Supplementary Text

Methods

Species nomenclature

The common names, Latin names, and abbreviations of the species analyzed by sequence, FISH, and/or PCR for this report are as follows: rhesus macaque, *Macaca mulatta*, MMU; Sumatran orangutan, *Pongo abelii*, PAB; gorilla, *Gorilla gorilla*, GGO; chimpanzee, *Pan troglodytes*, PTR; human, *Homo sapiens*, HSA.

BAC screening

High-density replica filters were prepared from BAC libraries CHORI-250, CHORI-253, CHORI-255, and CHORI-251 constructed from rhesus macaque, orangutan, gorilla, and chimpanzee DNA obtained from blood, respectively (http://bacpac.chori.org/) as described (Osoegawa et al. 2000). Seven overlapping oligonucleotide (overgo) probes (McPherson et al. 2001; Thomas et al. 2003) were designed to identify BAC clones containing sequence homologous to chromosome-specific sequence proximal of subtelomeric duplications on human 15qter and to block 3, a segment of human 15q duplicated in other human subtelomeres, but present only on 4q in chimpanzee and gorilla (Trask et al. 1998) (Supplementary Table 1). Human DNA sequence was aligned to orthologous sequences obtained from DNA fragments PCR-amplified from chimpanzee, gorilla and orangutan, and then sequence lacking mismatches was used to design the overgo probes.

Overgo probes were hybridized to BAC library filters as described (Osoegawa et al. 2000), and hybridization images were captured using a Storm 860 phosphoimager (Amersham) and analyzed with ArrayVision Ver6.0 (Imaging Research Inc). Hybridization-positive clones were re-arrayed into 384-well dishes and then used to prepare small colony filters for hybridization with individual probes. The hybridization-positive BAC DNA was purified using an Autogen960 robot, digested with *Hin*dIII, and separated on 1% agarose gels. The gels were stained with SYBR green and scanned with a Fluorimager 595 (Amersham). The gel images were analyzed with Image software 1.3, and contigs were assembled using FPC software with conditions: tolerance: 4 and cutoff: 1e-11 (Marra et al. 1997).

23 BACs were selected for further study based on their probe composition. We sequenced BAC ends as described (Kelley et al. 1999) and performed BLAT searches between BAC end sequences and the human genome assembly (Kent 2002). We performed FISH with BACs to metaphase chromosomes from human and the species from which the BAC was derived (see below). Eight BACs (CH253-41c18, CH253-150k13, CH250-15d24, CH250-483j10, CH255-96p10, CH251-237f19, CH251-35e6 and CH251-549l23) were chosen for sequencing based on their hybridization to the 4q or 15q subtelomeres. Two BACs had already been sequenced by other groups (CH251-35e6, AC150448 and CH251-549l23, AC183669).

As BAC sequence was generated, we performed BLAST searches to the high-throughput genomic sequences (htgs) and the genome survey sequences

(gss) databases (http://www.ncbi.nlm.nih.gov/) to identify overlapping clones. Sequenced macaque BACs CH250-246c20 (AC148620) and CH250-265f14 (AC148535) overlapped and extended our macaque contig. We extended the chimpanzee 4q contig proximally into 4q-specific sequence by sequencing PTB1-134p4 and RP43-179f17. PTB1-134p4 (AC197422) was identified by BLAST searches between human subtelomeric duplication sequences (Linardopoulou et al. 2005) and end sequences of chimpanzee PTB1 library clones (Fujiyama et al. 2002). We used FISH to determine the location(s) of each BAC with a BLAST match and assayed each BAC by PCR for additional subtelomerically duplicated sequences (data available on request). This effort yielded one BAC, PTB1-134p4, whose sequence provides contiguity between the portions of chimpanzee 4q with homology to human 4q and 15q. Van Geel et al. (2002b) identified chimpanzee BAC RP43-179f17 (AC205763) using a probe for sequence lying proximal of the *TUBB4Q* gene on the human 4q subtelomere. We sequenced PTB1-134p4 and RP43-179f17 and assembled their sequences into the chimpanzee 4q subtelomere contig.

Sequencing

All BAC sequences, except that of PTB1-134P4, were generated and refined to produce ‘comparative-grade’ finished sequence, as described (Blakesley et al. 2004). Such sequence data represent an 'enhanced' version of a 'Phase 2 submission'. Specifically, the indicated order and orientation of each sequence contig has been established using one or more of the following: read-pair data

from individual subclones, overlaps with neighboring clones, alignment with available reference sequence (e.g., human), and/or confirmation by PCR testing. In addition, the sequence assembly is generally based on at least 8-fold average coverage in Q20 bases and has been reviewed to rule out gross misassemblies, the low-quality ends of sequence contigs have been trimmed away, and each base is associated with a Phrap-derived quality score. PTB1-134P4 was sequenced to Phase 2-submission level, with 11.9-fold average coverage, but due to clone instability, the sequence is incomplete across the D4Z4 array and has multiple additional gaps. The order and orientation of sequence contigs was established by PCR.

Sequence of overlapping non-human BACs were assembled into contigs using results from BLAST (http://www.ncbi.nlm.nih.gov/BLAST/), cross_match (http://www.phrap.org/), and dotter (Sonnhammer and Durbin 1995) programs. We required overlapping BACs assembled into contigs to match at the level expected for alleles, i.e., to have contiguous overlap, except for sequencing gaps or N's, of ≥99.7% identity counting small insertions/deletions other than sequencing gaps as single events. Overlaps in macaque and orangutan contigs have ≥99.96% identity; overlaps in chimpanzee 4qter contig have 99.7% identity. We also required BACs assembled into contigs and their constituent segmentally duplicated blocks to map to the same chromosome location by FISH. Lower-identity links in chimpanzee 4q were confirmed by PCR and supported by FISH data, which indicate that 4q is the primary and sole common location of PTB-134p4 and the flanking BACs in the chimpanzee 4q contig.

Human sequence was extracted from the March 2006 hg18 assembly, build 36.1, using the UCSC browser (http://www.genome.ucsc.edu/) and corresponds to coordinates chr15: 100,089,153-100,338,915, chr4: 190,968,252-191,273,063, and chr 14: 19,681,735-19,682,661. We note that the current human assembly is an amalgam of 4qA and 4qB; our diagram of 4qA includes only those portions known to derive from a 4qA allele. Also, the browser shows a telomere at the end of the 4q sequence, but we can find no sequence in GenBank that supports this assembly. The distance from the D4Z4 array to the telomere has been estimated to be about 25 kb based on Southern blot analyses of a half-YAC clone (van Geel et al., 2002a).

Note also that we did not incorporate sequence of genome assemblies of 4q and 15q for other primates, because some of the whole-genome shotgun reads incorporated into these assemblies could have derived from paralogous sequences on other chromosomes.

Sequence and phylogenetic analysis

Matches to human subtelomerically duplicated sequences (called blocks) (Linardopoulou et al. 2005) were identified using cross_match (http://www.phrap.org/). OR genes were identified using methods described in (Young et al. 2002). Predicted OR amino acid sequences (with frameshifts corrected where necessary) were aligned using clustalw (http://www.ebi.ac.uk/clustalw/), and an unrooted neighbor-joining tree was constructed using PAUP (Swofford 2003). One thousand bootstrap replicates

were performed. We used Blast2Seq (Tatusova and Madden 1999) to align and calculate the percent amino-acid identity of pairs of genes (blastp option). dS values were generated using codeml (PAML package, version 4, Yang, 2007) in pairwise mode on an in-frame alignment of OR DNA sequences. This alignment was generated using estwise (Birney et al., 2004) to align each sequence to a profile hidden Markov model of OR proteins (built from a hand-curated alignment of all intact mouse and rat ORs using tools from HMMer suite, http://hmmer.janelia.org), processing estwise output using custom perl scripts. Full details and parameters will be provided on request. We used a custom program, PercentIDPlotQ.pl (developed and provided by Eleanor Williams and Elena Linardopoulou) to calculate the Jukes-Cantor corrected nucleotide divergence of orthologous sequences aligned using Mavid (Bray and Pachter 2004). Inverted repeats were identified using Inverted Repeats Finder (http://tandem.bu.edu/cgi-bin/irdb/irdb.exe) (Warburton et al. 2004). We used GeneCon (Sawyer, 1989) (http://www.math.wustl.edu/~sawyer/geneconv) to compute the statistical likelihood of gene conversion-like events. We used the most stringent setting (g0: gscale=0), which prohibits fragments with internal mismatches, and /es setting, which ignores insertions and deletions and scores only mismatches, and the more conservative Bonferroni-corrected Karlin-Altschul p value.

FISH

Cytogenetic locations of BACs were determined by FISH as described (Trask et al. 1998). We also TA-cloned ~5-kb portions of various subtelomerically duplicated sequences (TOPO TA cloning K4550-40, Invitrogen, USA) and pooled these plasmids to label as FISH probes for hybridization to primate metaphase chromosomes as described (Linardopoulou et al. 2005; Trask et al. 1998). PCR primers were designed to amplify segments of duplicated blocks from human BACs (Supplementary Table 5). Plasmids from blocks 6, 7 and 8 were pooled, as were plasmids from blocks 14 and 15. Block-5 and block-9 plasmid pools were hybridized separately. D4Z4 was isolated by digesting human BAC RP11-54d12 with KpnI and cloning the 3.3-kb insert into the pCR2.1-TOPO vector (Invitrogen, USA).

The primate cell lines for these analyses included GM10540, GM10494, GM10879, and a primary peripheral lymphocyte culture (PBL21) for human; AG16618, “Clint” (SOO6006), CRL1847 (“TANK”) and PT5, for chimpanzee; AG05251 and “Machi” for gorilla; GM06213 and CRL1850 (“Puti”) for Sumatran orangutan; and GM03443 for rhesus macaque. GM-, AG-, and “Clint” cell lines were obtained from the Coriell NIGMS cell repository (Camden, NJ, http://www.coriell.org). CRL- lines were obtained from the ATCC (Manassas, VA, http://www.atcc.org). Machi and PT5 were gifts from Huntington Willard's lab.

Chromosome sorting

Chromosome suspensions were prepared from orangutan CRL1850 cells, and chromosomes 14 and 15 were sorted using methods described in (Mefford et al. 2001) . Five thousand chromosomes were sorted into each tube and subjected to PCR assays as described in (Mefford et al. 2001) (Supplementary Table 3). The DNA in replicate tubes of sorted chromosomes was also universally amplified using DOP-PCR, labeled with biotin, and hybridized as a “paint” to cytogenetic preparations of orangutan metaphase cells using published methods (Trask et al. 1998) in order to confirm the identity and relative purity of the sorted chromosomes.

PCR

We performed PCR to assay for subtelomerically duplicated sequences in selected BACs and genomic DNA from orangutan, gorilla, chimpanzee and human as described (Linardopoulou et al. 2005) and with additional block-5 primers designed based on our orangutan sequence. We also performed PCR assays for OR genes L-U and sequences in the 50 kb between ORs S and U; primer design was based on our sequence from orangutan, rhesus macaque, and the human genome assembly. Primer sequences not previously published are given in Supplementary Table 3. Long-range PCR conditions were as described previously (Linardopoulou et al. 2005).

OR gene nomenclature

The official nomenclature for genes A-U is given in Supplementary Table 6. The multiple human paralogs of IJK were called OR-A, OR-B, and OR-C in previous publications (Linardopoulou et al. 2005; Mefford and Trask 2002; Trask et al. 1998). The family/subfamily designations for genes without a human ortholog were tentatively assigned as follows. If the amino-acid identity of the best match found by BLAST in the HORDE database was >60%, we assigned the gene to the same subfamily and family as its best match. If the amino-acid identity of the best match was <60% but >40%, we assigned the gene to the same family, but left the subfamily unassigned. If the best match was <40%, we left both family and subfamily unassigned (only the case for quite diverged pseudogenes). Note that for R, the orangutan ortholog, but not the macaque ortholog, received a subfamily assignment. They both match HORDE OR4F14P best, but the identity is just above 60% for orangutan and just below 60% for macaque.

References

Birney, E., Clamp, M., and R. Durban. 2004. GeneWise and Genomewise. *Genome Res* **14**:988-995.

Blakesley, R.W., N.F. Hansen, J.C. Mullikin, P.J. Thomas, J.C. McDowell, B. Maskeri, A.C. Young, B. Benjamin, S.Y. Brooks, and B.I. Coleman. 2004. An intermediate grade of finished genomic sequence suitable for comparative analyses. *Genome Res* **14:** 2235-2244.

Bray, N. and L. Pachter. 2004. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res* **14:** 693-699.

Fujiyama, A., H. Watanabe, A. Toyoda, T.D. Taylor, T. Itoh, S.F. Tsai, H.S. Park, M.L. Yaspo, H. Lehrach, and Z. Chen. 2002. Construction and analysis of a human-chimpanzee comparative clone map. *Science* **295:** 131-134.

Kelley, J.M., C.E. Field, M.B. Craven, D. Bocskai, U.J. Kim, S.D. Rounsley, and M.D. Adams. 1999. High throughput direct end sequencing of BAC clones. *Nucleic Acids Res* **27:** 1539-1546.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664.

Linardopoulou, E.V., E.M. Williams, Y. Fan, C. Friedman, J.M. Young, and B.J. Trask. 2005. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* **437:** 94-100.

Marra, M.A., T.A. Kucaba, N.L. Dietrich, E.D. Green, B. Brownstein, R.K. Wilson, K.M. McDonald, L.W. Hillier, J.D. McPherson, and R.H. Waterston. 1997. High throughput fingerprint analysis of large-insert clones. *Genome Res* **7:** 1072-1084.

McPherson, J.D., M. Marra, L. Hillier, R.H. Waterston, A. Chinwalla, J. Wallis, M. Sekhon, K. Wylie, E.R. Mardis, and R.K. Wilson. 2001. A physical map of the human genome. *Nature* **409:** 934-941.

Mefford, H.C., E. Linardopoulou, D. Coil, G. van den Engh, and B.J. Trask. 2001. Comparative sequencing of a multicopy subtelomeric region containing olfactory receptor genes reveals multiple interactions between non-homologous chromosomes. *Hum Mol Genet* **10:** 2363-2372.

Mefford, H.C. and B.J. Trask. 2002. The complex structure and dynamic evolution of human subtelomeres. *Nat Rev Genet* **3:** 91-102.

Osoegawa, K., M. Tateno, P.Y. Woon, E. Frengen, A.G. Mammoser, J.J. Catanese, Y. Hayashizaki, and P.J. de Jong. 2000. Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. *Genome Res* **10:** 116-128.

Sawyer, S. 1989. Statistical tests for detecting gene conversion. *Mol Biol Evol* **6**, 526-38.

Sonnhammer, E.L. and R. Durbin. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167:** GC1-10.

Swofford, D.L. 2003. *Phylogenetics Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates.

Tatusova, T.A. and T.L. Madden. 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* **174:** 247-250.

Thomas, J.W., J.W. Touchman, R.W. Blakesley, G.G. Bouffard, S.M. Beckstrom-Sternberg, E.H. Margulies, M. Blanchette, A.C. Siepel, P.J. Thomas, and J.C. McDowell. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788-793.

Trask, B.J., C. Friedman, A. Martin-Gallardo, L. Rowen, C. Akinbami, J. Blankenship, C. Collins, D. Giorgi, S. Iadonato, F. Johnson, and W.L. Kuo. 1998. Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum Mol Genet* **7:** 13-26.

van Geel, M., M.C. Dickson, A.F. Beck, D.J. Bolland, R.R. Frants, S.M. van der Maarel, P.J. de Jong, and J.E. Hewitt. 2002a. Genomic analysis of human chromosome 10q and 4q telomeres suggests a common origin. *Genomics* **79:** 210-217.

van Geel, M., E.E. Eichler, A.F. Beck, Z. Shan, T. Haaf, S.M. van der Maarel, R.R. Frants, and P.J. de Jong. 2002b. A cascade of complex subtelomeric duplications during the evolution of the hominoid and Old World monkey genomes. *Am J Hum Genet* **70:** 269-278.

Warburton, P.E., J. Giordano, F. Cheung, Y. Gelfand, and G. Benson. 2004. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res* **14:** 1861-1869.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**:1586-1591.

Young, J.M., C. Friedman, E.M. Williams, J.A. Ross, L. Tonnes-Priddy, and B.J. Trask. 2002. Different evolutionary processes shaped the mouse and human olfactory receptor gene families. *Hum Mol Genet* **11:** 535-546.