BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 10 2006, pages 1217–1224
doi:10.1093/bioinformatics/btl063

Sequence analysis

Refinement of optical map assemblies

Anton Valouev[1,2,*,†], Yu Zhang[3,†], David C. Schwartz[4] and Michael S. Waterman[2]

[1]MCB, 1050 Childs Way, Los Angeles, CA 90089-2910, USA, [2]Department of Mathematics, University of Southern California, Los Angeles, CA, USA, [3]Department of Statistics, Harvard University, MA, USA and [4]Laboratory for Molecular and Computational Genomics, Departments of Genetics and Chemistry, University of Wisconsin-Madison, WI, USA

Received on September 30, 2005; revised on February 2, 2006; accepted on February 19, 2006
Advance Access publication February 24, 2006
Associate Editor: Keith A Crandall

ABSTRACT

Motivation: Genomic mutations and variations provide insightful information about the functionality of sequence elements and their association with human diseases. Traditionally, variations are identified through analysis of short DNA sequences, usually shorter than 1000 bp per fragment. Optical maps provide both faster and more cost-efficient means for detecting such differences, because a single map can span over 1 million bp. Optical maps are assembled to cover the whole genome, and the accuracy of assembly is critical.

Results: We present a computationally efficient model-based method for improving quality of such assemblies. Our method provides very high accuracy even with moderate coverage (<20 ×). We utilize a hidden Markov model to represent the consensus map and use the expectation-Maximization algorithm to drive the refinement process. We also provide quality scores to assess the quality of the finished map.

Availability: Code is available from www.cmb.usc.edu/people/valouev/

Contact: valouev@usc.edu

1 INTRODUCTION

The comprehensive assessment of human genome polymorphism and somatic aberration drives meaningful association studies and insights into breakpoints that are prevalent in many types of cancer. As such, restriction maps reveal many types of differences that include apparent inserions/deletions, inversions, tandem duplications and even SNPs (S. Reslewic *et al*., manuscript submitted). Sequencing efforts also benefit from the use of accurate physical maps that span entire genomes. More specifically, physical maps provide scaffolds essential for sequence finishing and validation. Given this context, the optical mapping system (Valouev *et al*., 2005; Dimalanta *et al*., 2004; Ananthraman *et al*., 1999) constructs high-resolution ordered restriction maps from individual genomic DNA molecules that are assembled into map contigs spanning entire genomes.

Since optical maps are produced from single DNA molecules, a unique set of errors must be dealt with to ensure accuracy of inferred consensus map. In optical mapping, not only map assembly is more difficult compared with sequencing, but also correctly assembled draft consensus maps may contain inaccuracies that include false cuts, missing cuts, and fragment size inaccuracies. Our optical map assembler (A. Valouev *et al*., manuscript in preparation) works in the following fashion: (1) it identifies pairs of overlapping optical maps; (2) locates contigs as subsets of overlapping optical maps representing a particular portion of genome; (3) extracts a sequence of overlapping optical maps that represents the region from beginning to the end; and (4) then extracts a draft map by simply merging in order sequence of overlapping optical maps. The draft map represents a composite of a sequence of optical maps and outlines approximate positions of most restriction sites of corresponding region of the genome (Fig. 1). Notice from Figure 1 that at each position of draft map, fragments come from a single optical map, and also not all optical map measurements are incorporated, since only a sequence of overlapping optical maps is chosen for the construction. In sequence assembly, this problem is dealt with by constructing multiple alignment of all sequence reads and then inferring the consensus by majority vote across the columns of multiple alignment. We cannot do our consensus inference the same way because there is no multiple alignment algorithm for restriction maps. Instead, we choose to construct the draft map that only includes the subset of all measurements and then refine it by combining remaining majority of optical map measurements and voting off draft map inaccuracies. We proceed iteratively until no further improvements can be made to the draft map and all possible measurements are incorporated. This last step of improving the draft map accuracy is termed 'assembly refinement' and is considered in this paper.

We will describe the assembly refinement method that allows to take inaccurate draft consensus map and gradually improve its accuracy by introducing corrections. Our method takes advantage of the 'majority rule' strategy by combining the information from high-quality optical maps in order to purge draft consensus map of its inaccuracies. In our method, we employ a hidden Markov model (HMM) to represent the consensus map of interest. An expectation-maximization (EM) algorithm (Rabiner, 1989) is then used to eliminate the errors and update the consensus map within limited iterations. In our approach, the progress of the refinement is driven by changing the consensus towards the state where the final consensus map best fits the optical map data.

Another important benefit of our analysis is that it addresses the quality of the finished physical map. In sequencing efforts, quality scores are routinely provided for gauging base calls, as well as for finished sequence. For base calling, this is accomplished through analysis of sequence traces (Ewing and Green, 1998), while the quality assessment of finished consensus sequence is inferred through analysis of a set of sequence reads (Churchill and

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author 2006. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

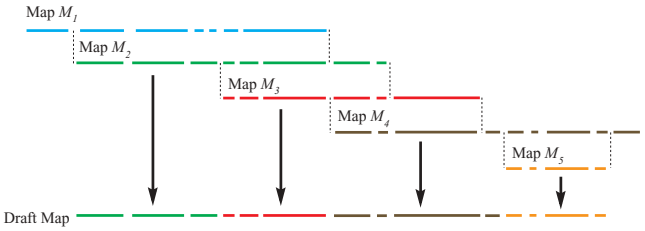

Fig. 1. Construction of draft consensus map from pairwise optical map alignments.

Waterman, 1992). The situation is quite different in optical mapping, since raw quality scores for each optical map will require the development of new set of metrics and analysis, specially contoured to single molecule datasets. Nonetheless, the problem remains of how to quantitatively assess the quality of finished physical map contigs in ways that would parallel techniques for sequence assembly. Here, we address this problem by providing consensus quality scores that are based on the optical maps contributing to the final assembly. These scores address correctness of estimated restriction sites, possibility of additional sites at each map position. Finally, we provide a simple method to find regions of potential mis-assembly by screening the optical map coverage and reporting regions where coverage is abnormally low.

2 OPTICAL MAPS AND ERROR MODELS

The optical mapping system (Dimalanta *et al*., 2004; Zhou *et al*., 2004) uses randomly sheared genomic DNA molecules (up to 4 Mb) as the mapping substrate. A microfluidic device is used to both elongate and deposit long DNA molecules onto charged glass surfaces for analysis (Dimalanta *et al*., 2004). After deposition, a solution containing buffer and a restriction enzyme is applied to cleave immobilized DNA molecules, which are then fluorochrome stained to reveal cleavage sites after imaging by fluorescence microscopy. Automated imaging enables large image datasets to be acquired and these are then analyzed by machine vision to produce ordered restriction maps from individual molecules. Restriction sites (cut sites) are characterized as punctuated dark gaps along a molecular backbone; likewise spurious cut sites bear a similar appearance. As previously mentioned, a unique set of errors is associated with single molecule datasets. Below we state the statistical models associated with the optical map inaccuracies. Interested readers can refer to Valouev *et al*. (2005) to get familiar with the detailed description and justification for the choice of distribution parameters.

- *Sizing errors*. Apparent length measurements are not used to estimate the size of restriction fragments since this would require uniform elongation of DNA molecules that would also retain its biochemical competence. Instead, integrated fluorescence intensity is used to estimate the size of each restriction fragment; however, such measurements suffer errors because of unequal distribution of fluorochromes. Our analysis indicates that for a DNA fragment of size Y, the estimated fragment size X follows a normal distribution $X \sim N(Y, \sigma^2 Y)$ for some constant σ (Valouev *et al*., 2005) This is due to integration of light intensity emitted along the span of the DNA fragment. The value of σ can vary slightly depending on experimental conditions, but is usually taken to be 0.6. Sizes of smaller fragments (<2 Kb) follow a different distribution which we take to be $X \sim N(Y, \eta^2)$ for some constant η.
- *Missing cuts*. Although the enzyme efficiency is high, most optical mapping datasets show that 20% of restriction sites remain undigested and therefore are not observed in the optical data. We treat the digestion of each restriction site as a Bernoulli event with probability of success $p = 0.8$. Furthermore, the digestion at different sites and different DNA molecules is assumed to be independent.
- *False cuts*. After the DNA is attached to the glass surface, it can break at random positions that do not contain restriction sites. The breakage process is assumed to be uniform, and therefore the number of the breakage sites follows a Poisson distribution with the rate $\zeta = 0.005 \times Kb^{-1}$.
- *Missing fragments*. Restriction fragments <1 Kb are not consistently immobilized on a optical mapping surfaces since electrostatic retention forces scale with DNA length—covalent attachment schemes would obviate this problem, but would also hinder enzymatic activity. Consequently, such fragments are not uniformly represented within an optical mapping dataset.
- *Molecular chimerism*. DNA molecules can cross paths upon deposition on an optical mapping surface, so that unambiguous resolution can be difficult in many of these instances. Thus sometimes, the image processing software may report a concatenate of two unrelated optical maps as a single map—we call this effect a molecular chimerism.

3 HIDDEN MARKOV MODEL

We use a HMM to represent the consensus map. Let $0 = c_0 < c_1 < \cdots < c_{n+1} = s_C$ mark the positions of n cut sites on the consensus map C, so that s_C is the total size of map C in base pairs ($s_C = \|C\|$). Here we include the beginning and the end of the map in the set of sites. Our HMM for physical mapping is shown in Figure 2.

In this model, there are:

(1) *Match states*. We have $n + 2$ match states corresponding to n internal cut sites c_i and the two ends of the consensus map. If an optical map contains a cut corresponding to the site c_i on the consensus, then we say that it goes through a match state M_i at that position.

(2) *Delete states*. We have n delete states corresponding to n internal cut sites of the consensus map. If an optical map covers the position c_i and has no cut site corresponding to c_i, then we say that it goes through a delete state D_i at that position.

(3) *Insert states*. We have $n + 1$ insert states corresponding to possible inserted sites between the adjacent cut sites of the consensus map. If an optical map covers an interval (c_i, c_{i+1}) between two adjacent sites in the consensus map and contains $l \in \{0, 1, \dots\}$ cut sites falling between the sites c_i and c_{i+1}, then we say it goes through the state I_i^l at that position.

If the consensus map is accurate, then match states represent the digested sites on optical maps delete states represent missing cuts on optical maps and insert states represent false cuts on optical maps. By aligning optical maps to the consensus, we

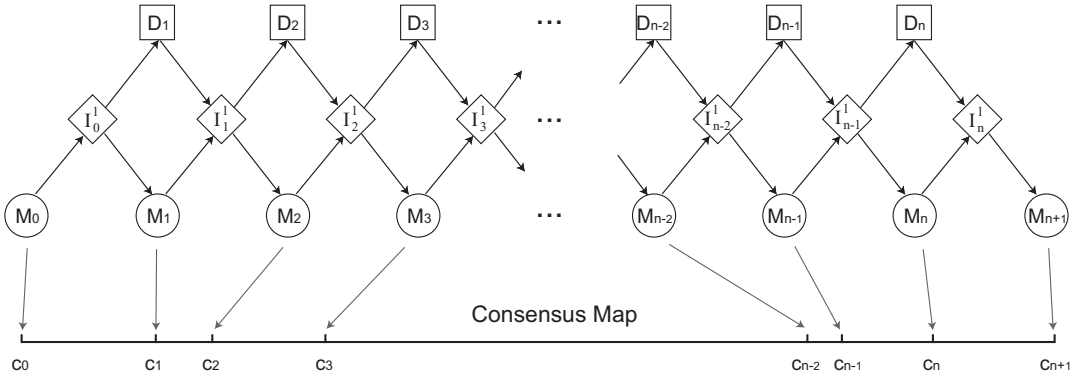

Fig. 2. HMM for physical mapping. Each restriction cut site has a group of corresponding states in the model.

represent each alignment as a path through the HMM. Further, we use this path information to detect and correct inaccuracies in the consensus map.

4 ALIGNMENTS

We represent alignments by pairs of matching sites of the consensus and optical maps. Suppose that optical map O is aligned to the consensus map C. Let $0 = o_0 < o_1 < \cdots < o_{m+1} = s_O$ mark the positions of cut sites on map O, where s_O gives the total size of the optical map O ($s_O = \|O\|$). Alignment between O and C is given by an ordered set of matching site pairs $(o_{\alpha_i}, c_{\beta_i})$ for $i \in \{1, \ldots, h\}$ with natural ordering given by $o_{\alpha_0} < \cdots < o_{\alpha_h}$ and $c_{\beta_0} < \cdots < c_{\beta_h}$.

To relate this to our model, we represent each alignment as the path through HMM. For these alignments, matching sites are represented by visiting match states. If the alignment path visits the delete state, the corresponding consensus site is missing in the optical map. Likewise, a visit of the insert state represents cuts in optical maps not present in the consensus. Figure 3 gives an example of an optical map aligned to a consensus map along with the representation of its alignment as a path through the model.

Given the alignment paths for all optical maps, we can calculate how many alignments go through a certain state at each consensus position c_i, and infer errors in the consensus map. For instance, if at some consensus location c_i most alignments go through the delete state D_i, it indicates that site c_i may not be present in the genome and therefore should be removed from the consensus map. When counting these alignments, we either use most likely alignments or expected number of alignments at each consensus map locus.

The transition probabilities for the Markov model are calculated depending on the size of the restriction fragment $s = c_{i+1} - c_i$ between the two adjacent consensus sites c_i and c_{i+1}. Conditioned on the correctness of the consensus map, the probability of going through the insert state I_i^l between two adjacent consensus sites c_i and c_{i+1} is given by the Poisson likelihood of having exactly l false cuts within a region of size s. Hence the transition probability from match to insert and from delete to insert is given by

$$\Pr(M_i \to I_i^l \,|\, M_i) = \Pr(D_i \to I_i^l \,|\, D_i) = \Pr(F = l \,|\, \zeta, c_{i+1} - c_i) = \mathrm{e}^{-\zeta(c_{i+1}-c_i)} \frac{\zeta^l (c_{i+1} - c_i)^l}{l!},$$

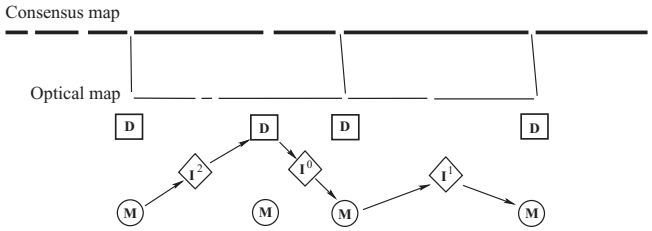

Fig. 3. Representation of optical map alignment by a path through the model states. In the top figure, only the matching sites are connected by lines. The bottom figure shows how the corresponding alignment can be represented as a path through the model states.

where for each optical map, l is the number of additionally optical map sites falling between consensus sites c_i and c_{i+1}. Notice here that case $l = 0$ corresponds to no internal cuts (insertions) between two match states. This is the case when there are no false cuts in the region. Since we use Poisson distribution to calculate a probability of having l false cuts between consensus sites c_i and c_{i+1}, the probability of going from I_i^l to a match state M_{i+1} is given by p, digestion efficiency. Therefore, transition probability of going from I_i^l to D_{i+1} is given by $1 - p$:

$$p = \Pr(I_i^l \to M_{i+1} \,|\, I_i^l) = 1 - \Pr(I_i^l \to D_{i+1} \,|\, I_i^l).$$

In our model, emissions correspond to observing particular fragment sizes in optical maps. Therefore emission likelihoods are calculated according to the size error model that we have formulated earlier in this paper. Missing and false cuts complicate these calculations since we need to account for regions rather than individual restriction fragments. Corresponding calculations require knowing the position of the last match state to determine the size of the matching region. And since the last match state can occur more than one model steps before a current profile position, our model can not be represented by a simple first order HMM. Instead, we keep a track of last δ model steps. In our implementation we take $\delta = 5$, i.e. the most recent site match to the consensus within the last five consensus sites.

The specified transition probabilities allow us to calculate the probability of aligning a particular optical map to a specified region of the consensus map. As an alternative to this approach, the most

likely alignment can be used instead. In this case, the most likely alignment is given by an optimal alignment score to the consensus map (Valouev *et al.*, 2005).

5 METHODS

Our method relies on iterative refinement of the consensus map to achieve the accurate map approximation that best fits the optical map data. Given the initial consensus map that may contain errors in the form of missing cuts false cuts and fragment size inaccuracies, we refine it by iterating over the following steps:

- deletion of some sites from the consensus map
- addition of some sites to the consensus map
- re-estimation of consensus fragment sizes.

Every time we update the model, we realign optical maps to the new consensus, and use this new alignment information for the next update step.

Our procedure uses EM algorithm for the HMM update. The expectation step (E-step) is accomplished by aligning optical maps to the profile. In this calculation, alignment probabilities enable finding the expected number of optical maps going through each of the model states. Alternatively, optimal alignment can be used for each optical map. This gives the exact number of best optical map alignments going through each of the model states. This information is used to update the model during the maximization step (M-step) when consensus errors are locally corrected by applying one of the model-update steps described above.

Our procedure can be easily scaled-up to handle large mammalian-sized genomes. The method complexity is $O(d \times n \times t)$, where d is the average depth of coverage by optical maps, n is the number of sites of the consensus map and t is the number of refinement iterations, The stopping criteria for the refinement procedure is usually taken to be a consensus state when no further changes can be made to the consensus map. Our experience shows that iterative process converges within 13–15 iterations (Table 2) and in some cases may depend on the coverage.

In the next section we describe the exact update procedure for each of the update steps.

5.1 Update: deletion of consensus sites

For each site on the current consensus, we perform a hypothesis test to determine whether the site should be removed. Suppose for the consensus site c_i, k optical maps cover this position. Furthermore, suppose x of them go through the corresponding delete state D_i. If the most likely alignments are used, x is given by the number of the most likely alignments going through this delete state. If the model profile is used instead, x is given by the expected number of optical maps going through corresponding delete state by adding probabilities of all transitions to delete state over all optical maps aligned to the profile.

The null hypothesis assumes that site c_i is a correct site. Therefore, x maps going through the delete state D_i must be due to failed enzyme digestion at the corresponding genomic position c_i. Given digestion efficiency p, we can calculate p-value of observing at least x undigested sites. This probability is given by the binomial distribution:

$$p_v = \Pr(X \geq x \,|\, k) = \sum_{j=x}^{k} \binom{k}{j} (1-p)^j p^{k-j}.$$

Another statistic that we use in addition to the p-values, employs a likelihood ratio. It controls the number of false positives produced by the first test. Under the H_0, the probability of the data is given by

$$L_{H_0} = \binom{k}{x} (1-p)^x p^{k-x},$$

since x out of k optical maps have undigested sites corresponding to c_i. The alternative hypothesis assumes that the site c_i is not correct, and hence $k - x$ matching to c_i optical maps must be because of false cuts at that position. This likelihood is given by the following expression:

$$L_{H_a} = \binom{k}{x} u^{k-x} (1-u)^x,$$

where u is the probability of having false cuts within a region of size $s = c_{i+1} - c_i$ and is given by $u = 1 - e^{-\zeta s}$. Now we can conclude that likelihood ratio for this test is given by

$$\text{lr} = \frac{L_{H_0}}{L_{H_a}} = \left(\frac{1-p}{1-u}\right)^x \left(\frac{p}{u}\right)^{k-x}.$$

The null hypothesis is rejected when the p-value and the likelihood ratio are both smaller than the specified thresholds. If H_0 is rejected, flanking consensus fragments are merged to produce a single updated fragment of the combined size.

5.2 Update: addition of consensus sites

For each interval (c_i, c_{i+1}) of the consensus map, we perform a hypothesis test to determine whether a cut needs to be added between consensus sites c_i and c_{i+1}. Suppose that k optical maps cover the interval (c_i, c_{i+1}), and x of them carry cuts that fall in between c_i and c_{i+1}.

As before, if most likely alignments are used, x is given by the number of the most likely alignments going through a corresponding insert state I_i^l for $l \geq 0$. If the model profile is used instead, x is given by the expected number of maps going through corresponding insert state I_i^l for $l \geq 0$. This number is given by adding probabilities of transitions to insert state of all alignments over all optical maps aligned to the profile.

Under the null hypothesis, the consensus map is correct, and hence there are no additional restriction sites between c_i and c_{i+1}. Therefore, the p-value of having at least x maps carrying such additional sites is given by

$$p_v = \Pr(X \geq x \,|\, k) = \sum_{j=x}^{k} \binom{k}{j} u^j (1-u)^{k-j},$$

since these additional cuts must be due to random DNA breakage. Here u gives the probability of observing at least one false cut within a given region of size $s = c_{i+1} - c_i$. Thus u can be calculated according to the following expression:

$$u = 1 - e^{-\zeta s},$$

where ζ is the rate of the random DNA breakage.

To control the false positive rate for the first test, we also use additional likelihood ratio test. Under H_0, extra sites on optical maps must be due to random DNA breaks. Under H_a, the consensus map is not accurate, and must contain an additional site between c_i and c_{i+1}. Corresponding optical map sites must be then due to the digestion of this restriction site by the endonuclease. With the digestion efficiency p this gives an expression for the likelihood ratio:

$$\text{lr} = \frac{L_{H_0}}{L_{H_a}} = \left(\frac{u}{p}\right)^x \left(\frac{1-u}{1-p}\right)^{k-x}.$$

Note that in order for the site to be inserted, cut locations on optical maps have to be consistent. To ensure this, additional tests may be applied. Our experience however shows that the first two tests are sufficient to make an accurate decision.

We reject the H_0 when the p-value and the likelihood ratio are smaller than their corresponding test thresholds. If the null hypothesis is rejected, the new site location is given by the maximum likelihood from the maps carrying inserts. Suppose y is the location of the new consensus site between c_i and c_{i+1}. Furthermore, suppose an optical map aligns exactly to the interval (c_i, c_{i+1}) so that optical map sites o_l and o_r are matching to c_i and c_{i+1} respectively. Let optical map site o_q correspond to the insertion within the interval (c_i, c_{i+1}). Hence we have $o_l < o_q < o_r$. Under the size

error model, we have $o_q - o_l \sim N(y - c_i, \sigma^2(y - c_i))$ and $o_r - o_q \sim N(c_{i+1} - y, \sigma^2(c_{i+1} - y))$. Thus the corresponding likelihood of observing o_q conditional on the position y of the new site is given by

$$l(y) = \frac{1}{2\pi\sigma^2\sqrt{(y - c_i)(c_{i+1} - y)}} \times \exp\left[-\frac{(o_q - o_l - y + c_i)^2}{2\sigma^2(y - c_i)} - \frac{(o_r - o_q - c_{i+1} + y)^2}{2\sigma^2(c_{i+1} - y)}\right]$$

Therefore, the total likelihood of the data is given by

$$L(y) = \prod_j l_j(y),$$

where the product is taken over x maps with insertions falling inside the interval (c_i, c_{i+1}). New site position y is then calculated using an iterative, gradient-based maximization of $L(y)$.

5.3 Update: fragment size re-estimation

In our approach, model update is driven towards maximization of the likelihood of the observed data. Therefore, we use maximum likelihood estimation to re-estimate sizes of the consensus fragments. Consider the i-th consensus fragment and optical maps that match to that fragment exactly. In other words, every optical map in our consideration must have sites matching to consensus sites flanking the i-th consensus fragment.

Let $x_1, \ldots, x_k$ be the sizes of regions of optical maps corresponding to the i-th consensus fragment. According to our error model, $x_j \sim N(y_i, \sigma^2 y_i)$ for all $j \in \{1, \ldots, k\}$, where y_i is the true underlying size. The maximum likelihood method gives the following expression for the size estimate of the i-th consensus fragment:

$$\hat{y}_i = \sqrt{\frac{\sigma^4}{4} + \frac{1}{k}\sum_{j=1}^{k} x_j^2} - \frac{\sigma^2}{2}.$$

To improve the estimation, we can take into account the quality of the optical map alignments from where $x_1, \ldots, x_k$ originate and weigh each x_j^2 by the matching likelihood.

5.4 Alignment selection

In order to ensure the convergence of our consensus map to the correct approximation, it is critical to minimize the number of spuriously aligned optical maps, so that they do not contribute to our estimations. We use the alignment score to assess whether the optical map comes from a particular consensus map region. The threshold for the alignment score is chosen based on the simulation such that the false positive rate is lower than the specified value.

Since draft consensus map may contain a multitude of errors, few optical maps may initially align well. However, with the progress of consensus map refinement, more consensus regions can be corrected, and thus more optical maps can be accurately aligned. For each refinement step, we realign the set of optical maps to the consensus map. Accurate alignments are then selected to be used for further error correction.

6 QUALITY SCORES

Quality score provides a practical means to assess the quality of the finished restriction map. Our quality scores address the accuracy of cut sites, possible additional sites between the adjacent consensus sites, sizing errors of consensus fragments and potential regions of mis-assembly.

- *Restriction site quality score*. Restriction site quality score provides comparison of likelihoods of two opposing hypothesis. H_0 asserts that consensus site c_i is correct and hence optical maps must have matching sites to the given consensus site. H_a asserts that the site is not correct, hence all corresponding matching optical map sites must be only due to false cuts. If x optical maps carry matching sites and l of them span the corresponding consensus position, the score is given by the following expression:

$$\text{score}(c_i) = x \cdot \log\left(\frac{p}{u}\right) + (k - x) \cdot \log\left(\frac{1 - p}{1 - u}\right),$$

  where $u = 1 - e^{-\zeta(c_{i+1} - c_{i-1})}$. This quality score is derived from the corresponding test likelihood ratio (see above). Positive scores indicate correctness of consensus site c_i.

- *Site addition score*. Site addition score between consensus sites c_i and c_{i+1} provides comparison of likelihoods of two opposing hypothesis. H_0 asserts that there should be no additional sites between c_i and c_{i+1}, and hence all optical map cuts falling between c_i and c_{i+1} must be explained by random DNA breakage. H_a on the other hand asserts that these cuts may be due to presence of restriction site between consensus sites c_i and c_{i+1}. If x optical maps carry cut sites falling between consensus sites c_i and c_{i+1} and total of k optical maps span this position by their alignments, the corresponding quality score is given by the following expression:

$$\text{score}(c_i, c_{i+1}) = x \cdot \log\left(\frac{1 - u}{1 - p}\right) + (k - x) \cdot \log\left(\frac{u}{p}\right),$$

  where $u = 1 - e^{-\zeta(c_{i+1} - c_i)}$. Thus a positive score is indicative of no additional sites between consensus sites c_i and c_{i+1}.

- *Mis-assembly p-values*. In case of correctness of consensus map, optical map coverage process (Waterman, 1995) at each consensus position must follow Poisson distribution with rate $c = LN/G$, where G is the size of the target genome, L is the average size of optical map, and N is the number of optical maps. Hence, mis-assembled regions must exhibit an abnormally low coverage across mis-assembled positions, since optical maps will fail to align well. If x optical maps span consensus map position c_i by their alignments, the corresponding p-value is given by the following expression:

$$p_{c_i} = \sum_{j=0}^{x} e^{-c}\frac{c^j}{j!}.$$

  This p-value may be used to find regions of abnormally low coverage resulting from incorrect assembly.

7 RESULTS

In order to evaluate the accuracy of our method, we selected the *Yersinia pestis* strain KIM bacterial genome to demonstrate how errors in the initial assemblies can be corrected using our refinement procedure.

The reference *Y.pestis XhoI* restriction map was produced by *in silico* digestion of 247 restriction sites of DNA sequence (Zhou *et al*., 2002). Our dataset comprised 251 optical maps, equivalent of roughly 50× coverage of the whole genome.

We first conducted initial assembly of optical maps. This involved the calculation of all pairwise alignments (or overlaps) of optical maps (Valouev *et al*., 2005). We then selected accurate overlaps based on the score and the matching measure of each overlap. This produced 691 overlaps that we consider accurate.

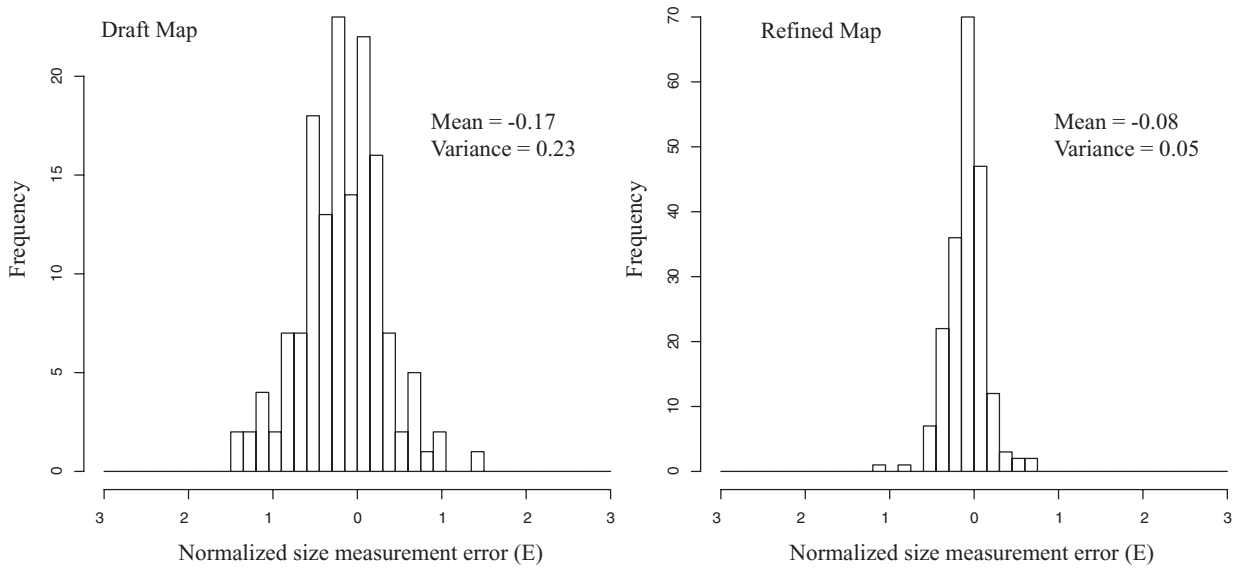

Fig. 4. Normalized size error before and after the refinement. Both graphs show histograms of normalized size error $E = (X - Y)/\sqrt{Y}$, where X are sizes of fragments on consensus map and Y are their underlying counterparts of the reference map. Note the diminished size dispersion after the refinement.

Contigs were formed from optical maps using pairwise overlap relations and extended until they covered the whole genome. Draft consensus map was produced by progressively merging overlapping maps.

This draft consensus map was then refined using our HMM-based method. To evaluate the number of errors present in the consensus map before and after the refinement, draft and refined consensus maps were compared with the reference map produced from the DNA sequence. Our draft consensus map contained 30 missing cuts and 12 false cuts. After the refinement, our consensus map contained only one missing cut, no false cuts and six missing fragments <2 Kb. The sizing difference was also reduced significantly (Fig. 4). Compared with the draft consensus map, fragment size variance was reduced by the factor of 5, indicating more accurate fragment size estimation. The refined consensus map was calculated based on 174 optical maps with accurate alignments. Site quality scores after the refinement are shown in Figure 5.

To evaluate the accuracy of our method for larger genomic assemblies with significant sizing inaccuracies, we performed a simulation. We took a 16 Mb reference map from human chromosome 1 and generated 500 optical maps using the statistical models associated with errors in optical maps. In particular, for those simulated optical maps, some cuts were deleted (20%), some random breaks were added (5 per 1 Mb) and sizing errors were introduced to all optical map fragments according to our size error model. These 500 optical maps contributed to an average coverage of 12 maps at each locus of the reference map. Draft consensus map of this 16 Mb region was produced by adding a large number of false cuts, missing cuts and size fragment differences. The initial map contained about 1200 restriction sites. Of these, we removed 87 sites, added 75 extra sites and changed sizes of 39 fragments in a random fashion. We then applied our refinement method to correct the errors we introduced. After the refinement, only two false cuts and six missing cuts were present in the refined consensus. Further, sizing errors were improved significantly (Fig. 6). Further analysis indicates that the method is most sensitive to significant changes of fragment sizes. In another simulation when a significant portions of fragments were removed or added (>30 Kb indel sizes), two regions failed to recover and were therefore considered to be breakdown regions.

Although breakdown regions were never observed for any actual optical assemblies that we have refined, the scenario is still plausible during assemblies of large mammalian-sized genomes. The reason the refinement failed for two regions during our simulation is that those regions were so inaccurate that optical maps were impossible to confidently align. Hence the necessary data that we needed to refine were simply missing. To overcome this limitation and recover regions with significant size inaccuracies (>30 Kb), the method can be further improved by employing alignment methods specifically designed to account for indels, or simply by recomputing breakdown regions using local map re-assembly.

We included the evaluation of run times for our refinement method using several synthetic datasets (Table 1). In particular, for three different human genomic contigs (1.9, 12.3 and 30.1 Mb), we have synthetically generated optical map coverage datasets to represent 20×, 100× and 500× copies of those regions. Inaccuracies were added to contigs to simulate draft maps and 15 iterations of refinement were applied. The running times are summarized in Table 1. Another question is how to choose a convergence criteria for stopping the refinement. Here we can select among the following possibilities: equilibrium for the number of accurately aligned maps, equilibrium for the alignment score per unit amount of consensus map and equilibrium state for the number of sites in the consensus map (Table 2). Our experience shows that all of these measures reach an equilibrium state at 13–15 iterations, so that can be chosen as a stopping criteria if desired.

8 CONCLUSION

The surface modalities used in optical mapping do not reliably retain restriction fragments below 1 Kb, consequently reducing

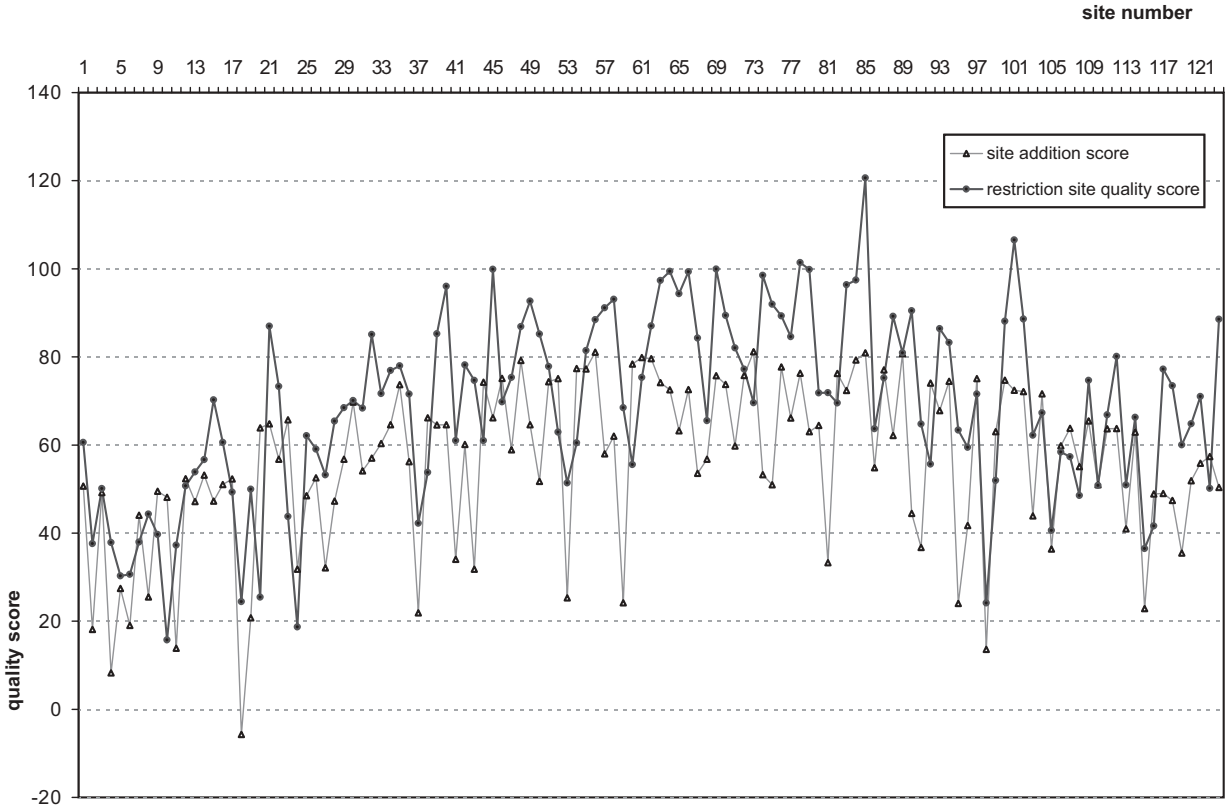

Fig. 5. Two site quality scores. Scores for the 122 cut sites after the refinement of *Y.pestis* assembly. Darker curve gives a quality score of the cuts of the consensus, lighter curve gives quality score for addition of cuts between adjacent consensus sites.

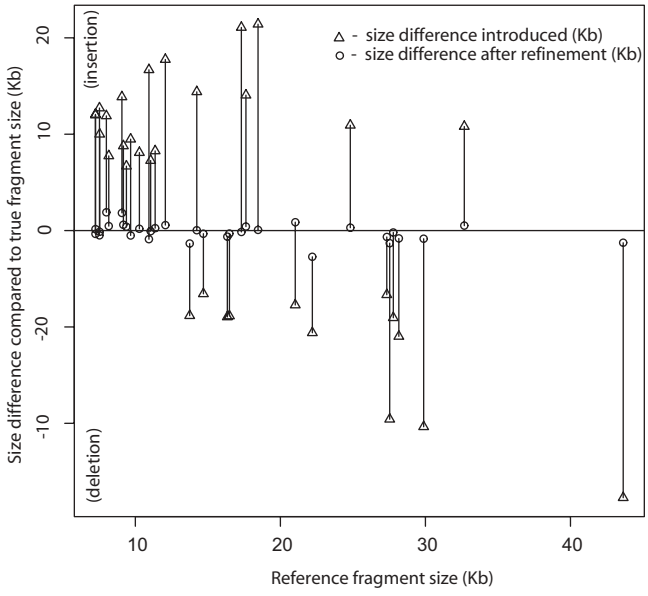

Fig. 6. Size errors before and after refinement. The horizontal axis marks sizes of fragments (in Kb) where indels were made. Vertical axis marks the sizes of indels made (also in Kb). Triangles mark the indels before the refinement, circles mark size discrepancies (compared with reference map) after the refinement.

Table 1. Refinement running time for three datasets at 20×, 100× and 500× optical map coverage

Reference region size (Mb)	Opt. maps	Coverage	Runing time on 3.4 Ghz CPU (s)
1.94	66	20×	67
1.94	316	100×	344
1.94	1577	500×	1894
12.33	387	20×	3431
12.33	1964	100×	17052
12.33	9787	500×	72949
30.11	953	20×	28993
30.11	4746	100×	124194
30.11	23819	500×	607389

their presence in any map dataset and attenuating overall map quality. Generally speaking, the quality of the consensus map at any given region depends the number of optical maps representing this region. More precisely, deep regions (represented by many optical maps) can be accurately corrected by reducing the overall number of false cuts and missing cuts, and increasing precision of restriction fragment size estimates. Shallow regions (covered by less than four optical maps), on the other hand, cannot be corrected unambiguously and hence are never modified.

Table 2. Convergence of EM algorithm for 1964 maps and 12.3 Mb reference region (100× coverage)

Iterations	1	2	3	4	5	6	7	8	9	13	14	15
Maps aligned	1312	1356	1607	1816	1816	1822	1720	1721	1740	1842	1844	1844
Average score (per 10 Kb)	0.93	0.938	0.848	1.27	1.27	1.275	1.422	1.442	1.247	1.136	1.137	1.137
Aligned map mass (Mb)	842	870	1022	1146	1146	1150	1089	1089	1102	1163	1164	1164

To summarize, our refinement method enables the accurate approximation of genomic restriction map through analysis of multiple optical maps representing this genome. In this context, consensus map quality scores can be used to evaluate the accuracy of the finished map and target low quality regions for reassembly.

There are other important applications of our method that go beyond the realm of *de novo* map construction. Structural alterations in the human genome are now being appreciated as forms of variation that complement SNPs. These events can be characterized at the map level using the method we have presented. More precisely, if a reference map, such as human Build 35, is taken as an approximation of the tested genome, refinement would create consensus maps by accurately mapping structural differences as apparent missing cuts, extra cuts or indels. Such developments would advance discovery of additional human structural differences and extend this analysis to populations.

The code of our refinement software is available for non-commercial purposes only from www.cmb.usc.edu/people/valouev.

ACKNOWLEDGEMENTS

This work is supported by grants NIH/CEGS P50 HG002790, NIH/NHGRI R01 HG000225, NSF DMR-0425880, NSF DBI-0501818, NSF EIA-0320708. A.V. was partially supported by the Preuss foundation fellowship at the time of this project. The authors want to thank Lei Li for inspiring conversations, John Nguyen and John McCrow for critical reading, and the anonymous reviewers for excellent comments.

Conflict of Interest: none declared.

REFERENCES

Ananthraman,T. (1999) Genomics via optical mapping III: contiging genomic DNA and variations. In *Proceedings 1th International Conference on Intelligent Systems for Molecular Biology*.

Churchill,G.A. and Wateman,M.S. (1992) The accuracy of DNA sequences: estimating sequence quality. *Genomics*, **14**, 89–98.

Dimalanta,E.T. *et al*. (2004) A microfluidic system for large DNA molecule arrays. *Anal. Chem*., **76**, 5293–5301.

Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res*., **8**, 186–194.

Lee,J.K., Dancik,V. and Waterman,M.S. (1998) Estimation for restriction sites observed by optical mapping using reversible jump Markov chain Monte Carlo. *J Comput Biol*., **5**, 505–15.

Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

Valouev,A. *et al*. (2005) Alignment of Optical Maps. In *Proceedings of the Annual International Conference on Computational Molecular Biology (RECOMB 2005)*, Cambridge, MA.

Waterman,M. (1995) *Introduction to Computational Biology*. Chapman and Hall/CRC, London.

Zhou,S. *et al*. (2002) A whole-genome shotgun optimal map of *Yerisinia pestis* strain KIM. *Appl. Environ. Microbiol*., **68** (12), 6321–31.

Zhou,S. *et al*. (2004) A single molecule approach to bacterial genomic comparisons via optical mapping. *J. Bacteriol*, **186**, 7773–7782.