Primate segmental duplications: crucibles of evolution, diversity and disease

Jeffrey A. Bailey and Evan E. Eichler‡*

Abstract | Compared with other mammals, the genomes of humans and other primates show an enrichment of large, interspersed segmental duplications (SDs) with high levels of sequence identity. Recent evidence has begun to shed light on the origin of primate SDs, pointing to a complex interplay of mechanisms and indicating that distinct waves of duplication took place during primate evolution. There is also evidence for a strong association between duplication, genomic instability and large-scale chromosomal rearrangements. Exciting new findings suggest that SDs have not only created novel primate gene families, but might have also influenced current human genic and phenotypic variation on a previously unappreciated scale. A growing number of examples link natural human genetic variation of these regions to susceptibility to common disease.

Low copy repeat
A term with variable meaning that is sometimes used synonymously with segmental duplication. It can denote a group of juxtaposed duplicons (duplication block), individual segmental duplication events or individual duplicons. The term emphasizes the low copy number of repeats (2–50 copies) relative to most transposable elements.

**Department of Pathology, Case Western University School of Medicine and University Hospitals of Cleveland, Ohio 44106, USA.*
‡Department of Genome Sciences, University of Washington School of Medicine and the Howard Hughes Medical Institute, 1705 NE Pacific Street, Seattle, Washington 98195, USA.
Correspondence to E.E.E.
e-mail: eee@gs.washington.edu
doi:10.1038/nrg1895
Published online 13 June 2006

Segmental duplications (SDs), also known as low copy repeats (LCRs), are continuous portions of DNA that map to two or more genomic locations. They can contain any standard constituent of genomic DNA, including genic sequence and common repeats, and can either be organized in tandem or map to interspersed locations. Since the term SD was originally coined — to distinguish the underlying events from whole-genome duplications — it has become clear that SDs are a common architectural feature of many genomes[1–7].

Human SDs were initially dismissed as rare peculiarities of certain genomic regions near centromeres and telomeres. However, interest in the biology of SDs has been ignited in recent years, as the initial draft sequencing of the human genome revealed an unexpectedly large number of SDs compared with previously sequenced species such as the fly and worm[1,2]. Here, we discuss subsequent findings that have been enabled by the availability of sequence data from an expanding range of species, which has provided important insights into the distribution, origins and mechanisms of evolution of primate SDs. Comparisons between and within species have also indicated that SDs have had important roles in primate evolution, in creating new genes and in shaping the human genetic variation that is thought to contribute significantly to disease susceptibility.

SD content in primates and other mammals

The recent generation of draft genome-sequence assemblies has made it possible to computationally detect and quantify the SD content of various genomes (TABLE 1). Most methods have used genome-wide pairwise comparisons, coupled with algorithms that capture large blocks of high sequence identity. Such approaches are designed to discriminate duplications of genomic sequence from uncharacterized transposons and retroelements that are also present at more than one copy. Most analyses are therefore limited to duplications that are ≥1–5 kb in size and contain ≥90% sequence identity[1,8,9]. If minimal gene conversion and a neutral molecular clock for evolution are assumed, primate SDs with 90% identity correspond to duplication events that occurred approximately 35–40 million years ago, roughly correlating with the divergence of the New and Old World monkeys. Although they can be nearly a megabase in size[10], most primate SDs that have been detected using common methods are <300 kb in length[1,11].

Initial whole-genome estimates of SD content in humans were complicated by sequence misassembly and collapse[1,8,12]. Duplications with high levels of sequence identity (>98%) were often confused with potential allelic overlap and compressed into a single locus during assembly. The first robust estimate of human SD content — 5.2% for SDs of ≥1 kb and ≥90% identity — was

Gene conversion
Used here in the general sense as the transfer of genetic information from one sequence to another based on homology — the strict definition is non-reciprocal meiotic exchange resulting in products with a 3:1 ratio of alleles.

Molecular clock
A molecular clock is said to exist when the rate of nucleotide change is approximately constant over evolutionary time; this rate can then be used to estimate the age of duplication or speciation events.

Whole-genome shotgun sequencing
A sequencing strategy that involves random fragmentation of the entire genome. The fragments are sequenced, and highly refined algorithms are used to reassemble the original genomic DNA sequence.

Fluorescence *in situ* hybridization
A technique in which a fluorescently labelled DNA probe is used to detect a particular chromosome region or gene by fluorescence microscopy. The intensity of the signal can be used to detect copy-number differences between the labelled chromosomal regions.

Hominoid
A primate superfamily that includes the great apes (orangutans, gorillas, chimpanzees and bonobos) and humans (hominids).

achieved using a measure of duplication obtained independently from the sequence assembly. This approach tested all reference sequence for overrepresentation within a set of whole-genome shotgun-sequence reads[13]. The resulting estimate, which was computed 2 years before the completion of the human-genome assembly[14], was remarkably consistent with others based on fluorescence *in situ* hybridization (FISH). It was also consistent with two recent calculations based on the nearly finished genome assembly (build 35), which estimated a content of 5.4% at ≥1 kb and ≥90% identity (analysis available at the Human Segmental Duplication Database) and 5.0% at ≥5 kb and ≥90% identity (analysis available at the Human Genome Segmental Duplication Database). As more than 50% of the remaining euchromatic gaps contain uncharacterized duplications[14,15], estimates of the content of euchromatin that is derived from recent duplication continue to rise incrementally, with an upper limit estimated at ~6.0%.

Although our understanding of human SDs is by far the most complete, information from non-human primate genomes is beginning to emerge. Analysis of the chimpanzee draft genome suggests a slightly higher level of recent duplication compared with the human genome, and also indicates that as many as a third of duplications with >94% identity differ in copy number or content between the two species[6] (analysis available at the Chimpanzee Segmental Duplication Database). Preliminary FISH analysis of randomly selected genomic clones indicates that the human, chimpanzee (a hominoid) and macaque (an Old World monkey) have significantly greater levels of SD (5–7%) than the marmoset (2%; a representative New World monkey)[7]. These preliminary results await validation in a more diverse array of primate species. However, the apparent increase in SD content among hominoids, and to a lesser extent Old World monkeys, might be accounted for by differences in effective ancestral population size, increased generation times among hominoids, chromosome structure and/or, most intriguingly — as we discuss later — adaptive evolution[7,9,13].

Compared with the average for human and chimpanzee, other sequenced mammalian genomes (rat, mouse and dog) show slightly less overall SD content (2–4%) (REFS 3–5; E.E.E., unpublished observations). Although these lower estimates are probably, in part, due to differences in the quality of the assemblies, the distribution of recent interspersed duplications seems to be constrained to fewer locations in these genomes. Other mammalian genomes show a predominance of tandem duplications when compared with humans and chimpanzees. Overall, SDs from all sequenced mammalian genomes are of similar size compared with human SDs, but are larger than other eukaryotes such as the worm and fly[2,7]. This difference probably reflects evolutionary constraints imposed by the smaller genome sizes of non-mammalian species[2].

A non-random, interspersed distribution

A non-random distribution of human SDs is apparent at the chromosome level. Chromosome 3 has the lowest proportion of duplicated sequence (1.7%), whereas chromosome 22 and the non-recombining Y chromosome have the greatest proportion (11.9% and 50.4%, respectively)[11]. Overall, chromosomes 7, 9, 10, 15, 16, 17, 22, X and Y are significantly enriched for duplication[9,11]. However, non-random distribution is most obvious within specific regions of chromosomes, termed duplication hubs or acceptor regions, that have been the target of multiple, independent duplication events and seem to have accumulated over time[16–21]. A comparison of the chimpanzee and human shows that new lineage-specific SDs preferentially map near shared ancestral duplications. This phenomenon, termed 'duplication shadowing'[6,22], means that unique regions flanking duplications are ~10 times as likely to become duplicated as other randomly distributed regions.

Other species also show clustering of duplications, but what truly distinguishes human and chimpanzee genomes from other sequenced species is the abundance of interchromosomal and interspersed intrachromosomal duplications. For instance, 48% of human alignments are interchromosomal, compared with 13% in mice and 15% in rats[4,5]. Similarly, tandem duplications with less than 1 Mb between them account for only ~45% of all human SDs but constitute 70–90% of duplications in mice, chickens and rats[7].

Based on their distribution and organization, SDs that are found in the genomes of humans and great apes (orangutans, gorillas, chimpanzees and bonobos) can be classified into three categories — pericentromeric, subtelomeric or interstitial regions of duplication — each of which is characterized by different frequencies and types of SD (FIG. 1).

Table 1 | **SD content of sequenced animal genomes**

	Caenorhabditis elegans	*Drosophila melanogaster*	Human	Mouse	Rat	Chicken	Chimpanzee*
SDs of >1 kb	4.3%	1.2%	5.2%	2.7%	1.6%	2.7%	N.D.
SDs of >10 kb	0.7%	0.1%	4.5%	2.2%	1.5%	0.3%	N.D.
SDs of >20 kb	N.D.	N.D.	4.0%	1.7%	0.9%	0.0%	~4.8%
Total SD content	5.0%	1.3%	13.7%	6.6%	4.0%	3.0%	~4.8%
Genome size	97	123	2,866	2,506	2,566	1,040	2,866

Data taken from REFS 2,7 for pairwise segmental duplications (SDs) with >90% identity. *Given the fragmented nature of SDs in the draft chimpanzee genome, the duplication content can only be estimated indirectly on the basis of human duplication content, adjusting for detected differences in SD compared with chimpanzee whole-genome shotgun sequencing[6]. DNA not assigned to a chromosome was not included in these calculations. Consequently, in other genomes the estimate of recent duplication might rise as the quality of the sequence assembly improves. N.D., not determined.

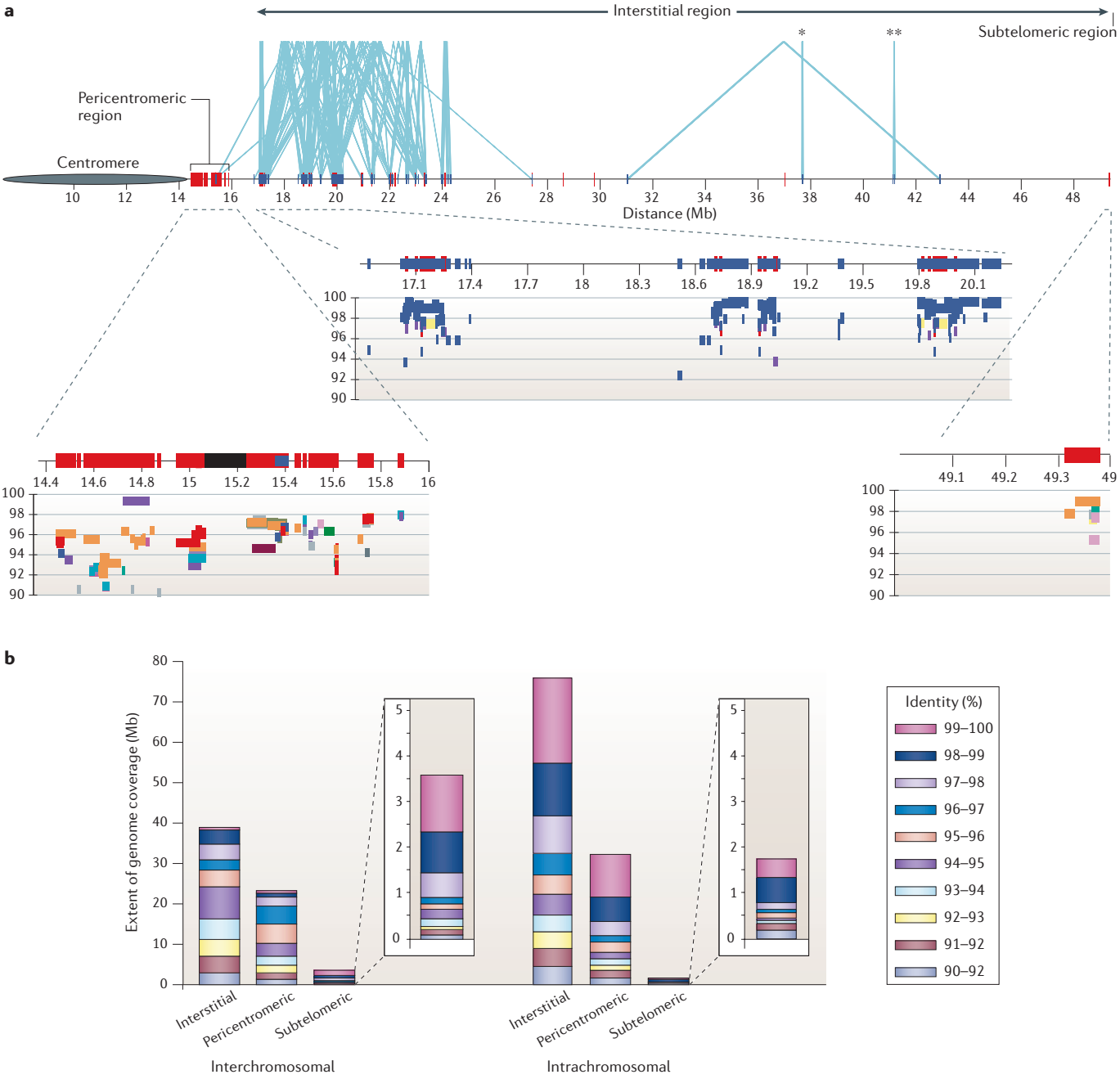

Figure 1 | **The distribution of segmental duplications (SDs) in the human genome. a** | The q-arm of chromosome 22 is used to show the distribution of the three classes of SD: pericentromeric, interstitial and subtelomeric. The overview (top panel) shows the position of SDs that are ≥10 kb in length and ≥90% identity. Interchromosomal and intrachromosomal pairwise SDs are shown in red and blue, respectively, with light blue lines joining homologous SDs. The expanded views of the main duplicated regions show their SD structures in terms of % identity (vertical scales) and the chromosomal location of the other pairwise copy (blue represents chromosome 22, other colours represent other individual chromosomes). The pericentromeric region consists of many juxtaposed duplicons that originate from diverse ancestral regions. Secondary duplications of larger segments consisting of multiple duplicons (duplication blocks) are distributed among non-homologous pericentromeric regions. The interstitial region shows examples of interspersed (*) and tandem (**) intrachromosomal duplications. The expanded view for this region delineates an interspersed cluster that spans 3.5 Mb and contains three large duplication blocks in which the average sequence identity is 98–99%. The subtelomeric region contains approximately 100 kb of interchromosomal SD sequence, which is shared with up to three other non-homologous subtelomeric regions. **b** | The distribution of interchromosomal and intrachromosomal duplications within the pericentromeric, subtelomeric and interstitial regions are shown in terms of % identity of alignments. The *y*-axis represents the total number of non-redundant base pairs within the genome (build 35) that consist of each type of SD. The distribution within each category was calculated as the proportion of pairwise alignments at each % identity.

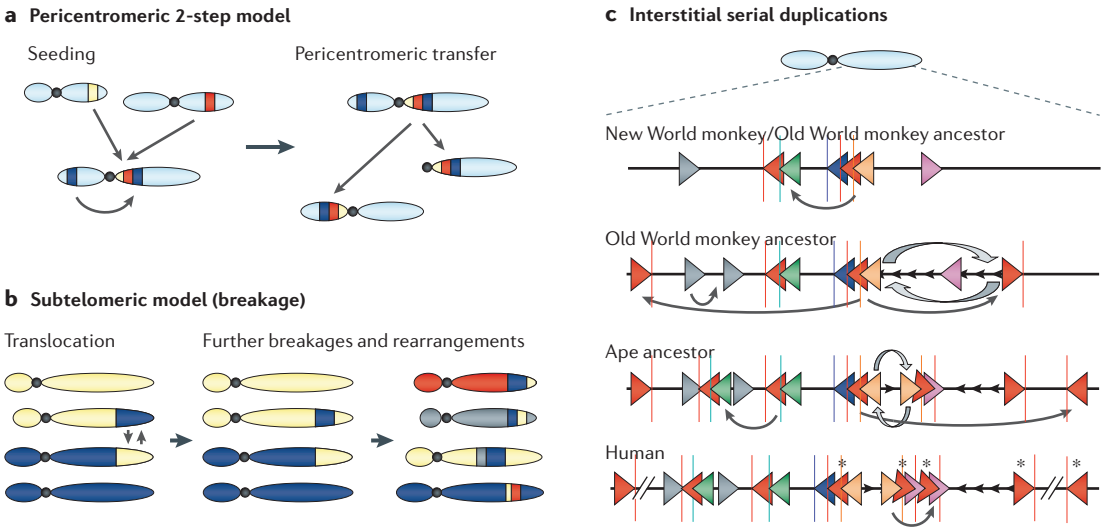

Figure 2 | **Models of segmental duplication (SD) formation. a** | Pericentromeric regions are thought to arise in a two-step process[29]. In the initial seeding event, SDs from different regions of the genome (blue, yellow and red bands) form juxtaposed duplicons. Blocks of multiple duplicons are subsequently transferred by duplication to non-homologous pericentromeric regions. **b** | SDs in subtelomeric regions are though to arise in a process that involves double-stranded breakage and repair, leading to translocations of subtelomeric regions[31]. With multiple rounds of breakage and repair, a mosaic pattern of juxtaposed duplicons is created, with SD copies maintaining the same orientation (different colours indicate segments from different chromosomes). **c** | Interstitial SDs result from multiple rounds of serial duplication, in which flanking sequence is often duplicated in subsequent events, leading to a complex pattern of duplication blocks. The example shown is a simplified version of the proposed sequence of duplication and rearrangement events that are thought to have occurred in the *BRCA1* region, which is based on comparative analysis of the human 17q11 region with other primate genomes[35]. Coloured arrowheads represent duplicons. Thin black arrows represent duplication events, whereas thick arrows represent inversions. Asterisks (*) denote the position of transcribed KIAA0563-related genes that have expanded through these duplication events.

Effective ancestral population size
Approximately the number of breeding individuals that produce offspring that live to reproductive age. It influences the rate of loss of genetic variation, the efficiency of natural selection and the accumulation of beneficial and deleterious mutations. It is frequently much smaller than the number of individuals in a population.

Pericentromeric region
The sequence adjacent to the centromere that is often defined as the first chromosomal sub-band or by defined physical distances of 2–5 Mb from the higher-order α-satellite arrays that comprise the centromere.

Interstitial region
An arbitrary name given to the euchromatic sequence within a chromosome arm that is bounded by the pericentromeric and subtelomeric regions.

Duplicon
A duplication, or portion thereof, that is traceable to an ancestral or donor location; a secondary duplication event can be composed of multiple duplicons. Also sometimes referred to as a low copy repeat.

SDs in pericentromeric regions. Pericentromeric regions are enriched for SDs, particularly within long tracts that are made up of many different ancestral duplication segments (termed duplicons). These regions also vary in duplication content from no SDs (for example, 16q11) to over 6 Mb of SDs (for example, 9q11) (REFS 11,23). Overall, 29 out of 43 pericentromeric regions (including those of the Y chromosome) have conspicuous SD tracts, totalling 47.6 Mb (REFS 11,24), which accounts for a third of all duplicated human genome sequence, and there is a 6:1 ratio of interchromosomal to intrachromosomal duplication in pericentromeric regions[11]. This bias towards interchromosomal SDs diminishes in a gradient-like fashion, falling to the genome average at 5 Mb from the centromere.

An interesting aspect of pericentromeric regions is that at least 30% of all sequence within them can be traced to duplicons originating from other chromosomes[11]. After initial rounds of duplication, larger segments are thought to be secondarily duplicated *en bloc* to other pericentromeric regions[25–28], a mode of pericentromeric duplication that has been referred to as the two-step model[29] (FIG. 2a). A detailed investigation of 700 kb of pericentromeric region 2p11 detected 14 duplication-seeding events that were accepted over 10–20 million years ago, with no subsequent seeding, although interchromosomal *en bloc* duplications to other pericentromeric regions have continued[17]. This punctuated pattern seems to apply to other pericentromeric regions: when all interchromosomal duplications in these regions are analysed, there is an abundance of alignments of 94–97% identity and a relative dearth of duplications with >99% identity[11] (FIG. 1b). Why these pericentromeric regions were permissive to interchromosomal duplications at specific time points during primate evolution remains to be investigated.

SDs in subtelomeric regions. Subtelomeric regions are similar to pericentromeric regions in that they are enriched in interchromosomal SDs, which are present in 30 out of 42 subtelomeric regions[30,31]. However, the total amount of subtelomeric duplication (2.6 Mb) is an order of magnitude less than that in pericentromeric regions. Typical subtelomeric SD regions range from 50 to 100 kb in length and, similar to those in pericentromeric regions, are organized as complex mosaics of duplications[31,32]. However, unlike pericentromeric regions, in which most SD sequence is derived from other regions, there is relatively little evidence for this with regard to subtelomeric SDs. Instead, their origin seems to involve duplication and exchange between subtelomeric regions. Subtelomeric duplicons also seem to have maintained the same relative orientation between non-homologous chromosomes to a much greater extent than pericentromeric regions[11,31]. Combined with detailed analysis of their breakpoints, a model of serial translocation between non-homologous chromosomes has been proposed to account for these features[31] (FIG. 2b). Finally,

recent interchromosomal exchanges (>99% identity) have occurred more frequently in subtelomeric than pericentromeric regions (FIG. 1b).

SDs in interstitial regions. Interstitial duplications are distributed within euchromatin between the pericentromeric and subtelomeric regions and account for the bulk of intrachromosomal duplications. In most sequenced species, clusters of tandem duplications predominate. By contrast, the human genome shows both a relative and an absolute increase in the amount of interspersed duplication in these regions. Furthermore, the average distance between the pairwise copies of intrachromosomal duplications is higher in the human genome, at 3 Mb compared with 57 kb in the mouse and 9 kb in the rat[4,5].

Similar to pericentromeric regions, particular chromosomes show significant enrichment for interstitial duplications, again suggesting a non-random distribution[9,11]. Phylogenetic analyses indicate that many interstitial regions of interspersed intrachromosomal duplication have emerged during evolution of the great ape through a complex series of events (FIG. 2c). These expansions are frequently associated with rearrangements such as inversions and microrearrangements[19,20,33–35], and have restructured as much as 10% of the euchromatin of some chromosomes (for example, chromosomes 2, 15 and 17). As we discuss later, they have also resulted in the evolution of novel primate gene families. Compared with interchromosomal duplication, interstitial duplications account for the vast majority of the largest and highest-identity human SDs[12] (FIG. 1b).

Mechanisms of SD origin and propagation

Tandem duplications are thought to occur mainly through non-allelic homologous recombination (NAHR) and replication error[36,37]. However, neither of these mechanisms accounts for the interspersed configuration of most human and chimpanzee SDs, which has prompted a search for sequence features that might provide clues to the prevalence of interspersed SDs in particular genomic regions.

Regional variations in GC content, repeat density and recombination rate explain only 4% of the variation in SD content[9]. However, detailed analyses of sequences at the boundaries of pairwise SD alignments have revealed some interesting patterns. *Alu* repetitive sequences are significantly enriched at these boundaries (representing 24% of the sequence at these regions compared with 10% elsewhere) and are restricted to younger subfamilies (*Alu*Y and *Alu*S) that have emerged recently during primate evolution[38]. This enrichment has been confirmed[39,40] and generally applies to pericentromeric and interstitial SDs, but not subtelomeric SDs[31]. *Alu* enrichment has been shown at SD breakpoints for a number of well-studied, interstitial interspersed clusters[21,33,35,41].

It should be emphasized that *Alu* elements are not found at all SD breakpoints and that, while 'chimeric' *Alu* elements consistent with NAHR having taken place are enriched at breakpoints, they do not predominate. This suggests that other non-homology-mediated mechanisms drive duplication events[38]. For example, *Alu* sequences might be preferential sites of double-strand breakage, with interspersed SDs arising owing to inaccurate repair. In this respect, it is interesting that satellite repeats and sites with increased DNA flexibility and low helix stability (markers of fragility) are enriched at duplication breakpoints[38,40].

Although global analyses support an involvement of both homologous and non-homologous processes in the origin and propagation of SDs, they do not completely explain their non-random distribution in the human genome. The preponderance of SDs within pericentromeric and subtelomeric regions indicates that chromosome structure is an important consideration in SD expansion and dispersal. In subtelomeric regions, non-reciprocal translocations are common[30,31] and, as explained above, it has been proposed that subtelomeric SDs arise through a mechanism of serial unequal subtelomeric translocations, resulting in the segmental transfer of blocks of sequence between non-homologous chromosomes (FIG. 2b). Further insight into this mechanism has come from a detailed analysis of the breakpoints of 41 subtelomeric duplicons, which found that 92% (49 out of 53) of the breakpoints were consistent with non-homologous end-joining (NHEJ)[31], whereas only 8% (4 out of 53) fit with a homology-based model such as *Alu–Alu* mediated NAHR. This indicates that SDs between non-homologous chromosomes arise primarily as a result of double-strand breaks (DSBs) that are subsequently repaired by NHEJ. However, this does not preclude the occurrence of homology-based processes between SDs, as evidenced by multiple blocks of differing sequence identity within particular interchromosomal duplications[31].

Analyses of pericentromeric regions support an association of SDs with *Alu*s and other repeat sequences that are pericentromere specific. For example, in one analysis of 700 kb of highly duplicated sequence on chromosome 2p11, both donor and acceptor duplications were enriched for *Alu* repeats (15 out of 38 donor and 16 out of 38 acceptor sites contained *Alu* repeats at the junctions). However, other conspicuous repeat sequences that are similar to immunoglobulin heavy-chain class-switch regions have been identified at the boundaries of transposed duplicated sequences, including CAGGG, CAAAAG, TAR and REP522 repeats[17,25,42]. One possibility is that *Alu* repeats have an important role in defining donor sequences, whereas potentially recombinogenic sequences promote *Alu* accumulation at specific locations.

The fact that the sizes and levels of identity of duplications decline in a gradient with distance from the centromere also implicates other centromeric repeats, such as α-satellites, in exchanges between non-homologous chromosomes that involve larger regions of DNA[10,11]. For example, the largest pericentromeric duplication at 22q11, which also has the highest level of identity, also maps adjacent to the centromere on chromosome 14 (the donor), and these chromosomes share very similar higher-order α-satellite sequences[10]. In this example, pericentromeric duplications might arise by a mechanism similar to that in subtelomeric regions, whereby a DSB in the pericentromeric region leads to a whole-arm translocation unless homology-mediated repair

Microrearrangements
Rearrangements that are less than a megabase in size.

Non-allelic homologous recombination
Homologous recombination between paralogous sequence (for example, segmental duplication and repetitive sequence); a major mechanism of recurrent rearrangements, also known as unequal crossing-over.

Immunoglobulin heavy-chain class-switch recombination
Non-homologous recombination that occurs during the development of B lymphocytes to produce difference classes (isotypes) of heavy-chain molecule.

α-Satellite
A class of ~170 bp repetitive sequences that are found at the centromeres of most primates. They are present in tandem arrays that constitute megabases of sequence.

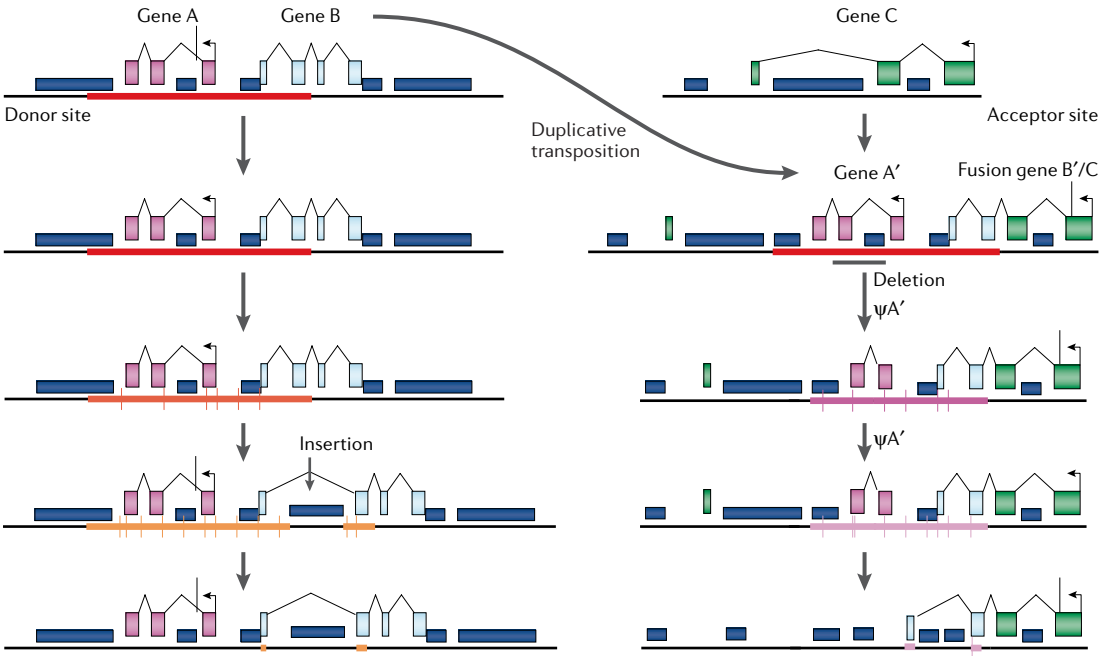

Figure 3 | **Mechanisms of segmental duplication (SD) divergence.** The formation of an SD during a duplicative transposition event. A chromosomal segment from a region that contains two genes, A and B, is translocated to a second region that contains a third gene, C. (Exons are shown above the coding sequence. The thick block in the sequence indicates the region that is duplicated; this is shown as changing colour over time to indicate sequence divergence. Blue boxes indicate repetitive elements (for example, Alus and LINEs), in which insertions occur over time.) SDs can duplicate whole genes, leading to entire gene copies (Gene A′), or create fusion genes (Gene B′/C). After duplication, both copies of an SD are initially identical. Over time, the SD copies are subjected to the same random forces of nucleotide substitution that apply throughout the genome to produce allelic variation and interspecies sequence divergence. Nucleotide substitutions (thin vertical lines) occur at a rate of approximately 1.5% change every 10 million years per copy (3% divergence between copies). When a substitution occurs, it represents both a rare single nucleotide variant between alleles and a paralogous sequence variant between SD copies. SDs also undergo differential deletion and insertion events. Combined, the forces of neutral substitution and insertion/deletion cause SD copies to diverge from each other. These mutational forces also alter duplicated genes, and usually result in nonsense and frameshift mutations, which lead to loss of function and the generation of unprocessed pseudogenes (ψA′). Purifying selection for functioning gene duplications maintains similarity between the duplicated exons (gene B and fusion gene B′/C). Only rarely do mutations occur that are innovative. Eventually, over tens to hundreds of millions of years, the copies will diverge to such an extent that their shared ancestry will be no longer recognizable except at regions where purifying selection has maintained them (shared B and B′ exons).

or rescue (a double-crossover event) occurs within the centromeric sequence.

In summary, the available data indicate that several mechanisms have been responsible for the origin and propagation of segmental duplications. Diverse sets of data support the involvement of specific common repeat classes — particularly *Alu* repeats — in this process. However, the mechanism by which blocks of >100 kb of DNA become integrated into new genomic locations over short evolutionary periods requires further investigation. Cell-culture experiments and studies of interindividual genetic variation that capture such novel insertions are needed to provide future progress[36,43].

SDs undergo homology-driven mutation

Once they have been formed, SDs are subject to the forces of evolution that affect all genomic sequences, including base-pair substitutions, insertions, deletions and retrotransposition (FIG. 3). Unlike unique regions, however, highly identical SDs also mutate by homology-driven processes (FIG. 4). Such processes contribute in several ways to structural differences in the architecture of SD regions, both within and between primate species.

First, homology between SDs can initiate NAHR, which occurs through the alignment of highly similar SDs followed by paralogous recombination. The type of rearrangement that occurs as a result of this (which can be duplication, deletion, inversion or translocation) depends on the location and orientation of the SDs. Such recombination in meiosis is well recognized as the major cause of genomic disorders, particularly recurrent ones[44] (for a recent review see REF. 45). Duplication-mediated rearrangement typically occurs between SDs that have alignments of >95% identity and >10 kb in length, with the largest and most identical SDs showing the highest rates of NAHR[13,46]. Within these regions of high sequence identity, rearrangement breakpoints are further restricted to intervals associated with particular repeats or AT-rich sequences[47–49]. Given the propensity of such regions to rearrange recurrently, it is perhaps not

Genomic disorder
A disease that results from the gain, loss or alteration of dosage-sensitive gene(s) as a result of genomic rearrangement (such as duplication, deletion and inversion).

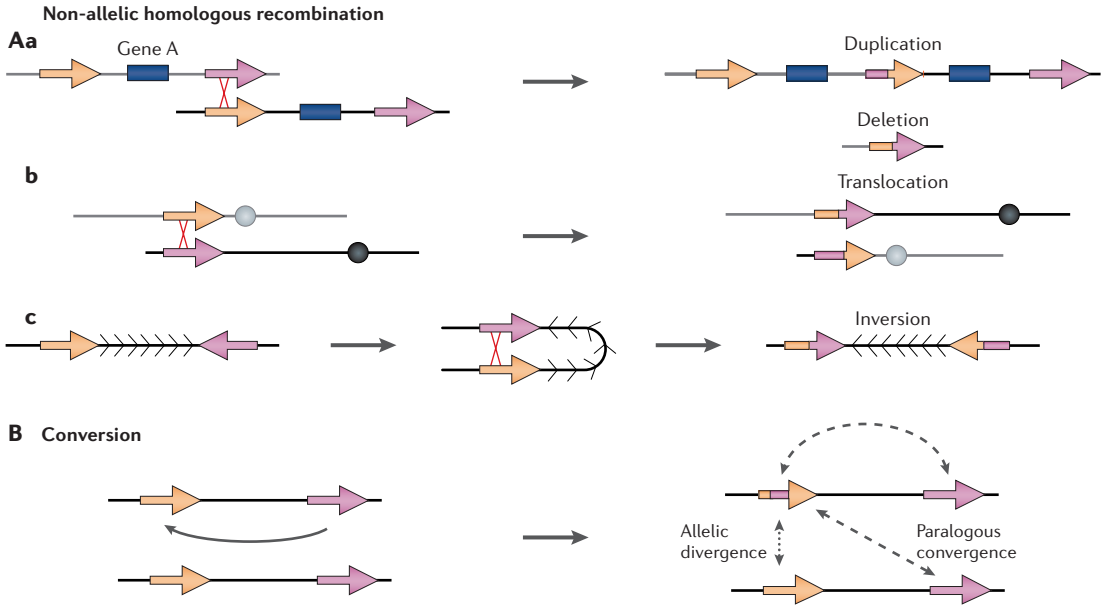

Figure 4 | **Non-random mechanisms of segmental duplication (SD) evolution.** Two major homology-driven forces affect the occurrence and evolution of SDs. **A** | Non-allelic homologous recombination (NAHR) between highly identical SDs causes further rearrangements depending on the location and orientation of the SD copies involved. Tandem duplications and intervening deletions can occur as a result of NAHR between adjacent duplicated sequences (**Aa**). Alternatively, translocations can result from exchange between SDs on non-homologous chromosomes (**Ab**). In both cases, the copies created are in the same orientation as the original SD. By contrast, inversions can occur as a consequence of recombination between inverted intrachromosomal duplications (**Ac**). **B** | Gene conversion between SDs occurs as a result of the transfer of sequence information between copies. Such events can increase allelic diversity at the converted copy and cause homogenization of the SD copies (in a process known as concerted evolution).

surprising that SDs are a source of structural variation among humans[32,50–54], as discussed later in more detail.

A second scenario in which homology-driven processes can lead to structural alterations is through the non-reciprocal transfer of sequence from one SD copy to another, which we refer to as gene conversion[55]. Evidence for gene conversion has been recently reported between several SDs that are frequently associated with genomic disorders[20,56–59]. This provides a link between meiotic NAHR — which underlies the rearrangements in these disorders — and conversion. This link is also apparent from pooled-sperm studies[60,61] that have allowed the individual detection of recombination and conversion events at the sub-kb level of resolution. Such investigations have shown that non-recombinant meiotic conversion rates equal or exceed the rate of meiotic recombination at several recombination hot spots.

Recently, novel computational algorithms have further facilitated the detection of gene conversion within SD sequence. A computational analysis of the SDs associated with velocardiofacial/DiGeorge syndrome, which is caused by NAHR between three highly identical (99%) duplication hubs on chromosome 22q11, found evidence for transfer of genetic information between the hubs[56]. However, this study did not exclude the possibility that some of the analysed sequence might have originated from an unfinished sequence gap within the middle duplication hub[14,15] and, therefore, sequences might be assigned incorrectly to the wrong hub as alleles[14,15] — emphasizing that care must be taken to avoid such confounding factors.

A study by Jackson and colleagues compared patterns of sequence variation among 10 major SD families that represent some of the human SDs with the highest identity. This provided evidence of widespread gene conversion (many transfers of <1 kb of sequence), which the authors estimate might have affected 20% of the sequence in these duplications[62]. Altogether, the data indicate that conversion is ubiquitous in SDs with the highest identity, and that both NAHR and gene conversion can alter the structure and sequence composition of a fraction of these regions over short periods of time.

Although conversion ultimately homogenizes SDs, prior to their fixation within a population it might also elevate allelic diversity at the converted locus; in essence, the donor copy acts as a reservoir of sequence variation. Initial human-genome analyses proposed that the observed twofold increase within SDs, compared with unique regions, of single nucleotide polymorphisms (SNPs) that had been catalogued in databases such as dbSNP was due not to the presence of different alleles, but to 'contamination' by paralogous sequence variants (PSVs)[8,13,63]. However, quantitative genotyping of a subset of single nucleotide variants within SDs (in a set of normal individuals and complete hydatidiform moles) allowed true SNPs and PSVs to be distinguished[50]. Only 23%, rather than the expected 50%, were consistent with PSVs. Interestingly, 12% (8 out of 79) of these variants fell into neither category; instead, there was evidence of the same SNP occurring within 2 or more copies of an SD — usually a signature of gene conversion[50].

Recombination hot spots
Regions of sequence that undergo increased meiotic rates of recombination.

Single nucleotide polymorphism
(SNP). A single nucleotide difference between orthologous sites. A strict definition requires a population frequency of at least 1% for the rare allele. SNPs represents 90% of the genetic variation within the human population in terms of numbers of variants.

Paralogous sequence variant
A single nucleotide difference between copies of a segmental duplication (or any paralogous sequence), which can be variant or fixed in a population.

Complete hydatidiform mole
Placental mass with uncontrolled growth due to the fertilization of an enucleated egg with one (90%) or two (10%) sperm — useful for genetic studies as those derived from single sperm represent a haploid genome.

So, although PSVs were misidentified as SNPs, there also seems to be an increased frequency of SNPs within SDs. This increased SNP density in SDs is also supported in detailed variation studies of SDs and repeat regions, which show high levels of nucleotide diversity[57,58,64].

Gene conversion in SDs blurs the distinction between allelic and paralogous sequence variation. If conversion operates over longer periods of evolutionary time, duplications might seem more recent (that is, more identical) than expected on the basis of the original timing of the duplication event. Phylogenetic studies of dozens of SDs and genome-wide comparisons of primate-genome sequences, however, do not generally suggest that the effect of gene conversion has been so longstanding to warrant the rejection of a molecular clock. Evidence of extensive conversion that is maintained in the long term has only been detected within a family of large palindromic SDs on the non-recombining portion of the Y chromosome[65] and rDNA portions of acrocentric chromosomes[66]. The former appear to have maintained allelic levels of identity (99.8%), despite having emerged before divergence from a common ancestor of the chimpanzee and human[65]. A limited study of high-quality orthologous sequence from humans, chimpanzees and macaques did show elevated rates of substitution in regions of human segmental duplication[7]. The effect, however, was relatively subtle (10–25% increase) and was not restricted to the duplicated segment, arguing against the involvement of gene conversion as the cause of the increased substitution rate. An examination of the sequence identity of SDs that are present in both the chimpanzee and human and that probably arose in the common ancestor of the two species showed that, overall, <10% of these have sequence identities greater than the expected average sequence divergence between the two species[6]. In summary, excluding the Y chromosome, such sequences have not been grossly homogenized by large-scale gene conversion within each lineage since divergence. This is consistent with evidence that most meiotic conversion events are small in scale (that is, they involve sequences of 50–300 bp)[60,61] and are restricted to particular recombinogenic hot spots[47–49,59,67–69]. It follows that meiotic conversion within SDs is likely to reinforce the formation of hot spots as recurrent conversion maintains small regions of high sequence identity.

SDs and primate genome architecture

For several decades, karyotype analyses, chromosome painting and gene-order studies supported a relatively random process of chromosomal evolution in mammals[70]. More recently, the submicroscopic resolution afforded by genomic-sequence availability suggests an overabundance of small rearrangements and reusage of rearrangement breakpoints during mammalian genome evolution, which is inconsistent with a model of random breakage[71–74]. These reused syntenic breakpoints show a strong association with SD content — it has been estimated that 25–53% of all breakpoints that are syntenic in the mouse and human map to a primate SD[75,76]. Similar associations are found among many mammalian species when sites of duplication and large-scale chromosomal rearrangement are compared[77,78]. Interestingly, in an analysis of mouse, human and rat genomes, syntenic breakpoints that have emerged specifically within the human or mouse lineages showed an equivalent enrichment of primate-specific SDs[76]. This indicates that certain regions of the ancestral genome might have been susceptible to both breakage and duplication since the mammalian radiation.

Important insights into the association between SDs and rearrangements are gained from the molecular resolution of hominoid large-scale chromosomal rearrangements[79,80]. Owing to their recent evolutionary origin, such events are unobscured by long-term evolutionary divergence and further rearrangement. Comparisons of the great ape and human genomes show that 8 out of 11 rearrangement events that have been precisely characterized associate specifically with SDs[20,22,81–91]. As uncharacterized breakpoints often map to poorly assembled and heavily duplicated regions, it is probable that this association will be strengthened by future analyses as these regions are resolved. Of particular interest will be the molecular characterization of chromosomal rearrangements during the evolution of the gibbon — a primate that has experienced an accelerated rate of rearrangement[92].

Given their large numbers relative to karyotypic rearrangements, a causal role for SDs has been suggested in the origin of rearrangements[78]. However, inversions and other rearrangements might actually be involved in the creation of SDs. For example, this was suggested by the analysis of an unusually complicated, inverted, dual chimpanzee-only duplication. This duplication is present at the site of a chimpanzee-specific pericentric inversion breakpoint on the orthologue of human chromosome 12. The authors suggest that DSBs followed by inversion and repair of long single-strand overhangs, account for the structure of this SD and the pericentric inversion in this region[87].

Large-scale sequence comparisons of the human and chimpanzee genomes have identified hundreds of smaller deletions, duplications and inversions that also show associations with SDs[22,93,94]. This is perhaps unsurprising: duplications, through NAHR, promote local genomic instability. Gains and losses of duplicated DNA seem to have been common in hominoids. For example, a survey of lineage-specific differences within hominoids using cDNA microarrays identified 1,005 genes that showed copy-number differences when compared with humans[95]. A disproportionate fraction of these mapped to sites of segmental duplication. Similarly, a duplication map of chimpanzee SDs (>20 kb and >94% identity) revealed 76.3 Mb of sequence that is differentially duplicated between chimpanzees and humans[6] (FIG. 5a). These data have been used to estimate a fixation rate for *de novo* duplications of 3–5 Mb per million years per lineage[6]. This level of change indicates that SD events have had a greater impact at the base-pair level (2.7%) in these two species than single-base-pair changes (1.2%) (REFS 6,93), suggesting that SDs have had a key role in shaping the architecture of primate genomes.

Chromosome painting
Visualization of individual, whole chromosomes by fluorescence *in situ* hybridization.

Synteny
A term originally meaning simply 'on the same chromosome' (regardless of linkage) which has now been co-opted to refer to a continuous block of sequence with a shared organization (conserved linkage) between two or more species.

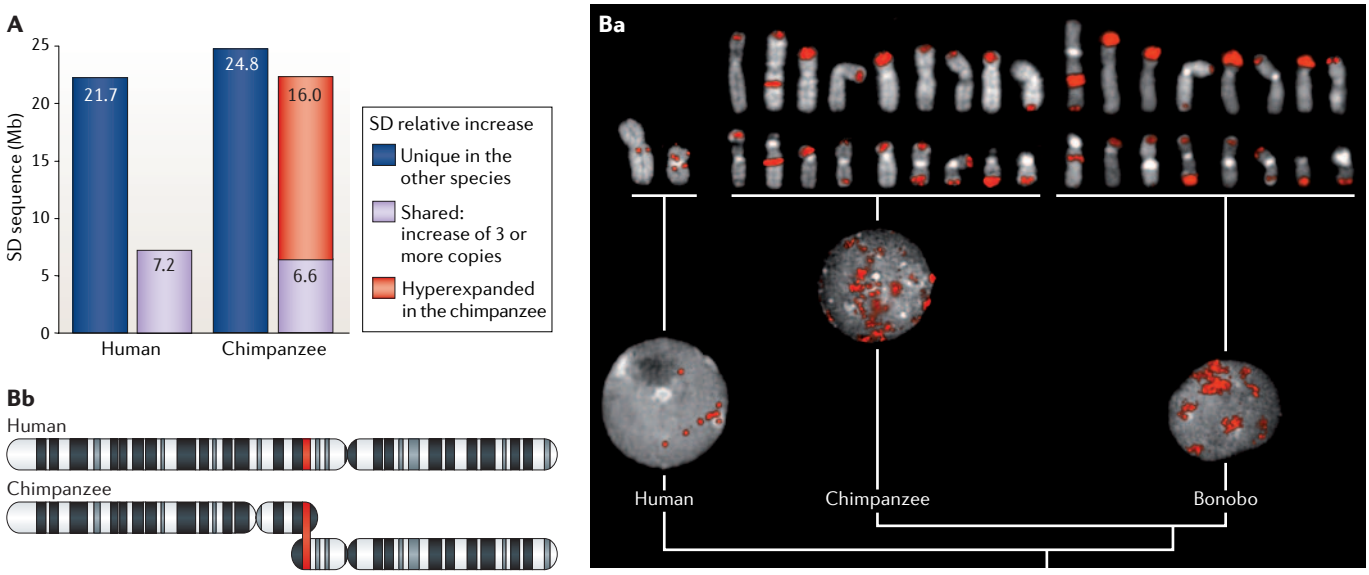

Figure 5 | **Variation in genomic segmental duplication (SD) content between chimpanzees and humans.** **A** | A comparison of duplicated sequence from the chimpanzee and human genomes allowed the identification of regions of shared duplication and those that contain human-specific or chimpanzee-specific multicopy sequence (>94% identity and >20 kb) (REF. 6). An estimated 60% represented duplicative gain, whereas 40% of the change occurred as a result of deletion of ancestral duplications. Overall, as shown in the graph, a minimal estimate of 76 Mb is differentially duplicated between humans and chimpanzees, corresponding to 3–5 Mb duplication gain per million years. This is a conservative estimate owing to limitations of the detectable differences to >94% identity, >20 kb and threefold or greater copy-number difference. **B** | Hyperexpansion of an SD in the chimpanzee lineage. Fluorescence *in situ* hybridization staining (red) is shown in panel **Ba** for an SD duplicon of ~40 kb, revealing an expansion to 400–500 copies of this sequence in the chimpanzee and bonobo, mainly near telomeres (only chromosomes that carry the duplicon are shown). This expansion added ~16 Mb of duplicated sequence in a common ancestor of chimpanzees and bonobos that is not present in gorillas and humans. As shown in panel **Bb**, the same SD underlies an association between duplication and rearrangement: it is associated with a large-scale chromosome fusion event that produced human chromosome 2 (this chromosome is shown next to the two orthologous chimpanzee chromosomes; the SD lies 40 kb proximal of the fusion point).

Roles in the evolution of new genes

The duplication of genomic sequences is one of the primary mechanisms for the creation of new genes[96–98], and the role of SDs in creating novel primate genes — by adaptive evolution, gene fusion or exon exaptation — has been the subject of many studies. Duplication of an entire gene, either in tandem or in an interspersed configuration, seems to be the most common method through which SDs create new genes. The exon shuffling and fusion transcripts that result from juxtaposing unrelated SDs occasionally produce transcripts that maintain an ORF (FIG. 6). Whole-genome analyses of human and chimpanzee SDs detect enrichment within these sequences of both genic features (such as exons) and expressed genes[7,9,13]. In fact, gene density is the greatest correlate of SD density (although even this correlation is weak) when compared with other factors such as GC content, repeat density and recombination rate[9]. However, not all SDs are enriched for genes, and not all duplication regions are likely to be equally innovative, owing to mechanistic and architectural constraints.

In humans, most of the gene enrichment that is seen within SDs is accounted for by interstitial intrachromosomal duplications. The subtelomeric regions also show some evidence for gene duplication, with well-documented interchromosomal exchanges leading to the expansion of multigene families including olfactory receptors and RAB-like and FOXD4 (REFS 32,99–101) genes. By contrast, pericentromeric interchromosomal duplications show reduced gene content and transcriptional activity when compared with the genome average for SDs[11,25,26,102]. Although there is ample evidence of many duplicated genic segments and an increased frequency of fusion transcripts in these regions, a minority maintain an ORF. However, gene annotation in these highly duplicated regions is particularly challenging.

Interestingly, a bias towards expression of transcripts in testes and cancer is a common feature of many of the fusion transcripts within duplicated regions[11,25,102]. However, this is more likely to be a consequence of chromatin structure[103], which might limit expression to testes or to malignancies (where chromosome structure is less tightly regulated), than it is to be a result of selection on expression patterns.

A growing list of hundreds of genes and gene families have been identified within interstitial SDs[11,13,21,93], but their functions require further investigation. This is especially true for genes in interspersed SDs, many of which show limited or no evidence of homology to genes in other mammals and model organisms. Their

Exaptation
When a gene or part of a gene (such as a domain or exon) is used for a new potentially adaptive function. Can also be used to describe cases where splicing occurs in non-genic sequence to create a new exon. This is a type of domain accretion, which is the addition of an exon or exons to a gene or transcript.

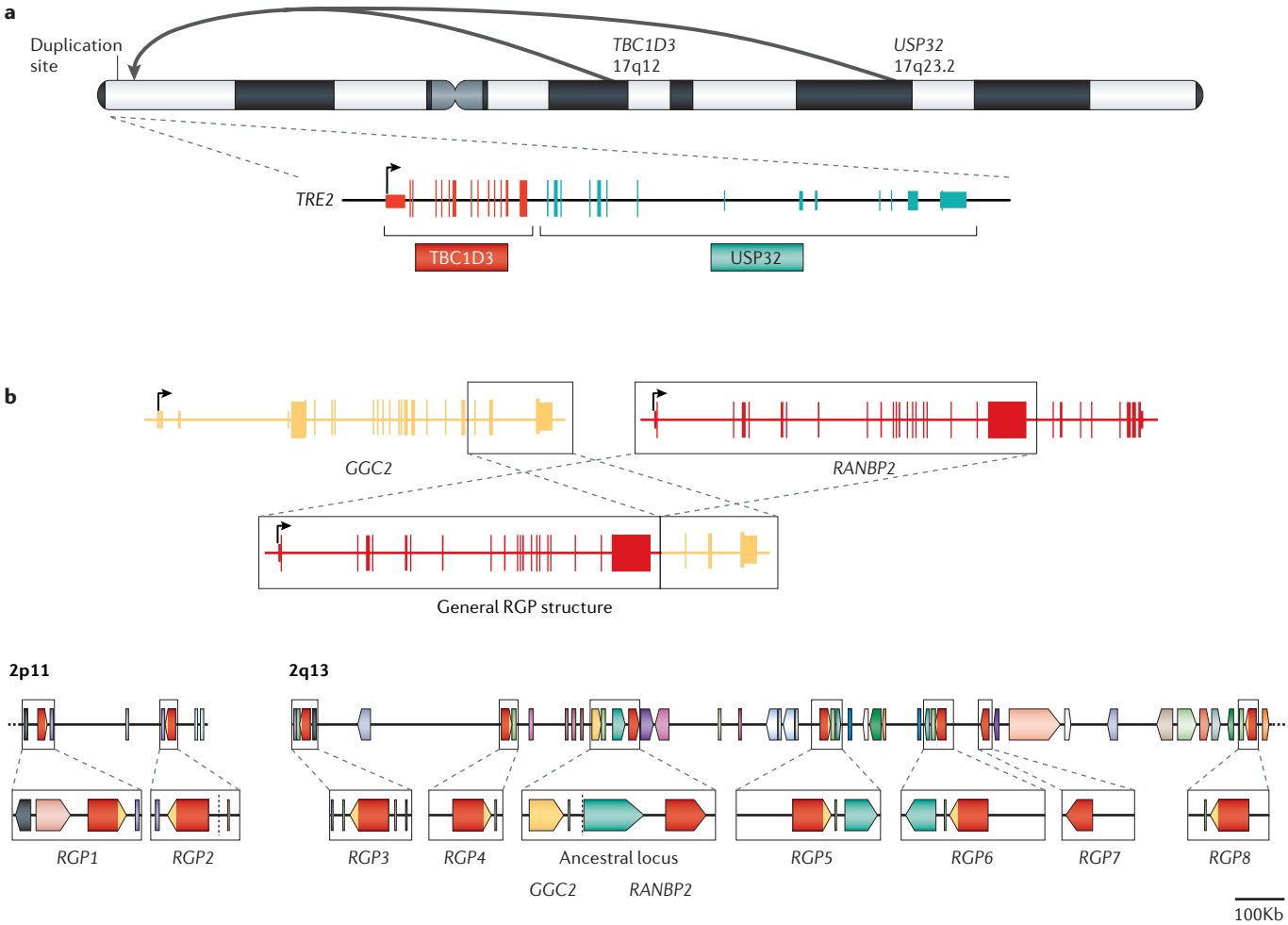

Figure 6 | **Gene innovation in segmental duplications (SDs).** Two examples of 'novel' primate-specific genes that have been created by SDs are shown. **a** | The *TRE2* oncogene (also known as ubiquitin-specific protease 6 (*USP6*)) is a hominoid-specific gene that is located at 17p13.2 in humans. This gene was formed from the fusion of two SDs, each carrying sequence from genes that are located on the q-arm of human chromosome 17, at a distance from the duplication site (TBC1 domain family member 3 (*TBC1D3*) and *USP32*). *TRE2* has derived exons 1–14 (red) from *TCB1D3* and exons 15–29 (green) from *USP32*. **b** | The RanBP2-like GRIP-domain-containing protein (RGP) gene family formed from the fusion of SDs of the genes *RANBP2* and *GGC2*, which are located on human chromosome 2q13. As shown in the top panel, a protypical RGP is composed of the first 20 exons of RANBP2 (red) and last three exons of GGC (yellow). This fusion sequence has been extensively duplicated as part of duplication hubs on chromosome 2, on both the q- and p-arms. Below this is a detailed view of the duplicon structure of the RGP-containing regions involved, illustrating the complex mosaic pattern that has arisen during the formation of this gene family. *GGC2* and *RANBP2* regions are shown in yellow and red, respectively; regions of other genes are shown in various colours. These duplication hubs show evidence for multiple functional copies under extensive positive selection[21]. Modified with permission from REF. 21 © (2005) Cold Spring Harbor Laboratory Press.

Duplication block
A group of juxtaposed duplicons that might be duplicated as part of a larger secondary duplication. Also sometimes referred to as a low copy repeat.

Duplication core
The central sequence around which a complex pattern of duplication forms common to a subset of interspersed duplications.

expansions have occurred within the last 25 million years and have been accompanied by the formation of complex interstitial duplication blocks at multiple chromosomal locations. Indeed, these rapidly evolving genes frequently map to the most duplicated segment (the duplication cores) of interspersed duplication blocks.

Genes in SDs share several properties. First, when compared with unique genes they are 5–10 times as likely to show interspecies and intraspecies structural and/or copy-number variation[6,53,95], which correlates with differences in mRNA expression levels[6,104,105]. Second, strong signatures of positive selection are common in segmentally duplicated genes[18,21,93,106,107]. Third, several functional categories are enriched among these genes, including immune response, xenobiotic recognition (both olfactory reception and drug detoxification), reproduction and nuclear functions. These three features suggest an important role for SDs in primate and human adaptive evolution: they might have facilitated adaptation to changes in food sources and to novel or altered infectious agents — particularly where a diversity of responses was advantageous. Elucidating the functions of these genes should be a key goal of future studies.

Consequences for variation within species

In humans, polymorphic insertions, deletions and inversion are non-randomly distributed, with a 4–12-fold greater frequency near sites of SD[51–54,108,109]. Similarly, in chimpanzee populations there is an almost 20-fold enrichment of copy-number variation for regions in which duplications arose in the human–chimpanzee ancestor[110]. As every SD begins its evolutionary life cycle as a structural variant, this link is perhaps unsurprising. Nevertheless, these data indicate that duplicated regions are continuing to rearrange and 'evolve' in contemporary primate populations.

Does such variation have phenotypic consequences? Insights into the relationship between duplications and phenotypic variation are longstanding. Some of the earliest human genetic phenotypic variation to be mapped — such as colour blindness, rhesus blood-group sensitivity and forms of α- and β-thalassaemia[111–113] — result from complex structural alterations of duplicated genes that show variation among individuals[114–116]. The effects of these alterations might be direct, for example owing to copy-number variation and concomitant differences in the expression or sequence of the gene, or indirect, such as recurrent structural rearrangements, as in the case of genomic disorders[117].

Recent data indicate that SDs and their associated structural variation might have protective or beneficial effects. For example, a common ~900-kb inversion polymorphism, which is mediated by SDs on 17q21.31, is associated with positive selection and, perhaps, increased fertility within the Icelandic population[118]. Similarly, increased copy number of CCL3L1 owing to duplication is associated with a significant reduction in susceptibility to HIV infection and progression to AIDS[119]. Most recently, an association has been found between increased copy number of the immunoglobulin Fc receptor gene *FCGR3* and a decreased risk of lupus glomerulonephritis. This structural variation in a duplicated gene accounts for much greater protection from disease than has been associated with single nucleotide variants[120].

An important lesson from human genetics is that copy number alone does not determine the genotype–phenotype relationship. In the case of rhesus-factor sensitization, colour blindness and α- and β-thalassaemias, it has been through understanding the precise structural details of the formation of fusion genes or the positions of duplicated segments with respect to functional promoters and transcriptional regulators that the most meaningful associations with phenotype or disease have been made. There is also a growing number of examples of rearrangements, including duplications, with long-range effects on gene transcription owing to the alteration of transcriptional regulators[121–123]. The same logic extends to non-human primates: to understand the phenotypic significance of these regions during evolution, genetic differences must be resolved at the base-pair level. This will require high-quality sequence as well as experiment-based gene and regulatory-sequence annotation in some of the regions of primate genomes that are the most difficult to study.

Xenobiotic
A compound that is foreign to biological systems, often referring to man-made compounds that are resistant to biodegradation.

Future directions

The copy number, location and structure of many regions that contain SDs seem to vary significantly between closely related primates, which might have evolutionary significance. However, the dynamic and complex architecture of many of these regions make them recalcitrant to standard methods of sequencing and analysis. Five non-human primate species — chimpanzee, gorilla, orangutan, macaque and marmoset — have been targeted for complete genome sequencing by the US National Institutes of Health and the Wellcome Trust. Accurate assembly of these genomes will require more than simply assembling whole-genome shotgun-sequence data[12]. Complementary sequencing of large-insert genomic clones for duplicated regions provides one potential approach to resolving their sequences. For example, we estimate that the addition of finished sequence data from ~1,000–1,500 BAC clones per genome would significantly improve sequence annotation of these regions, the identification of orthologous gene relationships and our understanding of primate genetic variation. Inevitably, novel sequencing technologies and the high-quality sequencing of additional species will be required to completely disentangle the structure and pattern of SDs in other non-human primate genomes.

In primate genomes, interspersed duplications have the potential to destabilize a large fraction of the genome through NAHR-mediated rearrangement. The association of such genomic duplications with gene-rich regions creates an evolutionary paradox: what drives the spread of these duplications during evolution, given their destabilizing effect? A clue comes from the fact that transcript diversity is increased within these regions. A common theme is emerging in which novel fusion transcripts and genes under positive selection map more frequently to the 'cores' of the most complex interspersed duplication blocks[18,21,124,125]. Based on known examples of positive selection[18,21], it is tempting to speculate that duplication fixation occurs as a result of positive selection acting on a small subset of emerging genes for which increased copy number or diversity of gene family members outweighs the negative consequences conferred by an increased probability of chromosomal rearrangement. The complex duplication architecture that has evolved flanking these cores might simply be a consequence of the expansion of newly minted adaptive gene families.

While the catalogue of primate-specific and hominoid-specific gene innovations is growing rapidly, understanding of their functions lags. Considerable effort will be required to dissect the biological roles of these duplicated genes owing to their redundant nature and absence of homologues in model organisms. Traditional transgenic knockout models of gene function are not an option, and the inherent copy-number redundancy might make it difficult to recognize subtle phenotypic effects that are due to point mutations. A key step to understanding their function will be to prove that that these genes encode functional products by developing antibodies to assess the localization of the encoded proteins.

Two-hybrid assays with such candidates could provide functional insights through the identification of more conserved binding partners that have known functions. In addition, the use of RNAi in cell culture will allow the knockdown of duplicated genes to determine the effects of this on transcriptional networks. Finally, the study of SD-associated structural variation and associated pathologies might provide the most useful information about the functions of duplicated genes, as associations between polymorphisms that affect them and disease susceptibility provide key phenotypic insights. Using these approaches, the study of SDs promises to greatly advance our understanding of both primate evolution and genetic susceptibility to human disease.

1. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001).
2. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
3. Cheung, J. *et al.* Recent segmental and gene duplications in the mouse genome. *Genome Biol.* **4**, R47 (2003).
4. Tuzun, E., Bailey, J. A. & Eichler, E. E. Recent segmental duplications in the working draft assembly of the brown Norway rat. *Genome Res.* **14**, 493–506 (2004).
5. Bailey, J. A., Church, D. M., Ventura, M., Rocchi, M. & Eichler, E. E. Analysis of segmental duplications and genome assembly in the mouse. *Genome Res.* **14**, 789–801 (2004).
6. Cheng, Z. *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**, 88–93 (2005).
7. 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).
8. Cheung, J. *et al.* Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.* **4**, R25 (2003).
9. 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).
10. 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).
11. She, X. *et al.* The structure and evolution of centromeric transition regions within the human genome. *Nature* **430**, 857–864 (2004).
12. She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927–930 (2004).
13. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
14. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
15. Eichler, E. E., Clark, R. A. & She, X. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nature Rev. Genet.* **5**, 345–354 (2004).
16. Courseaux, A. & Nahon, J. L. Birth of two chimeric genes in the Hominidae lineage. *Science* **291**, 1293–1297 (2001).
17. Horvath, J. E. *et al.* Punctuated duplication seeding events during the evolution of human chromosome 2p11. *Genome Res* **15**, 914–927 (2005).
18. Johnson, M. E. *et al.* Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519 (2001).
19. Courseaux, A. *et al.* Segmental duplications in euchromatic regions of human chromosome 5: a source of evolutionary instability and transcriptional innovation. *Genome Res.* **13**, 369–381 (2003).
20. Stankiewicz, P., Shaw, C. J., Withers, M., Inoue, K. & Lupski, J. R. Serial segmental duplications during primate evolution result in complex human genome architecture. *Genome Res.* **14**, 2209–2220 (2004).
21. Ciccarelli, F. D. *et al.* Complex genomic rearrangements lead to novel primate gene function. *Genome Res.* **15**, 343–351 (2005).
22. Newman, T. L. *et al.* A genome-wide survey of structural variation between human and chimpanzee. *Genome Res.* **15**, 1344–1356 (2005).
23. Humphray, S. J. *et al.* DNA sequence and analysis of human chromosome 9. *Nature* **429**, 369–374 (2004).
24. Kirsch, S. *et al.* Interchromosomal segmental duplications of the pericentromeric region on the human Y chromosome. *Genome Res.* **15**, 195–204 (2005).
25. Guy, J. *et al.* Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10p. *Genome Res.* **13**, 159–172 (2003).
26. Guy, J. *et al.* Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10q. *Hum. Mol. Genet.* **9**, 2029–2042 (2000).
27. Brun, M. E., Ruault, M., Ventura, M., Roizes, G. & De Sario, A. Juxtacentromeric region of human chromosome 21: a boundary between centromeric heterochromatin and euchromatic chromosome arms. *Gene* **312**, 41–50 (2003).
28. Horvath, J., Schwartz, S. & Eichler, E. The mosaic structure of a 2p11 pericentromeric segment: a strategy for characterizing complex regions of the human genome. *Genome Res* **10**, 839–852 (2000).
29. Eichler, E. E. *et al.* Interchromosomal duplications of the adrenoleukodystrophy locus: a phenomenon of pericentromeric plasticity. *Hum. Mol. Genet.* **6**, 991–1002 (1997).
30. Riethman, H. C. *et al.* Integration of telomere sequences with the draft human genome sequence. *Nature* **409**, 948–951 (2001).
31. Linardopoulou, E. V. *et al.* Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* **437**, 94–100 (2005).
32. Trask, B. *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).
33. Antonell, A., de Luis, O., Domingo-Roura, X. & Perez-Jurado, L. A. Evolutionary mechanisms shaping the genomic structure of the Williams–Beuren syndrome chromosomal region at human 7q11.23. *Genome Res.* **15**, 1179–1188 (2005).
34. Bradley, M. E. & Benner, S. A. Phylogenomic approaches to common problems encountered in the analysis of low copy repeats: the sulfotransferase 1A gene family example. *BMC Evol. Biol.* **5**, 22 (2005).
35. Jin, H. *et al.* Structural evolution of the *BRCA1* genomic region in primates. *Genomics* **84**, 1071–1082 (2004).
36. Koszul, R., Caburet, S., Dujon, B. & Fischer, G. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J.* **23**, 234–243 (2004).
37. Schacherer, J., de Montigny, J., Welcker, A., Souciet, J. L. & Potier, S. Duplication processes in *Saccharomyces cerevisiae* haploid strains. *Nucleic Acids Res.* **33**, 6319–6326 (2005).
38. Bailey, J. A., Liu, G. & Eichler, E. E. An *Alu* transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73**, 823–834 (2003).
39. Jurka, J., Kohany, O., Pavlicek, A., Kapitonov, V. V. & Jurka, M. V. Duplication, coclustering, and selection of human *Alu* retrotransposons. *Proc. Natl Acad. Sci. USA* **101**, 1268–1272 (2004).
40. Zhou, Y. & Mishra, B. Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. *Proc. Natl Acad. Sci. USA* **102**, 4051–4056 (2005).
41. Babcock, M. *et al.* Shuffling of genes within low-copy repeats on 22q11 (LCR22) by *Alu*-mediated recombination events during evolution. *Genome Res.* **13**, 2519–2532 (2003).
42. Eichler, E. E., Archidiacono, N. & Rocchi, M. CAGGG repeats and the pericentromeric duplication of the hominoid genome. *Genome Res.* **9**, 1048–1058 (1999).
43. Woodward, K. J. *et al.* Heterogeneous duplications in patients with Pelizaeus–Merzbacher disease suggest a mechanism of coupled homologous and nonhomologous recombination. *Am. J. Hum. Genet.* **77**, 966–987 (2005).
44. Pentao, L., Wise, C., Chinault, A., Patel, P. & Lupski, J. Charcot–Marie–Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1. 5 Mb monomer unit. *Nature Genet.* **2**, 292–300 (1992).
45. Shaw, C. J. & Lupski, J. R. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum. Mol. Genet.* **13**, R57–R64 (2004).
46. Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**, 74–82 (2002).
47. Reiter, L. T. *et al.* Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients. *Am. J. Hum. Genet.* **62**, 1023–1033 (1998).
48. Lopez-Correa, C. *et al.* Recombination hotspot in NF1 microdeletion patients. *Hum. Mol. Genet.* **10**, 1387–1392 (2001).
49. Bi, W. *et al.* Reciprocal crossovers and a positional preference for strand exchange in recombination events resulting in deletion or duplication of chromosome 17p11.2. *Am. J. Hum. Genet.* **73**, 1302–1315 (2003).
50. Fredman, D. *et al.* Complex SNP-related sequence variation in segmental genome duplications. *Nature Genet.* **36**, 861–866 (2004).
51. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
52. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genet.* **36**, 949–951 (2004).
53. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nature Genet.* **37**, 727–732 (2005).
54. Sharp, A. J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
55. Tayebi, N. *et al.* Gaucher disease with parkinsonian manifestations: does glucocerebrosidase deficiency contribute to a vulnerability to parkinsonism? *Mol. Genet. Metab.* **79**, 104–109 (2003).
56. Pavlicek, A., House, R., Gentles, A. J., Jurka, J. & Morrow, B. E. Traffic of genetic information between segmental duplications flanking the typical 22q11.2 deletion in velo-cardio-facial syndrome/DiGeorge syndrome. *Genome Res.* **15**, 1487–1495 (2005).
57. Verrelli, B. C. & Tishkoff, S. A. Signatures of selection and gene conversion associated with human color vision variation. *Am. J. Hum. Genet.* **75**, 363–375 (2004).
58. Bosch, E., Hurles, M. E., Navarro, A. & Jobling, M. A. Dynamics of a human interparalog gene conversion hotspot. *Genome Res.* **14**, 835–844 (2004).
59. Hurles, M. E., Willey, D., Matthews, L. & Hussain, S. S. Origins of chromosomal rearrangement hotspots in the human genome: evidence from the AZFa deletion hotspots. *Genome Biol.* **5**, R55 (2004).
60. Jeffreys, A. J. & May, C. A. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nature Genet.* **36**, 151–156 (2004).
61. Jeffreys, A. J. & Neumann, R. Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. *Hum. Mol. Genet.* **14**, 2277–2287 (2005).
62. Jackson, M. S. *et al.* Evidence for widespread reticulate evolution within human duplicons. *Am. J. Hum. Genet.* **77**, 824–840 (2005).
63. Estivill, X. *et al.* Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. *Hum. Mol. Genet.* **11**, 1987–1995 (2002).

64. Hallast, P., Nagirnaja, L., Margus, T. & Laan, M. Segmental duplications and gene conversion: human luteinizing hormone/chorionic gonadotropin β gene cluster. *Genome Res.* **15**, 1535–1546 (2005).
65. Rozen, S. *et al.* Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**, 873–876 (2003).
66. Gonzalez, I. L. *et al.* Variation among human 28S ribosomal RNA genes. *Proc. Natl Acad. Sci. USA* **82**, 7666–7670 (1985).
67. Reiter, L. *et al.* A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element. *Nature Genet.* **12**, 288–297 (1996).
68. Bayes, M., Magano, L. F., Rivera, N., Flores, R. & Perez Jurado, L. A. Mutational mechanisms of Williams–Beuren syndrome deletions. *Am. J. Hum. Genet.* **73**, 131–151 (2003).
69. Visser, R. *et al.* Identification of a 3.0-kb major recombination hotspot in patients with sotos syndrome who carry a common 1.9-Mb microdeletion. *Am. J. Hum. Genet.* **76**, 52–67 (2005).
70. Nadeau, J. H. & Taylor, B. A. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl Acad. Sci. USA* **81**, 814–818 (1984).
71. Pevzner, P. & Tesler, G. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.* **13**, 37–45 (2003).
72. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA* **100**, 11484–11489 (2003).
73. 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).
74. 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).
75. 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).
76. Bailey, J. A., Baertsch, R., Kent, W. J., Haussler, D. & Eichler, E. E. Hotspots of mammalian chromosomal evolution. *Genome Biol.* **5**, R23 (2004).
77. Armengol, L. *et al.* Murine segmental duplications are hot spots for chromosome and gene evolution. *Genomics* **86**, 692–700 (2005).
78. Murphy, W. J. *et al.* Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309**, 613–617 (2005).
79. Yunis, J. J. & Prakash, O. The origin of man: a chromosomal pictorial legacy. *Science* **215**, 1525–1530 (1982).
80. Yunis, J. J., Sawyer, J. R. & Dunham, K. The striking resemblance of high-resolution G-banded chromosomes of man and chimpanzee. *Science* **208**, 1145–1148 (1980).
81. Fan, Y., Linardopoulou, E., Friedman, C., Williams, E. & Trask, B. J. Genomic structure and evolution of the ancestral chromosome fusion site in 2q13–2q14.1 and paralogous regions on other human chromosomes. *Genome Res.* **12**, 1651–1662 (2002).
82. Yue, Y. *et al.* Genomic structure and paralogous regions of the inversion breakpoint occurring between human chromosome 3p12.3 and orangutan chromosome 2. *Cytogenet. Genome Res.* **108**, 98–105 (2005).
83. Kehrer-Sawatzki, H. *et al.* Breakpoint analysis of the pericentric inversion distinguishing human chromosome 4 from the homologous chromosome in the chimpanzee (*Pan troglodytes*). *Hum. Mutat.* **25**, 45–55 (2005).
84. Szamalek, J. M. *et al.* Molecular characterisation of the pericentric inversion that distinguishes human chromosome 5 from the homologous chimpanzee chromosome. *Hum. Genet.* **117**, 168–176 (2005).
85. Kehrer-Sawatzki, H., Szamalek, J. M., Tanzer, S., Platzer, M. & Hameister, H. Molecular characterization of the pericentric inversion of chimpanzee chromosome 11 homologous to human chromosome 9. *Genomics* **85**, 542–550 (2005).
86. Locke, D. P. *et al.* Refinement of a chimpanzee pericentric inversion breakpoint to a segmental duplication cluster. *Genome Biol.* **4**, R50 (2003).
87. Kehrer-Sawatzki, H., Sandig, C. A., Goidts, V. & Hameister, H. Breakpoint analysis of the pericentric inversion between chimpanzee chromosome 10 and the homologous chromosome 12 in humans. *Cytogenet. Genome Res.* **108**, 91–97 (2005).
88. Shimada, M. K. *et al.* Nucleotide sequence comparison of a chromosome rearrangement on human chromosome 12 and the corresponding ape chromosomes. *Cytogenet. Genome Res.* **108**, 83–90 (2005).
89. 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).
90. Kehrer-Sawatzki, H. *et al.* Molecular characterization of the pericentric inversion that causes differences between chimpanzee chromosome 19 and human chromosome 17. *Am. J. Hum. Genet.* **71**, 375–388 (2002).
91. Dennehey, B. K., Gutches, D. G., McConkey, E. H. & Krauter, K. S. Inversion, duplication, and changes in gene context are associated with human chromosome 18 evolution. *Genomics* **83**, 493–501 (2004).
92. Muller, S., Hollatz, M. & Wienberg, J. Chromosomal phylogeny and evolution of gibbons (Hylobatidae). *Hum. Genet.* **113**, 493–501 (2003).
93. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
94. Feuk, L. *et al.* Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet.* **1**, e56 (2005).
95. Fortna, A. *et al.* Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol.* **2**, e207 (2004).
96. Muller, H. J. Bar duplication. *Science* **83**, 528–530 (1936).
97. Ohno, S., Wolf, U. & Atkin, N. Evolution from fish to mammals by gene duplication. *Hereditas* **59**, 169–187 (1968).
98. Taylor, J. S. & Raes, J. Duplication and divergence: the evolution of new genes and old ideas. *Annu. Rev. Genet.* **38**, 615–643 (2004).
99. Brand-Arpon, V. *et al.* A genomic region encompassing a cluster of olfactory receptor genes and a myosin light chain kinase (*MYLK*) gene is duplicated on human chromosome regions 3q13–q21 and 3p13. *Genomics* **56**, 98–110 (1999).
100. Wong, A. C. *et al.* Two novel human RAB genes with near identical sequence each map to a telomere-associated region: the subtelomeric region of 22q13.3 and the ancestral telomere band 2q13. *Genomics* **59**, 326–334 (1999).
101. Wong, A. *et al.* Diverse fates of paralogs following segmental duplication of telomeric genes. *Genomics* **84**, 239–247 (2004).
102. Mudge, J. M. & Jackson, M. S. Evolutionary implications of pericentromeric gene expression in humans. *Cytogenet. Genome Res.* **108**, 47–57 (2005).
103. Grunau, C. *et al.* Frequent DNA hypomethylation of human juxtacentromeric BAGE loci in cancer. *Genes Chromosomes Cancer* **43**, 11–24 (2005).
104. Hollox, E. J., Armour, J. A. & Barber, J. C. Extensive normal copy number variation of a β-defensin antimicrobial-gene cluster. *Am. J. Hum. Genet.* **73**, 591–600 (2003).
105. Khaitovich, P. *et al.* Regional patterns of gene expression in human and chimpanzee brains. *Genome Res.* **14**, 1462–1473 (2004).
106. Birtle, Z., Goodstadt, L. & Ponting, C. Duplication and positive selection among hominin-specific PRAME genes. *BMC Genomics* **6**, 120 (2005).
107. Semple, C. A., Rolfe, M. & Dorin, J. R. Duplication and selection in the evolution of primate β-defensin genes. *Genome Biol.* **4**, R31 (2003).
108. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nature Rev. Genet.* **7**, 85–97 (2006).
109. Eichler, E. E. Widening the spectrum of human genetic variation. *Nature Genet.* **38**, 9–11 (2006).
110. Perry, G. H. *et al.* Hotspots for copy number variation in chimpanzees and humans. *Proc. Natl Acad. Sci. USA* **103**, 8006–8011 (2006).
111. Wilson, E. B. The sex chromosomes. *Arch. Mikrosk. Anat. Entwicklungsmech.* **77**, 249–271 (1911).
112. Levine, P., Katzin, E. M. & Burnham, L. Isoimmunizationin pregnancy: its possible bearing on the etiology of erythroblastosis foetalis. *JAMA* **116**, 825–827 (1941).
113. Cooley, T. B. & Lee, P. A series of cases of splenomegaly inchildren with anemia and peculiar bone changes. *Trans. Am. Pediatr. Soc.* **37**, 29 (1925).
114. Fucharoen, S. & Winichagoon, P. Thalassemia and abnormal hemoglobin. *Int. J. Hematol.* **76** (Suppl. 2), 83–89 (2002).
115. Wagner, F. F. & Flegel, W. A. Review: the molecular basis of the Rh blood group phenotypes. *Immunohematol.* **20**, 23–36 (2004).
116. Deeb, S. S. The molecular basis of variation in human color vision. *Clin. Genet.* **67**, 369–377 (2005).
117. Lupski, J. R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14**, 417–422 (1998).
118. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nature Genet* **37**, 129–137 (2005).
119. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–1440 (2005).
120. Aitman, T. J. *et al.* Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851–855 (2006).
121. Kleinjan, D. A. & van Heyningen, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* **76**, 8–32 (2005).
122. Lee, J. A. *et al.* Spastic paraplegia type 2 associated with axonal neuropathy and apparent PLP1 position effect. *Ann. Neurol.* **59**, 398–403 (2006).
123. Velagaleti, G. V. *et al.* Position effects due to chromosome breakpoints that map approximately 900 kb upstream and approximately 1.3 Mb downstream of *SOX9* in two patients with campomelic dysplasia. *Am. J. Hum. Genet.* **76**, 652–662 (2005).
124. 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).
125. Zody, M. C. *et al.* Analysis of the DNA sequence and duplication history of human chromosome 15. *Nature* **440**, 671–675 (2006).

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION

Chimpanzee Segmental Duplication Database, Genome Sciences, University of Washington: http://chimpparalogy.gs.washington.edu
Ensembl Genome Browsers: http://www.ensembl.org
Human Genome Segmental Duplication Database, Hospital for Sick Children, University of Toronto: http://projects.tcag.ca/humandup
Human Segmental Duplication Database, Genome Sciences, University of Washington: http://humanparalogy.gs.washington.edu
UCSC Genome Bioinformatics: http://genome.ucsc.edu
Access to this links box is available online.