SHE-RA
SHortread Error-Reducing Aligner
Preliminary Documentation for v1.0

Constructing composite sequences from overlapping mate-reads.
In order to align overlapping mate reads and produce accurate composite fragments, we developed the SHERA pacakge, a set of perl modules and scripts. The algorithm was designed to be broadly applicable to mated reads, independent of the sequencing technology. The source code is freely available at http://almlab.mit.edu/ALM/Software/Software.html. In brief, SHERA takes overlapping short reads as input, finds the best alignment, produces a composite read, calculates new quality scores for the overlapping bases, and scores the alignment confidence. A detailed description of each of these steps is provided below. We also describe our procedure for evaluating false alignment rates on both real and simulated reads.

Finding the best alignment.
For each mate pair, the algorithm scores all possible ungapped alignments. (This includes alignments longer than the read length, which are produced when a fragment is fully sequenced resulting in and additional synthesis reactions proceed past its 5-prime end, base-pairing with the Illumina adapters.) Each alignment is scored by summing over all matches and subtracting a penalty for each mismatch. For the alignment with the best score, the algorithm constructs the consensus sequence in this overlapping region by reporting the nucleotide with the higher quality value, or the informative base call if one nucleotide is 'N'. A Phred quality score is calculated for each consensus base (see below). The algorithm output includes sequence and quality of the composite read generated from each mate pair, as well as an alignment confidence score that enables filtering out probable mis-alignments.

Alignment confidence metric.
Alignment confidence was quantified in order to identify composite reads constructed from mis-alignments. Due to the relatively short alignment lengths with non-uniform base composition and quality, we did not model the statistical significance of an alignment analytically, but instead chose an empirical metric. Because mis-alignments of similar length have similar alignment scores (within noise) we use seven alignments of similar length of the same mate-pair to calculate a baseline, and divide by the range of the scores to estimate the alignment's significance. Alignments of all confidence levels are reported; we leave selection of this threshold to the user to achieve the sensitivity/specificity required of their downstream application.

Calculating a Phred Quality Score in the overlapped region.
A reports a Phred-scaled error probability for each base in the composite read. For overlapped bases, this is the posterior probability that the consensus base is wrong, given the base calls at that position in the alignment and their respective quality scores. Calculating this probability required making a few assumptions and approximations:
1. Given a basecall was in error, there is a uniform probability of reporting one of the other three bases.
2. The probability of correct alignment was approximated to be 1. (A decent approximation given true-positive rates of >99% in our quality filtered alignments, see Counting False Alignments below).
3. The basecalls of the forward and reverse read are independent observations, given the true base in the insert.
4. The prior probability of a base being present in an insert in this flowcell can be estimated from the low-error portion of the reads or approximated as uniform at the user's discretion (In our case, we used empirical observations to estimate %GC in single-genome samples, and a uniform distribution for the metagenomic sample.)
We first confirmed that the posterior error probability for the overlapped region was correct by counting the true error rate in composite reads constructed from simulated reads. Next, using real mate-reads as input, we confirmed that our Phred quality scores up to 32 correspond to the correct error rate as calculated by comparing to the reference. Above 32, the error rate does not improve (for neither the overlapped bases nor the original Illumina bases). This is due to errors in the reference and/or a very small percentage of Illumina reads with indels.

Counting False Alignments.
Minimizing the number of false alignments is an important step for some applications such as *de novo* assembly. Our goal was to find a threshold for alignment confidence that would allow no more than 1% misalignments to remain in our subset of composite reads (based on previous experience with *de novo* assembly of libraries containing hybrid reads). We used several approaches to assess our alignment accuracy and were pleased to find similar results.

First, we used the read-simulator from the MAQ package {{Li 2008}} to produce simulated reads that mimicked the error profiles observed in the quality values of our Illumina data as well as the distribution of overlap lengths. In the case of simulated reads, the true distance between mate pairs is known exactly and the reference is perfect. After filtering (confidence metric >= 0.5) we retain 89.7% of our composite reads, 0.5% of which were constructed from mis-alignments. This strategy is useful when no reference is available, or as a test case on new platforms.

Second, we used mate-reads from the control lane sequenced on our Illumina flowcell. Illumina provides a DNA library prepared from the PhiX174 bacteriophage genome for purposes of quality control, and we found that it had insert length 200bp+/-21 (estimated by mapping to the reference with MAQ). These mapped mate-pairs defined the "true" insert length of each sequenced fragment. After constructing the composite sequences and filtering, we plotted the difference in length between a composite fragment and its insert length as predicted by MAQ (Supplemental Figure 5). This histogram is a sum of two distributions, the overlapper software's misalignments (a broad gaussian) and a sharp peak of small (1-2bp) indels. We examined gapped alignments by eye to confirm this was the case and also found that many of the indels occur in homopolymer runs. We used a simple linear model to infer the number of misalignments; this yielded a false positive rate of 1.0% for PhiX174.

Finally, for the figures reported in the main text, we assessed alignment yield and accuracy on mate-pair reads sequenced from a dSPRI library. This library was generated from a single Prochlorococcus cell, for which a draft copy of the genome (N50=~7Kb) was constructed independently by de novo assembly of 454-Titanium reads from the same DNA sample. We mapped the Illumina mate-paired reads to these 454-Titanium contigs to define the "true" insert length of each sequenced insert. A small fraction of mate-pairs (<5%) were unmapped or mapped to different contigs; these were excluded from further analysis. As in the case of PhiX174, there were a number of small indels evident in the comparison between reads and reference which were excluded from our statistic. We estimate that 0.5% of our high confidence set of composite reads were mis-alignments introduced by SHERA.

SHE-RA was designed and written by Sonia Timberlake.
Paper in submission.
If you use this software, please cite this webpage.
For more information, contact soniat@mit.edu