Genome structural variation in human evolution

Edward J. Hollox (1), Luciana W. Zuccherato (2,3), Serena Tucci (4)

1. Department of Genetics and Genome Biology, University of Leicester, UK
2. Núcleo de Ensino e Pesquisa, Instituto Mário Penna, Belo Horizonte, Brazil
3. Departmento de Bioquímica e Imunologia, Universidade de Minas Gerais, Belo Horizonte, Brazil.
4. Department of Anthropology, Yale University, New Haven CT, USA

Corresponding author:

Dr Ed Hollox, University of Leicester

https://tinyurl.com/holloxgroup

@edhollox

Abstract

Structural variation (SV) is a large difference (typically >100bp) in the genomic structure of two genomes and includes both copy number variation (CNV) and variation that does not change copy number of a genomic region, such as an inversion. Improved reference genomes, combined with widespread genome sequencing using short read sequencing technology, and increasingly using long read sequencing, have re-ignited interest in SV. Recent large scale studies and functional focused analyses have highlighted the role of SV in human evolution. In this review we highlight human-specific SVs involved in changes in the brain, population-specific SVs that affect response to the environment, including dietary adaptation and adaptation to infectious diseases, and summarise the contribution of archaic hominin admixture to present-day human SV.

Keywords:

Structural variation, human evolution, adaptive evolution, introgression, CNV, infection

What is structural variation, and how do we measure it?

For a geneticist, evolution is the change of allele frequencies over time. Most studies of evolutionary genomics focus on single nucleotide variants (SNVs) but structural variants (SVs) are frequently neglected yet can confer important differences in phenotype between individuals, and result in evolutionary adaptations. Structural variants are differences in large sections of DNA, typically >100bp, between different genomes. The term structural variation includes several different types of variant, such as deletions, duplications, and more complex rearrangements such as large tandem repeats, inversions and retrotransposon insertions

(Figure 1). SVs that result in a change in the number of copies of a DNA sequence (DNA dosage) are often called copy number variants, or CNVs.

The development of array comparative hybridisation (aCGH), able to detect copy number changes in large numbers of individuals, revealed the extent of SV across the human genome [1]. Hybridisation intensity data from SNV genotyping chips was also used to detect copy number changes. Data from SNV genotyping chips were readily accessible and generated as part of standard SNV genotyping for genomewide association studies, but reproducibility across SV studies using SNP chip hybridisation data was found to be low [2]. Specialised microarray chips for aCGH, using either large-insert cones or oligonucleotides, generated data that were more reproducible, but had their own limits in terms of the sensitivity and specificity of detecting SVs of different sizes. High throughput sequencing improved reliability of SV detection, with three main approaches used today: sequence read-depth, split-read and discordant mate-pair (Figure 1). Challenges in analysing short-read sequence data for SVs remain, due to limitations in mapping these to a single complex reference genome, which is rich in >95% similar duplicated sequence. Incorrect mapping of sequences to the wrong copy of the duplicated region in the reference sequence can introduce or remove apparent SNV, and disrupt detection of SVs by any of the three approaches shown in figure 1.

The development of sequencing technologies from Pacific Biosciences and Oxford Nanopore that generate long sequence reads has accelerated SV discovery. Long sequence reads can either be mapped to the reference genome spanning complex regions, or contain sufficient sequence information to allow unambiguous mapping. Large consortia are generating data from human individuals using long-read sequencing, and making the data publicly accessible. For example, the Human PanGenome Reference Consortium (HPRC) aims to generate a diverse set of human reference genomes [3], by focusing on telomere-to-telomere assemblies of human chromosomes [4]. Furthermore, the Human Genome Structural Variation Consortium (HGSVC) has published 32 diploid genomes generated by long read sequencing and strand specific sequencing [5]. It suggests that although over 75% of SVs called by short-read approaches are detected in the long-read assembly approaches, only 30% of SVs detected by long-reads are detected by short-read approaches, with most of the SVs missed by short read approaches being smaller than 5kb.

Despite the advances in long-read technology, many more genomes have been sequenced using short read approaches, so SV detection using short sequencing reads is still very useful [6].

How much structural variation do human genomes have?

We now have a good estimate of the extent and nature of SVs in the human genome. As part of The 1000 Genomes Project, 2504 individuals from 26 populations were sequenced using Illumina short reads at medium sequencing coverage depth (~7.4x) providing a foundation for our understanding of SV across geographically diverse populations [7]. The population allele frequency distribution of SVs reflects that of SNVs - most variants are rare, with greater heterozygosity in sub-Saharan African populations [8]. This emphasises the importance of genetic drift and demographic processes in shaping the overall pattern of SV variation that we see today. Patterns of extreme population differences in allele frequency can be generated by directional selection in a particular geographical region or environment [9], but can also be generated by genetic drift. The influence of the high mutation rate of some SVs in generating large differences in allele frequency remains poorly understood . Despite this, by identifying SVs that show large differences in allele frequency in different populations, we can identify candidates for recent directional selection. In the 1000 Genomes cohort, 578 SVs show high allele frequency differences between populations. Genes involved in xenobiotic metabolism and cell-cell adhesion are overrepresented in multiallelic CNVs and duplications respectively, suggesting adaptation to novel environments [7].

A combination of short read sequencing and 10x Genomics linked read technology on 911 individuals from 54 populations from the Human Genome Diversity Panel identified many more

SVs, emphasising the importance of broad sampling across different human populations [10,11]. Moreover, an excess of SVs is shared between the genomes of present-day populations outside Africa, and those of archaic extinct hominins, such as Neanderthals and Denisovans [12,13] (Box 1).

How can structural variation affect function?

Structural variation can affect gene expression in a number of ways. The most straightforward is gain by duplication or loss by deletion of whole genes, causing a gene dosage effect, increasing or decreasing gene expression levels and levels of the encoded protein. An extreme example of this is homozygous gene deletion, causing complete loss of the transcript and encoded protein. However, other types of SV can have more subtle effects, for example the multiallelic CNV of the beta-defensin genes, where increased copy number leads to increase in mRNA expression, in protein levels secreted at the mucosal surface, and in an increase in antimicrobial efficacy [14,15]. This gene dosage effect has been observed more broadly in multiallelic CNV [16]. Alternatively, SV can indirectly affect expression levels of a gene by altering the spatial relationship between a regulatory element and a gene. This has been observed in mice at the HoxD locus, where SVs disrupt chromatin structure, exerting influence on genes at a distance from the SV [17,18]. SV can also generate novel fusion genes, which is seen at the genes encoding the immunity-related butyrophilins in humans, where a 56 kb deletion generates a novel fusion gene [19]. SV can also involve exons within a gene, and can either result in disrupting the reading frame of the transcript completely, or result in loss or gain of parts of the protein. An example is the *CR1* gene encoding complement receptor 1, where CNV of an 18 kb segment leads to extra copies of a block of 8 exons encoding extra complement-binding protein domains [20].

By comparing gene expression levels with SVs across different individuals, it is possible to correlate genotypes of SV with different expression levels of a particular gene, and define a particular SV as an expression quantitative trait locus (eQTL). Gene expression levels were measured in lymphoblastoid cell lines in a subset of the 1000 Genomes samples, and correlated to genotypes of SVs genomewide [5]. Expression levels of 1,500 genes are significantly correlated with the genotype of at least one SV. Similar approaches using induced pluripotent stem cells found similar levels of importance of SVs and gene expression, also showing that larger SVs are more likely to be an eQTL [21]. Many of these SVs that are eQTLs overlap distal chromatin loop anchors, suggesting that localised disruption of chromatin structure may be an important mechanism of how SVs affect gene expression levels. In a study involving measuring eQTLs of 13 tissues, SVs accounted for up to 6.8% of all eQTLs, and had on average larger effects than SNV eQTLs [22]. These studies show the importance of SVs in affecting expression levels of genes across cell types, and therefore highlight the functional roles of SVs.

Why should we look to structural variation as a source of adaptive variants?

At first glance, structural variation is just another class of genomic variant, and subject to the same rules of population genetics as single nucleotide variants (SNVs) (Box 2). These rules include a presumption (null hypothesis) that the allele frequency of the genetic variant changes by genetic drift effects alone, and is therefore evolving neutrally, with selection having no effect on the allele frequency. Early studies on SVs in humans suggested this, with the absence of negative selection allowing extensive gain of copy number of olfactory receptor (OR) genes encoding receptors involved in taste and smell [23]. Biases in gene content and SV distribution that have been attributed to selection [1] may be due to relaxation of functional constraint and therefore relaxation of negative selection [24], but SV ascertainment bias from aCGH data means that such inferences are not certain. Although relaxation from negative selection, and subsequent neutral evolution of SV, are likely to be major forces in many SVs; it does not exclude the possibility that some SVs are adaptive.

We know from medical genetics that structural variants disrupting genes can cause disease. For example, deletions and duplications involving *BRCA1* account for up to 40% of all inherited variants in that gene that cause breast cancer in some populations [25]. At a broader, population genetic level, the functional effect of structural variants is emphasised by a negative correlation between frequency of copy number variants and gene density [26].

There are examples of structural variants underlying adaptations in other animals. Large inversions can generate supergenes, where multiple genes cause an adaptive phenotype, with individuals heterozygous for the inversion unable to recombine across that region preventing breakdown of the set of co-adapted alleles [27,28]. In domesticated mammals, structural variants are responsible for some traits that have been selected by humans. For example, the number of copies of the pancreatic amylase gene *AMY2B* has increased in domestic dogs compared to wolves [29]. Because amylase is responsible for the first stage of digestion of amylose, the major component of starch, it is likely that this is an adaptation to a changing diet during the domestication of dogs and other animals [30,31].

Which structural variants help distinguish humans from other apes?

The largest differences between the human and chimpanzee genomes is at the ckromosomal level, where two pericentric inversions and a chromosome fusion, generating human chromosome 2, have occurred in the human lineage. A comparison between the sequences of reference chimpanzee genome and the human genome reveals 1.66 million variants, 38,773 non-synonymous substitutions and SVs accounting for 3% of genome differences [32]. Despite this extensive catalogue of information, in most cases it is still unclear which variants contribute to producing traits, such as larger brains and reduced muscular strength, that distinguish humans from other great apes. Structural variants that are unique to humans has been shown to have functional consequences that contribute to the differences between humans and other apes.

Male chimpanzees have penises studded with small hard spines, and it is thought that these have been lost in humans either because of reduced sperm competition or female sexual selection [33]. This developmental change is due to a 60 kb deletion near the Androgen Receptor (*AR*) gene on the X chromosome. This deletion removes a regulatory region which, in mice, is responsible for expression of the *AR* gene in penile spines. This regulatory region is also responsible for the development of sensory vibrissae - whiskers - that are absent in humans but present in other great apes [33,34].

The best evidence for SV being involved in human speciation is from variants involved in neuronal function and brain morphology. All mammalian brains have a cerebral cortex, the outer layer of the brain which in humans has a characteristic wrinkled appearance. The neocortex expanded rapidly during human evolution, and is three times the size of a chimpanzee's neocortex, folded to fit inside the human skull. It is thought that the neocortex is responsible for functions such as reasoning, language and sensory perception. The examples introduced below show that human-specific SVs contribute to the changes seen in the human neocortex.

The gene SLIT-ROBO Rho-GTPase-activating protein 2 (*SRGAP2*) has been repeatedly duplicated in the human lineage to form one full-length copy (*SRGAP2A*) and three partial copies in the human genome [35](Figure 2). Of these three partial copies, only *SRGAP2C* has a function, and *SRGAP2B* and *SRGAP2D* are pseudogenes. SRGAP2C heterodimerises and sequesters SRGAP2A, reducing the amount of SRGAP2A protein that can homodimerise. The SRGAP2A homodimer promotes maturation of dendritic spines in the brain neocortex, reducing the amount of this dimer in mice causes slower maturation and an increased density of dendritic spines, mirroring a characteristic of the human neocortex [36]. Dendritic spines enable connections of neurons with neighbouring excitatory axons, enhancing synapse connectivity in the brain neocortex [37]. Therefore, human-specific SRGAP2C slows development of dendritic spines in a form of neoteny, enabling higher plasticity of the brain neocortex.

Another duplicated gene is important in defining the distinct human brain neocortex. The *ARHGAP11B* gene arose by partial duplication from the *ARHGAP11A* gene in the human lineage

[38]. *ARHGAP11A* encodes a Rho-guanosine triphosphatase–activating protein (Rho_GAP), which is involved in cell regulation. ARHGAP11B is a truncated version of ARHGAP11A, does not function as a Rho-GAP, and does not inhibit ARHGAP11A activity, in contrast to the mechanism of SRGAP2C. Instead, ARGHGAP2B increases the number of basal progenitor cells in the neocortex in mice, and this novel function might be due to a small deletion introducing a frameshift, leading to a unique C-terminal end protein sequence. In the ferret brain and marmoset brain, human ARHGAP11B drives increases in the size of the neocortex, and in marmoset induces cortical folding, again mirroring results in mice [39,40].

Another set of duplications have generated new genes encoding human-specific proteins with neuronal function. The *NOTCH2* gene encodes for a receptor involved in signalling in response to ligand-binding and, as part of the Notch signalling system, has a variety of roles in neuronal development [41]. Prior to human-gorilla divergence, duplications and single nucleotide mutations generated multiple *NOTCH2NL* pseudogenes. However, in humans, three of these *NOTCH2NL* paralogues underwent gene conversion from the intact *NOTCH2* gene, reactivating them to form three human-specific paralogs *NOTCH2NLA, NOTCH2NLB, NOTCH2NLC*. These paralogues encode novel, functional NOTCH2NL proteins. Echoing the SRGAP2 mechanism, the human-specific *NOTCH2NL* paralogues were shown in mice and human organoids to delay neurogenesis by downregulating neuronal differentiation genes. The NOTCH2NL proteins appear to do this either by interacting directly with NOTCH2 increasing its signalling activity, or by blocking the expression of Notch-receptors on the cell membrane [42,43].

Taken together, a suite of human-specific SVs has produced important changes to the brain during human evolution. The mechanisms are all inhibitory, in that they slow or delay a particular aspect of development allowing the neocortex to increase in size, concomitant with a relative increase in brain case size (Figure 2b). This is an example of neoteny, long suggested to be important in human evolution.

Have structural variants allowed humans to adapt to different environments?

As humans have spread around the world, they have encountered different environments to which they adapted, culturally and genetically. Studies of selection have identified SNVs responsible for some adaptations in particular populations, such as the well-established persistence of the intestinal enzyme lactase as an adaptation to drinking fresh milk in pastoralist populations [44], and alteration in fatty acid metabolism caused by changes in fatty acid desaturases [45–47]. Both are genetic responses to diet, and are examples of gene-culture evolution, where a change in culture alters the environment, favouring selection of a novel allele.

SNVs also reveal that infectious diseases have contributed to shape population-specific adaptation. Examples include the Duffy negative allele that confers increased resistance to malaria caused by *Plasmodium vivax* [48], and the selection shown at the *LARGE* gene for an allele providing increased resistance to viral Lassa fever [49]. Where population-specific adaptations have been proposed for particular SVs, almost all have involved adaptations to diet or disease (Table 1).

Structural variation and infectious disease

Malaria is a disease borne by Anopheles mosquitoes, and caused by the protist parasites *Plasmodium falciparum* and *Plasmodium vivax*, and is likely to have originated from zoonoses within the past 10,000 years [50]. Malaria infection has been a strong agent of natural selection in the genome, and remains so today, as it is a major cause of childhood mortality in Africa. Alpha-globin, a component of haemoglobin, is encoded by two very closely related paralogous genes, *HBA1* and *HBA2*. Most people have two copies of both genes per diploid genome, but deletion and duplication alleles exist, generated by non-allelic homologous recombination (NAHR) between the two genes [51,52]. The deletion alleles are maintained at higher frequencies in sub-Saharan Africa (up to 20%) because they confer protection against severe malaria [53].

An SV involving two glycophorin genes also confers significant protection against severe malaria. Glycophorins are red blood cell surface receptors and two, glycophorin A and glycophorin B, are proteins used by *P. falciparum* to infect red blood cells. The reference genome has glycophorin A (*GYPA*), glycophorin B (*GYPB*) and the pseudogene *GYPE* on three 120kb tandem repeats (Figure 3), with NAHR between these repeats generating different deletions and duplications affecting one or more of the three glycophorin genes [54,55]. A duplication variant DUP4, known as the Dantu variant, is localised to East Africa, and is protective against severe malaria with a clinically significant odds ratio of between 0.5 and 0.72 [54]. The DUP4 variant has arisen recently and increased in frequency at a rate consistent with strong, recent, positive selection, according to extended haplotype tests [54]. The structure of the DUP4 variant results in two *GYPB-GYPA* fusion genes which encode a fusion protein at the cell surface (figure 3) [54,56]. Rather than directly limiting parasite infection, the fusion protein alters the surface tension of the red blood cell, resulting in a stiffer cell membrane, making the red blood cell less able to envelop the merozoite prior to cell entry and therefore increasing cell resistance to invasion by *P. falciparum* infectious merozoites [57].

As previous studies (Table 1) and genome-wide analyses have shown, often SV affects genes that are involved in the immune response to viral or bacterial infections. An example is *DMBT1*, a large secreted glycoprotein [58] with scavenger-receptor cysteine-rich (SRCR) domains, arranged in tandem, which bind bacteria and a variety of host molecules. The exons encoding the SRCR domains vary in copy number, and generate glycoproteins of different lengths, which differ in their ability to bind bacteria, including *Staphylococcus mutans*, the cause of dental caries [59]. Strong signatures of balancing selection have been identified at the start of *DMBT1* [61–63], and because of a correlation between DMBT1 SRCR domain copy number and historical dietary changes across populations, susceptibility to dental caries may drive evolution at this gene [60].

One particular gene family, those encoding Sialic acid-binding immunoglobulin-type lectins

(SIGLECs), have undergone many independent changes during primate evolution, with some important functional consequences [64]. SIGLECs encode lectins which bind specific sugar moieties on the surface of cells called sialic acids. SIGLEC5 and SIGLEC14 are paired receptors which show similar substrate-specificity yet cause opposite responses, with SIGLEC5 being inhibitory and SIGLEC14 being activatory. In the Mbuti population of the Democratic Republic of Congo, a deletion allele of the SIGLEC family member *SIGLEC5*, is at a frequency of 54% yet is absent from any other population [11]. The neighbouring gene, *SIGLEC14*, has a deletion allele in human populations, with frequencies ranging from 90% in Chinese individuals to 10% in northern Europeans [65]. Loss of *SIGLEC14* lowers the secretion of the inflammatory cytokine TNFalpha in response to bacterial lipopolysaccharide, reduces secretion of IL-1beta and confers susceptibility to Group B streptococcus infection [66]. Given that loss of *SIGLEC14* is the derived allele, what could be the possible advantage? It is possible that the resulting dampening of a pro-inflammatory response could be advantageous, as *SIGLEC14* loss reduces exacerbation in chronic obstructive pulmonary disease [67].

Structural variation and dietary changes in evolution

Humans developed different diets as they populated the world. Domesticated plants, such as wheat and rice, were selected for high starch content, as starch is a complex carbohydrate rich in energy if it can be digested efficiently. Multiple enzymes digest starch in humans, including amylase and maltase-glucoamylase. Amylase is encoded by *AMY1* (salivary amylase), expressed in the saliva, lacrimal glands and mammary glands, and *AMY2* (pancreatic amylase) secreted by the pancreas. The number of copies of both *AMY1* and *AMY2* varies and is positively correlated, although *AMY1* copy number has a broader range in the population (2-18) compared to *AMY2A* (0-4) [68].

There is a correlation between populations with starch-rich diets and higher copy numbers of *AMY1* which, together with the observation that archaic hominin genomes and chimpanzee genome have 1 copy *AMY1*, suggests that high copy number is a recent adaptation to a starch-rich diet produced by agriculture [69,70]. Although there is a clear gene dosage effect of *AMY1*, where a higher copy number is reflected in higher salivary amylase activity, the variation has small effect on the total activity in saliva [71] and it seems plausible that selection has increased *AMY2* copy number as well, since that is correlated with *AMY1* copy number.

Conclusions

Many SVs occur in clusters in the genome, generated by recurrent NAHR between segmental duplications (SDs), and are the cauldrons of new gene evolution, including those involved in human evolution. The formation of these SD-rich regions has been suggested to be due to elements known as core duplicons. These core duplicons are spread throughout the genome and act as substrates for NAHR, generating further SDs, which sponsor further NAHR and SVs. Ultimately, these SD-rich regions sponsor large SVs with clinical effects [72], such as large deletions and duplications of chromosome 16p11.21 [73], which result in a variety of clinical phenotypes including early-onset obesity and developmental delay [74,75] and neuroblastoma [76]. We speculate that this sponsoring of clinical phenotypes, which were likely to have been under negative selection in the past, places an upper limit on the runaway process of SD accrual in these regions by NAHR. Understanding the distribution of SVs in the genome and their evolution therefore is important for medical genetics as well as evolutionary genetics.

SVs are generated by mutation, and can carry new, or reassorted, gene sequences, which can either be caught and swept higher in frequency by adaptive selection, or be essentially neutral with a fate determined by genetic drift. Distinguishing these processes remains challenging, and requires knowledge of the functional consequences of different variants. Even then, any suggestion of adaptation needs to be plausible, and, the further back in time such adaptation occurred, the less we know about the environment that drove that adaptation. Nevertheless, as we show in this review, SVs are often excellent examples of adaptation throughout human evolution, and we anticipate more exciting examples to come.

Box 1 - Archaic hominin introgression of structural variants in modern human genomes

In the past decade, genome sequencing of Neanderthals revealed that they admixed with ancestors of present-day humans to an extent that they contributed to ~2% of the genome of populations outside Africa [80] Additionally, up to 5% of the genomes of Oceanic populations harbor ancestry inherited from Denisovans, a previously unknown hominin group whose remains were unearthed in the Denisovan Cave, in the Altai Mountains of Siberia [81,82]. Growing evidence suggests that Neanderthal and Denisovan introgression contributed to local adaptation and phenotypic variation in modern humans [83–85].Although studies of archaic hominin introgression have mainly focused on SNVs, recent advances in SVs analysis have suggested that archaic introgression had a role in shaping SVs diversity in human populations.

Sudmant and colleagues [12] used read-depth based digital comparative genomic hybridization (dCGH), SNVs and aCGH microarrays to compare 236 high coverage genomes from 125 geographically diverse populations to the Neanderthal and Denisovan high coverage genomes. They found that, while SVs inherited from Neanderthals appear to be shared among human populations outside Africa, Denisovan-inherited SVs alleles were uniquely found in present-day populations in Oceania, where they segregate at high frequencies [12]. In particular, two SDs on chromosome 16 are nearly fixed in Oceanic populations (allele frequency >80%), and are thought to be part of a larger composite duplication spanning ~225 kbp, that originated in the Denisovan lineage ~400,000 years ago and entered the human gene pool ~40,000 years ago through introgression. This Denisovan-Oceanic specific duplication has been recently found to segregate at high frequency also in a population living in Flores Island (Indonesia), likely as a consequence of gene flow with Oceanic populations [47]. To date, this complex duplication on chromosome 16 is the largest introgressed locus identified in modern human genomes. A more systematic characterization of introgressed SVs, using fluorescence in situ hybridization (FISH) experiments, bacterial artificial chromosome (BAC) cloning, and long-read sequencing to validate introgressed CNVs, contributed to sequence-resolve the structure and signatures of selection of the large Denisovan introgressed duplication at chromosome 16 [11], however the functional basis of any possible adaptation remains unclear.

Additionally, a large Neanderthal-derived haplotype encompassing a deletion has been found on chromosome 8p21.3 to segregate at 44% frequency in Oceanic populations [13]. The deletion has been shown to affect expression of the flanking *TNFRSF10D* genes, which encode cell-surface receptors that bind to TRAIL (TNF-related apoptosis-inducing ligand). Such receptors, known as “death-receptors”, are involved in apoptosis of viral-infected cell [86], suggesting that modulation of the immune response to viruses is a plausible explanation of this signal of adaptive introgression.

Box 2 - How are structural variants different from single nucleotide variants?

Structural variants are just another type of variant, so why are they so often treated separately? This is because of both technical and genetic reasons, which we will discuss in this box.

Large scale genotyping of single nucleotide variation was made possible by development of hybridisation-based SNV chips. Although a combination of hybridisation-intensity data and allelic ratio data could be used to detect and genotype copy number variation, this was generally reliable only for large changes, as such approaches had poor reliability [1] and often yielded data inconsistent across studies. The lack of a reliable genomewide approach for genotyping SVs meant that locus-specific assays were often used, which themselves had reliability issues, particularly for complex multiallelic SVs [87]. Genome sequencing studies have provided more reliable approaches of detecting SVs, but there remain problems in terms of reliability of calls and the degree of results filtering that is needed, meaning that SV calling requires more care and attention than SNV calling using standard approaches.

There are also challenges regarding determining genotype from copy number estimates, particularly for multiallelic regions. For example, a diploid individual with 4 copies across both genomes could have a genotype of 2-2 or 1-3. Using information of inheritance of pedigrees, together with genotype distribution across a population, can give an estimate of the relative probabilities of an individual having each genotype [88], but accuracy depends on the number and distribution of alleles in the population.

Both accuracy of SV calling and determination of multiallelic genotypes would be strengthened by accurate phasing/ imputation using patterns of linkage disequilibrium from flanking SNV genotypes. This approach has a lot of potential, and can often allow indirect imputation of SVs from just SNV data. Furthermore, because heterozygous inversions suppress recombination across the inverted region, imputation of inversions can use characteristic patterns of high LD within the inverted region itself [89]. However, these approaches depend on the SVs arising once, or at least only a few times, so that the same SV in different individuals is likely to show identity by descent (IBD). For some SVs, such as a large inversion on 8p23.1, or complex multiallelic variants, a high mutation rate means that the same allele will be identical by state (IBS) but not necessarily IBD . This limits LD with flanking SNVs, and makes accurate imputation much more challenging [16].

Box 3 - Measuring evolution at structural variants

Once SVs are accurately genotyped, we would like to test for evidence of selection. Some early analyses used standard population genetic statistics modified for use with copy number dosage data, such as Vst (an analogue of Fst) [1] and the K statistic (similar to the McDonald-Kreitman test) [90]. With reliable genotypes, a SV can be treated as any other variant for population genetic analysis, and population differentiation statistics are often used to suggest selection at a particular variant. However, by just treating a SV as any other variant means that important information from the sequence variation within, or surrounding, a SV is lost or misinterpreted. In particular, the high mutation rate of some SVs means that signatures of selection may be complex [91]. This can be potentially addressed using coalescent approaches, and approaches using coalescent analysis on fixed duplications have been developed [92], but with simplifying assumptions such as no gene conversion or no copy number variation. With more realistic models, coalescent approaches rapidly become complex [93]. Current forward-simulation approaches such as SliM explicitly forbid the length of a chromosome to vary, preventing SVs to be analysed as anything beyond a single variant with multiple alleles. Conversely, forward-simulation has been applied to sequence variation between duplicates, but they do not consider SVs [94]. There is clearly potential for forward simulation-based approaches to interrogate the patterns of evolution at SVs.

Glossary

Neoteny - evolution by retaining juvenile features in the adult, aften by slowing or delaying particular developmental processes.

Adaptive introgression - Acquisition of variants from archaic humans that have enabled adaptation in new environments.

Identity by descent - Two identical alleles that have arisen from a single mutational event. Segments of identity by descent are genomic regions over which a pair of individuals share a haplotype due to inheritance from a recent common ancestor

Identity by state - two identical alleles that have arisen in different mutational events

Expression quantitative trait locus (eQTL) - A variant that is correlated with levels of mRNA of a particular gene in a particular tissue or cell type.

Multiallelic copy number variation (mCNV) - a copy number variant with more that one allele in the population, usually each allele consists of a variable number of tandem repeats (VNTR)

Homodimer - a protein made of two identical subunits

Heterodimer - a protein hade of two different subunits

Phasing - determining the haplotype of multiple alleles from diploid genotypes

Imputation - Using the information from the known haplotypes present in a population to infer a genotype at a locus.

Linkage disequilibrium (LD) - the non-random association of alleles at two or more loci.

Non-allelic homologous recombination (NAHR) - a mutational process where unequal crossing over during meiosis between similar DNA sequences generates deletions or duplications. Also known as ectopic recombination.

Segmental duplication (SD) - a section of DNA that maps to at least two different genomic locations. Originally coined to distinguish shorter interspersed duplications from whole genome duplications.

Single nucleotide variants (SNVs) - a specific single nucleotide in the genome that differs between members of the same species - for example in some human genomes it could be an A, in others a C.

Single nucleotide polymorphisms (SNPs) - essentially synonymous with SNV, but sometimes used to imply a common SNV in a population: a SNP is an SNV which occurs at greater than 1%

frequency in a specific population.

Genetic drift - change in allele frequency from one generation to the next because of random variation in offspring number between different individuals in a finite population.

Acknowledgements

We would like to thank Dr Liza Felicoli for assistance in drawing figure 2.

References

1 Redon, R. *et al.* (2006) Global variation in copy number in the human genome. *Nature* 444, 444–54
2 Pinto, D. *et al.* (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol* 29, 512–20
3 Porubsky, D. *et al.* (2021) Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat Biotechnol* 39, 302–308
4 Logsdon, G.A. *et al.* (2021) The structure, function and evolution of a complete human chromosome 8. *Nature* DOI: 10.1038/s41586-021-03420-7
5 Ebert, P. *et al.* (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 372,
6 Mahmoud, M. *et al.* (2019) Structural variant calling: the long and the short of it. *Genome Biol* 20, 246
7 Sudmant, P.H. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature* 526, 75–81
8 1000 Genomes Project Consortium *et al.* (2015) A global reference for human genetic variation. *Nature* 526, 68–74
9 Coop, G. *et al.* (2009) The role of geography in human adaptation. *PLoS Genet* 5, e1000500
10 Bergstrom, A. *et al.* (2020) Insights into human genetic variation and population history from 929 diverse genomes. *Science* 367,
11 Almarri, M.A. *et al.* (2020) Population Structure, Stratification, and Introgression of Human Structural Variation. *Cell* 182, 189-199 e15
12 Sudmant, P.H. *et al.* (2015) Global diversity, population stratification, and selection of human copy-number variation. *Science* 349, aab3761
13 Hsieh, P. *et al.* (2019) Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science* 366,
14 Hollox, E.J. *et al.* (2003) Extensive normal copy number variation of a beta-

defensin antimicrobial-gene cluster. *Am J Hum Genet* 73, 591–600

15 James, C.P. *et al.* (2018) Human beta defensin (HBD) gene copy number affects HBD2 protein levels: impact on cervical bactericidal immunity in pregnancy. *Eur J Hum Genet* 26, 434–439

16 Handsaker, R.E. *et al.* (2015) Large multiallelic copy number variations in humans. *Nat. Genet.* 47, 296–303

17 Montavon, T. *et al.* (2012) Impact of copy number variations (CNVs) on long-range gene regulation at the HoxD locus. *Proc Natl Acad Sci U S A* 109, 20204–11

18 Spielmann, M. *et al.* (2018) Structural variation in the 3D genome. *Nat Rev Genet* 19, 453–467

19 Aigner, J. *et al.* (2013) A common 56-kilobase deletion in a primate-specific segmental duplication creates a novel butyrophilin-like protein. *BMC Genet* 14, 61

20 Vik, D.P. and Wong, W.W. (1993) Structure of the gene for the F allele of complement receptor type 1 and sequence of the coding region unique to the S allele. *The Journal of Immunology* 151, 6214–6224

21 Jakubosky, D. *et al.* (2020) Properties of structural variants and short tandem repeats associated with gene expression and complex traits. *Nat Commun* 11, 2927

22 Chiang, C. *et al.* (2017) The impact of structural variation on human gene expression. *Nat Genet* 49, 692–699

23 Young, J.M. *et al.* (2008) Extensive copy-number variation of the human olfactory receptor gene family. *Am J Hum Genet* 83, 228–42

24 Nguyen, D.Q. *et al.* (2008) Reduced purifying selection prevails over positive selection in human copy number variant evolution. *Genome Res* 18, 1711–23

25 Fackenthal, J.D. and Olopade, O.I. (2007) Breast cancer risk associated with BRCA1 and BRCA2 in diverse populations. *Nat Rev Cancer* 7, 937–48

26 Itsara, A. *et al.* (2009) Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* 84, 148–61

27 Kupper, C. *et al.* (2016) A supergene determines highly divergent male reproductive morphs in the ruff. *Nat Genet* 48, 79–83

28 Pettersson, M.E. *et al.* (2019) A chromosome-level assembly of the Atlantic herring genome-detection of a supergene and other signals of selection. *Genome Res* 29, 1919–1928

29 Freedman, A.H. *et al.* (2014) Genome sequencing highlights the dynamic early history of dogs. *PLoS Genet* 10, e1004016

30 Axelsson, E. *et al.* (2013) The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* 495, 360–4

31 Pajic, P. *et al.* (2019) Independent amylase gene copy number bursts correlate with dietary preferences in mammals. *Elife* 8,

32 Chimpanzee, S. and Analysis, C. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69–87

33 McLean, C.Y. *et al.* (2011) Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* 471, 216–9

34 Reno, P.L. *et al.* (2013) A penile spine/vibrissa enhancer sequence is missing in modern and extinct humans but is retained in multiple primates with penile spines and

sensory vibrissae. *PLoS One* 8, e84258

35 Dennis, M.Y. *et al.* (2012) Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* 149, 912–22

36 Charrier, C. *et al.* (2012) Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* 149, 923–35

37 Yuste, R. (2011) Dendritic spines and distributed circuits. *Neuron* 71, 772–81

38 Florio, M. *et al.* (2015) Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* 347, 1465–70

39 Kalebic, N. *et al.* (2018) Human-specific ARHGAP11B induces hallmarks of neocortical expansion in developing ferret neocortex. *Elife* 7,

40 Heide, M. *et al.* (2020) Human-specific ARHGAP11B increases size and folding of primate neocortex in the fetal marmoset. *Science* 369, 546–550

41 Ables, J.L. *et al.* (2011) Not(ch) just development: Notch signalling in the adult brain. *Nat Rev Neurosci* 12, 269–83

42 Fiddes, I.T. *et al.* (2018) Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. *Cell* 173, 1356-1369 e22

43 Suzuki, I.K. *et al.* (2018) Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch Regulation. *Cell* 173, 1370-1384 e16

44 Segurel, L. and Bon, C. (2017) On the Evolution of Lactase Persistence in Humans. *Annu Rev Genomics Hum Genet* 18, 297–319

45 Fumagalli, M. *et al.* (2015) Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* 349, 1343–7

46 Ye, K. *et al.* (2017) Dietary adaptation of FADS genes in Europe varied across time and geography. *Nat Ecol Evol* 1, 167

47 Tucci, S. *et al.* (2018) Evolutionary history and adaptation of a human pygmy population of Flores Island, Indonesia. *Science* 361, 511–516

48 Miller, L.H. *et al.* (1976) The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy. *N Engl J Med* 295, 302–4

49 Andersen, K.G. *et al.* (2012) Genome-wide scans provide evidence for positive selection of genes implicated in Lassa fever. *Philos Trans R Soc Lond B Biol Sci* 367, 868–77

50 Sharp, P.M. *et al.* (2020) Ape Origins of Human Malaria. *Annu Rev Microbiol* 74, 39–63

51 Kan, Y.W. *et al.* (1975) Deletion of alpha-globin genes in haemoglobin-H disease demonstrates multiple alpha-globin structural loci. *Nature* 255, 255–256

52 Lam, K.W. and Jeffreys, A.J. (2007) Processes of de novo duplication of human alpha-globin genes. *Proc Natl Acad Sci U S A* 104, 10950–5

53 Williams, T.N. *et al.* (2005) Both heterozygous and homozygous alpha+ thalassemias protect against severe and fatal Plasmodium falciparum malaria on the coast of Kenya. *Blood* 106, 368–71

54 Leffler, E.M. *et al.* (2017) Resistance to malaria through structural variation of red blood cell invasion receptors. *Science* 356,

55 Louzada, S. *et al.* (2020) Structural variation of the malaria-associated human glycophorin A-B-E region. *BMC Genomics* 21, 446

56 Algady, W. *et al.* (2018) The Malaria-Protective Human Glycophorin Structural

Variant DUP4 Shows Somatic Mosaicism and Association with Hemoglobin Levels. *Am. J. Hum. Genet.* 103, 769–776

57 Kariuki, S.N. *et al.* (2020) Red blood cell tension protects against severe malaria in the Dantu blood group. *Nature* 585, 579–583

58 Reichhardt, M.P. *et al.* (2017) SALSA-A dance on a slippery floor with changing partners. *Mol Immunol* 89, 100–110

59 Bikker, F.J. *et al.* (2017) The scavenging capacity of DMBT1 is impaired by germline deletions. *Immunogenetics* 69, 401–407

60 Polley, S. *et al.* (2015) Evolution of the rapidly mutating human salivary agglutinin gene (DMBT1) and population subsistence strategy. *Proc Natl Acad Sci U S A* 112, 5105–10

61 DeGiorgio, M. *et al.* (2014) A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genet* 10, e1004561

62 Siewert, K.M. and Voight, B.F. (2017) Detecting Long-Term Balancing Selection Using Allele Frequency Correlation. *Mol Biol Evol* 34, 2996–3005

63 Bitarello, B.D. *et al.* (2018) Signatures of Long-Term Balancing Selection in Human Genomes. *Genome Biology and Evolution* 10, 939–955

64 Cao, H. and Crocker, P.R. (2011) Evolution of CD33-related siglecs: regulating host immune functions and escaping pathogen exploitation? *Immunology* 132, 18–26

65 Yamanaka, M. *et al.* (2009) Deletion polymorphism of SIGLEC14 and its functional implications. *Glycobiology* 19, 841–6

66 Ali, S.R. *et al.* (2014) Siglec-5 and Siglec-14 are polymorphic paired receptors that modulate neutrophil and amnion signaling responses to group B Streptococcus. *J Exp Med* 211, 1231–42

67 Angata, T. *et al.* (2013) Loss of Siglec-14 reduces the risk of chronic obstructive pulmonary disease exacerbation. *Cell Mol Life Sci* 70, 3199–210

68 Carpenter, D. *et al.* (2015) Obesity, starch digestion and amylase: association between copy number variants at human salivary (AMY1) and pancreatic (AMY2) amylase genes. *Hum Mol Genet* 24, 3472–80

69 Perry, G.H. *et al.* (2007) Diet and the evolution of human amylase gene copy number variation. *Nat Genet* 39, 1256–60

70 Inchley, C.E. *et al.* (2016) Selective sweep on human amylase genes postdates the split with Neanderthals. *Sci Rep* 6, 37198

71 Carpenter, D. *et al.* (2017) Copy number variation of human AMY1 is a minor contributor to variation in salivary amylase expression and activity. *Hum Genomics* 11, 2

72 Lin, Y.-L. and Gokcumen, O. (2019) Fine-Scale Characterization of Genomic Structural Variation in the Human Genome Reveals Adaptive and Biomedically Relevant Hotspots. *Genome Biology and Evolution* 11, 1136–1151

73 Nuttle, X. *et al.* (2016) Emergence of a Homo sapiens-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* 536, 205–9

74 Jacquemont, S. *et al.* (2011) Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature* 478, 97–102

75 Bochukova, E.G. *et al.* (2010) Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* 463, 666–70

76 Egolf, L.E. *et al.* (2019) Germline 16p11.2 Microdeletion Predisposes to Neuroblastoma. *Am J Hum Genet* 105, 658–668

77 Collins, R.L. *et al.* (2020) A structural variation reference for medical and population genetics. *Nature* 581, 444–451

78 Gokhman, D. *et al.* (2021) Human-chimpanzee fused cells reveal cis-regulatory divergence underlying skeletal evolution. *Nat Genet* 53, 467–476

79 Trujillo, C.A. *et al.* (2021) Reintroduction of the archaic variant of NOVA1 in cortical organoids alters neurodevelopment. *Science* 371,

80 Prufer, K. *et al.* (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505, 43–9

81 Mallick, S. *et al.* (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538, 201–206

82 Meyer, M. *et al.* (2012) A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* 338, 222–226

83 Simonti, C.N. *et al.* (2016) The phenotypic legacy of admixture between modern humans and Neandertals. *Science* 351, 737–741

84 Vernot, B. *et al.* (2016) Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* 352, 235–239

85 Huerta-Sánchez, E. *et al.* (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512, 194–197

86 Solà-Riera, C. *et al.* (2019) Hantavirus Inhibits TRAIL-Mediated Killing of Infected Cells by Downregulating Death Receptor 5. *Cell Rep* 28, 2124-2139.e6

87 Cantsilieris, S. *et al.* (2014) Technical considerations for genotyping multi-allelic copy number variation (CNV), in regions of segmental duplication. *BMC Genomics* 15, 329

88 Zuccherato, L.W. *et al.* (2017) Population genetics of immune-related multilocus copy number variation in Native Americans. *J R Soc Interface* 14,

89 Ruiz-Arenas, C. *et al.* (2019) scoreInvHap: Inversion genotyping for genome-wide association studies. *PLoS Genet* 15, e1008203

90 Gokcumen, O. *et al.* (2011) Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection. *Genome Biol* 12, R52

91 Pennings, P.S. and Hermisson, J. (2006) Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genet* 2, e186

92 Thornton, K.R. (2007) The Neutral Coalescent Process for Recent Gene Duplications and Copy-Number Variants. *Genetics* 177, 987–1000

93 Teshima, K.M. and Innan, H. (2012) The coalescent with selection on copy number variants. *Genetics* 190, 1077–86

94 Hartasanchez, D.A. *et al.* (2016) SeDuS: segmental duplication simulator. *Bioinformatics* 32, 148–50

95 Xue, Y. *et al.* (2008) Adaptive evolution of UGT2B17 copy-number variation. *Am J Hum Genet* 83, 337–346

96 Saitou, M. and Gokcumen, O. (2019) Resolving the Insertion Sites of Polymorphic Duplications Reveals a HERC2 Haplotype under Selection. *Genome Biology and Evolution* 11, 1679–1690

97 Hardwick, R.J. *et al.* (2014) Haptoglobin (HP) and Haptoglobin-related protein (HPR) copy number variation, natural selection, and trypanosomiasis. *Hum Genet* 133, 69–83

98 Stefansson, H. *et al.* (2005) A common inversion under selection in Europeans. *Nat Genet* 37, 129–137

99 Boettger, L.M. *et al.* (2012) Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat Genet* 44, 881–885

100 Giannuzzi, G. *et al.* (2019) The Human-Specific BOLA2 Duplication Modifies Iron Homeostasis and Anemia Predisposition in Chromosome 16p11.2 Autism Individuals. *Am J Hum Genet* 105, 947–958

Figure Legends

Figure 1. Identifying structural variation in genomes from short-read sequencing

a) Different types of structural variant (SV). For each example, a representation of a heterozygous SV is shown, with the reference sequence is at the top , and arrows representing sequence orientation. The variation in terms of number of copies per diploid genome is shown on the right.

b) Different approaches in using short-read sequencing to detect SVs by aligning to a reference genome. The basis of using split-read, discordant pairs and read depth approaches to detect deletions, insertions and inversions is shown.

Figure 2. The evolution of *SRGAP2* genes in humans

a) Genomic organisation of *SRGAP2* genes in humans and chimpanzees, with the order and approximate timing of the duplication events in the human lineage. The different *SRGAP2* dimers generated in chimpanzees and humans are also shown.

b) An example of neoteny - comparison of cranial shape development in chimpanzees in humans and chimpanzees, highlighting the increased braincase size in humans and the maintenance of a high facial size/braincase size in humans during development.

Figure 3. The structure of the malaria-protective DUP4 variant encoded by a glycophorin B-A fusion gene

a) Fiber-fluorescent in-situ hybridisation (FISH) image showing the genomic organisation of the glycophorin region in humans, in the reference variant. Fluorescently-labelled fosmids, as indicated below the image, are used to detect extended DNA strands. The arrangement of fosmids is used to infer a structure for the region [56].

b) Fiber-FISH image for the glycophorin DUP4 variant, and inferred genomic structure.

c) GYPB/A fusion gene intron-exon structure, with GYPB derived exons shown in yellow and GYPA-derived exons shown in purple. Exon 3 is a pseudoexon not included in the transcript.

d) Predicted orientation of the GYPB/A fusion protein in the cell membrane, showing the external domain is glycophorin B-like, with the internal domain being glycophorin A-like.

Table 1 - Examples of structural variants with evidence of recent selection

Gene/locus	Region	Variant	Selection type	Environmental variable	Reference
Amylase	1p21.1	mCNV	Positive	High-starch diet	See text
Alpha-globin	16p13.3	Deletion	Balancing	Malaria infection	See text
UGT2B17	4q13.2	Deletion	Balancing (Europe) Positive (Southern Africa, East Asia)	Steroid hormone metabolism	[95]
Glycophorin locus	4q31.2	Complex duplication	Positive	Malaria infection	See text
DMBT1	8p23.1	mCNV	Balancing	Innate immunity/high-starch diet	See text
HERC2	15q13.1	Duplication	Negative	Pigmentation	[12,96]
HPR	16q22.2	Duplication	Positive	Trypanosomiasis infection	[16,97]
KANSL1	17q21.31	Inversion, Duplication	Positive	Fecundity	[12,98,99]

BOLA2	16p11.2	mCNV	Positive	Iron homeostasis/ diet	[73,100]

Outstanding questions (2000 characters).

Computational simulation of genomic variation is very important in testing evolutionary hypotheses explaining patterns of variation. However, incorporating models of mutation that accurately represent SVs, and the sequence variation within SVs, lags behind models of single nucleotide variation. How can we use population genetic simulations to more rigorously detect the effect of selection at SVs?

Will long read sequencing, and full assemblies of diploid genomes of humans and other apes from telomere to telomere uncover more examples of SVs likely to be adaptive at complex pericentromeric regions?

Analysis of exomes from the Genome Aggregation Database (gnomAD) collection has revealed many new SVs at low or intermediate frequencies [77]. What will analysis of SVs in even larger cohorts, such as UK Biobank, and in underrepresented populations reveal about the thier effect on human phenotypic variation and adaptation?

Functional characterisation of the effect of variants takes time and effort, but is important in understanding whether an SV can affect phenotype, and the mechanism behind that effect. In particular, rodent models provide a wide variety of tools to dissect gene function, but assessment of human specific variants in a human context can be difficult. Approaches using human-chimpanzee pluripotent cell models [78] or human cells modified to have ancestral variants [79] have excellent potential to explore genetic differences between the two species. Although challenges remain, these approaches may have potential in SV studies. How can we further improve approaches to test the effect of SVs on phenotype?

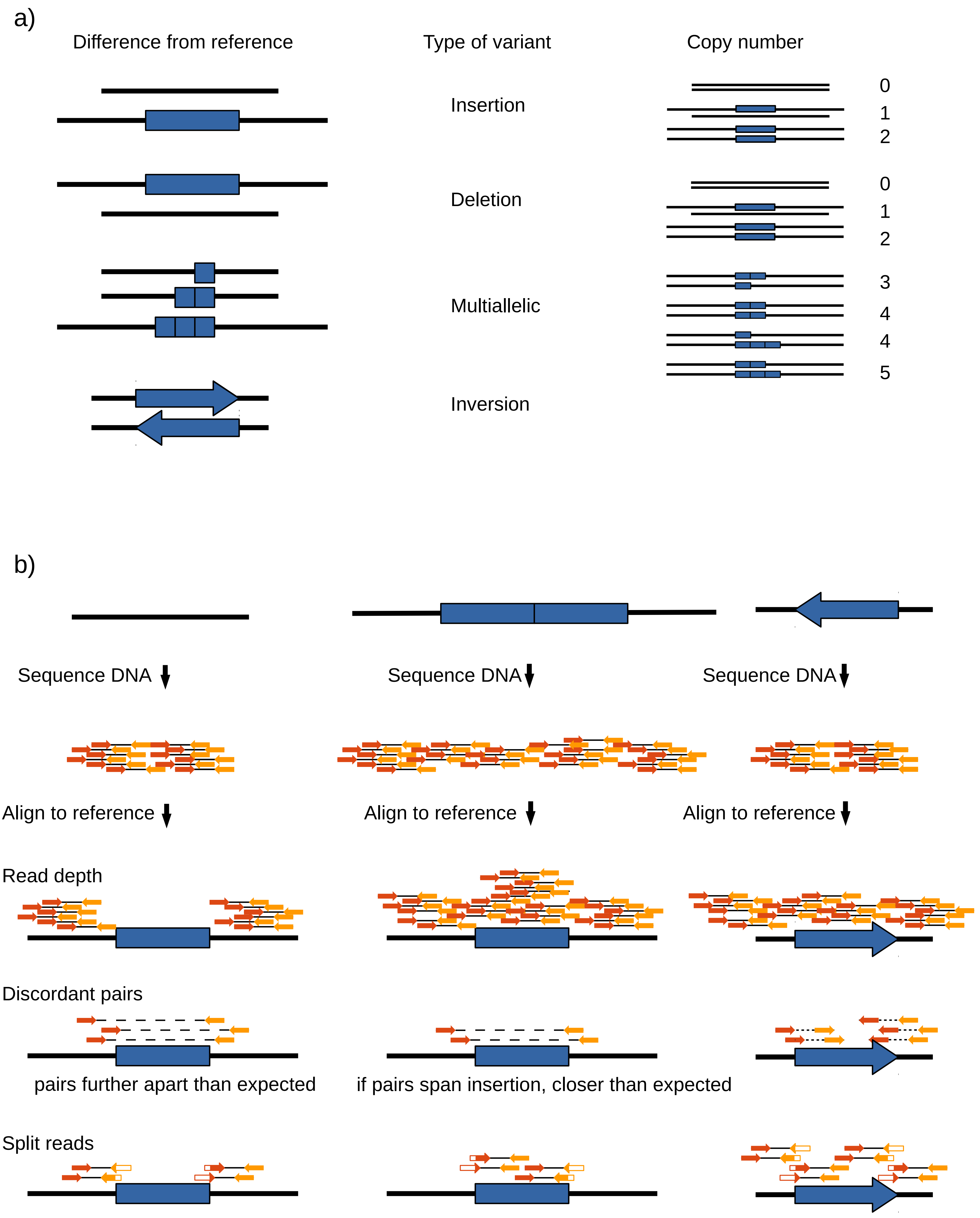

a)
Difference from reference
Type of variant
Copy number
Insertion
0
1
2
Deletion
0
1
2
Multiallelic
3
4
4
5
Inversion
b)
Sequence DNA
Sequence DNA
Sequence DNA
Align to reference
Align to reference
Align to reference
Read depth
Discordant pairs
pairs further apart than expected
if pairs span insertion, closer than expected
Split reads

Figure 2

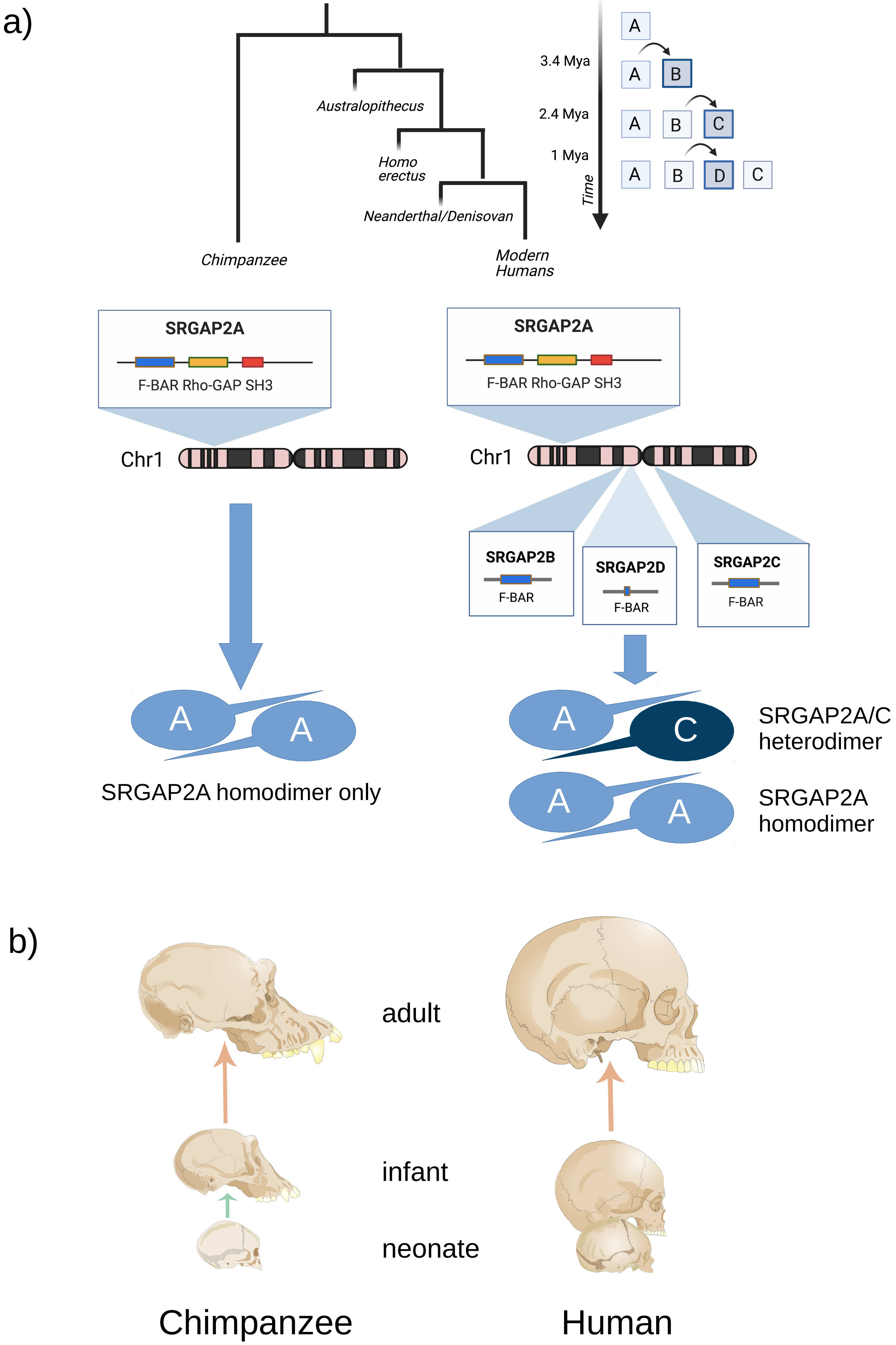
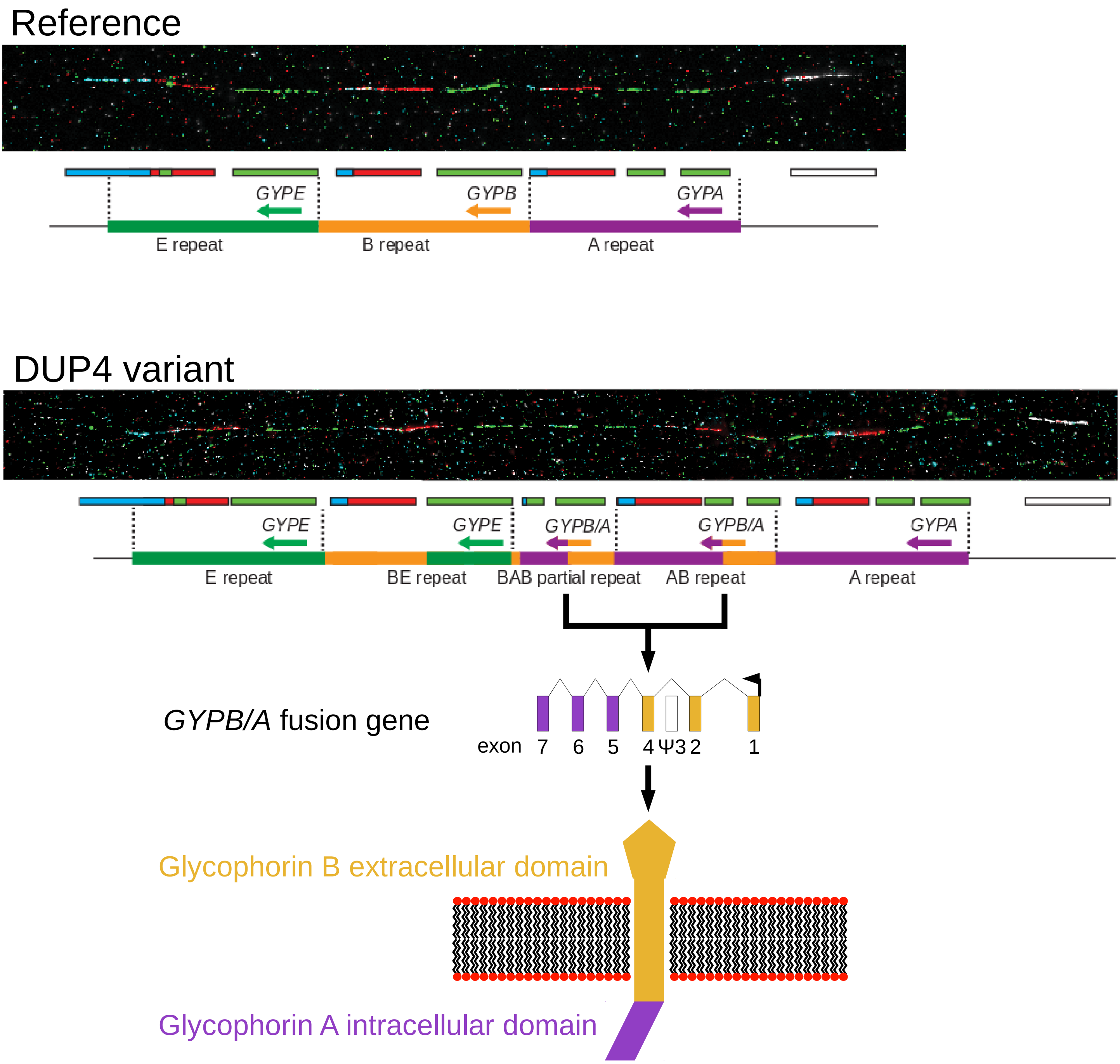

a)
Chimpanzee
Australopithecus
Homo erectus
Neanderthal/Denisovan
Modern Humans
3.4 Mya
2.4 Mya
1 Mya
Time
A
A B
A B C
A B D C
SRGAP2A
F-BAR Rho-GAP SH3
Chr1
SRGAP2A
F-BAR Rho-GAP SH3
Chr1
SRGAP2B
F-BAR
SRGAP2D
F-BAR
SRGAP2C
F-BAR
A A
SRGAP2A homodimer only
A C
SRGAP2A/C heterodimer
A A
SRGAP2A homodimer
b)
adult
infant
neonate
Chimpanzee
Human

Figure 3


Reference
GYPE
GYPB
GYPA
E repeat
B repeat
A repeat
DUP4 variant
GYPE
GYPE
GYPB/A
GYPB/A
GYPA
E repeat
BE repeat
BAB partial repeat
AB repeat
A repeat
GYPB/A fusion gene
exon 7 6 5 4 Ψ3 2 1
Glycophorin B extracellular domain
Glycophorin A intracellular domain