Comprehensive, multi-omic detection of somatic variants from the GIAB HG008 matched tumor-normal pair

Abstract # 7827

Alex Sockell[1], Khi Pin Chua[1], Christine Lambert[1], Matt Boitano[1], Melanie Wescott[1], Ian J McLaughlin[1], Primo Baybayan[1], Jennifer McDaniel[2], Justin Zook[2], Aaron M Wenger[1]
[1]PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025, [2]Material Measurement Laboratory, National Institute of Standards and Technology, 100 Bureau Dr., Gaithersburg, MD 20899, USA

Introduction

Disentangling the molecular drivers of cancer progression requires a precise understanding of the somatic alterations that take place at the DNA level during tumor development. These include not only small changes like SNVs and indels, but also structural variants, changes in repetitive elements, differential methylation, as well as the haplotype context in which these changes occur. Short-read sequencing methods lack the read length to comprehensively characterize structural variants or to span repetitive regions such as microsatellites and phase mutations into haplotypes, and require separate library prep to characterize methylation. Conversely, other long-read sequencing methods lack the per-base accuracy to call smaller variants like SNVs, while methods like optical genome mapping lack the resolution to precisely determine structural variant breakpoints. Accurate long-read sequencing with PacBio HiFi addresses these challenges, enabling comprehensive and multi-omic profiling of all types of variation leading to cancer development.

Here we apply PacBio HiFi to perform whole-genome sequencing of the newly described HG008 matched tumor-normal pair from the Genome in a Bottle (GIAB) consortium.[1] This reference sample includes an adherent, epithelial-like pancreatic adenocarcinoma (PDAC) cell line as the tumor material, with the matched normal obtained from adjacent duodenal and pancreatic tissue. We perform whole-genome sequencing of the tumor cell line and matched pancreatic normal tissue with PacBio HiFi, resulting in a more robust and comprehensive picture of somatic variation in this reference sample and contributing to the development of this novel benchmark.

Methods

Genomic DNA was extracted by GIAB using the Qiagen Puregene Cell kit for the tumor cell line and the PuregeneTissue kit for the normal tissue. SMRTbell library prep was performed by PacBio with the SMRTbell Prep kip 3.0. Sequencing was performed on the PacBio Revio system to 80x coverage (tumor) and 35x coverage (normal). Mapping and variant calling was performed using the HiFi-somatic-WDL workflow, using tools which have been validated and optimized for long-read data (Figure 1). Repeat expansions were called using TRGT/TRVZ.

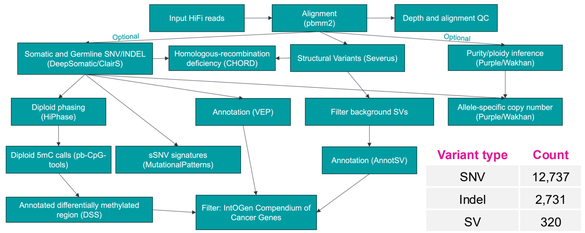

Figure 1. **Analysis workflow and variant count.** Schematic showing the workflow of the HiFi-somatic-WDL tool for alignment, QC, and variant calling results (inset) for different somatic variant types.

HG008 exhibits common pancreatic cancer CNVs

Copy number variants were called using Purple in the HiFi-somatic-WDL tool, which also estimates tumor purity and ploidy. Purity was very high (0.99), as expected for a tumor cell line. The estimated ploidy was 1.44, with chromosome-arm level deletions occurring on most chromosomes, leading to widespread loss of heterozygosity. A smaller number of amplifications were observed, including on chr17q which includes the HER2 locus, and on chr20q, which has been predicted to be an early tumorigenic event in pancreatic cancer.[2] We also observe subclonal deletion events on chr1q, chr2q, and both arms of chr10.

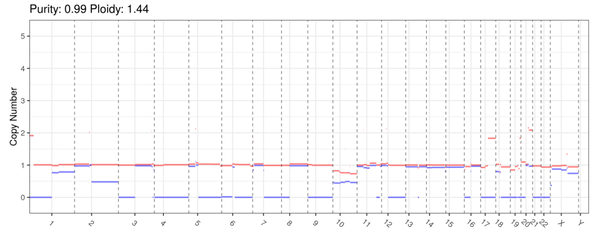

Figure 2. **Copy number profile.** Estimated copy number for each chromosome, with different haplotypes shown in red and blue.

HG008 displays SVs spanning key PDAC genes

A key benefit of long reads is the ability to call structural variants with high accuracy, and to precisely characterize structural variant breakpoints. HG008 was found to harbor 17 large structural variants involving one or more known cancer-causing genes in the IntOGen Cancer Gene Census (CGC), including genes commonly mutated in pancreatic cancer such as KRAS (88% of PDAC cases) and CDKN2A (18% of PDAC cases).[3]

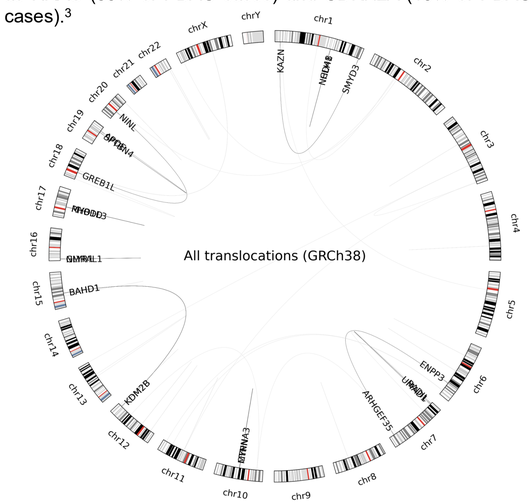

Figure 3. Circos plot of HG008 translocation events. Translocations are defined as all BND pairs that are at least 100 kb apart. Black lines indicate translocations involving known cancer-related genes.

Clonal deletion in CDKN2A tumor suppressor

Because the HG008 tumor cell line contains many chromosome-arm level deletions, a significant number of structural variants were located in regions that had experienced loss of heterozygosity. Some SVs were at or near 100% frequency in the tumor, indicating these were likely early truncal mutations that contributed to tumor development. For example, a 20 kb clonal deletion was found spanning the first 2 exons of the CDKN2A tumor suppressor, which is commonly lost in pancreatic cancer.[3]

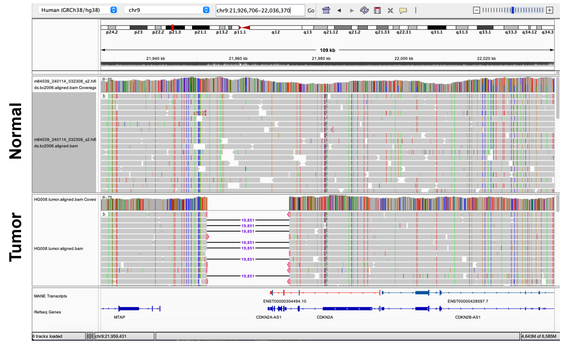

Figure 4. **Example CDKN2A deletion.** Tumor-specific clonal 20 kb deletion spanning multiple exons of the CDKN2A tumor suppressor gene.

Differential methylation of known cancer loci

HiFi sequencing enables direct detection of methylation by monitoring changes in polymerase kinetics caused by base modifications. HG008 exhibited tumor-specific methylation changes involving 58 known cancer related loci, including several genes known to be involved in pancreatic cancer. For example, the FAM135B gene, which was recently observed to be enriched in pancreatic cancer-derived extracellular vesicles, is hypomethylated in the tumor cell line relative to normal tissue, which is associated with both increased gene expression and increased chromatin accessibility.[4]

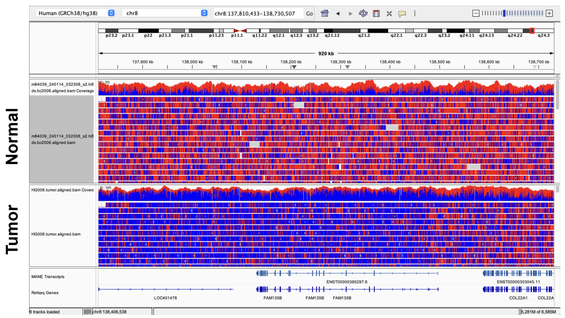

Figure 5. **Example of differential methylation.** IGV showing 5mC methylation status for the region surrounding the FAM135B locus. Red indicates hypermethylation, blue indicates hypomethylation.

Repeat contraction in NHEJ1 damage repair gene

A key advantage of HiFi sequencing is the ability to sequence through repetitive regions such as microsatellites which may harbor repeat expansions or contractions. The high per base accuracy enables precise mapping of intra-repeat variability, which can be easily visualized using the TRGT/TRVZ tool. We found that HG008 displayed a tumor-specific repeat contraction affecting the NHEJ1 gene, which is involved in the process of DNA repair via non-homologous end joining and has been implicated in several cancer types.[5] Because repeat expansions typically lead to reduced gene expression, the presence of a contraction in NHEJ1 in the tumor may result in higher expression of this gene, potentially shifting the DNA damage response in the tumor toward non-homologous end joining over homologous recombination.

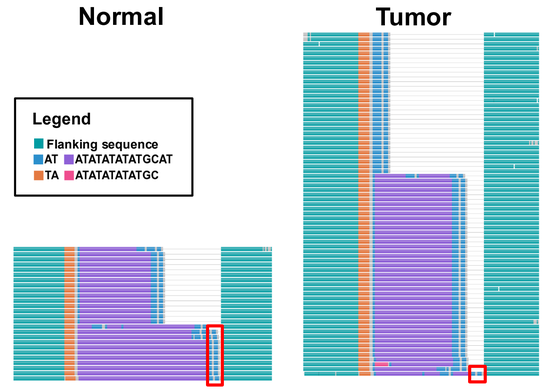

Figure 6. **Repeat expansion in NHEJ1.** TRVZ visualization of a candidate repeat expansion within the NHEJ1 gene. Flanking sequence is shown in green, with other colors indicating variable repeat motifs. Red box highlights presence of expansion in normal cells which is at much lower frequency in the tumor.

Conclusions and Next Steps

- HiFi sequencing enables comprehensive profiling of somatic events leading to tumor formation in HG008.
- Some alterations, including SVs, methylation changes, and repeat expansions, are only detectable using HiFi.
- Recapitulating results with Revio SPRQ chemistry will enable long-read somatic profiling with low DNA input.
- Multi-omic profiling with Kinnex and Fiber-seq may reveal mechanistic insights for these changes.

References

1. McDaniel J, et al. (2024). bioRxiv, 2024.09.18.613544.
2. Tabach Y, et al. (2011). PLoS One. 6(1), e14632.
3. Stefanoudakis D, et al. (2024) Curr. Issues Mol. Biol. 46(4), 2827-2844
4. Olou A, et al. (2025) Commun. Biol. 8, 638.