Molecular Ecology Resources

WILEY

MOLECULAR ECOLOGY RESOURCES

RESOURCE ARTICLE OPEN ACCESS

Benchmarking Imputed Low Coverage Genomes in a Human Population Genetics Context

Gludhug A. Purnomo[1,2,3] | João C. Teixeira[1,4,5,6] | Herawati Sudoyo[3] | Bastien Llamas[1,2,7,8] | Raymond Tobler[1,4,9]

[1]Australian Centre for Ancient DNA, School of Biological Sciences, University of Adelaide, Adelaide, South Australia, Australia | [2]Centre of Excellence for Australian Biodiversity and Heritage (CABAH), School of Biological Sciences, University of Adelaide, Adelaide, South Australia, Australia | [3]Genome Diversity and Disease Laboratory, Mochtar Riady Institute for Nanotechnology, Tangerang, Indonesia | [4]Evolution of Cultural Diversity Initiative, Australian National University, Canberra, Australian Capital Territory, Australia | [5]Australian Research Council Centre of Excellence for Australian Biodiversity and Heritage (CABAH), School of Culture, History and Language, College of Asia and the Pacific, Australian National University, Canberra, Australia | [6]Centre for Interdisciplinary Studies (CEIS20), University of Coimbra, Coimbra, Portugal | [7]Indigenous Genomics, Telethon Kids Institute, Adelaide, South Australia, Australia | [8]National Centre for Indigenous Genomics, Australian National University, Canberra, Australian Capital Territory, Australia | [9]Australian Research Council Centre of Excellence for Indigenous and Environmental Histories and Futures (CIEHF), School of Culture, History and Language, College of Asia and the Pacific, Australian National University, Canberra, Australian Capital Territory, Australia

Correspondence: Gludhug A. Purnomo (gludhug.purnomo@adelaide.edu.au) | Raymond Tobler (ray.tobler@anu.edu.au)

Received: 19 July 2024 | **Revised:** 2 May 2025 | **Accepted:** 5 June 2025

Handling Editor: Alana Alexander

Funding: This work was supported by Australian Research Council, CE1701000015, DE190101069, IN180100017.

Keywords: bioinfomatics/phyloinfomatics | genomics/proteomics | molecular evolution | population genetics—empirical

ABSTRACT

Ongoing advances in population genomic methodologies have recently enabled the study of millions of loci across hundreds of genomes at a relatively low cost, by leveraging a combination of low-coverage shotgun sequencing and innovative genotype imputation methods. This approach has the potential to provide abundant genotype information at low costs comparable to another widely used cost-effective genotyping approach—that is, SNP panels—while avoiding potential issues related to loci being ascertained in distantly related populations. Nonetheless, the wide adoption of imputation methods in humans and other species is currently constrained by the lack of publicly available reference panels that capture diversity representative of the target genomes—though the recent development of 'joint' imputation approaches, which allow genetic information from the target population to be used in genotype calling, may potentially mitigate this shortcoming. Here, we assess the performance of multiple genotyping approaches on eight low coverage genomes (range ~3× to ~5×) sourced from different Indonesian populations—including a joint imputation approach that leverages 248 additional low coverage genomes (mean ~2.4×) from related populations. The inclusion of these related genomes in the joint imputation process resulted in more accurate genotype calls and produced population genetic inferences with similar accuracy but improved precision compared to pseudohaploid calls—even though the reference panel was only weakly representative of the target genomes. These results highlight the enormous potential of joint imputation to enable economical population genetic research for taxa that are currently poorly represented in publicly available reference panels.

1 | Introduction

The past decade of population genomic research has seen a notable expansion in the taxonomic breadth of published studies in conjunction with an inflation in the number of individual genomes under study. Because costs associated with generating deeply sequenced genomes (≥ 30× coverage) at a population scale remain high, much of this growth has been driven by the availability of more affordable methods that provide reduced genomic sampling. In studies of human evolution and population genetic history, SNP arrays have been a particularly popular low-cost strategy for gathering population genomic data. Further, the increasing availability of large (e.g., (McCarthy et al. 2016; Taliun et al. 2021)) and region-specific (e.g., (Flanagan et al. 2024; Li et al. 2021)) panels of high-quality human genomes have made it possible for researchers to further expand the number of usable SNPs through genotype imputation against these genome panels. Despite the growth in publicly available resources, however, the widespread use of imputed SNP array datasets is currently hindered by inherent biases toward variants that are common in the populations used to design the arrays (which tend to be Eurasian groups)—a bias that can be amplified when imputing against a reference panel that lacks genomes representative of the target population (Lachance and Tishkoff 2013). Accordingly, while imputed SNP array datasets are an attractive option to economically generate millions of genotypes for human groups with Eurasian ancestry, this approach is less effective for other populations that remain poorly represented in existing genomic resources.

In the past decade, new imputation methods have emerged that are able to work with low coverage whole genome sequencing (WGS) data (e.g., Beagle (Browning et al. 2018), Minimac2 (Fuchsberger et al. 2015), Impute2 (Howie et al. 2012), GLIMPSE (Rubinacci et al. 2021)) in addition to SNP arrays. Imputed WGS datasets are less prone to ascertainment bias while remaining a cost effective way of generating large numbers of SNPs, and have become popular in genome-wide association studies (GWAS) where they outperform imputed SNP array data, especially at rare variants (CONVERGE consortium 2015; Gilly et al. 2016). Moreover, the development of 'joint' imputation methods has allowed genetic data from the target population genomes to be used alongside reference panel information in the imputation process—for example, Beagle (Browning et al. 2018), STITCH (Davies et al. 2016), GLIMPSE (Rubinacci et al. 2021)—a feature that facilitates accurate genotype calls even when representative genomes are missing from available reference panels (Rubinacci et al. 2021).

Another economical approach to population genomic inference that is capable of overcoming ascertainment biases makes use of genotype likelihoods (GLs) in place of definitive (i.e., 'hard') genotype calls (Korneliussen et al. 2014). GL approaches directly incorporate genotyping uncertainty into statistical inference procedures, with GL-based versions for many widely used population genetic analyses now available (Lou et al. 2021). Despite these advantages, GL approaches remain less popular than hard genotype calls in population genomic applications, possibly as a result of the unavailability of GL scores in large publicly available datasets as well as researchers' widespread familiarity with hard calls. The shortcomings associated with SNP array and GL-based approaches suggest that imputed low coverage WGS datasets could become an increasingly popular option for population genomics researchers in the near future – particularly if high genotyping accuracy is possible even when target populations are poorly represented in available reference panels. Despite this promise, detailed benchmarks for imputed genotype calls from low coverage WGS data in populations poorly represented among available reference panels are scarce, while systematic exploration of the efficacy of joint imputation for analyses of demographic and evolutionary history remain absent (Lou et al. 2021).

In this study, we evaluate the performance of jointly imputed low coverage human genomes obtained from eight different Indonesian populations, assessing the quality of genotyping calls along with inferences made in three widely used statistical procedures for inferring population structure and history—that is, PCA (Patterson et al. 2006; Price et al. 2006), ADMIXTURE (Alexander et al. 2009) and *f*4 statistics (Patterson et al. 2012). Because Indonesian populations currently lack suitable high quality genomic resources, the imputation was performed using a reference panel only weakly representative of the target genomes; however, we were able to assess potential performance gains by supplementing the eight target genomes with a cohort of ~250 additional low coverage genomes from related Indonesian groups. We show that the addition of these related genomes greatly improves both the number and accuracy of called genotypes and produces robust population genetic inferences with comparable accuracy to pseudohaploid calls (where the allelic state is determined by randomly sampling a single sequence read), while attaining even higher precision. Our results highlight the capacity for joint imputation to preserve genotype information in low coverage genomes, especially when a large set of related target genomes is used, emphasising the enormous potential of this economical approach for population genomic research in the future.

2 | Material and Methods

2.1 | Sample Collection and Ethics

The genetic data for this project comes from 256 individuals from 11 different populations across the Wallacea Archipelago (i.e., Kei, $n = 20$; Aru, $n = 23$; Tanimbar, $n = 22$; Seram, $n = 27$; Ternate, $n = 30$; Sanana, $n = 19$; Daa [in Central Sulawesi], $n = 22$; and Rote-Ndao, $n = 29$) and West Papua (Keerom, $n = 26$; Mappi, $n = 11$; and Sorong, $n = 27$). Permission to conduct the research was granted by the National Agency for Research and Innovation, under the auspices of the Indonesian State Ministry of Research and Technology. Informed consent for all 256 individuals was obtained for the collection and use of all biological samples during community visits that were overseen by the Indonesian Genome Diversity Project (IGDP) team, following the Protection of Human Subjects protocol established by Eijkman Institute Research Ethics Commission (EIREC). The study is also approved by The University of Adelaide Human Research Ethics Committee (Ethics approval no. H-2020-211).

2.2 | Whole Genome Sequencing, Read Processing and Alignment

DNA was extracted from whole blood samples for all 256 samples at the Eijkman Institute for Molecular Biology Jakarta using the Gentra Puregene Blood Core Kit C (QIAGEN) following the manufacturer's protocol. DNA sequence libraries were prepared using the Nextera DNA Flex Library Preparation Kit (Illumina) following the recommended protocol. After quantifying DNA concentrations for each sample using the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific), all 256 samples were multiplexed into a single pool and submitted to 150 bp paired-end sequencing across three lanes of the Illumina NovaSeq S4 flowcell. To obtain high (~30×) coverage samples for comparison, 8 of the 256 samples were chosen for sequencing on a single NovaSeq S4 lane, with one sample coming from each of the following populations (sample ID in parentheses along with specific island/region of origin if this is not denoted in the population name)—Seram (HLU013); Daa (Sulawesi; KAL007); Rote (RTE045); Sanana (TNT160); North Maluku (TNT172); Aru (ARU-LRK007); Sorong (West Papua; SRG059); and Keerom (West Papua; KRM048).

For all samples sequenced at high coverage, raw sequence reads in the fastq format were pre-processed with fastp to remove adapters and trim poly-G and poly-X tails, with the first 20 and last 5 nucleotides being trimmed if they fell below a quality threshold of 20 (Chen et al. 2018). The sequencing reads were then mapped and processed following the recently published protocol of the Human Genome Diversity Panel (HGDP) dataset, as outlined in Bergström et al. (Bergström et al. 2020). Briefly, processed reads were mapped to the human reference genome GRCh 38 (hg38) using BWA mem v0.7.17 with the -T 0 parameter (H. Li 2013). Mapped sequencing reads were sorted and duplicated reads marked using biobambam2 (Tischler and Leonard 2014), with nucleotide bases recalibrated using baseRecalibrator from the GATK software suite v3.5 (McKenna et al. 2010). These pre-processing and mapping protocols were also used for the low coverage genomes, with the exception that reads from all low coverage genomes were merged into a single file using samtools v1.9 (Li et al. 2009) prior to the sorting and duplicate marking step of the merged reads.

2.3 | Determining 'Truth' Genotype Set Using High Coverage Sequencing Data

Of the 256 individuals, eight were selected for both low- and high-coverage sequencing (see Table S1), with the latter being used to determine a set of 'true' genotypes against which the complementary low coverage genomes were assessed. To obtain a set of high quality 'truth' genotypes, we replicated the relevant protocols used for the recently published HGDP dataset, as outlined in (Bergström et al. 2020). Briefly, single nucleotide polymorphisms (SNPs) and indel variants were called using GATK Haplotypecaller and GenotypeGVCFs (Poplin et al. 2017). Genotypes were set to missing if the genotype quality (GQ) was equal to or lower than 20, or the coverage depth (DP) was equal to or greater than 1.65 times the genome-wide coverage for each sample. Next, the GATK Variant Quality Score Recalibration tool (VSQR) was used to compute call quality annotations (QD, MQRankSum,ReadPosRankSum,FS,MQ,VQSRMODE)toSNPs (files: hapmap_3.3.hg38.vcf.gz, 1000G_omni2.5.hg38.vcf.gz and 1000G_phase3.snps.highconfidence.hg38.vcf.gz) and indels (files: Mills_and_1000G_gold_standard.indels.hg38.vcf.gz and Homo_sapiens-assembly38.known_indels.vcf.gz) (all files available from: gatk.broadinstitute.org/hc/en-us/articles/360035890811-Resource-bundle). Regions with excess heterozygosity (ExcHet) were calculated on a per-allele basis using the bcftools fill-tags plugin (Danecek et al. 2021). All SNPs with a VQSR score below −8.3929, or indels with VQSR score below −1.0158 and ExcHet value of at least 60 (corresponding to a p-value of 10^{-6}) were set to missing.

2.4 | Genotyping Approaches on Low Coverage Genomes

The following four classes of genotype calls were assessed for all eight low coverage WGS genomes:

1. *Naive genotypes; denoted in results by 'Naive' label.*

For each of the eight tested individuals, standard genotype calls were made for all SNPs in the associated truth set. SNPs with genotyping quality (RGQ and GQ) of at least 20 were retained, with all other SNPs set as missing.

2. *Imputed genotypes jointly called on the eight target samples; denoted by 'Impute_8' label.*

Joint imputation was performed on the low coverage genomes of eight individuals that were also sequenced at high coverage using GLIMPSEv1.1 (Rubinacci et al. 2021) using a reference panel comprising 3202 phased genomes from the 1000 Genomes Project sample collection (Byrska-Bishop et al. 2022). For each imputed low coverage genome, SNPs with posterior genotype probabilities (GP) less than 90% were excluded from the final set of genotype calls for that individual.

3. *Imputed genotypes jointly called on all 256 low coverage samples; denoted by 'Impute_all' label.*

These genotypes were produced using the same joint imputation process as the Impute_8 genotypes, with an additional 248 low coverage samples being jointly imputed in combination with the eight tested individuals. Because the additional 248 genomes either come from the same source population as the eight tested genomes or exhibit significant shared ancestry (Purnomo et al. 2024), this approach could potentially improve genotyping accuracy in the eight tested individuals by leveraging genetic information from related individuals. Again, genotypes with posterior genotype probabilities (GP) less than 90 were excluded.

4. *Pseudohaploid calls; denoted in results by 'Pseudohap' label.*

For each low coverage genome from the eight tested individuals, a single read was randomly sampled for each non-missing SNP observed in the corresponding high coverage genome, creating 'pseudohaploid' calls at these sites. This is a standard practice in ancient DNA studies, where endogenous DNA yields are low, and is known to produce unbiased analyses at the cost of

halving the potential information available at each locus (Green et al. 2010). Pseudohaploid calls were made using the sequenceTools software (https://github.com/stschiff/sequenceTools.git), which is widely used in paleogenomic research, with only reads having both base and mapping quality ≥30 being used in the random sampling process.

For all low coverage genomes, variant discovery was performed using Haplotypecaller with parameters "-ERC GVCF" and "-includeNonVariantSites" to ensure that monomorphic sites were retained, and intermediate gVCFs were generated for individual samples. GATK CombineGVCFs was used to amalgamate all gVCFs on a population basis, with joint variant calling on each amalgamated gVCF being performed using GenotypeGVCFs, thereby allowing variant information from all samples in each population to be used in individual genotype calls. This step produces VCF files with Phred-scale Likelihood (PL), a normalised form of genotype likelihood, that are used in the subsequent imputation process.

Imputation was performed following the protocols outlined on the official GLIMPSE github repository website using default parameterisations (https://odelaneau.github.io/). First, genomic 'chunks' were defined by running the GLIMPSE_chunk algorithm on the 3202 phased genomes from the Thousand Genomes Project (TGP) that were used as the reference panel, with each chromosome being processed in parallel to expedite computation. Genotype imputation was performed for each resulting chromosome chunk using GLIMPSE_phase and the resulting VCF files for each chunk were then merged together using GLIMPSE_ligate. All imputed loci with genotype probabilities (GP) lower than 0.9 were set to missing using BCFtools and ignored in subsequent analyses.

2.5 | Performance Metrics

Genotype concordance was quantified for each of the eight tested low-coverage genomes as the proportion of true positive genotypes amongst all SNPs with genotypes called in the complementary truth set. Because this measure does not account for chance concordance events, we evaluated a second concordance metric (i.e., the imputation quality score; (Lin et al. 2010)) that makes this correction. Two classes of discordant genotypes were evaluated that are based on the number of allelic mismatches between the inferred and 'true' genotype at each SNP (Watowich et al. 2023).

We also benchmarked the performance of the four different genotyping approaches in three widely used population genetic methods: principal component analysis (PCA) using the smartpca function from EIGENSOFT v7.1.2 (Patterson et al. 2006; Price et al. 2006), ancestry estimation using ADMIXTURE v1.3.0 (Alexander et al. 2009) and *f*4 statistics using the ADMIXTOOLS2 v2.0.0 package (Maier et al. 2023; Maier and Patterson 2024). For these analyses, the high coverage genomes were merged with publicly available genomes from the Simons Genome Diversity Project (~300 genomes from 142 different ethnic groups; SGDP (Mallick et al. 2016)), along with data from Indonesia (Jacobs et al. 2019) and New Guinea (Malaspinas et al. 2016), to create a comparative global dataset (Table S1). Because all genomes in this global dataset had been mapped to an earlier reference genome version (GRCh37 with decoy sequences), all SNPs were converted to GRCh38 coordinates using the liftover tool available in Genozip v.12.0.34 (https://genozip.readthedocs.io/dvcf.html; (Lan et al. 2021; Lan et al. 2022)) with the relevant chain file obtained from the UCSC Genome Browser. In this merged dataset, SNPs missing in more than 5% of the combined samples, or having a minor allele frequency less than 1% across all samples, were removed using PLINK v.1.987 (Purcell et al. 2007), leaving a total of 5,166,352 SNPs available for further analysis.

The merged and masked dataset was further pruned to remove SNP windows exhibiting moderate to strong levels of linkage disequilibrium (LD) (i.e., $r > 0.4$). LD was measured across sliding windows containing 200 SNPs, with new measurements taken every 25 SNPs, using PLINK v.1.987 (i.e., parameter indep-pairwise 200 25 0.4; (Lazaridis et al. 2016)), which reduced the number of remaining SNPs to 528,617. Both PCA and ADMIXTURE analyses were performed using this final pruned SNP set, with *f*4 statistics measured on the final merged SNP set both before and after pruning.

For the PCA, the first 10 principal components (PCs) were estimated for the combined set of global samples and eight high coverage genomes using the smartpca function with no outlier removal step. Genotype calls from the eight low coverage samples were projected onto these 10 PCs. To test the optimal fit between low coverage genotypes and the truth set, Euclidean distances were measured between the low and high coverage genotypes for each sample.

For the ADMIXTURE analysis, ancestry components from $K = 3$ to $K = 12$ were estimated using the combined set of global samples and eight high coverage genomes, with cross validation indicating eight latent ancestry clusters provided the optimal fit for the genetic structure observed in our global dataset (Figures S1 and S2). The inferences from the optimal ADMIXTURE run were subsequently used to estimate the representation of the eight ancestry components in the low coverage genotype data. The Euclidean distance between the values of the eight inferred ancestry components was measured between low coverage genotypes and the truth set in order to infer the optimal genotype calls for each sample.

Finally, *f*4 statistics of the form *f*(Truth, Target = Low coverage; Test = Papuan or East Asian, Africa) were computed using the ADMIXTOOLS2 function *qpDstat* with f4mode = 'YES' (Maier et al. 2023). Under this *f*4 configuration, perfect agreement between high and low coverage genotypes would result in $f4 = 0$, with the magnitude of the *f*4 statistic expected to increase as the low coverage genotypes become less consistent with the truth set. The direction of the *f*4 statistic can also be informative about the causal nature of the underlying discrepancies and is discussed further in the results. In these tests, the combined set of Papuan Highlander samples (38 samples) and East Asian samples (32 samples) in the global dataset were used as proxies for Papuan and East Asian ancestry, respectively, with four Mbuti samples from the SGDP project being used to represent the African ancestry (Mallick et al. 2016). Separate analyses were performed on the complete and LD-pruned SNP sets.

3 | Results

We generated benchmarks for eight genomes that were sourced from different Indonesian populations in the Wallacean Archipelago (i.e., Seram [HLU013], Sulawesi [KAL007], Rote [RTE045]; Sanana [TNT160], Ternate [TNT172], Aru [LRK007]) and West Papua (i.e., Sorong [SRG059]; and Keerom [KRM048]) (see Table S1 for complete list). The mean coverage of the eight high coverage genomes used to generate the 'true' genotypes for each individual ranged from ~21× to ~39× (mean = 28.7×). The comparative low coverage genomes attained between ~3× and ~5× coverage (mean = 3.4×), with the extended set of 256 low coverage genomes used in the Impute_all approach ranging between ~1× and ~11× (mean = 2.4×) (Table S2). Merging the eight high coverage samples with the global dataset resulted in a joint variant set with 5,989,969 SNPs, with between ~5.27 M and ~5.53 M of these SNPs (i.e., ~88% to ~92%; Table S3) resulting in genotype calls for each of the eight tested individuals.

3.1 | Genotype Missingness

A fundamental utility of imputation is the recovery of genotypes at SNPs that would otherwise be missing when using naïve genotyping due to insufficient sequencing coverage. As expected, the proportion of missing genotypes was uniformly high across the eight low coverage genomes when naïve genotyping was used, with no calls being made for more than two thirds of the > 5 M SNPs in all eight samples (range 68.7% to 93.1%). In contrast, the proportion of missing genotypes decreased to below one third (range 31.2% to 24.3%) when joint imputation was performed on the eight low coverage genomes (i.e., Impute_8 approach) and impacted less than 10% of SNPs (range 6.6% to 9.4%) when the extended target cohort was used (i.e., Impute_8 approach). Notably, all approaches, including pseudohaploid calls, showed a relatively linear increase in genotype missingness as coverage decreased–though the Impute_all method tended to produce fewer missing genotypes than the pseudohaploid approach when coverage levels dropped below 4×. This result implies that joint imputation may have improved ability to recover allelic information in low coverage genomes even when using a distantly related reference panel—provided that a sizable cohort of target genomes is used.

3.2 | Genotype Accuracy

While joint imputation is able to recover a high proportion of genotypes in a reference panel, robust usage in downstream population genomic analyses also requires that genotype calls are highly accurate. To assess the accuracy of the different genotype calling methods, for each of the eight tested individuals we classified called genotypes as concordant when both alleles matched with those in the truth set, or as discordant genotypes otherwise. For discordant genotypes, we further distinguished cases where mismatches occurred at one allele (i.e., homozygotes vs. heterozygotes) or at both alleles (i.e., homozygotes of different classes), as the latter class has the highest potential to distort subsequent population genetic analyses (Watowich et al. 2023). To facilitate analogous truth set comparisons for pseudohaploids, we limited our comparisons to SNPs that were homozygous in the truth set and treated pseudohaploid calls at these loci as homozygous for a randomly sampled allele.

When examining the accuracy of the different genotyping approaches, the benefits of imputation for low coverage samples are once again clearly evident through their vastly improved concordance levels (> 97.6% for Impute_8 and > 99.3% for Impute_all) relative to naive genotype calls (between 55% and 80%; Figure 1). Indeed, the Impute_all approach displayed concordance levels approaching those observed for pseudohaploid calls (> 99.9%; Figure 1) and—while the majority of discordant genotypes were single allele mismatches for all four genotyping approaches—the lowest frequency of double mismatches is observed for Impute_all genotypes (< 0.01% for Impute_all vs. < 0.13% for Impute_8 and < 0.07% for pseudohaploids). Notably, the low concordance values observed for the naive genotyping were almost entirely due to large numbers of heterozygous loci that were incorrectly called as homozygous reference genotypes (Figure S3 and Table S3). While imputed genotypes also have their lowest precision at heterozygous sites—indicative of the general difficulty in imputing heterozygotes from low coverage data—these values achieved appreciably higher concordance (> 95% for Impute_8 and > 98% for Impute_all across all eight tested individuals; Figure S3) than naive calls.

The improvements in genotyping accuracy from imputation become even more apparent when using the imputation quality score (IQS), which accounts for chance concordance events (Lin et al. 2010). Naive genotypes produced scores that are ~20% lower than standard concordance measures, with IQS values ranging between 30% and 70%, whereas imputed genotypes had IQS values only a few percent lower at most than the corresponding concordance value for all eight tested individuals (Figure 1).

The accuracy of all genotype calling approaches also exhibits strong dependencies upon sequencing coverage, though this dependency was substantially reduced for imputed genotypes. When using naive genotyping, IQS values increase by ~20.67% per unit of coverage, whereas the Impute_all approach produced unit-wise increases of ~0.22% over the same coverage range, a 100-fold reduction in dependency that approached the values observed for pseudohaploid calls (~0.08%) (Figure 1). Sequencing coverage also had a major impact on the proportion of uncalled genotypes (i.e., missingness) in each genotyping approach, with the Impute_all approach showing the weakest dependency overall (i.e., missingness decreasing ~10.70% per unit of increase of coverage for naive genotypes, vs. ~1.25% and 7.22% for Impute_all and pseudohaploids, respectively) (Figure 1). As a result, the Impute_all approach returned between 94.3% and 97.2% of all callable SNPs in each individual, a significant improvement from the Impute_8 and pseudohaploid approaches (ranges: 71.9%–78.3% and 81.1%–96.1%, respectively), particularly at the lowest coverages (i.e., ~1×).

When assessing the accuracy of the imputed genotypes against the allele frequencies of all > 5 million callable SNPs for each individual, the Impute_all method once again outperformed the two other biallelic genotyping methods. For the Naive and

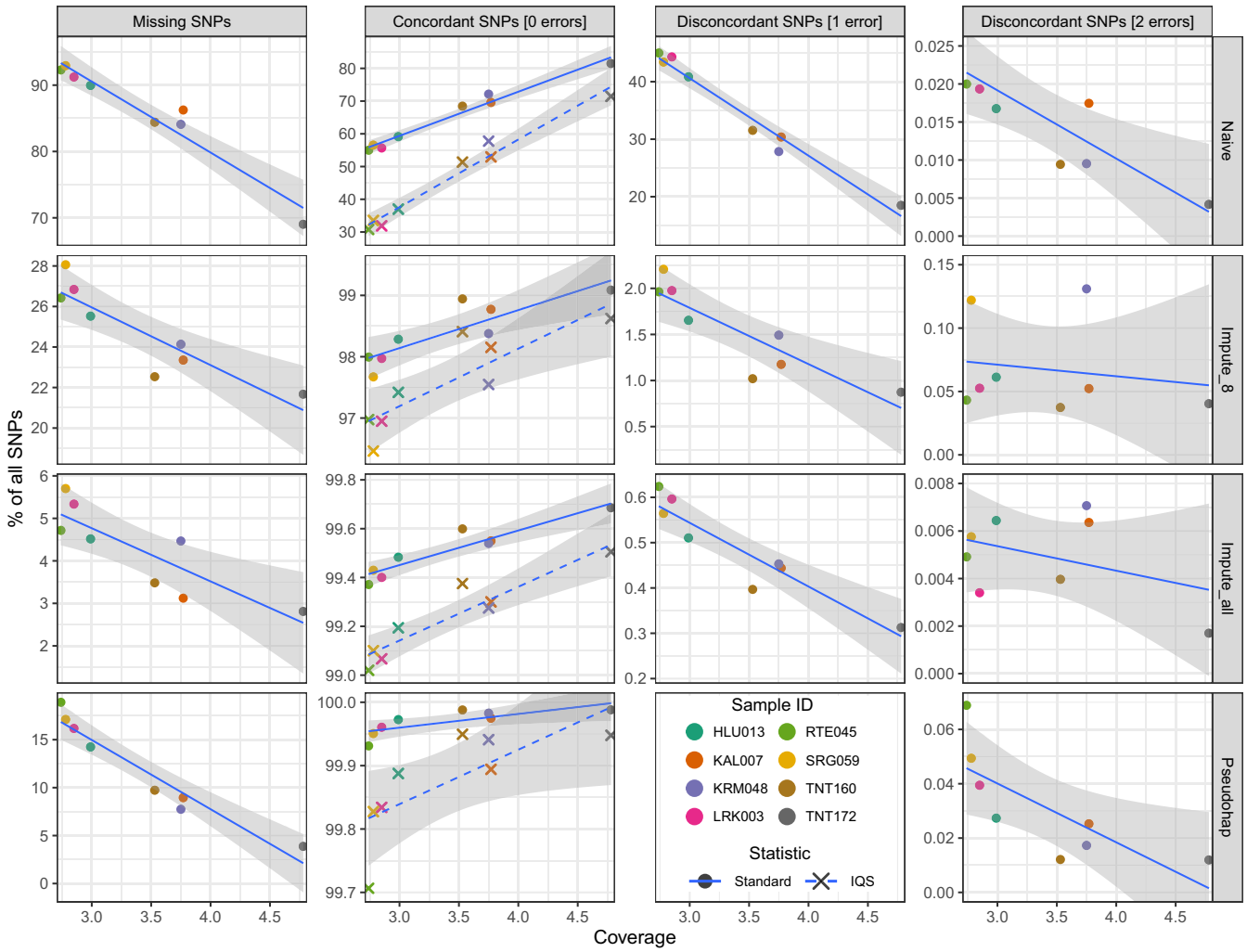

FIGURE 1 | Percentage of ~6 million genotypes called as missing, concordant, or discordant relative to sequencing coverage for the eight low coverage samples (see key) using four different genotyping methods (separate facet rows). Concordance was measured using a standard approach and a method that corrects for chance concordance events (IQS; see key), while discordance was measured according to the number of allelic 'errors' between the true and inferred genotypes. SNP concordance tended to improve proportionately with coverage, with discordance and missingness both exhibiting negative linear trends (blue lines; see key). Note that pseudohaploid calls were only compared against homozygote genotypes in the truth set, such that all discordant SNPs are classified as doubleton errors.

Impute_8 approaches, genotypes were least accurate at intermediate allele frequencies (Figure 2), as both approaches suffer from poor heterozygote calling accuracy and these genotypes are most frequent at intermediate frequencies (Figure S3). In contrast, genotypes called by the Impute_all approach are least accurate when one of the two alleles is rare in the reference panel, though this approach still manages to achieve consistently higher accuracies across all frequencies relative to the other imputation approaches (e.g., achieving >99% IQS values for minor allele frequencies in excess of 2% in the folded spectrum, vs. >95% and >23% for the Impute_8 and for Naive approaches, respectively).

Taken together, our results demonstrate that joint imputation is a particularly effective means of making highly accurate genotype calls across millions of sites, achieving low levels of genotype missingness and high levels of accuracy that are comparable to pseudohaploid methods, while also retaining information on both alleles.

3.3 | PCA and ADMIXTURE

To investigate the performance of the genotyping approaches in PCA, genotype calls from low coverage genomes were projected onto principal component axes defined by the eight high coverage genomes (the 'truth' set) and a global human genomic dataset and their positions compared to the truth set (Figure 3; see Methods). When measuring Euclidean distances between the projected and truth set genotypes, the pseudohaploid calls consistently produced the closest match to the truth set in the first PC (Figure 4), though the Impute_all method tended to be the best performing across all PCs, with similar results obtained after rescaling the distance by the eigenvalue associated with each PC (which captures the variation contributed to each PC and therefore places more emphasis on the first few PCs). Accordingly, while the Impute_all method provides the best performance overall, its advantage over the pseudohaploid approach is largely due to having improved

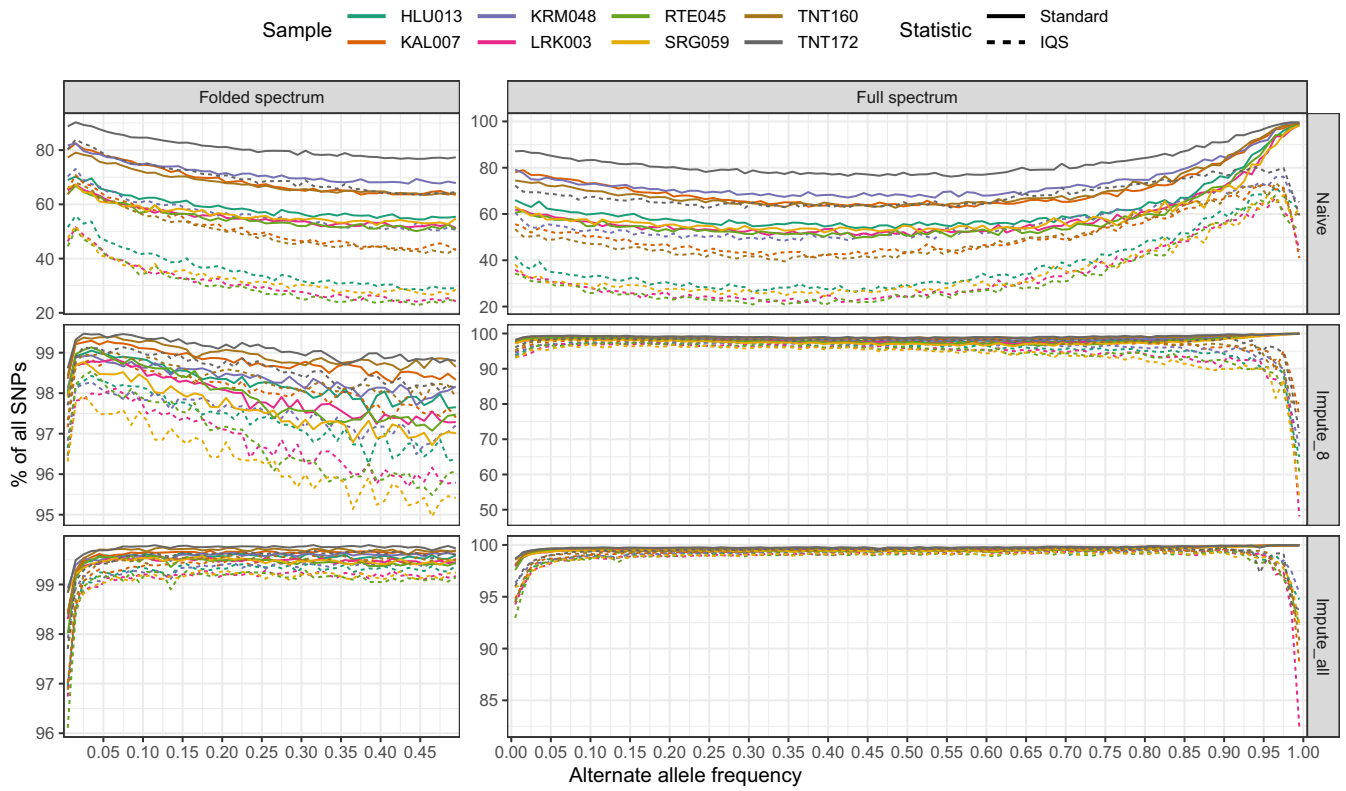

FIGURE 2 | Genotype concordance relative to the allele frequency in the reference panel for the three different imputation approaches (panel rows). Concordance was measured using a standard approach and a method that corrects for chance concordance events (IQS; see key). Results are shown for the folded and full frequency spectra (panel columns). Concordance decreases appreciably when using IQS to rule out chance concordance events, and this decline is particularly notable for imputed variants where the alternate allele is nearly fixed in the reference panel.

accuracy across lower PCs, and pseudohaploid calls may actually be preferable when visualising relationships in the first two PCs.

For the ADMIXTURE analyses, performance was evaluated by measuring the distance between the ancestry proportions estimated using the four different low coverage genotyping approaches against the values predicted by the truth set (Figure 5 and Figure S4; see Methods). Unlike the PCA results, the most accurate ADMIXTURE estimates were consistently produced by the pseudohaploid calls, though both the Impute_all and Impute_8 approaches once again performed markedly better than the naive genotype calls (Figure 6). Notably, the naive genotype calls tend to underestimate contributions of Papuan and East Asian ancestry and predict excessive amounts of African and South Asian ancestry relative to the truth set for all eight samples, reflecting PCA results where the naive genotypes are shifted away from the truth set and toward individuals with these ancestries in the first PC (Figure 6).

In contrast to the genotyping results, the accuracy of PCA and ADMIXTURE estimates was indifferent to the level of sequencing coverage (Figures 4 and 6). Only the naive genotype approach showed a significant decline in accuracy as coverage decreased, suggesting the other approaches are buffered from this effect across the narrow range of low coverage sequencing used in this study (i.e., ~3–5×).

3.4 | *f*4 Statistics

For each of the eight tested individuals, we evaluated *f*4 statistics of the form *f*(Truth, *Test* = Low coverage; *X* = Papuan or East Asian, *Y* = Mbuti), where *f*4 = 0 indicates perfect accuracy and higher magnitude *f*4 statistics indicate increasingly inaccurate genotyping calls (due to artefactual relationships between the low coverage sample and either the *X* or *Y* population). The observed *f*4 statistics reiterate the general trends observed in the PCA and ADMIXTURE analyses, with the naive genotype calls producing the largest *f*4 values, and the three other genotyping approaches producing *f*4 values close to 0 (Figure 7). Significantly positive *f*4 values (i.e., absolute standardised score greater than 3 standard errors from 0; i.e., $|Z| > 3$) were observed for the naive genotypes, regardless of whether the Papuan or East Asian populations were included as the *X* population, or whether the complete or LD-pruned SNP sets were used. This result is consistent with PCA and ADMIXTURE findings—where the naive genotypes are pulled toward African samples or show excess African ancestry, respectively—which suggests that low coverage samples may appear more 'African' relative to the truth set as more SNPs are evaluated, resulting in larger positive *f*4 values.

For the imputed genotypes, the impact of the SNP dataset on the *f*4 values depended on the composition of the *f*4 population quartet. When the *f*4 statistic was evaluated using the Papuan population in the *X* position, both imputed genotype methods

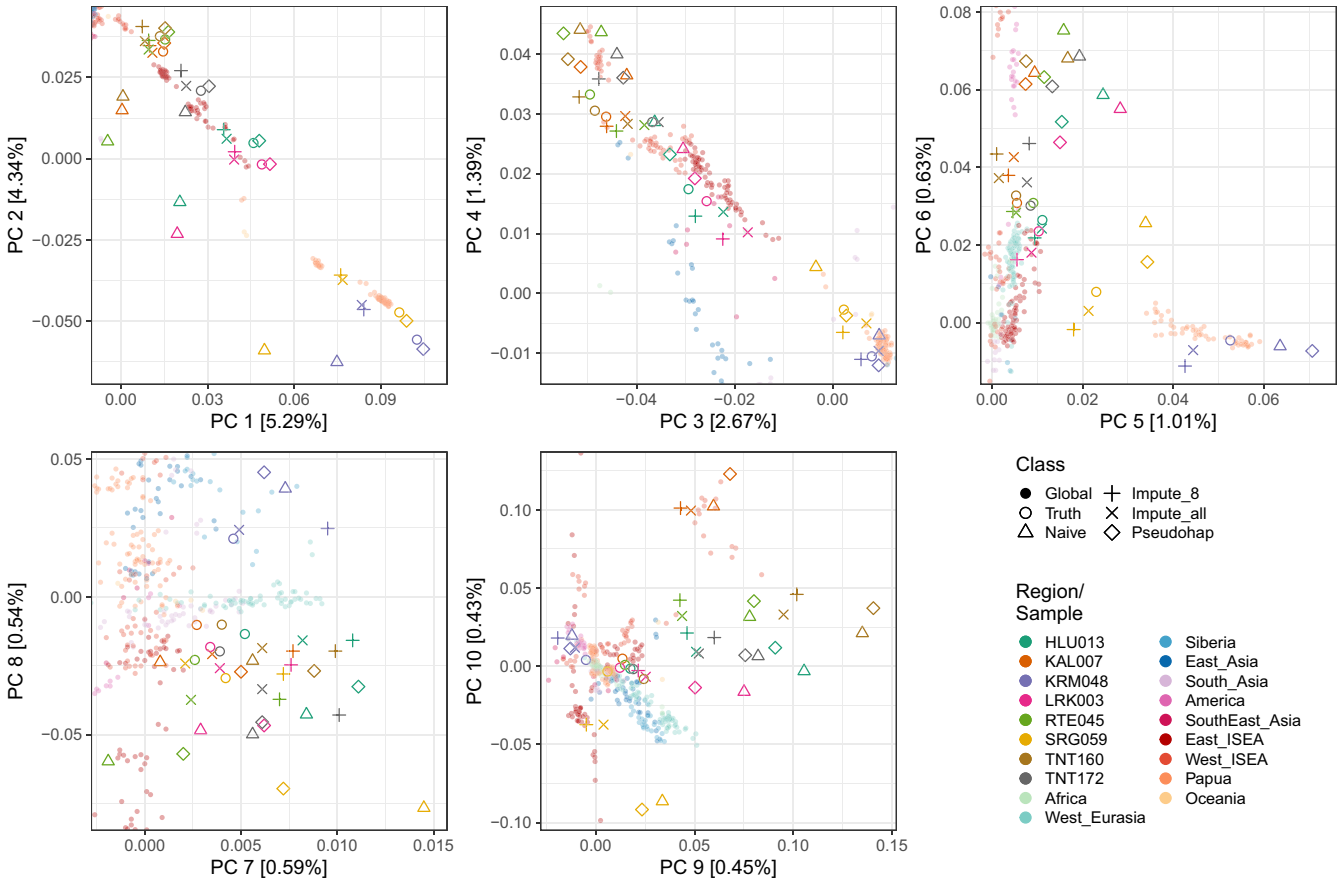

FIGURE 3 | The first 10 principal components from a PCA based on a global human dataset and the eight high coverage Wallacean and Papuan genomes. Genotypes for low coverage Wallacean and Papuan genomes were projected onto the PC space, with different symbols used for each genotyping approach (see key). To assist comparison of high and low coverage individuals, global samples lying beyond the PC space occupied by the projected samples are omitted (see Figure S4 for visualisation of the entire dataset).

returned positive *f*4 values, but these values were always lower when the full SNP set was used. In contrast, when the *X* position was occupied by an East Asian population, the sign of the *f*4 values depended on whether the pruned or unpruned SNP set was used (being +ve for the former and −ve for the latter; Figure 7). These results suggest that the imputation process tended to underestimate the true degree of Papuan ancestry amongst low coverage genomes but overestimated the East Asian component—possibly reflecting the lack of Papuan samples in the reference panel, such that increasing the number of SNPs either reduced or accentuated the bias in the *f*4 statistic, respectively. Echoing previous results, the Impute_all method tended to have *f*4 values much closer to zero than the Impute_8 method for all samples regardless of which SNP set was used, further emphasising the potential for improved population genetic estimation by leveraging genetic information from population cohorts.

Finally, the pseudohaploid calls displayed the least bias amongst all genotyping methods, having no significant *f*4 statistics amongst all evaluated scenarios (Figure 7). While being the least biased method overall, however, the pseudohaploid *f*4 statistics exhibited substantially more uncertainty than the Impute_all genotypes when using comparable SNP sets. This suggests that the pseudohaploid method was more accurate, but less precise, than the imputed data for measuring *f*4 statistics under the current testing framework.

4 | Discussion

Here we demonstrate the utility of low coverage WGS in population genomic research when using pseudohaploid or imputed genotypes, achieving high concordance (e.g., >98% for imputed genotypes across ~6 million SNPs) and robust inferences for PCA, ADMIXTURE and *f*4 analyses that align closely to the truth sets. In general, inferences from imputed genotypes, particularly those leveraging genetic information from large cohorts of related individuals in the target populations (i.e., Impute_all), were comparable to pseudohaploid calls—which are generally regarded to be highly robust to statistical artefacts resulting from SNP ascertainment (Patterson et al. 2012)—with the best choice differing depending on the statistic. Importantly, the results of the *f*4 tests in this study suggest improved accuracy of pseudohaploid calls over the imputed genotypes observed for some statistics (i.e., first few PC dimensions, ADMIXTURE, *f*4) may ultimately come at the cost of lower precision. In other words, while pseudohaploids produce results that are closer to the expected value on average than the imputed genotypes, they also exhibit more variation, meaning that a single estimate from

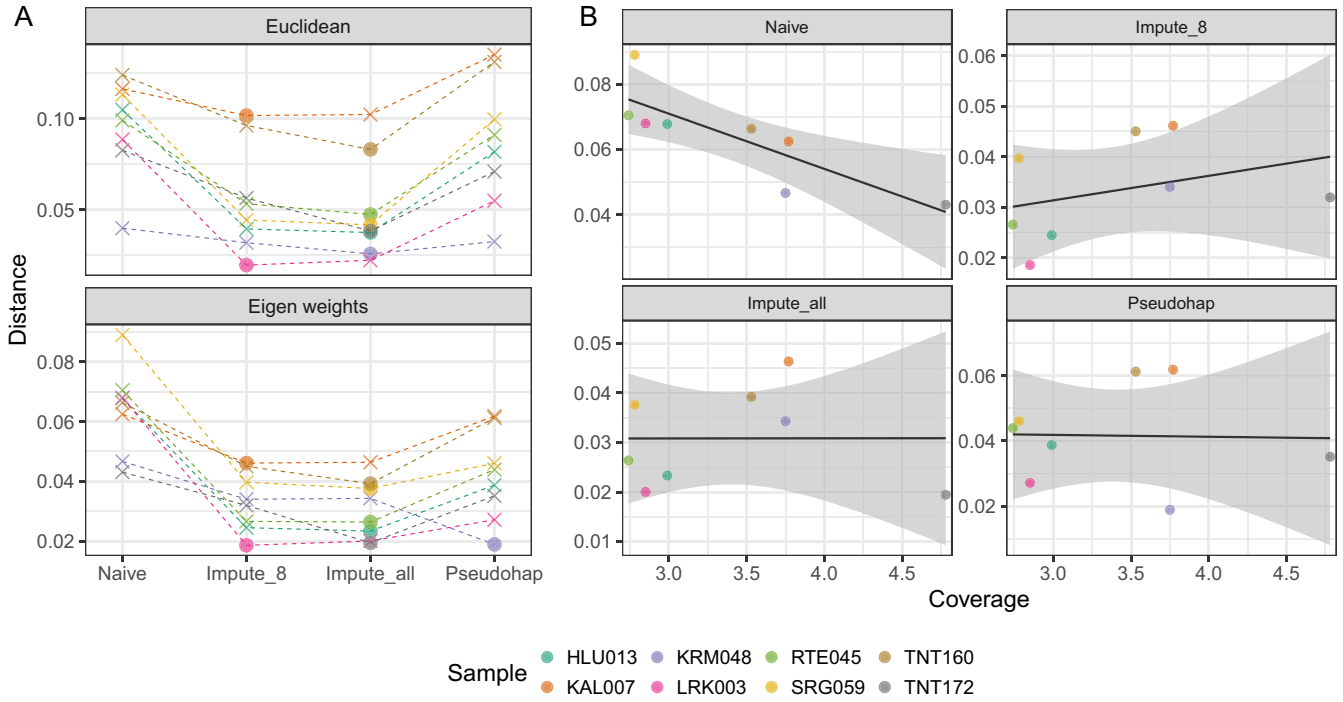

FIGURE 4 | Distances between truth sets and each low coverage genotyping method calculated across PCs 1 to 10. (A) Distances were measured directly (i.e., Euclidean; top panel) or reweighted according to the eigenvalue of each dimension (bottom panel), with minimal distances (circle symbols) typically being achieved by the Imput_all approach. (B) A negative linear dependency is present between the eigen-weighted distance (i.e., accuracy) and coverage for naive genotypes but is absent for other approaches.

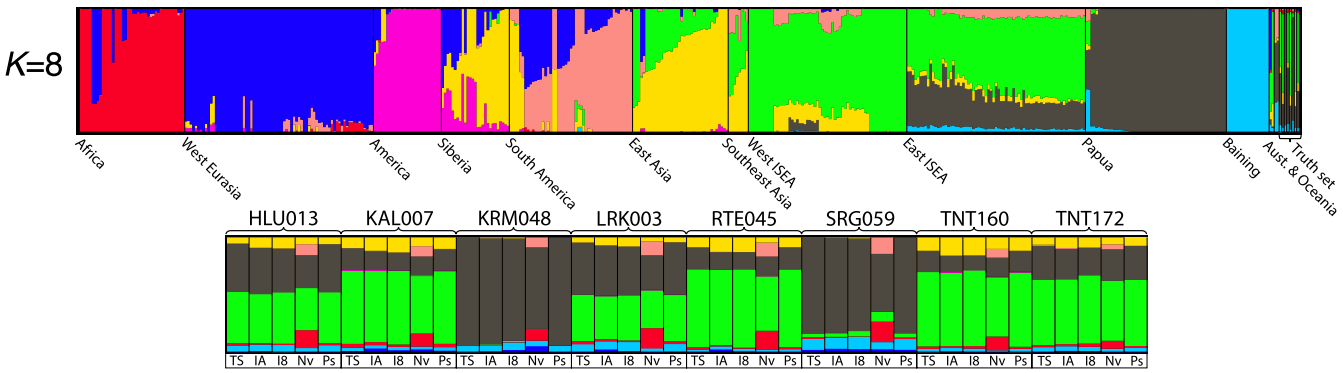

FIGURE 5 | Optimal ADMIXTURE results revealed eight ancestry components amongst worldwide human samples and eight high coverage Papuan genomes (top panel). These components were used to estimate ancestry proportions for the four low coverage genotyping methods (bottom panel; IA, Impute_All; I8, Impute_8; Nv, Naive; Ps, Pseudohaploid), with the truth set estimates (TS) also being included to facilitate comparison.

pseudohaploid data will frequently be further from the truth than an estimate made from imputed genotype data. This trade-off between precision and accuracy is a fundamental property of statistical estimation, with lowered precision of pseudohaploid data likely stemming from only having half the information of standard biallelic genotypes and the imputed genotypes being prone to bias that is dependent upon the composition of the reference data.

Importantly, analyses of GLIMPSE imputation on simulated genomic datasets (Rubinacci et al. 2021) indicate that more precise genotype imputation should be possible than what is reported here, simply by increasing the number of samples in the target population. While several hundred target samples were used in the current study, the number of samples per target population was relatively small (~20 to ~30). Thus, it remains to be seen what further improvements are possible when joint imputation is performed on 100s to 1000s of samples from the same population. In particular, it is important to understand the trade-off between sequencing coverage and sample size in target populations when using GLIMPSE—i.e., whether to sequence a handful of samples at reasonably high coverage versus sequencing more samples at proportionately lower coverage. This question has been examined using another low coverage WGS imputation algorithm that does not require a reference panel (i.e., STITCH; (Davies et al. 2016)), though, to our knowledge, this question has not been thoroughly explored for genotypes called from a joint imputation algorithm.

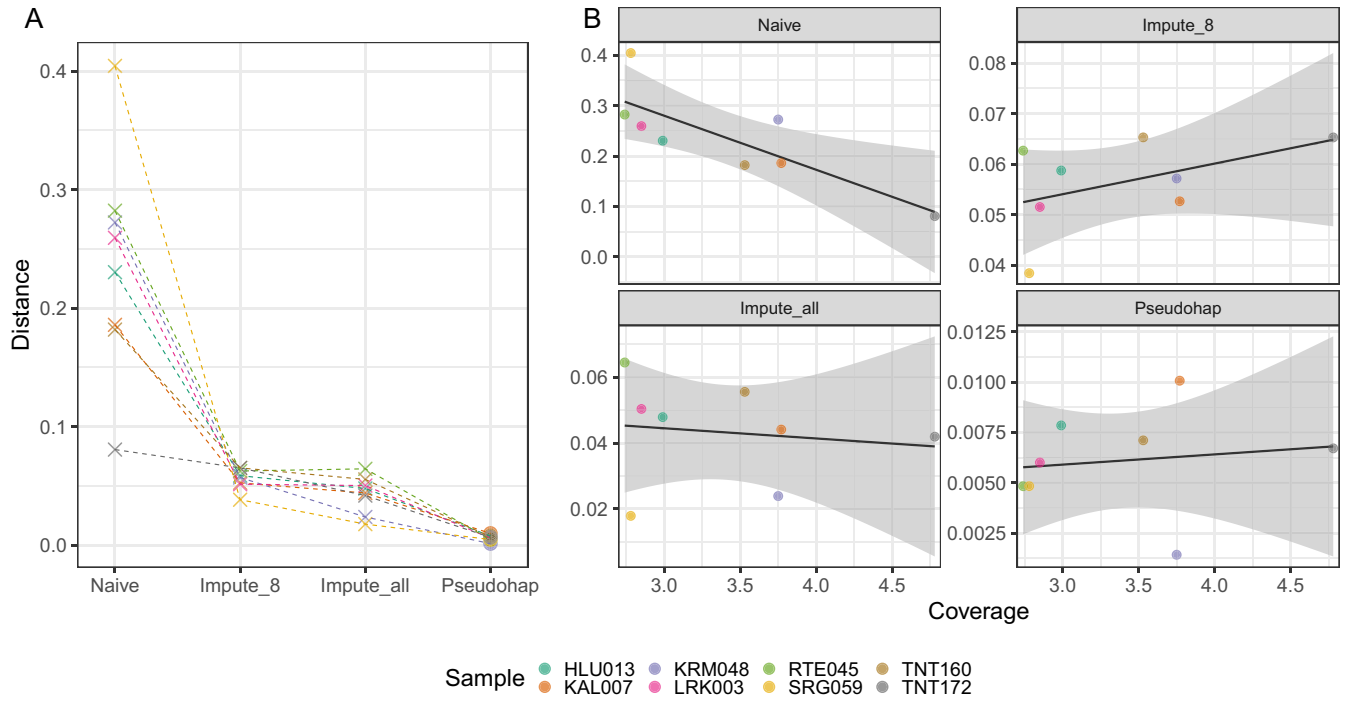

FIGURE 6 | Euclidean distances between truth sets and each low coverage genotyping method calculated for optimal ADMIXTURE results (i.e., $K = 8$). (A) Euclidean distance was minimal (circles) for the pseudohaploid method. (B) A negative linear dependency is present between the eigen-weighted distance (i.e., accuracy) and coverage for naive genotypes but is absent for other approaches.

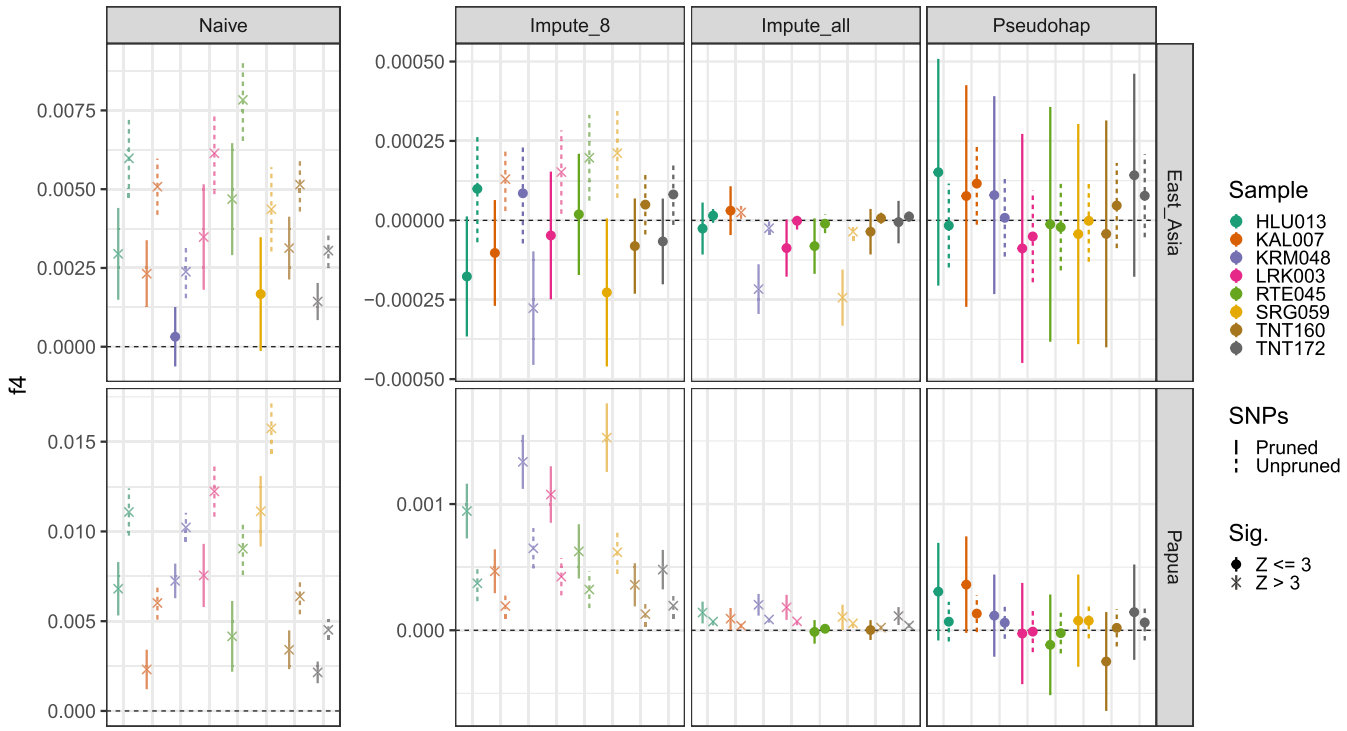

FIGURE 7 | The *f*4(Truth set, *Test* = Low coverage; *X* = Papua|East Asia, *Y* = Mbuti) statistics calculated for four different genotyping approaches using both LD-pruned and full (i.e., unpruned) SNP sets (see key). For this population configuration, $f4 = 0$ indicates perfect accuracy, and higher magnitude *f*4 statistics indicate increasingly inaccurate genotyping calls (due to artifactual relationships between the low coverage sample and either the *X* or *Y* population). Results for LD-pruned and unpruned SNP sets are included with confidence intervals included (see key)—absolute standardised *f*4 values less than 3 are indicated as circles; those greater than three are indicated as crosses.

Since undertaking this study, several new genomic panels have been released that cater for specific regional groups, including the Southeast Asian-Specific Reference Panel (SEARP; (Cengnata et al. 2024)) that comprises ~2,500 genomes that are more representative of the samples in our study. As expected, analyses of Southeast Asian genomes showed that imputed genotype calls

that leveraged the SEARP panel were much improved compared to those made from less representative resources such as the 1KGP reference panel, with rare alleles especially benefitting, though the authors did not explore imputation performance under a joint imputation approach. Accordingly, a key outstanding question concerns whether the gains obtained from including population-specific genomes in the reference panel can offset those acquired from using a large number of target population genomes under joint imputation. More work on this question is needed to help researchers to decide whether they should invest in supplementing existing reference panels with high quality population-specific genomes or allocate their resources to producing a larger number of target population genomes.

5 | Conclusion

Studies of increasingly larger genomic datasets are continuing to emerge as sequencing costs steadily decrease, though standard laboratory budgets mean that SNP panels have remained the most common way to produce human population genomic data for cohorts exceeding dozens of individuals. In recent years, the development of increasingly efficient imputation algorithms has led to genotype imputation on low coverage genomes becoming more cost effective than SNP panels, with recent improvements further facilitating highly accurate rare allele calls when using suitable reference datasets with 100,000's of genomes (Rubinacci et al. 2021). Our study highlights that the benefits of imputation can also be extended to populations that lack representation in available reference panels, by using joint imputation to draw upon the information available in the target genomes. When this information is obtained from a large set of related individuals, GLIMPSE's joint imputation facility produces highly accurate genotype calls and population genomic inferences, and the broad applicability of this approach suggests that it will play a growing role in population genomic research in humans and other taxa in the future.

Acknowledgements

Open access publishing facilitated by The University of Adelaide, as part of the Wiley - The University of Adelaide agreement via the Council of Australian University Librarians.

Disclosure

Benefit Statement: The data used in this study were collected as part of the research activities of the Indonesian Genome Diversity Project (IGDP). The IGDP has an overarching goal to build genomic resources for the development of precision medicine in Indonesia, an initiative that is further empowered by producing a robust understanding of Indonesia's genetic history and population genetic structure. Following the IGDP's informed consent process, all participants have been advised of these goals in acknowledgement that their consented genomes will facilitate population genetic and medical research that will ultimately benefit public health programmes in their community and across Indonesia more generally.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All requests to access the sequences presented in this work are managed through the Data Access Committee of the official data repository (accession EGAS50000000447) at the European Genome-phenome Archive (EGA; https://ega-archive.org/). Data Citation: GenomeAsia100K Consortium, The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature 576, 106-111 (2019). S. Carlhoff, et al., Genome of a middle Holocene hunter-gatherer from Wallacea. Nature 596, 543-547 (2021). S. Mallick, et al., The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538, 201-206 (2016). A-S. Malaspinas, et al., A genomic history of Aboriginal Australia. Nature 538, 207-214 (2016). G. S. Jacobs, et al., Multiple Deeply Divergent Denisovan Ancestries in Papuans. Cell 177, 1010-1021.e32 (2019).

References

Alexander, D. H., J. Novembre, and K. Lange. 2009. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." *Genome Research* 19, no. 9: 1655–1664.

Bergström, A., S. A. McCarthy, R. Hui, et al. 2020. "Insights Into Human Genetic Variation and Population History From 929 Diverse Genomes." *Science* 367, no. 6484: eaay5012. https://doi.org/10.1126/science.aay5012.

Browning, B. L., Y. Zhou, and S. R. Browning. 2018. "A One-Penny Imputed Genome From Next-Generation Reference Panels." *American Journal of Human Genetics* 103, no. 3: 338–348.

Byrska-Bishop, M., U. S. Evani, X. Zhao, et al. 2022. "High-Coverage Whole-Genome Sequencing of the Expanded 1000 Genomes Project Cohort Including 602 Trios." *Cell* 185, no. 18: 3426–3440.

Cengnata, A., L. Deng, W.-S. Yap, et al. 2024. "A Genotype Imputation Reference Panel Specific for Native Southeast Asian Populations." *NPJ Genomic Medicine* 9, no. 1: 47.

Chen, S., Y. Zhou, Y. Chen, and J. Gu. 2018. "Fastp: An Ultra-Fast All-In-One FASTQ Preprocessor." *Bioinformatics (Oxford, England)* 34, no. 17: i884–i890.

CONVERGE consortium. 2015. "Sparse Whole-Genome Sequencing Identifies Two Loci for Major Depressive Disorder." *Nature* 523, no. 7562: 588–591.

Danecek, P., J. K. Bonfield, J. Liddle, et al. 2021. "Twelve Years of SAMtools and BCFtools." *GigaScience* 10, no. 2: giab008. https://doi.org/10.1093/gigascience/giab008.

Davies, R. W., J. Flint, S. Myers, and R. Mott. 2016. "Rapid Genotype Imputation From Sequence Without Reference Panels." *Nature Genetics* 48, no. 8: 965–969.

Flanagan, J., X. Liu, D. Ortega-Reyes, et al. 2024. "Population-Specific Reference Panel Improves Imputation Quality for Genome-Wide Association Studies Conducted on the Japanese Population." *Communications Biology* 7, no. 1: 1665.

Fuchsberger, C., G. R. Abecasis, and D. A. Hinds. 2015. "minimac2: Faster Genotype Imputation." *Bioinformatics* 31, no. 5: 782–784.

Gilly, A., G. R. Ritchie, L. Southam, et al. 2016. "Very Low-Depth Sequencing in a Founder Population Identifies a Cardioprotective APOC3 Signal Missed by Genome-Wide Imputation." *Human Molecular Genetics* 25, no. 11: 2360–2365.

Green, R. E., J. Krause, A. W. Briggs, et al. 2010. "A Draft Sequence of the Neandertal Genome." *Science (New York, N.Y.)* 328, no. 5979: 710–722.

Howie, B., C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis. 2012. "Fast and Accurate Genotype Imputation in Genome-Wide Association Studies Through Pre-Phasing." *Nature Genetics* 44, no. 8: 955–959.

Jacobs, G. S., G. Hudjashov, L. Saag, et al. 2019. "Multiple Deeply Divergent Denisovan Ancestries in Papuans." *Cell* 177, no. 4: 1010–1021.e32.

Lachance, J., and S. A. Tishkoff. 2013. "SNP Ascertainment Bias in Population Genetic Analyses: Why It Is Important, and How to Correct It." *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 35, no. 9: 780–786.

Lan, D., R. Tobler, Y. Souilmi, and B. Llamas. 2021. "Genozip: A Universal Extensible Genomic Data Compressor." *Bioinformatics (Oxford, England)* 37, no. 16: 2225–2230.

Lan, D. M., G. Purnomo, R. Tobler, Y. Souilmi, and B. Llamas. 2022. "Genozip Dual-Coordinate VCF Format Enables Efficient Genomic Analyses and Alleviates Liftover Limitations." *bioRxiv*: 500374. https://doi.org/10.1101/2022.07.17.500374.

Lazaridis, I., D. Nadel, G. Rollefson, et al. 2016. "Genomic Insights Into the Origin of Farming in the Ancient Near East." *Nature* 536, no. 7617: 419–424.

Li, H. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs With BWA-MEM." *arXiv* 1303. http://arxiv.org/abs/1303.3997.

Li, H., B. Handsaker, A. Wysoker, et al. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics (Oxford, England)* 25, no. 16: 2078–2079.

Li, L., P. Huang, X. Sun, et al. 2021. "The ChinaMAP Reference Panel for the Accurate Genotype Imputation in Chinese Populations." *Cell Research* 31, no. 12: 1308–1310.

Lin, P., S. M. Hartz, Z. Zhang, et al. 2010. "A New Statistic to Evaluate Imputation Reliability." *PLoS One* 5, no. 3: e9697.

Lou, R. N., A. Jacobs, A. Wilder, and N. O. Therkildsen. 2021. "A Beginner's Guide to Low-Coverage Whole Genome Sequencing for Population Genomics." *Molecular Ecology* 30, no. 23: 5966–5993. https://doi.org/10.22541/au.160689616.68843086/v4.

Maier, R., P. Flegontov, O. Flegontova, U. Işıldak, P. Changmai, and D. Reich. 2023. "On the Limits of Fitting Complex Models of Population History to f-Statistics." *eLife* 12: e85492. https://doi.org/10.7554/eLife.85492.

Maier, R., and N. Patterson. 2024. "Admixtools: Inferring Demographic History From Genetic Data." https://github.com/uqrmaie1/admixtools.

Malaspinas, A.-S., M. C. Westaway, C. Muller, et al. 2016. "A Genomic History of Aboriginal Australia." *Nature* 538, no. 7624: 207–214.

Mallick, S., H. Li, M. Lipson, et al. 2016. "The Simons Genome Diversity Project: 300 Genomes From 142 Diverse Populations." *Nature* 538, no. 7624: 201–206.

McCarthy, S., S. Das, W. Kretzschmar, et al. 2016. "A Reference Panel of 64,976 Haplotypes for Genotype Imputation." *Nature Genetics* 48, no. 10: 1279–1283.

McKenna, A., M. Hanna, E. Banks, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data." *Genome Research* 20, no. 9: 1297–1303.

Patterson, N., P. Moorjani, Y. Luo, et al. 2012. "Ancient Admixture in Human History." *Genetics* 192, no. 3: 1065–1093.

Patterson, N., A. L. Price, and D. Reich. 2006. "Population Structure and Eigenanalysis." *PLoS Genetics* 2, no. 12: e190.

Poplin, R., V. Ruano-Rubio, M. A. DePristo, et al. 2017. "Scaling Accurate Genetic Variant Discovery to Tens of Thousands of Samples." *bioRxiv* 201178. https://doi.org/10.1101/201178.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies." *Nature Genetics* 38, no. 8: 904–909.

Purcell, S., B. Neale, K. Todd-Brown, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *American Journal of Human Genetics* 81, no. 3: 559–575.

Purnomo, G. A., S. Kealy, S. O'Connor, et al. 2024. "The Genetic Origins and Impacts of Historical Papuan Migrations Into Wallacea." *Proceedings of the National Academy of Sciences of the United States of America* 121, no. 52: e2412355121.

Rubinacci, S., D. M. Ribeiro, R. J. Hofmeister, and O. Delaneau. 2021. "Efficient Phasing and Imputation of Low-Coverage Sequencing Data Using Large Reference Panels." *Nature Genetics* 53, no. 1: 120–126.

Taliun, D., D. N. Harris, M. D. Kessler, et al. 2021. "Sequencing of 53,831 Diverse Genomes From the NHLBI TOPMed Program." *Nature* 590, no. 7845: 290–299.

Tischler, G., and S. Leonard. 2014. "Biobambam: Tools for Read Pair Collation Based Algorithms on BAM Files." *Source Code for Biology and Medicine* 9, no. 1: 13.

Watowich, M. M., K. L. Chiou, B. Graves, et al. 2023. "Best Practices for Genotype Imputation From Low-Coverage Sequencing Data in Natural Populations." *Molecular Ecology Resources* 25, no. 5: e13854. https://doi.org/10.1111/1755-0998.13854.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.