PERSPECTIVE

Against a Whole-Genome Shotgun

Philip Green[1]

Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195

The human genome project is entering its decisive final phase, in which the genome sequence will be determined in large-scale efforts in multiple laboratories worldwide. A number of sequencing groups are in the process of scaling up their throughput; over the next few years they will need to attain a collective capacity approaching half a gigabase per year to complete the 3-Gb genome sequence by the target date of 2005. At present, all contributing groups are using a clone-by-clone approach, in which mapped bacterial clones (typically 40–400 kb in size) from known chromosomal locations are sequenced to completion. Among other advantages, this permits a variety of alternative sequencing strategies and methods to be explored independently without redundancy of effort. Although it is not too late to consider implementing a different approach, any such approach must have as high a probability of success as the current one and offer significant advantages (such as decreased cost). I argue here that the whole-genome shotgun proposed by Weber and Myers satisfies neither condition.

Clone-by-Clone Sequencing

For purposes of comparison it is helpful to first outline a specific implementation of clone-by-clone sequencing. Although by no means the only one possible, this implementation is being used by several of the larger groups and seems likely to be the method of choice for the major part of the genome. One starts with a set of mapped sequence-tagged sites (STSs) (Olson et al. 1989) from a particular chromosomal region. These are screened against a bacterial artificial chromosome (BAC) (or other large bacterial clone) library (Kim et al. 1996) to obtain overlapping clusters of clones from that region. Since whole-genome mapping efforts are nearing the target density of 1 STS per 100 kb [Hudson et al. 1995; D.R. Cox and R.M. Myers et al. 1997, World Wide Web (WWW) site for the Stanford Human Genome Center, http://shgc.stanford.edu; E. Lander et al. 1997, WWW site for the Whitehead Institute/MIT Center for Genome Research, http://www-genome.wi.mit.edu], with several intensively mapped chromosomes already exceeding it (Nagaraja et al. 1997, Bouffard et al. 1997), and BACs average 130 kb or more in size in current libraries (Kim et al. 1996), this STS density should be adequate to obtain contiguous clone coverage of much of the genome; most gaps that remain should be closable by developing new STSs directly from the sequence adjacent to the gap and rescreening the library.

Restriction digests are performed on the clones obtained from the screens to determine their sizes and extent of overlap, and to eliminate anomalous clones, which generally have fingerprints inconsistent with other clones in the group. Selected clones are then sequenced using a two-stage strategy, consisting of a shotgun phase in which a number of reads are generated from random M13 or plasmid subclones, followed by a directed, or ''finishing'' phase. In the latter, the shotgun reads are assembled into contigs, the assembly is inspected and tested for correctness, additional data are collected to close gaps and resolve low-quality regions (e.g., compressions), and editing is performed to correct errors in assembly and to resolve discrepancies between reads and other data anomalies.

The amount of finishing effort required depends in part on the desired accuracy and completeness of the final sequence. In the case of the human genome, the goal that has been agreed upon by the U.S. funding agencies and essentially all of the major sequencing groups is a complete and highly accurate sequence with less than one error per 10 kb. There are several reasons for this target: The genome sequence should serve as a reference against which human variation can be cataloged, and consequently it should have an error rate substantially lower than the estimated polymorphism rate of one per kilobase; it should be accurate enough to permit genes to be identified and distinguished from pseudogenes, so only a minority of genes should have any errors in their coding regions (which average >1 kb in length); and it should be accurate enough to permit any region of the genome to be reliably obtained by PCR (in particular, gaps should be small, infrequent and of known size). Current experience

[1]E-MAIL phg@u.washington.edu; FAX (206) 685-7344.

7:410–417 ©1997 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/97 $5.00

indicates that this level of accuracy is attainable without unduly inflating the cost.

Not surprisingly (in view of the profound impact cloning has had upon molecular biology), a clone-based approach has important strengths. Clones provide modularity, which is a crucial consideration when analyzing something as large and complex as the human genome. In particular, they make it possible to target specific regions; to partition the project among multiple investigators without forcing them to interact with each other; to isolate problematic regions (e.g., repeats); and to adapt the sequencing strategy as needed in regions with unusual features (e.g., GC-richness, high repeat density). Importantly, clone-by-clone sequencing forces one to confront early on the issue of finishing and ensures that feedback regarding data quality is obtained quickly.

In addition, clones provide an important technical resource for sequencing. They permit efficient resequencing and gap-filling at the finishing stage, and make it possible to test the correctness of the assembly by means of restriction digests. Finally, because each clone represents a single haplotype, problems caused by the presence of polymorphisms are eliminated.

Whole-Genome Shotgun Sequencing

Weber and Myers propose whole-genome shotgun sequencing of the human genome as an alternative to clone-by-clone sequencing. Their approach would consist of a single whole-genome library construction and characterization phase (for the entire project), followed by a single shotgun phase, followed by a single finishing phase. In particular, finishing issues would not be addressed until fairly late in the project.

This is inherently a monolithic approach incompatible with clone-by-clone sequencing, and consequently it requires careful scrutiny. I will discuss a number of objections to it, but the most serious one is that for a variety of reasons (detailed below) the finishing stage has a high probability of failure; moreover, failure would not become evident until very late in the project when it would be too late to do anything about it. Even if the finishing could be made to succeed, it would almost certainly be much more expensive than in clone-by-clone sequencing, resulting in a significantly higher overall cost. As a result, other claimed advantages for the whole-genome shotgun (e.g., that it would yield a large supply of polymorphisms) become irrelevant, as that information could be obtained more cheaply by other approaches.

The prospect of obtaining an early, broad sampling of the genome with shotgun reads is at first sight appealing, and one might hope that even if the sequence could not be finished the read data itself would still be useful. However, it is clear upon reflection that unmapped genomic reads are an extremely inefficient way to obtain biological information and are virtually useless for most purposes. Essentially the only biologically interesting features recognizable from such reads would be exons with homology to previously known genes. However, only ~3% of the genome sequence is thought to code for protein; of this, a large fraction (at least half) has already been detected in the form of expressed sequence tags (ESTs) (Hillier et al. 1996) (which moreover have the advantage of being derived from clones with intact coding sequences); and a further large fraction is likely to lack detectable similarity to known genes (Green et al. 1993) and thus not even be reliably identifiable as a coding sequence. Moreover, when a homology is found, errors and incompleteness of the read sequence will make it impossible to tell without substantial additional work whether it represents a real gene or a pseudogene; and because it is unmapped, and the clone from which the sequence is derived would be unavailable (barring a massive clone-tracking effort), it would be useless for positional cloning efforts. For most molecular studies, sequence is useful only if it is reasonably accurate, mapped, and contiguous (and when this is not required, one gets most of what is needed from the EST databases). Sequence meeting these criteria will emerge much more slowly (if ever) from a whole-genome shotgun approach than from the clone-by-clone approach, which is already providing it.

Weber's and Myers' argument that the approach is feasible relies primarily on a greatly oversimplified computer simulation of the process of sequence reconstruction, which depends on incorrect assumptions about the nature of the genome (e.g., that repeats are uniformly distributed) and of sequence data and ignores a number of serious technical obstacles. It needs to be emphasized that what they have done was not an actual assembly of a simulated genome sequence; indeed, they could not do such an assembly, as software adequate to handle data on the required scale does not exist, nor do we have adequate knowledge of the sequence characteristics of the genome to permit a realistic simulation. Instead, they have idealized the process of assembly by simulating the *locations* of clones within

the genome (assuming they are randomly distributed), of reads within those clones, and of repeats, and then simply assuming that reads whose locations overlap would be correctly assembled together without difficulty unless the overlap occurs within a (simulated) repeat location, in which case forward–reverse read pair information could be used to determine the correct assembly. This procedure ignores the many complications that occur with real data and assumes that they will not cause disproportionate difficulties on the envisaged scale, which is highly questionable.

We simply do not know enough about the structure of the genome at the sequence level, or of the biases inherent in clone libraries, to simulate the sequencing process adequately. For the same reasons, it does not even seem possible to do a convincing pilot study to test the approach. Variation within the genome implies that library representation for any selected region cannot necessarily be extrapolated to the genome as a whole. Moreover, most finishing problems scale poorly with the size of the region being sequenced (see below), implying that no matter how large a region is tested it will not give an adequate picture of problems at the whole-genome level.

Nor does success of a whole-genome shotgun approach with bacterial genomes (Fleischmann et al. 1995; Bult et al. 1996) provide any confidence whatsoever that the same approach would work with the human genome. The fundamental difficulty in assembly is dealing with repeats, and bacterial genomes have very few of these. In addition, their size is such that an entire project can be completed within a year by a single laboratory, so logistical and data-quality monitoring issues are minimized. In contrast, the human genome is three orders of magnitude larger. Potential assembly problems caused by polymorphisms (see below) are not an issue with bacterial genomes. It is also worth noting that clone tracking, which appears infeasible for the whole human genome shotgun, was performed in the whole bacterial genome projects and played an important role in gap closure.

The arguments for the feasibility of the whole-genome approach are thus not persuasive. In contrast, there are a number of significant arguments against it, which are detailed below.

Finishing Issues

Finishing is the most difficult aspect of sequencing, because of the wide variety of problems encountered and the level of technical expertise required to deal with them. Careful consideration of finishing issues suggests that finishing would be much more difficult and expensive with a whole-genome shotgun than in clone-by-clone sequencing.

Gap-Filling and Other Finishing Data Collection

The accuracy requirements for the genome sequence entail that there be read coverage on both strands essentially everywhere and that regions of low data quality be resolved by the collection of additional data (e.g., dye-terminator reads to resolve compressions). In clone-by-clone sequencing, these criteria are met by retrieving relevant clones during the finishing phase to use as templates for additional data collection (e.g., primer walking), which in turn requires tracking all subclones during the shotgun phase because it is not known in advance which ones will be required. This is not particularly onerous or expensive, as they can be discarded as soon as the clone sequence is finished.

In contrast, clone tracking for the whole-genome shotgun would involve ~50 million clones (at multiple laboratories), because finishing is not done until the end of the project. This is impractical, and consequently Weber and Myers propose that all additional data collection at the finishing stage be done instead by the sequencing of PCR products. That approach has already been tried in clone sequencing and has been found to be significantly more expensive and less reliable than going back to the subclones. Sequencing of PCR products has higher reagent costs, is technically more demanding, yields lower data quality, and has a much higher failure rate than subclone-based sequencing. In a whole-genome approach the situation would be substantially worse, because of the fact that one would be amplifying by PCR from the entire genome rather than from a cosmid or BAC clone. Many gaps would be fairly large, and thus not capable of easy amplification, particularly as the choice of priming sites is constrained by the requirement that they lie in single-copy sequence. Because genomic PCR from a repeated region will amplify all copies of the repeats simultaneously, gap-filling within large, even moderately similar repeated regions would be extremely difficult, if not impossible. Even in nonrepeated regions genomic PCR is highly variable. These facts would inflate the cost of finishing enormously, relative to the clone-by-clone approach, and most likely there would be many failures, resulting in a final product of seriously degraded accuracy.

The amount of additional finishing data re-

quired will be substantial. In a 10× shotgun, the average coverage of each strand is only 5×, which results in an average gap frequency on each strand of about one per 15 kb (assuming 500-base reads), or every 7.5 kb for the two strands combined. This would require (in the whole-genome approach) on the order of 450,000 PCR sequencing reactions (assuming a perfect success rate!), not including sequencing to resolve compressions and other low-quality regions, which could easily double that number. Moreover, the distribution of clone locations in real libraries is never random; there are hot spots and cold spots. GC content and the nature and distribution of repetitive DNA (among other, unknown factors) appear to play a significant role in this, and the human genome with its wide variations in GC content and repeat density (for review, see Bernardi 1995) is likely to be represented unevenly in any given clone library. This is potentially quite serious, as it means that some regions are likely to have a low depth of coverage and thus have many more and larger gaps than predicted. Apparently unclonable regions have been found with most libraries, with significant regions failing to clone altogether. Complete absence from one cloning system as has been seen in the *Caenorhabditis elegans* project (Waterston and Sulston 1995), for example, creates severe problems for finishing.

The ability to go back to subclones is thus a significant advantage of the clone-by-clone approach, and the inability to do so with the whole-genome shotgun approach is a major disadvantage. Subclones often divide an otherwise intractable problem, allowing walking or other approaches that would be impossible if one were working on the whole clone, let alone the genome.

Repeats

Weber's and Myers' simulations assume that all repeats are members of known families (and thus relatively small in size) and are randomly distributed in the genome. However, repeat density varies widely, with some repeats (especially *Alu* repeats) often occurring in (apparently nonrandom) clusters. These can be quite difficult to sort out and could significantly affect Weber's and Myers' conclusions regarding assembly because they are often not spannable by a single read or by a forward–reverse pair from a plasmid. More seriously, there are numerous examples of ''local'' duplicated regions, which are not members of known families and vary widely in size (some extending over tens or hundreds of kilobases or more), evolutionary age (some very recent ones being >99.9% identical), and physical separation of the copies (from being immediately adjacent to being on different chromosomes). Some of these can cause problems for any approach, but the difficulties would be much worse with a whole-genome shotgun; with a clone-based approach one can often separate copies of the repeat into distinct clones, which then eliminates them as problems.

It should be emphasized that assembly in the presence of repeats is not a solved problem even at the single-cosmid scale and may require specialized data collection strategies as well as a significant amount of skilled editing; and that the larger the scale of the region being shotgunned, the worse the problem is—because the more likely one is to encounter large, near-perfect copies within the region. The complications caused by repeats go up roughly quadratically in the number of repeats in the region being assembled. Thus, they are already significantly worse for BACs than for cosmids, and would very likely be insuperable for the entire genome. Moreover, one would lack the most important resources (namely access to subclones of known location) necessary to solve them.

Polymorphisms

Having to deal with polymorphisms in the assembly presents significant problems for a whole-genome approach. The fundamental issue in assembly and editing is sorting out whether read discrepancies are the result of base-calling errors, of the presence of different repeats, or of cloning anomalies or other data artifacts. One generally can eliminate base-calling errors and clone anomalies relatively easily (and automatically), as they tend not to be confirmed by other reads, so the main problem is in detecting and resolving repeats. This is already difficult enough; adding in the complication of polymorphic differences (which may include the presence or absence of repeats or other DNA segments, in addition to simple single-base or microsatellite differences) makes the problem that much worse. With a clone-based approach one knows that similar but discrepant reads from a given clone are not allelic, because a single haplotype is represented in any one clone. In contrast, with the whole-genome shotgun method one will always have to consider two possibilities: that the reads are from different haplotypes, or that they are from different copies of a repeated sequence. Moreover, the rule of thumb, that an unconfirmed sequence feature is probably a data or clone error and can thus be ignored will no longer be valid, because with a 10× shotgun pre-

pared from multiple individuals (as Weber and Myers propose) it will often be the case that a given haplotype is represented only once at a particular site. Even when these issues can be sorted out (and it is not at all clear that they can), it will require an enormous increase in finishing effort to do so.

Data Anomalies

There are a number of anomalies of various types in real data sets that can cause problems in assembly. These include chimeric reads, which may arise either from chimeric or internally deleted clones or from gel-mistracking errors; low-quality reads; and mistracking or mislabeling of gel lanes, resulting in uncoupling of forward and reverse read pairs. There are at least three reasons why these are likely to be more problematic in a whole-genome shotgun than in a clone-by-clone approach. First, as noted above, the clone anomalies will be confounded with polymorphisms. Second, the potential for false joins will be far greater, because the number of opportunities (potential overlaps that must be considered) is far greater. Third, and most seriously, these errors are not easily detected prior to the assembly phase. In a clone-by-clone approach this assembly occurs quickly, when one still has access to the original subclones, and can determine relatively easily the source of the error, and can take steps to reduce the error rate with future clones. In the whole-genome shotgun approach, it would not occur until late in the project. Errors of all types are thus likely to invade the process during the shotgun phase, because they would not be detected for several years. A massive lane-mislabeling problem could easily escape detection until it was too late to do anything about it, and it would have drastic implications, as faithful whole-genome assembly in the presence of repeats is clearly impossible without reliable read–pair information.

Even on a small scale some of the above issues occasionally cause significant problems; on a large scale they are likely to be much worse, raising serious doubts whether the whole-genome approach is feasible. What is worse, one would not actually discover the extent of these problems until late in the project, when assembly and finishing commence. At this point, it would be too late to change strategy, and the entire project would have to be junked. A great advantage of the clone-by-clone approach is that the problems are confined to individual clones, and are detected early. The process of finishing gives one the best indication of potential problems with data quality, unusual sequence features, or library quality (e.g., chimera frequencies, cloning bias) and of whether the strategy requires modification. In a whole-genome approach one would not get this feedback until it was too late to do anything about it.

Cost

The claim that a whole-genome shotgun approach would be less expensive than clone-by-clone sequencing apparently is based on an assumption that clone mapping, making and tracking subclones, and inefficiencies caused by clone overlaps represent a major component of the cost of sequencing. This is false. Consideration of where the real costs of sequencing lie and of how these are likely to differ in the two approaches suggests that the whole-genome shotgun approach would be more expensive than clone-by-clone sequencing, quite possibly by a factor of two or more.

In the most efficient current clone-by-clone operations, the costs break down roughly as follows: <10% for clone mapping and subclone library preparation (these are inexpensive compared to the sequencing itself, because they need be carried out only once every 100 kb or so); 60%–70% for the shotgun phase; and 30%–40% for finishing. Reads from clone overlap regions and the cloning vector inflate the shotgun cost by perhaps 15% (such regions are not finished so they do not contribute to the finishing costs), and thus increase the total sequencing cost by ~70% × 15%, or only 10%. This can be reduced even further by carefully choosing clones from a well-mapped high-depth coverage map.

The cost of mapping and subclone library preparation (which is minimal in any case) would be partly eliminated in the whole-genome approach, but not entirely: There would need to be very extensive up-front testing of the whole genome λ and plasmid libraries with regard to chimera rates, rearrangements, insert sizes, and uniformity of genome coverage. This is critical, as the entire project depends on the integrity of these libraries, and it would be nontrivial. Such extensive testing of the subclone libraries in the clone-by-clone approach is unnecessary because feedback concerning them is obtained rapidly from the sequence assembly itself.

The important issue is therefore the shotgun and finishing costs for the two approaches. There is a tradeoff in shotgun vs. finishing effort: The higher the depth of the shotgun, the fewer the number of gaps that need filling. This tradeoff is not entirely simple, as coverage is not truly random, with some

gaps remaining despite a high depth of coverage; but it implies that the appropriate shotgun depth depends on the cost of gap filling relative to shotgun reads, which in turn depends on the gap-filling strategies that are available. The cost of closing a gap is substantially higher than the cost of a shotgun read, because it requires individual attention using a variety of specialized methods that involve more expensive reagents (e.g., alternative chemistries, custom primers) and have a higher failure rate. Shotgun depth in the clone-by-clone approach varies substantially between groups (depending on their preferred strategies), but typically is in the range 6×–8×. In the case of the whole-genome shotgun, where the gap-filling cost will be higher because of the complete reliance on PCR-based methods, Weber and Myers propose a higher shotgun depth of 10×. This appears unavoidable because with a lower coverage the difficulty of amplifying by PCR across gaps would become prohibitive.

The cost of the shotgun is essentially directly proportional to the number of reads that need to be obtained. It would be significantly higher in the whole-genome shotgun approach, because of two factors that increase the required number of reads. First, a higher depth of coverage (10×, vs. ~7×); as 70% × (10×/7×) = 100%, this factor alone ensures that the cost of the whole-genome shotgun raw data generation (without any finishing) will at least equal the entire cost of the clone-by-clone approach! Second, the whole-genome approach relies on double-stranded sequencing from λ and plasmid subclones, which although it certainly can yield reasonable data, on average (in most laboratories), has a higher failure rate, lower data quality, and shorter read length than sequencing of single-stranded M13 templates. As a result, the number of reads required for a given depth of coverage is higher than for a clone-by-clone approach based on M13 shotguns.

Apart from the above considerations, two other factors would further inflate the raw data collection costs for the whole-genome shotgun relative to the clone-by-clone approach.

First, it appears that accurate sizing of the λ inserts may be required to position contigs relative to each other, which is a minimal requirement when the finishing is entirely PCR-based. The issue is the following: Although Weber's and Myers' estimated contig sizes are on the order of 200 kb for a 10× shotgun (assuming random coverage), the contig size provided by the initial assembly will be much smaller, because any repeat that is not completely spanned by a read (or by a pair of overlapping forward–reverse reads from a plasmid) will produce an ambiguity that effectively terminates a contig. Given the density of repeats in the genome, many contigs will be 1 kb or less in size, consisting of the single-copy sequence between two repeats. The forward–reverse read pairs from the λ clones will often permit orienting such contigs with respect to each other; however, when there are two or more adjacent small contigs (which will often be the case), the only apparent way to order them with respect to each other would require fairly precise knowledge of the distance between the forward reverse read pairs (e.g., to order two 1-kb contigs with respect to each other one may need to determine the λ insert size within 1 kb.) Such information also seems necessary to assemble across the repeats reliably, as when one of the two reads from a λ clone lies within a repeat, there would be ambiguity about exactly which copy of the repeat it was.

Because insert sizes in λ clones vary substantially it would be necessary to obtain fairly precise insert size information by gel analysis. This is certainly doable, but because (absent tracking) it would need to be done for all of the λ clones it would add significantly to the cost of the project. Detecting 1-kb differences in λ molecules of ~50 kb is nontrivial; one would presumably have to do restriction digests and/or long-range PCR, and it is not clear that a single lane per λ clone would be adequate. This increases by 25% both the number of enzymatic reactions and the number of gel lanes for the project. (Although restriction fragment sizing also needs to be done for BACs in the clone-by-clone approach, the number of λ clones is 2 orders of magnitude larger because there is one λ clone per pair of reads, versus 10 BACs per 100 kb.) Note also that the presence of length polymorphisms or internally deleted clones would complicate the analysis significantly.

Second, the whole-genome approach requires that essentially all of the raw data be collected in the first part of the project. In contrast, with the conventional approach raw data collection will occur over the entire course of the project. As a result, the advances in sequencing technology that are anticipated to come on line over the next few years—including machines with much higher throughput, capillary electrophoresis (Kheterpal et al. 1995), chemistry improvements, automation advances, miniaturization technology—would occur too late to have much impact on the cost of a whole-genome shotgun but would potentially lower the cost of the conventional approach substantially. It is true that dramatic improvements in finishing technology would have the converse effect; but

(apart from improvements in software that have largely already been implemented) it is hard to see where these would come from, and in any case finishing is a smaller part of the overall costs, so the impact would be smaller.

Thus, the cost of the shotgun phase for the whole-genome approach would likely be significantly higher than the cost of the shotgun phases for clone-by-clone sequencing; similarly (even more so!) with finishing, for the reasons indicated previously. Because these constitute the great bulk of sequencing costs, a whole-genome approach would almost certainly be substantially more expensive.

Other Issues

In addition to the above objections, the whole-genome shotgun approach poses daunting logistic challenges. Each phase occurs separately and involves very different skills and different numbers of people. The shotgun phase for example requires relatively unskilled technical labor, whereas the finishing phase requires considerable experience in judging and manipulating DNA sequence data. It is not clear how one would deal with hiring, training, and laying off the relevant people on the massive scale required. It also is quite unclear how the project could be distributed among several laboratories: Problems with data quality in one laboratory would affect all laboratories, because any region of the genome would have shotgun reads generated at all labs.

Weber and Myers claim that the whole-genome shotgun method will avoid cloning artifacts. This advantage is entirely theoretical. There are no data to indicate that such artifacts are a significant problem with the conventional approach, provided one takes the precaution of requiring that any sequenced clone have a fingerprint that is consistent with other independent clones from the same region. Moreover, there are no data to indicate that (in the absence of fingerprinting) artifacts would not cause problems for the whole-genome shotgun method.

Although the whole-genome shotgun approach would undoubtedly yield many polymorphisms, this advantage is negated by the likely higher cost of the sequence itself. Given our prediction that the whole-genome approach would be significantly more expensive than clone-by-clone sequencing, it would be cheaper overall to obtain the reference sequence clone by clone, and then identify polymorphisms by doing an ~3× M13 shotgun of the genome from a mixture of other individuals and comparing these reads to the reference. This will also avoid the many problems that are caused when the assembly of the original sequence itself includes polymorphic reads. Even without such an effort, many polymorphisms will be automatically identified in the clone-by-clone approach by virtue of being present in clone overlaps involving clones from different haplotypes.

In summary, clone-by-clone sequencing works and is cost-effective, neither of which appears likely for the whole-genome shotgun method of sequencing. There is no reason to switch.

ACKNOWLEDGMENTS

This work was partly supported by grants from the National Human Genome Research Institute and the Department of Energy. I thank Bob Waterston and Maynard Olson for helpful comments on an earlier draft.

REFERENCES

Bernardi, G. 1995. The human genome: Organization and evolutionary history. *Annu. Rev. Genet.* 29: 445–476.

Bouffard, G.G., L.M. Iyer, J.R. Idol, V.V. Braden, A.F. Cunningham, L.A. Weintraub, R.M. Mohr-Tidwell, D.C. Peluso, R.S. Fulton, M.P. Leckie, and E.D. Green. 1997. A collection of 1814 human chromosome 7-specific STSs. *Genome Res.* 7: 59–64.

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 et al. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273: 1058–1073.

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 et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd.* *Science* 269: 496–512.

Green, P., D. Lipman, L. Hillier, R. Waterston, D. States, and J.M. Claverie. 1993. Ancient conserved regions in new gene sequences and the protein databases. *Science* 259: 1711–1716.

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.B. Soares, F. Tan, J. Thierry-Mieg, 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 et al.

1995. An STS-based map of the human genome. *Science* 270: 1945–1954.

Kheterpal, I., J.R. Scherer, S.M. Clark, A. Radhakrishnan, J.Y. Ju, C.L. Ginther, G.F. Sensabaugh, and R.A. Mathies. 1996. DNA sequencing using four-color confocal fluorescence capillary array scanner. *Electrophoresis* 17: 1852–1859.

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

Nagaraja, R., S. MacMillan, J. Kere, C. Jones, S. Griffin, M. Schmatz, J. Terrell, M. Shomaker, C. Jermak, C. Holt et al. 1997. X chromosome map at 75-kb STS resolution, revealing extremes of recombination and GC content. *Genome Res.* 7: 210–222.

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

Waterston, R. and J. Sulston. 1995. The genome of Caenorhabditis elegans. *Proc. Natl. Acad. Sci.* 92: 10836–10840.