Gap statistics for whole genome shotgun DNA sequencing projects

Michael C. Wendl* and Shiaw-Pyng Yang

Genome Sequencing Center, Washington University School of Medicine, Box 8501, 4444 Forest Park Blvd., Saint Louis, MO 63108 USA

Received on May 19, 2003; revised on December 12, 2003; accepted on January 5, 2004

Advance Access publication February 19, 2004

ABSTRACT

Motivation: Investigators utilize gap estimates for DNA sequencing projects. Standard theories assume sequences are independently and identically distributed, leading to appreciable under-prediction of gaps.

Results: Using a statistical scaling factor and data from 20 representative whole genome shotgun projects, we construct regression equations that relate coverage to a normalized gap measure. Prokaryotic genomes do not correlate to sequence coverage, while eukaryotes show strong correlation if the chaff is ignored. Gaps decrease at an exponential rate of only about one-third of that predicted via theory alone. Case studies suggest that departure from theory can largely be attributed to assembly difficulties for repeat-rich genomes, but bias and coverage anomalies are also important when repeats are sparse. Such factors cannot be readily characterized a priori, suggesting upper limits on the accuracy of gap prediction. We also find that diminishing coverage probability discussed in other studies is a theoretical artifact that does not arise for the typical project.

Contact: mwendl@wustl.edu

1 INTRODUCTION

Biomedical research increasingly relies on the availability of DNA sequences and several procedures for obtaining such sequences have been devised. Because of limitations on the amount of data that can be obtained per sequencing reaction, these methods all take a ‘shotgun’ approach at their most fundamental level (Sanger et al., 1980; Anderson, 1981; Deininger, 1983). The whole genome shotgun (WGS) technique extends this concept such that an entire genome is processed by way of random DNA fragments and then assembled directly (e.g. Myers et al., 2000; Wang et al., 2002; Jaffe et al., 2003; Huang et al., 2003; Mullikin and Ning, 2003). The method has been employed in many capacities, for example, to sequence bacterial genomes (e.g. Tettelin et al., 2001) and to obtain draft information for mammalian genomes (Venter et al., 2001). It is also used as the initial phase of hybrid approaches for large projects (Waterston et al., 2002).

Genome sequencing laboratories have become quite efficient at generating shotgun data via hardware and software automation (Lander et al., 2001). Although improvements are ongoing (Gordon et al., 2001), the rate-limiting step in finishing a project is closing the gaps that remain after shotgun completion. This phase, which consists predominantly of directed (non-random) reads, is largely manual and significantly more expensive per unit sequence than the shotgun component. Investigators, therefore, require estimates of the number of gaps that they will need to resolve. The most frequently employed analytical tool for this purpose is the theory by Lander and Waterman (1988), which was originally developed for clone fingerprinting. Its main limitation lies in the assumption of independent and identical distributions (IID), i.e. that everything is completely ‘random’. DNA does not necessarily satisfy this ideal; for example, violations include repeated sequences, nucleotide bias and cloning bias. Moreover, there are usually assembly difficulties, including good quality reads that cannot be integrated. Because such factors cannot be meaningfully quantified a priori, significant uncertainty may be associated with gap predictions. For example, estimates for the human genome project were particularly controversial (Marshall and Pennisi, 1998; Marshall, 1999). In general, the IID assumption leads to an idealized lower bound for gaps.

Another phenomenon mentioned in the literature is one that can be referred to as diminishing coverage probability. Theoretical studies of random coverage processes have shown that the expected cost of closing all gaps increases significantly as the unit of coverage becomes small compared to the target size (Flatto and Konheim, 1962). In the context of DNA sequencing, this observation implies progressively higher required coverages as the ratio of read length L to project size G decreases (Roach, 1995). Since the amount of information obtained per sequencing reaction is fixed, i.e. $L \approx 600$ base pairs, the result is essentially one of diminishing coverage probabilities for increasingly large genomes. However, the degree to which this factor impacts actual WGS projects remains unknown.

In principle, computer simulation can address these issues, although it raises additional concerns of its own. For instance, deducing expected values requires many

*To whom correspondence should be addressed.
numerically intensive simulations and a method to determine when statistical significance has been achieved. This suggests a certain motivation to extend the analytical theory, especially in light of the increased frequency with which the whole genome method is being applied. Empirical data from 20 representative finished genome projects are introduced into the theory, permitting derivation of correlation equations for gaps. Case studies based on current projects and assembly packages permit some assessment of the number of resulting gaps that can be attributed to computational difficulties in the assembly process. The results should be useful in refining gap estimates for future projects.

2 METHODS

The expected number of gaps in a random DNA shotgun sequencing process is (Wendl and Waterston, 2002)

\[ E = N e^{-\alpha(N-1)}, \]  

(1)

where \( e \approx 2.71828 \) is Euler’s number, \( N \) denotes the number of DNA fragments processed, \( \alpha = (L - T)/G \) is the effective fractional read length (Lander and Waterman, 1988) and \( T \) specifies the average number of bases required to unambiguously detect fragment overlap. To enable comparison of projects of vastly differing sizes in a normalized fashion, the problem of enumerating gaps is scaled according to the maximum of the expected number of gaps. In order to determine the value of \( N \) at which the maximum is expected to occur, Equation (1) can be differentiated with respect to \( N \) and set equal to zero to obtain

\[ \frac{dE}{dN} = 0 = -\alpha N e^{-\alpha(N-1)} + \alpha e^{-\alpha(N-1)}, \]  

(2)

Substituting this value, the maximum is found to be

\[ E_{\text{max}} = \frac{\alpha e^{-\frac{\alpha}{\alpha}}}{\alpha}. \]  

Equation (2) is a reasonable statistical measure of actual values of \( E_{\text{max}} \) because the coefficient of variation, \( C_v \), is small for WGS projects. For example, using results from Wendl and Waterston (2002), it can be shown that \( C_v \approx 1.31 \sqrt{\alpha} \) at \( N = \alpha^{-1} \). Thus, data scatter decreases roughly as the square root of the project size \( L/G \). Let us also define the point at which the shotgun phase is deemed complete as \( i = \sigma E_{\text{max}} \), where \( 0 \leq \sigma \leq 1 \) is a dimensionless gap-closing stringency factor and \( i \) represents the number of remaining gaps requiring manual

Table 1. Representative shotgun sequencing projects ordered by relative project size

<table>
<thead>
<tr>
<th>Project</th>
<th>( L/G )</th>
<th>( \delta )</th>
<th>( E_{\text{max}} )</th>
<th>( i_p )</th>
<th>GC</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse</td>
<td>2.28 × 10^{-7}</td>
<td>6.5</td>
<td>1 724 127</td>
<td>224 713</td>
<td>0.42</td>
<td>Waterston et al. (2002b)</td>
</tr>
<tr>
<td>Rice (indica)</td>
<td>1.09 × 10^{-6}</td>
<td>3.4</td>
<td>337 465</td>
<td>127 550</td>
<td>0.43</td>
<td>Yu et al. (2002)</td>
</tr>
<tr>
<td>Rice (japonica)</td>
<td>1.19 × 10^{-6}</td>
<td>6.4</td>
<td>309 019</td>
<td>42 109</td>
<td>0.44</td>
<td>Goff et al. (2002)</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>1.57 × 10^{-6}</td>
<td>5.25</td>
<td>250 515</td>
<td>77 645</td>
<td>0.44</td>
<td>Aparicio et al. (2002)</td>
</tr>
<tr>
<td>A. gambiae</td>
<td>2.12 × 10^{-6}</td>
<td>8.16</td>
<td>175 321</td>
<td>18 962</td>
<td>0.35</td>
<td>Holt et al. (2002)</td>
</tr>
<tr>
<td>C. intestinalis</td>
<td>4 × 10^{-6}</td>
<td>6.32</td>
<td>97 775</td>
<td>7623</td>
<td>0.35</td>
<td>Dehal et al. (2002)</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>4.6 × 10^{-6}</td>
<td>11.03</td>
<td>85 887</td>
<td>2483</td>
<td>0.42</td>
<td>Adams et al. (2000)</td>
</tr>
<tr>
<td>N. crassa</td>
<td>1.42 × 10^{-5}</td>
<td>17.6</td>
<td>27 870</td>
<td>958</td>
<td>0.50</td>
<td>Galagan et al. (2003)</td>
</tr>
<tr>
<td>P. yoelii</td>
<td>2.89 × 10^{-5}</td>
<td>6.41</td>
<td>13 560</td>
<td>5687</td>
<td>0.23</td>
<td>Carlton et al. (2002)</td>
</tr>
<tr>
<td>L. interspecifics</td>
<td>7.6 × 10^{-5}</td>
<td>8.5</td>
<td>5263</td>
<td>805</td>
<td>0.36</td>
<td>Ren et al. (2003)</td>
</tr>
<tr>
<td>B. melitensis</td>
<td>1.23 × 10^{-4}</td>
<td>8.88</td>
<td>3224</td>
<td>157</td>
<td>0.57</td>
<td>DelVecchio et al. (2002)</td>
</tr>
<tr>
<td>B. halodurans</td>
<td>1.25 × 10^{-4}</td>
<td>7.1</td>
<td>3124</td>
<td>656</td>
<td>0.44</td>
<td>Takami et al. (2000)</td>
</tr>
<tr>
<td>T. tengcongensis</td>
<td>1.3 × 10^{-4}</td>
<td>9.87</td>
<td>3098</td>
<td>273</td>
<td>0.38</td>
<td>Bao et al. (2002)</td>
</tr>
<tr>
<td>C. acetobutylicum</td>
<td>1.61 × 10^{-4}</td>
<td>8.3</td>
<td>2402</td>
<td>551</td>
<td>0.31</td>
<td>Nölling et al. (2001)</td>
</tr>
<tr>
<td>A. fulgidus</td>
<td>2.3 × 10^{-4}</td>
<td>6.66</td>
<td>1706</td>
<td>157</td>
<td>0.49</td>
<td>Klenk et al. (1997)</td>
</tr>
<tr>
<td>S. pneumoniae</td>
<td>2.4 × 10^{-4}</td>
<td>10.62</td>
<td>1629</td>
<td>390</td>
<td>0.40</td>
<td>Tettelin et al. (2001)</td>
</tr>
<tr>
<td>H. influenzae</td>
<td>2.6 × 10^{-4}</td>
<td>6.31</td>
<td>1513</td>
<td>140</td>
<td>0.38</td>
<td>Fleischmann et al. (1995)</td>
</tr>
<tr>
<td>F. nucleatum</td>
<td>2.78 × 10^{-4}</td>
<td>5</td>
<td>1394</td>
<td>387</td>
<td>0.27</td>
<td>Kapralov et al. (2002)</td>
</tr>
<tr>
<td>B. burgdorferi</td>
<td>3.5 × 10^{-4}</td>
<td>6.67</td>
<td>1118</td>
<td>524</td>
<td>0.29</td>
<td>Fraser et al. (1997)</td>
</tr>
<tr>
<td>T. pallidum</td>
<td>4.61 × 10^{-4}</td>
<td>6.67</td>
<td>846</td>
<td>130</td>
<td>0.33</td>
<td>Fraser et al. (1998)</td>
</tr>
</tbody>
</table>

*We did not include the landmark human WGS effort undertaken by Venter et al. (2001) because the coverage assessment remains in dispute. Specifically, Venter et al. (2001) report a coverage of approximately 6.1 × using a combination of their own data and that of the public sequencing project (Lander et al., 2001). However, Waterston et al. (2002a) argue that this figure under-represents the linking information apparently retained from public data in the assembly. According to their analysis, the effective coverage is closer to 12.6 ×.

Because the discrepancy is so substantial and because the issue remains unresolved (Myers et al., 2001), we omit this datum.

\( E_{\text{max}} \) is computed according to Equation (2), which requires an overlap parameter \( T \), in addition to \( L \) and \( G \). For the Drosophila project, we have used \( T = 38 \) as an approximation for the quoted rule of a minimum 40 bp overlap with 6% allowable mismatch (Myers et al., 2000). We have also assumed this value for the other eukaryote projects and for Plasmodium.

For all remaining cases, we have assumed \( T = 30 \) bp.

Strictly speaking, this is not a pure WGS project, but rather a hybrid one. The figures we quote can be taken as ‘pure WGS’ results, as we disregard additional continuity realized using map-based information.

The amount of ‘chaff’ was not evaluated for this project (Robert A. Holt, personal communication), so an estimate of 20% was made.
Whole genome shotgun gap statistics

![Figure 1](https://academic.oup.com/bioinformatics/article-abstract/20/10/1527/237170)

Fig. 1. Empirical and theoretical results for the spectrum of whole genome shotgun projects. Solid curves denote theoretical predictions of Equation (5) for a 90% confidence level. Points indicate empirical performance for individual projects in Table 1. The shallow dashed curves represent log-linearized least squares regressions of these data [Equations (6) and (7)], while the steep dashed curve represents standard Lander and Waterman (1988) theory for all projects using $T = 30$.

finishing. According to this model, the number of gaps is the product of two quantities, one of which depends upon the performance of random processing on the project and the other depends on the size of the project itself.

Using Equation (2), Lander and Waterman theory can be cast as $\sigma_{lw} = E_{lw}/E_{\text{max}} = \delta e^{1-\delta}$, where $\delta = \alpha N$ is the effective depth of coverage. Empirical data can also be framed with respect to $E_{\text{max}}$ by defining $\sigma_{\text{emp}} = i_e/E_{\text{max}}$, where $i_e$ is the actual number of gaps reported in draft assemblies (Table 1). For example, Adams et al. (2000) and Myers et al. (2000) cited $L = 551$ and $G = 120 \times 10^6$ for the *Drosophila* project and counted 2483 gaps in their shotgun assembly. These parameters yield $E_{\text{max}} = 85887$ gaps and a stringency of $\sigma_{\text{emp}} \approx 0.029$. Thus, roughly 97.1% of the gaps were closed at the end of the shotgun phase.

To examine the issue of diminishing coverage performance, we must resort to the probability density function (PDF) for gaps. Wendl and Waterston (2002) showed that typical projects are characterized by $p(k, N) = C_{N,k} \beta^k (1 - \beta)^{N-k}$, where $k$ represents the number of gaps, $\beta = E/N = e^{-\alpha(N-1)}$ and $C_{N,k}$ is the appropriate binomial coefficient. This expression is difficult to evaluate for whole genome sequencing projects because $N$ and $k$ can both be large. However, since $\beta$ is also small, we can replace the equation with an appropriate Poisson density function and recast it in logarithmic form to eliminate numerical problems. These steps yield

$$p(k, N) = \exp \left( k \ln E - E - \ln k! \right), \quad (3)$$

where $E$ is given by Equation (1). Equation (3) is useful for small $k$, but becomes difficult to evaluate for $k$ greater than order 100 because the factorial term tends to overflow. For larger $k$, we invoke Stirling’s approximation (Kreyszig, 1988), which yields

$$p(k, N) = \exp \left( k \ln E/k + k - E - \ln \sqrt{2\pi k} \right). \quad (4)$$

The parameter of interest is the probability of achieving $i$ or fewer gaps. This is calculated via summation of relevant discrete values as

$$P(k \leq i, N) = \sum_{k=0}^{i} p(k, N), \quad (5)$$

where we use Equation (3) for $k < 100$, otherwise we use Equation (4). To compute stringencies based on PDF, Equation (5) is solved for $N$ subject to a specific value of $i$ and a specific confidence value, e.g. $P = 0.9$. Here, we use the Newton–Raphson algorithm with a preconditioned guess derived from several iterations of the Bisection method (Kreyszig, 1988).

3 RESULTS

Figure 1 shows resulting gap statistics from various calculation methods over the parameter space ($\delta = NL/G, \sigma$) relevant to WGS sequencing. Here, $\delta$ is the nominal depth of coverage. Standard Lander and Waterman (1988) theory
appears as a single exponentially decreasing curve, reflecting the fact that it does not consider, in a normalized sense, the size of a project. The gap distribution function in Equation (5) is shown at a confidence level of $P = 0.9$ for several projects: *Treponema pallidum* at $G = 1.1$ Mb, *Vibrio cholerae* at $G = 4$ Mb and *Drosophila* at $G \approx 120$ Mb. As an idealized basis for comparison, the smallest class of the shotgun project is also shown, i.e. a BAC clone of roughly 0.1 Mb. Kinks in these curves indicate the coverage depth at which there exists a 90% chance that all gaps are closed, assuming IID libraries and perfect assembly algorithms.

Figure 1 also shows empirical data for the 20 representative WGS projects listed in Table 1. Ordinates are computed as the quotient of the number of gaps in an assembly and $E_{\text{max}}$, as given by Equation (2). Here, we take a gap to mean any break in the continuity of finished DNA sequence. In the nomenclature of Adams et al. (2000), we are counting both sequence breaks within scaffolds and physical separations between scaffolds as ‘gaps’. Abscissae are computed according to the number of reads in an assembly, not the total number of good quality reads that were processed. The difference, known by the common euphemism ‘chaff’, can be appreciable for eukaryotic projects, largely because of repeat-related assembly difficulties, etc. (Table 2). Two regression curves were modeled for these data via log-linearized least squares analysis. The first considers all the eukaryotic projects in Table 1, except for *Neurospora crassa* (Galagan et al., 2003), for which sequence coverage was uncharacteristically deep. This datum must be excluded as a statistical outlier. Specifically, WGS performance is not understood in the high-depth domain. Without further empirical corroboration, it is impossible to determine whether *N.crassa* taken alone is representative of the WGS process at such depths.

These projects, denoted by the shaded data points in Figure 1, lead to the least squares fit

$$\sigma_{\text{emp}} \approx 1.187e^{-0.334d},$$

with a Pearson’s correlation coefficient of $|\rho| \approx 0.924$. Using the standard two-tailed test for small samples (Student’s ‘$t$’ distribution), we find the correlation to be statistically significant to a level better than 0.003. In addition to the projects considered in Equation (6), the second regression accounts for the prokaryotic projects in Table 1 as well. The regression equation is

$$\sigma_{\text{emp}} \approx 0.701e^{-0.211d},$$

with a Pearson coefficient of $|\rho| \approx 0.549$ and a significance level of about 0.015.

### 4 DISCUSSION

Figure 1 shows that performance does in fact degrade with increasing project size as has been suggested by previous work (Roach, 1995). Specifically, the coverage required for a given confidence of closure increases with project size. In the ideal BAC case, the stringency just before closure is $\sigma \approx 0.014$, which can be taken as the optimal coverage configuration. That is, projects retaining gaps at stringencies less than 0.014 show degraded covering performance. In most instances, actual WGS projects will not be sequenced to depths great enough to realize this phenomenon. In particular, Equation (6) suggests that on the order of $13 \times$ coverage would be required for eukaryotic projects, not including chaff, while Equation (7) suggests an even higher figure. These estimates significantly exceed what investigators commonly consider to be a full, or complete shotgun phase. Diminishing coverage probability is therefore largely a theoretical artifact that does not arise for typical WGS projects. The only project attaining such depth is that for *N.crassa* (Galagan et al., 2003) at $17.6 \times$ coverage; however, this is considered to be a statistical outlier, as mentioned above. It is clear from Figure 1 that this datum does not follow exponential trends predicted by standard Lander and Waterman theory, Equation (5), etc. This further suggests the existence of a state of diminishing returns for projects that are carried significantly in excess of standard $10 \times$ coverage.

Figure 1 shows the well-known observation that Lander and Waterman theory under-predicts the actual number of gaps realized in WGS projects. This is a function not only of the IID assumption discussed above, but may also be due to additional considerations including assembly problems. The interaction of all these factors is project-specific, suggesting limitations on the accuracy of a priori predictions that one might ultimately expect. Indeed, some cases yield unanticipated results. For example, there are instances where a project known to be difficult to assemble exhibits fewer (normalized) gaps than ‘easier’ bacterial projects for at least the same coverage. In particular, *Anopheles gambiae* (Holt et al., 2002) shows appreciably better stringency than a number of bacterial projects, e.g. *Streptococcus pneumoniae* (Tettelin et al., 2001), even though its assembly was complicated by a high rate of polymorphic sequence. Moreover, bacterial projects have the highest number of per capita unresolved gaps, for instance *Borrelia burgdorferi* (Fraser et al., 1997) and
Whole genome shotgun gap statistics

Fig. 2. Gap-closing stringency versus overall GC content. Symbols correspond to those in Figure 1. Dashed line represents regression for prokaryotic projects (the 12 non-shaded points).

*Plasmodium* (Carlton et al., 2002). These projects are therefore more costly to finish per unit sequence than projects that are perceived to be more difficult to assemble.

Equations (6) and (7) are based on least squares analysis of WGS projects that are representative of both project size (spanning three orders of magnitude) and sequencing and assembly difficulties. The Pearson coefficient of \( |\rho| \approx 0.924 \) for Equation (6) indicates strong correlation between coverage and the resulting number of gaps for eukaryotic projects. The square of this number gives the fraction of variance shared by these two quantities. That is, \( \rho^2 \approx 0.85 \) indicates 85% of the variation in \( \sigma_{\text{emp}} \) can be attributed to variation in \( \delta \) for these particular data. Statisticians are careful to point out that \( \rho^2 \) does not imply causation. However, in this case it is well known that a cause and effect relationship between coverage and gaps does actually exist—increased coverage causes a reduction in gaps. One can then attribute about 85% of what determines the value of \( \sigma_{\text{emp}} \) directly to coverage for this sample of projects. The remaining 15% is collectively due to project-specific issues discussed above that are not considered by standard theory and that cannot, in general, be adequately characterized a priori. This further underscores limits on the accuracy of gap predictions, especially those that rely strictly on the IID idealization. Of course, this analysis is provisional and may change with the integration of additional data.

While the strong correlation for Equation (6) supports the Lander and Waterman prediction that gaps do in fact decay exponentially, it appears that for eukaryotic genomes the rate is only about one-third of that predicted by theory. This equation is proposed as a more appropriate tool for analyzing eukaryotic WGS projects in the range of approximately \( 4 \times < \delta < 10 \times \), with the proviso that there still remain significant factors that this model cannot account for. For example, the correlation for Equation (6) depends on the fact that ‘chaff’ reads have been excluded. The amount of chaff realized for a project varies from about 10 to over 25% (Table 2) and cannot be predicted a priori. Thus, there is an additional implicit factor of uncertainty relating the number of reads processed to the number of resulting gaps.

In examining Figure 1 more closely, we see that projects for prokaryotic genomes tend to be further displaced from the regression curve given by Equation (7) than eukaryotic projects. In other words, the prokaryotic data appear more scattered. These data (12 of the 20 projects in Table 1) do not yield a statistically significant correlation (\( P \approx 0.31 \)), a surprising result since such projects are presumed to be more straightforward than sequencing eukaryotic genomes. Some variation may be due to cloning and library bias. For example, prokaryote genomes are often poor in G–C sequence, which degrades clone stability (Carlton et al., 2002). Figure 2 shows a plot of \( \sigma_{\text{emp}} \) versus the overall GC content \( \gamma \) for each project in Table 1. Prokaryotic projects vary over a wide range of GC content, \( 0.23 \leq \gamma \leq 0.57 \), and show reasonably strong correlation (\( \sigma_{\text{emp}} \approx 1.129e^{-4.93\gamma}, |\rho| \approx 0.764, \) and a significance level better than 0.005). Of course, this result alone is not useful for predicting gaps. Conversely, eukaryotic projects cluster around a narrower range of GC content,
0.35 ≤ γ ≤ 0.50, and show no statistically significant correlation to γ (P ≈ 0.83). These projects appear to be less sensitive to GC content bias. Equation (7) will probably be a better indicator of gaps for prokaryote genome projects than pure Lander and Waterman theory, but not as relevant as Equation (6) for eukaryote genomes if chaff is neglected. In practical terms, the accuracy limitations inherent in Equation (7) are not too significant because prokaryote WGS sequencing is well established and the absolute number of gaps is relatively small as compared to eukaryote genome projects (Table 1).

Chaff statistics in Table 2 imply that difficulties in the assembly process have a strong influence on the resulting number of gaps. Gap counts would increase appreciably and the correlation strength of Equation (6) would be reduced if chaff reads were included in assemblies because many, if not most reads would appear as individual contigs. However, we would like to make a further assessment of how computational difficulties affect gap counts for data comprising the assembly itself. Mis-assemblies are often caused by repeats and manifest themselves as High Read Depth (HRD) regions, where HRD regions can be detected on a statistical basis by evaluating Poisson likelihoods of a given depth under the hypothesis that two or more repeats have not been erroneously collapsed into a single contig (e.g. Myers et al., 2000, the ‘A-statistic’). For example, a given base position has a nominal probability p = L/G of being covered by a single read under IID conditions. In an assembly of N reads, the distribution of coverage would follow a Poisson process having a rate of δ = NL/G. Using the resulting Poisson density function, one can count the number of reads covering each base position in an assembly and draw statistical inferences about the number and location of HRD regions.

Table 3 describes case studies for two ongoing projects at the Washington University Genome Sequencing Center, Histoplasma capsulatum strain G186AR (about 5% projected repeats) and Blastomyces dermatitidis (about 25% projected repeats). The WGS data for these efforts have been assembled using three well-known packages: Pcap (Huang et al., 2003), Arachne (Batzoglou et al., 2002), and Phrap (www.phrap.org). Based on an estimated genome size of G = 28.6 Mb, the Histoplasma project has a coverage of roughly δ = 3.3× from approximately 187 000 good quality reads as determined by Phred quality 20 scores (Ewing et al., 1998). Similarly, Blastomyces has an estimated size of 45 Mb and approximate coverage of δ = 2.7× from about 195 000 good quality reads. In addition to providing number of gaps, the table shows the so-called N50 statistic (Lander et al., 2001), which describes the quality of an assembly according to the length of a contig containing a ‘typical’ nucleotide. More exactly, the N50 length LN50 is defined such that half of the nucleotides reside in contigs or scaffolds having a length of at least LN50. Finally, the number of HRD regions and bases comprising these regions are shown. These figures were computed using a threshold depth of 4×. In other words, any neighborhood of at least 20 consecutive bases that is spanned by a number of reads greater than four times the nominal coverage would be extremely unlikely according to the Poisson coverage model. These neighborhoods are interpreted as collapsed regions of two or more repeats.

Mate–pair constraint statistics for contig and scaffold distances suggested the assemblies were of reasonable quality. For example, in the Pcap assembly for Histoplasma 5084 out of 102 384 constraints were not satisfied (about 5%). For Blastomyces, 3118 out of 144 994 constraints were not satisfied (about 2.2%). Results using the two other programs were similar. Because these figures include contributions that are independent of the assembly software, e.g. chimeric and mis-named reads, they represent upper bounds of genuine mis-assemblies.

Several notable observations can be made from Table 3. First, results for Histoplasma indicate gap number is not necessarily a direct indicator of long-range contiguity. Each assembly contains on the order of 6000 gaps, but the N50 contig lengths vary considerably. Gaps are therefore distributed differently among the three assemblies. The number of HRD regions in each assembly is fairly small compared to the number of gaps, as are the total number of HRD bases compared to the genome size. Thus, inability to resolve repeated sequences does not appear to have an overwhelming influence on the resulting number of gaps. This is largely what would be expected for a genome having low repeat density and suggests that scatter in the prokaryotic data are due to various biases and shortfalls in the coverage process in addition to difficulties in assembling the data.

The Blastomyces genome is about five times richer in repeats and shows somewhat different results. While the number of HRD regions and bases are small compared, respectively, to the number of gaps and genome size, the calculated number of gaps is quite different between the Pcap and Arachne algorithms. The Phrap package was unable to complete an assembly. These statistics suggest that sequence repeats coupled with the vagaries of how they are handled in different assembly algorithms can have a significant influence on gaps. The strong correlation

### Table 3. Case studies for assessment of assembly difficulties

<table>
<thead>
<tr>
<th>Project</th>
<th>Assembler</th>
<th>Gaps</th>
<th>LN50</th>
<th>HRD regions</th>
<th>HRD bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histo.</td>
<td>Pcap</td>
<td>6024</td>
<td>8215</td>
<td>294</td>
<td>68108</td>
</tr>
<tr>
<td>Histo.</td>
<td>Arachne</td>
<td>6468</td>
<td>5899</td>
<td>97</td>
<td>29175</td>
</tr>
<tr>
<td>Histo.</td>
<td>Phrap</td>
<td>6012</td>
<td>6966</td>
<td>304</td>
<td>86552</td>
</tr>
<tr>
<td>Blasto.</td>
<td>Pcap</td>
<td>9198</td>
<td>6859</td>
<td>1512</td>
<td>230628</td>
</tr>
<tr>
<td>Blasto.</td>
<td>Arachne</td>
<td>12279</td>
<td>4221</td>
<td>986</td>
<td>134490</td>
</tr>
<tr>
<td>Blasto.</td>
<td>Phrap</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>
shown by Equation (6) is surprising in light of this observation.

Based on what we have presented, it is clear that several factors impose limits on the accuracy of both analytical and computational models of the WGS coverage process. Assembly difficulties and various biases in nucleotide sequence and the cloning process seem to be especially significant. Moreover, variations in L or G can further skew results. For example, in calculating \( \sigma_{\text{emp}} \) for each project, \( E_{\text{max}} \) is based on an estimated \( G \), while \( i_e \) is associated with the actual value of \( G \). Most of these factors cannot be readily or accurately characterized a priori. Further modeling extensions are certainly possible, for example, in considering end-pairing information for specific clone types or anticipated base quality values for overlap calculations. However, these issues are probably subordinate to those we have mentioned, so that any resulting model may offer only limited improvement.

ACKNOWLEDGEMENTS

The authors are grateful to individuals of Washington University School of Medicine for critiques of various versions of the manuscript and/or informative discussions, especially Warren Gish, John Wallis, Gary Stormo, and Robert Waterston. We also thank John McPherson of Baylor College of Medicine and the anonymous reviewers whose suggestions improved the scope of the study. This work was partially supported by grants from the National Human Genome Research Institute (HG002042, principal investigator R.K. Wilson) and the National Institute of Allergy and Infectious Diseases (AI050934, principal investigator E.R. Mardis).

REFERENCES


