Ultra-accurate detection of rare mutations using duplex sequencing

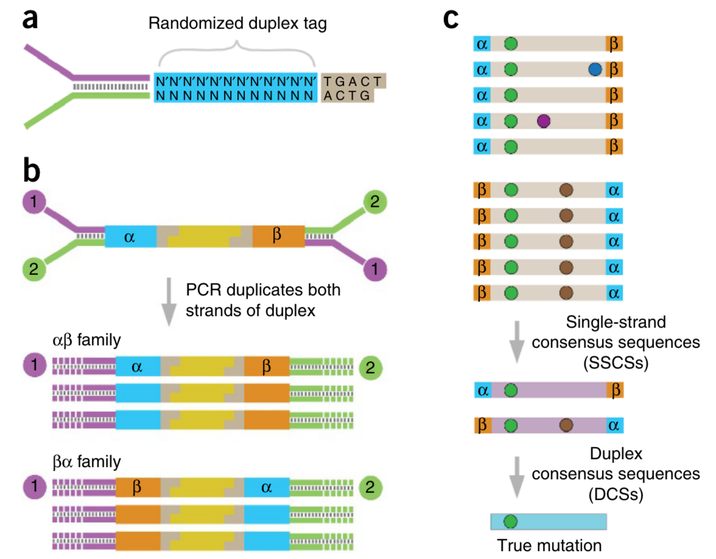
Marek Cmero, Jafar Jabbari, Rory Bowden

Motivation

- Detecting rare single-nucleotide variants in high throughput sequencing data is challenging due to the inherent error rate.
- Duplex sequencing dramatically increases sequencing accuracy (up to 20,000x). For 1x diploid human genome coverage this is the difference between:
  - 6.4 million errors using standard NGS
  - 30 errors(!) using duplex sequencing
- We developed an optimised end-to-end protocol and computational pipeline to perform ultra-accurate duplex sequencing tailored to your experimental requirements.

How it works

a. Double stranded duplex sequencing adapters contain a randomised duplex tag and an invariant spacer sequence.
b. Adapters are ligated to DNA fragments, resulting in unique tags at each end of the molecule. Each strand is then PCR amplified.
c. Reads containing unique α and β tags are grouped together (α-β and β-α 'families' represent different strands). Consensus sequences are obtained from single-strands and then both strands to obtain duplex consensus sequences.



Source: 1. Kennedy, S. R. et al. Detecting ultralow-frequency mutations by Duplex Sequencing. Nat Protoc 9, 2586–2606 (2014).

Computational pipeline

- We have developed a robust, open-source pipeline built using Snakemake.
- Automatically handles dependencies and supports cluster execution (SLURM).
- Outputs a detailed QC reports and family counts.

Extract UMIs — Extract UMIs using fgbio.

↓

Align — Align using BWA.

↓

Group by UMI — Group reads by UMI using fgbio.

↓

Correct barcodes (optional) — If barcodes are known (e.g., xGEN), correction can be applied using fgbio.

↓

Duplex consensus — Call single-strand and duplex consensus sequences.

↓

Align — Remap reads now that we have consensus sequences.

↓

Filter/clip — Filter based on error rate, base quality and required consensus read number using fgbio.

↓

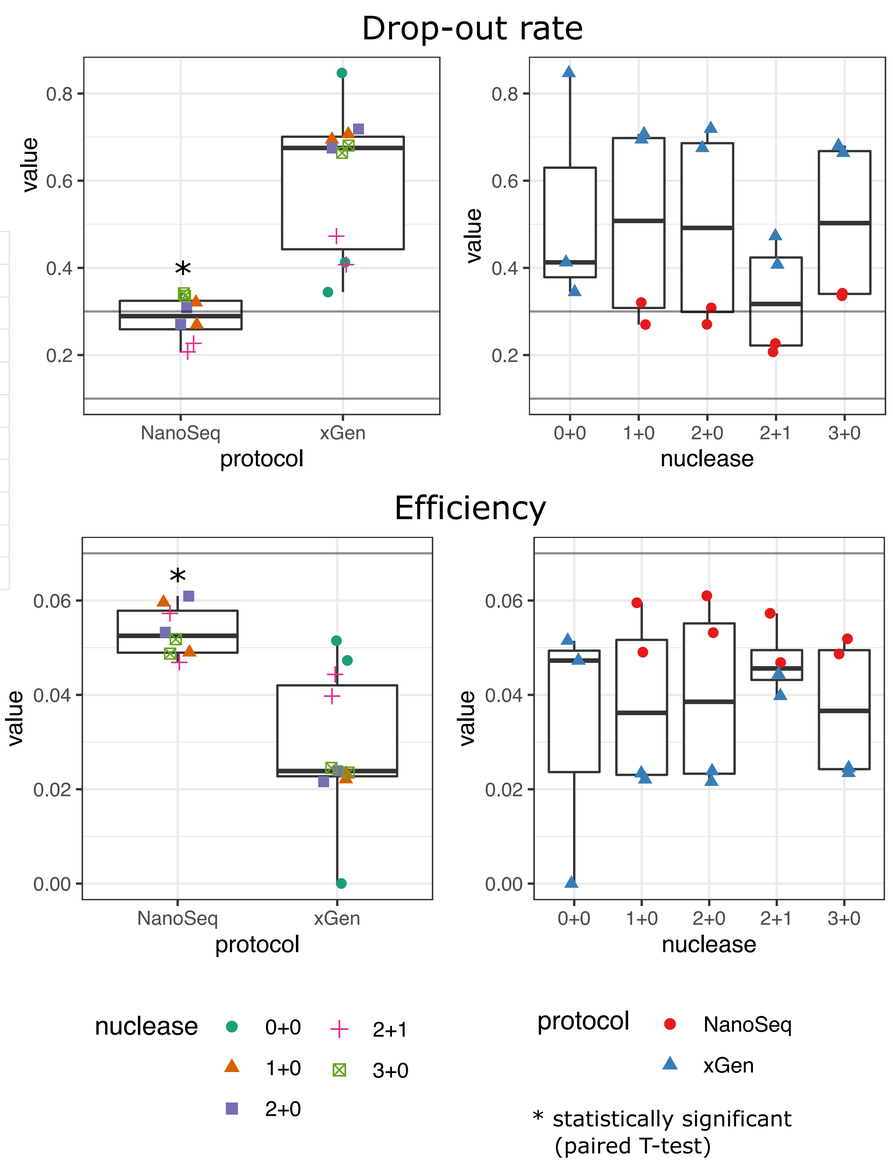
Call variants — Call variants using varscan.

Optimising the protocol

- Aimed to identify the optimum library preparation method, testing five end-repair treatments with two protocols in duplicate using E. coli.

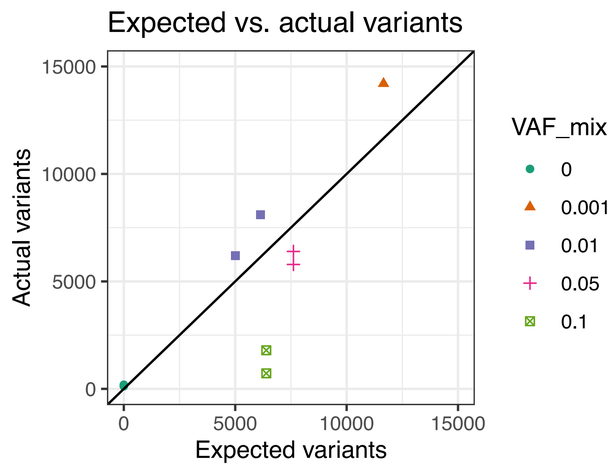
End-prep	Mung Bean nuclease units	S1 Nuclease units	Protocol	PCR	Samples	Key
MB1	1	0	NanoSeq	limited	2	1+0
MB2	2	0	NanoSeq	limited	2	2+0
MB3	3	0	Nanoseq	limited	2	3+0
MB-S1	2	1	Nanoseq	limited	2	2+1
MB1	1	0	xGEN	limited	2	1+0
MB2	2	0	xGEN	limited	2	2+0
MB3	3	0	xGEN	limited	2	3+0
MB-S1	2	1	xGEN	limited	2	2+1
xGEN	0	0	xGEN	limited	2	0+0
xGEN	0	0	xGEN	unlimited	1	0+0

- We found no significant difference in read quality, error rate or GC bias.
- Drop out rate was significantly lower and efficiency was sigificantly higher for the Nanoseq protocol.
  - **Drop out rate**: fraction of read family pairs missing one strand, beyond what would be expected with random sampling.
  - **Efficiency**: duplex bases / sequenced bases.
- We found no significant differences between nuclease units.



Variant detection sensitivity

- To test variant detection sensitivity, we performed spike-in experiments mixing two E. coli strains in different proportions.
- The number of detected variants showed good concordance with expectations.



Customising the assay

- We have developed a model to optimise duplex sequencing based on target allele frequency, capture region size and sample number.
- We can help taylor the assay to your experimental requirements.
- The table below shows some examples based on 95% confidence of detecting a variant (diploid organism) at the given allele frequency.

Target VAF	Capture size	Coverage (raw)	Coverage (duplex)	Sequenced bases	Reads*	Samples per lane*
0.01	2kb	9000x	300x	18Mb	60k	>33k
0.01	10kb	9000x	300x	90Mb	300k	>6.6k
0.01	5Mb	9000x	300x	45Gb	150m	13
0.01	37Mb (human exome)	9000x	300x	330Gb	1b	1
0.001	2kb	90,000x	3000x	180Mb	600k	>3.3k
0.001	10kb	90,000x	3000x	900Mb	3m	>660
0.001	5Mb	90,000x	3000x	450Gb	1.5b	1

*Figures based on NovaSeq S4 flow cell, 2x150 cycles

Genomics Lab, Advanced Technology and Biology Division
Contact: bowden.r@wehi.edu.au
Project inquiries: tinyurl.com/ftr9mn77
Code: github.com/WEHIGenomicsRnD/duplex-seq-pipeline