The complete sequence of human chromosome 5

Jeremy Schmutz*, Joel Martin⇓, Astrid Terry⇓, Olivier Couronne?, Jane Grimwood*, Steve Lowry⇓, Laurie A. Gordon#⇓, Duncan Scott⇓, Gary Xieδ⇓, Wayne Huang⇓, Uffe Hellsten⇓, Mary Tran-Gyamfi#⇓, Xinwei She@, Shyam Prabhakar?, Andrea Aerts⇓, Michael Altherrδ⇓, Eva Bajorek*, Stacey Black*, Elbert Branscomb#⇓, Chenier Caoile*, Jean F. Challacombeδ, Yee Man Chan*, Mirian Denys*, Chris Detter⇓, Julio Escobar*, Dave Flowers*, Dea Fotopulos*, Tijana Glavina⇓, Maria Gomez*, Eidelyn Gonzales*, David Goodstein⇓, Igor Grigoriev⇓, Matthew Groza#, Nancy Hammon⇓, Trevor Hawkins⇓, Lauren Haydu*, Sanjay Israni⇓, Jamie Jett⇓, Kristen Kadner⇓, Heather Kimball⇓, Arthur Kobayashi⇓#, Frederick Lopez*, Yunian Lou⇓, Diego Martinez⇓, Catherine Medina*, Jenna Morgan⇓, Richard Nandkeshwar#, James P. Noonan+, Sam Pitluck⇓, Martin Pollard⇓, Paul Predki⇓, James Priest?, Lucia Ramirez*, Sam Rash⇓, James Retterer*, Alex Rodriguez*, Stephanie Rogers*, Asaf Salamov⇓, Angelica Salazar*, Nina Thayerδ⇓, Hope Tice⇓, Ming Tsai*, Anna Ustaszewska⇓, Nu Vo*, Jeremy Wheeler*, Kevin Wu*, Joan Yang*, Mark Dickson*, Jan-Fang Cheng?, Evan E. Eichler@, Anne Olsen#⇓, Len A. Pennacchio?⇓, Daniel S. Rokhsar⇓, Paul Richardson⇓, Susan M. Lucas⇓, Richard M. Myers*, Edward M. Rubin?⇓

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

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

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

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

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

@ Department 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

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

Chromosome 5 is one of the largest human chromosomes yet has one of the lowest gene densities. This is partially explained by numerous gene-poor regions that display a remarkable degree of noncoding and syntenic conservation with non-mammalian vertebrates, suggesting they are functionally constrained. In total, we compiled 177.7 million base pairs of highly accurate finished sequence containing 923 manually curated protein-encoding genes including the protocadherin and interleukin gene families and the first complete versions of each of the large chromosome 5 specific internal duplications. These duplications are very recent evolutionary events and play a likely mechanistic role, since deletions of these regions are the cause of debilitating disorders including spinal muscular atrophy (SMA).

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. These studies of a 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 activities. 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 catalog of genes and evolutionarily conserved noncoding sequences. Many of these insights, as well as clues into disease causing deletions arising from the segmented duplication landscape of chromosome 5, can only now be appreciated with the finished sequence of this chromosome in hand.

Mapping and Sequencing

We initially seeded the chromosome with P1, PAC, and Caltech BAC clones anchored to a set of 1,645 radiation hybrid (RH) markers and known genes, mapping 5,392 clones to chromosome 5 with 4,943 of these localized by fluorescent *in situ* hybridization. After constructing a single enzyme restriction digest map, we chose an initial minimal tiling path to generate a draft sequence. After draft completion in 2001 we selected clones with an approach that integrated all of the publicly available draft sequence, previously reported clone contigs[1-3] including the Celera scaffolds[4], BAC and fosmid end sequences, and BACs isolated with a directed overgo hybridization strategy to close gaps between anchored contigs. The final version of the tiling path contains 1,763 large-insert clones, 96% of which are BACs with 4 gaps remaining, all in the long arm. None of these remaining physical gaps could be cloned in current vector systems or are part of large duplications.

Our standard strategy of seeding then walking based on restriction maps proved unworkable in the duplication region of 5q13 associated with SMA and led to mapping

errors due to the duplication with its primary insertion copy at 5p14 and a secondary copy in 5p13. We thus adopted a strategy of drafting high depth clone coverage from a BAC library (RPCI-11) built from the genome of a single individual to enable the construction of single haplotype paths spanning the duplicated regions. Hybridization probes were designed at 50 kb intervals across the working maps with additional probes for each uniquely identified duplicon and screened against RPCI-11 segments 3, 4 and 5. Probe results were then binned and ~40% of the positives selected for shotgun sequencing. Single haplotype local maps were constructed by sequence analysis, relying on large (>30 kb) alignments with zero or one discrepancy and multiple clone depth. This immediately resolved both 5p copies. To complete the sequence of the more complex 5q13 copy, we used an iterative cycle of probing, sequencing, direct repeat resolution, full clone finishing and reanalysis.

We generated sequence by using a clone-by-clone shotgun sequencing strategy[5] followed by finishing with a custom primer approach. Recalcitrant areas and difficult to sequence gaps were closed with additional sequence data derived from transposon sequencing, small insert shatter libraries[6], or PCR. Each clone was finished according to the agreed international standard for the human genome (http://genome.wustl.edu/Overview/g16stand.php). On the basis of internal and external quality checks, we estimate the accuracy of our finished sequence to exceed 99.99%[7]. In total, we finished 177,702,766 base pairs and estimate the total size of the chromosome, including the four clone gaps and the recalcitrant centromeric and subtelomeric regions, to be 180.8 Mb.

The finished sequence is estimated to cover 99.9% of the euchromatic sequence and to have captured all known genes that were previously mapped to chromosome 5 (T. Furey, personal communication). The Stanford v.4 G3 radiation hybrid (RH) map[8] was compared to the sequence and it matched the marker order well (see supplementary Fig. S1). Thirteen (out of 442 attempted marker placements) RH markers missing from the sequence were found to have been incorrectly assigned to chromosome 5. Recombination distances from the deCODE[9] meiotic maps were compared to physical distances and recombination rates accurately tracking physical distance (see supplementary Fig. S2), as reported for other chromosomes.

Gene Catalog

We placed gene model transcripts on the chromosome 5 genomic sequence and manually reviewed these models by using previously described methods[10.] Ultimately, a total of 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 text). Loci were placed in the following three categories: (1) 'known' genes, (2) 'novel' genes, and (3) 'pseudogenes', consistent with our previous definitions[10]. 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. 827 known loci were identified based on 2,203 RefSeq genes and other nearly full-length cDNA 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 EST variants. Evidence for 55 novel loci was supported by nearly full-length cDNA sequence, spliced ESTs, and/or similarity to known human or mouse gene sequences. Forty-one putative gene loci were modeled using orthologous mouse cDNA sequences. Twenty tRNA genes and four tRNA pseudogenes were predicted, similar in density to chromosome 19[10].

The extent of alternative splicing was characterized based on the existing cDNA and EST data. Considering only mRNA sequences in GenBank, 1,598 distinct transcripts were identified, providing an average coverage of 1.7 annotated transcripts per locus (see supplementary text). These mRNAs provide strong evidence for alternative splicing of 408 (44%) of the 923 loci, each having two or more associated transcripts. 577 pseudogenes and pseudogene fragments were also identified. These represent two classes: (i) 98 non-processed pseudogenes that display a structure similar to the parent locus and are therefore likely to have resulted from genomic duplication events: (ii) 479 processed pseudogenes that presumably resulted from viral retrotransposition of spliced mRNAs (see supplementary text). No significant bias toward over-representation of pseudogenes from a particular gene family was observed.

Chromosome 5 Genomic Duplications

We performed a detailed analysis of duplicated sequence (≥90% sequence identity and ≥1 kb in 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 Fig. S4). Chromosome 5 segmental duplications, however, show a higher degree of sequence identity (≥97.5%), especially with other regions of chromosome 5 (see supplementary Fig. S5) than do the duplications on other chromosomes. Intra-chromosomal duplications are clustered in 10 regions (Fig. 1) and represent the majority of the gene duplications on the chromosome (protocadherins, *PMCHL1*, *SMN1*-*SMN2*, *NYREN7*, etc). The high degree of sequence identity underlying most of these intra-chromosomal genomic duplications suggests that these structures are relatively recent duplications or gene conversion events that emerged during the separation of human and the great apes (see supplementary Fig. S3 and 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 on chromosome 5 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. In contrast, the telomeric regions show extensive interchromosomal duplications (Fig. 1 and see supplementary Fig. S4), with 25% (2.48/9.08 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 mapping to 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 disease[11, 12]. Analysis of carriers and controls suggests extreme variability of this locus, but the underlying structural variation has never been documented at the sequence level[13]. Within the assembled version of chromosome 5, we identified a complex arrangement of 311 pairwise alignments (one sixth of the total 1,769 alignments) (Fig. 1). On average, the duplications are long (~200 kb) and show a high degree of sequence identity (98.66%). Duplications in this region include inter-chromosomal duplications, the majority of which map to chromosome 6, with at least three very large tandem intrachromosomal duplications with high percent identity (>99.5%) and various interspersed intrachromosomal duplications to other regions (Fig. 2). Interestingly, this region is enriched in genes. We annotated fourteen gene loci in this region, including SERF1 (small EDRK-rich factor 1), BIRC1 (baculoviral IAP repeat-containing 1) 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 RPCI11-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 and SMAvar2) revealed extensive structural variation. At least two large-scale rearrangements (>100 kb) and multiple smaller insertion/deletion events are required to reconstruct an ancestral haplotype. Although there are several possible scenarios for the evolution of these two

variants, one explanation may be that a portion of the SMAvar1 region (69.8 Mb to 70.4 Mb) was duplicated (68.9 Mb to 69.4) and subsequently inverted (69.8 to 70.4 Mb) in the second haplotype (0.3 Mb to 0.9 Mb in SMAvar2). Such extensive structural variation between human haplotypes may not be uncommon in regions of extensive segmental duplication.

Protocadherin Gene Family

The largest gene family on chromosome 5 is the protocadherin (*PCDH*) gene cluster, which consists of 53 tandemly-arrayed, single-exon paralogous genes organized into three subclusters, designated α, β and γ[14]. Each protocadherin exon encodes an extracellular domain consisting of six cadherin-like ectodomain repeats, a transmembrane domain and a short cytoplasmic tail. At the 3' end of both the α and γ subclusters are an additional three short exons that are alternatively *cis*-spliced to each α and γ exon, providing a "constant" cytoplasmic region[14-16]. Each protocadherin gene is transcribed from its own promoter and all protocadherin cluster promoters share a highly conserved core motif[17, 18]. Promoter choice appears to determine the splicing of a particular α or γ variable exon to the first constant region exon, in that the splice donor site of the transcribed variable exon is used in *cis*-splicing[15].

Each neuron appears to express a distinct combination of protocadherin genes[19]. Protocadherin proteins are thought to form homophilic interactions at synapses, providing a molecular means to distinguish subsets of neurons based on the combinations of protocadherins they express[19, 20]. Protocadherin clusters are present in many vertebrate species, although the sequence content greatly differs between mammals and other

vertebrates. Protocadherin cluster genes in humans and other species also undergo frequent gene conversion events. These events are restricted to specific ectodomains, resulting in some ectodomains becoming nearly identical among paralogs while other ectodomains remain diverse. This process also generates allelic variants of human protocadherin cluster genes.

Comparative Biology

To understand further the evolution and functional sequences of human chromosome 5, we performed comparative analyses versus 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 non-randomly 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 likely regulate genes from a distance. Finally, we examined conservation in a comparative analysis of the extensively studied interleukin gene cluster region on 5q31.

Synteny

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. While recent experimental studies support that

large-scale rearrangements (40 kb to 175 kb) have frequently occurred during primate genome evolution[21], our comparison of finished human chromosome 5 and the recent chimpanzee draft genome sequence (Intl. Chimpanzee Genome Sequencing Consortium, 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). It has been proposed that these large-scale rearrangements create barriers to fertile mating and triggered the speciation that separated these two lineages[22]. Comparison versus the mouse genome sequence[23] yielded 142 chromosomal rearrangements ranging in size from 200 kb to 17 Mb. Between human and chicken, we found that one-third of human chromosome 5 is homologous to the sex chromosome Z[24], further supporting that sex chromosomes have evolved independently following the avian and mammalian split some 300 Mya[25].

Chimpanzee

In addition to exploring the syntenic relationship between human chromosome 5 and the recent draft assembly of the chimpanzee genome, we catalogued sequence changes between these two primate species. To explore the constraint on human-chimpanzee evolution in non-coding 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/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

that both of the former categories are under evolutionary constraint. This also supports that human/chimpanzee coding and noncoding sequences conserved in rodents have been under modern selective constraint since the last common ancestor of these two primates. We next exploited data that compared the patterns of variation within human and chimpanzee exons to identify genes potentially under positive selection in the human lineage[26]. We found that 21 genes randomly distributed over the length of chromosome 5 display a p-value less than 0.01 for an increased evolutionary rate in the human lineage. 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*), while one study links *SQSTM1* to Paget's disease of the bone[27]. As the chimpanzee genome reaches a further draft state, a similar complete re-analysis of the entire human gene set will likely yield a large number of quickly evolving genes, which may explain aspects of biology unique to *Homo sapiens*.

Vertebrate Conservation

To annotate functional elements, we identified slowly evolving regions, presumably under evolutionary constraint, through DNA comparison with rodent, chicken, *Xenopus* and *Fugu* (p-value < 0.01). A chromosome-wide analysis resulted in 15,325 discrete non-coding regions 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 is highly uneven along the length of the chromosome (Fig. 3b) with 42 centered around an Iroquois homeobox (*IRX*) gene family at 5p15. These discrete evolutionarily conserved

sequences provide an immediate substrate for functional sequences in noncoding DNA, including those important in gene regulation.

Gene poor regions

Recent work has shown that a significant fraction of non-coding elements conserved between human and *Fugu* have gene regulatory activity even though many are located at great distances from the genes whose expression they control[28]. In addition to their location between conserved flanking genes, evidence to support distant gene regulatory sequences is the maintenance of long syntenic blocks across distant evolutionary species[29]. 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 larger than 250 kb within all the species examined. Remarkably, despite this high level of conservation, these two large segments have very few genes, overlapping the extremely gene-poor regions at 5p15 and 5q34 that are 3.1 and 5.0 Mb in size, respectively. In addition, each is highly enriched for conserved noncoding sequences with distant non-mammalian vertebrates (Fig. 3c). In contrast to the Interleukin cluster (described below) and despite being gene poor, the 3.1 Mb 5p15 region contains 378, 220, and 42 noncoding elements conserved in rodents, chicken and *Fugu*, respectively[30]. A similar level of noncoding conservation was observed in the 5.0 Mb gene desert in the 5q34 region, which contains 1,087 noncoding elements conserved with rodents, 301 with chicken, but none in *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 play multiple roles during pattern formation of in vertebrate development. The high density of conserved noncoding elements with extended synteny in these gene poor regions suggests the existence of gene regulatory sequences that regulate the residing genes from a distance.

Interleukin Cluster

The interleukin gene cluster on 5q31 is a region of particular interest to immunologists because of the presence of five hematopoietic growth factor genes (*IL3*, *CSF2*, *IL5*, *IL13*, and *IL4*) and two quantitative trait loci (QTL) associated with atopic asthma and Crohn's disease susceptibility. From the comparative analysis of this 1 Mb, we found that 140 of the 190 (76%) human coding exons overlap regions conserved in mouse. This number decreased slightly to 126 (66%) when examining human-mouse-chicken conservation (p value < 0.01; Fig. 3d; see supplementary Table S4). Consistent with the fast evolutionary rate of the interleukin genes, the majority of the interleukin gene exons (18 of 21) are among the exon sequences that lack similarity in the analyzed species. Exons of two hypothetical protein genes (JGI_962 and LOC375468) in the 5q31 cluster also fall into this category. 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)[25]. One of these conserved 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*. These six conserved noncoding sequences display strong evolutionary constraint and

represent a prioritized substrate for future experimental studies to elucidate their function(s).

Human Disease

Not long after the concept of using anonymous polymorphic DNA markers to localize disease loci was proposed, linkages for many diseases on chromosome 5 were found, positional cloning and other strategies rapidly isolated the genes for these clearly segregating disorders. To date, mutations in 66 specific genes are known for Mendelian diseases, an additional 14 single-gene diseases have been finely mapped to the chromosome 5 and not yet linked to specific genes. In one of the first examples to take advantage of linkage disequilibrium to positionally clone a gene, Hästbacka *et al*. identified the DTD gene mutated in dyastrophic dysplasia in the Finnish population in 1994[31]. 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 prior to mapping the trait[32]. As mentioned previously, some of the more prominent disorders are caused by structural deletions in intrachromosomal duplication regions. Spinal muscular atrophy-1 (SMA), which has an incidence of 1 out of 6,000 newborns, can be caused by an inherited deletion in 5q13, mutations in the SMN-1 gene, or arise spontaneously[11]. Microdeletions in a duplicated region in 5q35 cause Sotos syndrome, a debilitating disorder that results in cranial overgrowth and mental retardation[33], in which the duplication is thought to mediate severity[34]. Chromosome 5 genes are also involved in obesity, deafness, epilepsy, a variety of eye disorders, muscular dystrophies, ataxias, startle disease, blood clotting

disorders, a wide variety of metabolic disorders, and diseases of many other organ systems have been cloned and used in diagnostic and functional studies to date (see supplementary Table S5). 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

Sequencing and finishing methods

BAC DNA was hydrodynamically sheared by using a Hydroshear Instrument (GeneMachines, San Carlos, CA), size selected (3-4kb) and subcloned into the plasmid vector pUC18. Randomly selected plasmid subclones were sequenced in both directions using universal primers and BigDye Terminator chemistry to an average sequence depth of 8x. Sequences were then assembled and edited by using the Phred/Phrap/Consed suite of programs[35, 36]. Following manual inspection of the assembled sequences, clones were finished by resequencing plasmid subclones and by walking on plasmid subclones or the large insert clone by using custom primers. All finishing reactions were performed with dGTP BigDye Terminator chemistry (Applied Biosystems, Foster City, CA). Finished clones contain no gaps and are estimated to contain less than one error per 10,000 base pairs. Clones with a very high repeat content or which showed considerable bias when cloned into the pUC derived vector had additional 8-10 kb libraries constructed in an alternate low copy number vector.

Marker Placement

Genetic markers were placed on the genomic sequence using E-PCR[37]. Markers were allowed to have up to 3 mismatches and were subsequently verified by placing the STS sequence, downloaded from UniSTS, by using NCBI Megablast using the parameters (-D3 –U T –F m –J F –X 180 –r 10 –q –20 –R T –W 22).

Pseudogene identification

Pseudogenes were defined as gene models built by homology to known human genes where alignment between the model and the homolog shows at least one stop codon or frameshift mutation. For the fragments of genomic sequence of the chromosome 5 masked of repeats by using RepeatMasker (A. Smit and P. Green, unpublished),[38] we identified homology to human IPI proteins by using NCBI BlastX. For each fragment of genomics 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 homolog was analyzed for the presence of stop codons and frameshifts. The models were then manually analyzed to confirm pseudogene status. Sequences of 431 processed pseudogenes earlier identified by Zhang *et al.*[39], were mapped to the genomic sequence of the 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 scheme[40] to identify all pairwise similarities representing duplicated regions (≥1 kb and ≥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 basepairs and 6.26 Mb of non-redundant duplicated bases were analyzed on chromosome 5. The program Parasight (Bailey, unpublished) was used to generate images of pairwise alignments. We also analyzed pairwise alignments for percent identity and the number of aligned bases. Satellite repeats were detected by using RepeatMasker (version: 2002/05/15) on slow settings. Analysis of haplotype structural variation was performed using the program *Miropeats* (threshold =7000)[41].

Comparative Analysis

In this work, we used the following genomes freezes: chimpanzee November 2003, mouse October 2003, rat June 2003, chicken February 2004 (downloaded at http://genome.ucsc.edu), *Xenopus tropicalis* v1.0 and *Fugu rubripes* v3.0 (downloaded at http://jgi.doe.gov/). All the segmental homology maps in *n*-dimensions are computed using PARAGON (v2.13; Couronne, unpublished work). As input for PARAGON, we used

BLASTZ (v6)[42] DNA pairwise alignments of all the species to human. Slow evolving regions are extracted from the alignments using GUMBY (p-value > 0.01; Prabhakar, unpublished work). Chimpanzee: We built a 4-dimension human/chimp/mouse/rat segmental homology map with PARAGON, aligned all the segments with MLAGAN (v12)[43] and computed the slow evolving conserved regions with GUMBY. Interleukin: Homology among species extracted from the PARAGON segmental map, the multiple alignments are done with MLAGAN and the slow evolving conserved regions are extracted with GUMBY.

References

1. 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).
2. 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).
3. Riethman, H. C. *et al*. Integration of telomere sequences with the draft human genome sequence. *Nature* **409,** 948-951 (2001).
4. Venter J. C. *et al*. The sequence of the human genome. Science **291**, 1304-1351 (2001).
5. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
6. 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).
7. Schmutz J. *et al*. Quality assessment of the human genome sequence. Nature. In press.
8. Olivier, M. *et al*. A High-Resolution Radiation Hybrid Map of the Human Genome Draft Sequence. *Science* **291**, 1298-1302 (2001).
9. Kong, A. *et al*. A high-resolution recombination map of the human genome. *Nature Genet*. **31**, 241-247 (2002).
10. Grimwood J. *et al*. The DNA sequence and biology of human chromosome 19. *Nature* **428**, 529-535 (2004).

11. Melki, J. *et al.* De novo and inherited deletions of the 5q13 region in spinal muscular atrophies. *Science* **264**, 1474-1477 (1994).

12. 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).

13. 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).

14. Wu Q. & Maniatis T. A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell* **97**, 779-790 (1999).

15. Tasic B. *et al*. Promoter choice determines splice site selection in protocadherin and pre-mRNA splicing. *Mol Cell* **10**, 21-33 (2002).

16. Wang, X, Su, H. & Bradley, A. Molecular mechanisms governing *Pcdh-* gene expression: evidence for a multiple promoter and *cis*-alternative splicing model. *Genes Dev*. **16,** 1890-1905 (2002).

17. Wu Q. *et al*. Comparative DNA sequence analysis of mouse and human protocadherin clusters. *Genome Res* **11**, 389-404 (2001).

18. Noonan, J.P. *et al*. Extensive linkage disequilibrium, a common 16.7 kilobase deletion, and evidence of balancing selection in the human protocadherin cluster. *Am. J. Hum.Genet.* **72**, 621-635. (2003).

19. Kohmura N. *et al*. Diversity revealed by a novel family of cadherins expressed in neurons at a synaptic complex. *Neuron* **20**, 1137-1151 (1998).

20. Obata, S. *et al*. Protocadherin Pcdh2 shows properties similar to, but distinct from, those of classical cadherins. *J. Cell Sci.* **108**, 3765-3773 (1995).

21. Locke, D.P. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Res.* **13,** 347-357 (2003).

22. Noor, M. A., Grams, K. L., Bertucci, L.A, & Reiland, J. Chromosomal inversions and the reproductive isolation of species. PNAS 98, 12084-8 (2001).

23. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).

24. Groenen *et al*. A consensus linkage map of the chicken genome. *Genome Res.* **10,** 137-147 (2000).

25. Nanda, I. *et al.* 300 million years of conserved synteny between chicken Z and human chromosome 9. *Nature Genet.* **21,** 258-259 (1999).

26. Clark, A. G. *et al*. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios, *Science* **302,** 1960-1963 (2003).

27. 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).

28. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302,** 413 (2003).

29. 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).

30. 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).

31. 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).

32. 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).

33. Kurotaki, N. *et al*. Haploinsufficiency of NSD1 causes Sotos syndrome. *Nature Genetics* **30**, 365-366 (2002).

34. Kurotaki, N. *et al*. Fifty microdeletions among 112 cases of Sotos syndrome: low copy repeats possibly mediate the common deletion. *Human Mutat*. **22**, 378-187 (2003).

35. 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).

36. Gordon, D., Abajian, C., and Green, P. Consed: A graphical tool for sequence finishing. *Genome Res.* **8**, 195-202 (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).

Acknowledgements We thank the International Chimpanzee Sequencing Consortium for pre-publication access to and permission to analyze 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 Mari Christensen, Paul Butler, and Elizabeth Fields for technical support, David Gordon of the University of Washington for his assistance in developing and customizing finishing tools, Terry Furey and Greg Schuler for their efforts toward assessing the quality and completeness of our assembly, and Pieter 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.

Correspondence and requests for material should be addressed to J. S. (jeremy@shgc.stanford.edu) and E.M.R (EMRubin@lbl.gov).

Table 1. Chromosome 5 sequence features

Sequence Length	177,702,766
GC content	39.5%

Gene Loci	923	with 1598 Full-length transcripts
Known	827	
Novel	55	
Putative	41	
Non-processed Pseudogenes	98	
Processed Pseudogenes	479	
tRNAs	20	
tRNA Pseudogenes	4	
Repeat content	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 & Low Complexity	2,594,624 (1.5%)	
Other	27,134,827 (15.3%)	

Figure Legends:

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 SMA region showing both SMAvar1, the published variant, and SMAvar2, the alternative RPC11 variant. **a**, Self_dot_plot[44] (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 is 200bp perfect match. The three largest near-identical repeats are colored pink, blue and yellow. The other colored boxes represent smaller identical repeats. **b**, RPCI-11 BAC clone path through SMAvar1 region, red clones are in the final tiling path, gray 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). The underlying pairwise alignments of segmental duplications (>95% >1kb) are depicted as a function of % identity below the horizontal line with different colors corresponding to the location of the pairwise alignment on different human chromosomes (light pink is chromosome 5; dark pink is chromosome 6; yellow is chromosome 3). **e,** A comparison the interhaplotype structure between the two variants using Miropeats[41] with threshold 7000. **f,** Gene content of SMAvar2. **g,** The duplication pattern for SMAvar2.

Figure 3. Comparative Biology. **a,** Segmental homology maps between human chromosome 5 and the mouse, rat, and chicken genomes (see Methods). **b,** Non-coding 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 shown expanded in subfigure c. **c**, The two largest human/mouse/rat/chicken homologuous segments overlap gene poor regions with a high density of conserved non-coding 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 enhancers[25]. These are conserved only in rodents (see text). For clarity only one isoform per gene is shown. In subfigures 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-value) (GUMBY, see Methods).