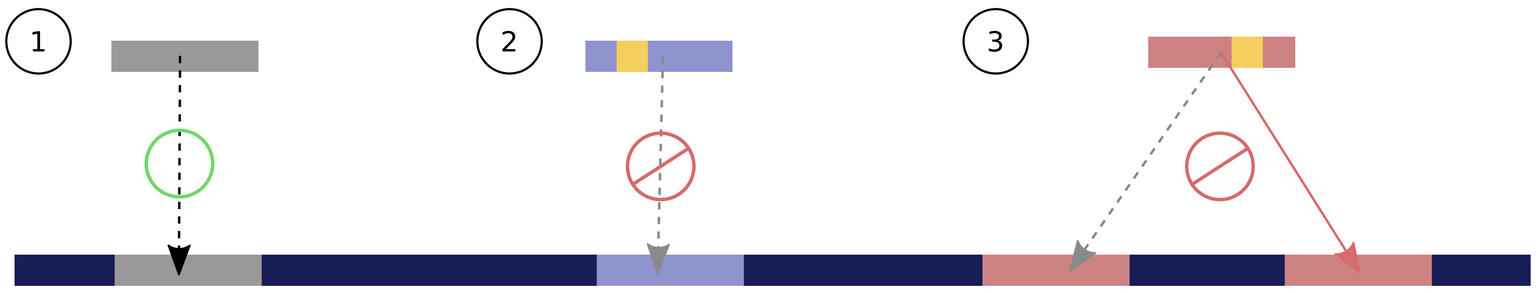
Up-front Imputation Improves Read Alignment

Taher Mun[1], Ben Langmead[1]

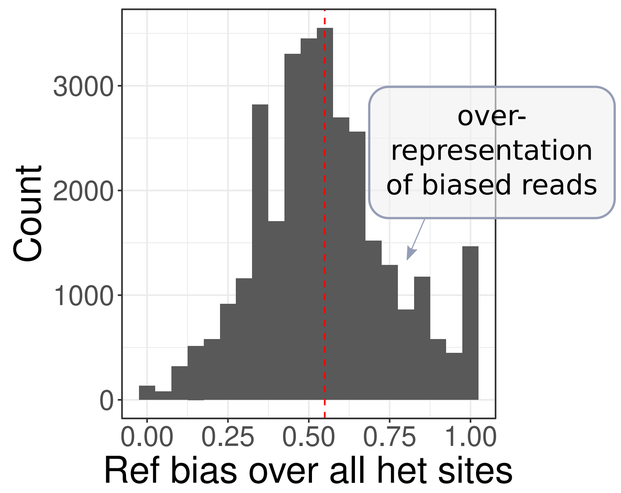
[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD

Reference Bias Affects Read Mapping Accuracy



1. Read is unique and contains no alt allele; no alignment error
2. Read contains ALT allele; more likely to be unaligned
3. Read contains ALT allele, originates in a repeat; likely to be aligned to wrong location

Bias Over Het Sites



10x NA12878 (ERR194147) alignments to chr21

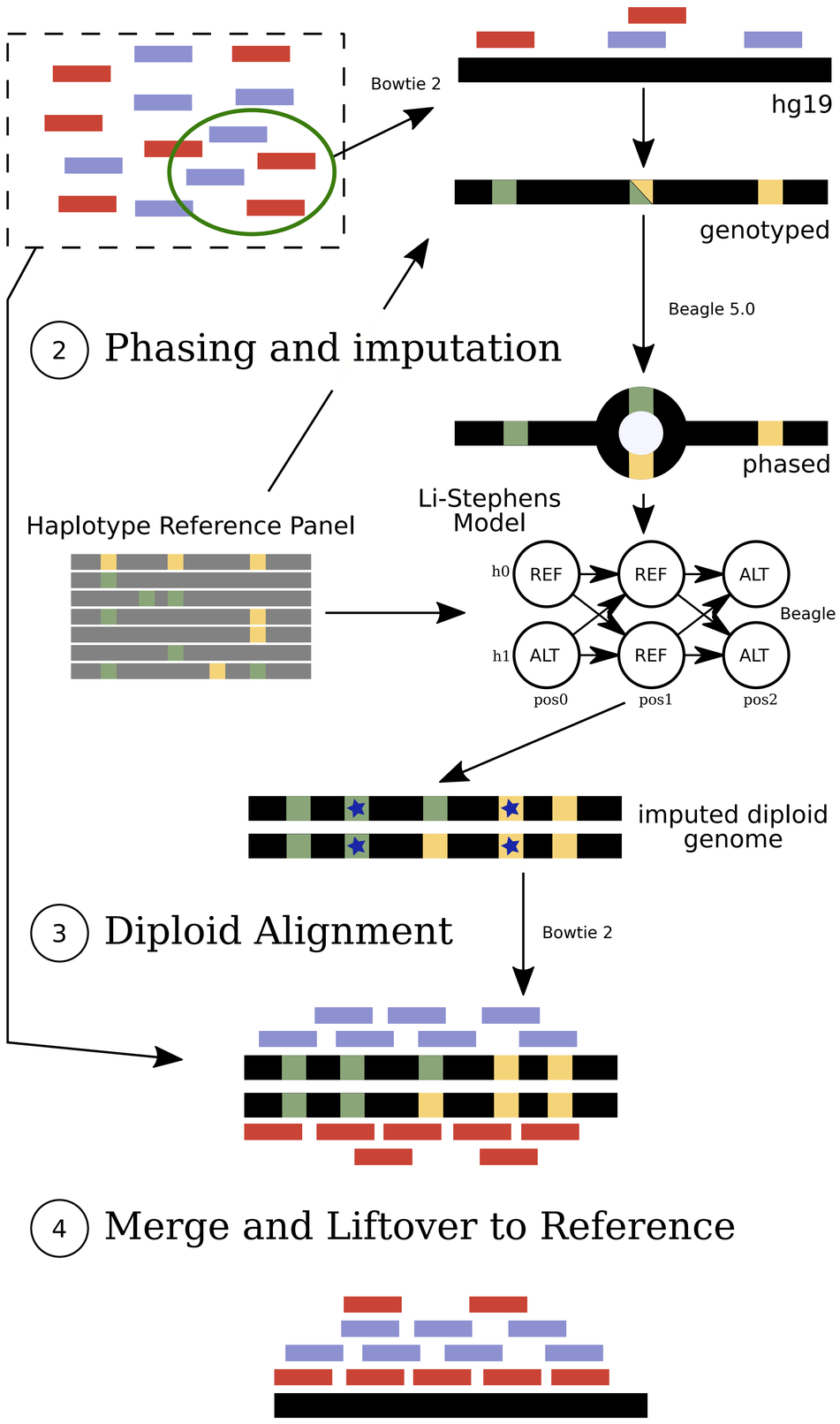
Methods

1 Simple low-coverage genotyping

2 Phasing and imputation

3 Diploid Alignment

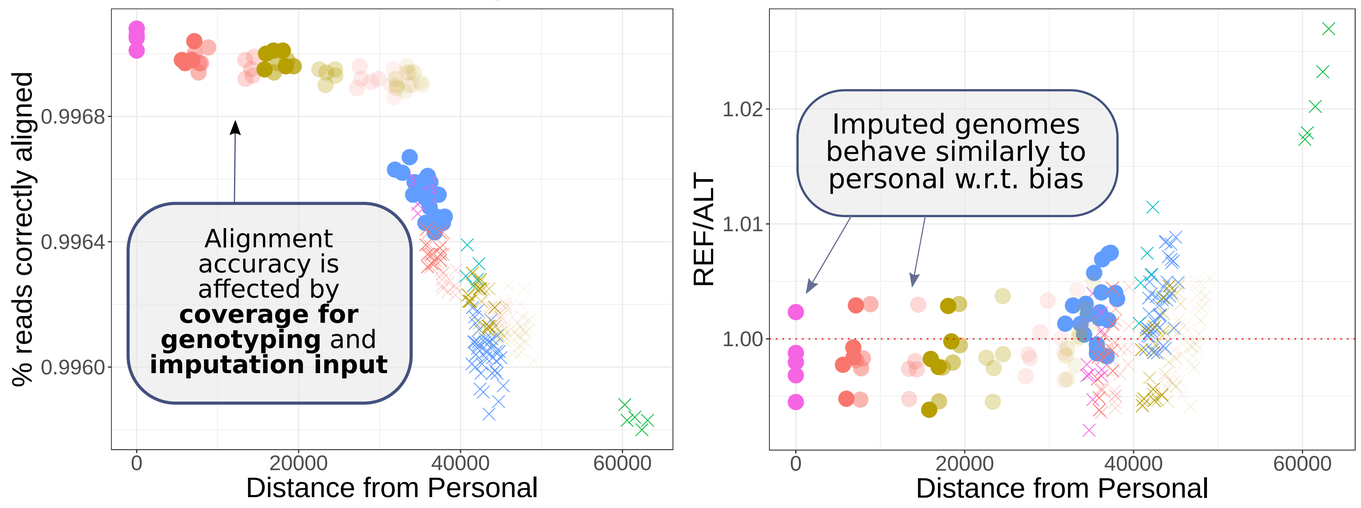
4 Merge and Liftover to Reference



Imputed Genomes Allow for:

1) More accurate alignments

2) Less reference bias



coverage: 5x, 10x, 20x, 30x

ploidy: diploid, haploid

reference: Imputed w/ hets & ALT/ALTs; Imputed w/hets only; GRCh37; Major Allele; Other 1KG Genome; Personal

Data: 30x simulated reads from chr21 from 5 samples in the 1000 Genomes Project

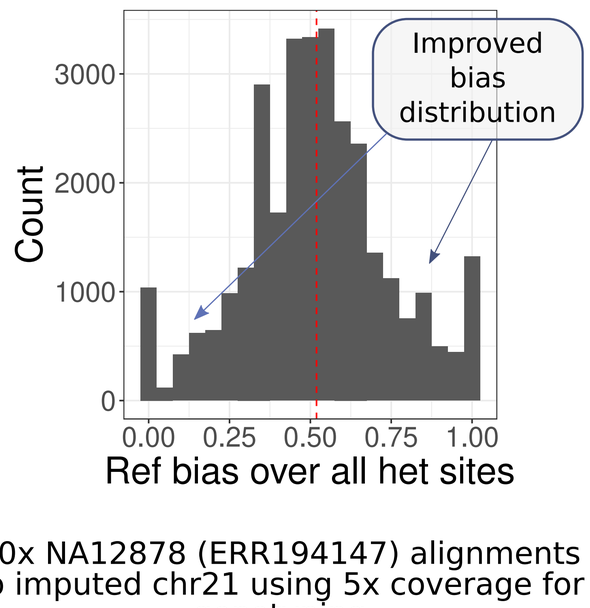
Graph or Linear?

The imputed diploid genome can be indexed as two linear sequences or as a graph. Shown below is the resource usage of both methods:

Step	Time (s)	Memory (GB)
genotyping + imputation	**158.98**	**13.07**
linear indexing (x2)	66.04	0.2
linear alignment (x2)	180.95	1.17
linear lifting (x2)*	139.13	0.02
linear total	**386.12**	**1.17**
graph indexing	807.35	6.59
graph alignment	184.43	0.92
graph surjection	39.21	0.27
graph total	**1030.99**	**6.59**

* threading not currently supported
30x simulated reads from 1KG sample NA18617
linear = Bowtie 2; graph = VG

Bias over Het Sites in Imputed Genome



10x NA12878 (ERR194147) alignments to imputed chr21 using 5x coverage for genotyping

Future Work

- Whole genome support
- Improving bottlenecks, esp. imputation & lifting alignments
- structural variants
- dynamic indexing

Try it out!

Contact Me!

tmun1@jhu.edu

@TaherMun

github.com/alshai

Acknowledgements

Thank you to Nae-Chyun Chen for his feedback related to reference bias

References

Beagle: Browning, et. al., doi:10.1016/j.ajhg.2018.07.015.
1000 Genomes Project: Consortium, doi:10.1038/nature15393.
Platinum Genomes: Eberle, et. al., doi:10.1101/gr.210500.116.
VG: Garrison, Erik, et al. doi:10.1038/nbt.4227.
Bowtie 2: Langmead and Salzberg, doi:10.1038/nmeth.1923.