Analysis of the pattern and trend of human genomic variations in the form of single nucleotide polymorphisms (SNPs) and small insertions and deletions (INDELs)

Vinay Kumar Chundi

Biotechnology

Submitted in partial fulfillment
of the requirements for the degree of

Master of Science

Faculty of Mathematics & Science, Brock University
St Catharines, Ontario

Abstract

Single nucleotide polymorphisms (SNPs) and small insertions/deletions (INDELs) are the most common genetic variations in the human genome. They have been shown to associate with phenotype variation including genetic disease. Based on data in a recent version of the NCBI dbSNP database (Build 150), there are 305,651,992 SNPs and 19,177,943 INDELs, and together as all small sequence variants, they represent approximately 11% of the human reference genome sequences. In this study, we aimed first to examine the characteristics of SNPs and INDELs based on their location and variation type. We then identified the ancestral alleles for these variants and examined the patterns of variation from the ancestral state. Our results show that the occurrence of small variants averages at 104 SNPs/kb and 6.5 INDELs/kb for a total of ~11% of the genome. Chromosome 16 and 21 represent the least and most conserved autosomes, respectively, while the sex chromosomes are shown to have a much lower density of SNPs and INDELs being more than 30% lower in the X chromosome and more than 85% lower in the Y chromosome. By gene context, SNPs are biased towards genic regions and INDELs are biased towards intergenic regions, and further, INDELs are biased towards protein-coding genes and intron regions within the genic regions and SNPs are biased towards non-coding genes in the genic regions. Within the coding regions, SNPs and INDELs are biased towards missense and frameshift variations, respectively. Some of the biases were due to biased sources of the variation data targeting at genic regions, while the bias towards intron regions is due to selection pressure. Further, genes with the highest level of variation showed enrichment in functions related to environmental sensing and immune responses, while those with least variation associate with critical processes such as mRNA splicing and processing. Through a comparative genomics

approach, we determined the ancestral state for most of these variants and our results indicate that ~0.79% of the genome has been subject to SNP and INDEL variation since the last common human ancestor.

Our study represents the first comprehensive data analysis of human variation in SNPs and INDELs and the determination of their ancestral state, providing useful resources for human genetics study and new insights into human evolution.

Acknowledgment

I am very grateful to my supervisor, Dr. Ping Liang for the opportunity to work in his lab. His encouragement, patience, guidance, and support have helped me understand the subject matter and gave me the opportunity to apply and broaden my skill set. I would like to thank my committee members, Dr. Charles Després and Dr. Jeffrey Atkinson for their valuable time and guidance in the process. I would like to acknowledge all my colleagues in the lab for helping me through my masters. Finally, I would like to thank my family and friends for their continued support in helping me achieve my goals.

Abbreviations

Base pairs (bp)

BLAST-like alignment tool (BLAT)

Coding DNA sequence (CDS)

Copy number variation (CNV)

Deoxyribonucleic acid (DNA)

Gene ontology (GO)

Kilo base pair (kb)

Structural variants (SV)

Single nucleotide polymorphism (SNP)

Small insertions/deletions (INDELs)

Mobile elements (MEs)

Next generation sequencing (NGS)

Ribonucleic acid (RNA)

Untranslated Region (UTR)

Table of Contents

Chapter 1: Introduction and related literature review

1.1.1 Introduction to human genetic variation

Variation in genetics refers to permanent alteration of a nucleotide sequence in a genome or extrachromosomal DNA of an organism (Logofet & Svirezhev, 1980). Genetic variations are widely seen among individuals between and within populations around the world. They are the basis of phenotypic variations, including disease states. There are many sources of variations, including DNA replication errors, DNA damage, recombination during meiosis, and DNA transpositions (Kidwell & Lisch, 2002). Based on the type and size, DNA variations can be divided into small sequence changes and large structural variants (SVs). Small sequence changes include single nucleotide polymorphism (SNPs), also known as single nucleotide variants (SNVs) and small insertions and deletions (INDELs) (≤50 bp in length). SVs can be divided into large insertions/deletions, copy number variations (CNVs), inversions, translocations, and mobile element insertions (MEIs) (Kidwell & Lisch, 2002).

The physical variation we observe among individuals is substantially contributed by genetic variation (Rotimi & Jorde, 2010). SNPs and INDELs are the most important components of human genetic variations (Bhangale, Rieder, Livingston, & Nickerson, 2005; Rotimi & Jorde, 2010). SNPs are single nucleotide base pairs that vary among individual DNA sequences (Rotimi & Jorde, 2010). For example, an individual will have a C-G at a given specific location in the haploid DNA sequence, whereas some other individual may have A-T at the same location. *Homo sapiens* (humans) have about half the number of genomic variation when compared to the genetic variations in the central African chimpanzees and gorillas (Bhangale et al., 2005; Yu, Jensen-Seaman, Chemnick, Ryder, & Li, 2004) and one-

tenth when compared to the fruit fly (*Drosophila pseudoobscura*) (W. H. Li & Sadler, 1991). The fact that humans have far less genomic variants when compared to some other primates, despite the much larger population size in humans suggests the occurrences of prehistoric bottlenecks in human evolution (Amos & Hoffman, 2010).

In humans, millions of SNPs have been identified, and previous surveys show that at the single nucleotide level on average any given two unrelated individuals differ at around 1 in 1000 bp (Sachidanandam et al., 2001). Common SNPs (i.e., those SNPs for which the presence of the lesser common allele is greater than 5%) are seen to be shared among populations around the world (Xing, Watkins, et al., 2009). This observed commonality shows continued gene flow and migration in human populations historically, in addition to our common origin (Rotimi & Jorde, 2010). Studies have shown that the majority of the genetic variations (around 85 to 90%) can be found within any human population (e.g., samples from Great Britain and from South Africa) (J. Z. Li et al., 2008). Occasionally, a SNP is present in one population but absent in another, and this is sometimes a result of a recent emergence of a variant which did not yet get enough time to spread to other populations (Merryweather-Clarke, Pointon, Shearman, & Robson, 1997). In such cases, this difference in prevalence might have been caused by natural selection. For example, hereditary lactase persistence is prevalent in African and European populations compared to other populations in the world, where consuming milk beyond childhood stage had a selective advantage (Tishkoff et al., 2007).

Many studies have shown SNPs in both the coding and non-coding regions of the genome to have a functional impact on the genes and also may associate with genetic diseases (Raitio et al., 2001). Most SNPs are located in the non-genic

locations of the DNA, which can be made use of as a biological marker for research due to less selection (Syvanen, 2005).

1. 1. 2 Functional impacts of SNPs and INDELs

When a SNP occurs in a gene or in a regulatory region near a gene, it might play a direct role in disease by altering the gene function. Most SNPs, however, do not play a significant role in altering health or development while other SNPs have been shown by studies to be directly involved or as an active participant to cause diseases (Syvanen, 2005). They influence a wide range of human diseases such as sickle-cell anaemia, cystic fibrosis and β-thalassemia (Ingram, 1956; Swersky, Chang, Wisoff, & Gorvoy, 1979). SNPs do not always function individually, rather they work together with other SNPs to manifest a disease condition as seen in osteoporosis (Singh, Singh, Juneja, Singh, & Kaur, 2011). Non-coding region SNPs (such as those in the UTRs) have been shown to manifest in high risk of cancer(Cheetham, Gruhl, Mattick, & Dinger, 2013) and also may affect the mRNA structure and susceptibility of disease(Wapinski & Chang, 2011). SNPs have also been showed to cause loss of function in proteins related to neurological disorders such as those causing amyotrophic lateral sclerosis (ALS) (Kamaraj, Rajendran, Sethumadhavan, Kumar, & Purohit, 2015).

Short insertions and deletions (INDELs) are the second most abundant variations known to contribute to human genetic variation and their influence on human phenotypes (Bhangale et al., 2005). When compared to SNPs and other structural variants (SVs), the origins and functional effects of INDELs are poorly understood at the population level (Montgomery et al., 2013). This lack of understanding is mainly due to the difficulties faced in discovery and genotyping of INDELs by methods other than direct sequencing (Montgomery et al., 2013). INDELs

occur at a lower density in genic regions when compared to the other genomic regions (MacArthur & Tyler-Smith, 2010).

INDELs have also been shown to be associated with disease in humans. For example, a frameshift mutation is responsible for Bloom Syndrome, a rare autosomal recessive disorder which is characterized by short stature, in Jewish or Japanese population (Kaneko, Tahara, & Matsuo, 1996). Both insertions and deletions can be used as genetic markers in natural populations for phylogenetic studies (Väli, Brandström, Johansson, & Ellegren, 2008). For example, short tandem repeats (STRs) have been used as markers for DNA fingerprinting in forensic science and paternity testing (Zamir, Springer, & Glattstein, 2015). INDELs have been shown to have functional effects while present in the CDS region of the genome. They are different from a point mutations, an INDEL insets or deletes nucleotides from a sequence, leading to changes in sequence length (Hill, Wang, Farwell, & Sommer, 2003). In the CDS region, unless the length of a given INDEL is a multiple of 3, it results in a functional effect known as the frameshift mutation. A frameshift mutation in the CDS region can result in the formation of a premature stop codon, leading to a truncated protein product (Wetterbom, Sevov, Cavelier, & Bergström, 2006).

1.1.3 Structural variants in the human genome

The term structural variation (SV) refers to large scale structural differences in the genomic DNA that are inherited and polymorphic in a species and involves DNA segments larger than 1kb (Redon et al., 2006). Since the development of high throughput DNA sequencing technologies with increased resolution, SVs have been found to be ubiquitous in all human DNA and often linked to disease association (Lee & Scherer, 2010). SVs can be classified into balanced and unbalanced based on whether alteration of sequence length is involved. Balanced SVs do not change the

total length of the nucleotide sequence, whereas the unbalanced SVs result in a change in a total length of the nucleotide sequence. The unbalanced SVs present in the human genome include large insertions/deletions, copy number variations (CNVs), and these events encompass an order of magnitudes of more nucleotides than SNPs and INDELs (Pang et al., 2010). The DNA variations that are balanced in nature include inversions and translocations and are less common in the human genome but can play an important role in chromosomal evolution and disease (Redon et al., 2006).

Copy number variations (CNVs) are defined as stretches of DNA larger than 1000 base pairs (bp), which are normally found only once on each chromosome in a person, but for some individuals, these are found as duplicates or triplicates or even higher copy number. In other words, there is a variation in the number of copies of the section of DNA from one individual to another (Choy, Setlur, Lee, & Lau, 2010). CNVs have several distinct features that support their role in disease pathogenesis. First, these SVs often encompass more than one gene and collectively include a higher number of nucleotide base pairs than SNPs (Redon et al., 2006). Due to spanning several thousand bases, CNVs often encompass DNA sequences that are functional in nature. Secondly, CNVs are enriched towards environmental sensor genes. They are genes that are significant to perceiving and interacting successfully with the changing environment (Sebat et al., 2004). This includes olfactory receptors enrichment, inflammatory and immune response genes, cell signalling molecules, structural proteins and ion channels (Tuzun et al., 2005). Similar to the other genetic variations, purifying and adaptive natural selective pressures appear to have influenced the distribution of the selective CNVs (Nguyen, Webber, & Ponting, 2006). Studies have shown their association with neuropsychiatric conditions such as autism and schizophrenia (Cook & Scherer, 2008). A recent comparison on the

relative impact of SNPs and CNVs on gene expression showed that a large portion, approximately 18% of gene expression variability was caused by known CNVs (Stranger et al., 2007). Around 53% of genes whose expression was influenced by CNVs had the corresponding CNV outside the actual gene, suggesting that many CNVs could have an effect on important regulatory sequences which are situated away from the target gene (Stranger et al., 2007).

Inversions are chromosomal rearrangements where a segment of a chromosome is reversed from end to end. Paracentric and pericentric are the two types of inversions that occur in the chromosomes, through breakage and rearrangement within a chromosome (SJÖDIN, 1971). Pericentric inversions are inversions where the centromere is included and includes a breakpoint in each arm of the chromosome (Muthuvel, Ravindran, Chander, & Subbian, 2016; SJÖDIN, 1971). A paracentric inversion, unlike pericentric inversion, does not include the centrosome and the breaks occur on the same arm (Muthuvel et al., 2016; SJÖDIN, 1971). Inversions have shown to be associated with diseases. The inversion polymorphism in chromosome 17q21.31 with approximately 900 kb in the population with European ancestry (de Jong et al., 2012)results in two divergent microtubule-associated protein tau (MAPT) haplotypes (H1 and H2). Haplotype 1 has been shown to associate with progressive corticobasal degeneration, supranuclear palsy, Parkinson's disease and Alzheimer's disease. Haplotype 2, on the other hand, is linked to deletion events associated with 17q21.31 microdeletion syndrome which is a disease characterized by learning disabilities (de Jong et al., 2012). Another study showed the association of pericentric inversion of chromosome 9 p12q13 found in children with dysmorphic features and congenital anomalies (Rao, Kerketta, Korgaonkar, & Ghosh, 2009).

Segmental duplication is a segment of DNA that is greater than 1 kb in size. Duplications have DNA segment with two or more copies per haploid genome. The duplicate segments have a sequence identity greater than 90%. Duplications can often be variable in copy number and hence are also CNVs (Redon et al., 2006). Chromosomal translocation is a chromosomal segment that is moved from one position to another, either within the same chromosome or to another chromosome (Redon et al., 2006). Duplications have been seen to associate with diseases, such as the chromosome Xp11.23-p11.22 duplication syndrome that is characterized by borderline to severe mental retardation and speech delay (Giorda et al., 2009).

Mobile elements (MEs), which are also known as transposable elements (TEs) are DNA elements that have the ability to move in the genome using either retrotransposition or by self-splicing (Xing, Zhang, et al., 2009). In the human genome, MEs account for 47-49% of the genome (Hancks & Kazazian, 2012; Tang, Mun, Joshi, Han, & Liang, 2018). Based on the method of transposition, they can be classified into two types, DNA transposons and retrotransposons. A DNA intermediate is used for transposition in the case of DNA transposons. DNA transposons are estimated to account for approximately 3 percent of the human genome (Cordaux & Batzer, 2009). Retrotransposons duplicate with the help of an RNA intermediate that is reverse transcribed before insertion of the DNA molecule into a new genomic location. They account for ~47% of the human genome (Cordaux & Batzer, 2009). There are two classes of retrotransposons, namely long terminal repeats (LTRs) and non-long terminal repeats (non-LTRs). The LTR retrotransposons have long terminal repeats (LTR). Endogenous retroviruses (ERVs) is a sub-category of LTR, which account for approximately 8% of the human genome. They comprise of two long terminal repeats (LTRs), ranging from 300–1,200 bp in length. The non-

LTRs make up for one third of the human genome and represents the majority of MEs (Cordaux & Batzer, 2009) in the human genome. Non-LTRs are of three types, long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs) and SINE-VNTR-Alu (SVAs). LINEs make up for approximately 17 percent of the human genome and are mainly represented by LINE 1 elements (L1s), while SINEs contribute to ~13% of the genome and are mainly represented by Alus (Cordaux & Batzer, 2009). Recent and ongoing ME transposition by certain members of retrotransposons, most notably, L1HS, AluYa5, AluYb8, AluYb9, SVA_E, and SVA_F, has led a to generation of a larger number of polymorphic MEs by having presence in some but not all human individuals. (Handsaker, Korn, Nemesh, & McCarroll, 2011; Xing, Zhang, et al., 2009).

Several studies have shown the connection between ME polymorphism and disease phenotypes. For example, the SVA insertion in the B4GALT1 gene results in the down-regulation of the gene. This results in Crohn's disease and systemic lupus erythematosus (Wang & Jordan, 2018). Other diseases that are caused by MEs range from haemophilia, muscular dystrophy and prostate cancer (J. M. Chen, Stenson, Cooper, & Férec, 2005; Hancks & Kazazian, 2012). Alu and L1 elements that make up for ~0.3% of the human diseases and are potential causative candidates for various health conditions, including obesity, multiple sclerosis, leukemia, psoriasis and breast cancer (Payer et al., 2017).

1.2 Major sources of human genetic variation data

1.2.1 Database SNP (dbSNP), history, and its current data

dbSNP was first introduced in December 1988 to address the need for a catalog of genomic variation which would facilitate the scientific efforts in gene association study, evolutionary biology and gene mapping (Coordinators et al., 2015).

Since dbSNP was developed long before the availability of the human reference genome, the initial human variants submitted to dbSNP were defined as a variant sequence in the context of flanking sequence, with often very little supporting evidence or validation of the data (Coordinators et al., 2015). With the advancement in sequencing technologies, dbSNP has grown at a fast pace and now includes validated data for over 300 organisms. The submissions include multiple independent submissions, genotype data, frequency data, and allele observations. dbSNP now also accepts clinical assertions for new and existing variants (Coordinators et al., 2015). The clinical assertion data must first be accepted by ClinVar database before being incorporated into dbSNP with an rs (refSNP) number (Landrum et al., 2016). The dbSNP data also carry information such as minor allele frequency, potential false positive status, and asserted allele origin.

dbSNP archives a collection of common and rare genetic variants including SNPs, INDELs, and multi-nucleotide polymorphism and INDELs. A unique variant accession identifier known as RS number or RSID is assigned to each dbSNP variant entry. The RSID is associated with aggregate information such as the associated gene, allele frequency and functional consequences (Wei et al., 2018). The RSID usage to refer a genomic variant in the dbSNP database serves several advantages, such as (i) the dbSNP record is updated on a regular basis, with accurate and precise locations and aggregated information from multiple submissions; (ii) it provides a stable ID as an unambiguous variant identifier in the publication; (iii) it adds to the convenience for sharing data and integrating with other data sources (Wei et al., 2018).

All dbSNP entries with the RSID are then grouped into a built. The most recent built released by dbSNP is build 151 (for human variants) that contains

335,215,764 new entries from the previous build 150, making it a total of 660,773,127 total entries (dbSNP build 151., http://www.ncbi.nlm.nih.gov/SNP/).

1.2.2 Variants data collected prior to large scale high throughput studies.

The first major breakthrough that helped the progress of DNA sequencing came with the development of Sanger's 'chain-termination' technique (Sanger, Nicklen, & Coulson, 1977). This technique makes use of deoxyribonucleotides (dNTPs) which are monomers of DNA strand molecules and modified dideoxynucleotides (ddNTPs) that terminate the DNA strand elongation. The ddNTPs lack the 3′-OH group that is required for the formation of a phosphodiester bond between two nucleotides and this causes the DNA strand extension to stop when a ddNTP is added (Men, Wilson, Siemering, & Forrest, 2008). It has become the most commonly used technology to sequence DNA for many decades until recently.

Sanger sequencing was used to sequence the sequence of interest, after sequencing, the reads generated are then aligned to the known reference genome sequence (Chaisson, Brinza, & Pevzner, 2009; Pop & Salzberg, 2008). Variants are then detected based on comparison to the reference genome. There are however limitations to the alignment approach, such as placing reads in the repetitive regions of the reference genome or in a corresponding region that may not exist in the reference sequence (Frazer, Murray, Schork, & Topol, 2009). Targeted Sanger sequencing was the only approach for identifying genomic variants before the availability of high throughput technologies, including microarray and next generations of sequencing (NGS). Due to its low throughput, despite its long span in use, Sanger sequencing contributed a small portion of variants but with high confidence among the collection of genomic variants we have today.

1.2.3 Importance of human genome project (HGP) and the reference genome in studying human genetic variations.

A human reference genome is essential to understand the blue print of human DNA. This enables researchers to learn more about the functions of genes and proteins, which can serve as critical information for the advancement in the fields of life science, biotechnology, and medicine. The human genome project (HGP) was designed to achieve these goals (Lander et al., 2001) by determining the nucleotide base pairs sequence that makes up human DNA, along with mapping and identifying all the genes present in the human genome (D. W. Collins & Jukes, 1994). The reference genome helps identify variants in a DNA sequence. The DNA of interest is sequenced and aligned to the human reference genome. A variant call file is then created by identifying where the aligned reads differ from the reference genome.

The human genome comprises 22 autosomal chromosomes (1-22), 2 sex chromosomes (X and Y) and mitochondrial DNA (mtDNA). Depending on the location in the human genome, the variant alleles have different features. In the diploid chromosome, out of the two alleles represented at any genomic position, one is inherited from each parent, the mtDNA, however, is maternally inherited. The sex chromosome ploidy differs based on gender. Females have two X chromosomes and males have an X and Y chromosomes. Due to this fact, we cannot see any heterozygous genotypes in the X chromosomes for males and Y chromosome variants for females (Guo et al., 2017). With the improvement in high throughput sequencing technologies, the phenomenon of heteroplasmy (presence of more than one type of organellar genome within a DNA) has been detected in humans mtDNA (Guo et al., 2012, 2013; P. et al., 2016; Ye, Samuels, Clark, & Guo, 2014).

The reference genome assembly is often referred to as a de novo assembly. For the reconstruction of the reference genome, DNA fragments of the target species are subjected to high quality sequencing. Contiguous segments (contig) DNA sequences can be assembled by merging and aligning overlapping sequences. Contig refers to a contiguous length of a genomic sequence in which the order of nucleotide bases is known to a high confidence level. When assembled together, multiple contigs form a scaffold, based on the positional relationship such as paired read information. A DNA scaffold is a portion of the genomic sequence that is composed of contigs but might contain gaps between the contigs (Guo et al., 2017). Several tools have been developed to carry out genome assembly from short reads (Simpson et al., 2009; Zerbino & Birney, 2008) and to close gaps between the scaffolds (Paulino et al., 2015; Pop, Kosack, & Salzberg, 2004). Multiple scaffolds together form a chromosome (Guo et al., 2017).

There are several challenges associated with reconstructing a complete and accurate human reference genome. Repetitive DNA regions such as the telomeres are one of the best known challenges (Moyzis et al., 1988), as they can convolute the consensus sequence considerably. Sequencing sensitivity to variable GC content can result in an uneven representation of the genome (Bentley et al., 2008); this can cause gaps between scaffolds. Scientists have been working to tackle these challenges and gradually improved the human genome.

The draft of the human genome sequences was first published in 2001 (H. G. S. Consortium et al., 2001; Venter et al., 2001). In 2009, the Genome Reference Consortium (GRC) announced the release of the reference genome version GRCh37 which is refereed as HG19 the UCSC (University of California, Santa Cruz) genome browser. The HG19 reference was used extensively in sequencing data analysis for

several years. GRCh38 (also known as HG38) was the 20th and the latest release of the human reference genome by the GRC (Karolchik et al., 2014), representing the most accurate version. The samples were from ethnically diverse donors when compared to the earlier version HG19 (Guo et al., 2017). The sequencing was conducted using gold standard Sanger sequencing, that has the ability to sequence reads that are up to 1000 nucleotides with 10 times more accuracy than the high throughput short read sequencing (Sanger et al., 1977). Compared to the HG19, HG38 altered 8000 nucleotide base pairs, and several misassembled regions were corrected, many gaps were filled in, while some sequence was added for centromeres. Improvement in representing the diversity of the reference was also made by including 261 alternate loci in 178 regions (Guo et al., 2017).

The reference genome is composed of 3 billion nucleotide base pairs, organized as 23 chromosomes. Genes only make up around 2-3% of the entire DNA sequence whereas ~50% of the sequence is made up of repetitive sequence. The remaining ~48% is noncoding non-repetitive DNA (Miklos & Rubin, 1996). Despite its incompleteness by missing coverage for some heterochromatin regions, the reference human genome holds many benefits in several fields from human evolution to molecular medicine (F. S. Collins & McKusick, 2001).

1.2.4 Hapmap project: Overview, contributions to our current understanding and data for human genetic variants.

When studying the association of phenotype with genotype, we can either follow a direct or indirect approach. The direct approach depends on the availability of a list of functional variants, which are tested for the association of the variants to the trait of interest from a number of phenotypically matched controls and cases (Deloukas & Bentley, 2004). The indirect approach involves testing genomic variants

across the entire genome with the assumption that the causative variants can be in linkage disequilibrium (LD) with one of the tested variants. The term "linkage disequilibrium" refers to the association of non-random alleles at two or more loci in a population. Haplotypes do not occur at the expected frequency when the alleles are at linkage disequilibrium (Deloukas & Bentley, 2004). Understanding the dynamics and characterisation of LD in the genome is necessary for enabling whole genome associated studies. The Hapmap project was launched in 2002 with the objective of constructing a genome-wide association map of LD and common haplotypes in four populations.

The international Hapmap consortium took the effort to catalog all common variants across the genome (variants that have a minor allele frequency (MAF) of at least 5% in more than one ethnic group) in order to be able to construct the haplotype map. The first two phases of the HapMap project included samples from four different population, namely 30 Yoruban (Nigeria) trios (family of three individuals), 30 trios in Utah families (North European descent), 48 unrelated Chan Chinese and 48 unrelated Japanese (Manolio & Collins, 2009). In phase 1, the HapMap consortium targeted to prioritize coding SNPs and genotyped 1 million SNPs. From phase 2, HapMap consortium prioritized non-synonymous SNPs in coding regions and genotyped 3 million SNPs. In phase 3 of the HapMap project samples from African ancestry in southwest USA (ASW), Chinese in metropolitan Danver, Colorado (CHD), Gujarati Indians in Houston, Texas (GIH), Luhya in Webuye, Kenya (LWK), Mexican ancestry in Los Angeles, California (MXL), Maasai in Kinyawa, Kenya (MKK) and Toscani in Italia (TSI) were added. Phase 3 of the HapMap project identified and released 1.4 million SNPs (Deloukas & Bentley, 2004).

For the scope of the HapMap project, the variants identified from the project are limited to common variants from a limited number of human populations.

1.2.5 1000 Genome Project: Overview, contributions to our current understanding and data for human genetic variants.

The 1000 Genome Project was launched in 2008 with the aim of establishing a deep catalogue of human genetic variations which would serve as a baseline for further research to understand the relationship between genotype and phenotype (Clarke, L., *et al*., 2012). Towards the conclusion of the data generation phase of the project, 92 terabases of whole genome and whole exome sequences were amassed. Sequencing centres submitted raw data to the sequence read archive as they were generated. Through a coordinated process, the data were assessed for quality, aligned to the human reference genome and a large number new sequence variants were identified (Clarke, L., *et al*., 2012). The pilot phase of the project included 180 individuals, phase 1 included 1092, followed by phase 2 with 1700 samples and phase 3 that included 2504 individuals covering 26 well recognized human populations from 5 continents. All the data collected in phase 2 and phase 3 were collectively published together in 2015 (1000 Genomes Project Consortium, 2012; 1000 Genomes Project Consortium et al., 2015, 2010).

The results published by the project collectively contained 88.3 million variants with 84.4 million as SNPs, 3.4 million as INDELs and around 60, 000 as SVs (1000 Genomes Project Consortium et al., 2015). Through the use of new algorithms for variant discovery, the phase 3 release also contains 475, 000 multi-allelic SNVs and INDELs. The variants identified by the project were deposited into dbSNP, contributing to 61% of the ~131 million entries included in dbSNP build 141 (Clarke et al., 2012).

1.2.6 Exome sequencing

Exome sequencing targets the exon regions instead of the entire genome. Even though exome sequencing does not consider the impact of non-coding regions, it is considered a well justified strategy for identifying rare variants associated with Mendelian phenotypes (Bamshad et al., 2011). One of the challenges for applying exome sequencing has been how to best define the set of targets that make up the exome. There is considerable uncertainty when it comes to selecting the sequence of the human genome that are truly protein coding. Initial efforts of exome sequencing were on the conservative side (for example, targeting the high confidence subset of a gene that is identified by the Consensus Coding Sequence (CCDS) Project). All the RefSeq collection and a large number of hypothetical proteins were (at a minimum) targeted by the commercial kits; however, this comes with certain limitations. First, the incomplete knowledge about all truly protein coding exons in the genome, hence making the capture probes (single stranded DNA molecules) only target exons that have been identified so far. Second, there is a considerable variation on the efficiency of capture probes in capturing the target sequences. Third, not all templates are sequenced with equal efficiency and hence not all sequences can be aligned with sufficient coverage to the reference genome to allow base calling. In fact, the effective coverage (for example, 50x coverage) of exons that can be achieved using commercial kits varies considerably. Finally, there is an issue on whether sequences other than exons should be targeted (example, microRNAs (miRNAs), promoters and ultra-conserved elements). Despite these caveats, exome sequencing has been shown to be a powerful new strategy for finding the cause of suspected or known Mendelian disorders for which the discovery of a genetic basis is unknown (Bamshad et al., 2011).

The Exome Aggregation Consortium project is a large scale exome sequencing project (ExAc; http://exac.broadinstitute.org), in which a total of 60,706 exomes were filtered from over 91,000 exomes. The project applied further genotype quality filters and defined a subset of 7,404,909 high-quality variants from 10,195,872 candidate variants that included 317,381 INDELs. This corresponds to one variant for every 8 base pairs (bp) of the exomes. A majority of the identified variants were low-frequency variants that were absent from previous high quality variants (Lek et al., 2016). 72% of the identified variants were absent in both 1000 GP as well as the ESP (Exome sequencing project) data sets that contained variants identified from 6, 503 exomes (Fu et al., 2013).

1.2.7 Other large-scale personal genome projects.

The association of genomic variation with disease and drug response along with the improvement in DNA sequencing technologies has given an optimistic impact on genomic medicine and personal genomics. Personal genomics, also known as consumer genetics is a branch of genomics that is concerned with the sequencing, analysis, and interpretation of an individual's genome (McGuire, Cho, McGuire, & Caulfield, 2007). Data obtained through personal genomics is a significant source of human variant data (McGuire et al., 2007). The first most well-known personal genome project that contributed to understanding human variant data is the research comparing Craig Venter's genome with the reference genome.

The project presented a complete genome sequence of an individual human. Approximately 32 million DNA fragments were used for the assembly, using Sanger sequencing (Levy et al., 2007). The DNA fragments were assembled into 4, 528 scaffolds which comprised 2, 810 million bases (Mb) of contiguous sequence. The coverage was approximately 7.5 fold coverage for all regions. After assembling the

genome, it was compared to the human reference genome and the comparison revealed over 4.1 million genomic variants, out of which 1,288,319 were novel. The data included 3,213,401 SNPs, 851,575 INDELs, 53,823 block substitutions as MEs, 90 inversions and 62 copy-number variations. The non SNP variations accounted for 22% of the total variants in number but they covered 74% of all variants in length (Levy et al., 2007). This suggests that the diploid genomic difference is mostly defined by non-SNP genetic alterations.

Another study completed about the same time was the sequencing of Jim Watson using the 454 NGS platform (Wheeler et al., 2008). Comparing this personal genome to the human reference genome, led to the identification of 3.3 million SNPs, of which 10,654 were present in the protein coding genes with many being non-synonymous. This study also identified small INDELs and CNVs ranged from 26,000 to 1.5 million base pairs (Wheeler et al., 2008).

The deep sequencing of 10,000 human genomes is one of the biggest personal genomic projects that have contributed to the identification of human genetic variants (Telenti et al., 2016). A total of 10,545 human genomes, covering four major populations (African, European, Asian and native American) were sequenced in the study at 30 to 40x coverage, emphasising quality and novel variant identification and sequence discovery. Confidently, 85% of the individual's human genome was sequenced. For all other samples, only the exons were sequenced. This study identified a total of 150 million SNPs, concluding that each newly sequenced genome contributes to an average of 8,579 novel genomic variants and 0.7 Mb of sequence that is not found in the most recent version of the reference genome (GRCh38) (Telenti et al., 2016). Another good example of a large scale project is the personal genome project (PGP). PGP was designed as a long term, large cohort study that aims

at sequencing 100, 000 volunteers (Church, 2006). Since 2005, the PGP has not published data summarizing their findings. In 2010 they revealed their intentions to use whole genome sequencing over exome sequencing for their future samples due to the fall in cost to sequence the whole genome (Lunshof et al., 2010). As of November 2017, the study has recruited more than 10,000 volunteers.

1.3 Differential functional impact of genetic variations in different genomic regions

The human genome consists of coding and non-coding genic regions and intergenic regions. Upon completing the first human genome sequencing project, the challenge was to identify the structures of all genes and functional elements in the genome. It was quickly identified that nearly 98.5% of the approximately 3.3 billion nucleotides in the human genome do not code for proteins, while the remaining 1.5% are genic sequences that code for a protein (Lander et al., 2001)

A gene requires many parts to function. First, a gene requires a promoter sequence which is identified and is attached by RNA polymerase and transcription factors for initiating the process of transcription. A consensus sequence like the TATA box helps recognize the promoter region (Sprouse et al., 2008). Thus, sequence variations in the promoter regions may alter the identification and binding of transcription factors, which in turn can have an influence on gene expression. For example, a SNP in the promoter region of the HLA-G gene, that inhibits maternal anti-fetal immune response, is associated with increased risk for miscarriage (Ober et al., 2003).

An added layer of regulation can occur for protein coding genes after mRNA processing to prepare it for protein translation. The final protein product is coded by

the sequence present between the start and stop codons. The flanking untranslated regions (UTRs) contain regulatory sequences (Guhaniyogi & Brewer, 2001). The 3'UTR has a terminator sequence that makes up for the end point of transcription and release RNA polymerase (Kuehner, Pearson, & Moore, 2011). The ribosome binds to the 5'UTR that translates the protein-coding region to a string of amino acids that fold to form a protein product (Palazzo & Lee, 2015).

The coding sequences (CDS) of a gene are the regions of the DNA sequence that codes for a protein starting from a start codon and terminating at the 3' stop codon. Only a subset of open reading frames are usually translated into proteins and hence making the identification of CDS in a genome difficult when compared to identifying the open reading frames within a DNA (Furuno et al., 2003).

SNPs present in CDS regions can cause either a synonymous or non-synonymous substitution. A synonymous substitution is when the resulting SNP codes for the same amino acid. Synonymous variants have no significant functional impact. In exceptional cases, these substitutions can affect protein function in other ways. For example, a study finds that a silent mutation in the multidrug resistance gene 1 (MDR1) which functions to expel drugs from the cell slows down translation resulting in the peptide chain folding in a different and unusual conformation. This resulted in the loss of function of the mutant pump (Kimchi-Sarfaty et al., 2007).

A non-synonymous substitution is a type of substitution where it leads to code for a different amino acid. The two types of non-synonymous substitutions are missense and nonsense substitutions. A missense substitution is a change in a single base resulting in a change in amino acid of protein and may lead to disease causing malfunction. For example, a missense mutation (P56S) in the protein vesicle

associated membrane protein (VAPB) results in protein aggregation and loss of function, which leads to amyotrophic lateral sclerosis (ALS) (Vinay Kumar, Kumar, Swetha, Ramaiah, & Anbarasu, 2014). Nonsense substitution is a point mutation which results in a premature stop codon, resulting in a truncated, incomplete and usually non-functional protein sequence. An example of nonsense mutation is cystic fibrosis caused by the cystic fibrosis transmembrane conductance regulator gene mutation G542X (Cordovado et al., 2012).

Introns are nucleotide sequences within a gene that are removed by RNA splicing during maturation of RNA products. SNPs and INDELs in the intron regions can have functional impacts and disease causing abilities by alternating RNA splicing leading to abnormal transcripts. Mutations in the splice sites have been shown to associate with diseases such as cystic fibrosis (Pagani & Baralle, 2004).

1.4 Limitations of the current reference genome and dbSNP data

Despite the undoubted value of the human reference genome, it still has limitations. The first limitation is that it is based on data collected from the HGP which focused on a small number of human DNA samples from a number of populations with the sample used to construct the majority of the genome being a Caucasian and hence it does not truly represent human diversity (Dolgin, 2009). The second limitation is that the most recent version (GRCh38) still has 603 ‘gaps’. The gaps represent those portions of the genome that are particularly difficult to sequence (Rosenfeld, Mason, & Smith, 2012). However, the gaps in the reference genome are expected to be gradually filled as scientists learn more ways of sequencing these regions. For example, the most recent build of the human reference genome was the first to include centromere sequences which are highly repetitive regions that are millions of base pairs long and known to have a structural role in the cell (Schneider

et al., 2017). Lastly, the reference genome is referred to as mosaic because it was built from multiple donors and is not representative of one complete set of chromosomes. This can cause issues when matching new sequences. Also, the human reference genome represents only a haploid version of a diploid genome (Guo et al., 2017).

There are several SNP public databases available for use. The largest is the dbSNP database, currently containing millions of variants (Aerts, Wetzels, Cohen, & Aerssens, 2002). While the dbSNP database has been continuously expanding, the completeness and quality of the submitted SNP data remain important. A recent evaluation of the dbSNP data has shown that around 6-12% of SNPs could not be validated (Reich, Gabriel, & Altshuler, 2003). These un-validated SNPs represent population-specific SNPs, rare variants and sequencing errors. dbSNP existed long before the completion of the HGP and the data contains all the variants submitted before HGP, which lack validation or has very less validating evidence (Schneider et al., 2017). The dbNSP data lack information regarding the ancestral state of the variants. With the data not reflecting the ancestral state, an identified variant is not always a new variant. There is always the possibility of the reference allele being the new variant. This makes it difficult to predict the functional impact of variants.

1.5 Research objectives.

This project was designed two major objectives: 1) to examine the distribution patterns of human genetic variants, focusing on SNPs and INDELs; 2) to determine the ancestral state of all SNP and INDEL variants through a comparative genomics approach. Knowing the ancestral state of all the variants can not only help better predict the functional impact of variants but also allow us to examine the pattern of human variations since the last common ancestor for all human populations.

Chapter 2: Materials and Methods

2.1 dbSNP dataset

The dataset of SNPs and INDELs were obtained from dbSNP build 150 at NCBI (https://www.ncbi.nlm.nih.gov/snp/) as a set of vcf files for each variant type. The dbSNP built 150 has a total of 305,651,992 SNPs and 19177943 INDELs or a total of 324,829,935 small variants, and they were used as the starting datasets for all analyses included in this study.

2.2 Analysis of genome distribution patterns of SNPs and INDELs

To examine the genome distribution of SNPs and INDELs, a density plot onto the chromosome ideogram for each variant type was made using an in-house perl script. For each chromosome, the average density of SNPs and INDELs were calculated as the number of variants per 1 kb of non-gap sequences. Further, t-test analysis was performed to evaluate the degree of differences of the variant density across chromosomes based on the densities of variants in a series of 1 kb sliding windows using R.

To obtain the functional impact and associated genes for all variants, SnpEff, an open source software tool, is used to annotate the variants (Cingolani et al., 2012). Variant data file in vcf format from dbSNP build 150 was used as input in the SnpEff with the default settings. The annotation output provides the genomic location by gene context, associated gene, and functional impact type, such as synonymous and non-synonymous SNPs (missense and nonsense), and frameshift for INDELs.

2.3 Functional characterization of genes with least and most variation

To obtain the list of genes with the least and most of variation, the density of SNPs and INDELs (variants/kb) for protein-coding genes were calculated based on the CDS regions of the genes, and all genes were sorted based on the variant density from high to low. The top 5% of genes by the number of genes were considered as genes with the most variation, while the bottom 5% of genes were considered as genes with least variation. For SNPs, we also obtained genes with the highest and lowest density of missense variants. These datasets were obtained by overlapping the location of variants and the location of gene coordinates using 'BEDtools' (Quinlan, 2014).

To obtain the functional categories of genes in gene ontology (GO) terms, each of the above gene lists was analysed Enrichr, an online tool for analysing GO term enrichment for a list of genes (E. Y. Chen et al., 2013). For each gene list, the significantly enriched GO terms in each of the GO domains (biological process, cellular component and molecular function) were collected (G. O. Consortium, 2004). The statistically significance of the GO term enrichment was based on a p-value adjusted for multi-test being 0.05 or lower. For those with more than 10 categories of significantly enriched GO terms, only the top 10 were kept.

2.4 Identification of the ancestral alleles for SNPs and INDELs

To identify which of the alleles between the reference and alternate allele represents the ancestral allele for each of the SNP and INDEL loci, we compared the human genomic sequences with the orthologous sequences in five other primate genomes. The 5 non-human primate genomes used in the study were Chimpanzee (*Pan troglodytes*/panTro4), Gorilla (*gorilla gorill*/gorGor4), Bonobo (*Pan paniscus*/panPan2), Orangutan (*Pongo pygmaeus abelii*/ponAbe3) and Gibbon

(*Nomascus leucogenys*/nomLeu3). These primate genome sequences were downloaded from the UCSC Genome Browser Website (http://genome.uscsc.edu). The strings associated with the scientific names of the species represent the genome sequence version identifications used by the UCSC genome browser.

To identify the orthologous sequences for each variant in the other primate genomes, the genomic sequences based the human reference genome (GRCh38) including a 50 bp flanking sequence on each side of a variant were extracted as fasta sequences for both the reference and alternate alleles of each SNP and INDEL, using an in-house perl script. These sequences were then matched to the genome sequences of the other 5 primates using BLAT, a tool designed for performing similarity search for sequences of high similarity (Kent, 2002).

The optional values for certain parameters of BLAT run were changed from the defaults to speed up the process. The minimum score, which is calculated as the number matches minus the mismatches minus gap penalty, was changed from the default value of 30 to 80. The minimum identity was set to 93 from the default of 90. The BLAT outputs in 'psl' format were processed to determine the orthologous position of each variant in the five non-human primate genomes based on the matches with the best BLAT score in each genome. The non-human primate sequences at the corresponding human genomic position were retrieved using an in-house perl script.

In the last step of the above process, a cutoff value for the blat score was set for each of the five other primate genomes to ensure that a sequence match between the human and the other primate genome is likely a true orthologous homology match. The blat cutoff value was set as a value, at and above which a match can be obtained in this genome for 95% of the human variants. This value was obtained for SNPs and

INDELs separately. To obtain these cutoff values, the mean and standard deviation values were calculated based on the best blat scores for all variants. The cutoff values at 95% coverage were then calculated by subtracting two standard deviations to the left of the mean value.

To call the ancestral allele for each human variant, the allele that matches the orthologous sequence in the chimpanzee genome is defined as the ancestral allele, while those with neither allele matching the chimpanzee sequences were removed from further analysis. As the final output of the process, a tab delimited text was generated for SNPs and INDELs, in which five columns of data are used to represent the information of a variant in the human genome as dbSNP ID, chromosome ID, base pair position, reference allele, alternate allele, and ancestral allele, and 3 columns for each of the other primate genomes as chromosome ID, position, and sequence. The data in this file were used for downstream analyses related to ancestral state.

2.5 Computational analysis

All the data processing except for some figure generation were performed on the computer systems on SHARCNET, which is part of the Compute Canada high performance computing facilities (https://www.sharcnet.ca). The BLAT sequence similarity search for more than 300 million human genomic sites, each with the reference and alternate alleles against 5 other primate genome sequences represented a major undertaking of computation in involving thousands of CPU hours, as well as intensive input and output (I/O) operation. It would not be possible without the use of the high performance computing facilities which have large CPU clusters and large file storage capacity.

Chapter 3. Results

3.1 The patterns of genomic variants in the human genome

3.1.1 Distribution patterns of SNPs and INDEL in the human genome

To understand the pattern of genomic variants in the human genome, we examined their distribution in the genome. For this, we first generated a density plot of SNPs on the chromosome ideogram, a standard graphical cytogenetic representation of chromosomes, to visually represent the distribution of SNPs in the genome (Figure 1). Overall, the SNPs are distributed throughout all chromosomes in a more or less similar density except for the two sex chromosomes. However, some regional differences are also seen within the autosomes. For example, a low density of SNPs can be noticed on the non-heterochromatin regions close to the centromere on chromosome 1 and chromosome 9, while a relatively higher density can be noticed towards the end of non-heterochromatin for most chromosomes (Figure 1). A very small number of SNPs are seen in the heterochromatin region of chromosome 13, 14, 21 and 22, more in part of these regions for chromosome 19, likely reflecting the variable level of available sequence, which might be completely lacking in other similar regions. The X and Y chromosomes show much lower SNP density with the Y chromosome being extremely low (Figure 1). In Y chromosome, the SNP density is also very uneven within the chromosome with a few small regions having a much higher density than the rest (Figure 1).

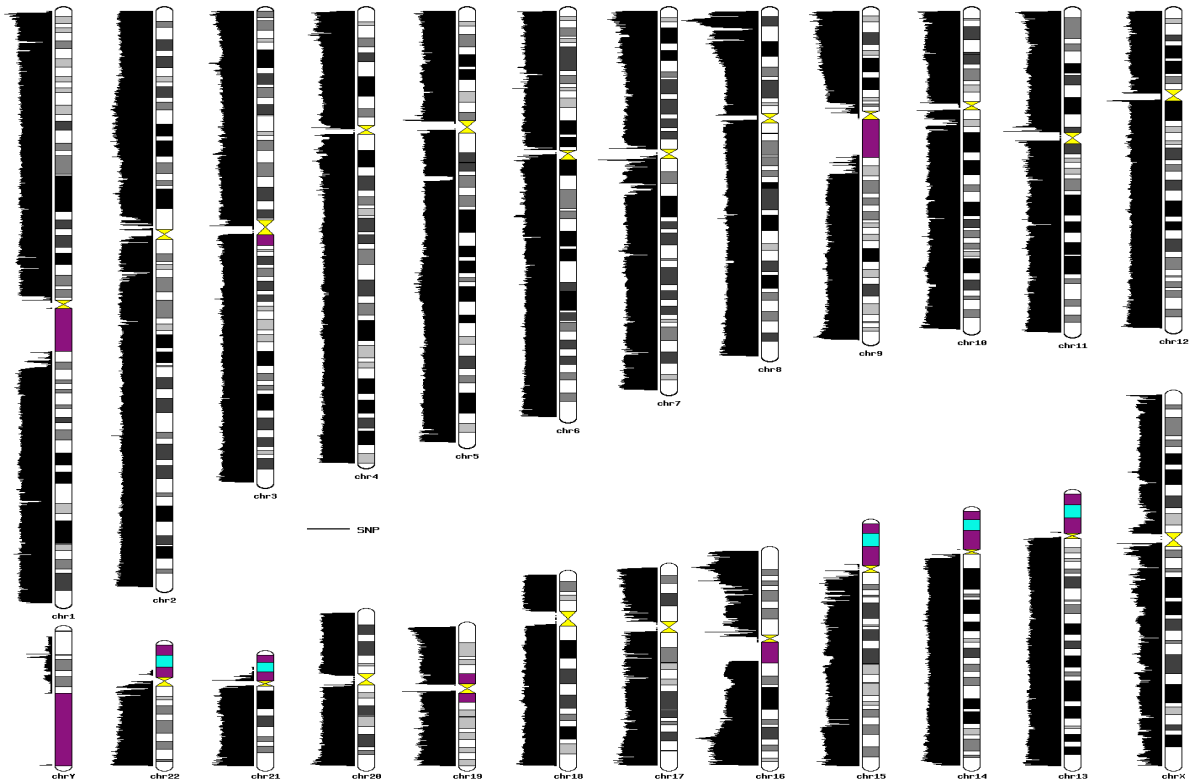

Figure 1. Density plots of single nucleotide polymorphism (SNPs) across the human chromosomes. Line plots along the chromosome ideogram show the density of SNPs in chromosomes (relative to the number of SNPs/700Kb). The regions in colour represent the heterochromatin regions in the genome.

To investigate this SNP density distribution further, we compared the average density (SNPs/kb) among chromosomes. As seen in Table 1, the genome average of SNPs is 104 SNPs/kb of non-gap sequence, which converts to an average frequency of variation at ~10%. The overall density (SNPs/kb) in the autosome is more or less similar across the autosomes. However, chromosome 16 has the highest density (119.5 SNPs/kb) followed by chromosome 8 (114.7) and chromosome 19 (114.2), while chromosome 21 having the lowest SNP density (98.6) followed by chromosome 22 (101.5) and chromosome 18 (101.5) (Table 1, Figure 2).

To determine whether the SNP densities among these autosomal chromosomes are statistically different, SNP densities using 1kb sliding windows were generated,

such that a t-test could be performed. The t-test results show that the SNP density differences between most pairwise comparisons among autosomes are statistically significant, except for some of those among chromosome 2, 3, 4, 5, 6, 7, 9 and 10, 13, 18 and 20 (Table S1).

Table 1: Density of single nucleotide polymorphisms (SNPs) and INDELs across the human chromosomes

chr ID	Sequence length (bp)*	SNP count	SNPs /kb	INDEL count	INDELs /kb	SNPs+ INDELs/kb
chr1	230481193	23586443	102.3	1510737	6.6	108.9
chr2	240548238	25932709	107.8	1590665	6.6	114.4
chr3	198100139	21290257	107.5	1306524	6.6	114.1
chr4	189752667	20563217	108.4	1275689	6.7	115.1
chr5	181265378	19260839	106.3	1169093	6.4	112.7
chr6	170078724	18098639	106.4	1168601	6.9	113.3
chr7	158970132	16842673	105.9	1062644	6.7	112.6
chr8	144768136	16609291	114.7	944736	6.5	121.3
chr9	121790590	13090916	107.5	792844	6.5	114.0
chr10	133263135	14153683	106.2	885116	6.6	112.9
chr11	134533742	14542735	108.1	877361	6.5	114.6
chr12	133138039	14076278	105.7	922071	6.9	112.7
chr13	97983128	10292378	105.0	663042	6.8	111.8
chr14	90568149	9593420	105.9	614811	6.8	112.7
chr15	84641348	8686511	102.6	566154	6.7	109.3
chr16	81805944	9776636	119.5	570201	7.0	126.5
chr17	82921074	8526660	102.8	593562	7.2	110.0
chr18	80089658	8131740	101.5	509689	6.4	107.9
chr19	58400758	6671959	114.2	498761	8.5	122.8
chr20	63944581	6678485	104.4	417613	6.5	111.0
chr21	40088683	3952073	98.6	274285	6.8	105.4
chr22	39159777	3973802	101.5	273995	7.0	108.5
chrX	154893130	10968603	70.8	672945	4.3	75.2
chrY	26415183	352045	13.3	16804	0.6	14.0
Autosomes	2756293213	294331344	106.8	18488194	6.7	113.5
Genome	2937601526	305651992	104.0	19177943	6.5	110.6

*, non-gap sequence in GRCh38; the red and green highlights indicate the highest and lowest densities, respectively.

In the sex chromosomes, the pattern seems to be different when compared to the autosomes. In chromosome X, the SNP density is below 66.3% of the average

density for the autosomes (Table 1; Figure 2). This seems to be correlated with the less number of chromosome copies (3/4 or 75% of autosomes) in the population gene pool.

In the case of Y chromosome, SNP density is 13. 3/kb, which is ~1/8 or 12.8% of these for the autosomes and less than 1/5 or 20% of that for X chromosome (Table 1; Figure 2). This low density can only be partially explained by the low number of Y chromosome copies (1/4 of autosomes and 1/3 of X chromosome) in the gene pool in the human populations. Therefore, there must be other factors contributing to this extremely low density of SNP variation.

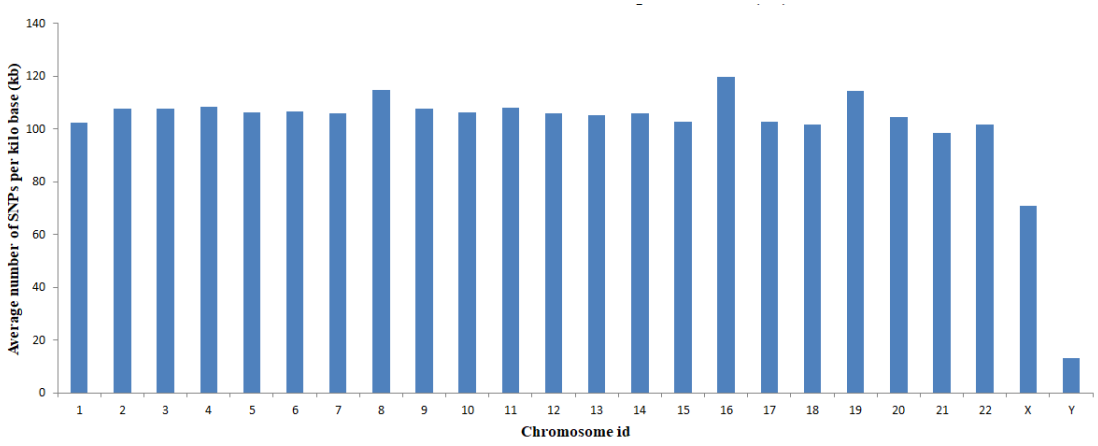

Figure 2. The density of single nucleotide polymorphisms (SNPs) in the human chromosomes shown. The height of the bars represents the average density of (SNPs/kb) in the non-gap chromosome sequence.

A similar analysis was performed for INDELs. As shown in Figure 3, the overall INDEL distribution pattern among chromosomes is very similar to that of SNPs (Figure 1). As seen in Table 1, the genome average density of INDELs is 6.5/Kb, which is more than 10X lower than that of SNPs. The overall relative density (INDELs/kb) in the autosomes is similar across the autosomes with chromosome 19 having the highest (8.5 INDELs/kb) and chromosomes 5 and 18 (6.4 INDELs/kb)

having the lowest (Table 1; Figure 3). It is interesting to notice that the chromosomes with the highest and lowest densities are different among autosomes for SNPs and INDEL, except for chromosome 19, which has a higher SNP density and the highest INDEL density and chromosome 18, which has the lowest INDEL density and lower SNP density.

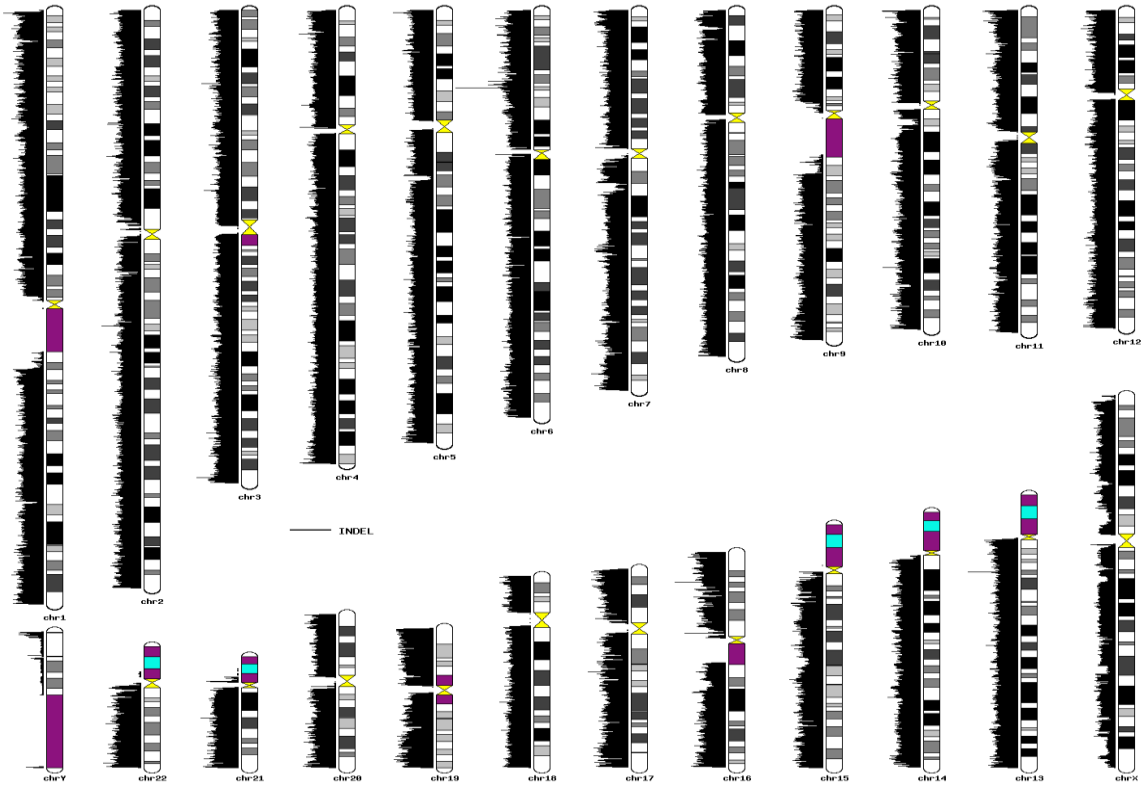

Figure 3. Distributions of small insertions and deletions (INDELs) across the human chromosomes. A density plot onto the chromosome ideogram shows the distribution of INDELs in the chromosomes. The regions in colour represent the heterochromatin regions in the genome and the length of the black lines on the left side of the chromosome indicates the relative density of INDELs (relative to the number of indels/65Kb).

Like for SNPs, the densities of INDELs in the sex chromosomes are also much lower than that of autosomes, with that of chromosome X being 4.3/kb or 64.7% of the autosome average, and the INDEL density for Y chromosome being 0.6 or 9.7% of the autosomes average (Table 1; Figure 3). These ratios are slightly lower than that

of SNPs (66.3% and 12.8% for X and Y chromosomes, respectively). The differences with the autosomes can be due to a similar reason as for the difference in the density of SNPs. When examining the detailed distribution within the chromosomes, it is interesting to notice that there is one small region on the short arm of chromosome 6 showing a very high density of INDELs (Figure 3). In Y chromosome the density profile (shape and location of the peaks) seems to be quite different between SNPs and INDELs even though their relative average densities are quite similar.

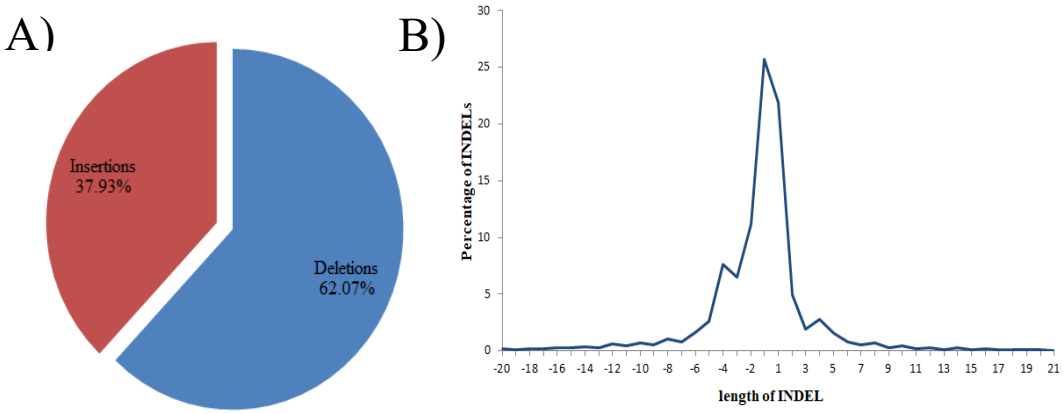

Figure 4. The distribution of INDELs in the genome by type and length. A) The distribution of INDELs defined as insertions and deletions based on the reference. B) The distribution of INDELs by length with deletions shown in negative values and insertions shown in positive values for their lengths.

We also checked the distribution of INDELs as small insertions and deletions based on the human reference allele. From Figure 4A, we can see that 37. 93% of the INDELs were small insertions and 62. 07% as small deletions. Variantion of 1 bp represents the dominant form as being ~25% of the small deletions and ~22% of small insertions (Figure 4B right).

When SNPs and INDELs were combined to represent all small variants outside of structural variants, the pattern of average density across chromosomes are

very similar to that of SNP density due to the much larger number of SNPs than INDELs. Therefore, by the density of SNPs or all small variations including INDELS, it can be stated that chromosome 16 is the least conserved among all chromosomes by having the highest density of variants, while chromosome 21 represents the most conserved chromosome by having the lowest density of variants among autosomes. However, the sex chromosomes, especially the Y chromosome has a much lower level of variation than the autosomes, likely due to less copy number in the population gene pool and some additional unknown factors.

3.1.2 Distribution of SNPs and INDELs among different genomic regions

We examined the distribution patterns of SNPs and INDELs within different genomic regions by gene context. For this, we divided the genome into genic vs inter-genic, coding gene vs non-coding gene, and different regions of protein coding genes, and compared the level of variations among these regions. As shown in Figure 5A, it can be seen that by number 58. 7% of SNPs are located in the genic regions, the remaining 41.3% located in the intergenic regions (Figure 5A left). For INDELs, the percentage in the genic and intergenic regions are 66% and 34% respectively (Figure 5B right). The average density of SNPs in the genic regions (103.11/kb) is slightly lower than in the intergenic regions (105.56/kb). For INDELs, the average density in the genic regions (6.92/kb) is higher than in the intergenic regions (5.88/kb) (Table S2). Unlike SNPs, the results for INDELs is a bit unexpected, as we would think that intergenic regions with less functional constrains should have a higher level of variation. It could be that the current dbSNP data for INDELs is still biased towards those in the genic regions, for example, with a larger proportion from the large-scale exome projects.

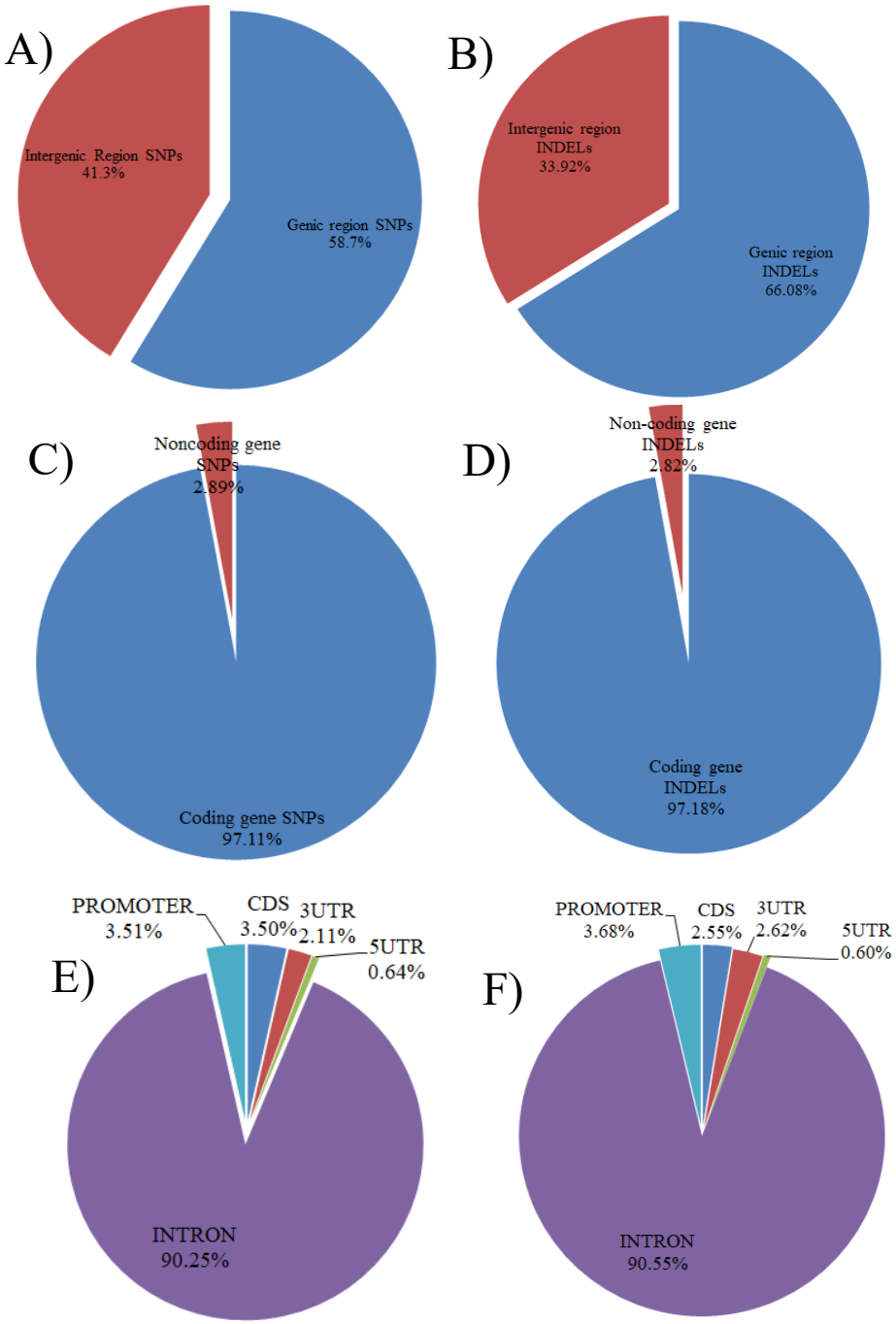

Figure 5. The composition of SNPs and INDELs between the genic and intergenic regions of the genome. A) A pie chart showing the percentage of SNPs in the genic and intergenic regions of the human genome. B) A pie chart showing the percentage of INDELs in the genic and intergenic regions of the human genome. C) A pie chart showing the percentage of SNPs among the coding and non-coding genes. D) A pie chart showing the distribution of INDELs among the coding and non-coding genes. E) A pie chart showing the distribution of SNPs in different regions of protein coding genes. F) A pie chart showing the distribution of INDELs in the different regions of protein coding genes.

Between the types of genes, 97.11% of the SNPs are associated with coding genes and only 2.89% located in the non-coding genes (Figure 5C). This is very similar for INDELs with 97. 18% in the coding genes and only 2.82% in the non-coding genes (Figure 5D). We further examined the density of SNPs and INDELs in these regions. From Table S2, we can see that the coding genes have an average density of SNPs (102.97/kb), which is lower than that in the non-coding genes (107.72/kb). Unlike for SNPs, the average density of INDELs (6.93/kb) in the coding genes is slightly higher than the density of INDELs (6.62/kb) in the non-coding genes (Table S2). The larger percentage of INDELs by number in the coding gene regions is mostly contributed by the much larger genomic regions in the genome than the non-coding gene regions, while the higher density in coding genes is probably attributed to the bias of dbSNP data towards the coding genes.

To investigate the distribution pattern further, we compared the distribution of SNPs and INDELs among the different regions of the protein coding genes by dividing them into core promotors, 5'-UTR, CDS, 3'-UTR, and introns. From Figure 5E & 5F, we can see that introns contain 90.25% of SNPs and the percentage of SNPs in the coding region (CDS), promoter, 5'-UTR, and 3'-UTR are 3. 50%, 3. 51%, 0.64% and 2.11%, respectively (Figure 5E). The distribution of INDELs in the coding genes is very similar to SNPs, with the percentages of INDELs in intron, promoter, 3'UTR, CDS and 5'UTR are 90.55%, 3. 68%, 2.62%, 2.55% and 0.60% of INDELs, respectively (Figure 5F), correlating with the genomic sizes of these regions.

The distribution of SNPs in different genic regions was also compared in density. From Figure S1, introns have the highest density of 119. 67 /kb, followed by promoter, CDS, 3'UTR and 5'UTR regions with average densities of 62.63, 50.18, 8.76 and 3.39 per kb, respectively. To determine whether the SNP density among the

different regions of the protein coding regions are statistically different, the SNP densities using a 1kb sliding window were generated, such that a t-test can be performed. The t-test showed the density of SNPs for all pairwise comparisons among the different gnomic regions are statistically different (Table S3). Therefore, by the density of SNPs, the UTR regions are the most conserved and the intron regions are the least conserved regions in the coding regions of the genome (Figure S1).

The distribution of SNPs and INDELs in the CDS regions was further analysed for the pattern by their functional impact as synonymous vs. non-synonymous for SNPs and, frameshift vs. non-frameshift for INDELs. As seen in Figure 6A, synonymous SNPs make up 37.31%, and among the non-synonymous SNPs, missense SNPs make up for 60.58 % and the non-sense SNPs make up for 2.11 %, indicating a relatively higher percentage of non-synonymous SNPs in the CDS region. For INDELs, 75.16% were frameshift variations and the remaining 24.84% are non-frameshift or in-frame INDELs (Figure 6B). While we were expecting higher percentage for synonymous SNPs and in-frame INDELs due to less selection pressure, the observed lower percentage of synonymous SNPs might be because more substitutions create non-synonymous rather than synonymous substitutions, we expect most new coding mutations to be non-synonymous. However, missense substitutions are more likely to be harmful and thus removed from the population by natural selection, so the rate of observed substitution per site is expected to be always higher at synonymous sites (Yang & Bielawski, 2000). However, the observed number of variants could often be higher for missense than synonymous variants, especially the low-frequency variants that haven't had much of a chance to be affected by natural selection.

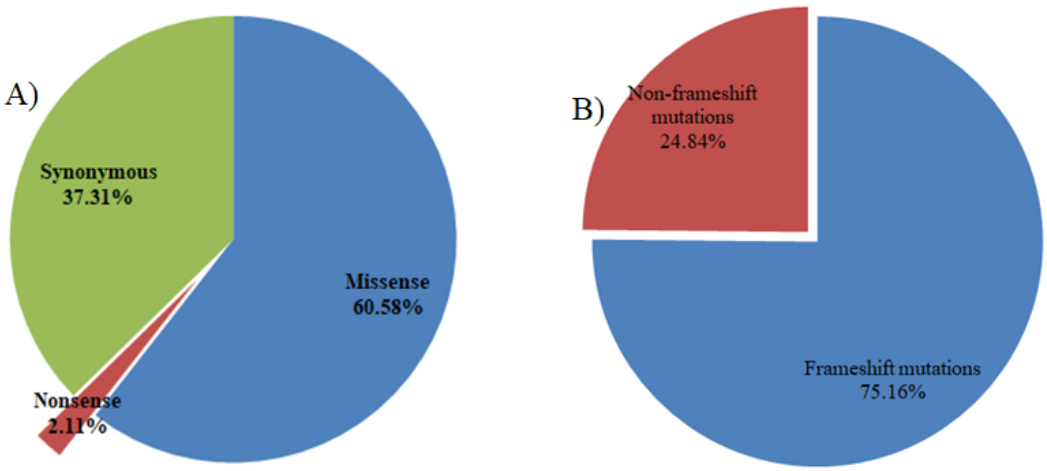

Figure 6. The composition by type of variation for SNPs and INDELs in the coding region (CDS). A) Distribution of SNPs in the coding region of the genome. B) The distribution of INDELs in CDS as reading frameshift and non-reading frameshift.

3.1.3 The functional characteristics of genes with the highest and lowest variability in the CDS regions

We are interested in knowing the characteristics of genes with the highest and lowest sequence variability. For this, we focused on the variability in the coding regions (CDS) of protein coding genes, since the functional impact of such variability is easier to be assessed and functions of these genes are better known than for the non-coding genes. All SNPs in the CDS regions are distributed in 19,644 genes, out of which 378 genes constitute the top 5% genes with the highest SNP density and 636 genes constitute the bottom 5% of genes with the lowest SNP density. All missense SNPs are distributed across 18,840 genes, out of which 340 and 635 genes represent the top 5% and bottom 5% of genes based on missense SNP density, respectively (Table S4). These gene lists were subjected to GO term enrichment analysis to examine the functional characteristics of these genes using Enrichr (E. Y. Chen et al., 2013). The GO terms covering the biological process, molecular function and cellular

components that are statistically enriched based on the adjusted p-value being 0.05 or less were selected and examined.

3. 1. 3. 1: Enriched GO terms by genes with the highest density of SNPs and missense SNPs.

Genes with most variability as shown to have the highest density of SNPs are predicted to involve in the biological processes that are less essential for cells or organisms or are those that require a high level of variability to deal with high variability of environment factors. As shown in Table 2, half of the top 10 biological processes GO terms are related to immune response, sensory or chemical detection, which seems to make perfect sense as these processes all need to deal with a high level of variability in antigens, chemical, and sensory substances. Aside from these processes, terms related to assembly chromosomes, nucleosomes, and protein-DNA complexes are among the remaining enriched biological process GO terms. This observation is a bit unexpected and interesting as we would think that these processes are critical for cells. However, the fact that they are associated with genes showing the highest variability may suggest that these processes could have a high level of flexibility, perhaps allowing the more intimate mechanism of gene regulation.

In comparison with the density of all SNPs, the density of missense SNPs is expected to have a better indication of variability constrain on protein function. The GO terms enriched among genes with the highest density of missense SNPs are more or less similar to that for all SNPs, but with a higher percentage of GO terms related to immune response, sensory perception and chemical detection with “immune response” related terms alone making up for more than half of the enriched biological process GO terms. Aside from these processes, peptide cross-linking, which is defined as the formation of covalent-link between or within protein chains is one of the

enriched terms (Table 2). This finding is surprising as the covalent bond formation within or between protein chains are essential for protein structure and function. However, on a closer examination, all seven genes (FLG; LCE2B; LCE2C; SPRR3; LCE2A; LCE1E; SPRR2D) associated with this process are associated with the formation of keratinocyte and epidermal cell differentiation processes. It is possible that peptide cross-linking is only associated with genes for the formation of epidermal cells and related components, which may also be related to an immune response.

Table 2. Biological process GO terms enriched among genes with the highest density of SNPs and missense SNPs

Term	Adjusted p-value
GO terms for genes with the highest density of SNPs	
chromatin assembly	0.00000000
innate immune response in mucosa	0.00000000
detection of chemical stimulus involved in sensory perception	0.00000000
nucleosome assembly	0.00000000
mucosal immune response	0.00000000
sensory perception of chemical stimulus	0.00000000
nucleosome organization	0.00000000
antibacterial humoral response	0.00000000
protein-DNA complex assembly	0.00000000
peptide cross-linking	0.00000000
GO terms for genes with the highest density of missense SNPs	
detection of chemical stimulus involved in sensory perception	0.00000003
sensory perception of chemical stimulus	0.00000003
type I interferon signalling pathway	0.00000003
regulation of peptidyl-serine phosphorylation of STAT protein	0.00000003
positive regulation of peptidyl-serine phosphorylation of STAT protein	0.00000003
peptide cross-linking	0.00000003
keratinocyte differentiation	0.00000003
cellular response to type I interferon	0.00000003
epidermal cell differentiation	0.00000003
natural killer cell activation involved in immune response	0.00000003

Similar to biological process, genes with the highest density of SNPs are predicted to associate with the molecular functions that are less essential for cells or organisms. As shown in Table 3, there are only four significant molecular function terms for the gene set with the highest density of SNPs. These terms are related to DNA binding and DNA repair. This is a surprising observation as DNA binding proteins such as transcription factors are considered to be conserved. DNA repair, on the other hand, can have higher variability than DNA binding. Moreover, this observation can be related to the enriched biological process GO terms for the same gene set, showing that these functions are possibly less conserved and more flexible than we might think. The gene set with the highest density of missense SNPs has five significantly enriched molecular function terms. Four of these terms are exactly the same as those seen in the genes with the highest density of all SNPs. With the additional term associated with immune response, which is in the agreement of the enriched biological process for immune response.

Table 3. Molecular function GO terms enriched among genes with the highest density of SNPs and missense SNPs

Term	Adjusted p-value
GO terms for genes with the highest density of SNPs	
mismatch repair complex binding	0.00000000
DNA binding	0.00000000
DNA insertion or deletion binding	0.02000000
oxidized DNA binding	0.04000000
GO terms for genes with highest density of missense SNPs	
mismatch repair complex binding	0.00137741
type I interferon receptor binding	0.00137741
DNA insertion or deletion binding	0.01000000
oxidized DNA binding	0.02000000
oxidized purine DNA binding	0.04000000

The cellular component GO terms for the genes with the highest density of SNPs and missense SNPs were seen to be enriched for the cellular components and structures occupied by macromolecules. As shown in Table 4, the four enriched terms are related to nuclear, luminal side of the endoplasmic reticulum and major histocompatibility complex (MHC), which seems to agree with the enriched GO terms for biological processes and molecular functions. MHC complex is one of the main cellular complexes associated with immune response (Janeway, Travers, Walport, & Shlomchik, 2001). Cellular components GO terms enriched among genes with the highest density of missense SNPs is very similar to that for all SNPs (Table 4).

Table 4. Cellular location GO terms enriched among genes with the highest density of SNPs and missense SNPs

Terms	**Adjusted p-value**
GO terms for genes with highest density of SNPs	
nuclear chromatin	0.00016103
nuclear chromosome part	0.00016103
MHC protein complex	0.00016103
integral component of lumenal side of endoplasmic reticulum membrane	0.00016103
GO terms for genes with highest density of missense SNPs	
MHC protein complex	0.00208215
integral component of lumenal side of endoplasmic reticulum membrane	0.01000000
Golgi lumen	0.01000000

3.1.3.2 GO terms enriched among genes with the lowest density of SNPs and missense SNPs

Genes with the least variability as having the lowest density of SNPs are predicted to be involved in biological processes that are very essential for cells or organisms. As shown in Table 5, spindle assembly was shown as the only enriched biological process with a statistical significance. This is expected as spindle assembly

is the aggregation, arrangement and bonding a set of components to form the spindle, array of microtubules and associated molecules that serve to separate the duplicated chromosomes precisely into daughter cells, and these are conserved and essential processes for cells.

The enriched biological processes GO terms for genes with the lowest density of missense SNPs have 110 significant ones. As shown in Table 5, over half of the top 10 terms are related to RNA/mRNA splicing and processing, pathways that involve regulatory proteins, and process of transcription termination. This observation is expected as these biological processes are critical for cells. It's also interesting to notice the non-canonical wnt signalling pathway, which controls the intercellular calcium levels, is among the enriched GO terms, indicating that this pathway is very conserved and essential for cells (Table 5).

Table 5. Biological process GO terms enriched among genes with the lowest density of SNPs and missense SNPs

Terms	**Adjusted p-value**
GO terms for genes with lowest density of SNPs	
spindle assembly	0.00000000
GO terms for genes with the lowest density of missense SNPs	
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	0.00000000
mRNA splicing, via spliceosome	0.00000000
mRNA processing	0.00000000
RNA splicing	0.00000000
RNA processing	0.00000000
mRNA metabolic process	0.00000000
non-canonical Wnt signaling pathway	0.00000000
ubiquitin-dependent protein catabolic process	0.00000000
termination of RNA polymerase II transcription	0.00000000
DNA-templated transcription, termination	0.00000000

Similar to biological process, genes with least variability are predicted to associate with the molecular functions that are critical for cells or organisms. As shown in Table 6, only 3 enriched molecular function GO terms were seen and all of them relate to ubiquitin protein activity, which targets substrates to the protein degradation pathways in mammalian cells (Clague & Urbé, 2010), indicating their importance for cellular function.

The gene set with the lowest density of missense SNPs has a total of 40 significant enriched molecular function terms, of which the top 10 terms are shown in Table 6. Similarly, these GO terms are associated with ubiquitin protein activity. Aside from this, molecular functions GO term related to the formation of mRNA from DNA make up for the remaining terms (Table 6). This result is very much expected as these functions are associated with the transcription process and are critical to all cells.

Table 6. Molecular function GO terms enriched among genes with the lowest density of SNPs and missense SNPs

Terms	Adjusted p-value
GO terms for genes with the lowest density of SNPs	
ubiquitin conjugating enzyme activity	0.00000000
ubiquitin-like protein conjugating enzyme activity	0.00000000
Lys63-specific deubiquitinase activity	0.04000000
GO terms for genes with the lowest density of missense SNPs	
RNA binding	0.00000000
ubiquitin protein ligase binding	0.00000000
GTP binding	0.00000000
nucleoside-triphosphatase activity	0.00000000
ubiquitin-like protein ligase binding	0.00000000
purine ribonucleoside binding	0.00000000
GTPase activity	0.00000000
guanyl ribonucleotide binding	0.00000000
purine ribonucleoside triphosphate binding	0.00000000
GDP binding	0.00000000

Among the enriched cellular component GO terms for the genes with the lowest density of overall SNPs (Table 7), there are only 4 significant enriched GO terms, all associated with the Golgi apparatus cellular component. A similar significant enrichment can be noticed for the gene set with the lowest density of missense SNPs. However, from the top 10 enriched terms of the 45 significant enriched cellular component terms for this gene set, aside from Golgi apparatus, we also see GO terms related nucleus and spliceosomal complex (Table 7). This makes sense as transcription occurs in the nucleus and spliceosomal complex plays a critical role in RNA splicing.

Table 7. Cellular location GO terms enriched among genes with the lowest density of SNPs and missense SNPs

Terms	Adjusted p-value
GO terms for genes with the lowest density of SNPs	
Golgi cisterna membrane	0.00000000
Golgi cis cisterna	0.00000000
Golgi cisterna	0.00000000
cis-Golgi network	0.00000000
GO terms for genes with the lowest density of missense SNPs	
Golgi cisterna membrane	0.00000000
spliceosomal snRNP complex	0.00000000
spliceosomal complex	0.00000000
Golgi cis cisterna	0.00000000
Golgi cisterna	0.00000000
spliceosomal tri-snRNP complex	0.00000000
U12-type spliceosomal complex	0.00000000
cytoplasmic ribonucleoprotein granule	0.00000000
nuclear speck	0.00000000
ficolin-1-rich granule	0.00000000

3.2 Determining the ancestral alleles for all human SNPs and INDELs.

The second objective of this study was to identify the ancestral state of all human SNP and INDEL variants considered in this study. This is important as one of the limitations of the current human reference genome sequence is that it does not reflect the ancestral state of the variants and that dbSNP contains defined ancestral state only for a very small number of variants. To achieve this goal, we relied on comparative genomics. Specifically, the process in this case involved comparing the human genome sequences with those of the 5 other primates closely related to human and identifying the sequences at the orthologous loci in these genomes as the basis to call the ancestral state for each human variant. Among the five other primate genomes, we used the chimpanzee genome the primary reference due to its closest distance from human, while other genomes were used for additional references at different evolutionary distances. The cutoff values for identifying the orthologous sequence in each genome for SNPs and INDELs can be found in Figure 7 and Table S5. As expected, the cutoff values for blat scores showed a graduate decreasing with the increasing evolutionary distance of the species from humans.

3.2.1 Common ancestral alleles across the primate genomes from the *Hominidae* group

The 305,651,992 SNP and 19,177,943 INDEL variants considered in this study make up 11.05% of the non-gap length genome sequence, of which 10.40% are from SNPs and 0.65% are from INDELs. Among these variants, 96.15% of SNPs and 94.18% of INDELs have an orthologous region in the chimpanzee genome. With SNPs and INDELs combined, 98.41% of these small variants have an orthologous region present in the chimpanzee. The ancestral alleles of a variant were identified as the allele shared with the orthologous sequence in the chimpanzee genome; all the

other genomes were used to confirm the confidence of the identified ancestral alleles and to count the shared ancestral alleles among different grouping primates by the evolutionary distance. The variants for which the ancestral alleles were not identifiable in this way were eliminated from further analysis.

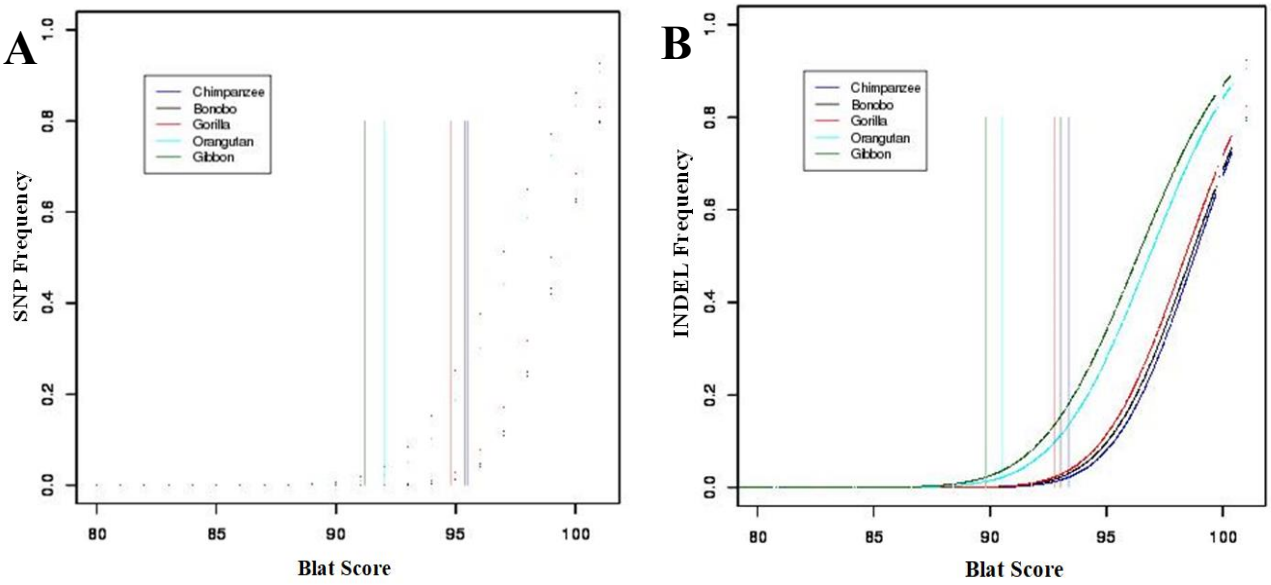

Figure 7. The cutoff values for minimal blat score in identifying orthologous regions of human SNPs and INDEL sites in other primate genomes. A). Cutoff values for SNPs. (B). Cutoff values for INDELs.

Between human and chimpanzee, 93.91% (287,030,300 SNPs) of the human SNPs have a shared allele with the chimpanzee genome (Figure 8) and 80.56% (15,451,232 INDELs) of human INDELs have a shared allele with the chimpanzee genome. Together, 302,481,532 of human SNPs and INDELs, representing 93.11% of all small variants, share an orthologous allele with the chimpanzee genome, i.e., with the ancestral state identifiable. These numbers start to gradually decrease when additional more distant genomes were included. 67.34% of the human variant sites share the ancestral allele with all five other primate genomes (Figure 8).

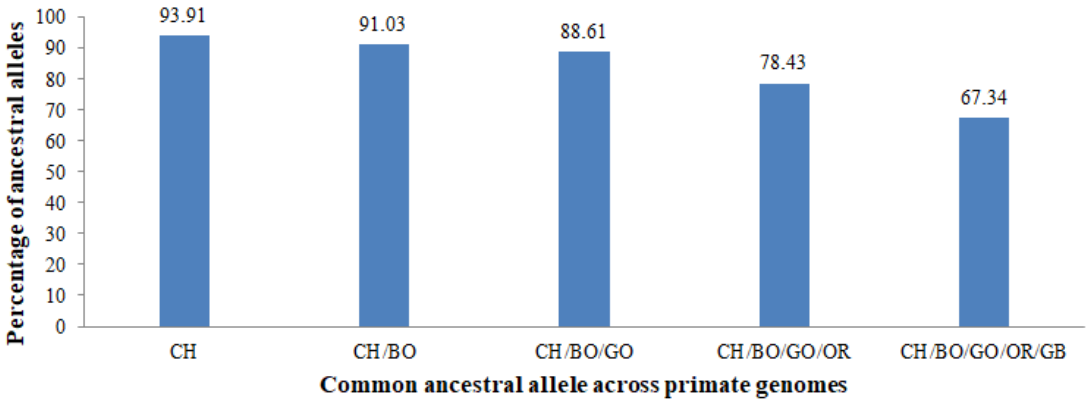

Figure 8. The ratios of ancestral alleles shared among the different primates for SNPs. Bar plots represent the percentages of shared ancestral alleles between human and different groups of other primates. CH: Chimpanzee; BO: Bonobo; GO: Gorilla; OR: Orangutan; GB: Gibbon.

3. 2. 2. Distribution of ancestral alleles.

After determining the ancestral state for each variant locus, we checked to see if the ancestral allele matched the human reference allele or the alternate allele. As shown in Figure 9, for SNPs, 95.34% of the ancestral alleles were the same as the human reference alleles, and the remaining 4.66% of the variant loci had the ancestral state matching the alternate allele. The latter group consists of 13,353,349 bp or 0.45% of the genome. For INDELs, 76.18% of the ancestral alleles matched the human reference allele and the remaining 23.82% matched the alternate allele. The latter constitutes a total of 10,255,109 bp in total or 0.34% of the entire genome. Therefore, the current human reference genome has approximately 0.45% difference in terms of SNPs and 0.34% difference in terms of INDELs for a total of 0.79% difference in small variants when compared to the last human common ancestor genome (Figure 9). In other words, a total of 0.79% of the human genome has been subject to small variation since the last common ancestor.

For INDELs, we also checked their distribution as insertions and deletions in reference to the ancestral allele. As shown in Figure 10A, 21.19% of the alleles were small insertions and 78.81% were small deletions, indicating small deletions were the prevalent INDELs since the last common ancestor. In comparison, this number is higher than the percentage of deletions based on the reference genome (60.07%, Figure 4). This is expected since the reference genome represents a younger genome, thus already carried a lot of new deletions since the ancestral genome. Among the INDELS based on the ancestral genome, deletions of 1 bp made up for ~38% of the small deletions, while insertion of 1bp made up approximately 13% of small insertions (Figure 10B), and in general the larger the size of the INDELs, the less frequently they are. This pattern is quite similar to that for the INDELs based on the reference genome. The data indicate that the INDEL variation from the ancestral genome is biased towards small size deletions.

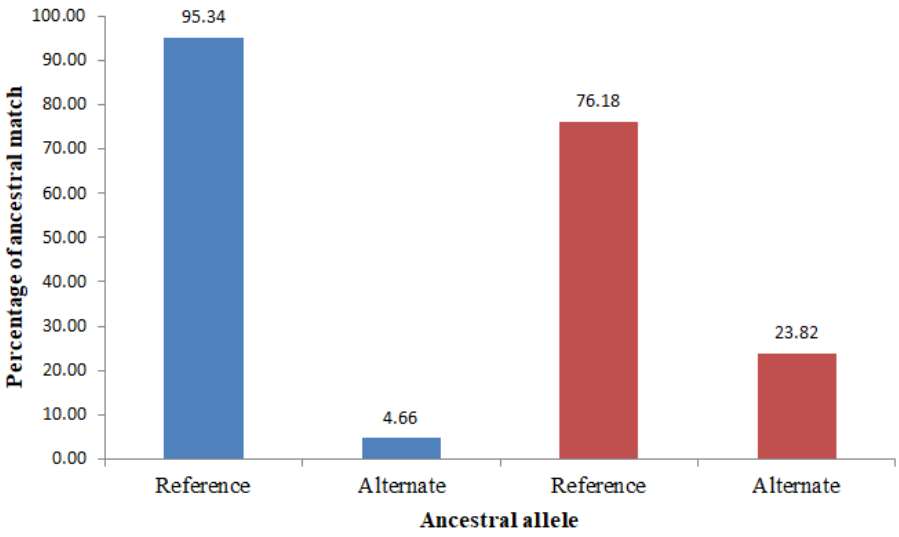

Figure 9. The percentage of the reference alleles and alternate alleles matching with the ancestral alleles. Bar plots showing the percentages of ancestral alleles as the reference alleles and alternate alleles for SNPs (blue) and INDELs (red).

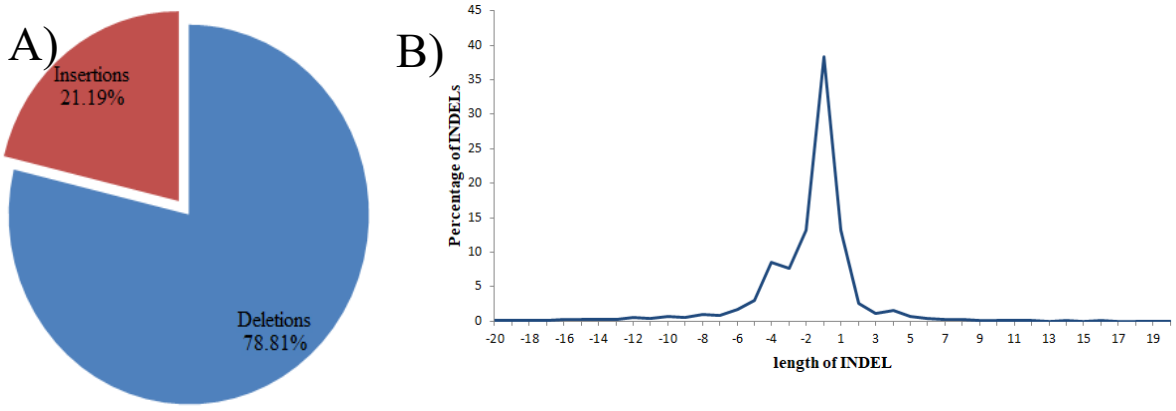

Figure 10. The percentage of INDELs in the genome from the ancestral allele. A) A pie chart showing the distributions of variations from the ancestral allele in insertions and deletions. B) A line plot showing the relative frequency of variation from the ancestral state by length with deletions showing in negative values and insertions in positive values for their lengths.

3.2.3 Pattern of nucleotide changes from the ancestral state.

The next analysis was to examine the pattern of variations by type using the ancestral state as the starting point. Specifically, we examined the ratio of SNPs among transitions and transversions. Transitions are interchange between two-ring purines (A ↔ G) or between the one-ring pyrimidines (C ↔ T), with the resulting nucleotides having a similar structure. Transversions are variations that occur as a result of interchanges between purine and pyrimidine bases, resulting in variants with a different structure. As shown in Figure 11A, 42.3% of the SNPs represent transitions from the ancestral state, while 57.7% represent transversions (Figure 11A). Among the transversions, the change from A to T stands out among all 8 possibilities by being more than 4 times higher (13.53) than the rest of 7 options, which showed a more or less rates from 3.4 to 4.2%. The reason for this high A->T ratio might be related to a large number of polyA sequences in the genome. However, we would

expect to see a similar rate between A->T and T->A because of the complementary nature of the double-stranded DNA.

All four types of transitions were approximately the same, but G->A and C->T are slightly higher than A->G and T->C, indicating that since the ancestral genome, more SNPs involved conversion from G/C pair to A/T paring than the other way around (Figure 11B). This might be contributed by DNA methylation, which specifically occurs on the cytosine bases in CpG dinucleotides, making it susceptible to conversion to the thymidine base (Jabbari & Bernardi, 2004). For this reason, we expect that CpG islands which are more resistant to DNA methylation should have a lower rate of C->T or G->A variations, so we further looked into transversion and transition changes between the CpG and non-CpG island regions from the ancestral alleles.

As shown in figure 11C, the profiles of nucleotide changes by type are different between in the CpG island regions and non-CpG island regions. Notably, the percentages of G->A and C->T changes (22.1 + 22.07=44. 17%) were higher in CpG island regions than these of non-CpG island regions (15.14+15.14=30.28%). A statistical t-test was performed and showed the rate of C->T change was significantly (p-value=0.0001) higher in the CpG (27.47 SNP/kb) island regions than the non-CpG (15.54 SNP/kb) island regions (Table S6). This is opposite to what was expected based on the lower rate of DNA methylation in CpG island. However, this might be related to the much higher ratio of CpG dinucleotides in the CpG islands. By definition, CpG islands are defined as genomic regions, which are more than 500 bp long and have a minimum 55% of GC content and a minimum of 0.6 for the observed/expected ratio of CpG dinucleotides, which is more than 4 times higher than the genome average (Takai & Jones, 2002). Therefore, the ratios of G->A and C->T

in CpG islands are higher due to the much higher density of CpG dinucleotides despite the lower rate of DNA methylation than in the non-CpG island regions.

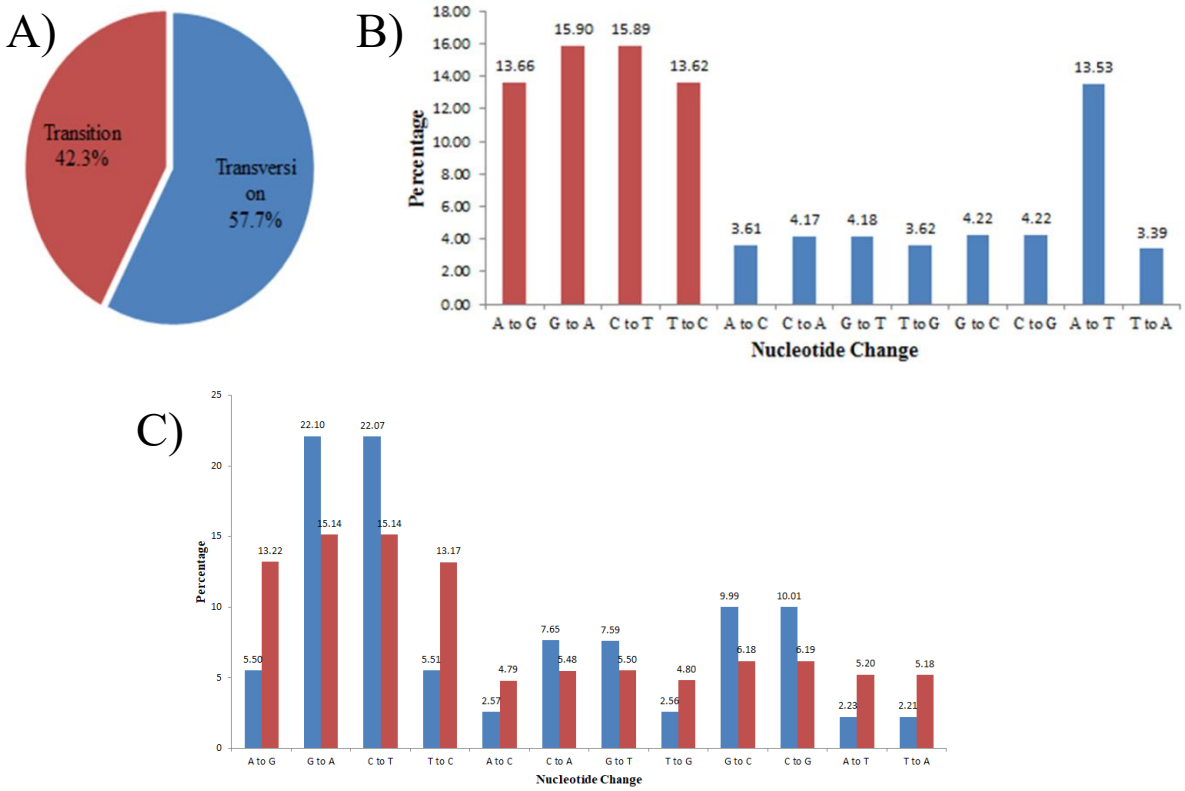

Figure 11. Distribution of SNPs between the transition and transversions variations from the ancestral state. A) A pie chart showing the breakdowns of variations from the ancestral state between transition and transversion in percentage. B) The distribution of SNPs in percentage among different types of transitions and transversions based on the ancestral allele represented. C) The distribution of SNPs in percentage among different types of transition and transversion for CpG island (blue) and non-CpG island (red) regions.

Chapter 4: Discussion

4.1 Overview of the study

The main objectives of this study were to systematically analyze the distribution pattern of SNPs/INDELs across the human genome and provide the ancestral state of the human sequences that are subject to variation. The data considered in this study was obtained from dbSNP build 150, which contains 305,651,992 and 19,177,943 INDELs identified mostly from large-scale international projects focusing on human genetic diversity, such as the HapMap project, 1000 genome project, exome sequencing project and other personal genome projects (1000 Genomes Project Consortium, 2012; 1000 Genomes Project Consortium et al., 2015, 2010; Deloukas & Bentley, 2004; Telenti et al., 2016). Looking into the distribution pattern of the variants helps better understand the trend of human genetic variation in association of gene function, while identifying the ancestral state of a variant not only helps more accurate assessment of the variant's functional impact but also allows to examine the trend of variations in the context of human evolution.

4.2. How are SNPs and INDELs distributed across the human genome?

We looked into the distribution pattern of genomic variants in several ways. First, we examined their distribution among chromosomes. Interestingly both SNPs and INDELs had a more or less similar distribution pattern across the autosomes (Figure 1 & 3). On the other hand, the sex chromosomes have a different distribution pattern (Figure 1&3).

By the average densities of SNPs and INDELs, chromosome 16 had the highest density of SNPs, while high densities of INDELs were found on chromosome 19, thus they are the most variable chromosomes. Chromosome 21 and 18 can be

considered as conserved chromosomes as they had the lowest density of SNPs and INDELs respectively (Table 1). This result is different from previous SNP density studies conducted by the HapMap project which showed chromosome X and Y to have the highest density of SNPs per kb (Sachidanandam et al., 2001). This could be due to the lower and incomplete coverage of variants in the HapMap project (Tantoso, Yang, & Li, 2006). In our study, we observed the X chromosome to have approximately on an average 25% less density of SNPs than the autosomes. This is expected as the autosomes have four copies of chromosomes and X chromosome only has 3 copies in the germline cells. For the Y chromosome, we expected the density to be ~25% of that for the autosomes because in the germline cells the Y chromosome only has one copy, but our results showed the Y chromosome had less than 10% of average density for autosomes. One of the reasons behind the low density of SNPs and INDELs in the Y chromosome can be due to the relatively low copy of Y chromosomes in the gene pool, i.e., 1/4 of the autosomes. There must be other factors involved for the lower than expected variation density for the Y chromosome. This trend is an interesting observation and our results can motivate further research on the Y chromosome variantion. Also, unlike the Hapmap project, we also include INDELs.

4.3 Significant functional association of SNPs and INDELs

The distribution and density of SNPs and INDELs in various functional regions of the genome were further analyzed. For both the genomic variants, the genic regions showed a higher percentage of variants than the intergenic regions (Figure 5A & 5B). In the genic region, SNPs and INDELs associated with the coding genes are staggeringly higher in number than the non-coding genes (Figure 5C & 5D). Further, within the coding genes, for both the variants, the introns had the highest percentage of variants followed by promoter, CDS and UTR regions (Figure 5E & 5F).

A good explanation to this bias in distribution, which is kind of opposite to what we would expect, is that for the regions that are sequenced, the analysis of variants over the years have been biased to the functional and conserved region of the gene that makes up for the genic region. Due to the price advantage, several large scale and personal genome projects used exome sequencing instead of whole genome sequencing, which might lead to a higher variant density in the coding region of the genome vs the noncoding region of the genome. For example, the deep sequencing of 10,000 human genomes, which is a personal genome project where 10,545 human genomes were sequenced using exome sequencing at 30-40x coverage and 150 million SNPs were reported. Exome sequencing contributed to 91.5% of these variants (Telenti et al., 2016). This explains why the SNPs are biased in the genic regions as oppose to the intergenic or intron regions.

When we looked into the SNP distribution in the coding region we can see that missense SNPs have contributed to 60.58% of the CDS regions being the highest in density followed by synonymous SNPs with the contribution of 37. 31%. Nonsense SNPs have the lowest density of SNPs in the coding SNPs (Figure 5C, 5D, 5E & 5F). This can be due to the fact that in CDS regions, variants at 2/3 of sites are likely to be missense, variants at less than 1/3 of the sites maybe synonymous, while the percentage a variant leading to a stop codon is even much smaller (<3/64). Also, this observation is different when compared to the ExAc data where more synonymous variants were reported than the missense and nonsense variants (Fu et al., 2013). Our results showed more non-synonymous SNPs compared to the synonymous SNPs. The difference between the ExAc data and our data is that ExAc only looked into 45 million SNPs whereas we looked into approximately 305 million SNPs, out of which 6,350,071 SNPs are in the CDS regions.

4.4 What genes are subjected to the highest and lowest variability?

In the study, we compared the functions of genes in GO terms between genes with the highest and lowest variability based on the density of SNPs and INDELs. The hypothesis is that genes that have the highest variability have the highest density of SNPs and are predicted to have an association with gene functions that require a high level of variability. On the other hand, the genes with the lowest variability are expected to associate with functions that are critical for cells and organisms. The gene sets included were the genes with the highest density and lowest density of SNPs, as well as genes with the highest and lowest density of missense SNPs. We looked into GO terms in biological process, molecular function and cellular components.

From the biological process for the two genes sets with the highest density of SNPs and missense SNPs, respectively, we can notice very similar biological process GO terms were enriched. The most common terms for both the gene sets include immune response, sensory or chemical detection. This finding is the same as a study conducted in 2009 that found enrichment of SNPs and CNVs towards "environmental sensor" genes that were defined as genes that are not necessarily critical for early embryonic development, but rather help perceive and interact successfully with ever-changing environment (Ionita-Laza, Rogers, Lange, Raby, & Lee, 2009). However, that study was conducted in 2009, where the data set used in the study involved a much lesser number of SNPs. In our study, we used the most recent data set for the analysis. We also analysed the GO terms for genes with the high density of missense SNPs which was not used in the earlier study (Table 2).

The genes with the lowest density of SNPs and missense SNPs showed the most association with RNA/mRNA splicing and processing and regulatory protein pathways being enriched. From this observation, we can say that the genes associated

with the transcription process are most conserved. It is also interesting to see that the intercellular calcium regulation is shown to be a conserved biological process (Table 5).

The enriched molecular function GO terms for the gene sets with the highest number of SNPs showed surprising enriched terms. Enrichment towards DNA binding and repair was associated with this gene set (Table 3). This is surprising as transcription factor proteins associated with DNA binding and repair are considered to be conserved (Sauer, Yocum, Doolittle, Lewis, & Pabo, 1982). The gene set with the highest density of missense SNPs showed the exact same enrichment association along with association with immune response terms (Table 3).

The enriched molecular function terms for genes with the lowest density of SNPs and missense SNPs were mainly associated with ubiquitin protein activity, which is the activity of a class of proteins that targets dysfunctional proteins for degradation (Clague & Urbé, 2010). Apart from ubiquitin protein activity, the missense gene set is associated with the formation of mRNA from DNA (Table 6), which is expected to be a very conserved process and are critical for cells like shown in the results.

4.5 The pattern of human variation since the last human common ancestor

The currently available variant data in dbSNP do not provide the ancestral allele for all variants, but only for a very small number based on an earlier limited study (Sherry et al., 2001). A study conducted in 2006 focused on identifying the ancestral variants by comparing the then available SNPs to the chimpanzee genome. At the time, only approximately 6 million SNPs existed in dbSNP (Spencer et al., 2006). The dbSNP build 150 used in this study has around 324 million SNPs and INDELs. Identifying the ancestral allele of a variant can be useful to better predict the

functional impact of variants and also allow researchers to examine the pattern of human variations since the last common ancestor for all human populations. The ancestral state of the human variants represents the genomic sequence in the last common ancestor of all human populations.

As the second objective, we looked into determining the ancestral state of all the human variants. But since the ancestral genome sequence is not available, we used the chimpanzee genome as a reference. Comparing to the early only study (Spencer et al., 2006), we analysed a total of 324,829,935 SNPs and INDELs and the ancestral allele of the human variants by comparing to the sequences of Chimpanzee, Gorilla, Orangutan, Bonobo and Gibbon genomes. 93.90% of SNPs and 80.56% of INDELs had shared alleles with the chimpanzee genome. 95.34% of SNPs had the ancestral allele match the reference allele and 4.66% of ancestral alleles match the alternate allele. For INDELs, 76.18% and 23.82% (Figure 9) of identified ancestral alleles matched the reference and alternate alleles respectively. Combining SNPs and INDELs, we determined that reference genome sequence differs in small variants from the ancestral genome by 0.79%.

The pattern of nucleotide changes was analysed using the identified ancestral allele as the reference. Results revealed 42.3% of transitions and 57.7% transversions, with G->A making up for the highest percentage of transitions and A->T had the highest percentage of transversion nucleotide changes (Figure 11A & 11B). Our results showed the variation since the ancestral genome had a bias towards deletions (78.81%)(Figure 10), and this bias is stronger compared to using the human reference genome as a base (62.07%) (Fig. 4). Earlier studies suggested that INDELs that are not in the multiples of 3 are relatively uncommon in the coding regions than the non-

coding regions (Bai, H., *et al*., 2013; Zheng, L. Y., *et al*., 2011). Our study contradicts this finding as INDELs showed a higher percentage of 1 bp deletion.

4.6 Summary and conclusion

In this study, through an unprecedented systematic analysis of human variants (SNPs and INDELs), that make up for 10.61% of the total non-gap human reference genome sequences, we examined the distribution and density pattern across the chromosomes and various regions of the genome, determined the ancestral state of these variant loci, as well as examined the pattern of variation since the last human common ancestor.

Through our study, we found the frequency of SNPs and INDELs averaged at 104 SNPs/kb and 6.5 INDELs/kb or ~11% of the genome has been subject to small variation. Further, by these small variations, chromosome 16 and 21 are shown to be most conserved autosomes, while the sex chromosomes show much lower variation with the Y chromosome showing extremely low variation. At the gene level, these small variants are biased towards genic regions, and especially protein coding genes. In the coding genes, SNPs and INDELs are biased towards missense and frameshift variations, respectively. Genes with a high level of variation are related to environment sensing and immune responses, while genes with the lowest variations are associated with critical processes such as RNA splicing and processing. Through a comparative genomic approach, we determined the ancestral state for most of these variants, and our data revealed that 0.79% of the current reference human genome is different from the hypothetical last common ancestral genome for humans.

Our study also has several limitations. First, we used dbSNP built 150 which contains 324,829,935 total number of SNPs and INDELs, and it was the latest built

when the project was started. However, a version of dbSNP (built 151) released on 22nd of March 2018 contains 660,773,127 total entries, which almost doubled in number. Nevertheless, we would expect that these newer variants are very likely to represent most rare variants that are very like to have the ancestral allele same as the reference genome. Second, the data collection methods and samples are biased away from the African population and does not reflect the genetic diversity of the African populations. For example, from the 2054 samples considered in the 1000 genome project, only 902 samples were from 7 African populations. Third, our study did not cover structural variants which are smaller in the number of entries but may involve a larger portion of the genome by sequence length. However, the methodology used in our study can be extended for performing a similar study with these SVs. Among SVs, the ancestral state of ME variants is always the absence of the MEs, thus are relatively easier to determine.

Appendixes

Table S1: T-test P values for pairwise comparison of SNP density among chromosomes

ChrID	chr1*	chr2	chr3	chr4	chr5	chr6	chr7	chr8	chr9	chr10	chr11	chr12
chr1	1	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	1.47E-07	2.20E-16	2.20E-16
chr2	2.20E-16	1	7.88E-03	4.61E-02	2.20E-16	2.28E-16	2.20E-16	2.20E-16	7.85E-06	2.14E-01	4.86E-01	2.20E-16
chr3	2.20E-16	7.88E-03	1	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.62E-04	4.89E-01	9.83E-05	2.20E-16
chr4	2.20E-16	4.61E-02	2.20E-16	1	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	8.81E-02	1.41E-07	2.20E-16
chr5	2.20E-16	2.20E-16	2.20E-16	2.20E-16	1	2.91E-02	5.70E-06	2.20E-16	7.08E-11	4.69E-01	2.20E-16	3.90E-07
chr6	2.20E-16	2.28E-16	2.20E-16	2.20E-16	2.91E-02	1	3.37E-11	2.20E-16	2.50E-07	6.06E-01	2.20E-16	1.52E-13
chr7	2.20E-16	2.20E-16	2.20E-16	2.20E-16	5.70E-06	3.37E-11	1	2.20E-16	2.20E-16	2.19E-01	2.20E-16	8.75E-01
chr8	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	1	2.20E-16	1.82E-15	2.20E-16	2.20E-16
chr9	2.20E-16	7.85E-06	2.62E-04	2.20E-16	7.08E-11	2.50E-07	2.20E-16	2.20E-16	1	8.49E-01	8.70E-10	2.20E-16
chr10	1.47E-07	2.14E-01	4.89E-01	8.81E-02	4.69E-01	6.06E-01	2.19E-01	1.82E-15	8.49E-01	1	2.68E-01	2.12E-01
chr11	2.20E-16	4.86E-01	9.83E-05	1.41E-07	2.20E-16	2.20E-16	2.20E-16	2.20E-16	8.70E-10	2.68E-01	1	2.20E-16
chr12	2.20E-16	2.20E-16	2.20E-16	2.20E-16	3.90E-07	1.52E-13	8.75E-01	2.20E-16	2.20E-16	2.12E-01	2.20E-16	1
chr13	2.84E-05	1.06E-01	2.60E-01	4.21E-02	8.91E-01	9.62E-01	5.55E-01	2.63E-14	4.98E-01	7.04E-01	1.34E-01	5.44E-01
chr14	2.20E-16	2.20E-16	2.20E-16	2.20E-16	4.46E-04	7.77E-08	5.50E-01	2.20E-16	2.20E-16	2.51E-01	2.20E-16	4.39E-01
chr15	3.22E-02	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.53E-05	2.20E-16	2.20E-16
chr16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
chr17	5.90E-02	2.02E-08	3.70E-07	3.65E-10	6.19E-04	2.44E-04	4.75E-03	2.20E-16	7.98E-06	5.43E-03	2.68E-08	5.02E-03
chr18	6.95E-02	8.93E-04	3.67E-03	2.08E-04	7.48E-02	5.15E-02	1.68E-01	2.20E-16	1.24E-02	6.44E-02	1.22E-03	1.72E-01
chr19	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	5.28E-07	2.20E-16	6.46E-12	2.20E-16	2.20E-16
chr20	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	3.06E-03	2.20E-16	2.20E-16
chr21	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
chr22	8.76E-09	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	4.99E-11	2.20E-16	2.20E-16
chrX	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
chrY	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16

ChrID	chr13	chr14	chr15	chr16	chr17	chr18	chr19	chr20	chr21	chr22	chrX	chrY
chr1	2.84E-05	2.20E-16	3.22E-02	2.20E-16	5.90E-02	6.95E-02	2.20E-16	2.20E-16	2.20E-16	8.76E-09	2.20E-16	2.20E-16
chr2	1.06E-01	2.20E-16	2.20E-16	2.20E-16	2.02E-08	8.93E-04	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
chr3	2.60E-01	2.20E-16	2.20E-16	2.20E-16	3.70E-07	3.67E-03	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
chr4	4.21E-02	2.20E-16	2.20E-16	2.20E-16	3.65E-10	2.08E-04	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
chr5	8.91E-01	4.46E-04	2.20E-16	2.20E-16	6.19E-04	7.48E-02	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
chr6	9.62E-01	7.77E-08	2.20E-16	2.20E-16	2.44E-04	5.15E-02	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
chr7	5.55E-01	5.50E-01	2.20E-16	2.20E-16	4.75E-03	1.68E-01	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
chr8	2.63E-14	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	5.28E-07	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
chr9	4.98E-01	2.20E-16	2.20E-16	2.20E-16	7.98E-06	1.24E-02	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
chr10	7.04E-01	2.51E-01	2.53E-05	2.20E-16	5.43E-03	6.44E-02	6.46E-12	3.06E-03	2.20E-16	4.99E-11	2.20E-16	2.20E-16
chr11	1.34E-01	2.20E-16	2.20E-16	2.20E-16	2.68E-08	1.22E-03	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
chr12	5.44E-01	4.39E-01	2.20E-16	2.20E-16	5.02E-03	1.72E-01	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
chr13	1	6.04E-01	8.98E-04	2.20E-16	2.85E-02	1.54E-01	2.82E-11	3.19E-02	6.60E-14	6.45E-08	2.20E-16	2.20E-16
chr14	6.04E-01	1	2.20E-16	2.20E-16	3.60E-03	1.50E-01	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
chr15	8.98E-04	2.20E-16	1	2.20E-16	3.73E-01	2.52E-01	2.20E-16	1.64E-04	2.20E-16	1.94E-07	2.20E-16	2.20E-16
chr16	2.20E-16	2.20E-16	2.20E-16	1	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
chr17	2.85E-02	3.60E-03	3.73E-01	2.20E-16	1	6.56E-01	2.20E-16	4.48E-01	5.52E-10	4.42E-04	2.20E-16	2.20E-16
chr18	1.54E-01	1.50E-01	2.52E-01	2.20E-16	6.56E-01	1	5.58E-15	9.87E-01	1.19E-06	3.12E-03	2.20E-16	2.20E-16
chr19	2.82E-11	2.20E-16	2.20E-16	2.20E-16	2.20E-16	5.58E-15	1	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
chr20	3.19E-02	2.20E-16	1.64E-04	2.20E-16	4.48E-01	9.87E-01	2.20E-16	1	2.20E-16	2.20E-16	2.20E-16	2.20E-16
chr21	6.60E-14	2.20E-16	2.20E-16	2.20E-16	5.52E-10	1.19E-06	2.20E-16	2.20E-16	1	4.87E-10	2.20E-16	2.20E-16
chr22	6.45E-08	2.20E-16	1.94E-07	2.20E-16	4.42E-04	3.12E-03	2.20E-16	2.20E-16	4.87E-10	1	2.20E-16	2.20E-16
chrX	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	1	2.20E-16
chrY	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	1

Table S2. The density of SNPs and INDELs in different genomic regions.

	Total length (kb)	Total SNPs	SNPs/kb	Total INDELs	INDELs/ kb	SNPs+INDELs /kb
Genic region	1816320887	187282994	103.1112	12574699	6.92317	110.03435
Intergenic region	1121280639	118368998	105.5659	6603244	5.88902	111.45492
Coding genes	1762757854	181512741	102.9709	12220035	6.93234	109.90323
Non-coding genes	53563033	5770253	107.7283	354664	6.62143	114.3497

Table S3: Statistical significance of SNP density differences among different genomic regions

Genomic region	**CDS region**	**5' UTR region**	**3' UTR region**	**Promoter region**	**Intergenic region**	**Introns**
CDS region	1*	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
5' UTR region	2.20E-16	1	2.20E-16	2.20E-16	2.20E-16	2.20E-16
3' UTR region	2.20E-16	2.20E-16	1	2.20E-16	2.20E-16	2.20E-16
Promoter region	2.20E-16	2.20E-16	2.20E-16	1	2.20E-16	2.20E-16
Intergenic region	2.20E-16	2.20E-16	2.20E-16	2.20E-16	1	2.20E-16
Introns	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	1
*, p-valu based on independent t-test						

Table S4: The top 5% and bottom 5% of genes with the highest and lowest SNP density

	Total SNPs	High density genes	Low density genes
All SNPs	19, 644	378	636
Missense SNPs	18, 840	340	636

Table S5. The sequence similarity cutoff values for identifying orthologous positions between human and other primate genomes

Primate genome	SNPs	INDELs
Chimpanzee	95. 05*	91. 82
Bonobo	94. 81	91. 41
Gorilla	94. 33	91. 24
Orangutan	91. 81	90. 78
Gibbon	91. 10	90. 23
*, the percent of sequence identity in blat sequence search		

Table S6. Average C->T density per kb between CpG island and non-CpG island regions

	Length	C->T SNPs	Density	p-value
CpG island	23640876	649559	27.4761	
non-CpG island	2913960650	45283608	15.5402	0.00001

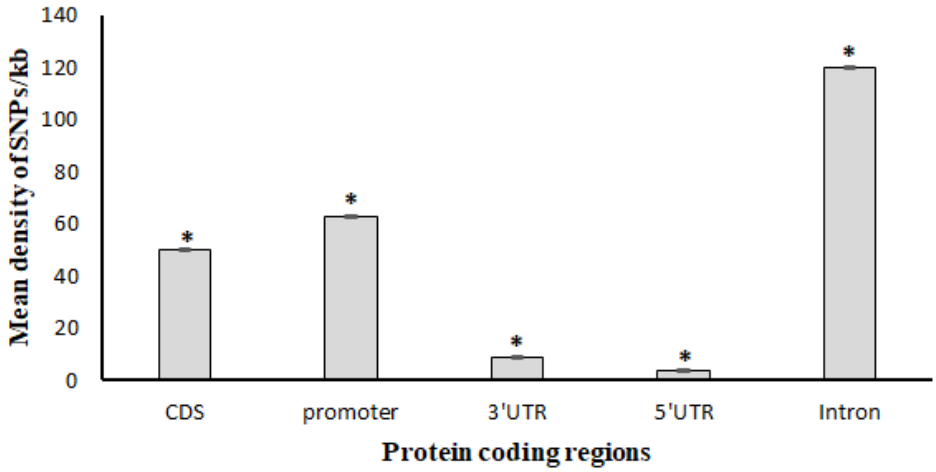

Figure S1: Average density of SNPs in the protein coding region. Bar plots show the mean density of SNPs/kb. The x-axis shows the different regions of the protein coding region. * shows the statistical difference between all the regions with p-value>0.05. The error bars on the bar chart represents standard error.

Reference

1000 Genomes Project Consortium. (2012). An integrated map of genetic variation. *Nature*. https://doi.org/10.1038/nature11632

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*. https://doi.org/10.1038/nature15393

1000 Genomes Project Consortium, G., Altshuler, D., Auton, A., Brooks, L., Durbin, R., Gibbs, R., … McVean, G. (2010). A map of human genome variation from population-scale sequencing. *Nature*. https://doi.org/10.1038/nature09991

Aerts, J., Wetzels, Y., Cohen, N., & Aerssens, J. (2002). Data mining of public SNP databases for the selection of intragenic SNPs. *Human Mutation*. https://doi.org/10.1002/humu.10107

Amos, W., & Hoffman, J. I. (2010). Evidence that two main bottleneck events shaped modern human genetic diversity. *Proceedings of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rspb.2009.1473

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3031

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. https://doi.org/10.1038/nature07517

Bhangale, T. R., Rieder, M. J., Livingston, R. J., & Nickerson, D. A. (2005). Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Human Molecular Genetics*. https://doi.org/10.1093/hmg/ddi006

Chaisson, M. J., Brinza, D., & Pevzner, P. A. (2009). De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Research*. https://doi.org/10.1101/gr.079053.108

Cheetham, S. W., Gruhl, F., Mattick, J. S., & Dinger, M. E. (2013). Long noncoding RNAs and the genetics of cancer. *British Journal of Cancer*. https://doi.org/10.1038/bjc.2013.233

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., … Ma'ayan, A. (2013). Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-14-128

Chen, J. M., Stenson, P. D., Cooper, D. N., & Férec, C. (2005). A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Human Genetics*. https://doi.org/10.1007/s00439-005-1321-0

Choy, K. W., Setlur, S. R., Lee, C., & Lau, T. K. (2010). The impact of human copy number variation on a new era of genetic testing. *BJOG: An International Journal of Obstetrics and Gynaecology*. https://doi.org/10.1111/j.1471-0528.2009.02470.x

Church, G. M. (2006). The Personal Genome Project. *Molecular Systems Biology*. https://doi.org/10.1038/msb4100040

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … 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*. https://doi.org/10.4161/fly.19695

Clague, M. J., & Urbé, S. (2010). Ubiquitin: Same molecule, different degradation pathways. *Cell*. https://doi.org/10.1016/j.cell.2010.11.012

Clarke, L., Zheng-Bradley, X., Smith, R., Kulesha, E., Xiao, C., Toneva, I., … Flicek, P. (2012). The 1000 Genomes Pproject: Data management and community access. *Nature Methods*. https://doi.org/10.1038/nmeth.1974

Collins, D. W., & Jukes, T. H. (1994). Rates of transition and transversion in coding sequences since the human- Rodent divergence. *Genomics*. https://doi.org/10.1006/geno.1994.1192

Collins, F. S., & McKusick, V. A. (2001). Implications of the human genome project for medical science. *Journal of the American Medical Association*. https://doi.org/10.1001/jama.285.5.540

Consortium, G. O. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkh036

Consortium, H. G. S., Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., … Chen, Y. J. (2001). Initial sequencing and analysis of the human genome. *Nature*. https://doi.org/10.1038/35057062

Cook, E. H., & Scherer, S. W. (2008). Copy-number variations associated with neuropsychiatric conditions. *Nature*. https://doi.org/10.1038/nature07458

Coordinators, N. R., T., B., J., B., D.A., B., C., B., E., B., … Coordinators, N. R. (2015). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gku1130

Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2640

Cordovado, S. K., Hendrix, M., Greene, C. N., Mochal, S., Earley, M. C., Farrell, P. M., … Mueller, P. W. (2012). CFTR mutation analysis and haplotype associations in CF patients. *Molecular Genetics and Metabolism*. https://doi.org/10.1016/j.ymgme.2011.10.013

de Jong, S., Chepelev, I., Janson, E., Strengman, E., van den Berg, L. H., Veldink, J. H., & Ophoff, R. A. (2012). Common inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue-specific manner. *BMC Genomics*. https://doi.org/10.1186/1471-2164-13-458

Deloukas, P., & Bentley, D. (2004). The HapMap project and its application to genetic studies of drug response. *Pharmacogenomics Journal*. https://doi.org/10.1038/sj.tpj.6500226

Dolgin, E. (2009). Human genomics: The genome finishers. *Nature*. https://doi.org/10.1038/462843a

Frazer, K. A., Murray, S. S., Schork, N. J., & Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2554

Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., … Akey, J. M. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*. https://doi.org/10.1038/nature11690

Furuno, M., Kasukawa, T., Saito, R., Adachi, J., Suzuki, H., Baldarelli, R., … Okazaki, Y. (2003). CDS annotation in full-length cDNA sequence. *Genome Research*. https://doi.org/10.1101/gr.1060303

Giorda, R., Bonaglia, M. C., Beri, S., Fichera, M., Novara, F., Magini, P., … Zuffardi, O. (2009). Complex Segmental Duplications Mediate a Recurrent dup(X)(p11.22-p11.23) Associated with Mental Retardation, Speech Delay, and EEG Anomalies in Males and Females. *American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2009.08.001

Guhaniyogi, J., & Brewer, G. (2001). Regulation of mRNA stability in mammalian cells. *Gene*. https://doi.org/10.1016/S0378-1119(01)00350-X

Guo, Y., Cai, Q., Samuels, D. C., Ye, F., Long, J., Li, C. I., … Boice, J. D. (2012). The use of next generation sequencing technology to study the effect of radiation therapy on mitochondrial DNA mutation. *Mutation Research - Genetic Toxicology and Environmental Mutagenesis*. https://doi.org/10.1016/j.mrgentox.2012.02.006

Guo, Y., Dai, Y., Yu, H., Zhao, S., Samuels, D. C., & Shyr, Y. (2017). Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*. https://doi.org/10.1016/j.ygeno.2017.01.005

Guo, Y., Sheng, Q., Li, J., Ye, F., Samuels, D. C., & Shyr, Y. (2013). Large Scale Comparison of Gene Expression Levels by Microarrays and RNAseq Using TCGA Data. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0071462

Hancks, D. C., & Kazazian, H. H. (2012). Active human retrotransposons: Variation and disease. *Current Opinion in Genetics and Development*. https://doi.org/10.1016/j.gde.2012.02.006

Handsaker, R. E., Korn, J. M., Nemesh, J., & McCarroll, S. A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature Genetics*. https://doi.org/10.1038/ng.768

Hill, K. A., Wang, J., Farwell, K. D., & Sommer, S. S. (2003). Spontaneous tandem-base mutations (TBM) show dramatic tissue, age, pattern and spectrum specificity. *Mutation Research - Genetic Toxicology and Environmental Mutagenesis*. https://doi.org/10.1016/S1383-5718(02)00277-2

Ingram, V. M. (1956). A specific chemical difference between the globins of normal human and sickle-cell anæmia hæmoglobin. *Nature*. https://doi.org/10.1038/178792a0

Ionita-Laza, I., Rogers, A. J., Lange, C., Raby, B. A., & Lee, C. (2009). Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics*. https://doi.org/10.1016/j.ygeno.2008.08.012

Jabbari, K., & Bernardi, G. (2004). Cytosine methylation and CpG, TpG (CpA) and TpA

frequencies. *Gene*. https://doi.org/10.1016/j.gene.2004.02.043

Janeway, C. A., Travers, P., Walport, M., & Shlomchik, M. (2001). The immune system in health and disease. Immunobiology. *Gerald Publishing, New York Jerne NK (1955) The Natural Selection Theory of Antibody Formation. In: Proc. Natl. Acad. Sci. USA*.

Kamaraj, B., Rajendran, V., Sethumadhavan, R., Kumar, C. V., & Purohit, R. (2015). Mutational analysis of FUS gene and its structural and functional role in amyotrophic lateral sclerosis 6. *Journal of Biomolecular Structure and Dynamics*. https://doi.org/10.1080/07391102.2014.915762

Kaneko, T., Tahara, S., & Matsuo, M. (1996). Non-linear accumulation of 8-hydroxy-2'-deoxyguanosine, a marker of oxidized DNA damage, during aging. *Mutation Research - DNAging Genetic Instability and Aging*. https://doi.org/10.1016/S0921-8734(96)90010-7

Karolchik, D., Barber, G. P., Casper, J., Clawson, H., Cline, M. S., Diekhans, M., … Kent, W. J. (2014). The UCSC Genome Browser database: 2014 update. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkt1168

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research*. https://doi.org/10.1101/gr.229202

Kidwell, M. G., & Lisch, D. (2002). Transposable elements as sources of variation in animals and plants. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.94.15.7704

Kimchi-Sarfaty, C., Oh, J. M., Kim, I. W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V., & Gottesman, M. M. (2007). A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science*. https://doi.org/10.1126/science.1135308

Kuehner, J. N., Pearson, E. L., & Moore, C. (2011). Unravelling the means to an end: RNA polymerase II transcription termination. *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm3098

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*. https://doi.org/10.1038/35057062

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., … Maglott, D. R. (2016). ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv1222

Lee, C., & Scherer, S. W. (2010). The clinical context of copy number variation in the human genome. *Expert Reviews in Molecular Medicine*. https://doi.org/10.1017/S1462399410001390

Lek, M., Karczewski, K. J., Minikel, E. V, Samocha, K. E., Banks, E., Fennell, T., … Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. https://doi.org/10.1038/nature19057

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., … Venter, J. C. (2007). The diploid genome sequence of an individual human. *PLoS Biology*. https://doi.org/10.1371/journal.pbio.0050254

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., … Myers, R. M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science*. https://doi.org/10.1126/science.1153717

Li, W. H., & Sadler, L. A. (1991). Low nucleotide diversity in man. *Genetics*.

Logofet, D. O., & Svirezhev, Y. M. (1980). The model for human population dynamics as a part of the global biosphere model: Some aspects of modelling in a dialogue regime. *Ecological Modelling*. https://doi.org/10.1016/0304-3800(80)90021-6

Lunshof, J. E., Bobe, J., Aach, J., Angrist, M., Thakuria, J. V., Vorhaus, D. B., … Church, G. M. (2010). Personal genomes in progress: From the human genome project to the Personal Genome Project. *Dialogues in Clinical Neuroscience*.

MacArthur, D. G., & Tyler-Smith, C. (2010). Loss-of-function variants in the genomes of healthy humans. *Human Molecular Genetics*. https://doi.org/10.1093/hmg/ddq365

Manolio, T. A., & Collins, F. S. (2009). The HapMap and Genome-Wide Association Studies in Diagnosis and Therapy. *Annual Review of Medicine*. https://doi.org/10.1146/annurev.med.60.061907.093117

McGuire, A. L., Cho, M. K., McGuire, S. E., & Caulfield, T. (2007). The future of personal genomics. *Science*. https://doi.org/10.1126/science.1147475

Men, A. E., Wilson, P., Siemering, K., & Forrest, S. (2008). Sanger DNA Sequencing. In *Next Generation Genome Sequencing: Towards Personalized Medicine*. https://doi.org/10.1002/9783527625130.ch1

Merryweather-Clarke, A. T., Pointon, J. J., Shearman, J. D., & Robson, K. J. (1997). Global prevalence of putative haemochromatosis mutations. *Journal of Medical Genetics*.

Miklos, G. L. G., & Rubin, G. M. (1996). The role of the genome project in determining gene function: Insights from model organisms. *Cell*. https://doi.org/10.1016/S0092-8674(00)80126-9

Montgomery, S. B., Goode, D. L., Kvikstad, E., Albers, C. A., Zhang, Z. D., Mu, X. J., … Lunter, G. (2013). The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Research*. https://doi.org/10.1101/gr.148718.112

Moyzis, R. K., Buckingham, J. M., Cram, L. S., Dani, M., Deaven, L. L., Jones, M. D., … Wu, J. R. (1988). A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*.

Muthuvel, A., Ravindran, M., Chander, A., & Subbian, C. (2016). Pericentric inversion of chromosome 9 causing infertility and subsequent successful in vitro fertilization. *Nigerian Medical Journal*. https://doi.org/10.4103/0300-1652.182080

Nguyen, D. Q., Webber, C., & Ponting, C. P. (2006). Bias of selection on human copy-number variants. *PLoS Genetics*. https://doi.org/10.1016/j.mssp.2006.01.060

Ober, C., Aldrich, C. L., Chervoneva, I., Billstrand, C., Rahimov, F., Gray, H. L., & Hyslop, T. (2003). Variation in the HLA-G Promoter Region Influences Miscarriage Rates. *The American Journal of Human Genetics*. https://doi.org/10.1086/375501

P., Z., D.C., S., B., L., T., S., J., P., & Y., S. (2016). Mitochondria sequence mapping strategies and practicability of mitochondria variant detection from exome and RNA sequencing data. *Briefings in Bioinformatics*. https://doi.org/http://dx.doi.org/10.1093/bib/bbv057

Pagani, F., & Baralle, F. E. (2004). Genomic variants in exons and introns: Identifying the splicing spoilers. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg1327

Palazzo, A. F., & Lee, E. S. (2015). Non-coding RNA: What is functional and what is junk? *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2015.00002

Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., … Scherer, S. W. (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome Biology*. https://doi.org/10.1186/gb-2010-11-5-r52

Paulino, D., Warren, R. L., Vandervalk, B. P., Raymond, A., Jackman, S. D., & Birol, I. (2015). Sealer: A scalable gap-closing application for finishing draft genomes. *BMC Bioinformatics*. https://doi.org/10.1186/s12859-015-0663-4

Payer, L. M., Steranka, J. P., Yang, W. R., Kryatova, M., Medabalimi, S., Ardeljan, D., … Burns, K. H. (2017). Structural variants caused by Alu insertions are associated with risks for many human diseases . *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1704117114

Pop, M., Kosack, D. S., & Salzberg, S. L. (2004). Hierarchical scaffolding with Bambus. *Genome Research*. https://doi.org/10.1101/gr.1536204

Pop, M., & Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. *Trends in Genetics*. https://doi.org/10.1016/j.tig.2007.12.006

Quinlan, A. R. (2014). BEDTools: The Swiss-Army tool for genome feature analysis. *Current Protocols in Bioinformatics*. https://doi.org/10.1002/0471250953.bi1112s47

Raitio, M., Lindroos, K., Laukkanen, M., Pastinen, T., Sistonen, P., Sajantila, A., & Syvänen, A. C. (2001). Y-chromosomal SNPs in finno-ugric-speaking populations analyzed by minisequencing on microarrays. *Genome Research*. https://doi.org/10.1101/gr.156301

Rao, B., Kerketta, L., Korgaonkar, S., & Ghosh, K. (2009). Pericentric inversion of chromosome 9[inv(9)(p12q13)]: Its association with genetic diseases. *Indian Journal of Human Genetics*. https://doi.org/10.4103/0971-6866.29856

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., … Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*. https://doi.org/10.1038/nature05329

Reich, D. E., Gabriel, S. B., & Altshuler, D. (2003). Quality and completeness of SNP databases. *Nature Genetics*. https://doi.org/10.1038/ng1133

Rosenfeld, J. A., Mason, C. E., & Smith, T. M. (2012). Limitations of the human reference genome for personalized genomics. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0040294

Rotimi, C. N., & Jorde, L. B. (2010). Ancestry and Disease in the Age of Genomic Medicine. *New England Journal of Medicine*. https://doi.org/10.1056/nejmra0911564

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., …

Altshuler, D. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*. https://doi.org/10.1038/35057149

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*.

Sauer, R. T., Yocum, R. R., Doolittle, R. F., Lewis, M., & Pabo, C. O. (1982). Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. *Nature*. https://doi.org/10.1038/298447a0

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., … Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*. https://doi.org/10.1101/gr.213611.116

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., … Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science*. https://doi.org/10.1126/science.1098918

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*. https://doi.org/10.1101/gr.089532.108

Singh, M., Singh, P., Juneja, P. K., Singh, S., & Kaur, T. (2011). SNP-SNP interactions within APOE gene influence plasma lipids in postmenopausal osteoporosis. *Rheumatology International*. https://doi.org/10.1007/s00296-010-1449-7

SJÖDIN, J. (1971). Induced paracentric and pericentric inversions in Vicia faba L. *Hereditas*. https://doi.org/10.1111/j.1601-5223.1971.tb02357.x

Spencer, C. C. A., Deloukas, P., Hunt, S., Mullikin, J., Myers, S., Silverman, B., … McVean, G. (2006). The influence of recombination on human genetic diversity. *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.0020148

Sprouse, R. O., Karpova, T. S., Mueller, F., Dasgupta, A., McNally, J. G., & Auble, D. T. (2008). Regulation of TATA-binding protein dynamics in living yeast cells. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.0801901105

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazlsy, C., Thorne, N., … Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene phenotypes. *Science*. https://doi.org/10.1126/science.1136678

Swersky, R. B., Chang, J. B., Wisoff, B. G., & Gorvoy, J. (1979). Endobronchial Balloon Tamponade of Hemoptysis in Patients with Cystic Fibrosis. *Annals of Thoracic Surgery*. https://doi.org/10.1016/S0003-4975(10)63289-4

Syvanen, A. C. (2005). Toward genome-wide snp genotyping. *Nature Genetics*. https://doi.org/10.1038/ng1558

Takai, D., & Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences*.

https://doi.org/10.1073/pnas.052410099

Tang, W., Mun, S., Joshi, A., Han, K., & Liang, P. (2018). Mobile elements contribute to the uniqueness of human genome with 15,000 human-specific insertions and 14 Mbp sequence increase. *DNA Research*. https://doi.org/10.1093/dnares/dsy022

Tantoso, E., Yang, Y., & Li, K. B. (2006). How well do HapMap SNPs capture the untyped SNPs? *BMC Genomics*. https://doi.org/10.1186/1471-2164-7-238

Telenti, A., Pierce, L. C. T., Biggs, W. H., di Iulio, J., Wong, E. H. M., Fabani, M. M., … Venter, J. C. (2016). Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1613365113

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., … Deloukas, P. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*. https://doi.org/10.1038/ng1946

Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M., … Eichler, E. E. (2005). Fine-scale structural variation of the human genome. *Nature Genetics*. https://doi.org/10.1038/ng1562

Väli, U., Brandström, M., Johansson, M., & Ellegren, H. (2008). Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. *BMC Genetics*. https://doi.org/10.1186/1471-2156-9-8

Venter, J. C., Adams MD Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, M. E. W., Zhang Q Zheng XQH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang JH, Miklos GLG, Nelson C, Broder S, Clark AG, Nadeau C, McKusick VA, Zinder, K. C. D., N Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, L. A. J., Mobarry C Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng ZM, Di, R. K., Francesco V Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge WM, Gong FC, Gu ZP, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke ZX, Ketchum KA, Lai ZW, Lei, D. P., … M Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen MY, Wu D, Wu M, Xia A, Zandieh A, Zhu XH, P. S. (2001). The sequence of the human genome. *Science*. https://doi.org/10.1126/science.1058040

Vinay Kumar, C., Kumar, K. M., Swetha, R., Ramaiah, S., & Anbarasu, A. (2014). Protein aggregation due to nsSNP resulting in P56S VABP protein is associated with amyotrophic lateral sclerosis. *Journal of Theoretical Biology*. https://doi.org/10.1016/j.jtbi.2014.03.027

Wang, L., & Jordan, I. K. (2018). Transposable element activity, genome regulation and human health. *Current Opinion in Genetics and Development*. https://doi.org/10.1016/j.gde.2018.02.006

Wapinski, O., & Chang, H. Y. (2011). Long noncoding RNAs and human disease. *Trends in Cell Biology*. https://doi.org/10.1016/j.tcb.2011.04.001

Wei, C. H., Phan, L., Feltz, J., Maiti, R., Hefferon, T., & Lu, Z. (2018). TmVar 2.0: Integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btx541

Wetterbom, A., Sevov, M., Cavelier, L., & Bergström, T. F. (2006). Comparative genomic analysis of human and chimpanzee indicates a key role for indels in primate evolution. *Journal of Molecular Evolution*. https://doi.org/10.1007/s00239-006-0045-7

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., … Rothberg, J. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*. https://doi.org/10.1038/nature06884

Xing, J., Watkins, W. S., Witherspoon, D. J., Zhang, Y., Guthery, S. L., Thara, R., … Jorde, L. B. (2009). Fine-scaled human genetic structure revealed by SNP microarrays. *Genome Research*. https://doi.org/10.1101/gr.085589.108

Xing, J., Zhang, Y., Han, K., Salem, A. H., Sen, S. K., Huff, C. D., … Jorde, L. B. (2009). Mobile elements create structural variation: Analysis of a complete human genome. *Genome Research*. https://doi.org/10.1101/gr.091827.109

Yang, Z., & Bielawski, J. R. (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution*. https://doi.org/10.1016/S0169-5347(00)01994-7

Ye, F., Samuels, D. C., Clark, T., & Guo, Y. (2014). High-throughput sequencing in mitochondrial DNA research. *Mitochondrion*. https://doi.org/10.1016/j.mito.2014.05.004

Yu, N., Jensen-Seaman, M. I., Chemnick, L., Ryder, O., & Li, W. H. (2004). Nucleotide Diversity in Gorillas. *Genetics*. https://doi.org/10.1534/genetics.166.3.1375

Zamir, A., Springer, B., & Glattstein, B. (2015). Fingerprints and DNA: STR Typing of DNA Extracted from Adhesive Tape after Processing for Fingerprints. *Journal of Forensic Sciences*. https://doi.org/10.1520/jfs14749j

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*. https://doi.org/10.1101/gr.074492.107