# Reducing reference bias using multiple population reference genomes

Nae-Chyun Chen[1], Brad Solomon[1], Taher Mun[1], Sheila Iyer[1], and Ben Langmead[1,*]

[1]Department of Computer Science, Johns Hopkins University
[*]*corresponding author;* langmea@cs.jhu.edu

March 3, 2020

**Abstract**

Most sequencing data analyses start by aligning sequencing reads to a linear reference genome. But failure to account for genetic variation causes reference bias and confounding of results downstream. Other approaches replace the linear reference with structures like graphs that can include genetic variation, incurring major computational overhead. We propose the "reference flow" alignment method that uses information from multiple population reference genomes to improve alignment accuracy and reduce reference bias. Compared to the graph aligner vg, reference flow exhibits a similar level of accuracy and bias avoidance, but with 13% of the memory footprint and 6 times the speed.

## 1 Introduction

Sequencing data analysis often begins with aligning reads to a reference genome, with the reference represented as a linear string of bases. Linear references such as the primary GRCh38 assembly[1] work naturally with efficient text indexes and sequence alignment algorithms. But linearity leads to *reference bias*: a tendency to miss alignments or report incorrect alignments for reads containing non-reference alleles. This can ultimately lead to confounding of scientific results, especially for analyses concerned with hypervariable regions[2] or allele-specific effects[3,4]. These problems can be more or less adverse depending on the individual under study; e.g. African-ancestry genomes contain more ALT alleles, and so can be more severely affected by reference bias[5].

While graph aligners[6-9] can reduce reference bias, linear aligners still perform better on certain classes of reads[10] and graph-aligner performance is sensitive to the number

of variants considered[11]. Other efforts have focused on elaborating the linear alignment paradigm to address reference bias. Some studies suggest replacing the typical linear reference with a "major-allele" version, with each variant set to its most common allele. This can increase alignment[10-12] and genotyping accuracy[13]. The major-allele reference is largely compatible with the standard reference (though indels can shift coordinates) and imposes little or no additional computational overhead.

We propose a new strategy called "reference flow" that uses a collection of references chosen so as to cover known genetic variants. The method proceeds in two passes where the first pass aligns reads to the "initial" reference and identifies unaligned reads and reads with ambiguous (low mapping-quality) alignments. The second pass re-aligns these reads to a set of five genomes corresponding to the "super populations" studied in the 1000 Genomes Project[14]. By merging results from both passes, we can achieve higher alignment accuracy and lower reference bias compared to methods that use a single reference. This method (a) can use existing, robust linear aligners like Bowtie 2[15] or BWA-MEM[16], (b) requires only a small number of pre-established linear reference genomes, and (c) imposes minimal computational overhead – with no possibility of exponential blowup – relative to linear aligners.

To contextualize the results, a diploid "personalized reference genome" – the genome from which reads are simulated – is used as the ideal reference genome for alignment. By considering the alignments to the diploid personalized reference as a rough upper bound on how well any method can do, we can express results in terms of the degree to which a method closes the gap between the standard linear reference and the personalized reference. When aligning simulated sequence reads, our method closed 94.13% of the gap in sensitivity on median compared to using GRCh38. Our method also reduced reference bias, reducing by 37% the number of strongly biased sites, and lowering the overall reference to alternate allele (REF-to-ALT) ratio from 1.014 to 1.004. When aligning real whole-genome sequencing reads from NA12878, our method reduced the number of strongly biased heterozygous (HET) sites by 9,861 (22%) and lowered the overall REF-to-ALT ratio from 1.072 to 1.017. It achieves similar gains as the vg graph aligner[6] in terms of alignment accuracy and reference bias avoidance while using just 13% of the memory and 17% of the CPU time.

## 2 Results

### 2.1 Standard and major-allele references

We built a global major-allele reference by modifying the GRCh38 primary reference[1] to contain the most common allele at each bi-allelic SNV or small indel. Common alleles were determined using the 1000 Genomes Project GRCh38 call set[14]. We call this the "global major" reference. We repeated this process but considering only the five subsets of individuals belonging to the five super populations labelled by the 1000 Genomes Project.

We call these "superpop major" references. Table 1 summarizes the variants included in each reference. All references were indexed for use with the Bowtie 2 aligner[15].

## 2.2 Simulations for major-allele reference flow

We studied the efficacy of a strategy we call "MajorFlow," which starts by aligning all reads to the global major reference. Reads that fail to align or align with low mapping quality are deferred to a second pass where they are realigned to each of the 5 superpop major references. For each read we report the best alignment to any reference. We performed all alignments using Bowtie 2 and default parameters[15], though the method is not restricted to a particular aligner or set of parameters (Section 3.5).

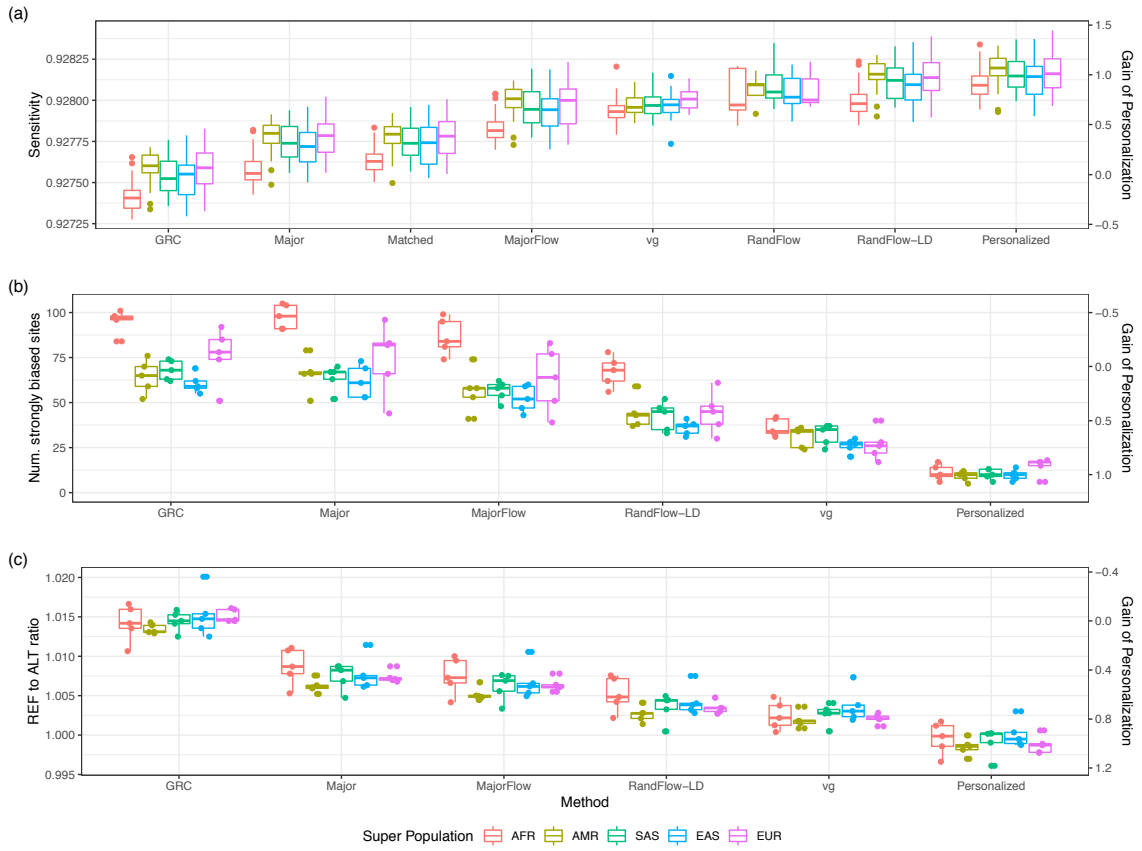


Figure 1: Alignment results using different methods. (a) Alignment sensitivity for 100 samples selected from the 1000 Genomes Project; 2 million reads are simulated from each sample. (b) The number of strongly biased heterozygous sites, and (c) the overall REF-to-ALT ratio for 25 samples; 20 million reads are simulated for each sample.

Table 1: Number and types of variants included in each major-allele reference of chromosome 21. Superpopulation labels are from the 1000 Genomes Project: AFR (African), AMR (admixed American), EAS (East Asian), EUR (European), and SAS (South Asian).

| Group | # samples | # SNVs | # indels | Longest DEL (bp) | Longest INS (bp) |
|---|---|---|---|---|---|
| Global | 2,548 | 25,437 | 4,006 | 44 | 29 |
| AFR | 671 | 27,542 | 4,195 | 44 | 52 |
| AMR | 348 | 26,424 | 4,175 | 44 | 29 |
| EAS | 515 | 26,394 | 4,114 | 44 | 52 |
| EUR | 522 | 25,884 | 4,132 | 32 | 29 |
| SAS | 492 | 26,748 | 4,200 | 44 | 29 |

We performed simulation experiments to compare MajorFlow to baselines that used Bowtie 2 to align to the GRCh38 primary assembly[1] or to major-allele references. We used Mason2[17] to simulate reads from GRCh38 chromosome 21 (Section 3.1). Starting from the 1000 Genomes Project GRCh38 call set[14,18], we randomly selected 100 individuals and built personalized, diploid references for each using phased variant calls (Table S2). We included single nucleotide variants (SNVs) and short insertions and deletions (indels). We simulated 1M reads from both haplotypes (2M total) of each individual. Since allelic-balance measurements require deeper coverage, we also simulated a larger set of 20M reads for 25 of the individuals (Table S2). We also assessed the alignment methods using an ideal, diploid personalized reference genome (Section 3.2). Results using the personalized reference serve as a rough upper bound on what is achievable with references that lack foreknowledge of donor genotypes[11,19,20].

We measured *sensitivity*, the fraction of input reads that aligned correctly. We called an alignment correct if its leftmost aligned base was within $\pm 10$ bases of its simulated point of origin (Section 3.3). A read aligning outside this window or failing to align was considered incorrect. We also measured *allelic balance* at HET SNVs, where we defined allelic balance as the number of alignments with the REF allele divided by the number with either the REF or ALT allele. We also counted the number of *strongly biased sites*, i.e. those with allelic balance $\leq 20\%$ or $\geq 80\%$. Finally, as an aggregate measure of balance, we measured the *overall REF-to-ALT* ratio totaled over all HET sites (Section 3.4).

The MajorFlow method ("MajorFlow" in Figure 1a) exhibited higher sensitivity than single-reference methods that used the standard reference ("GRC") or any of the major-allele references ("Major"). If we consider the increase in sensitivity relative to the sensitivity gap between the GRCh38 reference and the ideal personalized reference, MajorFlow's median sensitivity improvement closed about 67.29% of the gap. MajorFlow's sensitivity was still higher when we enhanced the major-allele strategy by always matching the ethnicity of the major-allele reference to that of the simulated sample ("Matched"). Alignments for reads simulated from the African super population (AFR) had lower sen-

sitivity compared to the others, even when aligned to the AFR superpop-major reference (Figure S1). This is consistent with AFR's greater genetic heterogeneity due to the out-of-Africa bottleneck. Consistent with past results, there was only a small difference in mapping sensitivity when using the global-major versus the superpop-major references, even when the simulated donor's ethnicity was matched with the reference (Figure 1a).

MajorFlow also reduced reference bias relative to the single linear references using the set of 25 deeper simulations. Overall REF-to-ALT ratio decreased from 1.0145 using the standard reference to 1.0075 using the global major reference, then further to 1.0064 using MajorFlow method (Figure 1c). The median number of strongly biased HET sites dropped from 70 for GRCh38 to 59 for MajorFlow (Figure 1c).

### 2.3 Simulations for stochastic reference flow

While MajorFlow outperformed the single-linear-reference strategies, we noticed it was less effective than the graph-based vg aligner at increasing sensitivity or reducing reference bias (Figure 1). We hypothesized this was because the major-allele references used by MajorFlow were too similar to each other, narrowing the genetic diversity visible to the method. Specifically, the mean edit distance between all pairs of superpop major references was 15,115 bp for chromosome 21, whereas the mean between all pairs of five individuals randomly drawn from the super populations was 47,966 bp (Figure S4).

We designed two alternative methods that draw on super-population-specific variation while also keeping the second-pass genomes genetically distinct. "RandFlow" generates a random reference haplotype for each super population by performing an independent draw at each variant site, choosing the ALT allele with probability equal to its frequency in the super population. "RandFlow-LD" is similar but additionally maintains some linkage disequilibrium (LD). RandFlow-LD begins by choosing one haplotype from the super population uniformly at random. Then, starting at the first (leftmost) polymorphic site in the haplotype and for the length of a 1,000-bp window, it selects alleles matching the chosen haplotype. At the next polymorphic site beyond the 1,000-bp window, the method chooses a new super-population haplotype uniformly at random and repeats the process. In this way, variant selection is still weighted by allele frequency (since haplotypes are selected uniformly at random) but a degree of LD is also maintained. Both strategies result in greater genetic distances between the second-pass references compared to MajorFlow, with mean pairwise distances on chromosome 21 being 47,316 for the RandFlow strategy and 46,326 for RandFlow-LD. Further details are in Section 3.5.

Using the chromosome-21 simulation data from the previous section, we observed that RandFlow and RandFlow-LD achieved higher sensitivity than MajorFlow. If we consider the increase relative to the sensitivity gap between the GRCh38 reference and the ideal personalized reference, RandFlow's and RandFlow-LD's median sensitivity improvement closed about 79.47% and 94.18% of the gap respectively (Figure 1a). While RandFlow had lower sensitivity than RandFlow-LD, we note that RandFlow does not require that

variants be phased, and so can benefit from larger compendia of unphased genotypes available through projects like GnomAD[21].

Using the set of 25 deeper simulations, RandFlow-LD reduced the median number of strongly biased HET sites to 44, from a median of 70 using the GRCh38 reference. RandFlow-LD also reduced the overall REF-to-ALT ratio to 1.0038, an improvement over GRCh38 (1.0145) and MajorFlow (1.0075).

We further compared the reference flow methods to vg[6]. vg aligns to a reference that is shaped as a graph rather than a string. Paths through the graph correspond to different combinations of REF and ALT alleles. Such methods can improve alignment accuracy and reduce reference bias by adding paths — thereby removing alignment-score penalties — for ALT alleles. We built a vg index using chromosome 21 of the GRCh38 primary assembly as the base, and including all variants from the 1000-Genomes GRCh38 callset having allele frequency at least 1%. There were 192,846 variants passing the threshold, about twice as many ALT alleles as we considered in our RandFlow (92,825) and RandFlow-LD (92,586) strategies. We found that RandFlow and RandFlow-LD had higher sensitivity than vg (Figure 1a), but that vg yielded a smaller number of stongly biased sites (30, versus 44 for RandFlow-LD) and a slightly more balanced overall REF-to-ALT ratio (1.0026, versus 1.0038 for RandFlow-LD). While neither approach is the clear winner in this comparison, the reference flow methods use substantially less time and memory, as discussed below.

### 2.4 Assessing reference bias with real data

We further assessed these methods using a deep whole-genome sequencing dataset from individual NA12878 (SRR622457) generated by the 1000 Genomes Project[14]. The dataset consisted of 1.4 billion Illumina HiSeq 2000 101-bp paired-end reads, though we used only the first end of the pair in these experiments. Since each read's true point of origin is unknown, we assess only allelic balance and not sensitivity. We assessed allelic balance only at sites where NA12878 is HET according to the 1000 Genomes Project GRCh38 call set. We also constructed and aligned to an ideal, diploid personalized reference using the phased variant calls for NA12878 from the GRCh38 call set. (Figure 2, Table 2). We assessed only the RandFlow-LD reference-flow method since it outperformed other reference-flow methods in the simulation experiments.

Consistent with the simulation experiments, we observed that RandFlow-LD and vg both reduced the number of strongly basied sites, from 44,810 in the case of GRCh38, to 34,939 (22% reduction) for RandFlow-LD and 31,781 (29% reduction) for vg (Table 2). Similarly, RandFlow-LD reduces the overall REF-to-ALT ratio from 1.0719 (GRCh38) to 1.0166 and vg reduces it to 1.0123. Both RandFlow-LD and vg substantially reduced reference bias compared to a method that aligned only to the global major reference ("Major").

Notably, the number of strongly biased sites was still as high as 23,704 when aligning to an ideal personalized reference (Table 2). In part this is because the 1000 Genomes

Project calls include only a subset of the variation present in the actual NA12878 genome. This is both because some genomic regions were excluded from the call set because of low mappability, and because the call set does not include larger-scale structural variants that can have an outside effect on sensitivity and bias. We also noted that the more strongly biased sites were biased toward REF (14,187) more often than toward ALT (9,517) when aligning to the personalized reference, supporting the argument that the call is missing variants that are truly present in NA12878 and that affect reference bias.

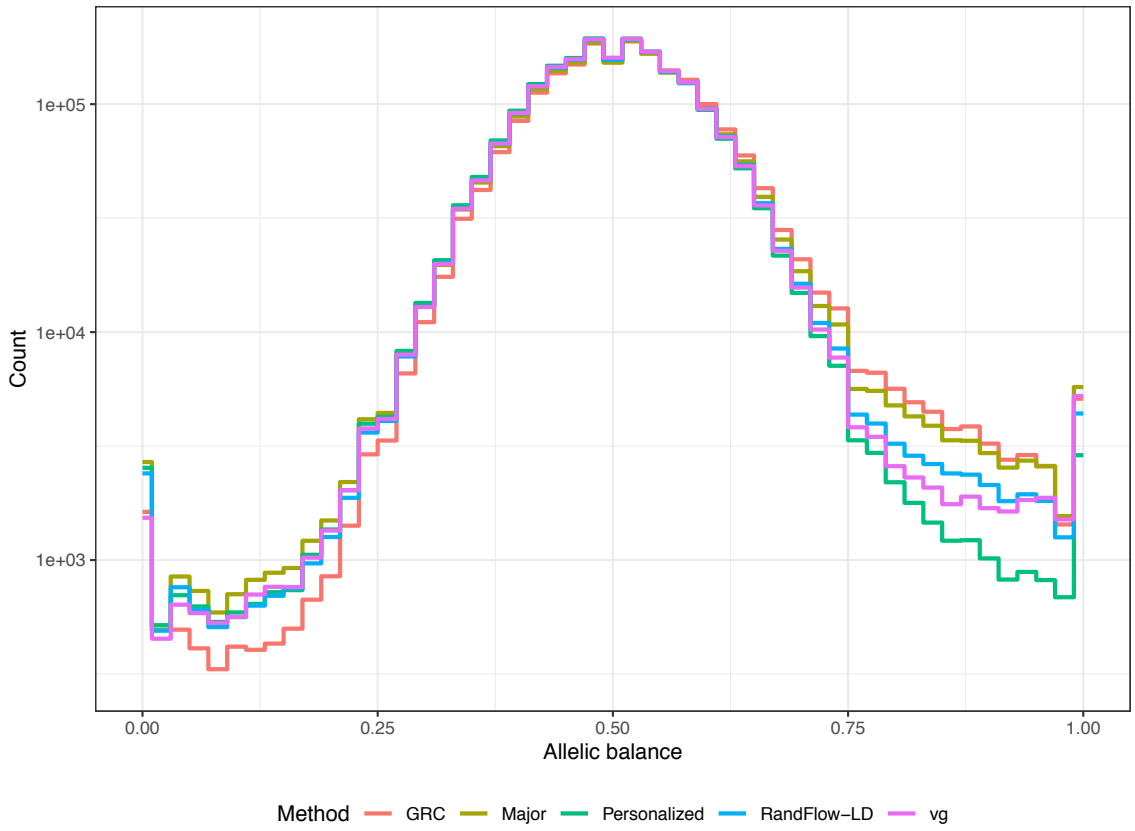


Figure 2: Histograms of allelic balance using a high-coverage real WGS dataset of individual NA12878 (SRR622457). Experiments are performed using GRCh38 (*GRC*), global major reference (*Major*), diploid personalized genome (*Personalized*), vg using alleles with frequency $\geq 10\%$ (*vg*) and reference flow using 1000-bp phased blocks (*RandFlow-LD*).

Table 2: Measures of allelic balance for NA12878 whole genome sequencing dataset. vg index includes variants with 10% or higher allele frequency in the 1000-Genomes Project GRCh38 call set.

| Method | # biased toward REF | # biased toward ALT | Total | REF-to-ALT ratio |
|---|---|---|---|---|
| GRCh38 | 38,735 | 6,075 | 44,810 | 1.0718 |
| Major | 36,059 | 10,893 | 46,952 | 1.0453 |
| RandFlow-LD | 25,773 | 9,176 | 34,949 | 1.0166 |
| vg | 23,391 | 8,390 | 31,781 | 1.0123 |
| Personalized | 14,187 | 9,517 | 23,704 | 0.9977 |

### 2.5 Computational efficiency

We constructed a dataset consisting of 10M single-end reads randomly sampled from the first end of the SRR622457 paired-end dataset. We ran each alignment method and measured the total size of index files on disk, the peak memory usage, and the CPU time (Table 3). We measured peak memory usage using the maximum resident set size reported by the GNU Time utility. We also measured CPU time using GNU Time. We performed the experiments on a computer with a 2.2 Ghz Intel Xeon CPU (E5-2650 v4) and 515GB memory. We configured all read-alignment jobs to use 16 simultaneous threads but otherwise left parameters at their defaults. Though RandFlow-LD was the only reference-flow approach we benchmarked here, we expect MajorFlow and RandFlow to perform similarly since they execute the same sequence of steps, using linear reference genomes of comparable length and complexity.

Compared to an alignment run against the GRCh38 primary assembly, RandFlow-LD used about 6.12 times as much disk space to store the reference index files, consistent with the fact that RandFlow-LD uses 1 reference in the first pass and 5 in the second (Table 3). vg used a similar amount similar size of disk space for its indexes (`.vg`, `.xg` and `.gcsa`). vg used 7.69 times as much peak memory usage compared to the linear-based alignment methods, including RandFlow-LD. The baseline approach used less than 8% of the CPU time as vg, while RandFlow-LD used closer to 17% of the CPU time as vg. Overall, RandFlow-LD used only about twice as much CPU time as the baseline.

We note that RandFlow-LD has similar peak memory footprint to the baseline because the reference-flow software runs the 6 alignment jobs serially. In other words, only one reference genome index is resident in memory at a time. Because the read aligners themselves are multithreaded, we can do this while using many simultaneous threads.

Table 3: Comparison of alignment methods using 10M single-end 101-bp reads from individual NA12878 (SRR622457). vg index includes variants with 10% or higher allele frequency in the 1000-Genomes Project GRCh38 call set. CPU time is compared to a baseline run of Bowtie 2 to the GRCh38 primary assembly.

| Method | Index size | Memory usage | CPU time |
|---|---|---|---|
| Bowtie2-GRCh38 | 3.9G | 3.5G | 1.00x (58m) |
| vg | 21.6G | 26.9G | 12.81x (743m) |
| RandFlow-LD | 23.9G | 3.5G | 2.14x (124m) |

# 3 Methods

## 3.1 DNA data simulation

We built diploid consensus genomes for the selected individuals (Table S2) using `bcftools`[22] based on the SNVs and indels specified by the 1000 Genome Project GRCh38 call set[18]. We used Mason2[17] to simulate paired-end Illumina 100-bp reads, but used only the first end in most experiments. Since variants were already included in the reference genomes we simulated from, we did not use Mason2's variation-adding feature. We enabled Mason2's features for generating random sequencing errors and quality values. We simulated reads independently from each haplotype to generate diploid read sets, keeping information about the haplotype, chromosome and offset of origin for downstream evaluations.

## 3.2 Building and aligning to the personalized reference

We built personalized, diploid reference genomes for each of the 100 randomly selected 1000 Genomes individuals[4,23] (Table S2). We used phased variant calls – including SNVs and indels and including sites with more than 2 ALT alleles – from both haplotypes of the selected individual to build FASTA files containing a personalized diploid reference genome. When aligning to the personalized diploid references, we aligned all reads separately to both haplotypes. We aligned to the haplotypes separately so that the mapping qualities could be informative; aligning to both together would have yielded consistently low mapping qualities. We then merged the resulting alignments. For a read that aligned to both haplotypes, we took the alignment with the higher alignment score. We broke ties by taking the alignment with higher mapping quality or, if the tie remained, at random.

## 3.3 Measuring sensitivity

In simulation experiments, we keep information about each read's haplotype, chromosome and offset of origin. We say a read aligns correctly if the alignment's leftmost mapped base is within $\pm 10$-bp of the leftmost base at the read's point of origin $p_s$. Since

we use Bowtie 2 with default alignment parameters, no "soft clipping" is possible and it does not affect the definition of correctness. Reads that fail to align or that align outside of the $\pm 10$-bp window are called incorrect. We define *sensitivity* as the fraction of reads that aligned correctly.

### 3.4 Allelic balance measurement

We measured allelic balance at each bi-allelic HET SNVs reported in the 1000 Genomes Project GRCh38 call set. HET SNVs that were contained within a larger deletion variant were excluded, whether or not the deletion was heterozygous. At each relevant HET, we considered the "pileup" of alleles at the site induced by overlapping read alignments. Let $a_{\text{ref}}$ and $a_{\text{alt}}$ denote the number of REF and ALT alleles overlapping the site:

$$\text{Allelic balance, } \beta = \frac{a_{\text{ref}}}{a_{\text{ref}} + a_{\text{alt}}} \tag{1}$$

We say a site is *strongly biased* when $\beta \leq 0.2$ or $\beta \geq 0.8$. For a collection of sites, we calculate the *overall REF-to-ALT ratio* as total number of REF alleles divided by the total number of ALT alleles across the sites:

$$\text{Overall REF-to-ALT ratio} = \frac{\sum_{i \in H} a_{\text{ref},i}}{\sum_{i \in H} a_{\text{alt},i}} \tag{2}$$

We ignore alleles besides REF and ALT, and we ignore alignments having a gap at the site. The assumption that on average $\beta$ should equal 0.5 at HET sites is well founded for simulated datasets. Real datasets have biases, which might be due to systematic sequencing errors or fragmentation bias, for example. Biases might also arise from errors in the set of sites we consider to be HET, e.g. if the variant caller that produced the HET calls was itself affected by allelic bias.

### 3.5 Reference flow

**Preparation** The reference-flow methods require that we first build read-alignment indexes and coordinate-translation indexes for the relevant species and populations. Both can be generated from a reference genome in FASTA format and a collection of population variants in VCF format. The reference-flow software (a) processes the VCF to select variants to include in the population reference genomes, (b) generates both the first-pass and the second-pass references based on the reference genome, and (c) builds Bowtie 2 indexes for all references.

For convenience, we provide pre-built RandFlow-LD genomes and indexes based on the GRCh38 reference and the 1000 Genomes Project GRCh38 call set (see Availability of data and materials).

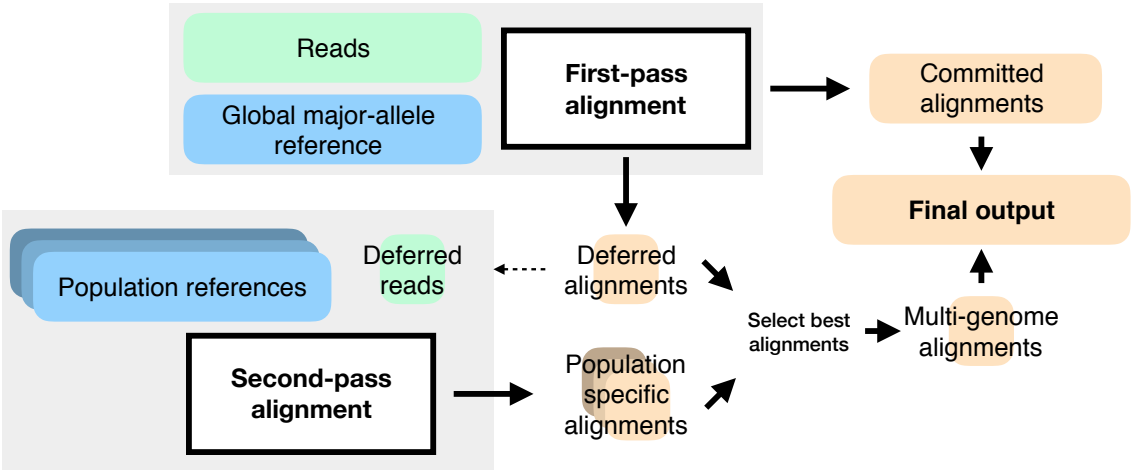


Figure 3: Reference flow workflow. All reads are aligned to major-allele reference in the first pass. High mapping quality alignments are "committed" and included in the final output. Unaligned reads or reads with low mapping quality are "deferred" and re-aligned against multiple population genomes in the second-pass. Second-pass alignments are merged into the final output.

**First pass** In the first pass, we align all reads to an initial reference genome. For the particular reference-flow strategies evaluated here (MajorFlow, RandFlow, and RandFlow-LD), we first aligned to the "global major" reference (Section 2.2). Reads that fail to align or that align with low mapping quality are "forwarded" to a second pass, whereas reads that align with high mapping quality are "committed" and are ultimately passed through to the final output. We use a mapping-quality threshold because it is readily available – reported by most popular read aligners – and because alignments with low MAPQ are the most likely to benefit from the second alignment pass. After empirical experiments, we selected a MAPQ threshold of 10 (Figure S2).

**Second pass** For reads forwarded to the second pass, we realign to a set of references that include a wider range of genetic variation. In the methods evaluated here, we use five second-pass references, each corresponding to a 1000 Genomes Project superpopulation: AFR (African), AMR (admixed American), EAS (East Asian), EUR (European), and SAS (South Asian). In the case of the MajorFlow method, the second-pass genomes are simply the major-allele references corresponding to each of these superpopulations (Section 2.2). In all cases, the second-pass references consist of a single haplotype.

**Stochastic references** In the RandFlow and RandFlow-LD strategies, second-pass references are designed to represent "random individuals" from the super populations. For

RandFlow, we construct the second-pass references by iterating through each polymorphic site $i$ and performing an independent random draw to choose the ALT allele with probability equal to its allele frequency $p_i$ in the super population:

$$GT_i = \begin{cases} \text{REF}_i & 1 - p_i \\ \text{ALT}_i & p_i \end{cases}$$

In the case of the RandFlow-LD strategy, for a variant site we select one haplotype in the super population uniformly at random. We then maintain the linkage disequilibrium (LD) relationship by selecting the genotypes from the same haplotype for the next 1000-bp region.

While we used the super population labels provided by the 1000 Genomes Project here, the reference-flow framework can work with any granularity of label. Further, neither the MajorFlow nor the RandFlow strategies require that genetic variants be phased. Those approaches could also work with larger, unphased compendia of genetic information such as GnomAD[21].

**Merging and lifting** For reads that aligned to more than one reference, we must choose a single "best" alignment to include in the ultimate SAM output. We select by choosing the alignment with the highest alignment score; roughly, this corresponds to the alignment with the fewest mismatches and gaps. If there is a tie for best alignment score, the alignment with higher mapping quality is selected. If there is a tie in both categories, we select at random from among the tied alignments.

For maximum compatibility with downstream tools, the SAM output from our reference-flow methods is with respect to the standard GRCh38 primary assembly. But since the reference genomes in our method – including the major-allele references – can have insertions of deletions with respect to the standard reference, we must translate ("lift over") these alignments to standard reference coordinates before outputting them. We implemented a simple method to lift over alignments that builds a succinct mapping of coordinates from a genome to the standard reference genome using a VCF file. We use the mapping to adjust the POS and CIGAR fields of a SAM file so as to be compatible with the standard reference. The time and memory used to lift the alignments were included in the benchmarking measurements discussed in Section 2.5.

# 4 Discussion

We proposed and evaluated a family of "reference-flow" alignment methods. These are based on the idea that reads that fail to align or align poorly to one reference might align well to another with a different complement of alleles. We first showed that a 2-pass method using superpopulation specific major-allele references (MajorFlow) outperformed

both a standard linear reference and individual major-allele references. As a further improvement, we proposed the RandFlow and RandFlow-LD methods that align to "random individuals" from each super population. These methods performed similarly to vg and approached the performance achieved using the ideal, personalized reference. The reference flow methods were much more computationally efficient than vg, running six times as fast and using 13% of the memory compared to vg when aligning to a graph containing all 1000 Genomes variants of frequency 10% or higher.

Our results complement key points from previous studies. Like the FORGe study[11], we also showed that alignment to a major-allele reference improves alignment accuracy compared to the standard linear reference. Also like FORGe, we showed that aligning to a super-population-matched major-allele reference did not substantially improve alignment accuracy compared to a global major-allele reference combining all super populations. Our results also reinforce that a well configured linear aligner can exhibit similar accuracy to a graph aligner[10].

For compatibility with downstream tools, the alignments that emerge from the reference-flow methods must have their reference coordinates translated back to the standard linear reference. Notably, this requires only a pairwise alignment from each of the reference-flow references to the standard reference. Thus, the approaches discussed here that use 5 references in the second pass require 6 pairwise whole-genome alignments: one from the first-pass major-allele reference to the standard reference, and 5 from each of the second-pass references. This can be advantageous in the situation where the reference-flow genomes are assemblies with no pre-existing multiple alignment (e.g. VCF file) describing their relationship. Algorithms for calculating genome-scale multiple alignments are resource intensive[24,25] and yield a more complex structure compared to a pairwise alignment. Reference flow's use of pairwise alignments also helps to solve an "N+1" problem; adding one additional reference to the second pass requires only that we index the new genome and obtain an additional whole-genome alignment (or otherwise infer such an alignment, e.g. from a VCF file) to the standard reference. This flexibility could be important in the coming era where long and ultra-long sequencing reads allow us to build many high quality human genome assemblies.

While we explored methods involving a single initial reference and 5 second-pass reference, we can also consider a wider class of possible architectures. For instance, considering that our method consistently performs worst on the AFR super population, we could imagine building a deeper "tree" of AFR-covering references. A read aligning poorly to the second-pass reference representing the AFR super population could, in a third pass, be aligned to an array of references for specific populations within AFR. We can imagine more complex architectures as well, leading to a general notion of a "reference flow graph" where nodes represent references and directed edges indicate which references might be attempted next. Whether a read should be forwarded along an edge would be dictated by a (possibly edge-specific) rule that uses alignment score, mapping quality, whether the alignment overlapped some genomic region, or other factors.

Since reference flow is essentially a "wrapper" that can be placed around an existing aligner, Bowtie 2 could be replaced by a different linear aligner such as BWA-MEM or even a graph aligner such as vg. It is even possible for different nodes in the graph to use different alignment tools. Since the wrapper is written using Snakemake[26], it is easily deployable both in one-sample single-computer scenarios and in scenarios involving many samples or a collection of networked computers.

In the future, it will be important to benchmark reference-flow methods when larger structural variants are included in the references. Structural variants have a disproportionate effect on alignment quality[27,28]. In principle are not difficult to include in the reference flow framework, though our lift over procedure is not currently robust enough to handle more complex structural variants like inversions or rearranges.

## 5 Acknowledgments

Part of this research project was conducted using computational resources at the Maryland Advanced Research Computing Center (MARCC).

## 6 Funding

NC, BS, TM and BL were supported by NIH grant R01GM118568 to BL and by NSF grant IIS-1349906 to BL.

## 7 Availability of data and materials

The reference flow software is available at: `https://github.com/langmead-lab/reference_flow` under the open source MIT license. The experiments described in this paper are available at: `https://github.com/langmead-lab/reference_flow-experiments` under the open source MIT license. Software versions used in the experiments are specified in Table S1.

## 8 Authors' contributions

NC, BS, and BL designed the method. NC and BS wrote the software and performed the experiment. TM wrote the liftover software. SI designed and performed the reference bias experiment. NC, BS, TM, and BL wrote the manuscript. All authors read and approved the final manuscript.

## 9 Ethics approval

Not applicable.

## 10 Consent for publication

Not applicable.

## 11 Competing interests

The authors declare that they have no competing interests.

## 12 Additional Files

### 12.1 Additional file 1 — Supplementary information

Contains Tables S1–S4, and Figures S1–S4. PDF, 206 KB.

## References


1. Church, D. M., Schneider, V. A., Steinberg, K. M., Schatz, M. C., Quinlan, A. R., Chin, C. S., Kitts, P. A., Aken, B., Marth, G. T., Hoffman, M. M., *et al.* Extending reference assembly models. *Genome Biol.* **16,** 13 (Jan. 2015).
2. Brandt, D. Y., Aguiar, V. R., Bitarello, B. D., Nunes, K., Goudet, J. & Meyer, D. Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. *G3: Genes, Genomes, Genetics* **5,** 931–941 (2015).
3. Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y. & Pritchard, J. K. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25,** 3207–3212 (2009).
4. Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology* **7** (2011).
5. Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M. O., Choudhury, A., *et al.* The African genome variation project shapes medical genetics in Africa. *Nature* **517,** 327 (2015).
6. Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature biotechnology* (2018).

7. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **37,** 907–915 (2019).

8. Rakocevic, G., Semenyuk, V., Lee, W.-P., Spencer, J., Browning, J., Johnson, I. J., Arsenijevic, V., Nadj, J., Ghose, K., Suciu, M. C., *et al. Fast and accurate genomic analyses using genome graphs* tech. rep. (Nature Publishing Group, 2019).

9. Rautiainen, M. & Marschall, T. GraphAligner: Rapid and Versatile Sequence-to-Graph Alignment. *BioRxiv,* 810812 (2019).

10. Grytten, I., Rand, K. D., Nederbragt, A. J. & Sandve, G. K. Assessing graph-based read mappers against a novel baseline approach highlights strengths and weaknesses of the current generation of methods. *BioRxiv,* 538066 (2019).

11. Pritt, J., Chen, N.-C. & Langmead, B. FORGe: prioritizing variants for graph genomes. *Genome biology* **19,** 220 (2018).

12. Shukla, H. G., Bawa, P. S. & Srinivasan, S. hg19KIndel: ethnicity normalized human reference genome. *BMC genomics* **20,** 459 (2019).

13. Dewey, F. E., Chen, R., Cordero, S. P., Ormond, K. E., Caleshu, C., Karczewski, K. J., Whirl-Carrillo, M., Wheeler, M. T., Dudley, J. T., Byrnes, J. K., *et al.* Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS genetics* **7,** e1002280 (2011).

14. Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., Abecasis, G. R., *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (Oct. 2015).

15. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9,** 357 (2012).

16. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).

17. Holtgrewe, M. Mason: a read simulator for second generation sequencing data (2010).

18. Lowy-Gallego, E., Fairley, S., Zheng-Bradley, X., Ruffier, M., Clarke, L., Flicek, P., Consortium, 1. G. P., *et al.* Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. *Wellcome Open Research* **4** (2019).

19. Ballouz, S., Dobin, A. & Gillis, J. A. Is it time to change the reference genome? *Genome Biol.* **20,** 159 (Aug. 2019).

20. Liu, X., MacLeod, J. N. & Liu, J. iMapSplice: Alleviating reference bias through personalized RNA-seq alignment. *PloS one* **13** (2018).

21. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv,* 531210 (2019).

22. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).

23. Yuan, S. & Qin, Z. *Read-mapping using personalized diploid reference genome for RNA sequencing data reduced bias for detecting allele-specific expression* in *2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops* (2012), 718–724.

24. Garriga, E., Di Tommaso, P., Magis, C., Erb, I., Mansouri, L., Baltzis, A., Laayouni, H., Kondrashov, F., Floden, E. & Notredame, C. Large multiple sequence alignments with a root-to-leaf regressive method. *Nature biotechnology* **37,** 1466–1470 (2019).

25. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology* **7** (2011).

26. Köster, J. & Rahmann, S. Snakemake – a scalable bioinformatics workflow engine. *Bioinformatics* **28,** 2520–2522 (2012).

27. Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M. P., Chavan, S., Vergara, C., Ortega, V. E., *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature genetics* **51,** 30–35 (2019).

28. Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., Dougherty, M. L., Nelson, B. J., Shah, A., Dutcher, S. K., *et al.* Characterizing the major structural variant alleles of the human genome. *Cell* **176,** 663–675 (2019).

29. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

30. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).