The Human Genome: Genes, Pseudogenes, and Variation on Chromosome 7

R.H. WATERSTON,* L.W. HILLIER,† L.A. FULTON,† R.S. FULTON,† T.A. GRAVES,† K.H. PEPIN,† P. BORK,‡ M. SUYAMA,‡ D. TORRENTS,‡ A.T. CHINWALLA,† E.R. MARDIS,† J.D. MCPHERSON,†¶ AND R.K. WILSON†

*Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195;
†Genome Sequencing Center, Department of Genetics, Washington University, St. Louis, Missouri 63108;
‡EMBL, Heidelberg 69117, Germany*

When the idea of sequencing the entire human genome was initially put forward (Dulbecco 1986; DeLisi 1988; Sinsheimer 1989), DNA sequencing was an expensive, laborious process. The largest genome to have been sequenced at that time was that of the Epstein-Barr virus, which at 172,282 bases, or about 1/20,000 the size of the human genome, had taken a group of about a dozen people several years to complete (Baer et al. 1984). The only systematic analyses of larger, more complex genomes had sought to produce ordered clone sets (clone-based physical maps) of the *Caenorhabditis elegans* and *Saccharomyces cerevisiae* genomes (Coulson et al. 1986; Olson et al. 1986). Today, less than 20 years later and just 2 1/2 years after publications describing draft sequences (Lander et al. 2001; Venter et al. 2001), we have in hand the essentially complete sequence of the human genome (Rogers, this volume). Groups from across the globe contributed to the sequence (Table 1) in a coordinated effort called the International Human Genome Project (IHGP).

Table 1. Contributions to the Finished Sequence by Center

Center	Bases
Wellcome Trust Sanger Institute	824,879,622
Washington University Genome Sequencing Center	584,507,988
U.S. DOE Joint Genome Institute	310,982,691
Whitehead Institute Center for Genome Research	374,404,412
Baylor College of Medicine Human Genome Sequencing Center	278,229,771
RIKEN Genomic Sciences Center	110,409,096
University of Washington Genome Center	108,221,396
Genoscope and CNRS UMR-8030	77,479,268
Keio University	21,285,289
GTC Sequencing Center	36,061,549
The Institute for Systems Biology	26,871,969
Max Planck Institute for Molecular Genetics	20,095,113
Beijing Genomics Institute/Human Genome Center	17,141,643
University of Oklahoma's Advanced Center for Genome Technology	5,444,166
Stanford Human Genome Center	8,374,342
Max Planck Institute for Molecular Genetics	5,024,546
GBF German Research Center for Biotechnology	5,547,223
Others	17,049,046

¶Present address: Department of Molecular and Human Genetics and Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, N1519, Houston, Texas 77030.

The groups exploited advances contributed by many laboratories for every step of the sequence process (Table 2), but none was more important than the steady advance in fluorescence-based DNA sequencing instruments (Smith et al. 1985; Ansorge et al. 1987; Prober et al. 1987; Brumbaugh et al. 1988). Initially these machines required sophisticated users and large amounts of carefully quantified, high-quality template DNA to produce sequence of limited accuracy over 300–400 bases, and capacity was limited—for example, 16 samples daily on the ABI373 machine. With steady improvements not only in the instruments, but also in each step of the process before and after sequence collection, a typical ABI3730 in a large genome center today produces 1300 samples daily with highly accurate reads of 750–900 bases with minimal attendance. Automation with sophisticated LIMS systems and sample tracking allow these machines to be fed continuously 24 hours a day, 7 days a week, 52 weeks a year, with little down time.

The clone-based hierarchical shotgun strategy employed by the IHGP has resulted in a sequence that contains more than 99% of the euchromatic sequence in highly accurate form. The details of the sequence are presented elsewhere, but it is clear that the sequence includes some large, complicated, repeated sequences that would

Table 2. Advances in Sequencing

Advance	Reference
Libraries	Shizuya et al. (1992)
DNA template prep	Hawkins et al. (1994)
Cycle sequencing	Axelrod and Majors (1989); Craxton (1993)
Fluorescent dyes	Ju et al. (1995)
Taq polymerase variants	Tabor and Richardson (1995)
Capillary sequencing	Lu et al. (1994)
Base calling	Ewing et al. (1998)
Assembly	P. Green and L. Hillier (unpubl.)
Databases	R. Durbin and J. Thierry-Mieg (unpubl.); FlyBase Consortium (1994)

The principal steps in the DNA sequence process are listed along with examples of early contributions to improvements to the process. The continual improvement of each step has been critical to the success of the Human Genome Project (Axelrod and Majors 1989; Shizuya et al. 1992; Craxton 1993; Hawkins et al. 1994; Lu et al. 1994; Ju et al. 1995; Tabor and Richardson 1995; Ewing et al. 1998).

have been difficult to obtain with the alternative whole-genome shotgun strategy. The single-base miscall rate is estimated at about 1/100,000 bases, and small deletion/insertion differences arising from errors in propagation and assembly occur on the order of 1 per 5 Mb (Schmutz et al., this volume), both rates much lower than observed polymorphism rates for these types of differences. Indeed, many of the small insertion/deletion artifacts occur in short, tandemly repeated sequences, which are often themselves polymorphic in the population.

Some regions of the genome are not accounted for in the current sequence. The short arms of the acrocentric chromosomes are not represented at all, and the large heterochromatic regions of Chromosomes 1, 9, and 16 and the centromeric α-satellite sequences are only sparsely represented. There are also infrequent gaps in the euchromatic regions, whose size generally is estimated from FISH or comparison with orthologous mouse/rat genome sequence. These gaps in sequence coverage reflect the absence of clones in available libraries and are particularly prevalent near the centromeres and telomeres and in heterochromatic regions (Dunham et al. 1999; Deloukas et al. 2001; Heilig et al. 2003; Hillier et al. 2003; Mungall et al. 2003; Skaletsky et al. 2003). These gaps are often flanked by locally highly repetitive sequence, by segmentally duplicated sequence, or by regions with exceptionally high GC content. Efforts to close these few remaining gaps and to correct detected errors continue at many of the centers.

With a highly accurate, essentially complete sequence of the genome now in hand, efforts are shifting to the understanding of its contents. A full understanding will take decades, but two immediate tasks face us now: defining those sequences that function in the specification of humans—the "parts list'"—and defining human variation. We comment in this paper on aspects of these tasks that we faced in annotating the finished Chromosome 7 sequence (Hillier et al. 2003). One aspect deals with the impact of human variation on the assembly of the human sequence and the extremes of variation observed in clone overlaps. These more highly variant regions complicated assembly of the genome, since such overlaps actually might have represented distinct regions of the genome that had resulted from segmental duplication, rather than the same region of the genome. With appropriate tests, they have been clearly shown to derive from a single region of the genome and thus represent curiously highly variant regions of the genome. Another aspect deals with efforts to improve the set of protein-coding gene predictions, the first and most critical element of the parts list. Gene-finding in mammalian genomes requires a combined approach, employing gene prediction programs, experimental evidence, and comparative sequence analysis. The best sources of experimental evidence currently are the RefSeq (Pruitt and Maglott 2001) and MGC (Strausberg et al. 1999) full-length cDNA sequence collections. However, in carefully aligning these sequences against the genome, we have noted discrepancies between the cDNA and genomic sequences that alter the reading frame. In investigating the source of these differences, we have discovered instances of sequence variants in the population that must alter the protein product of the gene. We have also used the mouse sequence for gene prediction in the human genome as a means of greatly reducing the problem of pseudogene contamination and other false-positive predictions in the predicted set.

ASSEMBLY AND VARIATION

Assembling the human genome sequence from individual BAC clones presented challenges not faced in assembling simpler genomes like those of yeast (Johnston et al. 1997; Mewes et al. 1997) and *C. elegans* (Consortium 1998). The extensive amount of repeated sequence, both from interspersed repeats (44% of the genome) and from segmental duplications (another 5% of the genome) (Lander et al. 2001), means that different BAC clones may contain similar sequence, but derive from entirely different regions of the genome. On the other hand, clones may derive from the same region of the genome, but because they derive from different genomes, they may differ in sequence. About 1 in 1300 bases is expected to differ between any two copies of the genome (Sachidanandam et al. 2001), and the sequenced clones came from multiple, different diploid individuals (with 70% coming from a single individual). Even within clones from a single library, only half the overlaps are expected to derive from the same chromosomal copy or haplotype. Thus, in judging whether two clones overlap on the basis of sequence, there is inevitably a balance to be found between accepting clones that truly should overlap and rejecting clones that derive from two similar but distinct regions of the genome.

Fortunately, the physical map largely mitigates the problem (McPherson et al. 2001). Here, the large size of the BAC clones (average insert length = 150–175 kb) and their extensive overlaps (>90% on average) allow all but the largest, most recent duplications to be sorted out. Even nearly identical repeats of greater than the BAC insert length may be recognized by the overabundance of clone coverage in a region and targeted for special attention. As a further check, each overlap can be examined at the sequence level. For Chromosomes 2, 4, and 7, we required that the overlap extend at least 2 kb through the ends of the clones with at least 99.8% sequence identity, with similar criteria applied for other chromosomes. Overlaps not meeting these criteria were subjected to additional scrutiny.

In extreme cases, such as the Y chromosome (Tilford et al. 2001; Skaletsky et al. 2003) and the Williams-Beuren Syndrome region (Hillier et al. 2003), the region could not be sorted out by the physical map alone, and its resolution required sequence information from clones from the same haplotype with extensive overlap. Using these procedures on Chromosome 7 allowed us to detect 8.2% of the sequence as segmentally duplicated, 7.0% within the chromosome and 2.2% between 7 and other chromosomes (Hillier et al. 2003).

In the course of the clone assembly, we encountered examples where the sequence variation between two clones was unusually high, and yet the physical map supported the join. To determine whether these clones de-

rived from the same region or instead arose from unrecognized large, highly similar duplications, we used PCR to recover a segment of the variant region from both the parent clones and the genome from 24 individuals, a subset of the DNA Polymorphism Discovery Resource (Collins et al. 1998). We reasoned that if the sequences derived from the same region, the two variants would be allelic and would segregate in the population, with heterozygotes and homozygotes of each allele in proportion to its frequency in the population. If, however, the sequences derived from distinct copies of a repeated sequence, both sequences would be present in all individuals. In rare instances, the same site might be variant within both copies; even here we would find unusual ratios of the two sequences in the population samples.

In toto for Chromosome 7 we analyzed 39 overlaps with a difference rate of at least 3 events per kilobase. These overlaps totaled 1911 kb, with an average variation of 4.5 differences per kilobase. In all cases, the variant bases were found to be polymorphic in the population sample, with homozygotes evident for at least one form. The proportion of heterozygotes to homozygotes was generally in accord with Hardy-Weinberg equilibrium. Furthermore, when several variants were present in the region assayed, they often behaved as a haplotype block; that is, the presence of a particular base in one variant position was predictive of the other variant bases assayed. Thus, these overlaps all appeared to sample highly variant regions of the genome.

To investigate more broadly the frequency and extent of such highly variant regions, we examined the pattern of variation within 2,718 overlaps between clones on Chromosomes 2, 4, and 7. For each overlap of at least 5 kb, we looked at the variation within 5-kb nonoverlapping windows. On average, just fewer than 4 differences in each window would be expected if the clones derived from different haplotypes. Although many segments conformed to expectation, we saw 302 intervals with more than 18 differences (2 standard deviations from the mean). Often we found multiple successive 5-kb segments with high variation (Fig. 1), as might be expected from a series of segments exhibiting linkage disequilibrium (Daly et al. 2001). To determine whether the set of variations arise from two distinct copies of a single haplotype block will require more experimental data across the region.

These investigations support two conclusions. First, the finding that none of the tested regions of high variability represented segmental duplications, but rather highly variable segments of the genome, suggests that

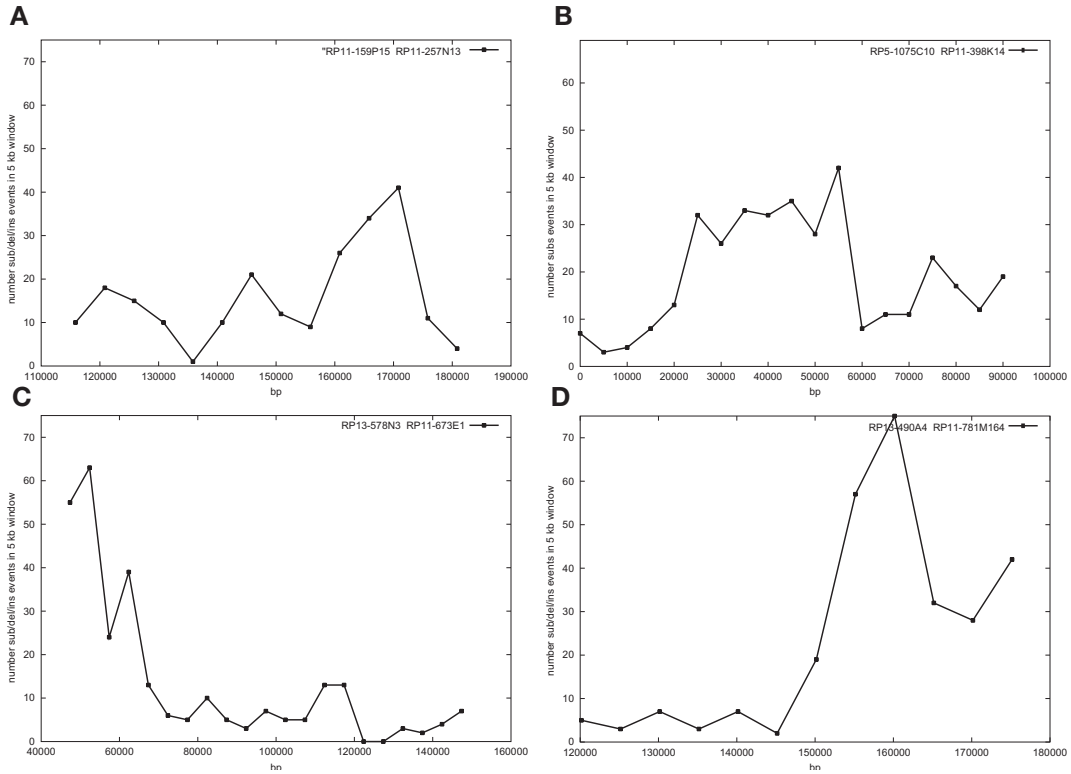

Figure 1. Regions of high variation in the overlap between adjacent clones. Four different overlaps are depicted. Each point represents the number of substitution events in 5-kb nonoverlapping windows, including insertion/deletion events for 3 of them. The full extent of the overlap is shown. The expected number of events per 5 kb between two clones of different origin would be about 4, and the mean number of events plus 2 standard deviations for a 5-kb window is 18 for this set (15 for substitutions alone). For much of the overlap the variation is within the normal limits, but in each case several successive 5-kb segments are present, with one case extending over about 40 kb. The region of high variation in *A* lies within a large intron of the LRP1B gene (low-density lipoprotein receptor-related protein 1B) and has several regions conserved in mouse and rat. The region of high variation in *B* shows no genes but does contain several regions conserved with mouse or rat. The region in *C* corresponds to the 5′ half of the GYPE gene (a surface glycophorin on red cells related to glycophorin A and B genes that specify the MN and Ss blood group antigens). The region in *D* contains no feature and exhibits no regions conserved with mouse and rat.

there are unlikely to be many large, recently duplicated segments of the genome yet to be identified in regions where the above criteria have been used to judge overlaps. Second, there are multiple regions of the genome where there is locally high variation, sometimes extending over tens of kilobases. Whether these are present because of balanced selection or because they represent by chance the tail end of the distribution from the last coalescent is uncertain (Charlesworth et al. 1997).

We have examined the regions for features that might suggest a basis for such balancing selection. Fifteen were in the region of some gene, including the known genes CRYPTIC, TSSC1, LRP1B, and GYPE. However, others contain no known features. Regardless of the basis, if variation has been accumulating at neutral rates, these regions have been maintained independently in the population since well before the founding of the species and, in some cases, almost equal to the time of the chimp–human divergence.

GENES

Finding genes in the draft human genome sequence was challenging, and the results were often inconsistent (Hogenesch et al. 2001). Even with finished, highly accurate genomic sequence, the extensive amount of noncoding sequence and the abundant pseudogenes present a significant challenge to gene prediction. The concerted efforts to obtain full-length sequences for cDNAs have been invaluable in annotation (Strausberg et al. 2002), providing direct evidence for an increasing number of genes. However, as yet, the collections for human remain incomplete and contain errors. Furthermore, the assignment of the sequence to its source in the genome may be complicated by the presence of close paralogs, recently formed pseudogenes, and polymorphisms. Another potentially powerful complementary approach comes from comparative sequence analysis, where conserved features such as genes stand out against neutrally evolving sequence. The recent release of a draft mouse sequence (Waterston et al. 2002) provides an opportunity to apply this approach genome-wide to human. As the first of many other mammalian sequences that will become available in the next few years, it provides a taste of the power to come from comparative analysis.

In annotating the finished sequence of Chromosome 7, we combined cDNAs, gene models from gene prediction programs, and comparison with the mouse sequence to derive a gene set that accurately reflects available experimental evidence and offers improved discrimination between genes and pseudogenes. In the course of these studies, we uncovered genes with apparently active and inactive forms present in the population. Such variants might be usefully exploited to learn more about the role of the gene in human biology.

KNOWN GENES: cDNAs

We began our efforts to define the genes on Chromosome 7 with the extensive RefSeq (Pruitt and Maglott 2001) and the MGC (Strausberg et al. 2002) cDNA collections. We aligned all 14,769 RefSeq and 10,047 MGC sequences against the entire genome sequence, allowing each cDNA to confirm only a single genomic locus. Spliced forms were favored over single or minimally spliced forms to avoid processed pseudogenes, and only the single match with the highest percent identity was kept to avoid confusion with recent paralogs.

In aligning these cDNAs against the genome, we noticed certain discrepancies that complicated interpretation. Alignment of the cDNAs to the genome, although often straightforward, was erroneous in many cases using any of the several programs. In our experience, Spidey (Wheelan et al. 2001) gave the most reliable results, aligning about 65% of the cDNAs and 79% of their exons accurately against the genomic sequence. But for any given gene, BLAT (Kent 2002) or EST_GENOME (Mott 1997) may give a better alignment (where the quality of the alignment is judged by minimizing base substitutions between the cDNA and genomic sequence and the avoidance of noncanonical splice sites). Every case was manually reviewed, and in some cases, manual review uncovered better alignments than did any of the available programs.

Even after exhaustive efforts to obtain an optimal alignment, some problems remained. For example, 23 mRNAs had no similarity to any mouse gene, and the translation product had no similarity to any known protein. Although these could be true genes, it seems more likely that they represent untranslated segments of bona fide genes. Nonetheless, these were kept in the current gene set. Eight others contained only a single, very short open reading frame (<20 amino acids), where again the translation product had no similarity to any known protein. Inspection of the genomic region generally showed that these instances were immediately downstream of another gene. We suspect these represent incomplete cDNAs from alternative 3′ exons, and we excluded these genes from the Chromosome 7 gene list.

Other genes contained differences between the cDNA and the aligned genomic sequence as expected, since the cDNA and genomic sequences derived from different individuals. Whereas most of these left the gene intact, at most changing a codon or two, some disrupted the reading frame of either the cDNA or the genomic sequence, either by changing the frame of the translation or by directly introducing a stop codon. For Chromosome 7, this amounted to some 60 genes (10% of the total genomic sites to which cDNAs aligned). To investigate which of the two sequences (the cDNA and the genomic) was correct, or if indeed both versions of the sequence are commonly represented in the population, we attempted to resequence the region in question from the original clone and related clones and also in a panel of 24 diverse individual DNAs from the DNA Polymorphism Discovery Resource (Collins et al. 1998). We also compared the sequence to the orthologous region from the mouse where available.

Because of repeated sequence and other technical issues, a few regions failed to give results. We found 35 of the cDNA/genomic pairs had a likely error in the cDNA sequence; that is, sequence from all tested individual

DNAs and the clones agreed with the original genomic sequence. Of these, 16 had multiple base insertion/deletion differences such that the reading frame between the two sequences was eventually restored. In such cases, comparison with the mouse sequence revealed that in each instance the reading frame of the genomic sequence yielded the amino acid sequence better conserved in the mouse translation. The other 19 had simple insertion/deletion or missense differences that shifted or truncated the reading frame in the cDNA relative to the genomic sequence. Again, translation of the orthologous mouse sequence supported the genomic reading frame.

We found eight instances where the genomic sequence was likely in error. For five of these, reexamination and/or resequencing revealed an error in the original BAC sequence, either in sequencing or during propagation of the clone. For an additional three, resequencing of the BAC agreed with the original genomic sequence, but all of the individual DNAs agreed with the cDNA sequence. No other clones from the same library were available in these cases to determine whether the difference might be an individual polymorphism.

Surprisingly, in three instances, we found the panel of 24 to contain both versions of the sequence; that is, the human population is polymorphic at the site. These cases involved different kinds of changes. In one, a single base deletion in the genomic sequence produced a frameshift in exon 31 of more than 40 exons in the gene encoding zonadhesin, a protein found on the surface of the sperm head (Fig. 2) (Hardy and Garbers 1995). In a second case, a premature stop codon (codon 60) was found in the genomic sequence where there is a glutamine in the cDNA for transmembrane protein induced by tumor necrosis factor (TMPIT, NM_031925) (Table 3). The consequence, if any, of these disrupted or altered translations on human biology will require further investigation.

The accurate alignment of 1,073 RefSeq and MGC cDNA sequences onto Chromosome 7 has provided a clear evidence for 605 genes. The alignment of the cDNAs against the finished genomic sequence also provides a valuable means of improving the cDNA resources. A few seem unlikely to represent independent or functional genes. Probable frameshift errors or other chain-terminating changes in genes have been recognized and investigated. The compensated frameshift errors found here have led to incorrect translation products and masked otherwise conserved regions of the protein. Finally, the careful comparison of cDNA and genomic sequence has led to the discovery of some variants that disrupt the reading frame. These genes may be in transition to becoming pseudogenes, or perhaps both forms of the protein may be of selective benefit in certain circumstances, leading to their longer-term persistence.

PREDICTED GENES

With a solid set of known genes established for Chromosome 7, we set out to find additional likely protein-coding genes. We wanted the set to be comprehensive, yet as free as possible of false predictions, particularly pseudogenes, which can be significant contaminants of predicted gene sets. For example, in initial analysis of the mouse genome, almost 20% of the identified genes were estimated to be pseudogenes. In one dramatic case, the mouse genome contains just one functional copy of the GAPDH gene and more than 400 related sequences. Of these, 118 were found contaminating the predicted gene set.

Pseudogenes are of two principal types, arising by distinct mechanisms. Processed pseudogenes result from the reverse transcription of mRNA and the subsequent insertion of the copy into the genome. Unprocessed pseudogenes arise from the duplication of segments of the genome that include functional genes or from the degradation of genes that are no longer subject to selection. Pseudogenes of either type may be complete or partial and, over time, will accumulate mutations that make it obvious they are inactive. Deletions, insertions, and other rearrangements can also obscure their origin.

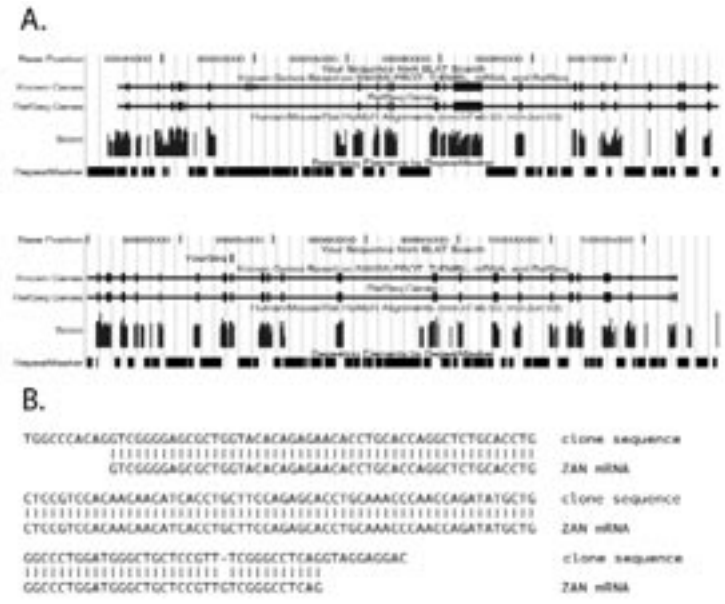

Figure 2. The gene encoding zonadhesin contains a variable base that alters the reading frame. (*A*) The gene structure as displayed on the UCSC browser (labeled "Known Genes" and "RefSeq Genes"), along with tracks showing the location along Chromosome 7, the position of the genomic region containing the frameshifting variation (labeled "Your sequence"), the regions conserved with mouse and rat, and the repeat content of the sequence. (*B*) Alignment of the genomic sequence and exon 31 of the zonadhesin cDNA, demonstrating the single-base deletion in the genomic sequence relative to the cDNA.

Table 3. Genes with Disruptive Variants

Gene ID RefSeq ID	Gene name	cDNA sequence	Genomic sequence	Codon	BACS gs/cdna	DNA PDR gs/het/cDNA
gi13994299 NM_031925	TMPIT	cttCAGaac L Q N	CttTAGacc L * N	42	3:3	0:0:24
gi15706488 BC012777	hypothetical	gaatgCGAgcc M R A	gaatgTGAgcc M * A	2	nd	2:11:9
gi16554448 NM_003386	zonadhesin	ccgtTGTcgg R C R	CcgtT*Ttcgg R F G	1,922	nd	9:7:7

To avoid pseudogenes and other false positives in the Chromosome 7 predicted gene set, we developed a protocol that exploited newly available draft genome sequence of the mouse. We reasoned that recently derived pseudogenes are problematic because they most closely resemble functional genes, whereas older pseudogenes have degraded through neutral drift and are less likely to be confused with genes. With 75 million years separating mouse and human from their last common ancestor, only pseudogenes that arose after the split would present problems. Since processed pseudogenes in general do not insert near their source gene, any processed pseudogene that arose in one organism thus would not have a counterpart in the corresponding portion of the other genome. (For convenience, we call regions in the mouse and human genomes that descended from a common ancestor orthologous, similar to the usage that has been adopted for genes.) Thus, for regions where the orthologous relationships have been established between the two genomes (and this covers 90% of the human genome), genes are likely to have counterparts in both organisms, whereas processed pseudogenes will not. Furthermore, having arisen since the divergence of the two species, the closest relative of the pseudogene will lie within the same genome, rather than in the second genome.

The situation is less clear with pseudogenes arising from duplication. Duplications may be at a distance, in which case orthology may be a useful discriminator, but often they are near and even tandemly arranged with respect to the source gene. In some cases, both copies may be functional members of a gene family, with one or both copies adopting new functions. Nonetheless, those copies that become pseudogenes may be recognized, because the copy may be incomplete or may have drifted significantly from the original, often with frameshift mutations or other chain-termination mutations disrupting the reading frame. The functional gene will remain most similar to the orthologous gene in the second organism.

With this screening procedure in mind, we set about to obtain a comprehensive set of putative genes, reasoning that false positives could be recognized and removed by careful examination of the orthologous regions between mouse and human. We used three different gene prediction programs, FGENESH2 (Solovyev 2001), TwinScan (Korf et al. 2001), and GeneWise (Birney and Durbin 2000), to recognize as many elements as possible. The first two use comparative sequence information (without using orthology) to modify ab initio predictions and improve accuracy, whereas GeneWise uses protein homologies to seed predictions. We used the mouse sequence as the informant sequence for TwinScan and FGENESH2 and all available protein predictions for GeneWise. TwinScan produced 1,350 gene models, FGENESH2 3,793 models, and GeneWise 22,326 models. (GeneWise predicts multiple, alternatively spliced models in any given region.) The combined output included 90% of all known exons and at least one exon from 98% of the known genes.

Although the combined output from the three programs is thus reasonably sensitive, there are undoubtedly a large number of false-positive predictions, including many pseudogenes. To eliminate most of these unwanted predictions, we used each prediction to search the orthologous region of the mouse genome for a matching gene. In turn, the matching mouse genes were used to search for matches in the orthologous region of the human genome. The matches in the orthologous regions had to be the best or close to the best in the entire genome. Furthermore, single-exon genes were removed if they had matches to multi-exon genes in either genome. Generally, at any one site only the best reciprocally matching pair was kept (near-best for local duplications). This left us with 728 FGENESH2, 284 TwinScan, and 400 GeneWise predictions.

To eliminate redundancy between the various gene models, for each region we accepted only one prediction, taking in order known genes, FGENESH2, TwinScan, and GeneWise (the order was based on the performance of the three programs against known genes) and giving priority to models with the best reciprocal matches. Models that showed signs of nonfunctionality (early truncation compared to closely related genes, absence of introns in gene models with closely related genes with exons) and high homology to L1/reverse transcriptase were also removed. This yielded an additional 545 predicted genes.

We examined this gene set in several ways to determine how complete and accurate it might be. On average, the predicted genes, when compared to the known genes, have fewer exons (6.7 vs. 9.5), have fewer coding bases (1,231 vs. 1,457), and span less of the genome (28.5 kb vs. 61.4 kb), suggesting that the predicted genes either lack terminal exons or include some fragmented genes that reduce the average. A high fraction of both known genes (92%) and predicted genes (95%) have reciprocal best matches in the mouse genome at the orthologous position, whereas the others have near-best matches. This is

expected, given the methods used to build the set, but nonetheless their conservation strongly supports their validity.

Since we had not used ESTs in building the gene set, we used them as an independent measure of the representation of the set. We found 41,399 spliced ESTs with their best match to Chromosome 7. Of these, 93% at least partially overlapped an exon of the gene set, and an additional 1% lay near or within existing genes, suggesting that they might represent alternative splice forms or exons missing from the predicted genes. The remainder lacked significant open reading frames, and none satisfied the reciprocal match criteria used in making the gene predictions. Only 5% of the remainder had any match to mouse sequence. These unmatched spliced ESTs may nonetheless represent missed genes, although at present there is little corroborating evidence that they derive from protein-coding genes.

PSEUDOGENES

We attempted to identify pseudogenes directly, adapting a method used previously (Waterston et al. 2002; Zdobnov et al. 2002). We inspected all the intervals between known and predicted genes for sequence that yielded translation products with similarity to known proteins. Altogether we identified 941 such regions. More than one pseudogene may lie in any one interval, and old, largely degraded pseudogenes would be missed using the thresholds used here. As a result, we probably have undercounted pseudogenes.

We then evaluated the validity of our classifications to determine how often likely pseudogenes were included in the gene set and how many excluded genes were later found in the pseudogene set. We reasoned that genes should largely be under purifying selection, and pseudogenes should be subject to neutral drift. These differences in evolutionary pressures would produce differences in the ratio of synonymous vs. nonsynonymous substitutions (K_a/K_s ratio) in the coding portion of the genes or pseudogenes (Ohta and Ina 1995). Positive selection acting on genes will increase the K_a/K_s ratio, but generally the positive selection is limited to specific domains. Only rarely will positive selection act so broadly across a gene as to elevate the K_a/K_s ratio to near or above that of neutrally evolving sequence.

Of the 941 regions identified as containing likely pseudogenes, nearly all (97% ± 3%) had K_a/K_s ratios consistent with neutrally evolving sequence, supporting our classification. As with the mouse genome analysis, a significant fraction of the predicted pseudogenes (33%) had as yet no disruption to the reading frame. Virtually all the predicted pseudogenes could be aligned to another region of the human genome with higher sequence identity than to any region of the mouse genome, consistent with an origin after the mouse–human divergence.

We also attempted to classify the pseudogenes by origin, by using the orthologous mouse region for related sequence. For 88% (573/654) of the identified pseudogenes, no related sequence in the orthologous mouse region was found; these are likely to represent processed pseudogenes and are broadly distributed through the chromosome. Another 12% (81/654) did have related sequence in the orthologous mouse region, suggesting they were derived by segmental duplication. Indeed, these lie predominantly in segmentally duplicated regions of the chromosome.

We carried out the same analysis on the 1,152 members of the gene set. Only 5% ± 3% had a ratio consistent with neutral selection, suggesting the set is relatively free of pseudogenes. The total of 1,152 genes is a relatively modest number. Extrapolating to the genome, this would suggest that the human genome contains some 25,000 genes. The total number of genes is only slightly more than the number of pseudogenes found and is about 40% less than the number of genes predicted in another analysis of Chromosome 7 sequence (Scherer et al. 2003). Our approach has been deliberately conservative, but several points of our analysis suggest that our count is fairly accurate. By K_a/K_s analysis, only a few (0–60) of the pseudogenes are likely to be functional. The gene set covers the vast majority of the ESTs that have their best match to Chromosome 7, and those few that fall outside the gene set do not seem likely to be protein-coding. Perhaps much of the difference between our estimate and that of others lies in our treatment of pseudogenes.

CONCLUSION

Our initial analysis of the content of human Chromosome 7 illustrates some of the challenges that lie ahead, even with an accurate, complete sequence. It also suggests some avenues available for understanding the genome.

An immediate goal must be defining the parts list, that is, all the functional elements of the genome. At present, obtaining even the protein-coding gene set remains a difficult and complex task. The available experimentally determined cDNA sequences remain incomplete, and both these and, to a lesser extent, the genome sequence contain errors. Alignment of the cDNA sequences to the genome is not always straightforward and is complicated by polymorphism. Gene prediction programs give only partial answers and are often confounded by the abundant pseudogenes in the human genome. Comparative sequence analysis, using the mouse as the informative sequence, improves the accuracy of exon prediction in new programs such as TwinScan and FGENESH2. Further processing of the results exploiting the conserved synteny between mouse and human to establish orthologous relationships helps substantially in distinguishing genes from pseudogenes. As the sequences of additional mammalian genomes become available over the next few years, the description of the gene set will become increasingly accurate and complete. These additional sequences will also facilitate the identification of other functional sequences, such as those for noncoding RNAs and regulation of gene expression. This pathway, combined with ongoing experimental testing and validation, holds the promise of a relatively complete parts list in just a few years' time.

Understanding how those parts function and how they contribute to human disease when they fail to function

normally will take much longer. Studies of homologous elements in experimentally tractable animals will be critical in this analysis, along with studies of human genes in tissue culture and in vitro. Ultimately, we will need to understand these elements in the context of the human. Natural variation in the human population provides a powerful tool for achieving this understanding. Thus, a key challenge in the coming years will be to define the variation in the human population, focusing first on common variations and then, as methods improve, to extend this to less common variations. In turn the impact of these variations, if any, on the human phenotype and in particular on health and disease must be established, leading to an understanding of the role of each element in the whole.

The complete sequence of the human genome is thus only a beginning. Our efforts on Chromosome 7 were, of course, focused on obtaining this reference sequence, but variation was encountered in the course of the analysis, and our results suggest two avenues of exploration.

In the comparison of the cDNAs to genomic sequence, we encountered examples of variations with predictable effects on gene function; that is, alterations to the reading frame that are expected to result in loss of function. This strong prediction is testable, and examination of the impact of such loss-of-function variations on the human phenotype would undoubtedly be highly informative about the normal role of the gene, just as loss-of-function mutations have been important in the genetic dissection of function in experimental animals.

The number of genes with such variations was small, but came at almost no extra cost and with the comparison of only two different sequences. As additional human genome sequences are described, other examples will come to light. Furthermore, for many genes, missense substitutions can have an equally predictable effect on protein function, expanding the range of changes that can be used. Comparative sequence analysis will reveal those residues under purifying selection, adding still more candidates. As resequencing of human genomes becomes routine, "acquisition by genotype," as this approach might be called, will become more widespread, complementing the more traditional approach of "acquisition by phenotype."

The regions with unusually high density of variation may point to functionally important regions of the genome. Differences in SNP density have been noted previously, with low density noted in some regions (Miller and Kwok 2001; Miller et al. 2001) and high density in HLA region (Mungall et al. 2003) and around the gene underlying the ABO blood group antigen (Yip 2002). The SNP "deserts" may reflect the relatively recent fixation of a single variant in the population, either through selective mechanisms or through founder effects and chance. The regions of high SNP density around HLA and the ABO gene are thought to represent instances of balanced selection, where multiple versions of the region have been maintained in the population over millions of years for HLA and hundreds of thousands of years for the ABO locus.

The regions described here have a lower density of variation than observed in the HLA region and are more similar in density to that observed in the ABO locus. Their size (tens of kilobases) is similar to that of haplotype blocks described elsewhere in the genome (Gabriel et al. 2002), but whether their boundaries correlate with the boundaries of haplotype blocks remains to be determined. Of course, such regions might simply represent the chance persistence of two different haplotype blocks over extensive evolutionary time. However, if balancing selection is responsible for the maintenance of these regions for the hundreds of thousands to millions of years required to accumulate this density of variation, understanding their role in the generation of the variable human phenotype will be important.

Both the enumeration of a parts list and a description of human variation require the high-quality, essentially complete reference human genome sequence that is now available to all without restriction. Distinguishing genes from pseudogenes was an almost impossible task with the draft human sequence and without a high-quality mouse sequence for comparison. The variants that alter the reading of genes are sufficiently unusual that errors in sequence would have obscured them in earlier versions, and even with a high-quality human genome, errors in the other data sets predominated. Recognition of areas of high SNP density requires accurate assembly, so that segmentally duplicated regions are not mistaken for such regions.

The completion of the human genome sequence marks a major milestone in science and comes just 50 years after the discovery of the structure of DNA. For the first time, we as a species have before us the genetic instruction set that molds us. Knowledge of the sequence presents the scientific community with the challenge to understand its content. The human genome sequence also presents society with enormous challenges, not the least of which is to use the knowledge for the betterment of humankind. Undoubtedly, the tasks will take longer than some have forecast, but the potential is clear and the implications profound.

ACKNOWLEDGMENTS

We thank the dedicated individuals at the Washington University Genome Sequencing Center and other centers across the world for their contributions throughout this project. The work was supported through grants from the National Human Genome Research Institute.

REFERENCES

Ansorge W., Sproat B., Stegemann J., Schwager C., and Zenke M. 1987. Automated DNA sequencing: Ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic Acids Res.* **15:** 4593.

Axelrod J.D. and Majors J. 1989. An improved method for photofootprinting yeast genes in vivo using Taq polymerase. *Nucleic Acids Res.* **17:** 171.

Baer R., Bankier A.T., Biggin M.D., Deininger P.L., Farrell P.J., Gibson T.J., Hatfull G., Hudson G.S., Satchwell S.C., and Seguin C., et al. 1984. DNA sequence and expression of the B95-8 Epstein-Barr virus genome. *Nature* **310:** 207.

Birney E. and R. Durbin R. 2000. Using GeneWise in the

Drosophila annotation experiment. *Genome Res.* **10:** 547.
Brumbaugh J.A., Middendorf L.R., Grone D.L., and Ruth J.L. 1988. Continuous, on-line DNA sequencing using oligodeoxynucleotide primers with multiple fluorophores. *Proc. Natl. Acad. Sci.* **85:** 5610.
Charlesworth B., Nordborg M., and Charlesworth D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet. Res.* **70:** 155.
Collins F.S., Brooks L.D., and Chakravarti A. 1998. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8:** 1229.
Consortium (The *C. elegans* Sequencing Consortium). 1998. Genome sequence of the nematode *C. elegans:* A platform for investigating biology. *Science* **282:** 2012.
Coulson A., Sulston J., Brenner S., and Karn J. 1986. Towards a physical map of the genome of the nematode *Caenorhabditis elegans. Proc. Natl. Acad. Sci.* **83:** 7831.
Craxton M. 1993. Cosmid sequencing. *Methods Mol. Biol.* **23:** 149.
Daly M.J., Rioux J.D., Schaffner S.F., Hudson T.J., and Lander E.S. 2001. High-resolution haplotype structure in the human genome. *Nat. Genet.* **29:** 229.
DeLisi C. 1988. The Human Genome Project. *Am. Sci.* **76:** 488.
Deloukas P., Matthews L.H., Ashurst J., Burton J., Gilbert J.G., Jones M., Stavrides G., Almeida J.P., Babbage A.K., Bagguley C.L., Bailey J., Barlow K.F., Bates K.N., Beard L.M., Beare D.M., Beasley O.P., Bird C.P., Blakey S.E., Bridgeman A.M., Brown A.J., Buck D., Burrill W., Butler A.P., Carder C., and Carter N.P., et al. 2001. The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414:** 865.
Dulbecco R. 1986. A turning point in cancer research: Sequencing the human genome. *Science* **231:** 1055.
Dunham I., Shimizu N., Roe B.A., Chissoe S., Hunt A.R., Collins J.E., Bruskiewich R., Beare D.M., Clamp M., Smink L.J., Ainscough R., Almeida J.P., Babbage A., Bagguley C., Bailey J., Barlow K., Bates K.N., Beasley O., Bird C.P., Blakey S., Bridgeman A.M., Buck D., Burgess J., Burrill W.D., and O'Brien K.P., et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489.
Ewing B., Hillier L., Wendl M.C., and Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175.
FlyBase Consortium. 1994. FlyBase—The *Drosophila* database (The FlyBase Consortium). *Nucleic Acids Res.* **22:** 3456.
Gabriel S.B., Schaffner S.F., Nguyen H., Moore J.M., Roy J., Blumenstiel B., Higgins J., DeFelice M., Lochner A., Faggart M., Liu-Cordero S.N., Rotimi C., Adeyemo A., Cooper R., Ward R., Lander E.S., Daly M.J., and Altshuler D. 2002. The structure of haplotype blocks in the human genome. *Science* **296:** 2225.
Hardy D.M. and Garbers D.L. 1995. A sperm membrane protein that binds in a species-specific manner to the egg extracellular matrix is homologous to von Willebrand factor. *J. Biol. Chem.* **270:** 26025.
Hawkins T.L., O'Connor-Morin T., Roy A., and Santillan C. 1994. DNA purification and isolation using a solid-phase. *Nucleic Acids Res.* **22:** 4543.
Heilig R., Eckenberg R., Petit J.L., Fonknechten N., Da Silva C., Cattolico L., Levy M., Barbe V., de Berardinis V., Ureta-Vidal A., Pelletier E., Vico V., Anthouard V., Rowen L., Madan A., Qin S., Sun H., Du H., Pepin K., Artiguenave F., Robert C., Cruaud C., Bruls T., Jaillon O., and Friedlander L., et al. 2003. The DNA sequence and analysis of human chromosome 14. *Nature* **421:** 601.
Hillier L.W., Fulton R.S., Fulton L.A., Graves T.A., Pepin K.H., Wagner-McPherson C., Layman D., Maas J., Jaeger S., Walker R., Wylie K., Sekhon M., Becker M.C., O'Laughlin M.D., Schaller M.E., Fewell G.A., Delehaunty K.D., Miner T.L., Nash W.E., Cordes M., Du H., Sun H., Edwards J., Bradshaw-Cordum H., and Ali J., et al. 2003. The DNA sequence of human chromosome 7. *Nature* **424:** 157.
Hogenesch J.B., Ching K.A., Batalov S., Su A.I., Walker J.R., Zhou Y., Kay S.A., Schultz P.G., and Cooke M.P. 2001. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106:** 413.
Johnston M., Hillier L., Riles L., Albermann K., Andre B., Ansorge W., Benes V., Bruckner M., Delius H., Dubois E., Dusterhoft A., Entian K.D., Floeth M., Goffeau A., Hebling U., Heumann K., Heuss-Neitzel D., Hilbert H., Hilger F., Kleine K., Kotter P., Louis E.J., Messenguy F., Mewes H.W., and Hoheisel J.D., et al. 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XII. *Nature* **387:** 87.
Ju J., Ruan C., Fuller C.W., Glazer A.N., and Mathies R.A. 1995. Fluorescence energy transfer dye-labeled primers for DNA sequencing and analysis. *Proc. Natl. Acad. Sci.* **92:** 4347.
Kent W.J. 2002. BLAT—The BLAST-like alignment tool. *Genome Res.* **12:** 656.
Korf I., Flicek P., Duan D, and Brent M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* (suppl. 1) **17:** S140.
Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L., Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov J.P., Miranda C., Morris W., and Naylor J., et al. (International Human Genome Sequencing Consortium). 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860.
Lu H., Arriaga E., Chen D.Y., and Dovichi N.J. 1994. High-speed and high-accuracy DNA sequencing by capillary gel electrophoresis in a simple, low cost instrument. Two-color peak-height encoded sequencing at 40 degrees C. *J. Chromatogr. A* **680:** 497.
McPherson J.D., Marra M., Hillier L., Waterston R.H., Chinwalla A., Wallis J., Sekhon M., Wylie K., Mardis E.R., Wilson R.K., Fulton R., Kucaba T.A., Wagner-McPherson C., Barbazuk W.B., Gregory S.G., Humphray S.J., French L., Evans R.S., Bethel G., Whittaker A., Holden J.L., McCann O.T., Dunham A., Soderlund C., and Scott C.E., et al. (International Human Genome Mapping Consortium). 2001. A physical map of the human genome. *Nature* **409:** 934.
Mewes H.W., Albermann K., Bahr M., Frishman D., Gleissner A., Hani J., Heumann K., Kleine K., Maierl A., Oliver S.G., Pfeiffer F., and Zollner A. 1997. Overview of the yeast genome. *Nature* **387:** 7.
Miller R.D. and Kwok P.Y. 2001. The birth and death of human single-nucleotide polymorphisms: New experimental evidence and implications for human history and medicine. *Hum. Mol. Genet.* **10:** 2195.
Miller R.D., Taillon-Miller P., and Kwok P.Y. 2001. Regions of low single-nucleotide polymorphism incidence in human and orangutan xq: Deserts and recent coalescences. *Genomics* **71:** 78.
Mott R. 1997. EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* **13:** 477.
Mungall A.J., Palmer S.A., Sims S.K., Edwards C.A., Ashurst K.L., Wilming L., Jones M.C., Horton R., Hunt S.E., Scott C.E., Gilbert J.G., Clamp M.E., Bethel G., Milne S., Ainscough R., Almeida J.P., Ambrose K.D., Andrews T.D., Ashwell R.I., Babbage A.K., Bagguley C.L., Bailey J., Banerjee R., Barker D.J., and Barlow K.F., et al. 2003. The DNA sequence and analysis of human chromosome 6. *Nature* **425:** 805.
Ohta T. and Ina Y. 1995. Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences. *J. Mol. Evol.* **41:** 717.
Olson M.V., Dutchik J.E., Graham M.Y., Brodeur G.M., Helms C., Frank M., MacCollin M., Scheinman R., and Frank T. 1986. Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl. Acad. Sci.* **83:** 7826.
Prober J.M., Trainor G.L., Dam R.J., Hobbs F.W., Robertson C.W., Zagursky R.J., Cocuzza A.J., Jensen M.A., and Baumeister K. 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238:** 336.

Pruitt K.D. and Maglott D.R. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29:** 137.

Sachidanandam R., Weissman D., Schmidt S.C., Kakol J.M., Stein L.D., Marth G., Sherry S., Mullikin J.C., Mortimore B.J., Willey D.L., Hunt S.E., Cole C.G., Coggill P.C., Rice C.M., Ning Z., Rogers J., Bentley D.R., Kwok P.Y., Mardis E.R., Yeh R.T., Schultz B., Cook L., Davenport R., Dante M., and Fulton L., et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928.

Scherer S.W., Cheung J., MacDonald J.R., Osborne L.R., Nakabayashi K., Herbrick J.A., Carson A.R., Parker-Katiraee L., Skaug J., Khaja R., Zhang J., Hudek A.K., Li M., Haddad M., Duggan G.E., Fernandez B.A., Kanematsu E., Gentles S., Christopoulos C.C., Choufani S., Kwasnicka D., Zheng X.H., Lai Z., Nusskern D., and Zhang Q., et al. 2003. Human chromosome 7: DNA sequence and biology. *Science* **300:** 767.

Shizuya H., Birren B., Kim U.J., Mancino V., Slepak T., Y. Tachiiri Y., and Simon M. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl. Acad. Sci.* **89:** 8794.

Sinsheimer, R.L. 1989. The Santa Cruz Workshop, May 1985. *Genomics* **5:** 954.

Skaletsky H., Kuroda-Kawaguchi T., Minx P.J., Cordum H.S., Hillier L., Brown L.G., Repping S., Pyntikova T., Ali J., Bieri T., Chinwalla A., Delehaunty A., Delehaunty K., Du H., Fewell G., Fulton L., Fulton R., Graves T., Hou S.F., Latreille P., Leonard S., Mardis E., Maupin R., McPherson J., and Miner T., et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423:** 825.

Smith L.M., Fung S., Hunkapiller M.W., Hunkapiller T.J., and Hood L.E. 1985. The synthesis of oligonucleotides containing an aliphatic amino group at the 5′ terminus: Synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Res.* **13:** 2399.

Solovyev V.V. 2001. Statistical approaches in eukaryotic gene prediction. In *Handbook of statistical genetics* (ed. D.J. Balding et al.), p. 83. Wiley, New York.

Strausberg, R.L., E.A. Feingold, R.D. Klausner, and F.S. Collins. 1999. The mammalian gene collection. *Science* **286:** 455-7.

Strausberg R.L., Feingold E.A., Grouse L.H., Derge J.G., Klausner R.D., Collins F.S., Wagner L., Shenmen C.M., Schuler G.D., Altschul S.F., Zeeberg B., Buetow K.H., Schaefer C.F., Bhat N.K., Hopkins R.F., Jordan H., Moore T., Max S.I., Wang J., Hsieh F., Diatchenko L., Marusina K., Farmer A.A., Rubin G.M., and Hong L., et al. 2002. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci.* **99:** 16899.

Tabor S. and Richardson C.C. 1995. A single residue in DNA polymerases of the *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc. Natl. Acad. Sci.* **92:** 6339.

Tilford C.A., Kuroda-Kawaguchi T., Skaletsky H., Rozen S., Brown L.G., Rosenberg M., McPherson J.D., Wylie K., Sekhon M., Kucaba T.A., Waterston R.H., and Page D.C. 2001. A physical map of the human Y chromosome. *Nature* **409:** 943.

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A., Gocayne J.D., Amanatides P., Ballew R.M., Huson D.H., Wortman J.R., Zhang Q., Kodira C.D., Zheng X.H., Chen L., Skupski M., Subramanian G., Thomas P.D., Zhang J., Gabor Miklos G.L., and Nelson C., et al. 2001. The sequence of the human genome. *Science* **291:** 1304.

Waterston R.H., Lindblad-Toh K., Birney E., Rogers J., Abril J.F., Agarwal P., Agarwala R., Ainscough R., Alexandersson M., An P., Antonarakis S.E., Attwood J., Baertsch R., Bailey J., Barlow K., Beck S., Berry E., Birren B., Bloom T., Bork P., Botcherby M., Bray N., Brent M.R., Brown D.G., and Brown S.D., et al. (Mouse Genome Sequencing Consortium). 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520.

Wheelan S.J., Church D.M., and Ostell J.M. 2001. Spidey: A tool for mRNA-to-genomic alignments. *Genome Res.* **11:** 1952.

Yip S.P. 2002. Sequence variation at the human ABO locus. *Ann. Hum. Genet.* **66:** 1.

Zdobnov E.M., von Mering C., Letunic I., Torrents D., Suyama M., Copley R.R., Christophides G.K., Thomasova D., Holt R.A., Subramanian G.M., Mueller H.M., Dimopoulos G., Law J.H., Wells M.A., Birney E., Charlab R., Halpern A.L., Kokoza E., Kraft C.L., Lai Z., Lewis S., Louis C., Barillas-Mury C., Nusskern D., and G.M. Rubin, et al. 2002. Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster. Science* **298:** 149.