# An improved method for hybrid correction of long-read, low-identity sequencing data

**James Gurtowski[1], Hayan Lee[1] and Michael C. Schatz[1]**

[1] Simons Center for Quantitative Biology, Cold Spring Harbor Lab, NY, USA

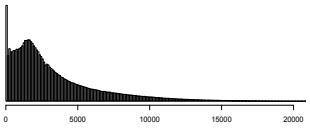
http://github.com/jgurtowski/ectools

## Abstract

Long-read technologies, such as Pacific Bioscience's SMRT sequencing, have greatly improved shotgun genome assemblies. Although methods now exist for non-hybrid correction of high error rate sequencing data (HGAP), these approaches have been limited to small microbial genomes where >50x coverage can be obtained relatively inexpensively. For larger Eukaryotic genomes, a hybrid correction approach using high-identity, low-cost libraries can still be more cost-effective. Our approach improves upon existing hybrid correction approaches in many cases improving final assembly results by as much as two to three fold according to the n50 metric. Our assembly of rice improved from an initial N50 contig size of 50kb, using existing correction techniques, to beyond 150kb using the new approach. This encouraging result lends to the future possibility of single contig chromosomes of eukaryotic genomes.

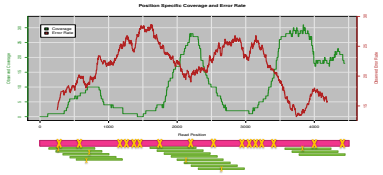
## Background

Current long read technologies differ from short read technologies in that they sequence a single molecule, rather than a clonal population of DNA fragments. The current leading long read technology from Pacific Biosciences, the RS, records the incorporation of nucleotides by a stationary polymerase in real time. The resulting read lengths are thus not uniform; instead tending toward an exponential distribution as seen below.



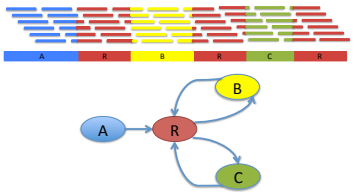

This is one of the few technologies that can produce reads greater than 20kb, but at the cost of reduced per-base accuracy. Currently, no assembler is designed to handle reads with an error rate as high as those produced by the RS making error correction very important for de novo assembly of this data.

Current hybrid error correction techniques align short high-identity reads to low-identity long reads and compute a consensus sequence. This approach is successful if the long-read error rate remains below a certain threshold. However, in certain circumstances, portions of reads may have more errors making the alignment of short reads to these regions difficult.




Above, a striking negative correlation appears between the increased per-base error rate and short read coverage. The alignment algorithm has difficulty placing reads in regions of high error rate. Regions that do not have short-read coverage will either remain uncorrected or cause the entire read to be split by the correction algorithm. When a read is split, vital information is lost and the ability to span repeats is severely diminished.

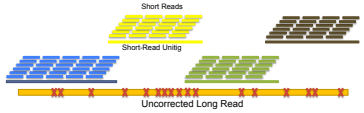
Below is an illustration of how repeats make assembly difficult. If repeats are longer than the read length, it is not possible to conclusively determine the ordering of unique contigs (unitigs).



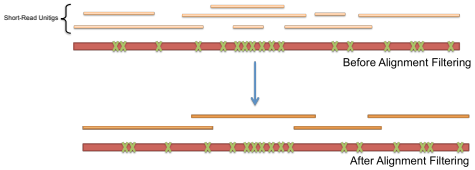

The above assembly graph is ambiguous because it is unclear whether R should be followed by B or C and how many times these contigs occur in succession.

## Methods

In all aspects of assembly long reads are preferred over short reads because they contain more contiguous information about the nucleotide sequence. Error correction is no exception. Rather than align raw short reads to a long-read backbone, the short reads are first preassembled into unitigs. These unitigs give more contiguous information and can span regions of high error rate. In order to align unitigs a very sensitive aligner is needed. Nucmer from the Mummer suite was chosen for its sensitivity and flexibility.

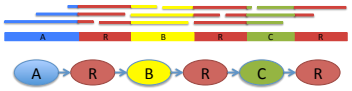



Above, unitigs composed of short reads are aligned to a raw long read. Because the short-read unitigs are the assembler's attempt at creating a unique representation of the genome, it is no longer appropriate to try to make a consensus sequence. Instead, we want to find the layout of the unitigs along the long read scaffold that maximizes long read coverage while minimizing short-read unitig overlap. This functionality is implemented in the delta-filter program which is also a part of the Mummer suite of tools. The functionality of delta-filter is depicted below where a large set of alignments are filtered down to a candidate set of unitigs that most likely come from the same region as the long read.




Using a variation of the longest increasing subsequence dynamic programming algorithm, delta-filter is able to filter unitig alignments to find the subset that are most likely to be from the same region as the long read. Once the set of unitigs is determined, the show-snps program is used to determine the difference between the unitig set and the raw long read. A simple python script uses this information to "correct" the long read, incorporating all of the information provided by the scaffold of unitigs.

Once the layout is determined and bases are corrected, and a conservative trimming algorithm removes regions of the long reads that did not have short-read unitig alignments. Depending on the context, the user may want to choose between shorter, well corrected reads and longer, lower identity reads. The user can specify a minimum identity and the splitting algorithm will make an effort to only output reads with a per-base identity greater than this number. It assumes all bases with a short-read unitig alignment can be corrected to within 99% identity while regions without coverage remain at the uncorrected 85% identity. Taking the average identity along the read and recursively splitting out sections without coverage eventually produces subreads with an identity that meets the user's requirements.




Once corrected, the long reads help to span repeats and conclusively determine the order of contigs in the assembly graph. The twisted assembly graph seen with shorter reads becomes a straight forward linear representation of the genomic sequence. It is for this reason that longer reads are so coveted, because they contain more information about the relative position of contiguous sequences.

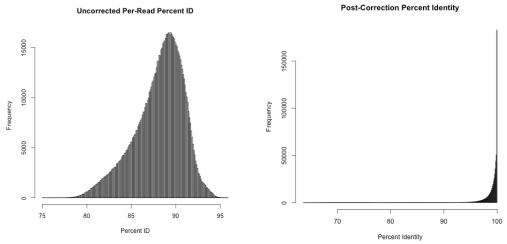
References:

Hybrid error correction and de novo assembly of single-molecule sequencing reads. Koren,S, Schatz, MC, et all. (2012) Nature Biotechnology. Doi:10.1038/nbt.2280

Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Chin, CS et al. (2013) Nature Methods. 10:563-569

Fast algorithms for large-scale genome alignment and comparison.
Delcher AL, Phillippy A, Carlton J, Salzberg SL, Nucleic Acids Res. 2002 Jun 1;30(11):2478-83.

## Results

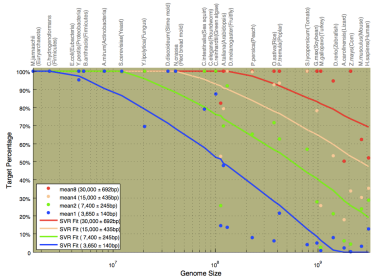



Error correction significantly improves the per-base identity. Most reads are corrected to within 1% error rate. These plots were made by aligning pre and post correction reads to the rice reference genome.

| Assembly | Contig NG50 |
|---|---|
| HiSeq Fragments<br>50x 2x100bp @ 180 | 3,925 |
| MiSeq Fragments<br>23x 459bp<br>8x 2x251bp @ 450 | 6,332 |
| "ALLPATHS-recipe"<br>50x 2x100bp @ 180<br>38x 2x50bp @ 2100<br>51x 2x50bp @ 4800 | 18,248 |
| PBeCR Reads<br>7x @ 3500 ** MiSeq for correction | 50,995 |
| Enchanced PBeCR<br>15x @ 3500 ** MiSeq for correction | 155,695 |

** Minimum read length

The table above shows the results of various rice assemblies. Assemblies using the long read data produced the best results. Our enhanced error correction pipeline was able to boost the n50 metric by nearly 3 fold over the correction with the existing pacbioToCA pipeline bundled with the Celera assembler.




Sequencing technology is improving rapidly. Over the past few years read length from the leading single molecule sequencing technology has increased at an exponential rate. Using simulation we are able to see how this exponential increase in throughput and read length will affect de novo assembly in the near future. The above graph has a selection of organisms from across the tree of life for which a reference genome exists. Various assemblies with different read lengths were conducted at 20x coverage to find how well different genomes assemble. The blue line shows what is possible with today's technology. The successive colors, green, peach and red are the predictions for next year, the year after and so on. This chart can be a great tool for planning future assembly projects.