Recent duplication, domain accretion and the dynamic mutation of the human genome

Evan E. Eichler

An estimated 5% of the human genome consists of interspersed duplications that have arisen over the past 35 million years of evolution. Two categories of such recently duplicated segments can be distinguished: segmental duplications between nonhomologous chromosomes (transchromosomal duplications) and duplications mainly restricted to a particular chromosome (chromosome-specific duplications). Many of these duplications exhibit an extraordinarily high degree of sequence identity at the nucleotide level (>95%) and span large genomic distances (1–100 kb). Preliminary analyses indicate that these same regions are targets for rapid evolutionary turnover among the genomes of closely related primates. The dynamic nature of these regions because of recurrent chromosomal rearrangement, and their ability to create fusion genes from juxtaposed cassettes suggest that duplicative transposition was an important force in the evolution of our genome.

The importance of gene duplication as a source of chromosomal variation, phenotypic difference and evolutionary change has been recognized since the 1930s (Refs 1,2). The seminal work of Susumu Ohno[3], later popularized in his monograph *Evolution by Gene Duplication* (1970), put forward polyploidization (whole-genome duplication) as the single most important mechanism by which vertebrate gene families have evolved[4]. Whole-genome duplication along with single base-pair mutation was heralded as the catalyst of vertebrate complexity. It allowed new genes to emerge unencumbered by the selective constraints of their ancestral function. The evolution of sex chromosome inheritance among mammals restricted whole-genome duplication events[5] to before the emergence of the vertebrate lineages (450–550 Myr ago). Since that time, only limited innovation through regional duplication of short stretches of genomic sequence is thought to have occurred (Box 1). Ohno stated[4]: 'Since polyploidy evolution was possible only at the initial stages of vertebrate evolution, it then follows that most of nature's experiments with gene duplication must have been done at the stages of fish and amphibians.'

The initial sequencing and analysis of the human genome reveals a remarkably complex pattern of both ancient and recent duplications[6–8]. With the sequence of a complete genome, it is now possible to begin to assess, without ascertainment bias, the contributions of various models of duplication (Box 1) to the architecture of the genome and ultimately the composition of the proteome. One surprising feature of the human genome analysis is the abundance of large blocks of genomic sequence that share a high degree of sequence identity (>90%). These blocks (termed segmental duplications) range in size from a few kb to hundreds of kb. They can include both exonic and intronic sequences and, unlike tandem duplicates, are interspersed throughout the genome. Clustering of segments of diverse origin seems to occur near PERICENTROMERIC (see Glossary) and subtelomeric regions. A similar genomic architecture has not been observed among the sequenced invertebrate genomes.

Glossary

α-satellite: A tandem repeated 171-bp sequence motif associated with the centromeric regions of all human chromosomes.
Domain accretion: The evolution of larger, multidomain proteins by the addition of DNA segments encoding distinct structural domains.
Duplicon: A duplicated genomic segment.
Duplication module: A duplicated segment in which the extent of the genomic duplication can be delineated by comparison to a progenitor sequence.
Dynamic mutation: Any mutation in which the probability of a second mutation changes with an initial mutation event.
Hardy–Weinberg equilibrium: Frequency of alleles will remain constant within a random-mating population unless acted upon by external forces.
Molecular-clock hypothesis: The rate of nucleotide change is constant over evolutionary time and is subject only to stochastic fluctuation. Among hominoids, the neutral rate of substitution has been estimated as 1.3×10^{-9} to 1.5×10^{-9} substitutions per site per year. Assuming that most of the duplicated sequence is neutral DNA (1.5×10^{-9} substitutions per site per year), paralogous sequence with 10% nucleotide divergence (corrected for multiple events) would have duplicated ~35 million years ago.
Neutral DNA: Sequence in which the pattern of nucleotide change is consistent with a model of no selection.
Orthologous: Sequence similarity due to a speciation event.
Paracentric inversion: A chromosomal inversion that does not span the centromere. (Pericentric inversions span the centromere.)
Paralogous: Sequence similarity due to a duplication event.
Pericentromeric DNA: DNA sequence flanking the primary constriction of chromosomes with defined centromeres. In the case of human chromosomes, the term has been used to characterize a 1–2 Mb region transition zone between euchromatin and α-satellite heterochromatic DNA.
Sympatric speciation: The formation of two separate species from a population occupying the same geographic location (as opposed to allopatric speciation where gene flow is interrupted by geographic separation).

Evan E. Eichler
Dept of Genetics and Center for Human Genetics, Case Western Reserve School of Medicine and University Hospitals of Cleveland, Cleveland, OH 44106, USA.
e-mail: eee@po.cwru.edu

http://tig.trends.com 0168-9525/01/$ – see front matter © 2001 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(01)02492-1

Box 1. Models of genomic duplication

Polyploidization

Duplication of the entire genome through a tetraploid (4N) evolutionary intermediate followed by re-establishment of the disomic state (2N). Leads to the apparent duplication of large portions of chromosomal regions. This process requires viability of a tetraploid ancestor and therefore among most vertebrates is restricted to antiquity. Two rounds of tetraploidization (2R hypothesis) were proposed to have occurred ~500 Myr ago[3,4,69].

Endoduplication

Tandem duplication of local chromosomal regions through unequal crossover events. Segments are capable of rapid homogenization through gene conversion. Leads to the formation of clustered gene families. Within the human genome, there is evidence for both recent (e.g. pregnancy-specific glycoprotein) and more ancient events (e.g. immunglobulin gene family).

Duplicative transposition

The duplication and transposition of a genomic block of material (1–100 kb) from one chromosomal region to another. Leads to an interspersed duplication of small genomic segments scattered throughout the genome (segmental duplications). Owing to their limited size, only recent events can be easily identified. Two different types of segmental duplications are distinguished: segments duplicated within a chromosome (chromosome-specific duplications) and between nonhomologous chromosomes (transchromosomal duplications).

The organization and recent emergence of these duplicated segments in the human genome suggests a third mechanism of duplication (duplicative transposition), independent from models of tandem duplication and polyploidization (Box 1). The object of this article is to summarize the structural features of this new class of duplicated segments and to discuss their potential impact on the evolution of our genome. This review will focus only on relatively recent duplication events (<35 Myr ago) for which the phylogenetic signal within NEUTRALLY EVOLVING, noncoding DNA can still be detected. Such events allow the extent of duplication to be resolved unambiguously and provide insight into duplicative forces currently shaping the structure of our genome.

Types of recent segmental duplications

Descriptions of segmental duplication within the human genome first appeared as a collection of anecdotal reports[9–15]. The duplications were identified either during routine physical mapping as part of the Human Genome Project or during characterization of breakpoints associated with recurrent chromosomal structural rearrangements. With the complete sequencing of the human genome, a more global view of segmental duplication has begun to emerge[7,8]. Operationally, two types of segmental duplications can be distinguished based largely on their distribution pattern within the genome. Chromosome-specific duplications are blocks of genomic sequence distributed in an interspersed fashion along a single chromosome. By contrast, transchromosomal duplications refer to segmental duplications that have spread among nonhomologous chromosomes (Fig. 1), with a peculiar bias towards pericentromeric and subtelomeric regions of humans chromosomes. As more finished human sequence and experimental data become available, it is apparent that overlap exists between these two groups. The mechanistic significance of this classification is therefore unknown, although structurally it appears that there are some important differences (Table 1).

Chromosome-specific duplications

Chromosome-specific duplications, also known as REPs (for ‘repeat regions’ after prokaryotic nomenclature) or LCR (‘low-copy repeat’ sequences)[16], were initially identified as unstable genomic regions associated with microdeletion and microduplication syndromes. In the past few years, a significant number of recurrent human genomic rearrangement breakpoints have been shown to lie within or in close proximity to duplicated segments[17–24]. The presence of large blocks of highly homologous sequence bracketing unique sequence predisposes these regions to translocation, deletion, inversion or duplication. This is believed to occur by processes of unequal crossover between the PARALOGOUS segments during meiosis (Fig. 2)[16].

Although the details regarding each of the chromosome-specific duplications vary, analysis of the composition and organization of these regions reveals several common features. In general, the duplications localize to a single chromosomal arm, with the paralogous segments separated by less than 10 Mb of intervening unique sequence. Many of the duplications are located within the proximal euchromatic regions of chromosomes. In the case of chromosome 22 (Fig. 1), >90% of the chromosome-specific duplications occur within the first 10 Mb of the long arm of chromosome 22 (Refs 7,21). The duplicated regions can be large – in excess of 400 kb in length – and the organization of the sequences within the blocks of chromosome-specific duplications is complex[25]. In most cases, the blocks are composed of smaller DUPLICONS. These MODULES correspond to fragments of genes or, in some cases, an entire complement of exons and introns that abut other modules of different ancestral origin (Fig. 3). Larger domains composed of multiple modules form the underlying structure of the chromosome-specific duplications. The organization and the distribution of particular modules can vary substantially among the chromosome-specific duplications (Fig. 3). These properties have made mapping and sequencing of these regions particularly problematic[8,26,27].

Chromosome-specific duplications can share a high degree of sequence identity[8]. A global analysis of the human genome shows that the majority of

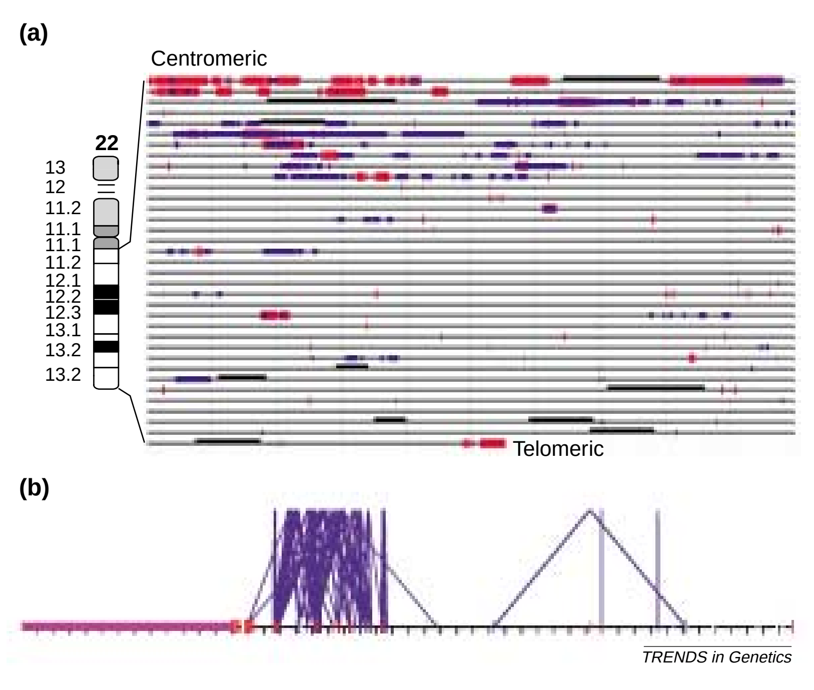

Fig. 1. Recent segmental duplications on chromosome 22q11. (a) An overview of recent (<35 Myr) duplicated sequence on the long (q) arm of chromosome 22 (Ref. 8). A total of 715 sequence alignments, which were >1 kb in length and >90% sequence identity, were identified after common repeats (e.g. Alu elements, LINEs) were removed. Each horizontal line represents 1 Mb. Top left-hand corner is the most centromeric sequence contig and at the bottom right is the most telomeric sequence. Black bars denote the 11 sequence gaps. Red, transchromosomal duplications between nonhomologous chromosomes; blue, intrachromosomal duplications. Overall 9.1% of the q arm is involved in recent large (>1 kb) duplications. Transchromosomal and intrachromosomal duplications constitute 3.9% and 6.4% of the total sequence, respectively. Of the overall duplicated sequence, 5% involves both inter- and intra-chromosomal duplications. This small degree of overlap suggests two distinct classes of duplication. More than 50% of interchromosomal alignments are restricted to the most centromeric 1.5-Mb and to the most telomeric 50-kb regions, suggesting that there is a positional bias for such events. (b) A reduced view showing the pattern of intrachromosomal duplications (blue lines) on chromosome 22. Each black tick denotes 10 Mb of sequence; purple bar, the short arm and centromeric region of chromosome 22, which were not sequenced as part of the Human Genome Project.

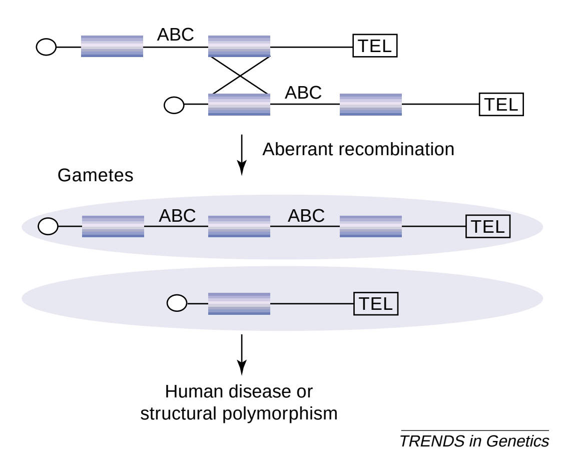

Fig. 2. Paralogous recombination and structural rearrangement. Unequal crossover between large chromosomal duplicons during meiosis can potentiate microduplication and microdeletion of large stretches of genomic DNA. Such events can lead to structural polymorphisms or disease, if the genes (A, B, C) flanked by the duplications are halploinsufficient, triplosensitive or imprinted. Open circle, centromere; TEL, telomere.

intrachromosomally duplicated bases share between 97.5% and 99% sequence identity (Table 1). In many cases, the degree of divergence among chromosome-specific duplications approaches levels of allelic variation (less than one nucleotide difference per kb)[17]. In the few cases where the recombination junctions have been characterized at the molecular level[28,29], rearrangement occurred within tracts of paralogy where perfect sequence identity extends beyond 400 bp – leading to both microdeletion and microduplication. Such molecular specificity for microdeletion and microduplications within paralogous sequence might relate to the minimal efficient processing segment required by the recombination machinery to initiate an unequal crossover[28]. Alternatively, the presence of hyper-recombinogenic sequences could explain the clustering of rearrangement breakpoints[29]. The high degree of sequence identity among the chromosome-specific duplications suggests a recent evolutionary origin. This has been confirmed in many cases by comparative fluorescence *in situ* hybridization (FISH) analysis[30,31].

In other examples, such as the Williams–Beuren syndrome repeat and the velocardio-facial/DiGeorge syndrome repeat[21,32], comparative FISH data show conserved duplication architecture that precedes the MOLECULAR CLOCK estimate based on sequence divergence. For example, the low level of sequence divergence among these duplications (<1.0%) would predict an origin after the separation of humans and chimpanzee lineages, in which comparison of neutral sequences show 1.3% divergence. An examination of primate species by comparative FISH, however, showed the preservation of duplication structure in several species of great ape and Old World monkey. Because the average divergence of neutral noncoding DNA is far in excess of 1% for these species, the analysis suggests that gene conversion events might be partly responsible for maintaining the high degree of sequence identity within the human genome.

Transchromosomal duplications

Segmental duplications distributed among nonhomologous chromosomes define a second class of recent duplications. The most notable property of transchromosomal duplications is their bias to accumulate near heterochromatic DNA – particularly within subtelomeric[13,33–39] and pericentromeric[8–12,14,15,40–46] regions of human chromosomes. Segmental duplications have been identified at the junctions of α-SATELLITE DNA, providing a structural transition between classically defined centromeric DNA and unique DNA[8,42,44]. Similarly, many interchromosomal duplicated blocks of sequence map to within <100 kb of telomeric repeat elements[39,47]. On the basis of the existing assembly of the human genome, there is approximately sevenfold more segmental

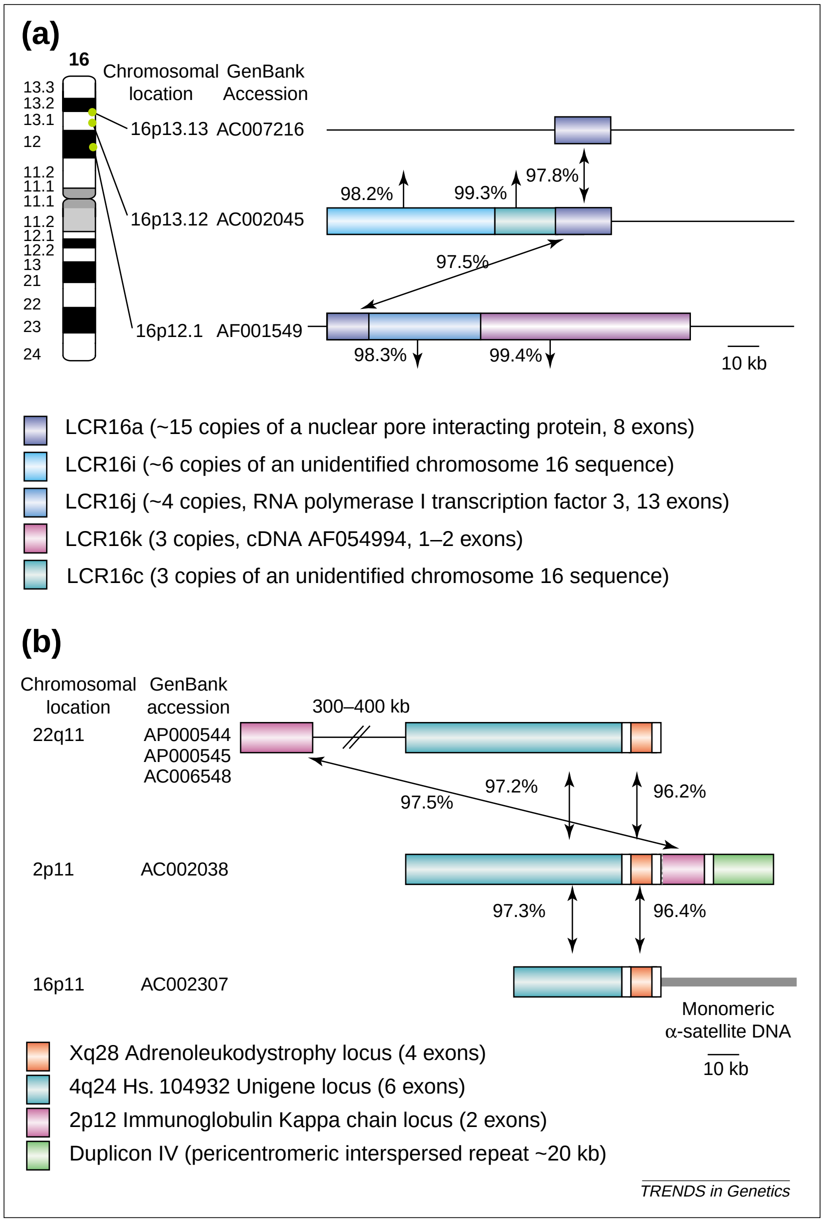
(a)
16
13.3
13.2
13.1
12
11.2
11.1
11.1
11.2
12.1
12.2
13
21
22
23
24
Chromosomal location
GenBank Accession
16p13.13 AC007216
16p13.12 AC002045
16p12.1 AF001549
98.2%
99.3%
97.8%
97.5%
98.3%
99.4%
10 kb
LCR16a (~15 copies of a nuclear pore interacting protein, 8 exons)
LCR16i (~6 copies of an unidentified chromosome 16 sequence)
LCR16j (~4 copies, RNA polymerase I transcription factor 3, 13 exons)
LCR16k (3 copies, cDNA AF054994, 1–2 exons)
LCR16c (3 copies of an unidentified chromosome 16 sequence)
(b)
Chromosomal location
GenBank accession
300–400 kb
22q11
AP000544
AP000545
AC006548
97.5%
97.2%
96.2%
2p11
AC002038
97.3%
96.4%
16p11
AC002307
Monomeric
α-satellite DNA
10 kb
Xq28 Adrenoleukodystrophy locus (4 exons)
4q24 Hs. 104932 Unigene locus (6 exons)
2p12 Immunoglobulin Kappa chain locus (2 exons)
Duplicon IV (pericentromeric interspersed repeat ~20 kb)
TRENDS in Genetics

Fig. 3. Mosaic structure of segmental duplications. (a) The structure of a subset of chromosome-specific duplications for chromosome 16 (LCR16) (Refs 25,57). In two of the three cases shown, the duplications are composed of smaller segments or modules. Several of these paralogous copies are expressed, and they represent recently emerging gene families. The duplications are spread throughout the chromosome (16q22, 16p11, 16p12, 16p13.1 and 16p13.3) in different combinations and copy numbers. The average degree of sequence identity at the genomic level is indicated for each of the segments: double-headed arrows, the percent identity between two specific copies; single-headed arrows, the average percent identity of this segment to all other copies. (b) The shared structure of pericentromeric duplications over a small segment (150 kb) is shown for 22q11, 16p11 and 2p11. The large transchromosomal duplicated blocks are composed of smaller segmental duplications that originate from gene-containing ancestral loci at 4q24, Xq28 and 2p12 (Refs 11,44,45). Arrows show the average degree of genomic sequence identity among these copies. Novel juxtapositions of exon–intron structure are generated. Duplicon IV represents a 20-kb segmental duplication for which no ancestral locus can be determined. At least 40 copies of this interspersed duplication localize exclusively to pericentromeric and subtelomeric regions.

duplications near centromeric and telomeric sequence markers[8]. The most pronounced effect is observed within pericentromeric regions, where an estimated 35% of all interchromosomally duplicated bases (>90% sequence identity and >1 kb in length) reside. Among chromosomes whose sequence is finished (chromosomes 21 and 22), the pericentromeric bias is, at first sight, more dramatic with more than 50% of all transchromosomal duplications localized to these regions. To date, only a subset of human chromosomes has been implicated in pericentromeric and subtelomeric duplications. The working-draft nature of the human genome sequence and the difficulties associated with mapping and sequencing these regions, however, might, at least partly, explain the apparent absence of pericentromeric and subtelomeric duplications for some chromosomes[8,27,47,48].

Detailed compositional analyses of a few pericentromeric and subtelomeric regions reveal a complex pattern of duplication within duplications. For the pericentromeric region of 10q11, 21q11, 2p11 and 16p11, it is apparent that zones of wall-to-wall duplications exist within the most proximal Mb of each of these chromosomal regions[8,12,42,44,45]. The individual duplicons range in size from a few kb to as large as 75 kb in length. These modules are concatenated to form complex arrays virtually devoid of unique DNA. In many cases, the ancestral segments originate from loci near the centromere, and they contain partial or complete gene structure. The organization of these regions is further complicated by duplications of larger segments (composed often of multiple modules) among different pericentromeric regions. These nonhomologous duplications appear to be restricted to particular subsets of human pericentromeric regions. For example, the proximal portions of 2p11, 10q11, 22q11 and 16p11 define one constellation of pericentromeric regions that share hundreds of kb of recently duplicated material[45]. Similar observations have been made for other groups of nonhomologous pericentromeres[46,49,50], as well as some subtelomeric regions[13,33,35,39].

Among pericentromeric regions, a two-step model has been proposed for the formation of the complex repeat structure[11,45,46]. Initially, duplicative transposition spreads the ancestral loci throughout the genome to pericentromeric regions. Subsequent rounds of exchange and duplication of pericentromeric segments follow, creating larger blocks of paralogy with several layers of duplicative history. Phylogenetic and comparative analyses of a few ancestral donor loci support this model. Many of the initial duplications arose from an ancestral loci at a time before the emergence of the human and African ape lineages, but after the separation of the orangutan from the hominoid lineage (5–10 Myr ago)[10,13,44]. The secondary exchanges among pericentromeric regions have been postulated to occur later during the separation of man and the great apes. Quantitative and qualitative differences in the distribution of pericentromeric duplications among humans and great apes support this model of events[10,44,51,52]. Indeed, a genome-wide study of the average degree of sequence identity among all

Table 1. Properties of the different duplication classes

Property	Chromosome-specific duplications	Transchromosomal duplications
Location	Interspersed within euchromatic regions of chromosomes	~Sevenfold bias toward pericentromeric/ subtelomeric locations
Disease	Association with recurrent chromosomal structural rearrangements[a]	No known association
Sequence identity[b]	96.5–97.5%	97.5–99.9%
Progenitor sequence	Ancestral copies not easily discerned	Ancestral copies – within euchromatic regions
Genes	Many examples of functional genes and gene families	Few examples of functional genes

[a]Examples of microdeletions and microduplications mediated through paralogous recombination include velocaridofacial/DiGeorge syndome, Prader-Willi syndrome, neurofibromatosis and spinal muscular atrophy.

[b]Ranges corresponding to modes of sequence identity observed within the October 2000 assembly.

transchromosomal duplications showed sequence divergence of approximately 3% (the mode)[8]. Based on neutral estimates of the molecular clock[53], this would correspond to a time after the Asian and African ape split (<12 Myr ago) (Ref. 54).

Does this suggest a burst of transchromosomal duplications during hominoid evolution? Although these data are intriguing, such an interpretation could be too simplistic. Detailed sequence analysis of several duplicated pericentromeric regions[42,55] provides evidence of more diverged paralogs (as low as 90% sequence identity). This indicates that more ancient events have occurred (>35 Myr ago). Larger deletion events might occur at a higher frequency within these gene-poor regions, effectively reducing the fraction of more divergent paralogs; that is, the probability that duplicated segment becomes deleted increases over time unless selection pressure emerges. Combined with single base-pair changes, the extent of the duplication events could become more difficult to resolve. Furthermore, paralogous segments with the highest degree of sequence identity (>98% sequence identity) might be under-represented within the current assembly of the human genome[8,27] owing to mis-assembly of these regions or the bias in selection against duplicated clones. These two effects would create the impression of a ‘transposition burst’ and not show a decreasing gradient of diverged genomic sequence as a function of time. Finally, among certain subtelomeric and pericentromeric regions polymorphic structural variability and transchromosomal duplication events continue to be documented within the human population[13,56]. This indicates that nonhomologous duplication events are an ongoing phenomenon in our species.

General properties of segmental duplications

A comparison of the transchromosomal and chromosome-specific duplications suggest several general properties:

(1) The basic building block of segmental duplication organization is the module – a minimal evolutionarily shared segment, many of which can be physically identified by sequence comparison with a functional ancestral locus. It should be pointed out that not all segments have gene-related sequences allowing the ancestral copy to be easily identified (Fig. 3).

(2) Segmental duplications are organized in a patchwork fashion in which different modules are concatenated to form larger complex arrays. These larger complex arrays are, in turn, duplicated and can be subjected to secondary rearrangement events. This creates a mosaic architecture of duplications within duplications.

(3) The majority of extant segmental duplications (at least as a fraction of bp) appear to have emerged or have undergone sequence conversion recently during hominoid evolution (within the past 10 Myr).

(4) The distribution pattern of segmental duplications is not random but localizes to specific heterochromatic and euchromatic regions of a subset of human chromosomes.

(5) Many of the segments contain intron–exon structure from ancestral loci that are juxtaposed to other exon-containing segments. Most of the duplications do not encode functional proteins because of apparent truncation of the ancestral gene structure. Transcription, however, has been observed for many of these putative unprocessed pseudogenes, including fusion transcripts between different segments (see below). Transcript expression and the emergence of novel gene families appear to occur more frequently among chromosome-specific duplications within euchromatic regions, as opposed to transchromosomal duplications localized to heterochromatic regions of the genome[25,57].

Implications of recent segmental duplications

In the case of the human genome, it is evident that ‘nature’s experiments’[4] with genome duplication are not limited to polyploidization and tandem duplication events. *In silico* analysis of human genome sequence and FISH analysis of bacterial artificial chromosomes (BACs) containing human sequences[8,26], both predict that >5% of the human genome is composed of recent segmental

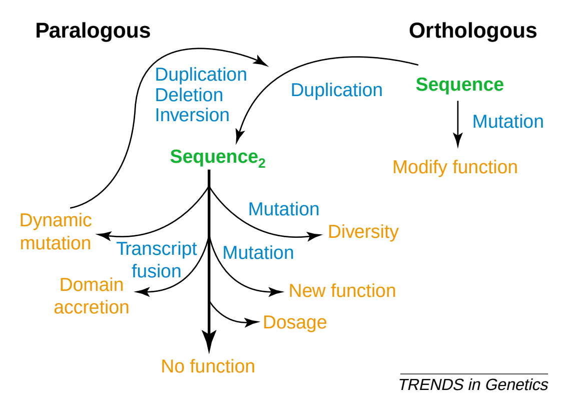

Fig. 4. Consequences of segmental duplication. The 'paralogous pathway' of gene evolution is contrasted with a genomic segment encoding a single-copy gene whose mutational diversity is restricted by selective constraint – the 'ORTHOLOGOUS pathway'. A duplicated genomic segment encoding a gene can lead to increased dosage of a particular product or, through subsequent mutation, to increased diversity in a family of related molecules or eventually to the evolution of a novel function. Mutational drift of the segmental duplication most often renders the duplicated gene nonfunctional. However, the presence of duplicated sequences promotes further structural rearrangements through paralogous recombination and/or to the juxtaposition of different modules leading to the generation of fusion transcripts. Because sequences flanking the duplicated segment can likewise be duplicated, this can lead to cycles of dynamic structural rearrangement at elevated frequency within these regions. These regions of elevated evolutionary turnover contrast with single-copy regions that are limited to a fixed-rate of mutational change under selective constraint. Duplicated genic segments provide many more avenues for innovation over short periods of evolutionary time.

duplications. Although it is too early to conclude whether this phenomenon is a unique property of higher primate genomes, studies of the architecture of invertebrate genomes do not reveal comparable levels of recent segmental duplication[7]. What are the potential ramifications of this unexpected complexity? The evolutionary virtues of sequence duplication have been lauded by scientists for more than 70 years[2,3]. They include an increase in gene dosage, generation of protein diversity and the evolution of new functions (Fig. 4). There is some evidence that the segmental duplications described in this review might have a similar effect, leading to the emergence of new hominoid genes[57]. The presence of widespread segmental duplications, however, has two other important implications that have been previously underestimated: DYNAMIC MUTATION and DOMAIN ACCRETION.

Dynamic structural evolution of chromosomes

The first important impact is exclusively structural in Nature. Both chromosome-specific and transchromosomal duplications increase the likelihood of secondary rearrangements leading to additional inversions, deletions and duplications (Fig. 4). In this regard, segmental duplication can be considered a dynamic process. Dynamic mutations were originally described during the study of microsatellites and triplet repeats as a class of mutations in which an initial event increased the probability of a secondary event[58]. Similarly, the presence of duplicated sequences increases the probability of secondary genomic rearrangement events[59]. The increased probability of secondary events is due to the presence of large blocks of identical or near identical sequence that can provide substrates for subsequent nonhomologous recombination events. This can, in turn, lead to duplication, deletion or inversion of unique flanking sequence (Fig. 2). Unless checked by selective constraint, this could potentiate cycles of segmental duplication. Indeed the mosaic architecture that we have observed for many of the segmental duplications could be the result of multiple rounds of recombination occurring in rapid evolutionary succession.

The many human genomic disorders[16,60] associated with segmental duplication are just one manifestation of such structural dynamism ongoing within the contemporary human species. Based on the combined incidence of known microdeletion and microduplication syndromes, an estimated one in every 1000 human births have duplication-mediated germline rearrangements. It might be argued that events that reduce genetic fitness are likely to be dead ends in terms of chromosomal evolution. However, recently large-scale rearrangements that have no immediate clinical consequence have been documented[61,62]. Duplication-mediated structural polymorphisms have been described ranging from 'small' deletions of 54 kb to inversions of >5 Mb. In each of these examples, a recent segmental duplication was identified at the breakpoints. The structural rearrangements appeared in HARDY–WEINBERG EQUILIBRIUM, at least within specific ethnic groups[61], and the rearrangements involved euchromatic and gene-rich regions of the genome[61,62]. Based on our current understanding of human genome, it is probable that many more large-scale structural polymorphisms will be discovered in the near future.

From an evolutionary perspective, such structural fluidity could provide the underlying mechanism for the construction of speciation barriers by creating regions with an inherent proclivity to rearrange. Rearranged chromosomes that share the identical pattern of rearrangement could occur at an elevated frequency in the population because of recurrent chromosomal structural rearrangements at sites of duplication. Individuals homozygous for such structurally variant chromosomes could, in theory, generate a genetic barrier for chromosomal segregation, creating an impetus for SYMPATRIC speciation. In this regard, it could be noteworthy that a recent sequence comparison between human chromosome 19 and the corresponding mouse chromosomes reveal regional gene-family duplications at ten out of 15 chromosome breakpoints between these two species[63]. These data suggest an association between duplicated regions and sites of chromosomal rearrangement between species.

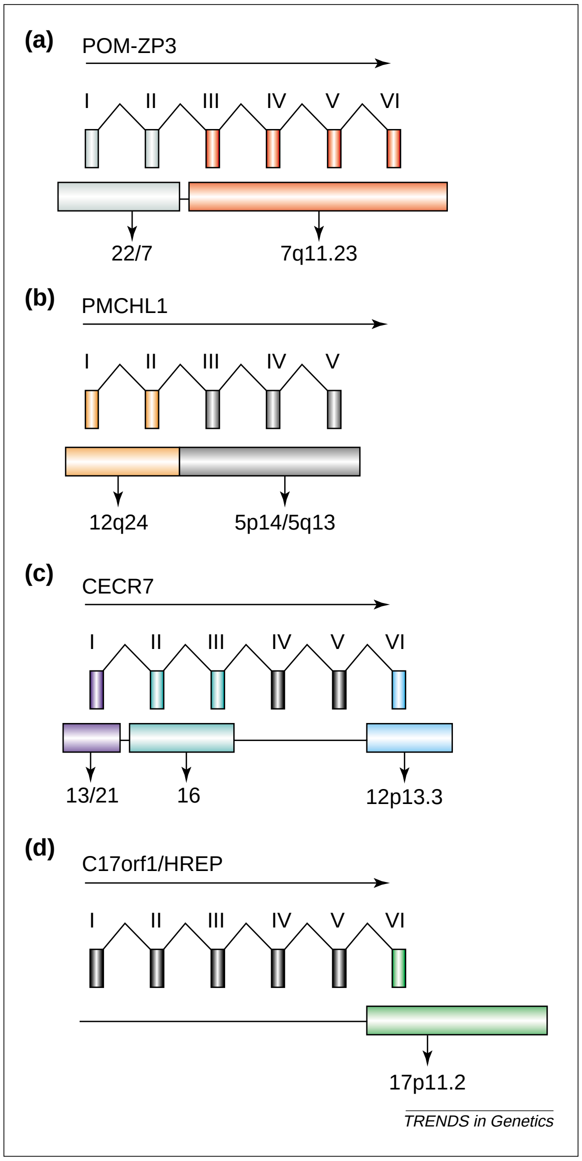

Fig. 5. Segmental duplications and fusion transcripts. Four examples of fusion transcripts that arose as a result of segmental duplication are shown[55,65–68]. (a) POM-ZP3, a 1.6-kb transcript from 7q11.23, consists of a chromosome-specific duplication of the ZP3A locus (zona pellucida glyocprotein gene 3A) juxtaposed to two exons of the POM121 (perinuclear outer membrane) locus. The duplication from ZP3A is believed to have occurred 3–5 Myr ago. Multiple copies of POM121 segmental duplications have been identified on chromosomes 7 and 22. The fusion transcript predicts a protein of 210 amino acids in which the first 76 amino acids are 83% identical to the rat POM121 gene, whereas the remaining 125 amino acids are 98% identical to the ZP3 gene[65]. (b) Mosaic structure of the PMCHL1 (promelanin corticotropin hormone-like 1) gene. The first two exons were recruited through duplication or retrotransposition from 12q24 to 5p14/5q13 before the Old World–hominoid divergence (~25 Myr ago) (Ref. 54). Three additional exons emerged from flanking sequence in a common ancestor of hominoids. A second duplication event generated two copies of this gene (PMCHL1 and PMCHL2) at 5p14 and 5q13 (Ref. 67). (c) The transcript CECR7 (Cat-Eye Syndrome critical region transcript number 7) appears to represent a mosaic of three different transchromosomal duplications from chromosome 16, 12p13.3 and 13/21. Three alternative terminal splice exons have been identified. No significant open reading frame has been detected for this transcript[55]. (d) c17orf1A (chromosome 17 open reading frame 1 A; also known as HREP) encodes a predicted protein of 999 amino acids. Its last exon was contributed by a segmental duplication (CMT1A-Rep), which emerged approximately 6 Myr ago[30] and represents the reverse complement of the last exon of another gene, COX10 (Refs 66,68). This segmental duplication resulted in the modification of the carboxy terminus of a conserved gene (AGIP, ancestral gene before the integration of proximal CMT1A Rep) between man and mouse[68].

Similarly, sequence characterization of the first PARACENTRIC INVERSION breakpoint between human and chimpanzee has identified segmental duplications at the site of the rearrangement[64].

Domain accretion

The process of segmental duplication that we have described is essentially a mechanism of genetic shuffling. It involves the mobilization of small tracts (1–100 kb) of genomic material from one region of the genome to another. These duplicated segments often include sequences with intron–exon structure (Fig. 3). Such duplications occur without the disruption of the ancestral locus. Furthermore, the juxtaposition of different segmental duplications with different exon-containing modules creates the potential for fusion transcripts from unrelated genes (Fig. 5). Recently, several such chimeric transcripts involving endogenous genes and exonic portions of segmental duplications have been described in the human genome[55,65–67]. In the case of the POM-ZP3 transcript, this led to the fusion of two different protein modules[65]. Although the biological significance of these fusion transcripts is unknown, at the least these data suggest that the mosaic genomic structure created by segmental duplications can lead to the formation of chimeric transcripts and proteins.

A major conclusion of both the private and public Human Genome Projects was that the human proteome contains a richer collection of multidomain proteins than sequenced invertebrate genomes[6,7]. A twofold increase in domain accretion has been described for the human when compared with the proteome of fly and invertebrate. Similarly, a tenfold increase in the relative proportion of recent segmental duplication was observed for the human when compared with invertebrate genomes. Could there be a correlation between an increased frequency of segmental duplication events and a proclivity to construct larger, multidomain proteins? In theory, segmental duplications provide an evolutionary vehicle for the mobilization of such protein domains. The dynamic evolutionary turnover within these regions, as demonstrated by comparative analyses of pericentromeric and subtelomeric regions, constantly churns out new juxtapositions of exon–intron modules. Such graveyards of genomic redundancy might occasionally produce a selectively advantageous chimera that results in a new gene innovation. Although it is true that the vast majority of such evolutionary experiments are probably failures at a functional level, the sheer abundance of these events in the past 35 Myr indicates that thousands of duplications and juxtapositions have occurred within the anthropoid lineage alone. Extrapolating the process of segmental duplication back to the emergence of vertebrates could explain much of multidomain diversity, even if the frequency of a successful event is a rare occurrence.

Conclusion

The studies of human genetics and modern molecular evolution have focused on the pattern and nature of changes in regions of conserved structure or function. The study of conservation is intuitive – conserved genes and evolutionary segments represent regions of biological importance. The converse is not necessarily true. Despite the evolutionary importance of conservation, regions of genomic hypervariability could also be structurally and functionally very important in an evolving species. The promise of genome sequence is that it allows an initial assessment of such regions. With respect to understanding mechanisms of genomic duplication, most of the current model has been built upon the study of conserved genic regions (i.e. conservation of gene order at the level of protein similarity). The finding of recent segmental duplications does not detract from the original models of duplication (polyploidization and tandem duplication). It simply adds another dimension of mutational change to the equation of genome evolution. It provides a means for extreme dynamism and genomic fluidity over very short periods of evolutionary time. However, much remains to be understood regarding segmental duplication, including a final resolution of its architecture in the human genome, the mechanism(s) by which it occurs and its occurrence among other vertebrate genomes. A much greater challenge, however, will be the design of methods and techniques to assess the function of these duplications and their products in the absence of a model organism.

Acknowledgements
I thank my colleagues Huntington F. Willard, Terry J. Hassold, Eugene V. Koonin and Joseph H. Nadeau, for helpful discussions in the preparation of this manuscript. I thank Jeffrey A. Bailey for technical assistance. Financial support was obtained from the National Institutes of Health, US Department of Energy and The March of Dimes. I apologize to my colleagues and the authors of relevant papers that could not be cited owing to space limitations.

References

1 Bridges, C. (1935) Salivary chromosome maps. *J. Heredit.* 26, 60–64
2 Muller, H.J. (1936) Bar duplication. *Science* 83, 528–530
3 Ohno, S. *et al.* (1968) Evolution from fish to mammals by gene duplication. *Hereditas* 59, 169–187
4 Ohno, S. (1970) *Evolution by Gene Duplication*, Springer Verlag
5 Muller, H.J. (1925) Why polyploidy is rarer in animals than plants. *Am. Nat.* 59, 346–353
6 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
7 International Human Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–920
8 Bailey, J.A. *et al.* (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* 11, 1005–1017
9 Tomlinson, I.M. *et al.* (1994) Human immunglobulin VH and D segments on chromosomes 15q11.2 and 16p11.2. *Hum. Mol. Genet.* 3, 853–860
10 Eichler, E.E. *et al.* (1996) 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
11 Eichler, E.E. *et al.* (1997) Interchromosomal duplications of the adrenoleukodystrophy locus: a phenomenon of pericentromeric plasticity. *Hum. Mol. Genet.* 6, 991–1002
12 Jackson, M.S. *et al.* (1999) Sequences flanking the centromere of human chromosome 10 are a complex patchwork of arm-specific sequences, stable duplications, and unstable sequences with homologies to telomeric and other centromeric locations. *Hum. Mol. Genet.* 8, 205–215
13 Trask, B. *et al.* (1998) 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
14 Zimonjic, D. *et al.* (1997) Fluorescence *in situ* hybridization analysis of keratinocyte growth factor gene amplification and dispersion in evolution of great apes and humans. *Proc. Natl. Acad. Sci. U. S. A.* 94, 11461–11465
15 Regnier, V. *et al.* (1997) Emergence and scattering of multiple neurofibromatosis (NF1)-related sequences during hominoid evolution suggest a process of pericentromeric interchromosomal transposition. *Hum. Mol. Genet.* 6, 9–16
16 Lupski, J.R. (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* 14, 417–422
17 Lefebvre, S. *et al.* (1995) Identification and characterization of a spinal muscular atrophy-determining gene. *Cell* 80, 155–165
18 Reiter, L. *et al.* (1996) A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element. *Nat. Genet.* 12, 288–297
19 Chen, K. *et al.* (1997) Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome. *Nat. Genet.* 17, 154–163
20 Amos-Landgraf, J.M. *et al.* (1999) Chromosome breakage in the Prader–Willi and Angelman syndromes involves recombination between large, transcribed repeats at proximal and distal breakpoints. *Am. J. Hum. Genet.* 65, 370–386
21 Shaikh, T.H. *et al.* (2000) Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum. Mol. Genet.* 9, 489–501
22 Peoples, R. *et al.* (2000) A physical map, including a BAC/PAC clone contig, of the Williams–Beuren syndrome – deletion region at 7q11.23. *Am. J. Hum. Genet.* 66, 47–68
23 Dorschner, M.O. *et al.* (2000) NF1 microdeletion breakpoints are clustered at flanking repetitive sequences. *Hum. Mol. Genet.* 9, 35–46
24 Saunier, S. *et al.* (2000) Characterization of the NPHP1 locus: mutational mechanism involved in deletions in familial juvenile nephronophthisis. *Am. J. Hum. Genet.* 66, 778–789
25 Loftus, B. *et al.* (1999) Genome duplications and other features in 12 Mbp of DNA sequence from human chromosome 16p and 16q. *Genomics* 60, 295–308
26 Cheung, V.G. *et al.* (2001) Integration of cytogenetic landmarks into the draft sequence of the human genome. The BAC Resource Consortium. *Nature* 409, 953–958
27 Eichler, E.E. (2001) Segmental duplications: what's missing, misassigned, and misassembled – and should we care? *Genome Res.* 11, 653–656
28 Reiter, L.T. *et al.* (1998) 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
29 Lopez-Correa, C. *et al.* (2001) Recombination hotspot in NF1 microdeletion patients. *Hum. Mol. Genet.* 10, 1387–1392
30 Keller, M.P. *et al.* (1999) Molecular evolution of the CMT1A-REP region: a human- and chimpanzee-specific repeat. *Mol. Biol. Evol.* 16, 1019–1026
31 Orti, R. *et al.* (1998) Conservation of pericentromeric duplications of a 200-kb part of the human 21q22.1 region in primates. *Cytogenet. Cell Genet.* 83, 262–265
32 DeSilva, U. *et al.* (1999) Comparative mapping of the region of human chromosome 7 deleted in Williams syndrome. *Genome Res.* 9, 428–436
33 Rouquier, S. *et al.* (1998) Distribution of olfactory receptor genes in the human genome. *Nat. Genet.* 18, 243–250
34 Trask, B. *et al.* (1993) Fluorescence *in situ* hybridization mapping of human chromosome 19: cytogenetic band location of 540 cosmids and 70 genes or DNA markers. *Genomics* 15, 133–145
35 Trask, B.J. *et al.* (1998) Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. *Hum. Mol. Genet.* 7, 2007–2020
36 Brand-Arpon, V. *et al.* (1999) 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
37 Wong, A.C. *et al.* (1999) 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

38 Grewal, P.K. *et al.* (1999) Recent amplification of the human FRG1 gene during primate evolution. *Gene* 227, 79–88

39 van Geel, M. *et al.* (1999) The FSHD region on human chromosome 4q35 contains potential coding regions among pseudogenes and a high density of repeat elements. *Genomics* 61, 55–65

40 Eichler, E. *et al.* (1999) CAGGG repeats and the pericentromeric duplication of the hominoid genome. *Genome Res.* 9, 1048–1058

41 Zimmer, F. *et al.* (1990) Transposition of human immunoglobulin V kappa genes within the same chromosome and the mechanism of their amplification. *EMBO J.* 9, 1535–1542

42 Guy, J. *et al.* (2000) Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10q. *Hum. Mol. Genet.* 9, 2029–2042

43 Potier, M. *et al.* (1998) Two sequence-ready contigs spanning the two copies of a 200-kb duplication on human 21q: partial sequence and polymorphisms. *Genomics* 51, 417–426

44 Horvath, J. *et al.* (2000) Molecular structure and evolution of an alpha/non-alpha satellite junction at 16p11. *Hum. Mol. Genet.* 9, 113–123

45 Horvath, J. *et al.* (2000) The mosaic structure of a 2p11 pericentromeric segment: a strategy for characterizing complex regions of the human genome. *Genome Res.* 10, 839–852

46 Luijten, M. *et al.* (2000) Mechanism of spreading of the highly related neurofibromatosis type 1 (NF1) pseudogenes on chromosomes 2, 14 and 22. *Eur. J. Hum. Genet.* 8, 209–214

47 Riethman, H.C. *et al.* (2001) Integration of telomere sequences with the draft human genome sequence. *Nature* 409, 948–951

48 Eichler, E.E. (1998) Masquerading repeats: paralogous pitfalls of the human genome. *Genome Res.* 8, 758–762

49 Hattori, M. *et al.* (2000) The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* 405, 311–319

50 Chen, H. *et al.* (1999) A testis-specific gene, TPTE, encodes a putative transmembrane tyrosine phosphatase and maps to the pericentromeric region of human chromosomes 21 and 13, and to chromosomes 15, 22, and Y. *Hum. Genet.* 105, 399–409

51 Arnold, N. *et al.* (1995) Comparative mapping of DNA probes derived from the Vk immunoglobulin gene regions on human and great ape chromosomes by fluorescence *in situ* hybridization. *Genomics* 26, 147–156

52 Ruault, M. *et al.* (1999) Juxta-centromeric region of human chromosome 21 is enriched for pseudogenes and gene fragments. *Gene* 239, 55–64

53 Chen, F.C. and Li, W.H. (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* 68, 444–456

54 Goodman, M. (1999) The genomic record of Humankind's evolutionary roots. *Am. J. Hum. Genet.* 64, 31–39

55 Footz, T.K. *et al.* (2001) Analysis of the cat eye syndrome critical region in humans and the region of conserved synteny in mice: a search for candidate genes at or near the human chromosome 22 pericentromere. *Genome Res.* 11, 1053–1070

56 Ritchie, R.J. *et al.* (1998) A large polymorphic repeat in the pericentromeric region of human chromosome 15q contains three partial gene duplications. *Hum. Mol. Genet.* 7, 1253–1260

57 Johnson, M.E. *et al.* Positive selection of a gene family during the emergence of humans and the great apes. *Nature* (in press)

58 Richards, R.I. and Sutherland, G.R. (1992) Dynamic mutations: a new class of mutations causing human disease. *Cell* 70, 709–712

59 Ohta, T. (1989) Role of gene duplication in evolution. *Genome* 31, 304–310

60 Ji, Y. *et al.* (2000) Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Res.* 10, 597–610

61 Sprenger, R. *et al.* (2000) Characterization of the glutathione *S*-transferase GSTT1 deletion: discrimination of all genotypes by polymerase chain reaction indicates a trimodular genotype–phenotype correlation. *Pharmacogenetics* 10, 557–565

62 Giglio, S. *et al.* (2001) Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am. J. Hum. Genet.* 68, 874–883

63 Dehal, P. *et al.* (2001) Human chromosome 19 and related regions in mouse: conservative and lineage specific evolution. *Science* 293, 104–111

64 Nickerson, E. *et al.* Breakpoint analysis of a pericentric inversion distinguishing the human and chimpanzee genomes. *Genome Res.* (in press)

65 Kipersztok, S. *et al.* (1995) POM-ZP3, a bipartite transcript derived from human ZP3 and a POM121 homologue. *Genomics* 25, 354–359

66 Kennerson, M.L. *et al.* (1998) Genomic structure and physical mapping of C17orf1: a gene associated with the proximal element of the CMT1A-REP binary repeat. *Genomics* 53, 110–112

67 Courseaux, A. and Nahon, J.L. (2001) Birth of two chimeric genes in the Hominidae lineage. *Science* 291, 1293–1297

68 Inoue, K. *et al.* (2001) The 1.4-mb CMT1A duplication/HNPP deletion genomic region reveals unique genome architectural features and provides insights into the recent evolution of new genes. *Genome Res.* 11, 1018–1033

69 Ohno, S. (1973) Ancient linkage groups and frozen accidents. *Nature* 244, 259–622

zzzzzzzz
AFTER 2.528.715 A's, T's, G's AND C's IT WAS TOO MUCH. 'THE BOOK OF LIFE' WAS EXTREMELY BORING.

Hanno Bolz is a genetic counsellor and research assistant in the Faculty of Medicine, University of Hamburg, Germany. He recently published a collection of his cartoons in *GenComics* (Wiley-VCH, 2001).