Published in final edited form as:

Trends Genet. 2012 June ; 28(6): 245–257. doi:10.1016/j.tig.2012.03.002.

Exploring the role of copy number variants in human adaptation

Rebecca C. Iskow^{1,2}, Omer Gokcumen^{1,2}, and Charles Lee^{1,2}

¹Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115, USA

²Harvard Medical School, Boston, MA 02115, USA

Abstract

Over the past decade, the ubiquity of copy number variants (CNVs, the gain or loss of genomic material) in the genomes of healthy humans has become apparent. Although some of these variants are associated with disorders, a handful of studies documented an adaptive advantage conferred by CNVs. In this review, we propose that CNVs are substrates for human evolution and adaptation. We discuss the possible mechanisms and evolutionary processes in which CNVs are selected, outline the current challenges in identifying these loci, and highlight that copy number variable regions allow for the creation of novel genes that may diversify the repertoire of such genes in response to rapidly changing environments. We expect that many more adaptive CNVs will be discovered in the coming years, and we believe that these new findings will contribute to our understanding of human-specific phenotypes.

Keywords

copy number variation; adaptation; human evolution

CNVs: a new consideration in adaptive evolution

Since diverging from the lineage leading to chimpanzees 6 million years ago, humans have evolved larger brains, gained an upright posture, lost body hair and developed complex spoken language, among other characteristics [1,2]. Even now, human evolution continues [3–9]. For example, humans are currently evolving to digest sugars commonly found in modern diets more efficiently [5,7], have higher birth rates [3,8] and are developing increased protection against infectious diseases, such as malaria [6,9]. Such evolutionary changes are adaptive because they affect an individual's likelihood of survival and reproduction. By studying the genetic variation influencing such characteristics, it is possible to understand adaptation in modern humans. Single nucleotide polymorphisms (SNPs; see Glossary), restriction fragment length polymorphisms (RFLPs) and protein polymorphisms have long been archetypes of human genetic variation, and many current tools for understanding evolution are based upon these variants (Box 1). However, little is known about the role of other forms of genetic variation in human adaptation. CNVs are a form of structural variation defined by the gain or loss of genomic material [10,11] and are common among healthy individuals [12–14]. In fact, any two persons will differ in copy number across 0.78% of their genomes on average [14]. When combining three of the most comprehensive CNV maps to date [12–14], approximately 7.6% of the human reference genome contains CNVs. CNVs can impact human phenotypic variation and contribute to diseases such as autism, psoriasis, schizophrenia, obesity and Crohn's disease [15–22].

However, the evolutionary impact of CNVs has yet to be comprehensively assessed. In this review, we explore the impact of fixed copy number differences (CNDs) between humans and other primate species on human evolution, and discuss the extent to which CNVs contribute to adaptive evolution in humans.

We argue that genomic copy number variation is a mechanism through which phenotypes are altered and upon which selective forces can act. With the recent advances in high-resolution array and second-generation sequencing technologies, CNVs can be detected with unprecedented sensitivity and vastly improved breakpoint resolution (Box 2). It is now possible to explore the sequence context and population dynamics of CNVs, permitting the study of their evolution. Herein, we describe the ways in which CNVs can be acted upon by selection and their unique features that should be considered when studying their evolutionary history.

Gene function can be altered by CNVs and affect adaptive potential

CNVs can have a phenotypic impact and, consequently, can alter the fitness of an allele through mechanisms including: (i) changing the coding sequence of a gene [23,24]; (ii) creating paralogs that can diverge from each other and take on new or specialized function (neofunctionalization) [25,26]; and (iii) altering the expression level of a gene [24,27]. Such scenarios are not necessarily mutually exclusive. For example, in theory, a CNV within a gene encoding a transcription factor can affect both the coding sequence of the gene itself as well as the expression of downstream targets of that gene. When a variant alters the function of a gene, whether through changing its expression levels or the encoded product, the variant then has the potential to alter the fitness of the organism and be acted upon by selection. Thus, it seems plausible that CNVs overlapping genes are more likely to be adaptive and affect fitness than are CNVs in intergenic regions [28]. Most functional regions in the human genome appear to be under purifying (negative) selection [29,30], but there are several examples of positive selection on SNPs [31]. The same is probably true for CNVs; that is, most CNVs overlapping genes are under purifying selection [14], but a handful of CNVs are thought to be under positive selection [28,32]. However, it is currently unclear to what extent positive selection acts upon CNVs and whether these few examples are exceptions, or indicative of a more general phenomenon. Recently, major resequencing projects have been accumulating massive amounts of genomic data at the population level, including data on human and non-human primates [12,33]. When these data are analyzed for the adaptive effect of copy number variation, the number of known CNVs that are under positive selection will probably increase and help researchers to better delineate the role of these variants in human evolution.

Non-darwinian evolutionary forces on CNVs

As described above, CNVs can be subject to either purifying or positive selection. However, such scenarios usually lead to fixation (either by the removal of a harmful CNV or the rapid rise in frequency of a beneficial CNV). Thus, such selective forces cannot account for the breadth of copy number variation currently present among human genomes. One explanation for the persistence and extent of CNVs among human genomes (especially for intergenic CNVs) is that most of these CNVs have evolved under neutral evolutionary pressures. Therefore, the frequency and sequence context of these neutral variants have been shaped entirely by demographic events, mutation rate and genetic drift. CNVs can be formed by a variety of mechanisms, including nonallelic homologous recombination (NAHR), nonhomologous end-joining and retrotransposition (for a detailed review on CNV formation mechanisms, see [34]). Once a CNV is formed by such mechanisms, it may be maintained in a population even in the absence of selection. In fact, it appears that many CNVs are

probably neutral because the genomes of healthy individuals are peppered with hundreds, if not thousands of such variants [28].

Purifying selection acts on detrimental variants that reduce fitness and, in general, eliminates those variants from the population [35]. The effect of purifying selection acting on CNVs is indirectly visible in the genomic distribution of these variants. There is an obvious depletion of CNVs that overlap with functional regions, such as genes and ultra-conserved elements [14,28,36]. Furthermore, there is a significant depletion of larger CNVs (>500 kb), which are more likely to overlap with functional regions and potentially disrupt epigenetic patterns, indicating strong purifying selection [37]. It is possible that purifying selection may act more potently on CNVs than on SNPs owing to the size of the variants. As such, variants that are under purifying selection (unless the selection is mild) are exceedingly rare, limiting the opportunity to study them directly.

Positive selection on CNVs

Positive selection acts on either a new variant or a previously neutral variant that has become advantageous owing to changes in the environment. Under positive selection, a variant can be carried to fixation in a population (i.e. selective sweep; Figure 1); however, only a fraction of the thousands of currently known CNVs exhibit empirical evidence of positive selection in humans (Table 1) [7,38–41]. As more genomes are sequenced and new analytical tools are developed to detect positive selection, we expect to observe more evidence supporting the role of adaptive CNVs.

Current challenges in detecting positive selection upon CNVs

Accurate integer copy number counts and nucleotide breakpoint resolution for CNVs are important to determine because they will help resolve the evolutionary history of individual loci. The primary challenge in studying the evolution of CNVs is the difficulty in obtaining accurate genotypes, especially for those CNVs with a large range of copy number. For instance, unlike SNPs, which are typically biallelic, the CNV gene encoding salivary amylase, *AMY1*, is multi-allelic and ranges in diploid copy number among humans from 2 to 15. Any given person can have many possible combinations of alleles, making CNVs overlapping this gene difficult to genotype. In fact, of 11,700 CNVs discovered recently, only approximately 5000 could be accurately genotyped using arrays for several reasons [14,42]. First, high sequence identity between paralogs or repeated sequences (such as retrotransposons) can cause cross-hybridization on the array, resulting in array signals that are complex to interpret. Second, DNA segments with high copy numbers are difficult to accurately quantify and distinguish from one another (e.g. if the array reference has a copy number of 12 for *AMY1* and the test has a copy number of 14, the expected $\log_2$ ratio of intensities would be 0.22, which is unlikely to meet the cut-off for some CNV calling algorithms). Finally, the precise boundaries of CNVs often cannot be determined because of the resolution of array-based technology and, consequently, overlapping CNVs with different breakpoints are difficult to distinguish from one another. Likewise, sequencing-based methods also have difficulty genotyping CNVs for several reasons including: (i) challenges in determining the precise breakpoints of tandem duplications; (ii) the location of the distant duplicated sequence is hard to identify with short reads; and (iii) duplications can have many possible integer copy number states. For these reasons, among others, it has been extremely difficult to genotype any duplications by sequencing-based methods alone [12]. Sequencing, similar to array-based technology, uses algorithms to estimate breakpoints, but the precise nucleotide resolution of CNVs cannot always be determined. Nevertheless, CNV analysis is improving, and two recent analytical methods have jump started the effort to create accurate, integer copy number genotypes for CNVs using sequencing data [43,44]. In

addition, the development of technologies that produce longer sequence reads will undoubtedly improve the detection and resolution of CNVs.

A second challenge in studying CNV evolution in humans is that traditional signatures of selection (or, more precisely, statistics that determine the likelihood of neutral evolution; Box 1), such as amino acid substitution based tests, allele frequency based tests and linkage based tests, are not applicable to all CNVs. As CNVs are more complex forms of genetic variation than are SNPs, tests of neutrality for CNVs have to be chosen carefully based on the specific context. In addition to having a large range of copy number, duplications of *AMY1* and many other CNVs were probably recurrent during evolutionary history [45,46]. Because similar CNVs can occur on different haplotype backgrounds, they are less likely to be in linkage disequilibrium (LD) with neighboring SNPs than are biallelic CNVs (Figure 1) [47–49]. Even some non-recurrent, biallelic CNVs, such as the common deletion in the cytosine deaminase-encoding *APOBEC3B* gene, are not in LD with surrounding SNPs [40]. In addition, duplications can initiate gene conversion events, which can then decrease the LD surrounding such variants [50]. For *AMY1* (and other complex CNVs that may have been under recent positive selection within human populations), tests of neutrality that examine LD, extended haplotype blocks and stretches of homozygosity will have diluted signals owing to multiple haplotypes containing the CNV that may be under selection (Figure 1B).

Another confounding factor in the identification of CNVs under positive selection is that high identity duplicates, such as the many copies of *AMY1*, are difficult to assemble unambiguously and so are underrepresented in annotated genomes [44]. Without the sequences of such duplications, sequence-dependent signatures of selection (such as K_d/K_s , Box 1) cannot be determined. Hence, the study of positive selection acting upon CNVs often relies on other signatures of positive selection, such as population differentiation (e.g. [49]). *AMY1* copy number exhibits high population differentiation. An increased copy number of *AMY1* permits more efficient breakdown of starch [7]. Populations consuming high quantities of starch, such as Japanese and Europeans, have significantly more copies of this gene than do those consuming lower quantities, such as Yakut and Biaka pygmy, indicating that copy number variation at *AMY1* is under positive selection in response to cultural change [7].

It is important to remember that the application of SNP-based tests of selection to CNVs assumes that the CNVs in question behave in a similar manner to SNPs [41]. Positive selection acting on duplications, as well as on recurrent and multi-allelic CNVs, are more difficult to test for selection and require a novel set of analytical tools [39,40]. Thus, studying CNVs for signs of selection is complicated twofold: first, the CNVs need to be genotyped accurately; and, second, tests of neutrality need to be developed that are appropriate for CNVs that are not in LD with surrounding variants.

Ancient positive selection on the copy number of genomic segments in humans compared with non-human primates

Whereas CNVs exhibit variation within species, CNDs are segments of genomic DNA that differ in copy number between species. Studying CNDs can help uncover signatures of ancient positive selection, which potentially led to the genetic and phenotypic divergence of closely related species. A surge in genomic duplication events is thought to have occurred in the ancestor of the great apes [51,52]. The ubiquity of the resulting segmental duplications present in the genomes of great apes set the stage for recurrent CNV formation by NAHR [46,53]. Thus, segmental duplications among apes have allowed for gene family expansions that can be acted upon by positive selection and, in turn, potentially contribute to the phenotypic divergence of species. Lineage-specific duplications are not necessarily copy

number variable. Many of these events have reached fixation and can be more accurately classified as segmental duplications or CNDs. Recent studies have documented human-specific gene family expansions for which the gene duplicates appear to be under positive selection [51,54–58]. Such gene families can be examined for high proportions of nonsynonymous substitutions (Box 1), which indicate that these genes are diversifying at the sequence level from each other and from their non-human primate orthologs. One of the most extreme examples of an amplified gene family with excessive nonsynonymous substitutions is the nuclear pore complex interacting protein (NPIP) family, also known as morpheus [56]. The genes in this family appear to be under positive selection in multiple primate species, including in humans [56]. Although the precise function of the morpheus genes is unknown, they are expressed in many tissues and their protein products appear to interact with the nuclear pore complex and may help shuttle mRNAs across the nuclear membrane [56]. Likewise, other comparative studies between primates have led to the identification of lineage-specific CNDs potentially associated with innate immunity [59], endurance running [54] and reproduction [60,61].

Interestingly, many duplicated genes in the human lineage appear to be involved in brain function. Another example of a highly amplified gene family in the human lineage, is the DUF1220 domain encoding neuroblastoma breakpoint family (NBPF) [44,51,54,62]. The members of this gene family are expressed primarily in the brain and, similar to morpheus, exhibit high proportions of nonsynonymous substitutions. Several other human-specific gene duplications thought to have evolved under positive selection are also involved in brain function and development [44,63–65]. For instance, the ancestral hydrocephalus inducing homolog (mouse) *HYDIN* gene on chromosome 16 has an additional copy on chromosome 1 in humans that appears to be involved in regulating brain size [66]. Individuals with deletions overlapping the human-specific copy of *HYDIN* on chromosome 1 tend to have microcephaly, whereas duplications tend to give rise to macrocephaly [67]. Furthermore, specific Gene Ontology (GO) categories appear to be enriched for positive selection among the lineage-specific CNDs, including inflammatory response and functions related to cell division [59]. Taken together, gene families related to brain function and cell growth have rapidly expanded and diversified in the human lineage. It could be postulated that these gene families have diverged in the human lineage as an adaptive change that is partially responsible for cognitive differences among primates.

Many of the copy number variable gene families in humans also include immune system-related genes that may have been under positive selection. Specifically, these genes seem to be subject to lineage-specific gene expansions and divergence [53,68]. One theoretical framework to explain this observation is that the duplication of immune system-related genes allows for a diverse 'reservoir' of genes from which to draw when encountering a new pathogen [69]. Such regions allow for the rapid adaptation to new environments and challenges; however, they also tend to be unstable and can predispose to disease. For example, β -defensins, major histocompatibility complex, class I (HLA) and killer cell immunoglobulin-like receptors (KIR) are copy number variable immune-related gene families that show evidence of positive selection in humans [39,68,70]. Such highly variable gene families have evolved in response to changing pathogens and environments. However, these gene families have been associated with certain disorders, such as psoriasis [16,71]. It is possible that the features of a locus that make it more 'adaptable' also, in some cases, enhance its predisposition to genomic instability and disease (reviewed in [72]).

In addition to gene duplications, human-specific deletions can also be acted upon by positive selection and increase in allelic frequency until fixation. Recently, a systematic study found over 500 human-specific deletions by identifying conserved sequences that were present in the chimpanzee and rhesus macaque reference genomes, but were missing in the human

reference genome [73]. In a similar fashion to lineage-specific duplications, human-specific deletions are preferentially located near genes involved with brain function. For example, one such deletion removes an enhancer near the growth-limiting gene, *GADD45G*, reducing the expression of this gene in the developing forebrain and so potentially allowing more growth of this brain region in humans [73]. Another one of the human-specific deletions identified appears to be involved in human reproduction. Humans have lost an enhancer element that regulates the expression of the androgen receptor (*AR*) gene, resulting in the prevention of growth of penile spines. The loss of this morphological characteristic in humans is coincident with increased copulation time in humans relative to chimpanzees, presumably owing to humans' tendency towards monogamy versus the tendency of chimpanzees towards promiscuity [73]. Thus, the deletion of the *AR* enhancer and its subsequent fixation may have been an adaptive response to changing reproductive behavior [73]. It is unknown how many of the other approximately 500 human-specific deletions are functional and under positive selection because fixed deletions are difficult to test formally for selection as they cannot be examined for substitution rates, linkage or allele frequencies.

Ancient positive selection acting on CNDs can also have modern-day medical relevance. For example, a higher copy number of the *CYP2D6* gene, which encodes an enzyme important for metabolism of xenobiotics, causes increased metabolism of many commonly prescribed drugs [74]. Indeed, patients with a high copy number of *CYP2D6* can metabolize several commonly-used drugs so rapidly that the medicine does not remain in the bloodstream long enough to benefit the patient [74]. Likewise, patients with a low copy number are hypersensitive to the same medications [75]. This gene varies widely in copy number among primates, including humans [76]. It is thought that the copy number variation of this gene and other members of its larger gene family (the cytochrome P450 genes) initially allowed for their expansion and neofunctionalization to combat various plant toxins [77]. One hypothesis to explain the large range in copy number in modern human genomes is that, after the initial increase in gene copy number, *CYP2D6* has since been released from tight selective pressure. This gene, and other cytochrome P450 gene family members, may currently be under neutral selection and, hence, randomly increasing or decreasing in copy number [78]. As humans started engaging in agriculture, the need for protection against toxic plants decreased [78]. Thus, the initial gene family expansion was driven by selection to prevent toxicity from the environment, but then the selective pressures upon this gene family subsequently lessened, leading to extensive copy number variation of *CYP2D6* in modern humans owing to genetic drift.

Recent natural selection on CNVs contributing to human adaptation

Unlike signatures of ancient positive selection on CNDs between species, signatures of recent positive selection can be detected on CNVs within species. Because CNVs vary within species, they can be examined using traditional tests of positive selection initially designed for use on SNPs. *UGT2B17* [41], which encodes an enzyme that breaks down steroids and has an important role in regulating the levels of androgens [79], has recently been described as a CNV under positive selection. *UGT2B17* ranges in copy number from 0 to 2, and the deletion allele causes differences in the levels of excreted testosterone in urine and serum [79,80], increased risk of graft-versus-host disease [81] and decreased bone density [82]. Despite these phenotypes, the *UGT2B17* gene deletion was found to be under positive selection based on population differentiation, the allele frequency distribution of linked SNPs and the unusual haplotype structure [41]. The deletion is common among Asian individuals who also have lower testosterone levels in general [41,83]. Thus, the deletion may help conserve testosterone in individuals whose overall level of testosterone is relatively low.

The *UGT2B17* gene deletion was one of the first CNVs to be described under recent positive selection within human populations not necessarily because of its associated phenotypes or strength of selection, but rather because it is a 'well-behaved' variant (i.e. the deletion overlapping this gene behaves like a SNP). Unlike many other CNVs, the *UGT2B17* CNV is common, biallelic, inherited in a Mendelian fashion and is in LD with nearby variants [41]. Positively selected variants generally rise in frequency quickly and there is not enough time for recombination to break up linkage between the selected variant and nearby variants (Figure 1). As such, the entire haplotype, and not just the selected variant, is indicative of a selective event. As described previously for the *AMY1* gene, owing to recurrence and multi-allelism, many CNVs are not in LD with nearby variants, making the study of their evolutionary histories difficult [47–49,84]. However, because the *UGT2B17* CNV is in LD with neighboring SNPs, established tests of neutrality on the nearby SNPs can be used as a proxy to study the haplotype as a whole, including the CNV.

Adaptation is not limited to positive selection; balancing selection can also explain some adaptive CNVs in humans. For example, the copy number of α -globin genes affects malarial morbidity and is a classic example of balancing selection in some human populations [9]. Most humans have four identical diploid copies of α -globin (represented as $\alpha\alpha/\alpha\alpha$ for each of two α -globin genes on two homologous chromosomes), but deletion polymorphisms and, in rare cases, duplication polymorphisms, exist, so that the diploid copy number can range from 0 to 6 [85]. Deletions of two copies of α -globin, whether in *cis* ($---/\alpha\alpha$) or in *trans* ($-\alpha/-\alpha$) cause thalassemia, a mild form of anemia. Loss of three copies of α -globin causes a more serious form of anemia, and having no copies is embryonically lethal. Despite this phenotype, in Southeast Asia, up to 5% of the population is heterozygous for the α -globin *cis* deletions ($---/\alpha\alpha$) [86]. These heterozygotes have reduced malarial morbidity compared with individuals with four copies of α -globin [9,87]. Together, the selective pressure to protect against malaria and the need for at least one intact α -globin gene to produce hemoglobin for survival has maintained the *cis* deletion allele in the population. A similar phenomenon has been observed in many other malaria-endemic populations, indicative of positive selection acting on the *trans* deletion genotype ($-\alpha/-\alpha$), despite its association with anemia [9]. There are only a few other examples of balancing selection in the human genome acting on CNVs. This is probably in part because of the difficulty of detecting such selection.

One of the most publicized and controversial examples of a CNV that is potentially adaptive is the copy number of the *CCL3L1* gene, encoding a ligand for the CCR5 receptor, which is used by HIV to enter its target cells. *CCL3L1* varies in copy number among humans from 0 to 14 [38]. It was suggested that an increased copy number of this gene would increase the amount of ligand, which could then compete with HIV for access to the cell receptors and prevent the entry of HIV into the cell [38]. This was borne out by an association with *CCL3L1* copy number and risk of acquiring HIV, where those with higher copy number appeared less susceptible to HIV infection [38]. Although intriguing, this finding should be interpreted with caution as it may have been subject to 'batch effects' because cases and controls were genotyped separately, and two groups have since not been able to replicate the original findings [88,89]. What remains undisputed is that this region exhibits both extensive variation in copy number and significant population differentiation. This indicates that the *CCL3L1* region has evolved under special circumstances, owing to either demography (such as population expansions or bottlenecks, which can alter the frequency of variants even under neutral conditions) or selection.

CNVs, adaptation and sensory perception

Gene duplications in humans that evolved under positive selection can have related but varied functions allowing for diverse sensory perception of the environment, including smell, taste and sight. Diverse gene families, such as the olfactory receptors (ORs), may create a reservoir of genes that can respond to varying environments. There are approximately 800 OR genes and pseudogenes in the human reference genome, a third of which are copy number variable [43]. After the teleost transition to land, there was a massive increase in the number of OR genes, presumably so that more airborne odorants could be detected [90]. This could be an indication of positive selection toward a population with a higher copy number and diversity of OR genes. However, apes and old world monkeys have three different receptors for color vision (trichromatism), whereas most mammals only have two (dichromatism). Having trichromatic vision is thought to lessen the need to sense the environment through smell, an idea known as the 'vision priority hypothesis' [91]. Therefore, OR genes in some primates may be under less functional constraint than in other organisms and could decay. This is reflected in humans by a reduction of intact ORs compared with mouse, cow and other placental mammals that have dichromatic vision [92]. Thus, the extreme variation of ORs (similar to the *CYP2D6*-containing gene family) may be primarily the result of an initial expansion under positive selection, followed by genomic drift [90], or the stochastic changes in gene copy number resulting from recurrent formation and deletion events. The highly variable nature of the OR genes may also affect the way in which humans perceive their environments. Each person has, on average, 25 copy number variable OR genes and pseudogenes, and a quarter of people have a homozygous deletion of at least one OR gene (some have as many as four) [43]. Because individuals perceive smells with different sensitivities from one another, it is possible that the varied number and repertoire of these genes create individualized perception of odors.

In a similar fashion, the varying copy number of *AMY1* and, consequently, the level of the salivary enzyme, amylase, may alter one's perception of starchy foods. Adding amylase or an amylase inhibitor to starchy food changes the perception of the texture of the food [93,94]. In fact, individuals with a low copy number of *AMY1* perceive starchy foods as more viscous than do those with higher copy numbers [93]. Such differences in perception may alter one's preference for certain foods (e.g. one might consider 'creamy' foods as more desirable, but having a higher copy number of *AMY1* causes creamy foods, such as custards, to digest faster in the mouth, making them seem watery and, subsequently, less desirable) [94].

Trichromatic vision, which arose from a duplication of an opsin gene on the X chromosome, is another example of a CNV that has ramifications on the perception of environment. Deletions and gene conversion events between the opsin gene paralogs can result in color blindness (reviewed in [28,95]). As CNVs have the potential to affect the senses of smell, taste and sight, it is tempting to speculate that certain multi-allelic copy number variable genes and gene families, whether under selection or neutral conditions, contribute to the way in which humans perceive their environments.

Concluding remarks

It is an exciting time for evolutionary geneticists to decipher the potential impact of CNVs on human evolution and adaptation. Gains and losses of genomic segments, which sometimes involve the creation of novel genes, can substantially impact phenotypes. Initial studies of the evolution of human CNVs and CNDs have now identified candidate regions under positive selection (Table 1) and have also demonstrated that most functionally

conserved regions lack CNVs (presumably because of negative selection) [14,36]. Regions that differ in copy number between humans and non-human primates can be informative about ancient adaptations that may have led to species-specific phenotypes. Such CNDs have altered sexual development and, possibly, brain development in the human lineage compared with chimpanzees. The copy number of variable regions within humans can inform about recent adaptations that may lead to genetic and phenotypic differences between individuals and populations. Recent positive selection acting on CNVs within human populations affects traits such as steroid metabolism, malarial morbidity and starch digestion. Notably, CNVs and CNDs exhibiting signatures of adaptation can have consequences regarding drug metabolism, immune response and sensory perception. Thus, it is warranted to study the copy number of medically relevant genes within an evolutionary framework.

In the coming years, the scientific community should anticipate the increasingly accurate discovery and analysis of CNVs, which, in turn, will highlight new regions of the human genome affecting adaptation. Identifying such CNVs will enhance our understanding of the evolutionary history of our species, delineate the genetic factors underlying our phenotypic variation and pinpoint the molecular reasons that predispose us to certain diseases.

Acknowledgments

We would like to thank Qihui Zhu, Sunita Setlur, George Perry, Upeka Samarakoon and Ryan E. Mills for helpful discussions and comments on a previous version of this manuscript. This work was funded by R01 GM0851533-04 and F32 AG 039979.

Glossary

Adaptation	an umbrella term describing any biological change (genetic or phenotypic) that increases the fitness of an individual in a given ecological setting. In this review, we use the term to describe hereditary genetic changes that increase the fitness of an individual by affecting its phenotype.
Copy number difference (CND)	loss or gain of genomic material between individuals of different species.
Copy number variant (CNV)	loss or gain of genomic material between individuals of the same species.
Fitness	the measure of evolutionary success, which can be formally described as the contribution of a particular individual (including his or her genotype and phenotype) to the next generation.
Haplotype	chromosomal regions for which there are more than one genetic variant in LD; these variants are often inherited together.
Linkage disequilibrium (LD)	the observation that some variants are inherited together more often than expected by random chance based on their individual frequencies. LD is used as an indicator of selection and demographic history.
Multi-allelism	the phenomenon of having more than two alleles in a given population, resulting in more than three diploid genotype states. SNPs are most frequently biallelic. By contrast, CNVs

	can exist in multiple allelic forms (e.g. 2–15 diploid copies for the amylase gene).
Neofunctionalization	the process by which a duplicated gene gains a different but often related function from its parent gene. It was hypothesized that the gene duplication event creates redundancy so that the new gene copy can accumulate mutations that then allow it to gain a new molecular function.
Neutral evolutionary pressures	non-selective forces, such as demography (e.g. population expansions and reductions) and stochastic events (e.g. genetic drift). Most of the variation in the genome is shaped by neutral evolutionary pressures.
Nonallelic homologous recombination (NAHR)	the genetic exchange between similar sequences, such as segmental duplications, as opposed to truly homologous sequences on paired chromosomes. Also known as 'ectopic' recombination.
Population differentiation	a greater difference in the frequency of a variant between than within populations. High population differentiation is a potential sign of positive selection acting on the variant.
Positive selection	the evolutionary process in which selective pressures favor a new or existing variant over other variants. As such, the frequency of the favored variant increases in the population.
Purifying (negative) selection	process favoring the pre-mutation genotype over any new mutations. Loci under purifying selection rarely vary in the population.
Segmental duplications	regions of the genome of a given species that are >1 kb and at least 90% identical to each other at the sequence level. They are at a fixed copy number within a species.
Selective sweep	fixation of a variant in a population or species caused by positive selection acting on it.
Single nucleotide polymorphism (SNP)	a single base-pair difference between the genomes of two individuals from the same species. Historically, SNPs were required to be at 1% allelic frequency; however, the term is often broadly used to describe any single nucleotide variant irrespective of its allele frequency. Single nucleotide substitutions, as opposed to SNPs, differ between species.
Ultra-conserved elements	regions of the genome with 200 consecutive nucleotides conserved with 100% identity across multiple species. Some of these regions have been argued to have functional relevance as enhancers and to have evolved under purifying selection [96].

References

1. Varki A, et al. Explaining human uniqueness: genome interactions with environment, behaviour and culture. *Nat. Rev. Genet.* 2008; 9:749–763. [PubMed: 18802414]
2. Zhang J, et al. Accelerated protein evolution and origins of human-specific features: Foxp2 as an example. *Genetics.* 2002; 162:1825–1835. [PubMed: 12524352]

3. Moreau C, et al. Deep human genealogies reveal a selective advantage to be on an expanding wave front. *Science*. 2011; 334:1148–1150. [PubMed: 22052972]
4. Milot E, et al. Evidence for evolution in response to natural selection in a contemporary human population. *Proc. Natl. Acad. Sci. U.S.A.* 2011; 108:17040–17045. [PubMed: 21969551]
5. Tishkoff SA, et al. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* 2007; 39:31–40. [PubMed: 17159977]
6. Tishkoff SA, et al. Haplotype diversity and linkage disequilibrium at human *G6PD*: recent origin of alleles that confer malarial resistance. *Science*. 2001; 293:455–462. [PubMed: 11423617]
7. Perry GH, et al. Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* 2007; 39:1256–1260. [PubMed: 17828263]
8. Stefansson H, et al. A common inversion under selection in Europeans. *Nat. Genet.* 2005; 37:129–137. [PubMed: 15654335]
9. Flint J, et al. High frequencies of alpha-thalassaemia are the result of natural selection by malaria. *Nature*. 1986; 321:744–750. [PubMed: 3713863]
10. Iafrate AJ, et al. Detection of large-scale variation in the human genome. *Nat. Genet.* 2004; 36:949–951. [PubMed: 15286789]
11. Sebat J, et al. Large-scale copy number polymorphism in the human genome. *Science*. 2004; 305:525–528. [PubMed: 15273396]
12. Mills RE, et al. Mapping copy number variation by population-scale genome sequencing. *Nature*. 2011; 470:59–65. [PubMed: 21293372]
13. Park H, et al. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat. Genet.* 2010; 42:400–405. [PubMed: 20364138]
14. Conrad DF, et al. Origins and functional impact of copy number variation in the human genome. *Nature*. 2010; 464:704–712. [PubMed: 19812545]
15. Cooper GM, et al. A copy number variation morbidity map of developmental delay. *Nat. Genet.* 2011; 43:838–846. [PubMed: 21841781]
16. de Cid R, et al. Deletion of the late cornified envelope *LCE3B* and *LCE3C* genes as a susceptibility factor for psoriasis. *Nat. Genet.* 2009; 41:211–215. [PubMed: 19169253]
17. Bochukova EG, et al. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature*. 2010; 463:666–670. [PubMed: 19966786]
18. Jacquemont S, et al. Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature*. 2011; 478:97–102. [PubMed: 21881559]
19. McCarroll SA, et al. Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease. *Nat. Genet.* 2008; 40:1107–1112. [PubMed: 19165925]
20. Sebat J, et al. Strong association of *de novo* copy number mutations with autism. *Science*. 2007; 316:445–449. [PubMed: 17363630]
21. Stefansson H, et al. Large recurrent microdeletions associated with schizophrenia. *Nature*. 2008; 455:232–236. [PubMed: 18668039]
22. The International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*. 2008; 455:237–241. [PubMed: 18668038]
23. Yamanaka M, et al. Deletion polymorphism of *SIGLEC14* and its functional implications. *Glycobiology*. 2009; 19:841–846. [PubMed: 19369701]
24. Schlattl A, et al. Relating CNVs to transcriptome data at fine-resolution: assessment of the effect of variant size, type, and overlap with functional regions. *Genome Res*. 2011; 21:2004–2013. [PubMed: 21862627]
25. Ohno, S. *Evolution by Gene Duplication*. Allen & Unwin; 1970.
26. Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* 2010; 11:97–108. [PubMed: 20051986]
27. Stranger BE, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*. 2007; 315:848–853. [PubMed: 17289997]
28. Cooper GM, et al. Mutational and selective effects on copy-number variants in the human genome. *Nat. Genet.* 2007; 39:S22–S29. [PubMed: 17597777]

29. Asthana S, et al. Widely distributed noncoding purifying selection in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* 2007; 104:12410–12415. [PubMed: 17640883]
30. Kryukov GV, et al. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* 2007; 80:727–739. [PubMed: 17357078]
31. Sabeti PC, et al. Genome-wide detection and characterization of positive selection in human populations. *Nature.* 2007; 449:913–918. [PubMed: 17943131]
32. Hurles ME, et al. The functional impact of structural variation in humans. *Genome.* 2010; 24:238–245.
33. Locke DP, et al. Comparative and demographic analysis of orang-utan genomes. *Nature.* 2011; 469:529–533. [PubMed: 21270892]
34. Hastings P, et al. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* 2009; 10:551–564. [PubMed: 19597530]
35. Charlesworth B, et al. The effect of deleterious mutations on neutral molecular variation. *Genetics.* 1993; 134:1289–1303. [PubMed: 8375663]
36. Derti A, et al. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat. Genet.* 2006; 38:1216–1220. [PubMed: 16998490]
37. Itsara A, et al. *De novo* rates and selection of large copy number variation. *Genome Res.* 2010; 20:1469–1481. [PubMed: 20841430]
38. Gonzalez E, et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science.* 2005; 307:1434–1440. [PubMed: 15637236]
39. Hardwick RJ, et al. A worldwide analysis of beta-defensin copy number variation suggests recent selection of a high-expressing DEFB103 gene copy in East Asia. *Hum. Mutat.* 2011; 32:743–750. [PubMed: 21387465]
40. Kidd JM, et al. Population stratification of a common *APOBEC* gene deletion polymorphism. *PLoS Genet.* 2007; 3:e63. [PubMed: 17447845]
41. Xue Y, et al. Adaptive evolution of *UGT2B17* copy-number variation. *Am. J. Hum. Genet.* 2008; 83:337–346. [PubMed: 18760392]
42. Craddock N, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature.* 2010; 464:713–720. [PubMed: 20360734]
43. Waszak SM, et al. Systematic inference of copy-number genotypes from personal genome sequencing data reveals extensive olfactory receptor gene content diversity. *PLoS Comput. Biol.* 2010; 6:e1000988. [PubMed: 21085617]
44. Sudmant PH, et al. Diversity of human copy number variation and multicopy genes. *Science.* 2010; 330:641–646. [PubMed: 21030649]
45. Gokcumen O, et al. Refinement of primate CNV hotspots identifies candidate genomic regions evolving under positive selection. *Genome Biol.* 2011; 12:R52. [PubMed: 21627829]
46. Fu W, et al. Identification of copy number variation hotspots in human populations. *Am. J. Hum. Genet.* 2010; 87:494–504.
47. Locke DP, et al. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* 2006; 79:275–290. [PubMed: 16826518]
48. McCarroll SA, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* 2008; 40:1166–1174. [PubMed: 18776908]
49. Campbell CD, et al. Population-genetic properties of differentiated human copy-number polymorphisms. *Am. J. Hum. Genet.* 2011; 88:317–332. [PubMed: 21397061]
50. Frisse L, et al. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am. J. Hum. Genet.* 2001; 69:831–843. [PubMed: 11533915]
51. Marques-Bonet T, et al. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature.* 2009; 457:877–881. [PubMed: 19212409]
52. She X, et al. A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. *Genome Res.* 2006; 16:576–583. [PubMed: 16606706]

53. Abu Bakar S, et al. Allelic recombination between distinct genomic locations generates copy number diversity in human beta-defensins. *Proc. Natl. Acad. Sci. U.S.A.* 2009; 106:853–858. [PubMed: 19131514]
54. Dumas L, et al. Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res.* 2007; 17:1266–1277. [PubMed: 17666543]
55. Han MV, et al. Adaptive evolution of young gene duplicates in mammals. *Genome Res.* 2009; 19:859–867. [PubMed: 19411603]
56. Johnson ME, et al. Positive selection of a gene family during the emergence of humans and African apes. *Nature.* 2001; 413:514–519. [PubMed: 11586358]
57. Gazave E, et al. Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. *Genome Res.* 2011; 21:1626–1639. [PubMed: 21824994]
58. Ciccarelli FD, et al. Complex genomic rearrangements lead to novel primate gene function. *Genome Res.* 2005; 15:343–351. [PubMed: 15710750]
59. Perry GH, et al. Copy number variation and evolution in humans and chimpanzees. *Genome Res.* 2008; 18:1698–1710. [PubMed: 18775914]
60. Niu A, et al. Rapid evolution and copy number variation of primate *RHOXF2*, an X-linked homeobox gene involved in male reproduction and possibly brain function. *BMC Evol. Biol.* 2011; 11:298. [PubMed: 21988730]
61. Yu Y, et al. Evolution of the *DAZ* gene and the *AZFc* region on primate Y chromosomes. *BMC Evol. Biol.* 2008; 8:96. [PubMed: 18366765]
62. Popesco MC, et al. Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. *Science.* 2006; 313:1304–1307. [PubMed: 16946073]
63. Sikela JM. The jewels of our genome: the search for the genomic changes underlying the evolutionarily unique capacities of the human brain. *PLoS Genet.* 2006; 2:e80. [PubMed: 16733552]
64. Fortna A, et al. Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol.* 2004; 2:E207. [PubMed: 15252450]
65. Chen LT, et al. A candidate target for G protein action in brain. *J. Biol. Chem.* 1999; 274:26931–26938. [PubMed: 10480904]
66. Doggett NA, et al. A 360-kb interchromosomal duplication of the human *HYDIN* locus. *Genomics.* 2006; 88:762–771. [PubMed: 16938426]
67. Brunetti-pierri N, et al. Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nat. Genet.* 2008; 40:1466–1471. [PubMed: 19029900]
68. Traherne JA, et al. Mechanisms of copy number variation and hybrid gene formation in the *KIR* immune gene complex. *Hum. Mol. Genet.* 2010; 19:737–751. [PubMed: 19959527]
69. Han K, et al. Identification of a genomic reservoir for new *TRIM* genes in primate genomes. *PLoS Genet.* 2011; 7:e1002388. [PubMed: 22144910]
70. Hirayasu K, et al. Evidence for natural selection on leukocyte immunoglobulin-like receptors for *HLA* class I in Northeast Asians. *Am. J. Hum. Genet.* 2008; 82:1075–1083. [PubMed: 18439545]
71. Hollox EJ, et al. Psoriasis is associated with increased beta-defensin genomic copy number. *Nat. Genet.* 2008; 40:23–25. [PubMed: 18059266]
72. Stankiewicz P, Lupski JR. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* 2002; 18:74–82. [PubMed: 11818139]
73. McLean CY, et al. Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature.* 2011; 471:216–219. [PubMed: 21390129]
74. Johansson I, et al. Inherited amplification of an active gene in the cytochrome P450 *CYP2D* locus as a cause of ultrarapid metabolism of debrisoquine. *Proc. Natl. Acad. Sci. U.S.A.* 1993; 90:11825–11829. [PubMed: 7903454]
75. Elkalioubie A, et al. Near-fatal tramadol cardiotoxicity in a *CYP2D6* ultrarapid metabolizer. *Eur. J. Clin. Pharmacol.* 2011; 67:855–858. [PubMed: 21691803]
76. Yasukochi Y, Satta Y. Evolution of the *CYP2D* gene cluster in humans and four non-human primates. *Genes Genet. Syst.* 2011; 86:109–116. [PubMed: 21670550]

77. Heim MH, Meyer U.a. Evolution of a highly polymorphic human cytochrome P450 gene cluster: *CYP2D6*. *Genomics*. 1992; 14:49–58. [PubMed: 1358797]
78. Kimura S, et al. The human debrisoquine 4-hydroxylase (*CYP2D*) locus: sequence and identification of the polymorphic *CYP2D6* gene, a related gene, and a pseudogene. *Am. J. Hum. Genet.* 1989; 45:889–904. [PubMed: 2574001]
79. Jakobsson J, et al. Large differences in testosterone excretion in Korean and Swedish men are strongly associated with a UDP-glucuronosyl transferase 2B17 polymorphism. *J. Clin. Endocrinol. Metabol.* 2006; 91:687–693.
80. Swanson C, et al. The uridine diphosphate glucuronosyl-transferase 2B15 D85Y and 2B17 deletion polymorphisms predict the glucuronidation pattern of androgens and fat mass in men. *J. Clin. Endocrinol. Metabol.* 2007; 92:4878–4882.
81. McCarroll SA, et al. Donor–recipient mismatch for common gene deletion polymorphisms in graft-versus-host disease. *Nature Genet.* 2009; 41:1341–1344. [PubMed: 19935662]
82. Yang T, et al. Genome-wide copy-number-variation study identified a susceptibility gene, *UGT2B17*, for osteoporosis. *Am. J. Hum. Genet.* 2008; 83:663–674. [PubMed: 18992858]
83. Biswas M, et al. Reduced total testosterone concentrations in young healthy South Asian men are partly explained by increased insulin resistance but not by altered adiposity. *Clin. Endocrinol.* 2010; 73:457–462.
84. Conrad DF, et al. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genetics.* 2006; 38:1251–1260. [PubMed: 17057719]
85. Goossens M, et al. Triplicated a-globin loci. *Genetics.* 1980; 77:518–521.
86. Lau YL, et al. Prevalence and genotypes of alpha- and beta-thalassemia carriers in Hong Kong – implications for population screening. *N. Engl. J. Med.* 1997; 336:1298–1301. [PubMed: 9113933]
87. May J, et al. Hemoglobin variants and disease manifestations in severe falciparum malaria. *J. Am. Med. Assoc.* 2007; 297:2220–2226.
88. Urban TJ, et al. *CCL3L1* and HIV/AIDS susceptibility. *Science.* 2010; 15:1110–1112.
89. Field SF, et al. Experimental aspects of copy number variant assays at *CCL3L1*. *Nat. Genet.* 2010; 15:1115–1117.
90. Nozawa M, et al. Genomic drift and copy number variation of sensory receptor genes in humans. *Proc. Natl. Acad. Sci. U.S.A.* 2007; 104:20421–20426. [PubMed: 18077390]
91. Gilad Y, et al. Loss of olfactory receptor genes coincides with the acquisition of full trichromatic vision in primates. *PLoS Biol.* 2004; 2:E5. [PubMed: 14737185]
92. Nei M, et al. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. *Genetics.* 2008; 9:951–963. [PubMed: 19002141]
93. Mandel AL, et al. Individual differences in *AMY1* gene copy number, salivary α -amylase levels, and the perception of oral starch. *PLoS ONE.* 2010; 5:e13352. [PubMed: 20967220]
94. de Wijk RA, et al. The role of alpha-amylase in the perception of oral texture and flavour in custards. *Physiol. Behav.* 2004; 83:81–91. [PubMed: 15501494]
95. Deeb SS. Genetics of variation in human color vision and the retinal cone mosaic. *Curr. Opin. Genet. Dev.* 2006; 16:301–307. [PubMed: 16647849]
96. Bejerano G, et al. Ultraconserved elements in the human genome. *Science.* 2004; 304:1321–1325. [PubMed: 15131266]
97. Grossman SR, et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science.* 2010; 327:883–886. [PubMed: 20056855]
98. Hurst LD. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* 2002; 18:486. [PubMed: 12175810]
99. McDonald JH, Kreitman M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature.* 1991; 351:652–654. [PubMed: 1904993]
100. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics.* 1989; 123:585–595. [PubMed: 2513255]
101. Fay J, Wu C. Hitchhiking under positive darwinian selection. *Genetics.* 2000; 155:1405–1413. [PubMed: 10880498]

102. Wright, S. *Evolution and the Genetics of Populations*. University of Chicago Press; 1984.
103. Redon R, et al. Global variation in copy number in the human genome. *Nature*. 2006; 444:444–454. [PubMed: 17122850]
104. Lewontin R, Kojima K. The evolutionary dynamics of complex polymorphisms. *Evolution*. 1960; 14:458–472.
105. Sabeti PC, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*. 2002; 419:832–837. [PubMed: 12397357]
106. Voight BF, et al. A map of recent positive selection in the human genome. *PLoS Biol*. 2006; 4:e72. [PubMed: 16494531]
107. Tuzun E, et al. Fine-scale structural variation of the human genome. *Nat. Genet*. 2005; 37:727–732. [PubMed: 15895083]
108. Kidd JM, et al. Mapping and sequencing of structural variation from eight human genomes. *Nature*. 2008; 453:56–64. [PubMed: 18451855]
109. Korbel JO, et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science*. 2007; 318:420–426. [PubMed: 17901297]
110. Mills RE, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res*. 2006; 16:1182–1190. [PubMed: 16902084]
111. Quinlan AR, Hall IM. Characterizing complex structural variation in germline and somatic genomes. *Trends Genet*. 2011; 28:43–53. [PubMed: 22094265]
112. Teague B, et al. High-resolution human genome structure by single-molecule analysis. *Proc. Natl. Acad. Sci. U.S.A.* 2010; 107:1–6.

Box 1. Tests of selection and their application to CNVs

Generally, variants are described as being neutral, under negative selection, under positive selection or, occasionally, under balancing selection. Most empirical tests for selection examine potential deviations from neutrality. Positive selection causes the allele frequency of a variant to rise, and it can drive sequence divergence and alter the relationship of the selected variant with nearby variants (Figure 1). These signatures can be used to detect positive selection. Often, more than one line of evidence is needed to convincingly state that a variant or gene is under positive selection [97]. The types of signatures that are examined can broadly be classified as frequency based (e.g. population differentiation), linkage based (e.g. extended homozygosity and LD), or substitution based (e.g. high ratio of nonsynonymous substitutions to synonymous substitutions) (Table I). Some tests of selection work better for detecting recent positive selection (i.e. within human populations), whereas others are better at detecting ancient positive selection (i.e. between human and non-human primates), as the signals of selection tend to decay over time. Such tests of selection need to be compared to a null model of neutrality. This null model is usually created from one of two sources: first, primarily neutral regions of the genome, such as introns or processed pseudogenes, are compared to the variant(s) of interest. Second, a genome-wide distribution of statistics can be created, and this is used to determine whether the variant(s) of interest are outliers in the distribution.

Perhaps the most convincing evidence of positive selection results from substitution based tests. Substitution based tests of positive selection (such as the K_d/K_s ratio) utilize the ratio of nonsynonymous and synonymous mutations to infer positive selection. The K_d/K_s ratio is a powerful test for species-specific population selection when cross-species data are available [98]. Under neutral conditions, the ratio should be similar to 1, and high proportions of nonsynonymous mutations can indicate positive selection acting on the gene. Also, a related test for cross-species comparisons is the McDonald-Kreitman test, which measures the genic nonsynonymous and synonymous mutations within and between species [99]. If the proportion of between-species nonsynonymous differences is high, but the within-species variation is very low, this indicates positive selection. There are several studies that successfully used these measures for SNPs, short tandem repeats (STRs), RFLPs, and more recently, CNVs (Table 1, main text) [56]. In special circumstances, individual studies have successfully applied tests of neutrality that were originally designed for other types of variants to understand selective pressures acting on CNVs [7,41,45].

Table 1

Select examples of tests of neutrality broadly split into three categories: frequency based, linkage based and substitution based

Test	Brief description	Refs
Frequency based		
Tajima's D	Measures the excess of high and low frequency alleles in a population. An excess of both high and low frequency alleles is indicative of recent positive selection.	[100]
Fay and Wu's H	Measures the excess of high frequency variants compared to intermediate frequency variants. An excess of high frequency alleles is indicative of recent positive selection.	[101]
F_{ST}	Determines whether a variant has differentiated allelic frequencies between populations more so than within populations. High population differentiation is indicative of positive selection acting in one or more populations.	[102]
V_{ST}	Determines whether a variant has differentiated array comparative genomic hybridization intensity data between populations, more so than within populations. High population differentiation is indicative of positive selection acting in one or more populations.	[103]
Linkage based		
LD	Determines whether multiple alleles occur together more often than expected by chance. Regions under positive selection tend to have high LD.	[104]
Extended haplotype homozygosity (EHH)	Measures the length of haplotypes for which variants exist in a homozygous state. Long stretches of homozygosity are indicative of recent selective sweeps.	[105]
Integrated haplotype score (iHS)	An extension of EHH that adjusts for whether the SNP in question is ancestral or derived for more sensitivity in detecting recent sweeps. Regions of positive selection tend to have long stretches of homozygosity.	[106]
Substitution based		
K_A/K_S	Determines the ratio of nonsynonymous to synonymous substitutions as those genes with high ratios are more likely to be under positive selection.	[98]
McDonald-Kreitman	Uses the ratio of fixed and polymorphic synonymous and nonsynonymous sites to determine the likelihood of neutrality. Genes with a high proportion of nonsynonymous fixed differences are indicative of positive selection.	[99]

Box 2. Genome-wide methods to detect CNVs

As genomics technologies, namely second-generation sequencing and array comparative genomic hybridization (aCGH), have advanced, so has the discovery of CNVs. CNVs were initially discovered using large insert clone aCGH, a technique based on hybridizing differentially labeled reference samples and test samples to large clone-derived DNAs that had been spotted on an array [103]. Subsequently, oligo-based aCGH allowed for high-throughput discovery of CNVs with unprecedented precision (Figure I). In addition, SNP arrays, in which a single fluorescently labeled sample is hybridized to millions of oligos, have been used to identify CNVs [48]. Two more recent studies used millions of oligo probes by aCGH to identify CNVs that are 500 base pairs (bp) or larger with a 50-bp resolution to detect the breakpoints [13,14]. Once CNVs are discovered, some can then be genotyped by either aCGH platforms with probes targeting the known CNVs or SNP-based arrays.

DNA sequencing, in addition to array-based technologies, has also been crucial in the accurate characterization of CNVs. Initially, first-generation (capillary) paired-end sequencing was used to determine whether two paired sequences from the same clone mapped to the reference genome further away or closer together than expected, indicating a loss or gain of genomic DNA in the sample, respectively [107–109]. In a similar fashion, split-reads, which map to separate locations in the reference genome, leaving a gap in between the two fragments of the read, have been utilized to discover CNVs [110]. These methodologies have recently been adapted to study CNVs using second-generation sequencing [12] (Figure I). Sequencing technologies are now being used to resolve even complex CNVs [111]. In addition to arrays and sequencing, visual-based technologies, such as optical mapping, hold promise for genome-wide methods to detect CNVs [112].

Figure 1.

Four methods for detecting copy number variants (CNVs). **(a)** Array comparative genomic hybridization (aCGH) is performed by shearing DNA, often by enzymatic cleavage, heat or sonication. Reference DNA is then labeled with one dye (e.g. Cy3) whereas test DNA is labeled with a different dye (e.g. Cy5). Test and reference DNA are then co-hybridized to an array spotted with thousands of oligonucleotide probes. The relative signal intensities of the dyes for each probe are normalized and then converted to $\log_2$ ratios (a positive $\log_2$ ratio indicates a gain in the test relative to the reference and a negative $\log_2$ ratio indicates a loss). Multiple probes in a region with $\log_2$ ratios deviating from 0 are used to make CNV calls. **(b)** Second-generation sequencing of DNA that has been sheared and ligated to specific adaptors. The resulting sequences are aligned to a reference genome. Read depth across the genome is normalized by GC-content and other parameters and converted to Z-scores. **(c)** DNA is either cloned into a vector and sequenced by traditional capillary-based methods or sequenced by second-generation technologies as described in **(b)**. Specific algorithms are used when aligning the reads to a reference genome that allow for a gap in the alignment. Such gaps indicate a loss in the test genome relative to the sequenced reference. **(d)** DNA is size selected and then either cloned into a vector for capillary sequencing or ligated to adaptors and subjected to second-generation sequencing. The ends of the insert are sequenced and aligned to a reference genome. If the ends align further away or closer together than expected based on the distribution of insert sizes, this indicates a loss or a gain in the test genome, respectively.

Figure 1.

Schematic depicting positive selection that alters the genomic context of the selected variant and can be detected by tests of neutrality. Lines represent haplotypes in a theoretical population; different colored circles represent single nucleotide polymorphisms (SNPs) on the haplotype; rectangles represent copy number variants (CNVs); and orange shapes represent substrates for positive selection. **(a)** Theoretical population with representative haplotypes (top). After this population undergoes neutral evolution at this locus for many generations (left), new mutations occur (black circles). The frequencies of some haplotypes and, therefore, some variants, change because of genetic drift. When positive selection occurs (right), the orange variant increases in allelic frequency and reaches fixation (selective sweep). In this example, fixation occurred quickly and recombination has not yet separated the selected variant from other variants on its shared haplotype. Thus, this population exhibits high linkage disequilibrium (LD) at this locus, as indicated by the dashed vertical lines. **(b)** Theoretical population [as in (a)] with a CNV derived from either a single ancestral event (left) or two independent occurrences in the same region (right). Next, a selective sweep occurs owing to positive selection acting upon the CNV. After the selective pressure has been lifted, the haplotypes are free to accumulate neutral mutations [such as the SNPs (black circles)]. Using measures of LD, extended haplotype homozygosity, population differentiation, K_a/K_s ratios, and the frequency of rare and common variants, this population can be tested for the occurrence of a selective event. Because two different haplotypes were selected for in the scenario on the right, some of these features of selection will be more difficult to ascertain.

Table I

Select examples of tests of neutrality broadly split into three categories: frequency based, linkage based and substitution based

Test	Brief description	Refs
Frequency based		
Tajima's D	Measures the excess of high and low frequency alleles in a population. An excess of both high and low frequency alleles is indicative of recent positive selection.	[100]
Fay and Wu's H	Measures the excess of high frequency variants compared to intermediate frequency variants. An excess of high frequency alleles is indicative of recent positive selection.	[101]
F_{ST}	Determines whether a variant has differentiated allelic frequencies between populations more so than within populations. High population differentiation is indicative of positive selection acting in one or more populations.	[102]
V_{ST}	Determines whether a variant has differentiated array comparative genomic hybridization intensity data between populations, more so than within populations. High population differentiation is indicative of positive selection acting in one or more populations.	[103]
Linkage based		
LD	Determines whether multiple alleles occur together more often than expected by chance. Regions under positive selection tend to have high LD.	[104]
Extended haplotype homozygosity (EHH)	Measures the length of haplotypes for which variants exist in a homozygous state. Long stretches of homozygosity are indicative of recent selective sweeps.	[105]
Integrated haplotype score (iHS)	An extension of EHH that adjusts for whether the SNP in question is ancestral or derived for more sensitivity in detecting recent sweeps. Regions of positive selection tend to have long stretches of homozygosity.	[106]
Substitution based		
K_a/K_s	Determines the ratio of nonsynonymous to synonymous substitutions as those genes with high ratios are more likely to be under positive selection.	[98]
McDonald-Kreitman	Uses the ratio of fixed and polymorphic synonymous and nonsynonymous sites to determine the likelihood of neutrality. Genes with a high proportion of nonsynonymous fixed differences are indicative of positive selection.	[99]

Table 1Examples of CNV genes showing evidence of positive selection in humans^a

Gene and/or gene family	Feature(s) of selection	Fixed or variable with humans	Refs
<i>ACOT1</i>	High population differentiation	Variable	[49]
<i>ACOT2</i>	High population differentiation	Variable	[44]
<i>ACTN2</i> ^b	High population differentiation	Variable	[14]
<i>ADAM5P</i>	High population differentiation	Variable	[44]
<i>ADRA1B</i>	High population differentiation	Variable	[103]
<i>AMAC1/AMAC1L1</i>	Nonsynonymous substitution	Fixed	[55]
<i>AMY1</i>	High population differentiation	Variable	[7]
<i>AMY2B</i>	Nonsynonymous substitution	Fixed	[55]
<i>ANKRD12</i>	High population differentiation	Variable	[49]
<i>ANKRD20A8P</i>	Nonsynonymous substitution	Fixed	[55]
<i>ANKRD36</i>	Nonsynonymous substitution, high population differentiation	Variable	[44,55]
<i>ANKRD62</i>	Nonsynonymous substitution	Fixed	[55]
<i>APOBEC3B</i>	High population differentiation	Variable	[40]
<i>APOL2/APOL4</i> ^c	Linkage	Variable	[14]
<i>AR</i> ^c	Functional experiments	Fixed	[73]
<i>ARHGEF5</i>	High population differentiation	Variable	[44]
<i>ATP6V1B2</i>	High population differentiation	Variable	[49]
<i>BSX</i>	High population differentiation	Variable	[103] ^d
<i>C17orf58</i>	High population differentiation	Variable	[44]
<i>C20orf191</i>	Nonsynonymous substitution	Fixed	[55]
<i>C2orf78</i>	High population differentiation	Variable	[44]
<i>CAMTA1</i>	High population differentiation	Variable	[103] ^d
<i>CCDC12</i>	High population differentiation	Variable	[103] ^d
<i>CCDC77/NM_001130148</i>	High population differentiation	Variable	[49]
<i>CCL3L1</i>	High population differentiation	Variable	[14,38,44,49]
<i>CCL4</i>	High population differentiation	Variable	[44,49]
<i>CCT8L1</i>	Nonsynonymous substitution	Variable	[55]
<i>CES1</i>	High population differentiation	Variable	[44]
<i>CFHR1</i>	High population differentiation	Variable	[44]
<i>CROCCL1</i>	High population differentiation	Variable	[44]
<i>CSAG1</i>	Nonsynonymous substitution	Fixed	[55]
<i>CYP2D6</i>	Nonsynonymous substitution	Variable	[77]
<i>DCTN2</i>	High population differentiation	Variable	[103] ^d
<i>DCTN4</i>	High population differentiation	Variable	[103] ^d

Gene and/or gene family	Feature(s) of selection	Fixed or variable with humans	Refs
<i>DDIT3</i>	High population differentiation	Variable	[103] ^d
<i>DEFB103</i>	High population differentiation	Variable	[39]
<i>DHFRL1</i>	High population differentiation	Variable	[44]
<i>DHX40P</i>	High population differentiation	Variable	[44]
<i>DTX3</i>	High population differentiation	Variable	[103] ^d
<i>DUSP11</i>	High population differentiation	Variable	[49]
<i>DUX4</i>	High population differentiation	Variable	[44]
<i>EFR3A</i>	High population differentiation	Variable	[49]
ENSG00000180015	Nonsynonymous substitution	Fixed	[55]
ENSG00000188532	Nonsynonymous substitution	Fixed	[55]
ENSP00000369576	Nonsynonymous substitution	Fixed	[55]
<i>ERC2</i>	High population differentiation	Variable	[44]
<i>ESPNP</i>	High population differentiation	Variable	[44]
<i>FAM118A</i>	High population differentiation	Variable	[49]
<i>FAM157A</i>	High population differentiation	Variable	[44]
<i>FAM66D</i>	High population differentiation	Variable	[44]
<i>FAM75A</i>	Nonsynonymous substitution	Fixed	[55]
<i>FAM90A</i>	Nonsynonymous substitution	Fixed	[55]
<i>FBXO40</i>	High population differentiation	Variable	[103] ^d
<i>FNBP4</i>	High population differentiation	Variable	[49]
<i>FUCA2</i>	High population differentiation	Variable	[103] ^d
<i>G antigen</i>	Nonsynonymous substitution	Fixed	[55]
<i>GADD45G</i> ^c	Functional experiments	Fixed	[73]
<i>GLII</i>	High population differentiation	Variable	[103] ^d
<i>Golgin subfamily A (ENSP00000289798)</i>	Nonsynonymous substitution	Fixed	[55]
<i>GSTT1</i>	High population differentiation	Variable	[44]
<i>GTF2H2</i>	High population differentiation	Variable	[44]
<i>GUSBL1</i>	High population differentiation	Variable	[44]
<i>HEATR4</i>	High population differentiation	Variable	[44]
<i>HLA-DPA1</i>	High population differentiation	Variable	[44]
<i>HP</i>	Nonsynonymous substitution	Fixed	[55]
<i>HYDIN</i>	Putative function	Variable	[46]
<i>IKBKB</i> ^c	Linkage	Variable	[14]
<i>IMMP2L</i>	High population differentiation	Variable	[103] ^d
<i>INHBC/E</i>	High population differentiation	Variable	[103] ^d
<i>KIAA1267</i>	High population differentiation	Variable	[44,49]
<i>KIF5A</i>	High population differentiation	Variable	[103] ^d
<i>KIR</i>	Putative function	Variable	[68]

Gene and/or gene family	Feature(s) of selection	Fixed or variable with humans	Refs
<i>KRCC1</i>	High population differentiation	Variable	[49]
<i>KRT14</i>	High population differentiation	Variable	[44]
<i>KRTAP9-9</i>	High population differentiation	Variable	[44]
<i>LGALS9</i>	High population differentiation	Variable	[44,49]
<i>LILRA3</i>	High population differentiation	Variable	[44,49]
<i>LOC100286793</i>	High population differentiation	Variable	[44]
<i>LOC392196</i>	High population differentiation	Variable	[44]
<i>LOC644172</i>	High population differentiation	Variable	[44]
<i>LRRC37A3</i>	High population differentiation	Variable	[44]
<i>MARS</i>	High population differentiation	Variable	[103] ^d
<i>MBD6</i>	High population differentiation	Variable	[103] ^d
<i>MDFI</i>	High population differentiation	Variable	[103] ^d
<i>Methyl CpG binding domain (MBD)</i>	Nonsynonymous substitution	Fixed	[55]
<i>METTL2B</i>	Nonsynonymous substitution	Fixed	[55]
<i>MGC57346</i>	High population differentiation	Variable	[44]
<i>Morpheus/NPIP</i>	Nonsynonymous substitution, high population differentiation	Variable	[44,55,56]
<i>MYL3</i>	High population differentiation	Variable	[103] ^d
<i>NBPF1/PDE4DIP</i>	High population differentiation, nonsynonymous substitution	Variable	[44,49,55,62]
<i>NBPF3</i>	High population differentiation	Variable	[44]
<i>NOTCH2</i>	High population differentiation	Variable	[44,49,103] ^d
<i>NPEPPS</i>	High population differentiation	Variable	[44,49]
<i>OCLN</i>	High population differentiation	Variable	[44,49]
<i>Olfactory receptors</i>	Nonsynonymous substitution	Fixed	[55]
<i>PCDHB13</i>	High population differentiation	Variable	[44]
<i>PDLIM3^b</i>	High population differentiation	Variable	[14]
<i>PDXDC1</i>	High population differentiation	Variable	[44,49]
<i>PGA5</i>	High population differentiation	Variable	[44]
<i>PIP5K2C</i>	High population differentiation	Variable	[103] ^d
<i>PLA2G10</i>	High population differentiation	Variable	[44]
<i>POLR2J4</i>	High population differentiation	Variable	[44]
<i>POM121L4P</i>	High population differentiation	Variable	[44]
<i>PRAMEF11</i>	Nonsynonymous substitution	Fixed	[55]
<i>PRH1</i>	High population differentiation	Variable	[44,103] ^d
<i>PRSS3</i>	High population differentiation	Variable	[44]
<i>PTHR1</i>	High population differentiation	Variable	[103] ^d
<i>Q8NBL2</i>	High population differentiation	Variable	[103] ^d
<i>RanBP2</i>	Nonsynonymous substitution	Fixed	[55,58]

Gene and/or gene family	Feature(s) of selection	Fixed or variable with humans	Refs
<i>RASA4P</i>	High population differentiation	Variable	[44]
<i>RDM1</i>	High population differentiation	Variable	[44]
<i>RFPL4A</i>	High population differentiation	Variable	[44]
<i>RHD</i>	High population differentiation	Variable	[44]
<i>RHOXF2</i>	Nonsynonymous substitution	Fixed	[60]
<i>RIM binding, PBR associated</i>	Nonsynonymous substitution	Fixed	[55]
<i>RNA exonuclease 1 homolog</i>	Nonsynonymous substitution	Fixed	[55]
<i>SAA1</i>	Nonsynonymous substitution	Fixed	[55]
<i>SERPINB4</i>	Nonsynonymous substitution	Fixed	[55]
<i>Sialic acid binding Ig lectin precursor</i>	Nonsynonymous substitution	Fixed	[55]
<i>Similar to nuclear pore membrane protein 121</i>	Nonsynonymous substitution	Fixed	[55]
<i>SIRPB1NM_001135844</i>	High population differentiation	Variable	[49]
<i>SLC25A24</i>	High population differentiation	Variable	[49]
<i>SMARCA2</i>	High population differentiation	Variable	[44]
<i>SNTB1</i>	High population differentiation	Variable	[103] ^d
<i>SPDYE6</i>	High population differentiation	Variable	[44]
<i>TADA2L/ACACA/NM_001166105</i>	High population differentiation	Variable	[49]
<i>TAS2R</i>	High population differentiation	Variable	[49,103] ^d
<i>TBC1D3C</i>	High population differentiation	Variable	[44]
<i>TCEB3C</i>	High population differentiation	Variable	[44,49]
<i>TESSP2</i>	High population differentiation	Variable	[103] ^d
<i>TMIE</i>	High population differentiation	Variable	[103] ^d
<i>TRIM family</i>	Nonsynonymous substitution	Variable	[69]
<i>TRY6</i>	High population differentiation	Variable	[44]
<i>TTC1</i>	High population differentiation	Variable	[103] ^d
<i>UBE2QP2</i>	High population differentiation	Variable	[44]
<i>Ubiquitin carboxyl-terminal hydrolase 17</i>	Nonsynonymous substitution	Fixed	[55]
<i>UDP glucuronosyltransferase precursor</i>	Nonsynonymous substitution	Fixed	[55]
<i>UGT2B17</i>	Linkage, high population differentiation	Variable	[41,44,49,103] ^d
<i>UPLP</i>	High population differentiation	Variable	[44]
<i>USP17</i>	High population differentiation	Variable	[44]
<i>USP18</i>	High population differentiation	Variable	[49]
<i>USP41</i>	Nonsynonymous substitution	Fixed	[55]
<i>ZNF267/TP53TG3</i>	High population differentiation	Variable	[44,49]
<i>ZNF705D</i>	High population differentiation	Variable	[44]
<i>ZNF717</i>	High population differentiation	Variable	[44]
<i>α-globin</i>	High population differentiation	Variable	[9]

^aThis list is not exhaustive.

^bIntronic CNVs.

^cNot overlapped by a CNV, but a CNV is nearby.

\$watermark-text

\$watermark-text

\$watermark-text