2006

A general coverage theory for shotgun DNA sequencing

Michael C. Wendl
Washington University School of Medicine

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 13, Number 6, 2006

Pp. 1177–1196

A General Coverage Theory for Shotgun DNA Sequencing

MICHAEL C. WENDL

ABSTRACT

The classical theory of shotgun DNA sequencing accounts for neither the placement dependencies that are a fundamental consequence of the forward-reverse sequencing strategy, nor the edge effect that arises for small to moderate-sized genomic targets. These phenomena are relevant to a number of sequencing scenarios, including large-insert BAC and fosmid clones, filtered genomic libraries, and macro-nuclear chromosomes. Here, we report a model that considers these two effects and provides both the expected value of coverage and its variance. Comparison to methyl-filtered maize data shows significant improvement over classical theory. The model is used to analyze coverage performance over a range of small to moderately-sized genomic targets. We find that the read pairing effect and the edge effect interact in a non-trivial fashion. Shorter reads give superior coverage per unit sequence depth relative to longer ones. In principle, end-sequences can be optimized with respect to template insert length; however, optimal performance is unlikely to be realized in most cases because of inherent size variation in any set of targets. Conversely, single-stranded reads exhibit roughly the same coverage attributes as optimized end-reads. Although linking information is lost, single-stranded data should not pose a significant assembly liability if the target represents predominantly low-copy sequence. We also find that random sequencing should be halted at substantially lower redundancies than those now associated with larger projects. Given the enormous amount of data generated per cycle on pyro-sequencing instruments, this observation suggests devising schemes to split each run cycle between two or more projects. This would prevent over-sequencing and would further leverage the pyro-sequencing method.

Key words: shotgun sequencing coverage, dimensional analysis, stopping problem, optimization, pyro-sequencing.

1. INTRODUCTION

RANDOM SHOTGUN SEQUENCING of genomic DNA is fast becoming commonplace and its importance in the larger context of biomedical research continues to grow. The technique is now routinely applied over a wide array of projects, for example, for individual BAC clones (Int. Human Genome Seq. Consortium, 2001), genic islands (Whitelaw *et al.*, 2003), small prokaryotic genomes (Fraser *et al.*, 1997), and

Genome Sequencing Center, Washington University, St. Louis, Missouri.

large eukaryotic genomes (Mouse Genome Seq. Consortium, 2002). Investigators recognize the need for broad classes of DNA sequence in addressing biomedical questions and this continues to place a growing demand on the rate of sequence production.

To some extent, the necessary increases in throughput are being met by advances in instruments, software, and chemistry. Commercialization of pyro-sequencing technology is especially exciting in this regard (Ronaghi, 2001). However, optimization remains an important aspect, as well. There are a number of weakly-constrained variables that can significantly affect the efficiency of a project, for example read length, insert length, and the stopping redundancy. Optimal settings for these variables depend on various factors, but cannot generally be determined in an exact sense because of dependencies on uncharacterized, sequence-specific phenomena. These include repeat structures, sequence and cloning biases, and assembly difficulties.

The earliest sequencing efforts, modest by today's standards, could exploit only a qualitative understanding of the shotgun method (Sanger *et al.*, 1977a). Later projects have benefited from an evolving body of theory that seeks to quantitatively model some of the relevant metrics. For example, the expected number of base positions covered after a given number of sequencing reads can readily be derived from the rudimentary representation probability reported by Clarke and Carbon (1976). Similarly, the Lander and Waterman (1988) model furnishes expected values of, for example, the number of gaps and the number of reads in a contig. Theorists have continued to refine these concepts (Roach, 1995; Wendl and Waterston, 2002) and have also developed more specialized models, for example, the A-statistic for repeat detection in assemblies (Myers *et al.*, 2000). Because sequence-specific effects cannot be easily characterized *a priori*, such descriptions are invariably based on the closure assumption of independently and identically distributed (IID) sequence reads.

Although the above results are well-established, there are a number of important issues that have yet to be adequately addressed. For example, one of the now-fundamental aspects of the shotgun method is sequencing randomly selected double-stranded DNA fragments from both ends. That is, "forward" and "reverse" reads are generated for each distinct insert (Fig. 1). This strategy creates linking information at a variety of scales that significantly facilitates the downstream phase of sequence assembly (Myers *et al.*, 2000). Yet, it also implies a strong placement dependence between paired reads, which classical theory does not consider.

Another relevant factor is the wide variation in target size. On one extreme lies the case where the target is a sizable genome, whose length γ is much greater than the read length λ. For example, λ/γ in the mouse project was on the order of 10^{-7} (Mouse Genome Seq. Consortium, 2002). Under these circumstances, each base position can be considered to have an essentially equal probability of being covered by a read, λ/γ. Indeed, this uniformity is a fundamental assumption in classical theory. Conversely, coverage probability is not constant for a small target, for example, a genic island (Whitelaw *et al.*, 2003). Here, λ/γ can be of order 10^{-1}, implying that the "edge effect" must be considered. This phenomenon is simply a position-dependent sampling bias. In particular, base positions in the neighborhood of a terminus have a lower covering probability than those further away.

Classical theory has firmly established itself in the roles of estimating and troubleshooting for shotgun sequencing projects. In what has largely been the typical scenario, λ/γ is very small, whereby pairing effects and edge effects are taken to be negligible. The main assumptions for applying classical theory are satisfied. However, investigators are increasingly interested in projects where read size and insert size are not vanishingly small compared to the genomic target. Examples include sequencing 40 kb fosmid clones with 2–3-kb inserts and full-length reads (Shimada *et al.*, 2005), sequencing 2–10-kb genomic islands from

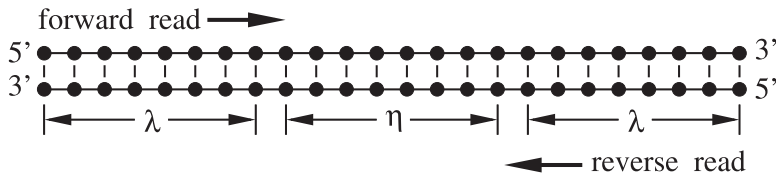

FIG. 1. Schematic of a double-stranded DNA insert sequenced from both 5′ ends. Each dot represents the position of a nucleotide.

filtered libraries (Whitelaw *et al.*, 2003), and sequencing small macro-nuclear chromosomes of average size <20 kb (Doak *et al.*, 2003). Standard models may be insufficient for these cases.

Some contextual background is helpful at this point. While early sequencing efforts also concentrated on covering segment lengths of order 10^4, these were all instances of a *single target*. Refinements of theory would hardly be justified here because such projects are now viewed as trivial. Conversely, the above scenarios involve *multiple target segments*. For instance, a mammalian genome project might traverse a minimum tiling path of order 10^5 fosmids and the number of macro-nuclear chromosomes in certain ciliates is of order 10^4. The number of genic islands in a filtered library can be similarly large. The multiplication of effort is obvious and invites deeper study of the covering process for the small target.

Our primary interest is in constructing a more comprehensive theory that integrates models for both the read-pairing effect and the edge effect. Although Wendl and Barbazuk (2005) considered edge effects for sequencing filtered libraries, the coverage component of their model is valid strictly for single-stranded sequencing. Current trends suggest that a significant fraction of future projects may employ end-sequencing from double-stranded templates. In fact, double-stranded sequencing is now largely the norm for large-scale projects. The model proposed here should therefore lead to better characterization of the shotgun process for the scenarios mentioned above.

We take the random variable C as the number of base positions in a genomic target covered by one or more reads. More formally, C is defined as the linear measure of the union of reads (Siegel, 1978). The obvious difficulty in quantifying C is the fact that reads can overlap one another. Robbins (1944) demonstrated how the moments of measure can be determined without knowing the underlying distribution of the measure itself and investigators have used this concept to determine the moment sets for a variety of covering problems. Here, we follow a variation of this concept for discrete configurations (Kolchin *et al.*, 1978) to construct the expected value and variance of coverage via the first and second moments. Higher moments could, in principle, be derived as well. However, the procedure becomes progressively more tedious, given pairing and edge effects. Moreover, higher moments are not typically of interest for practical sequencing calculations. Results are then applied to characterize the genomic coverage for some of the scenarios mentioned above.

2. RESULTS

Figure 1 shows a diagrammatic representation of forward and reverse sequencing reads from a cloned insert. Table 1 describes the notation used in constructing the theory. Sequencing reads have a length of λ base pairs (bp) and are separated by η non-participating bases. The latter are not elucidated by the forward or reverse reactions. Read length may include a reduction factor to account for the number of bases effectively lost in detecting overlaps (Lander and Waterman, 1988). For the typical project, $\eta > 0$ and forward and reverse read lengths are very nearly identical. For example, in examining the last 150 million sequencing reactions processed in our laboratory, we find an average length $\lambda = 600$ with a difference of only 2 bp between forwards and reverses.

TABLE 1. BASIC NOTATION USED IN THE MODEL

Variable	*Meaning*
γ	Length (bp) of genomic target
λ	Length (bp) of forward and reverse reads
η	Length (bp) of non-participating region between paired reads
τ	Length (bp) of insert templates ($\tau = 2\lambda + \eta$)
π	Number of possible insert placements ($\pi = \gamma - \tau + 1$)
μ_f	Right-most position reachable by a forward read ($\mu_f = \gamma - \lambda - \eta$)
μ_r	Left-most position reachable by a reverse read ($\mu_r = \lambda + \eta + 1$)
n	Number of inserts that have been processed
C	Number of bases covered by at least 1 read (random variable)
V	Number of bases not covered (random variable)
x, y	Independent target coordinates

Let the genomic target be a segment γ bp in length with a left-anchored coordinate system $x \in \{1, 2, 3, \ldots, \gamma\}$. Inserts are assumed to be IID over this region, although individual read pairs are dependent, as described above. For our purposes, an insert can be conceptualized according to a single idealized covering segment, as shown in Fig. 2. That is, the actual double-stranded structure is abstracted away. We presume, as an absolute upper bound, that the insert length is less than the target length, i.e., $\tau < \gamma$. Except for one special case treated below, we also neglect inserts that fall partially outside the target region. Such would occur, for example, where one read hits the target while the other lies in a flanking vector region, or where partial re-association in a filtering procedure leads to an insert straddling the target boundary. Under this assumption, there are $\pi = \gamma - \tau + 1$ possible placements of an insert on the target. Finally, let n be the number of inserts that have been processed, so that there are a total of $2n$ reads distributed over the target.

2.1. General theory

The general case having both pairing and edge effects is modeled as follows.

Theorem 1 (Expected Coverage). *The expected value of coverage is given by the first moment of C,*

$$E\langle C\rangle = \gamma - \sum_{x=1}^{\gamma}\left[\left(1 - \frac{f(x)}{\pi}\right)\left(1 - \frac{r(x)}{\pi - f(x)}\right)\right]^n,$$

where $f(x)$ and $r(x)$ are defined in Lemma 1.1.

Lemma 1.1. *The number of ways a forward read can cover an arbitrary target position x is*

$$f(x) = \begin{cases} x & : \text{ if } 1 \le x < \lambda \text{ else} \\ \lambda & : \text{ if } \lambda \le x \le \pi \text{ else} \\ \mu_f - x + 1 & : \text{ if } \pi < x \le \mu_f \text{ else} \\ 0 & : \text{ otherwise,} \end{cases}$$

where $\mu_f = \gamma - (\lambda + \eta)$ is the right-most possible position that can be reached by a forward read. The number of ways a reverse read can cover an arbitrary target position x is

$$r(x) = \begin{cases} 0 & : \text{ if } 1 \le x < \mu_r \text{ else} \\ x - (\lambda + \eta) & : \text{ if } \mu_r \le x < \tau \text{ else} \\ \lambda & : \text{ if } \tau \le x \le \gamma - \lambda + 1 \text{ else} \\ \gamma - x + 1 & : \text{ otherwise,} \end{cases}$$

where $\mu_r = \lambda + \eta + 1$ is the left-most possible position that can be reached by a reverse read.

Lemma 1.2. *Enumerations for $f(x)$ and $r(x)$ are governed by the bound $3\lambda + \eta - 2 \le \gamma$.*

Although the expected coverage is reported directly, the variance of coverage is more conveniently represented in terms of the random vacancy V, where $V + C = \gamma$ (Siegel, 1978; Kolchin *et al.*, 1978).

Theorem 2 (Variance of Coverage). *The variance of coverage is*

$$\sigma_c^2 = E\langle V^{[2]}\rangle + E\langle V\rangle - E^2\langle V\rangle$$

where $E\langle V^{[2]}\rangle$ is given by Lemma 2.1 and $E\langle V\rangle$ is the expected vacancy. The latter can be computed directly from Theorem 1 as $E\langle V\rangle = \gamma - E\langle C\rangle$.

Lemma 2.1. *The second "falling factorial" moment of vacancy, defined as $E\langle V^{[2]}\rangle = E\langle V^2 - V\rangle$, is given by the expression*

$$E\langle V^{[2]}\rangle = \sum_{x=1}^{\gamma}\sum_{y=1}^{\gamma}\left[\left(1 - \frac{\mathcal{F}(x,y)}{\pi}\right)\left(1 - \frac{\mathcal{R}(x,y)}{\pi - \mathcal{F}(x,y)}\right)\right]^n\left[1 - \delta_{xy}\right],$$

where

$$\mathcal{F}(x, y) = f(x) + f(y) - \phi(x, y)$$

and

$$\mathcal{R}(x, y) = r(x) + r(y) - \rho(x, y) - \xi(x, y).$$

Symbol δ_{xy} is the Kronecker delta and functions $\phi(x, y)$, $\rho(x, y)$, and $\xi(x, y)$ are defined in Lemmas 2.2 and 2.3.

Lemma 2.2. *Consider two arbitrary, but mutually exclusive target positions x and y, where $y > x$. The number of ways a forward read can cover both these positions is*

$$\phi(x, y) = \begin{cases} 0 & : \text{ if } y > \mu_f \text{ or } y - x \geq \lambda \text{ else} \\ x & : \text{ if } y \leq \lambda - 1 \text{ else} \\ \mu_f - y + 1 & : \text{ if } x \geq \mu_f - \lambda + 2 \text{ else} \\ \lambda - (y - x) & : \text{ otherwise.} \end{cases}$$

The number of ways a reverse read can cover both positions is

$$\rho(x, y) = \begin{cases} 0 & : \text{ if } x < \mu_r \text{ or } y - x \geq \lambda \text{ else} \\ x - (\lambda + \eta) & : \text{ if } y \leq \tau - 1 \text{ else} \\ \gamma - y + 1 & : \text{ if } x \geq \gamma - \lambda + 2 \text{ else} \\ \lambda - (y - x) & : \text{ otherwise.} \end{cases}$$

Both of these functions are symmetric with respect to the order of their arguments.

Lemma 2.3. *Consider two arbitrary, but mutually exclusive target positions x and y, where $y > x$. Let $\mathbf{A}$ denote the event whereby $0 \leq y - x - \lambda - \eta \leq \lambda - 1$ and $\bar{\mathbf{A}}$ the event whereby $1 \leq \lambda + \eta - y + x \leq \lambda - 1$. Then, the number of ways a forward read can cover x, while the corresponding reverse read covers y is*

$$\xi(x, y) = \begin{cases} 0 & : \text{ if } y - x \geq \tau \text{ or } y - x \leq \eta \text{ or } y < \mu_r \text{ or } x > \mu_f \text{ else} \\ x & : \text{ if } \mathbf{A} \text{ and } x \geq 1 \text{ and } y \leq \tau - 1 \text{ else} \\ y - \mu_r + 1 & : \text{ if } \bar{\mathbf{A}} \text{ and } y \geq \mu_r \text{ and } x \leq \lambda - 1 \text{ else} \\ \gamma - y + 1 & : \text{ if } \mathbf{A} \text{ and } x \geq \gamma - \tau + 2 \text{ and } y \leq \gamma \text{ else} \\ \mu_f - x + 1 & : \text{ if } \bar{\mathbf{A}} \text{ and } y \geq \gamma - \lambda + 2 \text{ and } x \leq \mu_f \text{ else} \\ \tau - y + x & : \text{ if } \mathbf{A} \text{ else} \\ y - x - \eta & : \text{ if } \bar{\mathbf{A}}. \end{cases}$$

Like $\phi(x, y)$ and $\rho(x, y)$, this function is symmetric with respect to the order of its arguments.

2.2. Special cases

The above results describe a general theory that accounts for both pairing and edge effects. Wendl and Barbazuk (2005) addressed the scenario in which only edge effects are important, that is, single-stranded sequencing on small targets. Here, we treat the complementary simplification where only the pairing effect remains.

Edge effects can vanish in two biologically-relevant ways. First, flanking regions can be attached to the target termini. These regions are not considered part of the target proper, but result in uniform coverage probability. Such a scenario is often invoked to avoid the complication of boundary conditions (Robbins, 1944). Assuming the length of each flanking region is τ, expected coverage and variance are given as follows.

Corollary 3. *The expected value of coverage in the absence of edge effects by way of added flanking regions is*

$$\frac{E\langle C_f \rangle}{\gamma} = 1 - \left[\left(1 - \frac{\lambda}{\pi_f} \right) \left(1 - \frac{\lambda}{\pi_f - \lambda} \right) \right]^n,$$

where $\pi_f = \gamma + \tau - 1$ represents the total number of ways an insert could be placed within the entire region.

Corollary 3 follows directly from Theorem 1 by noting that every base position has a fixed number of ways λ whereby it can be covered by an arbitrary read if edge effects are not present.

Corollary 4. *The variance of coverage in the absence of edge effects by way of added flanking regions is computed from Theorem 2 using the appropriately simplified results for vacancy, that is,* $E\langle V_f\rangle = \gamma - E\langle C_f\rangle$ *and*

$$E\langle V_f^{[2]}\rangle = \sum_{x=1}^{\gamma}\sum_{y=1}^{\gamma}\left[\left(1-\frac{2\lambda-\phi_f(x,y)}{\pi_f}\right)\left(1-\frac{2\lambda-\rho_f(x,y)-\xi_f(x,y)}{\pi_f-2\lambda+\phi_f(x,y)}\right)\right]^n\left[1-\delta_{xy}\right],$$

where

$$\phi_f(x,y)=\rho_f(x,y)=\begin{cases}0 & : \text{ if } y-x\geq\lambda \text{ else}\\ \lambda-(y-x) & : \text{ otherwise}\end{cases}$$

and

$$\xi_f(x,y)=\begin{cases}0 & : \text{ if } y-x\geq\tau \text{ or } y-x\leq\eta \text{ else}\\ \tau-y+x & : \text{ if } \mathbf{A} \text{ else}\\ y-x-\eta & : \text{ if } \bar{\mathbf{A}}.\end{cases}$$

The other scenario in which edge effects vanish is for "large" targets, as characterized by $\gamma \gg \tau$. Here, *almost* all base positions have the same covering probability, λ/γ, regardless of the read direction. Evidently, the pairing effect disappears as a by-product and we have:

Corollary 5 (Clarke and Carbon). *The expected value of coverage converges to standard theory as the target becomes large, in other words*

$$\frac{E\langle C_l\rangle}{\gamma}\sim 1-\left[1-\frac{\lambda}{\gamma}\right]^{2n}\sim 1-\exp\left(-\frac{2n\lambda}{\gamma}\right),$$

where $2n\lambda/\gamma$ *is the* sequence depth *(also called sequence redundancy). This expression is the well-known "Clarke and Carbon" (1976) formula.*

Another consequence of the large target scenario is that $\phi(x,y)$, $\rho(x,y)$, and $\xi(x,y)$ are all negligible. In light of their symmetry, each function specifies $\gamma(\gamma-1)/2$ elements, of which less than $\lambda\gamma$ are non-trivial for $\phi(x,y)$ and $\rho(x,y)$ and less than $2\lambda\gamma$ are non-trivial for $\xi(x,y)$. Consequently, almost all elements vanish, whereby coverage events for positions x and y can be treated exclusively of one another. Physically, this reflects the fact that, on a very large target, two arbitrarily-chosen coordinate positions will rarely be touched by either one or both reads of a given insert.

The phenomenon of exclusivity reduces this scenario to a variation of the classical occupancy problem (Kolchin *et al.*, 1978) with $2n$ independent trials and a success probability of λ/γ for each trial. This immediately implies all vacancy moments, and by extension, all coverage moments for standard theory.

Corollary 6 (Kolchin *et al.*). *The* k*th "falling factorial" moment of vacancy for large targets is*

$$E\langle V_l^{[k]}\rangle\sim\gamma^{[k]}\left(1-\frac{k\lambda}{\gamma}\right)^{2n}.$$

In particular, the $k=2$ *case can be used to estimate the variance of coverage from Theorem 2.*

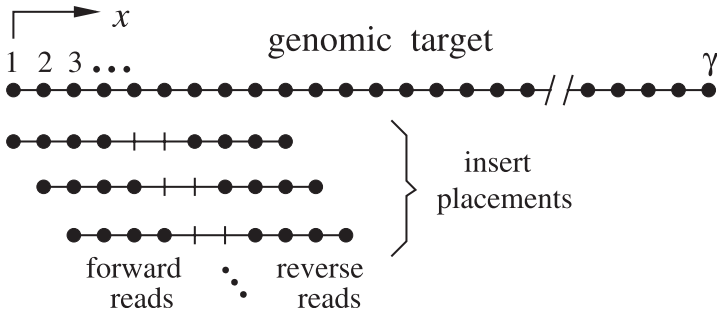

FIG. 2. Schematic of DNA target and its coordinate system and placements of idealized forward/reverse reads. Non-participating regions of the inserts are denoted by tick marks.

2.3. Computational implementation

Evaluations of Theorems 1 and 2 involve summations having the general form

$$E\langle\rangle = \sum_i \kappa_i^n. \tag{1}$$

This expression is taken over γ terms for the first moment and $\gamma(\gamma - 1)$ terms for the second. Here, the κ_i can be thought of as fixed "kernel" terms, which must simply be raised to a power appropriate to the number of reads that have been processed. Values of κ_i depend variously on functions f, r, ϕ, ρ, and ξ and need only to be computed once for each moment in each case of interest.

The number of distinct κ_i values grows very slowly with the target size; variation is limited to regions near the boundaries, which scale roughly with the insert size. Consequently, it is convenient to store only the distinct κ_i, along with their corresponding multiplicative factors, using hash tables. For example, Fig. 3 shows storage requirements versus target size for 600 bp reads derived from 2 kb inserts. The number of distinct κ_i for each moment is approximately constant for targets of roughly $\gamma > 2\tau$. This contrasts dramatically with a simple "raw storage" approach.

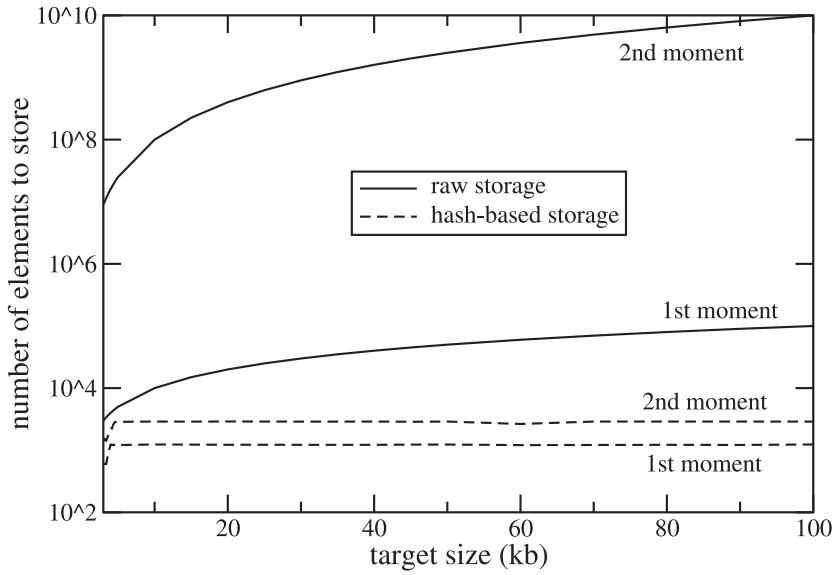

FIG. 3. Raw storage versus hash-based storage required to evaluate first and second moments for 600-bp reads derived from 2-kb inserts.

3. DISCUSSION

The shotgun approach for DNA sequencing has been employed for almost three decades now. Its ongoing use is necessitated by well-known limitations on read length. Standard chain-termination sequencing reactions (Sanger *et al.*, 1977b) can consistently deliver reads up to order 10^3 bp, however, target DNA molecules are often much larger. For example, BAC clones, microbial genomes, and mammalian chromosomes are of order 10^5, 10^6, and 10^8 bp in length, respectively. Although sequencing technologies continue to be developed, none is close to the ultimate goal of reliably reading whole molecules of arbitrary length.

Coverage evolution is well-understood in principle for large targets. It is basically described by the single-variable "standard theory" (Corollary 5). Here, sequence depth is the sole independent parameter and all projects collapse onto a single curve when plotted in terms of the fraction of the target covered. Earlier efforts for viral and cosmid targets (order 10^4 bp) could be characterized to a large extent with this model because read lengths were much shorter and single-stranded templates, e.g., M13, were used almost exclusively. Nowadays, small and intermediate-sized projects can take advantage of significantly longer reads and double-stranded templates, the latter of which permit read pairs to be separated by large distances. Here, edge effects and pairing effects can be important.

In order to establish some idea of its accuracy, we will first compare the theory to a data set of maize genes sequenced with methyl-filtered reads. We will then apply the theory to assess some of the relevant issues in sequencing small to intermediate-size targets (Table 2). Projects are all governed by certain constraints on both read length and insert length. Maximum read length has already been discussed above, but reads must also be at least about 100 bp long in order to be useful for overlap detection and assembly. Thus, we take the read length constraint as $100 \leq \lambda \leq 1000$. Insert size is constrained by the viability of cloning platforms on the upper end. Specifically, BAC clones are the largest stable clones that are routinely end-sequenced. On the lower end, inserts must exceed roughly twice the read length in order to generate distance-linking information for assembly. The practical limits on insert size are thus taken as $1000 \leq \tau \leq 2 \times 10^5$.

3.1. Methyl-filtered maize data: an example comparison

Methyl-filtering (MF) continues to be investigated as a way to sequence the gene sets of the most recalcitrant plant genomes (Palmer *et al.*, 2003; Rabinowicz *et al.*, 2003; Whitelaw *et al.*, 2003; Bedell *et al.*, 2005). The method exploits the fact that characteristic repeats are highly methylated, whereas genes are typically not. Certain cloning platforms, including *E. coli mcr*BC+, only amplify non-methylated DNA. The subsequent library consists predominantly of small genic islands that exhibit the edge effect (see especially Fig. 1 in Rabinowicz *et al.*, 2003).

Maize *(Zea mays)* is representative of high-repeat plant genomes and is being actively examined in this context. Investigators have been especially interested in assessing how well standard end-sequencing would capture its gene space (Springer *et al.*, 2004) and have assembled a set of 152 carefully validated, well-characterized genes for comparison (B. Barbazuk, unpublished data). Moreover, they have determined coverage for individual genes by aligning MF data from a pilot program of 500,000 MF sequences from strain B73 (Whitelaw *et al.*, 2003). These results serve as a useful test case.

Figure 4 shows comparisons of both the current theory (Thm. 1) and classical theory (Clarke and Carbon, 1976) to the maize MF data. The plot focuses on 12 genes that satisfy the following primary inclusion criteria:

1. A gene must be hit by more than two reads. Many showed either no hits, or only one or two hits. The latter scenario provides exact, but trivial agreement with theoretical models.
2. A gene must be hit by an even number of reads. Our model considers only read pairs, meaning that anomalous cases are excluded, for example, when one of the reads of a pair fails.
3. The sequencing parameters for a gene must satisfy Lemma 1.2. Because the maize MF data are derived from full-length reads, this condition eliminated many of the smallest genes.

The average gene in Fig. 4 is 5,600 bp long and is hit by eight pairs of reads. Coverage calculations use gene-specific average read lengths and assume an insert length of 2 kb.

TABLE 2. REPRESENTATIVE SMALL-TO-INTERMEDIATE SIZED TARGETS

γ (bp)	*Description*
9,784	Island 5482.m00088 on maize BAC AY530950 (Whitelaw *et al.*, 2003)
16,000	Macro-nuclear chromosome (Doak *et al.*, 2003)
40,000	Typical fosmid clone (Shimada *et al.*, 2005)
150,000	Typical BAC clone (Int. Human Genome Seq. Consortium, 2001)
910,725	Linear bacterial genome of *B. burgdorferi* (Fraser *et al.*, 1997)

Data and theory are plotted along the x-axis and y-axis, respectively, in Fig. 4. Perfect agreement would follow the line $y = x$ shown in the figure. However, we would not actually expect to realize any instances of such conformity using a small 12 member sample, even for a theory having absolute predictive power. Specifically, individual coverages are plotted on the x-axis, but the y-axis shows expected values. Consequently, there is an inherent "scatter" in x values that is not present in y values. Nevertheless, it seems clear that our model makes systematically better predictions (closer to $y = x$) in most cases. It is likely that much of the over-prediction error can be attributed to cloning bias in the MF process. Bias is a well-known deficiency of cloned libraries (Osoegawa *et al.*, 2001), but one which is essentially impossible to model without *a priori* characterization. The scatter effect noted above probably also factors into the error for the two left-most points because they both have sequence redundancies close to unity. This is in the neighborhood of maximum dispersion (see Fig. 7 below). Lastly, the alignment process itself is difficult (Springer *et al.*, 2004), and it is unclear whether any anomalies remain in the data.

In examining the plot more closely, we find notable contradictions at $x = 100\%$ coverage. Classical theory predicts $y > 98\%$ coverage, while current theory suggests just over 90%. These two cases could represent exceptions having no edge effects. Strictly speaking, it is gene islands that exhibit the edge effect, not the genes themselves. Genes can appear singly, or in clusters of two to four (Springer *et al.*, 2004). Any "interior" genes in a cluster would be expected to have significantly improved coverage, as suggested in the plot. Although this is the most likely explanation, it is also possible that the edge effect vanished by virtue of non-exonic flanking sequence that has survived the filtering process. For example, Palmer *et al.* (2003) report that about 7% of maize repeats are under-methylated and would therefore appear in an MF library. In summary, edge effects can be substantial in MF libraries and accounting for them in a

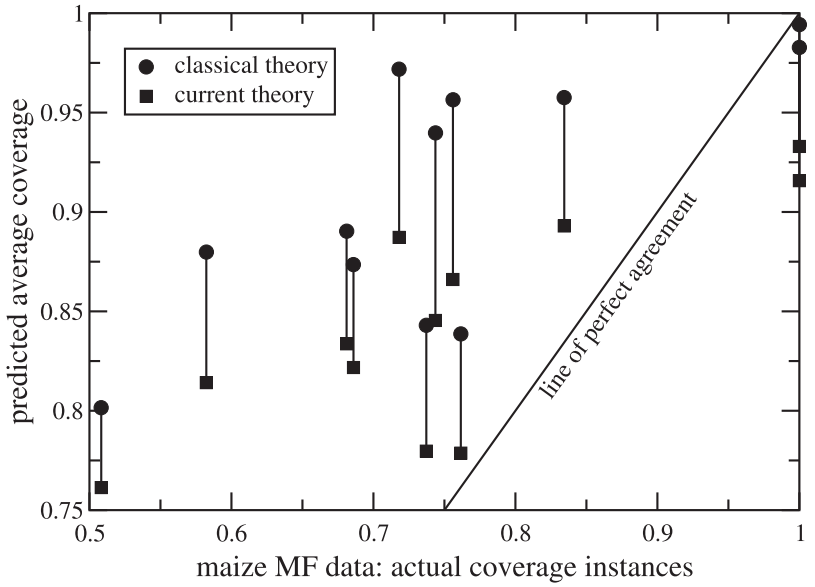

FIG. 4. Comparison of both current theory (Thm. 1) and classical theory (Clarke and Carbon, 1976) to maize methyl-filtered sequence data. Corresponding points from the two theoretical models are connected with vertical segments.

model can lead to significantly improved predictions. However, Thm. 1 is more accurate in the context of analyzing island coverage rather than gene coverage for MF sequencing.

3.2. Assessing the impact of read-pairing

Wendl and Barbazuk (2005) examined edge effects in the shotgun sequencing process. Theorem 1 indicates that edge effects and read pairing effects interact in a non-trivial way. Before studying this more general problem, we would like to make some assessment strictly of the read pairing effect. This phenomenon vanishes for large targets, but it is isolated for finite targets when flanking regions are present. For a flanked target of length π_f, standard theory (Corollary 5) yields

$$\frac{E\langle C_l\rangle}{\pi_f} = 1 - \left[1 - \frac{\lambda}{\pi_f}\right]^{2n}. \tag{2}$$

Conversely, Corollary 3 expands exactly (Mangulis, 1965) for π_f possible placements as

$$\frac{E\langle C_f\rangle}{\gamma} = 1 - \left[1 - \frac{2\lambda}{\pi_f}\right]^{n}. \tag{3}$$

Equations (2) and (3) are asymptotically equal for large n, and in fact, converge rather quickly for parameters characteristic of DNA sequencing ($n \gg 1$). In the worst case, standard theory under-predicts expected coverage for the projects listed in Table 2 by only a few percent (data not shown). The physical interpretation is a slightly enhanced coverage rate based on the fact that paired reads cannot overlap one another. Standard theory implicitly recognizes a probability of roughly $2\lambda/\gamma$ for this event because it ignores read pair placement constraints.

3.3. Dimensional analysis

Investigators are often concerned with the expected coverage for a shotgun sequencing project. The five variables $E\langle C\rangle$, n, λ, η, and γ, lead naturally to four dimensionless parameters (Barenblatt, 1987) that characterize the process (Table 3). The functional relationship for the fraction of the target covered can be written $T = T(D, L, I)$. This expression contains two parameters not found in standard theory: the insert length ratio, I, and the read length ratio, L. More exactly, the former is not considered by standard theory, while the latter is taken as vanishingly small.

Sequence depth, D, sometimes called sequence redundancy, is usually recognized as a relevant parameter, except in the context of monetary costs. Here, there has been a tendency to use the raw number of reads, $2n$, instead. This stems from the fact that chain-termination sequencing has a relatively high per-read cost. Consequently, the number of reads is often perceived to have a strong association with the overall project cost. Newer technologies, e.g., pyro-sequencing (Ronaghi, 2001), have dramatically lowered per-read costs, which shifts focus toward the more natural per-nucleotide basis. This is consistent with the concept of D, which is a dimensionless measure of the number of nucleotides processed. We will frame all results here in terms of D rather than n.

Dimensional analysis allows us to make some general observations for covering processes subject to the combined action of edge and pairing effects. Figure 5 shows how the predicted fractions of target coverage for all projects collapse onto curves of constant L. Values are all calculated at a sequence depth of $D = 3$ ("light shotgun redundancy") and the plot is arranged such that the single datum derived from standard

TABLE 3. DIMENSIONLESS PARAMETERS

Symbol	*Group*	*Meaning*
T	$E\langle C\rangle/\gamma$	Fraction of target covered
D	$2\lambda n/\gamma$	Sequence depth (redundancy)
L	λ/γ	Read length fraction
I	τ/γ	Insert length fraction

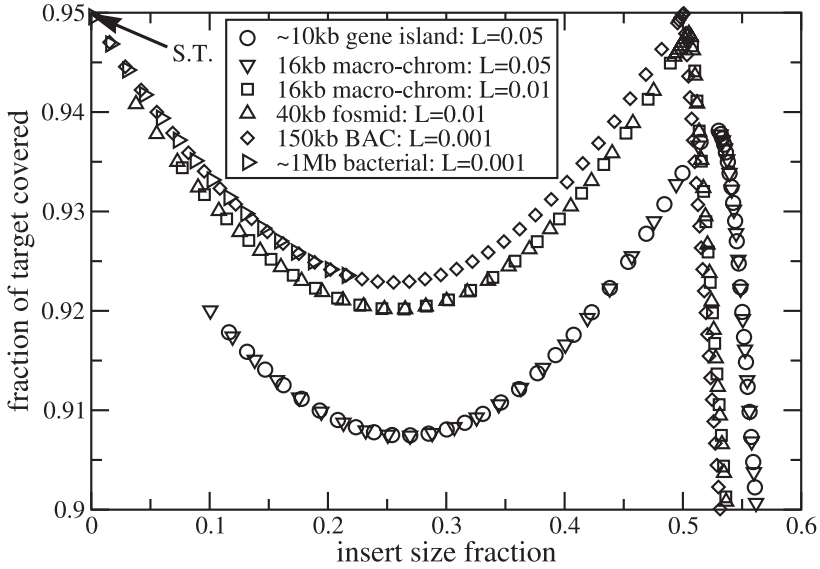

FIG. 5. Fraction of target coverage at threefold sequence redundancy for various read length fractions. The universal prediction from standard theory (S.T.) lies in the upper left corner.

theory ($L \to 0$) resides in the upper-left corner. Data are subject to the constraints on read length and insert length described above.

Before discussing specifics, let us provide some numerical context for this plot. The ordinate spans only a small range of the coverage domain, suggesting that the resulting variation is of minor importance. However, most of the resources for a sequencing project are actually expended in covering small "holdout" parts of the target. For example, standard theory predicts 90% coverage at about $D = 2.3$. A conventional "full shotgun" project ($D = 10$) devotes over three-fourths of its data to resolving the last 10% of the target. This is even more significant if the sequence is to be refined into so-called "base-perfect" form. Here, one approaches the problem from the opposite perspective of vacancy, i.e., how much remains to be finished. Because finishing methods are largely manual and significantly more expensive than random sequencing, differences of a few percent are substantial.

Referring now to Fig. 5, coverage performance clearly degrades for longer read lengths (higher L values). This is largely a consequence of the edge effect, which depends on several factors. First, it is clear from Lemma 1.1 that the fraction of the target affected by decreased edge-related covering probabilities is proportional to read length. Second, shorter read lengths imply a higher number of reads for a given sequence depth. Thus, there are more and better chances to cover problematic terminal regions using shorter reads.

Although coverage behavior for small targets is evidently best with smaller reads, there remains an element of ambiguity that is difficult to model. For example, read length is also important in the downstream assembly phase and longer reads are theoretically better at resolving repeat structures. However, such factors cannot be simulated *a priori* for specific genomes, so it is unclear how much weight should be assigned to such considerations. Recent hardware based on the pyro-sequencing method generates short reads, but is massively parallel (on the order of 10^5 reads per run cycle). From the assembly perspective, the usefulness of such instruments for *de novo* sequencing is currently a matter of debate (Chaisson *et al.*, 2004). However, it appears that coverage characteristics would be good.

Insert size also has an appreciable effect on coverage performance. Trends are similar over the 50-fold range of L shown in the figure. Specifically, there is a local/global maximum where I is lowest, followed consecutively by a local minimum at roughly $0.25 \leq I \leq 0.3$, a local/global maximum at roughly $I \approx 0.5$, and finally a global minimum for $I \to 1$. The maxima are target-specific and depend on the associated constraints for L and I. These considerations lead to different optimal strategies, as follows.

Targets of roughly 10^6 bp and larger show a global maximum at $I \to 0$. The standard model predicts a theoretically ideal coverage of $T = 0.95$ ("S.T." in top left corner) at threefold redundancy and this

milestone is evidently attainable for such targets. Edge effects are vanishingly small, but performance does degrade with increasing insert size. These observations effectively validate the present-day whole genome shotgun strategy, whereby the majority of sequence is generated from short inserts; sequences from large-insert clones are limited to the minimum number required to generate sufficient long-range linking information, typically around 1% of the total (Jaillon *et al.*, 2004). Historically, this strategy has been predicated on the observation that the ends of large-insert clones are substantially more difficult to sequence efficiently than those from small plasmids, but it now appears that plasmids provide optimal coverage performance, as well.

The opposite side of the spectrum offers a rather more surprising result. Targets in the range of 10^4 to 10^5 bp realize only a *local* maximum when using the smallest inserts. Instead, the global maximum lies roughly in the neighborhood $0.5 \leq I \leq 0.54$. In other words, the best strategy is to use short reads derived from inserts that are about half the target length. Stated more generally, coverage would be dramatically improved if insert length could be tailored to the target length for double-stranded templates.

This concept appears to be novel. Investigators usually use a fixed insert length, often in the neighborhood of 2 kb. Clearly, this can result in sub-optimal coverage. The effect becomes more pronounced as L is increased. Consider, for example, a 10-kb gene island sequenced with typical 2-kb inserts and 500-bp reads. The parameters are $I = 0.2$ and $L = 0.05$, meaning that predicted coverage is about 91% at 3-fold redundancy. By simply switching to 100-bp reads and 5-kb inserts ($I = 0.5$, $L = 0.01$), we obtain almost 95% coverage at the same depth. The basic phenomenon persists at higher redundancies, although it is less pronounced (data not shown).

In practical terms, it would be difficult to optimize the process along these lines for every individual target in a group because of inherent size variation in the lengths of reads, inserts, and targets. According to Fig. 5, coverage drops dramatically above the optimal insert size and some targets would invariably fall into this range. In fact, any set of targets is likely to contain some fraction of very small members that apparently could not be meaningfully covered by forward-reverse reads. Fine-tuning of the parameters is not really feasible because they are quite sensitive in the neighborhood of the optimum. Figure 6 demonstrates this assertion for the case of a 16-kb macro-chromosome. The optimum (top left corner) is realized roughly at an insert length of 8,000 and a read length of 100. Very slight increases in insert size dramatically reduce coverage. Likewise, increases in read length affect coverage, but the trend depends upon the insert size. In summary, any moderate change in parameters can result in a significant departure from the optimum.

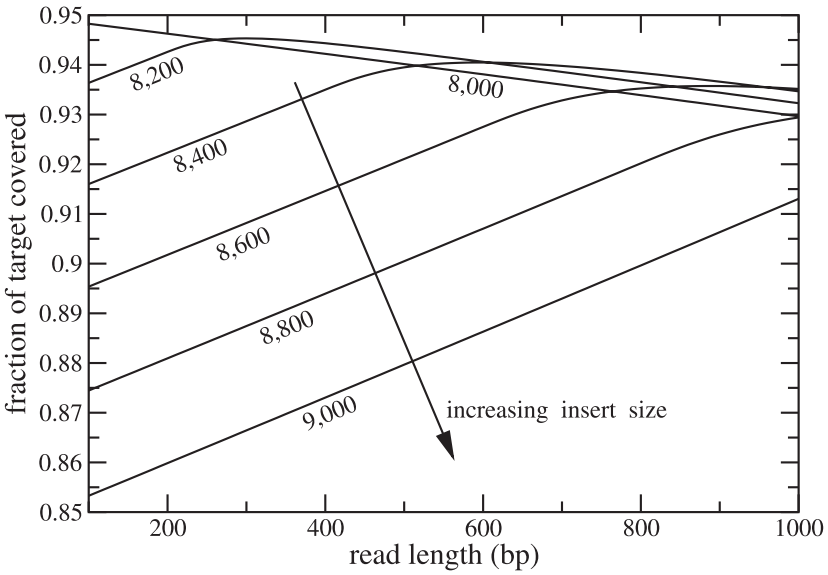

FIG. 6. Coverage response in the parameter neighborhood of the optimum for a 16-kb target at threefold sequence redundancy for a series of insert lengths.

3.4. Variability in the coverage process

Historically, investigators have been interested mainly in the first moment as a prediction and analysis tool for shotgun DNA sequencing. We have also reported the variance in Theorem 2, which provides a means for further analyzing the coverage process. Here, we will actually speak in terms of relative dispersion using the coefficient of variation, $COV = \sigma_c / E\langle C \rangle$.

Figure 7 shows *COV* plotted against sequence depth D for the insert length fractions 0.1, 0.3, and 0.5 and read length fractions 0.01 and 0.05. These parameters correspond roughly to the small-target milestones alluded to in Fig. 5, i.e., a local maximum, local minimum, and global maximum for "short" and "long" reads, respectively. The plot also shows results for *B. burgdorferi* (Table 2), which are discussed further below.

Curves for the three insert length fractions clearly group according to the read length fraction. It seems that variability depends much more on read length than on insert length. This is especially true as L becomes smaller, for example, curves for various I values cluster rather closely for $L = 0.01$. The other primary contributor is clearly the sequence depth. In general, variability reaches a maximum very early in a project ($D \leq 1$), after which it monotonically declines. Wendl and Waterston (2002) reported a similar trend for sequence gaps.

These observations can have a bearing on how deviations from the expected value can be realistically attributed to various factors. For example, the Poisson-based A-statistic (Myers *et al.*, 2000) has been used to infer collapsed regions in genome assemblies. Its premise is that if the statistic exceeds a calibrated, but fixed threshold, then one or more sequence repeats have been found. However, behavior clearly departs from the Poisson model as targets become smaller, suggesting the need for more sophisticated treatments. For example, a rudimentary extension might use the A-statistic, but allow for an adjustable threshold to account for large variances, where appropriate.

Wendl (2006) used the occupancy concept to model the probability density function (PDF) for coverage on large genomes. The so-called "intersection probability" $P_\cap$ was proposed as a way to identify reasonable upper bounds for D in shotgun sequencing projects. This concept cannot be applied directly to small or moderate-sized targets because our present theory does not furnish the density function. However, $P_\cap$ depends on the broadness of the PDF, so we might make some reasonable inferences based on *COV*.

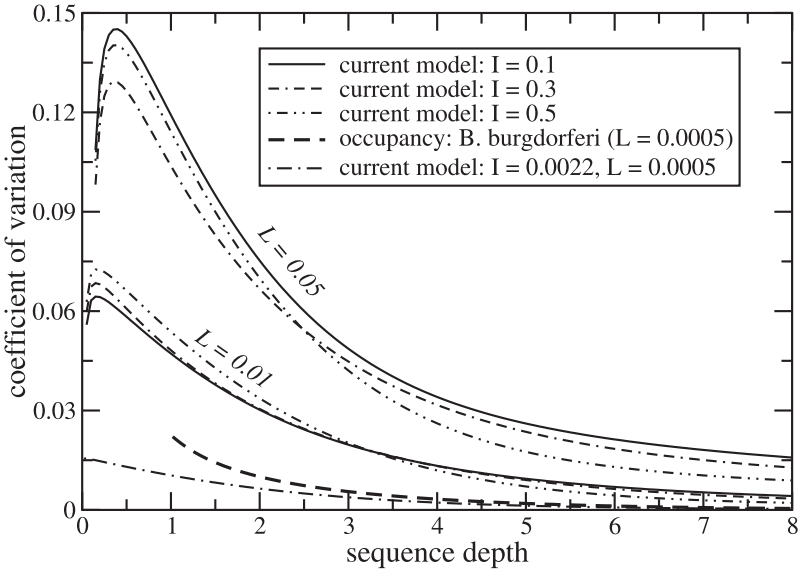

FIG. 7. Coefficient of variation (*COV*) for various insert length fractions. Curves for $L = 0.01$ group together, as do curves for $L = 0.05$. Also shown is the *COV* derived from the occupancy model of coverage (Wendl, 2006) for *B. burgdorferi* (Fraser *et al.*, 1997) and the corresponding *COV* derived from present theory.

Figure 7 shows *COV* derived from both the occupancy model and current theory for the $\gamma = 910{,}725$-bp linear chromosome of *B. burgdorferi* (Fraser *et al.*, 1997). Targets of order 10^6 are roughly the smallest for which the occupancy assumption remains reasonable (Wendl, 2006). The curves rapidly diverge for $D \leq 2$ because of properties inherent to the occupancy process. However, agreement is reasonable for $D > 2$, so that we might think of the *B. burgdorferi* chromosome as a reference point for considering smaller targets.

Basically, as L for a project is increased, the probability of realizing substantial coverage progress, say going from D to $D + 1$, decreases. This is what is estimated by $P_\cap$. Parameters from the *B. burgdorferi* project (Fraser *et al.*, 1997) imply $L \approx 0.0005$. Accordingly, we find that $D \approx 5.6$ would be a reasonable stopping point for random sequencing. This limit suggests that sequencing depth for the smaller targets listed in Table 2 should be substantially lower: almost certainly $D < 5$, but probably $D < 4$ in many cases.

3.5. Single-stranded versus double-stranded sequencing

Conventional shotgun sequencing can utilize either single-stranded or double-stranded templates. The former were used almost exclusively in the early days of sequencing, but the latter are somewhat more prevalent now because of the usefulness of linking information they provide in assembly. However, we have already established that the coverage process for double-stranded sequencing on smaller targets is very sensitive to the parameters, especially the insert length. Suboptimal performance appears to be unavoidable in most instances. We now examine the possibility that single-stranded sequencing might be the better covering alternative for such targets. The presumption is that inserts are sized such that they can be sequenced in their entirety.

Figure 8 shows a comparison of double-stranded and single-stranded coverage evolution, the latter of which is calculated from the model of Wendl and Barbazuk (2005). The theoretically ideal Clarke–Carbon curve is also shown for reference. We have again taken a representative 16-kb macro-chromosome as the example and plotted the fraction of the target covered against the sequence redundancy. The read length ratio is fixed at $L = 0.01$, but double-stranded results are computed for a number of insert length ratios around the optimal value of roughly $I = 0.5$.

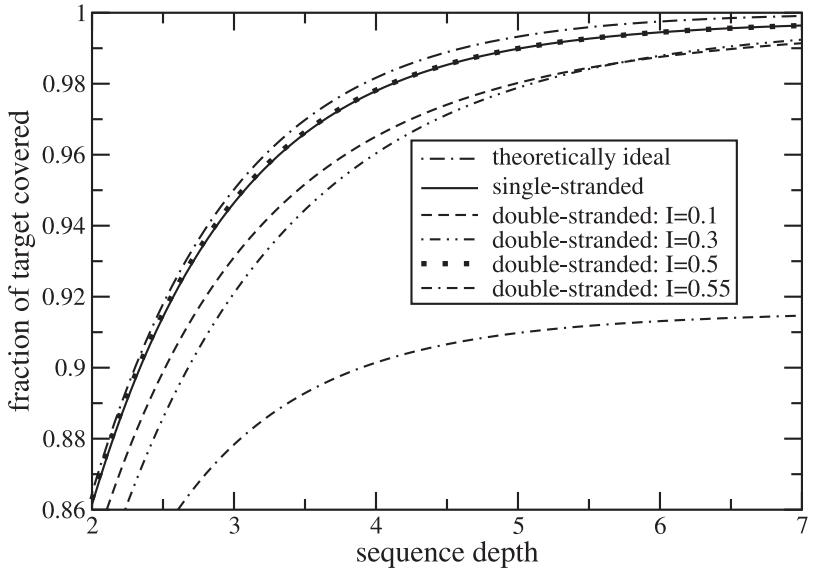

FIG. 8. Coverage as a function of sequence redundancy on a 16-kb target for a read length ratio of 0.01. Double-stranded sequencing results are shown at various insert length ratios, along with single-stranded results and the theoretically ideal Clarke–Carbon curve.

The sensitivity arising from insert length variation is clearly shown for double-stranded templates. Specifically, the curves in Fig. 8 follow trends we described above for varying I, except we see now that sequence depth also influences the process to some degree. For instance the rate of coverage-accrual (slope of the curve) decays significantly with sequence depth for $I = 0.55$ as compared to the smaller inserts. The consequences of exceeding the optimum insert size evidently worsen as more redundancy is generated.

The more notable feature in Fig. 8 is that the curve for single-stranded sequencing essentially coincides with the optimal double-stranded curve ($I = 0.5$). This implies that the best attainable coverage could be realized simply by using single-stranded templates. Of course, both of these curves remain below the theoretically ideal prediction given by standard theory. This is a consequence of edge effects for the finite read length ($L = 0.01$). Optimality of single-stranded sequencing persists over the range of L-values we have examined (data not shown).

These observations suggest that single-stranded sequencing is preferable for smaller target sizes, especially because there may be limited liability in losing the pairing information one would otherwise have obtained with end-sequences. Specifically, assembly difficulties posed by repeat structures generally diminish with target size. In fact, software designed for local assembly of fosmids and BACs, notably the Phrap package (<www.phrap.org>), does not typically use read-pairing information. Although investigators have shown that accuracy and contiguity can sometimes be improved by using pairing constraints (Pevzner *et al.*, 2004), the prevailing practice is still to disregard such information when the target length is the size of a BAC or less. The issue is arguably even less important for filtered DNA islands and macro-chromosomes. Here, target regions are highly correlated with low-copy genic sequence, which should pose a considerably more straightforward assembly problem.

3.6. Concluding Remarks

The observations we have made here suggest three general strategies for small to moderate targets, all of which run somewhat counter to conventional practice for large targets.

1. Shorter reads will tend to give more coverage per unit redundancy than longer reads. Electrophoresis-based sequencing instruments can readily be configured for short reads. Moreover, reads from pyro-sequencing instruments are currently two-thirds less than full-length Sanger reads (Ronaghi, 2001), making these instruments well-suited for small targets.
2. Coverage rates using single-stranded reads are roughly equal to those for double-stranded end-reads when insert length for the latter is set to its optimal value. Because double-stranded performance is highly sensitive within the neighborhood of the optimum, single-stranded reads will tend to give appreciably better results in practice. Pyro-sequencing instruments typically utilize single-stranded PCR-based templates and would be well-suited for small targets from this standpoint. Absence of linking information should not be a significant liability for assembly because of the small size and low-copy nature of the targets.
3. Sequence redundancy should be limited to values of roughly 4-fold to 5-fold. Many conventional whole genome shotgun projects have opted for dramatically higher depths (Wendl, 2006). Moreover, investigators have discussed depths of around 30-fold when short pyro-sequencing reads are utilized (Chaisson *et al.*, 2004). However, the coverage PDFs for small targets at high depth will not be much different from those in the fourfold to fivefold range. In particular, pyro-sequencing instruments are capable of generating enormous amounts of data per run cycle, suggesting the possibility of dividing a cycle between two or more sequencing projects. This would prevent "over-sequencing" and would further leverage the pyro-sequencing technique.

We have assumed that inserts are IID and have not made any sort of attempt to account for biases. The latter are invariably present in a number of forms for random shotgun sequencing. Results obtained for specific projects using this model should therefore be taken as upper bounds of performance. Actual coverages will likely fall somewhat short of predictions, as already illustrated in Fig. 4.

4. METHODS

Proofs for the mathematical results are shown in this section. The target has a left-anchored coordinate system $x \in \{1, 2, 3, \ldots, \gamma\}$ that represents the nucleotide positions (Fig. 2). Coordinates of reads and inserts refer to the starting positions of their respective left-most bases.

Proof of Lemma 1.1. The functions $f(x)$ and $r(x)$ can be demonstrated via elementary counting. For example, for $f(x)$ the left boundary directly constrains forward read placement, so that there are x ways to cover position x for $x < \lambda$ (Fig. 2). On the right end, forward read constraints are a function of how reverse reads impinge on the right boundary of the target. The right-most possible position that can be reached by a forward read is $\mu_f = \gamma - (\lambda + \eta)$. Symmetry dictates the same behavior for the right constraint, so that the number of placements is $\mu_f - x + 1$ for $x \leq \mu_f$. For the unconstrained region $\lambda \leq x \leq \pi$, any part of a read can cover x, giving λ possible placements. For $x > \mu_f$, there are no placements that will cover x.

Derivation of $r(x)$ follows along similar lines. The right boundary constrains reverse reads, giving $\gamma - x + 1$ ways to cover x when x is within λ of the boundary. The left-most position that can be reached by a reverse read is $\lambda + \eta + 1$. Positions to the left of this cannot be covered. However, within a read length moving rightward, x can be covered in $x - (\lambda + \eta)$ ways. As with $f(x)$, any part of a reverse read can cover x in the neighborhood between the constrained regions, giving λ possible placements. ■

Proof of Lemma 1.2. Functions $f(x)$ and $r(x)$ are predicated on the assumption that the left-constrained and right-constrained neighborhoods do not overlap. Specifically, for forward reads, the regions $x < \lambda$ and $x > \pi$ do not overlap. Similarly, for reverse reads, the regions $x < \tau$ and $x > \gamma - \lambda + 1$ do not overlap. Solving either of these yields the result. ■

Proof of Theorem 1. This can be demonstrated by way of the vacancy V, which is the complement of coverage, i.e., $V + C = \gamma$ (Siegel, 1978; Kolchin *et al.*, 1978). Let the random variable θ_x indicate the coverage status of position x after processing n inserts as

$$\theta_x = \begin{cases} 0 & : \quad x \text{ is covered} \\ 1 & : \quad x \text{ is not covered.} \end{cases}$$

Taking each position as a Bernoulli trial implies $E\langle\theta_x\rangle = P(\theta_x = 1)$. Then $E\langle V\rangle = \Sigma E\langle\theta_x\rangle = \Sigma P(\theta_x = 1)$.

Define $F_{i,x}$ as the event whereby the forward read of the ith insert does not cover position x. The counterpart event for the reverse read is $R_{i,x}$. The probability that x remains uncovered after n read pairs is then

$$P(\theta_x = 1) = P\left(F_{1,x} \cap R_{1,x} \cap F_{2,x} \cap R_{2,x} \cap \cdots \cap F_{n,x} \cap R_{n,x}\right).$$

Because inserts are independent of one another, this expression simplifies to $\left[P\left(F_{i,x} \cap R_{i,x}\right)\right]^n$. Without loss of generality, assume the forward read of any insert is always generated first, so that $P\left(F_{i,x} \cap R_{i,x}\right) = P\left(F_{i,x}\right) \cdot P\left(R_{i,x}|F_{i,x}\right)$. The forward-read probability is readily found as

$$P\left(F_{i,x}\right) = 1 - \frac{f(x)}{\pi}.$$

The reverse-read probability is similarly deduced, except that conditioning upon $F_{i,x}$ means that there are $f(x)$ fewer possible placements for the reverse read. Consequently

$$P\left(R_{i,x}|F_{i,x}\right) = 1 - \frac{r(x)}{\pi - f(x)}.$$

Theorem 1 then follows directly from the observation that $E\langle V\rangle + E\langle C\rangle = \gamma$. ■

Proof of Theorem 2. The second "falling factorial" of vacancy is $V^{[2]} = V\,(V-1)$, whereby

$$\begin{aligned}\sigma_c^2 &= E\langle V^2 - V\rangle + E\langle V\rangle - E^2\langle V\rangle \\ &= E\langle V^2\rangle - E^2\langle V\rangle \\ &= E\langle C^2\rangle - E^2\langle C\rangle.\end{aligned}$$

The last statement is an identity (Ross, 2000) and is obtained by substituting $V = \gamma - C$ in the preceding statement and simplifying. ■

Proof of Lemma 2.1. This argument largely follows the methodology of Kolchin *et al.* (1978). The second "falling factorial" of vacancy $V^{[2]} = V\,(V-1)$ implies

$$\begin{aligned}V^{[2]} &= \left(\theta_1 + \theta_2 + \cdots + \theta_\gamma\right)\left(\theta_1 + \theta_2 + \cdots + \theta_\gamma - 1\right) \\ &= \sum_{x=1}^{\gamma}\sum_{y=1}^{\gamma}\theta_x\theta_y\left[1 - \delta_{xy}\right].\end{aligned}$$

The latter statement is a consequence of the fact that $\theta_x^2 = \theta_x$. It is readily shown that $E\langle\theta_x\theta_y\rangle = P(\theta_x = \theta_y = 1)$, whereby

$$E\langle V^{[2]}\rangle = \sum_{x=1}^{\gamma}\sum_{y=1}^{\gamma}P(\theta_x = \theta_y = 1)\left[1 - \delta_{xy}\right].$$

Let $F_{i,x,y}$ be the event indicating that the forward read of the ith insert covers neither position x nor position y on the target. The counterpart event for the reverse read is $R_{i,x,y}$. Applying arguments similar to those in Theorem 1, we find $P(\theta_x = \theta_y = 1) = \left[P\left(F_{i,x,y}\right)\cdot P\left(R_{i,x,y}|F_{i,x,y}\right)\right]^n$. We can generically represent the probability that a read from the ith insert covers neither position x nor position y according to the identity

$$P_i(\theta_x = \theta_y = 1) = 1 - \left[P_i(\theta_x = 0) + P_i(\theta_y = 0) - P_i(\theta_x = \theta_y = 0)\right].$$

For forward reads, this expression takes the form

$$P\left(F_{i,x,y}\right) = 1 - \frac{f(x) + f(y) - \phi(x,y)}{\pi},$$

where $\phi(x,y)$ is the number of ways in which the forward read can cover both x and y. The reverse read is, once again, conditioned upon the forward read, so that by similar arguments made in the proof of Theorem 1, we find

$$P\left(R_{i,x,y}|F_{i,x,y}\right) = 1 - \frac{r(x) + r(y) - \rho(x,y) - \xi(x,y)}{\pi - [f(x) + f(y) - \phi(x,y)]},$$

where $\rho(x,y)$ is the number of ways in which the reverse read can cover both x and y. The extra term $\xi(x,y)$ accounts for the instances in which one read of the ith insert covers position x, while the other covers position y. Because these placements violate the conditioning upon $F_{i,x,y}$, they must be excluded. We note that $\rho(x,y)$ and $\xi(x,y)$ are mutually exclusive, since paired reads derived from the same insert cannot overlap one another. ■

Proof of Lemma 2.2. This proof is similar to the one for Lemma 1.1 in the sense that functions ϕ and ρ can be deduced via elementary counting. The process is conveniently illustrated by representing the

functions as $\gamma \times \gamma$ symmetric matrices. For example, ϕ takes the form

$x\backslash y$	1	2	3	4	$\cdots$							$\cdots$	μ_f-2	μ_f-1	μ_f
1	—	1	1	1	$\cdots$							$\cdots$	0	0	0
2		—	2	2	$\cdots$							$\cdots$	0	0	0
3			—	3	$\cdots$							$\cdots$	0	0	0
4				—	$\cdots$							$\iddots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$					$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\iddots$				
						—	$\lambda-1$	$\lambda-2$	$\lambda-3$	$\lambda-4$	$\cdots$				
							—	$\lambda-1$	$\lambda-2$	$\lambda-3$	$\cdots$				
								—	$\lambda-1$	$\lambda-2$	$\cdots$				
									—	$\lambda-1$	$\cdots$				
										—	$\cdots$				
											$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$												—	3	2	1
μ_f-2													—	2	1
μ_f-1														—	1
μ_f															—

where we show only the upper symmetric part for $y > x$ and have omitted the trivial elements for $\mu_f < x \leq \gamma$ and $\mu_f < y \leq \gamma$.

There is a clear banding pattern for the βth non-trivial band off the diagonal, where $\beta = y - x$. This consists of an interior region on the band, where ϕ is constant, and a boundary region on each end, where ϕ varies. The jth element from either end on a band is clearly $\phi = j$, where $j \in \{1, 2, 3, \ldots, \lambda - \beta - 1\}$. The remaining elements in the interior are all $\phi = \lambda - \beta$.

For the upper-left boundary, $\phi = j = x$, so that solving $x \leq \lambda - \beta - 1$, we find the limit $y \leq \lambda - 1$. For the lower-right boundary, $\phi = j = \mu_f - y + 1$, so that $\mu_f - y + 1 \leq \lambda - \beta - 1$ yields a limit of $x \geq \mu_f - \lambda + 2$. Elements vanish for bands $\beta \geq \lambda$, implying $\phi = 0$ for $y - x \geq \lambda$, or $\phi = \lambda - (y - x)$ otherwise. Finally, there are no ways to cover both positions if y lies out of bounds to the right, i.e., $y > \mu_f$.

The matrix for ρ is similar, although the representation shown above is displaced by $\lambda + \eta$ on both the rows and the columns. Thus, for the upper-left boundary, $\phi + \lambda + \eta = j = x$, so that solving $x - \lambda - \eta \leq \lambda - \beta - 1$ indicates the limit $y \leq \tau - 1$. For the lower-right boundary, $\phi = j = \gamma - y + 1$, so that $\gamma - y + 1 \leq \lambda - \beta - 1$ yields a limit of $x \geq \gamma - \lambda + 2$. Again, elements vanish for bands $\beta \geq \lambda$, implying $\phi = 0$ for $y - x \geq \lambda$, or $\phi = \lambda - (y - x)$ otherwise. Finally, there are no ways to cover both positions if x lies out of bounds to the left, i.e., $x < \mu_r$. ■

Proof of Lemma 2.3. Function $\xi(x, y)$ quantifies the number of ways one read in a pair can cover position x, while the other covers position y. Clearly, ξ vanishes if x and y are either farther apart than the length of an insert, or if they are closer together than the intervening distance between the reads, i.e., $y - x \geq \tau$ or $y - x \leq \eta$. The function also vanishes for regions unreachable by either the forward or the reverse read, i.e., $y < \mu_r$ or $x > \mu_f$.

Non-trivial values can be determined via elementary counting. For example, in regions far from the left or right boundaries of the target, the number of placements will be maximum if x and y are "in phase" relative to the two paired reads. That is, $\xi(x, y) = \lambda$ when $y - x = \lambda + \eta$. The immediate "out of phase" corollary is $\xi(x, y) = \lambda - j$ when $y - x = \lambda + \eta \pm j$, where $0 < j \leq \lambda - 1$.

Values in the boundary regions are similarly determined. For example, for the extreme left-placement of an insert, i.e., at $x = 1$, the reverse read covers positions μ_r through $\tau - 1$. Consequently, $\xi(x, y) = 1$ for $x = 1$ and $y \in \{\mu_r, \ldots, \tau - 1\}$. Likewise, for $x = 2$ this is the only placement that will also cover $y = \mu_r$. However, taken with the insert that begins at $x = 2$, there are now 2 ways to cover $x = 2$ and $y \in \{\mu_r + 1, \ldots, \tau - 1\}$. Following this procedure, we can enumerate ξ in the neighborhood of the left-end boundary, i.e., for $x < \lambda$ and $y < \tau$. A similar approach is used for the right-end boundary region defined as $x > \gamma - \tau + 1$ and $y > \gamma - \lambda + 1$.

Like ϕ and ρ, function ξ is conveniently cast as a banded $\gamma \times \gamma$ symmetric matrix of the enumerated values. Showing only the non-trivial entries, the matrix takes the form

$x\backslash y$	μ_r		$\cdots$		τ						$\gamma-\lambda+1$		$\cdots$		γ
1	1	1	1	$\cdots$	1										
	1	2	2	$\cdots$	2	$\ddots$									
$\vdots$	1	2	3	$\cdots$	3	$\ddots$									
	$\vdots$	$\vdots$	$\vdots$	$\ddots$		$\ddots$									
λ	1	2	3								1				
		$\ddots$	$\ddots$	$\ddots$		$\ddots$	$\ddots$	$\ddots$		$\iddots$					
						$\cdots$	λ	$\lambda-1$	$\lambda-2$	$\cdots$					
						$\cdots$	$\lambda-1$	λ	$\lambda-1$	$\cdots$					
						$\cdots$	$\lambda-2$	$\lambda-1$	λ	$\cdots$					
						$\iddots$		$\ddots$	$\ddots$	$\ddots$		$\ddots$	$\ddots$	$\ddots$	
$\gamma-\tau+1$					1								3	2	1
										$\ddots$		$\ddots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$										$\ddots$	3	$\cdots$	3	2	1
										$\ddots$	2	$\cdots$	2	2	1
μ_f											1	$\cdots$	1	1	1

The banded structure clearly shows a main diagonal, "superior off-diagonals" (upper-right relative to the main diagonal), and "inferior off-diagonals" (to the lower-left).

Extending the observation above, we can identify the coordinate of a band according to the index $j = y - x - \lambda - \eta$, where $j = 0$ represents the main diagonal, $j = 1$ represents the adjacent superior off-diagonal, etc. Consequently, $0 \leq j \leq \lambda - 1$ identifies the main diagonal and all the superior off-diagonals, while $1 \leq -j \leq \lambda - 1$ are the inferior off-diagonals. We have named these classifications **A** and $\bar{\mathbf{A}}$, respectively, in the statement of the lemma.

Specifying the non-trivial values of ξ is now straightforward. Within the left-boundary region, $\xi = x$ for **A**, otherwise $\xi = y - \mu_r + 1$. Within the right boundary region, $\xi = \gamma - y + 1$ for **A**, otherwise $\xi = \mu_f - x + 1$. We already observed that $\xi = \lambda - j$ in the interior region. Substituting parameters, we find $\xi = \tau - y + x$ for **A** and $\xi = y - x - \eta$ for $\bar{\mathbf{A}}$. ■

ACKNOWLEDGMENTS

I am grateful to J. Wallis of the Washington University Genome Sequencing Center and W. B. Barbazuk of the Donald Danforth Plant Science Center for discussions of the covering problem and especially to the latter for providing methyl-filtered maize data for comparison. This work was supported by a grant from the National Human Genome Research Institute (HG003079).

REFERENCES

Barenblatt, G.I. 1987. *Dimensional Analysis*, Gordon and Breach, New York.

Bedell, J.A., Budiman, M.A., Nunberg, A., *et al.* 2005. Sorghum genome sequencing by methylation filtration. *PLOS Biol.* 3, 103–115.

Chaisson, M., Pevzner, P., and Tang, H. 2004. Fragment assembly with short reads. *Bioinformatics* 20, 2067–2074.

Clarke, L., and Carbon, J. 1976. A colony bank containing synthetic Col El hybrid plasmids representative of the entire *E. coli* genome. *Cell* 9, 91–99.

Doak, T.G., Cavalcanti, A.R.O., Stover, N.A., *et al.* 2003. Sequencing the *Oxytricha trifallax* macronuclear genome: A pilot project. *Trends Genet.* 19, 603–607.

Fraser, C.M., Casjens, S., Huang, W.M., *et al.* 1997. Genomic sequence of a lyme disease spirochaete *Borrelia burgdorferi*. *Nature* 390, 580–586.

Int. Human Genome Seq. Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.

Jaillon, O., Aury, J.M., Brunet, F., *et al.* 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431, 946–957.

Kolchin, V.F., Sevastyanov, B.A., and Christyakov, V.P. 1978. *Random Allocations*, Wiley, New York.

Lander, E.S., and Waterman, M.S. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* 2, 231–239.

Mangulis, V. 1965. *Handbook of Series*, Academic Press, New York.

Mouse Genome Seq. Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562.

Myers, E.W., Sutton, G.G., Delcher, A.L., *et al.* 2000. A whole-genome assembly of *Drosophila*. *Science* 287, 2196–2204.

Osoegawa, K., Mammoser, A.G., Wu, C., *et al.* 2001. A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res.* 11, 483–496.

Palmer, L.E., Rabinowicz, P.D., O'Shaughnessy, A.L., *et al.* 2003. Maize genome sequencing by methylation filtration. *Science* 302, 2115–2117.

Pevzner, P.A., Tang, H., and Tessler, G. 2004. De novo repeat classification and fragment assembly. *Genome Res.* 14, 1786–1796.

Rabinowicz, P.D., McCombie, W.R., and Martienssen, R.A. 2003. Gene enrichment in plant genomic shotgun libraries. *Curr. Opin. Plant Biol.* 6, 150–156.

Roach, J.C. 1995. Random subcloning. *Genome Res.* 5, 464–473.

Robbins, H.E. 1944. On the measure of a random set. *Ann. Math. Statist.* 15, 70–74.

Ronaghi, M. 2001. Pyrosequencing sheds light on DNA sequencing. *Genome Res.* 11, 3–11.

Ross, S.M. 2000. *Introduction to Probability Models*, 7th ed., Academic Press, San Diego.

Sanger, F., Air, G.M., Barrell, B.G., *et al.* 1977a. Nucleotide sequence of bacteriophage phiX174 DNA. *Nature* 265, 687–695.

Sanger, F., Nicklen, S., and Coulson, A.R. 1977b. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 74, 5463–5467.

Shimada, M.K., Kim, C.G., Kitano, T., *et al.* 2005. Nucleotide sequence comparison of a chromosome rearrangement on human chromosome 12 and the corresponding ape chromosomes. *Cytogenet. Genome Res.* 108, 83–90.

Siegel, A.F. 1978. Random arcs on the circle. *J. Appl. Probabil.* 15, 774–789.

Springer, N.M., Xu, X.Q., and Barbazuk, W.B. 2004. Utility of different gene enrichment approaches toward identifying and sequencing the maize gene space. *Plant Physiol.* 136, 3023–3033.

Wendl, M.C. 2006. Occupancy modeling of coverage distribution for whole genome shotgun DNA sequencing. *Bull. Math. Biol.* 68, 179–196.

Wendl, M.C., and Barbazuk, W.B. 2005. Extension of Lander-Waterman theory for sequencing filtered DNA libraries. *BMC Bioinform.* 6, article no. 245.

Wendl, M.C., and Waterston, R.H. 2002. Generalized gap model for bacterial artificial chromosome clone fingerprint mapping and shotgun sequencing. *Genome Res.* 12, 1943–1949.

Whitelaw, C.A., Barbazuk, W.B., Pertea, G., *et al.* 2003. Enrichment of gene-coding sequences in maize by genome filtration. *Science* 302, 2118–2120.

Address correspondence to:
Dr. Michael C. Wendl
Genome Sequencing Center
Washington University
4444 Forest Park Blvd.
Campus Box 8501
St. Louis, MO 63108

E-mail: mwendl@wustl.edu