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

Jiannis Ragoussis[1], Timothée Revil[1], Nairi Pezeshkian[2], Lawrie Shabazian[3], Areej Al Hatib[3], Lucy Gilbert[3,4,5]

[1]McGill University Genome Centre, Quebec, QC, Canada. [2]Pacific Biosciences, Menlo Park, CA,USA. [3]McGill Research Institute - Center for Innovative Medicine, Women's Health Research Unit, MUHC - Royal Victoria Hospital. [4]Department of Obstetrics & Gynecology, Department of Oncology, McGill University. [5]Gynecologic Cancer Service, Cedars Cancer Centre, McGill University Health Centre, Quebec, Canada

DOvEEgene
Test Pap génomique
Genomic Pap Test

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.

These results are then analyzed by a Machine Learning Algorithm, developed based on an extensive training and test set of 481 samples which predicts the probability of a subject having cancer.

As these results of the algorithm are highly dependent on the quality of the variants detected, we were interested in testing the PacBio Onso sequencing by binding (SBB) technology which promises much higher sequencing qualities, thus should potentially increase specificity and sensitivity.

Target design

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	

Table 1. 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.

	brush	saliva
Total Probes	5837	6344
Total Probes Size (kbp)	134.039	149.756

Table 2. Sizes of the target panels.

Workflow

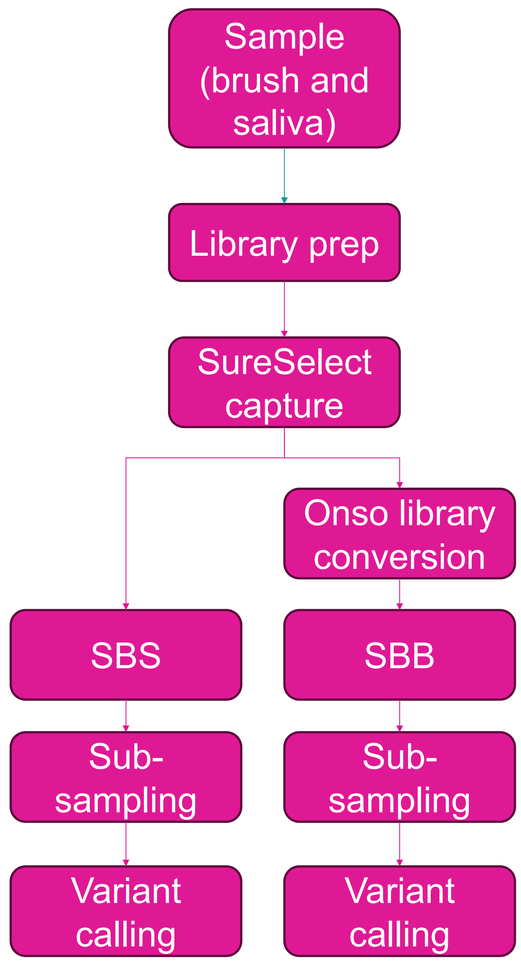

Figure 1. Workflow of bioinformatics analyses. Sample libraries are created then sequenced in parallel on Illumina's NovaSeq S4 flowcells and the PacBio Onso platform. Fastqs are then downsampled to the lowest common number of reads prior to analysis.

Match rates

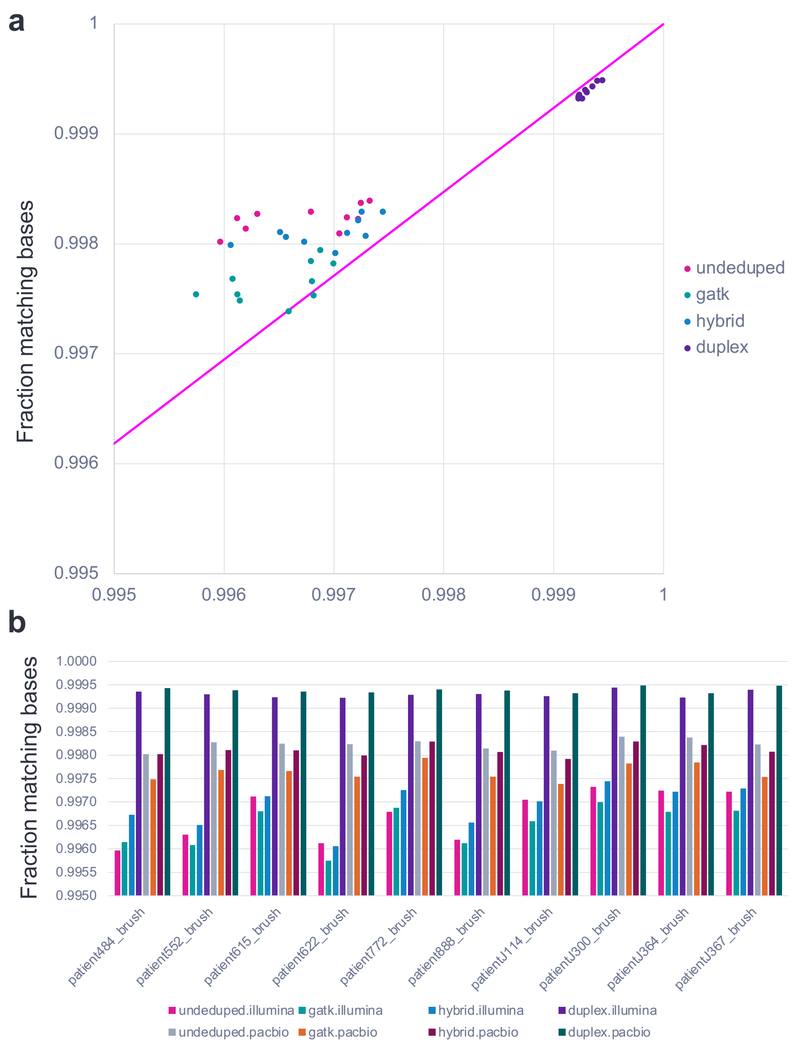

Figure 2a. Match rates of PacBio (y-axis) versus Illumina (x-axis). PacBio Onso had an average of 8.85E-05 fewer mismatches in overlapping bases from the same read pair, representing at least 25,000 fewer mismatches per sample. Values were measured using alfredQC on undeduped, GATK, hybrid or duplex deduplicated BAMs for all samples sequenced on either PacBio or Illumina. **Fig.2b Bar plot of match rates.** These show that in all situations, even duplex deduplication (using MBCs), PacBio had an advantage in mismatch rates.

MSI detection

Subject	Illumina	PacBio
subject484	0	4.65
subject552	0	2.27
subject615	0	0
subject622	4.55	4.55
subject772	2.22	4.44
subject888	0	0
subjectJ114	0	6.98
subjectJ300	0	0
subjectJ364	0	0
subjectJ367	0	2.27

Table 3. Percentage of unstable microsatellite sites in the DOvEEgene target regions. Preliminary results show that higher match rates with the PacBio Onso system allow a higher detection of unstable microsatellites in these subjects.

Lower false-positive variant rates

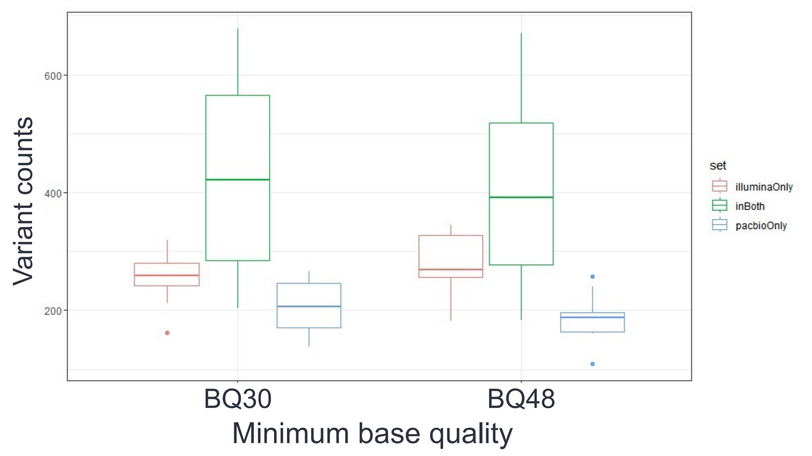

Figure 3. Common variants between PacBio and Illumina suggest lower false-positive rates for PacBio Onso sequencing. Left: Variants were called at a base quality (BQ) of 30 for both IL and PB at and identical sequencing depth and duplex deduplicated, then compared. There were consistently a lower number of variants found only in PB (blue) than in IL (red) for the 10 brush samples. **Right**: Same, but variants were called at BQ48 for PB, as the maximum BQ is 60 for this technology.

Conclusions

As the detection of cancer cells in our brush samples is highly dependent on the quality of the low-frequency variant calls (less than 1%), these preliminary results show great promise as they suggest:

- Lower mismatch rate, leading to:
- Better calling of unstable microsatellites
- Lower number of false positives in variant calls

References

Kennedy et al. (2014). Detecting ultralow-frequency mutations by Duplex Sequencing. *Nature Protocols,* 9 (2586-2606).