ARTICLES

A genome-wide comparison of recent chimpanzee and human segmental duplications

Ze Cheng[1], Mario Ventura[2], Xinwei She[1], Philipp Khaitovich[3], Tina Graves[4], Kazutoyo Osoegawa[5], Deanna Church[6], Pieter DeJong[5], Richard K. Wilson[4], Svante Pääbo[3], Mariano Rocchi[2] & Evan E. Eichler[1]

We present a global comparison of differences in content of segmental duplication between human and chimpanzee, and determine that 33% of human duplications (>94% sequence identity) are not duplicated in chimpanzee, including some human disease-causing duplications. Combining experimental and computational approaches, we estimate a genomic duplication rate of 4–5 megabases per million years since divergence. These changes have resulted in gene expression differences between the species. In terms of numbers of base pairs affected, we determine that *de novo* duplication has contributed most significantly to differences between the species, followed by deletion of ancestral duplications. Post-speciation gene conversion accounts for less than 10% of recent segmental duplication. Chimpanzee-specific hyperexpansion (>100 copies) of particular segments of DNA have resulted in marked quantitative differences and alterations in the genome landscape between chimpanzee and human. Almost all of the most extreme differences relate to changes in chromosome structure, including the emergence of African great ape subterminal heterochromatin. Nevertheless, base per base, large segmental duplication events have had a greater impact (2.7%) in altering the genomic landscape of these two species than single-base-pair substitution (1.2%).

Recent segmental duplications have had a pivotal role in the evolution of the architecture of the human genome[1–6], the emergence of new genes[7,8] and the adaptation of our species to its environment[9–12]. They contribute to large-scale structural polymorphism[13–17] and a host of genomic diseases[18]. Several gene and genomic-based analyses suggest that the human genome is particularly enriched for genes that have emerged as a result of recent duplication[11,19]. It is unknown whether slow rates of deletion, high rates of duplication or gene conversion are largely responsible for the evolutionary maintenance of these duplicates. We sought to understand the origin and impact of this fraction of the genome by performing a detailed comparison of the human and chimpanzee genomes for regions that showed evidence of shared and lineage-specific duplication.

Chimpanzee segmental duplications

We used two independent approaches to estimate the size and extent of chimpanzee (*Pan troglodytes*) duplications. We first performed a self-comparison of the chimpanzee genome assembly using the whole-genome assembly comparison method (WGAC)[20]. We noticed a significant (threefold) reduction of more divergent (94–95% sequence identity) chimpanzee interchromosomal pairwise alignments when compared to human (Supplementary Fig. S1). As expected, more recent duplications (>97% sequence identity) were five times as likely to be misassembled or fragmented when compared to unique chimpanzee sequence[21]. To identify these duplications using chimpanzee whole-genome shotgun (WGS) sequence data (3.5-fold sequence coverage), we implemented a second duplication detection method[11] that uses the depth of coverage of random sequence read data against a reference sequence to identify duplicated sequence. We applied the whole-genome shotgun sequence detection (WSSD) strategy by mapping 23.7 million reads from chimpanzee against the human genome reference (Fig. 1). Table 1 summarizes the results of the duplication analyses for the two genomes using the WSSD approach for regions >20 kilobases (kb) in length and >94% sequence identity.

We classified DNA into one of three possible categories: duplicated only in chimpanzee, duplicated only in human or shared between chimpanzee and human (Fig. 1b–d; see Methods). On the basis of five different computational and experimental analyses, including array comparative genomic hybridization and fluorescence *in situ* hybridization (FISH) (32 out of 34 validations; Supplementary Figs S2–S6, Supplementary Tables S1–S6 and Supplementary Methods), we estimate that we have detected >90% of all segmental duplications in the chimpanzee genome that are greater than 20 kb in length. A chimpanzee segmental duplication database as well as detailed chromosomal views for patterns of human and chimpanzee duplications are available (Supplementary Fig. S7; see also http://chimpparalogy.gs.washington.edu) based on mapping all four duplication tracks.

Gene and duplication structure analysis

Although most (66%) of the autosomal base pairs (bp) duplicated in humans are shared between human and chimpanzee (Table 1), a

[1]Howard Hughes Medical Institute, Department of Genome Sciences, University of Washington School of Medicine, 1705 NE Pacific Street, Seattle, Washington 98195, USA. [2]Department of Genetics and Microbiology, University of Bari, 70126 Bari, Italy. [3]Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany. [4]Washington University School of Medicine, 4444 Forest Park Blvd, St Louis, Missouri 63108, USA. [5]BACPAC Resources, Children's Hospital of Oakland Research Institute, Bruce Lyon Memorial Research Building, Oakland, California 94609, USA. [6]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, Maryland 20894, USA.

surprisingly large fraction (~33% or 26.5 out of 79.8 Mb) is duplicated in human but not chimpanzee (Table 1). These human-only duplication intervals map to 515 regions with an average length of 54.6 kb; there is a particular bias for human-specific duplications noted on chromosomes 5 and 15. Significant portions of the duplication architecture that predispose humans to Williams–Beuren syndrome, juvenile nephronophthisis, spinal muscular atrophy and Prader–Willi syndrome[18] appear to be single copies in the chimpanzee (Supplementary Fig. S5). Because non-allelic homologous recombination is thought to provide the molecular basis for recurrent chromosomal structural rearrangements associated with these diseases, these alterations in duplication architecture and the concomitant prevalence of these diseases may be a peculiarity of the human lineage of evolution.

From the perspective of the chimpanzee genome, we identified 11.4 Mb (202 regions) of human sequence that were duplicated in chimpanzee but not in human (approximately 224 kb of the 112 Mb of chimpanzee-specific sequence was also duplicated). If we correct for copy number in the chimpanzee genome, our analysis suggests that the two genomes show comparable levels of autosomal lineage-specific duplication (31.9 Mb in human versus 36.2 Mb in chimpanzee). In contrast, if we compare copy number estimates for shared duplications (588 regions, average length = 94.3 kb), we estimate that the chimpanzee genome has increased in size by as much as 26 Mb (1%), largely as a result of the hyperexpansion of a small number of chimpanzee segmental duplications (see below).

We determined that a chimpanzee-only and human-only duplication were 10.3 times more likely to be located in close proximity to a shared duplication than predicted based on a random simulation model (Supplementary Table S7). These data indicate that either lineage-specific deletion or duplication is occurring in proximity to regions of shared duplication. This effect, which we have termed 'duplication shadowing', suggests that loci near clusters of segmental duplication may be more susceptible to duplication/deletion, probably due to an increased frequency of non-allelic homologous recombination[18].

A total of 177 complete and partial genes (88 and 89 respectively) show evidence of duplication in human but not chimpanzee (for example, *SMN*, KARP1 binding protein, *N*-ethylmaleimide-sensitive factor, *CCL4L1*; Supplementary Table S8). In contrast, only 94 genes were duplicated in chimpanzee but not humans (for example, interleukin receptor-like 1, huntingtin interacting protein 1, and bone morphogenetic protein 2) (Supplementary Table S9). On the basis of U133 Affymetrix gene expression comparisons between human and chimpanzee for five tissues ($n = 30{,}323$ transcripts), we determined that 56% of the human-only gene duplicates showed significant differences in gene expression—83% of this gene expression difference was due to upregulation within human as opposed to chimpanzee ($P < 0.0001$). Similarly, 49% of chimpanzee duplications found in chimpanzee but not in human showed changes in gene expression within chimpanzee when compared to humans—57% was due to upregulation within chimpanzee as opposed to human ($P < 0.01$) (Supplementary Tables S10 and S11). These data indicate that a significant proportion of the lineage-specific duplications resulted in gene expression differences between the two species.

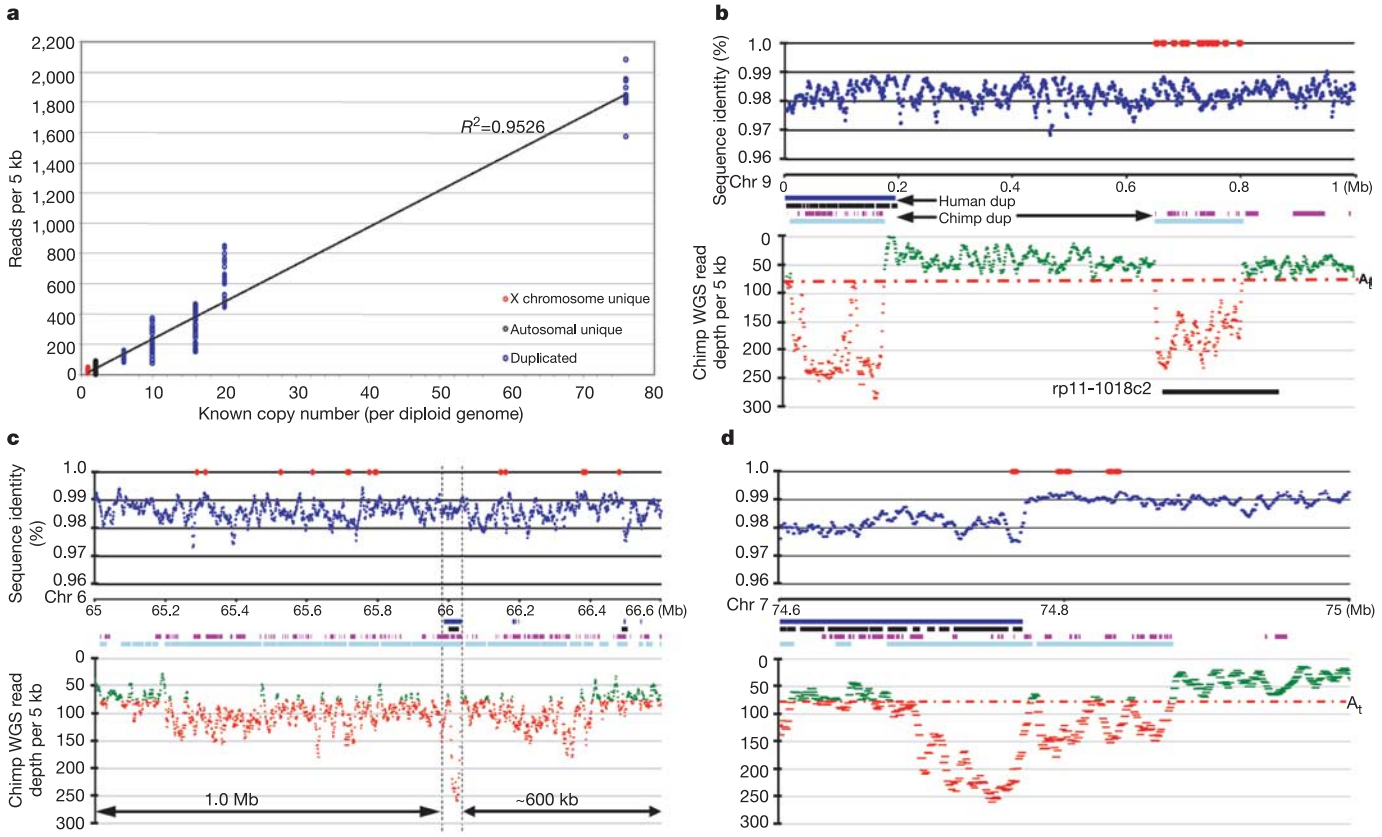

Figure 1 | Chimpanzee segmental duplication detection on human genome assembly NCBI-34 (build 34). **a**, Correlation of copy number and whole-genome shotgun sequence read coverage ($R^2 = 0.953$) is shown based on analysis of unique and duplicated chimpanzee loci of known copy number (Supplementary Table S1). **b**–**d**, Three examples of chimpanzee-only duplications are depicted based on comparison of the four duplication analyses (human WGAC, dark blue; human WSSD, black; chimpanzee WGAC, purple; chimpanzee WSSD, light blue). Significant departures (3 s.d.) in the depth-of-coverage of chimpanzee reads (5-kb windows) are shown below the tracks (red). Red dots indicate the position of 'triallelic' variants (Supplementary Methods).

Table 1 | Chromosome distribution of chimpanzee and human segmental duplications

Chromosome	Size	Human WSSD	Chimpanzee WSSD	Shared	Human only	Chimpanzee only	Shared duplications (human copy number corrected)	Shared duplications (chimpanzee copy number corrected)	Human only duplications (copy number corrected)	Chimpanzee only duplications (copy number corrected)
Chr 1	221.56	6.88	6.13	4.18	2.71	1.96	7.09	7.94	4.17	5.62
Chr 2*‡	237.54	7.95	5.42	4.86	3.09	0.57	5.11	8.42	3.78	1.25
Chr 3	194.47	1.52	1.25	1.06	0.46	0.19	1.21	1.28	0.72	0.55
Chr 4‡	186.84	2.86	2.45	2.29	0.57	0.15	4.91	9.01	0.87	0.55
Chr 5*	177.55	4.17	2.73	2.40	1.77	0.33	3.24	3.42	2.00	0.89
Chr 6†	167.26	1.19	2.86	0.92	0.28	1.95	1.31	1.31	0.38	5.25
Chr 7*†	154.68	7.94	7.02	5.19	2.75	1.83	5.18	6.38	2.93	4.26
Chr 8	142.35	2.19	2.47	1.88	0.31	0.58	2.97	3.78	0.29	1.66
Chr 9*‡	115.62	8.44	7.32	6.61	1.83	0.71	7.25	20.76	1.81	2.16
Chr 10†	131.17	4.73	3.76	3.43	1.30	0.33	3.37	3.41	1.76	3.47
Chr 11	130.91	2.47	2.18	1.78	0.69	0.40	1.63	1.87	0.78	0.88
Chr 12	129.83	0.89	0.79	0.65	0.24	0.14	0.65	0.65	0.21	0.33
Chr 13	95.56	0.99	0.70	0.58	0.41	0.12	0.75	0.74	0.43	1.25
Chr 14	87.19	0.72	0.55	0.25	0.47	0.30	0.25	0.25	0.46	0.70
Chr 15*	81.26	6.75	3.08	2.87	3.87	0.21	2.55	2.48	4.31	0.92
Chr 16*	79.93	8.00	6.82	5.99	2.01	0.83	6.34	6.72	2.31	3.87
Chr 17*	77.68	4.48	3.36	2.99	1.49	0.37	3.23	3.21	1.50	1.15
Chr 18	74.65	1.45	1.10	1.06	0.39	0.04	1.29	1.47	0.38	0.12
Chr 19	55.79	1.43	1.10	0.97	0.46	0.13	1.36	2.60	1.07	0.96
Chr 20	59.42	0.82	0.79	0.77	0.05	0.02	1.00	0.99	0.07	0.00
Chr 21	33.92	1.52	0.97	0.94	0.58	0.03	1.16	1.10	0.86	0.09
Chr 22	34.35	2.44	1.89	1.73	0.71	0.17	2.34	2.22	0.75	0.30
Total (autosome)	2,669.55	79.84	64.75	53.39	26.45	11.36	64.18	90.02	31.86	36.22
Chr X	149.22	6.17	1.96	1.63	4.54	0.33	2.64	2.02	4.98	1.41
Chr Y	24.65	10.56	3.88	3.56	7.00	0.32	4.00	4.77	7.06	1.45
Total	2,843.42	96.57	70.59	58.58	37.99	12.01	70.83	96.82	43.89	39.08

All values are in megabases. All segmental duplications (>94% identity, >20 kb in length) detected by WSSD were compared between chimpanzee and human based on the human genome sequence reference. Intervals were compared and duplications were classified as shared, chimpanzee only and human only (Methods). Copy number correction was performed based on factoring the number of redundant (duplicated) base pairs in the human genome and the estimated copy number of duplications as determined by WGS depth of coverage. The chimpanzee donor sequence was male. A detailed view for each region is available (http://chimpparalogy.gs.washington.edu and Supplementary Fig. S7). The average per cent lineage-specific duplication per autosome is 1.35 ± 1.22% and 1.33 ± 1.22% for human and chimpanzee, respectively.
*Chromosomes that show an excess of human-only duplications (>2.5% duplication).
†Chromosomes that show an excess of chimpanzee-only duplications (>2.5% duplication).
‡Three chromosomes (2, 4 and 9) account for 16 Mb of the increase (25.8 Mb in total) in shared autosomal duplication content in chimpanzee.

Rate estimates

Three possible scenarios have been put forward to explain the 'excess' of segmental duplications within the human–ape lineage when compared to other genomes[22,23]: frequent *de novo* duplication, a slow culling of duplications by deletion, and/or extensive gene conversion of ancient duplications[11,24,25]. Cross-species comparison of the chimpanzee-only duplications among humans and the great apes revealed that the majority (11 out of 17) of the duplications were restricted to the chimpanzee (multiple hybridization signals were not observed in human, gorilla or orang-utan; Supplementary Table S12). These probably emerged as a consequence of *de novo* segmental duplication after speciation. Six out of seventeen of the chimpanzee-only duplications, however, were also duplicated in the gorilla (and in one case orang-utan). We propose that these apparent duplications arose before the divergence of humans and great apes and have been subsequently deleted within the human lineage, although a small fraction of these (~30%) are expected to be due to lineage-specific sorting in the ancestral chimpanzee–gorilla population[26].

To address further this question and the potential for gene conversion[25], we compared the divergence patterns of shared chimpanzee and human duplications and human-only duplications. For the former case, we limited our analysis to those where only two copies of the duplications existed (binary duplicate patterns) (Supplementary Fig. S8). Using an estimate for chimpanzee–human sequence divergence (0.0131 ± 0.0045 nucleotide substitutions per site, Supplementary Table S13), we classified duplications as occurring before, after or near the time of speciation (Fig. 2; see also Supplementary Table S14). If one examines shared duplications, we note a very small fraction, ~8% by base pairs (3% by count), with a sequence identity consistent with post-speciation gene conversion events. For human-only duplications, 67.0% of the 'new' duplication base pairs show divergence consistent with a *de novo* duplication, whereas the remainder are more divergent, suggesting deletion of a more ancient duplication. Similar results were obtained if chimpanzee-only duplications were considered, although the number of alignments is larger due to the fragmented nature of the chimpanzee genome assembly. These findings closely parallel the results obtained by FISH and suggest that, at the base pair level, *de novo* duplication

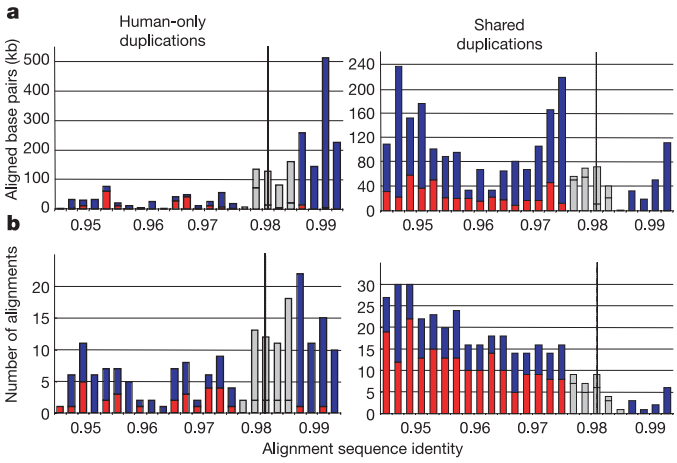

Figure 2 | Sequence identity spectra of human only versus shared duplications. a, b, The sequence identity (0.2% increments) of human only and shared duplication alignments is shown as a function of the total number of base pairs (**a**) and by count (**b**). Only single pairwise alignments were considered for shared duplications (Supplementary Fig. S8). Shared duplications were supported by human WGAC and either chimpanzee WSSD or chimpanzee WGAC. Interchromosomal (red), intrachromosomal (blue) and duplication alignments that lie within 1 s.d. of the chimpanzee–human sequence divergence (grey) are shown (Supplementary Table S13).

followed by deletion have contributed most significantly to the abundance of large, highly identical duplications within the human genome and not gene conversion of older duplications.

Hyperexpansion of chimpanzee segmental duplications

We examined all duplications for copy number differences between human and chimpanzee by estimating their representation in both genomes using the whole-genome shotgun sequencing detection method[11]. Regression analysis between duplications of known copy number and the depth of random sequence show excellent correspondence in both chimpanzee and human genome sequence libraries ($r^2 = 0.953$ and $r^2 = 0.96$, respectively) (Fig. 1a). For every shared duplication interval, we computed the differential copy number (δ) for each 5-kb window of human sequence (July 2003 Assembly) after correction for common repeat sequences (Methods). We limited our analysis to duplication intervals where ten or more consecutive windows (~14 kb in length) showed a copy number difference in either species greater than five ($\delta > 5$).

A total of 296 regions (7.2 Mb) were identified where the human genome showed significant increases in copy number when compared to chimpanzee (Supplementary Table S15). Thirty-three per cent of the human increase (98 out of 296 intervals) mapped within 5 Mb of the centromere, corresponding to 21 out of 29 pericentromeric duplication regions in the human genome. This was significantly different ($P = 0.0002$, Fisher's exact test) compared with the chimpanzee genome, which showed relatively little increase in pericentromeric duplication intervals (13 out of 92). Array comparative genomic hybridization (CGH) between human and chimpanzee genomes (Supplementary Fig. S6) confirms these results and suggests either a genome-wide global expansion of pericentromeric duplications in the human lineage or deletion of such duplication in the chimpanzee lineage.

In contrast, we identified only 92 regions (45 clusters) where the chimpanzee genome showed a significant increase in copy number when compared to human. Although these regions are fewer in number, they correspond to a more marked increase in the amount (22.6 Mb) of shared duplicated sequence that has occurred in the chimpanzee lineage. More than 70% (16.0 out of 22.6 Mb) of the chimpanzee increase mapped to two clusters on chromosomes 2, 4 and 9. One segmental duplication in particular was identified that showed an extraordinary increase in chimpanzee when compared to human. Our analysis indicated that this locus (~40 kb in size) mapped to four regions in human but was represented ~400 times within the chimpanzee genome (Fig. 3). The copies in human showed high sequence identity (99.2%) and mapped to human chromosome 9p24, 9q21 and near the ancestral centromere on 2q12.

Comparative FISH analysis (Fig. 4; see also Supplementary Fig. S9) revealed that the *Pan* hyperexpansion occurred in the common ancestor of bonobo (*Pan paniscus*) and chimpanzee (*Pan troglodytes*) (2–5 million years ago) and was targeted, with the exception of interstitial regions on chromosomes VII and XIII (phylogenetic group designation), exclusively to the subterminal portions of chimpanzee chromosomes. Sequence analysis of one chimpanzee locus (Fig. 3; AC150905) revealed the presence of a single copy of the 36-kb segmental duplication and a 14.5-kb cluster of tandem repetitive repeats (32-bp repeat unit clustered into larger 400–800-bp structures). Sequence similarity searches showed significant sequence identity with previously described subterminal satellite repeats (pCHT7 and pCHT13)[27].

We propose that most of the asymmetrical increase of duplicated DNA in the chimpanzee lineage has emerged as a mechanistic consequence of changes in chromosome structure and not selection. The subterminal caps are an idiosyncratic structural aspect of African great ape chromosomes[28], which are generally regarded as heterochromatic. Similar to human pericentromeric DNA, the regions have served as sinks for duplicative transposition and expansion of particular euchromatic segments. This process has led to an overall increase in chimpanzee genome size of at least 16 Mb since human and chimpanzee separated. It is interesting that the same region that represents the site of chromosome 2 fusion[29] in the human lineage has undergone a segmental duplication hyperexpansion within the subterminal region of chimpanzee chromosomes. This may suggest an inherent instability of this segment of DNA, further extending the association of segmental duplication and chromosomal rearrangement without a direct cause and effect relationship[1].

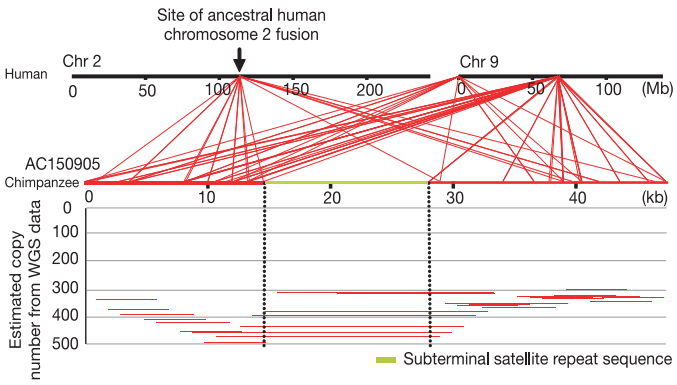

Figure 3 | Sequence structure of chimpanzee subterminal duplication. A schematic diagram depicting the organization of chimpanzee BAC 100G12 (AC150905) is shown. Segmental duplications (red) flank a 32-bp subterminal satellite repeat sequence associated with subterminal portions of great ape chromosomes. A large excess of chimpanzee reads map (on average 20,000 reads with 99.2% sequence identity) to each 5 kb of 'unique' sequence within the duplications, indicative of 300–500 copies of the segmental duplication. By comparison, the human genome assembly shows only four to five copies of this sequence, mapping to 9p24, 9q21 and 2q21.

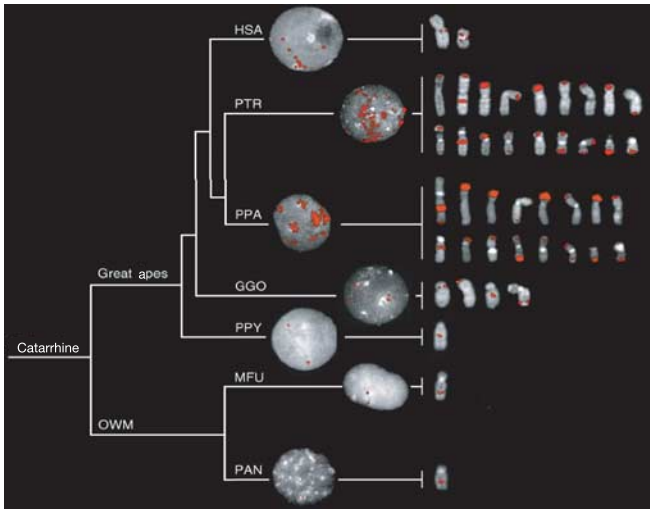

Figure 4 | A chimpanzee hyperexpansion of a shared segmental duplication. A human fosmid DNA clone (WIBR2-1785A6) corresponding to the duplicated region was hybridized against a series of primate chromosomes at metaphase, including human, common chimpanzee (*P. troglodytes*), bonobo (*P. paniscus*), gorilla, orang-utan, macaque and baboon. Hundreds of copies map to the subterminal portions of only chimpanzee and bonobo chromosomes, indicating a lineage-specific duplication expansion 2–6 million years ago. Interstitial chromosome signals are also noted on chromosomes VII and XIII and correspond to cross-hybridization with subterminal satellite repeat sequence. GGO, *Gorilla gorilla*; HSA, *Homo sapiens*; MFU, *Macaca fuscata*; OWM, Old World monkey; PAN, *Papio* anubis; PPA, *Pan paniscus*; PPY, *Pongo pygmaeus*; PTR, *Pan troglodytes*.

Discussion

Our analysis has revealed some important properties regarding the emergence and maintenance of segmental duplications within the human–ape lineage. First, we have determined that a significant fraction of the genome (1.5% or 46 Mb) is specifically duplicated in one lineage but not the other (Methods). Second, both FISH and sequence divergence data indicate that ~60% of these apparent lineage-specific differences are the result of *de novo* duplications, whereas most of the remainder is the result of deletion. In contrast to recent studies of the Y chromosome[25], the impact of gene conversion appears minimal for binary duplicates (<10%). Finally, in addition to qualitative differences in duplication content, we have identified significant (>5) copy number differences among shared human and chimpanzee duplications. These differences have contributed to a net gain of ~26 Mb of segmental duplication within the chimpanzee lineage (Table 1). In total, we conservatively estimate that 70 Mb (2.7%) of euchromatic sequence have been differentially duplicated between the chimpanzee and human, with 4.4 Mb of new genetic material being added on average per million years. Owing to limitations in our genomic duplication detection strategy (>20 kb), ours is almost certainly an underestimate (Methods). Nevertheless, when compared to single-base-pair differences, which account for 1.2% genetic difference, base per base, large segmental duplication events have had a greater impact (2.7%) in altering the genomic landscape of these two species.

METHODS

Duplication analyses. To detect chimpanzee duplications (>1 kb and >90% sequence identity), we performed a WGAC[20] on the Arachne November 2003 chimpanzee genome assembly[21]. We detected a total of 51,573 pairwise alignments corresponding to 136.7 Mb (35,453 non-redundant fragments) of 'duplicated' material. Forty per cent (54.3 Mb) of these fragments localized to unmapped portions of the chimpanzee genome (random assignment). As a second measure of chimpanzee duplication, independent from the genome assembly comparison, we modelled the depth of coverage of chimpanzee WSSD (23.7 million sequence reads)[11] against the human genome reference. The number of reads within 5-kb windows correlated strongly with copy number of duplication ($r^2 = 0.953$) (Fig. 1a). We set our thresholds of duplication detection at 75 reads per 5 kb for autosomes and 44 reads per 5 kb for the sex chromosomes (3 s.d. beyond the mean coverage based on our analysis of unique sequence). We defined a WSSD duplication interval as any region where six out of seven continuous windows showed read depth in excess of autosomal and sex chromosome thresholds. We focused on WSSD regions >20 kb in length (70.6 Mb in total) due to estimated false positive (<1.4%) and negative rates (<6.5%) at this length cut-off (Supplementary Figs S2 and S3). A corresponding chimpanzee segmental duplication database and UCSC genome browser track (http://chimpparalogy.gs.washington.edu) were developed.

WGAC and WSSD duplication intervals were compared between human and chimpanzee by mapping all four tracks onto the human genome reference. We initially categorized DNA as duplicated in human or chimpanzee based on a comparison of these four duplication analyses (chimpanzee WGAC, human WGAC, chimpanzee WSSD and human WSSD) (Fig. 1b–d). Chimpanzee duplication intervals were defined on the human reference genome as the longest interval of contiguous duplication that seeded within at least 20 kb of chimpanzee WSSD (see below). We similarly limited our analysis of human duplications to regions of >94% sequence identity and >20 kb in length. Regions were classified into one of three possible categories: duplicated only in chimpanzee, duplicated only in human or shared between chimpanzee and human.

Validation. Five separate analyses were performed to validate our database of chimpanzee segmental duplication and to provide estimates for false positive and negative detection rates[30] (Supplementary Methods, Supplementary Fig S2–S6 and Supplementary Tables S1–S6).

Expression analysis. Gene expression differences between human and chimpanzee were assessed for five tissues (heart, brain, liver, testis and kidney) using Affymetrix HG U133plus2 arrays (see Supplementary Methods) as described[31]. All primary expression data are publicly available at ArrayExpress (http://www.ebi.ac.uk/arrayexpress/), accession number E-AFMX-11.

Rate estimate. We estimated the amount of human-only (26.5 Mb) and chimpanzee-only (11.4 Mb) duplication and adjusted for copy number based on the WGS depth-of-coverage estimate (WSSD) of each corresponding region (31.9 Mb and 36.2 Mb, respectively) (Table 1). The amount of new duplicated material in chimpanzee was then simply 36.2 − 11.4 Mb (24.8 Mb), whereas the amount of new human autosomal material was corrected for the copy number of the reference human genome assembly (31.9 − (26.5/2.6) = 21.7 Mb). We estimate that there has been a minimum of 46.5 Mb of lineage-specific segmental duplication since separation of chimpanzee and human. We determined that there has been an increase of 7.2 Mb and 22.6 Mb of shared (chimpanzee and human) duplication in the human and chimpanzee lineages, respectively, for regions where the genomic copy number increased by five or more (Supplementary Table S15). Sixteen megabases of the chimpanzee increase is due to a lineage-specific expansion that occurred before the separation of *P. troglodytes* from *P. paniscus* (2 million years ago), but after the separation of human and chimpanzee (6 million years ago); 13.8 Mb (7.2 Mb in human and 6.6 Mb in chimpanzee) is the result of marked changes in copy number of a subset of segmental duplications between the two species. If we estimate that 60% of these bases have emerged by duplication, as opposed to deletion and gene conversion (Supplementary Table S14), we calculate that 0.6(13.8 + 46.5) + 16 Mb = 52.2 Mb have arisen as a result of *de novo* duplication since divergence of the two species. This corresponds to 4.4 Mb of duplication per million years or an effective fixation rate of 3.4 Mb of segmental duplication per million years (assuming chimpanzee–human separation at 6 million years ago and a polymorphism frequency of 0.2). This rate is a lower bound estimate, because sex chromosome duplications as well as autosomal duplications <20 kb in size were not considered owing to reduced power of detection in the chimpanzee lineage. If we extrapolate based on the analysis of duplications in human (Supplementary Table S2), we can calculate an upper bound of *de novo* duplication of 5.5 Mb per million years.

Received 31 March; accepted 30 June 2005.

1. Bailey, J. A., Baertsch, R., Kent, W. J., Haussler, D. & Eichler, E. E. Hotspots of mammalian chromosomal evolution. *Genome Biol.* **5,** R23 (2004).
2. She, X. *et al.* The structure and evolution of centromeric transition regions within the human genome. *Nature* **430,** 857–864 (2004).
3. Armengol, L., Pujana, M. A., Cheung, J., Scherer, S. W. & Estivill, X. Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Hum. Mol. Genet.* **12,** 2201–2208 (2003).
4. Trask, B. *et al.* Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum. Mol. Genet.* **7,** 13–26 (1998).
5. 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).
6. Ventura, M. *et al.* Neocentromeres in 15q24–26 map to duplicons which flanked an ancestral centromere in 15q25. *Genome Res.* **13,** 2059–2068 (2003).
7. Johnson, M. E. *et al.* Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413,** 514–519 (2001).
8. Courseaux, A. & Nahon, J. L. Birth of two chimeric genes in the Hominidae lineage. *Science* **291,** 1293–1297 (2001).
9. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nature Genet.* **37,** 129–137 (2005).
10. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307,** 1434–1440 (2005).
11. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297,** 1003–1007 (2002).
12. Khaitovich, P. *et al.* Regional patterns of gene expression in human and chimpanzee brains. *Genome Res.* **14,** 1462–1473 (2004).
13. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305,** 525–528 (2004).
14. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genet.* **36,** 949–951 (2004).
15. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nature Genet.* **37,** 727–732 (2005).
16. Sharp, A. J. *et al.* Segmental duplications and copy number variation in the human genome. *Am. J. Hum. Genet.* **77,** 78–88 (2005).
17. Fredman, D. *et al.* Complex SNP-related sequence variation in segmental genome duplications. *Nature Genet.* **36,** 861–866 (2004).
18. Stankiewicz, P. & Lupski, J. R. Genomic architecture, rearrangements and genomic disorders. *Trends Genet.* **18,** 74–82 (2002).
19. Zhang, L., Lu, H. H., Chung, W. Y., Yang, J. & Li, W. H. Patterns of segmental duplication in the human genome. *Mol. Biol. Evol.* **22,** 135–141 (2005).
20. 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).
21. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* doi:10.1038/nature04072 (this issue).

22. Tuzun, E., Bailey, J. A. & Eichler, E. E. Recent segmental duplications in the working draft assembly of the brown Norway rat. *Genome Res.* **14,** 493–506 (2004).
23. Bailey, J. A., Church, D. M., Ventura, M., Rocchi, M. & Eichler, E. E. Analysis of segmental duplications and genome assembly in the mouse. *Genome Res.* **14,** 789–801 (2004).
24. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
25. Rozen, S. *et al.* Abundant gene conversion between arms of massive palindromes in human and ape Y chromosomes. *Nature* **423,** 873–876 (2003).
26. Chen, F. C. & Li, W. H. 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 (2001).
27. Royle, N. J., Baird, D. M. & Jeffreys, A. J. A subterminal satellite located adjacent to telomeres in chimpanzees is absent from the human genome. *Nature Genet.* **6,** 52–56 (1994).
28. Yunis, J. J. & Prakash, O. The origin of man: a chromosomal pictorial legacy. *Science* **215,** 1525–1530 (1982).
29. Fan, Y., Linardopoulou, E., Friedman, C., Williams, E. & Trask, B. J. Genomic structure and evolution of the ancestral chromosome fusion site in 2q13–2q14.1 and paralogous regions on other human chromosomes. *Genome Res.* **12,** 1651–1662 (2002).
30. Fortna, A. *et al.* Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol.* **2,** E207 (2004).
31. Khaitovich, P. *et al.* Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* (in the press).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Lachmann, I. Hellman and G. Vessere for technical assistance; the Chimpanzee Sequencing and Analysis Consortium for access to the chimpanzee sequence data before publication; A. Force for discussions; and J. Pecotte, S. Warren and J. Rogers for providing some of the primate material used in this study. This work was supported by grants from the National Human Genome Research Institute, the National Institute of General Medical Sciences, Centro di Eccellenza Geni in campo Biosanitario e Agroalimentare, Ministero Italiano della Università e della Ricerca, the European Commission and the Bundesministerium für Bildung und Forschung.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.E.E. (eee@gs.washington.edu).