Research

A draft annotation and overview of the human genome

Fred A Wright*, William J Lemon*, Wei D Zhao*, Russell Sears*, Degen Zhuo*, Jian-Ping Wang*, Hee-Yung Yang†, Troy Baer‡, Don Stredney‡§, Joe Spitzner†, Al Stutz‡§, Ralf Krahe* and Bo Yuan*

Addresses: *Division of Human Cancer Genetics, The Ohio State University, 420 W. 12th Avenue, Columbus, OH 43210, USA. †LabBook.com, Busch Boulevard, Columbus, OH 43229, USA. ‡Ohio Supercomputer Center (OSC), Kinnear Road, Columbus, OH 43212, USA. §Department of Computer and Information Science, The Ohio State University, Neil Avenue, Columbus, OH 43210, USA.

Correspondence: Bo Yuan. E-mail: yuan.33@osu.edu

Published: 4 July 2001

Genome ***Biology*** 2001, **2(7)**:research0025.1–0025.18

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/7/research/0025

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 12 February 2001
Revised: 4 April 2001
Accepted: 1 June 2001

A previous version of this manuscript was made available before peer review at http://genomebiology.com/2001/2/3/preprint/0001/ (*Genome* ***Biology*** 2001, **2(3)**:preprint0001.1-0001.40)

Abstract

Background: The recent draft assembly of the human genome provides a unified basis for describing genomic structure and function. The draft is sufficiently accurate to provide useful annotation, enabling direct observations of previously inferred biological phenomena.

Results: We report here a functionally annotated human gene index placed directly on the genome. The index is based on the integration of public transcript, protein, and mapping information, supplemented with computational prediction. We describe numerous global features of the genome and examine the relationship of various genetic maps with the assembly. In addition, initial sequence analysis reveals highly ordered chromosomal landscapes associated with paralogous gene clusters and distinct functional compartments. Finally, these annotation data were synthesized to produce observations of gene density and number that accord well with historical estimates. Such a global approach had previously been described only for chromosomes 21 and 22, which together account for 2.2% of the genome.

Conclusions: We estimate that the genome contains 65,000-75,000 transcriptional units, with exon sequences comprising 4%. The creation of a comprehensive gene index requires the synthesis of all available computational and experimental evidence.

Background

The sequence of the human nuclear genome has been completed in draft form by an international public consortium consisting of 16 sequencing centers and associated computational facilities [1]. A private commercial version of the genome has also been sequenced and assembled using a whole genome shotgun approach [2]. Many lower organisms have been sequenced to date [3], but the 3.2 billion base pair (bp) human genome is approximately 25 times as large as the largest currently finished genomes - *Drosophila melanogaster* at 120 megabases (Mb) [4] and *Arabidopsis thaliana* at 115 Mb [5].

As of late 2000, the public human sequence was primarily based on approximately 24,000 accessioned bacterial artificial chromosome (BAC) clones covering 97% of the euchromatic portion of the genome [6]. The sequence of these clones is approximately 93% complete to at least 4-fold coverage [7]. Thirty percent of the genome is in finished form, including the entire sequence of chromosomes 21 and 22 [7].

These clones represent the most complete sequence information available, with overlapping clones positioned on a framework map using restriction fingerprinting [8]. However, reduction to a single consensus sequence permits placement of genes and other chromosomal structures in their proper positional context. Recently, the consortium has distributed a working draft assembly of the entire genome that removes redundancies, orients sequence fragments and clearly indicates gaps arising from sequencing and assembly. The total assembled length is 3.08 billion bp – about 4% smaller than estimates of genome size based on flow cytometry [9], presumably due to the exclusion of constitutive heterochromatic regions and centromeres. Major gaps (50-200 kilobases (kb)) comprise 16% of the assembly, whereas minor gaps (100 or fewer bp) and low-quality calls comprise 0.5%.

Large-scale sequencing will continue until at least 2003. The current coverage is, however, sufficient for the Human Genome Project to enter a new phase, in which the entire sequence can be analyzed to identify genes, regulatory regions and other genomic elements and structures. Linkage and genetic association studies can be immediately followed by investigation of candidate regions. The assembly provides simplified descriptions of the genome, as disparate data sources such as GenBank and numerous expressed sequence tag (EST) and protein databases are unified. Similarly, formerly independent maps, based on cytogenetic banding patterns, meiotic crossovers and radiation hybrids, may be placed within the single consensus sequence.

Results and discussion

Combine and conquer

Public attention surrounding completion of the draft human sequence has fostered the impression that we are entering a 'post-genomic' era, and that description of genes and their functions is straightforward. However, the challenges in genome annotation remain daunting [10], and the research community can anticipate years of additional work and manual curation to produce a true gene map of high quality.

Functional annotation of the genome is primarily hampered by the lack of a unified transcript index. Current transcript information still largely consists of anonymous and highly redundant ESTs. The situation is further complicated by extensive splicing variation and elusive gene expression. To address these problems, the Ensembl consortium relies initially on computational prediction, followed by confirmation with EST/protein alignments [11]. However, pure computational approaches can give differing results [12], and may miss 20% or more of transcript-supported exons [13]. Other gene identification approaches rely on selecting and grouping ESTs into putative gene indices [14,15], or consensus sequences [16,17]. These approaches emphasize internal consistency and result in limited EST populations that only partially overlap. The genome sequence serves as a powerful arbiter of the quality of EST evidence, and will enable consolidation of additional exons into transcriptional units. Thus, we adopt a more inclusive approach.

Our approach is to combine the major public cDNA, EST and protein databases, resolve redundancies, and place the resulting exonic sequences uniquely on the genome using the program Blast. We refer to these genomic segments (technically high-scoring segment pairs [18]) as 'exons', although the alignment evidence awaits future biological confirmation. Splicing evidence was carefully maintained within genomic clones, and across clones using the fingerprint map. For a given transcript, only the best match to genomic sequence (using splicing evidence, length and high sequence identity) was preserved, resulting in a unique location for each exonic unit within each database. We have successfully applied this approach to integrate UniGene consensus sequences into the human genome draft [19].

To compile a truly unique exonic index, redundancies must also be resolved across transcript databases. We grouped the databases into ranked categories and ordered them within categories. Transcripts with known boundary information (using the untranslated region database (UTR-DB)) [20] or full-length cDNAs in the human transcript database (HTDB) [21] were given precedence over other records. Consensus transcripts were given precedence over individual ESTs because they provide greatly improved specificity, splicing evidence and transcript integrity. We assembled UniGene-based human [19], mouse and rat consensus transcripts. Collectively, the databases represent almost all public information on known genes, transcripts and relevant homologous sequences. When aligned segments overlapped, only the segments from the highest-ranked categories were retained. After resolution of overlapping exons, a new exonic index of contiguous spliced components was formed. Each member of this new index inherited the rank of its highest-ranked exon, in order to facilitate subsequent identification of transcriptional units. Our approach also ensures that known genes are represented only once in the final gene map.

Table 1 describes the identification of exonic sequence via the public databases. Not all human transcript records could be placed on the genome, reflecting sequence gaps and the draft quality of the genomic clones. The percentage placement of known genes (80-89%) suggests that unsequenced regions will contribute substantial numbers of additional genes. The varying placement percentages among transcript databases reflect varying sequence quality and differing transcript lengths. Unique exons are those that have no overlap with those already placed by a higher-ranked database. Rodent transcripts provided a modest number of additional exons. Finally, additional placements were based on strong protein homology with supportive computer exon prediction. The percent placement was relatively low because all proteins from different species were considered, with specificity

Table 1

Identification of exons on the genome

Category	Database	Total records	Percent placed (%)	Total unique exons	Exons in complete ORFs	Exons in partial ORFs	Exon length (bp)	ORF length (bp)	Putative genes (non-splicing singletons)	Protein homology (Pfam hits)	CpG islands
Known genes	UTR-DB	40,258	80	19,195	5,075	1,895	6,925,762	1,990,818	10,007 (426)	5,701 (3,813)	3,866
	HTDB	15,305	89	48,477	12,077	7,706	11,893,081	4,043,544	4,816 (148)	2,938 (1,943)	1,960
Consensus transcripts	HINT	87,125	77	103,817	47,055	15,061	23,381,024	10,144,988	20,357 (959)	9,121 (6,453)	7,557
	EG	62,064	80	13,085	5,389	1,904	4,562,954	1,873,723	4,800 (154)	2,177 (1,679)	2,462
	THC	84,837	81	38,806	15,463	6,671	12,406,081	5,078,661	8,604 (322)	2,907 (2,026)	3,983
Transcripts	GenBank CDS	110,222	81	41,917	31,626	1,452	5,303,064	4,299,272	2,634 (227)	1,858 (1,607)	1,178
	dbEST Human	2,154,995	73	273,881	147,819	17,694	32,288,385	14,975,758	20,073 (7,136)	5,377 (3,745)	11,807
Rodent transcripts	MINT	92,531	30	8,284	5,433	120	866,046	780,566	777	123 (56)	486
	RINT	37,367	46	5,600	3,588	75	592,788	546,932	458	65 (32)	255
	EMBL	43,488	28	5,819	4,108	59	724,630	655,993	202	68 (72)	135
Protein homology	SWISS-PROT	86,593	38	27,526	12,072	1,163	9,858,797	7,784,205	1,648	1,648 (1,244)	158
	TrEMBL	351,834	13	22,670	8,134	1,677	4,385,497	2,886,034	1,185	1,185 (654)	92
	PIR	182,106	16	4,106	1,175	383	1,355,644	764,339	321	321 (132)	20
Total				613,183	299,014	55,860	114,543,753	55,824,833	75,982 (9,372)	33,489 (23,008)	33,959

Exons were identified after vector screening using transcript, rodent, and protein databases. The definition of a record varies according to the database, while 'exons' refer to high-scoring segment pairs in BlastN comparisons ($E < 10^{-15}$ and sequence identity > 90%) to the genome. Unique exons and all subsequent columns refer to placements that were possible after considering the preceding databases. Placement of rodent transcripts required evidence of splicing and sequence identity > 80%. ORFs were identified using getorf [84] using a minimum size of 30 bp to report. Protein homology required BlastX $E < 10^{-15}$. Pfam hits required score > 20 using hmmpfam [92]. Gene prediction programs are described in Table 2. CpG islands were identified using cpgreport [84] using standard criteria [45].

assured by using appropriately stringent criteria and exons confirmed by at least one gene prediction program.

When all of the databases are considered, 613,183 unique exons were placed, including 299,014 in complete open reading frames (ORFs) and 55,860 in partial ORFs. The total putative exonic lengths add to 106 Mb, or about 4% of the sequenced genome. About 50% of our described exonic sequences are in ORFs (Table 1). It is generally thought that the majority of exonic sequence is coding, suggesting that additional coding sequences remain to be discovered. This possible bias towards untranslated regions is to be expected, as current transcript information is largely derived from the 3′ or 5′ termini of cDNA libraries. At least 30-40% of the known genes or transcript indices contain one or more internal transcripts, suggesting alternative splicing, internal genes or occasional artifacts (misassembly or genomic contamination). The prevalence of alternative splicing remains unknown, but may occur frequently [22]. 'Sandwiched' transcripts were merged with their flanking indices, unless both the internal and the flanking sequences were from distinct known genes (< 150 apparent internal genes). In addition, we observed a small number of apparently overlapping exons (about 530 on opposite strands) [23].

We assessed three *ab initio* gene prediction methods by comparing their predicted exons to the ones identified by transcripts and proteins. Genscan, Grail and Fgene were used across the genomic clones to identify potential exons (Table 2). Approximately 70% of the 299,014 exons in ORFs with either transcript or protein support were identified by at least one of the programs, but a very large number (847,283) of unconfirmed exons were also predicted. The large apparent false negative and positive rates imply that pure computational gene prediction is not yet a practical alternative to experimental evidence.

Transcriptional units

Our consolidated exonic index is of inherent biological interest, but it is desirable to further identify transcriptional boundaries to create a putative gene index. We used an approach designed to minimize fragmentation of exons and provide conservative gene counts (see Materials and methods). The following criteria were used to identify gene boundaries: known 5′ or 3′ UTR sequences in UTR-DB; full-length cDNAs in HTDB; exons in partial ORFs as possible boundaries of coding regions; exons without continuous ORFs as additional UTR sequences; CpG islands; and gene boundaries predicted by Genscan. Multiple in-frame exons in a continuous ORF were always considered part of a single gene, an approach that tends to consolidate exons rather than create spurious additional genes. Additional consolidation resulted from extension of boundaries for multiple exons not residing in ORFs until the occurrence of genomic

Table 2

Genome-wide assessment of *ab initio* gene prediction methods

Genscan	Grail	Fgenes	Transcript-confirmed exons	Protein-supported exons	Unconfirmed exons
•			25,619	2,890	45,025
	•		52,644	14,685	434,409
		•	7,791	796	257,676
•	•		17,841	3,761	28,556
•		•	13,915	1,711	11,628
	•	•	3,990	450	49,420
•	•	•	53,566	9,871	20,569
Total exons			175,366	34,164	847,283

Three gene prediction programs, Genscan [93], Fgenes [94] and Grail1.3 [95] were used to screen individual genomic contigs. Exons consistently predicted by more than one program are merged into a unique exon index, which is then compared to transcript- and protein-based exons in complete ORFs. Transcript-confirmed exons, overlapping of predicted exons with transcript-based exons; protein-supported exons, predicted exons have at least strong protein homology ($E < 10^{-15}$); unconfirmed exons, predicted exons have no overlap with transcripts nor protein homology.

landmarks described above. The success of this approach depends largely on the extension and consolidation of overlapping transcripts, and the integrity of ORFs and other genomic landmarks provided by the draft sequences.

Table 1 lists the number of genes added by each database to the cumulative sum. The total number of known genes in UTR-DB, HTDB and HINT is 16,673. This compares with 11,191 entries with at least partial functional annotation in UniGene (May 2000 build) and 11,863 entries in the HUGO Human Gene Nomenclature database [24]. Approximately 48% of the transcriptional units were based on consensus transcripts and 28% based on individual ESTs. A total of 9,372 transcriptional units were based on singleton transcripts without splicing evidence. Single-exon (intronless) genes occur with appreciable frequency in the human genome [25]. At least one third of the singletons in our gene index contain intact ORFs, and are predominantly histones, G-protein receptors, olfactory receptors, and cytokines or their homologs/paralogs. The remaining two-thirds of the singletons do not have intact ORFs, and possibly represent pseudogenes, genomic contamination or other artifacts. It is thought that most intronless genes originated from retrotranspositions of SINEs and LINEs [26]. Thus, the total number of single-exon genes might be under-represented in this study, because of the repeat masking process necessary prior to our analysis. This also applies to the tRNA, rRNA and other snRNAs in the human genome, which have been similarly masked. A total of 1,437 units were supported only by rodent transcript consensus, predominantly derived from cDNA libraries of early embryogenesis or tissues of the central nervous system. An additional 3,154 units were identified on the basis of protein homology, with exons supported by at least one gene prediction program. Our approach yields an overall estimate of 75,982 transcriptional units, with 66,610 supported by multiple transcripts or individual transcripts with splicing evidence. Therefore, the consolidation and integration of mainly the transcript information into the genomic consensus assures that our putative gene index is largely based on experimental evidence, rather than *ab initio* gene prediction.

It is important to note that pseudogenes are common in the human genome, and are thought to largely originate from gene duplication or retrotransposition [27]. The extent to which pseudogenes remain transcriptionally active is still largely unknown, however. It is also difficult to identify pseudogenes computationally. Although nonfunctional pseudogenes can have characteristic structural features, some functional genes can also exhibit such features [27].

We observed that 45% of the gene units were associated with CpG islands (defined as 10 kb upstream or within the gene). For the 6,500 known genes with known 5′ boundaries, the value was 40%. The average genomic size of each of our transcriptional units from the first to last identified exon (including only transcript or protein-based exons) is approximately 12 kb. The overall average gene length is likely to be significantly longer, but full-length cDNA information is not yet available for most genes.

Comparison of gene counts

Our count of 66,000-75,000 transcriptional units on the genome is consistent with gene count estimates [28,29] that had held sway until recent widely varying estimates [17,30,31].

Ewing and Green [17] examined 680 assumed genes on chromosome 22 and found matches to 2% of a selected set of assembled EST contigs. The sampling approach assumes that the 680 genes represent 2% of all genes, resulting in an overall count of 34,000. An examination of evolutionarily conserved regions in known genes on chromosome 22 in humans compared to the fish *Tetraodon nigroviridis* [30] results in an estimate of around 30,000 genes, assuming a uniform rate of conserved regions per true gene. These approaches resulted in similar estimates when applied to larger sets of mRNAs or known genes, and are similar to the current 33,000 genes reported by Ensembl as having Genscan computational support and EST confirmation. All of these estimates are carefully constructed and remarkably concordant, and we propose possible explanations for the difference from our results. The differences do not result entirely from the reliance on transcriptional evidence, as has been proposed [32].

Our estimate of 854 genes on chromosome 22 is 25% greater than that of Ewing and Green [17], but represents only 1.4% (rather than 2%) of our gene total. It was noted [17] that high

gene expression on chromosome 22 could result in low gene count estimates by biasing the reference sample. In addition, known genes may be more highly expressed than unknown genes, which presumably aided their initial identification and characterization. Our evaluation of EST evidence supports the existence of both forms of bias. We have found that 5% of Ewing and Green's original set of EST contigs (selected with less stringent criteria than those used to estimate gene counts) map to chromosome 22. An examination of UniGene transcripts (May 2000) reveals that the known genes contain a median of 41 entries, whereas anonymous transcripts contain a median of just two entries. This is not entirely explained by the greater length of the known gene-like transcripts (having been correctly assembled as a single unit). In dividing the number of ESTs in the consensus by its length, we obtain a median of 0.017 entries per base pair for known genes and 0.005 entries per base pair for anonymous transcripts. On chromosome 22, the median number of ESTs per anonymous transcript is three, which is significantly higher than that among other transcripts on the genome (geometric mean 3.76 versus 3.11 for other chromosomes, $p < 0.0001$, Wilcoxon rank-sum test). The estimate based on conserved regions [30] is calibrated using known genes. This approach also introduces bias, as such genes appear more likely to belong to the evolutionary core proteome. Known genes comprise 22% of all of our transcriptional units, but comprise 71% of our units which are conserved with rodents, *Drosophila* and *Caenorhabditis elegans*. A recent high gene estimate based on transcript evidence [31], again using chromosome 22, appears to result from less stringent alignment criteria, resulting in many putative genes.

As genomic annotation proceeds, the number of protein-coding genes will become clearer. Our approach seems to rule out artifactual or genomic contamination as the predominant explanation for transcriptional units with unknown function or protein homology. Ensembl has recently listed a count of 170,160 'confirmed' exons, whereas we report 299,014 in complete ORFs and many more in untranslated regions, suggesting that our approach identifies considerable additional transcription. We point out that only 58% of known genes exhibit protein homology (Table 1) and, for example, a large proportion of transcriptional units have not been functionally classified in *Drosophila* [4]. We therefore propose that most of the unclassified transcriptional units are in fact coding - the lack of protein homology may reflect difficulty in studying these proteins, or rapid gene evolution, and some portion is likely to function at the RNA level [33].

Gene map

The placement of transcriptional units is not without error, as most genomic clones are unfinished and the restriction fingerprint map can be subject to misassembly. To resolve placement errors, we used a relational database to integrate information from several independent maps, including Genemap'99, assembled genomic contigs, and fingerprint, radiation hybrid and cytogenetic maps (see Materials and methods). Placement required a minimum of three concordant criteria. Together, a total of 75,982 transcriptional units were placed on the genome, providing an initial glimpse of a complete gene map. The map and associated functional annotation are available as additional data files online.

Functional annotation

SWISS-PROT, TrEMBL, PIR (Protein Information Resource) and Pfam (Protein Families database) were used to annotate our unified gene index, because functional keywords in these databases are standardized [34] (Table 3). We used the classification schema developed by the International Gene Ontology Consortium to assign each keyword to an appropriate ontological description ([35] and see additional data files for keyword assignments). When more than one unrelated protein was identified within one gene unit, clear functional roles and biological processes were given priority over other keyword designations. Similarly, protein-based annotation was performed for HINT consensus transcripts. The transcriptional units resulted in a greater number of annotations (around 23,000) than HINT transcripts (around 11,000) because of the increased length of exonic sequences from other transcript databases and the included genomic sequence. It is also important to note at least 12,000 of the gene units had more than one conserved protein domain as evidenced by Pfam hits. Additional functional repertoire and biological complexity might be derived from shuffling, and other recombinant events of individual exons during genome evolution.

The annotation also allows us to assess the protein composition of human versus other species. A BlastX result of $E < 10^{-20}$ was required in cross-species DNA-protein alignments to be considered homologous. A total of 20,892 human transcriptional units (30% of all units) are homologous with at least one other species; 5,792 (10%) were conserved across mammals (mouse or rat), *Drosophila*, and *C. elegans*. A total of 1,759 (3%) were conserved across all of these species and yeast. These values are very consistent with a recent comparative genomic survey [36].

Global tissue expression profiles

During the assembly of UniGene [19], we retained the library source for each EST, via links provided by UniGene to the IMAGE consortium [37]. Most of the 2,500 libraries comprising UniGene ESTs were derived from single tissues or embryonic stages, and we further standardized the library source annotation into 102 categories. Keywords and derived categories are available as additional data files online. The most highly represented categories were various types of tumors (15.0% of all ESTs), fetal tissue (10.7%), embryo (6.2%), infant (5.1%), and testis (4.3%). We reasoned that some genes might exhibit highly tissue-specific expression, such that most of the ESTs comprising a transcript would be derived from the tissue. The identified genes

Table 3

Ontological classification of 22,339 human gene products

Biological function	Number of transcripts	Biological process	Number of transcripts
Transcription factor	958 (306)	Carbohydrate metabolism	281 (84)
Translation factor	62 (27)	Nucleotide and nucleic acid metabolism	173 (51)
RNA binding	142 (41)	DNA replication	240 (126)
Ribosomal protein	232 (130)	Transcription	1,059 (651)
Cell cycle regulator	42 (16)	RNA processing	204 (59)
Structural protein	145 (48)	Amino acid and derivative metabolism	87 (29)
Cytoskeleton structural protein	329 (181)	Protein biosynthesis	264 (162)
Extracellular matrix	361 (87)	Protein modification	235 (88)
Actin binding	66 (25)	Protein targeting	26 (5)
Motor protein	245 (77)	Protein degradation	136 (45)
Chaperone	87 (27)	Proteolysis and peptidolysis	96 (36)
Enzyme	2,664 (1,404)	Lipid metabolism	424 (187)
Protein kinase	895 (484)	Monocarbon compound metabolism	9 (3)
Protein kinase inhibitor	19 (12)	Coenzyme and prosthetic group metabolism	92 (29)
Protein phosphatase	43 (7)	Steroid compound metabolism	40 (10)
Protein phosphatase inhibitor	17 (3)	Prostaglandin metabolism	12 (3)
Protease	441 (255)	Transport	549 (288)
Protease inhibitor	92 (37)	Electron transport	491 (273)
Enzyme activator	18 (3)	Ion transport	302 (90)
Enzyme inhibitor	14 (4)	Small molecular transport	19 (9)
Alkyl transfer	17 (3)	Neurotransmitter transport	9 (3)
Amide transfer	15 (3)	Ion homeostasis	201 (57)
Carbonyl transfer	191 (38)	Organelle organization and biogenesis	408 (254)
Hydroxyl transfer	13 (6)	Nuclear organization and biogenesis	1,380 (647)
Phosphoryl transfer	823 (281)	Cytoplasm organization and biogenesis	42 (20)
Oxireduction	148 (76)	Meiosis	15 (2)
Transmembrane protein	184 (48)	Mitosis	25 (6)
Receptor	921 (478)	Cell cycle	271 (100)
G protein-linked receptor	164 (106)	DNA packaging	15 (6)
Defense/immunity protein	353 (164)	DNA repair	132 (41)
Ligand binding or carrier	691 (331)	DNA recombination	31 (3)
Ion channel	245 (141)	Methylation	185 (53)
Oncogene	128 (42)	Signal transduction	1,231 (383)
Tumor suppressor	8 (6)	Growth regulation	15 (4)
Growth factor	95 (40)	Differentiation	24 (6)
Hormone	42 (14)	Apoptosis	160 (49)
Cell communication	247 (84)	Angiogenesis	11 (4)
Cell adhesion	433 (252)	Defense/immunity	112 (49)
		Detoxification	33 (15)
		Stress response	90 (41)
		Developmental process	278 (99)
		Neurogenesis and regeneration	147 (43)
		Physiological process	159 (43)
		Sensory perception	292 (65)
Functionally classified	12,334 (5,204)	Process classified	10,005 (4,225)

Each transcriptional unit and HINT transcript (in parentheses) was assigned to a unique biological function or process.

are potential candidates for diseases of the involved tissues. Similar approaches have been used to identify candidate genes for pathologies of the prostate [38] and retina [39]. We explore here the global nature of tissue/source specificity. The result was 7,459 HINT transcripts with highly significant tissue-specificity (11%). Many of these are known genes, and an examination of the most specific transcripts revealed clear relationships to the associated tissue. For example, a search for retina-specific genes revealed that the ten most significantly associated with retina include five known genes, all related to retina function. Four are implicated in retina pathology: *GNAT1* and *ARR* (night blindness), *RHO* (retinitis pigmentosa), and *GUCA1A* (cone dystrophy). Similar results were observed in numerous other tissues, although not as obviously related to pathology. The results appear especially striking for tissues with substantial EST representation, including brain, lung, liver, kidney, and testis, suggesting that putative tissue involvement can be inferred for many anonymous ESTs. Where possible, the tissue expression profile has been incorporated into the annotation of our gene index. Approximately half (50.5%) of the tissue-specific clusters were from embryonic tissue libraries (while such tissue contributed 6.2% of all UniGene ESTs). This striking result is consistent with the highly regulated and specific nature of embryonic development [40]. The embryo category is followed by brain (9.7% brain-specific versus 3.8% of ESTs) in number of tissue-specific clusters, kidney (5.5% versus 3.5%), and testis (6.1% versus 4.3%). We also examined the locations of the tissue-specific transcripts on the genome, and found no evidence of regional clustering (see description of regional functional clustering in Materials and methods).

A global view of the human genome

In keeping with the long-standing clinical importance of cytogenetics, it is important to align Giemsa-staining G (dark) cytobands versus R (pale) bands (ISCN 1995) to the assembly [41]. Cytoband boundaries on genomic sequence have been depicted with apparent precision [13,42] but in fact are largely unknown. With only a few-fold genomic coverage, the gap sizes in unfinished sequence are difficult to estimate precisely. Thus, it is preferable to align the cytoband positions to the fixed assembly rather than the reverse. Such an 'assembly-corrected' alignment was performed using genes/ESTs that have been mapped cytogenetically and also placed on the assembly. This alignment is approximate, as the resolution of conventional staining techniques and fluorescence *in situ* hybridization (FISH) is limited to 1-3 Mb [43].

Density of genomic features

The resulting corrected ideograms and six major genomic features are plotted across the genome in Figure 1. Unique exons (as determined above), CpG islands, genomic GC content, *Alu* and LINE1 elements, and minisatellites are plotted as densities (proportion of bases belonging to feature) in 1 Mb intervals. The assembly-corrected ideogram clearly differs from the standard ideogram - for example, in our representation 1p is longer than 1q. This may reflect more complete sequencing on 1p, or perhaps differing DNA-packing densities on the two chromosome arms. Many of the chromosomes show a suggestive relationship between cytobands and exon density, consistent with the expectation that R bands are relatively gene rich. A more striking result is the expected positive correlation among exons, CpG islands, GC content and minisatellites, which track each other closely on most chromosomes. Exon density is relatively high on chromosomes known to be gene rich (for example, 17 and 19) [44], and low on chromosomes 4, 13, X, and Y.

A total of 48,000 CpG islands were found on the assembly using standard criteria [45] (see Figure 1 legend), with a median length of 336 bp. As sequencing gaps are filled, this number may increase. Considering the varying definitions of CpG islands (especially the minimum length of CpG-rich region), this number is in close agreement with the estimate of 45,000 obtained by Antequera and Bird [28] using methylation-sensitive restriction enzymes. The CpG island density

Figure 1 *(see figure on following two pages)*
Overview map of features on the entire human genome, based on the working draft assembly (15 June 2000 release) and finished sequences for chromosomes 21 and 22. Ideograms are oriented with the p-arm at the top, and are assembly-corrected to form an approximate cytogenetic alignment with the features of the draft assembly depicted to the right of each ideogram. Sequencing gaps at the centromeres and contiguous heterochromatic regions are represented by horizontal lines. Chromosome 19 is an exception, for which evidence suggests that both heterochromatic regions are at least partially sequenced. Genomic features are presented as densities (that is, proportion of base pairs occupied by each feature) in nonoverlapping 1 Mb intervals. The densities are corrected for sequencing gaps, indicated in the draft assembly as 50-200 kb segments of Ns (unsequenced nucleotides), but (with the exception of GC content) are not corrected for sporadic Ns of lower-quality base calls, because these would not interfere with assignment of the feature to the assembly. Exon density (red) is based on high-scoring pairs from Table 1, not necessarily in ORFs. CpG island density (blue) is based on standard definitions [45] of a run of at least 200 bases with GC content > 50% and observed over expected CpG > 0.6, and implemented using the program cpg [90]. GC content (green) is the number of G or C bases divided by the number of non-N bases in the 1 Mb interval. LINE1 (blue) and *Alu* (black) repeat elements were determined using RepeatMasker [91] and minisatellites of repeat size 20-50 bp by the etandem program of the EMBOSS suite [84]. Density ranges were selected to illuminate features across the genome while preserving a common scale to facilitate comparison. A number of values exceed the range for the feature and are truncated, with a small dot of the corresponding color placed under the ordinate. The data points for the figure are available in the additional data file online.

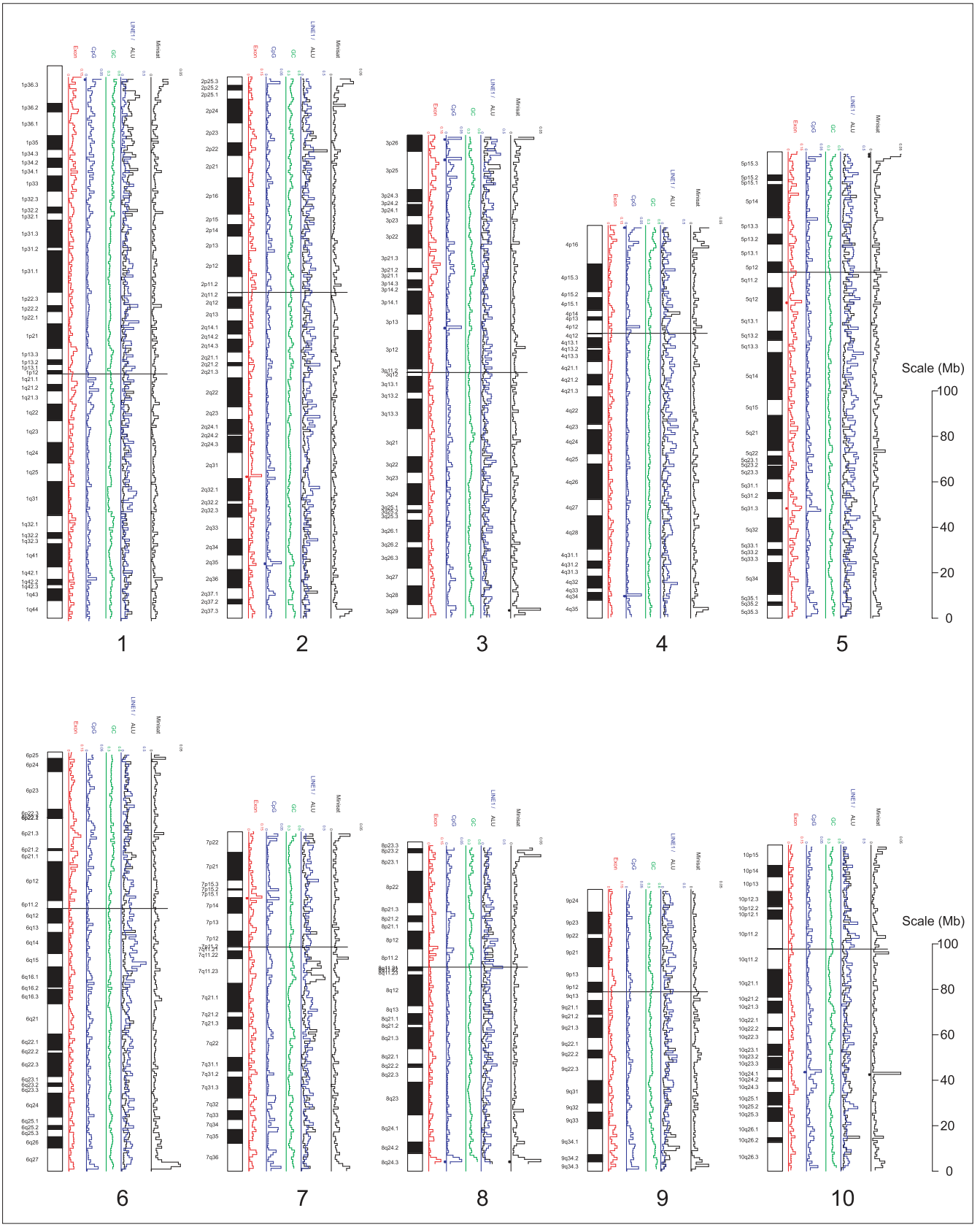

Figure 1 *(continued on following page, legend see previous page)*

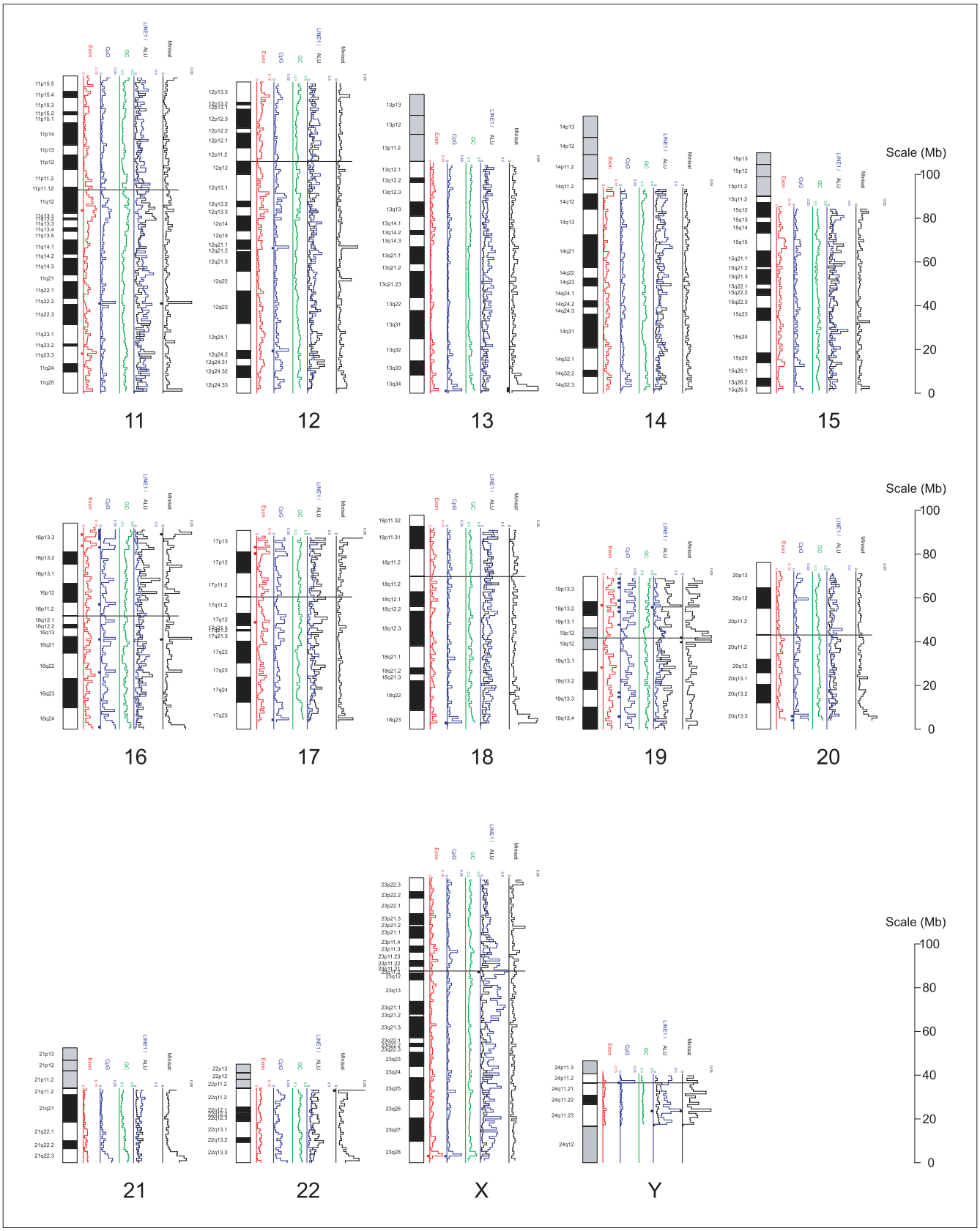

Figure 1 *(continued from previous page)*

is also in agreement with a report of FISH karyotypes using CpG island probes [46] with contrasting fluorescent signal in late-replicating regions. Extended regions of high CpG island density, such as the terminus of 1p and 1q21-q22, are apparent in the FISH assay. Short spikes of CpG islands (for example, in 3p26 and 3p25 of Figure 1) do not obviously appear in the assay, perhaps because they are below the resolution of FISH or are part of transcriptionally active regions.

In contrast to exon and CpG island density, GC content shows limited variation - in the range 35-55% for most 1 Mb intervals. The overall GC content is 41.1%. This compares with estimates in the range of 40-41% based on density gradient centrifugation [47] and flow cytometry [48].

Consistent with previous reports [49] *Alu* repeats show an apparent positive correlation with exon, CpG and GC densities, while LINE1 densities do not show such correlation. Approximately 1.1 million *Alu* repeats were identified, as expected [50]. However, a total of 758,000 LINE1 repeats were identified - 40% higher than estimates based on a sampling of sequenced regions [50]. Minisatellites of the hypervariable family (20-50 bp repeat size) are dispersed throughout the genome but, as expected [51], show sharp spikes in subtelomeric regions of most chromosomes.

Comparison of cytogenetic bands

We next examined the overall correspondence between cytobands and exonic density and other genomic features. Table 4 gives the average densities of features in the R bands versus G bands based on the assembly-corrected alignment. Genomic intervals residing in R bands were significantly richer in exons, CpG islands, GC content, *Alu* repeats and minisatellites than those in G bands. The reverse is true for LINE1 elements. These observations accord with predictions based on a variety of indirect methods [52], or a selected set of genes [53], but only now can be investigated directly using the sequence of the entire genome. The increased exonic density in R bands was fairly modest (approximately 30%), and may reflect attenuation due to alignment error. In addition, the analysis did not account for variation in staining intensity in G bands [41]. The results across the chromosomes were fairly consistent, however, and the R/G exonic density ratio exceeded 2.0 on two chromosomes (13 and 21) and was below 1.0 on only one chromosome (Y). The increased density of CpG islands in R bands was more striking (59%), whereas GC content was only a few percent higher (42.2 versus 39.8% in G bands), again consistent with previous observations [54]. The results for the cytobands are also reflected in pairwise correlations of the genomic features across 1 Mb intervals. These correlations do not depend on the cytoband alignment, and most features were positively correlated. LINE1 elements again differed from other features, showing a negative correlation with exons, CpG islands, GC content and *Alu* repeats.

Table 4

(a) Density of features per megabase in Giemsa-staining cytogenetic bands

	R	G	R/G ratio
Exons	0.0415	0.0319	1.30
CpG islands	0.0119	0.0075	1.59
GC content	42.23%	39.76%	1.06
LINE1 repeats	0.1435	0.1602	0.90
Alu repeats	0.1204	0.0937	1.28
Minisatellites	0.0090	0.0078	1.15

(b) Correlation of features in 1 Mb intervals

	Exon	CpG	GC	LINE1	*Alu*	Minisatellite
Exon	1.00	0.65	0.64	-0.26	0.73	0.19
CpG		1.00	0.73	-0.42	0.58	0.16
GC			1.00	-0.54	0.61	0.13
LINE1				1.00	-0.20	0.28
Alu					1.00	0.23
Minisatellite						1.00

(a) Pale-staining (R) and dark-staining (G) bands are compared, with alignment of cytogenetic bands to sequence as described in the text. All of the features except LINE1 elements are denser in the R bands. The true differences are likely to be larger, as errors in cytoband alignment will tend to understate the differences in the band types. The differences in the bands are highly significant at $p < 0.001$ for all features except for minisatellites ($p = 0.006$). **(b)** Rank correlations of features, in 1 Mb intervals ($p = 0.03$, corrected for multiple comparisons).

Gene density

We analyzed the exonic sequence for each chromosome as given in Table 1. Figure 2a shows the density of exonic sequence per chromosome. Chromosomes 19 and 17 are the richest (that is, densest) in exonic sequence [44], by factors of 2.04 and 1.62, respectively, compared to the average for the genome. Chromosomes 4, 13, 21, X and Y are exon-poor. A similar pattern emerges in the density of transcriptional units across the chromosomes, as shown in Figure 2b [19]. Reports based on integrated radiation hybrid (RH) maps of ESTs [55,56] indicated that chromosomes 1 and 22 were more gene-rich, but otherwise broadly agree with our results.

An intriguing clinical observation follows from these data and the tissue-specific observations. It had been noted [52] that the aneuploidies that are compatible with survival until birth (trisomies 13, 18 and 21, as well as X and Y aneuploidy) appeared to occur in relatively gene-poor chromosomes. Our data confirm these observations. However, the most obvious models for the deleterious effects of aneuploidy should instead depend on the total number of genes. In examining our HINT transcripts we have found that in fact the total number of embryo-specific transcripts is lowest on these five chromosomes (Figure 3). We suggest that trisomy of other

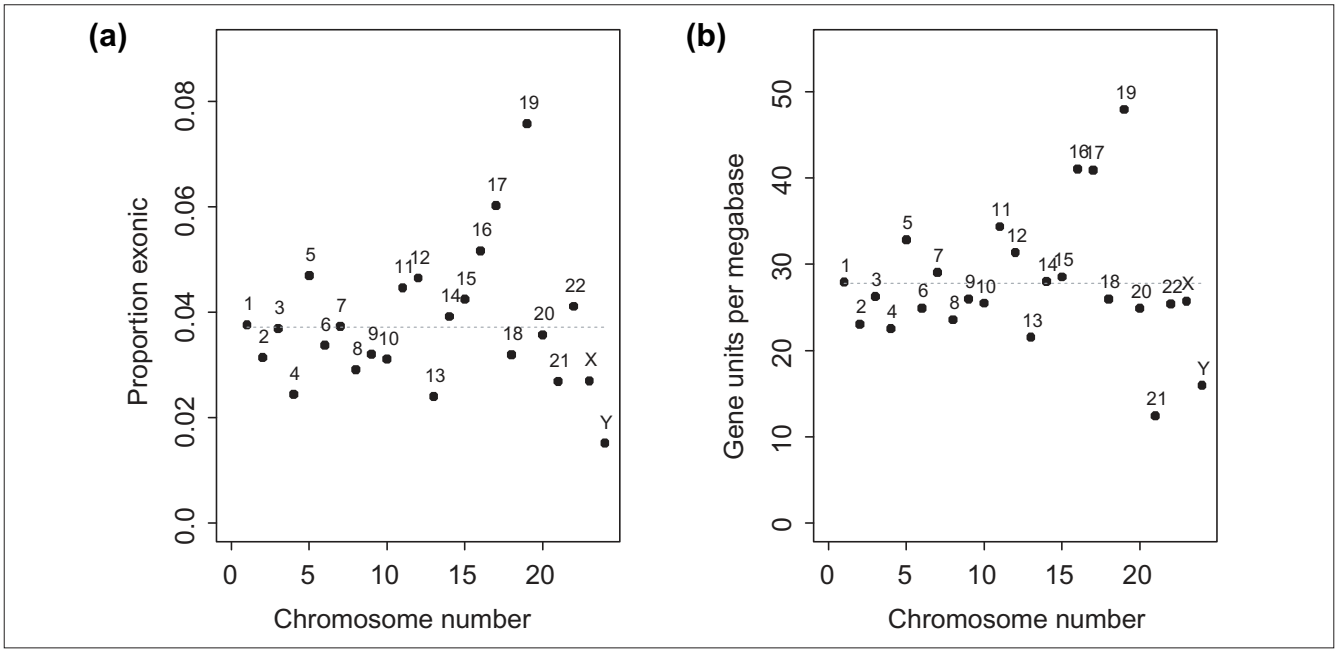

Figure 2
Coding sequence density for human chromosomes. **(a)** The proportion of assembled sequence that is exonic provides direct confirmation of previously hypothesized patterns of gene density. **(b)** Transcriptional units per megabase. Additional plots and data are in the additional data files online.

chromosomes may exceed a limit of survivable dosage compensation during development.

Comparisons to genetic and radiation hybrid maps

A total of 3,628 Genethon markers from the Marshfield map were localized via e-PCR [57] on the assembly, along with 28,350 Genebridge 4 markers/ESTs and 4,688 Stanford G3 markers appearing in Genemap'99. Figure 4 shows the positions of markers on the Chromosome 1 assembly. The curves are nearly monotonically increasing, showing that the assembly is broadly correct, although localized orientation errors and outliers remain (see additional data files online for plots for all chromosomes). These plots are immediately useful as they enable the placement of new markers on genetic maps without the need for mapping experiments. Some of the variation is likely to reflect estimation error in the published maps, and the curves are not completely monotonic for finished chromosomes 21 and 22. However, other regions are likely to reflect errors in assembly, as the genetic and RH maps agree with each other but disagree with the assembly (for example, the 130-148 Mb region is reversed on chromosome 5; a 15 Mb region of Xqter belongs at Xpter; numerous other isolated reversals and extensive reversals appear on chromosome 16). The genetic map shows a higher recombination rate per unit physical distance (that is, higher slope) at the telomeres, and a low male recombination rate (and thus sex-averaged rate) near the centromere (approximately 130 Mb). Similar patterns hold for the entire genome. These observations agree with previous studies which had been limited to comparisons of genetic and RH maps [58], male/female meiotic ratios [59], or relatively few markers on well-sequenced chromosomes [59]. The plots offer an interesting perspective on positional cloning efforts. For example, examination of the plots reveals that the hemochromatosis gene *HFE*, at 28 Mb on 6p, lies at the edge of a recombination 'cold spot' from 28-40 Mb. This fact complicated efforts to map the gene via linkage disequilibrium [60]. In contrast, the *NIDDM1* gene at 2qter (a region with higher recombination rate) was initially mapped to a 7 centimorgan (cM) region, which fortunately was discovered to be only 1.7 Mb of sequence [61].

The radiation hybrid plots tend to be more linear, which is consistent with the model that radiation induces chromosomal breakpoints essentially uniformly [62]. However, jumps in the Genebridge 4 (GB4) map occur at the centromere on most chromosomes. This may result from incomplete centromeric sequencing and assembly, so that a large centromeric gap might not appear as such. Alternatively, the jumps may reflect statistical difficulties in estimating breakpoint rates across the centromere. We note that no jump occurs in the G3 map, apparently because the higher radiation intensity produces insufficient marker pairs in the rescued hybrids that span the centromere. Thus, the jump cannot be accurately estimated and was simply suppressed in the published map [63]. The GB4 jump is strikingly large on several chromosomes, and we propose that the jumps might reflect increased radiation sensitivity at the centromere. This hypothesis is worth additional investigation.

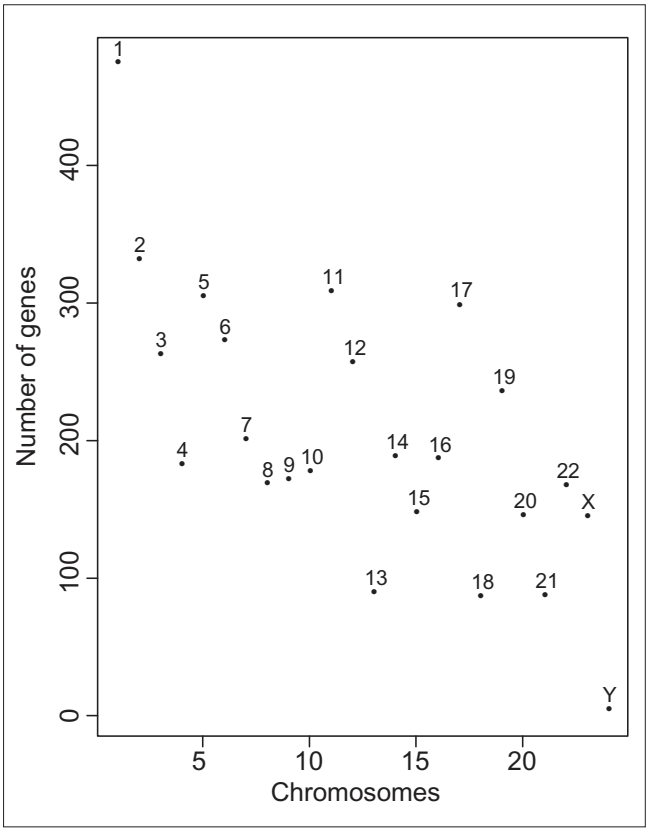

Figure 3
Total number of embryo-specific genes (based on HINT clusters) for each chromosome. Chromosomes 13, 18, 21 and Y clearly have lower numbers than other chromosomes.

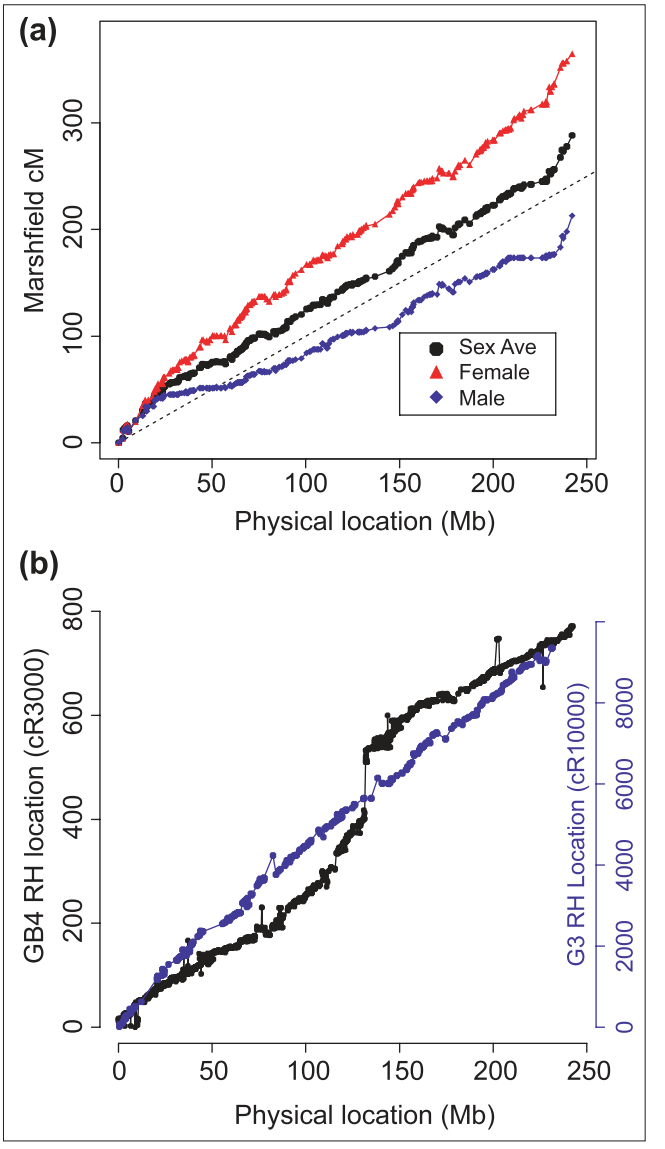

Figure 4
The correspondence between physical location and maps constructed using different mapping methods. **(a)** Correspondence between the genetic map and physical location. **(b)** Correspondence between radiation hybrid maps versus physical location. The GB4 (black) radiation hybrid map shows a jump at the centromere, reflecting a sequencing gap and possible increased radiation sensitivity in the region. The jump for the Stanford G3 map (blue) is not easily estimated and is suppressed in the published map. Chromosome 1 is shown here for illustration, and the corresponding figures and data points for the entire genome are available in the additional data files online.

Clusters and compartments

The availability of the full assembly enables a comparison of the entire genome to itself for evidence of homology arising from duplications or insertions. We emphasize that the genome is still in draft form, and a complete description of these features will be a large and ongoing scientific and computational task. We used BlastN [64] to identify intrachromosomal homology and to provide an initial look at the genomic landscape. Local duplication is a feature common to all chromosomes, as evidenced by the near-diagonal runs in dot-matrix plots in which the line of complete identity has been removed (Figure 5, and see additional data files online for full-page plots for each chromosome). These runs vary across the chromosomes, and tend to be of high sequence identity, indicative of recent origin. More distant duplications also occur, and include large repetitive regions of high identity on chromosomes 10 and 17. The Y chromosome shows strong internal sequence similarity, some of which arises from strikingly long duplications (from several of the order of 100 kb to a duplication of almost 1 Mb near the q-terminus of the euchromatic region). Near-duplicate sequences appear throughout the genome, producing a 'plaid' appearance on many chromosomes. These sequences tend to have lower sequence similarity (blue in Figure 5), consistent with an ancient origin and accumulated mutations.

As an example of functional duplication, we note that more than 60% of the entire zinc-finger (ZNF) families are mapped to chromosome 19, restricted to six large tandemly duplicated

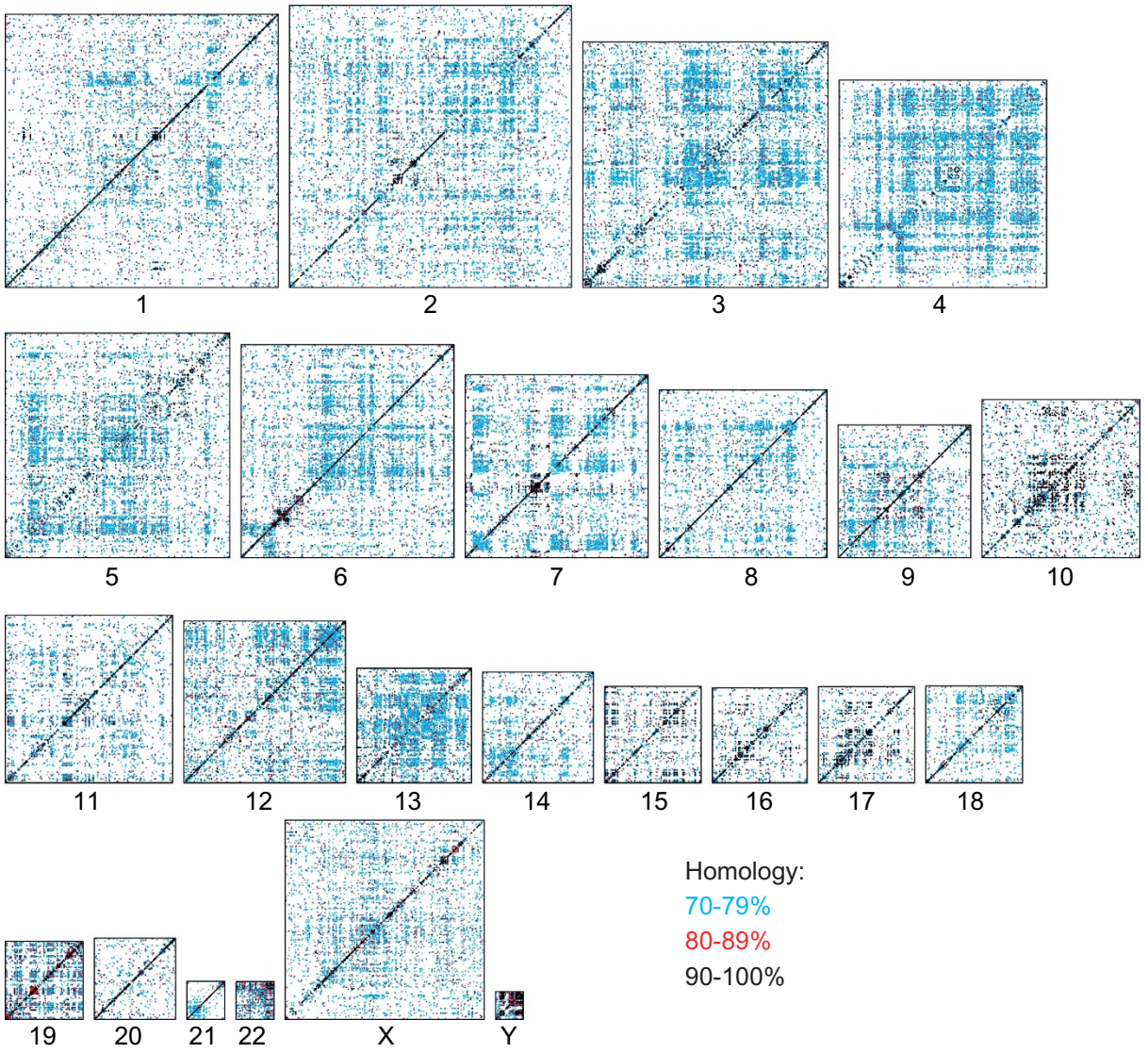

Figure 5
Repeat-masked chromosome sequences were divided into 1 Mb segments and analyzed against the entire chromosomal sequence. Matches of at least 70% identity (both forward and reverse) and $E < 10^{-25}$ are plotted. The diagonal line of complete identity has been removed to clarify features near the diagonal. Plots for each chromosome are available in the additional data files online.

gene clusters spanning the chromosome. More than one type of ZNF is found within each cluster, presumably as a result of sequence divergence. A majority of these ZNFs densely populate the 22-27 Mb region (see Figure 5). The remaining ZNFs are mapped to 15q21 (bZIP), 7q11 (KRAB), 11q13 (C_3HC_4), 11q23 (C_3HC_4), 6p21 (C_2H_2), 10p11 (KRAB), 10q11 (C_2H_2), 16p11 (C_2H_2), 9q22 (C_2H_2), and 3p21 (C_2H_2).

The largest functional group is related to phosphoryl transfer and protein kinases. Interestingly, many of the biological functions involving phosphoryl transfer form large gene clusters as well. For example, the mitogen-activated protein kinase family, phosphatidylinositol-4 phosphate 5-kinase family, protein kinase C family and at least 55 other diverse protein kinases are distributed in five gene clusters on chromosome 1, only about one third of which have been previously described. Similar gene clusters are also found on chromosomes 2, 3, 6, 19, 22 and X. In addition, DNA repair genes form gene clusters on different chromosomes, with postmeiotic segregation proteins (PMS) on chromosome 7, glycosidases on chromosome 12, MutS homologs on chromosome 6, MutT homologs on chromosome X, MutL

homologs on chromosome 2, Rad1/Rec1/Rad7 homologs on chromosome 10, excision repair on chromosome 11, and repair for single-strand nicks on chromosome 19. Additional regions of high and striking sequence similarity and the list of matching sequences with protein homology are provided in the additional data files online.

Paralogous genes resulting from recent gene duplication might preserve the same functionality and regulatory apparatus as their progenitors. We used chromosome 19 as the model to test this hypothesis by comparing the cDNA library profiles of spatially adjacent paralogous genes. At least one of the ZNF clusters (22-27 Mb region, Figure 5) appears to be more recent than the remaining clusters on the same chromosome (> 80% sequence identity). Intriguingly, two distinct tissue library profiles were scored for a total of 38 mapped ZNF paralogs, with the telomeric portion of the cluster predominantly expressed in germ cells (589/622 ESTs). The remaining members of the cluster were primarily expressed in embryos 9-19 weeks of age (145/167 ESTs). The same phenomenon did not hold for the ZNF clusters, where sequence similarity is lower. We were motivated to find additional paralogous genes, with their regulation similarly preserved. We mapped gene indices on duplicated genomic sequences. Alcohol dehydrogenases (1, 2, 3, 4, 5 and 7) are tandemly duplicated on 4q21, with their transcripts consistently being over-represented in embryonic and fetal cDNA libraries. Similar observations were obtained for other gene clusters, including amylases on 1p21, annexins on 4q21, homeobox proteins on 7p15 and 17q21, metallothioneins on 16q13, crystalline proteins on 2q33, glutathione-*S*-transferases (m1, m2, m3, m4 and m5) on 1p13, histone families (H2A/H2B/H3/H4) on 6p21, killer cell lectin-like receptors on 12p13, proline-rich proteins on 12p13, protocadherins on 5p15, s100 calcium-binding proteins on 1q21, keratins on 17q12, ADP-ribosylation factors (3, 4 and 5) on 10q22, and the major histocompatibility complex on 6p21. Together, these observations strongly support the notion that much of the regional clustering of functionally related proteins originates from gene duplication.

Clustering of ontological groups

We also examined the locations of all transcriptional units that had been classified according to a gene ontology-derived schema (Table 3, and see Materials and methods) for evidence of regional clustering of functionally related proteins. We applied a test that corrected for regional gene density, and found substantial evidence for regional clustering among the transcripts belonging to the same category (see additional data files online for location plots for the top 60 ontological categories). Such clustering is pervasive - much of it is likely to have arisen from duplication in which functional units have been preserved.

As an additional demonstration of the duplication phenomenon, we considered the occurrence of Pfam motifs within ORFs, with only the best Pfam match retained per ORF (around 1,930 of the 2,011 Pfam categories were represented). Matching successive runs of four or more (that occur at least three times on the genome) appear in the additional data files online. Many of the runs occur on the near-diagonal. Most involve four identical Pfam categories in succession, or a double run of two categories, again pointing to local duplication.

We also examined the runs of six or more gene units in which the ontological classifications occur in the same order (or the reverse) in multiple locations on the genome. A dot-matrix plot across the genome appears in the additional data files online. The plot shows clear evidence of local duplication, while the distant matches (even across chromosomes) are under investigation in the context of the complete sequence. We have noticed interesting associations among membrane proteins, ion channels, electron transporters, ATP-binding cassettes, and genes involving metabolism on chromosomes 2, 5 and 7. Proximity may be important for regulating functionally coupled genes, and intriguing observations of this phenomenon are well established in prokaryotic organisms [65] and recently reported in yeast [66]. We are investigating the possibility that at least some of the positional-functional coupling may be due to regulated mechanisms other than gene duplication.

Conclusions

The human genome is a capacious resource that will support years of intensive investigation. The quality of the draft sequence has now reached the point that genetic maps can truly be integrated into the genome. Analysis at the sequence level shows pervasive local and distant duplication, much of which preserves function. We have found evidence for a large number of transcriptional units (65,000-75,000) and performed initial annotation and classification. The effective study of transcription and protein function requires the compilation of all available evidence of transcription and protein homology.

Since the initial submission of this manuscript, two reports have appeared [67,68] in which the human genome is analyzed and described. The raw genomic sequence used for the present study was generated by the Human Genome Consortium (HGC), and is principally the same as that used in their report [67]. In our effort, we have benefited from early-release and open data policies adopted by the consortium. The three reports offer insights into varying aspects of the genome. In our report we have emphasized the compilation of all available transcriptional evidence, and have made observations of functional-positional clustering of genes and global tissue specificity that will provide the basis for in-depth investigation.

The other reports have arrived at estimates of approximately 30,000 genes in the genome, and the ensuing attention

leaves the unfortunate impression that genomic analysis consists primarily of gene counting. While our estimate of 65,000-75,000 transcriptional units clearly differs from these estimates, it is useful to consider that there is much independent agreement in the reports. We have placed around 15,000 known genes in the genome, and described 33,000 that have substantial homology to the major protein databases. These figures are very comparable to the known and 'confirmed' genes, respectively, placed on the sequence by Celera Genomics and HGC. The Celera and HGC reports rely on a combination of computational gene prediction and transcriptional evidence to identify genes, in a manner that may be highly conservative. Our approach to identifying genes and exons is more heavily transcript-based, with the presumption that pure computational gene prediction may still have substantial false-negative rates. In applying this approach, we find experimental evidence for many novel exons with clear splicing evidence, on the order of twice as many as described in the other reports. This fact largely explains our greater number of transcriptional units, as our mean number of exons per transcriptional unit (613,183/759,822 = 8.07) is very comparable to the Celera and HGC reports.

We hold that the reliance on *ab initio* approaches presents difficulties, as known genes may be more highly expressed and may present a biased sample for calibration. Another difference in the approaches is that we have used pre-assembled consensus transcripts whenever possible in alignment to the genome. The use of such consensus sequences is likely to improve alignment, increase splicing evidence, and has been shown to improve the detection of protein homology [19]. We propose that the remaining 32,000-42,000 units we have identified will represent a useful resource for additional investigation as genomic annotation proceeds. A comparison of all three approaches will surely yield new insights.

Materials and methods

Exon identification

The 26 June 2000 version of the repeat-masked draft sequences was downloaded from the Ensembl Genome Server [69] and blasted against cDNA and protein sequences by using the Blast program compiled from the NCBI toolkit (6.1) on a 32-node SGI Linux/Intel Cluster, with four 550 MHz Pentium III Xeon processors and 2 GB of RAM on each node. The following databases were used: Human UTR-DB (EBI) (version 13) [70]; Human Transcript Database (Baylor University) (version 1) [71]; GenBank CDS (NCBI) (only PRI mRNA sequences were used, version 119) [72]; HINT (Ohio State University) [73]; EST Assembly Project (University of Washington) [74]; TIGR Human Gene Index (version 4.5) [75]; dbEST (NCBI) (version 119) [76]; MINT and RINT (Ohio State University) [73]; EMBL Rodent (EMBL) (version 63) [77]; SWISS-PROT (EMBL) (version 39) [78]; TrEMBL (EMBL) (version 14) [78]; PIR (MIPS-JIPID) (version 65) [79]; and Pfam (Sanger Centre) (version 5.4) [80]. Gene prediction programs used are described in Table 2. The Mouse and Rat Indices of Non-redundant Transcripts (MINT and RINT) were derived from Mouse and Rat UniGene [81] using the same approach that we have applied to human UniGene [19]. Briefly, chimeric sequences were removed, UniGene transcripts were assembled into sequence contigs, and links to progenitor records retained.

The genome-wide hit expectation value was set at $E < 10^{-25}$ (BlastN) or $E < 10^{-15}$ (BlastX) to filter out nonspecific high-scoring segment pairs (HSPs). Default parameters of Blast were used. The Blast report was parsed into field-specific tables using the program MSPcrunch Version 2.3 [82]. The resulting table was processed using a set of Perl scripts by first retaining only the HSPs that were spliced from the same transcripts on the same genomic contig. The same process was then applied to the HSPs on the genomic sequences. Spliced HSPs from the same transcripts were retained, followed by the singleton HSPs that were both longer and higher in sequence identity over their overlapping counterparts, resulting in a unique placement for each cDNA segment on the genomic sequence.

Prediction of transcriptional units

A set of Perl scripts was used to implement the algorithm described below. Genomic clones were ordered and oriented using the fingerprint map and draft assembly. Within unfinished clones, sequence contigs were further ordered and oriented according to Ensembl's assembly [83]. This mapping produced the positional context necessary for consolidating fragmented exon units. Where necessary, small sequencing gaps (100 bp or fewer) were ignored and genomic clones were considered contiguous except where a large gap was indicated in the draft assembly (> 50 kb). ORFs were determined using the program getorf [84]. The exon index is an integrated table generated by a Sybase relational database, consisting of chromosome number, fingerprinted contig (FPC) ID, FPC contig order, BAC contig ID, BAC contig order, BAC contig orientation, starting position of exon on BAC contig, end position of exon on BAC contig, exon orientation, transcript orientation (available from GenBank, IMAGE, UniGene, HINT and dbEST), evidence (transcript, protein, gene prediction, ORF, Pfam), database name (Table 1), feature (poly(A) signal, CpG island, Genscan boundary), starting position on exon (or feature), end position on exon (or feature), score (BlastN, BlastX). The index was first ordered and oriented for the individual BAC contigs according to Ensembl or UCSC [85] maps. The resulting contigs were then ordered and oriented according to the FPC order and orientation information in the UCSC genome assembly, resulting in a numeric sorting order for all the individual contigs. In addition, large gap information (> 50 kb) available from the UCSC assembly was incorporated into the same index, where no overlapping information was available between presumably adjacent BACs.

The consolidation algorithm follows a hierarchy in which unit boundaries are respected for the highest-ranking feature. The features in descending ranking were: UTRs based on known UTR indices; exons containing no ORFs or incomplete ORFs; boundaries of known full-length cDNAs (HTDB-based indices); EST orientation information (5′ or 3′ origin from the original IMAGE, UniGene/HINT, and dbEST databases); and Genscan-predicted poly(A) signals. When clear boundary indicators were not available, information from the transcript indices HINT (assembled from UniGene) [19] and EG [17] were used directly as secondary evidence for potential gene boundaries. The rationale is that each UniGene cluster has at least one known gene, or two sets of ESTs representing both the 5′ and 3′ termini of a gene, or at least one EST containing a poly(A) signal [81]. Similar stringent criteria were used in Ewing and Green's EST assemblies [17]. Multiple exons not residing in intact ORFs were consolidated until the occurrence of exons in a partial or complete ORF. Multiple in-frame exons in a continuous ORF were always considered part of a single gene. To prevent any overconsolidation as a result of lack of transitional exons (in partial or complete ORFs) for adjacent genes, CpG islands, large gaps (> 50 kb) between exons and Genscan prediction were used as gene boundaries when higher-ranking boundary information was unavailable. In such instances, HINT and EG index identity was respected. Although a variety of criteria were used for determining transcriptional unit boundaries, the vast majority of the consolidation was achieved on the basis of terminal information from gene indices and transition and continuation of open reading frames.

Gene mapping

A relational database was used to integrate multiple largely independent maps for the genomic clones, where transcripts had been placed. This integration thus results in a transcript map based on the order and position of genomic clones. Individual sequencing contigs within each unfinished clone were oriented using the Ensembl contig map [83]. The fingerprint (see [8], version 15 June 2000), GoldenPath assembly (Versions 15 June and 5 September 2000), and radiation hybrid maps [86] were used to place genomic clones into their chromosomal context. As a substantial number of the clones in the working draft had not been physically typed with RH or genetic markers, the program e-PCR [48] and primers collected in the RHdb [87] and Genethon [88] were used under stringent criteria (mismatch = 0, margin = 50, and word size = 7). Genetic mapping information was obtained from the Marshfield map [89]. In addition, Genemap'99 for cDNA was integrated into the genomic clones harboring HINT consensus transcripts. For the HINT consensus with more than one mapped EST, an averaged RH position was used. Cytogenetic bands were inherited from the original UniGene database. Furthermore, we incorporated a weighted composite quality score for the following four maps: Genemap'99 (the number of consistently mapped ESTs and their associated genomic clones), e-PCR (the number of consistently mapped sequence-tagged sites (STSs) in a genomic clone), FPC (the supporting evidence in the original database), Blast (evidence of splicing). On the basis of such an integrated database schema, mapping information from sequence, clone, contigs, radiation hybrid and cytogenetic positions for a given transcript could be obtained through a SQL (Structured Query Language) join statement.

Tissue-specific transcripts

We noted the total number of ESTs contributed by each tissue to compute an expected proportion. For each HINT consensus transcript, we identified the tissue/source contributing the most ESTs to the consensus. The expected binomial distribution for the fixed number of ESTs in the consensus was used to compute a p-value, which was then Bonferroni-corrected for the 81 tissues x 67,000 HINT consensus transcripts.

Cytoband alignment

G bands are known to be relatively AT-rich, but the precise relationship between sequence and cytoband position is too poorly understood to be used for alignment. Genes/ESTs with cytoband position appearing in UniGene were placed on the full genome assembly. Cytoband cutpoints were used to create a scatterplot with the center of the cytoband forming the x-coordinate, and assembly position as the y-coordinate. Outliers were identified as points lying more than 2.5 standard errors outside of prediction intervals from a third-degree polynomial regression fit. A Loess regression fit was used on the remaining points to estimate cytoband boundaries, with p and q arms fitted separately. Centromeres and heterochromatic regions were assumed not sequenced, on the basis of a review of current clone frameworks. Primary sources for assignments of genes to heterochromatic regions were examined and in most cases deemed inconclusive. An exception is chromosome 19, which has a considerable number of genes assigned to 19q12 and finished sequence in the region. Scatterplots and regression fits for the entire genome are in the additional data files online.

Genomic feature correlations

All 1 Mb intervals were combined to produce Table 4, but statistical tests were performed by computing ratios and correlations within each chromosome separately, in order to account for correlation of features within each chromosome. These statistics were then compared across the chromosomes to an appropriate null value using single sample t-tests. Some of the features were skewed, and pairwise comparisons were performed using Spearman rank correlations. A Bonferroni multiple-comparison procedure was applied to the 15 unique correlations.

Regional functional clustering

Apparently significant clustering can arise from the fact that genes exhibit regional clustering. To correct for this, we con-

sidered the physical order of all mapped transcripts and calculated the distances (in ranked location) between transcripts belonging to the same ontological category. Under the null hypothesis, the transcripts in a category should be distributed uniformly among all mapped transcripts with ontological classification, and the successive distances are approximately truncated exponential. On this basis we compared the observed tenth percentile of successive distances to that under the null hypothesis to compute a p-value. All tests were highly significant, with $p < 0.0001$ for 59 of the 60 largest categories, and quantile-quantile plots with observed versus expected distributions showed striking evidence of clustering. These tests were confirmed with permutation tests with empirical generations under the null hypothesis. As a conservative correction for the possibility that separate transcriptional units might belong to the same gene, we considered successive distances for every other transcript. These tests were also significant, with $p < 0.01$ for the 60 categories.

Additional data files

Additional data files are available online as follows: transcriptional units; computational prediction; FPC, e-PCR, RH and assembly maps; functional annotation; tissue expression profiles; a global view of the human genome; density of genomic features; comparisons to genetic and RH maps; clusters and compartments; and clustering of ontological groups.

Acknowledgements

We thank the numerous investigators of the Human Genome Project for sequence availability and for open-data policies, Albert de la Chapelle for support and encouragement, Jian-Ping Guo, Solomon Gibbs, Dara Goodheart, Daolong Wang and Anthony Jakubisin for assistance, the Ohio Supercomputer Center (OSC) for invaluable assistance and computational resources, the Institute for Pure and Applied Mathematics at UCLA for provision of technical facilities, and LabBook.Com for database and user interface support. This work was supported in part by the Solove Research Foundation and NIH grant GM58934 (F.A.W.).

References

1. **International Human Genome Consortium** [http://www.nhgri.nih.gov/genome_sequence.html]
2. Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO, Hunkapiller M: **Shotgun sequencing of the human genome.** *Science* 1998, **280:**1540-1542.
3. **TIGR Microbial Database** [http://www.tigr.org/tdb/mdb/mdbcomplete.html]
4. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al.*: **The genome sequence of *Drosophila melanogaster*.** *Science* 2000, **287:**2185-2195.
5. The Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408:**796-815.
6. **Washington University GSC Human Genome Project BAC and Accession Maps** [http://genome.wustl.edu/gsc/human/Mapping]
7. **NCBI Human Genome Sequencing Progress** [http://www.ncbi.nlm.nih.gov/genome/seq]
8. Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, Hillier LW, McPherson JD, Waterston RH: **High throughput fingerprint analysis of large-insert clones.** *Genome Res* 1997, **7:**1072-1084.
9. Morton NE: **Parameters of the human genome.** *Proc Natl Acad Sci USA* 1991, **88:**7474-7476.
10. Boguski MS: **Biosequence exegesis.** *Science* 1999, **286:**453-455.
11. **Ensembl Science Documentation** [http://www.ensembl.org/Docs/wiki/html/EnsemblDocs/ScienceDocumentation.html]
12. Murakami K, Takagi T: **Gene recognition by combination of several gene-finding programs.** *Bioinformatics* 1998, **14:**665-675.
13. Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, Smink LJ, *et al.*: **The DNA sequence of human chromosome 22.** *Nature* 1999, **402:**489-495.
14. Boguski MS, Schuler GD: **ESTablishing a human transcript map.** *Nat Genet* 1995, **10:**369-371.
15. Miller RT, Christoffels AG, Gopalakrishnan C, Burke J, Ptitsyn AA, Broveak TR, Hide WA: **A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base.** *Genome Res* 1999, **9:**1143-1155.
16. Quackenbush J, Liang F, Holt I, Pertea G, Upton J: **The TIGR gene indices: reconstruction and representation of expressed gene sequences.** *Nucleic Acids Res* 2000, **28:**141-145.
17. Ewing B, Green P: **Analysis of expressed sequence tags indicates 35,000 human genes.** *Nat Genet* 2000, **25:**232-234.
18. **Blast Help Manual** [http://www.ncbi.nlm.nih.gov/BLAST/blast_help.html]
19. Zhuo D, Zhao WD, Wright FA, Yang H-Y, Wang J-P, Sears S, Baer T, Kwon D-H, Gordon D, Gibbs S, *et al.*: **Assembly, annotation, and integration of UniGene clusters into the human genome draft.** *Genome Res* 2001, **11:**904-918.
20. Pesole G, Sabino L, Grillo G, Licciulli F, Larizza A, Makalowski W, Saccone C: **UTRdb and UTRsite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs.** *Nucleic Acids Res* 2000, **28:**193-196.
21. Bouck J, McLeod MP, Worley K, Gibbs RA: **The human transcript database: a catalogue of full length cDNA inserts.** *Bioinformatics* 2000, **16:**176-177.
22. Mironov AA, Fickett JW, Gelfand MS: **Frequent alternative splicing of human genes.** *Genome Res* 1999, **9:**1288-1293.
23. Burke J, Wang H, Hide W, Davison DB: **Alternative gene form discovery and candidate gene selection from gene indexing projects.** *Genome Res* 1998, **8:**276-290.
24. **Guidelines for Human Gene Nomenclature** [http://www.gene.ucl.ac.uk/nomenclature/guidelines.html]
25. Gentles AJ, Karlin S: **Why are human G-protein-coupled receptors predominantly intronless?** *Trends Genet* 1999, **15:**47-49.
26. Brosius J: **Many G-protein-coupled receptors are encoded by retrogenes.** *Trends Genet* 1999, **15:**304-305.
27. Mighell AJ, Smith NR, Robinson PA, Markham AF: **Vertebrate pseudogenes.** *FEBS Lett* 2000, **468:**109-114.
28. Antequera F, Bird A: **Number of CpG islands and genes in human and mouse.** *Proc Natl Acad Sci USA* 1993, **90:**11995-11999.
29. Fields C, Adams MD, White O, Venter JC: **How many genes in the human genome?** *Nat Genet* 1994, **7:**345-346.
30. Roest Crollius H, Jaillon O, Bernot A, Dasilva C, Bouneau L, Fischer C, Fizames C, Wincker P, Brottier P, Quetier F, *et al.*: **Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence.** *Nat Genet* 2000, **25:**235-238.
31. Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J: **Gene index analysis of the human genome estimates approximately 120,000 genes.** *Nat Genet* 2000, **25:**239-240.
32. Aparicio AAJR: **How to count ... human genes.** *Nat Genet* 2000, **25:**129-130.
33. Erdmann VA, Barciszewska MZ, Szymanski M, Hochberg A, Groot N, Barciszewski J: **The non-coding RNAs as riboregulators.** *Nucleic Acids Res* 2001, **29:**189-193.
34. Junker VL, Apweiler R, Bairoch A: **Representation of functional information in the SWISS-PROT data bank.** *Bioinformatics* 1999, **15:**1066-1067.
35. **International Gene Ontology Consortium** [http://www.geneontology.org]
36. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, *et al.*: **Comparative genomics of the eukaryotes.** *Science* 2000, **287:**2204-2215.
37. **The IMAGE Consortium** [http://image.llnl.gov]

38. Walker MG, Volkmuth W, Sprinzak E, Hodgson D, Klinger T: **Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes.** *Genome Res* 1999, **9:**1198-1203.
39. Sohocki MM, Malone KA, Sullivan LS, Daiger SP: **Localization of retina/pineal-expressed sequences: identification of novel candidate genes for inherited retinal disorders.** *Genomics* 1999, **58:**29-33.
40. Mannervik M, Nibu Y, Zhang H, Levine M: **Transcriptional coregulators in development.** *Science* 1999, **284**:606-609.
41. Francke U: **Digitized and differentially shaded human chromosome ideograms for genomic applications.** *Cytogenet Cell Genet* 1994, **65:**206-218.
42. Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS, Toyoda A, Ishii K, Totoki Y, Choi DK, *et al.*: **The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium.** *Nature* 2000, **405:**311-319.
43. Trask BJ: **Fluorescence *in situ* hybridization: applications in cytogenetics and gene mapping.** *Trends Genet* 1991, **7:**149-154.
44. Inglehearn CF: **Intelligent linkage analysis using gene density estimates.** *Nat Genet* 1997, **16:**15.
45. Larsen F, Gundersen G, Lopez R, Prydz H: **CpG islands as gene markers in the human genome.** *Genomics* 1992, **13:**1095-1107.
46. Craig JM, Bickmore WA: **The distribution of CpG islands in mammalian chromosomes**. *Nat Genet* 1994, 7:376-382.
47. Thiery JP, Macaya G, Bernardi G: **An analysis of eukaryotic genomes by density gradient centrifugation.** *J Mol Biol* 1976, **108:**219-235.
48. Vinogradov: **Measurement by flow cytometry of genomic AT/GC ratio and genome size.** *Cytometry* 1994, **16:**34-40.
49. Korenberg JR, Rykowski MC: **Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands.** *Cell* 1988, **53:**391-400.
50. Smit AF: **The origin of interspersed repeats in the human genome.** *Curr Opin Genet Dev* 1996, **6**:743-748.
51. Amarger V, Gauguier D, Yerle M, Apiou F, Pinton P, Giraudeau F, Monfouilloux S, Lathrop M, Dutrillaux B, Buard J, *et al.*: **Analysis of distribution in the human, pig, and rat genomes points toward a general subtelomeric origin of minisatellite structures.** *Genomics* 1998, **52:**62-71.
52. Strachan T, Read AP: *Human Molecular Genetics*, 2nd edn. New York: BIOS Scientific Publishers Ltd, 1999.
53. Craig JM, Bickmore WA: **Chromosome bands-flavours to savour**. *BioEssays* 1993, **15**:349-354.
54. Saitoh Y, Laemmli U: **Metaphase chromosome structure: bands arise from a differential folding path of the highly AT-rich scaffold.** *Cell* 1994, **76:**609-622.
55. Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, *et al.*: **A gene map of the human genome.** *Science* 1996, **274:**540-546.
56. Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tome P, Hui L, Matise TC, McKusick KB, Beckmann JS, *et al.*: **A physical map of 30,000 human genes.** *Science* 1998, **282:**744-746.
57. Schuler GD: **Sequence mapping by electronic PCR.** *Genome Res* 1997, **7:**541-550.
58. Collins A, Frezal J, Teague J, Morton NE: **A metric map of humans: 23,500 loci in 850 bands.** *Proc Natl Acad Sci USA* 1996, **93:**14771-14775.
59. Broman KW, Murray JC, Sheffield VC, White RL, Weber JL: **Comprehensive human genetic maps: individual and sex-specific variation in recombination.** *Am J Hum Genet* 1998, **63**:861-869.
60. Thomas W, Fullan A, Loeb DB, McClelland EE, Bacon BR, Wolff RK: **A haplotype and linkage disequilibrium analysis of the hereditary hemochromatosis region.** *Hum Genet* 1998, **102:**517-525.
61. Horikawa Y, Oda N, Cox N, Li X, Orho-Melander M, Hara M, Hinokio Y, Lindner TH, Mashima H, Schwarz PEH, *et al.*: **Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus.** *Nat Genet* 2000, **26:**163-175.
62. Lawrence S, Morton NE, Cox DR: **Radiation hybrid mapping.** *Proc Natl Acad Sci USA* 1991, **88:**7477-74788.
63. **Stanford Human Genome Center** [http://www-shgc.stanford.edu/Mapping]
64. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
65. Kihara D, Kanehisa M: **Tandem clusters of membrane proteins in complete genome sequences.** *Genome Res* 2000, **10:**731-743.
66. Cohen BA, Mitra RD, Hughes JD, Church GM: **A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression.** *Nat Genet* 2000, **26:**183-186.
67. The International Human Genome Sequencing Consortium: **Initial sequence and analysis of the human genome.** *Nature* 2001, **409:**860-921.
68. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith H, Yandell M, *et al.*: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.
69. **The Ensembl Human Genome Server** [http://www.ensembl.org]
70. **UTR DB at European Bioinformatics Insitute** [ftp://ftp.ebi.ac.uk/pub/database/UTR]
71. **Baylor Human Transcript Database** [http://www.hgsc.bcm.tmc.edu/HTDB]
72. **GenBank CDS at National Center for Biotechnology Information** [ftp://ncbi.nlm.nih.gov/blast/db/nt.Z]
73. **Assembly, annotation and integration of UniGene clusters into the human genome draft** [http://pandora.med.ohio-state.edu/HINT]
74. **EST Assembly Projects** [http://www.phrap.org/est_assembly]
75. **TIGR Human Gene Index** [http://www.tigr.org/tdb/hgi]
76. **dbEST at National Center for Biotechnology Information** [ftp://ncbi.nlm.nih/blast/db/est_human.Z]
77. **EMBL Rodent at European Bioinformatics Institute** [ftp://ftp.ebi.ac.uk/pub/databases/embl/release/rod.dat.gz]
78. **SWISS-PROT/TrEMBL database** [http://www.ebi.ac.uk/swissprot]
79. **Protein Information Resource** [http://pir.georgetown.edu]
80. **Protein families database of alignments and hidden Markov models** [http://www.sanger.ac.uk/Software/Pfam]
81. **NCBI UniGene Resources** [http://www.ncbi.nlm.nih.gov/UniGene/]
82. **MSPcrunch 2.1** [ftp://ftp.cgr.ki.se/pub/prog/MSPcrunch+Blixem]
83. **Ensembl Assembly at the Sanger Centre** [ftp://ftp.sanger.ac.uk/pub/ensembl/data/mysql/contig.txt.table.gz]
84. **The European Molecular Biology Open Software Suite** [http://www.emboss.org]
85. **Human Genome Project Working Draft at UCSC** [http://genome.ucsc.edu]
86. **Genemap at National Center for Biotechnology Information** [ftp://ncbi.nlm.nih.gov/repository/genemap/Mar1999]
87. **Radiation Hybrid Database** [http://corba.ebi.ac.uk/RHdb]
88. **Genethon** [http://www.genethon.fr/genethon_en.html]
89. **Marshfield Clinic: Center for Medical Genetics** [http://research.marshfieldclinic.org/genetics]
90. **Sanger Centre Software** [http://www.sanger.ac.uk/software]
91. **RepeatMasker** [http://www.genome.washington.edu/UWGC/analysistools/repeatmask.htm]
92. **HMMER 2.1.1** [http://hmmer.wustl.edu]
93. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268:**78-94.
94. Solovyov V, Salamov A: **The Gene-Finder computer tools for analysis of human and model organisms genome sequences.** In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*. Edited by Gaasterland T, Karp P, Karplus K, Ouzounis C, Sander C, Valencia A. Menlo Park: American Association for Artificial Intelligence Press, 1997.
95. **Grail 1.3** [http://compbio.ornl.gov/Grail-1.3]