Enzymatic Methyl-seq enables simultaneous analysis of methylation and germline variants from whole genome and target capture data

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

Introduction

NEBNext® Enzymatic Methyl-seq generates consistent and accurate methylation calls using various sample types across a wide range of inputs. The use of base conversion in EM-seq results in a three base genome which poses a challenge to variant detection, however, this can be resolved bioinformatically. Without additional processing, the deamination of unmodified cytosine to uracil (sequenced as thymine) within EM-seq libraries will result in incorrect mutation calls. However, 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, 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 germline SNPs can be called with high recall and precision. With this 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, 10, and 1 ng (EM-seq v1 and v2) and 100 ng (Covaris-sheared Ultra II DNA libraries) of NA12878 gDNA was used for whole genome library construction. Libraries were sequenced using Illumina NovaSeq 6000 with 2x150 bp reads.

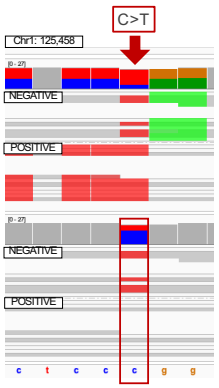
100 ng of EM-seq v2 libraries constructed using Covaris or NEBNext UltraShear® for fragmentation with both standard 3-hour deamination and the shortened 30-minute protocol were used for performing Twist Human Methylome panel target capture.

Analysis

Whole genome datasets were downsampled to 910M total reads and capture datasets were downsampled to 248M reads. Trimmed (fastp) reads were aligned to T2T with bwa-meth (EM-seq) or bwa-mem (Ultra II) and duplicates marked (Picard). EM-seq bams were masked with Revelio[1]. Variants for whole genome libraries were called with Strelka2[2] (using default passing variants) and Freebayes[3] (--min-base-quality 1, Qual > 15). Hap.py[4] was used to assess concordance of SNPs with Ultra II in confident regions from NA12878 GIAB (lifted over from GRCh38 to T2T). For capture libraries, variants were called with Strelka2 for targeted regions and filtered for variants within the GIAB benchmark regions and those with Qual >15.

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 Watson (+) 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 Watson strand.



Original Alignment

T's (red) in the positive alignments and A's (green) in the negative alignments could be the result of conversion and are not informative for variant calling.

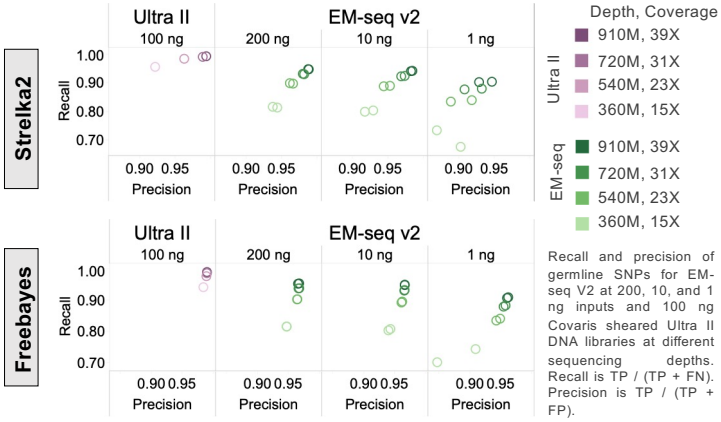
Masked Alignment

Base quality is set to 0 (now shaded gray) for T and A mismatches to reduce the confounded signal. The negative strand alignments can then be used to determine the signal from true variants; for example, a homozygous C-to-T mutation (red box).

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.

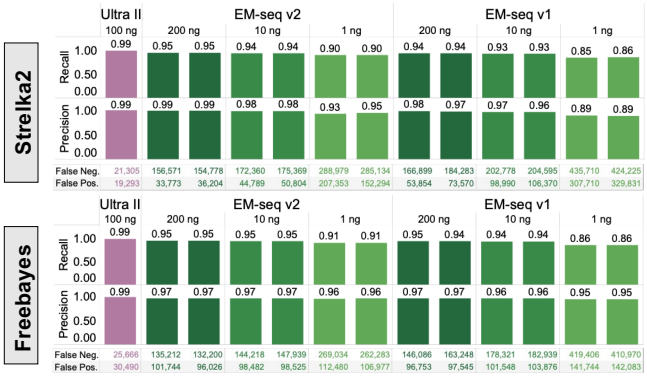
```
Masked    T A C G T C A G A C G
Original  T A C G T T A G A C G
5' - T G T A C G T C A G A C G A A - 3' Watson (+)
3' - A C A T G C A G T C T G C T T - 5' Crick (-)
       C G T C A A C G A  Original
       C G T C A G A C G A  Masked
```

Results

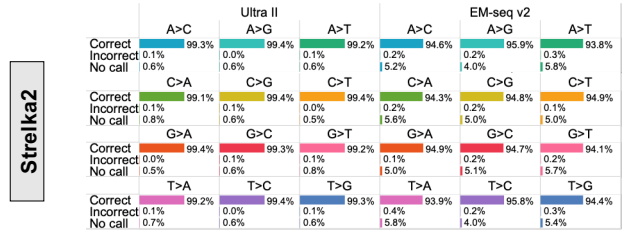


Recall and precision of germline SNPs for EM-seq V2 at 200, 10, and 1 ng inputs and 100 ng Covaris sheared Ultra II DNA libraries at different sequencing depths. Recall is TP / (TP + FN). Precision is TP / (TP + FP).

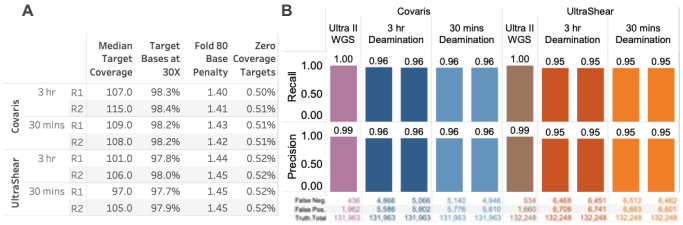
Results, continued



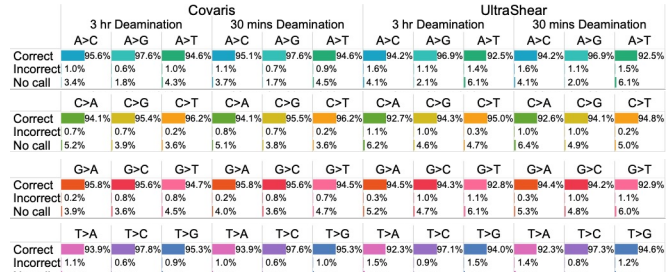
Recall and precision of germline variant calling compared to 2.9M SNPs called in a 100 ng Covaris sheared NA12878 Ultra II DNA library, using Strelka2 and Freebayes are shown. All libraries were sequenced to approximately 910M total reads. Recall is TP / (TP + FN). Precision is TP / (TP + FP). Two reps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



Proportion of SNPs called correctly using Strelka2 across mutation types for 200 ng input EM-seq v2 libraries and 100 ng input Ultra II library at 910M total reads. Variant calls were evaluated against 2.9M SNPs passing default quality filtering for one Ultra II 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.



EM-seq v2 libraries were made using 100 ng of NA12878 DNA fragmented with UltraShear or Covaris with 3 hour and 30 min. deamination conditions. Target capture was performed using Twist Human Methylome Panel. (A) HS metrics for libraries downsampled to 248M reads, all with >99% mapping rate. (B) Recall and precision of germline variant calling compared to 131K SNPs in targeted regions using Strelka2 and Freebayes. Recall is TP / (TP + FN). Precision is TP / (TP + FP). Two replicates are shown, except for Ultra II controls where one rep was used as the truth set.



Proportion of SNPs called correctly using Strelka2 across mutation types for 100 ng input EM-seq v2 target-capture libraries sequenced to 248M reads compared with a 100 ng input Ultra II library sequenced to 910M reads (shown above in B). Variant calls were evaluated against 131K SNPs in targeted regions. 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 in WGS and target-capture libraries.
- Bases that may have resulted from conversion are ignored, and informative alignments are used to call variants with conventional software.
- Approximately half the alignments carry methylation information and are thus not used for variant calling, hence, higher sequencing depth is needed to achieve the same power as with standard DNA libraries.
- Variant callers may assign a high strand bias score, which is expected due to masking, but can cause some variants not to pass default quality filters. Optimized filtering could rescue some of these variants.

Authors would like to acknowledge the technical assistance provided by Dora Posfai, Kristen Augulewicz, Harry Bell, and Rebecca Gawron at the New England Biolabs' Sequencing Core Facility.

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/s12864-022-08692-y
2. Kim, S., Scheffler, K., Halpern, A.L. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591–594 (2018). https://doi.org/10.1038/s41592-018-0051-x
3. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907* [q-bio.GN] 2012
4. Krusche, P. *et al.* Hap.py: Haplotype VCF comparison tools. https://github.com/Illumina/hap.py