LETTERS

Analysis of the DNA sequence and duplication history of human chromosome 15

Michael C. Zody[1], Manuel Garber[1], Ted Sharpe[1], Sarah K. Young[1], Lee Rowen[2], Keith O'Neill[1], Charles A. Whittaker[1]†, Michael Kamal[1], Jean L. Chang[1], Christina A. Cuomo[1], Ken Dewar[1]†, Michael G. FitzGerald[1], Chinnappa D. Kodira[1], Anup Madan[2]†, Shizhen Qin[2], Xiaoping Yang[1], Nissa Abbasi[2]†, Amr Abouelleil[1], Harindra M. Arachchi[1], Lida Baradarani[2]†, Brian Birditt[2]†, Scott Bloom[2], Toby Bloom[1], Mark L. Borowsky[1], Jeremy Burke[2], Jonathan Butler[1], April Cook[1], Kurt DeArellano[1], David DeCaprio[1], Lester Dorris III[1], Monica Dors[2], Evan E. Eichler[3], Reinhard Engels[1], Jessica Fahey[2]†, Peter Fleetwood[2]†, Cynthia Friedman[4], Gary Gearin[1], Jennifer L. Hall[1], Grace Hensley[2]†, Ericka Johnson[2]†, Charlien Jones[1], Asha Kamat[1], Amardeep Kaur[2], Devin P. Locke[3], Anuradha Madan[2]†, Glen Munson[1], David B. Jaffe[1], Annie Lui[1], Pendexter Macdonald[1], Evan Mauceli[1], Jerome W. Naylor[1], Ryan Nesbitt[2], Robert Nicol[1], Sinéad B. O'Leary[1], Amber Ratcliffe[2]†, Steven Rounsley[1], Xinwei She[3], Katherine M. B. Sneddon[5], Sandra Stewart[2], Carrie Sougnez[1], Sabrina M. Stone[1], Kerri Topham[1], Dascena Vincent[2]†, Shunguang Wang[1], Andrew R. Zimmer[1], Bruce W. Birren[1], Leroy Hood[2], Eric S. Lander[1] & Chad Nusbaum[1]

Here we present a finished sequence of human chromosome 15, together with a high-quality gene catalogue. As chromosome 15 is one of seven human chromosomes with a high rate of segmental duplication[1], we have carried out a detailed analysis of the duplication structure of the chromosome. Segmental duplications in chromosome 15 are largely clustered in two regions, on proximal and distal 15q; the proximal region is notable because recombination among the segmental duplications can result in deletions causing Prader-Willi and Angelman syndromes[2,3]. Sequence analysis shows that the proximal and distal regions of 15q share extensive ancient similarity[4]. Using a simple approach, we have been able to reconstruct many of the events by which the current duplication structure arose. We find that most of the intrachromosomal duplications seem to share a common ancestry. Finally, we demonstrate that some remaining gaps in the genome sequence are probably due to structural polymorphisms between haplotypes; this may explain a significant fraction of the gaps remaining in the human genome.

The present work describes the completion of a physical map, high-quality finished sequence, and gene catalogue for the euchromatic q arm of human chromosome 15, representing 2.9% of the human genome. The finished sequence contains 81,871,010 bases and is interrupted by nine euchromatic gaps and one gap containing the heterochromatic p arm and centromere regions (Fig. 1). The total size of the euchromatic gaps is estimated at 544 kilobases (kb) (Methods and Supplementary Table S1). These gaps remain despite the screening of genomic libraries containing a combined ~53-fold physical coverage, and are refractory to current cloning and mapping technology; six are within or adjacent to large duplicated regions. Of the finished sequence, 74% was generated by the Broad Institute of MIT and Harvard (formerly the Whitehead Institute/MIT Center for Genome Research (WICGR)), 25% by the Multimegabase Sequencing Center (initially at the University of Washington, currently at the Institute for Systems Biology), and the remaining ~1% by three other groups (Supplementary Table S2). The analyses here are referenced to NCBI Build 35; however, we have slightly improved this sequence (including closing one of the euchromatic gaps), and provide the updated clone path in Supplementary Table S3. Details of construction of the clone map and sequencing are described in the Supplementary Information. The short arm of chromosome 15, as in other acrocentric human chromosomes (chromosomes 13, 14, 21 and 22), is heterochromatic and was not sequenced as part of the Human Genome Project; it is estimated at 17 Mb (ref. 5) and contains arrays of ribosomal RNA genes, satellite sequences and other repeated sequences[6].

We assessed the local accuracy of the clone path by aligning paired-end sequences from a human fosmid library (designated WIBR2, representing 10 × physical coverage) to the finished sequence[7,8]. This analysis revealed no aberrant clones. In addition, an independent quality assessment exercise commissioned by the National Human Genome Research Institute[9] estimated the accuracy of the finished sequence to be better than one error in 100,000 bases (J. Schmutz, personal communication).

Several analyses suggest that nearly the entire euchromatic region

[1]Broad Institute of MIT and Harvard, 320 Charles Street, Cambridge, Massachusetts 02141, USA. [2]Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103, USA. [3]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA. [4]Division of Human Biology, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, 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.); Neurogenomics Research Lab, 200 B EMRB, University of Iowa, Iowa City, Iowa 52242, USA (Anup Madan); Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA (N.A.); Blue Heron Technologies, Bothell, Washington 98021, USA (L.B.); Department of Microbiology, Box 358070, University of Washington, Seattle, Washington 98195, USA (B.B.); Seattle University School of Nursing, Seattle, Washington 98122, USA (J.F.); Corbis Corporation, Seattle, Washington 98104, USA (P.F.); Geospiza, 100 West Harrison North Tower, Suite 330, Seattle, Washington 98119, USA (G.H.); Division of Medical Genetics, Box 357720, University of Washington, Seattle, Washington 98195, USA (E.J.); 3095 Medical Laboratories, Department of Neurosurgery, University of Iowa, Iowa City, Iowa 52242, USA (Anurhadha Madan); Nanostring Technologies, 201 Elliott Avenue West, Suite 300, Seattle, Washington 98119, USA (A.R.); Genelex Corporation, 3000 1st Avenue, Suite 1, Seattle, Washington 98121, USA (D.V.).

of chromosome 15 is present and accurately represented in the finished sequence. All genes in the RefSeq[10] database (596 loci, 742 transcripts) previously mapped to chromosome 15 are present and complete in the finished sequence. Furthermore, the finished sequence shows excellent alignment to genetic and radiation hybrid maps (Supplementary Fig. S1). The genetic map[11] shows perfect alignment, with no discrepancies among 125 sequence-based genetic markers (Supplementary Table S4). The radiation hybrid map[12] contains only local discrepancies, owing to its lower resolution (Supplementary Table S5). A large gap in the radiation hybrid coordinates (254–280 cR) at ~74 Mb in the physical map, near a region where chromosome breakage has been observed independently in multiple mammalian lineages (see below), is probably the result of non-random breakage in the generation of the radiation hybrid panel.

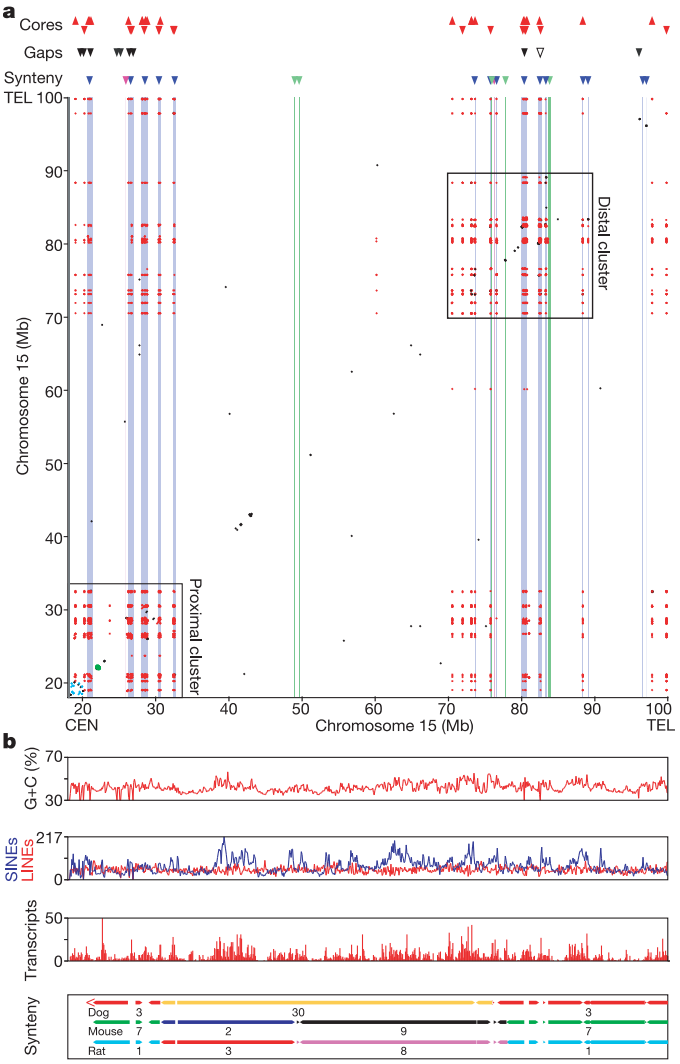

Figure 1 | Overview and duplication content of human chromosome 15. **a**, Dot-plot of duplicons on human chromosome 15, showing association with species-specific breaks in conserved synteny. Class 1 duplications are shown in red; other coloured dots indicate alignments between minor duplication classes. Vertical bands topped by arrows represent breaks in synteny (human-specific in purple, rodent-specific in green, dog-specific in pink). Black arrows at the top denote gaps in the human sequence (open arrow indicates a gap that was closed after Build 35 was made). Red triangles at the top show the locations and strand of class 1 core elements. The 15q telomere (TEL) and the centromere (CEN) are indicated. **b**, The following features are represented in discrete windows of 100 kb (top to bottom): G+C content on a scale from 30–70%; densities of LINEs (red) and SINEs (blue) (long and short interspersed elements, respectively); and transcripts as counts of elements. The bottom panel shows blocks of conserved synteny (100-kb resolution) with dog, mouse and rat. Chromosomes are numbered, and are coloured arbitrarily for ease of distinction.

We produced a manually curated[8] catalogue of genes, containing 695 gene loci (including all genes in RefSeq) and 250 pseudogene loci on chromosome 15. Table 1 classifies the genes according to standardized categories. The 3% of genes in the 'novel' and 'putative' categories were annotated based only on spliced expressed-sequence-tag (EST) evidence; some of these may prove to be pseudogenes. The full-length transcripts of known genes have an average length of 3,267 bp, with an average of 11.6 exons. Internal exon lengths average 156 bp. Gene loci have an average of 4.6 distinct transcripts, with 66% having at least two transcripts. These gene statistics are similar to recent reports[8,13–16]. Examples of genes that represent extremes of these distributions are described in the Supplementary Information. Most (74%) of the 250 pseudogenes are processed. In addition, we identified 9 transfer RNA genes (Supplementary Table S6) and found six known microRNAs mapping to chromosome 15 (Supplementary Table S7).

In most aspects of its landscape, chromosome 15 is close to genome-wide averages[7]. The overall gene density is 8.6 genes per Mb. There are 18 gene deserts (defined as 500 kb without an identified coding gene, Supplementary Table S8) comprising 14.9 Mb (~18.3% of the chromosome). The overall G+C content is 42.2%, but varies substantially across the chromosome (Fig. 1b). Transposable element fossils cover 38.3%. Chromosome 15 is also typical in its content of non-coding sequence conservation (see Supplementary Information).

Chromosome 15 is, however, one of seven autosomes that are significantly enriched in segmental duplications (defined as regions >1 kb that are not high-copy repeats and have >90% identity to another region in the genome[17]), with 8.8% of its euchromatin composed of such sequence (Supplementary Fig. S2). As with other heavily duplicated chromosomes, chromosome 15 has a large fraction of intrachromosomal duplication: 50% is strictly intrachromosomal, 30% is both intra- and interchromosomal, and 20% is solely interchromosomal (largely in the proximal 1.5 Mb). The proportion of purely interchromosomal duplication might be even lower, as some undetected tandem duplication may exist near the centromere (see below). Recombination among segmental duplications within the region 15q11–q13 gives rise to deletions that are known to cause Prader-Willi and Angelman syndromes[2,3] (Supplementary Information).

We sought to investigate the duplication landscape of chromosome 15 by studying the relationships among the duplicated segments. Previous work has shown that a sequence within the Prader-Willi/Angelman syndrome region, termed LCR15 (ref. 4), is also duplicated on distal 15q (Supplementary Fig. S2). By extending our analysis to detect more ancient relationships (sequence identity less than 90%), we found much more extensive similarity among the duplicated sequences in both proximal and distal 15q (Fig. 1a). We clustered together segmental duplications containing related sequence (Methods) and found that most fell into a single large cluster, which we refer to as 'class 1'. The class includes 67% of all bases in segmental duplications and 91% of all pairwise duplication events (as some bases reside within multiple independent events) (Supplementary Table S9).

Although the segmental duplications are related to one another in a complex fashion, we sought to identify a 'core element' that was present in many of the class 1 elements. We took the longest duplicated class 1 region (213 kb starting at 18.89 Mb, within the Prader-Willi/Angelman syndrome region) and aligned all duplicated regions of the chromosome to it, counting the number of different duplication regions that aligned to each base. We selected a core element that includes the highest peak of coverage (Supplementary Fig. S3); the element is 2,920 bp long and lies within the ~15-kb LCR15 element.

Table 1 | Chromosome 15 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 genes	532	76	66,994	4.6	3,267	11.6	156 ($n = 6,471$)	6,157 ($n = 8,277$)	76
Novel CDS	73	11	40,090	2.1	1,185	5.2	154 ($n = 278$)	8,108 ($n = 384$)	35
Novel transcripts	68	10	29,855	1.8	867	3.5	146 ($n = 182$)	8,851 ($n = 351$)	46
Putative genes	15	2	10,074	1.5	1,070	2.9	109 ($n = 10$)	6,700 ($n = 35$)	47
Gene fragments	7	1	1,563	1.0	425	2.3			
Total	**695**		**57,963**						
Pseudogenes	250	26	3,297	1.0	1,091	2.2	195 ($n = 234$)	1,878 ($n = 294$)	27

Categorization is according to Hawk2 standards (http://www.sanger.ac.uk/Info/workshops/hawk2; see Supplementary Information). CDS, coding sequences.
*Average chromosomal distance from the beginning of the 5′-most exon to the end of the 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. (Sample size given in parentheses.)
‖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. (Sample size given in parentheses.)
¶Percentage of genes with a transcript having a CpG island (as assessed by FirstEF) within −2 kb and +1 kb of the transcription start.

The human genome contains 41 nearly full-length copies of the core element: there are 37 on chromosome 15, two on the Y chromosome, and one each on chromosomes 2 and 10. To understand the origins of the element, we compared the core element to the dog[18] and mouse[19] genomes. The dog and mouse genomes each contain a single copy of the element, which is orthologous to the copy on human chromosome 2. The similarity among the sequences is shown in a phylogenetic tree (Fig. 2, see Methods). The copy on chromosome 2 is at the root of the human duplications, closest to mouse and dog, as would be expected from conserved synteny. The duplications on chromosome 15 fall into two distinct and well-separated branches: a proximal branch containing all the elements in the Prader-Willi/Angelman syndrome region (chromosome position 18–32 Mb), and a distal branch containing all the elements from 73 to 88 Mb, with a tight clustering of elements around 80–83 Mb. A further two repeats in the subtelomeric region (98–100 Mb) are closely related to the proximal branch. Pairwise divergence between elements in the two branches is ~11%, indicating that they share an ancient origin followed by local duplications, but with no recent interaction between branches.

From the tree, it is possible to reconstruct the likely history of the core element. The sequence on chromosome 2 lies in the 3′ untranslated region (UTR) of a splice variant of the gene intersectin 2 (*ITSN2*). This sequence seems to have moved by retroposition to chromosome 10 (at 30.68 Mb), inserting immediately downstream of the 5′ coding sequence of an interchromosomally duplicated copy of *GOLGA2* (the origin of which is on chromosome 9). A combined unit (15 kb, consisting of *GOLGA2* and the *ITSN2* UTR) then was copied to chromosome 15, where it has duplicated extensively. Finally, two copies exist on the arms of a large palindrome on the Y chromosome, and seem to have moved to the Y chromosome by segmental duplication of ~40 kb of chromosome 15 (at 82.7 Mb).

We next sought to understand why the large regions of segmental duplication in proximal 15q (denoted 'A') and distal 15q (denoted 'C') are separated by a large stretch that contains almost no duplicated sequence (denoted 'B'). Analysis of conserved synteny with other species allows a reconstruction of the history of chromosome 15 (Fig. 3). Briefly, the three segments were adjacent in the boreoeutherian ancestor (the common ancestor of Euarchontoglires and Laurasiatheria), but were found in the order A–C–B. In the primate lineage, the chromosome apparently underwent a single large inversion that separated segments A and C. (Details of the reconstruction and comparison to recent reports[20,21] can be found in the Supplementary Information and Supplementary Fig. S4.) This suggests that the core element was transferred to chromosome 15 before the divergence of apes and Old World monkeys, and expanded locally (in the originally contiguous A–C region). The inversion subsequently separated regions A and C, and the element continued to expand separately in each region.

To test this hypothesis, we examined the current draft assembly of the rhesus macaque genome (rheMac1; R. Gibbs, personal communication). We found at least 12 nearly full-length copies of the core element that we added to the evolutionary tree (Supplementary Fig. S5). We also found unique orthologues of the copies on human chromosomes 2 and 10. The remaining macaque elements were split between the proximal and distal clusters, confirming that the element

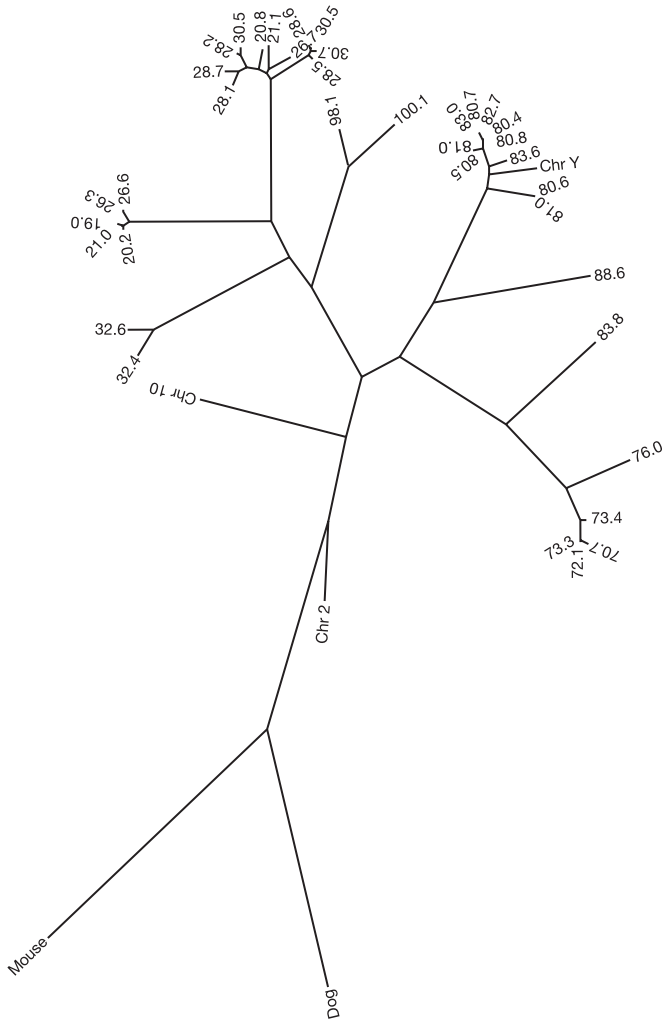

Figure 2 | Phylogenetic tree of the 41 human copies and the unique dog and mouse copies of the conserved core element. Chromosome 15 copies are distinguished by their physical position (in Mb). Chr, chromosome.

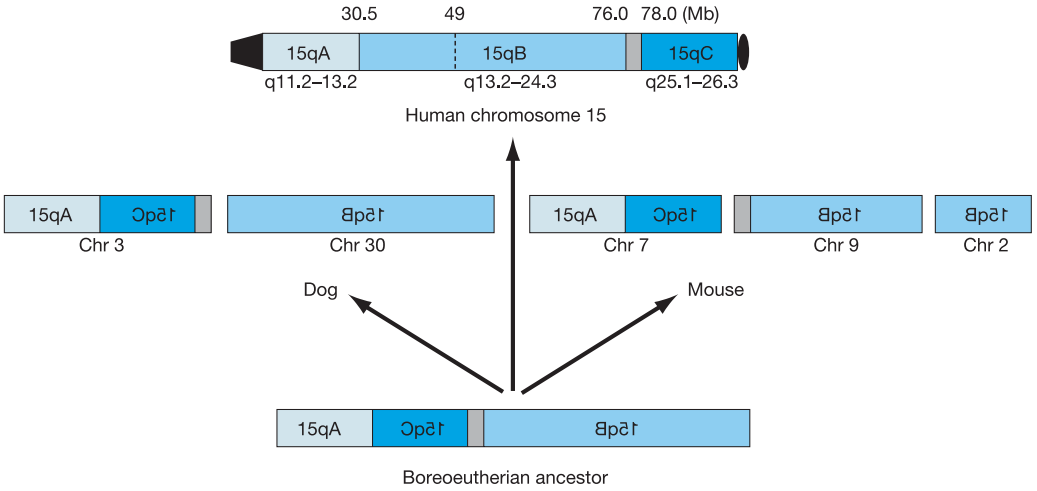

Figure 3 | History of the structural organization of human chromosome 15. For simplicity, we have depicted the q arm as three large segments—A, B and C—that have travelled together. Internal rearrangements exist within these segments, but do not cross between them in mammals. Rat (not shown) is identical to mouse except for the chromosome numbers. The breakpoints between distal B and proximal C differ by 2 Mb in rodent and dog (grey box). Segments labelled with backwards text are inverted with respect to the modern human chromosome. Chr, chromosome.

had already appeared and begun to duplicate on chromosome 15 before the divergence of Old World monkeys and apes. The human and macaque elements are grouped into separate clusters in both the proximal and distal branches, indicating that local duplication has continued to occur in both the human and macaque lineages.

The analysis of conserved synteny also reveals that the segmental duplications are closely associated with chromosomal rearrangements. Chromosome 15 has 15 human-specific breakpoints of conserved synteny, all of which are inversions. Of these, 13 occur in regions containing class 1 duplications. This suggests that the segmental duplications may have mediated the inversions and that these inversions may have helped to disperse the elements.

The class 1 core element serves as a useful marker for tracing chromosomal history. However, the ubiquity of the core element raises the possibility that it had a causal role in the process of segmental duplication on chromosome 15. The element is derived from a UTR on chromosome 2, of which at least 500 bases are highly conserved across mammals and thus are presumably functional. Moreover, many of the copies on chromosome 15 are transcribed: 13 known genes on chromosome 15 (all golgins or golgin-like proteins) contain this duplicated UTR, and another 16 transcripts stop just short of it (Supplementary Table S10). It will be interesting to investigate whether functional properties of the fusion element on chromosome 15 promote local duplication, and to explore whether this had significant implications for primate evolution.

Finally, we note that the segmental duplications represent the main challenge in closing the remaining gaps in the sequence of chromosome 15. Build 35 contains ten gaps, seven of which lie within or immediately adjacent to class 1 duplications (Fig. 1). In some cases, the duplicated sequences flanking the gaps are so similar (>99.7% identity) that they may represent allelic variants. Moreover, six of the seven duplication-associated gaps are adjacent to or within reported sites of copy-number polymorphism[22,23] (Supplementary Table S1). We have recently been able to close one gap (at 82.7 Mb) (decreasing the number of gaps to nine) by finding previously missed overlap between two flanking clones; another clone spanning this gap carries an alternative haplotype with an additional 100 kb, including an 80-kb near-perfect duplication. Examination of three of the other gaps suggests that they might also be due to structural variation, although more work will be required to confirm this.

The finished sequence of chromosome 15 offers a window into the natural history of segmental duplications and the structural history of chromosomes. Notably, most of the intrachromosomal duplication involves a single class of duplicons. On the basis of these results, we suggest an important role for such duplicons in structural evolution and gene diversification.

METHODS

Production of gene catalogue and annotation. The gene catalogue was produced as described previously[8]. Gene symbols were assigned by the HUGO Gene Nomenclature Committee for biologically characterized loci. A complete list of gene symbols from this paper can be found in Supplementary Table S11. Annotation was performed as described previously[8]. Our annotations are available from the Vertebrate Genome Annotation database (VEGA, http://vega.sanger.ac.uk/Homo_sapiens).

Segmental duplications. Segmental duplications were defined as pairs of regions of 90% or greater identity (excluding repeat-masked bases) that extend for 1 kb or more. The map of segmental duplications was prepared using a method adapted from ref. 17, by concatenating all-against-all MegaBlast[24] alignments. A genome database was built using hard-masked sequence. This same hard-masked sequence was presented to MegaBlast as a probe, chromosome by chromosome. All alignments of 80% or better identity with expectation $<10^{-4}$ were kept. Alignments were then concatenated if they were contiguous except for masked repeats. Unmasked gaps could be crossed but were penalized to prevent over-merging by being treated as bases of 50% identity. Final segments meeting the 1-kb length and 90% identity criteria were retained.

Duplication class clustering. Pairwise intrachromosomal duplications were defined as above. A pairwise duplication A ~ A′ was considered to be in the same class as another pairwise duplication B ~ B′ if B or B′ overlapped A or A′ by 150 bp or more. We extended this by transitive closure to build maximally linked sets (that is, if A ~ A′ linked to B ~ B′ and C ~ C′, all were clustered, even if B ~ B′ did not overlap C ~ C′). The number of duplications in a class is counted as the number of distinct pairwise alignments X ~ X′ that were clustered. The number of bases in a class is counted as the number of distinct bases covered by at least one pairwise duplication in that class.

Construction of core element phylogeny. Full-length or nearly full-length copies of the core element in human were identified by MegaBlast (release 2.2.11). Copies in the mouse and dog genomes were identified by MegaBlast followed by blastn (release 2.2.11) to refine the boundaries and extend the regions. Multiple alignments of the elements were generated with ClustalW (v.1.83). Pairwise and multiple alignment parameters were adjusted by reducing the gap extension penalty to 0.1 and replacing the standard DNA matrix with a custom matrix scoring 10 for any match, −5 for any mismatch, and 0 for any alignment to an unknown base (N). The trees were output in phylip format and all gaps of length >1 converted to single indels by substitution of '?' characters for all but the first '−' in the gap to avoid generating disproportionately long branches for element copies with substantial deletion. Terminal gaps were also treated this way. Trees were built with the dnapars parsimony module of phylip (v.3.65)[25]. The tree represented is the first of 15 equally likely trees that differ only

in the leaf placement of the seven nearly identical copies of the element at 80 and 82 Mb on chromosome 15.

Received 9 November 2005; accepted 26 January 2006.

1. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297,** 1003–1007 (2002).
2. Emanuel, B. S. & Shaikh, T. H. Segmental duplications: an 'expanding' role in genomic instability and disease. *Nature Rev. Genet.* **2,** 791–800 (2001).
3. Soejima, H. & Wagstaff, J. Imprinting centers, chromatin structure, and disease. *J. Cell. Biochem.* **95,** 226–233 (2005).
4. Pujana, M. A. *et al.* Additional complexity on human chromosome 15q: Identification of a set of newly recognized duplicons (LCR15) on 15q11-q13, 15q24, and 15q26. *Genome Res.* **11,** 98–111 (2001).
5. Morton, N. Parameters of the human genome. *Proc. Natl Acad. Sci. USA* **88,** 7474–7476 (1991).
6. Kehrer-Sawatzki, H. *et al.* Mapping of members of the low-copy-number repetitive DNA sequence family from chAB4 within the p arms of human acrocentric chromosomes: characterization of Robertsonian translocations. *Chromosome Res.* **6,** 429–435 (1998).
7. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
8. Nusbaum, C. *et al.* DNA sequence and analysis of human chromosome 18. *Nature* **437,** 551–555 (2005).
9. Schmutz, J. *et al.* Quality assessment of the human genome sequence. *Nature* **429,** 365–368 (2004).
10. 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).
11. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31,** 241–247 (2002).
12. Schuler, G. D. *et al.* A gene map of the human genome. *Science* **274,** 540–546 (1996).
13. Nusbaum, C. *et al.* DNA sequence and analysis of human chromosome 8. *Nature* **439,** 331–335 (2006).
14. Hillier, L. W. *et al.* Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature* **434,** 724–731 (2005).
15. Grimwood, J. *et al.* The DNA sequence and biology of human chromosome 19. *Nature* **428,** 529–535 (2004).
16. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 10. *Nature* **429,** 375–381 (2004).
17. Bailey, J. A. *et al.* Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11,** 1005–1017 (2001).
18. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438,** 803–819 (2005).
19. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
20. Murphy, W. J. *et al.* Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309,** 613–617 (2005).
21. Bourque, G., Zdobnov, E. M., Bork, P., Pevzner, P. A. & Tesler, G. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res.* **15,** 98–110 (2005).
22. Sharp, A. J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77,** 78–88 (2005).
23. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305,** 525–528 (2004).
24. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *NAR* **25,** 3389 (1997).
25. Felsenstein, J. PHYLIP—Phylogeny Inference Package (version 3.2). *Cladistics* **5,** 164–166 (1989).

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 text. We are grateful to T. Furey for help with lists of genetic markers and placement of RefSeqs, and to K. Lindblad-Toh for sharing data from the genome projects of dog and opossum. Fluorescence *in situ* hybridization (FISH) data for opossum were provided by M. Breen. We thank the members of the Baylor College of Medicine Human Genome Sequencing Center, the J. Craig Venter Institute Joint Technology Center, and the Washington University Genome Sequencing Center for generation and early release of the assembly of the rhesus macaque genome. We thank the Sanger Institute for gap sizing by FISH. We also acknowledge the HUGO Gene Nomenclature Committee (S. Povey (chair), E. A. Bruford, V. K. Khodiyar, R. C. Lovering, M. J. Lush, T. P. Sneddon, C. C. Talbot Jr and M. W. Wright) for assigning official gene symbols. We are grateful to all members, present and past, of the Broad (and Whitehead) sequencing platform for their dedication and the consistent high quality of their data.

Author Information Accession numbers for all clones contributing to the finished sequence of human chromosome 15 can be found in Supplementary Table S3. The updated human chromosome 15 sequence can be accessed through GenBank accession number NC_000015. 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 M.C.Z. (mczody@broad.mit.edu) or C.N. (chad@broad.mit.edu).