Genomics **68,** 237–246 (2000)
doi:10.1006/geno.2000.6303, available online at http://www.idealibrary.com on IDEAL®

Modeling the Feasibility of Whole Genome Shotgun Sequencing Using a Pairwise End Strategy

Andrew F. Siegel,*,†,‡,§,¶,1 Ger van den Engh,* Leroy Hood,*,¶
Barbara Trask,* and Jared C. Roach¶

*Department of Molecular Biotechnology, †Department of Management Science, ‡Department of Finance, and §Department of Statistics, University of Washington, Seattle, Washington 98195; and ¶The Institute for Systems Biology, Seattle, Washington 98105

Received October 5, 1999; accepted June 27, 2000

In pairwise end sequencing, sequences are determined from both ends of random subclones derived from a DNA target. Sufficiently similar overlapping end sequences are identified and grouped into contigs. When a clone's paired end sequences fall in different contigs, the contigs are connected together to form scaffolds. Increasingly, the goals of pairwise strategies are large and highly repetitive genomic targets. Here, we consider large-scale pairwise strategies that employ mixtures of subclone sizes. We explore the properties of scaffold formation within a hybrid theory/simulation mathematical model of a genomic target that contains many repeat families. Using this model, we evaluate problems that may arise, such as falsely linked end sequences (due either to random matches or to homologous repeats) and scaffolds that terminate without extending the full length of the target. We illustrate our model with an exploration of a strategy for sequencing the human genome. Our results show that, for a strategy that generates 10-fold sequence coverage derived from the ends of clones ranging in length from 2 to 150 kb, using an appropriate rule for detecting overlaps, we expect few false links while obtaining a single scaffold extending the length of each chromosome. © 2000 Academic Press

INTRODUCTION

The defining feature of a pairwise end strategy is that both ends of a large number of randomly selected fragments (subclones) of the target genome are sequenced. Each sequence read from a subclone end, on the order of 550 bp, is termed an "end sequence." Once determined, overlapping end sequences are identified and assembled into contigs. The two reads in each pair are in opposite orientation, and the distance between them is known approximately. This information is used to determine connections between contigs, resulting in "scaffolds," which are maximally connected sets of contigs. Scaffolds can be oriented and assigned to chromosome locations by combining pairwise strategies with other mapping techniques, such as aligning contigs along radiation hybrid, genetic, or cytogenetic maps.

The pairwise end-sequencing strategy was described by Edwards and Caskey (1991), and variants of the strategy have been developed by several groups (Chen *et al.*, 1993; Richards *et al.*, 1994; Smith *et al.*, 1994; Venter *et al.*, 1996, 1998). Simulations and analysis by Roach *et al.* (1995) illustrated the feasibility of pairwise strategies at low redundancies and demonstrated the utility of combinations of subclone sizes for scaffold building. Weber and Myers (1997) proposed to sequence the human genome with a "map-based" double-barrel shotgun approach, similar to the strategy of Roach *et al.* (1995), but on a genomic scale and using mapped sequence-tagged sites (STSs) as an aid to assembly. The initiative by Venter *et al.* (1998) implements a variation of this proposal. Roughly 10-fold total sequence coverage will be obtained by sequencing the ends of three sets of subclones, whose inserts average 2, 10, and 150 kb. The sequence redundancy (the average number of sequence reads covering a random base in the genome) and mapping redundancy (the average number of clone inserts that encompass a given base in the genome) provided by these three sets of subclones are indicated in Table 1.

The pairwise end strategy differs from map-based strategies (e.g., Blattner *et al.*, 1997; Goffeau *et al.*, 1996) for producing contiguous genomic sequence. In map-based strategies, a minimally overlapping "tiling" path of clones is first identified by any of a variety of mapping techniques, and each clone is sequenced separately. The pairwise end strategy eliminates the preliminary mapping phase, which can be advantageous. Pairwise end sequencing has been used successfully to sequence small genomic targets, such as microbial genomes and large-insert subclones of large genomes (e.g., Edwards *et al.*, 1990; Fleischmann *et al.*, 1995; Fraser *et al.*, 1995). The 125-Mb *Drosophila melanogaster* euchromatic genome has also been successfully

[1] To whom correspondence should be addressed at Box 353200, University of Washington, Seattle, WA 98195-3200. Telephone: (206) 543-4476. Fax: (206) 685-9392. E-mail: asiegel@u.washington.edu.

0888-7543/00 $35.00

TABLE 1

Subclone Library Characteristics

Vector type	Insert size (kb)	Number of clones	Number of sequences	Coverage	
				Sequences	Clones
High-copy plasmid	2	30,000,000	60,000,000	8.6	17
Low-copy plasmid	10	5,000,000	10,000,000	1.4	14
BAC	150	300,000	600,000	0.1	13
Total		35,300,000	70,600,000	10	44

scaffolded (Myers *et al.*, 2000). Despite these successes, the feasibility of the strategy for large genomes has been challenged (Green, 1997; Eichler, 1998). The main criticisms and drawbacks of pairwise end sequencing stem from a lack of compartmentalization, which would be provided by a clone-by-clone approach (Green, 1997).

Pairwise end sequencing is likely to be a major strategy for most large-scale sequencing projects in the foreseeable future. Target organisms include the human, microbes, model vertebrates, and plants. Some of these organisms have genomes that are larger than that of the human and that contain complex repeat families. Methods to predict the success of pairwise strategies on target genomes are therefore needed. Success can be measured in terms of (i) the (small) incidence of falsely declared end-sequence overlaps, (ii) the (small) number and size of gaps in sequence coverage, and (iii) the (large) length of the longest scaffold (i.e., does it extend nearly the length of the target).

Myers and Weber (1997) have simulated aspects of whole genome pairwise sequencing. Their simulations assume that the only genome-wide repeats are SINES and LINES and that these are all indistinguishable, so they focus only on assembly of unique sequence. An STS map of 100-kb resolution is also assumed, but no data from paired BAC end sequences are utilized. They conclude that, under these assumptions, a scaffold will span adjacent STS markers over 99% of the time and that, therefore, whole genome pairwise sequencing is feasible.

Anson and Myers (1999) simulate algorithms for "intermarker assembly." These simulations focus on assembly of paired reads from short- and medium-length inserts over 100-kb regions. Their simulations indicate that piecewise assembly of the genome is possible without testing all $n(n-1)/2$ comparisons of raw sequence reads from the entire genome. These simulations also indicate that assembly of unique sequence is possible even under the constraint that all SINE and LINE repeats are treated as identical. Their simulations are limited, however, by allowing for the presence of only a moderate number of longer low-copy-number repeats. These proposed algorithms do not utilize any information from paired BAC end sequences, which leaves open the strong possibility that such information can be leveraged to provide algorithms for complete genomic assembly, including repeated regions.

Recently, a data set generator has been constructed that permits nearly exact simulation of all parameters of shotgun strategies (Myers, 1999). Computing power exists to simulate entire strategies on these data sets completely. Such simulations will complement theory, and vice versa.

In this paper, we construct a tractable mathematical model to capture the key properties of the pairwise end-sequencing strategy, including the presence of long uncharacterized low-copy-number repeats in the target genome. We assume that information from all paired sequences, including BAC ends, will be used for scaffold assembly. We apply our model to predict the outcome of the strategy when applied to the human genome. Our mathematical model permits optimization of certain parameters that are not easily optimized with simulations.

We model the target, such as the human genome, as a string of bases independently chosen from a given nucleotide distribution. A variety of repeat families is superimposed on these independent bases. Subclones containing fragments of different size classes are assumed to be located independently and uniformly at random over the target. Sequences at each subclone end are sequenced with a specified error rate. We develop a decision rule for declaring an overlap between two end sequences, based on the number of matching and nonmatching bases when testing a potentially overlapping alignment and using information about identified repeat segments. False overlaps are counted using a theoretical model that considers all possible alignments of all pairs of end sequences, while the size of scaffolds is obtained by using only the true declared overlaps.

We use the theoretical model to derive the probability, as a function of the decision rule, of detecting a true overlap as well as the expected number of falsely declared end-sequence overlaps. False overlaps can be due to random similarity or can be a result of both sequences overlapping members of a repeat family. The simulation model then places repeat segments and subclones randomly over the target and declares overlap of adjacent end-sequence pairs according to the probabilities derived by the theoretical model. Contigs of declared end sequences are linked by subclones to form scaffolds, whose properties are evaluated.

In the following, we present our notation and assumptions in detail and then outline calculations for the theoretical and the simulation models. Results for sequencing on the scale of the human genome then follow.

METHODS

Notation, Assumptions, and Theoretical Results for True and False Overlap Declaration

We now outline model specifications for the target size, subclone library characteristics, end-sequence length, the rule for declaring end-sequence overlap, the probability distribution for randomness of nucleotides, sequencing-error rates, and repeat families. We also derive theoretical formulae for true and false overlaps, specifically: (i) the probability that two overlapping end sequences will be (correctly) declared to overlap, (ii) the expected number of falsely declared overlaps that would be found by random coincidence, and (iii) the expected number of false overlaps due to repeat-family homology.

- Target and library specifications are denoted as follows:
  - T is the length of the genomic target in basepairs. For the human genome, T is 3.5×10^9.
  - θ denotes the number of subclone groups. Within a subclone group, all subclones are modeled as having a fixed length. For the strategy modeled here, $\theta = 3$.
  - C_i is the clone insert length in basepairs for the ith group of subclones, with $i = 1, \ldots, \theta$.
  - N_i is the number of subclones from group i to be analyzed to form the library of end sequences. The N_i clones of length C_i are assumed to be randomly and independently located along the target.

- End-sequence number, length, and decision rule are specified as follows:
  - $n = 2 \sum_{i=1}^{\theta} N_i$ is the number of end sequences.
  - m is the number of bases that define an end sequence at each end of each subclone, corresponding to the sequencing read length.
  - We use a compound *decision rule* for declaring overlap of two end sequences, with respect to a proposed alignment, where the decision depends upon whether or not there are known repeat segments. This rule is to be used in the core phase of an assembly when all reads are being compared to all other reads. The rule proceeds as follows:
  - If the overlap region includes at least 50 bp of contiguous known repeat sequence, then:
    - If all such repeat sequence is in alignment and there is also at least 50 bp of aligned contiguous unique sequence, then overlap is declared (because the combination of substantial unique sequence agreement with one or more aligned repeat segments essentially guarantees true overlap).
    - Otherwise overlap is *not* declared (because either the overlap region occurs entirely or almost entirely within a repeat segment or the alignment is insufficient).
  - Otherwise (i.e., in the absence of known repeat homology) we use the function $k(j)$ to specify the rule used to decide overlap, $j = 1, \ldots, m$. If the two end sequences are aligned so that each has j bases in the overlap region, and if *more than* $k(j)$ of these bases are read as identical, they will be declared to overlap. While this model includes only substitution errors in the equations, our high chosen error rate is intended to account for both substitution and indel errors, approximating the indels as substitutions.

- Nucleotides are assumed to be drawn from the following probability distribution:
  - α_A, α_C, α_G, and α_T specify the probabilities of choosing each base at random (so that $\alpha_A + \alpha_C + \alpha_G + \alpha_T = 1$). We assume that the target consists of independently selected random bases from this distribution with interspersed repeats.
  - $\alpha = \alpha_A^2 + \alpha_C^2 + \alpha_G^2 + \alpha_T^2$ is then the probability that two independently selected bases are identical by coincidence.

- Errors in sequencing clone ends are modeled as follows:
  - ϵ denotes the sequencing *problem rate* (which will be used to define the observed error rate). We assume that bases are read independently. A base is read initially correctly with probability $1 - \epsilon$. With probability ϵ, the reading is instead an independently sampled random base (with distribution specified by α_A, α_C, α_G, and α_T) that may, by coincidence, be the correct base.
  - ϵ^* denotes the sequencing *error rate*, i.e., the probability that a given base has been read incorrectly. Note that $\epsilon^* = \epsilon(1 - \alpha)$, so that the observed error rate is less than the problem rate ϵ.

- Probabilities and frequencies of true and false overlaps are as follows:
  - $p_{\text{true}}(j)$ denotes the probability that two end sequences that overlap by j basepairs will be (correctly) declared to overlap. We may write (following Theorem 4 of Siegel *et al.*, 1999)

$$p_{\text{true}}(j) = B^{+}[j,\ k(j),\ \alpha + (1 - \epsilon)^2(1 - \alpha)], \qquad [1]$$

where B^+ denotes the following upper cumulative binomial probability

$$B^{+}(j,\ k,\ p) = P(X_j > k) = \sum_{i=k+1}^{j} \binom{j}{i} p^{i}(1 - p)^{j-i}, \qquad [2]$$

where X_j has a binomial distribution with j trials and probability p of success on each trial.

  - λ_{false} denotes the expected number of false overlaps (not due to repeat homology, which is considered separately) that would be obtained if all false alignments of all pairs of end-sequences were examined and may be computed conservatively as follows. The *random* number of false overlaps is the sum (over all end sequence pairs, not necessarily from the same subclone) of the sum (over all possible alignments of such a pair of end sequences) of the indicator function that two random end sequences with this alignment will be declared to overlap. An extra factor of 2 is due to the fact that the two end sequences might have initially come from complementary strands. The expected number of false overlaps is then twice the number $n(n - 1)/2$ of end sequence pairs times the sum (over all possible alignments) of the probability that two end sequences evaluated at this alignment will be declared to overlap

$$\lambda_{\text{false}} = n(n - 1) \sum_{i=1}^{2m-1} B^{+}\{i - 2(i - m)_{+},\ k[i - 2(i - m)_{+}],\ \alpha\}, \qquad [3]$$

where "positive part" notation x_+ is defined to be x if $x > 0$ and 0 otherwise. Note that λ_{false} is a slight overcount of the number of falsely declared overlaps because we have not excluded true alignments from the calculation.

- Each *simple repeat family* is modeled as a group of randomly dispersed segments with similar sequences, with these segments conditionally independent given a family-prototype segment. This specification model expresses both similarity and randomness in a tractable manner. For family f (where $f = 1, \ldots, \phi$), where ϕ denotes the number of families, we define the following:
  - L_f is the length, in bases, of each segment of the family. For example, for MIR repeats, L_f is 260.
  - R_f is the number of segments in the family. For example, with MIR repeats, R_f is 11,000. We will also allow random R_f from a known probability distribution. Allowing a random R_f permits the modeling of repeat families with approximately known or unknown numbers of members.
  - We assume that each family has a prototype segment (not necessarily present in the genome) consisting of L_f bases selected independently at random from the α_A, α_C, α_G, α_T distribution. For tractability, we ignore any deviations in the GC content of repeats from the genomic average.
  - ϵ_f is the *problem rate* for family f (using the same terminology

as earlier, even though these are not really "problems"). We assume that each segment's bases are independently determined. A base is identical to the homologous prototype base with probability $1 - \epsilon_f$. With probability ϵ_f, the base is instead an independently sampled random base (with distribution specified by α_A, α_C, α_G, and α_T) that may, by coincidence, be the same base as in the prototype.

- ϵ_f^* is the *difference rate* for family *f*, i.e., the probability that a given base in one segment differs from the homologous base in another segment from the same family. Relationships are $\epsilon_f^* = [1 - (1 - \epsilon_f)^2](1 - \alpha)$ and $\epsilon_f = 1 - \sqrt{1 - \epsilon_f^*/(1 - \alpha)}$, because differences can occur whenever either or both segments differ from the prototype. The *mean similarity* of a family is $1 - \epsilon_f^*$.
- p_f denotes the probability that two homologous bases within family *f* will be read as identical. Taking account of reading errors, we may write (following Theorem 6 of Siegel *et al.,* 1999):

$$p_f = \alpha + (1 - \epsilon_f)^2(1 - \epsilon)^2(1 - \alpha). \qquad [4]$$

- $\lambda_{\text{family,false}}$ denotes the expected number of false overlaps declared due to unidentified repeat-family homology, which can happen when two end sequences contain a homologous, but not a true, overlap. Identified repeat families are treated separately in the decision rule, which automatically rejects overlap declaration within known repeat regions. The total expected number of false overlaps due to unidentified homology will be found by summing the rate of false overlaps over all families

$$\lambda_{\text{family,false}} = \sum_{f=1}^{\phi} \lambda_{f,\text{false}}, \qquad [5]$$

where formulas for $\lambda_{f,\text{false}}$ in two cases are now developed.

- First consider unidentified families where the size L_f of the segment is large compared to the size *m* of the end sequence. In this case, for tractability, we will deal conservatively with edge effects by computing as though each end sequence that overlaps a segment is entirely within it. To count the number of such false overlaps, express it as the sum [over all $R_f(R_f - 1)/2$ pairs of segments] of the sum (over the number of end sequences overlapping the first segment of the pair) of the sum (over the number of end sequences overlapping the second segment of the pair and homologously overlapping the first end sequence) of the indicator function that these two end sequences are declared to overlap. The expected number of false matches, if $L_f \gg m$, is therefore

$$\lambda_{f,\text{false}} = \left(\frac{R_f(R_f - 1)}{2}\right)\left(\frac{L_f + m - 1}{T - m + 1}\, n\right)\left(\frac{n - 1}{T - m + 1}\right)$$
$$\times \sum_{i=1}^{2m-1} B^+\{i - 2(i - m)_+,\ k[i - 2(i - m)_+],\ p_f\}. \qquad [6]$$

Note that if a repeat family has a random number of segments R_f, we may count the expected number of false overlaps by using the expected value $E\{[R_f(R_f - 1)]/2\}$ in place of the first term in Eq. [6]. Similarly, we may use the mean value if L_f is random and uncorrelated with R_f.

- Next, consider unidentified families for which $L_f < m$. Computing $\lambda_{f,\text{false}}$ becomes more complex because convolutions of binomial distributions must be considered to assess a proposed overlap that includes some homologous bases and some random bases. Theorem 1 in Appendix A derives an upper bound on $\lambda_{f,\text{false}}$ for the case in which $L_f < m$.
- Each *compound repeat family* is modeled as two interacting subfamilies, called types A and B. Each subfamily is a simple repeat family characterized by a difference rate that applies when comparing two segments within that subfamily. In addition, a difference rate may be specified for use when comparing a segment of type A to a segment of type B within the same compound repeat family. This model represents repeat families that have evolved with periodic seeding resulting in a subfamily structure. Such families include both the *Alu* and the LINE families. For unidentified compound repeat families (recall that identified families are considered directly in the decision rule), in addition to using Eqs. [6] and [8] to count false matches within each subfamily, we can change the term $R_f(R_f - 1)/2$ in these equations to $R_{f,\text{A}}R_{f,\text{B}}$ (the product of the number of segments in each subfamily, representing the number of homologous comparisons) to count false matches due to interaction within the compound family. Modification of Eq. [9] would be more complex.
- We can now, as our final theoretical goal, specify the particular functional form used for the decision rule $k(j)$ that is used in the absence of identified repeat homology

$$k(j) = \text{int}\{\min[j,\ jp_{\max} + \gamma\sqrt{jp_{\max}(1 - p_{\max})}]\}, \qquad [7]$$

where "int" is the integer part function, $p_{\max} = \max_{f=1,\ldots,\phi} p_f$, and γ is a parameter that can be used to control the probability of declaring overlap. The functional form is motivated by a family of one-sided statistical hypothesis tests with type I error controlled by γ, with the null hypothesis corresponding to the maximum repeat family similarity and with the research hypothesis representing identical bases observed with reading errors. This decision rule is set at γ standard deviations above the mean of the binomially distributed "number of aligned bases read as identical" for the repeat family with the most similarity. This stringency protects against falsely declaring unidentified homology as overlap.

The Simulation Model for Scaffold Size

A computer-simulation model was developed to study the properties of scaffolds constructed from true, declared end-sequence overlaps. We simulated random subclone locations chosen along a connected portion of the target (e.g., a chromosome) and the resulting linkage of end sequences into contigs and scaffolds. False overlaps (whether due to homology or not) were not considered within the simulation model for three reasons: (i) the occurrence of such false overlaps has already been accounted for and controlled by the theoretical model (and, by adjusting γ, may be set at less than one expected false overlap in the entire genome), (ii) if false overlaps were added to the simulation, the resulting scaffolds could not be smaller, and (iii) assembly algorithms are highly likely to identify and remove false overlaps by recognizing topological inconsistencies and identifying mismatches either by correlating multiple paired sequences or identifying inconsistencies in multiple alignments. Since our model labels a potential overlap as false based solely on an isolated comparison of two reads (in the absence of known repeat homology), a project with 10-fold sequencing redundancy is highly likely to be able to resolve such false overlaps by considering multiple alignments.

Each trial of the simulation model proceeds as follows:

1. Repeat segment locations are chosen uniformly at random within a chromosome of the target, representing the identified repeat families, subject to the constraint that they not overlap one another and reflecting the rates of occurrence of each family.
2. Subclone locations are randomly selected, subject to constraints relating to the insert size of each subclone group.
3. End sequences are identified and sorted by left end point.
4. Contigs are formed. Beginning with the leftmost end sequence, each sequence is tested independently to see if it is declared to overlap the sequence to its right, using the decision rule identified earlier (so that overlaps entirely within known repeat families are rejected, but overlaps with sufficient known repeat alignment and unique sequence alignment are accepted). In the absence of known repeat homology, this test is performed using the binomial conditional probability $p_{\text{true}}(j)$ of detecting an overlap given the true size *j* of the overlap (if any) in basepairs. Note that this procedure is conservative because only immediate overlaps are considered. If an overlap of two neighboring end sequences is not declared, then a gap

TABLE 2

Repeat Family Characteristics Used in Calculations

Name	Families	Bases (L_f)	Segments in each subfamily[a]		Assumed identified?
			Type A: ($R_{f,A}$)	Type B: ($R_{f,B}$)	
Alu	1	320	678,000	273,000	Yes
MIR	1	260	11,000	0	Yes
LINE1	1	Exponential with mean 1,000	52,000	64,000	Yes
LTR element	75	4,000 to 12,000	25	5	Yes
LTR, isolated	75[b]	10% of corresponding LTR element length	500	100	Yes
DNA transposon	50	80 to 3,000	70	0	No
Highly similar blocks	200	1,000 to 40,000	0	2 to 30	Yes
Moderately similar blocks	200	1,000 to 40,000	2 to 30	0	No

[a] Within a subfamily the mean similarity is 85% for two segments of type A and is 95% for two segments of type B. The mean AB similarity across subfamilies within a family is 85%. Similarities are binomially distributed: for example, there are 2,628,203 *Alu* pairs expected with greater than 99% similarity. Characteristics of interspersed repeats are modeled on Smit (1999).

[b] Each isolated LTR family is paired with a LTR element family and has homology with both ends of the LTR elements.

is considered to exist even if the next end sequence would be declared to overlap them both.

5. Scaffolds are identified by considering the connections among contigs implied by paired subclone ends.

RESULTS

Sequencing on the Scale of the Human Genome

We investigated the properties of the end-sequence strategy for sequencing a target of length $T = 3.5$ Gb, using libraries of subclones of three lengths, as shown in Table 1, and end sequences of length $m = 550$. These libraries therefore cover the target to 10-fold sequencing redundancy and 44-fold mapping redundancy. The choices of parameters in this example are intended to model the human genomic project proposed by Venter *et al.* (1998). Our theory can also be applied to lower redundancy projects. We set the base distribution to $\alpha_A = \alpha_T = 30\%$ and $\alpha_C = \alpha_G = 20\%$. Sequencing errors were assumed to occur $\epsilon^* = 1.5\%$ of the time. These sequencing-error rates are higher than is typical for errors in single reads, but were chosen to improve robustness of the results to polymorphism, as subclones from different haplotypes will be included in data for the Human Genome Project. By comparison, the sequencing error rate in the *Drosophila* project was 0.5% for an average read length of 551 bp (Myers *et al.*, 2000).

We included eight nominal repeat families with mean similarity of 85 and 95%, with characteristics as shown in Table 2. We omitted families with mean similarity below 85%, as these are not likely to result in false declared overlaps. The families in Table 2 model the repeat families in the human genome (Smit, 1999, and references below). Where a range is given, a value is chosen uniformly for each family (for example, there are 50 DNA transposon families, each one with 70 segments of identical length chosen from the uniform distribution from 80 to 3000 bp). Some of these families are assumed to be identified during contig assembly, as indicated in Table 2. For unidentified families, our model presumes a worst-case situation in which repeats are not masked.

LINE1 family members are usually present as truncated forms in the genome, and they are modeled here as with length chosen as independent exponential random variables each with mean segment length 1000 bp and in register at one end.

LTR-element families are bounded on both ends by LTRs. In addition to LTR-element family members, there are many isolated LTRs in the genome (Smit, 1999). We model 75 LTR element families, with each family modeled as a compound repeat family with constant segment length drawn from the uniform distribution. Each of the 75 LTR-element families is paired with a set of isolated LTRs with segment length equal to 10% of the LTR-element length.

Because the DNA transposon families are not assumed to be identified, we compute the number of expected false matches using Eq. [6] and the mean segment length $L_f = 1{,}540$, which results in less than one false overlap expected for all 50 such families (the computed number of false overlaps is 0.002 using $\gamma = 3$ in the decision rule).

In addition, there are a number of highly similar blocks in the human genome. These include pericentromeric and subtelomeric paralogous duplications (reviewed in Eichler, 1998; Trask *et al.*, 1998a), multigene families such as the olfactory receptor family, and small blocks of the antigen receptor families (Trask *et al.*, 1998b; Hood *et al.*, 1995), as well as less-well-characterized blocks. Within the moderately similar gene and homology-block families, we model 200 families, which are not assumed to be identified during assembly, using the mean segment length of 20,500 bp for L_f in Eq. [6]. In addition, because the number of segments per family varies, we replaced the term $R_f(R_f - 1)/2$ in Eq. [6] with its expectation 155 derived when R_f has a uniform discrete distribution from 2 to

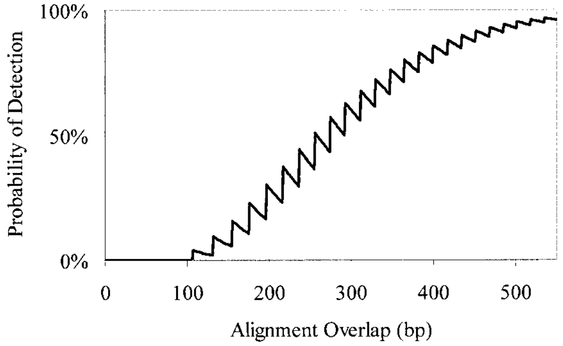

FIG. 1. The probability of detecting a true overlap (choosing γ = 3.0 to keep the expected number of false overlaps well below one for the entire genome) as a function of the size of the aligned overlap being tested. Because ends occur about every 50 bp, adjacent ends will overlap by about 500 bp and are likely to be detected. Discontinuities are due to the discrete nature of the decision function k.

30. Use of both adjustments simultaneously (length per segment and number of segments) may be justified by assuming that the length and the number of segments are uncorrelated. The result from Eq. [6] is that less than one false overlap is expected for all 200 moderately similar block families (the computed number of false overlaps is 0.006 using $\gamma = 3$ in the decision rule).

Overall, less than one falsely declared overlap, on average, was found to occur in the entire 3.5-Gb project, according to the theoretical model, with the choice $\gamma = 3$ to specify the overlap decision rule in Eq. [7]. Using this rule, the probability of detecting a true overlap varies as a function of the length of aligned overlap being tested (Fig. 1). Adjacent end sequences are likely, but not guaranteed, to be declared to overlap. For a given end sequence not involving identified repeat homology, the probability that the first true overlap (in a given direction) is detected is 0.927.

We simulated scaffold formation 100 times in a 250-Mb chromosomally sized target (human chromosomes vary from ~50 to ~250 Mb). Of these 100 simulations, in 99 cases there was a single main scaffold across the entire chromosome, while in the single remaining case there were two main scaffolds (one covering 89% of the chromosome, the other covering 11%) that overlapped each other but did not connect within a local region with multiple blocks of repeat-segment homology. After following our initial assembly algorithm, relying mostly on all pairwise comparisons, it is expected that one would reexamine missed true overlaps. It is likely that these missed overlaps will then be detected within the context of the much simpler problem of linking a small number of scaffolds. For this reason, the two main scaffolds of this one simulation would be merged, and a single main scaffold would be found in all 100 simulations.

Averaging over the 100 simulations, we found that 97.5% of all end sequences were in the main scaffold, which included 99.0% of all bases in the target and left 2730 actual sequence-mapped gaps across the chromosome.

Sequencing the Drosophila Genome

We modeled scaffold formation and false matches for the *Drosophila* Genomic Project, choosing parameter values to match those implemented by Adams *et al.* (2000). This genome has a high percentage of highly similar repeat families and is smaller than the human genome. We investigated the properties of the end-sequence strategy on a target genome of length $T = 125$ Mb, using libraries of subclones of three lengths, with end sequences of length $m = 551$. We set the base distribution to $\alpha_A = \alpha_T = 28.8\%$ and $\alpha_C = \alpha_G = 21.2\%$. Sequencing errors were assumed to occur at rate $\epsilon^* = 0.5\%$. We included 100 repeat families, half identified and half not identified, of varying lengths with varying numbers of segments and similarity. We adjusted the decision rule γ to obtain less than one false match for the unidentified homologous families across the entire genome. Scaffold simulation created repeat segments for the identified families for a target chromosome arm of size 30 Mb and found a single scaffold across nearly the target chromosome arm in 99 of 100 simulations, with one simulation needing two overlapping scaffolds (covering 59.6 and 40.6% of the target chromosome arm) that failed to connect across a repeat-rich region using our conservative overlap detection rules. Averaging across all 100 simulations (and merging the two main scaffolds in one simulation), we find that scaffolds cover 99.997% of the target chromosome arm from start to finish, including 99.94% of all bases with 102 actual sequence-mapped gaps.

DISCUSSION

We have presented a hybrid theory/simulation model for analyzing genomic sequencing performed with a pairwise end-sequencing strategy. Our theoretical model derives the expected number of falsely linked clone ends due to random coincidence and unidentified homology, while our simulation model shows how the true detected overlaps form contigs and scaffolds. Within the context of our model, library parameters may be changed to study their effect on false overlaps and on scaffold size. Using this model, various parameters—such as the number of end sequences determined, the average insert size, the overlap decision rule, the costs of various steps in the procedure—can be varied to assess the effect on overall costs or success of the approach. We anticipate that the flexibility of our mathematical model will permit it to be used generally to optimize parameters not only for the Human Genome Project, but also for future genome sequencing projects. Furthermore, our model is capable of guiding the refinement and development of assembly algorithms.

Our results are not tied to a particular assembly algorithm. We model and simulate the sequence of the target genome and the sequence and locations of subclones. We then infer that assembly algorithms with certain characteristics will produce scaffolds spanning this target genome. Our model assumes that algorithms will be able to recognize the presence of mislabeled clones. With sufficient data, most algorithms will be able to exclude such clones based on an excess of contradicting correct data. Furthermore, as is standard and necessary for modeling genome strategies, we do not model unclonable (or difficult to clone) regions; each of these regions is likely to cause a break in the continuity of a scaffold. Centromeres are examples of such regions.

Our theoretical calculations find an upper bound on the rate of occurrence of false matches under the assumption that an all-by-all comparison is performed on raw sequence reads. In some cases, this task could be computationally slow. However, a number of computational approaches are likely to make such comparisons feasible even on very large data sets. These include hashing and sorting techniques, binning based on external mapping data, and phased merging of raw reads into high-confidence contigs. In the presence of such computational refinements, our results continue to provide an upper bound on the false-match rate.

Myers *et al.* (2000) provide the results of a particular assembly algorithm for the *Drosophila* genome. In this assembly, the four *Drosophila* chromosomes gel into 6 to 25 major scaffolds, with the gaps between these scaffolds occurring near centromeres and telomeres, as expected for unclonable regions. The results of the *Drosophila* Genome Project conform to the predictions of our model, and vice versa. The coverage in sequence reads and clone lengths for this project are analogous to the parameters of the Human Genome Project that we have modeled.

Based on our model, we conclude that, with careful choice of the overlap decision rule, a 3.5-Gb target, such as the human genome, with repeat families having mean similarity of 85 and 95% (and many pairs with much greater similarities), may be successfully sequenced using a sequencing redundancy of about 10-fold. Success is here defined as producing a single scaffold across each of the human chromosomes, with more than 99% of the genome covered in assembled contigs. This result provides independent confirmation of the results of other simulation models for human whole genome shotgunning (e.g., Weber and Myers, 1997).

The cost of a pairwise project is, to a first approximation, the cost of raw sequence read production. The rate of a pairwise project is, to a first approximation, proportional to the rate of raw sequence read production. If scaffolds are computed continuously with the production of raw data, gelling of small scaffolds into large scaffolds tends to occur during a short period of sequence production. The redundancy at which this occurs is currently predictable only by simulation (Roach *et al.*, 1995; Roach, 1998). At this redundancy, projects can be described as undergoing a phase transition from a state of many small scaffolds to a state of one or few large scaffolds.

Assembly problems arise almost exclusively from the presence of repeat families in the genome. The severity of these problems is directly related to the degree of similarity, the copy number of each given family, and the sequencing error rate. For this paper, we have attempted to model the repeat families of the human genome as reasonably as possible. However, our knowledge of repeat families is still incomplete. We have chosen our assumptions to be conservative, so our model will tend to predict an upper bound on errors. One source of our conservatism lies in our consideration of only adjacent reads for establishing contig connectivity. Actual algorithms employ overlapping sets of reads to resolve ambiguity and thus will be able to resolve many of the errors that we predict. We assume that most repeats can be identified before or during all-by-all comparisons of sequence reads. This task is usually accomplished by comparing sequences to known repeat families, such as with the program RepeatMasker (A. Smit, San Diego, CA, pers. comm., 1999). Previously uncharacterized repeats may also be identified by statistical consideration of the number of expected reads over a given region.

Assembly problems due to repeats are not unique to pairwise shotgun sequencing. All random subcloning strategies are susceptible to confoundment by repeats. In general, a repeat longer than the length of an effective mapping unit may represent an insurmountable problem if the repeat is of sufficiently high similarity relative to the sequencing error rate. For traditional shotgun sequencing, the mapping unit is the length of a sequence read. For pairwise end sequencing, the mapping unit is the length of the subclone. This feature drives the motivation for choosing clones of varying lengths, which allows one to circumvent regions of the lengths of all the genomic repeat families (Roach *et al.*, 1995). This is a major distinction between pairwise end sequencing and traditional shotgun sequencing.

Aside from their utility in resolving repeats, the use of subclones of longer lengths is also critical for closure. Mapping redundancy is the most important factor driving the coalescence of contigs into a single scaffold per chromosome (Roach *et al.*, 1995). Plasmid end sequences provide little mapping redundancy, but are the most economical and accurate contribution to sequencing redundancy. Therefore, to avoid gaps in subclone coverage of the target, there is a need for longer clones. The addition of BAC end sequences provides an excess of mapping redundancy. Intermediate length subclones contribute to both mapping and sequencing redundancy and also facilitate certain computations involved in scaffold assembly.

Our analysis predicts that whole-genome pairwise end strategies can yield extensive sequence coverage of the human genome, assuming that the subclone inserts are random. Pairwise end sequencing is cost-effective, in that it eliminates the need to precede sequencing with a mapping phase. Producing high-quality, random subclone libraries is the main up-front cost. Following this, sequence generation is automatable. For the human genome, much of the necessary sequencing is complete: over 800,000 sequences from the ends of human BAC clones are available (http://www.ornl.gov/meetings/bacpac) as are a substantial number of sequences from smaller clones (see Smaglik and Butler, 2000).

APPENDIX A

Theorem 1. For an unidentified homologous family f with R_f segments each of length L_f for which $L_f < m$, an upper bound on the expected number $\lambda_{f,\text{false}}$ of false overlaps due to homology may be computed as the sum of the two equations

$$\frac{n(n-1)R_f(R_f-1)}{2(T-m+1)^2}\{2\sum_{i=1}^{L_f}\sum_{j=1}^{L_f}\text{convolute}_f[\min(i,j), m-\max(i,j), k(m-|i-j|)] + 4\sum_{i=1}^{L_f}\sum_{j=L_f+1}^{m-1}\text{convolute}_f[i, m-j, k(m+i-j)] + 2\sum_{i=1}^{L_f}\sum_{j=m}^{m+i-1}B^+[m+i-j, k(m+i-j), p_f] + \sum_{i=L_f+1}^{m-1}\sum_{j=L_f+1}^{m-1}\text{convolute}_f[L_f, m-|i-j|-L_f, k(m-|i-j|)]\} \quad [8]$$

$$+ \frac{n(n-1)R_f^2(R_f-1)^2}{2(T-L_f+1)^2(T-m+1)^2} \times \sum_{d=0}^{m-2}\sum_{i=1}^{m-d-1}\sum_{j=1}^{m-d-1}\text{convolute}_f\{\min[L_f, m-d-\max(i,j)] + \min(L_f, i, j), m-|i-j|-\min[L_f, m-d-\max(i,j)]-\min(L_f, i, j), k(m-|i-j|)\}, \quad [9]$$

where the convolution function is defined as the following upper cumulative distribution function of the sum of two binomial distributions, $X_{j,f}$ based on homologous-base matches and X_r based upon random matches

$$\text{convolute}_f(j, r, k) = P(X_{j,f} + X_r > k) = \sum_{i=(k-r)_+}^{j} b(j, i, p_f)B^+(r, k-i, \alpha) \quad [10]$$

and where the binomial probability b is defined as

$$b(j, i, p) = \binom{j}{i} p^i(1-p)^{j-i}. \quad [11]$$

Proof. Equation [8] represents the expected number of homologous overlaps involving two end sequences whose overlap region includes a single repeat segment, computed as the expected value of the sum (over all repeat-segment pairs) of the sum (over all end-sequence pairs) of the sum (over all placements of each end sequence to overlap its repeat segment) of the indicator function that a false overlap is declared due to homology. The first two of these sums are represented in the initial multiplier, which counts the number of ways two end sequences can each be associated with a different repeat segment, then multiplies by the probability $1/(T-m+1)^2$ that each end sequence falls randomly at a particular location of its repeat segment. The four individual double summations represent the four possible overlap configurations: Case 1: both end sequences have an end point in their respective repeat segments and either both end points are 5′ ends or both are 3′ ends (factor of 2 is because there are two possible assignments of end sequences to 5′/3′ ends). Case 2: one end sequence has an end point in its repeat segment, while the other has both end points outside of (therefore completely containing) its repeat segment (the factor of 4 is because either end sequence could have its end point in its repeat segment, and this end point could be either 5′ or 3′). Case 3: both end sequences have an end point in their respective repeat segments but one end point is 5′ and the other is 3′ (hence the factor of 2). Case 4: both end sequences have both end points outside of (therefore completely containing) their respective repeat segments. In each of these four cases, the summations extend over all placements of end sequences overlapping repeat segments. The probabilities being summed in each case represent the likelihood of falsely declaring an overlap when the end sequences are aligned to match the homologous repeat segment region [for which two bases will be read as identical with probability $p_f = \alpha + (1-\epsilon_f)^2(1-\epsilon)^2(1-\alpha)$]. Note that the nonhomologous portion (if any) of the overlap being tested will have two bases read as identical with probability α.

Equation [9] represents the expected number of homologous overlaps involving two end sequences whose overlap region includes part of two repeat segment sequences, each separated by d unique-sequence bases, as shown in Fig. 2. This expectation is computed as the expected value of the sum (over all pairs of ordered repeat-segment pairs with exactly d unique-sequence bases in between) of the sum (over all end-sequence pairs) of the sum (over all placements of each end sequence to overlap its repeat-segment pair) of the indicator function that a false overlap is declared due to homology. The first two of these sums are represented in the initial multiplier, which counts the ex-

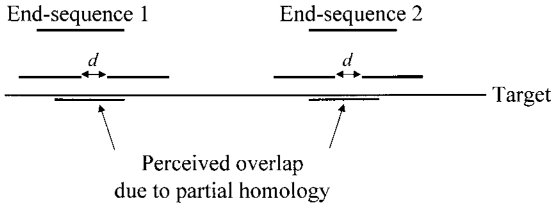

FIG. 2. Two end sequences, each overlapping a pair of small unidentified repeat segments separated by *d* bases. When aligned according to the perceived overlap due to partial homology, these end sequences may falsely be declared to overlap.

pected number of ways two end sequences can each be associated with a different repeat-segment pair, then multiplies by the probability $1/(T - m + 1)^2$ that each end sequence falls randomly at a particular location of its repeat-segment pair [note that if the number of repeat-segment pairs separated by *d* bases has a Poisson distribution with mean $\lambda = R_f(R_f - 1)/(T - L_f + 1)$, then the expected number of *pairs* of such repeat-segment pairs is $\lambda^2/2$]. The triple sum extends over all possible repeat-segment-pair separations *d*, as well as placements of the two end sequences on their respective repeat-segment pairs.

The result is an upper bound because an end sequence that overlaps two repeat segments will be counted separately for each of these two repeat segments in Eq. [8] and because two end sequences that each overlap two repeat segments may be counted in both [8] and [9]. □

ACKNOWLEDGMENTS

This work has been supported in part by NIH Grant GM57070. Andrew F. Siegel holds the Grant I. Butterbaugh Professorship at the University of Washington.

REFERENCES

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. C., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Miklos, G. L. G., Abril, J. F., Agbayani, A., An, H., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., Pablos, B. D., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S. M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D. C., Scheeler, F., Shen, H., Shue, B. C., Sidén-Kiamos, I., Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., Woodage, T., Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M., and Venter, J. C. (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Anson, E., and Myers, E. W. (1999). Algorithms for whole genome shotgun sequencing. *In* "Proceedings of the Third International Conference on Computational Molecular Biology" (S. Istrail, P. Pevzner, and M. Waterman, Eds.), pp. 1–9, ACM Press, New York.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1474.

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

Edwards, A., and Caskey, T. (1991). Closure strategies for random DNA sequencing. *Methods Comp. Methods Enzymol.* **3:** 41–47.

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

Eichler, E. (1998). Masquerading repeats: Paralogous pitfalls of the human genome. *Genome Res.* **8:** 758–762.

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

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

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. (1996). Life with 6000 genes. *Science* **274:** 546, 563–567.

Green, P. (1997). Against a whole-genome shotgun. *Genome Res.* **7:** 410–417.

Hood, L., Rowen, L., and Koop, B. F. (1995). Human and mouse T-cell receptor loci: Genomics, evolution, diversity, and serendipity. *Ann. N. Y. Acad. Sci.* **758:** 390–412.

Myers, E. W. (1999). A dataset generator for whole genome shotgun sequencing. *In* "Proceedings, Seventh International Conference on

Intelligent Systems for Molecular Biology" (T. Lengauer, R. Schneider, P. Bork, D. Brutlag, J. Glasgow, H. Mewes, and R. Zimmer, Eds.), pp. 202–210, AAAI Press, Menlo Park, CA.

Myers, E. W., Sutton, G. G., Delcher, A. A., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., Anson, E. L., Bolanos, R. A., Chou, H.-H., Jordan, C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C., Chen, L., Dunn, P. J., Lai, Z., Liang, Y., Nusskern, D. R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G. M., Adams, M. D., and Venter, J. C. (2000). A whole-genome assembly of *Drosophila*. *Science* **287:** 2196–2204.

Myers, E. W., and Weber, J. L. (1997). Is whole genome sequencing feasible? *In* "Computational Methods in Genome Research" (S. Suhai, Ed.), pp. 73–89, Plenum, New York.

Richards, S., Muzny, D. M., Civitello, D. M., Lu, F., and Gibbs, R. A. (1994). Sequence map gaps and directed reverse sequencing for the completion of large sequencing projects. *In* "Automated DNA Sequencing and Analysis" (M. Adams, C. Fields, and J. Venter, Eds.), pp. 191–198, Academic Press, New York.

Roach, J. C. (1998). "Random Subcloning, Pairwise End Sequencing, and the Molecular Evolution of the Vertebrate Trypsinogens," Doctoral dissertation, University of Washington, Seattle.

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

Siegel, A. F., Trask, B. J., Roach, J., Mahairas, G. G., Hood, L., and van den Engh, G. (1999). Analysis of sequence-tagged connector (STC) strategies for DNA sequencing. *Genome Res.* **9:** 297–307.

Smaglik, P., and Butler, D. (2000). Celera turns to public genome data to speed up endgame. *Nature* **403:** 119.

Smit, A. F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9:** 657–663.

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

Trask, B. J., Friedman, C., Martin-Gallardo, A., Rowen, L., Akinbami, C., Blankenship, J., Collins, C., Giorgi, D., Iadonato, S., Johnson, F., Kuo, W. L., Massa, H., Morrish, T., Naylor, S., Nguyen, O. T., Rouquier, S., Smith, T., Wong, D. J., Youngblom, J., and van den Engh, G. (1998a). Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum. Mol. Genet.* **7:** 13–26.

Trask, B. J., Massa, H., Brand-Arpon, V., Chan, K., Friedman, C., Nguyen, O. T., Eichler, E., van den Engh, G., Rouquier, S., Shizuya, H., and Giorgi, D. (1998b). Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. *Hum. Mol. Genet.* **7:** 2007–2020.

Venter, J. C., Adams, M. D., Sutton, G. G., Kerlavage, A. R., Smith, H. O., and Hunkapiller, M. (1998). Shotgun sequencing for the human genome. *Science* **280:** 1540–1542.

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

Weber, J. L., and Myers, E. W. (1997). Human whole-genome shotgun sequencing. *Genome Res.* **7:** 401–409.