The DNA sequence and biology of human chromosome 19

Jane Grimwood[1], Laurie A. Gordon[2,3], Anne Olsen[2,3], Astrid Terry[2], Jeremy Schmutz[1], Jane Lamerdin[2,3], Uffe Hellsten[2], David Goodstein[2], Olivier Couronne[2], Mary Tran-Gyamfi[2,3], Andrea Aerts[2], Michael Altherr[2,4], Linda Ashworth[2,3], Eva Bajorek[1], Stacey Black[1], Elbert Branscomb[2,3], Sean Caenepeel[2], Anthony Carrano[2,3], Chenier Caoile[1], Yee Man Chan[1], Mari Christensen[2,3], Catherine A. Cleland[2,4], Alex Copeland[2], Eileen Dalin[2], Paramvir Dehal[2], Mirian Denys[1], John C. Detter[2], Julio Escobar[1], Dave Flowers[1], Dea Fotopulos[1], Carmen Garcia[1], Anca M. Georgescu[2,3], Tijana Glavina[2], Maria Gomez[1], Eidelyn Gonzales[1], Matthew Groza[2,3], Nancy Hammon[2], Trevor Hawkins[2], Lauren Haydu[1], Isaac Ho[2], Wayne Huang[2], Sanjay Israni[2], Jamie Jett[2], Kristen Kadner[2], Heather Kimball[2], Arthur Kobayashi[2,3], Vladimer Larionov[5], Sun-Hee Leem[5], Frederick Lopez[1], Yunian Lou[2], Steve Lowry[2], Stephanie Malfatti[2,3], Diego Martinez[2], Paula McCready[2,3], Catherine Medina[1], Jenna Morgan[2], Kathryn Nelson[2,4], Matt Nolan[2], Ivan Ovcharenko[2,3], Sam Pitluck[2], Martin Pollard[2], Anthony P. Popkie[6], Paul Predki[2], Glenda Quan[2,3], Lucia Ramirez[1], Sam Rash[2], James Retterer[1], Alex Rodriguez[1], Stephanine Rogers[1], Asaf Salamov[2], Angelica Salazar[1], Xinwei She[6], Doug Smith[2], Tom Slezak[2,3], Victor Solovyev[2], Nina Thayer[2,4], Hope Tice[2], Ming Tsai[1], Anna Ustaszewska[2], Nu Vo[1], Mark Wagner[2,3], Jeremy Wheeler[1], Kevin Wu[1], Gary Xie[2,4], Joan Yang[1], Inna Dubchak[2], Terrence S. Furey[7], Pieter DeJong[8], Mark Dickson[1], David Gordon[9], Evan E. Eichler[6], Len A. Pennacchio[2], Paul Richardson[2], Lisa Stubbs[2,3], Daniel S. Rokhsar[2], Richard M. Myers[1], Edward M. Rubin[2] & Susan M. Lucas[2]

[1]*Stanford Human Genome Center, Department of Genetics, Stanford University School of Medicine, 975 California Avenue, Palo Alto, California 94304, USA*
[2]*DOE's Joint Genome Institute, 2800 Mitchell Avenue, Walnut Creek, California 94598, USA*
[3]*Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, California 94550, USA*
[4]*Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[5]*Laboratory of Biosystems and Cancer, National Cancer Institute, NIH, Bethesda, Maryland 20892, USA*
[6]*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*
[7]*Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA*
[8]*Children's Hospital Oakland, Oakland, California 94609, USA*
[9]*Howard Hughes Medical Institute at the Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA*

Chromosome 19 has the highest gene density of all human chromosomes, more than double the genome-wide average. The large clustered gene families, corresponding high G+C content, CpG islands and density of repetitive DNA indicate a chromosome rich in biological and evolutionary significance. Here we describe 55.8 million base pairs of highly accurate finished sequence representing 99.9% of the euchromatin portion of the chromosome. Manual curation of gene loci reveals 1,461 protein-coding genes and 321 pseudogenes. Among these are genes directly implicated in mendelian disorders, including familial hypercholesterolaemia and insulin-resistant diabetes. Nearly one-quarter of these genes belong to tandemly arranged families, encompassing more than 25% of the chromosome. Comparative analyses show a fascinating picture of conservation and divergence, revealing large blocks of gene orthology with rodents, scattered regions with more recent gene family expansions and deletions, and segments of coding and non-coding conservation with the distant fish species *Takifugu*.

The finished human chromosome 19 sequence, comprising a gene density more than double the genome-wide average[1,2], marks the culmination of 18 yr of research spanning the history of modern genomics. The Department of Energy's (DOE) role in the project was initiated through an effort to understand how the body responds to and repairs radiation damage[3,4]. A link between unrepaired DNA damage and human cancer[5], and the subsequent mapping of multiple DNA repair genes to the chromosome[6], provided the impetus for the DOE to select the chromosome as one of its sequencing targets[7]. The physical mapping was completed at Lawrence Livermore National Laboratory (LLNL) in 1995 with the generation of the first high-quality metric map of a human chromosome[8]. In 1999, sequencing and finishing was transferred from LLNL to the DOE's Joint Genome Institute's Production Sequencing Facility and the Stanford Human Genome Center.

The clone map and finished sequence

A complete *Eco*RI restriction map of chromosome-19-specific cosmids[9] provided the foundation for sequencing human chromosome 19. A physical metric across the entire map was generated by estimating the distances between specific points in the cosmid contigs, using fluorescence *in situ* hybridization in sperm pronuclei[10,11]. These initial efforts provided a metrically resolved scaffold framework of 216 cosmid reference points defining contig order and orientation, gap locations and gap sizes, and efficiently directed future map closure efforts (see Supplementary Methods). Gaps in this map were subsequently filled by screening 100-fold clone coverage of the genome using available large-insert genomic libraries (see Supplementary S1). The final tiling path consists of 860 clones that span the entire euchromatin regions of both the p and q arms of the chromosome. The path consists of 511 chromosome-19-specific cosmids, 333 bacterial artificial chromosomes, 11 fosmids, three P1-derived artificial chromosomes, one yeast artificial chromosome clone and one genomic polymerase chain reaction (PCR) product. The chromosome is represented in four contigs, one of which covers the entire q arm, and the clone map is estimated to cover 99.9% of the euchromatin sequence.

Sequence was generated using a clone-by-clone shotgun sequencing strategy[1] followed by finishing using a custom primer approach. Recalcitrant areas or hard gaps were closed with additional sequence data derived from transposon sequencing, small insert shatter libraries 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%[12]. In total, we finished 55,785,651 base pairs (bp) and estimate the total size of the chromosome, including the two clone gaps and the recalcitrant centromeric and subtelomeric regions, to be 63.8 megabases (Mb).

Comparison to physical and genetic maps

We observed strong concordance between the chromosome 19 sequence and previously existing physical and genetic maps. All sequence-tagged sites from the Genethon microsatellite-based genetic map[13], the deCODE map[14] and the Marshfield genetic maps[15] were present in the chromosome 19 sequence (see Supplementary Methods).

We compared recombination distances in the deCODE female, male and sex-averaged meiotic maps[14] with physical distance as determined from the sequence assembly. Recombination statistics for chromosome 19 are similar to other human chromosomes, with the female and sex-averaged comparisons showing a relatively linear relationship between recombination and physical distances, with an average of 2.1 cM Mb^{-1} (Fig. 1). The male meiotic map, however, shows striking differences, particularly in the q arm, showing a long 'desert' with a meiotic distance of only 2.7 cM in the 20.7–44.1 Mb euchromatic region surrounding the centromere (<0.18 cM Mb^{-1}). Although the basis for this result is not clear, it is interesting that this segment of the chromosome is particularly rich in long interspersed nuclear element (LINE) sequences, a finding in agreement with other chromosomes[16]. Also consistent with other chromosomes is a marked increase in male recombination near both telomeres[16].

The chromosome landscape

A hallmark feature of chromosome 19 is its unusually high density of genes. On average, 26 protein-coding gene loci were found per megabase, and exons cover 3.55 Mb of the sequence (6.4%), a percentage significantly higher than the genome-wide average of 1.5%[2]. Protein-coding loci (exons plus introns) span 28.1 Mb (50% of the chromosome) and the average annotated mature transcript contains 58% and 42% coding and untranslated regions, respectively.

Chromosome 19 is also unusual in its density of repeat sequences. Nearly 55% of this chromosome consists of repetitive elements (Table 1), whereas chromosomes 6, 7, 14, 20, 21 and 22 all have repeat contents ranging from 40% to 46% (the genome average is 44.8%)[1,17–22]. This difference is due mainly to an unusually high content of short interspersed nuclear elements (SINEs) on chromosome 19. Specifically, *Alu* repeats make up 25.8% of the chromosome, compared with 13.8%, 13.3%, 9.5% and 16.8% on chromosomes 7, 14, 21 and 22, respectively.

In addition, the G+C content of the chromosome is unusually high, with an average of 48%. This compares with 41% as reported in the whole human genome analysis[1]. G+C content can be separated into two distinct chromosomal categories: regions of human-specific gene family expansions and regions of 1:1 gene orthology blocks with mouse (Fig. 2). In the 20 large duplicated gene family regions (covering a combined 15 Mb of the chromosome) the average G+C content is 43.1%. In contrast, the G+C content is significantly higher (50.3%) in the remaining 38 Mb of the chromosome where there is a clear 1:1 human/mouse orthologous relationship (see 'Comparative biology' section). This high G+C content correlates with gene and *Alu* repeat density, and is negatively correlated with LINE density. Finally, two-thirds of genes have at least one CpG island and 1,623 (88%) of these 1,841 elements are found within 2,000 bp upstream to 1,000 bp downstream of the putative transcription start sites of annotated genes.

The gene and protein catalogue of chromosome 19

An automated pipeline of evidence-based and *ab initio* methods was used to place gene model transcripts on the underlying genomic sequence. Subsequent to this, each transcript was manually reviewed using a combination of human-expressed sequence evidence (messenger RNA and expressed sequence tags (ESTs)) and homology to known genes in humans, mice and other organisms. Additional genes were identified manually from underlying experimental data. Ultimately, a total of 1,461 protein-coding regions were verified as valid gene loci (see Supplementary S2). These loci contain 2,341 full-length (or nearly full-length) transcripts, as well as partial evidence for additional splice variants as discussed below. We placed loci in the following three categories: (1) 'known' genes that correspond to RefSeq genes[23], human complementary DNA or protein sequences; (2) 'novel' or previously unidentified loci that have an open reading frame (ORF) greater than 100 amino acids, and/or are identical to a spliced human EST, and/or have homology to known genes or proteins (all species); and (3) 'pseudogenes', which have sequence similar to genes or proteins found elsewhere in the genome but lack the introns present in the original version (processed) and/or have a disrupted or partial ORF. Transcripts for which a unique ORF could not be determined and putative genes (*ab initio* models) with no supporting experimental evidence were not considered valid.

We identified 1,320 known loci based on 1,551 RefSeq genes and other nearly full-length cDNA sequences in GenBank that mapped to chromosome 19. On the basis of EST evidence, we were able to extend 60% of these RefSeq transcripts by more than 50 nucleotides at the 5′ and/or 3′ ends while maintaining the original ORF. A total of 41% of the RefSeq loci were extended at the 5′ end, more accurately locating the transcriptional start site for these transcripts.

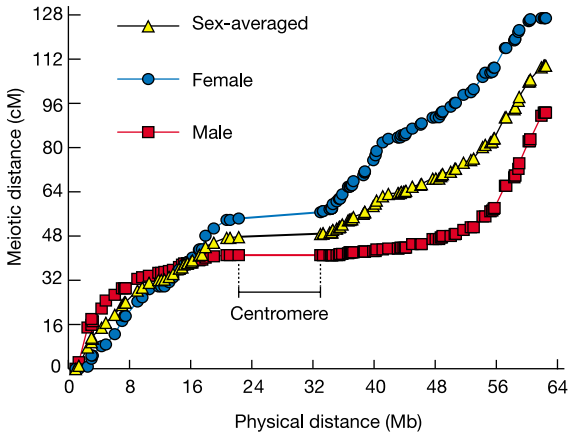

Figure 1 Chromosome 19 meiotic distance versus sequence-based physical distance. The genetic and physical maps were aligned from the telomere of the short arm to the telomere of the long arm. The position of each genetic marker on the female, male and the sex-averaged genetic map is indicated.

Table 1 **Interspersed repetitive elements**

Element	Coverage (bp)	Coverage (%)
Alu	14,415,071	25.83
L1	5,551,771	9.95
L2	1,215,945	2.18
LTR	1,178,970	2.11
MER	1,837,372	3.29
MIR	864,988	1.55
HERV	1,057,852	1.90
MLT	9,270,34	1.66
Simple	781,731	1.4
Low complexity	439,717	0.79
Other	2,838,415	5.09
Total	31,108,866	55.75

HERV, human endogenous retrovirus-like; LTR, long terminal repeat; MER, medium reiterated frequency repeat; MIR, mammalian interspersed repeat; MLT, mammalian LTR transposon.

A total of 88% of the transcripts end with a stop in the final exon/untranslated region.

We found evidence for 141 novel loci for which RefSeq genes were not available. These loci are supported by other nearly full-length cDNA sequences, one or more spliced ESTs and/or similarity to known human or mouse sequences. Within this group are 58 human loci modelled using orthologous mouse cDNA sequences. Transfer RNA genes are one of the best-understood non-protein-coding RNA genes with respect to function. We confidently predicted 11 tRNA genes and three tRNA pseudogenes, in stark contrast with the 157 tRNAs found on the p arm of chromosome 6 (ref. 17).

The largest gene on chromosome 19 is the alpha 1A subunit of the P/Q-type voltage-dependent calcium channel (*CACNA1A*), which extends over more than 300 kilobases (kb) and contains 47 exons. The transcript with the most exons is the skeletal muscle ryanodine receptor (*RYR1*), which has 105 exons spread over nearly 154 kb. The largest exon on the chromosome is 21,693 bp and is found within the *MUC16* gene, a gene encoding an unusually large transmembrane glycoprotein with a role in embryonic development and neoplastic transformation[24].

Alternative splicing

We characterized the extent of alternative splicing based on the existing cDNA/EST data. Considering only mRNA sequences in GenBank, we identified 2,341 distinct chromosome 19 transcripts that provided an average coverage of 1.6 annotated transcripts per locus (see Supplementary S2). These mRNAs provide strong evidence for alternative splicing of 568 (39%) of the 1,461 annotated gene loci, with each having two or more associated transcripts. Furthermore, an additional 452 genes have at least one EST sequence overlapping with non-annotated exons and also contain

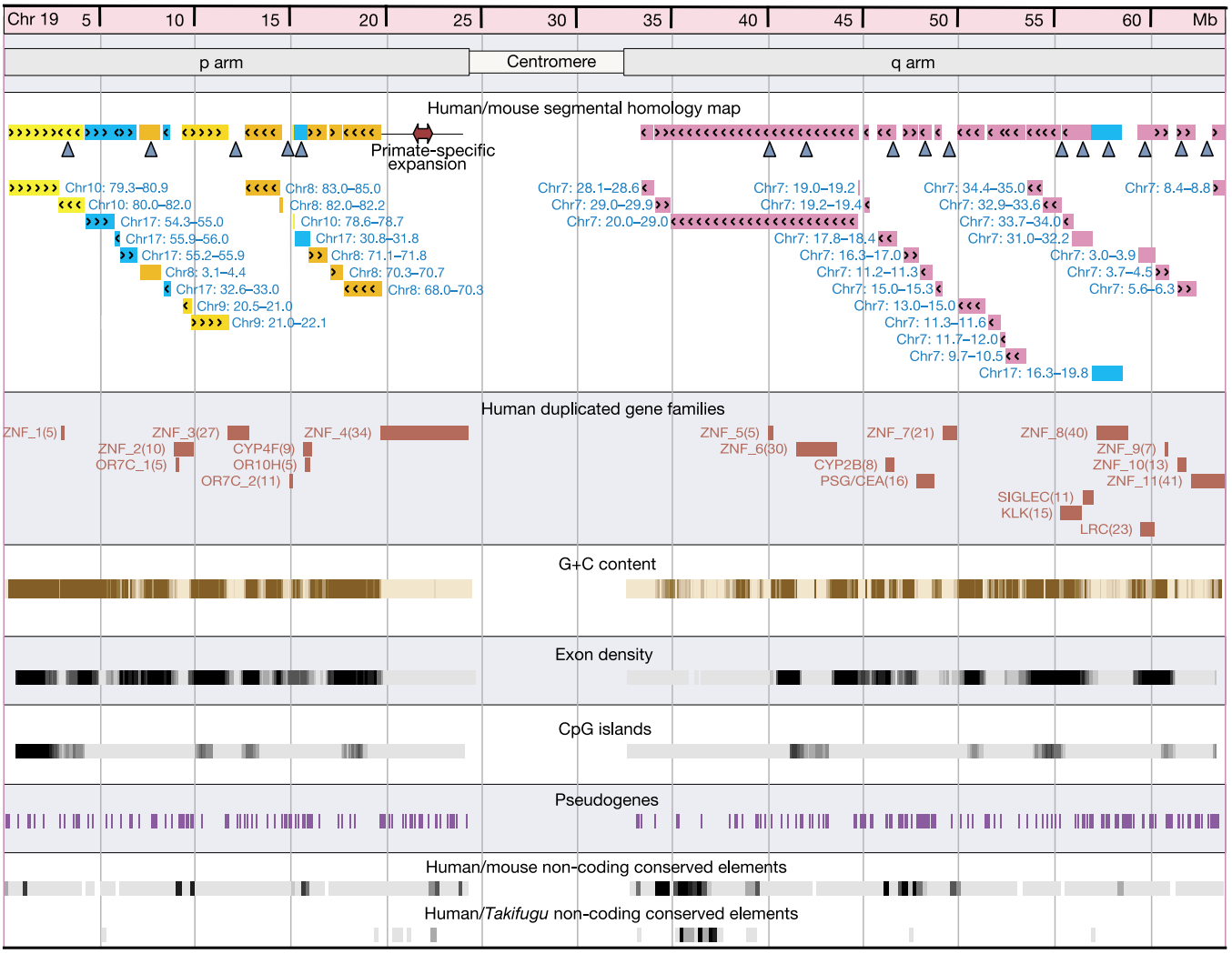

Figure 2 The chromosome 19 landscape. The following sections appear in order from top to bottom: (1) human/mouse blocks of homology larger that 100 kb. The grey triangles indicate regions where significant lineage-specific and no 1:1 orthology can be identified. (2) Same as (1) but showing positions in the mouse genome. This map shows extensive intrachromosomal and interchromosomal rearrangement between human and mouse. (3) The duplicated gene families covering more than 25% of the chromosome (see 'Gene families and duplication analysis' section). Numbers in parentheses are the number of functional genes in the cluster (see also Table 2). (4) G+C content (see 'Chromosome landscape' section). Note that most of the zinc finger and olfactory receptor gene families have a lower G+C content whereas the rest of the chromosome has an unusually high G+C content (although the G+C content is not high in the corresponding mouse syntenic regions). (5) Exon density using a 1-Mb sliding window. The exon density correlates with the G+C content fluctuations. (6) The location of CpG islands on chromosome 19 (see 'Chromosome landscape' section). (7) Pseudogenes identified during the annotation of chromosome 19 (see 'Pseudogene' section). (8) Human/mouse and human/*Takifugu* non-coding DNA element density. This density is uneven over the chromosome with the 5-Mb region in the proximal portion of the q arm containing the highest density of human/mouse non-coding elements and the majority of the non-coding human/*Takifugu* elements. These regions have a low G+C content, a low gene density and fewer mouse/human breakpoints (see 'Comparative biology' section).

flanking canonical splice sites at the genomic locus. Thus, existing experimental data support alternative splicing for a minimum of 1,020 of the genes (70%) on chromosome 19.

It is likely that an even larger fraction of chromosome 19 genes are subject to some form of alternative splicing. As most of our conclusions are based on existing transcribed sequence data, the depth of the EST database seems to be a limiting factor. In fact, of the 184 genes with a total of 500 or more overlapping ESTs, 181 (98%) displayed low-frequency alternative transcripts. Thus deeper EST clone coverage would probably show that a very large fraction of loci can exhibit alternate splicing. Recent studies support this conclusion[25].

Pseudogenes

We identified 321 pseudogenes on chromosome 19 (see Supplementary S3). Of these, 177 (55%) were classified as 'processed' pseudogenes, that is, products of viral retrotransposition events involving spliced messenger RNAs that can frequently be identified by the absence of introns that are present in the parent locus and by the presence of poly(A) stretches embedded in the adjacent genomic sequence. Forty-six (14%) pseudogenes probably arose from genomic duplication events, displaying remnants of introns from the parent locus. An additional 98 (31%) pseudogenes were classified as potential pseudogene fragments owing to their partial nature.

A total of 70 (22%) of the 321 pseudogenes on chromosome 19 contain uninterrupted, but partial, ORFs and probably represent young processed pseudogenes that have not had sufficient time to accumulate mutations to disrupt their ORF, but may also have retained some function. Of the 22 olfactory receptor loci annotated as pseudogenes, four contain a complete ORF. Recent studies have shown that a significant fraction of putative olfactory receptor pseudogenes in the genome are segregating as alleles with intact, presumably functional copies in the human population[26]. Whether any olfactory receptor, or other, pseudogenes on chromosome 19 also vary in humans between such potential functional and non-functional states remains to be explored experimentally.

Gene families and duplication analysis

Chromosome 19 is notable for the prevalence of duplication structures of two types: tandemly clustered gene families and large segmental duplications. As to the latter, chromosome 19 shows evidence of extensive genomic duplication with 7.35% of the sequence sharing sequence homology to more than one location in the genome (Fig. 3; see also Supplementary S4). In contrast, whole-genome estimates of segmental duplication predict 5–6% duplicated sequence using the same alignment parameters (≥1 kb length; ≥90% sequence identity)[1]. This enrichment on chromosome 19 is predominantly due to an increase in intrachromosomal duplications (6.20% of the sequence) rather than interchromosomal duplications (1.69% of the sequence). Using sequence divergence as a surrogate of evolutionary age, these data indicate that most of the tandem expansions of duplications on chromosome 19 occurred much earlier (30–40 million years ago) when compared with the more recent interchromosomal duplication events. The most marked feature of the segmental duplication pattern on chromosome 19 is the pattern of large intrachromosomal duplications (>90% sequence identity) clustered in tandem arrangement (Fig. 3; see also Supplementary S5).

More than 25% of the genes on chromosome 19 are members of large, well-defined, tandemly clustered gene families (Fig. 4 and Table 2). Considerable evidence has documented the existence of lineage-specific changes within these and other tandemly clustered families due to ongoing gene duplication, deletion and mutational

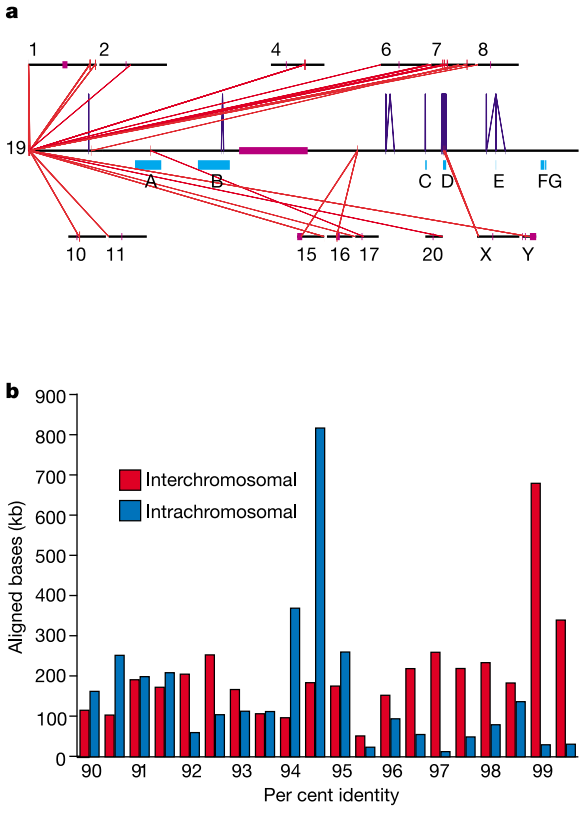

Figure 3 Recent segmental duplications on chromosome 19. **a**, Large (>20 kb), highly similar (>95%) intrachromosomal (blue) and interchromosomal (red) segmental duplications are shown for chromosome 19. Chromosome 19 is drawn at a greater scale relative to the other chromosomes. Gene clusters detected in duplicated sequences (>1 kb with identity >90%) are represented as light blue bars below the chromosome 19 sequence. A, B, ZNF genes; C, cytochrome P450; D, pregnancy-specific α-1-glycoprotein (PSG); E, chorionic gonadotropin β-peptide (CGB); F, leukocyte immunoglobulin-like receptor; G, killer cell immunoglobulin-like receptor. **b**, Sequence similarity of segmental duplications. For all pairwise alignments, the total number of aligned bases was calculated and binned based on per cent sequence identity. Sequence identity distributions for interchromosomally (red) and intrachromosomally (blue) duplicated bases are shown.

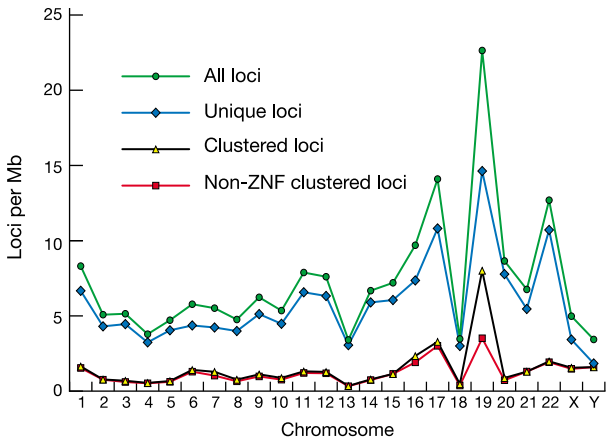

Figure 4 Chromosome 19 is extremely gene dense relative to other human chromosomes. Known genes (November 2003) were downloaded from the UCSC browser (http://genome.ucsc.edu) and plotted relative to the relative chromosome size (minus the centromere). Yellow triangles (clustered loci) indicate genes in tandem duplications, whereas blue diamonds (unique loci) indicate loci that are not duplicated in a tandem manner. Green circles (all loci) represent the sum of clustered and unique loci. A subset of the clustered loci that does not involve the KRAB-Kruppel ZNF genes is shown as red squares (non-ZNF clustered loci). As well as having the highest number of genes, chromosome 19 also has the largest number of genes contained in tandem gene families.

events[27–34]. These clustered sets of paralogues therefore represent a potentially rich source of genetic diversity, and because of their prevalence, chromosome 19 has an especially dynamic evolutionary history. The largest group of such genes on chromosome 19 encodes Krüppel-type (or C2H2) zinc finger transcription factor (ZNF) proteins, with 266 of the approximately 800 total human genes of this type located primarily within 11 large familial clusters[27,29]. Chromosome 19 contains members of several different ZNF gene subfamilies but most of the clustered genes belong to the KRAB-ZNF subtype (Table 2). One large cluster of ZNF genes, located in the pericentromeric region of the short arm, is exclusive to and arose early in the primate lineage[30]. A unique aspect of this region is the admixture of classical centromeric β-satellite sequences with the ZNF genes. A total of 27 blocks of β-satellite repeat (each ranging from 10 to 50 kb; mean = 22 ± 8 kb) map immediately distal to the centromeric α-satellite sequence. These blocks are located throughout the first 4 Mb of the pericentromeric region of 19p12 with an average of 114 kb separating each β-satellite block. Embedded between most of these β-satellite blocks are 1 to 2 KRAB-ZNF genes, indicating that these structures were co-ordinately duplicated[30].

Human chromosome 19 also carries a large collection of genes encoding receptor proteins with immunoglobulin-like domains. Genes of the closely related leukocyte immunoglobulin-like receptors (LILRA and LILRB), the leukocyte-associated immunoglobulin-like receptors (LAIR) and the killer cell immunoglobulin receptors (KIR) are found together in the leukocyte cluster region (LCR) of 19q13.4. The proteins encoded by these loci function as receptors for specific classes of antigens on the surface of various types of immune cells. As with the ZNF and olfactory receptor (OR) genes, the LAIR, LILR and KIR gene families differ extensively in their relative numbers and types between different vertebrate lineages. The variety of different immunoglobulin-like receptor proteins may define some of the major differences in strategies adopted by particular lineages to combat infectious agents and antigens encountered in different environments[31]. The KIR gene family arose recently in the primate lineage, and consistent with this, the repertoire of this gene family varies both in gene number and type even between individual humans. Specific KIR haplotypes have been shown to determine differential susceptibility to immune-related diseases, and are associated with differential rates of progression to AIDS in HIV-infected individuals[32]. The KIR haplotype represented in the public human genome sequence corresponds to the most commonly occurring haplotype in the Caucasian population, called A-1D (ref. 32). This carries nine of the seventeen previously described KIR family members, including a known deletion variant of the *KIRDS4* locus, *KIR2DS4*.

Several other large and evolutionarily diverse gene clusters exist on the chromosome, all with diverse evolutionary histories and involvement in various medical conditions. In addition to the LRC family genes, of particular medical importance are the rapidly evolving cytochrome P450 subfamily II genes (CYP2)[33] involved in the metabolism of steroid hormones, carcinogens and other substances, and the kallikrein (KLK)[34] serine protease family, associated with tumour progression. The positions, size and functions of these gene families are summarized in Table 2.

Comparative biology

To define further the chromosomal landscape and annotate putative functional sequence elements, we performed a comparative analysis of finished human chromosome 19 versus the draft mouse and *Takifugu* (fish) genomes. Using DNA alignments we constructed a homology map refining the map of syntenic homology between chromosome 19 and the mouse genome. Alignments of chromosome 19 with orthologous mouse sequence reveals regions of two general types: vast regions of synteny with 1:1 gene orthology, and significant segments where, owing to lineage-specific duplication and rearrangement events, 1:1 orthology is rare. Intervals in which unambiguous 1:1 orthologous relationships can be defined span over 45 Mb or 82% of chromosome 19 (excluding the centromere). By scanning these regions for contiguous collinear nucleotide similarity, 38 blocks larger than 100 kb were identified, the longest segment being 9.7 Mb (Fig. 2; see also Supplementary S6). All of these blocks lie within the boundaries of larger, previously characterized segments of human–mouse syntenic homology[27]. Homology scanning and manual curation permitted a more precise definition of syntenic borders and identified previously undetected internal rearrangements within the larger segments. Within the p arm, extensive rearrangements have occurred since the time of primate–rodent divergence, with 16 major rearrangements located in nine distinct syntenic segments derived from four mouse chromosomes (mouse chromosomes 8, 9, 10 and 17). In contrast,

Table 2 **Chromosome 19 gene family clusters**

Family/subfamily	Number of functional genes	Chromosome 19 location (Mb)	Function
Krüppel-type zinc finger genes	266		Transcriptional activation and repression
KRAB-ZNF subfamily	202		
Other subfamilies	64		
Cluster 1	5	2.7–2.9	
Cluster 2	10	8.7–9.7	
Cluster 3	27	11.5–12.6	
Cluster 4	34	19.5–24.1	
Cluster 5	5	34.8–35.1	
Cluster 6	30	36.3–38.4	
Cluster 7	21	44.0–44.7	
Cluster 8	40	52.0–53.7	
Cluster 9*	7	55.6–55.8	
Cluster 10	13	56.2–56.7	
Cluster 11	41	57.0–58.7	
Olfactory receptors			Odorant detection
Cluster 1	5	8.8–9.0	
Cluster 2	11	14.7–14.9	
OR10H family	5	15.6–15.8	
Cytochrome P450			
CYP4F family	9	15.5–15.9	Leukotrine receptors/inflammation response
CYP2 family	8	40.9–41.4	Metabolism of drugs, toxins and steroid hormones
Pregnancy-specific glycoproteins/carcinoembryonic antigens	16	42.6–43.5	Maintenance of pregnancy
Sialic acid glycoprotein lectins	11	51.3–51.8	Sialic-acid-recognizing cell surface lectins
Kallikrein proteins	15	50.1–51.2	Tissue-specific serine proteases
Immunoglobulin-like receptors LAIR, LILR, ILT and KIR	23	54.3–55	HLA antigen receptors on T- and B-cell surfaces

*All ZNF clusters except cluster 9 contain at least some KRAB-ZNF family members and many clusters contain ZNF and mixed types; cluster 9 is comprised entirely of SCAN-ZNF genes.

the q arm aligns almost entirely with mouse chromosome 7 and rat chromosome 1, the single exception being a 3-Mb region related to mouse chromosome 17. However, within this relatively stable synteny group are found more than 21 intrachromosomal breakpoints involving transposition or inversion events that have occurred since the primate–rodent divergence (Fig. 2).

The second type of chromosome 19 regions corresponds to tandem gene families, in which extensive lineage-specific gene duplication and loss have occurred (Table 2). Reflecting the recent duplication and high divergence rates within these clusters, only approximately 51% of gene-family exons display an apparent match in the mouse or the rat genomes compared with about 84.7% when the human genome is interrogated for matches[35]. In some of these rapidly evolving regions, especially the ZNF and OR clusters from which both dispersed and tandem gene copies have been generated independently in each lineage, human–mouse homology relationships are less obvious. This picture is complicated by the fact that many chromosome 19 gene families are localized at the boundaries of human–rodent syntenic rearrangements; in several cases, family members have been split by these events onto separate chromosomes in mouse[27]. However, members of most conserved chromosome 19 gene clusters are clearly anchored within syntenically conserved positions by identities of flanking unique genes and evolutionary analyses[27–29].

Comparisons of human chromosome 19 and mouse DNA sequences confirm extensive coding and non-coding conservation with 4,586 discrete fragments showing conservation (>70% identity with a score of match-mismatch >60) but lacking evidence of encoding protein or being transcribed. The putative conserved non-coding elements are unevenly distributed along the chromosome with several high-density clusters in the proximal portion of the q arm in a large region with low G+C content, low gene density and low mouse/human breakpoint density (Fig. 2). It is unclear whether the majority of these are functional elements under strong evolutionary pressure or whether they are simply conserved owing to a low local mutation rate.

Additional DNA comparisons of human chromosome 19 against *Takifugu* revealed extensive coding and significantly less non-coding conservation. A total of 57% of human exons are covered by *Takifugu* conservation. In contrast, only 66 chromosome 19 sequences were conserved with *Takifugu* (>70% identity and match-mismatch score >50) for which no coding or transcribed evidence could be found. Three sequences showed evidence of being non-coding RNA genes[36] (http://www.genetics.wustl.edu/eddy/software/#qrna) whereas the remaining 63 appear to be untranscribed in nature. Similar to human–mouse non-coding conserved elements, these human–*Takifugu* non-coding conserved elements are also enriched in the low-gene-dense proximal portion of the q arm. This finding of ancient evolutionary constraints on a small fraction of non-coding chromosome 19 DNA suggests a probable role in basic vertebrate biological functions, and recent work has shown that a significant fraction of non-coding elements conserved between human and *Takifugu* can have gene regulatory activity even when located great distances from genes[37].

Conclusions and implications for human disease

From the beginning of the Human Genome Project (HGP), one of the major goals was to facilitate the identification and study of the functions of genes that are involved in genetic diseases, both simple mendelian forms and those with more complex inheritance. Now that the sequence is complete, the process of mapping and identifying a disease gene is no longer limited by experimental molecular biology but essentially only by the number of meioses in accurately diagnosed families. As chromosome 19 is crowded with genes, it is not surprising that it has a large number of well-mapped and cloned genes for single-gene disorders, as well as many loci with evidence for association with complex traits. Currently, there are at least 97 single-gene mendelian traits, most of which correspond to rare genetic diseases, that have been localized to specific loci or regions on the chromosome by meiotic mapping in families (see Supplementary S7 and http://www.ncbi.nlm.nih.gov/Omim/).

The genes and associated mutations have been identified for about 75% of these traits. Some of these genes were identified through classical biochemical and molecular approaches before the formal start of the HGP, including the low-density lipoprotein receptor gene[38], which leads to familial hypercholesterolaemia when mutated[39], and the gene encoding the erythropoietin receptor, which is important for erythroid and myeloid cell differentiation and results in autosomal-dominant erythrocytosis when mutations that increase its expression are present[40]. However, most of the genes contributing to altered phenotypes were found with positional cloning approaches that became increasingly facile as the maps and sequence of the chromosome were produced and refined. Even so, there remain at least 20 mendelian diseases mapped on chromosome 19 for which the genes have not yet been identified. Finally, the neurturin gene lying on chromosome 19, which encodes the ligand of a tyrosine kinase receptor (RET), represents one of the limited examples of a gene that has convincingly been demonstrated to contribute to a multigenic trait. Mutations in both the neurturin and the receptor genes together lead to Hirschsprung's disease, whereas mutations in neurturin alone are insufficient to cause the disorder[41]. Biology has truly entered a new era with the completion of the sequence of the entire human genome. The finished sequence of chromosome 19 will facilitate the identification of additional genes contributing to single-gene disorders as well as complex traits. In addition, the large number of evolutionarily conserved non-coding sequences shared with other vertebrates suggest new candidate regions for searching for the genetic basis of human disorders and quantitative traits where sequence alterations of gene regulatory elements have been suggested as a frequent molecular mechanism[42]. □

Methods

Annotation

All human and mouse expressed sequences in GenBank (February 2003) were aligned to their 'best-in-genome position' against repeat-masked chromosome 19 sequence (version 1, February 2003) using BLAT[43]. Alignments were post-processed to correct alignment errors and enforce common splice signals[44]. The mRNA and EST evidence was analysed to retain information about original noted start, CDS, protein sequence and any indicated poly(A) site/signal. In addition, putative transcripts were produced using the evidence-based gene finder FgenesH++ (ref. 44) and by aligning syntenic mouse cDNAs to the chromosome with GeneWise[45]. cDNA and computational predictions that consistently overlapped ESTs at the same locus were automatically extended. Other analyses including MZFEF, GenomeScan and CpGseek were run in parallel. Cases where a human mRNA was wholly contained within another human mRNA as a distinct transcript were not reported. All results were loaded into a MySQL database. Finally, each predicted model was analysed for domain content with InterPro and aligned using a double affine Smith–Waterman algorithm against human and mouse proteomes, as well as Swissprot and other GenBank proteins. The resulting gene structures were inspected relative to EST evidence and protein homology/domain content using a web-based interface. With a combination of the browser, local tools and the Apollo editing system[46], a distributed group of annotators corrected apparent errors as necessary. For gene families with well-defined domain content, custom gene models were constructed with respect to the known structures of these gene families, using the following hierarchy of criteria: (1) matches to cDNA sequence data; (2) protein homology; and (3) gene prediction data. If there was evidence for additional genes or transcripts that were not previously represented by the auto-promoted models, they were also indicated. We report only those loci with model transcripts confirmed by two or more methods. After the completion of the manual curation, the resultant gene catalogue was aligned to the NCBI July 2003 build 34 of chromosome 19. The browser interface can be found at http://genome.jgi-psf.org/Chr19/Chr19.home.html.

Pseudogenes were identified by means of a two-pronged approach. First, the gene structure (exon number) of all FgenesH models was compared with the gene structure of the models' closest human homologues. Those FgenesH models containing fewer exons than their closest human homologue were deemed potential processed or partial pseudogenes. Each of these sequences was manually analysed for pseudogene status. The chromosome 19 sequence was then masked of all repeats (http://ftp.genome.washington.edu/RM/RepeatMasker.html) and of all coding exons from the 1,461 protein-coding genes. The most current version of the human IPI protein data set was then aligned against the masked sequence using two alignment programs: tblastn and

prot_map. Protein alignments were filtered to select the longest and best-scoring alignment for a given genomic locus with the top alignment at each locus being manually analysed to determine pseudogene status. Single-exon pseudogenes (including those interrupted by repetitive elements or short inserts, that is <100 bp) that shared similarity to a protein sequence from the most current version of the Ensembl database, such that the pseudogene sequence covers at least 70% of the coding sequence, where classified as processed pseudogenes; any single-exon pseudogenes aligning to less than 70% of the length of the closest Ensembl protein sequence were classified as potential pseudogene fragments.

Comparative analysis

The mouse and rat genomes were taken from http://genome.ucsc.edu; we used the repeat masked version of mouse February 2003 freeze (mm3) and the rat June 2003 freeze (rn3). The *Takifugu* genome assembly is available from the JGI website (http://www.jgi.doe.gov). The cross-species alignments were performed using BLASTZ[47]. The rodent genomes were chopped into 1-Mb pieces with a 10-kb overlap. Each piece was then aligned to the entire chromosome 19 sequence. The raw DNA BLASTZ alignments were then used to create the segmental maps, include duplications, and extract the conserved regions using the PARAGON software (O.C., unpublished data). The mRNA and EST alignments (http://genome.ucsc.edu) were used to filter out the non-coding set from coding evidence (human/mouse EST, spliced and non-spliced, non-human/mouse mRNA and EST aligned to each genome using BLAT). Some of the coverage statistics data were done with software developed by J. Kent (http://www.cse.ucsc.edu/~kent). Parts of Fig. 2 were made using a local modified installation of the UCSC genome browser.

Segmental duplication analysis

We used a BLAST-based detection scheme[48] to identify all pairwise similarities representing duplicated regions ($\geq$1 kb and $\geq$90% identity) within chromosome 19 (version 1, February 2003) and compared them to all other chromosomes in the NCBI genome assembly (build 31). We compared these BLAST-based pairwise similarities to the whole-genome shotgun sequence detection database of segmental duplications, which was ascertained via an assembly-free method[48]. 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. β-Satellite repeats were detected using RepeatMasker (15 May 2002 version) on sensitive settings (A. Smit and P. Green, unpublished data).

Received 1 December 2003; accepted 10 February 2004; doi:10.1038/nature02399.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
3. Setlow, R. B. & Setlow, J. K. Evidence that ultraviolet light induced thymine dimers in DNA cause biological damage. *Proc. Natl Acad. Sci. USA* **48,** 1250–1257 (1962).
4. Setlow, R. B. & Carrier, W. L. The disappearance of the thymine dimers from DNA: An error correcting mechanism. *Proc. Natl Acad. Sci. USA* **51,** 226–231 (1964).
5. Cleaver, J. E. Defective repair replication of DNA in xeroderma pigmentosum. *Nature* **218,** 652–656 (1968).
6. Mohrenweiser, H. W. *et al.* Refined mapping of the three DNA repair genes, ERCC1, ERCC2, and XRCC1, on human chromosome 19. *Cytogenet. Cell Genet.* **52,** 11–14 (1989).
7. Patrinos, A. & Drell, D. W. The Human Genome Project: view from the Department of Energy. *J. Am. Med. Womens Assoc.* **52,** 8–10 (1997).
8. Ashworth, L. K. *et al.* An integrated metric physical map of human chromosome 19. *Nature Genet.* **11,** 422–427 (1995).
9. de Jong, P. *et al.* Human chromosome-specific partial digest libraries in lambda and cosmid vectors. *Cytogenet. Cell Genet.* **51,** 985 (1989).
10. Brandiff, B. F. *et al.* Human chromosome 19p: A fluorescence *in situ* hybridization map with genomic estimates for 79 intervals spanning 20 Mbp. *Genomics* **23,** 582–591 (1994).
11. Gordon, L. A. *et al.* A 30-Mb metric fluorescence *in situ* hybridization map of human chromosome 19q. *Genomics* **30,** 187–192 (1995).
12. Felsenfeld, A., Peterson, J., Schloss, J. & Guyer, M. Assessing the quality of the DNA sequence from the Human Genome Project. *Genome Res.* **9,** 1–4 (1999).
13. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380,** 152–154 (1996).
14. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31,** 241–247 (2002).
15. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63,** 861–869 (1998).
16. Yu, A. *et al.* Comparison of human genetic and sequence-based physical maps. *Nature* **409,** 951–953 (2001).
17. Mungall, A. J. *et al.* The DNA sequence and analysis of human chromosome 6. *Nature* **425,** 805–811 (2003).
18. Hillier, L. W. *et al.* The DNA sequence of human chromosome 7. *Nature* **424,** 157–164 (2003).
19. Hellig, R. *et al.* The DNA sequence and analysis of human chromosome 14. *Nature* **421,** 601–607 (2003).
20. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414,** 865–871 (2001).
21. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405,** 311–319 (2000).
22. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402,** 489–495 (1999).
23. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29,** 137–140 (2001).
24. O'Brien, T. J., Beard, J. B.,, Underwood, L. J. & Shigemasa, K. The CA 125 gene: a newly discovered extension of the glycosylated N-terminal domain doubles the size of this extracellular superstructure. *Tumour Biol.* **23,** 154–169 (2002).
25. Kan, Z., States, D. & Gish, W. Selecting for functional alternative splices in ESTs. *Genome Res.* **12,** 1837–1845 (2002).
26. Menashe, I., Man, O., Lancet, D. & Gilad, Y. Different noses for different people. *Nature Genet.* **34,** 143–144 (2003).
27. Dehal, P. *et al.* Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* **293,** 104–111 (2001).
28. Newman, T. & Trask, B. J. Complex evolution of 7E olfactory receptor genes in segmental duplications. *Genome Res.* **13,** 781–793 (2003).
29. Shannon, M., Hamilton, A. T., Gordon, L., Branscomb, E. & Stubbs, L. Differential expansion of zinc-finger transcription factor loci in homologous human and mouse gene clusters. *Genome Res.* **13,** 1097–1110 (2003).
30. Eichler, E. E. *et al.* Complex β-satellite repeat structures and the expansion of the zinc finger gene cluster in 19p12. *Genome Res.* **8,** 791–808 (1998).
31. Trowsdale, J. *et al.* The genomic context of natural killer receptor extended gene families. *Immunol. Rev.* **181,** 20–38 (2001).
32. Hsu, K. C. *et al.* The killer cell immunoglobulin-like receptor (KIR) genomic region: gene-order, haplotypes and allelic polymorphism. *Immunol. Rev.* **190,** 40–52 (2002).
33. Hoffmann, S. M., Nelson, D. R. & Keeney, D. S. Organization, structure and evolution of the CYP2 gene cluster on human chromosome 19. *Pharmacogenetics* **11,** 687–698 (2001).
34. Yousef, G. M. & Diamandis, E. P. The new human tissue kallikrein gene family: structure, function, and association to disease. *Endocr. Rev.* **22,** 184–204 (2001).
35. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
36. Rivas, E. & Eddy, S. R. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2,** 8 (2001).
37. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302,** 413 (2003).
38. Yamamoto, T. *et al.* The human LDL receptor: a cysteine-rich protein with multiple Alu sequences in its mRNA. *Cell* **39,** 27–38 (1984).
39. Hobbs, H. H., Russell, D. W., Brown, M. S. & Goldstein, J. L. The LDL receptor locus in familial hypercholesterolemia: mutational analysis of a membrane protein. *Annu. Rev. Genet.* **24,** 133–170 (1990).
40. Juvonen, E., Ikkala, E., Fyhrquist, F. & Ruutu, T. Autosomal dominant erythrocytosis caused by increased sensitivity to erythropoietin. *Blood* **78,** 3066–3069 (1991).
41. Doray, B. *et al.* Mutation of the RET ligand, neurturin, supports multigenic inheritance in Hirschsprung disease. *Hum. Mol. Genet.* **7,** 1449–1452 (1998).
42. Glazier, A. M., Nadeau, J. H. & Aitman, T. J. Finding genes that underlie complex traits. *Science* **298,** 2345–2349 (2002).
43. Kent, W. J. The BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
44. Solovyev, V. V. in *Bioinformatics From Genomes To Drugs. V.1. Basic Technologies* (ed. Lengauer, T.) 59–111 (Wiley-VCH, Weinheim, 2002).
45. Birney, E. & Durbin, R. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10,** 547–548 (2002).
46. Lewis, S. E. *et al.* Apollo: a sequence annotation editor. *Genome Biol.* **3,** research0082.1–0082.14 (2002).
47. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13,** 103–107 (2003).
48. 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).

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

Acknowledgements We are grateful to our colleagues, S. Scherer, L. French and The Whitehead Institute/MIT Center for Genome Research for assistance with the screening of map gaps. We also acknowledge the HUGO Gene Nomenclature Committee for determining chromosome 19 gene symbols, H. Riethman for guidance regarding the telomeric regions of the chromosome, and U. Francke for helping to compile Supplementary Table S7 as well as for useful discussions regarding disease gene locations. 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.G. (jane@shgc.stanford.edu) and E.M.R (emrubin@lbl.gov). The entire sequence for the chromosome is deposited in GenBank under accession numbers NT_00255, NT_ 077812, NT_011295 and NT_011109.