

Gaps in the Human Genome Project

If the multinational Human Genome Project is to continue its successful start, sequencing strategies must be changed.

Jared C. Roach, Andrew F. Siegel, Ger van den Eng, Barbara Trask and Leroy Hood

The adage 'Buy now, pay later' reflects the current operational strategy of the Human Genome Project (HGP). Thirteen per cent of the sequence is now available, and the project seems to be ahead of schedule and under budget¹. But the looming costs associated with gap closure in the final phases of the HGP have yet to be fully addressed. Here we evaluate and quantify those deferred costs.

Currently, the HGP emphasizes highly parallel production of first-draft sequence. 'Highly parallel' refers to the simultaneous initiation or extension of known sequence from many different seed locations. This emphasis represents an acceleration of a subset of the original goals of the project's US arm, stated in 1990, which focused on contiguous high-quality sequence. The current shift in strategy aims for a first draft of 90% of the human genome by the end of next year². Contiguous high-quality sequence has been deferred until 2003. The motivation for this shift initially stemmed from the announcement of a private initiative from Celera Genomics Corporation to compete with the public HGP³. Increased recognition of the public-health importance of raw genomic data, which can facilitate molecular diagnosis and drug discovery, has further stimulated this policy.

The acceleration in sequencing was made possible by improvements in the throughput of automated sequencing machines, both in the number of machines dedicated to the HGP worldwide and in the capacity of each machine. Public capacity is concentrated in five major centres: the UK Sanger Centre in Cambridge, the US Department of Energy's Joint Genome Institute in California and, with funding from the National Institutes of Health, the triumvirate of Washington University in St Louis, the Whitehead Institute at the Massachusetts Institute of Technology and Baylor College of Medicine in Texas. Worldwide capacity, as measured by the amount of high-quality sequence deposited in GenBank, is now 23 megabases per month, and is expected to increase.

The chromosomes of the human genome are 50–250 megabases long, too large to be sequenced directly. They must first be fragmented into intermediate clones about 150 kilobases (kb) long, such as bacterial artificial chromosomes (BACs; see Glossary for a

definition). The resulting collection of fragments, or BAC 'library', compiled from many genomes, is highly redundant. The exact location of each BAC on the target genome is not initially known, but can be found with additional effort and expense. Algorithms for choosing BACs are called BAC selection strategies (Fig. 1).

Improving BAC selection

We show here that the current HGP strategy for BAC selection is inefficient, and we suggest how it can be improved. We evaluate four BAC selection strategies: mapping, random BAC, limited seeding and parking. Following selection, regardless of strategy, each BAC is sequenced by 'shotgun' strategies initially to first-draft quality, and then eventually to high quality, at a total cost of

US\$30,000–40,000 per BAC, although some optimistic estimates of this cost are as low as \$4,000. All these strategies can be implemented if the shotgun sequencing of the BACs is only to first-draft accuracy, but we assume here for simplicity that they are completely sequenced to the high-quality level. Celera's whole-genome pairwise shotgun approach, which does not use a BAC selection strategy, has been discussed elsewhere^{4,5}.

The choice of BAC selection strategy directly affects the overall cost of the HGP. Redundant sequencing of the genome is the main source of inefficiency that can vary with the choice of strategy. Thus, the amount of overlap among selected BACs is a key parameter for evaluating any strategy (Fig. 2). However, the advantages of choosing BACs with minimal overlap must be balanced

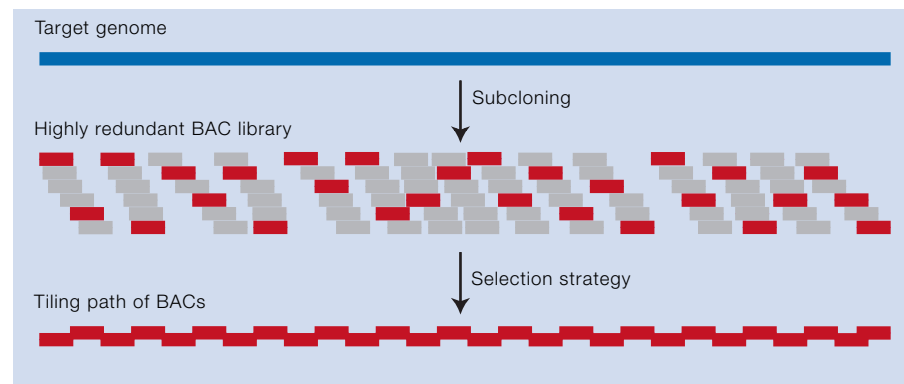


Figure 1 Overview of the strategy for the Human Genome Project. Many replicates of the human genome are fractionated into a highly redundant library of intermediate-size clones (usually BACs). One of several strategies is then used to select a subset of this library for sequencing. In practice, a composite library is used, representative of the genetic diversity of all humans. The actual redundancy of the library is around 30-fold, about ten times that depicted here. (Figure not to scale.)

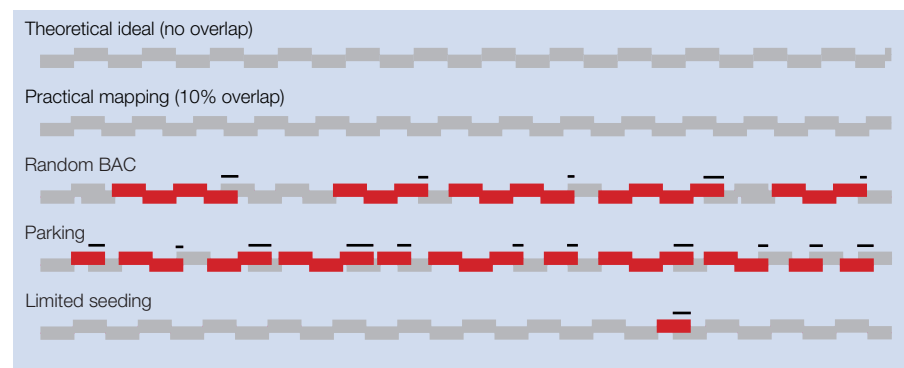


Figure 2 Gap closing for basic genome-sequencing strategies. Grey, BACs sequenced before the initiation of gap closing; red, BACs sequenced for closing the resulting gaps, together with the last BAC filling a gap as part of a limited seeding strategy. Black bars indicate inefficient overlaps caused by mismatches between clone lengths and gap lengths. Implied gap sizes and distributions are roughly to scale.

against the costs of selecting them from the BAC library. Further, the effect of strategy choice on the expense of gap closing must also be considered.

A theoretically ideal strategy would be to completely sequence a minimal tiling path of BACs placed end-to-end across the genome; however, this would be prohibitively expensive (more than \$400 million). Rather, the original strategy for the HGP was to identify a tiling path of BACs with mutual overlaps of about 10%. This up-front mapping approach was more costly and time-consuming than expected and has been abandoned, although mapping data remain valuable for other purposes. The cost of mapping the entire genome with 10% overlapping BACs would be approximately \$240 million⁶.

Without mapping data, BACs must initially be sequenced at random. If no map is then generated from sequence data as they are produced, the strategy is very inefficient. Such a 'random BAC' strategy is not contemplated for the human genome, but we include it here for reference⁷.

A more efficient strategy is to choose a subset of seed BACs, shotgun sequence them and use these sequenced seed BACs to initiate an iterative walking strategy to produce a map. The most likely walking strategy is the sequence-tagged-connector (STC) strategy, which involves selecting BACs that will extend contigs from among a large set of BACs whose ends have been previously sequenced (an STC library)^{8,9}. In practice, the BAC seeds need not be random, as there are enough mapping data to allow about 1,000 evenly spaced seeds to be chosen along the human chromosomes. This combination of limited seed placement with subsequent walking is dubbed a 'limited seeding' strategy.

A variant seeding strategy is to place a very large number of seeds before beginning walking. One such approach screens each seed by limited sequencing (for example, to 0.5-fold redundancy) and rejects the seed if it significantly overlaps any previous seed. This strategy is called a 'parking' strategy, by analogy to cars parked serially along a curb¹⁰.

Gap sizes

Each of these BAC selection strategies produces a different number of expected gaps and a different distribution of gap sizes (Table 1). The parking strategy is the current *de facto* strategy for the HGP, although most centres supplement it with limited regional walking based on previous mapping data. As a result, the HGP now has more than 1,000 seed contigs, with a distribution of gaps between them consistent with either a random or a parking strategy, biased slightly because of the presence of a few deliberately selected seeds. Many HGP statistics are archived at the National Center for Biotechnology Information (<http://ray.nlm.nih.gov/genome/seq>)².

Table 1 Characteristics of basic genome-sequencing strategies

Strategy	Up-front cost (\$ million)*	BACs sequenced to reach 50% coverage	Screening cost (\$ million)	Gaps at 50% coverage	BACs required for gap closing	Overall cost estimate (\$ million)†
Theoretical ideal	>400	10,000	0	NA	10,000	>1,100
Practical mapping	240	11,100	0	NA	11,100	1,017
Random BAC	19	13,900	0	6,900	14,400	1,044
Parking	19	10,000	33	10,000	16,200	1,019
Limited seeding	19	10,500	0	1,000	10,500	759

Numbers are expectations derived from mathematical models and confirmed by simulation^{6,7,10,11}. NA, not applicable. * See text. † Estimates for the cost of sequencing a BAC vary from \$4,000 to \$40,000. We use \$35,000 here, plus \$5,000 per gap to account for loss of production-line efficiencies. Library management costs, most significant for the parking strategy, are not included.

The main advantage of the parking strategy is that minimal coordination is required within and between sequencing centres. The only necessary coordination is a central database that can be checked for duplicate effort. Additionally, if the sequencing capacity devoted to the HGP ever exceeded the number of available sites for walking, the additional capacity could be devoted to new seeds rather than sitting idle. One early advantage of the parking strategy was its independence

from any STC library, which meant that BAC selection could be started before the highly redundant STC library became available in mid-1999. A disadvantage of the parking strategy is that few contigs longer than 150 kb are initially produced, which delays analyses requiring longer contigs.

Neither the random BAC nor the parking strategy can be used indefinitely. Once the HGP is half completed, the average random BAC will overlap previous BACs to a

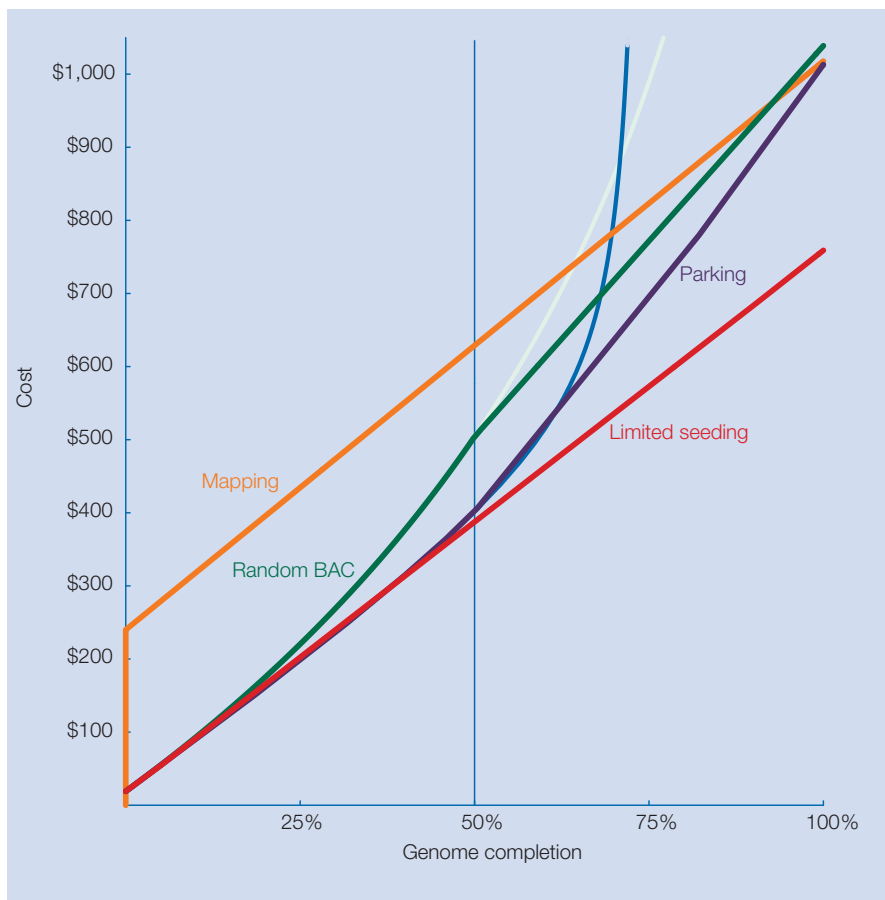


Figure 3 Efficiencies of the basic strategies for sequencing the human genome. Yellow, practical mapping strategy; green, random BAC strategy; red, limited-seeding strategy; blue, parking strategy. Faded curves for the random BAC and parking strategies indicate the consequences of not switching to a gap-closing strategy; solid curves represent a switch at 50% completion. The screening cost for parking is chosen to be the cost of 0.5-fold BAC sequencing. This choice has little effect on our conclusions, as the slope of the parking curve is largely insensitive to the screening cost. The faded curves for random BAC and parking strategies are asymptotic to 100% and Renyi's number (~74.8%), respectively⁹.

prohibitive extent (Fig. 3). For parking, finding non-overlapping seeds becomes increasingly difficult, and eventually impossible at a 'jamming limit' near 75% completion¹⁰. Therefore, after approximately half of these strategies, there must be a switch to a walking strategy. The STC library is therefore needed for all strategies, except mapping, requiring an investment of about \$19 million¹¹.

The cost of closing gaps varies between strategies because there are no perfect mechanisms for choosing clones that exactly match the lengths of gaps. Extremely short gaps can be closed by PCR (polymerase chain reaction), but these are rare. For example, only 5.4% of the gaps are shorter than 10 kb for the parking strategy described here. When gaps exist that are not integer multiples of the effective length of a BAC, a considerable portion of the last (or often the only) BAC used for closing the gap may overlap already known sequence. This source of inefficiency occurs once per gap and thus becomes significant in strategies that result in many gaps. The accumulated inefficiency is the main drawback of the parking strategy, which splinters the unsequenced portion of the genome and maximizes the number of gaps. The amount of inefficiency varies depending on the size of each gap, but averages at about half the length of a BAC per gap.

Costs for each strategy

The mapping and random BAC strategies are more expensive than limited seeding at all stages (Fig. 3). At less than 32% completion, the parking strategy is marginally less expensive than limited seeding, as the screening costs to avoid overlap in parking are initially less than the cost of sequencing overlaps for the seeding strategy. However, even if the parking strategy were to switch to a walking strategy before 32% completion, the final cost of the completed genome would still be greater, because of the larger number of gaps that must be closed. Above 32% completion, the limited seeding strategy is the least expensive strategy at all phases. We estimate the overall cost difference between the parking strategy and the limited seeding strategy to be \$260 million.

Several modifications to the parking strategy can reduce costs. Any reduction in the total number of seeds will reduce the final number of gaps to be closed, improving the parking strategy by making it a form of the limited seeding strategy. Hence, walking based on previous map data or with BACs from an STC library should be implemented for each seed as soon as the seed is produced. Additional seeds should be sequenced only when sequencing capacity exceeds the number of BACs available for walking. Sequencing capacity is currently about 150 BACs per month, and it is unlikely that capacity will exceed 2,000 simultaneous walking steps

Glossary

Bacterial artificial

chromosome (BAC): A device for propagating a small segment of the human genome within an *Escherichia coli* cell, facilitating further analysis. The acronym BAC is used interchangeably to refer to the device itself or to the segment of cloned DNA that it propagates.

Contig: a contiguous tract of known sequence.

Gap: a tract of unknown sequence between adjacent contigs.

First-draft sequence: nearly contiguous sequence with an error rate of about 1:1,000.

High-quality sequence: contiguous sequence with an error rate of less than 1:10,000.

Shotgun: an algorithm that blindly selects clones from a random library for further analysis, such as sequencing. 'Shotgunning' describes the initial creation of a random library.

Mapping: identifying the location of every BAC in a library before sequencing, using methods such as restriction digests or PCR assays.

Random BAC: a tiered shotgun strategy. First, BACs are blindly selected from a genomic library; they are then shotgunned into

smaller plasmid libraries, from which plasmids are blindly selected for sequencing.

Seeding: selection of a set of BACs that will serve as the initiation sites for a walking algorithm.

Parking: a version of the seeding algorithm in which seeds are chosen such that no seed overlaps any previous seed.

Walking: an iterative algorithm for extending a contig. The algorithm 'walks' along the target by using known sequence from the contig as physical or virtual bait to fish for overlapping BACs, then characterizing these BACs so that they may in turn be used as bait.

Sequence-tagged connector (STC): the sequenced end of a BAC (usually produced in pairs). The 'STC strategy' is a walking strategy in which the entire sequence of a contig is used to identify the STCs of overlapping BACs. The minimally overlapping BAC from each end of the contig is then shotgun sequenced, extending the contig.

Minimal tiling path: given a particular strategy, the set of clones with the shortest overall length that spans the genome (see Fig. 1).

Finishing: the final stages of sequencing a BAC. Typically, BACs are shotgunned into plasmid libraries, from which a three- to sevenfold excess of raw sequence data is produced, resulting in first-draft sequence. Finishing eliminates errors and closes gaps in first-draft sequence.

Gap closing: an algorithm for closing large gaps between BACs. By contrast, finishing includes the process of filling in small gaps within the sequence of a particular BAC.

Whole-genome pairwise shotgun: an algorithm for genome sequencing that bypasses a BAC selection strategy. Paired ends of clones of varying sizes are sequenced, producing almost all the target sequence, but fragmented into a very large number of ordered contigs. Finishing would require closing all of the short gaps between these contigs.

Completion: the goal of the Human Genome Project, which includes the complete high-quality sequence of all of the unique DNA in the genome, as well as representative sequences from repetitive DNA. The non-redundant length of the genome is approximately three gigabases.

before the HGP is completed. Innovative methods for tailoring the length of sequence necessary for closing a gap to the length of that gap could also eliminate much inefficiency. PCR is one such method for very short gaps. Constructing an additional STC resource consisting of small insert BACs (for example, 30–100 kb) would reduce certain overlaps, allowing a modest gain in efficiency. Another strategy modification allows small overlaps during the parking screening process. Some modifications are already being implemented. Nevertheless, even with further modifications, the cost differential between limited seeding and parking will remain very high.

The scale of the HGP, as well as its multinational and highly collaborative nature, dictate that a number of political, economic and technical factors influence strategy choice. Our data indicate that some choices may result in false economies. To reduce redundancies, we recommend greater coordination at all levels and that centres agree

on a minimal set of evenly spaced seed BACs. In particular, the current default parking strategy should be abandoned, and the HGP should be directed towards a strategy that reduces the ultimate number of gaps in the genome. ■

The authors are in the Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195, USA. A.F.S. is also in the Departments of Management Science, Finance and Statistics.

Correspondence should be addressed to L.H. e-mail: leehood@u.washington.edu

- Collins, F. S. *et al.* *Science* **282**, 682–689 (1998).
- Jang, W. *et al.* *Trends Genet.* **15**, 284–286 (1999).
- Venter, J. C. *Science* **280**, 1540–1542 (1998).
- Roach, J. C. *et al.* *Genomics* **26**, 345–353 (1995).
- Weber, J. L. & Myers, E. W. *Genome Res.* **7**, 401–409 (1997).
- Siegel, A. F., Roach, J. C., Magness, C., Thayer, E. & van den Eng, G. *J. Comput. Biol.* **5**, 113–126 (1998).
- Roach, J. C. *Genome Res.* **5**, 464–473 (1995).
- Venter, J. C., Smith, H. O. & Hood, L. *Nature* **381**, 364–366 (1996).
- Mahairas, G. G. *et al.* *Proc. Natl Acad. Sci. USA* **96**, 9739–9744 (1999).
- Krapivsky, P. L. *J. Statist. Phys.* **69**, 135–150 (1992).
- Siegel, A. F. *et al.* *Genome Res.* **9**, 297–307 (1999).