

WHOLE-GENOME DNA SEQUENCING

Computation is integrally and powerfully involved with the DNA sequencing technology that promises to reveal the complete human DNA sequence in the next several years. After introducing the latest DNA sequencing methods, this article describes three current approaches for completing the sequencing.

The prevailing method of determining the sequence of a long DNA segment is the *shotgun sequencing* approach, in which a random sampling of short fragment sequences is acquired and then assembled by a computer program to infer the sampled segment's sequence. In the early 1980s, such segments were typically on the order of 5,000 to 10,000 base pairs (5 to 10 kbp). By 1990, this method was sequencing segments on the order of 40 kbp, and by 1995, the entire 1,800-kbp *H. Influenzae* bacterium had been sequenced.¹ The source segment is clearly becoming extremely large without any concomitant increase in the length of the sampled fragment sequences. Soon, shotgun data sets could well consist of millions of sampled fragment sequences and will require significant computational resources to assemble.

The whole-genome shotgun sequencing of *H. Influenzae* in 1995 showed that direct shotgun sequencing could handle a much larger source sequence segment than biologists had commonly thought. Before that, cosmid-sized clones of 30 to 50 kbps were considered this approach's up-

per limit. (A *cosmid* is a type of vector for manipulating and replicating inserted pieces of DNA.) Now, the shotgun sequencing of 200 kbp *bacterial artificial chromosomes* (BACs) is a given. This achievement inspired Jim Weber and me to propose the use of a shotgun approach to sequence the human genome,² after which we requested funding for a pilot project from the US National Institutes of Health. The established community rejected our controversial proposal,³ but in May of 1998, Craig Venter and the Perkin-Elmer Corporation announced a new private venture, Celera Genomics, aimed at using a whole-genome shotgun approach to sequence the fruit fly *Drosophila* (≈ 120 Mbp) in 1999 and the human genome (≈ 3.5 Gbp) by 2001.⁴

After introducing the basic technology of shotgun DNA sequencing and briefly summarizing the computational and algorithmic results to date on the problem of assembling shotgun data sets, this article characterizes the nature of DNA sequences and improvements in sequencing technology that affect the computational problem. It then analyzes three current proposals for sequencing the human genome, including the one we're pursuing at Celera Genomics.

1521-9615/99/\$10.00 © 1999 IEEE

GENE MYERS
Celera Genomics

Shotgun DNA sequencing basics

In the late 1970s, the first procedures emerged for determining the sequence of nucleotides

along a given DNA strand.^{5,6} In the version used most commonly today, a biochemical sequencing reaction produces a collection of geometrically distributed copies of every prefix of the given DNA strand such that the last nucleotide or base—A, C, G, or T—of each prefix is known. In a process called *gel electrophoresis*, this material passes through a permeable gel under an applied voltage, which separates the prefixes in order of length, letting either a technician or a combination of a laser, charge-coupled device detector, and software determine the sequence of nucleotides along one end of the source strand. How much of the source we can determine is limited by the fact that the size ratio between consecutive prefixes approaches 1 and the number of long prefixes is diminishing geometrically. With today's technology, biologists can resolve on average the source's first 500 nucleotides and upwards of 800 to 900 bases for a particularly clean reaction. The result of such a sequencing experiment is called a *read*. Over the last 20 years, special machines and robots have been developed that automate much of this process.

To determine the sequence of much longer stretches of DNA, Frederick Sanger and his colleagues devised the *shotgun DNA sequencing* strategy.⁷ This approach entails sampling DNA fragments as randomly as possible from the source sequence and then producing a sequencing read of the first 300 to 900 bases of one end of each fragment. To maximize the sequence produced from each fragment, such experiments involve sampling fragments whose length is longer than a read's maximum expected length. If enough fragments are sequenced and their sampling is sufficiently random across the source, the process should let us determine the source by finding sequence overlaps among the reads of fragments that were sampled from overlapping stretches. This basic shotgun approach is at the heart of all current approaches to genome sequencing.

As currently practiced in many DNA sequencing centers, the basic shotgun protocol starts with a pure sample of a large number of copies of the source DNA whose sequence is to be determined, typically a segment of 100 kbp or longer.

1. Technicians randomly fracture the sample either using sound (*sonication*) or passing it through a nozzle under pressure (*nebulation*), which produces a uniformly random partitioning of each copy of the source strand into a collection of DNA fragments.

2. To remove fragments that are too large or too small, this pool of fragments is size-selected, typically using size separation under gel electrophoresis and then simply excising a band of the gel containing the desired size. With care, this procedure produces a normally distributed collection of fragment sizes with a 10% variance.
3. The technicians then insert the size-selected fragments into the DNA of a genetically engineered bacterial virus (phage), called a *vector*. Usually, at most one fragment is inserted at a predetermined point, called the *cloning site*, in the vector. Typically, the number of vectors where more than one fragment gets inserted is less than 1%, but can be as low as 0.01% for some meticulously executed protocols. The fragments at this point are often called *inserts* and the collection of inserts is a *library*.
4. A bacterium is then infected with a single vector, which reproduces to produce a bacterial colony containing millions of copies of the vector and its associated insert. The procedure thus has effectively *cloned* a pure sample of the given insert. This procedure repeats simultaneously for as many inserts as desired for sequencing in the final step.
5. By design, the vector then permits a sequencing reaction to be performed, starting just to the left or right of a source fragment's insertion point. The sequencing reaction produces a read of the first 300 to 900 bases of one end of the insert.

A key failure in this process occurs if the sampled reads are not randomly sampled but biased to come from particular regions of the source. This can happen for three reasons: the fracturing of the fragments might be biased, the insertion of fragments into vectors might be biased, or some insert/vector combinations might not clone properly because the insert has reacted toxically with the host/vector environment. Anecdotal evidence suggests that the first two biases are minimal in well-performed experiments, but the third bias definitely exists. Picking host/vector combinations for which the insert DNA will be relatively inert will reduce this toxicity bias.

Sequencing reactions tend to fail for a variety of reasons. In a production context, investigators consider a 70 to 80% success rate to be a very good yield. In initially processing the sequencing information, technicians must screen these failed reactions and also screen reads from vec-

The fragment-assembly problem

Given the reads obtained from a shotgun protocol, the computational problem, called *fragment assembly*, is to infer the source sequence given the collection of reads. For the purposes of illustration, we might parameterize a typical problem occurring in practice today as follows (see the “Definitions” box for help with the terminology). For a source strand of length $G = 100$ Kbp, we would then typically sequence $R = 1,500$ reads of average length $\bar{L} = 500$. Thus, we would collect altogether $N = R\bar{L} = 750$ Kbps of data, so that we have sequenced on average every base pair in the source $\bar{c} = N/G = 7.5$ times. The quantity \bar{c} is the average sequencing coverage, and practitioners say that the source has been sequenced to 7.5X coverage. In practice, an investigator will decide on a given level of coverage and then sequence inserts until a total of $N = G\bar{c}$ base pairs of data have been collected. Software for fragment

assembly must account for the following essential characteristics of the data:

- *Incomplete coverage*: Not every source base pair is sequenced exactly \bar{c} times due to both the stochastic nature of the sampling and cloning bias I’ve mentioned. Some portions of the source might be covered by more than \bar{c} reads, and others might not be covered at all. In general, there can be several such *gaps* or maximal contiguous regions where the source sequence has not been sampled. Gaps necessarily dictate a fragmented, incomplete solution to the problem.
- *Sequencing errors*: The gel-electrophoretic experiment yielding a read, like most physical experiments, is prone to error, especially near the end of a read where the signal strength and separation of consecutive prefix fragments become small. In a very stringently

controlled production environment, the error rate is less than 1% in the first 500 or so bases. Thereafter, the error rate increases rapidly, reaching more than 15% from 650 to 900 bases into the read, after which the resulting sequence is effectively unusable. (However, I have seen data sets where an error rate of 5% occurs in the “sweet” part of the read, consisting of the first 500 bases.)

- *Unknown orientation*: DNA is a double-stranded helix. Which of the source sequence’s two strands is actually read depends on the arbitrary way the given insert orients itself in the vector. Thus we do not know whether to use a read or its Watson-Crick complement in the reconstruction. The Watson-Crick complement $(a_1 a_2 \dots a_n)^c$ of a sequence $a_1 a_2 \dots a_n$ is $wc(a_n) \dots wc(a_2) wc(a_1)$ where $wc(A) = T$, $wc(T) = A$, $wc(C) = G$, and $wc(G) = C$.

I will now develop a mathematical formulation of the fragment-assembly problem. For input, we have a collection of reads

$$\mathcal{F} = \{f_i\}_{i=1}^R$$

that are sequences over the four-letter alphabet $\Sigma = \{A, C, G, T\}$. An ε -layout is a string S over Σ and a collection of R pairs of integers, $(s_i, e_i)_{i \in [1, R]}$, such that

- if $s_i < e_i$ then f_i can be aligned to the substring $S[s_i, e_i]$ with less than

Definitions

| | |
|-----------|---|
| G | Length of target sequence |
| \bar{L} | Average length of sequence read |
| R | Number of sequencing reads in shotgun data set |
| N | $R\bar{L}$, total number of base pairs sequenced |
| \bar{I} | Average length of a clone insert |
| \bar{c} | N/G , average sequence coverage |
| \bar{m} | $R\bar{I}/2G$, average clone or map coverage |

tors where no insert occurred. Moreover, because the sequencing reaction begins in the vector at one end of the insert location or the other, the initial part of a read can consist of the vector DNA sequence leading up to the beginning of the insert. This bit of vector sequence must be carefully identified and removed. Similarly, if an insert is particularly short, the technicians might need to trim vector sequence from the end of a read. After taking these steps, the process will have produced a set of sequence reads randomly sampled from the source sequence.

See the sidebar, “The fragment-assembly problem,” for a discussion of the computational problem associated with shotgun sequencing.

DNA sequence characteristics

As we’ve seen, when researchers first began employing shotgun sequencing in the early 1980s, a typical source sequence size was 5 to 10 kbp. By 1990, they were shotgun-sequencing cosmid-sized sources for which $G \approx 40$ kbp, and in 1995 the bacteria *H. Influenzae* of length 1.8 Mbp was successfully shotgun sequenced. In the past three years, 20 bacterial genomes in this size range have been shotgun-sequenced. In August 1998, Celera Genomics was formed to shotgun-sequence the entirety of the fruit fly *Drosophila* in 1999 ($G \approx 120$ Mbp) and the human genome by 2001 ($G \approx 3.0$ Gbp).

With the trend toward sequencing higher or-

- $\varepsilon |f_i|$ differences, and
- if $s_i > e_i$, then f_i can be aligned to the substring $S[e_i, s_i]^c$ with less than $\varepsilon |f_i|$ differences, then
- $\cup_i [\min(s_i, e_i), \max(s_i, e_i)] = [1, |S|]$.

The string S represents the reconstruction of the source strand, and the integer pairs indicate the substrings of S that gave rise to each read. The order of s_i and e_i encode the *orientation* of the fragment read in the layout—that is, whether f_i was sampled from S or its complement strand. The parameter $\varepsilon \in [0, 1]$ models the maximum error rate of the sequencing process.

The set of ε -layouts models the set of all possible solutions to the fragment-assembly problem. Of course, there are many such solutions, so the computational problem is to find one that is in some sense best. Traditionally, the fragment-assembly problem has been phrased as one of finding a shortest common superstring (SCS) of the fragment reads within error rate ε ; that is, find an ε -layout for which S is as short as possible. Unfortunately, as Figure A illustrates, this appeal to parsimony often produces over-compressed results when the source sequence contains repeated subsegments. This tendency has prompted the proposal of maximum-likelihood criteria based on the distribution of fragment start points in the layout.¹ While such a criteria provides a better objective function, algorithm designs for computing it have proven elusive.

A common computational architecture for fragment assembly, advocated by several authors,^{2–4} divides the problem into three phases: *overlap*, *layout*,

and *consensus*. The overlap phase compares every fragment read against every other read (in both orientations) to determine if they overlap. Given the presence of sequencing errors, an overlap is necessarily approximate in that not all characters in the overlapping region coincide. This problem is a variation on traditional sequence comparison where the degree of difference permitted is bounded by ε . The best deterministic designs for finding all ε -overlaps lets us solve problems on the order of $N = 1$ to 5 Mbp in a matter of minutes on a typical workstation.⁵ For contexts requiring even greater speed, most investigators resort to heuristics that detect overlapping reads by finding exact common substrings of some length k using a hashing scheme. Typically, they choose k to provide the best compromise between sensitivity and

speed for a given N and ε . Conceptually, we can think of the result of the overlap phase as producing an *overlap graph* in which every vertex models a read and every edge an ε -overlap between two reads.

The layout phase determines the pairs (s_i, e_i) that position every fragment in the assembly. In graph theoretic terms, we accomplish this by selecting a spanning forest of the overlap graph; such a subset positions every fragment with respect to every other, transitively, through the overlaps on the path between them. Finding a spanning forest that optimizes a criterion such as shortest or most likely is known to be NP-hard.⁶ Investigators have proposed greedy algorithms that come within a given factor of optimal,^{7,8} simulated annealing⁹ and genetic algorithms,¹⁰ relaxation methods based on generat-

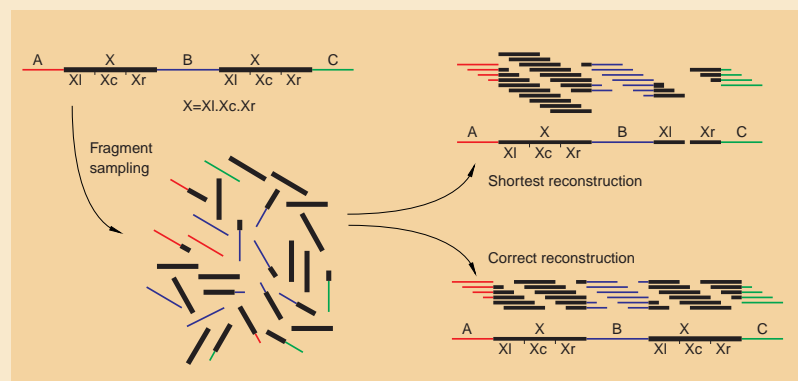


Figure A. The shortest answer isn't always the correct one. A DNA source at the upper left consists of unique stretches A, B, and C separated by a repeated sequence X. Below it, the source has been sampled perfectly uniformly across the target, as evidenced by the correct reconstruction of the pieces shown at lower right. But note the result in the upper right of a program that produces the minimum-length reconstruction. The interior portion Xc of the repeat sequence, which is covered only by reads completely interior to X, is overcompressed.

ganisms (which have an extensive repeat structure not found in lower-order organisms) and toward larger and larger source sizes, investigators commonly see several repetitive substrings in a source sequence of even moderate size. Before 1990, this was rarely considered an impediment to sequencing as it was practiced then, but it is now clearly a major computational difficulty. Repeats occur at several scales. For example, in

the human T-cell receptor locus, there is a five-fold repeat of a trypsinogen gene that is 4 kbp long and that varies 3 to 5% between copies. Three of these were close enough together that they appeared in a single shotgun-sequenced cosmid source.⁸ Such large-scale repeats are problematic for shotgun approaches because reads with unique portions outside the repeat cannot span it. Smaller repeated elements such

ing either spanning forests or weighted matchings in order of score,⁴ problem simplification by chordal graph collapsing,¹ and a reduction to greedy Eulerian tour.¹¹ Ultimately, the complicating factor is the presence of repeated strings within the source, which has led to the use of quality values assessing the accuracy of each base in a read, in an attempt to distinguish ε -overlaps that are true from those induced by repeats. Currently, such an edge discriminator coupled with the basic greedy algorithm is employed in the most widely used *phrap* program.¹²

Finally, the consensus phase forms a consensus-measure multiple alignment of the reads in all regions where the coverage is two or greater. The resulting consensus character for each position of the multiple alignment gives the ultimate reconstruction *S*. Like pairwise sequence comparison, sequence multiple alignment has been extensively studied. In most formulations, investigators start with the initial multiple alignment obtained by pairwise merging the alignments between reads using the overlaps selected for the spanning forest by the overlap stage.¹³ They then refine this initial multiple alignment using either a window-sweep optimization, a Hidden-Markov model gradient-descent algorithm,¹⁴ or round-robin realignment.¹⁵

Before we return to the discussion of sequencing, it behooves us to appreciate some statistics of shotgun sampling. In an analysis that is essentially the dual of that for packet collision on an Ethernet (as here we want packets to collide), Michael Waterman and Eric Lander determined that if

sampling were perfectly uniform, we should expect to see¹⁶

- $1 - e^{-\bar{c}}$ of the source strand covered by some read,
- $Fe^{-\bar{c}}$ gaps in the coverage of the source,
- gap-free segments or *contigs* of average length $(\bar{L}/\bar{c})e^{\bar{c}}$, and
- gaps of average length \bar{L}/\bar{c} .

There are several interesting things to note about these results. First, the percentage of the genome covered depends only on \bar{c} and not on the size of the reads or length of the source. Second, the number of gaps rises to a maximum at $\bar{c} = 1$ and declines with an exponentially vanishing tail thereafter. Contig lengths rise exponentially in \bar{c} , and gaps quickly become very small.

References

1. E. Myers, "Toward Simplifying and Accurately Formulating Fragment Assembly," *J. Computational Biology*, Vol. 2, No. 2, 1995, p. 275–290.
2. H. Peltola, H. Soderlund, and E. Ukkonen, "SEQAID: A DNA Sequence Assembly Program Based on a Mathematical Model," *Nucleic Acids Research*, Vol. 12, No. 1, pp. 307–321.
3. X. Huang, "A Contig Assembly Program Based on Sensitive Detection of Fragment Overlaps," *Genomics*, Vol. 14, 1992, pp. 18–25.
4. J. Kececiloglu and E. Myers, "Exact and Approximate Algorithms for the Sequence Reconstruction Problem," *Algorithmica*, Vol. 13, Nos. 1–2, 1995, pp. 7–51.
5. E. Myers, "A Sublinear Algorithm for Approximate Keyword Matching," *Algorithmica*, Vol. 12, Nos. 4–5, 1994, pp. 345–374.
6. J. Turner, "Approximation Algorithms for the Shortest Common Superstring Problem," *Information and Computation*, Vol. 83, 1989, pp. 1–20.
7. J. Tarhio and E. Ukkonen, "A Greedy Approximation Algorithm for Constructing Shortest Common Superstrings," *Theoretical Computer Science*, Vol. 57, 1988, pp. 131–145.
8. A. Blum et al., "Linear Approximation of Shortest Superstrings," *J. ACM*, Vol. 41, No. 4, 1994, pp. 630–647.
9. C. Burks et al., "Stochastic Optimization Tools for Genomic Sequence Assembly," *Automated DNA Sequencing and Analysis*, M.D. Adams, C. Fields, and J.C. Venter, eds., Academic Press, New York, 1994, pp. 249–259.
10. R. Parsons, S. Forrest, and C. Burks, "Genetic Algorithms for DNA Sequence Assembly," *Proc. First Conf. Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, Calif., 1993, pp. 310–318.
11. R. Idury and M.S. Waterman, "A New Algorithm for Shotgun Sequencing," *J. Computational Biology*, Vol. 2, No. 2, 1995, pp. 291–306.
12. B. Ewing et al., "Base-Calling of Automated Sequencer Traces Using *phred*: Accuracy Assessment," *Genome Research*, Vol. 8, No. 3, 1998, pp. 175–185.
13. D. Feng and R. Doolittle, "Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees," *J. Molecular Evolution*, Vol. 25, No. 4, 1987, pp. 351–360.
14. A. Krogh et al., "Hidden Markov Models in Computational Biology," *J. Molecular Biology*, Vol. 235, No. 5, 1994, pp. 1501–1531.
15. E. Anson and E. Myers, "ReAligner: A Program for Refining DNA Sequence Multialignments," *J. Computational Biology*, Vol. 4, No. 3, 1997, pp. 369–383.
16. E.S. Lander and M.S. Waterman, "Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis," *Genomics*, Vol. 2, No. 3, 1988, pp. 231–239.

as Alus that are small retrotransposons of length approximately 300 bp do not share this feature but are still problematic because they cluster and can constitute up to 50 or 60% of the source sequence, with copies varying from 5 to 15% between each other.^{9,10} Finally, in telomeric and centromeric regions, microsatellite repeats of the form x^n are common.⁹ The repeated "satellite" x is three to six bases long, n is very large, and

the motif has 1 to 2% variation within it.

Repeats have three characterizing dimensions: length, copy number, and fidelity between copies. As the examples above demonstrate, repeats found in DNA cover a wide range along each of these dimensions. From a computational perspective, it is the long, high-fidelity repeats of low copy numbers that cause the greatest difficulty. On a whole-genome scale, the problem

initially looks quite daunting. For example, consider human DNA. It contains a number of ubiquitous repeats such as the Alu above and the longer LINE (long interspersed nucleotide element) elements that have an average length of 1,000 base pairs. The human genome contains an estimated one million Alus and 200,000 line elements, making it roughly 10% Alu and 5% LINE in terms of total content. We further estimate that there are roughly 80,000 distinct genes in the human genome, and probably 25% of these have two to five copies within the genome. There are also large 43-kbp-long RNA pseudogene arrays that occur in tandem clusters and that vary by only 2 to 3% between copies. Finally, there have been large 50- to 150-kbp-long genome duplications where a section of one chromosome has been duplicated near the centromere of another. Any attempt to directly shotgun a large portion or the entirety of a genome as a single source thus must carefully contemplate the impact of repeats on its underlying algorithms.

While practitioners have ambitiously increased the size of the source sequences, the technology for obtaining a read has not improved the length of a read \bar{L} at a corresponding rate, leading to greater and greater ratios of $\omega = G/\bar{L}$. Thus, the expected number of gaps grows as $\omega \bar{c} e^{-\bar{c}}$, ignoring the exacerbating effect of clone bias. Fragmentation of the solution into a collection of gap-separated contigs therefore increases at least linearly with source size for a fixed level of sequencing coverage. This, combined with the increasing difficulty of correctly resolving repetitive elements in the source, has led investigators to develop enhancements to the shotgun sequencing protocol.

“Double-barreled” shotgun sequencing

In the predominant variation on shotgun sequencing, inserts are size-selected so that their average length \bar{I} is at least $2\bar{L}$ or longer and both ends of the insert are sequenced.¹¹ This procedure gives rise to a pair of reads, called *mates*, that are in opposite orientations and at a distance from each other approximately equal to the insert length. While these mate pairings could operate in an integral way within the fragment-assembly software, this information typically serves instead to confirm the assembly and most importantly to order contigs with respect to each other. (A *contig* is a maximal overlapping arrangement of frag-

ments covering a contiguous region of the reconstructed fragment.) That is, if a read in one contig has a mate in another contig, we know the orientation of the contigs to each other and have an idea of the distance between them. At 7.5X coverage, for example, contigs tend to be quite large, at an average of 66 kbp, and gaps quite small, at an average of 66 bp. Because there are typically many mated pairs between a pair of adjacent contigs, we can quite reliably order the contigs. Such a maximally linked and ordered set of contigs is called a *scaffold* (see Figure 1). The next step is to sequence the small gaps between adjacent contigs by amplifying a sample of the sequence between the contigs with a process called PCR (for polymerase chain reaction) that only requires knowing 18 to 25 unique bases on either side of the gap to be amplified.

With the one exception of the TIGR (from The Institute of Genetic Research) assembler,¹² investigators have used mate information only for confirmations, primarily because it can be quite unreliable, with on average 10% of reported pairs proving unrelated. There are three sources of such false positives.

- Two small fragments from distant parts of the source might get inserted into the vector. For such a *chimeric clone*, the reads at both ends thus come from uncorrelated parts of the genome. Appropriate care—such as size-selecting clones or using asymmetric linkers in the insertion step—can keep this source of false pairings to as low as 0.01%.
- A sample can simply be mistracked as it flows through the sequencing factory. For example, a technician might place a microtiter plate in the wrong orientation within a stack of plates or transfer materials to an incorrect destination. Simple precautions such as using asymmetric plates and dual-bar scanning any transfer can also keep this source of false positives under 0.1%.
- In slab gel-sequencing machines, the material often does not migrate along a straight line but gently undulates, causing the optical-scanning software to misnumber the 32 to 96 lanes of sequencing reactions that run simultaneously on a given slab. This predominant source accounts for 10% of the false-positive rate.

How then should we choose the average size \bar{I} of the inserts in such a strategy? We can define

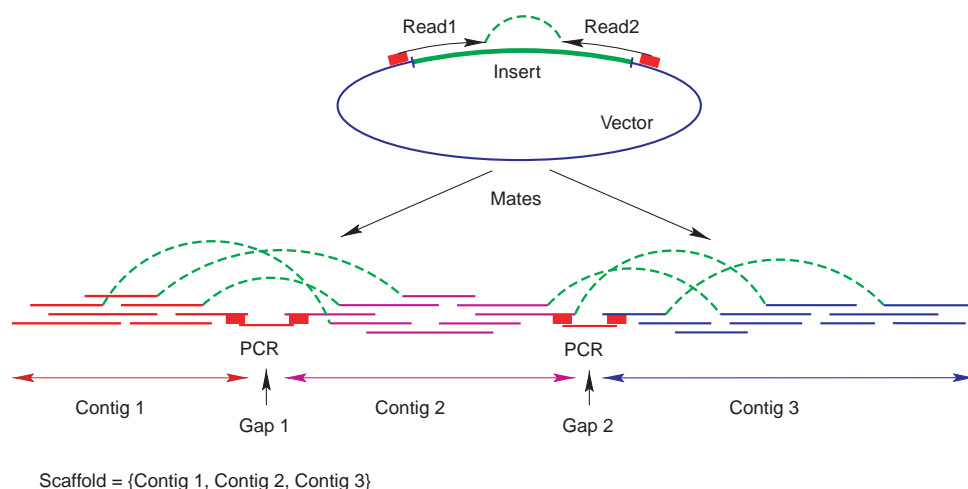


Figure 1. Mates, contigs, gaps, and scaffolds. The top of the figure shows a blue vector with a green insert for which read reactions are primed at both ends. A light green dashed arc depicts the relationship between the reads and is used within the assembly shown below it. The relative order of the differently colored three contigs is fixed by the mate pairings. We then prime PCR reactions across the two gaps (primers in red, polymerase chain reactions sequence in gold). The three contigs in aggregate constitute a single scaffold.

the *map* or *clone coverage* of such a project as $\bar{m} = C\bar{l}/G$, where the number of clones C is $R/2$ in the current context. From the definitions it follows that $\bar{m} = \bar{c}(\bar{l}/2\bar{L})$ is larger than \bar{c} , so there are a factor of $e^{-1/12\bar{L}}$ fewer gaps in the coverage of the source by inserts than there are gaps in the coverage of the source by reads. For example, if inserts are 5 kbp long, there are a factor of e^{-5} or 148 fewer clone gaps than sequence gaps. From another viewpoint, scaffolds are on average 148 times larger than contigs, so that for 7.5X sequencing project of a 200-kbp source, we would expect all the contigs to be ordered by the mate information.

Recent simulation studies have indicated that from a purely informatic perspective, there is an advantage in using long inserts and no advantage in having some percentage of the reads be unpaired.¹³ However, this finding must be tempered against the experimental fact that because of the different cloning vehicles required to serve as the vector as the insert becomes larger, it is more difficult to sequence the ends of long inserts, and greater care must be taken to avoid chimeric clones. Counterbalancing economic pressure thus encourages the use of single reads and shorter inserts. Fortunately, we lose little of the benefits of having long end-sequenced inserts in hybrid schemas where a sizable fraction of a project is single reads and where the paired reads are from inserts over a distribution of insert lengths skewed to the shorter lengths.

Sequencing the human and other whole genomes

After the idea that the human genome could be sequenced began to be discussed in the early to

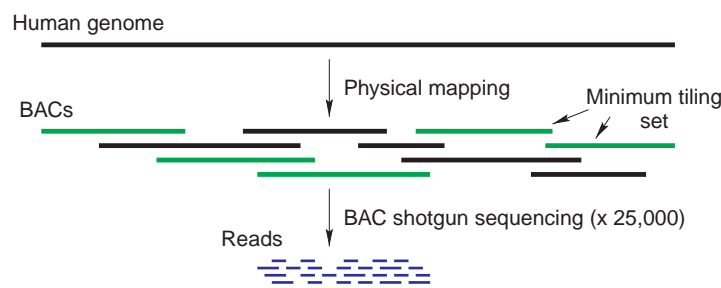
mid 1980s, the US National Institutes of Health and Department of Energy announced the start of the Human Genome Program (HGP) in 1990, with an objective to do so by 2005 in concert with the UK's Sanger Centre and other laboratories in Europe and Japan.¹⁴ A single approach, described next, was adopted and continues to be followed. In the last few years, several interesting alternative strategies have emerged, and I describe two of these as well, the last of which Celera Genomics is actually pursuing. This latter plan has a potential to produce the entire sequence in two years time—by 2001—at one tenth the cost of the HGP.

The clone-by-clone approach

The HGP proposal involves a hierarchical two-tiered approach. This approach first randomly fractures the whole human DNA sequence into 50- to 300-kbp pieces and inserts them into BACs, which are a vector mechanism designed to accommodate such large DNA segments. The resulting collection of BAC inserts is maintained in a *library* from which investigators can select a particular BAC insert to amplify for further experimentation. The first step consists of determining an assembly, or *physical map*, of these large inserts that covers the human genome. Given a physical map, the investigators then pick a minimal tiling set of the inserts that covers the genome. At the second level, they shotgun-sequence each of the inserts in the tiling set. This has been coined a *clone-by-clone* approach because once we have the tiling set of BAC clones, we conceptually imagine sequencing each tiling clone in a march across the genome (see Figure 2).

The term physical map stems from the obser-

Figure 2. The Human Genome Project's two-tiered approach. After first fragmenting the genome into large bacterial-artificial-chromosome-sized segments, the investigators build a physical map of them. They then select a minimum tiling set of the BACs in the map (shown in green) and shotgun sequence each of these.



vation that such an assembly gives the physical location of each segment in the genome. Unlike the fragment-assembly problem, where the complete sequence of the inserts is used to determine overlaps between them, overlaps between BAC inserts are determined on the basis of *fingerprint* data about each insert, which is necessarily less informative than knowing the entire sequence of the insert. Here are several types of fingerprints that various research groups have used and the conceptual nature of the information they convey.

- *Restriction length digests*: The approximate lengths of the pieces that result when an insert is split at each occurrence of a particular substrings of length 4, 6, or 8. The agents that perform the cutting are called *restriction enzymes*.¹⁵
- *Restriction maps*: The approximate locations along the insert of a selectable set of substrings of length 4, 6, or 8, cut by restriction enzymes.¹⁶
- *Oligo probe hybridization*: The presence or absence of each of a set of 12- to 24-length substrings.¹⁷
- *STS probes*: The presence or absence of a pair of 18-length substrings between 200 and 1,000 bases apart in the insert.¹⁸ (A region characterized by such a pair is a *sequence tagged site*, or STS.)

The STS probe is currently the most widely used because of the ease, cost, reliability, and automatability of determining the information. Even for these experiments, investigators must deal with fairly high error rates—roughly 2% false positives (a probe is reported for an insert when it does not contain it) and 10 to 20% false negatives (a probe is not reported when it should be). Most false negatives are due to experimental failures, while the false positives are expected to be induced by repetitions in the genome. Such sparse

information of such moderate reliability leaves us with a problem that is computationally very difficult to solve optimally and for which there is considerable ambiguity in the answers delivered.^{19,20}

The HGP approach has the advantage that the outcome is understood and portends to deliver most of the genome. Shotgun sequencing of BACs is now fairly routine. Reliable software is available, and centers capable of rapidly sequencing BACs continue to gear up. Physical maps, while hard to build, have been prepared for a number of chromosomes. While not complete, they do cover a significant percentage of the chromosomes involved. Thus we are certain to see a reasonable return on continued investment in the HGP.

HGP's shortcomings are in terms of cost, efficiency, and, to a lesser extent, the completeness of what will be determined. Sequencing at this scale is basically an issue of designing a medium-sized factory. Issues are simplicity, automatability, and cost of each step, and scalability of the overall process. The HGP design has the drawback of involving two separate processes: sequencing and physical mapping. While sequencing is heavily automatable once an insert library of fragments has been prepared, investigators must prepare a minimum of 30,000 clone libraries of BACs by hand and must continue to laboriously build and try to complete physical maps of each of the chromosomes. Originally, all the physical maps were to be completed, at a modest cost, in the project's first five years. The cost has been much heavier than anticipated, and eight years into the project, maps are available for only a few chromosomes—and most of these maps have on the order of hundreds of gaps, some of considerable size. Also, it is difficult to construct BAC clones that are not chimeric. By some estimates, 1 to 5% of the BAC sequences being sequenced are actually two or more unrelated segments of the human genome that have been inserted together into the BAC.

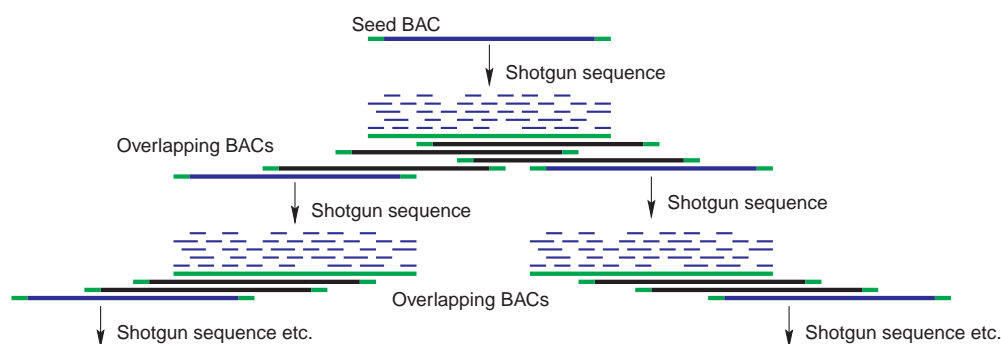


Figure 3. Ordered shotgun sequencing. Starting at the top, we shotgun-sequence a selected seed BAC whose sequenced ends are shown in green. Once the entire sequence of this BAC (shown as a solid green line) is revealed, we observe overlaps with a number of end-sequences of other BACs in the library. We then shotgun-sequence the left- and rightmost of these (shown with a purple interior). The process continues iteratively, giving a BAC-by-BAC walk across the genome.

The sequence-tagged connector approach

An interesting proposal that circumvents the physical-mapping step involves initially sequencing both ends of approximately 600,000 BACs.²¹ BAC clones have an average size of $\bar{L} = 150$ kbp, implying a clone coverage of the human genome of $\bar{m} = 30X$. Sampling theory tells us that there will thus be roughly $600,000 \times e^{-30} \approx 10^{-7}$ gaps in the genome's coverage by BAC clones; that is, with good probability there will be no gaps. Unfortunately, the BAC inserts are produced by partial digestion with restriction enzymes, implying that BAC endpoints are not particularly random. Estimating the effect of this is difficult, but the implication is that there might be a few gaps despite the high clone coverage. On the other hand, without any further effort, few of the BACs can be assembled, because their end sequences constitute a coverage in sequence of only $\bar{c} = 0.1$. On average, there is one BAC-end sequence in every 5 kbp segment of the genome, and few of them overlap.

The next step involves randomly selecting a few of the BAC clones as seeds of an ordered clone-by-clone walk. Each of the selected BAC clones is shotgun-sequenced. Most notably, once a BAC is sequenced, on average 30 end-sequences of other BACs will be discovered to overlap the BAC's interior. Half will extend into the genome in each direction, with one having an overlap, on average, of only 7.5 kbp with the sequenced BAC. The next step is to shotgun-sequence the two minimally overlapping BACs in each direction and then in turn determine minimally overlapping BACs in each direction to sequence next. The investigator is therefore effectively discovering how to continue a seeded set of bidirectional, clone-by-clone walks across the genome as each clone in each walk is sequenced. Figure 3 illustrates the process.

The sequence-tagged connector approach

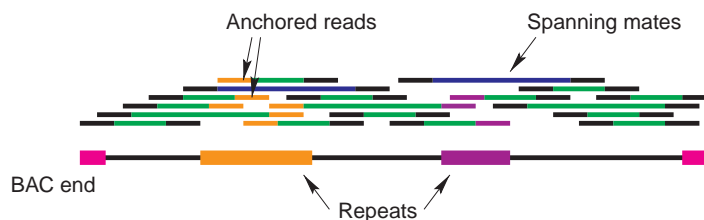
eliminates the physical mapping step. Moreover, the organization of the sequencing factory is simplified because sequencing BAC ends and the smaller shotgun inserts are similar sequencing processes. But, the approach still suffers from needing to make at least 25,000 BAC libraries and because the BAC clones must be maintained for the length of the entire project. Getting a consistently successful sequencing reaction for the end of a BAC is also more difficult, so greater effort and expense must go to end-sequence BACs. The quality of the end reads is poorer as well—on the order of 2 to 5% error.

The whole-genome shotgun approach

The plan we have developed at Celera Genomics involves collecting 60 to 70 million high-quality sequencing reads for a 10X coverage of the genome. We will use only those portions of a read that have an error rate of 1% or less, in contrast to current practice with BAC shotgunning where as much of a read as possible is used to get better coverage with only 6 to 7X and thus reduce cost. For a problem on the scale of the human genome, the ends of such reads, at a 10 to 15% error rate, are too noisy to detect overlaps. We must use only the high-quality parts of a read and collect 10X to compensate for the shorter length. Even so, without any additional information, assembling this large set of reads is effectively impossible given the genome's repetitive nature.

Recall, however, that we can end-sequence inserts to produce mate pairs. Typically, the insert lengths are on average 2 kbp. With care, we can use inserts as long as 10 kbp, although the success rate of reactions on such longer clones is lower, so they are somewhat more expensive to collect. The plan is for 80% of the reads to be in

Figure 4. Whole-genome shotgun assembly. Mated pairs of fragments are black segments with an intervening green segment connecting them. Given two BAC end sequences shown in red, where for the purposes of illustration we assume there is a gold and purple repeat in the BAC, the problem is to determine the set of mated reads that cover the BAC. Mate pairs that span repeats have their connecting line colored blue, and the reads completely interior to a repeat are given the repeat's color. Such reads are often anchored in the sense that their mate is not in a repeat.



2-kbp mate pairs and the remaining 20% to be in 10-kbp mate pairs. I noted earlier that on current slab gel-sequencing machines, false pairings of mates occur at about a 10% rate because of lane-tracking errors. The current plan also uses next-generation capillary gel-sequencing machines in which the material of each sequencing reaction migrates down its own physically separate microcapillary tube. Thus for these machines, the lane-tracking problem disappears and we can now expect mate-pairing errors to be less than 1% and possibly as good as 0.01%. With information of this quality, investigators can now use mate-pairing information as a key component of the assembly algorithm.

Intuitively, we understand that mate pairs can resolve any repeat whose length is shorter than the distance between mates as follows. Imagine building an assembly by progressively adding fragments at a given end (see Figure 4). As long as you are in a unique stretch of sequence, the placement of the next maximally overlapping fragment is obvious and correct. However, when you enter a repeat of sufficiently high fidelity with other copies, you begin to place fragments from many of the copies together. Notice, however, that while fragments are being incorrectly incorporated, you are still effectively putting together a facsimile of the repeat's sequence. The real problem develops when you exit the repeat at the other end: if there are 100 copies of the repeat that are intertwined at this point, there are 100 unique flanks into which you could walk and you don't know which to take. However, there is very likely a mate pair that spans the repeat in that it has a read in the unique flanking sequence on each side of the repeat. Such a spanning mate indicates which of the 100 options to take. Moreover, you can resolve the tangle of reads from different copies within the repeat

stretch by observing that most of these reads have a mate that is anchored in the sense of being in a unique part of the genome. From the anchored mates on the flanks of the repeat, you can generally determine enough of the reads actually sampled from that copy of the repeat to infer its exact sequence.

Of course, there will be many repeats in a genome longer than 10 kbp. To resolve these, we need mated pairs of reads at longer lengths. Fortunately, there will be 600,000 BAC end sequences produced in anticipation of the ordered shotgun approach described above. These BAC end pairs essentially serve as very long-range mates, albeit of less reliability. Moreover, in separate radiation-hybrid mapping efforts,²² STS marker maps that place and linearly order read-sized sequences roughly every 200 kbp along the genome have already been constructed. While these maps are not very accurate, they do give additional long-range mate pairings of up to any length required to resolve a repeat. Originally, we conceived of solving the computational problem by solving a series of *intermarker assembly problems* that require assembling the sequence between a pair of STS markers or BAC-end sequences given the 60 to 70 million reads in the whole-genome shotgun data set. Simulation work has shown that with 99.8% probability, we can unambiguously assemble 99.7% of the sequence between the markers.

I've spent a fair bit of time discussing how to assemble a whole-genome data set because this is the component of the proposal that most critics think is impossible. In terms of a sequencing factory, this approach provides the greatest simplicity because we only have to set up a sequencing pipeline. Moreover, we have to build only two sequencing libraries from whole human DNA. We can therefore expend great effort to insure that these libraries do not contain undesirable artifacts and can completely automate all the remaining steps, thus making the manpower required to run the factory very small. Finally, there is no need to store BAC or other clones for

any length of time, because once the BACs have been end-sequenced they are no longer needed except as PCR templates for gap filling. Coupled with the new-generation capillary gel-sequencing machines that give us greater speed and capacity, the plan will be very efficient in terms of both time and cost.

A final issue centers on understanding diversity in the human genome. Each human cell, with the exception of sperm and egg cells, has two copies of each chromosome, one version from each parent. The complement of DNA inherited from one parent is called a *haplotype*. Each human haplotype varies from another by 0.1%, and the total number of sites of variation over the human population takes an estimated 0.3% of the genome.²³ In the clone-by-clone approach, each assembly of a BAC clone is of a given haplotype, so the HGP effort will produce a series of overlapping clones, each representing some haplotype. This is to be contrasted to the whole-shotgun approach, where fragments from different haplotypes come together to give the overall assembly. In this case, we can detect many of the sites of genetic variation between haplotypes. Even if we sequence only one pair of haplotypes, we will detect an estimated three million sites of single nucleotide variation or polymorphism during the course of the project. ■

References

1. R.D. Fleischmann et al., "Whole-Genome Random Sequencing and Assembly of *H. Influenzae*," *Science*, Vol. 269, No. 5,223, 1995, pp. 496-512.
2. J. Weber and W. Myers, "Human Whole Genome Shotgun Sequencing," *Genome Research*, Vol. 7, No. 5, 1997, pp. 401-409.
3. P. Green, "Against a Whole-Genome Shotgun," *Genome Research*, Vol. 7, No. 5, 1997, pp. 410-417.
4. J.C. Venter et al., "Shotgun Sequencing of the Human Genome," *Science*, Vol. 280, No. 5,369, 1998, pp. 1540-1542.
5. F. Sanger, S. Nicklen, and A.R. Coulson, "DNA Sequencing with Chain-Terminating Inhibitors," *Proc. Nat'l Academy of Science*, Vol. 74, No. 12, 1977, pp. 5463-5467.
6. A.M. Maxam and W. Gilbert, "A New Method for Sequencing DNA," *Proc. Nat'l Academy of Science*, No. 74, No. 2, 1977, pp. 560-564.
7. F. Sanger et al., "Nucleotide Sequence of Bacteriophage λ DNA," *J. Molecular Biology*, Vol. 162, No. 4, 1982, pp. 729-773.
8. L. Rowen, B.F. Koop, and L. Hood, "The Complete 685-Kilobase DNA Sequence of the Human Beta T cell Receptor Locus," *Science*, Vol. 272, No. 5,269, 1996, pp. 1755-1762.
9. G.I. Bell, "Roles of Repetitive Sequences," *Computers Chemistry*, Vol. 16, 1992, pp. 135-143.
10. F.J.M. Iris, "Optimized Methods for Large-Scale Sequencing in Alu-Rich Genomic Regions," *Automated DNA Sequencing and Analysis*, M.D. Adams, C. Fields, and J.C. Venter, eds., Academic Press, London, 1994, pp. 199-210.
11. A. Edwards and C.T. Caskey, "Closure Strategies for Random DNA Sequencing," *Methods: A Companion to Methods Enzymology* 3, Academic Press, New York, 1991, pp. 41-47.
12. G.G. Sutton et al., "TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects," *Genome Science & Technology*, Vol. 1, No. 1, 1995, pp. 9-19.
13. J.C. Roach et al., "Pairwise End Sequencing: A Unified Approach to Genomic Mapping and Sequencing," *Genomics*, Vol. 26, No. 26, 1995, p. 345.
14. F. Collins and D. Galas, "A New Five-Year Plan for the U.S. Human Genome Project," *Science*, Vol. 262, No. 5,130, 1993, pp. 43-46.
15. M.R. Olson et al., "Random Clone Strategy for Genomic Restriction Mapping in Yeast," *Proc. Nat'l Academy of Science*, No. 83, No. 20, 1986, pp. 7826-7830.
16. Y. Kohara, A. Akiyama, and K. Isono, "The Physical Map of the *E. Coli* Chromosome: Application of a New Strategy for Rapid Analysis and Sorting of a Large Genomic Library," *Cell*, Vol. 50, No. 3, 1987, pp. 495-508.
17. A. Coulson et al., "Toward a Physical Map of the Genome of the Nematode, *C. Elegans*," *Proc. Nat'l Academy of Science*, Vol. 83, 1986, pp. 7821-7825.
18. M.R. Olson et al., "A Common Language for Physical Mapping of the Human Genome," *Science*, Vol. 245, No. 4,925, 1989, pp. 1434-1435.
19. F. Alizadeh et al., "Physical Mapping of Chromosomes: A Combinatorial Problem in Molecular Biology," *Algorithmica*, Vol. 13, Nos. 1 and 2, 1995, pp. 52-76.
20. M. Jain and E. Myers, "Algorithms for Computing and Integrating Physical Maps Using Unique Probes," *J. Computational Biology*, Vol. 4, No. 4, 1997, pp. 449-466.
21. T.J. Hudson et al., "An STS-Based Map of the Human Genome," *Science*, Vol. 270, No. 5,244, 1995, pp. 1945-1954.
22. J.C. Venter, H.O. Smith, and L. Hood, "A New Strategy for Genome Sequencing," *Nature*, Vol. 381, No. 6,581, 1996, pp. 364-366.
23. A.G. Clark et al., "Haplotype Structure and Population Genetic Inferences from Nucleotide Sequence Variation in Human Lipoprotein Lipase," *American J. Human Genetics*, Vol. 63, No. 2, 1998, pp. 595-612.

Gene Myers is the director of Informatics Research at Celera Genomics and a professor currently on leave from the Department of Computer Science at the University of Arizona. His research interests include algorithm design, pattern matching, computer graphics, and computational molecular biology. He received his PhD in computer science from the University of Colorado. He is an associate editor of the *Journal of Computational Biology*. Contact him at Celera Genomics, 45 W. Gude Dr., Rockville, MD 20850; myersgw@celera.com.