sample using GRCh38, thus indicating a private variant. This is a confirmation that the last portion of *MAP2K3* lost his unbalanced heterozygosity in GRCh38.

target region name	exon	reads with all variants		reads with none variant		GRCh38 confirmed variants
		GRCh37	GRCh38	GRCh37	GRCh38	
MAP2K3_158294.12020	3	gene	gene	gene	gene	9
MAP2K3_158295.17245	4	gene	gene	gene	gene	6
MAP2K3_158296.5164	5	gene	gene	gene	gene	10
MAP2K3_158296.5164	9	patch	new region	gene	gene	0
MAP2K3_158296.5164	10	patch	new region	gene	gene	0
MAP2K3_158296.5164	10	patch	new region	gene	gene	0
MAP2K3_158296.5164	11	patch	new region	gene	gene	0
MAP2K3_158296.5164	12	patch	new region	gene	gene	1

Table 7. BLAST results of read realignments. The results for reads with all variants or none of them are reported separately: for each group the alignments on both the references used (GRCh37 and GRCh38 full assemblies) are shown; 'gene' corresponds to the target gene, 'new region' stands for a different region within the same chromosome and 'patch' refers to HG987 patch. The number of variants confirmed in the last release of the human reference genome is also reported.

4.9 Analysis of the physical coverage in mate pair whole genome data

To further confirm that the *MAP2K3* genomic regions hid a duplication not reported in GRCh37, I analyzed genome sequencing mate pair data downloaded from the Genome In A Bottle project[94]. Reads from parents of an Ashkenazi Trio and a Chinese trio were aligned against the GRCh37 and GRCh38 primary assemblies. Figure 9 shows the physical coverage of a 570 kb region of chromosome 17, containing *MAP2K3* and *KCNJ2* (Potassium Voltage-Gated Channel, J2), another gene classified as partially duplicated (Table 5, column 2). In agreement with our findings, the coverage in these two genes was at least doubled than the average when mapped on GRCh37. Using GRCh38, the physical coverage of *MAP2K3* and *KCNJ12* genes was still higher than the flanking regions, but to a lesser extent than what was observed with GRCh37. In the box of Figure 9 the region of *MAP2K3* is enlarged to show in more detail the reduction of physical coverage starting from exon 9 which actually is the last part of the gene, known to be duplicated in GRCh38.

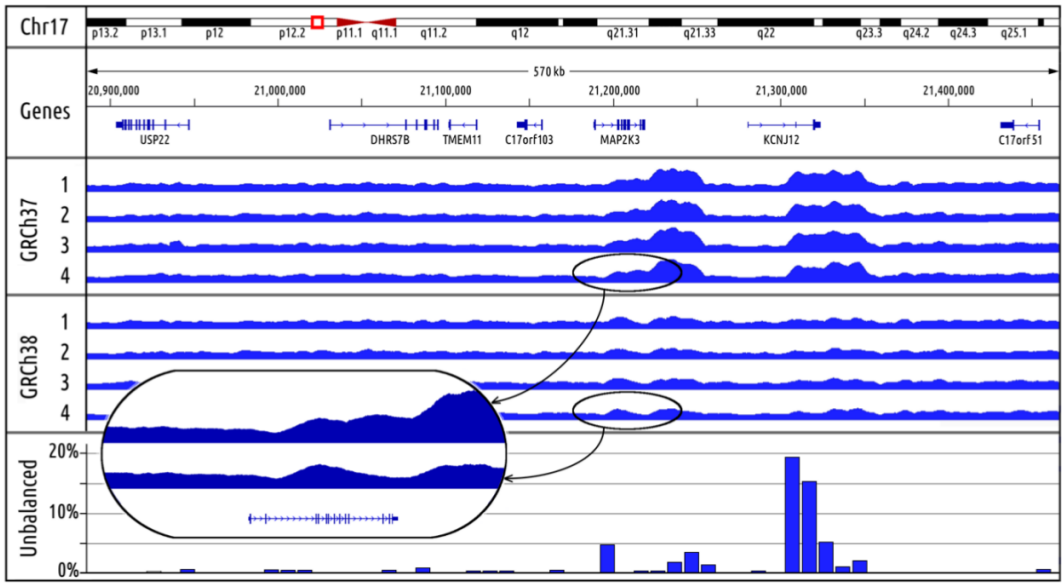

Figure 9. Physical coverage profiles of a 570 kb region of chromosome 17. The set of mate pair from the Genome In A Bottle project was aligned on the GRCh37 and GRCh38 primary assemblies; numbers 1, 2 refer to two Ashkenazi individuals, whereas 3 and 4 refer to two Chinese individuals. The frame at the bottom shows the percentage of variants with unbalanced heterozygosity based on the total number of biallelic variants in non-overlapping 10 kb sized windows. In the box the region of *MAP2K3* is enlarged to show the reduction of physical coverage starting from exon 9 in GRCh38.

4.10 Recap of recurrent exome variants analyses

This paragraph aims to summarize the results obtained with the analyses of the 9,313 variants that I unexpectedly found in more than 90% of the Ion Proton samples. First of all, I compared the Ion Proton dataset with independent Illumina and SOLiD datasets. In addition, I identified variants falling in MAiR positions. Finally, I performed a statistical test on allele and genotype frequencies to recognize 'unbalanced' regions and I re-mapped reads on different reference assemblies. Through this wide set of analyses, I demonstrated that: i) 8,680 variants of the 9,313 fell in MAiR positions, meaning that the reference does not carry the most frequent allele in the population; ii) 316 were possible indicators of gene or region duplications; iii) 82 were both MAiRs and with an unbalanced heterozygous genotype, thus involving the issues of points i and ii; iv) 16, among which 1 was also MAiR, could be Ion Proton specific errors as they were absent in Illumina and SOLiD samples, v) only 219 stand without a clear explanation. Among the latter, 41 variants have never been previously reported, while for the 178 remaining I hypothesized they might be population specific polymorphisms.

In fact, the frequencies of these 178 variants are significantly higher for the Europeans compared to the total population, as reported in Figure 10.

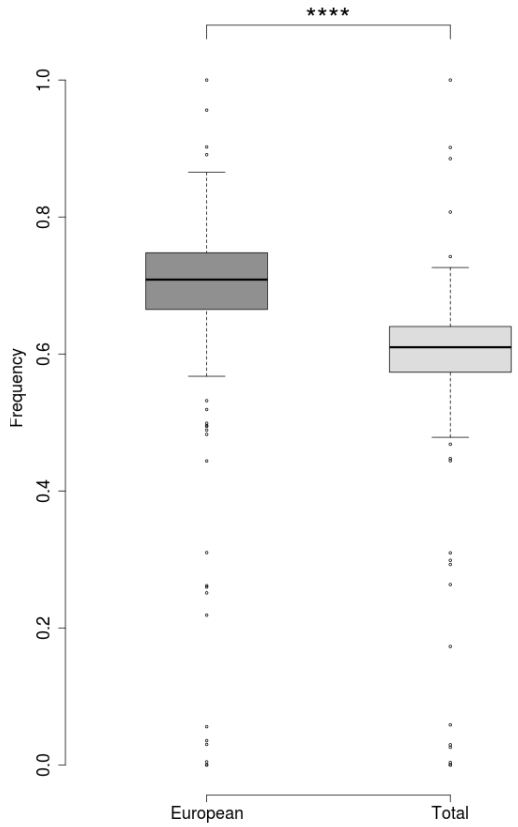

Figure 10. European and Total frequencies of the 178 possible population specific polymorphisms. The difference between allele frequencies in the two analyzed populations is highly significant (*p-value*<0.0001).

4.11 From exomes to genomes

Results described so far highlighted two important aspects: i) all resequencing analyses should take into account that false positive variants could originate from the reference used; ii) although the improvements of GRCh38, some reference driven problems are still detectable. Exome data analyses allowed me to accurately identify some of these problems, for example unreported gene duplications and genomic positions that do not represent the most frequent alleles.

Being aware of the importance of these findings, I decided to move my analyses towards a new and wider direction, the analysis of whole genome sequencing (WGS) data. As discussed in the Introduction, thanks to the reduction of DNA sequencing cost, together with several advantages of sequencing the entire genome

- uniform coverage, detection of non-coding variants and copy number variations -, WGS is becoming the leading strategy routinely used not only in the research field but also in the clinical one[7]. As a result, the non-coding regions are increasingly analyzed to understand their functional roles and to discover non-coding variants involved in determining human traits and complex diseases. From this perspective, it is now essential to have a reference genome that accurately represents the entire human DNA sequence.

4.12 Minor Alleles in GRCh37 and GRCh38 reference genomes

In 2017 the Genome Reference Consortium (GRC) published a paper to describe the assembly updates in GRCh38, including the correction of 8,248 erroneous bases[1]. The graph in Figure 11 shows the alternative allele frequency of these sites in GRCh37: for the large majority of the 8,248 positions the reported alternative allele frequency was higher than 90%, indicating that the corresponding reference allele was very rare in the population. In these MAiR positions the reference allele was replaced with the most common allele in the population in GRCh38.

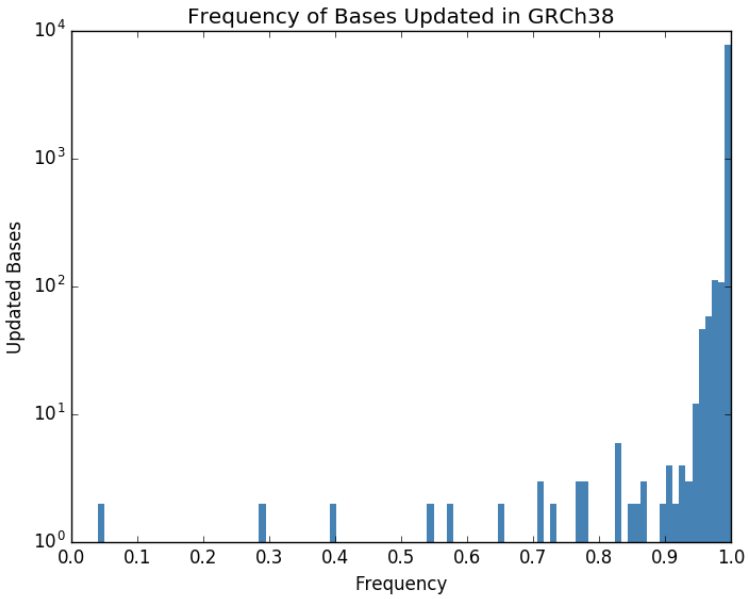

Figure 11: Frequency of bases updated in GRCh38 by the GRC. The authors identified 8,248 erroneous positions in GRCh37 and corrected them in GRCh38[1]. The histogram shows the alternative allele frequency of these positions in GRCh37. The large majority of corrected positions have a frequency higher than 0.9, thus confirming the authors' choice to report in the reference genome the most common allele in the population. Frequencies derive from 1000 Genomes Project Phase1 data. Note the logarithmic scale on the y-axis.

Since the analysis of the Ion Proton exome dataset revealed that the number of variants mapped on MAiRs in GRCh37 was 18,839, I expected that much more than 8,248 genomic positions required a revision by the GRC. To better clarify how many minor alleles are present in the reference genome, I screened the Phase3 VCF files of the 1000 Genomes Project, consisting of 2,504 whole genome sequences from 26 populations, aligned on GRCh37[112]. The results of this analysis are shown in Figure 12A.

A total number of 84,801,880 variants were present in the GRCh37 Phase3 VCF file. I found 436,700 (about 145 variants per Mbp) with an allelic frequency equal or greater than 90% and 32,105 variants (about 11 variants per Mbp) with 100% allelic frequency, in homozygosity in all the individuals. These findings showed that the surprisingly high number of HF variants observed in exomes was also confirmed in whole genomes.

Recently, the European Bioinformatics Institute has re-aligned the 1000 Genomes Project sequencing data on the GRCh38 reference genome[58]. A total number of 82,218,941 variants were present in the GRCh38 Phase3 VCF file, confirming the previously discussed results of Guo and colleagues[37].

More importantly, I were interested in evaluating the number of HF variants found in GRCh37 that have been amended in GRCh38. 2,198,258 HF variant positions were present in the GRCh37 Phase3 VCF file. For each position I checked if in the GRCh38 release the reference allele was substituted with the most common allele in the population or if the position was not included in GRCh38. I found that this correction occurred only in 70,497 cases, while 2,127,761 positions maintained the less frequent allele in the population.

Figure 12B shows these results in detail. It can be seen that the rate of correction is practically the same, around 3%, in the range between 0% and 95%. Some improvement is seen for the alleles with a frequency between 95%-100%, that were corrected in 5.6% of the cases. Nevertheless, the large majority of these loci still maintain the minor allele in the GRCh38 reference genome, including 8,593 alleles with a frequency of 100%. These findings clearly suggested that a further deep revision of base pair level errors is necessary to make the reference genome the accurate representation of the most common DNA sequence in the population.

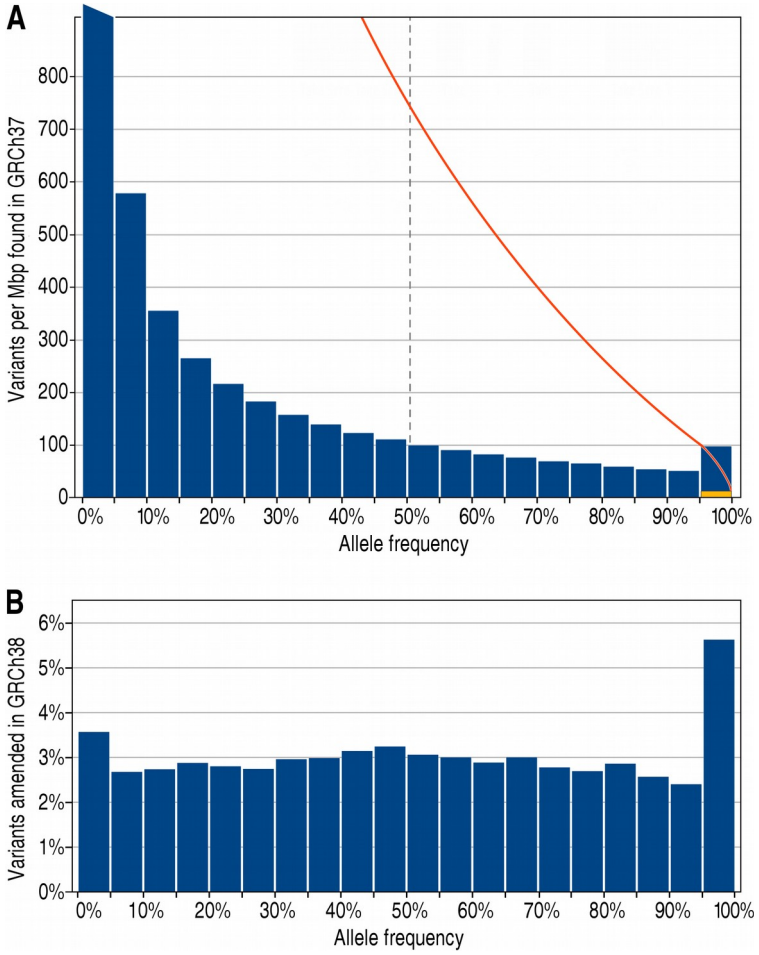

Figure 12. Allelic frequencies of variants found in GRCh37 and amended in GRCh38. The variants of 2,504 genomes (1000 Genomes Project, Phase3) were divided into classes according to their allelic frequency. Frame A: the blue blocks indicate the average number of variants per Mbp of each class. Note that the first bar is outside the range of the Y-axis. The red line indicates the sum of values from a given allele frequency to the right end, that is the number of variants with at least the indicated allele frequency. It can be seen that there are about 730 variants/Mbp with an allele frequency >50%. The yellow sector at the bottom of the 95-100% block corresponds to variants found in homozygosity in 100% of the individuals (about 11 variants / Mbp). Frame B shows the percentage of variants that have been amended in the GRCh38 release.

4.13 Genome regions with unbalanced heterozygosity

In addition to base pair level errors, I detected several exomic regions hiding duplications not reported in GRCh37. In search of all the possible unreported duplicated regions in the entire genome, I performed the previously described statistical test for the unbalanced heterozygosity on the 1000 Genomes Project Phase3 data. For each chromosome I considered non-overlapping 100 kb sized windows and for each window I calculated the percentage of unbalanced variants

based on the total number of biallelic variants. For each variant I considered its allele frequency in the whole 1000 Genomes Project population and the observed frequency of heterozygous genotypes. As previously described, I performed a one-tailed binomial test, corrected for false discovery, to identify all deviations from the Hardy-Weinberg equation. Variants were considered significantly unbalanced if their corrected *p-values* were lower than 0.01. The results of this analysis are shown in Figure 13. Many regions with unbalanced heterozygosity can be clearly detected. All these unbalanced regions might conceal a duplication.

As shown in Table 5, column 4 some genes with unbalanced heterozygosity in the GRCh37 lost their imbalance thanks to the introduction of alternate loci in the GRCh38 full assembly. However, as stated above, highly polymorphic alleles should not produce unbalanced heterozygosity when aligned on the reference genome. This is proved by the portion of chromosome 6 corresponding to the Major Histocompatibility Complex (MHC), indicated by an asterisk in Figure 13: although it is possibly the most polymorphic region of the genome, it is not particularly associated to unbalanced heterozygosity, suggesting that the nature of this genetic inconsistency should be found elsewhere.

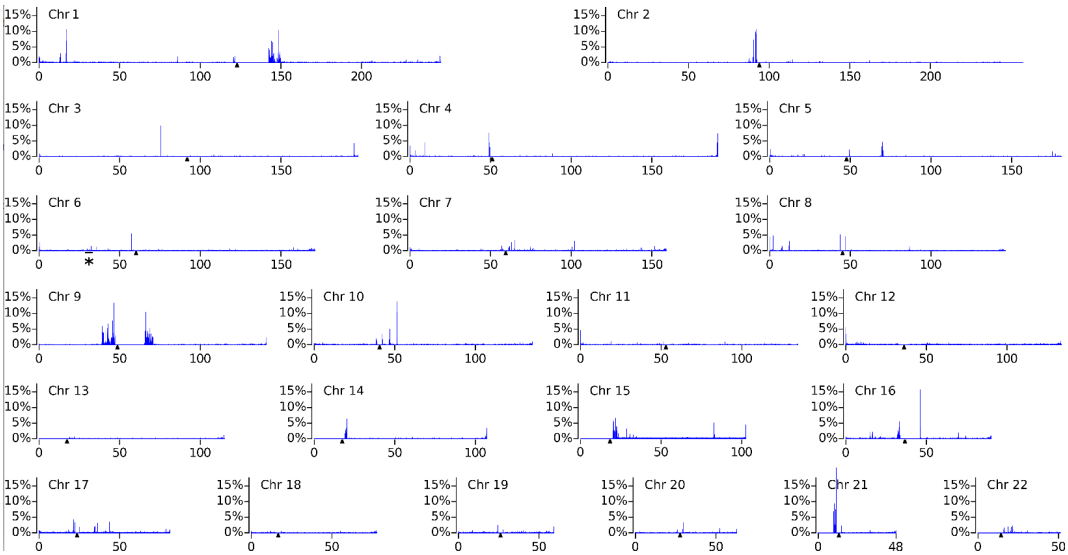

Figura 13. Genome wide analysis of regions with unbalanced heterozygosity in GRCh37. For each non-overlapping 100 kb window I considered the percentage of biallelic variants with a significant unbalanced heterozygosity. Centromeres are indicated by a small triangle below the baseline. The region marked by the asterisk in chromosome 6 indicates the MHC.

I detected a total of 86,649 unbalanced variants that I believe to be mostly due to unreported genomic duplications. An interesting issue is to understand how many HF variants are produced by these putative unreported genomic duplications in each individual. To answer this question I analyzed the VCF file of the 2,504 individuals belonging to the 1000 Genomes Project and I found that on average each individual carries 4,491,760 variants of which 1,950,334 are HF; I also found that on average each individual carries 62,308 unbalanced variants; finally I found that in each individual, on average 24,768 HF variants are unbalanced. The diagram in Figure 14 confirm that HF variants are not only due to MAiRs, but also to regions with unbalanced heterozygosity, possibly derived from unreported genomic duplications.

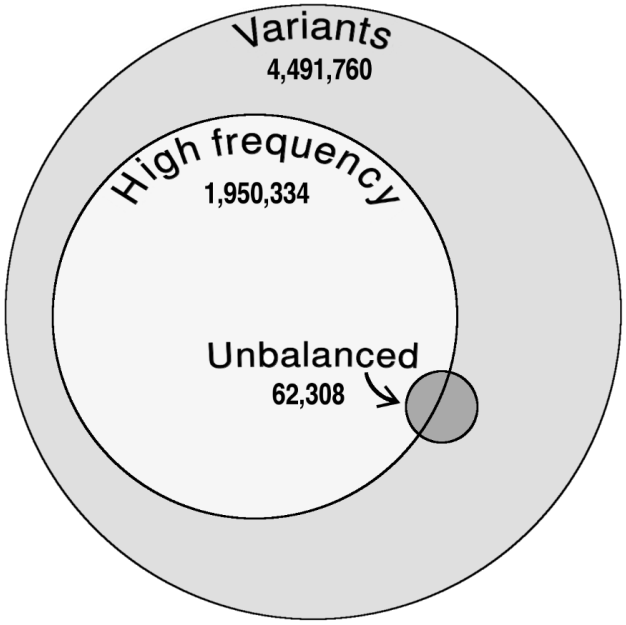

Figure 14. Variants per individual in the 1000 Genomes Project. Average number of variants per individual found in the population of 2,504 people studied in the 1000 Genomes Project. Variants have been further subdivided in High Frequency and Unbalanced variants.

4.14 Variants distribution in exomes and genomes

The results obtained from exomes and whole genomes are slightly different in terms of the number of variants per Mbp. Typically, in the genome of a single individual about 4.5 million variants are detected (Figure 14), equivalent to ~1500

variants/Mbp, whereas in exomes the average number of variants is 48,785, over a target region of 57 Mbp, equivalent to only 856 variants/Mbp. This is not surprising as it is known that protein coding sequences tend to be more conserved than other genomic regions. Similarly, the reference genome seems to be slightly more accurate in coding regions. This can be reckoned by considering the number of minor alleles in the reference genome. In this respect, in whole genomes I found ~145 variants per Mbp with an allelic frequency greater than 90%, whereas in the exomes I found only 68 such variants per Mbp. Surprisingly, the number of variants with 100% allelic frequency (meaning that the allele reported in the genome was never found in my analyses) is slightly greater in exomes (15/Mbp) than in whole genomes (11/Mbp). Unfortunately, I found that only a small percentage of these genomic positions, about 3%, have been corrected in the GRCh38 release, rising to ~5% for the variants found with a frequency above 95% (see Figure 12).

4.15 Brief summary of main results

In this work I presented a comprehensive study of variants found in two datasets, the former composed by 222 exomes sequenced at the CRIBI center and the latter by 2,504 genomes included in the 1000 Genomes Project. I focused my analyses on the possible explanations for the presence of anomalous variants both in exomes and genomes and on the strategies adopted to individuate, characterize and filter them. These strategies allowed me to discriminate variants associated to reference genome errors, including uncorrected bases and misassemblies.

I firstly analyzed all the samples with two different releases of the human reference genome, GRCh37 and GRCh38. I saw, as already reported in literature[37], that the number of variants identified using the latest reference was reduced, dropping from 264,303 to 255,124 for the exomes and from 84,801,88 to 82,218,941 for the genomes, using respectively GRCh37 and GRCh38.

Nevertheless, despite the upgrade in the reference, many positions still carry the minor allele, instead of the major one as it should be expected. I found that the 90.40% of variants mapped in MAiR exome positions in GRCh37 were kept in GRCh38 and that this percentage increased to 96,80% in whole genome data.

These results indicated that, although more than 8,000 bases have been corrected in the last release of human genome[1], others efforts are necessary to further reduce the base-pair-level errors.

On the other hand, to investigate the presence of misassemblies, in particular gene duplications, I selected variants with an unbalanced heterozygous genotype. This characteristic is consistent with the hypothesis that some genomic regions are still not reported as duplicated. With this analysis I identified 45 genes that could have undergone to gene duplication events not reported in GRCh37. As Schneider *et al.* in 2017 claimed that GRCh38 provides reference assembly representation for previously missing human-specific and paralogous sequences[1], I checked whether the new regions added to GRCh38 contained the sequences of the 45 suspicious genes marked as duplicated in GRCh37. Only 15 genes turned out to be 'fully amended' in GRCh38 (Table 5, column 1), while the remaining 30 genes are still unresolved in GRCh38. Moreover, the analysis of unbalanced heterozygosity on whole genome data revealed that unreported duplications might concern numerous and wide regions of the genome.

Chapter 5

Conclusions

Whole exome sequencing and whole genome sequencing are powerful tools for analyzing human genetic variation and rare hereditary diseases. Nevertheless, the big amount of data obtained in exome and genome sequencing projects may be difficult to handle and researchers and geneticists can fall in misleading interpretations of the results. In order to reduce errors, the performed analyses must be as reliable as possible.

A crucial role in determining the accuracy of exome and genome sequencing analyses is played by the human reference genome. The reference assembly affects the read alignment process and the variant calling step, as well as it serves as the foundation for variants annotation. Ideally, the human reference genome should be representative of the total sequence variations and, as a result, it is a very dynamic resource: as our comprehension of the global human diversity evolves, the reference assembly also evolves and new findings lead to the need of continuous ameliorations.

GRCh38 represents the current and most updated version of the reference genome. However, it is still scarcely used in exome and genome studies[1]. My work confirmed that GRCh38 is more complete and accurate than GRCh37, even if my results indicated that some inconsistencies are still there. Many positions of the reference genome should be amended to avoid the call of high frequency (HF) variants in exome and genome sequencing analyses. Most of these HF variants map on the minor alleles in the reference (MAiRs) and some map on misassembled regions with unbalanced heterozygosity.

In this thesis different methods to identify these inexact positions have been presented. In particular, the described statistical test on the heterozygous genotypes has proved to be powerful in unveiling all regions with unbalanced heterozigosity; importantly, duplications for some of the resulting regions have been confirmed by read realignment and coverage analysis of mate pair whole genome data. These important results should draw attention of the Genome Reference Consortium and

encourage a further revision of repetitive and segmentally duplicated regions. At present the assembly of these complex regions is not a trivial task. Hopefully, it might be facilitated by technological advancements leading to an improved assembly contiguity, for example a further increase in read length or the use of longer library inserts, as stated by the Genome Reference Consortium[1].

In addition, a more deep integration of information from different databases could be useful both in defining the most common alleles in the global population – and therefore the consensus sequence of the human reference genome – and in increasing our knowledge on human diversity. As discussed in the Introduction, the number of entirely sequenced genomes is strongly growing and data from them should serve as additional information useful to update the current reference genome. In this context, an important resource is now represented by the Genome Aggregation Database (gnomAD)[86], developed for aggregating data from 123,136 exomes and 15,496 whole genomes belonging to unrelated individuals sequenced as part of various disease-specific and population genetic studies. The number of sequenced individuals, extremely higher than any other exome and genome database, makes gnomAD a very precious resource in collecting the global human variation.

The inclusion of the entire known human variation in the reference genome, although desired, poses some practical problems. According to the assembly model developed by the Genome Reference Consortium, the additional variation is included in the full assembly in the form of patches and alternate loci. In my work I considered GRCh37 and GRCh38 both as primary and full assembly versions. The full assembly references with patches and alternate loci could be useful to improve the accuracy in read mapping and, therefore, to reduce the number of false positive calls. However, their huge size (31 and 46 Gb for GRCh37 and GRCh38, respectively) and the lack of suitable bioinformatic tools make their use very impracticable for most practical applications.

For this reason, it is now spreading the idea that a comprehensive graph-based representation of genome-wide population variation would be more appropriate than a single reference genome[68]. Although this means to revolutionize the model that supports the infrastructure and the tools used in the sequencing data analysis,

this effort would allow to develop a more robust analysis framework.

Even if the graph-based assembly will be successful, it will be a long time before the new model enters in the common practice. GRCh38 will continue to be the most comprehensive and highest quality representation of the human genome usable in resequencing analyses, more and more demanded in the clinical field. As a result, its correctness and completeness are fundamental.

Chapter 6

Future perspectives: variant prioritization in lncRNAs

As stated at the very beginning of this thesis, I spent part of my PhD period addressing the problem of interpreting the functional effect of nucleotide variants in non-coding regions of the human genome. This part of the project is here described as future perspectives for two reasons. Firstly, although preliminary results seem to be very promising, this study requires further work. Secondly, this project deals with the problem of understanding the meaning of DNA alterations far beyond the limited portion of protein coding genes and this actually represents one of the major future challenges in the human genetics field.

It has become largely accepted that the non-coding portion of the human genome accounts for the regulation of gene expression, a complex process involving many different factors and levels of control[113]. Data from genome-wide association studies (GWAS) suggested that more than 90% of disease-associated SNPs are located in functional non-coding regions of the human genome, for example in promoter regions, enhancers elements or in non-coding RNA genes[114]. These data indicate that many disease-causing variants are likely to exert their effect by altering the regulation of genes rather than by directly affecting genes and protein functions.

It is thus evident that the analysis of the human genome, both to better understand the mechanisms behind gene regulation and to discover new genetic alterations driving diseases development, can no longer afford to neglect variants in non-coding regions. In this context, the usage of Whole Genome Sequencing (WGS) approaches has rapidly grown over the last few years, also thanks to the remarkable reduction in DNA sequencing costs (see the Introduction, Figure 2).

However, the prioritization and the functional interpretation of non-coding variants are still challenging. While non-synonymous variants and their effects on protein functions can be predicted by computational methods based on protein sequence homology and physical properties of amino acids, such an approach cannot be applied to non-coding variants. Alternative types of computational methods that

use various genomic and epigenomic annotations have been developed to allow their prioritization. Examples of these tools are Genome-Wide Annotation of VAriants (GWAVA)[35], Combined Annotation-Dependent Depletion (CADD)[36] and FunSeq2[115], specifically developed for prioritizing non-coding regulatory variants in cancer. All these approaches integrate a wide range of variant-specific annotations of different classes: conservation metrics, regulatory sites information (for example DNase hypersensitivity sites and transcription factor binding sites) or transcript information (for example distance to exon-intron boundaries and expression levels in cell lines). A more recent model for the prediction of functional effects of non-coding variants is DeepSEA[116], a deep learning-based algorithm requiring large-scale chromatin-profiling data to train the model. All these methods are therefore based on a wide *a priori* knowledge derived from several sources.

Here I describe the first steps towards the development of a new method for prioritizing non-coding variants based on an alternative strategy: the comparative genomics approach. The idea is that nucleotide variants located in conserved domains are more likely involved in disrupting the functional role of non-coding elements. It was reported that conserved domains are generally involved in determining the secondary structures of non-coding RNAs as well as in interacting with targets, for example mRNAs or DNA double helix, or in the splicing process[34]. In my method, these conserved functional domains are searched by comparing orthologous sequences found in phylogenetically related organisms and looking for regions that are conserved across species.

Given the extent and the heterogeneity of the non-coding portion of the human genome, I decided to focus first on a single class of non-coding elements, the long non-coding RNAs. Among non-coding elements, long non-coding RNAs are emerging as central players in cell biology, but these functional components of the human genome are still largely unexplored.

6.1 Brief introduction on long non-coding RNAs

Long non-coding RNAs (lncRNAs) are a heterogeneous class defined as transcripts more than 200 nucleotides in length with absent or low protein coding

ability. They are also called long intergenic non-coding RNAs (lincRNAs) when they do not overlap with any protein coding transcription unit.

With the advancement in DNA sequencing techniques, thousands of lncRNAs have been identified in the human genome - the estimated number is over 100,000[117]. To date, only a limited number of human lncRNAs has been functionally characterized. Among them, *H19* and *Xist* (X-inactive specific transcript) were discovered in the early 1990s[118,119]. Other two well studied lncRNAs, *HOTAIR* (HOX antisense intergenic RNA) and *HOTTIP* (HOXA transcript at the distal tip), were described only several years later[120,121]. However, the number of characterized lncRNAs is expected to grow very quickly.

Although the detailed mechanism of action is known only for a few dozen of annotated lncRNAs, the available examples show the complexity of their biology: they act as crucial regulators in many different cellular processes by interacting with DNA, RNA and proteins; they are involved in post-transcriptional gene regulation by controlling protein synthesis, RNA maturation and RNA transport; they have been also implicated in transcriptional gene silencing via epigenetic regulation and chromatin remodeling[117].

Given their role in so many different processes, lncRNAs are involved in the etiopathology of numerous human disorders, including hepatocellular carcinoma, Alzheimer's disease and diabetes[122]. Two types of alterations can affect lncRNAs function and drive diseases development: large chromosomal rearrangements (translocations, amplifications or deletions) and small mutations (small insertions/deletions or SNPs). While larger alterations usually alter the expression of lncRNAs, understanding how small mutations are involved in disease etiopathology can be more challenging. Also small alterations can affect the expression level of lncRNAs, for example if they are located in promoter sequences. Moreover, they may have other consequences on the alternative splicing process of the transcript or on the secondary structure determination.

To date, more than 7 millions SNPs in human lncRNAs have been identified and some of them have been described[123]. For example, it was reported that the expression of the previously mentioned *HOTAIR* lncRNA, an oncogene involved in gastric cancer development, is altered by SNP rs920778, contributing to

increase cancer susceptibility[124]. On the contrary, SNP rs2839698 in *H19* gene is associated with a significantly decreased risk of bladder cancer[125]. Nevertheless, a very high number of variants in lncRNAs has not yet been investigated and their prioritization represents the fundamental first step to face this task.

6.2 Identification of conserved domains in lncRNAs

As mentioned above, my method for prioritizing non-coding variants in lncRNAs is based on the identification of conserved domains by using a comparative genomics approach. The pipeline structure, described in detail in Material and Methods chapter, consists of three different steps. The first one identifies the orthologous genes of the human lncRNA in the genomes of 28 primates by performing two BLAST alignments in search of the 'reciprocal best hits' (RBHs). In the second step, the orthologous sequences are aligned with a multiple sequence alignment tool called T-Coffee[100]. Finally, the conserved domains in the human lncRNA are identified with Gblocks[105].

6.3 Pipeline validation by comparison with published data

The UCSC Genome Browser website provides WIG* format files containing a 'conservation score' for each base of the human reference genome[126]. These data derived from the multiple alignment of the human reference genome with 32 placental mammal genomes; the multiple alignment is then used to compute base-by-base conservation scores. These scores were calculated with phastCons, a program able to identify conserved elements in the genome of different organisms using genome-wide multiple alignments[127].

I used these available data to assess the accuracy of my pipeline. In particular, I checked whether the conserved blocks identified with my pipeline correspond to the positions with the highest conservation scores in the WIG files.

Compared to the published data based on the alignment of the whole human

*WIG (wiggle) format: The WIG format is designed for display of dense continuous data such as probability scores. A WIG file consists of one of more blocks, each containing a declaration line followed by lines defining data elements. There are two main formatting options: fixedStep and variableStep. VariableStep format is designed for data with irregular intervals between data points and is the more commonly used format. It begins with a declaration line, followed by two columns containing chromosome positions and data values. FixedStep format is designed for data with regular intervals between data points and is the more compact of the two wiggle formats. It begins with a declaration line, followed by a single column of data values.

genome with other 32 genomes, my pipeline aligns only a specific region of the human genome, the lncRNA of interest, with the other genomes; this should guarantee a more accurate alignment of the region of interest. A second advantage of my pipeline is that any organism can be included in the analysis.

6.4 *LINCMD1* as positive control

As initial test to assess the accuracy of my pipeline, I chose a well studied human lncRNA, called Long Intergenic Non-Protein Coding RNA Muscle Differentiation 1 (*LINCMD1*, NCBI Gene ID: 101154644, Ensembl Gene ID: ENSG00000225613). *LINCMD1* is a long non-coding cytoplasmic RNA expressed during myoblast differentiation.

The gene is localized on the reverse strand of chromosome 6 p-arm and it is structured in three exons and two introns, for a total length of 4,306 bases. The first intron of *LINCMD1* hosts the *MIR133B* sequence (NCBI Gene ID: 442890, Ensembl Gene ID: ENSG00000199080).

The functional role of *LINCMD1* in mouse has been extensively described in two different works[128,129]. In mouse *LINCMD1* has two different binding sites for *MIR135* and one single binding site for *MIR133*[128]. It was found that *LINCMD1* acts in the cytoplasm as a competing endogenous RNA (ceRNA) for *MIR133* and *MIR135*, thus limiting their binding to their natural mRNA targets: the Myocyte-specific enhancer factor 2C (*MEF2C*) targeted by *MIR135* and Mastermind-like-1 (*MAML1*) controlled by *MIR133*[128]. Both these proteins have a relevant function in myogenesis. MEF2C protein belongs to a family of transcription factors that activate the expression of numerous muscle-specific genes[130]; moreover, it was shown to play a key role in differentiation of muscle cells[131]. *MAML1* encodes critical transcriptional coactivators for Notch signaling, that have documented roles in myogenesis[132]. Even more importantly, a crosstalk between *MAML1* and *MEF2C* have been described in muscle cells[133] and their expression is regulated by *MIR133* and *MIR135*[128,129]. According to the role of *LINCMD1* as decoy for *MIR133* and *MIR135*, Cesana *et al.* observed that a *LINCMD1* depletion decreases the levels of both MAML1 and MEF2C proteins, while a *LINCMD1* over-expression leads to their accumulation[128].

LINCMD1 sequence was found to be conserved in human myoblasts, in particular around the recognition motifs for *MIR135* and *MIR133*; also its function is maintaned in human muscle cells[128]. For this reason, I chose this lncRNA to test my pipeline: I expected to find the miRNA binding sites within the conserved blocks identified with my pipeline.

6.4.1 Results for LINCMD1

The first step of my pipeline, which aims to identify orthologous genes of a human lncRNA in the primate genomes, allows to find the orthologous sequences of *LINCMD1* in all the 28 organisms. I found that, compared to the human gene (4,306 bases long), the sequences of 7 primates are significantly shorter: 1,169 for *Carlito syrichta*, 1,137 for *Daubentonia madagascariensis*, 1,324 for *Eulemur flavifrons* and *Eulemur macaco*, 1,258 for *Microcebus murinus*, 732 for *Otolemur garnettii* and 1,326 for *Propithecus coquereli*. This could be explained by the lack of *LINCMD1* gene conservation in these species, but also by the possible incompleteness of genome sequences for these organisms. In fact, except for *Microcebus murinus*, genome sequences of these 7 primates are not assembled in chromosomes, but published as scaffolds.

Since the multiple alignment of sequences with very different lengths may be problematic, I performed the T-Coffee[100] alignment both including and excluding the shorter orthologous sequences. Even if in some regions the alignment of shorter sequences was very fragmented, with stretches of few nucleotides aligned far a part, I observed that a well aligned block of at least 500 nucleotides was present for all orthologous sequences. Thus, I chose to continue the analysis including shorter sequences.

The resulting alignment file in FASTA format (.fasta_aln) was used as input file for Gblocks[105], with the minimum number of sequences for a conserved position (*b1*) equal to 15 (default parameter, 50% of the number of sequences + 1), the minimum number of sequences for a flank position (*b2*) equal to 22 (75% of the number of sequences) and the minimum length of a block (*b4*) equal to 3. In fact, also very short sequences can be crucial in determining secondary structures or in binding other molecules. A total number of 20 blocks was found with these parameters.

The most interesting result is the extremely high conservation of 374 bases of the human LINCMD1 intron 1, from position 52,148,824 to position 52,149,197 of chromosome 6 (genomic coordinates refer to GRCh38). This region hosts the *MIR133B* sequence, from position 52,148,923 to position 52,149,041. Interestingly, the WIG file, derived from the multiple alignment of the human reference genome with 32 placental mammal genomes[126], reports the highest conservation scores for the genomic region extending from position 52,148,902 to position 52,149,050. This region is included in the most conserved domain identified with my pipeline, which, however, is longer on both sides (see Figure 15, where the dark blue profile represents the conservation score for each base of *LINCMD1* and the two light blue rectangles represent two conserved blocks identified with my pipeline).

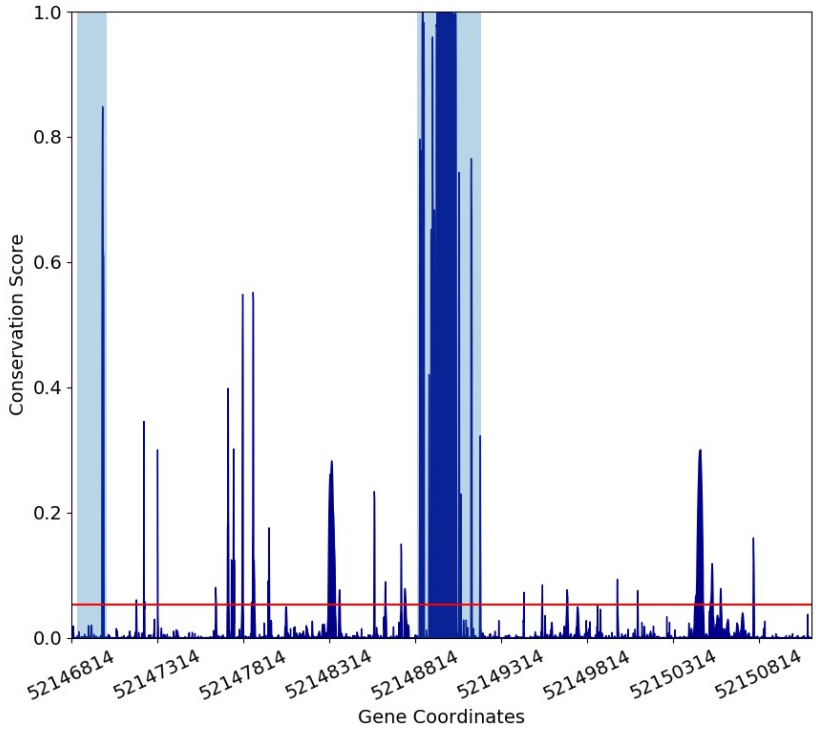

Figure 15. Conserved domains of *LINCMD1*. The dark blue profile represents the conservation scores reported in the WIG file available at the UCSC Genome Browser website[126] and derived from the multiple alignment of the human reference genome with 32 placental mammal genomes. Conservation scores go from 0 (no conservation) to 1.00 (maximum conservation). The two light blue rectangles represent two conserved blocks identified with my method. The red line represents the average conservation score of the entire gene. According to the WIG file, the central part of *LINCMD1* is the most conserved, even if also a very short conserved sequence is present in the first portion. Both these regions are included in wider conserved domains identified with Gblocks.

The WIG file reported also a short but well conserved sequence of 12 bases (GGGAGGACATGT, from position 52,146,989 to position 52,147,000), represented by the first dark blue pick in Figure 15. Once again, this region is included in a wider conserved block identified with my pipeline.

These results seem to indicate that my pipeline is less accurate in identifying short conserved regions. It should be considered that my method compares orthologous sequences belonging to primates, while data in the WIG file derive from the alignment of the human reference genome with 32 placental mammal genomes. As a consequence, considering phylogenetically closer organisms, resulting conserved regions are wider.

6.5 Future improvements of the pipeline

In light of the above, some improvements could make my pipeline more accurate in identifying conserved domains of lncRNAs. Firstly, by gradually including less related organisms, the degree of conservation could be evaluated depending on the phylogenetic distance. The identification of few nucleotides highly conserved in very different species would be more informative than the identification of long sequences conserved only in close organisms.

Secondly, I should select only organisms with as much as possible complete genomes, assembled in chromosomes and not only released as scaffolds. This would make possible to know if shorter orthologous sequences are due to the lack of conservation along the entire gene or to an incomplete gene sequence in the selected organism. Among available genomes, only 12 of the 28 primate genomes[98] and only 37 of the 178 mammal genomes[134] are assembled in chromosomes (data referred to April 2018). Thanks to advancement in DNA sequencing techniques and genome assembly approaches, these numbers are expected to grow and genomes are expected to be more complete in the near future.

Once the process of identification of conserved domains will be improved, the following steps will consist in integrating information from other sources. For example *LINCMD1* is known to interact with two miRNAs, *MIR135* and *MIR133*[128]. Different programs allow to predict miRNAs binding sites on target genes. For example the StarMir[135] tool provides, for each position of the lncRNA,

the probability that the position is involved in miRNAs binding. This will be particularly useful to check if positions with the highest probability to bind miRNAs correspond to the most conserved regions. In these positions, nucleotide variants may affect the ability to bind miRNAs, thus compromising the *LINCMD1* function.

Although the approach presented in this last chapter is still under development and results are very preliminary, some general considerations can be done. The comprehension of the whole human genome and, in particular, of the still unexplored non-coding portions, depends also on the availability of genomic data referred to other species. More complete genomes of related organisms could help to perform more accurate comparative analyses and transcriptome data of different organisms could open the way for better understanding the non-coding transcripts world. These data are expected to be available in the near future as, according to what discussed in the previous chapters of this thesis, the revolution started one decade ago in DNA sequencing techniques is still ongoing.

Bibliography

1. Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P.A., Murphy, T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D., *et al.* (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. *27*, 849–864.
2. Consortium, I.H.G.S. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.
3. The Cost of Sequencing a Human Genome Natl. Hum. Genome Res. Inst. NHGRI. Available at: https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/.
4. DNA Sequencing Costs: Data Natl. Hum. Genome Res. Inst. NHGRI. Available at: https://www.genome.gov/27541954/dna-sequencing-costs-data/ [Accessed June 25, 2018].
5. Illumina Press Release. Available at: https://emea.illumina.com/company/news-center/press-releases/press-release-details.html?newsid=2236383 [Accessed July 18, 2018].
6. Scopus Home Page. Available at: https://www.scopus.com/ [Accessed June 25, 2018].
7. Berberich, A.J., Ho, R., and Hegele, R.A. (2018). Whole genome sequencing in the clinic: empowerment or too much information? CMAJ Can. Med. Assoc. J. *190*, E124–E125.
8. Towbin, J.A. (2014). INHERITED CARDIOMYOPATHIES. Circ. J. Off. J. Jpn. Circ. Soc. *78*, 2347–2356.
9. Oliveira, T.G.M., Mitne-Neto, M., Cerdeira, L.T., Marsiglia, J.D.C., Arteaga-Fernandez, E., Krieger, J.E., and Pereira, A.C. (2015). A Variant Detection Pipeline for Inherited Cardiomyopathy-Associated Genes Using Next-Generation Sequencing. J. Mol. Diagn. JMD *17*, 420–430.
10. Poloni, G., De Bortoli, M., Calore, M., Rampazzo, A., and Lorenzon, A. (2016). Arrhythmogenic right-ventricular cardiomyopathy: molecular genetics into clinical practice in the era of next generation sequencing. J. Cardiovasc. Med. Hagerstown Md *17*, 399–407.
11. Meienberg, J., Bruggmann, R., Oexle, K., and Matyas, G. (2016). Clinical sequencing: is WGS the better WES? Hum. Genet. *135*, 359–362.
12. Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Ozen, S., Sanjad, S., *et al.* (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc. Natl. Acad. Sci. U. S. A. *106*, 19096–19101.
13. Li, M.H., Abrudan, J.L., Dulik, M.C., Sasson, A., Brunton, J., Jayaraman, V., Dugan, N., Haley, D., Rajagopalan, R., Biswas, S., *et al.* (2015). Utility and limitations of exome sequencing as a genetic diagnostic tool for conditions associated with pediatric sudden cardiac arrest/sudden cardiac death. Hum. Genomics *9*.
14. Meynert, A.M., Ansari, M., FitzPatrick, D.R., and Taylor, M.S. (2014). Variant

detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics *15*.
15. Meienberg, J., Zerjavic, K., Keller, I., Okoniewski, M., Patrignani, A., Ludin, K., Xu, Z., Steinmann, B., Carrel, T., Röthlisberger, B., *et al.* (2015). New insights into the performance of human whole-exome capture platforms. Nucleic Acids Res. *43*, e76.
16. Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q.B., Antipenko, A., Shang, L., Boisson, B., Casanova, J.-L., and Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc. Natl. Acad. Sci. U. S. A. *112*, 5473–5478.
17. Hegde, M., Santani, A., Mao, R., Ferreira-Gonzalez, A., Weck, K.E., and Voelkerding, K.V. (2017). Development and Validation of Clinical Whole-Exome and Whole-Genome Sequencing for Detection of Germline Variants in Inherited Disease. Arch. Pathol. Lab. Med. *141*, 798–805.
18. Stavropoulos, D.J., Merico, D., Jobling, R., Bowdin, S., Monfared, N., Thiruvahindrapuram, B., Nalpathamkalam, T., Pellecchia, G., Yuen, R.K.C., Szego, M.J., *et al.* (2016). Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine. Npj Genomic Med. *1*, 15012.
19. Lionel, A.C., Costain, G., Monfared, N., Walker, S., Reuter, M.S., Hosseini, S.M., Thiruvahindrapuram, B., Merico, D., Jobling, R., Nalpathamkalam, T., *et al.* (2018). Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genet. Med. *20*, 435–443.
20. Farnaes, L., Hildreth, A., Sweeney, N.M., Clark, M.M., Chowdhury, S., Nahas, S., Cakici, J.A., Benson, W., Kaplan, R.H., Kronick, R., *et al.* (2018). Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. Npj Genomic Med. *3*, 10.
21. Li, R., Li, Y., Zheng, H., Luo, R., Zhu, H., Li, Q., Qian, W., Ren, Y., Tian, G., Li, J., *et al.* (2010). Building the sequence map of the human pan-genome. Nat. Biotechnol. *28*, 57–63.
22. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. *25*, 1754–1760.
23. GATK Home Page. Available at: https://software.broadinstitute.org/gatk/.
24. MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A., *et al.* (2014). Guidelines for investigating causality of sequence variants in human disease. Nature *508*, 469–476.
25. Bao, R., Huang, L., Andrade, J., Tan, W., Kibbe, W.A., Jiang, H., and Feng, G. (2014). Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. Cancer Inform. *13*, 67–82.
26. Kassahn, K.S., Scott, H.S., and Caramins, M.C. (2014). Integrating massively parallel sequencing into diagnostic workflows and managing the annotation and clinical interpretation challenge. Hum. Mutat. *35*, 413–423.

27. Bertoldi, L., Forcato, C., Vitulo, N., Birolo, G., De Pascale, F., Feltrin, E., Schiavon, R., Anglani, F., Negrisolo, S., Zanetti, A., *et al.* (2017). QueryOR: a comprehensive web platform for genetic variant analysis and prioritization. BMC Bioinformatics *18*, 225.
28. Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. *39*, e118.
29. Ng, P.C., and Henikoff, S. (2001). Predicting Deleterious Amino Acid Substitutions. Genome Res. *11*, 863–874.
30. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. *Chapter 7*, Unit7.20.
31. Makrythanasis, P., and Antonarakis, S.E. (2013). Pathogenic variants in non-protein-coding sequences. Clin. Genet. *84*, 422–428.
32. Hrdlickova, B., de Almeida, R.C., Borek, Z., and Withoff, S. (2014). Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. Biochim. Biophys. Acta *1842*, 1910–1922.
33. Zhang, F., and Lupski, J.R. (2015). Non-coding genetic variants in human disease. Hum. Mol. Genet. *24*, R102-110.
34. Li, H., He, Z., Gu, Y., Fang, L., and Lv, X. (2016). Prioritization of non-coding disease-causing variants and long non-coding RNAs in liver cancer. Oncol. Lett. *12*, 3987–3994.
35. Ritchie, G.R.S., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation of noncoding sequence variants. Nat. Methods *11*, 294–296.
36. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.
37. Guo, Y., Dai, Y., Yu, H., Zhao, S., Samuels, D.C., and Shyr, Y. (2017). Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. Genomics *109*, 83–90.
38. Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinforma. Oxf. Engl. *30*, 2843–2851.
39. Miga, K.H., Eisenhart, C., and Kent, W.J. (2015). Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments. Nucleic Acids Res. *43*, e133.
40. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., *et al.* (2001). The sequence of the human genome. Science *291*, 1304–1351.
41. NCBI Genome Assembly. Available at: https://www.ncbi.nlm.nih.gov/assembly/.
42. NCBI Genome Assembly Model. Available at: https://www.ncbi.nlm.nih.gov/assembly/model/.
43. Church, D.M., Schneider, V.A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W.M., Ritchie, G.R.S., *et al.* (2011).

Modernizing reference genome assemblies. PLoS Biol. *9*, e1001091.
44. NCBI Genome Assembly GRCh37.p13. Available at: https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.25/.
45. NCBI Genome Assembly GRCh38. Available at: https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/.
46. NCBI Genome Assembly GRCh38.p12. Available at: https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.38/.
47. Kidd, J.M., Sampas, N., Antonacci, F., Graves, T., Fulton, R., Hayden, H.S., Alkan, C., Malig, M., Ventura, M., Giannuzzi, G., *et al.* (2010). Characterization of Missing Human Genome Sequences and Copy-number Polymorphic Insertions. Nat. Methods *7*, 365–371.
48. Hajirasouliha, I., Hormozdiari, F., Alkan, C., Kidd, J.M., Birol, I., Eichler, E.E., and Sahinalp, S.C. (2010). Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. Bioinformatics *26*, 1277–1283.
49. Alkan, C., Sajjadian, S., and Eichler, E.E. (2011). Limitations of next-generation genome sequence assembly. Nat. Methods *8*, 61–65.
50. Genovese, G., Handsaker, R.E., Li, H., Altemose, N., Lindgren, A.M., Chambert, K., Pasaniuc, B., Price, A.L., Reich, D., Morton, C.C., *et al.* (2013). Using population admixture to help complete maps of the human genome. Nat. Genet. *45*, 406–414, 414e1-2.
51. Genovese, G., Handsaker, R.E., Li, H., Kenny, E.E., and McCarroll, S.A. (2013). Mapping the human reference genome's missing sequence by three-way admixture in Latino genomes. Am. J. Hum. Genet. *93*, 411–421.
52. Chen, G., Wang, C., Shi, L., Tong, W., Qu, X., Chen, J., Yang, J., Shi, C., Chen, L., Zhou, P., *et al.* (2013). Comprehensively identifying and characterizing the missing gene sequences in human reference genome with integrated analytic approaches. Hum. Genet. *132*, 899–911.
53. Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., *et al.* (2007). The diploid genome sequence of an individual human. PLoS Biol. *5*, e254.
54. Liu, Y., Koyutürk, M., Maxwell, S., Xiang, M., Veigl, M., Cooper, R.S., Tayo, B.O., Li, L., LaFramboise, T., Wang, Z., *et al.* (2014). Discovery of common sequences absent in the human reference genome using pooled samples from next generation sequencing. BMC Genomics *15*, 685.
55. Chaisson, M.J.P., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., *et al.* (2015). Resolving the complexity of the human genome using single-molecule sequencing. Nature *517*, 608–611.
56. Anshul Kundaje (2010). A comprehensive collection of signal artifact blacklist regions in the human genome. Available at: https://personal.broadinstitute.org/anshul/projects/encode/rawdata/blacklists/hg19-blacklist-README.pdf.
57. NCBI Genome Assembly hs38d1. Available at:

https://www.ncbi.nlm.nih.gov/assembly/GCA_000786075.2/.
58. Zheng-Bradley, X., Streeter, I., Fairley, S., Richardson, D., Clarke, L., and Flicek, P. (2017). Alignment of 1000 Genomes Project reads to reference assembly GRCh38. GigaScience *6*, 1–8.
59. Sudmant, P.H., Kitzman, J.O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., 1000 Genomes Project, *et al.* (2010). Diversity of human copy number variation and multicopy genes. Science *330*, 641–646.
60. Sudmant, P.H., Mallick, S., Nelson, B.J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B.P., Baker, C., Nordenfelt, S., Bamshad, M., *et al.* (2015). Global diversity, population stratification, and selection of human copy-number variation. Science *349*, aab3761.
61. Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science *290*, 1151–1155.
62. Ho, M.-R., Tsai, K.-W., Chen, C., and Lin, W. (2011). dbDNV: a resource of duplicated gene nucleotide variants in human genome. Nucleic Acids Res. *39*, D920-925.
63. Duplicated-gene Necleotide Variants. Available at: http://goods.ibms.sinica.edu.tw/DNVs/.
64. Nuttle, X., Huddleston, J., O'Roak, B.J., Antonacci, F., Fichera, M., Romano, C., Shendure, J., and Eichler, E.E. (2013). Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions. Nat. Methods *10*, 903–909.
65. Zook, J.M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., and Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat. Biotechnol. *32*, 246–251.
66. Goldfeder, R.L., Priest, J.R., Zook, J.M., Grove, M.E., Waggott, D., Wheeler, M.T., Salit, M., and Ashley, E.A. (2016). Medical implications of technical accuracy in genome sequencing. Genome Med. *8*, 24.
67. Torrent Suite™ Software v5.2.1 - Software Release Notes. Available at: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/TorrentSuite_v521_ReleaseNotes.pdf.
68. Paten, B., Novak, A.M., Eizenga, J.M., and Garrison, E. (2017). Genome graphs and the evolution of genome inference. Genome Res., gr.214155.116.
69. Dilthey, A., Cox, C., Iqbal, Z., Nelson, M.R., and McVean, G. (2015). Improved genome inference in the MHC using a population reference graph. Nat. Genet. *47*, 682–688.
70. Novak, A.M., Hickey, G., Garrison, E., Blum, S., Connelly, A., Dilthey, A., Eizenga, J., Elmohamed, M.A.S., Guthrie, S., Kahles, A., *et al.* (2017). Genome Graphs. bioRxiv, 101378.
71. Rand, K.D., Grytten, I., Nederbragt, A.J., Storvik, G.O., Glad, I.K., and Sandve, G.K. (2017). Coordinates and intervals in graph-based reference genomes. BMC Bioinformatics *18*.
72. Rakocevic, G., Semenyuk, V., Spencer, J., Browning, J., Johnson, I.,

Arsenijevic, V., Nadj, J., Ghose, K., Suciu, M.C., Ji, S.-G., *et al.* (2018). Fast and Accurate Genomic Analyses using Genome Graphs. bioRxiv, 194530.
73. Graph - Seven Bridges. Available at: https://www.sevenbridges.com/graph/.
74. Barbitoff, Y.A., Bezdvornykh, I.V., Polev, D.E., Serebryakova, E.A., Glotov, A.S., Glotov, O.S., and Predeus, A.V. (2018). Catching hidden variation: systematic correction of reference minor allele annotation in clinical variant calling. Genet. Med. *20*, 360–364.
75. Data | 1000 Genomes. Available at: http://www.internationalgenome.org/data.
76. Giorgio Valle Group | Genomics and Bioinformatics Unit – University of Padua. Available at: http://genomics.cribi.unipd.it/main/.
77. de Ligt, J., Willemsen, M.H., van Bon, B.W.M., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., *et al.* (2012). Diagnostic exome sequencing in persons with severe intellectual disability. N. Engl. J. Med. *367*, 1921–1929.
78. European Genome-phenome Archive - Study EGAS00001000287, Dataset EGAD00001000277. Available at: https://www.ebi.ac.uk/ega/datasets/EGAD00001000277.
79. Tan, A., Abecasis, G.R., and Kang, H.M. (2015). Unified representation of genetic variants. Bioinforma. Oxf. Engl. *31*, 2202–2204.
80. Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.-P., and Wang, L. (2014). CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinforma. Oxf. Engl. *30*, 1006–1007.
81. Introduction to the GATK Best Practices. Available at: https://software.broadinstitute.org/gatk/best-practices/.
82. dbSNP Home Page. Available at: https://www.ncbi.nlm.nih.gov/SNP/.
83. Sherry, S.T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.
84. Exome Variant Server. Available at: http://evs.gs.washington.edu/EVS/.
85. ExAC Browser. Available at: http://exac.broadinstitute.org/.
86. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
87. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92.
88. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.
89. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. *57*, 289–300.

90. 1000 Genomes Project - Phase1. Available at: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated_call_sets/.
91. 1000 Genomes Project - Phase 3, GRCh37. Available at: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.
92. Genome Research Supplemental Material. Available at: https://genome.cshlp.org/content/suppl/2017/04/10/gr.213611.116.DC1/Supplemental_VCF_S1_S2.zip.
93. 1000 Genomes Project - Phase 3, GRCh38. Available at: ftp://ftp.ensembl.org/pub/release-90/variation/vcf/homo_sapiens/1000GENOMES- phase_3.vcf.gz.
94. GIAB - The Joint Initiative for Metrology in Biology. Available at: http://jimb.stanford.edu/giab [Accessed August 5, 2018].
95. Ashkenazim Trio Whole Genome Data. Available at: ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/.
96. Cinese Trio Whole Genome Data. Available at: ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/ChineseTrio/.
97. Zook, J.M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C.E., Alexander, N., *et al.* (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci. Data *3*, 160025.
98. Primate Genomes List - Genome – NCBI. Available at: https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/primates.
99. Moreno-Hagelsieb, G., and Latimer, K. (2008). Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinforma. Oxf. Engl. *24*, 319–324.
100. Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. *302*, 205–217.
101. Multiple Sequence Alignment - CLUSTALW. Available at: https://www.genome.jp/tools-bin/clustalw.
102. T-Coffee 11 and related packages documentation. Available at: https://tcoffee.readthedocs.io/en/latest/.
103. MUSCLE. Available at: https://www.drive5.com/muscle/.
104. MAFFT version7. Available at: https://mafft.cbrc.jp/alignment/server/.
105. Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. *17*, 540–552.
106. Gblocks Documentation. Available at: http://molevol.cmima.csic.es/castresana/Gblocks/Gblocks_documentation.html.
107. Genome Reference Consortium Assembly Terminology. Available at: https://www.ncbi.nlm.nih.gov/grc/help/definitions/.
108. Keaney, L., Williams, F., Meenagh, A., Sleator, C., and Middleton, D. (2004). Investigation of killer cell immunoglobulin-like receptor gene diversity III. KIR2DL3. Tissue Antigens *64*, 188–194.
109. Middleton, D., and Gonzelez, F. (2010). The extensive polymorphism of KIR

genes. Immunology *129*, 8–19.
110. Wysk, M., Yang, D.D., Lu, H.T., Flavell, R.A., and Davis, R.J. (1999). Requirement of mitogen-activated protein kinase kinase 3 (MKK3) for tumor necrosis factor-induced cytokine expression. Proc. Natl. Acad. Sci. U. S. A. *96*, 3763–3768.
111. Bossi, G. (2016). MKK3 as oncotarget. Aging *8*, 1–2.
112. The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
113. Barrett, L.W., Fletcher, S., and Wilton, S.D. (2012). Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. Cell. Mol. Life Sci. *69*, 3613–3634.
114. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., *et al.* (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.
115. Fu, Y., Liu, Z., Lou, S., Bedford, J., Mu, X.J., Yip, K.Y., Khurana, E., and Gerstein, M. (2014). FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol. *15*, 480.
116. Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods *12*, 931–934.
117. Kopp, F., and Mendell, J.T. (2018). Functional Classification and Experimental Dissection of Long Noncoding RNAs. Cell *172*, 393–407.
118. Brannan, C.I., Dees, E.C., Ingram, R.S., and Tilghman, S.M. (1990). The product of the H19 gene may function as an RNA. Mol. Cell. Biol. *10*, 28–36.
119. Borsani, G., Tonlorenzi, R., Simmler, M.C., Dandolo, L., Arnaud, D., Capra, V., Grompe, M., Pizzuti, A., Muzny, D., Lawrence, C., *et al.* (1991). Characterization of a murine gene expressed from the inactive X chromosome. Nature *351*, 325–329.
120. Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., *et al.* (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311–1323.
121. Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., *et al.* (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120–124.
122. DiStefano, J.K. (2018). The Emerging Role of Long Noncoding RNAs in Human Disease. Methods Mol. Biol. Clifton NJ *1706*, 91–110.
123. Miao, Y.-R., Liu, W., Zhang, Q., and Guo, A.-Y. (2018). lncRNASNP2: an updated database of functional SNPs and mutations in human and mouse lncRNAs. Nucleic Acids Res. *46*, D276–D280.
124. Pan, W., Liu, L., Wei, J., Ge, Y., Zhang, J., Chen, H., Zhou, L., Yuan, Q., Zhou, C., and Yang, M. (2016). A functional lncRNA HOTAIR genetic variant contributes to gastric cancer susceptibility. Mol. Carcinog. *55*, 90–96.

125. Verhaegh, G.W., Verkleij, L., Vermeulen, S.H.H.M., den Heijer, M., Witjes, J.A., and Kiemeney, L.A. (2008). Polymorphisms in the H19 gene and the risk of bladder cancer. Eur. Urol. *54*, 1118–1126.
126. PhastCons scores for multiple alignments of 33 placental mammal genomes to the human genome. Available at: http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/placentalMammals/.
127. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.
128. Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., and Bozzoni, I. (2011). A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell *147*, 358–369.
129. Legnini, I., Morlando, M., Mangiavacchi, A., Fatica, A., and Bozzoni, I. (2014). A feedforward regulatory loop between HuR and the long noncoding RNA linc-MD1 controls early phases of myogenesis. Mol. Cell *53*, 506–514.
130. Lin, Q., Schwarz, J., Bucana, C., and Olson, E.N. (1997). Control of mouse cardiac morphogenesis and myogenesis by transcription factor MEF2C. Science *276*, 1404–1407.
131. Lilly, B., Zhao, B., Ranganayakulu, G., Paterson, B.M., Schulz, R.A., and Olson, E.N. (1995). Requirement of MADS domain transcription factor D-MEF2 for muscle formation in Drosophila. Science *267*, 688–693.
132. Luo, D., Renault, V.M., and Rando, T.A. (2005). The regulation of Notch signaling in muscle stem cell activation and postnatal myogenesis. Semin. Cell Dev. Biol. *16*, 612–622.
133. Wilson-Rawls, J., Molkentin, J.D., Black, B.L., and Olson, E.N. (1999). Activated notch inhibits myogenic activity of the MADS-Box transcription factor myocyte enhancer factor 2C. Mol. Cell. Biol. *19*, 2853–2862.
134. Mammal Genomes List - Genome – NCBI. Available at: https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/mammals.
135. Rennie, W., Liu, C., Carmack, C.S., Wolenc, A., Kanoria, S., Lu, J., Long, D., and Ding, Y. (2014). STarMir: a web server for prediction of microRNA binding sites. Nucleic Acids Res. *42*, W114-118.

Manuscript draft

Data mining of recurrent variants reveals inconsistencies in the human reference genome

Running title: Mining inconsistencies in human reference genome

Fabio De Pascale,* Margherita Ferrarini,* Loris Bertoldi,* Alessandro Vezzi, and Giorgio Valle[1]

Department of Biology, Università degli Studi di Padova, Padova, 35131, Italy

1) To whom correspondence should be addressed. Tel: +390498276281; email: giorgio.valle@unipd.it; Dept. Biology via Ugo Bassi 58B Padova, 35131, Italy

* These authors contributed equally to this work.

The authors declare that they have no competing interests.

This work was supported by the University of Padova, BIOINFOGEN project.

Abstract

In human genetics several problems are responsible for the call of false-positive variants occurring with high frequency in exome and genome analyses. It is known that the reference genome does not always represent the real consensus sequence of the human population, due to the inclusion of rare alleles and sequencing errors. In particular, genomic duplications are often misassembled and as a result they may be found in the reference genome as a collapsed consensus, thus generating false variants. In this work we performed a thorough search for conflicting information between the human reference genome (GRCh37 and GRCh38) and some of the most popular human genetic resources such as the 1000 Genomes Project, to disclose minor alleles and to mine genetic inconsistencies. To search for unreported genomic duplications, we performed a genome wide screening for unbalanced heterozygosity. We found that inaccuracies and errors are much higher

than expected. Minor alleles occurring with a frequency <10% are found on average every ~7,000 bases and include many rare variants that are never found elsewhere, producing high numbers of false positives as well as possible false negatives. The systematic screening for unbalanced heterozygosity revealed ~86,000 variants that are likely the result of unreported genomic duplications, involving functionally relevant genes such as MAP2K3 and KCNJ12. Our findings may help the ongoing quest to obtain a highly accurate human genome reference sequence. Moreover, the results presented in this study will be useful to human geneticists in the process of filtering and selecting causative variants.

Key words: whole exome sequencing; whole genome sequencing; human reference genome; recurrent variants; GRCh37; GRCh38

Introduction

Since its first draft, released in 2001[1], the reference sequence of the human genome has undergone several updates and improvements. Notably, in 2009 the Human Genome Reference Consortium made available the GRCh37 release (also known as hg19) that was followed by the GRCh38 release in 2013 and further updated in the following years, in the form of "patches" and alternative loci.

Interestingly, many users are still adopting GRCh37 for their studies[2]. This can be partly explained by the difficulties in updating tools and pipelines when a new version of the genome becomes available. Indeed, many commercially available exome kits, for instance the "Ion AmpliSeq Exome RDY Kit" from Thermo Fisher Scientific or the "Nextera Rapid Capture Exome" from Illumina are still based on the old GRCh37 release. As a consequence, GRCh37 is also recommended for bioinformatics analyses.

The reluctance to update to the new release of the genome has several drawbacks because GRCh38 contains important improvements. It was derived from many donors instead of a few and includes many amendments; furthermore, GRCh38 supports the representation of complex haplotypes with the introduction of

alternative loci as well as many regions that were missing in the previous release such as segmental duplications, centromeres and telomeres[2].

The problems arising from using the old reference genome for next generation sequencing (NGS) data analysis have already been widely discussed in the literature. Two different studies demonstrated that the poor representation of repeated sequences in GRCh37 produces read misalignments and false positive variants[3,4]. To solve this problem it is possible to include in the analysis "decoy" sequence[3] or "sponge" database[4] that allow an improvement in read mapping and in the resolution of false heterozygous calls[3].

More recent studies on GRCh38 showed a further improvement of read mappability and a decrease of false positive single-nucleotide variants[5,6]. However we observed that the problem of many false positive variants remains also in GRCh38. In general, both GRCh37 and GRCh38 produce false positive recurrent variants that are found with an allelic frequency >50%. Generally, in a typical exome several thousand such variants are expected, while in whole genomes they can be well over one million. In many cases these high frequency (HF) variants are due to minor alleles in reference genome (MAiR), which can be easily filtered out with appropriate tools[7]. However, even after removing what is reported as common variant in the databases, many shared variants still remain. Therefore, the presence of some HF variants cannot be explained only by MAiRs. An interesting hypothesis is that there may be instrument-specific sequencing errors, but as will be further detailed, we found that exomes obtained by different technologies such as Ion Proton, Illumina and SOLiD, exhibit a largely overlapping set of HF variants, indicating that the problem is not due to artifacts of a particular chemistry or sequencing platform.

The study presented in this paper has two main aims: firstly we want to evaluate and classify HF variants, both in GRCh37 and GRCh38. We believe that a clear repertoire of the recurrent miscalls will help geneticists in analyzing exome data, facilitating the process of variant prioritization.

A second, but not less important scope of this paper is to better understand the nature of this problem and to verify the hypothesis that some of these false

positives may originate from duplicated regions that are not reported in the reference genome. This can be experimentally verified because any "collapsed" repeated sequence in the reference genome would be the target for reads derived from two or more real genomic regions, resulting in a disproportion between frequency, heterozygosity and homozygosity of alleles.

With this premise we analyzed exomes from different platforms, as well as whole genome sequencing (WGS) data, using both GRCh37 and GRCh38. We found that the problem of collapsed repeats is indeed responsible for the call of many false positive variants, several of which are still present in GRCh38. Furthermore, we suggest several positions of the reference genome that require a revision in future updates.

Materials and Methods

Exome Datasets

In this study, three different exome datasets were used. The main dataset is composed by 222 exomes enriched with the Ion AmpliSeq Exome panel and sequenced at the CRIBI facility, University of Padua (genomics.cribi.unipd.it)[8], with the Ion Proton system (Thermo Fisher Scientific). These samples came from a wide range of projects including cohorts of individuals, trios and individual patients. The second dataset includes 22 exomes enriched with Illumina TruSeq Exome panel and sequenced with Illumina NextSeq 500 platform at CRIBI[8]. The third dataset refers to the study published by de Ligt *et al.*[9], on 300 exomes enriched with SOLiD-optimized target enrichment and sequenced with SOLiD 4 System (Life Technologies), belonging to 100 trios composed of patients with unexplained severe intellectual disability and their unaffected parents (European Genome-phenome Archive, https://www.ebi.ac.uk/ega: study EGAS0000100028, dataset EGAD00001000277).

Alignment and variant calling

All the exomes were aligned against the primary assembly of GRCh37, as recommended by the manufacturers of the enrichment kits. A detailed description

of the analysis workflow of each dataset is available in Supplementary File S1. The 222 Ion Proton exomes were also aligned on GRCh38.p10, downloaded from Ensembl. Alignment and variant calling were performed according to the Torrent Suite 5.0 exome analysis pipeline. The target regions (release 2014) were migrated from GRCh37 to GRCh38 coordinates using CrossMap[10].

Identification of exome variants mapped on MAiRs

Minor Allele in Reference (MAiR) are those positions of the human reference genome with an allele that is not the most frequent in the population[7]. To identify variants in the Ion Proton dataset falling in these positions we analyzed their allelic frequencies in 3 different databases, namely dbSNP[11] release 144 , ExAC[12] release 0.3.1 and ESP6500SI-V2. A genomic position was marked as MAiR if the reference allele frequency was lower than any alternative allele frequency in all three databases. More details are reported in Supplementary File S1.

Impact of MAiR positions at the protein level

Variants in GRCh37 MAiR positions confirmed in GRCh38 genome were annotated using both SnpEff v.4.2[13] and VEP v.84[14], employing respectively UCSC and RefSeq transcripts. These two different annotations were chosen to avoid transcript-dependent biases. Missense variants were selected from the two annotated VCF files and analyzed with an in-house developed Python script to allocate them to one of the following three classes: i) match to the manually reviewed human protein sequence of SwissProt, ii) match to the Human Polymorphisms and Disease Mutations release 2017_05 of UniProt, iii) neither of the above.

HF variants in 1000 Genomes on GRCh37 and GRCh38

1000 Genomes Project VCF files Phase3 for both GRCh37 and GRCh38 were collected respectively from 1000 Genomes ftp website and Ensembl ftp website. Variants reported with frequencies higher than 50% were marked as high frequency (HF). A HF variant was considered amended if it was called against GRCh37, but not in GRCh38. These analyses were performed using an in-house Python script.

Statistical test on heterozygous genotype frequencies

The GRCh37 dataset of Phase3 1000 Genomes Project was searched for variants with unbalanced heterozygous genotype frequency. Each variant was tested with a one-tailed binomial test between observed and expected allele frequencies. Observed frequencies were computed as the number of heterozygous genotypes divided by the total number of genomes. The expected frequencies were computed with the Hardy-Weinberg formula in which the heterozygous genotype frequency can be calculated as $x=2pq$, where q is the alternative allele frequency reported in the VCF file and p is the reference frequency calculated as $1-q$. Resulting probabilities were corrected for false discovery rate using the Benjamini-Hochberg procedure[15]. Variants were considered significantly unbalanced if the corrected probability was lower than 0.01. This analysis was performed only for biallelic variants.

Results

Ion Proton exome dataset

The work presented in this paper originated from the analysis of a heterogeneous dataset of 222 exomes produced in our laboratory using Ion Proton technology and Ampliseq chemistry, as detailed in the Material and Methods.

The overall analysis of the data led to the identification of 264,303 variants called against the GRCh37 reference genome, including 239,255 SNPs and 25,048 small INDELs (14,075 deletions and 10,973 insertions). Among the total variants, 245,088 were detected as biallelic in the analyzed population, whereas 19,215 were multiallelic.

We should expect that the reference genome reports the most common alleles in the population; therefore, the variants found in the exomes should be minor alleles. Instead we found that 9,313 variants were present in more than 90% of the individuals. A full list of these variants is provided in the Supplementary File S2. Remarkably, 2,349 variants were found in 100% of the individuals. This finding is not completely unexpected because Minor Alleles In Reference (MAiRs) are a

known problem[7,16]; however the large number of their occurrences is notable. Interestingly, we realigned the exomic sequences on GRCh38 and found that 8,132 out of 9,313 remained uncorrected.

The above values refer to the presence of variants in a diploid genome. In terms of allelic frequency, we found that 3,898 variants had an allelic frequency equal or greater than 90%; of these, 841 scored 100% allelic frequency, being homozygous in all of the samples. This number of recurrent variants is unexpectedly high, considering that the average number of variants in each sample was 48,785. This finding is even more surprising since a considerable number of recurrent variants are reported with a low frequency in databases and in some cases they are not reported at all as known variants, making difficult the process of recognizing them as false positives.

It should be considered that the reference genome should ideally contain the major alleles and therefore as a result we should not find variants with allelic frequencies above 50%. Fluctuations due to subsampling and/or ethnicity are certainly possible, but cannot explain this high number of HF variants.

Comparison with Illumina and SOLiD exome datasets

To verify whether the recurrent variants found in at least 90% of the exomes could result from Ion Proton specific errors, we analyzed the exomes of two independent datasets produced with Illumina and SOLiD, using their respective enrichment, sequencing and analysis pipelines, as detailed in the Materials and Methods. In Table 1 it can be seen that the large majority of variants that occur in >90% of the exomes was confirmed also with the Illumina and SOLiD technologies.

Unfortunately, the exomic target regions captured with different technologies are not precisely overlapping; thus, of the 9,313 Ion Proton variants, only 6,085 fell in regions covered by the Illumina target and 7,046 in regions covered by the SOLiD target. Variants were considered “confirmed” if they were respectively present in at least 50% of Illumina or SOLiD analyses. As shown in Table 1, the very large majority (99%) of the variants on target were confirmed by Illumina, while 81% were confirmed by SOLiD. Unconfirmed variants could either be false positives of the Ion Proton or false negatives of the Illumina and SOLiD. Only 41 variants that

were localized in common target regions were not confirmed in both Illumina and SOLiD and therefore these could be Ion Proton specific systematic errors.

Minor Alleles in GRCh37 and GRCh38 reference genomes

To better clarify how many minor alleles are present in the reference genomes, we screened the Phase 3 VCF files of the 1000 Genomes Project (1KGP), consisting of 2,504 whole genome sequences, from 26 populations, aligned on GRCh37[17]. The results of this analysis are shown in Figure 1A.

We found 436,700 (about 145 variants per Mbp) with an allelic frequency equal or greater than 90% and 32,105 variants (about 11 variants per Mbp) with 100% allelic frequency, in homozygosity in all the individuals. These findings show that the surprisingly high number of HF variants observed in exomes was also confirmed in whole genomes.

Recently, the European Bioinformatics Institute has re-aligned the 1KGP sequencing data on the GRCh38 reference genome[6]. Therefore, we were interested in evaluating the number of HF variants found in GRCh37 that have been amended in GRCh38. A total number of 2,198,258 HF variants positions were present in the GRCh37 VCF file. For each position we checked if in the GRCh38 release the reference allele was substituted with the most common allele in the population or if the position was not included in GRCh38. We found that this correction occurred only in 70,497 cases, while 2,127,761 positions maintained the less frequent allele in the population. Figure 1B shows these results in more detail. It can be seen that the rate of correction is practically the same, around 3%, in the range between 0% and 95%. Some improvement is seen for the alleles with a frequency between 95%-100%, that were corrected in 5.6% of the cases. Nevertheless, the large majority of these loci still maintain the minor allele in the GRCh38 reference genome, including 8,593 alleles with a frequency of 100%.

Mining for incongruities

MAiRs undoubtedly provide a very easy and satisfactory explanation of variants that are found in the population with a high frequency. However, a closer look at some of the HF variants revealed that assembly errors in the reference genome may also be involved in the problem. In particular, we observed that the sequence

coverage of WGS is consistently higher than average in some specific regions of all individuals. A possible explanation is that there may be genomic duplications that are reported as single regions in the reference, which could be the source of false variant calling[18]. Indeed, any difference between the two repeats would be seen in all individuals as a heterozygous variant mapping on the "collapsed" reference. Figure 2 provides a schematic representation of a tandem duplication that in the reference is collapsed into a single region.

To verify the extent of the above problem, we performed a genome-wide analysis, aiming at the detection of regions with unbalanced genotypes. For each chromosome we considered non-overlapping 100kb sized windows and for each window we calculated the percentage of unbalanced variants based on the total number of biallelic variants. For each variant we considered its allele frequency in the whole 1KGP population and the observed frequency of homozygous and heterozygous genotypes. We considered that according to the Hardy-Weinberg equation we should expect that $p^2+2pq+q^2=1$; therefore, to verify this null hypothesis, we performed a one-tailed binomial test, corrected for false discovery as discussed in the Material and Methods. Variants were considered significantly unbalanced if their corrected p-values were lower than 0.01. The results of this analysis are shown in Figure 3. Many regions with unbalanced heterozygosity can be clearly detected. The portion of chromosome 6 corresponding to the MHC is indicated by an asterisk. It can be seen that although the MHC is possibly the most polymorphic region of the genome, it is not particularly associated to unbalanced heterozygosity, suggesting that the nature of this genetic inconsistency should be found elsewhere. A detailed list of unbalanced regions is supplied in the Supplementary File S3.

We detected 86,649 unbalanced variants that we believe to be mostly due to unreported genomic duplications. An interesting issue is to understand how many HF variants are produced by these putative unreported genomic duplications in each individual. To answer this question we analyzed the VCF file of the 2,504 individuals belonging to the 1KGP and we found that on average each individual carries 4,491,760 variants of which 1,950,334 are HF; we also found that on average each individual carries 62,308 unbalanced variants; finally we found that

in each individual, on average 24,768 HF variants are unbalanced. The diagram in Figure S1 summarizes these observations.

Analysis of the physical coverage in mate pair whole genome data

To further verify whether the genomic regions with unbalanced heterozygosity hide duplications, we focused on 570kb of chromosome 17, particularly remarkable for the presence of important genes such as *MAP2K3* (MAP Kinase Kinase 3) and *KCNJ2* (Potassium Voltage-Gated Channel, J2). We downloaded mate pair whole genome sequencing data from "The Genome in a Bottle" project[19] and aligned them both on the GRCh37 and on toplevel GRCh38 (more details in Supplementary File S1, Figure S2). Figure 4 shows the physical coverage of this genomic region. In agreement with our findings, the coverage in these two genes is at least doubled than the average when mapped on GRCh37. Using GRCh38, the physical coverage of *MAP2K3* and *KCNJ12* genes is still higher than the flanking regions, but to a lesser extent than what was observed with GRCh37.

Discussion

The results obtained from exomes and whole genomes are slightly different in terms of the number of variants per Mbp. Typically, in the genome of a single individual we detect about 4.5 million variants (Figure S1), equivalent to ~1500 variants/Mbp, whereas in exomes the average number of variants is 48,785, over a target region of 57Mbp, equivalent to only 856 variants/Mbp. This is not surprising as it is known that protein-coding sequences tend to be more conserved than other genomic regions. Similarly, it is not surprising that the reference genome seems to be slightly more accurate in coding regions. This can be reckoned by considering the number of high frequency variants, that reflects the presence of a minor allele in the reference genome. In this respect, in whole genomes we found ~145 variants per Mbp with an allelic frequency greater than 90%, whereas in the exomes we found only 68 such variants per Mbp. Surprisingly, the number of variants with 100% allelic frequency (meaning that the allele reported in the genome was never found in our analyses) is slightly greater in exomes (15/Mbp) than in whole genomes (11/Mbp). Unfortunately, we found that only a small percentage of these

genomic positions, about 3%, have been corrected in the GRCh38 release, rising to ~5% for the variants found with a frequency above 95% (see Figure 1).

As shown in Figure S1, an interesting result of our analysis is that HF variants are not only due to MAiRs but also to regions with unbalanced heterozygosity, possibly derived from unreported genomic duplications (Figures 2 and 4). To better understand this point, we focused on the impact of HF variants on protein coding regions; moreover, we analyzed whether the corresponding position of the genome had been corrected in GRCh38. To obtain an answer to this question we focused our analysis on the exomic Ion Proton regions. Using GRCh37 we found 18,839 HF variants that have the alternative allele frequency higher than the reference in all three major databases (dbSNP, ExAC, ESV, see Materials and Methods). We further checked whether they had been corrected in GRCh38 and found that this occurred only for 1,808 HF variants, while the remaining 17,031 were unchanged.

We also examined whether any of the 17,031 HF variants retained in GRCh38 matched their corresponding protein in the UniProt variation database. We annotated the variants both with SnpEff[13] and VEP[14], employing respectively UCSC and RefSeq transcripts. We detected a comparable number of missense variants in the two databases: 3,814 with RefSeq and 3,761 with UCSC. We also found that ~2,800 of them were already known, but as minor protein variants, while only ~100 corresponded to the most frequent amino acid in the reference protein; finally, ~880 alleles found with high frequency in exomes did not show any known counterpart at the protein level. These data confirm a previous finding by Barbitoff *et al.*[16] and indicate that some general revision of the reference is required also at the protein level.

A further point that we want to expand in this discussion is about unbalanced heterozygosity on protein coding sequences. This is particularly interesting because it is very likely that these genomic regions may have been wrongly assembled and the resulting proteins may be different from what was indicated in the reference genome. Overall, in the Ion Proton exome target we found 753 variants with unbalanced heterozygosity. In particular we found 145 target regions

(amplicons) containing more than one unbalanced variant, for a total of 560 variants spanning over 45 genes.

We also investigated whether the genome reference of these 45 "unbalanced" genes was modified in the GRCh38 release. As mentioned before, unless otherwise specified, in this paper we took as a reference the GRCh37 primary assembly, that is about 3.1 Gbp long, containing one single consensus base per position. However, toplevel assemblies are also available, which include unmappable scaffolds, haplotypes and patches. In this respect, the toplevel GRCh37 and GRCh38 releases contain respectively 31 and 46 Gbp. To better understand the progress of the current human reference genome, we selected the reads previously mapped on the 45 unbalanced genes and we re-mapped them on toplevel GRCh38 using BLAST. We found that 11 genes out of the 45 still presented the same heterozygosity problem (Table 2, column 5). Of the remaining genes, only 15 were duplicated in the chromosomal primary sequence (Table 2, column 1); 5 genes were only partially duplicated and they still had some regions with heterozygosity (Table 2, column 2); 4 genes were duplicated on scaffolds not placed in chromosomal sequences (Table 2, column 3) and 10 genes were not duplicated, but reported as different haplotypes in the toplevel assembly (Table 2, column 4). It should be noticed that haplotypes should not produce unbalanced heterozygosity when aligned on the primary reference genome, as seen for the MHC locus on chromosome 6 (Figure 3). Therefore, some of the genes reported in column 4 should be revised as they could be duplications rather than haplotypes.

Among the genes with unbalanced heterozygosity in GRCh37 we found *PRIM2*. This was expected since it had been previously reported as a paralog gene misassembled in GRCh37[20], which was fully amended in GRCh38[2]. In fact, our screening process placed *PRIM2* as an amended gene in GRCh38.

Another interesting case is the *MAP2K3* gene (Figure 4), also known as *MKK3*, which encodes the mitogen-activated protein kinase kinase 3. This protein participates in the MAP kinase-mediated signaling cascade and has a well known role in tumor invasion and progression[21,22]. The gene maps on chromosome 17 and includes 12 exons. When we re-aligned the reads on toplevel GRCh38 we found that they mapped on the first portion of the gene, in particular exons 3, 4 and 5,

still on the same position of *MAP2K3*, maintaining the unbalanced heterozygosity. Instead, the reads from exons 9, 10, 11, 12 have a more complex behavior: they mapped on the same position if they did not carry the variants, otherwise they mapped on a nearby region that was not present on GRCh37. Therefore we can conclude that *MAP2K3* has been partially amended in GRCh38, with the insertion of a duplication of the last part of the gene, including exons 9 to 12, while it remains with its original unbalanced heterozygosity at the beginning of the gene, as seen in exons 3 to 5. This can also be appreciated in Figure 4, reporting shotgun coverage of four individuals, showing a considerable normalization of the coverage in GRCh38 as compared to GRCh37. On the other hand it can also be noted that some problems are still there as the coverage remains relatively high in the first part of the gene, thus confirming the indications resulting from the unbalanced heterozygosity about the residual inconsistency in GRCh38. A detailed description of the *MAP2K3* data is reported in the Supplementary File S1.

Conclusions

The results of our analysis show that there are many positions of the reference genome that could be amended to avoid the call of so many high frequency variants in exome and genome sequencing. Most of these HF variants map on the minor allele in the reference (MAiRs) and some map on misassembled regions with unbalanced heterozygosity. We also considered both GRCh37 and GRCh38, as primary and toplevel versions. Although there are several improvements in GRCh38, many amendments could be easily done to further improve it. The toplevel references with haplotypes and patches could be useful, but their huge size and the lack of suitable bioinformatic tools make their use very impracticable for most practical applications.

Acknowledgments

The authors are grateful to their colleagues Stefano Campanaro, Georgine Faulkner, Chiara Romualdi, Luca Pagani, Riccardo Schiavon, Stefano Toppo, Nicola Vitulo for useful suggestions.

The authors would thank the University of Padua for funding this study.

The authors would also like to thank the Exome Aggregation Consortium and the groups that provided exome variant data for comparison. A full list of contributing groups can be found at http://exac.broadinstitute.org/about.

Funding

This work was supported by the University of Padova, [Strategic project BIOINFOGEN].

Manuscript references

1 International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 2001; **409**: 860–921.
2 Schneider VA, Graves-Lindsay T, Howe K *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* 2017; **27**: 849–864.
3 Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 2014; **30**: 2843–2851.
4 Miga KH, Eisenhart C, Kent WJ. Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments. *Nucleic Acids Res* 2015; **43**: e133–e133.
5 Guo Y, Dai Y, Yu H, Zhao S, Samuels DC, Shyr Y. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics* 2017; **109**: 83–90.
6 Zheng-Bradley X, Streeter I, Fairley S, Richardson D, Clarke L, Flicek P. Alignment of 1000 Genomes Project reads to reference assembly GRCh38. *GigaScience* 2017; **6**: 1–8.
7 Bertoldi L, Forcato C, Vitulo N *et al.* QueryOR: a comprehensive web platform for genetic variant analysis and prioritization. *BMC Bioinformatics* 2017; **18**: 225.
8 CRIBI C. Genomics Facility. genomics.cribi.unipd.it.
9 de Ligt J, Willemsen MH, van Bon BWM *et al.* Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *N Engl J Med* 2012; **367**: 1921–1929.
10 Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* 2014; **30**: 1006–1007.
11 Sherry ST, Ward M-H, Kholodov M *et al.* dbSNP: the NCBI database of

genetic variation. *Nucleic Acids Res* 2001; **29**: 308–311.
12 Lek M, Karczewski KJ, Minikel EV *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016; **536**: 285–291.
13 Cingolani P, Platts A, Wang LL *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* 2012; **6**: 80–92.
14 McLaren W, Gil L, Hunt SE *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* 2016; **17**: 122.
15 Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B Methodol* 1995; **57**: 289–300.
16 Barbitoff YA, Bezdvornykh IV, Polev DE *et al.* Catching hidden variation: systematic correction of reference minor allele annotation in clinical variant calling. *Genet Med* 2017. doi:10.1038/gim.2017.168.
17 Consortium T 1000 GP. A global reference for human genetic variation. *Nature* 2015; **526**: 68.
18 Ho M-R, Tsai K-W, Chen C, Lin W. dbDNV: a resource of duplicated gene nucleotide variants in human genome. *Nucleic Acids Res* 2011; **39**: D920–D925.
19 Zook JM, Chapman B, Wang J *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* 2014; **32**: 246.
20 Genovese G, Handsaker RE, Li H *et al.* Using population admixture to help complete maps of the human genome. *Nat Genet* 2013; **45**: 406–414.
21 Bossi G. MKK3 as oncotarget. *Aging* 2016; **8**: 1–2.
22 Wysk M, Yang DD, Lu H-T, Flavell RA, Davis RJ. Requirement of mitogen-activated protein kinase kinase 3 (MKK3) for tumor necrosis factor-induced cytokine expression. *Proc Natl Acad Sci* 1999; **96**: 3763–3768.

Figure and legends of manuscript

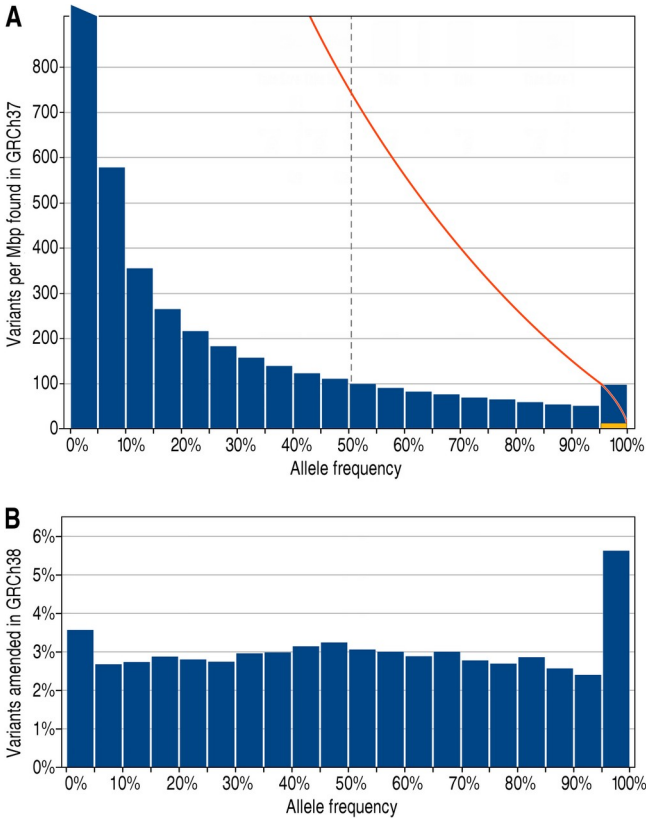

Figure 1. Allelic frequencies of variants found in GRCh37 and amended in GRCh38. The variants of 2,504 genomes (1000 Genomes Project, phase 3) were divided into classes according to their allelic frequency. Frame A: the blue blocks indicate the average number of variants per Mbp of each class. The red line indicates the sum of values from a given allele frequency to the right end, that is the number of variants with at least the indicated allele frequency. It can be seen that there are about 730 variants/Mbp with an allele frequency >50%. The yellow sector at the bottom of the 95-100% block corresponds to variants found in homozygosity in 100% of the individuals (about 11 variants / Mbp). Frame B shows the percentage of variants that have been amended in the GRCh38 release.

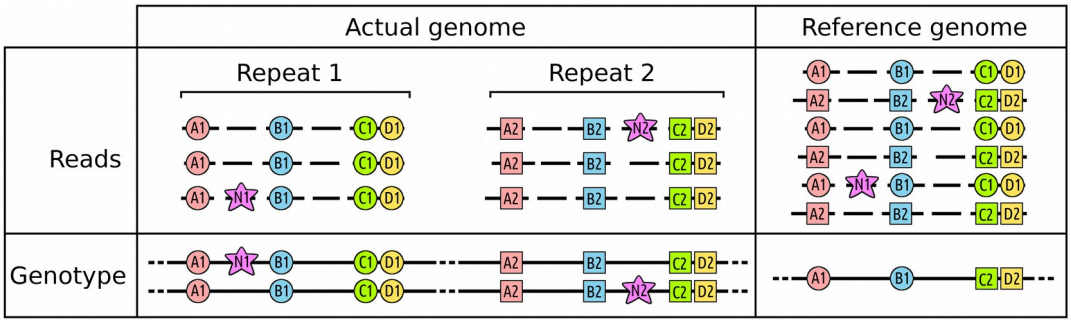

Figure 2. Hypothetical genomic region with a tandem repeat. The hypothetical tandem repeat is almost identical with the exception of four positions: A, B, C and D. This condition may be ancestral and shared by the entire population, repeat 1 having A1, B1, C1, D1 and repeat 2 having A2, B2, C2, D2. Two new variants are also shown as N1 and N2. Sometimes this kind of repeat may be misassembled in the reference genome, being reported as a single collapsed sequence, as shown in the bottom frame on the right. As a result, the four loci A, B, C and D will show a heterozygous genotype in all the individuals and the consequent variant call in all the loci, which is incompatible with the genetics.

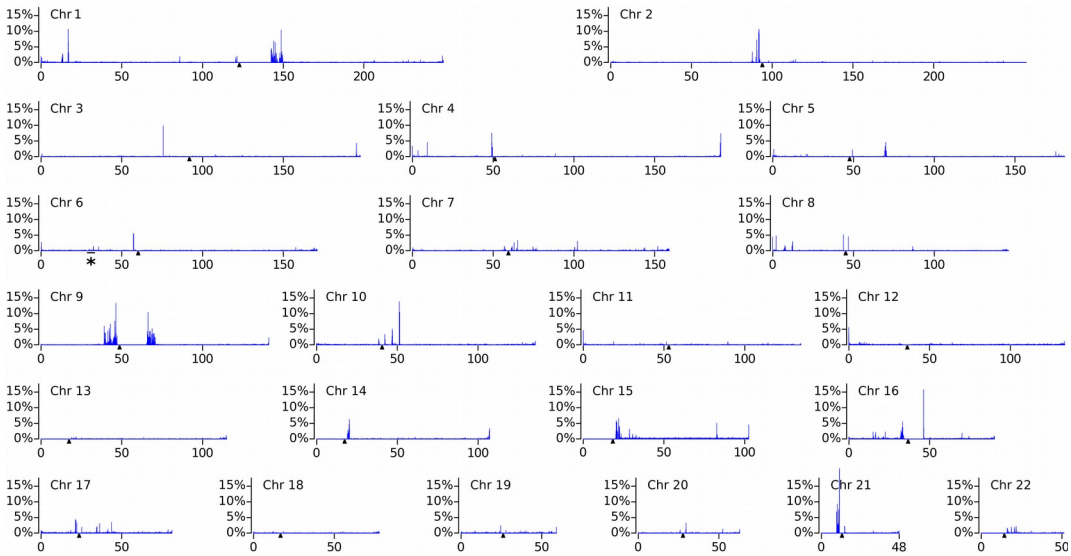

Figure 3. Genome wide analysis of regions with unbalanced heterozygosity. For each non-overlapping 100kb window we considered the percentage of biallelic variants with a significant unbalanced heterozygosity. Centromeres are indicated by a small triangle below the baseline. The region marked by the asterisk in chromosome 6 indicates the MHC.

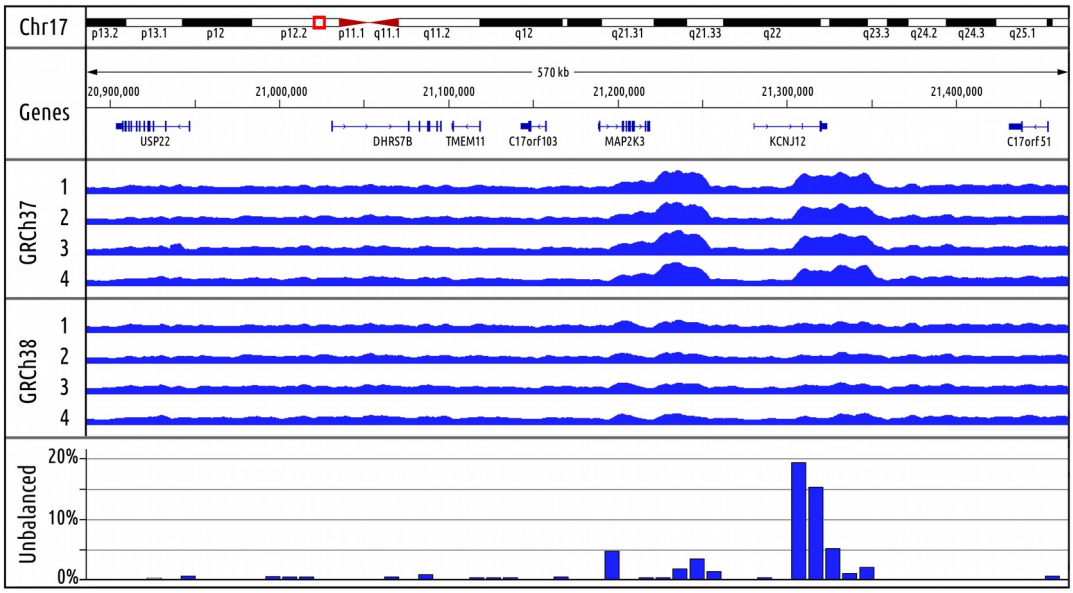

Figure 4. Physical coverage profiles of a 570kb region of chromosome 17. The set of mate pairs from "The genome in a bottle" project were aligned on GRCh37 (3.1 Gbp) and on toplevel GRCh38 (48.7 Gbp); numbers 1, 2 refer to two Ashkenazim individuals, whereas 3 and 4 refer to two Chinese individuals. The frame at the bottom shows the percentage of variants with unbalanced heterozygosity.

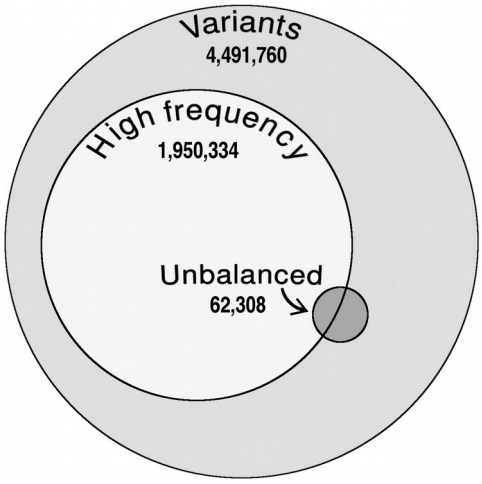

Figure S1. Variants per individual in the 1000 Genomes Project. Average number of variants per individual found in the population of 2,504 people studied in the 1000 Genomes Project. Variants have been further subdivided in High Frequency and Unbalanced variants.

Table and legends of manuscript

Ion Proton	Illumina control dataset			SOLiD control dataset		
	on target		off target	on target		off target
9313	6085		3228	7046		2267
	confirmed	not confirmed		confirmed	not confirmed	
	6008	77		5733	1313	

Table 4. Recurrent variants in Ion Proton exomes and their sharing in Illumina and SOLiD datasets. The 9,313 variants found in more than 90% of the Ion Proton exomes were analyzed to verify whether they were also present in at least 50% of the exomes obtained with Illumina and SOLiD technologies. As the exomic target regions captured with the three technologies did not precisely overlap, *'confirmed'* and *'not confirmed'* refer to variants falling in target regions shared between Ion Proton and Illumina or between Ion Proton and SOLiD (*'on target'* variants). However, also the number of Ion Proton variants outside Illumina and SOLiD target regions is reported (*'off target'* variants). A large percentage of variants that shared the exomic target was confirmed: 6,008 of the 6,085 on target variants were confirmed by Illumina (99%) and 5,733 of the 7,046 on target variants were confirmed by SOLiD (81%).

1 **Fully amended**	2 **Partially amended**	3 **Unplaced scaffold**	4 **Alternative loci**	5 **Unchanged**
BCLAF1	*FRG2B*	*CTBP2*	*CES1*	*ALG1L2*
CCDC144NL	*FRG2C*	*FAM104B*	*HLA-DQA2*	*ANKRD36*
FRG1	*KCNJ12*	*MLL3*	*HNRNPCL1*	*FAM131C*
HYDIN	*KRT6B*	*NBPF1*	*KIR2DL3*	*FAM194B*
KRTAP4-11	*MAP2K3*		*KIR2DS4*	*GPRIN2*
*LOC653486**			*KRTAP9-2*	*OR1D5*
NBPF10			*MUC20*	*PCDH11X*
NOTCH2NL			*OR9G1*	*PDPR*
OR4C3			*PRSS3*	*PER3*
OR4C45			*TNXB*	*TPTE*
OR4M2				*ZDHHC11*
PDE4DIP				
*PPYR1**				
PRIM2				
SEC22B				

Table 5. Genes with unbalanced heterozygosity in GRCh37 and their status in GRCh38. Column 1: Fully amended genes that have been duplicated within chromosomes in GRCh38 and as a result lost the variants with unbalanced heterozygosity. Column 2: Partially amended genes that are still showing unbalanced variants in some of the exons. Column 3: Genes whose duplication was found on extra chromosomal scaffolds in the primary assembly and as a result lost the variants with unbalanced heterozygosity. Column 4: Genes that have not been duplicated, but reported as different alternative loci in the full assembly. Column 5: Unchanged heterozygosity in GRCh38. **LOC653486* and *PPYR1* have changed name in GRCh38 respectively to *SCGB1C1* and *NPY4R*. More details are given in the text.

Supplementary File S1

Supplementary Materials and Methods

Ion Proton dataset, alignment and variant calling. Exomes were sequenced to reach a final mean coverage of 80x and a target uniformity higher than 90%. Alignment and variant calling were carried out according to the Torrent Suite 5.0 exome analysis pipeline, as suggested by the manufacturer. Variants were merged into a unique file using CombineVariants of Genome Analysis Toolkit (GATK v. 3.6) and then normalized applying the method proposed by Tan and colleagues (Tan *et al.* 2015) in order to eliminate different representations of the same variant. Variant annotation, based on GRCh37.82 version of Ensembl transcripts, was performed using *in-house* software. In 2014 Life Technologies provided a new smaller exome BED file, the Ion AmpliSeq Exome Hi-Q Effective Regions, without actually changing the AmpliSeq Exome panel. In this file poor performing regions are masked during the variant calling step. According to the manufacturer, the usage of this file should guarantee high confidence variant calling. For this study we considered only the variants covered by the new BED file.

Illumina control dataset, alignment and variant calling. Each sample was sequenced with 75 bp paired-end reads by Illumina NextSeq 500 to a final average coverage of 103x. Alignment and variant calling were performed according to the recommendations of the GATK Best Practices. Briefly, reads were aligned using BWA mem (v. 0.7.12) with default parameters. The resulting BAM files were further processed by Picard MarkDuplicatesWithMateCigar (Picard v. 1.55) and GATK BaseRecalibrator (GATK v. 3.6). Variant calling was performed using GATK HaplotypeCaller (GATK v. 3.6) with default parameters. The collected variants were firstly filtered using GATK VariantRecalibrator (GATK v. 3.6) and then normalized as previously described.

SOLiD control dataset. VCF files of de Ligt *et al*. (de Ligt *et al.* 2012) were downloaded from The European Genome-phenome Archive. Variant normalization was performed as indicated above.

Identification of exome variants mapped on MAIRs. The varinats databases used were: i) the Single Nucleotide Polymorphism database (dbSNP) (Sherry *et al.* 2001) version 144, modified to recover old variants excluded from this release but present in the online version; ii) the NHLBI Exome Sequencing Project (ESP) database version ESP6500SI-V2; iii) the Exome Aggregation Consortium (ExAC) database version 0.3.1 (Lek *et al.* 2016). When different populations frequencies were present, only the total one was considered.

Supplementary File 3. This is a wig format file obtained as follow: for each chromosome the percentage of unbalanced variants on the total number of biallelic variants was calculated in non-overlapping 10 kb sized windows. Values range from 0.0 to 100, with 0.0 indicating the absence of unbalanced variants in the given window; NaN values indicate that any biallelic variant was found in the given window. Data refer to GRCh37. This wig file can be downloaded at NAR online.

Confirmation of unbalanced variants of *MAP2K3*. VCF files containing variants in chromosome 17 were downloaded from two different databases: Genome Aggregation Database (gnomAD) version 2.0.1 (Lek *et al.* 2016) and 1000 Genomes Project (1KGP) database Phase1 release and Phase3 release. Using the statistical test described in the main text (see paragraph Statistical test on heterozygous genotype frequencies), variants with a heterozygous genotype frequency significantly higher than the expected were selected and subsequently compared with variants identified in *MAP2K3* in the Ion Proton dataset.

Analysis of the physical coverage in mate pair whole genome data. Whole genome sequencing mate pair data were downloaded from the Genome In A Bottle project. Samples were parents of an Ashkenazi trio and a Chinese trio. For details on libraries preparation and sequencing refer to the work of Zook and colleagues (Zook *et al.* 2016). Reads were aligned against GRCh37 (primary assembly) and GRCh38 (toplevel) with BWA mem (v. 0.7.12 with default parameters). In-house script was used to produce a physical coverage profile in *MAP2K3* and *KCNJ2* regions on chromosome 17.

Supplementary Results

***MAP2K3* gene.** Globally, 54 unbalanced variants were localized in *MAP2K3*. We checked if these variants were unbalanced in three genomes databases: gnomAD, 1KGP Phase1 release and 1KGP Phase3 release. See Material and Methods in Supplementary information for details. The majority of variants were confirmed to have an unbalanced heterozygosity in gnomAD and 1KGP Phase1 release. The not confirmed variants were absent in the databases or present with multiple alternative alleles (these variants were not included in the statistical test). Variants in 1KGP Phase3 release were collected using the hs37d5 genome reference, which corresponds to the GRCh37 primary assembly (chromosomal plus unlocalized and unplaced contigs) integrated with rCRS mitochondrial sequence, Human herpesvirus 4 type 1 and the concatenated decoy sequences. As reported by Li *et al*. (Li 2014), the integration in standard pipelines of decoy sequences allows the resolution of false heterozygous calls. In fact, the majority of unbalanced variants in *MAP2K3* were not present in the 1KGP Phase3 release because reads with variants align to the decoy sequences. As a result, none of *MAP2K3* variants were unbalanced in 1KGP Phase3 release. Results are summarized in Supplementary File 4.

We focused on 8 enriched regions of *MAP2K3* carrying more than one unbalanced variant: the first three regions match exons 3, 4 and 5, while the last five regions match exons 9, 10, 11, 12 (see Table S1). Reads from these regions were realigned against the GRCh37 and GRCh38 toplevel human genome assemblies.

Results in Table S1 show that exons 3, 4 and 5 behaved very differently from exons 9, 10, 11 and 12. In the former, reads aligned only to the original target gene independently of the reference used. In the latter, the amelioration of the reference genome lead to different results: in GRCh37 reads carrying all variants aligned to a fix patch, called HG987_PATCH, while in the GRCh38 they aligned to a new region added in the chromosome 17; on the other side, in both the references reads with none variant aligned to the original target gene.

target region name	exon	reads with all variants		reads with none variant		GRCh38 confirmed variants
		GRCh37	GRCh38	GRCh37	GRCh38	
MAP2K3_158294.12020	3	gene	gene	gene	gene	9
MAP2K3_158295.17245	4	gene	gene	gene	gene	6
MAP2K3_158296.5164	5	gene	gene	gene	gene	10
MAP2K3_158300.13718	9	patch	new region	gene	gene	0
MAP2K3_158301.11673	10	patch	new region	gene	gene	0
MAP2K3_158301.16511	10	patch	new region	gene	gene	0
MAP2K3_158302.11191	11	patch	new region	gene	gene	0
MAP2K3_158303.8516	12	patch	new region	gene	gene	1

Table S1. **BLAST result of read realignments.** The results for reads with all or none variants are reported separately: for each group the alignments on both the references used (toplevel GRCh37 and GRCh38) are shown; "gene" corresponds to the original target gene, "new region" stands for a different region within the same chromosome and "patch" refers to HG987 patch. The number of variants confirmed in the last release of the human genome reference is also reported.

These results indicate that a sequence very similar to the last portion of *MAP2K3* was included in the HG987_PATCH added in the GRCh37 toplevel release. This sequence is indeed absent in the GRCh37 primary assembly. This patch was inserted in the GRCh38 release and its coordinates correspond to the new region in the chromosome 17 where reads aligned.

We suggest a similar solution also for exons 3, 4 and 5. In fact, we saw two groups of reads, one carrying all the variants and the other any of them, but they both aligned only to the target gene. Of the 51 unbalanced variants localized in the analyzed exons of *MAP2K3*, 26 variants were confirmed using GRCh38 for variant calling. Of these, 25 were localized in exons 3, 4 and 5 and only one variant was localized in exon 12 - it should be pointed out that this variant was identified in only one sample using GRCh38, thus indicating a private variant. We can conclude that the duplication inserted in chromosome 17 with the HG987_PATCH resolution could be extended also to include the first three exons. In fact, a BLAST search confirmed that the duplication span only exons from 8 to 12.

These findings are supported also by the physical coverage analysis of mate pair reads. As shown in the Figure S2, the physical coverage increases for the first part of the gene, while it decreases starting from exon 9 which actually is the last part of the gene, known to be duplicated in GRCh38.

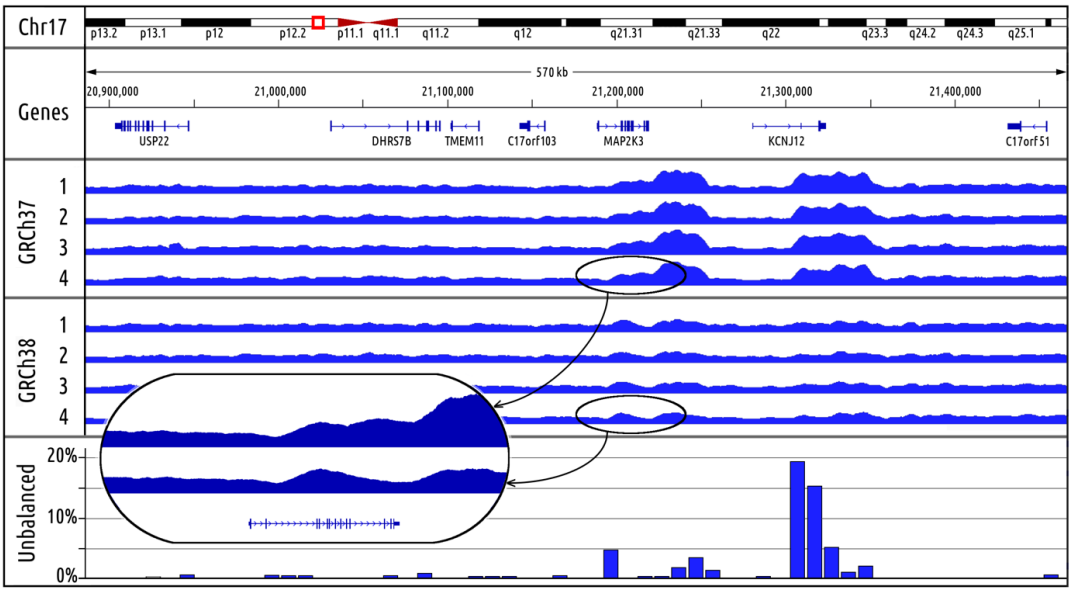

Figure S2. **Physical coverage profiles of a 570 kb region of chromosome 17.** The set of mate pairs from "The genome in a bottle" project was aligned on GRCh37 (3.1 Gb) and on toplevel GRCh38 (48.7 Gb); numbers 1, 2 refer to two Ashkenazi individuals, whereas 3 and 4 refer to two Chinese individuals. The frame at the bottom shows the percentage of variants with unbalanced heterozygosity. In the box the region of MAP2K3 is enlarged to show the reduction of physical coverage starting from exon 9 in the toplevel GRCh38 reference.

Supplementary File S1 references

de Ligt J, Willemsen MH, van Bon BWM, Kleefstra T, Yntema HG, Kroes T, Vulto-van Silfhout AT, Koolen DA, de Vries P, Gilissen C, *et al.* 2012. Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *N Engl J Med* 367: 1921–1929.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, *et al.* 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536: 285–291.

Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30: 2843–2851.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29: 308–311.

Tan A, Abecasis GR, Kang HM. 2015. Unified representation of genetic variants. *Bioinforma Oxf Engl* 31: 2202–2204.

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, *et al.* 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* 3: 160025.