Human Whole-Genome Shotgun Sequencing

James L. Weber and Eugene W. Myers

Genome Res. 1997 7: 401-409
Access the most recent version at doi:10.1101/gr.7.5.401

PERSPECTIVE

Human Whole-Genome Shotgun Sequencing

James L. Weber[1,3] and Eugene W. Myers[2]

[1]Center for Medical Genetics, Marshfield Medical Research Foundation, Marshfield, Wisconsin 54449;
[2]Department of Computer Science, University of Arizona, Tucson, Arizona 85721

Large-scale sequencing of the human genome is now under way (Boguski et al. 1996; Marshall and Pennisi 1996). Although at the beginning of the Genome Project, many doubted the scientific value of sequencing the entire human genome, these doubts have evaporated almost entirely (Gibbs 1995; Olson 1995). Primary reasons for generating the human genomic sequence are listed in Table 1.

The approach being taken for human genomic sequencing is the same as that used for the *Saccharomyces cerevisiae* and *Caenorhabditis elegans* genomes, namely construction of overlapping arrays of large insert *Escherichia coli* clones, followed by complete sequencing of these clones one at a time. In this article, we outline an alternative approach to sequencing the human and other large genomes, which we argue is less costly and more informative than the clone-by-clone approach.

A Plan for Human Whole-Genome Shotgun Sequencing

Although there are many conceivable variations, the crux of our plan involves high-quality, semiautomated sequencing from both ends of very large numbers of randomly selected human genomic DNA fragments. DNA of high molecular weight purified from at least a few different human donors would be sheared, size-selected, and cloned into *E. coli*. Insert sizes would fall into two classes. Long inserts would be 5–20 kb in size and would be cloned into plasmid, phage, or possibly cosmid vectors. Short inserts would be 0.4–1.2 kb in size and would be cloned into plasmid vectors. Read lengths would be of sufficient magnitude so that the two sequence reads from the ends of the short inserts overlap. The ratio of long to short inserts would be ≥ 1. Standard, gel-based methods would be utilized to generate at least 30 billion nucleotides of raw sequence (10-fold coverage of the genome). Many laboratories throughout the world could participate in raw sequence generation, but all sequences would be deposited in a common, public database, and only a few or possibly even one large informatics group would assay the primary task of sequence assembly. Following initial assembly, gaps in sequence coverage would need to be filled and uncertainties in assembly would need to be resolved.

Sequencing from both ends of relatively long insert subclones is an essential feature of the plan. Initially, Edwards and colleagues (1990) and, more recently, several other groups (Chen et al. 1993; Smith et al. 1994; Kupfer et al. 1995; Roach et al. 1995; Nurminsky and Hartl 1996) recognized that sequence information from both ends of relatively long inserts dramatically improves the efficiency of sequence assembly. In contrast to single sequence reads from one end of shotgun subclones, the pairs of sequence reads from both ends have known spacing and orientation. Use of relatively long insert subclones also aids in the assembly of sequences containing interspersed repetitive elements. Roach and colleagues (1995) showed that use of a mixture of long and short inserts can be as effective in enhancing assembly as use of only long inserts. Precise knowledge of the length of the long insert clones is not required to realize the advantages of end sequencing.

Another essential feature of the plan is the attachment of quality values to the raw sequences. The quality values would indicate the likelihood that each base call is correct. Quality values would aid sequence assembly (Churchill and Waterman 1992; Giddings et al. 1993; Lawrence and Solovyev 1994; Lipshutz et al. 1994), would help to distinguish true DNA polymorphisms from sequencing errors, and would also label uncertain sequences. Quality values would not obviate the need for relatively low error rates in the sequencing (Fleischmann et al. 1995). Low error rates would minimize the number of overlapping nucleotides required for sequence joining and also the ultimate sequence redundancy that is required. Frequent and appropriate quality controls would need to be utilized to ensure that the raw sequence generated is high quality. The quality of the combined sequences from the ends of the short inserts would be enhanced because

[3]Corresponding author.
E-MAIL weberj@mfldclin.edu; FAX (715) 389-3808.

Table 1. Primary Reasons for Sequencing Human Genomic DNA

Complete sequencing of all genes
Determine intron/exon structure of all genes
Map genes and other sequences
Reveal noncoding regulatory sequences
Identify polymorphisms
Develop methodology for other genomes
Uncover the unexpected

the overlapping segment occurs at the ends of the sequence reads where base calling is typically least reliable.

Feasibility of Whole-Genome Shotgun Sequencing

The feasibility of human whole-genome shotgun sequencing was evaluated by computer simulation designed to determine whether sufficient coverage and linkage information would result from such an approach. The simulation considered sequencing from both ends of two classes of inserts, long and short. The simulation also modeled both short and long interspersed repetitive elements (SINEs and LINEs). To be conservative, all interspersed repeats were considered to be identical in sequence so that overlaps in reads that fell within repetitive elements were useless for joining sequences. Many parameters such as fold coverage of the genome, sequence read length, amount of repetitive DNA, ratio of long to short inserts, and nucleotides of overlap required to join sequences were varied in the simulations. Default parameters (Table 2) are assumed to be in force unless otherwise stated. The default value for LINE length was conservatively chosen to be 1.5 kb, because although full-length LINE-1 (L1) elements are 6–7 kb in length, the vast majority of human L1 elements are truncated with average length ~0.7 kb (Smit et al. 1995; A. Smit, pers. comm.). Note that the simulation does not solve an assembly problem over simulated data, but instead analyzes the nature of the sampling obtained. Details of the simulation, including source code, can be obtained from Gene Myers (gene@cs.arizona.edu).

Two outcomes of the simulation, contig length and scaffold length, were monitored particularly closely. Contigs are defined as sequence assemblies without any discontinuities. Scaffolds (Roach et al. 1995) are defined as collections of two or more contigs joined by long inserts whose ends are in different contigs. Scaffolds, by definition, contain discontinuities, but the positions and approximate sizes of the discontinuities are known. The simulation confirmed that coverage of the genome is largely a function of the amount of raw sequence generated (Lander and Waterman 1988; Fleischmann et al. 1995). As shown in Table 3, the average simulated contig length increased dramatically as the fold coverage of the genome increased from 0.5 to 10. Average contig length was also dependent on the amount of interspersed repetitive DNA and the ratio of long to short inserts (Fig. 1). Increasing amounts of repetitive DNA led to shorter average contigs. Even at 50% total repetitive DNA, however, maximum contig length was still near 100 kb. When long-to-short insert ratios were greater than 1, con-

Table 2. Simulation Default Parameters

35-nucleotide overlap required for sequence joining
10-fold genome coverage
400-nucleotide read lengths
15% variation in insert sizes
10,000-nucleotide average size for long inserts
700-nucleotide average size for short inserts
1:1 ratio of long to short inserts
100 kb spacing between STSs
300-nucleotide STS length
20% of genome comprised of SINEs with 300-nucleotide lengths
5% of genome comprised of LINEs with 1500-nucleotide lengths
4:1 ratio of SINEs to LINEs

Table 3. Simulated Effects of Genome Coverage

Fold coverage	Average contig length (kb)
0.5	0.85
1	1.0
2	1.5
4	4.8
6	17.7
8	65.8
10	226

All simulation parameters other than fold coverage were set to default values (see Table 2). Average contig length excluded those contigs consisting of only single reads. The single-read contigs comprised only ~0.1% of all reads.

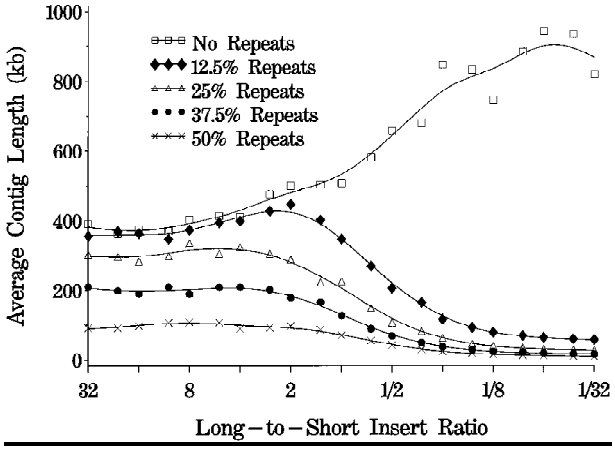

Figure 1 Average simulation contig length as a function of repeat density and long-to-short insert ratio. At each level of repetitive DNA, 80% of the repeats were assumed to be SINES and 20% LINES. All simulation parameters not specified in the plots were set to default values (see Table 2). Average contig length excluded those contigs consisting of only single reads. The single-read contigs comprised only ~0.1% of all reads.

tig length was largely independent of the ratio. These results were only modestly affected by read length (from 200 to 800 bases) and by the minimum overlap required for sequence joining (from 20 to 60 bases) (data not shown).

Given the large number of contigs that would be generated with the whole-genome shotgun approach, a pivotal question is whether the simulation contigs could be ordered into scaffolds. For a hypothetical human chromosome, 400 Mb in size, one scaffold spanning the entire chromosome length was obtained in each of 100 simulation iterations. After assembly, an average of 160 contigs and six small scaffolds remained unconnected to the single, very large scaffold (scaffolds can overlap without being connected by common sequence).

Using the default parameters, only ~16,000 gaps between contigs (0.04% of the genome) with average size of ~70 bp and maximum size <1700 bp remained after assembly. Although filling these gaps would certainly require a large effort, because the gaps are short, it should be possible to fill virtually all of them using PCR. Additional effort, if deemed necessary, would be required to sequence the complementary strand of segments with only single-strand coverage. Simulation results indicate that under default conditions, 616,000 of these single-stranded regions would exist with an average size of 106 bases.

Although a large amount of computing power would be required to perform the sequence similarity searches necessary for assembly, such power is already available. Using conservative and sensitive overlap detection algorithms, it would currently be possible to span sequence-tagged sites (STSs) spaced at 100 kb at a rate of at least one STS pair per day per 100 mips (million instructions per second) workstation. With a cluster of 100 such workstations the assembly of the entire human genome would take 300 days. By using less sensitive, but faster, overlap detection software, this time could be reduced by nearly a factor of 10. Note also that the power of computer processors has doubled every 18 months for many years, and this trend is likely to continue (Patterson 1995). If contemplated machines such as the 3-teraflop supercomputer planned in 1998 for Lawrence Livermore National Laboratory (Macilwain 1996) were recruited to the task of assembly, then the human genome could be assembled, in principle, in 4 min.

It is important to realize that because of significant progress in the genetic and physical mapping of STSs (Olson et al. 1989), the real task of shotgun sequence assembly would be greatly simplified to the task of building contigs and scaffolds that span adjacent STSs. Each of the STSs would serve as a nucleation site for this linking process. Already >30,000 total human STSs, including >16,000 genes, have been physically mapped, and the tally is increasing rapidly (Cox et al. 1994; Hudson et al. 1995; Schuler et al. 1996 and Web sites listed therein). Expressed sequence tags (ESTs) (Adams et al. 1991, 1995; Hillier et al. 1996) are particularly valuable for sequence assembly because the coding sequences are often interrupted by introns. For the purposes of assembly, a single EST will therefore usually be the equivalent of an array of ordered STSs, a nearly ideal framework for assembly. Plans to generate full-length cDNA sequences (Marshall 1996) will only enhance the utility of these sequences for assembly. Some genes like the dystrophin and neurofibromatosis I genes, for example, cover enormous segments of the genome (2.3 and 0.35 Mb, respectively) (Heim et al. 1995; Prior et al. 1995). Assuming, conservatively, a total of 80,000 human ESTs and an average of three exons per sequence, a grand total of >250,000 STSs with an average spacing of only 12 kb is already available for assembly (Table 4).

At present, the process for human whole-genome shotgun sequence assembly can only be projected. Nevertheless, a possible senario for assembly would be to begin with all existing mapped

Table 4. Human STSs

Type	Number
STRPs	10,000
Nonpolymorphic anonymous	5,000
Genes (ESTs)	80,000 (×3)
Total	255,000

STSs (including ESTs) within a specific chromosomal interval, to add shotgun reads in a very conservative fashion utilizing only sequence overlaps of high probability, to meld these growing assemblies to unmapped STSs within the database, and then to add in lower probability overlapping sequences. The sequence assemblies would continually be examined for disagreements with EST structure or with existing map information and also for the presence of forks or loops, which would indicate the presence of unrecognized interspersed (forks) or tandem (loops) repeats, or other errors in assembly or cloning artifacts. Software for assembly on this scale does not exist, but we have begun work in this direction. Our initial perception is that STS anchors provide sufficient directional information to allow resolution of low copy number repeats (of any scale) and that high copy number repeats can be factored as a consensus sequence that can be resolved at specific sites on a case-by-case basis. The development of such software poses difficult technical questions, but we believe these are surmountable in a several man–year horizon. We note, for example, that human coding sequences have been assembled from individual reads by several groups despite the presence of sequence errors, polymorphisms, alternative splicing, and repetitive elements (Schuler et al. 1996). Also, software developed for assembly of human sequences would be applied in the future to many other organisms.

Whole-genome shotgun sequencing would not result in a single unbroken sequence for entire chromosomes. Even using recombination and restriction-deficient *E. coli* strains (Chalker et al. 1988; Raleigh et al. 1988; Doherty et al. 1993), a small portion of the genome would likely be resistant to cloning or would not yield stable clones. Sequences from long arrays of tandem repeats such as centromeric satellite DNA, rDNA repeats, and some minisatellites would not be able to be assembled perfectly. Note, however, that these limitations apply to both whole-genome shotgun and clone-by-clone sequencing approaches.

The feasibility of whole-genome shotgun sequencing was also supported by the recent success achieved by Venter and colleagues in sequencing three bacterial genomes with sizes ranging from 0.6 to 1.8 Mb (Fleischman et al. 1995; Fraser et al. 1995; Bult et al. 1996). Neither raw sequence generation, sequence assembly, nor sequence finishing was an impediment to the shotgun sequencing of the bacterial chromosomes. Distances between human STSs are much smaller than the sizes of the bacterial genomes.

Our strategy for whole-genome shotgun sequencing is also entirely consistent with the bacterial artificial chromosome (BAC) end sequencing strategy proposed recently by Venter et al. (1996). Although we feel that large-scale BAC end sequencing would probably not be absolutely required, it would certainly assist in the assembly of the shotgun sequence fragments. BAC clones would likely span some arrays of tandem repeats that are too large for our ''long insert'' clones.

Advantages of Whole-Genome Shotgun Sequencing

Whole-genome shotgun sequencing of human genomic DNA holds a number of important advantages compared to conventional clone-by-clone sequencing. Foremost among these advantages are detection of large numbers of DNA polymorphisms, more complete and less artifactual coverage of the genome, and improved speed and cost.

A significant fraction of all common human DNA polymorphisms can be detected through shotgun sequencing. Polymorphisms are important because they are used to map genes through linkage analysis (Terwilliger and Ott, 1994), to presymptomatically predict disease status (Antonarakis 1989; Weber 1994), to detect submicroscopic chromosomal rearrangements (Lupski et al. 1991), to identify individuals in, for example, paternity and forensic testing (Hagelberg et al. 1991; Frigeau and Fourney 1993; Smith 1995; Urquhart et al. 1995), and to study a wide range of biological phenomena such as evolution (Bowcock and Cavalli-Sforza 1991; Bowcock et al. 1994; Jorde et al. 1995), population biology (Edwards et al. 1992; Deka et al. 1995; Morell et al. 1995), and recombination (Tanzi et al. 1992; Weber et al. 1993). Polymorphisms within coding and regulatory elements are also the source of relative risk for many common diseases. Common variants of the apolipoprotein E gene on chromosome 19, for example, strongly influence an in-

dividual's risk of developing late onset Alzheimer's disease (Saunders et al. 1993; Kamboh 1995; Kamboh et al. 1995). Many highly informative human DNA polymorphisms based on short tandem repeats have already been identified, but the vast majority of the much more frequent biallelic base substitution and short insertion/deletion polymorphisms remain unknown (Kwok et al. 1994, 1996). Although allele frequencies vary widely, most human DNA polymorphisms are common to all populations (Bowcock and Cavalli-Sforza 1991; Jorde et al. 1995; Bowcock et al. 1994; Deka et al. 1995; Edwards et al. 1992; Morell et al. 1995).

DNA polymorphisms would not usually be detected through clone-by-clone sequencing because only one variant for each genomic region would be sampled. If the genome is sequenced through the clone-by-clone approach, then much additional funding would be required to identify the polymorphisms at a later date and many years would be lost. Calculation of the exact fraction of polymorphisms that would be identified through whole-genome shotgun sequencing requires a distribution of polymorphisms as a function of informativeness, which is not yet known. However by generating 6 billion nucleotides of raw sequence from each of five unrelated individuals, it can be calculated that ~65% of all 20% heterozygosity biallelic polymorphisms and >99% of all 80% multiallelic polymorphisms would, for example, be detected. To optimize polymorphism detection, DNA should ideally be sequenced from donors with widely differing geographic ancestry.

Sequencing errors would likely be encountered much more frequently in whole-genome shotgun sequencing than true polymorphisms. Sequencing error rates would likely be at least 1%, whereas the rate of polymorphisms would likely be on the order of 0.1%. Although confirmation may be necessary in many cases, several factors would allow many of the polymorphisms to be identified despite the background of sequencing errors. True polymorphisms would often have multiple sequence reads per allele, true polymorphisms would usually have high-quality values attached to each allele, and true polymorphisms do not occur randomly thoughout the genome. Specific sequence features will spotlight polymorphisms. For example, it has been known for many years that CpG dinucleotides are more commonly polymorphic than other dinucleotides (Schumm et al. 1988; Deininger and Batzer 1993; Becker et al. 1996; Sommer and Ketterling 1996).

Rearrangements in the large insert contig clones and biases in the coverage of these clones will, to a large degree, be eliminated by whole-genome shotgun sequencing. Many of the cosmid clones projected for use in sequencing have been developed from hybrid tissue culture cell lines which, themselves, have been propagated for many cell generations. Rearrangements and artifacts have undoubtedly been introduced into the cloned material during this process. Although BACs/PACs (P1-derived artificial chromosomes) appear to be more stable than cosmids, artifacts such as chimeras and deletions still occur at a significant frequency (Kim et al. 1996; Boysen et al. 1997). By starting with total human genomic DNA, many of these artifacts will be eliminated. The cosmid or BAC/PAC assemblies will also likely exclude at least some long arrays of tandem repeats. The genome will be more equally represented with shotgun sequencing using small inserts. In addition, overlaps between large insert clones will lead to largely unproductive duplicative sequencing or to the expenditure of resources to avoid this duplication.

Whole-genome shotgun sequencing would also be less expensive and therefore faster than the clone-by-clone approach. The steps of preparation, mapping, storage, and tracking of tens of thousands of sequence-ready large-insert clones; parallel generation, storage and tracking of subclones for each of the large insert clones; and avoidance of large-insert clone overlap would be entirely eliminated with shotgun sequencing. The processes of sequence assembly and sequence finishing could be carried out much more efficiently in central facilities. Reducing the process of DNA sequencing to the core task of raw sequence generation would also allow efforts to be focused on driving down the costs of a few relatively straightforward procedures in large factory-like operations. With shotgun sequencing there would be no need to wait for expensive, sequence-ready large-insert clone assemblies to be generated and no need to sequence one chromosome or one chromosomal segment at a time. To date, no one has generated overlapping cosmid or BAC/PAC assemblies that span even significant portions of human chromosomes without many gaps (Ashworth et al. 1995; Doggett et al. 1995). Perhaps this can be accomplished eventually but only through great effort, time, and cost. The assertion that collection of large-insert templates for sequencing is trivial is simply wrong. Although initiation of genome-wide sequence assembly would probably not be worthwhile until ~2.5-fold sequence coverage was obtained, completion of partial cDNA sequences, identification of regulatory regions, defini-

tion of intron/exon boundaries, and identification of polymorphisms are all tasks that could be undertaken continuously from the start of shotgun sequence generation. The large number of laboratories worldwide undertaking position cloning projects, for example, could utilize the shotgun sequences from the outset.

Estimating the actual costs of human genomic sequencing is certainly hazardous. Nevertheless, our best effort is summarized in Table 5. Assuming optimistically that clone-by-clone sequencing of human DNA can be completed for $0.30 per finished base, and assuming that sequencing is completed by the end of the year 2003, an average cost per year of $130 million is calculated. Assuming conservatively a cost of $0.01 for generation of a single base of raw sequence, spending of $130 million per year would give 10-fold coverage by about the end of the millennium with $90 million remaining for software development and computer assembly. Filling gaps and resolving uncertainties would add additional costs to whole-genome shotgun sequencing in the next century.

We assert that the goals listed in Table 1 are the true motivation for sequencing the human genome, not the accomplishment of some arbitrary, mythical goal of 99.99% accuracy of a single, artifactual (in places) and nonrepresentative copy of the genome. Most research laboratories, both public and private, want discrete genomic sequence information, and they want it as early as possible. They are interested in information such as the intron/exon structure of specific genes, the polymorphisms that may occur in specific coding and regulatory sequences, and lists of coding sequences that lie within specific chromosomal intervals. The sooner this critical information is available, the sooner it can be applied to accelerating research progress.

Table 5. Costs of Human Genomic Sequencing

Clone by clone
$0.30 per finished base
$130 million per year for 7 years
Total $900 million spent by end of 2003
Shotgun
$0.01 per raw base
$130 million for 3 years would provide
10× coverage plus an additional $90 million for informatics

Americans spend ~$35 billion per year, public and private, on biomedical research (Silverstein et al. 1995). If the efficiency of this research is improved by even 1%, and this is probably a gross underestimate, then savings would be $350 million per year, far more than the cost of sequencing. Whole-genome shotgun sequencing will allow these savings to be realized far sooner than with clone-by-clone sequencing. We should generate as much of the critical sequence information as rapidly as possible and leave cleanup of gaps and problematic regions for future years.

It is not too late to change strategies for sequencing the human genome. Only a few percent of the sequence has been generated at this time. Even if the human genome is not sequenced via the shotgun approach, there are still many other large genomes that will be sequenced in the future, including many agriculturally important species. It will likely be too expensive to sequence other large genomes via the clone-by-clone approach. A possible general strategy for sequencing other large genomes would be a random cDNA sequencing project, followed possibly by some radiation hybrid physical mapping of the ESTs, followed by whole-genome shotgunning.

About a decade ago, when the Genome Project was just being contemplated, Fred Blattner proposed whole-genome shotgun sequencing of both the *E. coli* and human genomes. His proposals were neglected. Today, no one considers for a moment sequencing bacterial genomes by any method other than whole-genome shotgun sequencing. Even at several dollars per finished base the human sequence is probably one of the greatest bargains in human history. We laud efforts now under way in several large sequencing centers to generate human genomic sequence. The reality, however, is that research dollars are always limited. We should sequence the human and other eukaryotic genomes using the most rapid, cost effective, and productive strategy.

REFERENCES

Adams, M.D., J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R.F. Morenco, A.R. Kerlavage, W.R. McCombie, and J.C. Venter. 1991. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* 252: 1651–1656.

Adams, M.D., A.R. Kerlavage, R.D. Fleischmann, R.A. Fuldner, C.J. Bult, N.H. Lee, E.F. Kirkness, K.G. Welnstock, J.D. Gocayne, O. White, G. Sutton, J.A. Blake, R.C. Brandon, M.-W. Chlu, R.A. Clayton, R.T. Cline, M.D. Cotton, J.

Earle-Hughes, L.D. Fine, L.M. Fitzgerald, W.M. FitzHugh, J.L. Fritchman, N.S.M. Geoghagen, A. Glodek, C.L. Gnehm, M.C. Hanna, E. Hedblom, P.S. Hinkie, Jr., J.M. Kelley, K.M. Kilmek, J.C. Kelley et al. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* (Suppl. 3) 377: 3–17.

Antonarakis, S.E. 1989. Diagnosis of genetic disorders at the DNA level. *N. Engl. J. Med.* 320: 153–163.

Ashworth, L.K., M.A. Batzer, B. Brandriff, E. Branscomb, P. de Jong, E. Garcia, J.A. Garnes, L.A. Gordon, J.E. Lamerdin, G. Lennon, H. Mohrenweiser, A.S. Olsen, T. Slezak, and A.V. Carrano. 1995. An integrated metric physical map of human chromosome 19. *Nature Genet.* 11: 422–427.

Becker, J., R. Schwaab, A. Moller-Taube, U. Schwaab, W. Schmidt, H.H. Brackmann, T. Grimm, K. Olek, and J. Oldenburg. 1996. Characterization of the factor VIII defect in 147 patients with sporadic hemophilia A. *Am. J. Hum. Genet.* 58: 657–670.

Boguski, M., A. Chakravarti, R. Gibbs, E. Green, and R.M. Myers. 1996. The end of the beginning: The race to begin human genomic sequencing. *Genome Res.* 6: 771–772.

Bowcock, A. and L.C. Cavalli-Sforza. 1991. The study of variation in the human genome. *Genomics* 11: 491–498.

Bowcock, A.M., A. Rulz-Linares, J. Tomfohrde, E. Minch, J.R. Kidd, and L.L. Cavalli-Sforza. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368: 455–457.

Boysen, C., I. Simon, and L. Hood. 1997. Analysis of the 1.1-Mb human α/δ T-cell receptor locus with bacterial artificial chromosome clones. *Genome Res.* 7: 330–338.

Bult, C.J., O. White, G.J. Olsen, L. Zhou, R.D. Fleischmann, G.G. Sutton, J.A. Blake, L.M. FitzGerald, R.A. Clayton, J.D. Gocayne, A.R. Kerlavage, B.A. Dougherty, J.-F. Tomb, M.D. Adams, C.I. Reich, R. Overbeek, E.F. Kirkness, K.G. Weinstock, J.M. Merrick, A. Glodek, J.L. Scott, N.S.M. Geoghagen, J.F. Weidman, J.L. Fuhrmann, D. Nguyen, T.R. Utterback, J.M. Kelley, J.D. Peterson, P.W. Sadow, M.C. Hanna, M.D. Cotton, K.M. Roberts, M.A. Hurst, B.P. Kaine, M. Borodovsky, H.-P. Klenk, C.M. Fraser, H.O. Smith, C.R. Woese, and J.C. Venter. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii.* *Science* 273: 1058–1073.

Chalker, A.F., D.R.F. Leach, and R.G. Lloyd. 1988. *Escherichia coli* sbcC mutants permit stable propagation of DNA replicons containing a long palindrome. *Gene* 71: 201–205.

Chen, E.Y., D. Schlessinger, and J. Kere. 1993. Ordered shotgun sequencing, a strategy for integrated mapping and sequencing of YAC clones. *Genomics* 17: 651–656.

Churchill, G.A. and M.S. Waterman. 1992. The accuracy of DNA sequences: Estimating sequence quality. *Genomics* 14: 89–98.

Cox, D.R., E.D. Green, E.S. Lander, D. Cohen, and R.M. Myers. 1994. Assessing mapping progress in the human genome project. *Science* 265: 2031–2032.

Deininger, P.L. and M.A. Batzer. 1993. Evolution of retroposons. In *Evolutionary Biology* (ed. M.K. Hect, R.H. MacIntyre, and M.T. Clegg), pp. 157–196. Plenum, New York, NY.

Deka, R., L. Jin, M.D. Shriver, L.M. Yu, S. DeCroo, J. Hundrieser, C.H. Bunker, R.E. Ferrell, and R. Chakraborty. 1995. Population genetics of dinucleotide (dC-dA)n.(dG-dT)n polymorphisms in world populations. *Am. J. Hum. Genet.* 56: 461–474.

Doggett, N.A., L.A. Goodwin, J.G. Tesmer, L.J. Melncke, D.C. Bruce, L.M. Clark, M.R. Altherr, A.A. Ford, H.-C. Chi, B.L. Marrone, J.L. Logmire, S.A. Lane, S.A. Whitmore, M.G. Lowenstein, R.D. Sutherland, M.O. Mundt, E.H. Knill, W.J. Bruno, C.A. Macken, D.C. Torney, J.-R. Wu, J. Griffith, Sutherland, L.L. Deaven, D.F. Callen, and R.K. Moyzis. 1995. An integrated physical map of human chromosome 16. *Nature* (Suppl.) 377: 335–365.

Doherty, J.P., R. Lindeman, R.J. Trent, M.W. Graham, and D.M. Woodcock. 1993. *Escherichia coli* host strains SURE™ and SRB fail to preserve a palindrome cloned in lambda phage: Improved alternate host strains. *Gene* 124: 29–35.

Edwards, A., H. Voss, P. Rice, A. Civitello, J. Stegemann, C. Schwager, J. Zimmermann, H. Erfle, C.T. Caskey, and W. Ansorge. 1990. Automated DNA sequencing of the human HPRT locus. *Genomics* 6: 593–608.

Edwards, A., H.A. Hammond, L. Jin, C.T. Caskey, and R. Chakraborty. 1992. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics* 12: 241–253.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.-F. Tomb, B.A. Dougherty, J.M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J.D. Gocayne, J. Scott, R. Shirley, L.-I. Liu, A. Glodek, J.M. Kelley, J.F. Weidman, C.A. Phillips, T. Spriggs, E. Hedblom, M.D. Cotton, T.R. Utterback, M.C. Hanna, D.T. Nguyen, D.M. Saudek, R.C. Brandon, L.D. Fine, J.L. Fritchman, J.L. Fuhrmann, N.S.M. Geoghagen, C.L. Gnehm, L.A. McDonald, K.V. Small, C.M. Fraser, H.O. Smith, and J.C. Venter. 1995 Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496–512.

Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G. Sutton, J.M. Kelley, J.L. Fritchman, J.F. Weidman, K.V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T.R. Utterback, D.M. Saudek, C.A. Phillips, J.M. Merrick, J.-F. Tomb, B.A. Dougherty, K.F. Bott, P.-C. Hu, T.S. Lucier, S.N. Peterson, H.O. Smith, C.A. Hutchison III, and J.C. Venter. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* 270: 397–403.

Frégeau, C.J. and R.M. Fourney. 1993. DNA typing with fluorescently tagged short tandem repeats: A sensitive and

accurate approach to human identification. *BioTechniques* 15: 100–119.

Gibbs, R.A. 1995. Pressing ahead with human genome sequencing. *Nature Genet.* 11: 121–125.

Giddings, M.C., R.L. Brumley, Jr., M. Haker, and L.M. Smith. 1993. An adaptive, object oriented strategy for base calling in DNA sequence analysis. *Nucleic Acids Res.* 21: 4530–4540.

Hagelberg, E., I.C. Gray, and A.J. Jeffreys. 1991. Identification of the skeletal remains of a murder victim by DNA analysis. *Nature* 352: 427–429.

Heim, R.A., L.N.W. Kam-Morgan, C.G. Binnie, D.D. Corns, M.C. Cayouette, R.A. Farber, A.S. Aylsworth, and L.M. Silverman. 1995. Distribution of 13 truncating mutations in the neurofibromatosis 1 gene. *Hum. Mol. Genet.* 4: 975–981.

Hillier, L., G. Lennon, M. Becker, M.F. Bonaldo, B. Chiapelli, S. Chissoe, N. Dietrich, T. DuBuque, A. Favello, W. Gish, M. Hawkins, M. Hultman, T. Kucaba, M. Lacy, M. Le, N. Le, E. Mardis, B. Moore, M. Morris, J. Parsons, C. Prange, L. Rifkin, T. Rohlfing, K. Schellenberg, M. Bento Soares, F. Tan, J. Thierry-Meg, E. Trevaskis, K. Underwood, P. Wohldman, R. Waterston, R. Wilson, and M. Marra. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* 6: 807–828.

Hudson, T.J., L.D. Stein, S.S. Gerety, J. Ma, A.B. Castle, J. Silva, D.K. Slonim, R. Baptista, L. Kruglyak, S.-H. Xu, X. Hu, A.M.E. Colbert, C. Rosenberg, M. Reeve-Daly, S. Rozen, L. Hui, X. Wu, C. Vestergaard, K.M. Wilson, J.S. Bae, S. Maitra, S. Ganiatsas, C.A. Evans, M.M. DeAngelis, K.A. Ingalls, R.W. Nahf, L.T. Horton, Jr., M. Oskin et al. 1995. An STS-based map of the human genome. *Science* 270: 1945–1954.

Jorde, L.B., M.J. Bamshad, W.S. Watkins, R. Zenger, A.E. Fraley, P.A. Krakowiak, K.D. Carpenter, H. Soodyall, T. Jenkins, and A.R. Rogers. 1995. Origins and affinities of modern humans: A comparison of mitochondrial and nuclear genetic data. *Am. J. Hum. Genet.* 57: 523–538.

Kamboh, M.I. 1995. Apolipoprotein E polymorphism and susceptibility to Alzheimer's disease. *Hum. Biol.* 67: 195–215.

Kamboh, M.I., D.K. Sanghera, R.E. Ferrell, and S.T. DeKosky. 1995. APOE*4-associated Alzheimer's disease risk is modified by α1-antichymotrypsin polymorphism. *Nature Genet.* 10: 486–488.

Kim, U.-J., B.W. Birren, T. Slepak, V. Mancino, C. Boysen, H.-L. Kang, M.I. Simon, and H. Shzuya. 1996. Construction and characterization of a human bacterial artificial chromosome library. *Genomics* 34: 213–218.

Kupfer, K., M.W. Smith, J. Quackenbush, and G.A. Evans. 1995. Physical mapping of complex genomes by sampled sequencing: A theoretical analysis. *Genomics* 27: 90–100.

Kwok, P.Y., C. Carlson, T.D. Yager, W. Ankener, and D.A. Nickerson. 1994. Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. *Genomics* 23: 138–144.

Kwok, P.-Y., Q. Deng, H. Zakeri, and D.A. Nickerson. 1996. Increasing the information content of STS-based genome maps: Identifying polymorphisms in mapped STSs. *Genomics* 31: 123–126.

Lander, E.S. and M.S. Waterman. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* 2: 231–239.

Lawrence, C.B. and V.V. Solovyev. 1994. Assignment of position-specific error probability to primary DNA sequence data. *Nucleic Acids Res.* 22: 1272–1280.

Lipshutz, R.J., F. Taverner, K. Hennessy, G. Hartzell, and R. Davis. 1994. DNA sequence confidence estimation. *Genomics* 19: 417–424.

Lupski, J.R, R.M. de Oca-Luna, S. Slaugenhaupt, L. Pentao, V. Guzzetta, B.J. Trask, O. Saucedo-Cardenas, D.F. Barker, J.M. Killian, C.A. Garcia, A. Chakravarti, and P.I. Patel. 1991. DNA duplication associated with Charcot-Marie-Tooth Disease Type 1A. *Cell* 66: 219–131.

Macilwain, C. 1996. Livermore to get $100m supercomputer. *Nature* 382: 385.

Marshall, E. 1996. The human gene hunt scales up. *Science* 274: 1356.

Marshall, E. and E. Pennisi. 1996. NIH launches the final push to sequence the genome. *Science* 272: 188–189.

Morell, R., Y. Liang, J.H. Asher, Jr., J.L. Weber, J.T. Hinnant, S. Winata, I.N. Arhya, and T.B. Friedman. 1995. Analysis of short tandem repeat (STR) allele frequency distributions in a Balinese population. *Hum. Mol. Genet.* 4: 85–91.

Nurminsky, D.I. and D.L. Hartl. 1996. Sequence scanning: A method for rapid sequence acquisition from large-fragment DNA clones. *Proc. Natl. Acad. Sci.* 93: 1694–1698.

Olson, M.V. 1995. A time to sequence. *Science* 270: 394–396.

Olson, M., L. Hood, C. Cantor, and D. Botstein. 1989. A common language for physical mapping the human genome. *Science* 245: 1434–1435.

Patterson, D.A. 1995. Microprocessors in 2020. *Sci. Am.* 273: 62–67.

Prior, T.W., C. Bartolo, D.K. Pearl, A.C. Papp, P.J. Snyder, M.S. Sedra, A.H.M. Burghes, and J.R. Mendell. 1995. Spectrum of small mutations in the dystrophin coding region. *Am. J. Hum. Genet.* 57: 22–33.

Raleigh, E.A., N.E. Murray, H. Revel, R.M. Blumenthal, D. Westaway, A.D. Reith, P.W.J. Rigby, J. Elhai, and D. Hanahan. 1988. McrA and McrB restriction phenotypes of some *E. coli* strains and implications for gene cloning. *Nucleic Acids Res.* 16: 1563–1575.

Roach, J.C., C. Boysen, K. Wang, and L. Hood. 1995. Pairwise end sequencing: A unified approach to genomic mapping and sequencing. *Genomics* 26: 345–353.

Saunders, A.M., W.J. Strittmatter, D. Schmechel, P.H. St. George-Hyslop, M.A. Pericak-Vance, S.H. Joo, B.L. Rosi, J.F. Gusella, D.R. Crapper-MacLachlan, M.J. Alberts, C. Hulette, B. Crain, D. Goldgaber, and A.D. Roses. 1993. Association of apolipoprotein E allele ϵ4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* 43: 1467–1472.

Schuler, G.D., M.S. Boguski, E.A. Stewart, L.D. Stein, G. Gyapay, K. Rice, R.E. White, P. Rodriguez-Tomi, A. Aggarwal, E. Bajorek, S. Bentolila, B.B. Birren, A. Butler, A.B. Castle, N. Chiannilkulchai, A. Chu, C. Clee, S. Cowles, P.J.R. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, C. Edwards, J.-B. Fan, N. Fang, C. Fizames, C. Garrett, L. Green, D. Hadley, M. Harris, P. Harrison, S. Brady, A. Hicks, E. Holloway, L. Hui, S. Hussain, C. Louis-Dit-Sully, J. Ma, A. MacGilvery, C. Mader, A. Maratukulam, T.C. Matise, K.B. McKusick, J. Morissette, A. Mungall, D. Muselet, H.C. Nusbaum, D.C. Page, A. Peck, S. Perkins, M. Piercy, F. Qin, J. Quackenbush, S. Ranby, T. Reif, S. Rozen, C. Sanders, X. She, J. Silva, D.K. Slonim, C. Soderlund, W.-L. Sun, P. Tabar, T. Thangarajah, N. Vega-Czarny, D. Vollrath, S. Voyticky, T. Wilmer, X. Wu, M.D. Adams, C. Auffray, N.A.R. Walter, R. Brandon, A. Dehejia, P.N. Goodfellow, R. Houlgatte, J.R. Hudson, Jr., S.E. Ide, K.R. Iorio, W.Y. Lee, N. Seki, T. Nagase, K. Ishikawa, N. Nomura, C. Phillips, M.H. Polymeropoulos, M. Sandusky, K. Schmitt, R. Berry, K. Swanson, R. Torres, J.C. Venter, J.M. Sikela, J.S. Beckmann, J. Weissenbach, R.M. Myers, D.R. Cox, M.R. James, D. Bentley, P. Deloukas, E.S. Lander, and T.J. Hudson. 1996. A gene map of the human genome. *Science* 274: 540–546.

Schumm, J.W., R.G. Knowlton, J.C. Braman, D.F. Barker, D. Botsein, G. Akots, V.A. Brown, T.C. Gravius, C. Helms, K. Hsiao, K. Rediker, J.G. Thurston, and H. Donis-Keller. 1988. Identification of more than 500 RFLPs by screening random genomic clones. *Am. J. Hum. Genet.* 42: 143–159.

Silverstein, S.C., H.H. Garrison, and S.J. Heinig. 1995. A few basic economic facts about research in the medical and related life sciences. *FASEB J.* 9: 833–840.

Smit, A.F.A., G. Toth, A.D. Riggs, and J. Jurka. 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J. Mol. Biol.* 246: 401–417.

Smith, M.W., A.L. Holmsen, Y.H. Wei, M. Peterson, and G.A. Evans. 1994. Genomic sequence sampling: A strategy for high resolution sequence-based physical mapping of complex genomes. *Nature Genet.* 7: 40–47.

Smith, R.N. 1995. Accurate size comparison of short tandem repeat alleles amplified by PCR. *Biotechniques* 18: 122–128.

Sommer, S.S. and R.P. Ketterling. 1996. The factor IX gene as a model for analysis of human germline mutations. *Hum. Mol. Genet.* 5: 1505–1514.

Tanzi, T.E., P.C. Watkins, G.D. Stewart, N.S. Wexler, J.F. Gusella, and J.L. Haines. 1992. A genetic linkage map of human chromosome 21: Analysis of recombination as a function of sex and age. *Am. J. Hum. Genet.* 50: 551–558.

Terwilliger, J.D. and J. Ott. 1994. *Handbook of human genetic linkage*. Johns Hopkins University Press, Baltimore, MD.

Urquhart, A., N.J. Oldroyd, C.P. Kimpton, and P. Gill. 1995. Highly discriminating heptaplex short tandem repeat PCR system for forensic identification. *Biotechniques* 18: 116–121.

Venter, J.C., H.O. Smith, and L. Hood. 1996. A new strategy for genome sequencing. *Nature* 381: 364–366.

Weber, J.L. 1994. Know thy genome. *Nature Genet.* 7: 343–344.

Weber, J.L., Z. Wang, K. Hansen, M. Stevenson, C. Kappel, S. Salzman, P.J. Wilkie, B. Keats, N.C. Dracopoli, B.F. Brandriff, and A.S. Olsen. 1993. Evidence for human meiotic crossover interference obtained through construction of a short tandem repeat polymorphism linkage map of chromosome 19. *Am. J. Hum. Genet.* 53: 1079–1095.