nature genetics

Mutational and selective effects on copy-number variants in the human genome

Gregory M Cooper, Deborah A Nickerson & Evan E Eichler

Comprehensive descriptions of large insertion/deletion or segmental duplication polymorphisms (SDs) in the human genome have recently been generated. These annotations, known collectively as structural or copy-number variants (CNVs), include thousands of discrete genomic regions and span hundreds of millions of nucleotides. Here we review the genomic distribution of CNVs, which is strongly correlated with gene, repeat and segmental duplication content. We explore the evolutionary mechanisms giving rise to this nonrandom distribution, considering the available data on both human polymorphisms and the fixed changes that differentiate humans from other species. It is likely that mutational biases, selective effects and interactions between these forces all contribute substantially to the spectrum of human copy-number variation. Although defining these variants with nucleotide-level precision remains a largely unmet but critical challenge, our understanding of their potential medical impact and evolutionary importance is rapidly emerging.

The HapMap project[1] provides a powerful resource for studying the relationship between genetic and phenotypic variation in humans as well as the evolutionary and genealogical history of modern human populations. Technical and practical concerns have led to a variation map composed almost exclusively of single nucleotide polymorphisms (SNPs), despite the fact that other types of variation in the human genome are likely to be of considerable importance. Recent advances, however, have facilitated insights into a number of other variants in the human genome (**Table 1**, see http://genome.ucsc.edu and http://projects.tcag.ca/variation/). These include small insertion and deletion (indel) polymorphisms identified in sequence traces[2–4], common deletion polymorphisms mined from HapMap genotype data[5,6] and variants identified by comparing clone paired-end sequence data to the reference human assembly[7]. There have also been studies using comparative genomic hybridization with microarrays (array-CGH) to identify large (>50 kb) CNVs among dozens of individuals[8–10]. Finally, two recent studies using genome-wide BAC and oligonucleotide arrays have identified large CNVs in hundreds of individuals[11,12], including the reference HapMap samples.

The authors are in the Department of Genome Sciences and Evan E. Eichler is also at the Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA. e-mail: coopergm@u.washington.edu or eee@gs.washington.edu

Published online 27 June 2007; doi:10.1038/ng2054

Based on these studies, there are now over 400,000 annotated indel polymorphisms smaller than 1,000 nucleotides in length[2–4] (**Table 1**). These polymorphisms, particularly for the smallest events affecting from one to a handful of nucleotides, are the second most commonly occurring form of variation in the human genome. They behave similarly to SNPs in terms of their broad genomic distribution, allelic associations with other variants, frequency distributions in human populations and contributions to evolutionary divergence among mammals[2–4,13]. The sequence properties of the larger (>1 kb) variants are less clear. Thus far, more than 4,000 distinct regions of the reference genome assembly have been annotated to harbor CNVs (along with a small number of inversions, which are more difficult to detect and presently less explored), and many are frequently polymorphic in human populations. Although they affect dramatically fewer regions compared with SNPs or small indels, these variants span an estimated 600 Mb of the human genome (see **Table 1** and references therein). We note that this number should be interpreted cautiously, as it is likely to be too large; most known CNVs were identified through BAC array-CGH, a method that lacks precision in identifying breakpoints and is likely to substantially overestimate variant size. Also, there are different degrees of quality control among the different studies and only a minority have been validated across platforms. However, even assuming an order of magnitude difference between currently estimated and true variant sizes, an estimated 60 Mb would still be within CNVs, representing a sizable fraction of the genome (~2%).

Comparisons between chimpanzee and human genomic sequences confirm a similar relative influence of copy-number changes when compared with single-nucleotide substitutions[14–17] (**Table 2**). Whereas ~35 million fixed nucleotide substitutions contribute to a human-chimp sequence divergence of ~1.2%, ~5 million mutations involving gain or loss of DNA raise the total amount of human-chimp genomic divergence to ~5%. Even after excluding small indels, mobile-element insertions and microsatellite expansions, several thousand indel differences (>5 kb) spanning ~115 Mb remain between the two species[15]. Although it is possible that assembly errors may contribute to this estimate, it should be emphasized that many of the larger events were annotated using approaches independent from chimpanzee assembly (namely mapping of whole-genome shotgun sequence and paired-end sequence data from the chimpanzee genome against the high-quality human genome assembly to identify duplications and structurally variant regions[15–17]). Furthermore, the limited number and the size of these events facilitated identification and validation for many of them, and efforts are underway to provide high quality BAC-based sequence data from these regions.

Table 1 Summary of recent analyses of structural variation in the human genome

Reference	Coverage	Analysis	No. of individuals	No. events or regions	Size range (bp)	Average size (bp) Median size (bp)	Total bp
Mills *et al.*, *Genome Res.* 2006	16 million whole-genome shotgun traces	Alignment of sequence traces from SNP Consortium resequencing	36	415,434	1–9,989	20 2	8,360,235
Conrad *et al.*, *Nat. Genet.* 2006	1.3 million SNPs	HapMap SNP-genotyping data mining based on Mendelian inconsistencies	180 (60 parent-offspring trios from CEU and YRI populations)	609	25–993,000	34,996 17,217	21,313,127
McCarroll *et al.*, *Nat. Genet.* 2006	1.3 million SNPs	HapMap SNP-genotyping data mining based on null genotypes, Mendelian inconsistencies and deviations from Hardy-Weinberg equilibrium	269 (CEU, YRI and CHB+JPT populations, including 60 trios)	538	96–745,418	16,874 6,887	9,078,084
Hinds *et al.*, *Nat. Genet.* 2006	100 million to 200 million bp	Oligonucleotide array hybridization	24	1,000	72–8,001	1,379 947	137,912
Tuzun *et al.*, *Nat. Genet.* 2005	×8 coverage fosmid library	Paired end-sequencing	1	297	700–1,944,156	55,706 25,230	14,984,826
Iafrate *et al.*, *Nat. Genet.* 2004	5,264 BACs	BAC array-CGH	55[a]	246	19,597–337,967	146,189 150,395	35,962,540
Sharp *et al.*, *Am. J. Hum. Genet.* 2005	1,986 BACs	BAC array-CGH	47	124	29,514–410,301	170,019 164,704	21,082,320
Sebat *et al.*, *Science* 2004	85,000 oligonucleotides	ROMA-CGH	20	72	754–1,698,859	350,670 199,800	25,248,203
Wong *et al.*, *Am. J. Hum. Genet.* 2007	26,363 BACs	BAC array-CGH	105	1,365[b]	50,459–1,037,332	185,504 175,314	253,212,685
Redon *et al.*, *Nature* 2006	26,574 BACs	BAC array-CGH	269 (CEU, YRI and CHB+JPT populations, including 60 trios)	913	2,639–7,378,760	349,880 227,889	319,440,476
Redon *et al.*, *Nature* 2006	500,000 SNPs	Affyx 500K SNP array analysis	269 (CEU, YRI and CHB+JPT populations, including 60 trios)	980	1,033–3,605,436	165,996 63,140	162,675,683
All variations	NA	NA	NA	323,573	1–7,442,054	1,901 2	615,095,095
All variations >1 kb	NA	NA	NA	4,131	1,004–7,442,053	148,578 93,356	613,774,371

NA, not applicable; ROMA, representational oligonucleotide microarray analysis; CEU (Utah residents with ancestry from northern and western Europe); YRI, Yoruba in Ibadan, Nigeria; CHB+JPT, Han Chinese in Beijing, China and Japanese in Tokyo.
[a]39 healthy controls, 16 with karyotype abnormalities. [b]Accounting for only those sites that showed in two or more individuals.

Thus, CNVs contribute substantially, if not predominantly, to per-nucleotide heterozygosity in human populations and divergence between humans and other species. CNVs are likely to inform all aspects of human genetic analysis, including associations with both rare and common traits, clinical diagnostics and treatments, population demographics, and the molecular and phenotypic evolution of the human species. In this perspective, we focus our analysis on the large (>1 kb) variants identified in recent genome-wide surveys. Although still incomplete, the available data allow us to assess the genomic landscape of CNVs and provide insight into the likely influences of biased mutational mechanisms and natural selection.

Distribution of CNVs in the human genome

We combined published human CNVs (along with a small number of inversions) into a single nonredundant set (**Table 1** and **Supplementary Table 1** online) and evaluated the density of these variants across the reference human genome assembly in 1-Mb nonoverlapping windows (see **Supplementary Note** and **Supplementary Table 2** online). The average genome-wide per-nucleotide density is ~21%, and the median 1-Mb window has a density of ~16.6%. Consistent with previous analyses[9–12], the distribution of CNVs in the genome is nonrandom and highly correlated with other genomic features, including exons ($P = 2 \times 10^{-12}$), segmental duplications (SDs; $P < 2 \times 10^{-16}$) and mobile elements such as *Alu* repeats ($P = 3 \times 10^{-7}$). These correlations are at least partially transitive, but remain very significant even when considered simultaneously (see **Supplementary Note**). Additionally, there are both 'cold' and 'hot' regions of copy-number variation: there are 250 1-Mb regions in which more than half of all bases are within an annotated CNV and 60 regions where more than 90% of all bases are annotated. Not surprisingly, these hotspots cluster in pericentromeric and subtelomeric regions of chromosomes, regions previously recognized as evolutionarily unstable and highly polymorphic[18–20]. This accounts for only ~1/3 of all hotspots, however, with the remainder occurring elsewhere in the genome (**Fig. 1**; **Supplementary Table 2**).

CNVs and segmental duplications. There is a strong relationship between duplicated sequences in the reference genome assembly and copy-number variation in the human genome. More than half of all nucleotides annotated to be within SDs (http://humanparalogy.gs.washington.edu)[21] overlap with CNVs. The average SD density in the most CNV-rich fraction of the genome is ~25%, in contrast to a genome-wide average density of 4–5% and a density of 2–3% in CNV-poor regions (see **Supplementary Note** online). These findings confirm

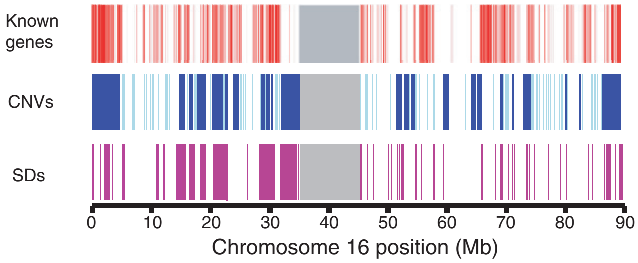


Figure 1 Copy-number variation along human chromosome 16. Regional densities of segmental duplications (SDs), CNVs and exons vary and co-vary across the genome, as seen here for human chromosome 16 (*x* axis corresponds to coordinates in the hg17 assembly). On the bottom row, the positions of all SDs are marked in purple. In the middle row, positions of all CNVs (nonredundant set of all CNVs > 1 kb described in **Table 1** and **Supplementary Table 1**) are marked in light blue and common (>3% frequency as estimated in refs. 11,12) CNVs are marked in darker blue. On the top row, exons of known genes are colored red. Heterochromatic sequence is gray. Note that CNVs are enriched in telomeric (for example, near 0 Mb) and pericentromeric (for example, near 35 Mb) regions and that there is a strong correlation between SDs and CNVs, particularly for common variants. Also note that CNVs and SDs often overlap gene-rich regions (for example, near 30 Mb), but that CNVs may also reside in regions that are poor in both gene and segmental duplication content (for example, 60–66 Mb).

previous estimates that duplicated regions of the genome are enriched 4–10-fold for copy-number variation[9–12] (**Fig. 1**). This strong correlation in part reflects the fundamentally similar nature of CNVs and SDs. Indeed, many SDs represent the reference assembly alleles of CNVs rather than duplication events that are fixed in the human population. To ameliorate this circularity, we stratified SDs using sequence identity as a surrogate for evolutionary age (**Fig. 2**). We reasoned that older duplications (as measured by divergence) are less likely to be presently polymorphic. Consistent with previous analyses[11], the correlation between SDs and CNVs is most striking when only young, high-identity duplications are considered (**Fig. 2a**, red bars). However, the relationship remains pronounced even when only relatively old, divergent SDs are considered (**Fig. 2a**, blue and purple bars). This indicates that the relationship between these two features is not strictly dependent upon sequence identity—that is, longer blocks of nearly perfect sequence identity predisposing to nonallelic homologous recombination—and that the presence of ancient (and probably fixed) copy-number mutations is associated with the presence of currently polymorphic CNVs. Indeed, it is known that more divergent repeat sequences such as *Alu* elements and LINEs may contribute to nonallelic homologous recombination, although in these cases it is the proximity of the homologous repeats that predisposes to the arrangement. Such events are less likely to be recurrent.

However, not all CNVs map to duplicated regions of the reference genome assembly, as approximately half of all CNV nucleotides are within annotations that do not overlap SDs. Interestingly, CNVs that do not overlap SDs are characterized by significantly fewer individuals that show copy number deviation[11,12]. For example, from the BAC array-CGH data presented in Redon *et al.*[12], the average frequency for CNVs associated with SDs is 10.9%, in contrast to only 3.2% for variants that are not associated with SDs ($P = 1.9 \times 10^{-12}$). This difference is even larger when comparing CNVs overlapping young SDs (98–100% identical) versus the remaining sites (20% versus 3.7%, $P = 7.4 \times 10^{-15}$). Thus, both the density and population frequency of copy-number variation are correlated with the presence of duplicated sequences in the reference genome assembly. We note, however, that for many variants it is not allele frequency *per se* that is measured, but rather a measure of how many individuals show some deviation in copy number from the reference sample or assembly, which effectively prohibits distinguishing common alleles from recurrent mutational events and allelic heterogeneity.

CNVs and gene content. There is also a significant relationship between the genomic regions affected by CNVs and gene content. The exon density (including both coding and untranslated regions) in the most CNV-rich regions of the genome is over 2.7%, in contrast to the genome-wide average of 2.1% ($P = 0.0013$; see **Supplementary Note**). Conversely, the density of CNVs in the most gene-rich regions of the genome is over 30%, compared with a genome-wide average of 21% ($P = 3.3 \times 10^{-9}$). Thus, gene-rich regions tend to be rich in copy-number variation and vice versa. However, given that there is a strong correlation between SDs and gene content[21–23], we asked whether the association between copy-number variation and gene content could be separated from its association with SDs (**Fig. 2b**). We partitioned variants that do not overlap SDs from those that do; we still find a significant ($P = 0.019$), albeit weaker, enrichment for copy-number variation in gene-rich regions. However, unlike CNVs that overlap SDs (**Fig. 2b**, red bars), CNVs that are not associated with SDs are also enriched in the most gene-poor regions of the genome ($P = 2.6 \times 10^{-6}$; **Fig. 2b**, blue bars). Thus, the spectrum of copy-number variation in the genome can be separated based on overlap with SDs, and these two groups have distinct gene density profiles.

We also sought to characterize the types of genes seen in CNVs, considering the two groups independently. GO-term[24] and PANTHER[25] analyses reveal that CNVs associated with segmental duplication are highly enriched for genes involved in sensory perception (for example, olfactory receptors) and the immune response, consistent with previous studies[19,20,26–28] (**Fig. 3** and **Supplementary Note**). Other functional categories such as 'cell adhesion' and 'structural proteins' are also enriched, although these observations are driven largely by a few gene clusters, namely the *LCE* and keratin (structural proteins) and protocadherin (cell adhesion) loci (that is, these enrichment values result from a 'jackpot' effect in which one or a few CNVs overlap with dozens of distinct but functionally related genes because these genes reside in a genomic cluster). Interestingly, the protocadherin cluster on human chromosome 5, thought to be important for generating combinatorial complexity in synaptic connections in the brain[29], is particularly rich in copy-number variation, harboring gain and loss annotations from many of the CNV annotation efforts. In fact, detailed analyses of this cluster show that it has been prone to frequent copy-number mutations and gene-conversion events throughout mammalian and vertebrate evolution[30].

CNVs that do not overlap SDs show no enrichment for olfactory receptors and only weak enrichment for defense and immunity proteins (probably as a consequence of the tight association of olfactory receptors and immunoglobulin genes with SDs). These regions are enriched, however, for genes involved in signaling ($P = 2 \times 10^{-7}$; **Fig. 3**). This functional category includes members of the *FGF*, *EGF*, *WNT* and *BMP* families that collectively play a wide variety of roles in regulating organismal

Table 2 Human versus chimpanzee genetic variation

Type	Size	No. events	Mb
Substitution	1 bp	35,000,000	35
Structural	<80 bp	4,930,000	18.1
	80 bp–15 kb	70,000	48.9
	>15 kb	~1,000	21
	SD (LS)	~940	46
	SD (QD)	~590	26
Total structural	>1 bp	5,000,000	160

Between-species differences based principally on a comparison of two representative individuals. QD, quantitative differences between shared SDs; LS, lineage-specific duplications (see refs. 14–17).

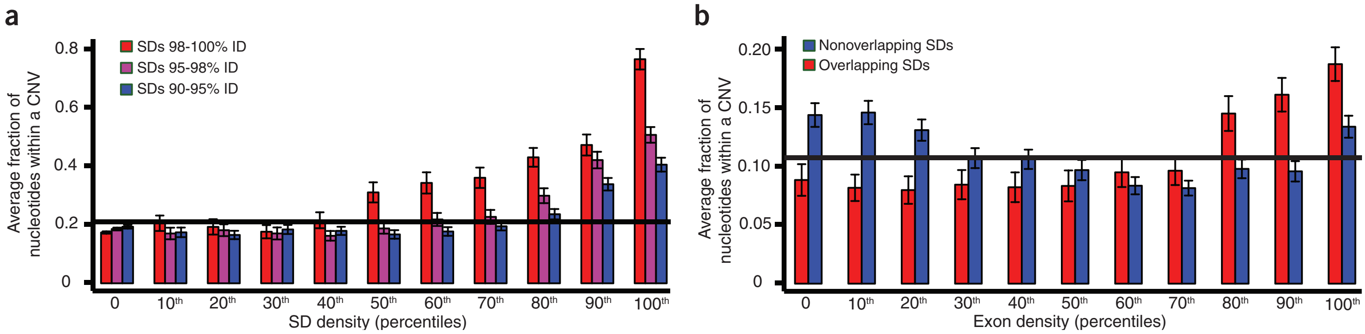


Figure 2 Relationships among copy-number variants, segmental duplications, and genes. (**a**) Copy-number variation and segmental duplication similarity. Nonoverlapping 1-Mb windows were binned according to their segmental duplication content, treating SDs of each of the three indicated levels of percent identity separately (see **Supplementary Note**). Those windows with no SDs of the given similarity level (excluding heterochromatic sequences) were binned into one group ('0'), and the remainder were ranked and binned into deciles (0–10th percentile, 10th–20th, etc.). For the windows within each bin, the average number of nucleotides that are within a CNV is plotted as a vertical bar. For example, nearly 80% of nucleotides within the windows containing the highest density of young (98–100% identity) SDs are within a CNV (red column, 100th-percentile bin). The genome-wide average density of CNVs (~21%) is plotted as a horizontal line; error bars, s.e.m. (**b**) Copy-number variation gene content in duplicated and unique regions of the reference genome assembly. Nonoverlapping 1-Mb windows were binned as described above according to their exon density; those with no annotations (excluding heterochromatic sequences) were binned into one group ('0'), and the remainder were ranked and placed into deciles. For the windows within each bin, the average number of nucleotides that are within a CNV is plotted as a vertical bar, with those CNVs that overlap SDs separated from those that do. For example, nearly 20% of nucleotides within the windows containing the highest density of exons are within a CNV that overlaps an SD (red column, 100th-percentile bin). A horizontal line is drawn at ~10.5%, close to the average density of each class of CNV (that is, those that overlap SDs and those that do not); error bars, s.e.m.

development and cellular growth, proliferation and differentiation. Interestingly, this observation is in direct contrast with previous results[12]; this difference is likely to be due to the rapidly expanding set of CNVs identified in the human genome (~70% increase in total genomic coverage) and the focus on CNVs that do not overlap SDs. We further note that this enrichment is unlikely to be driven by a single study, as there is still a weak enrichment for signaling molecules even after eliminating genes that overlap a CNV from only one study (not shown). We also find a significant ($P = 0.0023$) enrichment for ion channel genes specifically for CNVs that do not overlap duplicated sequence. Although it is unclear at present whether the copy-number variation occurs at these genes *per se* or in their sequence vicinity, this bias suggests two things: (i) forces other than nonallelic homologous recombination of duplicated sequence are likely to be responsible for this variation and (ii) individual humans may differ considerably in the expression or actual complement of such genes. As these genes are involved in a variety of physiological and developmental processes, including medically relevant ones such as cancer (for example, receptor tyrosine kinases) and are frequently targets of therapeutic drugs (for example, voltage-gated ion channels[31]), understanding such differences in the human population is likely to be important in our understanding and treatment of diseases.

Evolutionary mechanisms influencing CNVs

The nonrandom genomic distribution of large CNVs led us to explore the evolutionary processes impacting these variants. Although there is considerable work on the evolutionary fates of gene duplicates[32–34] that is relevant to characterizing many individual loci, these analyses do not explain the spectrum of CNVs in the genome *per se*. Furthermore, these models generally assume that a fundamentally random process underlies the initial generation of duplicated sequence. If particular regions are systematically more or less prone to copy-number mutations for mechanistic reasons, then evolutionary inferences must account for such bias when defining the neutral expectation. Therefore, we undertake a more general exploration and consider to what extent this genomic distribution is likely to be driven by neutral evolutionary processes versus natural selection. Although enforcing a dichotomy between these two is artificial, it provides a context in which to discuss the evolution and population dynamics of CNVs and relate evidence from other analyses. We consider each in turn.

Neutrality of CNVs. A key tenet of a neutralist view of CNVs is that such mutations can exist with weak or no phenotypic consequences. Although it is difficult to establish with certainty that any given mutation does not contribute to a phenotype, several observations support the validity of this prediction. The existence of many CNVs in a large sample of 'normal' individuals indicates that many such mutations confer minimal to no phenotypic consequence within humans. At the very least such variants do not have substantial deleterious effects. More specific albeit anecdotal studies also provide support. An analysis of autistic children and their unaffected parents, for example, found no phenotypic link with a common deletion that effectively eliminates three of the protocadherin genes (discussed above) on chromosome 5 (ref. 35). Also, studies in mice in which large (~1-Mb) deletions are shown to have no major phenotype confirm that even very large variants may be of minimal, or at the very least not seriously deleterious, influence[36]. Finally, very high levels of indel and genomic structural polymorphism are seen in wild-type individuals of the tunicate *Ciona savignyi*, in which 15–20% of all nucleotides are either allele specific or in an inverted orientation when comparing any two homologous chromosomes[37]. This demonstrates that even extreme levels of genomic structural variation are tolerable, if not phenotypically inconsequential, in a healthy, wild population.

Another prediction of the neutralist hypothesis is that the mechanisms that give rise to copy-number mutations are strongly correlated with local genomic features, making particular regions systematically more prone to such mutations. Evidence confirming this prediction is abundant. For example, recent studies show directly that regions flanked by SDs of high sequence similarity are much more likely to harbor copy-number variation than other genomic sites, probably as a result of nonallelic homologous recombination[9,12]. Further evidence for this effect comes from comparisons between the human and chimpanzee genomes: lineage-specific duplications are substantially (~10-fold) more likely to occur in regions near ancestrally duplicated sequence, a phenomenon termed 'duplication shadowing'[17]. The strong correlation between the presence of SDs and the complete collection of known CNVs genome-wide provides further corroboration for a tight mechanistic link.

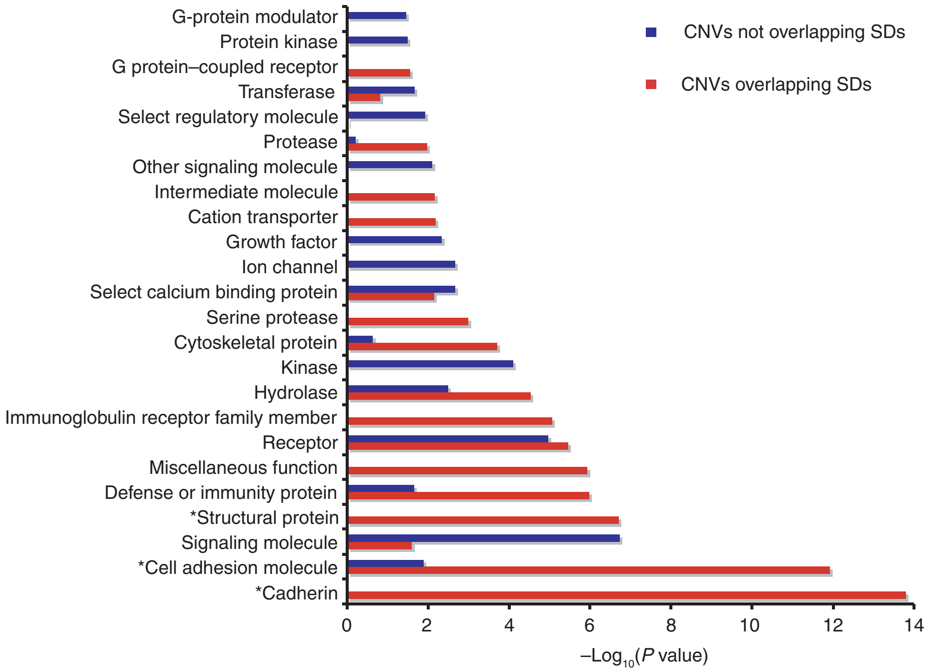


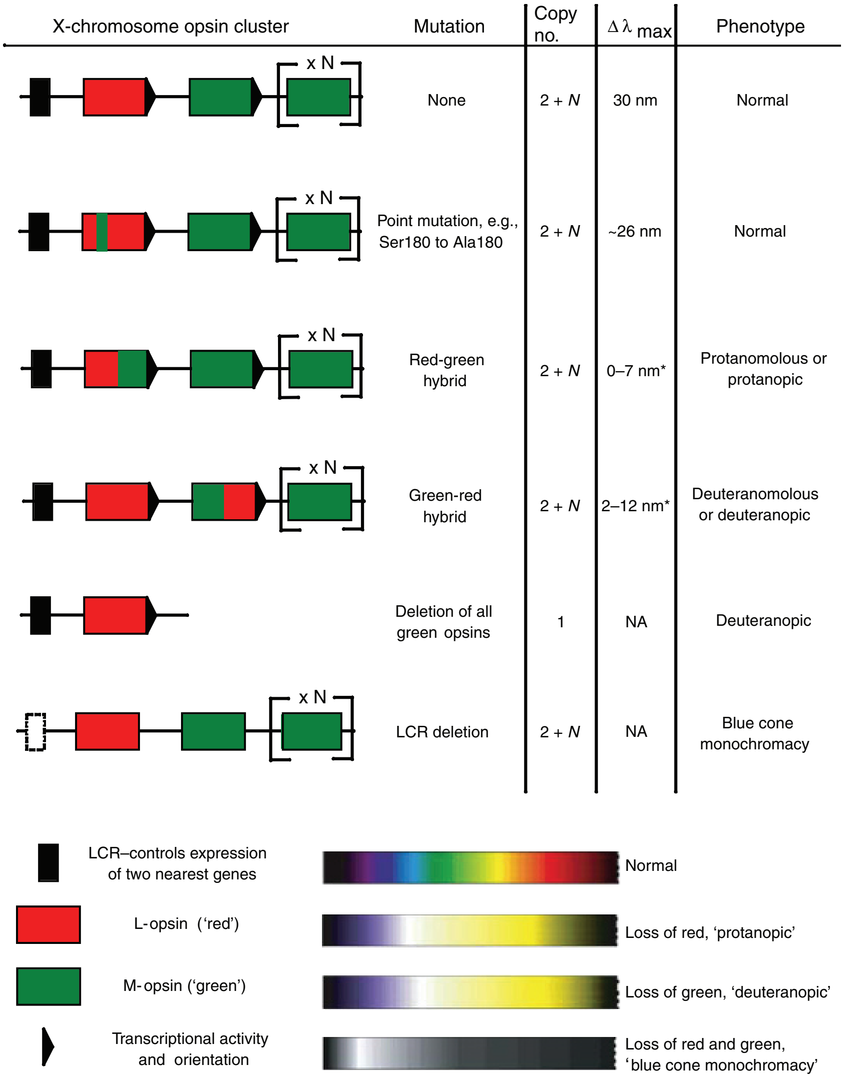
Figure 3 Functional annotation of copy-number variation gene content. Molecular function annotations for the genes affected by either CNVs that overlap SDs (red) or CNVs that do not overlap SDs (blue) were obtained using the PANTHER[25] classification system (see **Supplementary Note**). Negative log transformed *P* values, after Bonferroni adjustment, for the most significantly enriched functional groups are plotted as horizontal bars for each of the molecular functions indicated along the *y* axis. Some molecular functions are commonly enriched (for example, 'receptor'), but others are unique to either category of CNVs, such as 'signaling molecule' or 'serine protease'. Functional annotations with an asterisk show enrichment largely as a result of a few gene clusters (for example, clusters of keratin or protocadherin genes) that harbor CNVs. Note also that these categories are not independent as they share substantial overlap in some cases (for example, cytoskeletal protein, intermediate filament and miscellaneous function).

Insights from studies of chromosome evolution. Also important to the 'neutralist' perspective are results from analyses of chromosome evolution within mammals. Genome sequence assemblies[38–41] and high-resolution genetic maps of mammalian species have led to the reconstruction of ancestral mammalian karyotypes. Concomitant with such reconstructions are inferences regarding the numbers, types and locations of mutational events that must have occurred to generate the extant karyotypes[42–44]. The most relevant conclusion from these analyses is that the distribution of rearrangement breakpoints throughout evolution is nonrandom, highlighted by the existence of long stretches of genomic sequence being stable and short stretches, known as 'fragile sites', being prone to breakage[45,46]. Some regions are sufficiently unstable as to yield recurrent changes even among the closely related great apes, where a pericentromeric region of human chromosome 16 has been independently inverted in both chimpanzees and gorillas[47]. Furthermore, sites subject to rearrangement 'reuse' throughout mammalian chromosomal evolution are strongly correlated with segmental duplication content[48]. Interestingly, however, it appears that the relationship between SDs and evolutionarily 'fragile sites' is not necessarily causative[49] and may in fact be correlated through another mechanism. In any case, the strong correlations that exist among SDs, CNVs and chromosomal rearrangement hotspots provide substantial evidence that mutational mechanisms can explain much of the distribution of copy-number variation in the genome.

Selective pressure on CNVs: a boost in prior probability. An alternative to the neutralist perspective suggests that natural selection actively influences the distribution of CNVs in the genome. Intuitively, CNVs are highly likely to be subjected to selective pressure. Consider that large variants, in contrast with SNPs and small indels, often affect entire protein-coding genes and substantial amounts of flanking DNA. They may alter gene expression levels through a copy-number effect, duplicate or delete transcriptional regulatory elements for genes both within and near the variant, and also result in hybrid or truncated transcripts[27,28,50–52]. It is also clear that some CNVs contribute to human phenotypes, including many genetic disorders[28,53], color vision[54] (**Fig. 4**), glomerulonephritis[55], Parkinson's disease[56], Alzheimer's disease[57], Crohn's disease[58], hereditary pancreatitis[59], autism[60,61] and HIV-AIDS susceptibility[62]. The fact that large copy-number mutations have a higher likelihood, relative to SNPs and microindels, of altering genomic functionality and contributing to a phenotype increases the prior probability that selection has affected any particular mutation. Although difficult to quantify precisely, this increase in prior probability is likely to be at least several orders of magnitude. For example, of the 297 variant sites identified by Tuzun *et al.*, 6 or 7 (~2%) are thought to influence disease or disease susceptibility, a rate far greater than would be seen for any similar evaluation of SNPs[7].

Evidence for negative and positive selection on CNVs. The effects of negative selection are particularly visible for deletions, as might be expected for loss-of-function alleles: they tend to be at lower frequencies in human populations than other types of variants[6,63] and also tend to be biased away from genes[12]. Negative selection is also likely to contribute to the enrichment in gene-poor regions for CNVs not overlapping SDs (**Fig. 2b**), because such variants would have a lower likelihood of disrupting protein-coding sequence than mutations that arise in gene-rich regions. In fact, although it is possible that mutational mechanisms can explain the bias toward both gene-poor and gene-rich regions of the genome seen for CNVs that are not associated with segmental duplication (**Fig. 2b**, blue bars), such a pattern may instead reflect both negative and positive selective pressures (see below). Furthermore, regions of the genome known to be under intense purifying selection ('ultraconserved' elements[64]) are significantly depleted within CNVs[65]; deletion or amplification of these functionally critical regions probably results in deleterious consequences. Finally, the 'fragile-site' model of chromosome evolution[46] may be at least partially a consequence of purifying selection against breakpoints that disrupt blocks of coregulated genes (for example, *HOX* clusters[66]) or regulatory domains of genes encoding key developmental proteins (for example, *DACH*[67]); this is more of a 'stable-site' model than a 'fragile-site' model, but either mechanism would result in a nonrandom distribution of chromosomal breakpoints.

Adaptive selection is also likely to be a prominent influence on copy-number polymorphisms. One study found evidence for positive selection acting on a large (900-kb) inversion polymorphism, potentially as a consequence of increased fertility[68]. Functional biases in the genes that are associated with CNVs provide an indication of adaptive selection. Many genes affected by CNVs, particularly those variants that overlap


X-chromosome opsin cluster
x N
x N
x N
x N
x N
LCR–controls expression of two nearest genes
L-opsin ('red')
M-opsin ('green')
Transcriptional activity and orientation
Normal
Loss of red, 'protanopic'
Loss of green, 'deuteranopic'
Loss of red and green, 'blue cone monochromacy'

Mutation	Copy no.	Δλ max	Phenotype
None	2 + *N*	30 nm	Normal
Point mutation, e.g., Ser180 to Ala180	2 + *N*	~26 nm	Normal
Red-green hybrid	2 + *N*	0–7 nm*	Protanomolous or protanopic
Green-red hybrid	2 + *N*	2–12 nm*	Deuteranomolous or deuteranopic
Deletion of all green opsins	1	NA	Deuteranopic
LCR deletion	2 + *N*	NA	Blue cone monochromacy

Figure 4 Variation in the X-chromosome opsin locus sequence, structure and resulting phenotype. A locus control region (LCR; black rectangle) activates transcription of the two nearest, and only the two nearest, genes, as indicated by black arrowheads. Normal individuals (first row) have both red and green photopigment genes immediately downstream of the LCR, and a variable number of downstream green opsins. Total copy number (2 + *N*) is polymorphic among humans and typically ranges from 2 to 6; however, variation in copy number by itself does not dictate phenotype. The difference in maximal wavelength absorption ($\Delta\lambda_{max}$; third column) between the two expressed proteins is generally indicative of a person's ability to discriminate colors (fourth column): defects arise when $\Delta\lambda_{max}$ values are small, ranging from moderate (protanomolous and deuteranomolous) to severe (protanopic and deuteranopic) when $\Delta\lambda_{max}$ is near 0. The legend includes simulated images of how these individuals view the color spectrum (from ref. 54). A common SNP alters the red photopigment such that it responds to a wavelength closer to that of the green photopigment (second row). Hybrid opsin genes generated by unequal crossing-over commonly cause defects in color vision (rows three and four); note that a range of $\Delta\lambda_{max}$ values (asterisks in the third column) are observed depending on the crossover breakpoints. Some mutation events result in loss of all green opsins (fifth row). Deletion of the LCR, which is outside of the copy-number variable region, eliminates expression altogether (last row). Figure adapted with permission from figures in ref. 54 and with advice from Samir Deeb.

SDs, belong to categories of environmentally responsive functions such as sensory perception and immunity. These genes have long been considered, and in many cases shown, to be subject to rapid adaptive changes throughout mammalian evolution[39,40]. A variety of studies have also found signatures of positive selection at the level of amino acid replacements within recently duplicated gene families, including the morpheus[69], RanBP2 (ref. 70) and DUF1220 (ref. 71) families, as well as more global elevations of amino acid replacement rates among CNV-affected genes[26].

The associations between CNVs, SDs and genes may reflect a combination of positive and negative selective pressures. Consider the distinctions in gene density observed when CNVs are partitioned according to whether or not they overlap SDs (**Fig. 2b**). CNVs that overlap SDs are enriched for high-frequency events, and SDs are themselves a mixture of both high-frequency CNVs and fixed duplication events. Their lack of enrichment in gene-poor regions, in contrast to that seen for CNVs not overlapping SDs, may be due to a fixation bias between CNVs that arise in gene-poor versus gene-rich regions. Although, on the one hand, copy-number mutations are less likely to be deleterious (and removed by negative selection) in gene-poor regions, resulting in an enrichment in these areas, such events are also less likely to be given a boost in fixation probability by adaptive selection. Those CNVs that affect gene-rich regions without resulting in deleterious effects, on the other hand, would more likely be subjected to adaptive selection and therefore pushed to higher frequency and fixed more often. Although this argument must be tempered by the ability of SDs to induce recurrent mutation events, making them appear to be 'high frequency', such a scenario would explain why SDs and high-frequency CNVs would be enriched in gene-rich regions but not gene-poor regions. At the same time, other copy-number mutations would be enriched in both gene-rich regions (owing to the presence of very young but beneficial variants) and gene-poor regions (owing to variants that are not deleterious but subject to genetic drift).

Unifying the hypotheses. Of course, the reality is that the distribution of CNVs is likely to be a complex product of both mutational and selective effects. Several recent studies highlight this concept. First, one expects that the interplay between neutral and selective effects would be at least partially mediated by variant size. This dynamic has in fact been observed directly in a recent study of children affected by idiopathic mental retardation, where CNVs found in the unaffected parents are several times smaller than the putatively causative variants identified in the affected children[72]. Second, in another study of idiopathic mental retardation, several duplication-mediated regions of genomic instability were discovered in which large, recurrent deletions consistently result in a well defined and dramatic disorder[73]. One of these regions, located on chromosome 17q21.31, corresponded to the same region that is frequently inverted in European populations (see above) and is associated with positive selection and apparently increased fecundity in Icelandic populations[68]. Interestingly, in every case studied thus far, the germline rearrangement associated with disease (under negative selection) occurred in a parent who carried the inverted haplotype (under positive selection)[73–75]. These data indicate that the inversion structure or some other property of the inversion haplotype may act as premutation state for disease. Collectively, these studies imply that interactions between biased mutational effects and selective pressures are key influences on the evolution and population dynamics of human genomic structural variation (including both CNV and inversion polymorphisms).

Future directions

Although an impressive knowledge of the copy-number variation landscape has emerged, many unresolved questions and technological challenges remain. First, our current view of this variation is, for the most part, top-down and largely impressionistic. Our reliance on BAC-based arrays or commercial SNP platforms for detecting many of the known variants limits our ability to define the true extent of variation. This has several ramifications. First, many of the common smaller variants (<50 kb) have simply not yet been discovered. Second, lack of resolution in inferring variant breakpoints means that the annotations for many known variants include more DNA than is in reality affected. It will be important to determine which genomic functional elements are internal and external to a given variant, rather than knowing only that they are in the vicinity. The lack of breakpoint resolution also precludes many analyses, particularly evolutionary and population genetic, which need to distinguish between recurrent mutation events and legitimate allele-frequency differences. It is very difficult, if not impossible, to accurately delineate neutral and selective effects on particular loci without such information. Third, although current approaches have had some success in classifying CNVs in the range of 0, 1 or 2 copies, multicopy expansions remain more opaque. Distinguishing between 3, 4 or more copies will ultimately be necessary to characterize variants as to both their potential phenotypic effects and their population genetic and evolutionary histories. Fourth, most of the current technology has not been able to discover balanced rearrangement events, such as inversions, or sequences that are not represented within the human reference genome assembly. Thus, our understanding of genomic structural variation has largely been limited to regions of copy-number difference with respect to the reference genome assembly.

Ultimately, a more careful, sequence-based description of allelic states will be necessary. Consider the examples of gene amplifications that have been subject to adaptive selection pressure[69–71], in which a variety of smaller-scale nucleotide changes (for example, amino acid replacements) accompany copy-number mutations. The red-green opsin gene family also highlights this issue[54] (**Fig. 4**). Although mutation breakpoints and copy number differences are important to predicting the color-vision phenotype, allelic states that differ by single-nucleotide changes, gene order and status of nearby sequence elements are also critical (**Fig. 4**). Mutations at all scales contribute to the genotype, which is the unit that contributes (or fails to contribute) to a change in phenotype. Ultimately, it is the sequence that matters. Future technology development that cost-effectively and comprehensively captures both single-nucleotide and structural variation remains one of the most important goals of human genetics.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

The authors acknowledge support from National Human Genome Research Institute Interdisciplinary Training in Genomic Sciences grant T32 HG00035 and National Heart, Lung, and Blood Institute Program for Genomic Applications grant HL066682. E.E.E is an investigator of the Howard Hughes Medical Institute.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics

Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
2. Bhangale, T.R., Rieder, M.J., Livingston, R.J. & Nickerson, D.A. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum. Mol. Genet.* **14**, 59–69 (2005).
3. Bhangale, T.R., Stephens, M. & Nickerson, D.A. Automating resequencing-based detection of insertion-deletion polymorphisms. *Nat. Genet.* **38**, 1457–1462 (2006).
4. Mills, R.E. *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16**, 1182–1190 (2006).
5. McCarroll, S.A. *et al.* Common deletion polymorphisms in the human genome. *Nat. Genet.* **38**, 86–92 (2006).
6. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. & Pritchard, J.K. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* **38**, 75–81 (2006).
7. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
8. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
9. Sharp, A.J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
10. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
11. Wong, K.K. *et al.* A comprehensive analysis of common copy-number variations in the human genome. *Am. J. Hum. Genet.* **80**, 91–104 (2007).
12. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
13. Cooper, G.M. *et al.* Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res.* **14**, 539–548 (2004).
14. Feuk, L. *et al.* Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet* [online] **1**, e56 (2005).
15. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
16. Newman, T.L. *et al.* A genome-wide survey of structural variation between human and chimpanzee. *Genome Res.* **15**, 1344–1356 (2005).
17. Cheng, Z. *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**, 88–93 (2005).
18. She, X. *et al.* The structure and evolution of centromeric transition regions within the human genome. *Nature* **430**, 857–864 (2004).
19. Trask, B.J. *et al.* Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum. Mol. Genet.* **7**, 13–26 (1998).
20. Eichler, E.E. *et al.* Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. *Hum. Mol. Genet.* **5**, 899–912 (1996).
21. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
22. 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.* **16**, 576–583 (2006).
23. Zhang, L., Lu, H.H., Chung, W.Y., Yang, J. & Li, W.H. Patterns of segmental duplication in the human genome. *Mol. Biol. Evol.* **22**, 135–141 (2005).
24. Beiβbarth, T. & Speed, T.P. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20**, 1464–1465 (2004).
25. Thomas, P.D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13**, 2129–2141 (2003).
26. Nguyen, D.Q., Webber, C. & Ponting, C.P. Bias of selection on human copy-number variants. *PLoS Genet* [online] **2**, e20 (2006).
27. Sharp, A.J., Cheng, Z. & Eichler, E.E. Structural variation of the human genome. *Annu. Rev. Genomics Hum. Genet.* **7**, 407–442 (2006).
28. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
29. Wu, Q. & Maniatis, T. A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell* **97**, 779–790 (1999).
30. Noonan, J.P., Grimwood, J., Schmutz, J., Dickson, M. & Myers, R.M. Gene conversion and the evolution of protocadherin gene cluster diversity. *Genome Res.* **14**, 354–366 (2004).
31. Yu, F.H., Yarov-Yarovoy, V., Gutman, G.A. & Catterall, W.A. Overview of molecular relationships in the voltage-gated ion channel superfamily. *Pharmacol. Rev.* **57**, 387–395 (2005).
32. Ohno, S., Wolf, U. & Atkin, N.B. Evolution from fish to mammals by gene duplication. *Hereditas* **59**, 169–187 (1968).
33. Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
34. Lynch, M. & Conery, J.S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
35. Noonan, J.P. *et al.* Extensive linkage disequilibrium, a common 16.7-kilobase deletion, and evidence of balancing selection in the human protocadherin α cluster. *Am. J. Hum. Genet.* **72**, 621–635 (2003).
36. Nobrega, M.A., Zhu, Y., Plajzer-Frick, I., Afzal, V. & Rubin, E.M. Megabase deletions of gene deserts result in viable mice. *Nature* **431**, 988–993 (2004).
37. Small, K.S., Brudno, M., Hill, M.M. & Sidow, A. Extreme genomic variation in a natural population. *Proc. Natl. Acad. Sci. USA* **104**, 5698–5703 (2007).

38. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
39. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
40. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
41. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
42. Bourque, G., Pevzner, P.A. & Tesler, G. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res.* **14**, 507–516 (2004).
43. Murphy, W.J., Bourque, G., Tesler, G., Pevzner, P. & O'Brien, S.J. Reconstructing the genomic architecture of mammalian ancestors using multispecies comparative maps. *Hum. Genomics* **1**, 30–40 (2003).
44. Murphy, W.J. *et al.* Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309**, 613–617 (2005).
45. Peng, Q., Pevzner, P.A. & Tesler, G. The fragile breakage versus random breakage models of chromosome evolution. *PLoS Comput. Biol.* [online] **2**, e14 (2006).
46. Pevzner, P. & Tesler, G. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci. USA* **100**, 7672–7677 (2003).
47. Goidts, V. *et al.* Independent intrachromosomal recombination events underlie the pericentric inversions of chimpanzee and gorilla chromosomes homologous to human chromosome 16. *Genome Res.* **15**, 1232–1242 (2005).
48. Armengol, L., Pujana, M.A., Cheung, J., Scherer, S.W. & Estivill, X. Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Hum. Mol. Genet.* **12**, 2201–2208 (2003).
49. Bailey, J.A., Baertsch, R., Kent, W.J., Haussler, D. & Eichler, E.E. Hotspots of mammalian chromosomal evolution. *Genome Biol.* [online] **5**, R23 (2004).
50. Bailey, J.A. *et al.* Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *Am. J. Hum. Genet.* **70**, 83–100 (2002).
51. Paulding, C.A., Ruvolo, M. & Haber, D.A. The *Tre2* (*USP6*) oncogene is a hominoid-specific gene. *Proc. Natl. Acad. Sci. USA* **100**, 2507–2511 (2003).
52. Courseaux, A. & Nahon, J.L. Birth of two chimeric genes in the *Hominidae* lineage. *Science* **291**, 1293–1297 (2001).
53. Inoue, K. & Lupski, J.R. Molecular mechanisms for genomic disorders. *Annu. Rev. Genomics Hum. Genet.* **3**, 199–242 (2002).
54. Deeb, S.S. Genetics of variation in human color vision and the retinal cone mosaic. *Curr. Opin. Genet. Dev.* **16**, 301–307 (2006).
55. Aitman, T.J. *et al.* Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851–855 (2006).
56. Singleton, A.B. *et al.* α-Synuclein locus triplication causes Parkinson's disease. *Science* **302**, 841 (2003).
57. Rovelet-Lecrux, A. *et al.* APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat. Genet.* **38**, 24–26 (2006).
58. Fellermann, K. *et al.* A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am. J. Hum. Genet.* **79**, 439–448 (2006).
59. Le Marechal, C. *et al.* Hereditary pancreatitis caused by triplication of the trypsinogen locus. *Nat. Genet.* **38**, 1372–1374 (2006).
60. The Autism Genome Project Consortium. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat. Genet.* **39**, 319–328 (2007).
61. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
62. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–1440 (2005).
63. Locke, D.P. *et al.* Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79**, 275–290 (2006).
64. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
65. Derti, A., Roth, F.P., Church, G.M. & Wu, C.T. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat. Genet.* **38**, 1216–1220 (2006).
66. Spitz, F., Gonzalez, F. & Duboule, D. A global control region defines a chromosomal regulatory landscape containing the *HoxD* cluster. *Cell* **113**, 405–417 (2003).
67. Nobrega, M.A., Ovcharenko, I., Afzal, V. & Rubin, E.M. Scanning human gene deserts for long-range enhancers. *Science* **302**, 413 (2003).
68. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).
69. Johnson, M.E. *et al.* Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519 (2001).
70. Ciccarelli, F.D. *et al.* Complex genomic rearrangements lead to novel primate gene function. *Genome Res.* **15**, 343–351 (2005).
71. Popesco, M.C. *et al.* Human lineage–specific amplification, selection, and neuronal expression of DUF1220 domains. *Science* **313**, 1304–1307 (2006).
72. de Vries, B.B. *et al.* Diagnostic genome profiling in mental retardation. *Am. J. Hum. Genet.* **77**, 606–616 (2005).
73. Sharp, A.J. *et al.* Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat. Genet.* **38**, 1038–1042 (2006).
74. Shaw-Smith, C. *et al.* Microdeletion encompassing *MAPT* at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nat. Genet.* **38**, 1032–1037 (2006).
75. Koolen, D.A. *et al.* A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat. Genet.* **38**, 999–1001 (2006).