LETTERS

DNA sequence and analysis of human chromosome 8

Chad Nusbaum[1], Tarjei S. Mikkelsen[1], Michael C. Zody[1], Shuichi Asakawa[2], Stefan Taudien[3], Manuel Garber[1], Chinnappa D. Kodira[1], Mary G. Schueler[4], Atsushi Shimizu[2], Charles A. Whittaker[1]†, Jean L. Chang[1], Christina A. Cuomo[1], Ken Dewar[1]†, Michael G. FitzGerald[1], Xiaoping Yang[1], Nicole R. Allen[1], Scott Anderson[1], Teruyo Asakawa[2], Karin Blechschmidt[3], Toby Bloom[1], Mark L. Borowsky[1], Jonathan Butler[1], April Cook[1], Benjamin Corum[1], Kurt DeArellano[1], David DeCaprio[1], Kathleen T. Dooley[1], Lester Dorris III[1], Reinhard Engels[1], Gernot Glöckner[3], Nabil Hafez[1], Daniel S. Hagopian[1], Jennifer L. Hall[1], Sabine K. Ishikawa[2], David B. Jaffe[1], Asha Kamat[1], Jun Kudoh[2], Rüdiger Lehmann[3], Tashi Lokitsang[1], Pendexter Macdonald[1], John E. Major[1], Charles D. Matthews[1], Evan Mauceli[1], Uwe Menzel[3]†, Atanas H. Mihalev[1], Shinsei Minoshima[2]†, Yuji Murayama[2], Jerome W. Naylor[1], Robert Nicol[1], Cindy Nguyen[1], Sinéad B. O'Leary[1], Keith O'Neill[1], Stephen C. J. Parker[1]†, Andreas Polley[3]†, Christina K. Raymond[1], Kathrin Reichwald[3]†, Joseph Rodriguez[1], Takashi Sasaki[2], Markus Schilhabel[3], Roman Siddiqui[3], Cherylyn L Smith[1], Tam P. Sneddon[5], Jessica A. Talamas[1], Pema Tenzin[1], Kerri Topham[1], Vijay Venkataraman[1], Gaiping Wen[3]†, Satoru Yamazaki[2], Sarah K. Young[1], Qiandong Zeng[1], Andrew R. Zimmer[1], Andre Rosenthal[3]†, Bruce W. Birren[1], Matthias Platzer[3], Nobuyoshi Shimizu[2] & Eric S. Lander[1]

The International Human Genome Sequencing Consortium (IHGSC) recently completed a sequence of the human genome[1]. As part of this project, we have focused on chromosome 8. Although some chromosomes exhibit extreme characteristics in terms of length, gene content, repeat content and fraction segmentally duplicated, chromosome 8 is distinctly typical in character, being very close to the genome median in each of these aspects. This work describes a finished sequence and gene catalogue for the chromosome, which represents just over 5% of the euchromatic human genome. A unique feature of the chromosome is a vast region of ~15 megabases on distal 8p that appears to have a strikingly high mutation rate, which has accelerated in the hominids relative to other sequenced mammals. This fast-evolving region contains a number of genes related to innate immunity and the nervous system, including loci that appear to be under positive selection[2]—these include the major defensin (DEF) gene cluster[3,4] and *MCPH1*[5,6], a gene that may have contributed to the evolution of expanded brain size in the great apes. The data from chromosome 8 should allow a better understanding of both normal and disease biology and genome evolution.

The finished sequence of chromosome 8 contains 145,556,489 bases and is interrupted by only four euchromatic gaps, one gap at the 8p telomere and one gap containing the centromeric heterochromatin (Fig. 1 and Supplementary Table S1). These gaps are refractory to current cloning and mapping technology. The estimated total size of the euchromatic gaps is 427 kilobases (kb), based on direct sizing of three gaps and estimation of the remaining two gaps at the genome-wide average of ~100 kb each. This corresponds to ~0.3% of the euchromatic length of the chromosome, similar to the genome average[1,7–11]. In all, 182.3 megabases (Mb) of finished sequence were generated by the Broad Institute of MIT and Harvard (formerly Whitehead Institute/MIT Center for Genome Research (WICGR)), 27.9 Mb by Keio University School of Medicine, 8.4 Mb by the Institute of Molecular Biotechnology in Jena, and 5.8 Mb by 10 other groups (Supplementary Tables S2 and S3). These sequences (which include overlap) were combined to yield the finished path (see Methods).

We assessed the local accuracy of the clone path by aligning paired-end sequences from a human Fosmid library (WIBR2, representing ×10 physical coverage) to the finished sequence[7]. Errors in the clone path were detected by identifying discrepancies between the predicted and observed distances between Fosmid ends[7]. This revealed two deleted clones, which were replaced. Finally, an independent quality assessment exercise commissioned by NHGRI estimated the accuracy of the finished sequence at less than 1 error in 100,000 bases[12] (J. Schmutz, personal communication).

Several analyses support the idea that nearly the entire euchromatic region of chromosome 8 is present and accurately represented. From the well-curated RefSeq[13] data set 681 transcripts (from 573 unique genes) mapped to chromosome 8. All but one of these are present and complete in the finished sequence. The finished sequence shows excellent co-linearity with the genetic map[14] (Supplementary

[1]Broad Institute of MIT and Harvard, 320 Charles St, Cambridge, Massachusetts 02141, USA. [2]Department of Molecular Biology, Keio University School of Medicine, 35 Shinanomachi, Shinjuku-ku, Tokyo 160-8582, Japan. [3]Genome Analysis, Institute of Molecular Biotechnology, Beutenbergstrasse 11, Jena 07745, Germany. [4]National Human Genome Research Institute, National Institutes of Health, 50 South Drive Rm 5529, Bethesda, Maryland 20982, USA. [5]HUGO Gene Nomenclature Committee (HGNC), The Galton Laboratory, Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK. †Present addresses: MIT Center for Cancer Research, 77 Massachusetts Avenue E18-570, Cambridge, Massachusetts 02139, USA (C.A.W.); McGill University and Genome Quebec Innovation Centre, Montreal, Quebec H3A 1A4, Canada (K.D.); Department of Genetics and Pathology, Uppsala University, SE-751 85 Uppsala, Sweden (U.M.); Photon Medical Research Center, Hamamatsu University School of Medicine, Handayama, Hamamatsu, Shizuoka 431-3192, Japan (S.M.); Boston University Bioinformatics and Systems Biology Program, 24 Cummington St, Boston, Massachusetts 02215, USA (S.C.J.P.); TraitGenetics GmbH, Am Schwabeplan 1b, 06466 Gatersleben, Germany (A.P.); University Clinic for Child and Adolescent Psychiatry, University of Duisburg-Essen, Virchowstr. 174, 45147 Essen, Germany (K.R.); GSF-Forschungszentrum für Umwelt und Gesundheit, Ingolstädter Landstraße 1, 85674 Neuherberg, Germany (G.W.); Signature Diagnostics AG, Voltaireweg 4B, 14469 Potsdam, Germany (A.R.).

Fig. S1). Among 247 sequence-based genetic markers (Supplementary Table S4) there are six discrepancies. One discrepancy consists of eight markers and spans a region in 8p23 known to be the site of a polymorphic inversion in the human population[15,16] (see below). Five discrepancies each consist of single markers out of order by one position; all occur in small regions where the genetic map shows no recombination in one of the two sexes (Supplementary Table S4). The sequence also shows good agreement with the radiation hybrid (RH) map[17] (Supplementary Table S5).

We produced a manually curated gene catalogue, containing 793 gene loci and 301 pseudogene loci (see Methods). The catalogue includes all previously known genes on chromosome 8 (Table 1). According to the Hawk2 categorization scheme[18], there are 614 'known' genes, 109 'novel CDS', 43 'novel transcripts', 14 'putatives' and 13 'gene fragments'. The small set of novel and putative categories were annotated by spliced expressed sequence tag (EST) evidence only; some 'putative novel' loci may prove to be pseudogenes. Comparison of manual annotation performed at the Broad Institute of MIT and Harvard to manual annotation for specific regions done at Jena and Keio indicated that they were largely the same, and that virtually all differences could be attributable to edge effects (see Supplementary Information).

Full-length transcripts of known genes contain an average of 9.9 exons, comparable to recently published reports[8–11,19], have an average length of 3,056 base pairs (bp), and internal exons have an average length of 155 bp. There is evidence of extensive alternate splicing. Gene loci have an average of 4.1 distinct transcripts, with 63% having at least two transcripts, values that are similar to recent reports[8,9,11,20]. Of the 301 pseudogenes on chromosome 8, ~84% are processed pseudogenes arising from retrotransposition; the remaining 16% are unprocessed. We also identified 13 tRNA genes (Supplementary Table S6). Examples of genes that represent extremes from these averages are described in Supplementary Information.

Several aspects of the genome landscape are notable. The overall gene density is 5.6 genes Mb^{-1}, below the genome average of ~10 genes Mb^{-1}. Gene distribution is highly heterogeneous, with 44 gene deserts (500 kb without a coding gene, Supplementary Table S7) that together comprise 41.9 Mb or ~29% the total length. The overall G+C content is 39.2%, but varies substantially across the chromosome (Fig. 1). Nearly half of the chromosome is composed of repeat sequences, with transposable element fossils comprising 44.5%, low complexity sequence (including simple sequence repeats and satellite sequences) comprising 1.8%, and segmental duplications comprising ~2.1% (with interchromosomal and intrachromosomal duplications at ~1.5% each, with some sequence included in both categories) (E. Eichler and X. She, personal communication).

Chromosome 8 is the first human autosome and one of only two chromosomes (the other being chromosome X[20]) for which sequences span the entire pericentromeric region. The regions on both arms stretch from unique euchromatin through pericentromeric satellites and into the higher-order alpha-satellite array (Fig. 2). Three variant higher-order repeat units populate the chromosome 8 higher-order array, *D8Z2* (ref. 21 and Supplementary Information). The proximal termini of both the 8p and 8q sequence contigs are comprised of nine copies of the 1.9-kb unit. The p and q arm higher-order units are highly identical to each other (96–98%) and occur in the same head-to-tail orientation, indicating that these sequences sample the edges of the chromosome 8-specific array. Analysis of the finished pericentromeric sequence of chromosome 8 is essential to test and further develop primate centromere evolution hypotheses using an autosomal model.

The most striking feature on chromosome 8 emerges from evolutionary and population genetic comparisons (Fig. 3). The most distal 15 Mb on chromosome 8p show an extremely high divergence between human and chimpanzee (0.021 substitutions per site, 4.0 s.d. above the mean of 0.012). The region also shows a strikingly high polymorphism rate in the human population (0.0018, 3.2 s.d. above the mean of 0.0010). The peak divergence reaches 0.032 (8.6 s.d.), and diversity 0.0028 (7.1 s.d.), across a 1-Mb region (3.3–4.3 Mb) overlapping the *CSMD1* gene. This is the highest divergence level seen across all autosomes and chromosome X. Only regions of chromosome Y may be more rapidly diverging, driven by the high mutation rate in the male germ line. We excluded trivial explanations for this observation, such as unresolved segmental duplications (Supplementary Information). Diversity is also locally high in the chimpanzee, although the data are more limited.

The high rate of divergence and diversity at distal 8p might reflect either an extraordinary mutation rate or population genetic history. The latter alternative would require an unusually long coalescence time to the most recent common ancestor over a very large region; this would be remarkable inasmuch as local coalescence times tend to be correlated over short distances, as the correlation falls below 0.5 within 20 kb (ref. 22). We sought to resolve the issue by examining the divergence rates with more distant mammalian species, where the impact of population genetic history should be negligible.

Comparison of ancestral interspersed repeats in the human, dog[23]

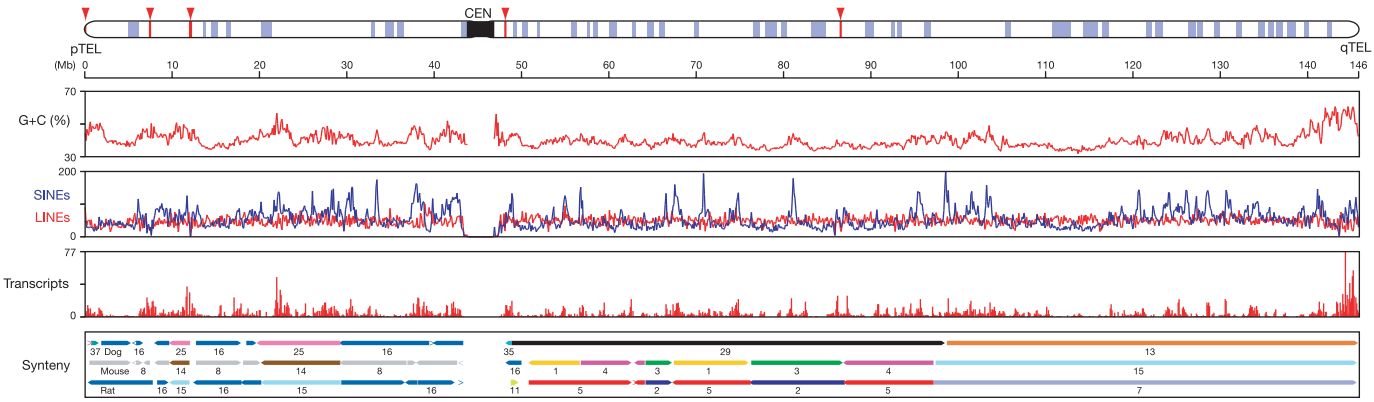

Figure 1 | Overview of human chromosome 8. The features are addressed in the order of top to bottom. In the cartoon, blue shading indicates gene deserts (≥500 kb with no transcript, Supplementary Table S7); telomeres (pTEL and qTEL), the centromere (CEN) and euchromatic sequence gaps (red lines) are indicated. The following features are represented in discrete windows of 100 kb: G+C content (on a scale from 30–70%); densities of LINEs (long interspersed nucleotide elements; red) and SINEs (short interspersed nucleotide elements; blue); and densities of transcripts (all are counts of elements). The box at the bottom shows blocks of conserved synteny (100-kb resolution) with dog, mouse and rat as determined for this work. Chromosomes are numbered, and are coloured arbitrarily for ease of distinction.

Table 1 | Chromosome 8 gene content

Category	Gene number	Gene percentage	Gene length (bp)*	Number of alternative transcripts	Transcript length (bp)†	Number of exons per transcript‡	Internal exon length (bp)§	Intron length (bp)‖	CpG-5′ association¶
Known gene	614	77	81,744	4.1	3,056	9.9	155 ($n = 5,725$)	9,630 ($n = 7,710$)	77
Novel transcript	43	5	96,268	1.8	1,116	3.8	146 ($n = 127$)	27,207 ($n = 248$)	42
Putative gene	14	2	45,433	1.2	714	2.4	123 ($n = 21$)	23,787 ($n = 57$)	36
Novel CDS	109	14	21,890	1.9	1,142	5.0	138 ($n = 487$)	5,103 ($n = 625$)	29
Gene fragment	13	2	648	1.0	648	1.0	-	-	8
Total	793	-	72,334	-	-	-	-	-	-
Pseudogene	301	28	1,334	1.0	875	1.3	195 ($n = 50$)	1,430 ($n = 97$)	5

* Average chromosomal distance from beginning of 5′-most exon to 3′-most exon in all transcripts in a gene.
† Average length summed across the footprint of all exons in all transcripts in a gene—total exon space per gene.
‡ Average number of exons in transcripts. Exons common to different transcripts were counted once per transcript.
§ Average length of exons using the footprint of all non-terminal exons of all transcripts in a gene. Unique overlapping exons or contained exons are counted separately, making this an average length of unique exons in a gene.
‖ Average length of unique introns in a gene. In the case of exon skipping, both the shorter and longer versions of the overlapping introns were counted towards the average.
¶ Percentage of genes with a transcript having a CpG island (as assessed by FirstEF) within −2 kb and +1 kb of transcription start.

and mouse[24] genomes reveals that the region exhibits above-average lineage-specific divergence rates on all three lineages across 100 million years of evolution, but that the rate is the most elevated relative to the genome-wide mean in the lineage leading to humans. The greatest elevation is seen in the most distal 6 Mb of 8p, where the ancestral interspersed repeat divergence rates in the orthologous sequences have been 0.19 (3.3 s.d. above the mean of 0.14) on the human lineage and 0.41 (1.0 s.d. above the mean of 0.38) in the mouse lineage since the primate–rodent split, and 0.24 (1.9 s.d. above the mean of 0.20) in the dog lineage since the divergence from the common boreo-eutherian ancestor.

The biological basis for the apparently high mutation rate is unclear. Three major factors have been associated with high mutation rates in the human genome: proximity to telomeres, high recombination rate and high A+T content[25,26]. The region on chromosome 8p has all three factors. The mean sex-averaged recombination rate across the first 6 Mb is 2.7 cM Mb^{-1}, with a 1-Mb window peak of 3.5, as compared to the genome-wide average of 1.2. The region from 2.5–6 Mb is 62% A+T, as compared to a genome-wide average of 59%. It is unusual in this regard, because subtelomeric regions with high recombination rates are typically (A+T)-poor. Notably, the region is not subtelomeric in the mouse, where the lowest rate elevation is observed.

The distal region on chromosome 8p also contains at least two loci that appear to be undergoing positive selection (Fig. 3). The first locus is the major cluster of defensin genes, which lies within the region of high mutation (5.5–7.5 Mb), although ~2.5 Mb from the peak. The defensin genes express small cationic antimicrobial

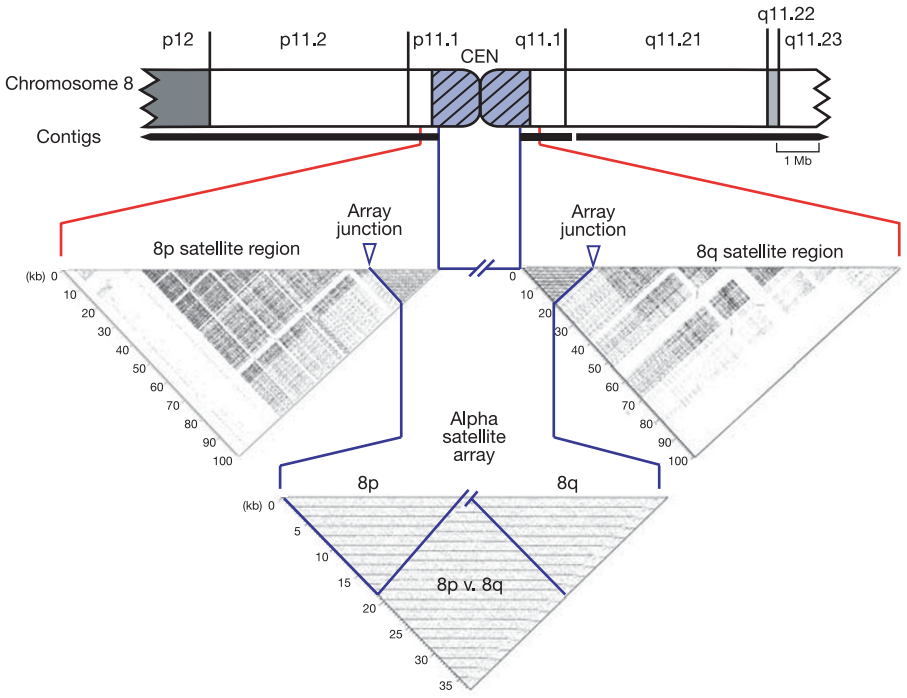

Figure 2 | 8p and 8q pericentromeric contigs extend into chromosome 8-specific higher-order alpha satellite, *D8Z2*. The pericentromeric region of chromosome 8 is shown as a truncated ideogram with the extent of sequence coverage shown below by black bars. Dotter plots show self–self alignments of the most proximal ~100 kb from each arm including ~36 kb of the chromosome-specific alpha satellite array (*D8Z2*). Junctions between the arm-specific satellite region and *D8Z2* are marked with blue arrows. Dark blocks indicate the highly repetitive nature of the satellite region and mark similarity between monomers within each satellite family. Gaps in the dark blocks occur where interspersed elements (LINEs, SINEs and long terminal repeats) interrupt the satellite sequences. In the alpha satellite array dotter plot (bottom), *D8Z2* from 8p (~18 kb) is joined with that of 8q (~18 kb). The plot reveals the periodic nature of the centromeric, higher-order alpha satellite array with black horizontal lines indicating near identity of sequences spaced at ~1.9-kb intervals. The regions outlined in blue are self–self alignments ('8p' and '8q'), whereas the remaining rectangular region of the plot is an alignment of 8p versus 8q *D8Z2*.

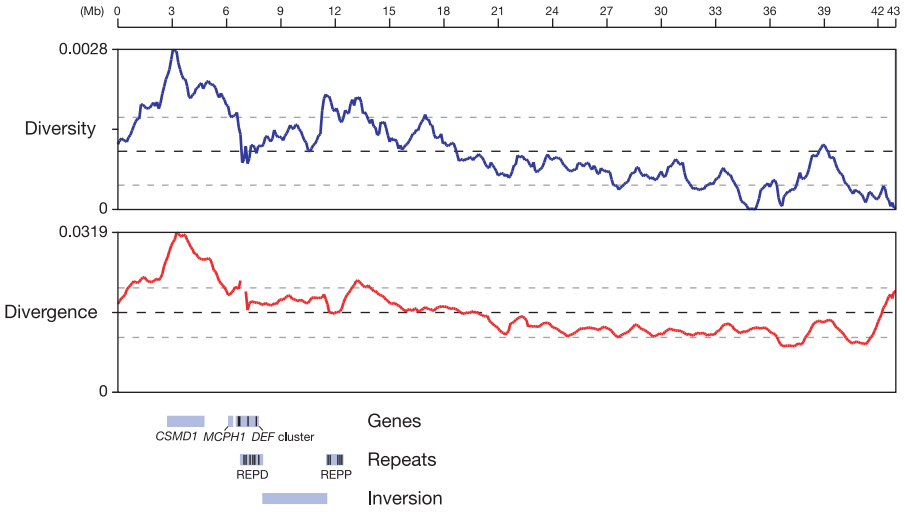

Figure 3 | Diversity and divergence on 8p. Coloured lines indicate the distribution of human diversity (blue) and human–chimpanzee divergence (red). Values of genome averages and of 2 standard deviations from the means are indicated (dark and light dashed lines, respectively). Features mentioned in the text are indicated in the bottom panel, including genes, two low copy repeats (LCRs) and the common 8p23 inversion. Vertical ticks in the LCR boxes indicate olfactory receptor genes or pseudogenes, and vertical ticks in the DEF cluster boxes represent individual defensin (DEF) genes. There is a discontinuity in the divergence plot from 6.98 to 8.13 Mb. This region, corresponding to the REPD repeat, is also highly duplicated in the chimpanzee, making it impossible to align sequence with high enough confidence to call divergence.

peptides crucial to the innate immune response[27]. Studies[2,3] have suggested that defensins have been under positive selection, with a high ratio of non-synonymous to synonymous changes detected in the mature peptide coding exon. Moreover, gene and segmental duplication within the cluster have led to extensive copy number[28,29] and haplotype[30] polymorphism within and across populations, which are thought to influence variation in disease susceptibility and contribute to ongoing adaptive evolution in both the human and chimpanzee species. The second locus showing positive selection is *MCPH1*, mutations in which cause microcephaly (Online Mendelian Inheritance in Man (OMIM): 251200); there is clear evidence of accelerated non-synonymous divergence correlating with the expansion of brain size throughout the lineage from simian ancestors to the human and chimpanzee[4,5].

To investigate the diversity of copy number in the defensin clusters, we resequenced several dozen polymerase chain reaction (PCR) products from representative intervals from *DEFB105A* (beta-defensin cluster) and *DEFA1* (alpha-defensin cluster) in 14 chimpanzees, 1 gibbon, 1 macaque and 4 breeds of dog (see Methods and Supplementary Information). In all species studied, the gene family has multiple members, and the members are more similar within a species than across species. Thus, the defensin clusters have either independently duplicated in each species or have undergone gene conversion events within species.

Finally, we note that the majority of the genes in the region of high divergence in distal 8p play important roles in development or signalling in the nervous system. Notably, the extremely large *CSMD1* gene, which lies at the peak of divergence and diversity, is widely expressed in brain tissues. High regional mutation rates and positive selection are generally assumed to be distinct, but it is possible that the former may facilitate the latter by increasing the rate of appearance of potentially advantageous single, or interacting, alleles (see also ref. 31). It is intriguing to speculate whether the accelerated divergence rate of this region has contributed to the rapid expansion and evolution of the primate brain.

METHODS

See Supplementary Information for details on clone path building, generation of sequence map, sizing of gaps and gene annotation. The final version of the clone path is available in AGP format (see http://www.ncbi.nlm.nih.gov/genome/guide/glossary.htm) at http://www.broad.mit.edu/tools/data/data-human.html.

Gene amplification and sequencing. TBLASTN (http://www.ncbi.nlm.nih.gov/BLAST) was used to identify *DEFB105* and *DEFA1* orthologues in 16 chimpanzees, 1 gibbon, 1 macaque and 4 dog breeds (akita, golden retriever, greyhound and mastiff). PCR primers for gene amplification were designed using Primer3 (http://frodo.wi.mit.edu/primer3) based on the species reference sequence. Human and macaque primers were used for gibbon. Amplified products were cloned, and for each individual/gene combination, 48 or 96 clones were sequenced.

Haplotype analysis. Neighbourhood Quality Standard[32] (NQS) scores were computed for all sequenced products using the published constraints[32]. Reads were trimmed to the first and last three consecutive NQS bases, and aligned to the reference sequence using PatternHunter (http://www.bioinformaticssolutions.com). Multiple sequence alignments were built from the pairwise alignments and inspected to find SNPs that were: at NQS bases, supported by at least two reads, and in a ten base window where not more than two other variations were observed. To minimize false positives due to errors during PCR amplification, we restricted our analysis to haplotypes that differed in >3 bases.

Received 5 August; accepted 6 October 2005.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2. Vallender, E. J. & Lahn, B. T. Positive selection on the human genome. *Hum. Mol. Genet.* **13** (suppl. 2), R245–R254 (2004).
3. Maxwell, A. I., Morrison, G. M. & Dorin, J. R. Rapid sequence divergence in mammalian β-defensins by adaptive evolution. *Mol. Immunol.* **40,** 413–421 (2003).
4. Xiao, Y. *et al.* A genome-wide screen identifies a single β-defensin gene cluster in the chicken: implications for the origin and evolution of mammalian defensins. *BMC Genom.* **5,** 56 (2004).
5. Evans, P. D., Anderson, J. R., Vallender, E. J., Choi, S. S. & Lahn, B. T. Reconstructing the evolutionary history of microcephalin, a gene controlling human brain size. *Hum. Mol. Genet.* **13,** 1139–1145 (2004).
6. Evans, P. D. *et al.* Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. *Science* **309,** 1717–1720 (2005).
7. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
8. Grimwood, J. *et al.* The DNA sequence and biology of human chromosome 19. *Nature* **428,** 529–535 (2004).
9. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 10. *Nature* **429,** 375–381 (2004).
10. Martin, J. *et al.* The sequence and analysis of duplication-rich human chromosome 16. *Nature* **432,** 988–994 (2004).
11. Nusbaum, C. *et al.* DNA sequence and analysis of human chromosome 18. *Nature* **437,** 551–555 (2005).
12. Schmutz, J. *et al.* Quality assessment of the human genome sequence. *Nature* **429,** 365–368 (2004).
13. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq):

a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33,** D501–D504 (2005).
14. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31,** 225–226 (2002).
15. Giglio, S. *et al.* Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am. J. Hum. Genet.* **68,** 874–883 (2001).
16. Shimokawa, O. *et al.* Molecular characterization of inv dup del(8p): analysis of five cases. *Am. J. Med. Genet. A* **128,** 133–137 (2004).
17. Deloukas, P. *et al.* A physical map of 30,000 genes. *Science* **282,** 744–746 (1998).
18. Ashurst, J. L. *et al.* The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res.* **33,** D459–D465 (2005).
19. Hillier, L. W. *et al.* Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature* **434,** 724–731 (2005).
20. Ross, M. T. *et al.* The DNA sequence of the human X chromosome. *Nature* **434,** 325–337 (2005).
21. Ge, Y., Wagner, M. J., Siciliano, M. & Wells, D. E. Sequence, higher order repeat structure, and long-range organization of alpha satellite DNA specific to human chromosome 8. *Genomics* **13,** 585–593 (1992).
22. Reich, D. E. *et al.* Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genet.* **32,** 135–142 (2002).
23. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438,** 803–819 (2005).
24. Mouse Genome Sequencing Consortium, Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
25. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
26. Hellmann, I. *et al.* Why do human diversity levels vary at a megabase scale? *Genome Res.* **15,** 1222–1231 (2005).
27. Lehrer, R. I. Primate defensins. *Nature Rev. Microbiol.* **2,** 727–738 (2004).
28. Hollox, E. J., Armour, J. A. & Barber, J. C. Extensive normal copy number variation of a β-defensin antimicrobial-gene cluster. *Am. J. Hum. Genet.* **73,** 591–600 (2003).
29. Mars, W. M. *et al.* Inheritance of unequal numbers of the genes encoding the human neutrophil defensins HP-1 and HP-3. *J. Biol. Chem.* **270,** 30371–30376 (1995).
30. Taudien, S. *et al.* Polymorphic segmental duplications at 8p23.1 challenge the determination of individual defensin gene repertoires and the assembly of a contiguous human reference sequence. *BMC Genom.* **5,** 92 (2004).
31. Wyckoff, G. J., Malcom, C. M., Vallender, E. J. & Lahn, B. T. A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate. *Trends Genet.* **21,** 381–385 (2005).
32. Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407,** 513–516 (2000).

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

Acknowledgements We thank L. Gaffney for help with figures and tables; L. French and her group at the Sanger Institute for attempting fibre FISH analysis to size some clone gaps in the tiling path of chromosome 8; E. Eichler and X. She for sharing their data on segmental duplications; T. Furey for help with lists of genetic markers and placement of RefSeq genes; M. Kamal for assistance and advice with synteny analysis; and K. Lindblad-Toh for sharing data from the dog genome project. We also acknowledge the HUGO Gene Nomenclature Committee (S. Povey, chair) for assigning official gene symbols. We are deeply grateful to all the members, present and past, of the Genome Sequencing Platform of the Broad Institute (and Whitehead Center for Genome Research), Keio University School of Medicine and the Institute of Molecular Biology at Jena for their dedication and for the consistent high quality of their data that made this work possible. This work was supported by grants from the National Human Genome Research Institute, RIKEN, the 'Research for the Future' Program from the Japan Society for the Promotion of Science (JSPS), the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT), the Federal German Ministry of Education, Research and Technology, and the Thüringer Kultusministerium.

Author Information Accession numbers for all clones contributing to the finished sequence of human chromosome 8 can be found in Supplementary Table S2. The updated human chromosome 8 sequence can be accessed through GenBank accession number NC_000008. 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 C.N. (chad@broad.mit.edu) or N.S. (shimizu@dmb.med.keio.ac.jp).