

The DNA sequence and comparative analysis of human chromosome 5

Jeremy Schmutz¹, Joel Martin², Astrid Terry², Olivier Couronne³, Jane Grimwood¹, Steve Lowry², Laurie A. Gordon^{2,4}, Duncan Scott², Gary Xie^{2,5}, Wayne Huang², Uffe Hellsten², Mary Tran-Gyamfi^{2,4}, Xinwei She⁶, Shyam Prabhakar³, Andrea Aerts², Michael Altherr^{2,5}, Eva Bajorek¹, Stacey Black¹, Elbert Branscomb^{2,4}, Chenier Caoile¹, Jean F. Challacombe⁵, Yee Man Chan¹, Mirian Denys¹, John C. Detter², Julio Escobar¹, Dave Flowers¹, Dea Fotopulos¹, Tijana Glavina², Maria Gomez¹, Eidelyn Gonzales¹, David Goodstein², Igor Grigoriev², Matthew Groza⁴, Nancy Hammon², Trevor Hawkins², Lauren Haydu¹, Sanjay Israni², Jamie Jett², Kristen Kadner², Heather Kimball², Arthur Kobayashi^{2,4}, Frederick Lopez¹, Yunian Lou², Diego Martinez², Catherine Medina¹, Jenna Morgan², Richard Nandkeshwar⁴, James P. Noonan⁷, Sam Pitluck², Martin Pollard², Paul Predki², James Priest³, Lucia Ramirez¹, James Retterer¹, Alex Rodriguez¹, Stephanie Rogers¹, Asaf Salamov², Angelica Salazar¹, Nina Thayer^{2,5}, Hope Tice², Ming Tsai¹, Anna Ustaszewska², Nu Vo¹, Jeremy Wheeler¹, Kevin Wu¹, Joan Yang¹, Mark Dickson¹, Jan-Fang Cheng³, Evan E. Eichler⁶, Anne Olsen^{2,4}, Len A. Pennacchio^{2,3}, Daniel S. Rokhsar², Paul Richardson², Susan M. Lucas², Richard M. Myers¹ & Edward M. Rubin^{2,3}

¹Stanford Human Genome Center, Department of Genetics, Stanford University School of Medicine, 975 California Ave, Palo Alto, California 94304, USA

²DOE's Joint Genome Institute, 2800 Mitchell Avenue, Walnut Creek, California 94598, USA

³Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, California 94720, USA

⁴Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, California 94550, USA

⁵Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

⁶Department of Genetics, Center for Computational Genomics and Center for Human Genetics, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, Cleveland, Ohio 44106, USA

⁷Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA

Chromosome 5 is one of the largest human chromosomes and contains numerous intrachromosomal duplications, yet it has one of the lowest gene densities. This is partially explained by numerous gene-poor regions that display a remarkable degree of noncoding conservation with non-mammalian vertebrates, suggesting that they are functionally constrained. In total, we compiled 177.7 million base pairs of highly accurate finished sequence containing 923 manually curated protein-coding genes including the protocadherin and interleukin gene families. We also completely sequenced versions of the large chromosome-5-specific internal duplications. These duplications are very recent evolutionary events and probably have a mechanistic role in human physiological variation, as deletions in these regions are the cause of debilitating disorders including spinal muscular atrophy.

The US Department of Energy's interest in chromosome 5 emerged from a series of pilot studies begun at the Lawrence Berkeley National Laboratory focusing on a cluster of interleukin genes located at human 5q31. The insights gained from these detailed analyses of a single megabase of chromosome 5 illustrated how finished human sequence could contribute to gene annotation and how multi-mammalian sequence comparisons could lead to the sequence-based identification of noncoding elements possessing gene regulatory activities¹⁻³. The finished sequence of chromosome 5 and its analysis alone and in comparison to orthologous regions in other vertebrate genomes now provides a chromosome-wide catalogue of genes and evolutionarily conserved noncoding sequences. Many of these observations, as well as clues into disease-causing deletions arising from the segmented duplication landscape of chromosome 5, can only now be appreciated upon finishing the sequence of this chromosome.

Mapping and sequencing

After the completion of the initial draft sequencing in 2001 we selected clones with an approach that integrated all of the public sequence, previously reported clone contigs⁴⁻⁶ including the Celera scaffolds⁷, bacterial artificial chromosome (BAC) and fosmid end sequences, and BACs isolated with an overgo hybridization strategy to close gaps between anchored contigs. The final version of the tiling path contains 1,763 clones, (96% BACs) with four gaps remaining, all in the long arm. None of these remaining gaps are part of the large chromosome 5 duplications, and they appear to be unclonable in current vector systems. In addition, our standard strategy of seeding and then walking into gaps based on restriction

maps proved unworkable in the duplication region of 5q13 associated with spinal muscular atrophy (SMA), and led to mapping errors with its primary insertion copy at 5p14 and secondary copy at 5p13. Therefore, we adopted a strategy of drafting high depth clone coverage from the single individual RPCI-11 BAC library in order to construct single haplotype paths spanning the duplications.

On the basis of internal and external quality checks, we estimate the accuracy of our finished sequence to exceed 99.99%⁸. In total, we finished 177,702,766 base pairs (bp) and estimate the total chromosome size, including the clone gaps and the recalcitrant centromeric and subtelomeric regions, to be 180.8 megabases (Mb). The finished

Table 1 **Chromosome 5 sequence features**

Feature	Value
Sequence length (bp)	177,702,766
G+C content (%)	39.5
Gene loci	923*
Known	827
Novel	55
Putative	41
Non-processed pseudogenes	98
Processed pseudogenes	479
tRNAs	20
tRNA pseudogenes	4
Total repeat content (bp)	82,349,155 (46.3%)
Alu	14,998,401 (8.4%)
LINE 1	32,864,033 (18.5%)
LINE 2	4,757,270 (2.7%)
Simple and low complexity	2,594,624 (1.5%)
Other	27,134,827 (15.3%)

*With 1,598 full-length transcripts.

sequence covers 99.9% of the euchromatic sequence and captures all known genes that were previously mapped to chromosome 5 (T. Furey, personal communication). The Stanford v.4 G3 radiation hybrid map⁹ was compared to the sequence and it matched the marker order well (see Supplementary Fig. S1). Thirteen (out of 442) unplaced markers were found to have been originally incorrectly assigned to chromosome 5. Recombination distances from the deCODE¹⁰ meiotic maps were compared to physical distances with recombination rates accurately tracking physical distance (see Supplementary Fig. S2), as previously reported for other chromosomes^{11–13}.

Gene catalogue

We placed gene model transcripts on the chromosome 5 sequence and manually reviewed these models using previously described methods¹¹ (Table 1). Ultimately, 923 protein-coding regions were verified as gene loci (see Supplementary Table S1 and http://www.jgi.doe.gov/human_chr5). These loci contain 1,598 full-length (or nearly full-length) transcripts, including partial evidence for additional splice variants (see Supplementary Information). Loci were placed in three categories: ‘known’, ‘novel’ and ‘pseudogenes’, consistent with our previous definitions¹¹. Transcripts for which a unique open reading frame (ORF) could not be determined and putative genes defined by *ab initio* models but with no supporting experimental evidence were not considered valid. A total of 827 known loci were identified based on 2,203 RefSeq genes and other full-length complementary DNA sequences in GenBank, extending 36% of RefSeq transcripts by more than 50 bp at the 5′ end and 18% at the 3′ end, while maintaining the original ORF. Gene loci 3′ ends were not extended when the only evidence was from rare expressed sequence tag (EST) variants. Evidence for 55 novel loci was supported by full-length cDNA sequence, spliced ESTs, and/or similarity to known human or mouse gene sequences. Forty-one putative gene loci were modelled using orthologous mouse cDNA sequences. Twenty transfer RNA genes and four tRNA pseudogenes were predicted, similar in density to other finished chromosomes^{11–13}.

The extent of alternative splicing was characterized based on the existing cDNA and EST data. Considering only messenger RNA sequences in GenBank, 1,598 distinct transcripts were identified, providing an average coverage of 1.7 annotated transcripts per locus (see Supplementary Information). These mRNAs provide strong evidence for alternative splicing of 408 (44%) of the 923 loci, each having two or more associated transcripts. A total of 577 pseudogenes and pseudogene fragments were also identified, representing two classes: (1) 98 non-processed pseudogenes that display a structure similar to the parent locus and probably resulted from genomic duplication events; (2) 479 processed pseudogenes that

presumably resulted from viral retrotransposition of spliced mRNAs (see Supplementary Information). No significant bias towards over-representation of pseudogenes from a particular gene family was observed.

Chromosome 5 genomic duplications

We performed a detailed analysis of duplicated sequence ($\geq 90\%$ identity and ≥ 1 kilobase (kb) length) by comparing chromosome 5 against the July 2003 human genome assembly. An estimated 3.49% (6.26 Mb) of the chromosome consists of segmental duplications, lower than the genome-wide average of 5.3% (see Supplementary Table S2 and Supplementary Fig. S4). Chromosome 5 segmental duplications, however, show a higher degree of sequence identity ($\geq 97.5\%$), especially with other regions of chromosome 5 (see Supplementary Fig. S5), than do the duplications on other chromosomes. Intrachromosomal duplications are clustered in ten regions (Fig. 1) and represent the majority of the gene duplications, including the largest gene family: the protocadherins (see Supplementary Information). The high degree of sequence identity underlying most of these intrachromosomal genomic duplications suggests that these structures are relatively recent duplications or gene conversion events that emerged during the separation of humans and the great apes (see Supplementary Fig. S3 and Supplementary Table S2).

Subtelomeric and pericentromeric biases have been reported for segmental duplications for other human chromosomes. Despite the fact that large tracts of alpha-satellite DNA have been sequenced on both chromosomal arms near the centromere, there is little evidence for extensive pericentromeric duplication, with 5p11 showing almost a complete absence of duplications. A single duplication in 5q11 (96% identity over 250 kb) between chromosomes 1 and 5 accounts for nearly all pericentromeric duplicated bases. The pericentromeric region of chromosome 5, along with 19q11, may define a duplication-quiescent model of pericentromeric organization. The telomeric regions do show extensive interchromosomal duplications (Fig. 1), with 25% (2.48 out of 9.08 Mb) of all interchromosomal alignments occurring within 2 Mb of the long arm telomeric repeat sequence (see Supplementary Table S3).

SMA duplication region

One of the most duplicated regions on chromosome 5 occurs in a 1–2-Mb interval in 5q13.3. Homozygous deletions of the *SMN1* gene and variable copies of the *SMN2* duplication in this region have been associated with various forms of spinal muscular atrophy and susceptibility to the disease^{14,15}. Analysis of carriers and controls suggests extreme locus variability, but the underlying structural variation has never been documented at the sequence level¹⁶. We identified a complex arrangement of 311 pairwise alignments

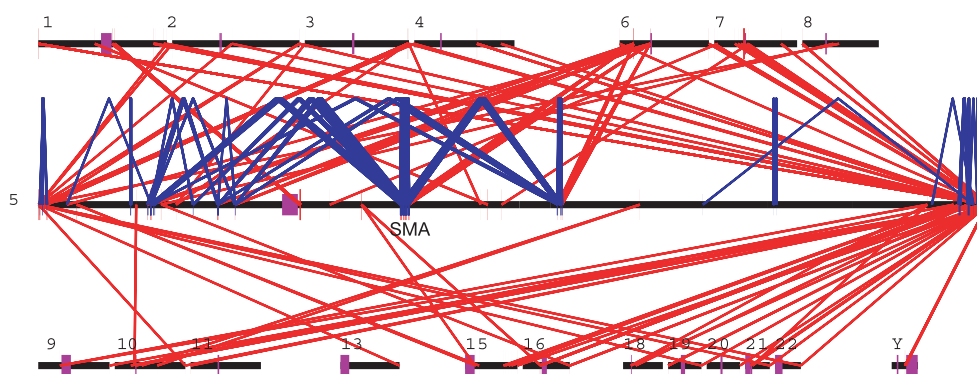


Figure 1 Distribution of segmental duplications on chromosome 5. Large (>5 kb) highly similar (>90%) intrachromosomal (blue) and interchromosomal (red) segmental

duplications are shown for chromosome 5. Chromosome 5 is drawn at a greater scale than the other chromosomes. The centromeres are depicted as purple bars.

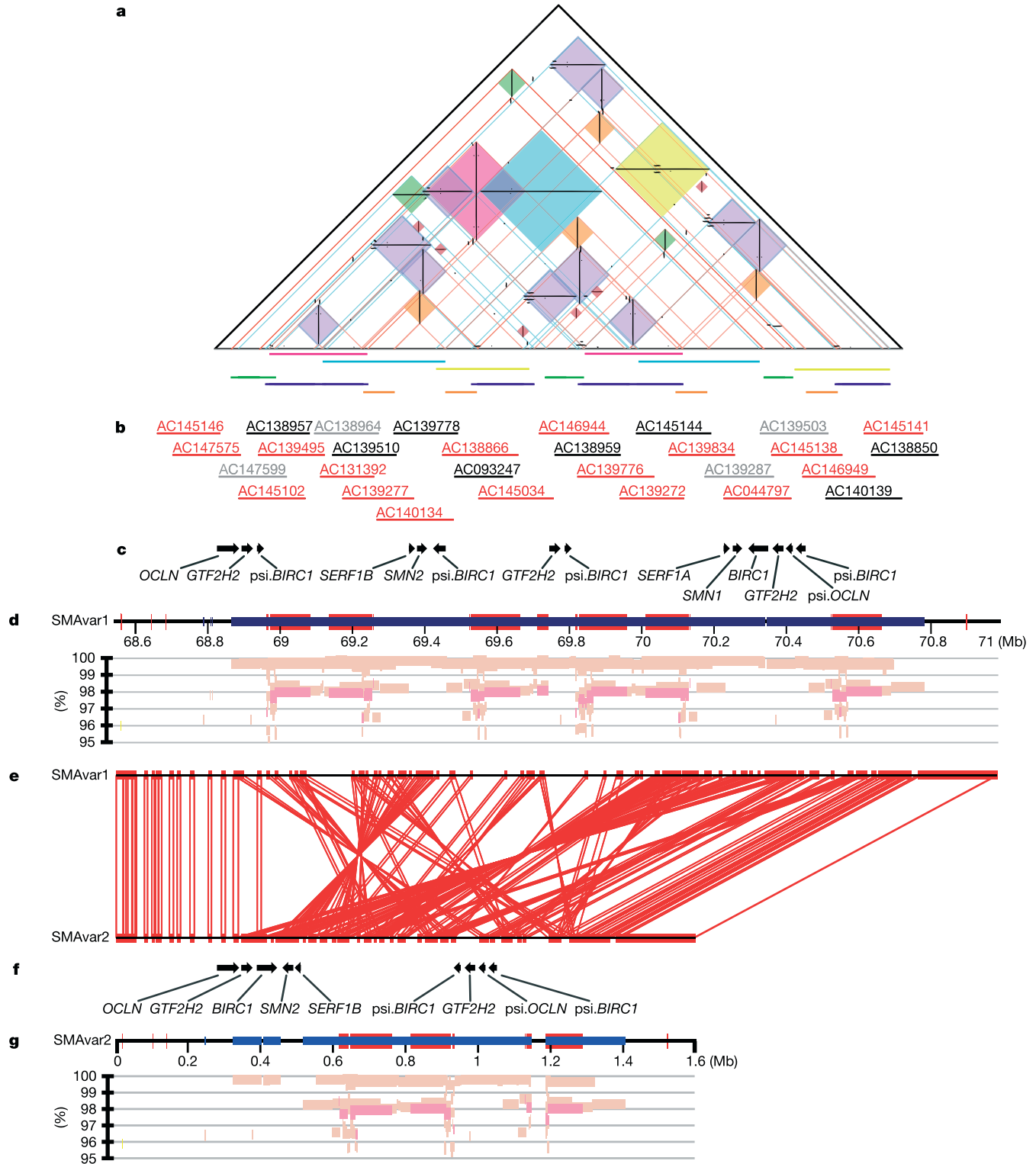


Figure 2 Diagram of the SMA region showing both SMAVar1, the published variant, and SMAVar2, the alternative RPC11 variant. **a**, Self_dot_plot⁴⁴ (http://staffa.wi.mit.edu/page/Y/azfc/self_dot_plot.pl) of SMAVar1. Vertical bars represent inverted repeats, horizontal bars direct repeats. Each dot represents a 200-bp perfect match. The three largest repeats are coloured pink, blue and yellow. **b**, RPCI-11 BAC clone path through SMAVar1. Red clones are in the final path, black clones are finished, grey clones are unfinished. **c**, Gene content of SMAVar1. **d**, The duplication pattern for SMAVar1 is shown along the

scale. Interchromosomal (red) and intrachromosomal duplications (blue) are indicated. The underlying pairwise alignments of segmental duplications (>95% > 1 kb) are depicted as a function of per cent identity (below the horizontal line) with different colours corresponding to the location of the pairwise alignment on different human chromosomes (light pink = 5; dark pink = 6; yellow = 3). **e**, A comparison of the interhaplotype structure between the two variants using Miropeats⁴¹ with a threshold of 7,000. **f**, Gene content of SMAVar2. **g**, Duplication pattern for SMAVar2.

representing the SMA region (Fig. 1). On average, the duplications are long (~200 kb) and show a high degree of identity (98.66%). Duplications in this region include interchromosomal duplications, all of which map to chromosome 6, with three very large tandem

(>99.5% identity) and other various interspersed intrachromosomal duplications (Fig. 2). Interestingly, this region is enriched in genes. We annotated 14 loci in this region, including *SERF1* (small EDRK-rich factor 1), *BIRC1* (baculoviral IAP repeat-containing 1)

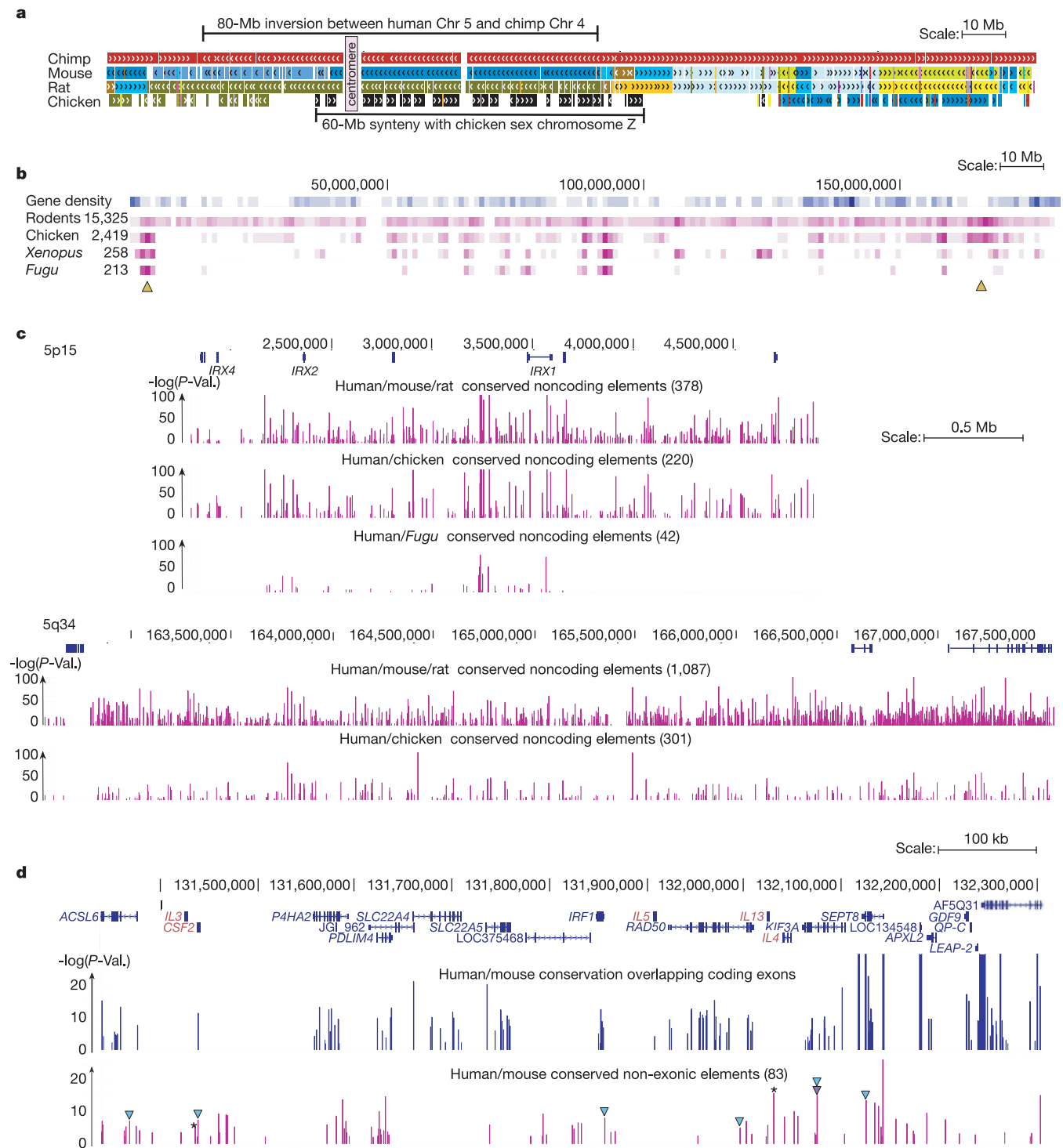


Figure 3 Comparative biology. **a**, Segmental homology maps between human chromosome 5 and the mouse, rat and chicken genomes (see Methods). **b**, Noncoding conservation density. The plot shows the normalized density of the human/mouse/rat, human/mouse/chicken, human/mouse/*Xenopus* and human/mouse/*Fugu* conserved elements. Yellow triangles indicate the location of regions expanded in panels **c** and **d**. **c**, The two largest human/mouse/rat/chicken homologous segments overlap gene-poor regions with a high density of conserved noncoding elements (see text). **d**, Interleukin

region. The first plot shows conservation overlapping coding exons, the second plot shows non-exonic conservation. Blue triangles indicate uncharacterized elements conserved in chicken; Purple triangles show uncharacterized elements conserved in *Xenopus*; asterisks are known interleukin enhancers³. These are conserved only in rodents (see text). For clarity only one isoform per gene is shown. In **c** and **d** conserved elements are ranked by their statistical significance relative to the local neutral mutation rate. The height of the bars is proportional to $-\log(P\text{-value})$ (PEAK-VISTA; see Methods).

and *SMN* (survival of motor neuron), the gene for *SMA*.

During the sequencing and assembly of this region, we generated a consensus sequence for a second haplotype variant from the RPCI-11 BAC library. Both haplotypes represent high-quality finished sequence and differ only by a remaining ~50-kb clone gap within *SMAvar2*. Sequence comparison of these regions (*SMAvar1* against *SMAvar2*) revealed extensive variation. At least two large-scale rearrangements (>400 kb) and multiple smaller insertion/deletion events are required to reconstruct an ancestral haplotype. Although there are many scenarios for the evolution of these variants, one explanation may be that a portion of the *SMAvar2* region (0.3–0.9 Mb) was inverted (68.9–69.4 Mb) and subsequently duplicated in *SMAvar1* (69.8–70.4 Mb). Such extensive structural variation between haplotypes may not be uncommon in regions of extensive segmental duplication.

Comparative biology

To understand further the evolution and functional sequences of human chromosome 5, we performed comparative analyses against the available chimpanzee, mouse, rat, chicken, frog (*Xenopus tropicalis*) and fish (*Fugu rubripes*) draft genomes. These comparisons revealed numerous large-scale chromosomal rearrangement events occurring since each of these species last common ancestor with humans, as well as a variety of nonrandomly distributed conserved noncoding regions (Fig. 3a). Additional analyses of the distribution of genes and conserved noncoding sequences along the length of the chromosome support the existence of large gene-poor regions with highly conserved noncoding sequences that may regulate genes from a distance. Furthermore, we examined conservation in a comparative analysis of the extensively studied interleukin gene cluster.

Synten

By building segmental maps from DNA alignments of all the vertebrate species described above, we were able to confirm and extend previous homologous chromosomal relationships with human chromosome 5. Whereas recent experimental studies support that large-scale rearrangements (40–175 kb) have frequently occurred during primate genome evolution¹⁷, our comparison of chromosome 5 and the recent chimpanzee draft genome sequence (International Chimpanzee Genome Sequencing Consortium, manuscript in preparation) uncovered even larger-scale events. For example, we found a large 80-Mb inversion in comparison to the chimpanzee genome, homologous to almost half of human chromosome 5 between 5p14 and 5q15 (Fig. 3a). This finding using the genomic draft data independently confirms previous fluorescence *in situ* hybridization (FISH) experiments¹⁸. It has been proposed that these large-scale rearrangements create barriers to fertile mating and triggered the speciation that separated these two lineages¹⁹. Comparison with the mouse genome sequence²⁰ yielded 142 chromosomal rearrangements ranging in size from 200 kb to 17 Mb. Between human and chicken, we found that one-third of chromosome 5 is homologous to the chicken sex chromosome Z²¹, further indicating that sex chromosomes have evolved independently after the avian and mammalian split some 300 million years ago²².

Chimpanzee

In addition to exploring the syntenic relationship between chromosome 5 and the chimpanzee draft assembly, we catalogued sequence changes between these two primates. To explore the constraint on human–chimpanzee evolution in noncoding regions, we compared the number of nucleotide substitutions in coding sequences, as well as noncoding regions conserved and not conserved in rodents. We found a substitution rate of 0.0067 changes per nucleotide in coding sequences, 0.0091 in noncoding regions conserved in rodents, and 0.015 in noncoding regions not conserved in rodents. The decreased

substitution rate in coding sequences and noncoding sequences conserved in rodents (compared to noncoding regions not conserved in rodents) support the theory that both of the former categories are under evolutionary constraint. This also supports the theory that human/chimpanzee coding and noncoding sequences conserved in rodents have been under moderate selective constraint since the last common human/chimpanzee ancestor. We next compared the patterns of variation within human and chimpanzee exons to identify genes potentially under positive selection in the human lineage as reported in ref. 23. We found 21 genes randomly distributed over chromosome 5 displaying a *P*-value less than 0.01 for an increased evolutionary rate in humans. Of note is that the two highest ranked genes (*FBN2* and *SQSTM1*) are both linked to human diseases. Mutations in *FBN2* cause pathologies similar to Marfan syndrome (*FBN1*), whereas *SQSTM1* has been linked to Paget's disease of the bone²⁴. As the chimpanzee genome reaches a further draft state, a similar complete re-analysis of the entire human gene set will probably yield large numbers of quickly evolving genes, which may explain unique aspects of human biology.

Vertebrate conservation

To annotate functional elements, we identified slowly evolving regions, presumably under evolutionary constraint^{25,26}, through DNA comparison with rodent, chicken, *Xenopus* and *Fugu* (*P*-value <0.01). A chromosome-wide analysis resulted in 15,325 discrete noncoding regions conserved between human/mouse/rat, 2,429 between human/mouse/chicken, 258 between human/mouse/*Xenopus* and 213 between human/mouse/*Fugu*. We found that the distribution of human/mouse/*Fugu* conserved noncoding sequences was highly uneven along the chromosome (Fig. 3b), with 42 centred in 5p15 around an Iroquois homeobox (*IRX*) gene family. These discrete evolutionarily conserved sequences represent a prioritized substrate for future experimental studies to elucidate their function and potential role in gene regulation.

Gene-poor regions

Recent work has shown that a significant fraction of noncoding elements conserved between human and *Fugu* has gene regulatory activity even though many are located at great distances from the genes whose expression they control²⁷. In addition to their location between conserved flanking genes, evidence to support distant gene regulatory sequences is found in the maintenance of long syntenic blocks across distant evolutionary species²⁸. To determine whether such regions exist on human chromosome 5, we built a segmental homology map between human, chimp, mouse, rat and chicken. This map revealed two segments larger than 3 Mb that do not contain any evolutionary break points or insertions (>250 kb) within all examined species. Notably, despite this high level of conservation, these two large segments have very few known genes and overlap the extreme gene-poor regions at 5p15 (3.1 Mb) and 5q34 (5.0 Mb). In addition, each is highly enriched for conserved noncoding sequences with distantly related non-mammalian vertebrates (Fig. 3c). In contrast to the interleukin cluster (described below) and despite being gene poor, the 5p15 region contains 378, 220 and 42 noncoding elements conserved in rodents, chicken and *Fugu*, respectively³. A similar level of noncoding conservation was observed in the 5q34 gene desert region containing 1,087 noncoding elements conserved with rodents, 301 with chicken, but none with *Fugu*. Although functional studies are needed to determine whether these ancient conserved sequences regulate the limited number of genes in these regions, it is interesting to note that the 5p15 region contains a cluster of *IRX* genes that have multiple roles during pattern formation in vertebrate development. The high density of conserved noncoding elements with extended synten in these gene-poor regions suggests that these regions contain elements that regulate distant genes.

Interleukin cluster

The interleukin gene cluster on 5q31 is a region of particular interest to immunologists because of the presence of five haematopoietic growth factor genes (*IL3*, *CSF2*, *IL5*, *IL13* and *IL4*) and two quantitative trait loci associated with atopic asthma and Crohn's disease susceptibility. From the comparative analysis of this 1 Mb of sequence, we found that 140 of the 190 (76%) human coding exons overlap regions conserved in mouse. This number decreased to 126 (66%) when examining human/mouse/chicken conservation (P -value <0.01 ; Fig. 3d; see also Supplementary Table S4). Consistent with the known fast evolutionary rate of the interleukin genes, most of the interleukin exons (18 of 21) are among the exon sequences that lack similarity between the species. In the analysis of noncoding sequences, we found 83 conserved human/mouse elements that include two previously characterized gene enhancers (CNS-1 and CNS-7)²². One of these elements is more highly conserved than CNS-1 and CNS-7, yet remains functionally undefined. In addition, we found six human/mouse/chicken conserved noncoding sequences, one of which is also conserved in *Xenopus*.

Human disease

Not long after the concept of using anonymous polymorphic DNA markers to localize disease loci was proposed, linkage to many diseases on chromosome 5 was found, and positional cloning and other strategies rapidly isolated the genes for these clearly segregating disorders. So far, mutations in 66 specific genes are known for mendelian diseases (see Supplementary Table S5); an additional 14 single-gene diseases have been mapped to chromosome 5 but have not yet been linked to specific genes. In one of the first examples of a study taking advantage of linkage disequilibrium to positionally clone a gene, ref. 29 identified the DTD gene mutated in diastrophic dysplasia in the Finnish population in 1994. Identification of mutations in the growth hormone receptor gene, at 5p12-p13, in Laron dwarfism was an early case of 'positional candidate cloning', in which the gene was cloned and its location known before mapping the trait³⁰. In addition to SMA, microdeletions in a duplicated region in 5q35 cause Sotos syndrome, a debilitating disorder that results in cranial overgrowth and mental retardation³¹, in which the duplication is thought to mediate severity³². The availability of this completed sequence will further advance our understanding of human disease, and the rate at which disease genes are identified and cloned with causative mutations should be greatly accelerated. □

Methods

Mapping and sequencing

We seeded chromosome 5 with P1, PAC and Caltech BAC clones anchored to a set of 1,645 radiation hybrid markers and known genes, mapping 5,392 clones to chromosome 5 and with 4,943 of these localized by FISH. After constructing a single enzyme restriction digest map, we chose a minimal tiling path. For the SMA duplication regions, hybridization probes were designed at 50-kb intervals across the working maps with additional probes for each uniquely identified duplcon and screened against RPCI-11. Results were binned and ~40% of positives selected for sequencing. Single haplotype maps were constructed by sequence analysis, relying on >30-kb alignments with zero or one discrepancy and multiple clone depth. For the complex 5q13 copy, we used an iterative cycle of probing, sequencing, direct repeat resolution, finishing and re-analysis.

We generated sequence by using a clone-by-clone shotgun sequencing strategy³³ followed by finishing with a custom primer approach. BAC DNA was sheared by using a Hydroshear Instrument (GeneMachines), size selected (3–4 kb) and subcloned into the vector pUC18. Randomly selected subclones were sequenced in both directions using universal primers and BigDye Terminator chemistry to an average depth of $\times 8$. Sequences were assembled and edited by using the Phred/Phrap/Consed suite of programs^{34,35}. After manual inspection of the assembled sequences, clones were finished by re-sequencing and by sequencing off of plasmid subclones or the large insert clone by using custom primers. All finishing reactions were performed with dGTP BigDye Terminator chemistry (Applied Biosystems). Clones with high repeat content or that showed considerable bias when cloned into pUC18 had additional 8–10-kb libraries constructed in a low copy number vector. Recalcitrant areas and difficult to sequence gaps were closed with sequence data derived from transposon sequencing, small insert shatter

libraries³⁶, or PCR. Each clone was finished according to the agreed international standard for the human genome (<http://genome.wustl.edu/Overview/g16stand.php>).

Marker placement

Genetic markers were placed on the genomic sequence using electronic PCR³⁷. Markers were allowed to have up to three mismatches and were subsequently verified by placing the STS sequence (downloaded from UniSTS) via NCBI Megablast using a drop-off value of 180, a match reward of 10, a gap penalty of -20, and a word size of 22.

Pseudogene identification

Pseudogenes were defined as gene models built by homology to known human genes where alignment between the model and the homologue shows at least one stop codon or frameshift mutation. For the fragments of chromosome 5 genomic sequence that were masked of repeats by using RepeatMasker (A. Smit and P. Green, unpublished data)³⁸, we identified homology to human IPI proteins by using NCBI BLASTX. For each fragment of genomic sequence homologous to an IPI protein, we built gene models by using the GeneWise program. The overlapping gene models were clustered and the alignment of the top-scoring model with its human homologue was analysed for the presence of stop codons and frameshifts. The models were then manually analysed to confirm pseudogene status. Sequences of 431 processed pseudogenes that had been identified previously³⁹ were mapped to the genomic sequence of chromosome 5 by using the BLAT tool. Loci with multi-exon mapping, overlaps with the pseudogenes described above, and simple repeats identified by RepeatMasker were eliminated. Pseudogene status of the remaining sequences was manually validated.

Segmental duplication analysis

We used a BLAST-based detection scheme⁴⁰ to identify all pairwise similarities representing duplicated regions (≥ 1 kb and $\geq 90\%$ identity) within the finished sequence of chromosome 5 and compared to all other chromosomes in the NCBI genome assembly (build 34). A total of 1,818 pairwise alignments representing 16.57 Mb of aligned base pairs and 6.26 Mb of non-redundant duplicated bases were analysed on chromosome 5. The program Parasight (J. A. Bailey, unpublished data) was used to generate images of pairwise alignments. We also analysed pairwise alignments for per cent identity and the number of aligned bases. Satellite repeats were detected by using RepeatMasker (version 15 May 2002) on slow settings. Analysis of haplotype structural variation was performed using the program Miropeats (threshold = 7,000)⁴¹.

Comparative analysis

In this work, we used the following genomic assembly builds: chimpanzee November 2003, mouse October 2003, rat June 2003, chicken February 2004 (from <http://genome.ucsc.edu>), *X. tropicalis* v1.0 and *F. rubripes* v3.0 (from <http://jgi.doe.gov/>). All the segmental homology maps in n -dimensions are computed using PARAGON (v2.13; O. Couronne, unpublished data). As input for PARAGON, we used BLASTZ (v6)⁴² DNA pairwise alignments of all the species to human. Slowly evolving regions are extracted from the alignments using PEAK-VISTA (P -value >0.01 ; S. Prabhakar, unpublished data). We built a four-dimension human/chimp/mouse/rat segmental homology map with PARAGON, aligned all the segments with MLAGAN (v12)⁴³ and computed the slowly evolving conserved regions with PEAK-VISTA. Interleukin homology among species was extracted from the PARAGON segmental map, built with MLAGAN multiple alignments; the slowly evolving conserved regions were extracted with RANK-VISTA.

Received 10 May; accepted 2 August 2004; doi:10.1038/nature02919.

1. Frazer, K. A. *et al.* Computational and biological analysis of 680 kb of DNA sequence from the human 5q31 cytokine gene cluster region. *Genome Res.* **7**, 495–512 (1997).
2. Symula, D. J. *et al.* Functional screening of an asthma QTL in YAC transgenic mice. *Nature Genet.* **23**, 241–244 (1999).
3. Loots, G. G. *et al.* Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136–140 (2000).
4. Church, D. M., Yang, J., Bocian, M., Shiang, S. & Wasmuth, J. J. A High-resolution physical and transcript map of the Cri du Chat region of human chromosome 5p. *Genome Res.* **7**, 787–801 (1997).
5. Puechberty, J. *et al.* Genetic and physical analyses of the centromeric and pericentromeric regions of human chromosome 5: Recombination across 5cen. *Genomics* **56**, 274–287 (1999).
6. Riethman, H. C. *et al.* Integration of telomere sequences with the draft human genome sequence. *Nature* **409**, 948–951 (2001).
7. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
8. Schmutz, J. *et al.* Quality assessment of the human genome sequence. *Nature* **429**, 365–368 (2004).
9. Olivier, M. *et al.* A high-resolution radiation hybrid map of the human genome draft sequence. *Science* **291**, 1298–1302 (2001).
10. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31**, 241–247 (2002).
11. Grimwood, J. *et al.* The DNA sequence and biology of human chromosome 19. *Nature* **428**, 529–535 (2004).
12. Heilig, R. *et al.* The DNA sequence and analysis of human chromosome 14. *Nature* **421**, 601–607 (2003).
13. Hiller, L. W. *et al.* The DNA sequence of human chromosome 7. *Nature* **424**, 157–164 (2003).
14. Melki, J. *et al.* De novo and inherited deletions of the 5q13 region in spinal muscular atrophies. *Science* **264**, 1474–1477 (1994).
15. Monani, U. *et al.* A single nucleotide difference that alters splicing patterns distinguishes the SMA gene SMN1 from the copy gene SMN2. *Hum. Mol. Genet.* **8**, 1177–1183 (1999).
16. Chen, Q. *et al.* Sequence of a 131-kb region of 5q13.1 containing the spinal muscular atrophy candidate genes SMN and NAIP. *Genomics* **48**, 121–127 (1998).
17. Locke, D. P. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Res.* **13**, 347–357 (2003).

18. Yunis, J. J. & Prakash, O. The origin of man: a chromosomal pictorial legacy. *Science* **215**, 1525–1530 (1982).
19. Noor, M. A., Grams, K. L., Bertucci, L. A. & Reiland, J. Chromosomal inversions and the reproductive isolation of species. *Proc. Natl Acad. Sci. USA* **98**, 12084–12088 (2001).
20. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
21. Groenen, M. A. *et al.* A consensus linkage map of the chicken genome. *Genome Res.* **10**, 137–147 (2000).
22. Nanda, I. *et al.* 300 million years of conserved synteny between chicken Z and human chromosome 9. *Nature Genet.* **21**, 258–259 (1999).
23. Clark, A. G. *et al.* Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**, 1960–1963 (2003).
24. Hocking, L. J. *et al.* Domain-specific mutations in sequestosome 1 (SQSTM1) cause familial and sporadic Paget's disease. *Hum. Mol. Genet.* **11**, 2735–2739 (2002).
25. Pennacchio, L. A. & Rubin, E. M. Genomic strategies to identify mammalian regulatory sequences. *Nature Rev. Genet.* **2**, 100–109 (2001).
26. Ghanem, N. *et al.* Regulatory roles of conserved intergenic domains in vertebrate Dlx bigene clusters. *Genome Res.* **13**, 533–543 (2003).
27. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302**, 413 (2003).
28. Flint, J. *et al.* Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster. *Hum. Mol. Genet.* **10**, 371–382 (2001).
29. Hästbacka, J. *et al.* The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. *Cell* **78**, 1073–1087 (1994).
30. Barton, D. E., Foellmer, B. E., Wood, W. I. & Francke, U. Chromosome mapping of the growth hormone receptor gene in man and mouse. *Cytogenet. Cell Genet.* **50**, 137–141 (1989).
31. Kurotaki, N. *et al.* Haploinsufficiency of NSD1 causes Sotos syndrome. *Nature Genet.* **30**, 365–366 (2002).
32. Kurotaki, N. *et al.* Fifty microdeletions among 112 cases of Sotos syndrome: low copy repeats possibly mediate the common deletion. *Hum. Mutat.* **22**, 378–387 (2003).
33. International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
34. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using Phred. I. accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
35. Gordon, D., Abajian, C. & Green, P. Consed: A graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998).
36. McMurray, A. A., Sulston, J. E. & Quail, M. A. Short insert libraries as a method of problem solving in genome sequencing. *Genome Res.* **8**, 562–566 (1998).
37. Schuler, G. D. Sequence mapping by electronic PCR. *Genome Res.* **7**, 541–550 (1997).
38. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).
39. Zhang, Z., Harrison, P. M., Liu, Y. & Gerstein, M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.* **13**, 2541–2558 (2003).
40. 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).
41. Parsons, J. D. Miropeats: graphical DNA sequence comparisons. *Comput. Appl. Biosci.* **11**, 615–619 (1995).
42. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
43. Brudno, M. *et al.* LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**, 721–731 (2003).
44. Kurdoa-Kawaguchi, T. *et al.* The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nature Genet.* **29**, 279–286 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the International Chimpanzee Sequencing Consortium for pre-publication access to and permission to analyse the relevant portions of the chimpanzee genomic sequence, and the Washington University Genome Sequencing Center for pre-publication access to the chicken genomic assembly. We also thank M. Christensen, P. Butler and E. Fields for technical support, D. Gordon of the University of Washington for his assistance in developing and customizing finishing tools, T. Furey and G. Schuler for their efforts towards assessing the quality and completeness of our assembly, and P. DeJong for the construction of genomic resources. This work was performed under the auspices of the US DOE's Office of Science, Biological and Environmental Research Program, by the University of California, Lawrence Livermore National Laboratory, Lawrence Berkeley National Laboratory and Stanford University.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.S. (jeremy@shgc.stanford.edu) and E.M.R. (EMRubin@lbl.gov). The chromosome 5 sequence has been deposited in GenBank under accession number NC_000005.