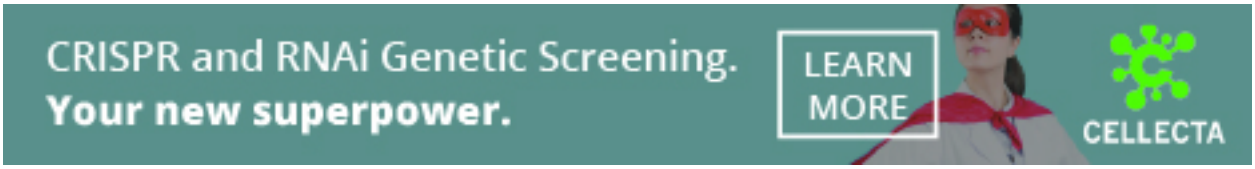
The Human Genome Sequence Expedition: Views from the "Base Camp"

Eric D. Green and Aravinda Chakravarti

Genome Res. 2001 11: 645-651
Access the most recent version at doi:10.1101/gr.188701



Commentary

The Human Genome Sequence Expedition: Views from the "Base Camp"

Eric D. Green[1,3] and Aravinda Chakravarti[2]

[1]*Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA;* [2]*McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21287, USA*

The past year has brought unprecedented public attention to biomedical research, with a particularly intense focus on the Human Genome Project and the completion of a first-generation ~3-billion-basepair human genome sequence. Much of this attention related to the competition between the two parallel, yet separate, efforts of the publicly-funded International Human Genome Sequencing Consortium and the private company Celera Genomics. Despite the apparent rancor between the groups, two celebratory events notably punctuate the past year: the joint media announcement in late June 2000 that both groups had generated a "working draft" sequence of the human genome and the two landmark scientific publications in February 2001 that describe the efforts of each project (International Human Genome Sequencing Consortium 2001; Venter et al. 2001).

Numerous grandiose clichés and metaphors have been used to convey the magnitude of these accomplishments and their associated implications for biomedical research and clinical medicine. Here we add one more to this list. Our choice for capturing the essence of contemporary human genome analysis is an analogy to a mountain climbing expedition, one where significant progress has been made to provide a spectacular view of the genetic landscape. But this is not an expedition that is complete, with uncertain—yet exciting—genomic terrain ahead. Indeed, the Human Genome Project is now firmly at the "base camp" of the expedition to elucidate the human genetic blueprint and to begin to understand its content. Nevertheless, this is a milestone of tremendous significance and excitement.

Here we outline some of the key lessons learned during the initial analysis of the human genome sequence. We highlight a few of the many remaining questions in understanding the genome's structure and function, with most answers likely becoming available later in the expedition. Finally, we preview the anticipated climb to the final summit and the ascent to a complete and finished sequence.

[3]Corresponding author.
E-MAIL egreen@nhgri.nih.gov; FAX (301) 402-4735.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.188701.

Detailed Reports from the Base Camp

Two papers (International Human Genome Sequencing Consortium 2001; Venter et al. 2001) report on the draft human genome sequence and provide the first detailed views from the base camp. The rigor and presentation of both papers is outstanding, and the respective groups should be commended for their scholarly contributions to the scientific literature. Of course, these papers and the related companion articles in the corresponding issues of *Nature* and *Science* represent the tip of a literature iceberg that will be revealed over time, as more complete sequence becomes available and other investigators develop more insightful ways of studying the genome's structure and function. Indeed, much of this can be followed daily on various Web sites that provide browser-based views of the human genome (e.g., see genome.cse.ucsc.edu, www.ensembl.org, www.ncbi.nlm.nih.gov/genome/guide/human). Even keeping track of the most informative Web sites can be challenging; toward that end, electronic "hubs" have been created that provide electronic pointers to the most relevant and, in some cases, new sites (e.g., see www.nhgri.nih.gov/genome_hub).

We make no attempt to summarize all of the analyses presented in the two major and other companion papers. Rather, we provide some clues about the major findings uncovered and actively encourage all readers to carefully read through these important publications.

Views from the Initial Ascent

The Human Gene Inventory

A very small fraction of the human genome sequence encodes protein. Perhaps a bit surprising to some, this figure is 1–2%. Generating a complete inventory of protein-coding sequences represents a high priority in the analysis of the human genome. Even a year ago, there remained significant debate about the total number of human genes, with estimates differing by threefold or more (Ewing and Green 2000; Crollius et al.

Generating the Working Draft Human Genome Sequence—By the Numbers[1]

15	Number of months for the Human Genome Project to increase the percentage of the human genome sequenced from 10% to 90%
1000	Approximate number of raw bases of DNA sequence generated per second (24 hours per day, 7 days per week) by the Human Genome Project sequencing centers during the above 15-month interval of intense data generation
9	Number of months for Celera Genomics to generate its whole-genome shotgun data for the human genome
23,147 and 7.5	Total megabases of raw DNA sequence generated from bacterial (mostly BAC) clones by the Human Genome Project in generating the working draft human sequence and the redundant coverage (on average) across the human genome this data provided
14,800 and 5.1	Total megabases of raw DNA sequence generated in a whole-genome shotgun fashion by Celera Genomics in generating the working draft human sequence and the redundant coverage (on average) across the human genome this data provided
2693 and 2910	Number of megabases of working draft human sequence generated by the Human Genome Project and Celera Genomics, respectively
249	Number of individuals given on the 'partial' list of authors for the International Human Genome Sequencing Consortium's (i.e., the Human Genome Project's) human sequence publication
274	Number of individuals listed on the Celera Genomics' human sequence publication

[1]Derived from information provided in the International Human Genome Sequencing Consortium (2001) and Venter et al. (2001).

2000; Liang et al. 2000). Guesses by genome scientists have differed to an even greater extent (see www.ensembl.org/Genesweep). The base camp provides a clear view of the human gene inventory, allowing for improved, albeit imprecise, estimates about the human gene number.

Gratifyingly, very similar conclusions about the total number of human genes were reached by the two sequencing groups (International Human Genome Sequencing Consortium 2001; Venter et al. 2001). In both cases, rigorous approaches were used to derive estimates and features about the human gene repertoire. Several salient features of the analyses deserve mention. First, it is now clear that the lower estimates of gene number are the correct ones, with the human genome containing ~30,000–40,000 genes. Of course, numerous pitfalls in the computational approaches used for gene counting may exist, and significant caution should be exercised when considering the seemingly precise estimates (Bork and Copley 2001). Second, the new estimates emphasize the existence of a great degree of alternate splicing of human transcripts, with perhaps >50% of genes producing more than a single mRNA species. Third, there is striking nonhomogeneity in the distribution of genes across the genome. For example, GC-rich regions tend to be densely packed with genes that typically contain smaller introns, while GC-poor regions tend to be relatively lacking in genes but these contain larger introns.

With refinement of the human gene inventory, interest is turning to the corresponding human protein collection (the proteome). A key notion that emerges from the initial analyses is that mammalian proteomic complexity seems to have been achieved as a consequence of multiple, incremental changes as opposed toindividual major upheavals. Thus, humans (and likely other vertebrates) seem to have gained a complex physiology by deriving a larger number of unique protein products from genes (through alternate splicing) and not through the development of numerous vertebrate-specific protein and protein-domain families. Indeed, the human genome sequence is a remarkable example of the combinatorial reuse of available, evolutionarily conserved protein domains in sophisticated ways to form yet more complex proteins. Thus, innovation in protein architecture, leading to a more complex proteome, has been created despite having only a slightly larger number of genes as compared to simpler organisms.

Tandem Versus Dispersed Gene Duplications

If new proteins have arisen from the old, then how did the new biological functions arise and evolve? Susumu Ohno (Ohno 1970) postulated that although the ancestral vertebrate genome had expanded by whole-genome duplication, new genes usually arise by tandem duplication with subsequent divergence of function. The draft human genome sequence has illuminated much of this history, demonstrating that the mechanisms for gene birth are quite varied. Three types of duplications are evident. First, tandem duplications, the key premise of Ohno's thesis based on knowledge in the 1960's, are uncommon. These duplications are created from sequence repeats in a local region that enable unequal crossing over. Although SINE and LINE repeats are common in the human genome, they do not appear to mediate these events with any frequency. Second, retrotransposition is the most common mechanism for gene birth. In these instances, processed mRNAs are retrotransposed, leading to intronless paralogs that are present at one or many new sites. Intriguingly, many of these genes have a role in the protein translational process. Other DNA-mediated retrotransposition events are also common, usually in-

volving partial gene sequences and thus creating pseudogenes at the new site. At least 2909 such cases of duplications have been identified, resulting anywhere from a few copies (e.g., globin genes) to >1000 copies (e.g., olfactory receptor genes). Third, segmental duplications, involving the transfer of entire genomic segments to a new site, with subsequent rearrangement of order and gene loss, have also been identified. The largest block involves the transfer of 33 genes from chromosome 2 to chromosome 14, supported by the recognition that the entire common segment maps to a single site in the mouse genome. Duplication is not an isolated event and can be modified by both the loss and gain of additional DNA sequences. A complete catalog of such duplicated genes will be important since the paralogous genes are often mutant for the same disease phenotype.

The invasion of new duplicated segments throughout the human genome has occurred over evolutionary time, and these times can be estimated from sequence comparisons of the paralogs. Deciphering the sequence of other vertebrate genomes will allow more detailed comparisons, providing important insight into how and when such duplicated sequences originated and how new functions evolved. For more on the evolution of gene duplications, see Hughes et al. (2001), Makałowski (2001), and Robinson-Rechavi et al. (2001).

Horizontal Gene Transfer

One major surprise from the human sequence (International Human Genome Sequencing Consortium 2001) is the presence of >200 proteins with significant similarity to bacterial proteins but no apparent homology to proteins from yeast, worm, fly, or other nonvertebrate eukaryotes. Most of these proteins are intracellular enzymes. The corresponding genes are not artifactual and have genomic counterparts confirmed by PCR testing. In some cases, other vertebrates were found to contain orthologous genes. The majority of these genes are present in numerous bacteria but limited to vertebrates among the eukaryotes. A likely possibility is that these genes entered the vertebrate lineage by horizontal transfer from bacteria, with the subsequent acquisition of intronic sequences. Among such horizontally-transferred genes are at least two that encode paralogs of monoamine oxidase, a well-studied enzyme that is a target of psychiatric drugs. The vertebrate-specific acquisition of bacterial genes raises a number of intriguing questions (e.g., when did such transfers occur? what is the mechanism for such transfer?) that can be addressed as additional vertebrate genomes are sequenced.

Nature of Non–Coding DNA

More than 98% of the human genome does not encode protein. Of this, just under one-half corresponds to interspersed repetitive DNA, reflecting remnants of transposons. The human genome contains the following major classes of interspersed repeats (in the indicated proportions of the total draft sequence): SINES (13%), LINES (20%), LTR retrotransposons (8%), and DNA transposons (3%) (International Human Genome Sequencing Consortium 2001). In addition to these interspersed repeats, the genome contains a number of other classes of repetitive elements, including processed pseudogenes, simple sequence repeats, segmental duplications, and tandemly repeated sequence blocks (such as those at centromeres and telomeres). When all repeats are considered, well over one-half of the human genome corresponds to repetitive sequences.

One of the most compelling lessons learned at the human genome base camp is that the use of the phrase "junk DNA" is näive. All genomic sequences are part of the molecular fossil record and studying these regions in humans and other mammals will provide unprecedented insight into genome evolution. For example, despite the plethora of transposon-derived repetitive sequences, there is evidence for a massive decline in transposon activity in the human lineage. This is in sharp contrast to the mouse genome, where transposon activity has not declined. LTR retrotransposons appear to remain active in the mouse, while these elements are nearly extinct in the human genome. The striking difference in the dynamic behavior of such a large fraction of two mammalian genomes is intriguing, perhaps pointing to key and fundamental forces in the evolution of these different mammals.

Long–Range Distribution of Repetitive Sequences

Consistent with the general nonhomogeneous nature of features across the human genome, the distribution of repetitive elements varies greatly across each human chromosome. Some regions contain upwards of 90% repetitive sequences in intervals >500 kb, whereas others (e.g., the *HOX* gene clusters) are nearly devoid of repeats (<2% of the sequence). The possible functional significance of both extremes is of great interest: Do they arise from variation in the invasion or the retention of repeat sequences? Similarly, certain repetitive elements tend to cluster within specific genomic neighborhoods, such as *Alu* elements within the GC-rich regions of the genome.

Specialized Chromosomal Regions

There is now increased acceptance of the view that repeated sequences can impart biological functions. Nowhere is this more evident than in the structures of centromeres and telomeres as well as their neighboring locations. The function of the ~5% of the human genome recognized as containing α-satellite sequences is

now widely thought to be associated with centromere function, while the telomere-associated repeat motif TTAGGG is critical for telomere structure and function (Riethman et al. 2001). Importantly, the pericentromeric and subtelomeric regions are also rich in specific types of repeats arising from segmental duplications.

Segmental duplications, typically involving the transfer of 1–200 kb of DNA from a target region to one or more other sites, is remarkably common (see above). These regions arise from DNA transfer, not from unequal recombination, and are most often recent events. Most often, these duplications are absent in closely related species. There are two types of segmental duplications. Intrachromosomal duplications account for 2.0%–2.3% of the human genome, can be as large as 100 kb or more, and masquerade as low-copy chromosome-specific repeats with upwards of 99% sequence identity. When in close proximity, these repeats are substrates for unequal crossing over and the deletion or duplication of the intervening genomic region. Some of these regions are implicated in human Mendelian disorders as a result of the deletion or duplication of dosage-sensitive gene(s) in the target interval [e.g., Charcot-Marie-Tooth syndrome type 1A from duplications on 17p (Lupski et al. 1991) and hereditary neuropathy with liability to pressure palsies from complementary deletions on 17p (Chance et al. 1993)]. Interchromosomal duplications account for 1.5%–2.0% of the human genome and involve segments that are as large as 10–50 kb or more with ~97% sequence identity. Surprisingly, the majority of these events translocate specifically to the pericentromeric regions and, somewhat less frequently, to the subtelomeric regions of human chromosomes. Thus, the pericentromeric regions are preferential sites of location for interchromosomal duplications, which appear to result from minisatellite-like sequence motifs that target sequence insertions. Sequences from these regions frequently move to other pericentromeric targets as well. These pericentromeric regions are young in the human genome and polymorphic among individuals. In addition, the sequence divergence between the duplicated segments appears as sequence differences, confounding polymorphism analysis involving those regions. The above features of segmental duplications reveal the great plasticity in the human genome. Since many such duplications contain bona fide genes, it is suspected that these regions are a crucible for new gene evolution, with exon and protein domain shuffling and accretion.

Patterns of Polymorphism across the Human Genome

Most sequence differences between individuals arise from alleleic sequence changes, and understanding the patterns of this genomic variation is crucial for the study of molecular alterations in human disease. Simple sequence repeat (SSR) motifs are highly polymorphic and have been the workhorses of genetic disease mapping. The draft human genome sequence reveals a SSR every 2 kb, (on average) accounting for ~3% of the human genome. Of these, dinucleotide repeats are the most abundant (with AC and AT repeats being the most common), but trinucleotide repeats, including those directly involved in human disease, are less frequent than expected. The source of this intriguing difference is unknown. Previous mapping indicated the presence of fewer X chromosome SSRs than expected, either because they are less frequent or are less variable. The draft sequence shows an incidence of SSRs on the sex chromosomes equal to that on autosomes, indicating that X chromosome SSRs are less polymorphic. This can be explained by the unique genetic behavior of the sex chromosomes.

The draft human sequence has also driven the discovery of the most common type of human sequence variant, the single nucleotide polymorphism (SNP). To date, >2.3 million SNPs have been identified (The International SNP Map Working Group 2001), among which transition differences outnumber transversions by 2:1. Only a minute fraction of these SNPs occur within coding sequences, among which missense SNPs and silent SNPs are approximately equal. The frequency of variant sites (SNPs) between any two human genomes, termed nucleotide diversity, is ~8×10^{-4}. This value appears to hold for entire autosomal chromosomes but is severely reduced on sex chromosomes. The latter finding is expected because of the special transmission biology of sex chromosomes. The most important finding is the highly significant local genomic variation in SNP diversity; thus, when 200-kb segments across the genome are compared, the nucleotide diversity varies from 0 to 60×10^{-4}. These data are as expected, given the current thinking of human genetic and genome history (i.e., the founding of the human population from ~10,000 founders ~150,000 years ago). Importantly, the SNP data shows that there is greater variation within a given genome than between a given human genomes. With the majority of SNPs in non-coding DNA and unlikely to affect biological function, the challenge remains to identify the ~1% of SNPs that affect protein function, so that they can be used as direct probes of common human disease. Nonetheless, the majority of all known SNPs remain crucial tools for linkage and association markers, helping to identify the subset of common SNPs that are functionally relevant.

Patterns of Recombination across the Genome

One major factor affecting SNP frequency across the genome is the variation in recombination frequency within the human genome. As with other organisms, the frequency of meiotic recombination shows varia-

tion at several levels. The mapping of >8000 SSRs, and comparison of the resulting meiotic map to the physical map, allow for a detailed accounting of this variation (Yu et al. 2001). The genetic map in the human is 2730 cM in male meioses and 4250 cM in female meioses (a 1.6:1 gender difference), so that 2–3 mapped SSRs occur per cM. There is also at least 10-fold greater variation in recombination across the genome. Thus far, 19 meiotic "deserts" (cold spots of recombination that are <5 Mb in length and with rates <0.3 cM/Mb) and 12 meiotic "jungles" (hot spots of recombination that are <6 Mb in length and with rates >3 cM/Mb) have been identified. Unlike nucleotide diversity, recombination does show chromosome-specific patterns of variability, with recombination suppressed at centromeres, increased at telomeres, and decreasing with increasing chromosomal size. These features assure the occurrence of an obligate crossover on each chromosome arm, which is needed so that segregation of homologs can occur. From on the analyses so far, there are no strong predictors of recombination patterns based on repeat type and content or recognizable sequence motifs. Such observations are consistent with a model involving recombination initiation by double-stranded breaks that themselves depend on chromatin configuration; it is presently unknown if local chromatin accessibility is sequence dependent.

The patterns of recombination across the genome have implications for the choice of genetic markers for performing linkage mapping and association studies. With linkage disequilibrium between SSR markers being greater in recombination deserts than in jungles, an accurate linkage map can be immensely helpful in the choice of a higher density of SNP markers and the design of association studies.

Genome–Wide Mutation Rates

SNP patterns also depend on local mutation frequencies. Evolution occurs through the selection of individual mutations, and so it is natural to ask whether the mutation rate is the same in the compositionally heterogeneous human genome. The draft human sequence shows remarkable regional bias in substitution patterns based on GC content. Specifically, GC/CG pairs mutate to AT/TA pairs at a higher rate in AT-rich regions compared to GC-rich regions. This bias appears to rise from the earlier replication of GC-rich regions and the corresponding depletion of guanine pools. It is also likely that differences in DNA repair associated with transcriptional activity (e.g., gene-rich regions) may also contribute to the variation. Consequently, the human genome is not at equilibrium and has ~7% greater GC-content than expected. This feature may be due to natural selection or the invasion of transposable elements that prefer GC-rich regions. It has been suggested that much of non-repeated DNA may be remnants of ancient repeats no longer recognizable through mutation. Two other major features emerge: the youth of the Y chromosome, where DNA can be both gained and lost with little functional consequence (as assessed by younger-than-average LINE and other repeat elements) and the 2:1 excess of mutations in the male versus female germline. This last feature could either reflect inefficient repair in the male germline, analogous to mtDNA, or the greater number of cell divisions in male meioses.

Strategies for Genome Exploration

Significant attention has been given, in the scientific and popular press, to the two different strategies adopted by the Human Genome Project and Celera Genomics for their respective human genome expeditions. The main questions revolved around which was the most efficient method, whether repeats compromised the assembly of whole-genome shotgun data, and the accuracy of the assembly prior to finishing. While perhaps not yielding a particularly good "story," the assessment made at the base camp reveals remarkable convergence in these respective approaches, especially relevant for the planning of future ascents and expeditions of the human and other genomes.

In their paper (International Human Genome Sequencing Consortium 2001), the Human Genome Project explorers review the rationale for their choice of a hierarchical (i.e., clone-by-clone) shotgun sequencing strategy. Indeed, this reflects a combination of scientific, logistical, historical, and other factors. Importantly, this group recognized the value of supplementary sequence data generated in a whole-genome fashion, taking advantage of both BAC-end sequences as well as whole-genome sequence reads generated as part of a large SNP-discovery program (The International SNP Map Working Group 2001). Meanwhile, the Celera Genomics effort (Venter et al. 2001) employed a whole-genome shotgun approach for generating their data, but they made extensive use of the hierarchical information provided by the public Human Genome Project. Specifically, to supplement their own ~27 million sequence reads, the Celera Genomics group took the Human Genome Project's draft sequence (available in GenBank) and artificially shredded the assemblies into a series of 550-bp reads that formed perfect coverage across each sequenced BAC. This yielded just over 16 million synthetic reads (termed "faux reads" by Celera) that provided an additional ~3-fold coverage for use in conjunction with their own ~5-fold coverage to perform the whole-genome sequence assembly. Indeed, because these faux reads arose from mapped and ordered BACs, the estimated coverage provided is much greater than ~3-fold.

Based on these experiences, it is now clear that the most effective strategy for sequence-based genome ex-

ploration involves a "hybrid approach," whereby sequence reads are generated in both a hierarchical and a whole-genome shotgun fashion. Numerous subtle details (such as the optimal ratio of reads derived by each means, the differences among genomes with respect to how to implement a hybrid approach, and how to use information from one sequenced genome during the assembly of a related genome) can and will be clarified in future studies. The key point is that the past rhetoric about the dichotomy between the two strategies will quiet down with the recognition that the state-of-the-art approach for whole-genome sequencing involves the use of desirable elements of both strategies (see also Benos et al. [2001] for a comparison of sequencing strategies in *Drosophila*). Indeed, this is precisely how the next set of genome expeditions are already being performed (see below).

To the Summit of the Human Genome . . . and Beyond

The human genome has surely not been the first genome to be sequenced; nevertheless, it has been the most remarkable. This is not an anthropocentric view. The interest in this genome, from a strictly scientific view, stems from what it is teaching us about how information is retained and modified, as well as how it evolves within a genome. Among all genomes sequenced, the human genome has the lowest gene density; in other words, has the largest noncoding-to-coding ratio. Only a tiny fraction of the genome is gene coding, including the surprising fistful that we "inherited" from bacteria. Although no new protein families are evident, the genome has been remarkably adept at increasing protein diversity both by combining new domains and by invoking alternate splicing. Surprisingly, beyond this stable set of DNA, the rest, and majority, of the human genome is a tempest. The genome appears to routinely both duplicate segments and haul them away to both neighboring and distant parts and be a "rooming house" for transposons of various sorts and for various lengths, although it has been quiet lately. Much of the transposon-derived sequence changes quickly, and thus is now beyond recognition and appears as unique non-coding DNA. When these elements go for a ride, so do genes in the neighborhood. These features explain both how new functions usually emerge and how the C-value paradox can be explained. The C-value paradox arose from an attempt to explain the lack of correlation between organismal complexity and DNA content: Many salamanders have DNA contents that would put the human to shame. Thus, saltational DNA expansion may indicate such transposon invasions, which themselves can mould the genome in the future but not necessarily lead to new gene functions or proteins. Although non-genic, these elements lead to human disease. Thus, the non-homogeneous nature of the human genome may arise from very different and distinct processes shaping the evolution of the genome. Finishing the human genome and comparing it to other vertebrate genomes should clarify whether or not the above is a realistic scenario (see below).

Features of the Human Genome—A Scorecard[1]

Genetic maps	
Total male length	2730 cM
Total female length	4250 cM
No. genetic markers	~8000
No. recombination cold spots	19 regions
No. recombination hot spots	12 regions
Physical maps	
No. mapped BAC clones	372,264
Depth	20-fold
Sequence	
Estimated genome size	
Human Genome Project	3200 Mb (2900 Mb euchromatic)
Celera Genomics	2910 Mb
Total coverage in working draft sequence	~90%
GC content	~40% (range 30–65%)
Repeats	~44%
Coding	~1.5%
Unique, noncoding	~54.5%
Estimated no. genes	
Human Genome Project	32,000
Celera Genomics	39,114
Average gene size	27 kb
No. protein families	1695

[1]Derived from information provided in the International Human Genome Sequencing Consortium (2001) and Venter et al. (2001).

As with any expedition, arrival at the base camp also intensifies the preparation for the final ascent. Although the initial view of the human genome is fascinating, the scientific data from the completed and finished sequence will be even more so. Thus, the highest current priority is to finish the human genome sequence to a high accuracy and as completely as possible. Currently, greater than one-third (>1 Gb) of the human genome is finished to an accuracy of <1 error per 10,000 bp, with the goal of finishing the remaining <2 Gb within the next two years. Among the many tasks are the following: Numerous gaps, both in the clone map and in the sequence of individual clones, must be filled; minor (although, irritating) instances of clone-to-clone contamination within the draft sequence must be rectified; and, long-range ambiguities, especially with regard to large segmental duplications, (see Eichler 2001) need to be resolved.

With a finished human genome sequence in hand, the view from the summit will be even more spectacular because some of the current views will undergo revision. Activities at the summit will aggressively focus

on further refinement of the human gene inventory. These will be aided by ever-improving computational tools for gene prediction (e.g., Kan et al. 2001; Rojic et al. 2001; Yeh et al. 2001; Zhou et al. 2001), comparative analyses with other vertebrate genome sequence (see below), and the generation of complete sets of mammalian full-length cDNA sequences (Strausberg et al. 1999; The RIKEN Genome Exploration Research Group Phase II Team and The FANTOM Consortium 2001) (e.g., see mgc.nci.nih.gov). In parallel, a more complete understanding of the corresponding and inevitably more complex human protein inventory will be pursued. Almost certainly this will involve new initiatives to study protein structure and function on a large scale, similar to the industrialization of DNA sequencing that has occurred over the past five years. In addition, systematic efforts to identify regulatory elements that orchestrate the complex expression of genes will begin (Pennacchio and Rubin 2001). Finally, genetic variation of the sequence will teach us to what extent protein and regulatory functions are impacted by inter-individual sequence differences (Chakravarti 2001).

A key component of the above activities, especially with respect to the cataloging of genes and their regulatory elements, will be the comparative sequencing of multiple other vertebrate genomes by the Human Genome Project. In contrast to the sequencing of the human genome, smaller groups of sequencing centers will come together to elucidate the sequence of other vertebrates. For example, hybrid sequencing strategies are being actively used to sequence the mouse, rat, and zebrafish genomes. Sequencing efforts involving the generation of whole-genome shotgun data are also ongoing for two pufferfish species. The selection and prioritization of other vertebrate genomes for sequencing by the Human Genome Project represents an active and often lively topic of discussion [e.g., in contemplating the sequencing of the chimpanzee genome (VandeBerg et al. 2000; Varki 2000; McConkey and Varki 2000)]. A critical issue is the desired level of completeness and accuracy for comparative sequencing, although it has been the experience to date that the most definitive and compelling conclusions emerge only from the analysis of highly accurate sequence data. With the available sequencing capacity, the continued decline in the costs of sequencing, and the increasing recognition of the value of comparative sequence data, it can be confidently anticipated that numerous other parallel expeditions will be initiated for sequencing myriad vertebrate genomes in the years ahead.

REFERENCES

Benos, P.V., Gatt, M.K., Murphy, L., Harris, D., Barrell, B., Ferraz, C., Vidal, S., Brun, C., Demaille, J., Cadieu, E., et al. 2001. *Genome Res.* **11:** XXX.

Bork, P. and Copley, R. 2001. *Nature* **409:** 818–820.

Chakravarti, A. 2001. *Nature* **409:** 822–823.

Chance, P.F., Alderson, M.K., Leppig, K.A., Lensch, M.W., Matsunami, N., Smith, B., Swanson, P.D., Odelberg, S.J., Disteche, C.M., and Bird, T.D. 1993. *Cell* **15:** 143–151.

Crollius, H.R., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., et al. 2000. *Nature Genet.* **25:** 235–238.

Eichler, E. 2001. *Genome Res.* **11:** 653–656.

Ewing, B. and Green, P. 2000. Nature Genet. **25:** 232–234.

Glusman, G., Yanai, I., Rubin, I., and Lancet, D. 2001. *Genome Res.* **11:** 685–702.

Hughes, A., daSilva, J., and Friedman, R. 2001. *Genome Res.* **11:** XXX

International Human Genome Sequencing Consortium. 2001. Nature **409:** 860–921.

Kan, Z., Rouchka, E.C., Gish, W.R., and States, D.J. 2001. *Genome Res.* **11:** 889–900.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L., and Quackenbush, J. 2000. *Nature Genet.* **25:** 239–240.

Lupski, J.R., de Oca-Luna, R.M., Slaugenhaupt, S., Pentao, L., Guzzetta, V., Trask, B.J., Saucedo-Cardenas, O., Barker, D.F., Killian, J.M., and Garcia, C.A. 1991. *Cell* **66:** 219–232.

Makałowski, W. 2001. *Genome Res.* **11:** 667–670.

McConkey, E.H. and Varki, A. 2000. *Science* **289:** 1295–1296.

Ohno, S. 1970. Springer-Verlag, Berlin.

Pennacchio, L.A. and Rubin, E.M. 2001. *Nat. Rev. Genet.* **2:** 100–109.

Riethman, H.C., Xiang, Z., Paul, S., Morse, E., Hu, X.-L., Flint, J., Chi, H.-C., Grady, D.L., and Moyzis, R.K. 2001. **409:** 948–951.

Robinson-Rechavi, M., Marchand, O., Escrival, H., Bardet, P.-L., Zelus, D., Hughes, S., and Laudet, V. 2001. *Genome Res.* **11:** 781–788.

Rojic, S., Mackworth, A.K., and Oullette, F.B.F. 2001. *Genome Res.* **11:** 817–832.

Strausberg, R.L., Feingold, E.A., Klausner, R.D., and Collins, F.S. 1999. *Science* **286:** 455–457.

The International SNP Map Working Group. 2001. *Nature* **409:** 928–933.

The RIKEN Genome Exploration Research Group Phase II Team and The FANTOM Consortium. 2001. *Nature* **409:** 685–690.

VandeBerg, J.L., Williams-Blangero, S., Dyke, B., and Rogers, J. 2000. *Science* **290:** 1504–1505.

Varki, A. 2000. *Genome Res.* **10:** 1065–1070.

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., et al. 2001. *Science* **291:** 1304–1351.

Yeh, R.F., Lim, L.P., and Burge, C.B. 2001. *Genome Res.* **11:** 803–816.

Yu, A., Zhao, C., Fan, Y., Jang, W., Mungall, A.J., Deloukas, P., Olsen, A., Doggett, N.A., Ghebranious, N., Broman, K.W., et al. 2001. *Nature* **409:** 951–953.

Zhou, D., Zhao., W.D., Wright, F.A., Yang, H.-Y., Wang, J.P., Sears, R., Baer, T., Kwon, D.-H., Gordon, D., Gibbs, S., et al. 2001. *Genome Res.* **11:** 904–918.