Mobelwrap: Accurate and sensitive mobile element detection using short read sequencing

Emma M. Rath[1], Mark Pinese[1], David M. Thomas[1]

[1]Kinghorn Cancer Centre, Garvan Institute of Medical Research

Introduction

Mobile elements are genomic features with the ability to move within the genome, affect gene function, and cause disease. However, the study of mobile element insertions (MEIs) in human health has been hampered by the challenge of detecting these repetitive events from short read sequencing data. The tool Mobster[1] has been recently described to identify MEIs from short read sequencing data, but does not perform genotyping. To extend this functionality, we have developed the toolset Mobelwrap, a wrapper to Mobster that increases sensitivity and specificity of base Mobster calls, and enables genotyping of MEIs. We have benchmarked Mobelwrap on over 1,000 Caucasian cohort whole genome sequencing samples from the Garvan Institute.

Methods

The Mobster[1] software identifies novel MEIs in the genome of a given person or sample by identifying:

* split reads in the sample's BAM file where part of the sequenced DNA read is clearly part of one sequence region in the genome and the adjacent part of the read is not from the same genomic region, and

* discordant read pairs in the sample's BAM file where the DNA reads are from the two ends of the same fixed-length DNA fragment and yet they map to different parts of the genome,

where one of the partial sequences pattern matches that of a known mobile element.

This approach successfully homes in on the regions in large genomes containing novel mobile element insertions. To avoid mistakenly identifying existing known mobile element regions as novel MEIs, the Mobster pipeline ignores MEI calls that fall within 90 basepairs of existing mobile element regions.

Mobelwrap

Our newly developed Mobelwrap software carries out further processing on the Mobster MEI calls to:

* genotype the MEI calls,
* identify novel MEIs in regions of the genome that are known to already contain an existing mobile element,
* remove false-positive MEI calls,
* merge MEIs from multiple samples in the same genomic region.

Mobelwrap achieves this by inspecting the DNA reads mapped to the genomic regions identified by Mobster as containing a novel MEI, and then concatenating and clustering those reads. This finer-grained approach on a greatly reduced data volume allows Mobelwrap to identify whether a concatenated reference sequence is observed, thus permitting genotyping, and to confirm the presence of multiple reads containing the non-reference sequence mobile element, to allow removal of false positives. This complementary approach to the Mobster strategy allows Mobelwrap to consider MEIs identified as falling in regions that already contain known mobile elements, and confidently identify whether the sample's DNA reads simply match the reference genome or represent a new MEI inside an existing mobile element region.

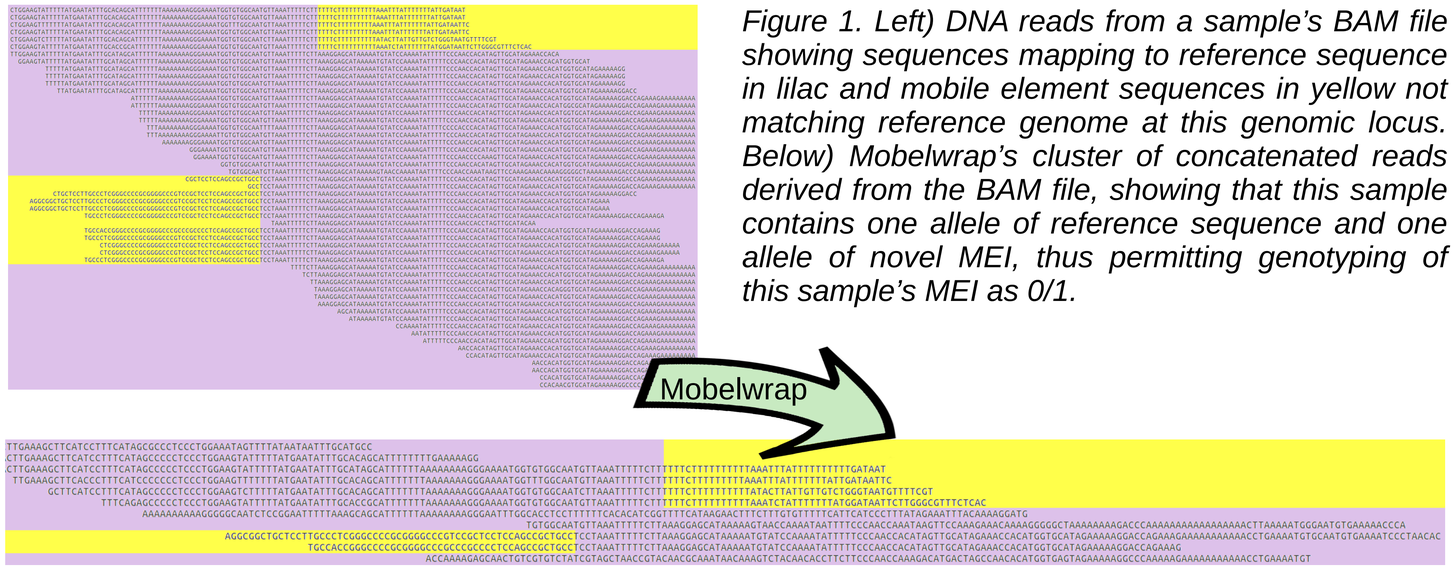
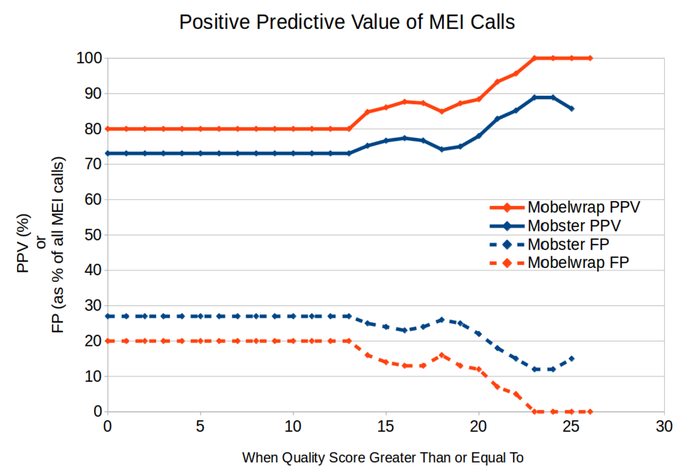
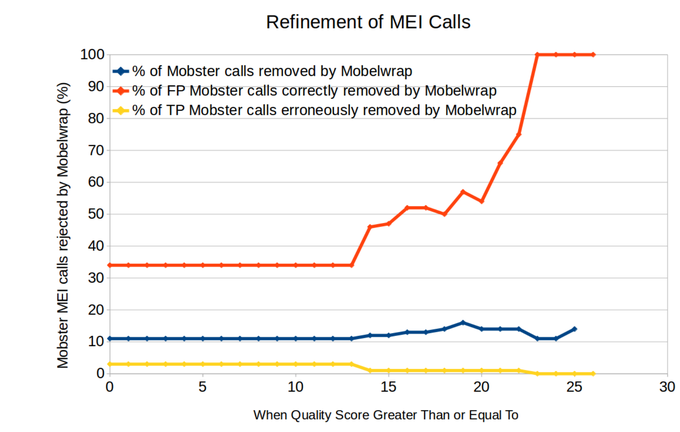


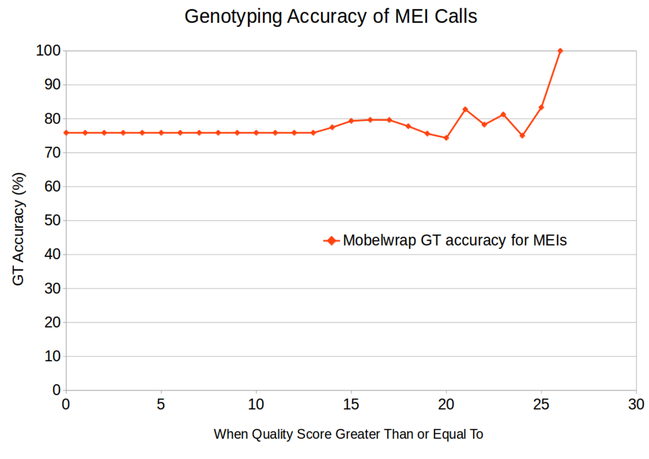
Figure 1. Left) DNA reads from a sample's BAM file showing sequences mapping to reference sequence in lilac and mobile element sequences in yellow not matching reference genome at this genomic locus. Below) Mobelwrap's cluster of concatenated reads derived from the BAM file, showing that this sample contains one allele of reference sequence and one allele of novel MEI, thus permitting genotyping of this sample's MEI as 0/1.

Tool Reliability Results

The accuracy of the genotyping produced by Mobelwrap is 76%. Our Mobelwrap toolset increases the Positive Predictive Power (PPV) of Mobster MEI calls from 82% to 92%, decreasing the False Positive rate from 22% to 11%, when the quality cut-off filter is at an intermediate level. When high quality filtering is used, PPV can reach 100%. Overall, Mobelwrap significantly increases the sensitivity, precision, and utility of Mobster for MEI detection, and we anticipate discovering significant features of mobile elements and their links to disease.







Preliminary Results for a Large Australian Healthy Cohort

The genotyping and merging by our Mobelwrap toolset results in a reduction of separate MEI events to 31% of the initial results, indicating the existence of a significant volume of Caucasian-specific mobile elements. Our preliminary results indicate that a significant amount (41%) of new MEIs occur in regions of the genome that already have older known mobile elements.

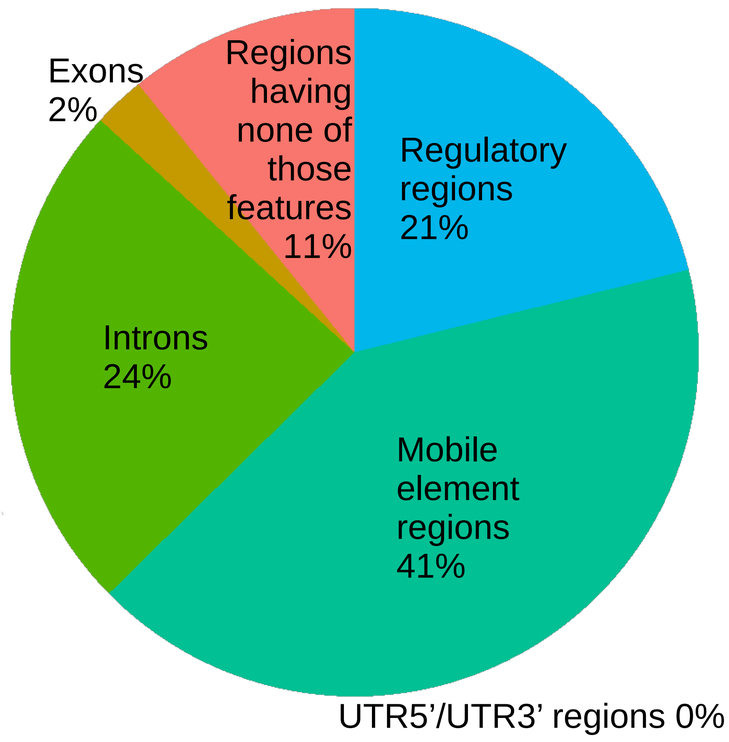


Figure 3. Above) Types of genomic regions where the novel MEIs occur in a cohort of over 1,000 healthy Australian Causasians.

References

1) Thung DT, de Ligt J, Vissers LE, Steehouwer M, Kroon M, de Vries P, Slagboom EP, Ye K, Veltman JA, Hehir-Kwa JY. Mobster: Accurate detection of mobile element insertions in next generation sequencing data. Genome Biol. 2014;15(10):488.

Figure 2. Left) Graphs showing reliability of results from Mobster and Mobelwrap according to levels of quality filtering. Left top) Sensitivity as measured by PPV. Left middle) Movelwrap improvement in specificity as measured by false positives. Left bottom) Accuracy of genotyping by Mobelwrap.