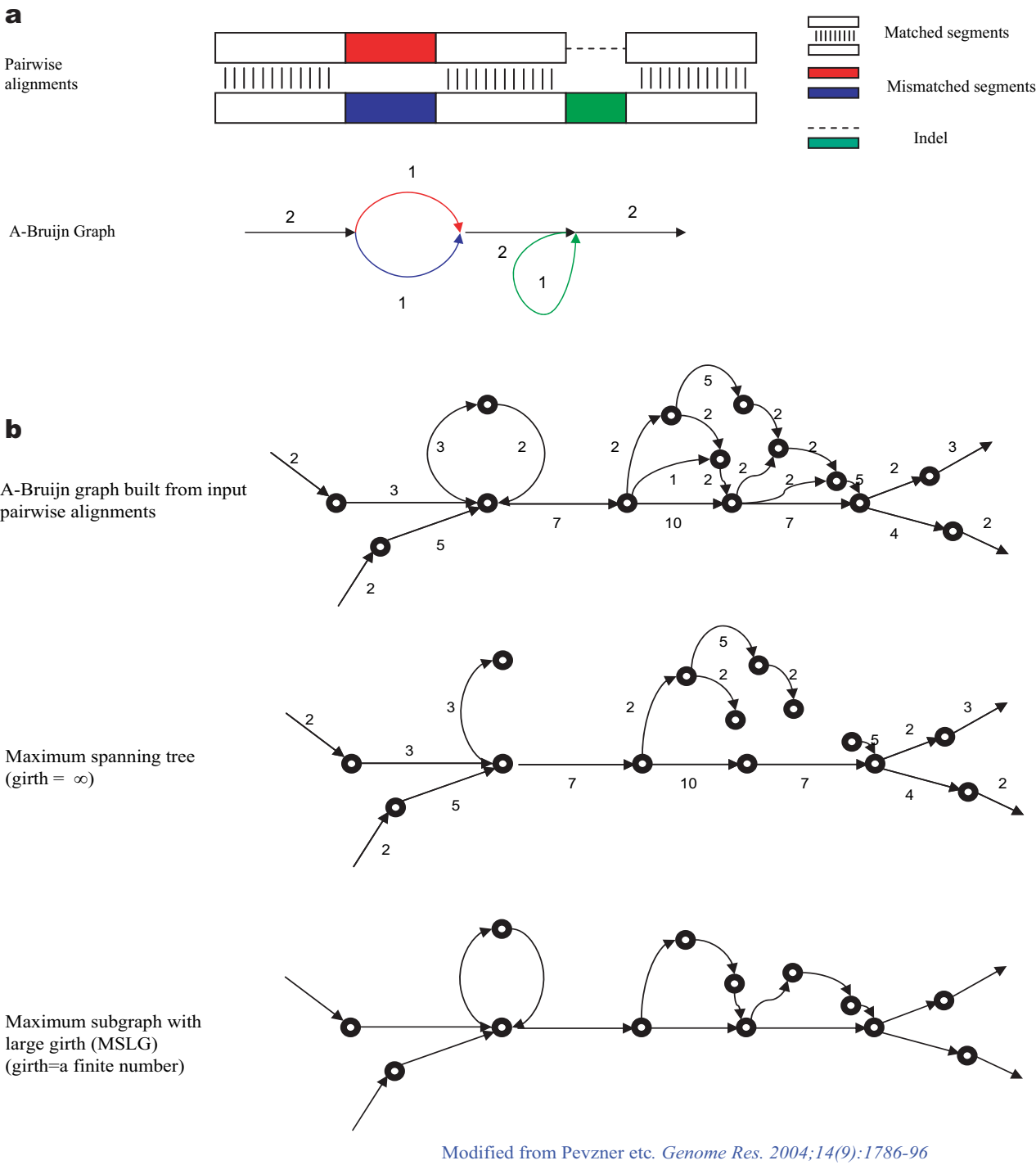
Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution

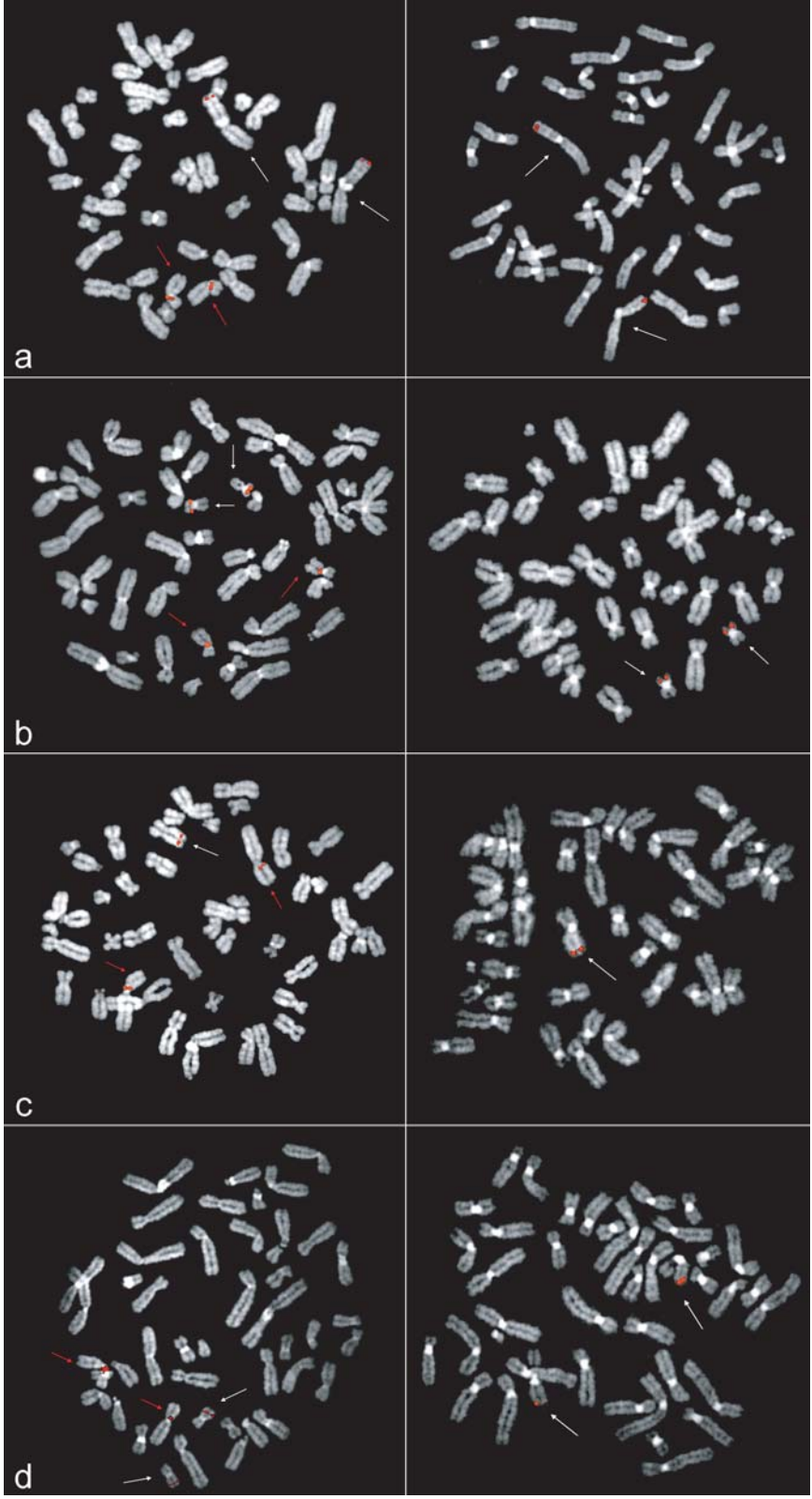
Zhaoshi Jiang[1], Haixu Tang[2], Mario Ventura[3], Maria Francesca Cardone[3], Tomas Marques-Bonet[1], Xinwei She[1], Pavel A. Pevzner[4], Evan E. Eichler[1,5]†.



Modified from Pevzner etc. *Genome Res. 2004;14(9):1786-96*

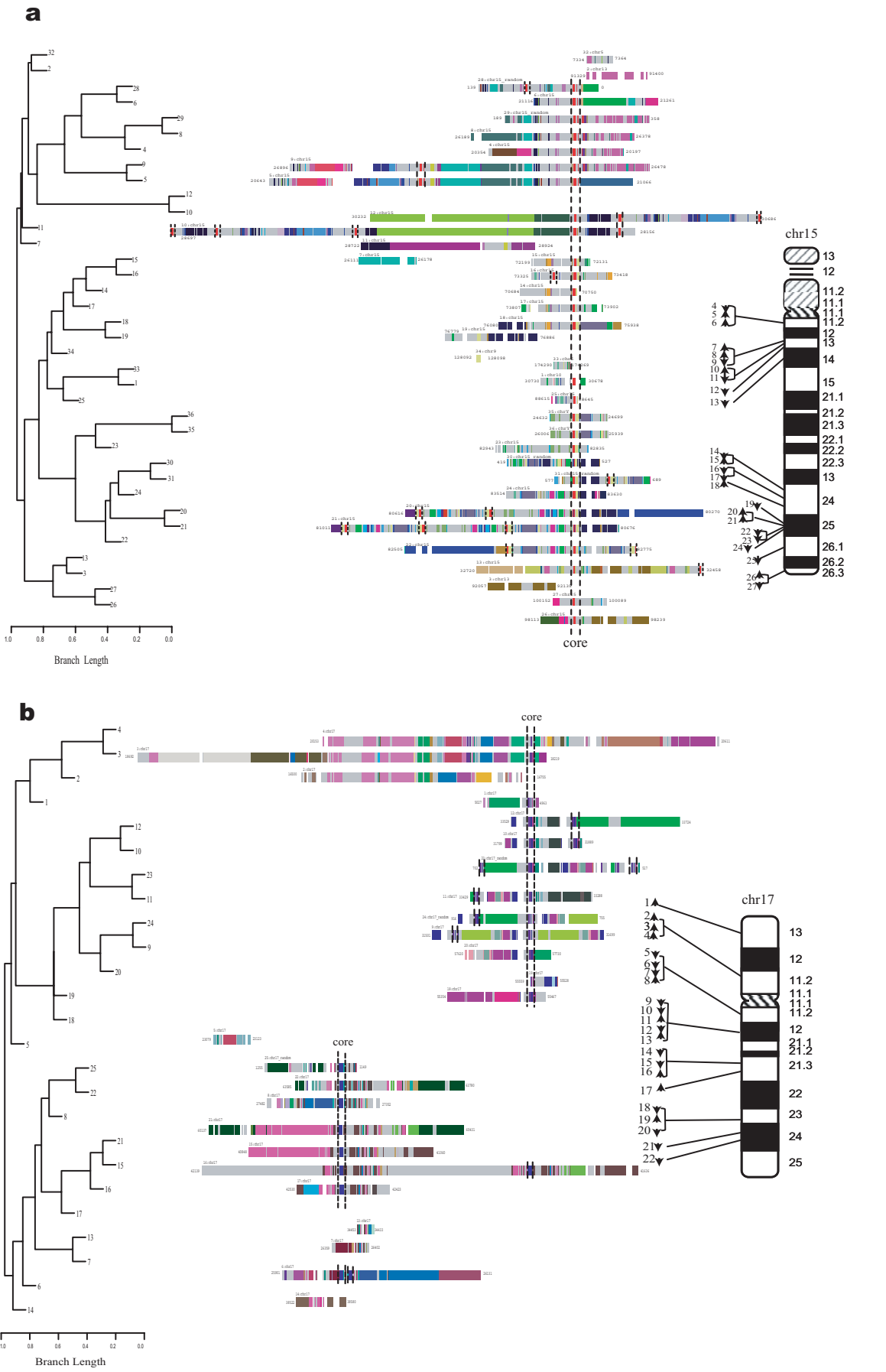
Supplementary Figure 1 Modified A-Bruijn graph and duplication subunit definition.

(a) A schematic depicting A-Bruijn graph construction from imperfect repeat structures. A segmental duplication pairwise alignment will consist of matched (empty boxes) and mismatched (colored boxes) portions as well as unaligned segments due to insertion/deletion events (dashed line). When constructing of the A-Bruijn graph, the matched portions of alignment are first glued and collapsed into a single edge, while the mismatched and indel portions form whirls and bulges in the graph. (b) Simplification of A-Bruijn graph by solving of the Maximum Subgraph with Large Girth (MSLG) problem. The numbers associated with edges indicates the multiplicity (copy number) of each edge. When girth = ∞, the problem is reduced to Maximum Spanning Tree Problem. For a finite large girth, we use the maximum spanning tree to arrive at an approximate solution of MSLG, in which edges are added to the graph in the decreasing order of their multiplicities, and an edge is added if and only if it dose not form a short cycle (shorter than the girth (n=25 bp) with existing edges.



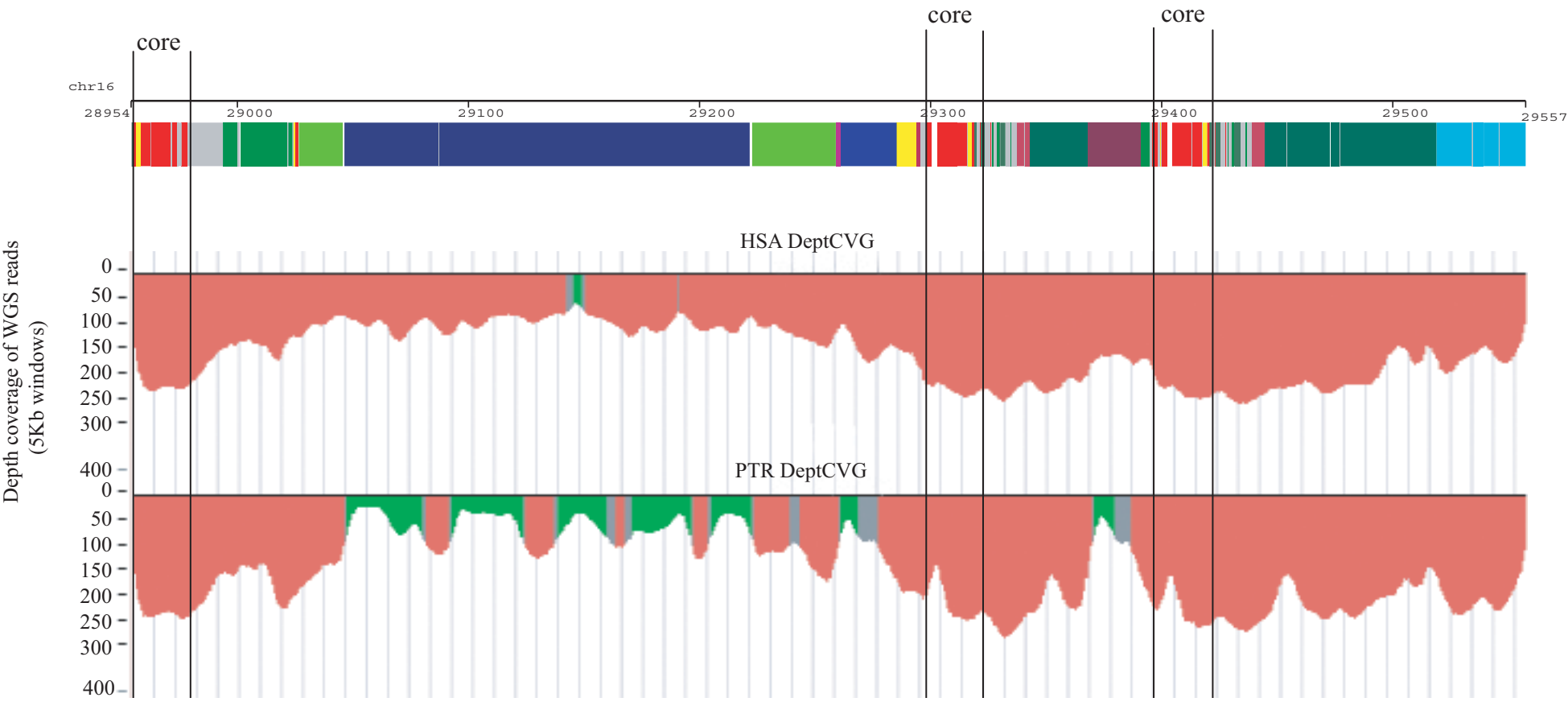
Supplementary Figure 2 Metaphase results of comparative FISH analysis.

The figure illustrates the underlying metaphases for the comparative FISH analysis that we used to generate Fig. 5. A human fosmid clone corresponding to one of the predicted derivative locus (red arrows) was used as a FISH probe on metaphase chromosomes from both human and macaque lymphoblastoid cells (left and right panels respectively). FISH results on macaque shows a single positive signal corresponding to the syntenic region of ancestral loci predicted by the computational method (white arrows), confirming the ancestral position.



Supplementary Figure 3 Duplication structures of chr15 and chr17 groups.

(a-b) Figure depicts the mosaic structure of *complex duplication blocks* for chromosome 15 (a) and chr17 (b). The duplication blocks were numbered according to genomic location of a locus in the chromosome. Different colors denote distinct ancestral loci. A "core element" shared by 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.



Supplementary Figure 4 Core vs. flanking duplicon comparative primate analysis.

Duplication block from chromosome 16 is shown with their core vs. flanking duplicon architecture (the location of core duplicons is marked). These regions were assessed for evidence of segmental duplication in different species by measuring the excess of whole genome shotgun sequence from human (HSA) and chimpanzee (PTR) [13, 29] aligned to the human reference genome (% identity threshold cutoff >94%). Significant excess (3 s.d.) in the depth of coverage (red) represents evidence of duplication in each species. The cores (as determined computationally) represent regions of shared duplication between HAS and PTR while the flanking duplicons are much more likely to be younger and human-specific events.

Supporting Note

The following sections detail the computational algorithm and the genome-wide analyses that were performed to reconstruct the evolutionary history of human segmental duplications.

1. Duplication Subunit Definition

Currently, the dataset of known human segmental duplications is represented as a set of 28,856 pairwise alignments [1, 2] (≥ 1kbp in length and ≥ 90% sequence identity) (http://www.genome.ucsc.edu). It offers no direct information regarding the order and directionality of the duplication events or the origin of the ancestral region (duplicon). Detailing the substructure is notoriously difficult due to the complex pattern created by larger, secondary duplications and subsequent rearrangements at different times during evolution [3-5]. Both processes have generated many boundaries which no longer correspond to the initial ancestral duplication events [2, 6]. Consequently, the mosaic structure of complex duplication regions can not be readily deduced from a simple analysis of pairwise alignments (Fig. 1 and Fig. 2).

Existing algorithms that classify transposon/retrotransposon repeat families [6-8] cannot be directly applied to delineating the boundaries of segmental duplications due to the complex pattern of duplication alignments. In theory, resolution of these structures can be achieved by application of the classical de Bruijn graph approach [9] or the suffix tree approach [10], if different copies of the same subunits perfectly align to each other. Human segmental duplications, however, represent imperfectly aligned repeats, containing both

mismatches and indels in their alignments (Supplementary Fig. 1a). Pevzner et. al. generalized the concept of the de Bruijn graph to classify nearly identical repeat copies, in which A-Bruijn graph, the counterpart of de Bruijn graph in the case of imperfectly matched repeats, is constructed from pairwise alignments of repeat copies using a modified maximum spanning tree algorithm [11]. This algorithm was recently applied to various bioinformatics problems including construction of multiple alignment of proteins with rearranged domains [12], comparative repeat analysis [13], and protein sequencing via tandem mass spectrometry [14]. Here, we adapt this algorithm to derive the mosaic structure of subunits for the segmental duplications in the entire human genome.

We essentially implemented the repeat-graph algorithm formally described in Pevzner et al., 2004. et al as follows: "Let S be a genomic sequence of length n and $A = (a_{ij})$ be a binary $n \times n$ "similarity matrix" representing the set $\mathcal{A}$ of all significant local pairwise alignments between regions from S. The matrix A is defined as $a_{ij} = 1$ if and only if the positions i and j are aligned in at least one of the pairwise alignments and $a_{ij} = 0$ otherwise (note that insertions and deletions are not recorded in A). Matrix A represents an "adjacency matrix" of a graph (called the A-graph) on n vertices 1,..., n (vertices i and j are connected iff $a_{ij} = 1$). Let V be the set of connected components of this graph and let $v_i \in V$ be the connected component containing vertex i ($1 \leq i \leq n$). The A-Bruijn graph $G(V, E)$ is defined as the multi-graph on the vertex set V with (n-1) directed edges (v_i, v_{i+1}) for $1 \leq i < n$. One can view the A-Bruijn graph as the Eulerian path obtained from the path (1,..., n) after contracting each connected component into a single vertex (Supplementary Fig.1). Vertices v_1 and v_n are called the source and sink."

.

Using these underlying segmental duplication pairwise alignments, we constructed an A-Bruijn graph that summarizes all possible extant sequence relationships among recent duplications as follows: Pairwise alignments are first binned into families (n=665) if there is any evidence of shared duplication alignments between them. We classify these alignments into non-intersected groups as follows. Each pairwise alignment consists of two (non-overlapping) segments S_i and S_j from the genome. We map these aligned segments onto the genome and define duplication region as the continuous regions that are covered by the segments. Many aligned segments do not overlap but come close to each other (e.g. separated by Alu or LINE element) and we have chosen to represent them within the same duplication region. We therefore combine non-overlapping segments if the interval between them is shorter than an overlap threshold (default 500bp). We view all duplication regions as vertices in the segment duplication graph and connect vertices *v* and *w* if there exists a pairwise alignment such that one segment of this alignment is a part of duplication region *v* and the other segment is a part of duplication region *w*. The connected components in the segmental duplication graph define the duplication family. We assume that the different groups are not evolutionarily related (as there is no shared duplicated sequence nor proximity of sequence) and, therefore, can be analyzed separately.

Next, the procedure takes the underlying pairwise alignments within each group as input and threads through each alignment (RepeatGluer)[11] to define the edges of the repeat graph—a procedure that is reiteratively performed for each alignment until all

components of a duplication family are represented as a path consisting of edges of known copy number (i.e. process is transitive such that "if x is glued to vertex y and vertex y is glued to vertex z, then vertices x an z are also glued" defining a single set of linkage clusters[13]. The edges of the graph correspond to continuous genomic segments for which no breakpoint exists—these are defined as the duplication subunits—while the vertices correspond to the alignment breakpoints. We applied the RepeatGluer algorithm to build the repeat graph for each duplication family. Assume there are n duplicated regions, S^1, S^2,..., S^n, in the genome, and any given pairwise alignment involves two segments from these regions. An A-Bruijn graph, as a directed graph with n sources and n sinks, can be constructed to represent these pairwise alignments. We first model each duplicated region $S = S_1 ... S_m$ as a directed path on m vertices. Two vertices in duplication regions S^i and S^j are then glued into a single vertex if they correspond to a pair of matched positions in the pairwise alignment between segments of S^i and S^j. Supplementary Fig. 1 schematically depicts the construction of an A-Bruijn graph from a hypothetical pairwise alignment. Matched (colored) and mismatched (empty) regions for a pairwise alignment are identified, and putative deletions and insertions are flagged (dashed lines). During A-Bruijn graph construction, the matched regions of alignment are "glued' or collapsed into a single edge (Supplementary Fig. 1a), while the mismatched and indel regions are transformed into whirls and bulges in the graph, respectively. The resulting graph from this initial step is often complicated by an excess of short whirls and bulges, and cannot directly be used for delineating the boundaries of duplication subunits (Supplementary Fig. 1b).

During the process of A-Bruijn graph construction, the perfect matched portion of duplication alignment are merged and represented as a single edge in the graph; while mismatches (e.g. indels) are represented in our graph as multiple edges, constituting whirls and bulges. The A-Bruijn graph is further simplified by eliminating these short whirls and bulges, computationally formulated as the Maximum Subgraph with Large Girth (MSLG) problem. The aim of MSLG is to remove the minimum number of edges from the A-Bruijn graph to find a maximum subgraph that does not contain any cycles shorter than a specified length. The length of cycle is a predefined parameter called girth. The MSLG problem with girth = ∞ is similar to the well-known Maximum Spanning Tree Problem [15], which can be solved by a greedy algorithm in linear time. However, for an arbitrary girth, the MSLG problem is NP-hard. We approximate solution of the MSLG problem by adding edges to the constructed maximum spanning tree. This algorithm analyzes edges that did not make it into the maximum spanning tree in order of their decreasing multiplicities (i.e. copy numbers). An edge is added to the graph if and only if it does not form a short cycle (shorter than the specified girth) with existing edges; otherwise, it is deleted from the graph (Supplementary Fig. 1b) [11].

Thus, we initially sought to minimize the girth in an effort to capture the maximum number of breakpoints and to ensure that each subunit is as homogenous as possible regarding its sequence origin. (Note: After a duplication event both ancestors and derivative loci can be further fragmented by secondary events, such as a new insertion of transposable elements or a secondary partial duplication event). We examined the repeat graph of human segmental duplications at various girths reaching a computational limit at

25 bp (the algorithm is memory intensive requiring 29 Gigabytes of memory from our 32 Gigabyte computational cluster). No significant differences in the repeat graph topology were observed for girths ranging from 150 to 25 bp suggesting that an optimum cycle length had been obtained. At this resolution (girth=25 bp), we obtained a total of 15,548 non-redundant duplication subunits which is sufficient to encode ~98% (148.6 Mb/152.2 Mb) of all duplicated basepairs of the human genome. Since the majority duplicated basepairs (147.9 Mb/ 152.2 Mb or 97%) correspond to subunits ≥100 bp in length (n=11,951), we selected this subset for further analyses.

2. Duplicon Definition Based on Comparative Sequence Analysis

In order to identify the ancestral location of each duplication subunit (termed *duplicon*) [16], we took advantage of published genome sequence of outgroup mammalian species (macaque, mouse, rat and dog) and the expectation that the majority of the segmental duplications emerged recently during human primate evolution (see below). Due to the multi-step process of segmental duplications, an ancestral locus will typically share a larger homologous synteny block (HSB) in an outgroup species because sequence anchors extend beyond the boundaries of the duplication (Fig. 4). We, therefore, examined reciprocal best-hit for each duplication subunit using the program of liftOver, based on the underlying cross-species chain data from UCSC genome browser (http://genome.ucsc.edu) [17]. The cross-species chain data were derived from BLASTZ alignments, an algorithm using a gap scoring system that allows longer gaps than traditional affine gap scoring systems. Consequently, it more effectively tolerates sequence gaps (indels) in both lineages and allows larger syntenic alignment between

different genomes to be constructed. Using this underlying cross-species alignment data, the program of liftOver, which tracks coordinate systems between different genome assemblies, converts genomic coordinates between species to find the best homologous synteny block[18].

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 (Fig. 4). If more than one locus with an equivalent number of outgroup species reciprocal best-hit was identified, the ancestral state was classified as "not determined". Using this method we determined a likely ancestral state for 6,999/11,951 subunits (47.2%) or 102.4 Mb/152.2 Mb (67.3%) of all human duplicated basepairs. This analysis provided the first genome-wide prediction of ancestral and derivative duplication loci and provides directionality to the initial segmental duplication events within the human genome.

In order to guarantee each duplication subunit is homogenous in terms of its sequence origin, we deliberately overfragment the duplications by using a girth (n=25bp) as short as computationally possible during the process of A-Bruijn graph reconstruction. We chain duplicons into larger duplicons if one of two conditions is met. During repeat graph construction the boundaries of a duplication subunit can shift maximally by 25 bp on either side due to the presence of bulges or whirls in the repeat graph (Supplementary Fig. 1a). Therefore, once the ancestral origin of the duplication subunit was identified, we chained any adjacent ancestral duplicons that mapped within 50 bp of one another. In

addition, common repeat sequences would potentially disrupt an ancestral duplication subunit. We, therefore, chained adjacent subunits if the intervening sequence was completely composed of common repeat sequences such as Alu's and L1s. This procedure reduced the number of duplication subunits from 6,999 to 4,692. We consider this chaining threshold conservative as the probability of chaining two independent ancestral duplications separated precisely by a common repeat sequence is rare. Therefore, we consider 4,692 to represent an upper bound to the number of ancestral loci. However, we recognize that duplication subunits (especially more ancient ones) incur rearrangement events that alter the order and proximity with respect to the ancestral locus, leading to fragmentation (false negatives) by this method. We, therefore, also considered a series of arbitrary chaining parameters based strictly on the length of intervening sequence between two duplicons, irrespective of orientation. Our results showed that the number of duplicons reduced incrementally as chain length increased from 10 to 100 kb. Based on this asymptotic relationship we estimate a lower bound to the number of ancestral duplicons of about 2,200.

3. Validation

To test the overall efficacy of our approach, we compared our *in silico* results (n=4,692 duplicons) to three experimental datasets where the ancestral states had been determined from comparative sequence, comparative FISH and phylogenetic analysis on chromosomes 15q11, 16p12/13 and 2p11. [2, 3, 19-21]. We found an excellent correspondence 37/41 ancestral loci were consistent between the two methods. In other words, the location of an ancestral duplication and the extent of the segmental duplication

boundaries were within 500 bp of that predicted by the experimental FISH and comparative sequence analysis. Four regions (indicated by ND) are regions where ancestral loci were not determined by *in silico* method but had been identified by experimental analyses. Our computational approach predicted an additional 19 duplicons that had not been detected by the experimental analysis—most of these corresponded to duplicons of insufficient length to perform comparative FISH analysis.

As a second test of validation, we performed comparative primate FISH. Human fosmid probes (~40 kb in length) were selected corresponding to the derived locus (n=12) and used as probes for comparative FISH purposes. If we had correctly identified the ancestral locus using our computational approach, then the probe, when hybridized to an outgroup primate species, should hybridize to the ancestral locus as opposed to the derived locus. Since derived duplicated loci are typically mosaic in their organization, we required that the duplicon have a minimum length of 40 kb so that the probe would correspond only to the duplicon and not flanking duplicons which would complicate the analysis. Within the confines of this length threshold, we selected regions randomly (See Supplementary Table 1 for result).

4. Limitations

The validation experiments predict that >90% of the computationally defined ancestral subunits are accurately identified. There are, however, several limitations of our approach:

1) **Incomplete underlying pairwise alignments:** The graph theory approach is dependent upon a complete set of accurate pairwise alignments to identify duplication subunits. Optimal global alignments (Needleman-Wunsch) can occasionally fail due to large insertion/deletions or inversion events. We identified 2% of the initial alignment as incomplete and corrected these by altering affine gap parameters to produce more consistent alignment results (uniform sequence identity). However, there is still missing sequence from the human genome assembly. Since this gap sequence is enriched in segmental duplications, it is likely we are missing a fraction of edges and vertices in our repeat graph. As more sequence becomes available from these regions, a more complete repeat graph may be generated.

2) **Girth (25 bp):** During repeat graph construction of human segmental duplications, we assessed various girths reaching a computational limit at 25 bp (the algorithm is memory intensive requiring 29 Gigabytes of memory from our 32 Gigabyte computational cluster). Ideally, a girth of 1 bp would be preferred but this is not computationally feasible at this time.

3) **Highly fragmented subunits:** In this study we only considered duplication subunits ≥ 100 bp in length. While this constitutes the majority (147.9 Mb/ 152.2 Mb or 97%) of duplicated basepairs, we did not consider shorter duplication subunits since there is frequently insufficient phylogenetic signal to map these by reciprocal BLASTZ alignments between human and outgroup genomes. The remaining short duplication subunits correspond to regions of high-frequency of rearrangement (HFR) and were accordingly classified as HFR duplication subunits in our analysis. Consequently, duplication subunits less that 100 bp in length are not considered in this analysis.

4) **90% sequence identity threshold:** Based on a neutral rate of evolutionary decay and a molecular clock established based on primate resequencing of segmental duplications16, a threshold of 10% sequence divergence was chosen in order to focus on duplications that emerged within the last 40 million years of human genome evolution. The segmental duplications are, therefore expected to be primate-specific and to have emerged since the separation of the Old World Monkey and New World Monkey lineages (< 40 million years ago). In some rare cases, however, two lineages of duplication may have arisen from a more ancient duplication that existed before this % identity threshold. Based on our operational threshold of 90%, such duplications would currently be represented as two distinct duplication subunits and therefore as two separate ancestral loci. The "grandfather" locus in such cases would not be identified.

5) **Missing Ancestral Sequences:** Only 67% of the ancestral loci could be assigned to the duplication subunits. There are three circumstances where an ancestral locus will not be identified or potentially misassigned in this analysis: a) the ancestral locus has been destroyed in the outgroup or human genomes (i.e. deletion or gene conversion during evolution; tandem segmental duplications are particularly prone to this effect as the evolutionary history has been erased by sequence homogenization); b) the duplications have independently duplicated in outgroup species such that no clear reciprocal best-hit can be identified; and c) the ancestral locus maps to the Y chromosome (the genome sequence from all outgroup species has been obtained from female individuals and as such syntenic relationships to the Y chromosome can not be

defined). Duplicons that originated from the Y chromosome and were duplicated to the X chromosome or an autosome would not be identified.

5. Primate Duplication Analyses Comparisons

All human segmental duplications (WGAC alignments) were reduced into a non-redundant set of duplication subunits (as described above) consisting of 49.12 Mb of sequence. Using the human genome sequence as a reference, we selected only those human duplication subunits for which there was evidence of recent duplication (>94% sequence identity) as determined by the whole genome shotgun sequence detection method in human [1]. We classified each of these human duplication subunits into one of three categories based on the duplication maps established for chimpanzee [22] and macaque [23]. In the case of macaque, segmental duplications were detected by mapping macaque WGS reads against the macaque genome assembly and then using the UCSC liftOver tool to map coordinates back against the human reference sequence (hg17). 90.6% (13.55 Mb/14.96 Mb) of the macaque duplications could be reliably mapped back against the human genome. Human segmental duplications were then classified into one of three categories: HSA-only (detected as duplicated only in the human lineage); HSA+PTR (detected in human and chimpanzee but not in macaque); and HSA+PTR+MMU (detected in all three species). The sequence identity spectrum for each of the three categories (based on human sequence alignments) was compared. This three-way comparison clearly indicates an excess of high-sequence identity (>98%) duplications that are specific to the human lineage. The duplications of HSA and PTR were also compared for core and non-core regions. And we found that the cores represent

regions of shared duplication among human and chimpanzee while the non-core duplicons are much more likely to be younger and human-specific events (Supplementary Fig. 4).

6. Duplication Divergence and Simulation Analyses

Sequence divergence between derivative and ancestor pairs was computed using Kimura's two parameter model [24]. For any given duplication block/clade, we tested whether the observed distribution differed significantly from a random distribution model as follows: For a specific duplication block composed of a certain number (N) of duplicon subunits, we first computed pairwise sequence divergence (K_{2m}) for each ancestor-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 ancestor-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). A similar analysis was repeated based on our analysis of the variance. Based on 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 are only counted once in this analysis. Similarly, we repeated the analysis at the level of duplication groups (see below) where each ancestral duplicon is only counted once during the simulation. 10/24 duplication groups showed evidence of non-random distribution of genetic divergence.

7. 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 is present within a duplication block, we assigned a "1", otherwise a "0" to that block generating a binary *phylogenetic profile* for each block. *Complex duplication blocks were* then clustered into 24 *duplication groups* by hierarchical clustering based on the similarity of their *phylogenetic profile* [25-27]. A duplication group is a cluster of complex duplication blocks grouped based on shared duplicon or duplication subunit content. There is no fixed definition for assigning duplication blocks into duplication groups, but by definition if a duplication block has no subunits in common with another duplication block it will have a maximal branch length of 1.0 (see branch length Fig. 7). If there is no shared content, we infer there is no related evolutionary history. As the branch length approaches 0, the duplication blocks share more duplicons in common and the structure of the two duplication blocks being compared becomes nearly identical (100% shared duplicon content). Our approach was to cluster duplication blocks requiring different numbers of duplicons and to analyze consistency in the topology of the tree. We found that the composition of the duplication blocks remain largely invariant irrespective of the number of subunits that are considered (i.e.5, 10, 15, 20, etc, data not shown). Internode branch lengths approximating 1.0 , thus, were used to partition the duplication groups. A chromosome name will be assigned to a group, if the majority of blocks (>50%) in that clade are belong to a homologous chromosome, otherwise the group is designated as a mixed (M) clade (Fig. 7a).

8. Core Duplicon Definition

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, some groups were further divided into subgroups. 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 =0.67~1) from non-core duplicons (Ci<0.67). We, therefore, operationally 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. 7 b and Supplementary Fig. 3 a-b).

9. Gene/Transcript Analysis

Non-redundant genes (Refseq gene n=22,589) and spliced ESTs (n=4,246,559) were assigned to core and non-core locations based on the highest alignment score (http://www.genome.ucsc.edu). If a transcript mapped to two or more locations with an equivalent score, one was selected at random. Each transcript was assigned once and only once to the genome. Alternative splice variants were eliminated by clustering exons that overlapped and counting it as a single exon. Fusion transcripts were defined as Refseq genes or spliced ESTs where different exons within the same transcript mapped to two or more distinct ancestral positions. We required that the ancestral duplicon loci are separated by more than a Mb or map to different chromosomes.

Differences in exon density between core and non-core duplicons were tested by simulation as follows: Core and non-core duplicons were partitioned randomly among duplicated regions of the genome. The enrichment of Refseq gene and EST exon density (exon/Mb) was then computed both based on observed data and the simulated data (n=1,000 replicates) and significance established empirically by assessing the number of times the core/non-core enrichment was observed. We considered the different levels of Refseq annotation as part of our analysis (data not shown). The enrichment of exons in core region was observed for genes assigned to the "reviewed", "provisional" and "predicted" categories. Both reviewed and provisional curations constitute the majority of the genes mapping to cores. These are thought to be well-supported and "represent valid transcript and proteins" (http://www.ncbi.nlm.nih.gov/RefSeq/key.html#status). We did not find an enrichment of exons for genes assigned to the "validated" category, although this represented the minority (~10%).

10. Positive Selection Analysis on Genes Mapping to the Core Duplicons

There are several metrics that may be applied to assess positive selection. For those gene families for which positive selection had not been demonstrated previously, we considered four different measures: a likelihood method assessing dN/dS ratios (omega) for different branches in the phylogenetic tree (Branch Analysis)[28]; a distance-based approach that provides branch estimates of the number of synonymous (bs) and non-synonymous (bn) substitutions per site (Bn/Bs Analysis) [29]; a sites model test which tests selection over individual codons (Sites Analysis)[28] and an overall formal Likelihood Ratio Test (LRT) comparing a partially neutral vs. a positive model of likelihood for the

evolution of the entire gene family. The latter is considered one of the most stringent tests of positive selection. For every gene family, we constructed an amino acid alignment using gene RefSeq models from human (HSA), chimpanzee (PTR) and macaque (MMU) gene copies where available. A posterior back-translation to DNA was applied in the construction of the optimal multiple sequence alignment. We manually inspected all alignments and we conservatively retained only those regions where the CDS aligned according to the gene model. As a positive control we used two portions of the NPIP gene family which had been shown previously to be under positive selection during primate evolution [30]. Two independent parts of this gene were used to ensure optimal amino acid alignments (exon 5 shows considerable alternative splicing). Some evidence for positive selection was observed for the different gene families mapping to the core duplicons (esp. GOLGA). However, none showed as robust signal as the positive control nor could a neutral model be formally rejected for the overall evolution of these three gene families (Supplementary Table 5). We take this as weak evidence of positive selection. Additional analyses, such as subcloning and sequencing cDNA from outgroup species, will need to be performed to more formally confirm or reject positive selection.

References:

1. Bailey, J. A. et al. Recent segmental duplications in the human genome. Science 297, 1003-7 (2002).
2. She, X. et al. The structure and evolution of centromeric transition regions within the human genome. Nature 430, 857-64 (2004).
3. Horvath, J. E. et al. Punctuated duplication seeding events during the evolution of human chromosome 2p11. Genome Res (2005).
4. Samonte, R. V. & Eichler, E. E. Segmental duplications and the evolution of the primate genome. Nat Rev Genet 3, 65-72 (2002).

5. Stankiewicz, P., Shaw, C. J., Withers, M., Inoue, K. & Lupski, J. R. Serial segmental duplications during primate evolution result in complex human genome architecture. Genome Res 14, 2209-20 (2004).
6. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12, 1269-76 (2002).
7. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. Bioinformatics 21 Suppl 1, i351-i358 (2005).
8. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. Bioinformatics 21 Suppl 1, i152-8 (2005).
9. de Bruijn, N. A combinatorial problem. Koninklije Nedderlandse Academie van Wetenshappen Proc. A49, 758-764 (1946).
10. McCreight, E. M. A Space-Economical Suffix Tree Construction Algorithm. Journal of the ACM (JACM) 23, 262-272 (1976).
11. Pevzner, P. A., Tang, H. & Tesler, G. De novo repeat classification and fragment assembly. Genome Res 14, 1786-96 (2004).
12. Raphael, B., Zhi, D., Tang, H. & Pevzner, P. A novel method for multiple alignment of sequences with repeated and shuffled elements. Genome Res 14, 2336-46 (2004).
13. Zhi, D., Krishna, S. S., Cao, H., Pevzner, P. & Godzik, A. Representing and comparing protein structures as paths in three-dimensional space. BMC Bioinformatics 7, 460 (2006).
14. Bandeira, N., Tang, H., Bafna, V. & Pevzner, P. Shotgun protein sequencing by tandem mass spectra assembly. Anal Chem 76, 7221-33 (2004).
15. Cormen, T., Leiserson, C., and Rivest, R. *Introduction to algorithms.* (1989).
16. Bailey, J. A. & Eichler, E. E. Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet 7, 552-64 (2006).
17. 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 U S A 100, 11484-9 (2003).
18. Schwartz, S. et al. Human-mouse alignments with BLASTZ. Genome Res 13, 103-7 (2003).
19. Horvath, J. E. et al. Using a Pericentromeric Interspersed Repeat to Recapitulate the Phylogeny and Expansion of Human Centromeric Segmental Duplications. Mol Biol Evol (2003).
20. Locke, D. P. et al. Molecular evolution of the human chromosome 15 pericentromeric region. Cytogenet Genome Res 108, 73-82 (2005).
21. Johnson, M. E. et al. Recurrent duplication-driven transposition of DNA during hominoid evolution. Proc Natl Acad Sci U S A (2006).
22. Cheng, Z. et al. A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature 437, 88-93 (2005).
23. Gibbs, R. A. et al. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316, 222-34 (2007).
24. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16, 111-20 (1980).
25. Bowers, P. M., Cokus, S. J., Eisenberg, D. & Yeates, T. O. Use of logic relationships to decipher protein network organization. Science 306, 2246-9 (2004).
26. Rivera, M. C. & Lake, J. A. The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature 431, 152-5 (2004).
27. 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, 681-90 (2004).
28. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13, 555-6 (1997).
29. Zhang, J., Rosenberg, H. F. & Nei, M. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci U S A 95, 3708-13 (1998).
30. Johnson, M. E. et al. Positive selection of a gene family during the emergence of humans and African apes. Nature 413, 514-9. (2001).

Glossary of Terms

Segmental duplication: Continuous portions of genomic DNA that can be mapped to two or more locations within a genome. Segmental duplications are operationally defined as pairwise alignments between genomic regions, typically ≥1kb in length and ≥90% in sequence identity.

Duplication subunit: The smallest continuous genomic segment within a segmental duplication for which no breakpoint exists when compared to all other pairwise alignments of the region. Computationally it is defined as two adjacent vertices of the A-Bruijn repeat graph defining the start and end points of the subunit with the edge of the graph representing the continuously aligned sequence. Duplication subunits by definition are non-redundant. All segmental duplications can be decomposed into a series of duplication subunits.

Girth: The length of an inconsistency within an alignment (deletion, insertion and mismatch) which can be tolerated during repeat graph construction. Such an inconsistency will generate a whirl and bulge in A-Bruijn graph and will be subsequently collapsed or simplified during duplication subunit definition (in this study a girth of 25 bp was selected).

Duplicon (ancestor): A duplication, or portion, thereof that can be tracked to an ancestral or ancestor locus. The extent of sequence overlap with the ancestor locus defines the boundaries of the duplicon. In this study, the ancestral locus was defined by comparison to mammalian outgroup sequence and chaining the duplication subunits that correspond to the ancestral locus.

Complex duplication blocks: A larger duplication region composed of multiple (≥10) adjacent duplicons that originate from different areas of the genome.

Phylogenetic profile: phylogenetic profile of each complex duplications block was constructed based on the extent of shared duplicon content for each complex duplication block. If a duplicon is present within a duplication block, we assigned a "1", otherwise a "0" to that block generating a binary *phylogenetic profile* for each block.

Duplication family: All connected components in a segmental duplication graph. A term is applied only to the construction of the repeat graph. In this analysis there were 665 duplication families for which individual repeat graphs were constructed.

Duplication groups: A cluster of complex duplication blocks grouped based on shared duplicon or duplication subunit content.

Core duplicon: A duplicon or a series of adjacent duplicons that are shared by the majority of the members of a duplication group; they represent the central sequence

around which a complex pattern of duplication forms. Operationally, they were defined by a core index (see Methods) of at least 67% (top 10% in terms of duplicon abundance).

Flanking duplicons or duplication subunits: Duplications subunits or duplicons in a complex duplication block that flank the core duplicon.

Fusion transcripts/gene: A transcript/gene (as indicated by non-redundant EST /refSeq gene) derived from two evolutionary distinct duplicons (duplicons must either be located in different chromosome or separated from each other at least 1 Mb).

Collinear duplicon pairs (Methods only): After duplication seeding events (first step), both ancestors and derivatives may be fragmented by secondary genomic rearrangement events. This generates multiple pairs of ancestors (duplicons) and derivatives, which originate from single seeding event. In order to reduce this overestimation of ancestral loci, we introduced the concept of collinear duplication" as follows: Let's assume that we have N ancestors (D_1 to D_N) that donated to n derivatives (A_1 to An). If we find the physical distance between D_1 to D_N and A_1 to An are within a specified range (10kb), we consider all these duplication pairs as collinear duplicon pairs and their sequence divergences (K_{2m}) were averaged and counted only once in our simulation analysis (Fig. 6) .