Improving linear alignment accuracy and reducing bias using reference flow

Nae-Chyun Chen [#] Brad Solomon Taher Mun Sheila Iyer Ben Langmead [*]

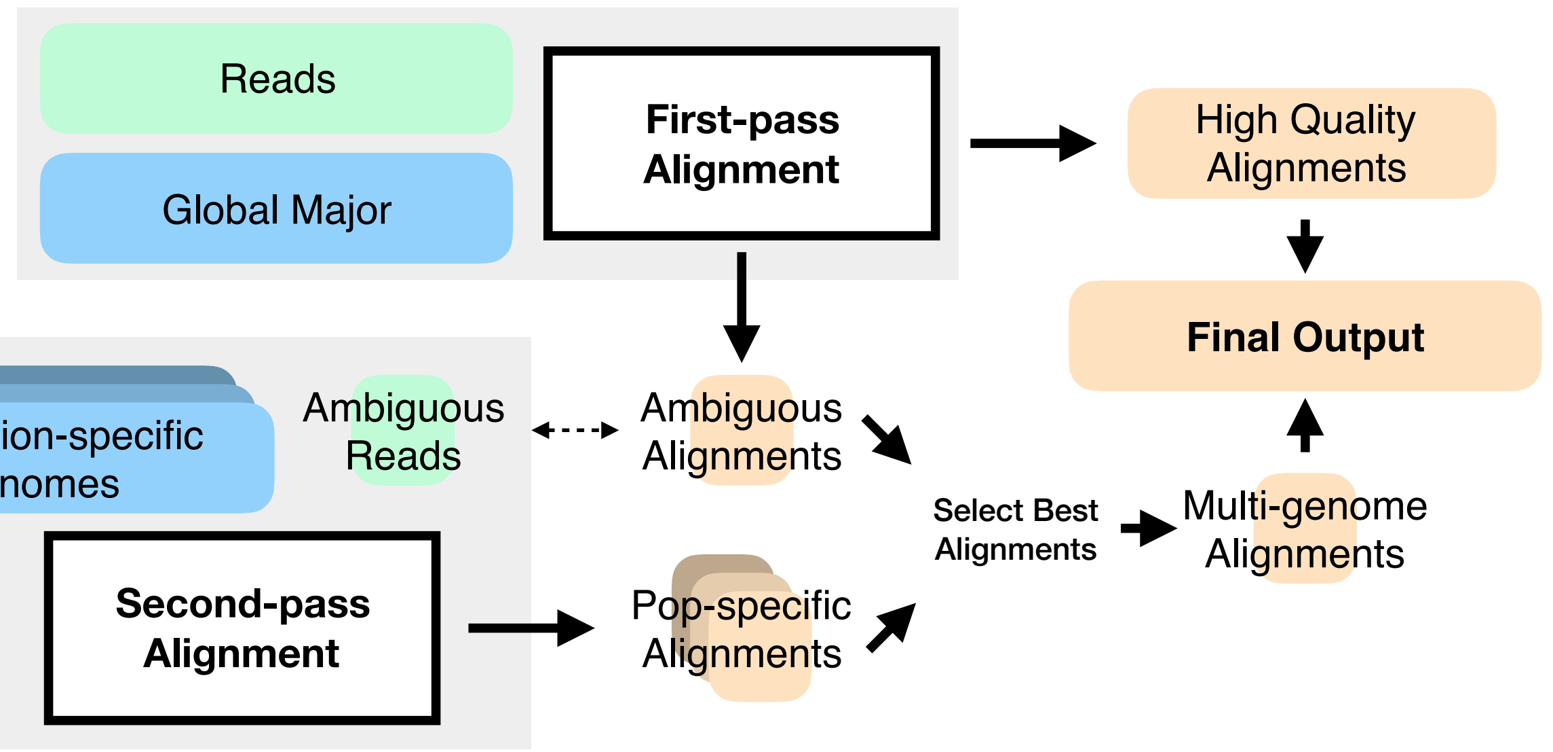
Department of Computer Science, Johns Hopkins University, Baltimore, USA [#]cnaechy1@jhu.edu [*]langmea@cs.jhu.edu

JOHNS HOPKINS UNIVERSITY

Abstract

Reference flow uses a two-pass strategy that identifies ambiguous read alignments in the first pass and re-aligns them to population-specific alternative genomes. Relative to the gain from personalization, reference flow improves **86%** in read mapping sensitivity and reducing **56%** of highly biased sites. It is 5.6x faster and uses 0.12x less memory than a graph aligner.

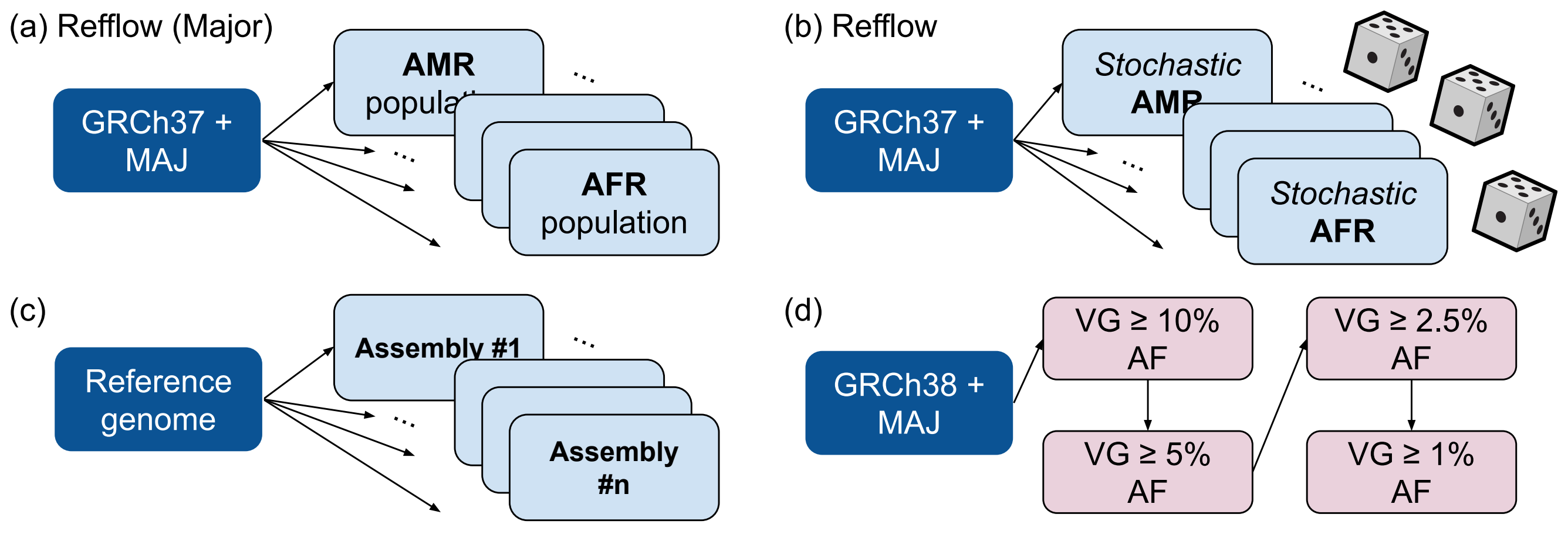
Reference flow: a multi-pass alignment framework enabled by read selection



- **First-pass**: major allele reference is the "centroid" of population

- **Second-pass**: population-specific reference genomes. Stochastic update increases variant diversity and improves performance

- **Selection**: empirically decided mapping quality cutoff can "commit" 80+% reads at whole human genome scale



Reference flow can be generalized to draft assemblies, or combined with other pan-genome-based aligners

Data

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

- All samples were used to build the global major allele genome and population-specific genomes

- Personalized genomes were constructed for 100 random individuals; Mason 2 [3] was used for reads simulation

Deeply sequenced real reads for NA12878 (SRR622457) were used for the real data experiment

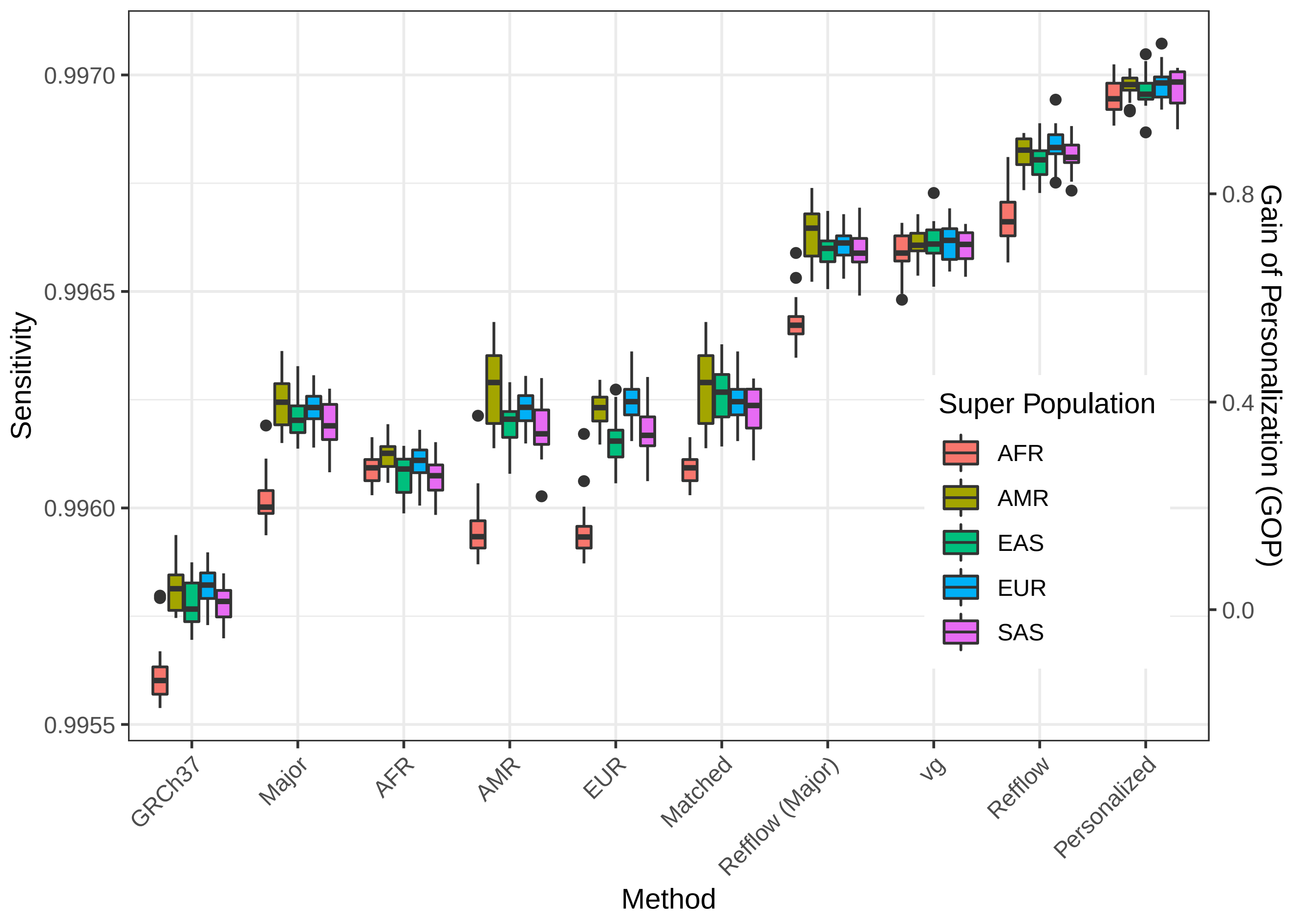
References

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

[2] Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, et al.
Variation graph toolkit improves read mapping by representing genetic variation in the reference.
Nature biotechnology, 2018.

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

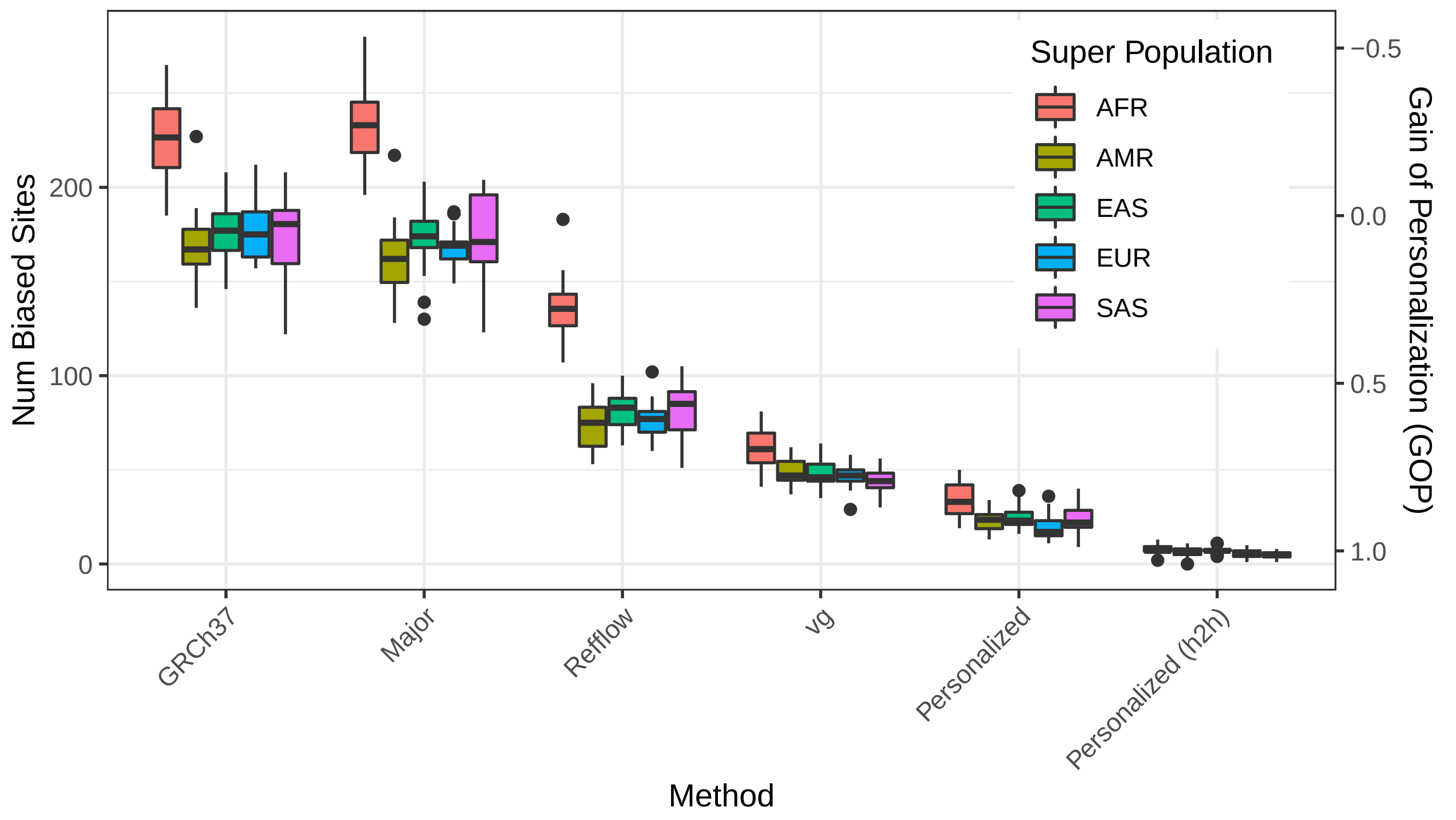
More accurate read mapping than vg [2]



- Major allele reference is a more effective single-haplotype reference in terms of read mapping sensitivity (35.6% GOP)

- Reference flow further improves alignment by integrating multiple population-specific genomes (86.4% GOP)

Reference flow reduces allelic bias

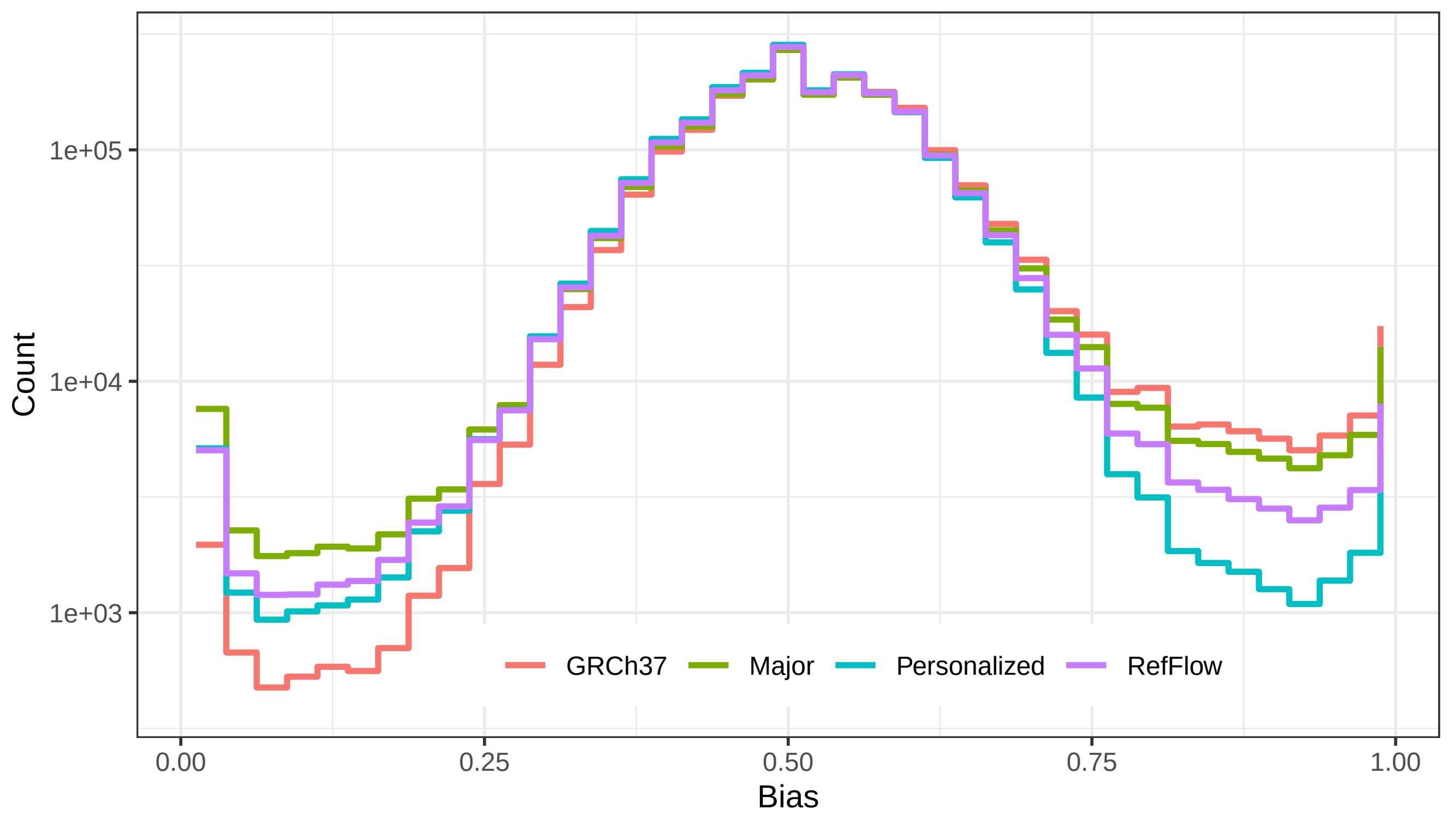


- Major allele reference is limited in reducing allelic bias

- Reference flow recovers 55.9% GOP (haplotype to haplotype)

- Personalized (haplotype to haplotype) aligns reads from each haplotype separately and reduces cross-mapping bias

Reference flow reduces bias for real reads



- Reference flow can reduce bias when real reads are used

- Even personalized is still slightly in favor of the reference alleles

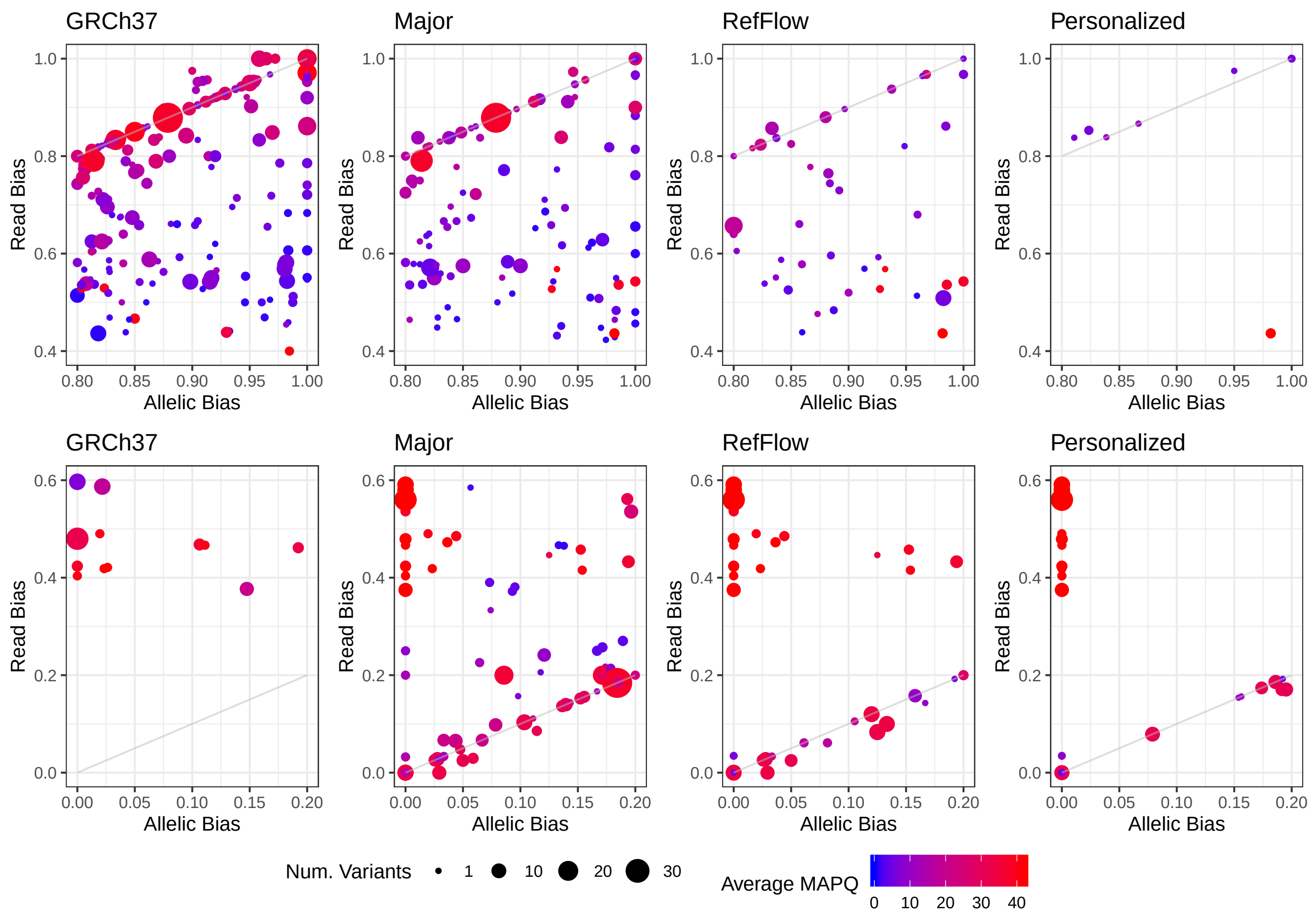
Reference flow is computationally efficient

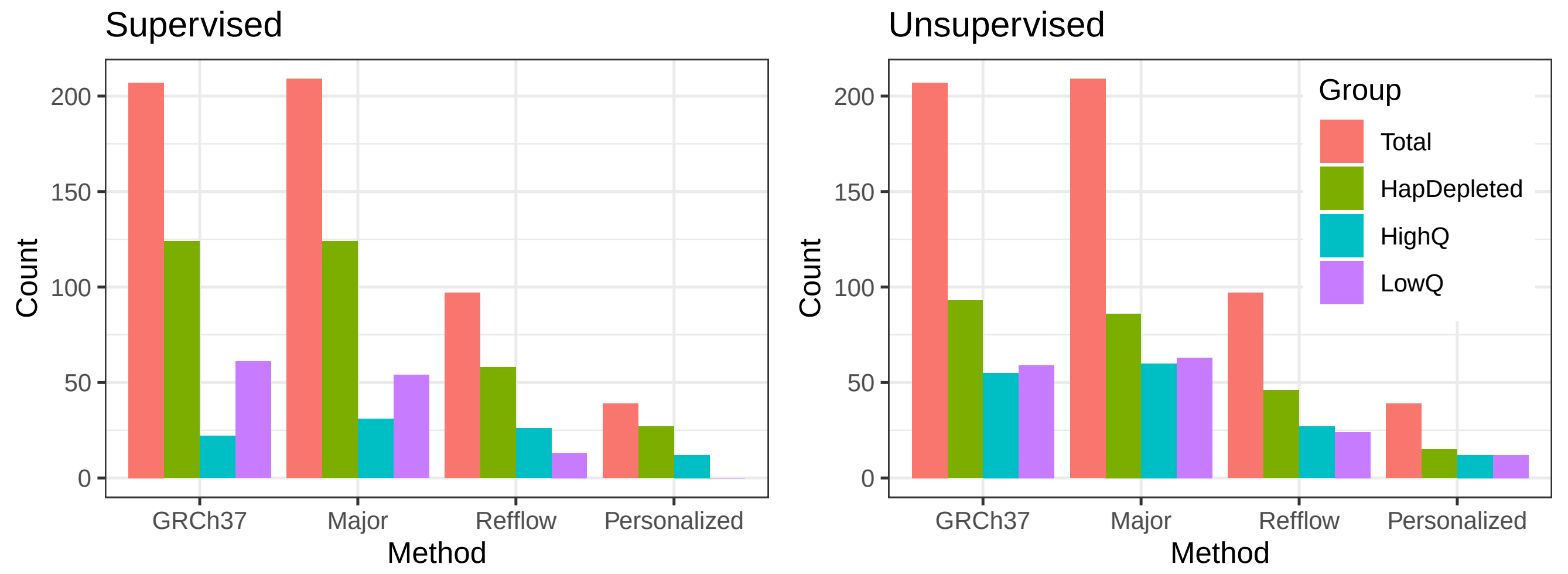
Method	Index size	Memory usage	CPU time
Bowtie2-GRCh37	3.6G	3.5G	1x (54m)
vg*	19.6G	26.9G	13.59x (734m)
Reference flow	21.6G	3.3G	2.42x (131m)

- 10M randomly sampled 101-bp real reads are aligned to whole human genome using 16 threads

* Reads are aligned to GRCh38 with allele frequency > 0.1 variants using vg (we were unable to index vg using GRCh37)

Read bias and allelic bias





- 20M reads are simulated using NA12878 chr21 data

- *HapDepleted*: reads from one haplotype are mis-mapped
- *HighQ*: high MAPQ alignments with balanced read assignment
- *LowQ*: low MAPQ alignments with balanced read assignment

- Unsupervised analysis achieves high correlation without synthetic information. Pearson correlation (p-value): 0.99 (0.007)/0.75 (0.254)/0.99 (0.013) for HapDep./HighQ/LowQ

- Can be further applied for real data analysis

Comparison Metrics

- Gain Of Personalization (GOP)(x)

$$\equiv (x - x_{\text{GRCh37}})/(x_{\text{personalized}} - x_{\text{GRCh37}})$$

Mapping accuracy measurement

- Sensitivity $\equiv |\text{pos}_{\text{mapped}} - \text{pos}_{\text{simulation}}| \leq 10$-bp

Allelic bias measurement

- Only bi-allelic heterozygous SNV sites are considered

- Bias $\equiv$ REF/(REF+ALT+others)

- Biased Site $\equiv$ (Bias ≥ 0.8) $\vee$ (Bias ≤ 0.2)

- Ratio REF to ALT $\equiv \sum$REF/$\sum$ALT