1 Results

1.1 Coverage-normalized data can be used to locate errors in high-coverage shotgun sequencing data

K-mer spectral analysis is a powerful approach to error detection and correction in shotgun sequencing data that uses k-mer abundances to determine likely errors (cite Pevzner, Quake, khmer-counting paper). While approaches derived from spectral analysis can be very accurate – Zhang et al. (2014) suggest that spectral analysis is considerably more effective at finding errors than quality-based approaches – it is also very compute intensive: most implementations count k-mers across entire sequencing data sets, which can be memory- or I/O-intensive for large data sets.

Brown et al. (2012) introduced a single-pass algorithm for normalizing k-mer abundance spectra, termed "digital normalization" (abbreviated as "diginorm"). This procedure estimates the k-mer coverage of each read in an online algorithm, by calculating the median k-mer abundance of the read given all previous reads; reads above a certain estimated coverage are set aside and their k-mers are not tracked. This algorithm is both online and *streaming* because it only collects k-mers in reads with a low estimated coverage; for the same reason, it is sublinear in memory for high coverage data sets. The net effect of diginorm is to reduce the data set size that must be considered for downstream processing, such as *de novo* assembly (cite trinity, elijah, etc.)

To determine if digital normalization could be applied prior to k-mer spectral error detection, we first generated a synthetic data set from a simulated low-complexity genome ("simple genome"; see Methods for generation and Table XXX for data set details). We then applied digital normalization to a median 20-mer coverage of 20 (k=20, C=20) and used k-mer counts from the downsampled read set to detect errors by looking for bases at the beginning or ends of low-abundance regions in each read; we used a "trusted k-mer" cutoff of $C_0 = 3$ as our abundance cutoff, below which we assumed k-mers were erroneous.

Of the 531 reads in the simple genome with one or more errors, predicted errors matched exactly for 317 of them (true positives), and 466 reads were correctly predicted to contain no errors (true negatives). 90 reads were falsely predicted to have no errors (false negatives). The errors in 124 reads were miscalled – while the reads each had one or more errors, the positions were not correctly called – and three reads did not have errors but were predicted to have errors, leading to a total of 127 false positives. Using the above definitions, we calculated the prediction sensitivity to be 77.9the prediction specificity to be 71.4%.

@@put in picture of k-mer abundance spectrum before and after?

@list in a table: matched exactly, etc. rundown.

@next, E. coli?

1.2 Coverage-normalized data can be used to locate errors in variable coverage shotgun sequencing data

One of the drawbacks of spectral abundance analysis is that it cannot be applied to metagenomic or transcriptomic data sets, which frequently contain reads from both high-abundance and low-abundance molecules. This variability in coverage confounds naive spectral analysis for two reasons: first, errors in very high abundance regions can accumulate and increase over the threshold for trusted k-mers, thus appearing to be correct; and second, correct reads from low coverage regions yield k-mers below the trusted k-mer threshold and appear to be incorrect. In practice, therefore, error correction for metagenomic and transcriptome data uses more complex approaches than a single threshold.

If we apply diginorm to variable coverage data, we address the problem of accumulated errors from high-abundance regions by eliminating the majority of the reads from these regions; this is a form of error correction, detailed in the original diginorm paper (cite diginorm). However, we can also address the issue of low-abundance regions by using the median k-mer abundance to estimate which reads are too low-abundance to analyze; in effect, we can ignore reads that are outside of a desired coverage rage (Figure XXX).

@graph with read abundance spectrum ,showing which reads we will call errors in.

We generated two more synthetic data sets, "simple metagenome" and "simple mRNAseq," which contain both high- and low-abundance species (see Table XXX for data set details). After generating synthetic reads with a 1% error rate and applying digital normalization to k=20/C=20, we again applied our spectral error detection approach using the normalized counts, but with the modification that only reads with a median k-mer abundance of 20 or greater were examined. For the simple metagenome data set, 2254 of 2347 reads (96.0%) met the coverage criterion. Of the 2347 reads total, the errors in 658 erroneous reads were called perfectly (TP) and 1126 of the reads with no errors were correctly called as error-free (TN). 297 reads were incorrectly determined to be error free (FN; including the reads that were too low coverage to be considered). Of the remaining 266 reads, 261 were miscalled (errors existed but were not exactly called) and 5 were incorrectly called as erroneous when they were in fact correct. We calculated the prediction sensitivity to be 68.9% and the prediction specificity to be 71.2%. For the simple mRNAseq data set, 524 of 568 reads (92.3%) met the coverage criterion, with 153 true positives, 255 true negatives, 87 false negatives (including the low coverage reads), and 73 false positives, for a prediction sensitivity of 63.8% and prediction specificity of 67.7%.