Germline variant calling with EM-seq data

Laura N. Blum, Louise Williams, Keerthana Krishnan, V K Chaithanya Ponnaluri,
Matthew A. Campbell, Bradley W. Langhorst | New England Biolabs, Inc.

Introduction

Although the use of base conversion in NEBNext® EM-seq™ library prep poses a challenge to variant detection, this can be resolved bioinformatically. Without additional processing, the deamination of unmodified cytosine to uracil (sequenced as thymine) within EM-seq libraries would result in extraneous called mutations. Because methylation information is preserved on only one strand in EM-seq libraries, the other strand can be used to detect genetic variation. To accomplish this, we utilize a standalone tool, Revelio[1], to mask bases that may have come from conversion prior to variant calling. We can then apply a conventional variant caller to analyze the masked data. Using this method, we can call germline SNPs with high recall and precision. With the ability to assess methylation state and genetic mutations from a single library we can maximize the utility of our sequencing datasets.

Methods

Library Construction

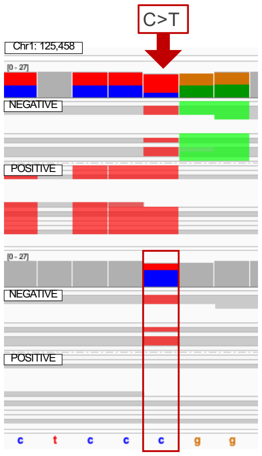
200 ng, 10 ng, and 1 ng (EM-seq v1 and v2) and 100 ng (Covaris-sheared UltraII) of NA12878 DNA was used for library construction. Libraries were sequenced using Illumina NovaSeq 6000 with 2x150 bp reads.

Analysis

Datasets were downsampled to 910M total reads (seqtk). Trimmed (fastp) reads were aligned to T2T with bwa-meth (EM-seq) or bwa-mem (UltraII) and duplicates marked (Picard). EM-seq bams were masked with Revelio. Variants for all libraries were called with Strelka2 (using default passing variants) and Freebayes (--min-base-quality 1, Qual > 15). Hap.py was used to assess concordance of SNPs with UltraII in confident regions from NA12878 GIAB (lifted over from GRCh38 to T2T).

Masking converted bases

Bases which could be the result of either conversion or mutation are set to the reference base and their base quality assigned to 0, so that they can be ignored for variant calling. This applies in C-to-T context for forward alignments to the original (+) strand and reverse complement alignments to the crick (-) strand. It applies in G-to-A context for forward alignments to the crick strand and reverse complement alignments to the original strand.



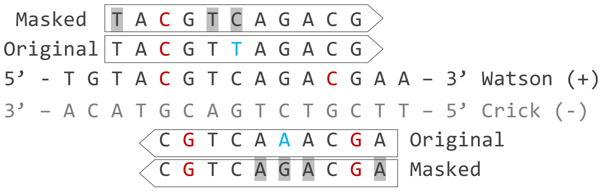
Original Alignment

Ts (red) in the positive alignments and As (green) in the negative alignments could be the result of conversion and are not informative for variant calling.

Masked Alignment

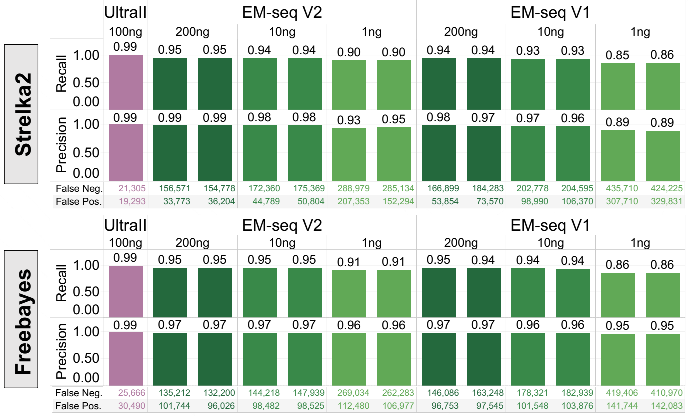
Base quality is set to 0 (now shaded gray) for T and A mismatches and they are ignored for variant calling. The homozygous C-to-T mutation (red box) is called correctly using only negative strand alignments.

Example alignments: red indicates modified C, blue indicates C>T conversion. Gray shading shows BQ=0. Ts and As that could have resulted from mutation followed by conversion are also set to BQ=0.

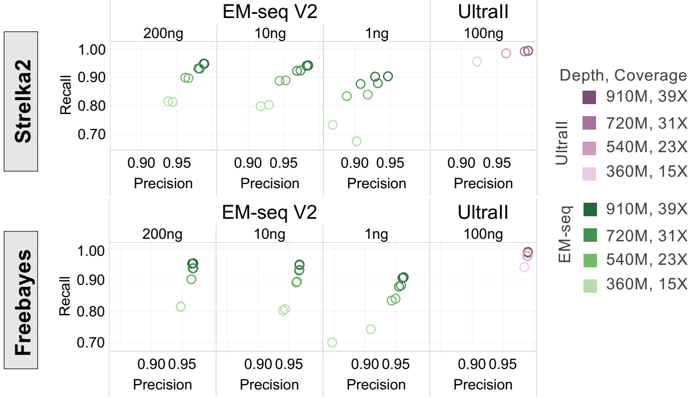


```
 Masked   T A C G T C A G A C G
Original  T A C G T T A G A C G
5’ - T G T A C G T C A G A C G A A – 3’ Watson (+)
3’ – A C A T G C A G T C T G C T T – 5’ Crick (-)
               C G T C A A A C G A  Original
               C G T C A G A C G A  Masked
```

Results



Recall and precision of germline variant calling compared to 2.9M SNPs called in a 100 ng UltraII NA12878 library, using Strelka2 and Freebayes. All libraries were sequenced to approximately 910M total reads. Recall is TP / (TP + FN). Precision is TP / (TP + FP). Two reps are shown, except for UltraII where one rep was used as the truth set.



Recall and precision of germline SNPs for EM-seq V2 at 200 ng, 10 ng, and 1 ng inputs and UltraII at 100 ng input at different sequencing depths.

Strelka2

	EM-seq V2			UltraII		
	A>C	A>G	A>T	A>C	A>G	A>T
Correct	94.6%	95.9%	93.8%	99.3%	99.4%	99.2%
Incorrect	0.2%	0.2%	0.3%	0.1%	0.0%	0.1%
No call	5.2%	4.0%	5.8%	0.6%	0.6%	0.6%
	C>A	C>G	C>T	C>A	C>G	C>T
Correct	94.3%	94.8%	94.9%	99.1%	99.4%	99.4%
Incorrect	0.2%	0.2%	0.1%	0.1%	0.1%	0.0%
No call	5.6%	5.0%	5.0%	0.8%	0.6%	0.5%
	G>A	G>C	G>T	G>A	G>C	G>T
Correct	94.9%	94.7%	94.1%	99.4%	99.3%	99.2%
Incorrect	0.1%	0.2%	0.2%	0.0%	0.1%	0.1%
No call	5.0%	5.1%	5.7%	0.5%	0.6%	0.8%
	T>A	T>C	T>G	T>A	T>C	T>G
Correct	93.9%	95.8%	94.4%	99.2%	99.4%	99.3%
Incorrect	0.4%	0.2%	0.3%	0.1%	0.0%	0.1%
No call	5.8%	4.0%	5.4%	0.7%	0.6%	0.6%

Proportion of SNPs called correctly using Strelka2 across mutation types for 200 ng input EM-seq V2 libraries and a 100 ng input UltraII library at 910M total reads. Variant calls were evaluated against 2.9M SNPs passing default quality filtering for one UltraII replicate at 900M reads. Incorrect means the genotype was wrong, and ‘No call’ means no call was made or it did not pass the quality threshold.

Conclusions

- The problem of distinguishing converted bases from real mutations can be resolved bioinformatically by leveraging the strand-specific nature of EM-seq libraries, allowing us to call germline SNPs with a high degree of accuracy.
- Bases that may have resulted from conversion are ignored, and informative alignments are used to call variants with conventional software.
- Because approximately half of the alignments carry methylation information and are thus not used for variant calling, higher sequencing depth is needed to achieve the same power as with standard DNA libraries.

References

1. Nunn, A., Otto, C., Fasold, M. *et al.* Manipulating base quality scores enables variant calling from bisulfite sequencing alignments using conventional bayesian approaches.*BMC Genomics* **23**, 477 (2022). https://doi.org/10.1186/s12864-022-08691-6