Identification of mixups among DNA sequencing plates

Nikola Stojanovic¹,*, Jean L. Chang¹, Jessica Lehoczky¹,², Michael C. Zody¹ and Ken Dewar¹, ³

¹Whitehead Institute, Center for Genome Research, 320 Charles Street, Cambridge, MA 02141, USA, ²Department of Human Genetics, University of Michigan, 1241 East Catherine Street, Ann Arbor, MI 48109, USA and ³Montreal Genome Centre, The Research Institute of the McGill University Health Centre, 1650 Cedar Avenue, Montreal, Quebec, Canada H3G 1A4

Received on January 10, 2002; revised on April 13, 2002; accepted on May 9, 2002

ABSTRACT

Motivation: During the process of high-throughput genome sequencing there are opportunities for mixups of reagents and data associated with particular projects. The sequencing templates or sequence data generated for an assembly may become contaminated with reagents or sequences from another project, resulting in poorer quality and inaccurate assemblies.

Results: We have developed a system to assess sequence assemblies and monitor for laboratory mixups. We describe several methods for testing the consistency of assemblies and resolving mixed ones. We use statistical tests to evaluate the distribution of sequencing reads from different plates into contigs, and a graph-based approach to resolve situations where data has been inappropriately combined. While these methods have been designed for use in a high-throughput DNA sequencing environment processing thousands of clones, they can be applied in any situation where distinct sequencing projects are performed at redundant coverage.

Availability: Our software is available for downloading from: ftp-genome.wi.mit.edu/distribution/mixups_detection.

Contact: nick@genome.wi.mit.edu

INTRODUCTION

The challenge to generate very high quality DNA sequence data has been present since the beginning of the sequencing phase of the Human Genome Project. Three different quality measures have been established, referred to as accuracy, contiguity and fidelity (Felsenfeld et al., 1999). While methods have been established to measure accuracy and contiguity (Felsenfeld et al., 1999; Seluja et al., 1999; Ewing et al., 1998; Ewing and Green, 1998; Bouck et al., 1998), the standards for fidelity have been less articulated. Fidelity has two aspects: (1) the composition, order, and orientation of the derived sequence as compared to the true sequence of the clone, and (2) the clone’s sequence as compared to the true sequence of the corresponding region in the human genome. Preliminary examination has estimated that during the draft phase of the Human Genome Project (publicly funded effort) at least 2.2% of the sequence was foreign to the clone to which it had been assigned, and another 2.13% was estimated to be misassembled (Venter, 2001). Studies of GenBank sequence contamination (Seluja et al., 1999, and references within) appeared to pay little attention to these issues. We shall provide some answers to the first fidelity problem, when it is caused by laboratory errors.

The detection of these errors usually occurs after the production sequencing of clones previously associated with certain known sequences. The identification of problematic projects involves manual checking, by running Blast (Altschul et al., 1990) on assembled sequence data, and on our internal database of markers. In certain cases, we have also found it helpful to check for the presence of organism-specific sequences. However, these techniques involve considerable work by data analysts, and in order for it to be initiated, a mixed-up project needs to be detected first.

The continuing acceleration of human genome sequencing has introduced potential new problems. In order to make more data available earlier, the sequencing has been conducted in phases, working draft followed by deep shotgun, then by complete, finished sequence (International Human Genome Sequencing Consortium, 2001). Working draft consists of consensus sequence with at least 3-fold coverage of bases with a Phred (Ewing et al., 1998) score of 20 or higher (less than 1 error per 100 bases, (Ewing

*To whom correspondence should be addressed.

¹The standard required by GenBank is actually 4×, a little higher than reported by Bentley (2000).
Identification of mixups among DNA sequencing plates

and Green, 1998), and it provides about 95% coverage of each BAC or PAC clone (Bentley, 2000). In the deep shotgun phase, paired end plasmid sequencing is typically performed to a depth of 8–10×, which leads to fully ordered and oriented assembled sequence with few remaining sequence gaps (2.1 per 100 Kb, on average), of which even fewer (3–4 per Mb) are not spanned by plasmid subclones. The deep shotgun product is then the template for directed finishing techniques which lead to the final complete sequence assembly, and final GenBank submission. This two phase production sequencing protocol provides new opportunities for errors, starting with selecting the wrong clone to provide additional sequence coverage. However, sequence mixups are not unique to a two-tier sequencing clone to provide additional sequence coverage. Mixed DNA errors are extremely rare, but subsequent re-sequencing (or deep shotgun sequencing) of the clone is unlikely to experience the same mixup. Thus, correct sequencing in the deep shotgun phase results in a mixed sequence-type error.

Mixed sequence errors vary from a single plate being mistakenly included in a sequencing project, to entire subsets of production runs being added to the first phase sequencing of a different genomic segment. New information is incorporated into a project, but it does not appear consistent with its other data. If only a minor portion of the reads are due to mixed sequence errors the assembly may not be affected, but when significant portions of the reads are incorrectly assigned (possibly due to existing name errors), the resulting assembly will be incorrect.

Mixed DNA errors arise from a contamination from a neighboring well in the glycerol stock containing the original large insert clone library, or from an accidental combination of DNA preparations during the purification of the large insert clone library. Inoculation from single colonies and subsequent fingerprint analysis of these preparations ensure that mixed DNA errors are extremely rare, but when present the resulting assembly resembles a mixed sequence error and remains intractable to our methods, as the underlying physical template of the clone is mixed.

In practice, any combination of these errors can happen. A remedial action is often possible, but occasionally a return to an early step in the process, for a re-run, is required. In certain cases even that may not be sufficient—chimeric clones have presented problems in genome mapping (Arratia et al., 1991), and cannot be effectively dealt with in sequencing.

An overview of the sequencing procedures in the Whitehead Institute’s Center for Genome Research, annotated with the types of errors that can happen at any stage, is provided in Figure 1.

**ALGORITHM**

The quality control of sequencing projects is an ongoing process, taking place at various phases. It involves the use of test plates and checking for marker content and clone overlap, comparing virtual restriction enzyme digestions with laboratory data etc. The analysis we describe takes place after the individual sequencing reads have been
collected, and the assembly of these sequences has taken place. We study the relationship between the contents of sequencing plates and the assembled contigs, and we are interested in certain parameters of the distribution of the reads.

Our proportion of high-quality sequencing reads has consistently been above 80% for a (96-well or 384-well) sequencing plate (internal QC data—we consider the sequence whose middle segment has Phred (Ewing et al., 1998; Ewing and Green, 1998) score of 20 or better, on average, a high-quality read). This still leaves us with a proportion of poor quality data. While there are reasons to include such reads in the assembly, they may skew our analysis. Poor quality sequences may become concentrated at the ends of contigs or form singlets (un-assembled reads), so we made a decision to base our analysis on high-quality reads only, even when the assembly has been done using all available data.

Studies have established that sequence depths above \(3 \times\) generate assemblies where contig lengths are larger than individual sequence lengths (Bouck et al., 1998), but they may still be too low for our purposes. Presently, our sequence assemblies are done by Phrap (Green, 1996). It is comparable, if not better than other currently available programs, but it has been pointed out (Chen and Skiena, 2000; Huang and Madan, 1999) that it can be overzealous in extending contigs. Our attention to certain contig characteristics, described below, has been in large part motivated by our awareness of this feature of Phrap.

Before proceeding with our methods, we shall provide notation which we shall consistently follow in the remainder of this paper. We shall refer to high-quality reads simply as ‘reads’, and if we talk about all reads we shall explicitly state that. We shall denote the total number of contigs forming the assembly by \(K_0\) and the number of large contigs in the assembly (as defined below) by \(K_1\). We shall denote the total number of plates participating in the assembly by \(N\), and the total number of reads in the assembly by \(M\). We shall refer to the number of reads in contig \(C_i\) as \(n_i\), to the number of reads in plate \(P_j\) as \(n_i\), and to the number of reads from plate \(P_j\) in contig \(C_i\) as \(n_{ij}\). We shall denote the probability that a read comes from plate \(P_i\) by \(p_i\), where \(p_i = n_i/M\). Since a read in an assembly must be in some contig \(C_j\) (we can view singlets as contigs of their own), the probability that a read in an unspecified contig comes from plate \(P_j\) is \(p_j\). We shall denote clone coverage (redundancy) by \(c\), average read length by \(L\), total length of a clone by \(S\) and the fraction of reads needed to overlap in order to merge them in a contig by \(\theta\). The variables defined here will remain in effect throughout this text, and we shall introduce additional ones as needed.

Since we would like to study the relationship between the plates used to sequence a clone, and the contigs formed by Phrap assembly, we want to make sure that any contig we consider provides sufficient information. We thus consider only contigs that are large enough such that the probability that they contain at least one read from each sequencing plate is 95% or greater, i.e. those \(C_j, j \in [1, K_0]\) for which \(P(n_{ij} \geq 1) \geq 0.95, \forall i \in [1, N]\). The number of reads \(n_{ij}\) from plate \(P_i\) in contig \(C_j\) can be approximated by binomial distribution, thus \(P(n_{ij} \geq 1) = 1 - (1 - p_i)^{k_j}\), for any \(j \in [1, K_0]\).

If \(P(n_{ij} \geq 1) \geq 0.95\) then \(1 - (1 - p_i)^{k_j} \geq 0.95\) or \(k_j \geq \ln(0.05)/\ln(1 - p_i)\). For very small values of \(p_i\),
which would occur when \( n_i \) is very small relative to \( M \), the value of \( k_j \) may grow unacceptably large. We thus consider only plates with at least 20 high-quality reads (out of 96), i.e., require that \( n_i \geq 20 \). We have found this restriction harmless, as our data indicates that very poor quality plates show up only occasionally and under special circumstances.

When \( n_i \geq 20 \) we have \( p_i \geq 20/M \). If we estimate \( c = M/\ell \) then \( M = c\ell/L \). Thus \( p_i \geq 20\ell/cS \), and, if we define

\[
k_0 = \frac{\ln 0.05}{\ln \left(1 - \frac{20\ell}{cS}\right)}
\]

we have that whenever \( k_j \geq k_0 \) it must also be that \( P[n_i^{(j)} \geq 1] \geq 0.95 \). The problem is then to estimate the clone coverage \( c \) so that we expect to find useful contigs in every assembly we examine.

Assuming an average read length \( \ell \) of 600 bases and average clone length \( S \) of 150 Kb, for 6× coverage equation (1) yields \( k_0 = 223.18 \). Lander–Waterman formulas (Lander and Waterman, 1988) predict the expected number of reads in a contig to be \( e^{(1-\theta)} \). We have established that in the construction of contigs \( \theta \) lies approximately between 0.1 and 0.3. For \( \theta = 0.1 \) and 6× coverage we obtain the expected number of reads in a contig to be 221.41, close to our \( k_0 \). Consequently, we decided to trigger the project checking procedure when the coverage of a clone reaches 6× or better. Due to generally high number of good quality reads on a plate, our analysis can be actually performed on coverages of 4× or more, even if higher values of \( \theta \) are assumed.

Before describing the full protocol, we shall outline several individual components of our tests.

**Distribution of reads in contigs**

When looking for a proper distribution of reads, we want to make sure that each sequencing plate contributes to the contigs as expected.

For each large contig, we assume that the number of reads from plates that build it has a multinomial distribution. If we estimate the parameters of this distribution as a set of probabilities \( p_i \) that a read has come from a plate \( P_i \), then the statistic

\[
Q = \frac{\sum_{i=1}^{N} (n_i^{(j)} - k_j p_i)^2}{k_j p_i}
\]

has a limiting chi-square distribution with \( N - 1 \) degrees of freedom, for every \( C_j \) we consider. For our size range of \( k_j \) this distribution should be adequate whenever \( N > 2 \).

We use statistic \( Q \) to do simple testing of the hypothesis that the composition of every given contig is in good agreement with that expected by completely random placement, resulting in approximately \( k_j p_i \) reads from \( P_i \) (\( i = 1, N \)) in \( C_j \). We consider the reads well distributed if we cannot reject the hypothesis at 0.95 significance level.

However, an assembly being considered may not contain a single contig large enough for our tests. This would indicate a possible problem, as the Lander–Waterman prediction is violated in this case. It would thus be likely that the coverage is less than declared, either because of a combination of projects, or an error in the clone size measurement. Our software can deal with this situation, but the theoretical model would clearly be violated.

After obtaining the individual \( p \)-values \( P_i \) for the examined contigs, we use the \( \texttt{qfast} \) algorithm (Bailey and Gribskov, 1998) to estimate the probability that their product has the value less than or equal \( p \), as:

\[
P \left\{ \prod_{i=1}^{K_i} P_i \leq p \right\} = p \sum_{i=0}^{K_i-1} \frac{(-\ln p)^i}{i!}
\]

Again, we assume that values above 0.95 suggest a problem with the assembly.

After we examine the contigs separately, we concatenate all \( C_i \) where \( k_i \geq 2 \) into a large, artificial ‘contig’. We then repeat our analysis on this one because the local analysis (contig–by–contig) may not discover low contributing plates systematically absent from large assemblies. Such plates would have most of their reads in singlets, and that would be caught in this step. Here we insist on a very good fit, at 0.05 or better, as the number of singlets in a reasonable assembly should be very small relative to the number of reads in contigs—the distribution should be almost exact. However, due to the stringency of our requirement, failing this test would make an assembly suspect (if it is not already) rather than declaring it a mixture.

**Fairness of contigs**

Even when the reads building a contig come from all plates, in expected proportions, they may be distributed in an uneven way. This could occur when the contig itself is chimeric, joined from two or more different segments (due to the already discussed greediness of Phrap). At worst, these segments may come from different clones, in which case we would detect a pattern where reads from specific plates fall in limited regions of the contig.

In our experience, chimeric contigs do appear in mixed-up projects, but it is difficult to estimate the actual rate. In artificially introduced mixups, almost every assembly contained chimeras, but they tended to appear at lower coverages and smaller contigs. In addition, if a contig contains a few foreign reads at its ends, it may, but need not, be considered chimeric. Of particular concern are large contigs consisting of segments belonging to different projects that become merged together either over the
and variance of
considering each
$P_i$ is approximately normally distributed with mean of $yl_i$
array with $P_i'$. We label
significance.

common sequence of the cloning vector, or incorrectly
resolved sequence repeats (Figure 2). In the course of
hundreds of tests we have seen only two or three such
contigs, a low occurrence rate, but sufficient to warrant
attention.

In order to estimate the fairness of a contig, we analyze
its composition with respect to each $P_i$. For each $P_i$ and
$C_j$, we form an array $x_l$, $l \in [1, n^{(j)}]$ of the expected
starting positions of reads from $P_i$ in $C_j$. We compare this
array with $y_l$, $l \in [1, n^{(j)}]$ of observed starting positions,
using the Wilcoxon–Mann–Whitney rank sum test. After
generating the sum $T$ of the ranks of $yl_i$ in the sorted array
of all $x_l$ and $y_l$, we postulate that the statistic
$$U = n_l^2 + n_i^{(j)}(n_i^{(j)} + 1) - T$$
is approximately normally distributed with mean of $n_l^{(j)}/2$
and variance of $n_l^{(j)^2}/12$. We then apply a two-
tailed test of the hypothesis that $x_l$ and $y_l$ follow the same
distribution pattern.

After analyzing each $P_i$, we normalize the data (convert
the two-tailed test to one-tailed), and apply the method
of Bailey and Gribskov (described above). We label $C_j$
unfair if our meta-analysis directs us to challenge it at 0.95
significance.

Plate separation and grouping

If it is suspected that a project might be mixed, it is desirable
to determine whether an association between subsets of
plates and subsets of contigs can be discovered, inde-
pendently of previous tests. This check may corroborate
their results or indicate that the doubt about the structure
of the assembled project was not well grounded.

In order to do the separation, we build a graph $G$,
considering each $P_i$, a vertex. We build the edges between
$P_i$ and $P_j$, such that we assign them weights $w_{ij}$ based
on their ‘distance’, defined as Pearson Product-Moment

coefficient of correlation
\[
\frac{\sum_{i=1}^{K_0} u(i) x(i) - \frac{1}{K_0} \left( \sum_{i=1}^{K_0} u(i)^2 \right) \left( \sum_{i=1}^{K_0} x(i)^2 \right)}{\sqrt{\sum_{i=1}^{K_0} u(i)^2 - \frac{1}{K_0} \left( \sum_{i=1}^{K_0} u(i)^2 \right)^2} \sqrt{\sum_{i=1}^{K_0} x(i)^2 - \frac{1}{K_0} \left( \sum_{i=1}^{K_0} x(i)^2 \right)^2}}
\]

with $K_0$ modified so that we leave the suspected unfair
contigs out of the consideration. In rare cases of negative
correlation we set $w_{ij} = 0$.

After forming $G$, we proceed to cluster its vertices into
groups of plates that correlate to each other. We iterate
over the graph, merging one vertex in each step, until we
either reach the state in which the correlation between
any two clusters is less than 0.51, or we reduce the graph
to 3, or less, vertices total. In every step we update the
distances using Lance–Williams formulas for hierarchical
clustering, where:
\[
w_{ik,j,k} = \alpha_i w_{ik} + \alpha_j w_{jk} + \beta w_{ij} + \gamma |w_{ik} - w_{jk}|
\]

Specifically, we use group average clustering, character-
ized by $\alpha_i = \frac{|i|}{|i| + |j|}$ (where $|i|$ denotes the number of
plates in cluster $i$), $\beta = 0$ and $\gamma = 0$.

At the end of this process we may have the whole graph
collapsed, or two or more distinct super-nodes. In the
former case we conclude that we face either a consistent
close data set or a preparation mixture (among samples,
not plates), and in the latter we separate plates containing
data from different clones.

IMPLEMENTATION

The analysis of the assembled sequence data is done in
stages, functioning as a finite state machine. The state of
a project at any time depends on the outcome of previous
tests, and has its own criteria for the transitions. The states
can be roughly divided into three groups: passing track,
suspect track and failure track. If a project remains on the
passing track, it is declared clean. If it ever ‘slips’ onto the
suspect track, the examination continues, but the criteria it
must pass in the subsequent tests becomes more stringent
and if it fails again, it moves to the failing track. A
suspect project must be ultimately declared either passing
or failing, and if additional evidence of problems could
not have been found, it can be declared clean. Otherwise,
its fails. Once a project is on the failure track it cannot be
transitioned back to either passing or suspect path—at that
point the goal is to establish the nature of the problem.
All projects that fail at the end are labeled for inspection
by a data analyst. An outline of this process is shown in
Figure 3.

We first check the size of the assembly, and how well
it correlates to the estimated length of the clone. If it is
not within 20% of the estimate, we look whether its size
is such that it could have come from the library that the
clone is presumed to belong to. If it is not likely (too big,
or too small), the project is placed on the failure track. Otherwise, a mistake could have been made during the size measurement, and the project is declared a suspect.

The next test is the distribution of reads. If the correct distribution cannot be rejected at 0.99 significance, or declared a suspect at 0.95 significance (0.05 for merged assembly), large contigs are examined for fairness, at 0.99 significance (0.95 to suspect). A properly distributed clone of expected size, composed of fair contigs, is considered correct, and can be approved for further work, and finishing. Due to the nature of these tests, this verification can be done efficiently, with minimal use of processor time and computer memory.

For a project on the suspect track (incorrect clone size) the distribution and contig fairness must be confirmed at 0.95 significance. In addition, a suspect project must be checked for the presence of multiple vector-insert junctions, and the plate separation (based on contigs whose fairness cannot be challenged at 0.95 significance, whenever possible) must be attempted.

If any of the tests cannot be applied, due to the absence of large contigs, for example, the project is transferred to the failure track. Once there, it is checked for multiple SP6 or T7 occurrences, and a plate separation attempt is made. The results of these tests are stored together with the assembly.

While this process can detect different mixups, not all can be resolved. If the plates do not separate, we may suspect that we are dealing with mixed DNA (sample mixup), but in general the situations where the data do not look right, without a clear indication of what has happened, must remain uncharacterized, and handled by other methods.

The full description of our decision tree is available with the implementation for downloading. Our software is organized in a hierarchical way, so that tools and criterions can be flexibly combined, according to the needs of a particular laboratory.

**RESULTS**

In order to evaluate our methods, and calibrate thresholds, we simulated assemblies and introduced mixups to real projects. We repeated all tests for coverages between 6× and 12×. As we have not noticed significant variations in the data for varying redundancies, we shall present them together.

We measured the distribution of goodness-of-fit data on assemblies of clean sequencing projects, and on assemblies to which we added 1, 2, and more foreign plates, up to having 50% of sequence from a different clone. Each arrangement was evaluated on 10 different datasets for each coverage, and the results are shown in Tables 1 and 2.

As we could not effectively induce the creation of chimeric contigs in Phrap assemblies, we measured the results of our contig fairness test using a simulated model. Data have been organized in groups of 300 artificial assemblies at about 8× coverage. We simulated the assemblies with 1, 3 and 5 large contigs, with the largest one making 80, 60 and 30% of the consensus, respectively.

Table 3 lists the results of this simulation, performed under different contig compositions. In order to control variations, randomizations (contig shuffle) of the layout of reads from different plates were performed at 5, 50 and 100% levels. Reads were placed sequentially by their plates, reads from $P_1$ first, followed by reads from $P_2$ etc. This order was then disrupted by random shuffling, where $z\%$ shuffle of a contig of size $S$ denotes $zS/100$ swaps of reads at random positions. Thus, 5% shuffle yields a very organized (biased) placement, while 100% shuffle leads to a completely random layout (and the config should thus be fair).

Finally, we sampled a large number of projects from our sequencing pipeline, and re-assembled them in the following manner: for each coverage we counted the plates necessary to achieve it; we mixed in the desired number of plates from a different project, varying the number of...
Table 1. Results of the read distribution goodness-of-fit for sets of 10 projects to whose assembly a varying number of foreign plates has been added, and coverages of 6× through 12×

<table>
<thead>
<tr>
<th>Other plates</th>
<th>Clone count</th>
<th>0.00–0.05</th>
<th>0.05–0.25</th>
<th>0.25–0.75</th>
<th>0.75–0.95</th>
<th>0.95–0.99</th>
<th>0.99–1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>70</td>
<td>38</td>
<td>15</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>70</td>
<td>30</td>
<td>15</td>
<td>12</td>
<td>2</td>
<td>15</td>
<td>38</td>
</tr>
<tr>
<td>2</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>60</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>49</td>
</tr>
<tr>
<td>12</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>49</td>
</tr>
<tr>
<td>13</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>37</td>
</tr>
<tr>
<td>14</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>29</td>
</tr>
<tr>
<td>15</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>17</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>18</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 2. Results of the read distribution goodness-of-fit for artificially merged contings (all contings with 2 or more reads), for sets of 10 projects to whose assembly a varying number of foreign plates has been added, and coverages of 6× through 12×

<table>
<thead>
<tr>
<th>Other plates</th>
<th>Clone count</th>
<th>0.00–0.05</th>
<th>0.05–0.25</th>
<th>0.25–0.50</th>
<th>0.50–0.75</th>
<th>0.75–0.95</th>
<th>0.95–1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>70</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>70</td>
<td>30</td>
<td>15</td>
<td>12</td>
<td>8</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>70</td>
<td>54</td>
<td>8</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>70</td>
<td>66</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
<td>68</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>70</td>
<td>66</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>70</td>
<td>66</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>70</td>
<td>67</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>70</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>70</td>
<td>69</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>60</td>
<td>59</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>50</td>
<td>46</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>50</td>
<td>48</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>40</td>
<td>39</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>40</td>
<td>39</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>30</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

As summarized in Table 4, clean assemblies were declared a mixup in 8 out of 70 cases (11.43%), and the introduction of a single foreign plate was not recognized in 7 out of 70 (10%). An addition of 2 foreign plates has not been detected once (1.43%, at 6× coverage), and once they could not have been isolated. Otherwise, all mixups involving more than 1 plate of DNA foreign to the clone were discovered and properly characterized.

foreign plates from 0 to 50% of the total, repeating the procedure 10 times for each type of mix and coverage. At each step, and each repetition, we selected the projects and plates at random, gathering a new data set and a different assembly. We took care that the clones were correct, and that the criteria to decide which projects were clean were as independent as possible from the logic of our tests.
Identification of mixups among DNA sequencing plates

Table 3. Results of the simulation of contig fairness, made on 300 artificial assemblies, and varying composition of the contigs.

<table>
<thead>
<tr>
<th>Largest contig read number</th>
<th>Largest number of clones</th>
<th>Large contigs</th>
<th>Contig shuffle</th>
<th>0.00–0.50</th>
<th>0.50–0.75</th>
<th>0.75–0.85</th>
<th>0.85–0.95</th>
<th>0.95–0.99</th>
<th>0.99–1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>80%</td>
<td>1613</td>
<td>300</td>
<td>300</td>
<td>5%</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>80%</td>
<td>1613</td>
<td>300</td>
<td>300</td>
<td>50%</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>80%</td>
<td>1613</td>
<td>300</td>
<td>300</td>
<td>100%</td>
<td>300</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>60%</td>
<td>1210</td>
<td>300</td>
<td>858</td>
<td>5%</td>
<td>91</td>
<td>14</td>
<td>9</td>
<td>15</td>
<td>40</td>
</tr>
<tr>
<td>60%</td>
<td>1210</td>
<td>300</td>
<td>863</td>
<td>50%</td>
<td>443</td>
<td>30</td>
<td>16</td>
<td>38</td>
<td>88</td>
</tr>
<tr>
<td>60%</td>
<td>1210</td>
<td>300</td>
<td>870</td>
<td>100%</td>
<td>841</td>
<td>22</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>30%</td>
<td>605</td>
<td>300</td>
<td>1500</td>
<td>5%</td>
<td>279</td>
<td>14</td>
<td>20</td>
<td>43</td>
<td>86</td>
</tr>
<tr>
<td>30%</td>
<td>605</td>
<td>300</td>
<td>1500</td>
<td>50%</td>
<td>1083</td>
<td>40</td>
<td>14</td>
<td>46</td>
<td>138</td>
</tr>
<tr>
<td>30%</td>
<td>605</td>
<td>300</td>
<td>1500</td>
<td>100%</td>
<td>1446</td>
<td>39</td>
<td>8</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 4. Results of the application of the composite suite of tests, for various degrees of mixups, on sets of 10 projects for each of the coverages 6 × through 12 ×.

<table>
<thead>
<tr>
<th>Foreign plates</th>
<th>Number of clones</th>
<th>Clone correct</th>
<th>Sample mixup</th>
<th>Plates mixup</th>
<th>Uncharacterized mixup</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>70</td>
<td>62</td>
<td>0</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>70</td>
<td>7</td>
<td>0</td>
<td>63</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>70</td>
<td>1</td>
<td>0</td>
<td>68</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>60</td>
<td>0</td>
<td>0</td>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

DISCUSSION

Our integration of tools has been tuned towards rejection of faulty assemblies. However, other applications may have different priorities, and our software can be configured to perform differently.

Because of our calibration, one can be quite sure that an assembly that has been declared correct is indeed correct (at less than two plates worth of contaminating samples), but there is a non-negligible number of rejected correct clones (due to a somewhat unbalanced assembly, or fragmented contigs). It is thus important to understand the meaning of the tests, so that a data analyst can examine the logs and make decisions. A single foreign plate will rarely harm the assembly—it may generate a larger number of singlets and small contigs, but is unlikely to affect the final consensus sequence. However, our tools will detect even a single plate contamination about 90% of time.

Sometimes one should be cautious of a single foreign plate. We routinely use test plates to evaluate a clone, checking their reads for known markers before proceeding with deep coverage. If a name error happens at this stage, it would lead to a mixed sequence error which might not be detected, leading us to finish an irrelevant clone. Therefore, if there is a single plate of special importance in a sequencing project, the marker content of large contigs should be checked after the assembly, even if our tests declare the project clean.

The real performance of our checking is expected to be somewhat worse than reported above (on our production datasets the number of reported mixups, mostly false positives, was about 15%). Our test set consisted of clean clones with a graceful assembly, large contigs, few singlets and high sequence quality. There are regions in DNA for which statistics cannot help. One such example is a clone rich in low-complexity sequence, with sites where it is difficult to get reads. The assembly of such clone would be fragmented, with no contigs of expected length, given the coverage. This situation would force our software to take a cautious path, due to small sample size, and an otherwise acceptable imbalance in the distribution of reads would cause this project to fail. However, while settings like this increase the failure rate of correct projects, we do not expect an increase in the number of unrecognized mixups.

Our decision to perform the tests only after 6 × coverage has been achieved, and our requirement of a minimal number of high-quality reads are not essential. Our goal was to maintain a consistent model, rather than compromising it to handle scarce data. If the need arises, this can be changed, as our methods proved to be robust enough to perform in less favorable settings.

1425
ACKNOWLEDGEMENTS
This work was supported by the genome sequencing grant from the National Institutes of Health, to the Whitehead Institute Center for Genome Research (Eric S. Lander). The software for the statistical distributions has been downloaded from Netlib (http://www.netlib.com) at the University of Tennessee and the Oak Ridge National Laboratory. We are thankful to the Netlib maintainers and the authors of the package: Barry W. Brown, James Lovato and Kathy Russell, for permitting us to use and re-distribute their software. We would like to thank Dr John Lehoczky, of Carnegie–Mellon University, for helpful comments, and Dr Eric Lander, of MIT and the Whitehead Institute, for his support in making this work possible. We are indebted to the anonymous referees, as their suggestions helped us to substantially improve our methods.

REFERENCES