Improved detection of low frequency mutations in ovarian and endometrial cancers by utilizing a highly accurate sequencing platform

Abstract # 1873619
Timothée Revil[1], Nairi Pezeshkian[2], Dan Nasko[2], Lucy Gilbert[1], Alexandra Sockell[2], **Jiannis Ragoussis[1]**
[1]McGill University, Quebec, QC, Canada,[2]Pacific Biosciences, Menlo Park, CA

Introduction

Ovarian and endometrial cancers are the 4th highest (combined) cancer killer of Canadian women. In 2020, over 3000 women were diagnosed with an ovarian cancer, of which 75% were in the later stages. The goal of the DOvEEgene (Detecting Ovarian and Endometrial cancer Early using Genomics) project is to detect these cancers as early as the first stage through a low-cost, low invasiveness and widely available test, similar to what the Pap test has done for cervical cancers.

In this assay, for each subject, an intra-uterine brush sample is collected along with a saliva sample. The genomic DNA is extracted from both these samples, captured using probes with a total size of 146.46 kb using SureSelect XT HS (see target design), sequenced at 20 million reads to a median DNA fragment depth of at least 80% at 1000x, and deduplicated using UMIs. In parallel, uncaptured libraries are also used for Low-pass whole genome sequencing (LP-WGS). Somatic and copy number variants are called, as well as germline variants for 10 genes, and microsatellite instability (MSI) status is determined for known microsatellite loci within the target region. Separately, clinical MSI testing is performed on each sample using a PCR-based assay.

As the ability to detect early stage cancers relies on high sensitivity and specificity, we were interested in testing the PacBio Onso sequencing by binding (SBB) technology which promises much higher sequencing qualities and better performance in homopolymer regions, thus should potentially increase variant detection and MSI calling performance,.

DOvEEgene experimental workflow

Library prep and hybrid capture were performed for all samples, then the post-capture library was split into 2 aliquots and sequenced in parallel on Illumina's NovaSeq S4 flowcells and the PacBio Onso platform following Onso library conversion. Fastqs were then downsampled to the lowest common number of reads prior to analysis and variant calling performance was compared for different types of variants.

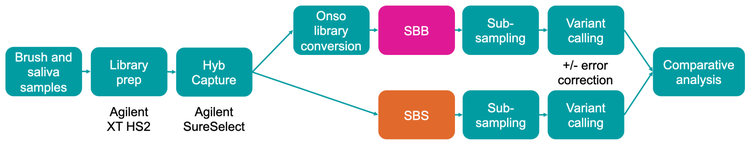

Figure 1. Overview of experimental workflow. Library prep was performed with the Agilent XT HS2 kit, and hybridization-based capture was performed using a custom Agilent SureSelect panel. Library conversion was done using a 5-cycle PCR reaction. Error correction methods tested include GATK start/stop, hybrid GATK + single UMI, and duplex UMI.

Increased raw accuracy and empirical error rates for PacBio SBB

PacBio displayed improved Q scores across both reads compared to Illumina, with the majority of Onso data above Q50 (predicted error rate ≤ 1 in 100,000). PacBio also displayed lower mismatch rates, regardless of error correction method applied.

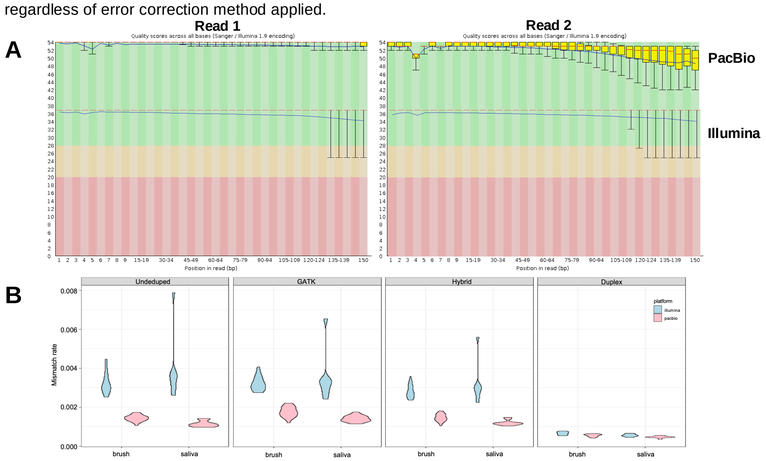

Figure 2. A) Q scores by sequencing cycle for read 1 and read 2. **B)** Comparison of mismatch rates measured as the fraction of bases different from the reference. Comparison was performed for reads with no error correction (undeduped), or after error correction with GATK, AGeNT in hybrid mode, or AGeNT in full duplex mode.

PacBio displays improved performance for duplex correction

We next examined how PacBio Onso and Illumina NovaSeq performed after duplex correction. We found that the increased accuracy of the PacBio Onso resulted in a greater percentage of reads being kept after duplex correction, leading to greater post-correction depth. Variant calling after duplex correction showed good agreement between platforms, with all 3 Illumina-only calls determined to be a result of multi-allelic low-frequency mismatches, likely caused by sequence context that are challenging for Illumina but not PacBio.

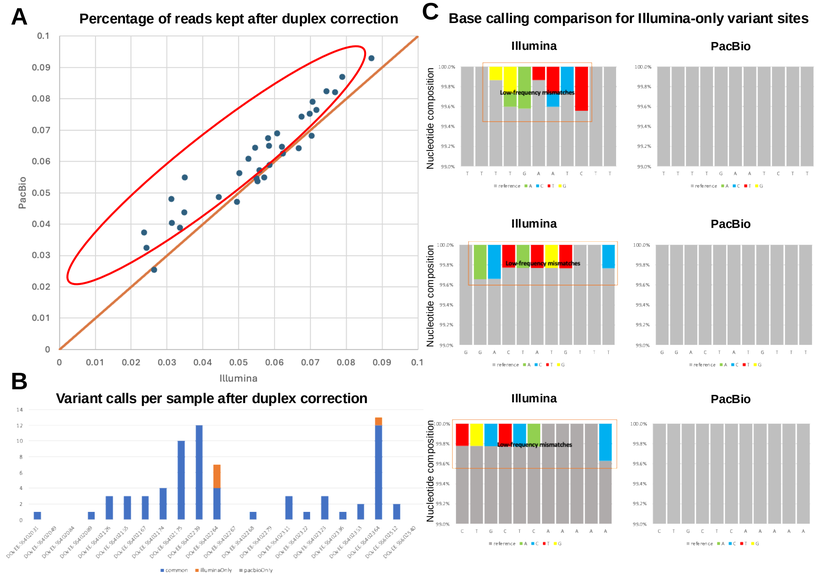

Figure 3. Comparison of performance after duplex error correction. **A)** Percent of reads kept by each technology after duplex error correction. Red circle indicates samples for whom more reads were kept by PacBio. **B)** Variant calls per sample after duplex correction, for variants called by both technologies (blue), or only by Illumina (orange). **C)** Nucleotide composition surrounding the 3 Illumina-only variants, illustrating they fall in potentially problematic regions for Illumina.

Improved sequencing performance by PacBio at microsatellites

We next compared the performance of PacBio Onso and Illumina NovaSeq at known microsatellite loci within the targeted region. PacBio displayed significantly better sequencing performance in these regions, as shown by the increased percentage of reads with the correct deletion start point, as well as by the significant reduction in mismatch errors surrounding the microsatellite locus.

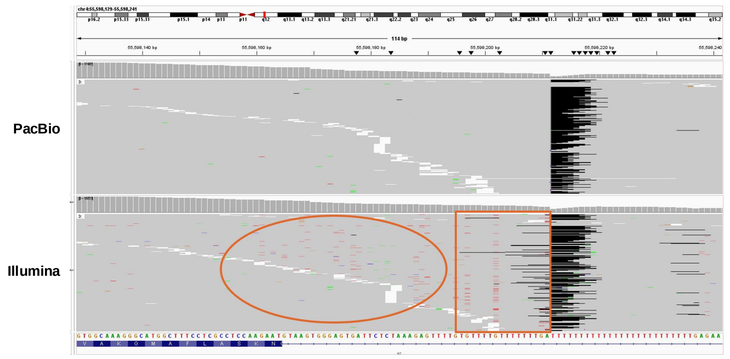

Figure 4. Representative IGV plot showing improved sequencing performance of PacBio Onso (top) compared to Illumina Novaseq (bottom). Orange rectangle: Illumina reads showing incorrect deletion start point. Orange oval: Increased mismatches in Illumina reads adjacent to microsatellite.

PacBio identifies more unstable microsatellites in MSI+ samples

We next compared MSI calling performance using MSIsensor-pro for PacBio Onso vs. Illumina NovaSeq. This tool calculates per-locus thresholds for each microsatellite based on the amount of noise in the corresponding normal saliva sample. PacBio and Illumina thresholds were highly correlated ($R^2 = 0.999$), with PacBio thresholds tending to be slightly lower on average, which may be a result of reduced noise in the saliva samples. PacBio called slightly more microsatellites as unstable on average across samples, despite having a similar number of callable loci across technologies.

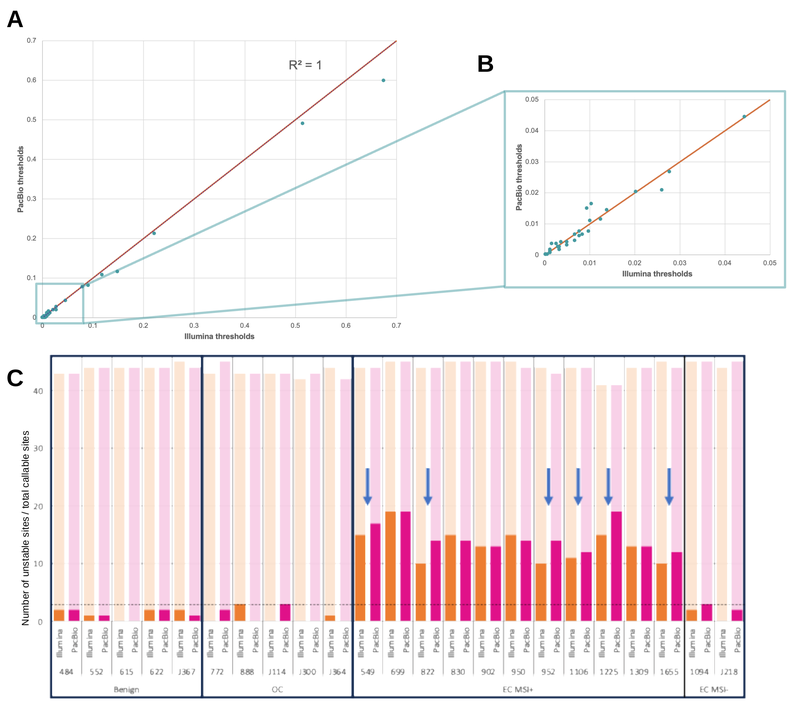

Figure 5. A) Correlation of PacBio vs. Illumina thresholds calculated by MSIsensor-pro for each microsatellite locus. **B)** Zoom-in showing lowest sites with the lowest thresholds. **C)** Number of callable sites (transparent) and sites called as unstable (opaque) for PacBio (magenta) and Illumina (orange). Samples for which PacBio identified more sites as unstable are indicated with a blue arrow.

Conclusion

- PacBio Onso displays higher Q scores and lower empirical error rates, regardless of error correction method.
- Mismatch rates for non-error corrected PacBio approach duplex-corrected Illumina.
- A greater percentage of reads are kept after duplex correction for PacBio compared to Illumina.
- Variant calls were highly concordant between technologies, with Illumina-only variant calls determined to be located in problematic regions containing multiple low-frequency mismatches.
- Improved performance at microsatellite loci by PacBio Onso results in increased detection of unstable microsatellites in known MSI+ samples

References

1. Kennedy et al. (2014). Detecting ultralow-frequency mutations by Duplex Sequencing. *Nature Protocols,* 9 (2586-2606).
2. Jia et al. (2020) MSIsensor-pro: Fast, Accurate, and Matched-normal-sample-free Detection of Microsatellite Instability. *Genomics, Proteomics & Bioinformatics,* 18:1 (65-71)

Acknowledgements

The authors would like to thank everyone who helped generate data for the poster.