A Scalable Liquid Biopsy Pipeline Using Duplex Sequencing

MARK FLEHARTY, MADELEINE DURAN, YOSSI FARJOUN, TAKUTO SATO, DAVID BENJAMIN, CARRIE CIBULSKIS, BRENDAN BLUMENSTIEL, MATTHEW DEFELICE, JUSTIN RHOADES, MICHELLE CIPICCHIO, JUSTIN ABREAU, KATIE LARKIN, VIKTOR ADALSTEINSSON, STACEY GABRIEL, NIALL LENNON

Broad Institute of MIT and Harvard, 415 Main St Cambridge, MA 02142

Abstract

Accurate detection of somatic events using liquid biopsies has the potential to revolutionize precision medicine in cancer. Given the ease of access to patient specimens and the ability to characterize somatic variants repeatedly over time, we aim to offer an end-to-end pipeline to deliver high quality liquid biopsy results to support translational research. Effective processing of liquid biopsies requires a scalable pipeline that meets the following specifications:

- 90% Sensitivity at 1% allele fraction
- Less than 1 False Positive per Mega-base

Future applications presently in development aim for 90% sensitivity at 0.1% allele fraction. Using a custom data pipeline and a lab process that incorporates duplex unique molecular indices (UMI) we have called variants over our custom 396 gene pan-cancer panel with 2Mb of target territory and custom designed panels for glioma and multiple myeloma. This pipeline makes use of UMIs for increasing the available depth of reads and reduces base calling errors by utilizing duplex-consensus calling. We have benchmarked this technology using pooled sample analysis to simulate somatic variants from a tumor and normal-normal analysis as an independent measure of FPR. The pooled samples were spiked in at 5%, 2.5% and 1%. The normal-normal replicates were taken from the biological source material. Sensitivity for events at 1% observed allele fraction (1.0% spike in) exceeds 90% with FPR < 1/Mb.

Laboratory Methods

Cell free DNA (cfDNA) is extracted from fractionated blood plasma using the QIAsymphony platform (Qiagen). Illumina sequencing ready libraries are constructed using a Kapa Hyper Prep kit during which duplex UMI adapters (IDT) are incorporated. The libraries are enriched using custom hybrid selection baits (Twist Bioscience) and sequenced on the HiSeqX platform.

Informatics Methods

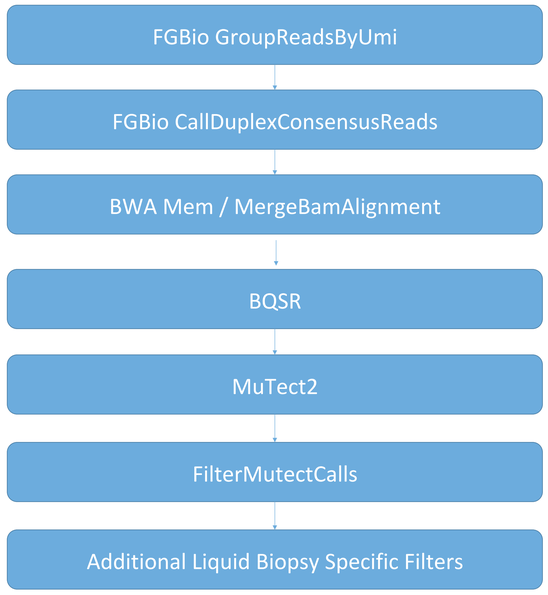

Figure 1, Overview of pipeline.

Duplex sequencing dramatically increases the quality of sequenced bases. Variants in non-duplex sequenced bams often have large numbers of alt bases at every loci with ultra deep (> 1000X) sequencing making variant calling difficult. The bases sequenced using duplex sequencing are much higher quality making variant calls much more reliable.

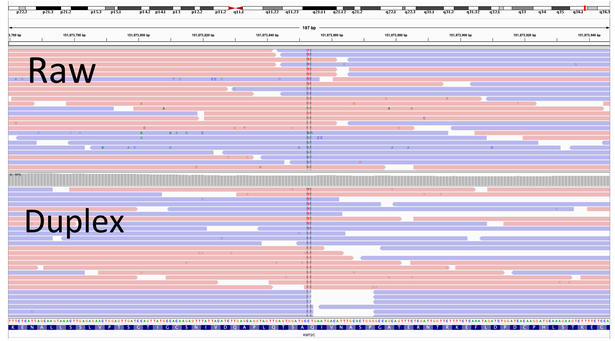

Figure 2, Side by side comparison of base calls from sequencer "Raw", and duplex consensus called reads, "Duplex". Total counts in raw; A:14, C:18934, G:12, T:113, N:1. Total counts in duplex; A: 0, C:1450, G:0, T:9, N: 12.

Filters Used in Duplex Consensus Pipeline

- Present In Normal

Observations indicate even a single alt read in the normal are a sign the variant is a false positive.

- Stricter Strand Bias

Observations indicate that support from an alt allele should come from both strands at least once.

- N ratio

A high ratio of Ns to alt reads indicates a likely false positive. Currently we filter variants whose N / alt ratio is greater than 4.

- Custom mapping filter

Alt reads are compared to reference using BLAST to ensure alt reads were not misaligned.

Sensitivity and False Positive Rate of Assay Over Different Panels

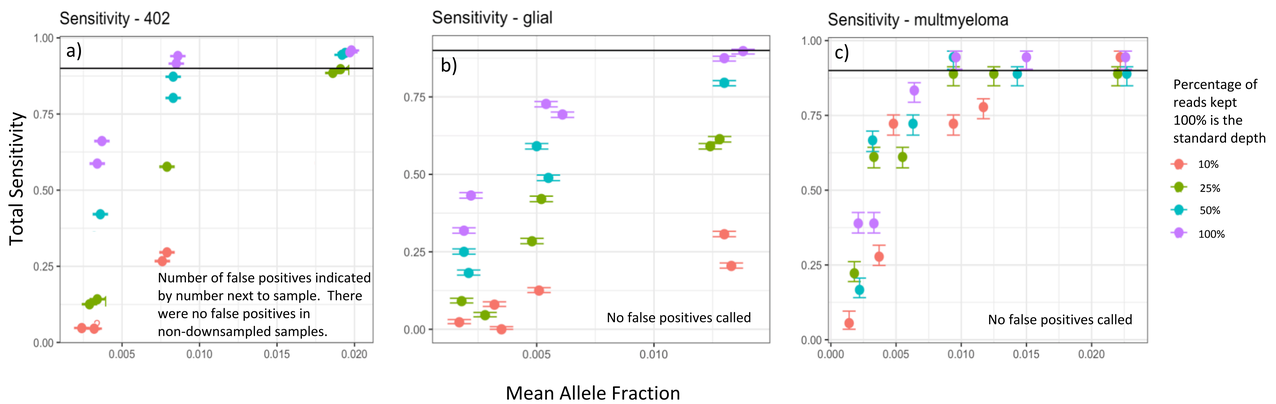

Figure 3, Sensitivity of assay at various depths and allele fractions. Heterozygous variants were used for sensitivity measurements, so the observed alt AF should be ½ of the spike in, differences are thought to be due to variances introduced by titrating small volumes of DNA. ***a*** *plot of sensitivities vs. allele fraction for the 396 gene pan-cancer panel.* ***b*** *glial panel,* ***c*** *multiply myeloma panel. 95% confidence intervals were determined using the Clopper-Pearson method. Downsampling is performed to simulate the effect of sequencing at lower depths.*

Current Development Work

Duplex consensus reads are formed by combining duplicate reads from the αβ and βα families of reads. The presence of only one group of these reads results in 'singletons'. These are reads that could contribute to library complexity, but are left out because we cannot find both strands.

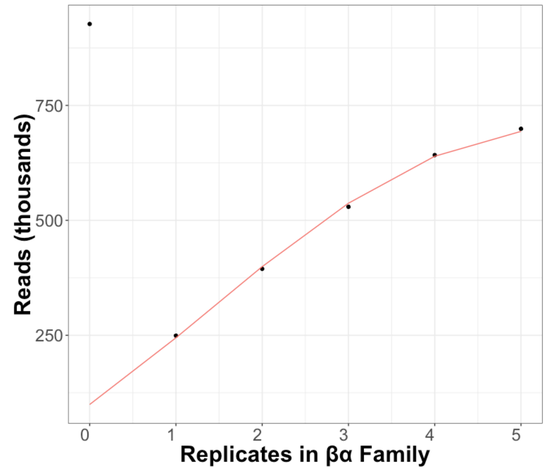

Figure 4, Reads by number of replicates in minority strand. We expect the number of duplicates in the minority strand to follow a beta-binomial model. The extreme outlier at x = 0 is due to the presence of a larger than expected number of 'singleton' reads.

We are presently investigating ways to improve the efficiency of duplex capture so that we can increase the sensitivity of the assay.

Conclusion

We have developed an end-to-end pipeline for delivering results from a liquid biopsy assay using duplex sequencing from multiple custom panels. Our sensitivity is limited by library complexity which limits the achievable depth. Consensus calling is of critical importance for increasing the sensitivity of the assay because detection of variants at the allele fractions of interest requires higher base qualities than are typically reported by the Illumina platform.

Acknowledgments

Data used in this poster was generated at the Broad Institute, for more information please visit: http://genomics.broadinstitute.org/