

# Sequencing a Genome by Walking With Clone-end Sequences: A Mathematical Analysis

Serafim Batzoglou, Bonnie Berger, Jill Mesirov and Eric S. Lander

## ABSTRACT

One important approach to sequencing a large genome is (i) to sequence a collection of non-overlapping 'seeds' chosen from a genomic library of large-insert clones (such as bacterial artificial chromosomes (BACs)) and then (ii) to take successive 'walking' steps by selecting and sequencing minimally overlapping clones, using information such as clone-end sequences to identify the overlaps. We analyze the strategic issues involved in using this approach. We derive formulas

showing how two key factors, the initial density of seed clones and the depth of the genomic library used for walking, affect the cost and time of a sequencing project—that is, the amount of redundant sequencing and the number of steps to cover the vast majority of the genome. We also discuss a variant strategy in which a second genomic library with clones having a somewhat smaller insert size is used to close gaps. This approach can dramatically decrease the amount of redundant sequencing, without affecting the rate at which the genome is covered.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.  
RECOMB 2000 Tokyo Japan USA  
Copyright ACM 2000 1-58113-186-0/00/04 \$5 00

1 Laboratory for Computer Science, MIT, Cambridge, MA 02139.

2 Mathematics Department, MIT, Cambridge, MA 02139.

3 Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142.

4 Department of Biology, Massachusetts Institute of Technology, Cambridge MA 02139.

\* To whom correspondence should be addressed.