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

Abstract # 2443
Timothée Revil[1], Nairi Pezeshkian[2], Lucy Gilbert[1], **Alexandra Sockell[2]**, Jiannis Ragoussis[1]
[1]McGill University, Quebec, QC, Canada,[2]Pacific Biosciences, Menlo Park, CA

Intro

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 an IHC-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 panel and experimental workflow

A

AKT1	PMS2
APC	PPP2R1A
ARID1A	PTEN
ARID5B	RB1
BRCA1	RNF43
BRCA2	TP53
CDK12	MSH2 (germline)
CDKN2A	MSH6 (germline)
CTNNB1	PALB2 (germline)
FBXW7	BRAF-hotspots
FGFR2	EGFR-hotspots
KRAS	MAPK1-hotspots
MLH1	POLE-hotspots
NF1	MED12 exon 2
NRAS	48 genotyping SNPs
PIK3CA	9 microsatellites
PIK3R1	

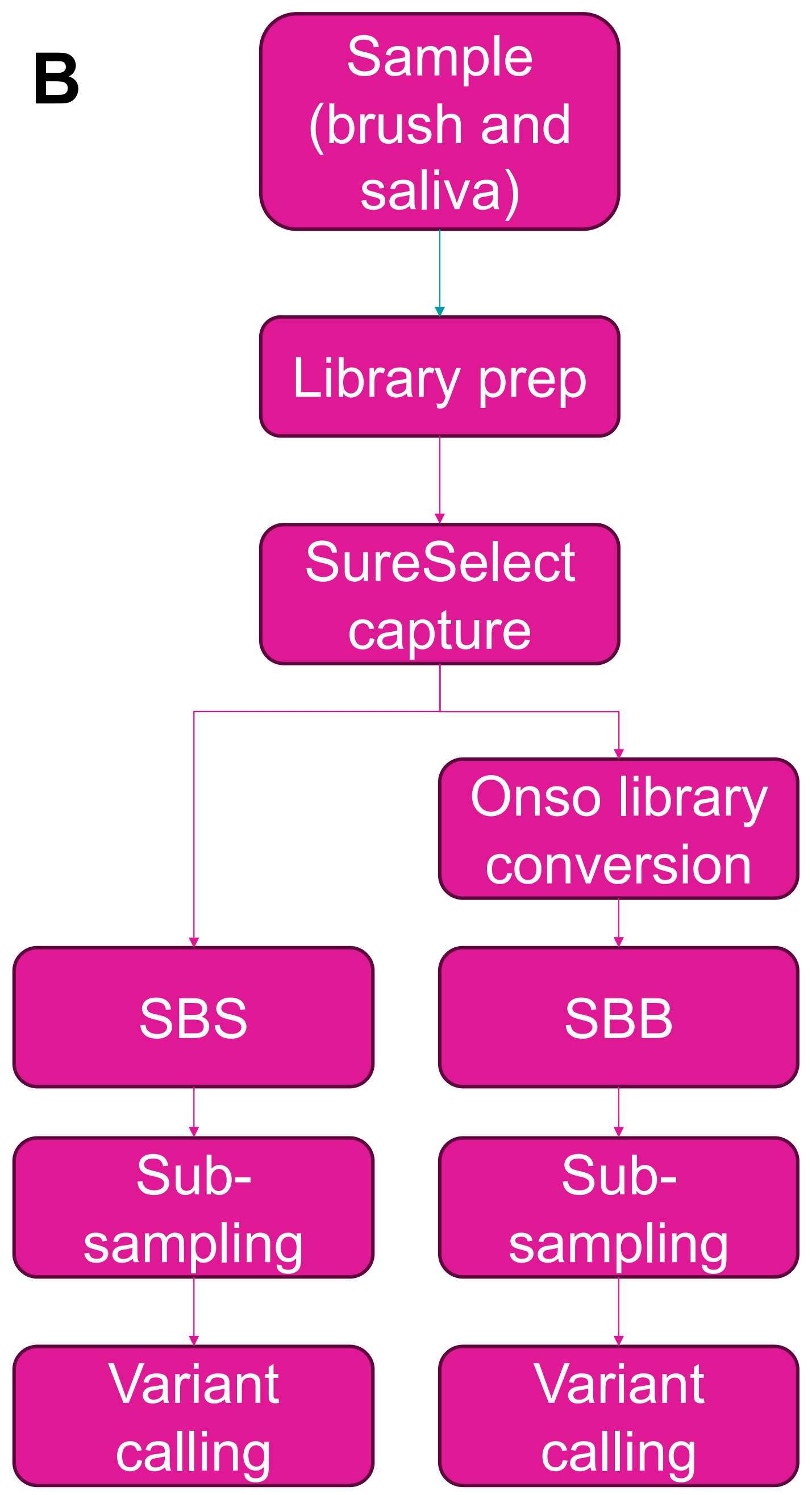

Figure 1. A) Genes captured using Agilent's SureSelect XT HS2. Genes in blue: all coding exons were captured for both saliva and brush samples. Purple: same, but were also used for germline variant calling. Green: only used for germline variant calling. Orange: only hotspots were captured. Yellow: additional panel information. **B)** Workflow of bioinformatics analyses. Sample libraries are created then sequenced in parallel on Illumina's NovaSeq S4 flowcells and the PacBio Onso platform. Fastqs are then downsampled to the lowest common number of reads prior to analysis.

PacBio displays lower total empirical error rates

We compared total error rates (includes mismatches, insertions, and deletions) for PacBio and Illumina. As expected, PacBio displayed lower total error rates regardless of the error correction method applied.

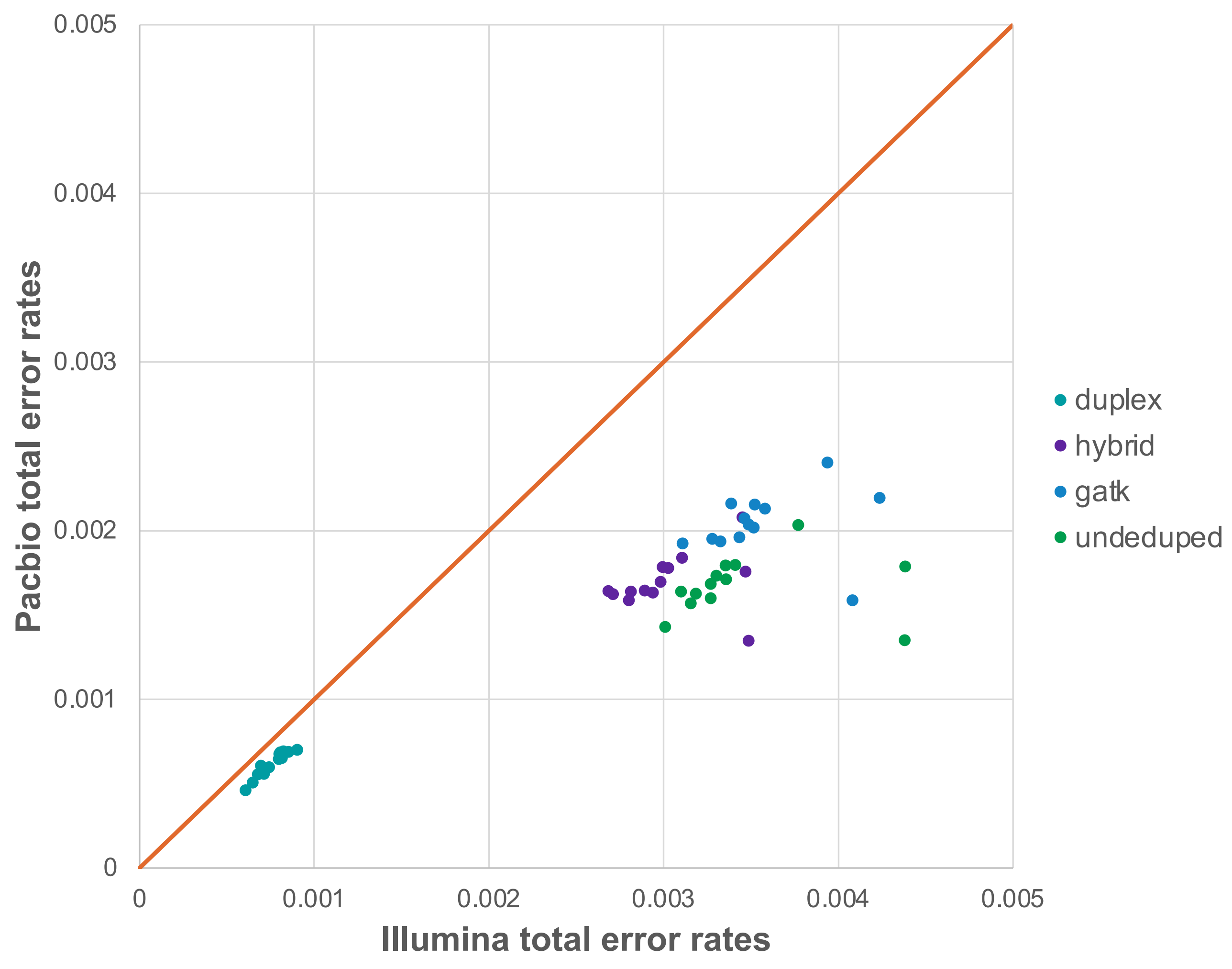

Figure 2. Comparison of total empirical error rates, measured as the fraction of bases different from the reference and including mismatches, insertions, and deletions. Comparison was performed for reads with no error correction (undeduped), or after error correction with GATK, AGeNT in hybrid mode, or AGeNT in full duplex mode.

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.

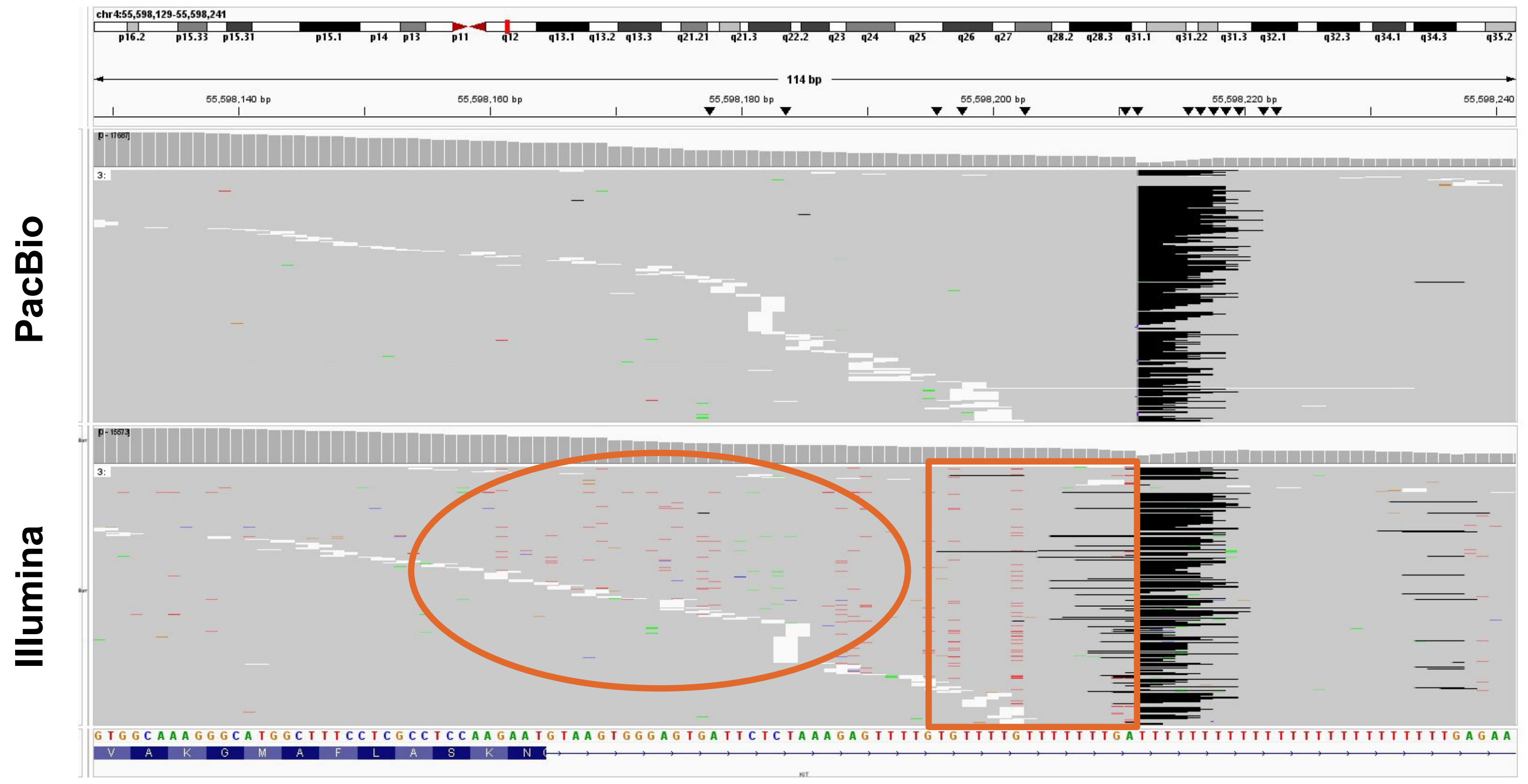

Figure 3. Representative IGV plot showing improved sequencing performance of PacBio Onso (top) compared to Illumina Novaseq (bottom). Orange rectangle: Illumina reads showing incorrect deletion start point. Orange oval: Increased mismatches in Illumina reads adjacent to microsatellite.

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.

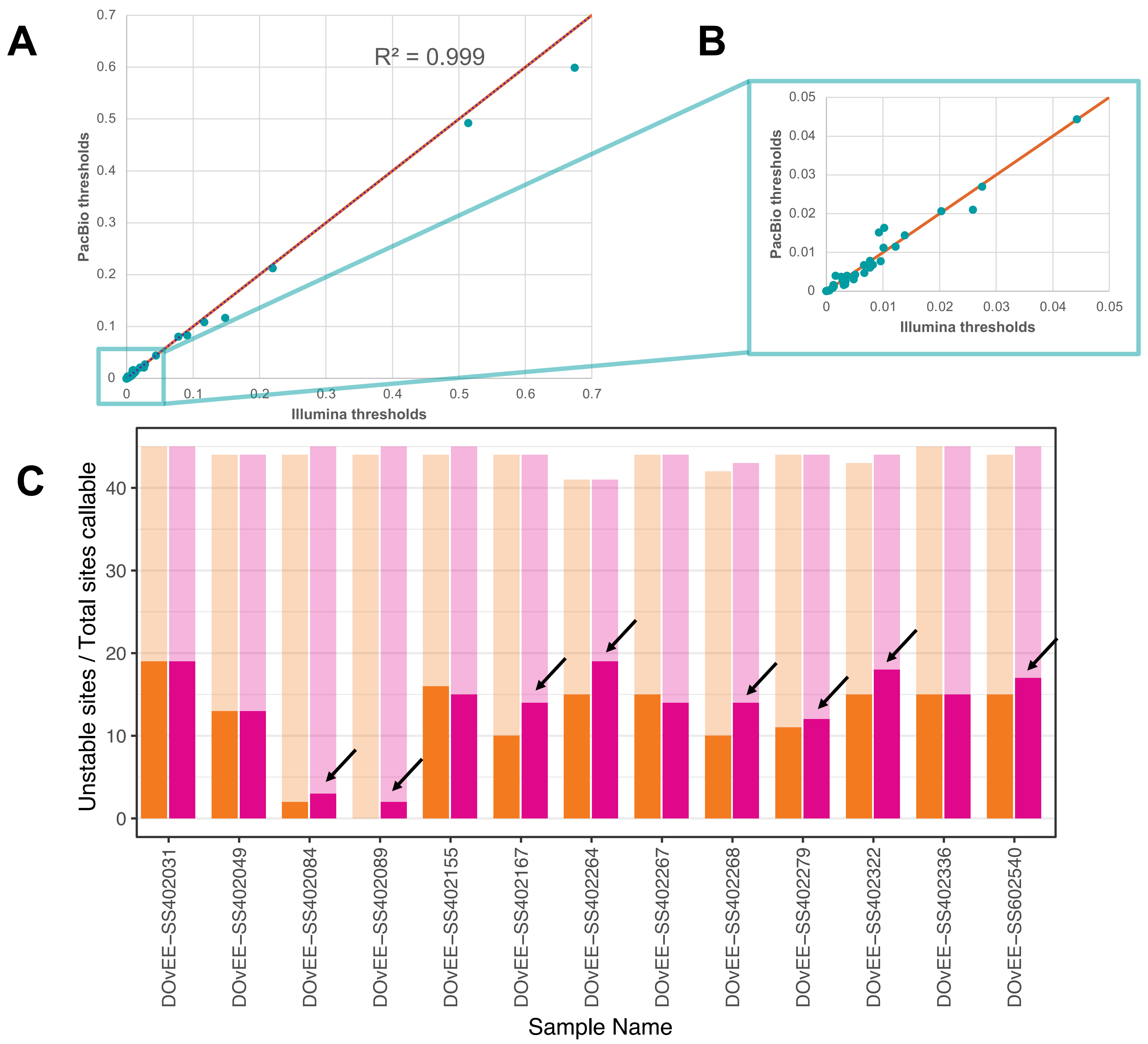

Figure 4. A) Correlation of PacBio vs. Illumina thresholds calculated by MSIsensor-pro for each microsatellite locus. **B)** Zoom-in showing lowest sites with the lowest thresholds. **C)** Number of callable sites (transparent) and sites called as unstable (opaque) for PacBio (magenta) and Illumina (orange). Samples for which PacBio identified more sites as unstable are indicated with a black arrow.

Conclusion

- PacBio Onso displays **lower total empirical error rates**, regardless of error correction method.
- 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.