Improved linear alignments through selective re-alignment of diverse references

Nae-Chyun Chen [1] Brad Solomon [1] Ben Langmead [1*]

[1]Department of Computer Science, Johns Hopkins University, Baltimore, USA *langmea@cs.jhu.edu

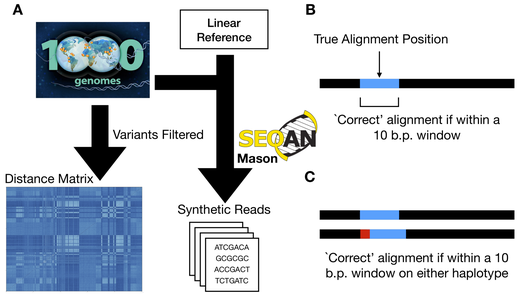
Abstract

Alignment accuracy can be improved by using known genetic variants to remove undesirable alignment penalties. However, the choice of which variants to include substantially affects alignment accuracy [4]. Here we present a novel strategy, selective re-alignment, which uses a variant-free major-allele linear reference to produce an unbiased core alignment and an ERG-based algorithm for the gradual addition of variants. In addition, by "committing" over 90% of reads aligned to a variant-free linear reference, we are capable of testing a wide range of potential variant sets in a fraction of the standard alignment time and compute resources. By "merging" the combined alignment across many possible variant sets, we are capable of exceeding the accuracy of a personalized reference on synthetic data.

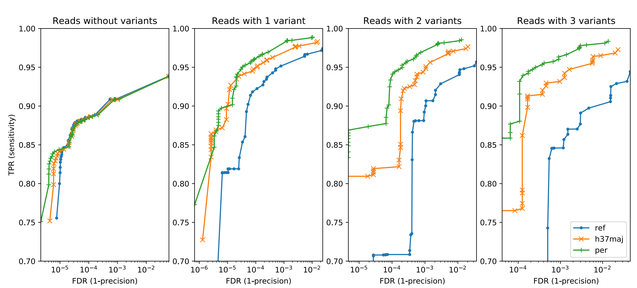
Data and Comparison Metrics

2504 samples from the **Phase 3 1000 Genomes Project** [1] were processed as follows:

- Variants present in ≥ 20% of the dataset were compiled into a pairwise distance matrix
- Haploid and diploid synthetic reads were constructed using **Mason** [3]
- All **gold standards** were selected randomly from each of the five super-populations
- All results tested were compared against **NA18278 (EUR)**



Alignment accuracy for read subsets

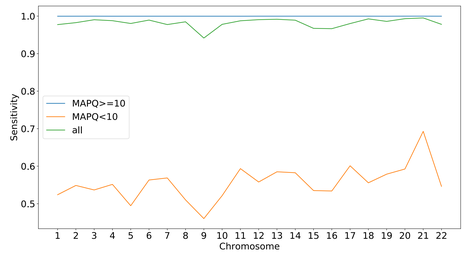


Many reads don't benefit from the inclusion of variants.

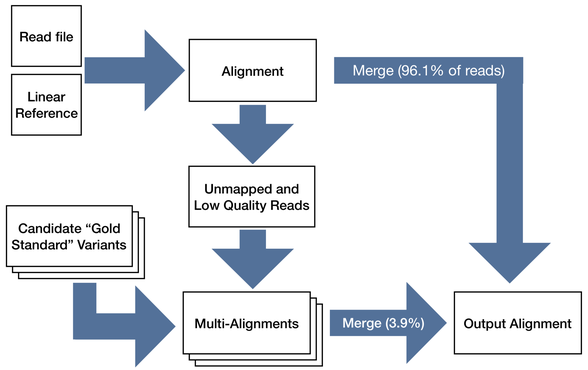
As more variants are included, references which diverge from the standard perform better.

Mapping quality is a good predictor of read accuracy

For more than 95% high quality reads, a linear alignment yields a sensitivity of 99.9%.

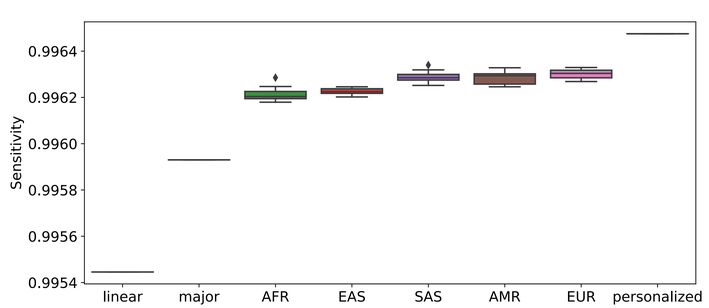


Selective re-alignment and merging



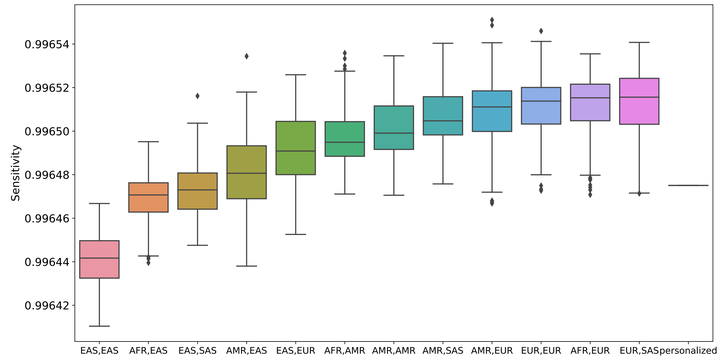
Two-pass linear alignment outperforms both linear and major-allele

Both a major-allele reference and a second-pass alignment using a random individual's variants in the reference are more accurate than a linear alignment.



Selective re-alignment is more accurate then a personalized reference

Using a **first-pass major allele alignment** and a **second-pass alignment with two individuals**, over **74%** of the tested second-pass pairs exceeded the personalized linear reference.



Merging was performed by selecting the read alignment with better alignment score

Selective re-alignment is as efficient as a linear alignment

Normalized index size and run-times using the Bowtie2 aligner for all references.

Reference	Major	Two-pass	Personal	Selective re-alignment
Index size	1	2.12	1.14	3.25
Runtime	1	1.20	1.06	1.33

Selection of better ``gold standards'' is ongoing

A preliminary set of "gold standard" variant sets were selected by taking individuals from each of the major annotated super-populations. We are exploring several alternative annotation-free approaches such as taking cluster centers generated by:

- The **Ward's minimum variance criterion** for hierarchical clustering
- The **Uniform Manifold Approximation and Projection** dimensional reduction

References

[1] 1000 Genomes Project Consortium et al.
A global reference for human genetic variation.
Nature, 526(7571):68, 2015.

[2] Ivar Grytten, Knut Dagestad Rand, Alexander J Nederbragt, and Geir Kjetil Sandve.
Assessing graph-based read mappers against a novel baseline approach highlights strengths and weaknesses of the current generation of methods.
BioRxiv, page 538066, 2019.

[3] Manuel Holtgrewe.
Mason–a read simulator for second generation sequencing data.
Technical Report FU Berlin, 2010.

[4] Jacob Pritt, Nae-Chyun Chen, and Ben Langmead.
Forge: prioritizing variants for graph genomes.
Genome biology, 19(1):220, 2018.