Integrated heteroduplex correction in PacBio's circular consensus algorithm

Derek Barnett[1], John Harting[1], Walter Lee[1], Armin Töpfer[1], Fritz Sedlazeck[2,3], Jenny Ekholm[1], Nina Gonzaludo[1], Justin Blethrow[1], James Drake[1], Zev Kronenberg[1]

[1]PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025, USA
[2]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, 77030, USA
[3]Department of Computer Science, Rice University, 6100 Main Street, Houston, TX, 77005,USA

Background and objective

A heteroduplex (HD) (**Fig. 1**) is a double-stranded sequence comprised of two non-complementary strands that can form during PCR. These mixed-template artifacts produce misleading results in downstream analysis, e.g., false haplotypes during diplotyping. Unlike short-read technologies, PacBio Single-Molecule Real-Time sequencing produces strand-level base calls. Heteroduplex signatures can be directly observed and corrected using the stranded sub-read data. Our new method is integrated in the circular consensus sequence algorithm which generates accurate HiFi data from sub-reads.

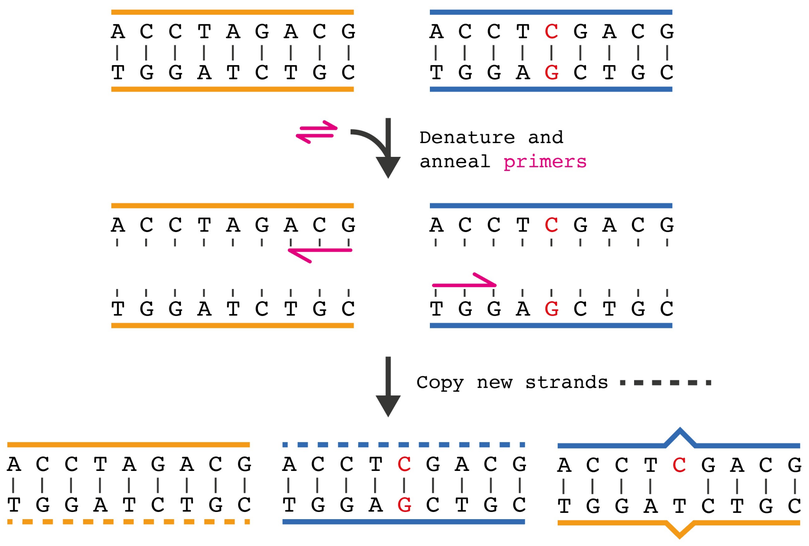

Fig. 1. The formation of a heteroduplex during PCR. Red text highlights the mismatch between DNA strands.

Methods

The transformation of PacBio subreads into high accuracy HiFi reads is done by the circular consensus sequence (CCS) algorithm (**Fig. 2**). During CCS, an intermediate draft sequence is generated, and subreads are mapped and aligned to the draft. The heteroduplex algorithm (hd-finder) takes the subread alignments and generates a read pileup whereby variants are identified. At each site, the bases are sorted and counted by strand. The 2x2 count data is subjected to a Fisher's exact test. The fraction of significant sites across the draft is used to determine if a read contains heteroduplex. Heteroduplex flagged reads are split by strand and reprocessed resulting in two HiFi reads, one for each strand.

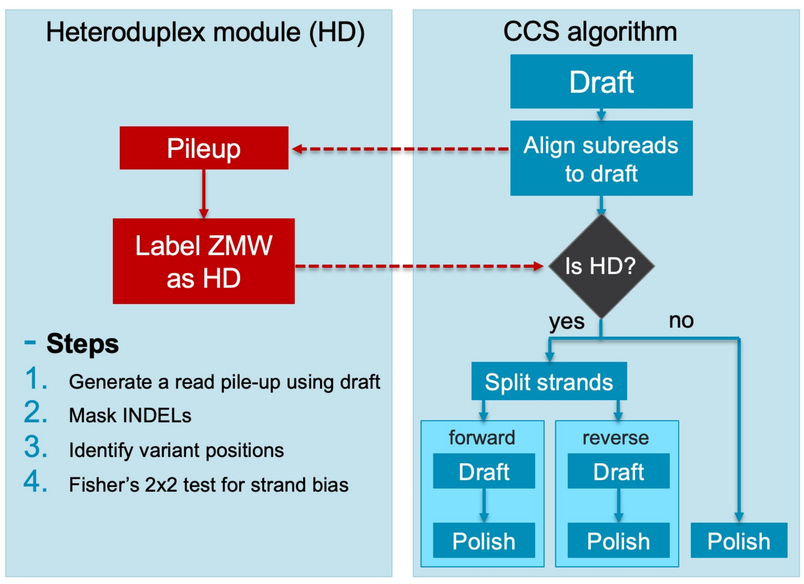

Fig. 2. Heteroduplex CCS workflow

Heteroduplex enriched library

A heteroduplex (HD) enriched library was designed to provide a training dataset. The library contains ~90% heteroduplex molecules.

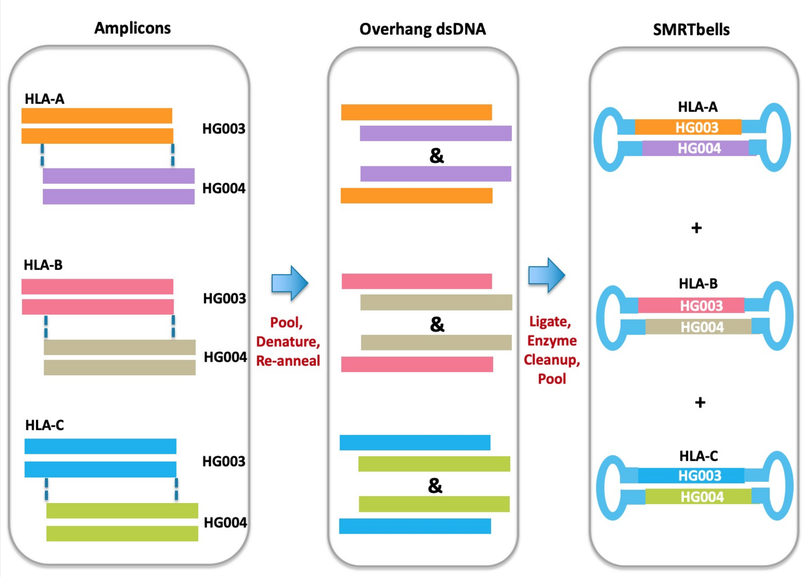

Figure 3. Heteroduplex enrichment library preparation. HG003 and HG004 genomic DNA was extracted (Lucigen Masterpure) from cell pellets purchased from Coriell Repositories. PCR primers specific to HLA-A, HLA-B, and HLA-C were designed (Integrated DNA Technologies). For a given locus, a primer pair was used for PCR on one genomic source, and a second primer pair with a four base offset was used for the second genomic source. The resulting amplification products for a given locus were pooled, heated, and reannealed. The annealed products that produced a four-base overhang on both sides were ligated to corresponding adapter sequences. Exonuclease treatment removed amplicons that did not contain a ligated adapter on both ends, resulting in an enrichment of SMRTbell templates that contain a single strand from each of the two genomic sources.

Visualizing heteroduplex

- Heteroduplex shown for amplified *HLA-A* data (**Fig. 4**)
- In heteroduplex HiFi data (**Fig. 4A**) Single Nucleotide Variants (SNVs) are shadowed with INDELs
- In heteroduplex subread data (**Fig. 4B**) extreme strand bias can be seen at SNV sites

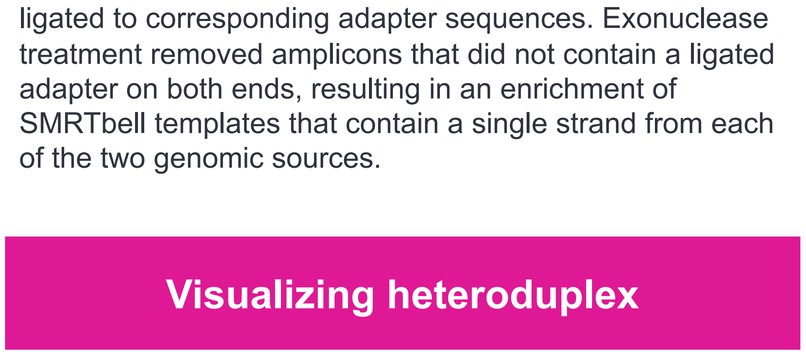

Fig. 4. Heteroduplex shown in HiFi and subread data. **(A)** HiFi data aligned to HLA-A. Note the non-random Association of SNVs and INDELS. **(B)** Subread data Aligned to HLA-A, sorted by strand. Note that the alleles strongly associate with strand.

Results

We demonstrate the accuracy of the hd-finder algorithm is ~94% (**Table 1**) by using a heteroduplex enriched amplicon library. We also show that applying the hd-finder to amplified datasets improves the quality of downstream analysis of important human genes (**Fig. 5**).

TP ZMWs	TN ZMWs	FP ZMWs	FN ZMWs	Sensitivity	Specificity
881	948	54	52	0.94	0.95

Table 1. Accuracy of the HD detection method. True positive reads came from the special HD library and manually curated to a list of 881 reliable reads. The true negative dataset came from a genomic no-amp library.

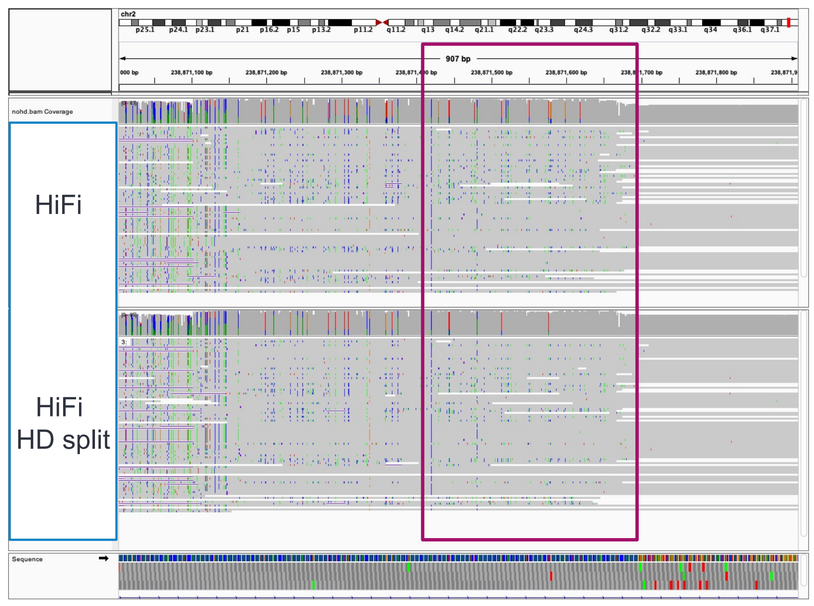

Figure 5. Variant calling differences in the *TWIST2* gene before/after using the hd-finder algorithm. *TWIST2* is associated with two rare genetic diseases. The region above is GC rich which likely increases the chances of HD formation. The top panel is HiFi data and the bottom panel shows HiFi data after the hd-finder algorithm was applied. The box highlights allele frequency differences after running hd-finder.

Conclusion

- Heteroduplexes are common amplification artifacts, independent of sequencing technology
- PacBio sequencing can identify and resolve heteroduplex on the fly.
- The heteroduplex algorithm is a powerful new method for improving HiFi amplicon targets.
- Learn more: https://ccs.how/faq/mode-heteroduplex-filtering.html

Acknowledgements

The authors would like to thank Jason Underwood for library design suggestions, and PacBio users who provide useful feedback.