nature genetics

Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution

Zhaoshi Jiang[1], Haixu Tang[2], Mario Ventura[3], Maria Francesca Cardone[3], Tomas Marques-Bonet[1], Xinwei She[1], Pavel A Pevzner[4] & Evan E Eichler[1]

Human segmental duplications are hotspots for nonallelic homologous recombination leading to genomic disorders, copy-number polymorphisms and gene and transcript innovations. The complex structure and history of these regions have precluded a global evolutionary analysis. Combining a modified A-Bruijn graph algorithm with comparative genome sequence data, we identify the origin of 4,692 ancestral duplication loci and use these to cluster 437 complex duplication blocks into 24 distinct groups. The sequence-divergence data between ancestral-derivative pairs and a comparison with the chimpanzee and macaque genome support a 'punctuated' model of evolution. Our analysis reveals that human segmental duplications are frequently organized around 'core' duplicons, which are enriched for transcripts and, in some cases, encode primate-specific genes undergoing positive selection. We hypothesize that the rapid expansion and fixation of some intrachromosomal segmental duplications during great-ape evolution has been due to the selective advantage conferred by these genes and transcripts embedded within these core duplications.

Human duplication architecture differs from other sequenced mammalian genomes in its complexity and in the frequency of large blocks of interspersed duplication[1,2]. Approximately 400 blocks of the human genome have been identified that appear to have been targeted by multiple duplication events during the course of primate evolution[2]. Phylogenetic and comparative analyses of a few of these complex duplication regions indicate a multistep model for their origin[3–12]. These data suggest that a series of independent ancestral loci (duplicons) transposed to specific genomic regions, creating duplication blocks with a mosaic architecture of juxtaposed duplicated segments. Subsequent duplications of larger segments among these blocks resulted in a complex pattern of duplication-within-duplication pairwise alignments (**Figs. 1** and **2**). These properties have considerably complicated ancestral reconstruction, making traditional multiple-sequence alignment approaches impractical. In this study, we aimed to systematically pinpoint the ancestral origin of each human segmental duplication and to organize duplication blocks based on their shared evolutionary history.

RESULTS

Duplication subunit definition

Currently, the dataset of known human segmental duplications is represented by a collection of 28,856 pairwise alignments ($\geq$90% sequence identity, $\geq$1 kb) corresponding to 152.2 Mb of genomic sequence[13,14] (http://www.genome.ucsc.edu). The data do not offer any direct information regarding the directionality of the duplication events or the origin of the ancestral locus. We began by grouping all duplication blocks that shared any sequence homology into 665 bins and constructing a repeat graph for each of these bins using a modified A-Bruijn graph approach[15]. The procedure takes the underlying pairwise alignments within each group as input and threads through each alignment (using RepeatGluer)[15] to define the edges of the repeat graph. The edges of the graph correspond to continuous genomic segments for which no breakpoint exists—these are defined as the duplication subunits—and the vertices correspond to the alignment breakpoints (**Supplementary Fig. 1** online). Using this procedure, we decomposed the 28,856 pairwise alignments into 11,951 nonredundant duplication subunits with a minimum length of 100 bp, representing 97% (147.9 Mb/152.2 Mb) of all duplicated base pairs (**Figs. 1** and **3**). This analysis provided a controlled 'genomic vocabulary' to compare the content and organization of duplication blocks across the genome, identifying all loci with a potentially shared evolutionary history.

Ancestral origin of human segmental duplications

Many mutational processes (such as deletions, duplications and retrotranspositions) will generate breakpoints (vertices) in the repeat graph, leading to over-fragmentation of the ancestral subunits. Thus, the duplication subunit defined by the repeat graph may not correspond to the true extent of the ancestral duplication. To identify the

[1]Department of Genome Sciences, University of Washington School of Medicine and the Howard Hughes Medical Institute, 1705 NE Pacific Street, Seattle, Washington, 98195. [2]School of Informatics and Center for Genomics and Bioinformatics, Indiana University, Bloomington, Indiana 47408. [3]Department of Genetics and Microbiology, University of Bari, 70126 Bari, Italy. [4]Department of Computer Science and Engineering, University of California Santa Cruz, La Jolla, California 92093. Correspondence should be addressed to E.E.E. (eee@gs.washington.edu).

Received 26 April; accepted 7 August; published online 7 October 2007; doi:10.1038/ng.2007.9

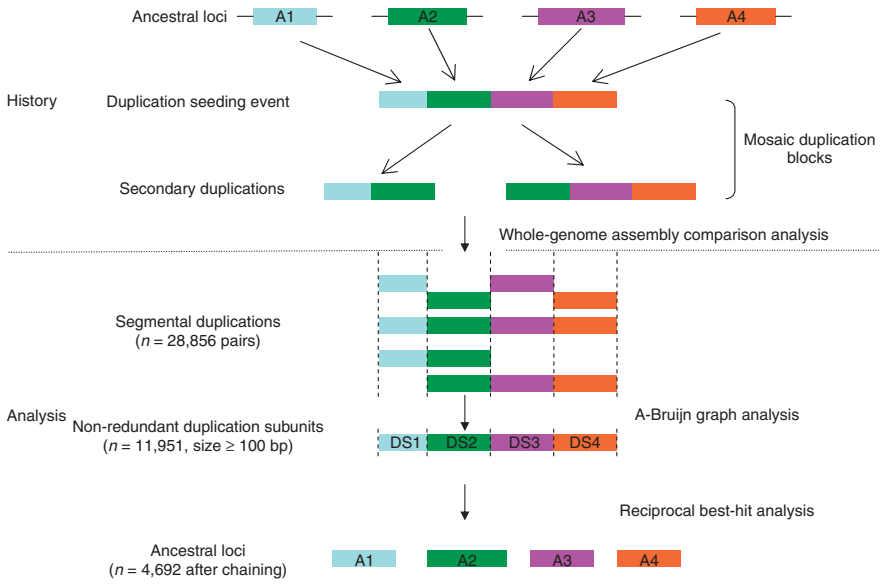

Figure 1 Ancestral-state determination of duplication blocks. The figure schematically illustrates the history of segmental duplications and the computational process of ancestral-state determination. Individual pairwise alignments (WGAC) are decomposed into duplication subunits (for example, DS1 and DS2) by the modified A-Bruijn graph method. Reciprocal comparisons of each human subunit and its flanking sequence to other outgroup mammalian genomes are used to determine the likely ancestral state.

ancestral location of each duplication subunit (duplicon), we took advantage of published genome sequences of outgroup mammalian species (macaque, mouse, rat and dog)[16–19] and the observation that the majority of the segmental duplications emerged recently within human evolution[3,6,10,20–23]. As a result of the multistep process of segmental duplications, an ancestral locus will typically share a larger homologous synteny block in an outgroup species, because synteny extends beyond the boundaries of the duplicated portion (**Fig. 4**). We therefore examined all reciprocal best hits for each duplication subunit using the liftOver program and cross-species chain data, from the University of California Santa Cruz genome browser (http://genome.ucsc.edu)[24]. We defined the human ancestral locus parsimoniously as follows: for any duplication subunit with a given number of copies, the duplicon is defined as the majority-rule reciprocal best hit for all individual human-to-outgroup species comparisons. If we identified more than one locus with an equivalent number of reciprocal best hits, we classified the ancestral state as 'not determined'. After chaining across common repeat sequences, we converted the 11,951 subunits into 4,692 likely

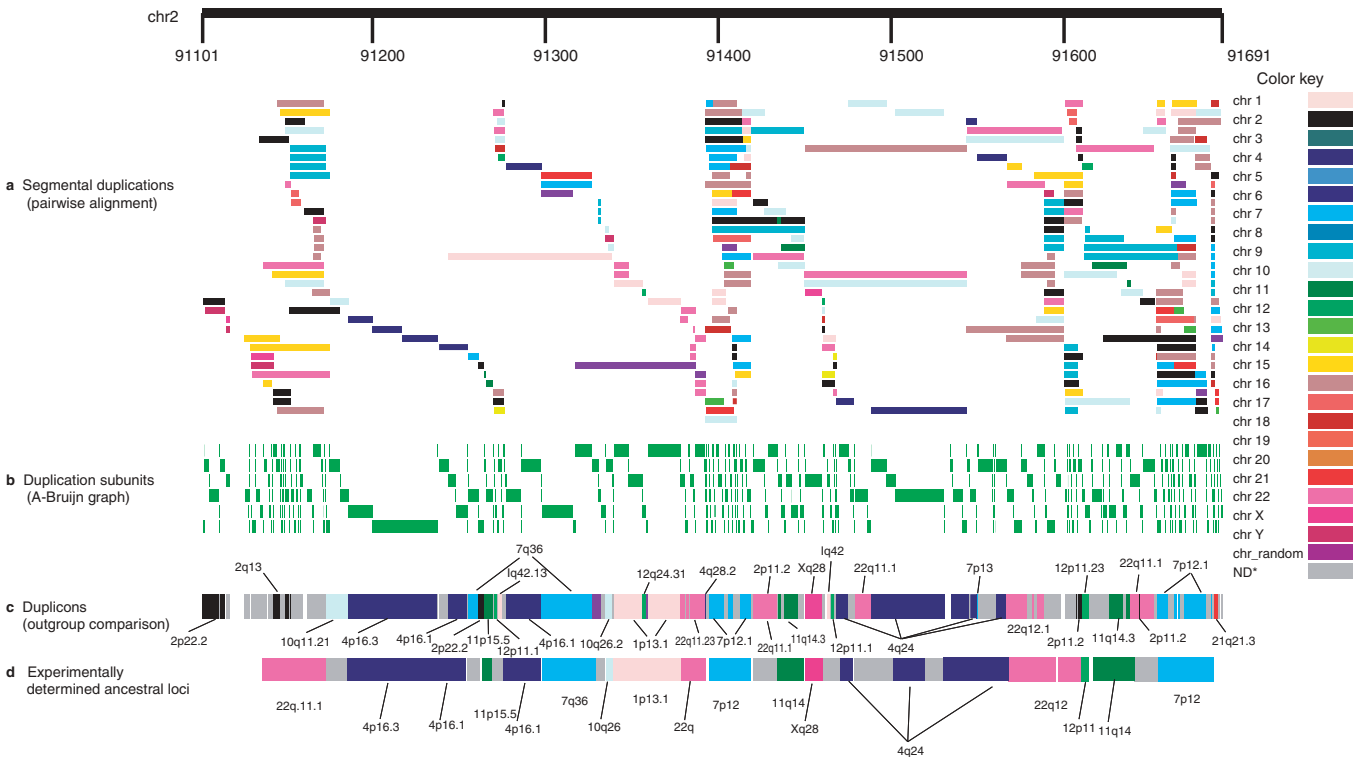

Figure 2 Ancestral-state determination of 2p11 region. Ancestral-duplicon determination for one 750-kb duplication block on human chromosome 2p11. (**a**,**b**) Segmental duplications (multicolor track indicating chromosomal location; (**a**) are converted into a set of nonredundant (**b**) duplication subunits (based on breakpoints in the graph). (**c**) Ancestral duplicons (colored bars) are then predicted based on reciprocal best-hit analysis between human and outgroup mammalian genomes (see Methods). (**d**) These results are compared against experimental results of ancestral duplicons from comparative primate FISH and phylogenetic analyses[10,23]. In this example, 15 of 16 ancestral loci were accurately predicted by the computational method.

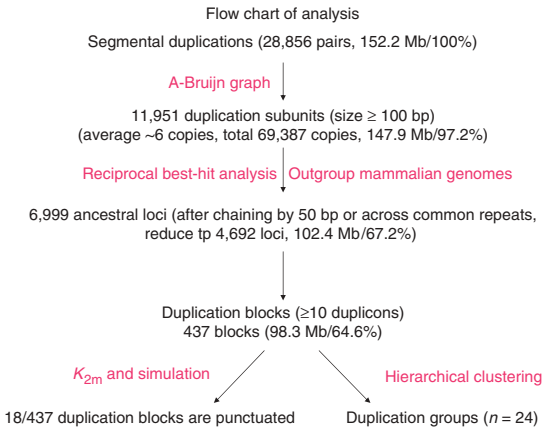

Figure 3 Flowchart of computational analysis. The flow chart summarizes the computational analysis. The number of pairwise alignments, duplication subunits and duplicons identified at each step are indicated.

ancestral loci corresponding to 102.4/152.2 Mb (67.3%) of duplicated base pairs. Of these putative ancestral duplicons, 80% (by base pair) or 65.3% (by count) were supported by at least two or more outgroup species. We also compared those ancestral loci with macaque duplication data (whole-genome shotgun sequence detection data) and found that only ~10% (by count) represent shared duplication between macaque and human. This analysis provided the first genome-wide prediction of ancestral versus derivative duplication loci for the human genome.

We performed two different tests to assess the validity of this approach. First, we tested a subset of larger duplicons (>40 kb) by comparative FISH. In this experiment, we used a fosmid genomic clone corresponding to a derivative locus in human as a probe against a chromosomal metaphase from an outgroup primate species, for example, macaque (**Fig. 5** and **Supplementary Fig. 2** online). If the ancestral locus were correctly identified and the duplication had occurred after the separation of macaque and human, the human probe from the derived locus should hybridize to a single site that is syntenic to the computationally-inferred ancestral site. Nine out of 12 ancestral loci predicted *in silico* were confirmed by FISH (**Supplementary Table 1** online).

Second, we compared the computationally inferred ancestral locus with three published datasets in which ancestral origins had been determined by phylogenetic and comparative sequencing analyses[10,11,25]. **Figure 2** graphically depicts a comparison of one region from 2p11. From our FISH analysis and comparative analysis of the 2p11 region and two other regions (chromosomes 16p12 and 15q11), we determined that 46 of 51 (90%) ancestral loci (duplicons) are consistent between *in silico* prediction and experimental results (**Figs. 2** and **5** and **Supplementary Table 1**). Notably, despite this correspondence, in some cases our method predicted a more refined fragmented substructure than that suggested by the experimental approach. In particular, our method defined 19 additional duplicons corresponding to uncharacterized regions in the original experimental datasets (data not shown). Most of these previously undetected duplicons were relatively short in length (<7 kb) and below the level of resolution typically obtained from cosmid or BAC FISH probes to define ancestral loci.

Temporal and spatial biases of segmental duplications

The delineation between ancestral loci and their duplicates (termed derivative loci) (**Supplementary Table 2** and **Supplementary Table 3** online) provided, for the first time, the opportunity to assess the genome-wide spatial and temporal distribution of historical duplication events. We found that human chromosomes 1q, 7, 9p, 10q, 15q, 16p, 17, 19, 22q, X and Yq are considerably enriched for both ancestral and derivative duplications as a result of extensive intrachromosomal duplication. By contrast, human chromosomes 2, 3p, 4, 5q, 6q, 8q, 12 and 18q have been particularly quiescent. The spectrum of pairwise sequence identity between ancestral and derivative loci (data not shown) confirmed notable differences between inter- and intrachromosomal duplications[1,2]. Our nonredundant analysis based on the ancestral origin showed that intrachromosomal duplication events vastly outpace interchromosomal events when sequence identity thresholds exceed 99%. This excess of intrachromosomal duplication 'seeding events' occurred primarily on a subset of chromosomes during the course of human and great-ape evolution[1,2].

We analyzed all genomic regions that appeared to have been the target of multiple, independent duplication events during the course of human evolution (**Supplementary Table 3**). We identified 437 regions, termed complex duplication blocks, consisting of ten or more duplicons. These complex duplication blocks included almost all

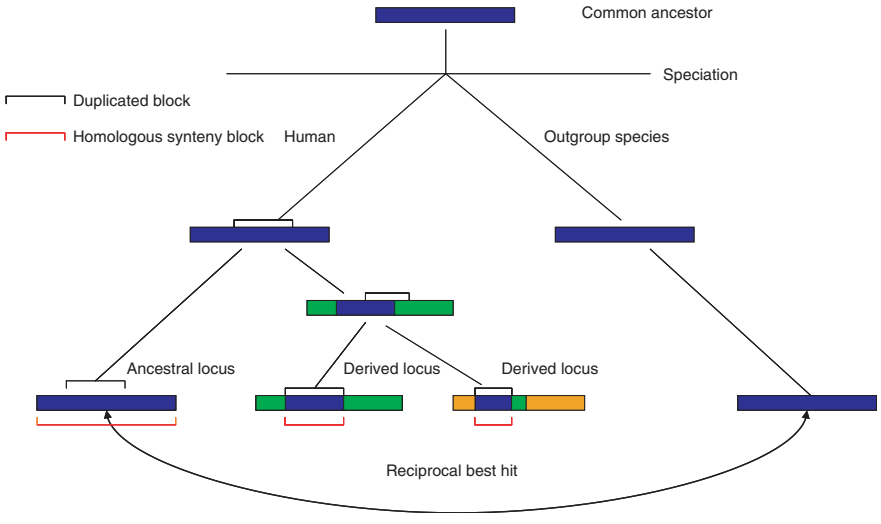

Figure 4 Definition of the ancestral loci by reciprocal best hit. This figure schematically illustrates definition of a primate ancestral locus by reciprocal best-hit analysis between human and an outgroup species. Because of the multistep process of segmental duplications, an ancestral locus will typically share a larger homologous synteny block (highlighted by red brackets) when compared with derivative duplication loci. Consequently, orthologous sequence anchors will extend beyond the boundaries of the duplicated sequence when all human loci are compared with an outgroup genome. The ancestral locus was defined as the reciprocal best-hit locus between the human and outgroup species. We examined the reciprocal best hit for each duplication subunit by using the program of liftOver and cross-species alignment data to distinguish the ancestral loci from their secondary derived loci.

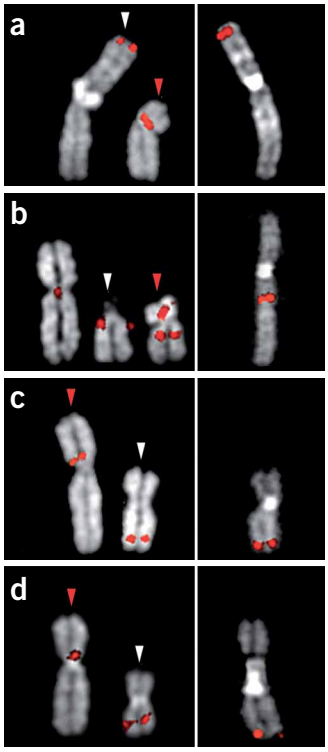

Figure 5 Validation of duplicons by comparative FISH analysis. (**a**–**d**) The figure shows four examples of cross-species FISH validation of ancestral origin of duplications. Human fosmid clones WIBR2-1306D23 (**a**), WIBR2-0929I18 (**b**), WIBR2-1802G13 (**c**) and WIBR2-0996M18 (**d**), corresponding to a predicted derivative duplicated locus (red arrows), were used as FISH probes on metaphase chromosomes from both human (left) and macaque lymphoblastoid cells (right). FISH results from macaque showed a single positive signal corresponding to the syntenic region of ancestral loci predicted by the computational method (white arrows). (See Methods and **Supplementary Table 1**.)

regions associated with recurrent chromosomal structural rearrangements and diseases. As a surrogate for evolutionary age, pairwise sequence divergence (K_{2m}) was computed between ancestral and derivative loci as a function of duplication block location in the genome (**Fig. 6**). Visual inspection revealed a nonuniform distribution of divergence for specific blocks and for specific chromosomes. To confirm whether this nonuniform distribution differed substantively from a random-distribution model, we carried out a simulation (see Methods) for each of the 437 duplication blocks, controlling for potential artifacts such as over-fragmentation of ancestral loci and tandem or redundant duplications. For each duplication block composed of a number of duplicons (N), we computed pairwise sequence divergence (K_{2m}) for each ancestral-derivative duplication pair and then calculated the mean K_{2m} and associated variance for all pairs within a block. We randomly selected the same number (N) of ancestral-derivative duplication pairs from the whole-genome K_{2m} dataset and computed the mean K_{2m} from those random pairs (10,000 replicates). From this distribution of simulated means, we determined an empirical P value based on the number of replicates that were greater or lower than the mean of the simulated data (one-tailed test). Based on conservative criteria (Bonferroni correction for multiple tests), we identified 18 duplication blocks where K_{2m} is significantly lower than the expected average from simulated data ($P < 0.0001$). When we extended our analysis to different interrelated groups of duplication blocks, we found evidence of clustered sequence divergence for 9 of the 24 groups of duplication blocks. Eight of these nine were composed mainly of intrachromosomal segmental duplications.

Reconstructing the evolutionary relationship of duplicated sequences on the basis of nucleotide divergence is complicated by processes such as gene conversion, which homogenizes paralogous sequences, resulting in an underestimate of the age of a duplication event[26–28]. To assess the effect of gene conversion, we independently estimated the timing of each duplication subunit by carrying out a genome-wide analysis of segmental duplications for both the macaque[16] and the chimpanzee genomes[29]. We categorized all human duplication subunits on the basis of their duplication status within these three species and compared each category to the distribution of sequence identity alignments in human. This three-way comparison indicates an excess of high-sequence-identity (>98%) duplications that are specific to the human lineage. Of the duplication subunits that have emerged since the divergence of human and chimpanzee, 68.7% by base pair (7.66 of 11.15 Mb) and 71.3% by count are intrachromosomal in origin. The most notable effect is observed for the 18 duplication blocks identified as recently punctuated, where the majority of base pairs (93.7%) have emerged since the divergence of human and chimpanzee from the macaque lineage (~25 million years ago).

Clustering complex duplication blocks identifies cores

In the human genome, duplicons are organized into larger duplication blocks spanning hundreds of kilobases (**Fig. 2**) where different duplicons are arranged as mosaic structures. Numerous studies have shown that secondary duplication events have occurred among the blocks, with more recent events sharing larger duplications and, consequently, more duplicons in common. This complex interrelationship complicates traditional phylogenetic analyses to reconstruct the evolution of these regions, but we took advantage of this property to cluster duplication blocks based on their ancestral duplicon content and thus avoid complications arising from sequence homogenization and phylogenetic distance–based estimates. We examined all regions with at least ten duplicons to provide sufficient phylogenetic signal and generated phylogenetic profiles depending on the presence or absence of ancestral subunits (see Methods). A genome-wide hierarchical clustering tree of those blocks was constructed based on the presence of shared duplicons (phylogenetic profiles)[30–32]. The tree was constructed with each terminal node in the tree representing a complex duplication block. In total, we clustered the 437 duplication blocks into 24 distinct duplication groups (**Fig. 7a**), revealing a complex yet decipherable higher-order hierarchical architecture.

An examination of the chromosomal distribution of the duplication groups distinguishes two different categories, those in which duplication blocks ($n = 10$) are distributed among multiple nonhomologous chromosomes (mixed groups) and those that are primarily restricted to a specific chromosome ($n = 14$) (intrachromosomal, **Fig. 7a** and **Supplementary Table 4** online). We analyzed the structural relationship of the duplication blocks within each group and identified 'core' duplicons shared by the majority of blocks within a specific group. Core duplicons were defined structurally as ancestral duplicons that were represented in more than >67% of the blocks within a given clade (upper 10% of core indices corresponding to 2.2 Mb of human genome sequence). By this definition, we identified a total of 14 core duplicons. The remaining ancestral duplicons were designated as non-core duplicons (96.1 Mb). A comparison of the structure of duplication blocks within specific chromosomes showed that the core duplicons pinpoint the focal point of each duplication group

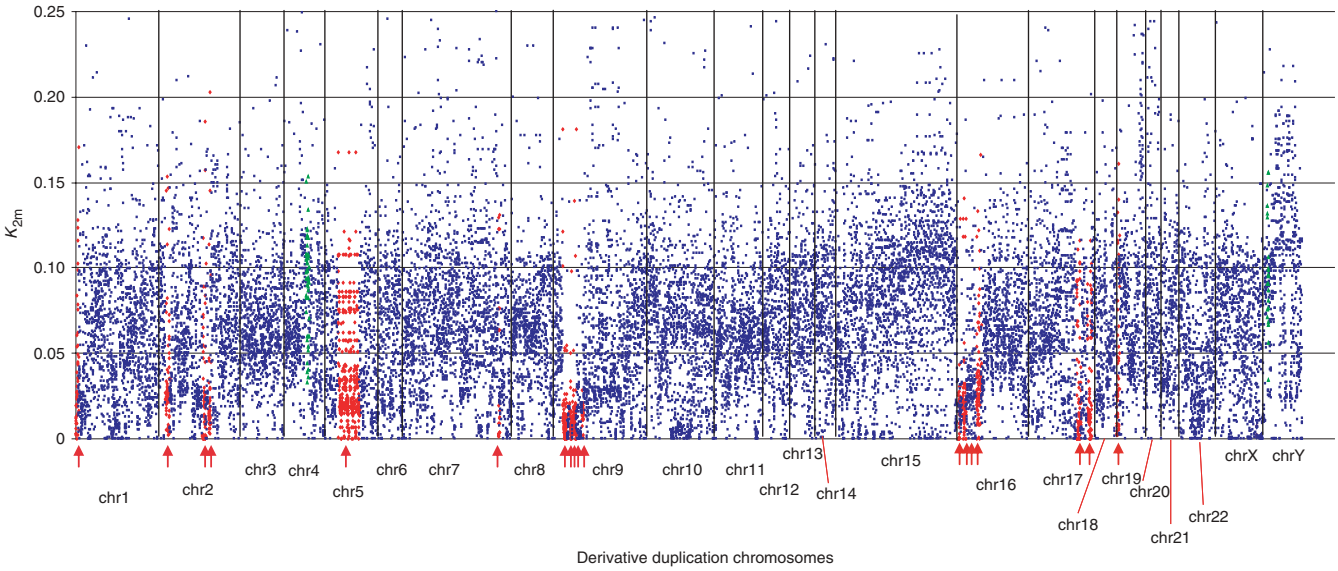

Figure 6 Nonrandom distribution of sequence divergence. The distribution of sequence divergence between ancestral and derivative loci is shown as a function of the location of duplication blocks in the human genome. For each duplication block, we tested the significance of divergence clustering by developing a random simulation model (see Methods). We found 20 of 437 duplication blocks that significantly depart from a continuous genomic duplication model. Eighteen blocks have an excess of low K_{2m} values (suggesting a preponderance of evolutionary younger events; red, $P < 0.0001$), and two duplication blocks show a significant enrichment of higher K_{2m} ancestor-derivative values (suggestive that duplication activity occurred and then ceased; green, $P < 0.0001$). The effect predominates for particular chromosomes (for example, chr2, chr4, chr5, chr9, chr16 and chrY).

architecture (**Fig. 7b** and **Supplementary Fig. 3a**,**b** online), with flanking duplicons showing decreasing copy number and sequence divergence (data not shown).

We also compared the core and non-core duplications between human and chimpanzee. We found that the cores represent regions of shared duplication among human and chimpanzee, whereas the flanking duplicons are much more likely to represent more recent and human-specific events (**Supplementary Fig. 4** online).

We compared the gene and spliced EST content between core and remaining non-core duplicons by measuring the density of exons (number of exons per megabase). We observed a significant ($P < 0.001$) twofold excess of both RefSeq genes and spliced ESTs in core duplicons when compared with non-core regions (**Table 1**) after controlling for sequence redundancy. The RefSeq gene density of core duplicons (74.2 exons per Mb) was higher than the unique nonduplicated regions of the genome (63.0 exons per Mb), whereas the non-core regions were greatly depleted (35.1 exons per Mb). In contrast, core duplicons are substantially enriched for spliced ESTs (1,095.4 exons per Mb), even when compared to unique regions of the genome (383.0 exons per Mb). Previous studies[13,33] have noted that duplicated regions were frequently gene and transcript rich. This analysis, however, restricts the bulk of that enrichment to a small portion (2.2/98.3 Mb = 2.2%) of core segmental duplications.

We note that in 4 of 14 cases, there is compelling evidence that the genes embedded within the cores are associated with previously unknown human gene innovations (**Supplementary Table 5** online). In two cases, the core duplicon has been part of fusion genes whose functions seem to be notably different from those of their antecedents (for example, the USP6 (also known as TRE2) and NBPF11 gene families on chromosomes 17 and 1, respectively)[34,35]. In the case of NBPF11 (also known as DUF1220) and two other cases (for example, NPIP and RANBP2 gene families) there has been evidence of substantial, if not extreme, positive selection occurring in conjunction with the expansion of novel gene families embedded within the core[8,36]. In these two cases, the propagation of these core duplications corresponds precisely with duplication blocks that show evidence of 'punctuated evolution' (**Fig. 6**).

DISCUSSION

In summary, we have developed a global framework to reconstruct the evolutionary history of human segmental duplications. We systematically codified its architecture, identified the ancestral origin (corresponding to 67.3% of duplicated base pairs) and provided a higher-order hierarchical structure for block relationships. These data provide a biologically relevant vocabulary to examine the composition, structural variation and evolution of segmental duplications as other primate genomes become sequenced to a high-quality standard. Two features are noteworthy. First, our analysis shows that a subset of duplication blocks have significant, nonrandom clustering of sequence divergences with respect to ancestral duplicons. This suggests temporal biases in duplication activity for specific regions of our genome. Most models of genomic duplication, in contrast, suggest a continuous model for much smaller, more recent duplications[37,38].

Second, a large fraction of the recent duplication architecture centers around a rather small subset of core duplicons. These cores are focal points of human gene and transcript innovations. Both effects predominate among intrachromosomal duplicates that have expanded during human and great-ape evolution[1,2]. We showed that these regions are enriched for genes and transcripts when compared to non-core duplicons. Several of these gene families show evidence of substantial, and in some cases extreme, positive selection. The data indicate that these core regions have led to the emergence of new gene families that are either unique to hominoids or considerably diverged when compared with other mammalian species[8,34–36]. The function of these core gene families is largely unknown.

The organization and structure of these core elements with respect to flanking duplicons suggest that they may have driven the evolution

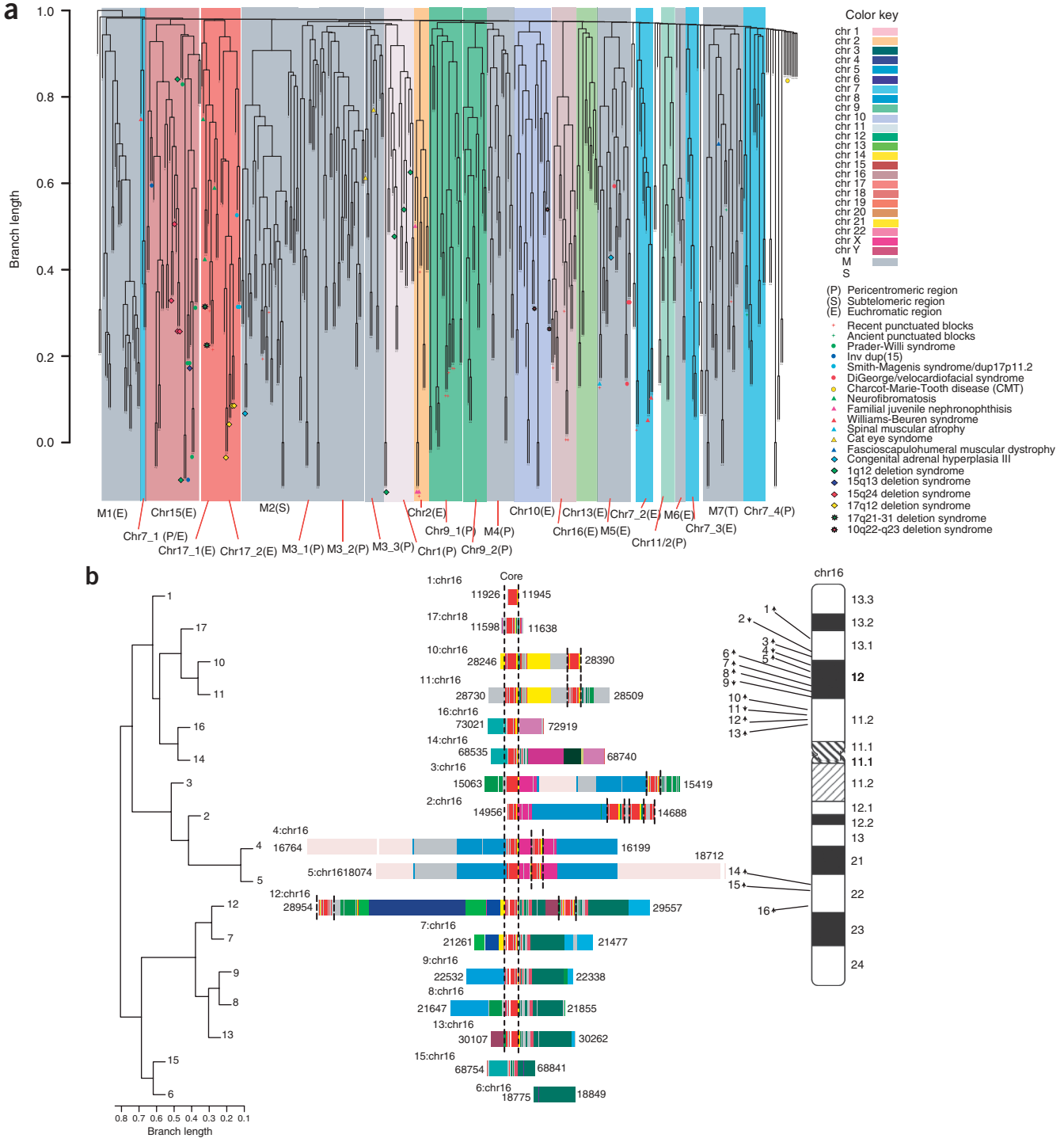

Figure 7 Genome-wide hierarchical clustering of duplication blocks and core structure. (**a**) Complex duplication blocks with more than ten duplicons are clustered on the basis of the similarity of their phylogenetic profile. Each of the 437 duplication blocks is represented as a terminus in the tree (see Methods; **Supplementary Table 5**). Twenty-four duplication groups or clades can be distinguished. There are 14 intrachromosomal duplication groups (color) and 10 mixed clades ('M') that have blocks among multiple nonhomologous chromosomes (gray). Letter designations further define blocks: 'P' blocks are primarily pericentromeric and map within 5 Mb of the centromere, 'S' blocks are primarily subtelomeric and map within 500 kb of the telomere and 'E' duplication blocks map mainly from euchromatic regions. Duplication blocks associated with genomic disorders and that show evidence of punctuated duplication are indicated according to the key. (**b**) Depiction of the mosaic structure of complex duplication blocks for chromosome 16. The duplication blocks were numbered according to genomic location of a locus in the chromosome. All block coordinates are in kb. Different colors denote distinct ancestral loci. A 'core element' shared by a majority of the blocks is highlighted by vertical dash lines. The branch length indicates the percentage difference between pairwise complex duplication blocks (terminus) based on shared duplicon content.

Table 1 Exon density of core versus non-core segmental duplications

		Exon density (exon/Mb)		Core/non-core ratio	
		Core	Non-core	Ratio	*P* value
RefSeq	All transcripts	74.2	35.1	2.1	<0.001
	Fusion transcripts	40.1	17.4	2.3	<0.001
EST	All transcripts	1095.4	489.2	2.2	<0.001
	Fusion transcripts	332.3	77.2	4.3	<0.001

RefSeq genes and ESTs were assigned to either core or non-core duplicons based on the best alignment score. Each transcript was assigned once and only once to the genome. Alternative splice variants were eliminated by clustering exons that overlapped. The exon density (number of exons per Mb) was computed on the basis of the number of exons and the length of duplication region that contained those exons. The exon density of core was found to be significantly higher than that of non-core for both RefSeq genes and ESTs. Fusion transcripts were defined where different exons within the same transcript mapped to distinct duplicons that mapped to different chromosomes or were separated by more than one Mb. We calculated the significance by randomly assigning core and non-core regions to the duplicated genomic regions and computed the exon density of Refseq and EST. The simulation shows that exon enrichment within the core is highly significant.

of the duplication blocks accounting for, in part, the surge of human and great-ape intrachromosomal segmental duplications. When compared with other sequenced genomes, one of the unique aspects of hominoid genome architecture is the abundance of large, highly identical segmental duplications that are interspersed throughout the genome[2,14,29]. It is paradoxical that a large number of such events are predicted to have occurred and have been fixed recently during evolution, despite their known association with diseases and genome rearrangements. We speculate that the formation of a novel gene within a genomically unstable core region was a predisposing event, and that subsequent duplications and selections have operated in a coordinated fashion to increase copy number of these fusion genes within different genomic environments.

METHODS

Ancestral duplication definition. Currently, the dataset of known human segmental duplications is represented by a set of 28,856 pairwise alignments[13,23] (http://www.genome.ucsc.edu). Using these underlying pairwise alignments, we constructed an A-Bruijn graph (Using RepeatGluer)[15] based on a partitioned set of these pairwise alignments (**Figs. 1** and **2** and **Supplementary Note** online). The sequence segment between two adjacent vertexes (the edge of an A-Bruijn graph) is defined as a duplication subunit. We considered all breakpoints with a predefined resolution parameter (*girth* = 25 bp) and generated 15,548 distinct duplication subunits. To establish synteny between human and outgroup species, we limited our subsequent analyses to duplication subunits ($n = 11{,}951$) of sufficient length ($\geq$100 bp), which corresponds to 97.2% of all segmental duplications in the human genome. The most likely ancestral origin of the duplication subunit was determined based on reciprocal best hit for each duplication subunit using the program liftOver and the cross-species chain (BLASTZ whole-genome alignments between human and mouse, rat, dog and macaque) data from University of California Santa Cruz genome browser (http://www.genome.ucsc.edu). We validated a subset of larger duplicons (>40 kb) by comparative FISH analysis (**Fig. 5** and **Supplementary Table 1**). When we compared these results to three sets of experimentally characterized ancestral duplicons[10,11,23,39,40], we found an excellent correspondence (~90%) between predicted and experimentally validated ancestral duplicons.

Duplication divergence and simulation analyses. Sequence divergence between ancestral-derivative duplication pairs was computed using Kimura's two-parameter model[41]. For any given duplication block or clade, we tested whether the observed distribution differed significantly from a random distribution model. For a specific duplication block composed of a certain number (N) of duplicon subunits, we first computed pairwise sequence divergence (K_{2m}) for each ancestral-derivative pair and then calculated the mean K_{2m} and associated variance for all pairs within a block. We randomly selected the same number (N) of ancestral-derivative pairs from the whole-genome K_{2m} dataset and computed the mean K_{2m} from those random pairs (10,000 replicates). Based on this distribution of simulated means, we determined an empirical *P* value based on the number of replicates that were greater or lower than the mean of the simulated data (one-tailed test). Using a Bonferroni correction for multiple testing, we applied a strict threshold for significance ($P < 0.0001$). To eliminate potential artifacts, collinear duplication pairs and local tandem duplications were only counted once in this analysis. Similarly, we repeated the analysis at the level of duplication groups where each ancestral duplicon is only counted once during the simulation. Ten out of 24 duplication groups showed evidence of punctuated duplications of sequence divergence.

Hierarchical clustering of duplication blocks. A binary phylogenetic profile was constructed based on the extent of shared duplicon content for each duplication block composed of ten or more duplicons. If a duplicon was present within a duplication block, we assigned a '1', and if otherwise, we assigned a '0,' generating a binary phylogenetic profile for each block. Complex duplication blocks were then clustered into duplication groups by hierarchical clustering on the basis of the similarity of their phylogenetic profiles[30–32]. A chromosome name was assigned to a clade if the majority of blocks (>50%) in that clade belonged to a homologous chromosome; otherwise the clade was designated as a mixed (M) clade (**Fig. 7a**).

Core duplicon definition. We identified core duplicons as an ancestral duplicon or a series of adjacent ancestral duplicons where subunits are shared by the majority (~67%) of the members of a group (**Fig. 7a,b** and **Supplementary Fig. 3a,b**). For every duplicon, we calculated a core index, $C_i = N_s/N_t$, where N_s is the number of duplication blocks that contain that subunit and N_t is the total number of duplication blocks within a group. For all duplicons, we determined the mean core index ($C_i = 0.40 \pm 0.18$; median = 0.38). A threshold of 0.67 (top 10% values for the core index) was selected to distinguish cores ($C_i = 0.67 \sim 1$) from non-core duplicons ($C_i < 0.67$) (**Fig. 7a,b** and **Supplementary Fig. 3a,b**).

Validation of ancestral duplicons. To validate the origin of the computationally inferred ancestral locus (duplicon), we selected a subset of independent duplication subunits ($n = 12$) for confirmation by comparative FISH. A human probe (fosmid) corresponding to the derivative locus was purified and used as a FISH probe against metaphase chromosomal spreads from human lymphoblast and macaque lymphoblastoid cell lines. Results were classified into three categories: confirmed, in which the FISH result in macaque showed a single positive signal corresponding to the syntenic region predicted by the computational analysis; ambiguous, in which the FISH result showed multiple signals in macaque including the predicted ancestral locus; and not confirmed, in which the FISH result mapped to a cytogenetic band position that did not correspond to the predicted locus. We found that 9 of 12 predicted duplicons were confirmed, 2 were ambiguous and 1 was not confirmed (**Supplementary Table 1**).

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

ACKNOWLEDGMENTS

We thank P. Green, J. Felsenstein, T. Newman, C. Alkan and Z. Bao for useful comments and valuable discussions in the preparation of this manuscript, and E. Tüzün and Z. Cheng for computational assistance. This work was supported by a US National Institutes of Health grant GM58815 to E.E.E. and a Rosetta Inpharmatics fellowship (Merck Laboratories) to Z.J. T.M.-B. is a research fellow supported by Departament d'Educació i Universitats de la Generalitat de Catalunya. E.E.E. is an investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

Z.J. performed the analyses and drafted the manuscript. H.T. implemented the program package and performed part of the analyses. M.V. and M.F.C. performed the FISH validation experiment. T.M.-B. performed the positive selection analysis on the core genes. X.S. was involved in part of the fusion gene analysis. P.A.P. and E.E.E. designed the study, and E.E.E. finalized the manuscript.

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

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. 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).
3. Eichler, E.E. *et al.* Interchromosomal duplications of the adrenoleukodystrophy locus: a phenomenon of pericentromeric plasticity. *Hum. Mol. Genet.* **6**, 991–1002 (1997).
4. Orti, R. *et al.* Conservation of pericentromeric duplications of a 200-kb part of the human 21q22.1 region in primates. *Cytogenet. Cell Genet.* **83**, 262–265 (1998).
5. Jackson, M.S. *et al.* 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 (1999).
6. 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).
7. Horvath, J. *et al.* Molecular structure and evolution of an alpha/non-alpha satellite junction at 16p11. *Hum. Mol. Genet.* **9**, 113–123 (2000).
8. Johnson, M.E. *et al.* Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519 (2001).
9. Stankiewicz, P. & Lupski, J.R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**, 74–82 (2002).
10. Horvath, J.E. *et al.* Punctuated duplication seeding events during the evolution of human chromosome 2p11. *Genome Res.* **15**, 914–927 (2005).
11. Locke, D.P. *et al.* Molecular evolution of the human chromosome 15 pericentromeric region. *Cytogenet. Genome Res.* **108**, 73–82 (2005).
12. Linardopoulou, E.V. *et al.* Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* **437**, 94–100 (2005).
13. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
14. She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927–930 (2004).
15. Pevzner, P.A., Tang, H. & Tesler, G. De novo repeat classification and fragment assembly. *Genome Res.* **14**, 1786–1796 (2004).
16. Gibbs, R.A. *et al.* Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222–234 (2007).
17. Waterston, R. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
18. Gibbs, R.A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
19. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
20. Eichler, E.E. *et al.* Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. *Hum. Mol. Genet.* **5**, 899–912 (1996).
21. Regnier, V. *et al.* 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 (1997).
22. Potier, M. *et al.* Two sequence-ready contigs spanning the two copies of a 200-kb duplication on human 21q: partial sequence and polymorphisms. *Genomics* **51**, 417–426 (1998).
23. She, X. *et al.* The structure and evolution of centromeric transition regions within the human genome. *Nature* **430**, 857–864 (2004).
24. 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).
25. Eichler, E.E. *et al.* Divergent origins and concerted expansion of two segmental duplications on chromosome 16. *J. Hered.* **92**, 462–468 (2001).
26. Jackson, M.S. *et al.* Evidence for widespread reticulate evolution within human duplicons. *Am. J. Hum. Genet.* **77**, 824–840 (2005).
27. Hurles, M.E. Gene conversion homogenizes the CMT1A paralogous repeats. *BMC Genomics* **2**, 11 (2001).
28. 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).
29. Cheng, Z. *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**, 88–93 (2005).
30. Bowers, P.M., Cokus, S.J., Eisenberg, D. & Yeates, T.O. Use of logic relationships to decipher protein network organization. *Science* **306**, 2246–2249 (2004).
31. Rivera, M.C. & Lake, J.A. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* **431**, 152–155 (2004).
32. Lake, J.A. & Rivera, M.C. Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Mol. Biol. Evol.* **21**, 681–690 (2004).
33. Rual, J.F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
34. 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).
35. Vandepoele, K., Van Roy, N., Staes, K., Speleman, F. & van Roy, F. A novel gene family NBPF: intricate structure generated by gene duplications during primate evolution. *Mol. Biol. Evol.* **22**, 2265–2274 (2005).
36. Ciccarelli, F.D. *et al.* Complex genomic rearrangements lead to novel primate gene function. *Genome Res.* **15**, 343–351 (2005).
37. Gu, X., Wang, Y. & Gu, J. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat. Genet.* **31**, 205–209 (2002).
38. Lynch, M. & Conery, J.S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
39. Horvath, J.E. *et al.* Using a pericentromeric interspersed repeat to recapitulate the phylogeny and expansion of human centromeric segmental duplications. *Mol. Biol. Evol.* **20**, 1463–1479 (2003).
40. Johnson, M.E. *et al.* Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proc. Natl. Acad. Sci. USA* **103**, 17626–17631 (2006).
41. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980).