Evaluating genetic-ancestry inference from single-cell transcriptomic datasets

Jianing Yao[1,2,3], Steven Gazal[1,2,4]

[1] Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
[2] Center for Genetic Epidemiology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
[3] Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA
[4] Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA
Correspondence should be addressed to S.G. (gazal@usc.edu).

Abstract

Characterizing the ancestry of donors in single-cell transcriptomic studies is crucial to ensure genetic homogeneity, reduce biases in analyses, identify ancestry-specific regulatory mechanisms and their downstream roles in disease, and ensure that existing datasets are representative of human genetic diversity. While these datasets are now widely available, information on the ancestry of donors is often missing, hindering further analysis. Here, we propose a framework to evaluate methods for inferring genetic-ancestry from genetic polymorphisms detected in single-cell sequencing reads. We demonstrate that widely used tools (e.g., ADMIXTURE) provide accurate inference of genetic-ancestry and admixture proportions, despite the limited number of genetic polymorphisms identified and imperfect variant calling from sequencing reads. We infer genetic-ancestry for 401 donors from ten Human Cell Atlas datasets and report a high proportion of donors of European ancestry in this resource. For researchers generating single-cell transcriptomic datasets, we recommend reporting genetic-ancestry inference for all donors and generating datasets that represent diverse ancestries.

Keywords: Genetic-ancestry inference; admixture estimation; single-cell transcriptomics; variant calling; Human Cell Atlas

Introduction

Gene expression can vary across individuals from diverse human ancestries due to heterogeneous genetic and environmental backgrounds [1–10], potentially contributing to disparities in disease mechanisms and susceptibility [10]. Therefore, it is critical to characterize the ancestry of donors in transcriptomic studies for several reasons. First, it helps reduce confounding issues by ensuring genetic homogeneity of the dataset. For example, as expression quantitative trait loci (eQTLs) can have different allele frequencies across populations, ancestry can induce false positives in differential expression analyses. Second, it enables the identification of ancestry-specific regulatory mechanisms, such as eQTLs with ancestry-specific effect sizes or alleles that are enriched in one population. Finally, it allows researchers to interpret genome-wide association study (GWAS) results using ancestry-matched functional data, ensuring accurate interpretation of underlying biological mechanisms.

Single-cell transcriptomic sequencing technologies, such as single-cell RNA sequencing (scRNA-seq) and single-nucleus RNA sequencing (snRNA-seq), have emerged as powerful techniques to quantify gene expression within individual cells or their nuclei, and have enabled the characterization of transcriptome profiles across multiple human cell types [11]. Hundreds of datasets across diverse tissues and conditions are now publicly available, for example, through the Human Cell Atlas [11] (HCA) portal [12]. However, the lack of donor ancestry information impedes further investigative work, such as integration with non-European GWAS. Detecting genetic polymorphisms in sequencing reads offers a unique opportunity to characterize genetic-ancestry (defined as the biological descendance from various ancestral groups) of donors without access to additional genetic data [13,14]. However, the accuracy of this strategy remains unclear due to the limited fraction of the genome that is transcribed and the noise in single-cell/nuclei data that impacts variant calling accuracy [13].

Here, we propose a framework to evaluate methods for inferring genetic-ancestry from single-cell transcriptomic data, using individuals from both the Human Genome Diversity Project [15] (HGDP) and the 1000 Genomes Project [16] (1kGP) as a reference dataset [17]. In this work, we focus on global genetic-ancestry, defined as genome-wide ancestry proportions for each donor, rather than local ancestry, which assigns ancestry labels to specific genomic segments. We demonstrate that widely used tools (e.g., ADMIXTURE [18]) provide accurate inference of genetic-ancestry and admixture proportions, despite the limited number of variants and imperfect variant calling from sequencing reads. We inferred genetic-ancestry for 401 donors from ten HCA datasets [19–28] (including five with unreported ancestry) and highlighted a strikingly high proportion of donors of European ancestry in this resource. We recommend that researchers generating single-cell datasets document the genetic-ancestry of the donors and prioritize including donors from diverse ancestries.

Material and methods

Common SNP calling in single-cell transcriptomic data

We identified genetic variants from single-cell transcriptomic data using GATK [29,30] following the best-practice pipeline recommended by Schnepp et al. [31], and as it is to our knowledge, the most widely maintained and computationally optimized tool for this purpose. This pipeline also provided the highest genotyping accuracy in snRNA-seq data compared with other methods [13]. We used the RNAseq short variant discovery workflow as the following. First, processed sequencing reads were aligned to the human hg19 reference genome using STARsolo v2.7.11a in two-pass mode to achieve better alignments around novel splice junctions (SJs). In the first pass, SJs were identified across all samples by combining unique and multi-mapped reads, and then filtered to exclude SJs supported only by multi-mapped reads or with a unique read count below two. The filtered SJ file was then used in the second pass, with the STAR output coordinate-sorted. Second, we combined the BAM files of all cells from each individual to create a pseudo-bulk BAM file for each donor. Duplicate reads were flagged using GATK's MarkDuplicates, and GATK's SplitNCigarReads was applied to reformat alignments spanning introns, ensuring compatibility with HaplotypeCaller for variant calling. Third, variants for each donor were called using GATK's HaplotypeCaller. Finally, joint genotyping of all donors from each HCA dataset was performed using GATK's GenotypeGVCFs, followed by filtering with GATK's VariantFiltration, which removed sites with low QualByDepth <2, high FisherStrand > 60, StrandOddsRatio > 3, RMSMappingQuality < 40, MappingQualityRankSumTest < -12.5, and ReadPosRankSumTest < -8.

We performed additional quality control (QC) steps to select high-quality SNPs and samples. To retain only SNPs that are informative for genetic-ancestry inference while limiting false-positive genotypes, we kept SNPs that were common (minor allele frequency (MAF) > 5%) in the QCed HGDP+1kGP dataset (see below) and were located within coding exons and

88 untranslated regions (UTRs), defined using UCSC reference files and extended by 150 base pairs
89 upstream and downstream. We used PLINK to exclude SNPs with genotype missingness greater
90 than 10% and donors with genotype missingness greater than 10%.

91 We have released an open-source pipeline implementing our framework (see **Code**
92 **availability**).

93

94 **Single-cell transcriptomics datasets**

95 To evaluate the genotype error rate of our pipeline, we leveraged 25 snRNA-seq samples
96 from GTEx, obtained from 16 donors across 8 different tissues [32]. We compared genotypes
97 inferred by our pipeline with those obtained by GTEx via whole-genome sequencing, which we
98 considered as ground truth. We also analyzed 12 individuals with both scRNA-seq data from
99 peripheral blood mononuclear cells (PBMCs) and SNP-array data [33], considering the SNP-array
100 genotypes as ground truth.

101 We next investigated the 50 HCA projects with the largest number of donors (as of July
102 2025). We report their sample size, sequencing technology, availability of FASTQ files on the
103 HCA website, tissue, and reported ancestry information in **Supp. Table 1**. We identified 15
104 projects with no or limited information about ancestry and inferred genetic-ancestry for the five of
105 these with available FASTQ files (datasets labeled Airway epithelium, Breast, Lung, Nose, and
106 Skin; **Table 1**). To enable a comprehensive evaluation of genetic-ancestry inference across
107 varying combinations of genes expressed and technical platforms, we also selected five additional
108 datasets with available FASTQ files that encompass diverse tissues and cell types, employ
109 different sequencing technologies (i.e., Smart-seq, Drop-seq, Microwell-seq, and 10x Genomics
110 3' sequencing), and include donors of different ancestries (i.e., African-American, East Asian, and
111 European) (**Table 1**). Three datasets had ancestry information available (datasets labeled
112 CD45+, Cell landscape, and Heart), while two datasets were collected from European countries

113 (England and Sweden), where donor ancestry can be reasonably inferred from population
114 demographics (datasets labeled Bone marrow and Embryo). One dataset included both scRNA-
115 seq and snRNA-seq data; individuals profiled with these different technologies were analyzed
116 separately. After QC, these ten datasets comprised 401 individuals.

117

118 **Genetic-ancestry inference approaches**

119 We considered the harmonized HGDP+1kGP dataset [17] as a reference population dataset for
120 genetic-ancestry inference and selected six genetic-ancestry groups (Africa, America, Europe,
121 Middle East, East Asia, and South Asia; note that the Finnish population was included within the
122 European group). Several QC steps were applied before downstream analysis. Specifically, we
123 excluded individuals who were outliers in principal component analyses within their reported
124 genetic-ancestry groups to obtain homogeneous reference populations. The final dataset
125 consisted of 3,481 individuals from 67 populations (**Supp. Table 2**). Further analyses were
126 restricted to common SNPs (MAF > 5%) in this dataset.

127 We considered three commonly used approaches to infer the genetic-ancestry of study
128 samples (i.e., donors from single-cell studies) using a reference dataset (i.e., HGDP+1kGP). Two
129 approaches relied on principal component analysis (PCA) performed on the HGDP+1kGP
130 dataset, with projection of donors onto the first five principal components (PCs). First, we assigned
131 each donor to the genetic-ancestry group whose centroid had the smallest Euclidean distance to
132 that donor (method labeled PCA-Distance). Second, we used a random forest classifier trained
133 on the five PCs to predict genetic-ancestry probabilities for each study sample, as previously
134 described by Koenig et al. [17] (method labeled PCA-RandomForest). Finally, we applied
135 ADMIXTURE [18] in supervised mode, using the HGDP+1kGP genetic-ancestry groups as
136 reference (e.g., K = 6) to estimate the proportions of ancestral populations for each donor. All
137 three approaches were run after linkage disequilibrium (LD) pruning performed with PLINK [34]

138 (using the --indep-pairwise option with a window size of 50 SNPs, a step size of 10 SNPs, and an
139 r^2 threshold of 0.1) on the HGDP+1kGP dataset restricted to the investigated SNPs derived from
140 HCA single-cell datasets (between 2,968 and 17,544 SNPs after LD pruning; see **Table 1**). We
141 also performed additional analyses relaxing the MAF cutoff (1% and 2%) and the r^2 threshold
142 (0.15 and 0.2).

143 Other methods exist to infer genetic-ancestry, but we did not investigate them in this study.
144 For instance, STRUCTURE [35] employs a similar probabilistic model to that of ADMIXTURE, but
145 is substantially less computationally efficient for large datasets. Furthermore, methods such as
146 RFmix [36] infer local genetic-ancestry, but rely on haplotype information and require phasing of
147 genetic data; we opted against these methods, as phasing genetic variants inferred from single-
148 cell data presents a considerable technical challenge due to data sparsity and may compromise
149 accuracy. Finally, we acknowledge the utility of recent tools that work directly with single-cell
150 measurements for variant calling and genetic-ancestry inference, such as Monopogen [13], which
151 leverages linkage disequilibrium information from external reference panels to call germline and
152 somatic SNVs from sparse single-cell profiles and support global and local genetic-ancestry
153 mapping, and scAI-SNP [14], which is trained on ancestry-informative SNPs derived from the 1000
154 Genomes Project to estimate ancestry contributions for each donor. Nonetheless, in this work,
155 we restrict our evaluations to variants called using a standard GATK-based workflow and to global
156 genetic-ancestry inference approaches based on PCA and ADMIXTURE, as this combination is
157 widely used in practice and provides a simple baseline that can be readily adopted in existing
158 population-genetic and single-cell analysis pipelines.

159

160 **Evaluating genetic-ancestry inference using HGDP+1kGP**

161 We evaluated the three genetic-ancestry inference approaches using a leave-one-population-out
162 strategy on HGDP+1kGP. Specifically, for each of the 67 HGDP+1kGP populations, we removed

163 the corresponding individuals from the reference dataset and inferred their genetic-ancestry using
164 each approach to assess the robustness of the inference methods when prior information about
165 the excluded population was unavailable. We repeated this procedure for every population. For
166 PCA-RandomForest and ADMIXTURE, we assigned each individual from the removed population
167 to the genetic-ancestry category with the highest predicted proportion. Finally, we compared the
168 inferred genetic-ancestry to the labels provided in the HGDP+1kGP dataset and evaluated the
169 inference error rate both within each genetic-ancestry group and across all individuals.

170 We evaluated each approach using different sets of genotypes. First, to define a gold
171 standard, we considered genotypes from all common SNPs in the HGDP+1kGP dataset,
172 performed LD pruning, and retained 295,434 SNPs; we labeled this genotype dataset "all-SNPs".
173 Second, to evaluate genetic-ancestry inference from SNPs identified in single-cell datasets, we
174 considered eleven distinct sets of SNPs corresponding to the LD-pruned SNPs identified in the
175 ten HCA datasets; one dataset had two sets of SNPs from scRNA-seq and snRNA-seq data,
176 leading to eleven sets of sc-SNPs. We labeled these genotype datasets "sc-SNPs". Finally, to
177 account for genotype detection error using single-cell data, we leveraged the error rate estimated
178 from the GTEx and PBMCs datasets (8%; see **Results**) and simulated an 8% genotype error rate
179 within each sc-SNPs dataset for individuals in the removed population; we labeled these genotype
180 datasets "sc-SNPs-8%error". Genotype error was simulated by randomly selecting 8% of the
181 genotypes of each individual and changing each selected genotype to an alternative value. We
182 also explored additional analyses using a range of simulated genotype error rates (see **Results**).
183 We note that all the investigated methods have negligible computation time when using sc-SNPs
184 (<1 minute for PCA and PCA-RandomForest, <5 minutes for ADMIXTURE). In comparison,
185 ADMIXTURE is more computationally intensive when using all-SNPs (~1 hour).

186 Finally, we evaluated the estimation of genetic-admixture proportions with ADMIXTURE
187 in four admixed populations from HGDP+1kGP. First, we considered African Americans from the

188 Southwest US (ASW; 68 individuals) and African Caribbeans in Barbados (ACB; 112 individuals),
189 and focused on their proportions of African and European genetic-ancestry [37]. Second, we
190 considered Mexican Ancestry individuals from Los Angeles (MXL; 96 individuals), and focused
191 on their proportions of African, American, European, and Middle Eastern genetic-ancestry. Finally,
192 we considered Balochi individuals from HGDP (21 individuals), and focused on their proportions
193 of African, European, Middle Eastern, and South Asian genetic-ancestry. We first ran
194 ADMIXTURE using the set of 295,434 LD-pruned SNPs (all-SNPs) and treated the resulting
195 genetic-ancestry proportions as the gold standard. We then reran ADMIXTURE on these
196 individuals using genotypes from each set of sc-SNPs, as well as on genotypes in which we
197 simulated an 8% genotype error rate (sc-SNPs-8%error).

198 ## Results

199 ### Estimating genotype error rate

200 We first assessed the precision of SNP detection by GATK in single-cell transcriptomic datasets
201 by comparing genotypes called in GTEx snRNA-seq data from 8 tissues and in scRNA-seq data
202 from PBMCs to those obtained via WGS and SNP-array, respectively. Across these 9 datasets,
203 we observed an average genotype error rate of 8.02% (**Figure 1** and **Supp. Table 3**). The
204 genotype error rate varied across tissue types, ranging from 5.98% in heart left ventricle samples
205 to 12.64% in skin sun-exposed lower leg samples from GTEx, while the error rate remained similar
206 across polymorphism types (e.g., SNPs with A and C alleles) or across alleles (e.g., A-to-C error)
207 (**Supp. Figure 1**). We also note that restricting SNPs to coding exons and UTRs reduced the
208 genotype error rate from 12.05% to 8.02% (**Supp. Figure 1**), confirming that this QC step helps
209 remove potential false-positive genotypes. Based on these results, we used a genotype error rate
210 of 8% for the main analyses and varied the genotype error rate from 1% to 13% in additional
211 analyses.

212

213 ### Detecting common genetic polymorphisms in HCA single-cell transcriptomic datasets

214 We next ran our pipeline on ten datasets from the HCA encompassing diverse tissues and cell
215 types, employing different sequencing technologies, and including donors with different
216 ancestries. After QC, we detected between 6,589 and 115,931 SNPs that were also present in the
217 HGDP+1kGP dataset (average: 34,682), which were further reduced to between 2,968 and
218 17,544 common SNPs (average: 8,609) after LD pruning (**Table 1**). To assess the number of
219 ancestry-informative SNPs derived from each HCA dataset, we examined SNPs that were
220 significantly associated with the first five PCs computed using all-SNPs. We observed that about
221 6,000 SNPs were associated with the first and second PCs, about 3,000 SNPs were associated
222 with the third and fourth PCs, and about 300 SNPs were associated with the fifth PC; on average,

223 this represents a ~30-fold decrease in ancestry-informative SNPs compared with using all-SNPs
224 (**Supp. Figure 2**).
225

226 **Quantifying genetic-ancestry inference accuracy**

227 We estimated genetic-ancestry inference error using a leave-one-population-out approach on the
228 HGDP+1kGP dataset; results within the six genetic-ancestry groups, along with the overall
229 average, are reported in **Figure 2** and **Supp. Table 4**. Using the all-SNPs dataset as the gold
230 standard, we observed that ADMIXTURE provided a very low error rate (0.11% across the
231 groups), lower than that obtained with PCA-Distance (0.59%) and PCA-RandomForest (4.84%).
232 When restricting these analyses to SNPs identified in single-cell datasets (sc-SNPs),
233 ADMIXTURE still yielded an extremely low error rate (0.21% across the eleven sets of sc-SNPs),
234 more than six times lower than those of PCA-Distance and PCA-RandomForest (1.37% and
235 5.05%, respectively). After simulating genotype error in these datasets (sc-SNPs-8%error), the
236 error rate for ADMIXTURE remained lower than that for PCA-Distance and PCA-RandomForest
237 (1.42%, 2.24%, and 4.30%, respectively). The higher error rate for PCA-RandomForest was
238 driven by misclassification of individuals from the Druze and Mozabite populations in the Middle
239 Eastern group. The increased error rate for ADMIXTURE from sc-SNPs to sc-SNPs-8%error was
240 driven by misclassification of individuals from the Adygei and Toscani populations, who were
241 inferred as having admixture of European and Middle Eastern genetic-ancestries (**Supp. Table**
242 **5**). We observed similar trends across the eleven sets of sc-SNPs (**Supp. Figure 3**).

243 We next investigated the impact of genotype error rate on genetic-ancestry inference and
244 found that the ADMIXTURE error rate increased exponentially, whereas the error rate for PCA-
245 based methods remained stable (PCA-RandomForest) or increased linearly (PCA-Distance)
246 (**Figure 3** and **Supp. Table 6**). The elevated ADMIXTURE inference error under high simulated

247 genotype error was driven by misclassification of individuals from European populations (Adygei,
248 Iberian, and Toscani; **Supp. Table 5**).

249 To further assess the robustness and limitations of our approach, we performed two
250 supplementary analyses. First, we investigated whether including SNPs with lower MAF and
251 relaxing the LD pruning thresholds would improve genetic-ancestry inference. We found that
252 using less stringent values for these parameters did not affect the error rate when using all-SNPs
253 and sc-SNPs; however, including SNPs with MAF < 5% increased the error rate of all methods
254 when using sc-SNPs-8%error (**Supp. Figure 4**). Second, we investigated whether SNPs
255 identified in single-cell datasets allow for more fine-grained genetic-ancestry inference.
256 Specifically, we tested whether subpopulations within an ancestry group could be inferred (e.g.,
257 identifying Japanese rather than simply East Asian genetic-ancestry) [38]. We observed that
258 although ADMIXTURE can yield an acceptable error rate within some ancestry groups when using
259 sc-SNPs, all inference approaches produced high error rates when using sc-SNPs-8%error
260 (**Supp. Figure 5** and **Supp. Figure 6**).

261 Overall, our results indicate that, although inferring continental genetic-ancestry from
262 SNPs in single-cell data yields extremely low error rates with ADMIXTURE when no genotype
263 error is assumed, the error rate increases to modest but non-negligible values when accounting
264 for an 8% genotype error rate, primarily due to misclassification of some European populations
265 as partially Middle Eastern. We therefore recommend using ADMIXTURE to infer ancestry from
266 single-cell data, while exercising caution when interpreting individuals inferred as a genetic-
267 admixture of Europeans and Middle Easterns.

268

269 **Quantifying genetic-admixture inference accuracy**

270 Donors from single-cell datasets are likely to be genetically admixed. We therefore evaluated
271 genetic-admixture inference with ADMIXTURE by leveraging admixed individuals from the ASW,

272 ACB, MXL, and Balochi populations from HGDP+1kGP (**Figure 4** and **Supp. Table 7**). Results
273 obtained with sc-SNPs were nearly identical to those obtained with all-SNPs (used here as the
274 gold standard). Results obtained with sc-SNPs-8%error were highly correlated with those from
275 all-SNPs, although they slightly underestimated some admixture proportions in certain cases. For
276 example, European genetic-admixture was well estimated in ASW and ACB populations
277 (regression slopes = 1.01 and 1.14, respectively), but underestimated in the MXL population
278 (regression slope = 0.79). We reached similar conclusions across the eleven sets of sc-SNPs
279 (**Supp. Figure 7** and **Supp. Table 8**). Overall, these results suggest that ADMIXTURE can also
280 provide robust estimates of genetic-admixture proportions.

281

282 **Characterizing the ancestry of donors from the HCA**

283 Finally, we aimed to characterize the ancestry of donors represented in the HCA by investigating
284 the 50 HCA projects with the largest numbers of donors (**Supp. Table 1**). Ancestry information
285 for donors was explicitly available for 20 datasets, and could be deduced for 3 additional datasets
286 in which the corresponding publications reported a single ancestry origin. These 23 datasets
287 contained 3,505 unique donors, including 2,283 (65.1%) of European ancestry, 806 of Asian
288 ancestry (467 East Asian, 111 South Asian, 228 unreported/other; 23.0%), and 158 African
289 American donors (4.5%); ancestry for the remaining individuals was mostly missing (5.1%). An
290 additional 12 datasets were expected to include donors from European or East Asian countries
291 based on sample demographics. Assuming that these individuals have European or East Asian
292 ancestries, the numbers of European and Asian donors would increase to 2,741 (65.2%) and
293 1,046 (24.9%), respectively, across 4,203 donors. Overall, these results indicate that donors
294 contributing to HCA projects are overwhelmingly of European and Asian ancestry. Among the 15
295 remaining projects with no or limited ancestry information, five had FASTQ files available on the
296 HCA website, allowing us to infer genetic-ancestry.

297 To benchmark our pipeline on real data, we first inferred genetic-ancestry for three
298 datasets in which ancestry information was available for 10 donors of African ancestry, 52 donors
299 of East Asian ancestry, and 66 donors of European ancestry (**Supp. Figure 8** and **Supp. Table**
300 **9**). ADMIXTURE results were largely consistent with reported ancestry, with a mean African
301 genetic-admixture proportion of 73.9% for donors of African ancestry, a mean East Asian genetic-
302 admixture proportion of 99.2% for donors of East Asian ancestry, and a mean European genetic-
303 admixture proportion of 91.0% for donors of European ancestry. Only three individuals from the
304 same dataset (Heart) had genetic-ancestry that did not match the reported ancestry. One donor
305 of African ancestry was inferred as European and one donor of European ancestry was inferred
306 as African, suggesting either sample mix-up or mislabeled ancestry. Another donor of European
307 ancestry was inferred as having American and European genetic-admixture (62.9% and 27.0%,
308 respectively). Similar conclusions were obtained with PCA-Distance and PCA-RandomForest.

309 We next applied ADMIXTURE to infer genetic-ancestry for 118 individuals from two
310 datasets generated in European countries (England and Sweden), for which European genetic-
311 ancestry is expected (**Supp. Figure 8** and **Supp. Table 9**). Although the inferred European
312 ancestry proportion was high across samples (mean = 90.3%), six individuals had <50%
313 European genetic-ancestry. Two individuals had primarily African or South Asian ancestry, and
314 one individual had a genetic-admixture of European and East Asian ancestry (57.7% and 34.3%,
315 respectively). The three remaining individuals had European and Middle Eastern genetic-
316 admixture, but we caution against overinterpreting the genetic-ancestry of these donors based on
317 our previous results. We obtained consistent conclusions with PCA-Distance and PCA-
318 RandomForest.

319 Finally, we applied ADMIXTURE to infer genetic-ancestry for 155 individuals from five
320 projects where ancestry was unreported but FASTQ files were available (**Figure 5** and **Supp.**
321 **Table 9**). Overall, we observed that these datasets were also dominated by donors with European

genetic-ancestry: the mean European genetic-admixture proportion was 75.4%, and 105 individuals had predominantly (>75%) European genetic-admixture. Across the five datasets, 16 individuals had predominantly African genetic-admixture, two had American, one had East Asian, and two had South Asian genetic-admixture. Fifteen individuals had at least 25% genetic-admixture from these ancestries. Seven individuals were inferred as nearly half European and half Middle Eastern, but we caution against overinterpreting the ancestry of these donors in light of our previous results. We obtained consistent conclusions with PCA-Distance and PCA-RandomForest.

330 ## Discussion

331 Characterizing the ancestry of donors in single-cell transcriptomic studies is essential to ensure
332 genetic homogeneity of the dataset and to identify ancestry-specific regulatory mechanisms linked
333 to disease risk. Here, we propose a framework to evaluate methods for inferring genetic-ancestry
334 from single-cell data. We considered eleven realistic sets of sc-SNPs that capture the variability
335 in SNP detection across different single-cell technologies and human tissues, and we simulated
336 genotype errors to reflect imperfect variant calling in single-cell data. We demonstrated that
337 ADMIXTURE provides accurate inference of genetic-ancestry and genetic-admixture, despite the
338 limited number of genetic polymorphisms and the presence of variant-calling errors in single-cell
339 sequencing reads. However, we recommend careful interpretation of individuals inferred as a
340 genetic-admixture of European and Middle Eastern ancestries. We anticipate that our easy-to-
341 use pipeline (see **Code availability**) will enable researchers to infer and document the genetic-
342 ancestry of their samples and facilitate the characterization of genetic-ancestry across publicly
343 available single-cell datasets.

344 We observed an over-representation of donors with European genetic-ancestry in HCA
345 datasets, consistent with observations in RNA-seq datasets [39] and genome-wide association
346 studies [40]. Although recent studies have generated large single-cell datasets using hundreds of
347 Asian donors [33,41], there remains a need to generate datasets in more genetically diverse
348 populations to ensure that existing datasets are representative of human genetic diversity. As
349 recent methods that integrate GWAS results with single-cell data from a few donors have
350 identified precise cellular processes impacting disease [42,43], generating single-cell data for even
351 a small number of non-European donors will substantially improve the interpretation of GWAS in
352 these populations.

We note some limitations of our work. First, we considered six genetic-ancestry groups that are not fully representative of human genetic diversity. In addition, categorizing genetic-ancestry oversimplifies human genetic diversity and disregards important aspects of human demographic history [44]. Second, we restricted our inferences to common SNPs identified in sequencing reads by GATK, as calling variants at known loci decreases the probability of false positives, and GATK is a widely maintained tool that provides high genotyping accuracy compared to other methods [13,31]. More sophisticated variant-calling methods [13] might improve the number of SNPs detected and/or genotype calling accuracy. However, even when simulating a conservative 8% genotype error rate on common variants, we observed that ADMIXTURE provides reliable results and can achieve extremely high genetic-ancestry inference accuracy at lower genotype error rates. Third, our empirical analyses were restricted to ten HCA datasets, selected to include projects with no or limited ancestry information and a set of additional studies spanning diverse tissues, cell types, and technologies. Applying our pipeline to the complete HCA resource (~24K samples from ~500 projects, as of November 14, 2025) would enable a comprehensive genetic-ancestry characterization of publicly available single-cell transcriptomic data. Despite these limitations, our study provides a robust framework to evaluate methods for inferring genetic-ancestry from genetic polymorphisms detected in single-cell sequencing reads.

Code availability

Our pipeline and the code to replicate our analyses are available at https://github.com/JianingYao/scRNA-seq-genetic-ancestry.

Acknowledgment

We thank members of the Gazal and Mancuso labs for helpful discussions. This research has been funded by the National Institutes of Health grant R35 GM147789.

Competing Interest

S.G reports consulting fees from Eleven Therapeutics unrelated to the present work. The other authors declare no competing interests.

381 **References**

382 1. Nédélec, Y. *et al.* Genetic Ancestry and Natural Selection Drive Population Differences in
383 Immune Responses to Pathogens. *Cell* vol. 167 657–669.e21 Preprint at
384 https://doi.org/10.1016/j.cell.2016.09.025 (2016).
385 2. Quach, H. *et al.* Genetic Adaptation and Neandertal Admixture Shaped the Immune System
386 of Human Populations. *Cell* vol. 167 643–656.e17 Preprint at
387 https://doi.org/10.1016/j.cell.2016.09.024 (2016).
388 3. Idaghdour, Y. *et al.* Geographical genomics of human leukocyte gene expression variation
389 in southern Morocco. *Nat. Genet.* **42**, 62–67 (2010).
390 4. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation
391 in humans. *Nature* **501**, 506–511 (2013).
392 5. Martin, A. R. *et al.* Transcriptome sequencing from diverse human populations reveals
393 differentiated regulatory architecture. *PLoS Genet.* **10**, e1004549 (2014).
394 6. Hughes, D. A. *et al.* Evaluating intra- and inter-individual variation in the human placental
395 transcriptome. *Genome Biol.* **16**, 54 (2015).
396 7. Melé, M. *et al.* Human genomics. The human transcriptome across tissues and individuals.
397 *Science* **348**, 660–665 (2015).
398 8. Randolph, H. E. *et al.* Genetic ancestry effects on the response to viral infection are
399 pervasive but cell type specific. *Science* **374**, 1127–1133 (2021).
400 9. Aquino, Y. *et al.* Dissecting human population variation in single-cell responses to SARS-
401 CoV-2. *Nature* **621**, 120–128 (2023).
402 10. Wang, J., Zhang, Z., Lu, Z., Mancuso, N. & Gazal, S. Genes with differential expression
403 across ancestries are enriched in ancestry-specific disease effects likely due to gene-by-
404 environment interactions. *Am J Hum Genet* **111**, 2117–2128 (2024).
405 11. Regev, A. *et al.* Science Forum: The Human Cell Atlas. (2017) doi:10.7554/eLife.27041.

406 12. HCA Data Portal. https://data.humancellatlas.org/.
407 13. Dou, J. *et al.* Single-nucleotide variant calling in single-cell sequencing data with
408 Monopogen. *Nature Biotechnology* **42**, 803–812 (2023).
409 14. Hong, S. C., Muyas, F., Cortés-Ciriano, I. & Hormoz, S. scAI-SNP: a method for inferring
410 ancestry from single-cell data. *bioRxiv* 2024.05.14.594208 (2024)
411 doi:10.1101/2024.05.14.594208.
412 15. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of
413 variation. *Science (New York, N.Y.)* **319**, (2008).
414 16. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, (2015).
415 17. Koenig, Z. *et al.* A harmonized public resource of deeply sequenced diverse human
416 genomes. *Genome Res.* **34**, 796–809 (2024).
417 18. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in
418 unrelated individuals. *Genome research* **19**, (2009).
419 19. Carraro, G. *et al.* Transcriptional analysis of cystic fibrosis airways at single-cell resolution
420 reveals altered epithelial cell states and composition. *Nature Medicine* **27**, 806–814 (2021).
421 20. Murrow, L. M. *et al.* Mapping hormone-regulated cell-cell interaction networks in the human
422 breast at single-cell resolution. *Cell Syst* **13**, 644–664.e8 (2022).
423 21. Leader, A. M. *et al.* Single-cell analysis of human non-small cell lung cancer lesions refines
424 tumor classification and patient stratification. *Cancer Cell* **39**, 1594–1609.e12 (2021).
425 22. Winkley, K. *et al.* Immune cell residency in the nasal mucosa may partially explain
426 respiratory disease severity across the age range. *Sci Rep* **11**, 15927 (2021).
427 23. Ganier, C. *et al.* Multiscale spatial mapping of cell populations across anatomical sites in
428 healthy human skin and basal cell carcinoma. *Proc Natl Acad Sci U S A* **121**, e2313326120
429 (2024).
430 24. Giustacchini, A. *et al.* Single-cell transcriptomics uncovers distinct molecular signatures of
431 stem cells in chronic myeloid leukemia. *Nature Medicine* **23**, 692–702 (2017).

25. Petropoulos, S. *et al.* Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* **167**, 285 (2016).
26. Cillo, A. R. *et al.* Immune Landscape of Viral- and Carcinogen-Driven Head and Neck Cancer. *Immunity* **52**, (2020).
27. Han, X. *et al.* Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
28. Koenig, A. L. *et al.* Single-cell transcriptomics reveals cell-type-specific diversification in human heart failure. *Nature Cardiovascular Research* **1**, 263–280 (2022).
29. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303 (2010).
30. GitHub - gatk-workflows/gatk4-rnaseq-germline-snps-indels: Workflows for processing RNA data for germline short variant discovery with GATK v4 and related tools. *GitHub* https://github.com/gatk-workflows/gatk4-rnaseq-germline-snps-indels.
31. Schnepp, P. M., Chen, M., Keller, E. T. & Zhou, X. SNV identification from single-cell RNA sequencing data. *Hum Mol Genet* **28**, 3569–3583 (2019).
32. Eraslan, G. *et al.* Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* **376**, eabl4290 (2022).
33. Perez, R. K. *et al.* Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science* **376**, eabf1970 (2022).
34. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, (2007).
35. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
36. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet* **93**, 278–288 (2013).

458 37. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across
459 Diverse Populations. *Am J Hum Genet* **107**, 788–789 (2020).
460 38. Bollas, A. E. *et al.* SNVstory: inferring genetic ancestry from genome sequencing data.
461 *BMC Bioinformatics* **25**, 1–18 (2024).
462 39. Romero, I. G. *et al.* Public RNA-seq data are not representative of global human diversity.
463 *bioRxiv* 2024.10.11.617967 (2024) doi:10.1101/2024.10.11.617967.
464 40. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health
465 disparities. *Nat Genet* **51**, 584–591 (2019).
466 41. Tian, C. *et al.* Single-cell RNA sequencing of peripheral blood links cell-type-specific
467 regulation of splicing to autoimmune and inflammatory diseases. *Nature Genetics* **56**,
468 2739–2752 (2024).
469 42. Zhang, M. J. *et al.* Polygenic enrichment distinguishes disease associations of individual
470 cells in single-cell RNA-seq data. *Nat Genet* **54**, 1572–1580 (2022).
471 43. Jagadeesh, K. A. *et al.* Identifying disease-critical cell types and cellular processes by
472 integrating single-cell RNA-sequencing and human genetics. *Nat Genet* **54**, 1479–1492
473 (2022).
474 44. Lewis, A. C. F. *et al.* Getting genetic ancestry right for science and society. *Science* (2022)
475 doi:10.1126/science.abm7530.

476

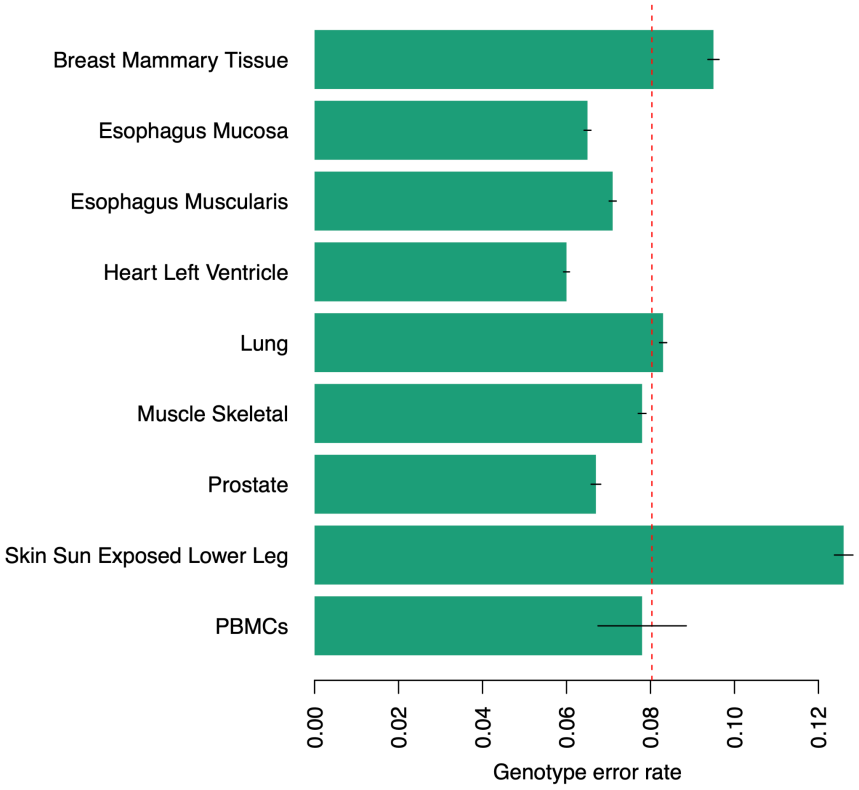
477 **Table**

Dataset	Technology	Tissue	# donors before / after QC	# cells	# Qced SNPs before / after pruning	Reported ancestry	Ref.
Airway epithelium	10x 3' v2/v3, Drop-seq	Lung epithelium, submucosal gland	50 / 43	Not reported	9,100 / 4,515	Not reported	[19]
Breast	10x 3' v2/v3	Basal, Epithelial, Luminal cells	35 / 33	103K	26,603 / 8,757	Not reported	[20]
Lung	10x 3' v1/v2/v3, 10x 5' v1,	Blood, lung	24 / 24	362K	22,416 / 8,058	Not reported	[21]
Nose	10x 3' v3	Nasal cavity mucosa	39 / 34	Not reported	13,106 / 5,846	Not reported	[22]
Skin	10x 3' v2	Diverse skin regions	22 / 21	233K	65,252 / 13,594	Not reported	[23]
Bone marrow	Smart-seq2	Bone marrow	42 / 37	2.3K	6,589 / 2,968	European *	[24]
Embryo	Smart-seq2	Embryonic cell, inner cell mass cell	88 / 81	1.5K	14,039 / 5,014	European *	[25]
CD45+	10x 3' v2	CD 45+ cells from diverse tissues	37 / 35	126.0K	21,895 / 7,614	European	[26]
Cell landscape	Microwell-Seq	Major human organs	58 / 53	703K	21,049 / 7,504	Asian, European	[27]
Heart	10x 5' v1 (cell; nuclei)	Apical region of left ventricle	7 / 7; 36 / 33	270.5K	65,518 / 13,284; 115,931 / 17,544	African, European	[28]

478

479 **Table 1. HCA single-cell transcriptomic datasets analyzed in this study.** These ten datasets
480 provided eleven sets of sc-SNPs, as scRNA-seq and snRNA-seq data from the Heart dataset
481 were analyzed separately. * Ancestry not provided, but expected from demographics.

482 **Figures**



483

484 **Figure 1. Genotype error rate across tissues.** We report genotype error rates across eight
485 tissues from GTEx and PBMCs. Genotypes identified in the GTEx datasets were compared to
486 whole-genome sequencing genotypes; genotypes identified in the PBMCs dataset were
487 compared to SNP-array genotypes. The dashed vertical red line represents the mean, and was
488 used to simulate genotype error in downstream analyses. Error bars represent 95% confidence
489 intervals. Numerical results are reported in **Supp. Table 3**.

490

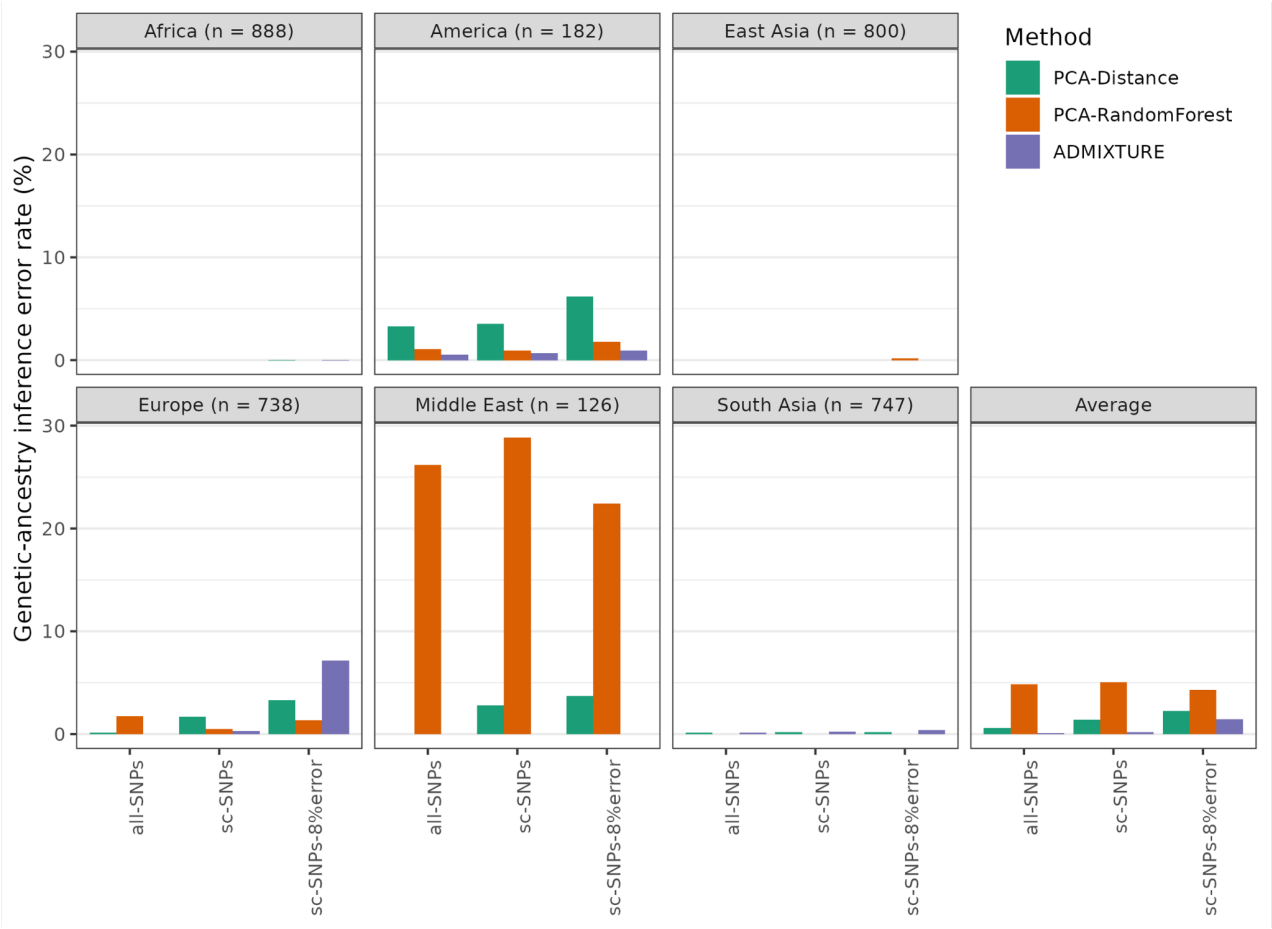
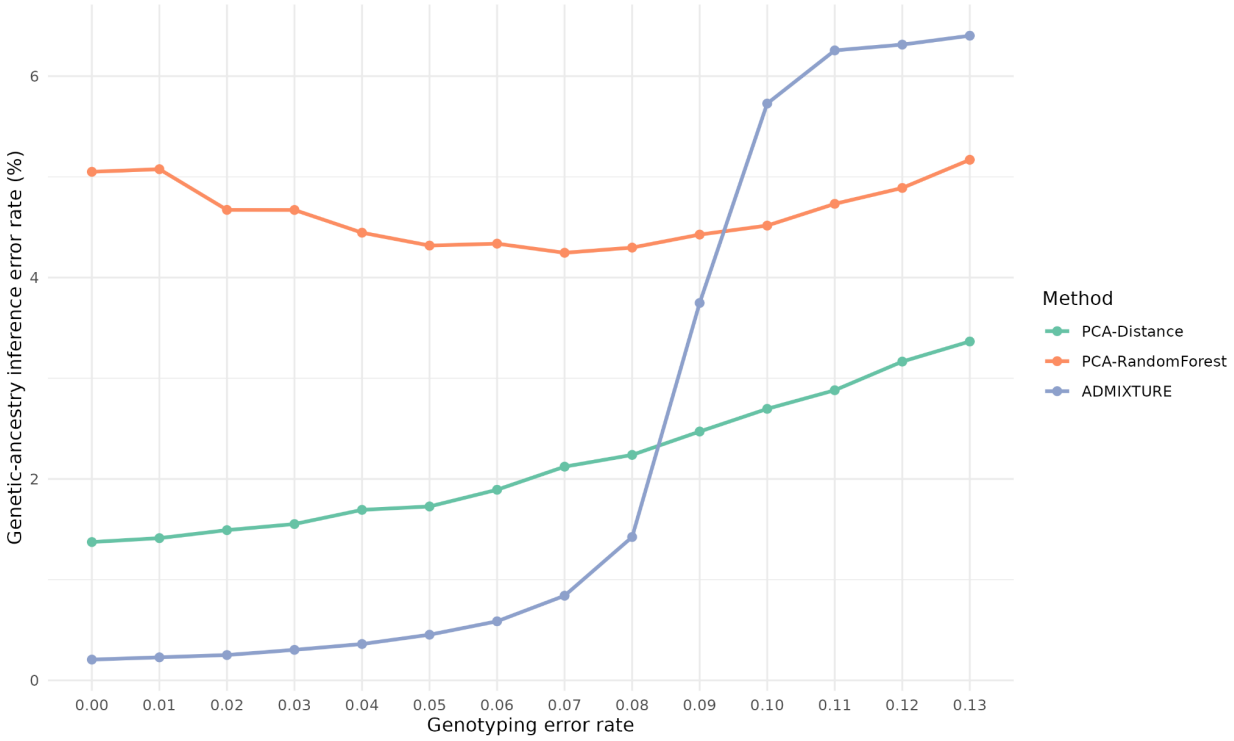


Figure 2. Genetic-ancestry inference error rate in HGDP+1kGP. We report genetic-ancestry inference error rates estimated using 3,481 HGDP+1kGP individuals from 67 populations spanning six ancestry groups. We considered three sets of genotypes for genetic-ancestry inference: genotypes from all common SNPs (all-SNPs; 295,434 common SNPs after LD pruning), genotypes from SNPs observed in single-cell datasets (sc-SNPs), and genotypes from SNPs observed in single-cell datasets for which we simulated an 8% genotype error rate (sc-SNPs-8%error). Genetic-ancestry inference error rates estimated using genotypes from sc-SNPs and sc-SNPs-8%error were averaged across eleven sets of sc-SNPs (see **Table 1)**. Error rates estimated using genotypes from all-SNPs were considered the gold standard. Error rates stratified by single-cell dataset are reported in **Supp. Figure 3**. Numerical results are reported in **Supp. Table 4**.

503



504

505 **Figure 3. Genetic-ancestry inference error rate as a function of genotyping error rate.** For
506 various genotyping error rates, we simulated errors within each sc-SNPs dataset for individuals
507 in the removed population and reported the corresponding genetic-ancestry inference error rate.
508 Numerical results are reported in **Supp. Table 6**.

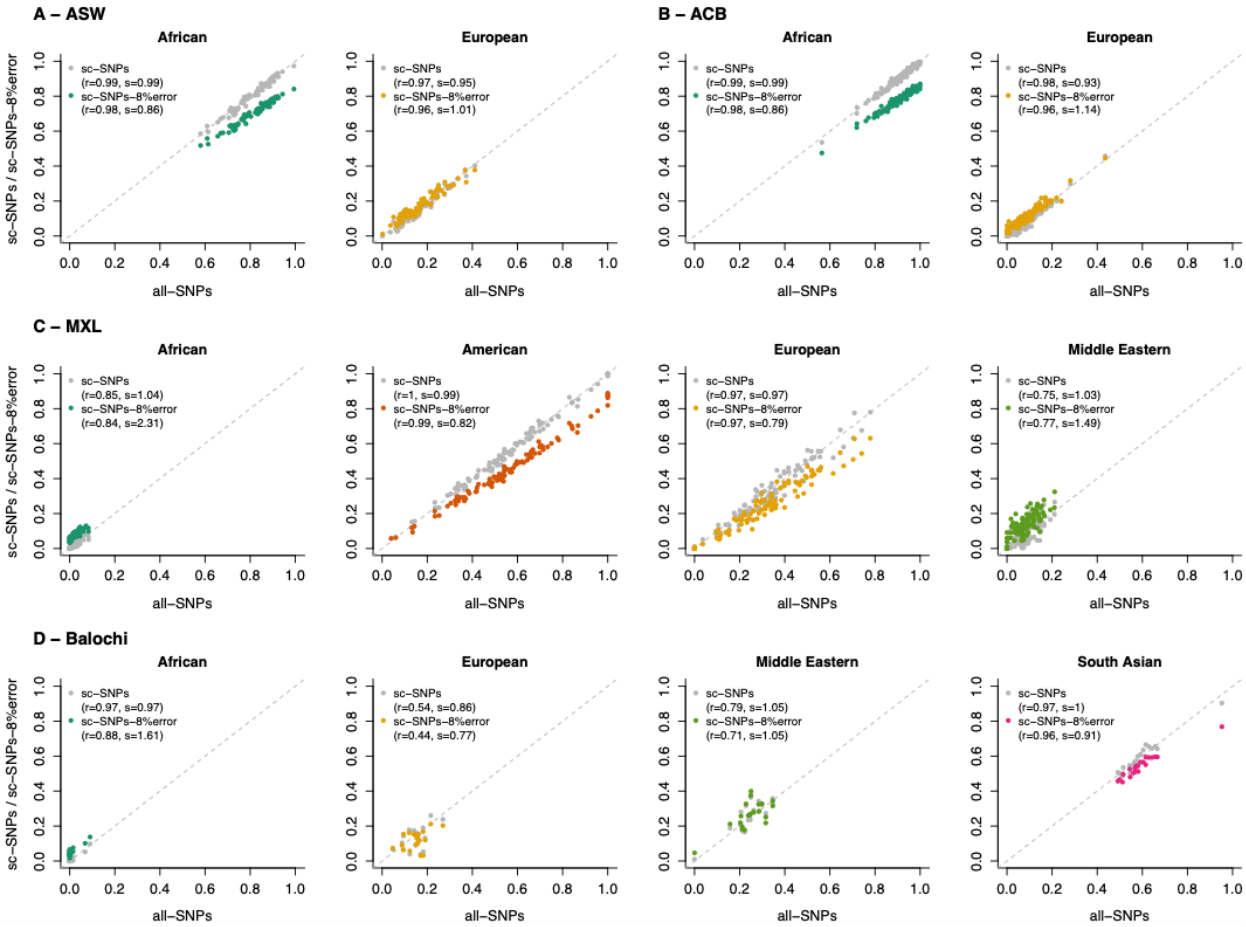
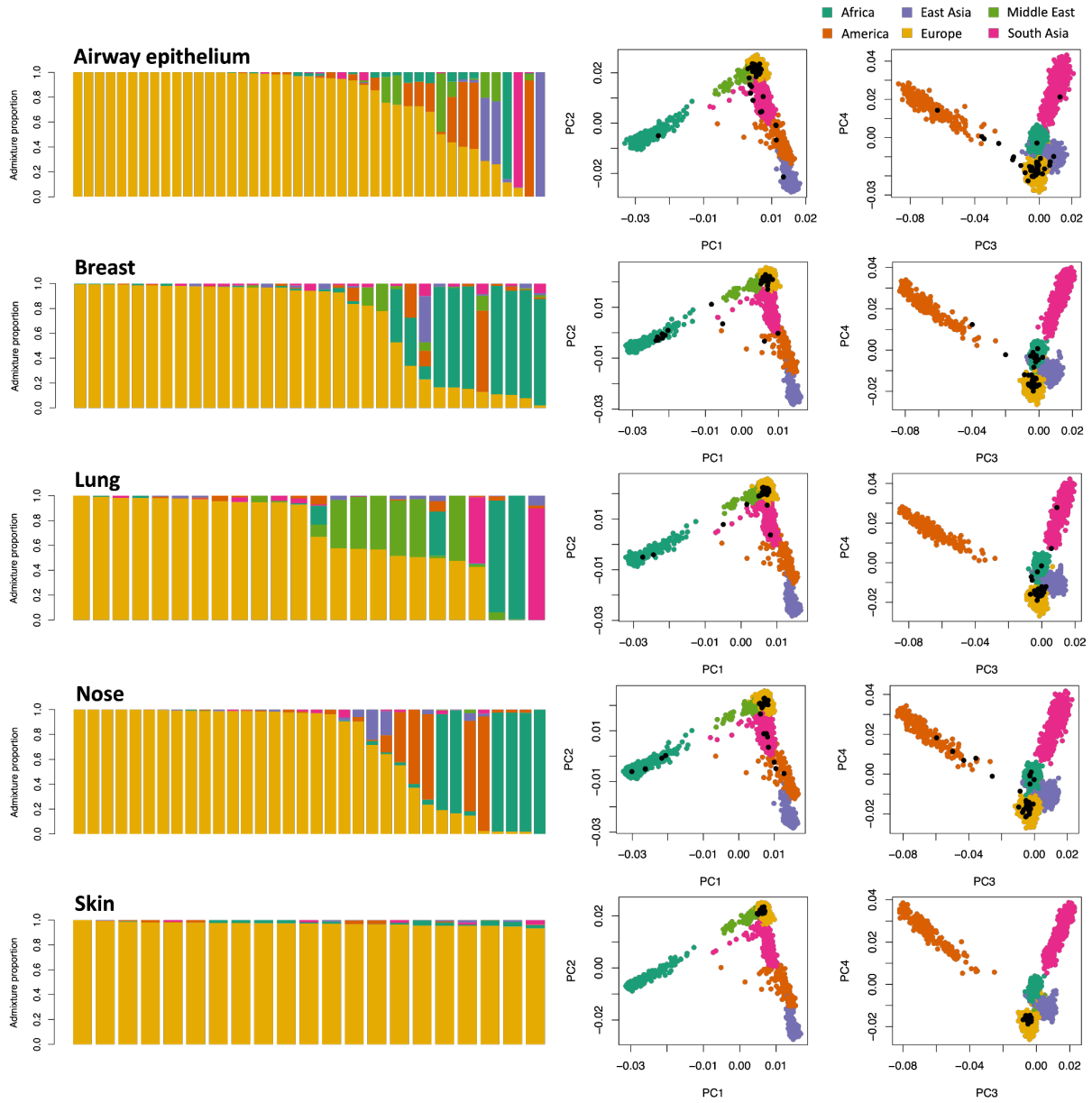


Figure 4. ADMIXTURE estimates among four admixed populations. We report ADMIXTURE estimates for African Americans from the Southwest US (ASW), African Caribbeans in Barbados (ACB), Mexican Ancestry individuals from Los Angeles (MXL), and Balochi individuals from HGDP. Genetic-admixture was estimated using genotypes from all-SNPs (x-axis), sc-SNPs (y-axis, grey dots), and sc-SNPs-8%error (y-axis, blue dots). Proportions estimated using sc-SNPs and sc-SNPs-8%error were averaged across eleven sets of sc-SNPs. For estimates obtained with sc-SNPs and sc-SNPs-8%error, we report their correlation (*r*) and regression slope (*s*) with those obtained using all-SNPs (considered the gold standard). Numerical results are reported in **Supp. Table 7**; estimates using SNPs from each single-cell dataset are reported in **Supp. Table 8**.



519

520 **Figure 5. ADMIXTURE and PCA analyses of five HCA single-cell datasets with unreported**
521 **ancestry.** The left panels show admixture proportions for HCA donors from each dataset. The
522 right panels display PCA plots of HGDP+1kGP individuals (colored dots) with projected HCA
523 donors (black dots). Results for five additional HCA datasets are reported in **Supp. Figure 8.**
524 Numerical results are reported in **Supp. Table 9**.

525 **Supplementary Table Legends**

526

527 **Supplementary Table 1. Description of the 50 HCA projects with the largest numbers of**
528 **donors.** For each dataset, we report sample size, sequencing technology, availability of FASTQ
529 files, tissue type, and reported ancestry information. We note that the Asian Immune Diversity
530 Atlas (AIDA) has two projects corresponding to different phases; we only included individuals from
531 the latest phase.

532

533 **Supplementary Table 2. Description of the HGDP+1kGP reference dataset.** We report the
534 3,481 HGDP+1kGP individuals used in our reference dataset, their genetic-ancestry groups (i.e.,
535 Africa, America, Europe, Middle East, East Asia, or South Asia), and their populations.

536

537 **Supplementary Table 3. Estimation of genotype error rate in single-cell data.** We evaluated
538 the genotype error rate of our pipeline by comparing genotypes called in GTEx snRNA-seq data
539 from eight tissues and in scRNA-seq data from PBMCs to those obtained via WGS and SNP-
540 array, respectively. We report genotype error rates by tissue, polymorphism type (e.g., SNPs with
541 A and C alleles), and allele-specific errors (e.g., A-to-C errors).

542

543 **Supplementary Table 4. Genetic-ancestry inference error rate in HGDP+1kGP.** We report
544 genetic-ancestry inference error rates estimated using 3,481 HGDP+1kGP individuals from 67
545 populations across six ancestry groups.

546

547 **Supplementary Table 5. ADMIXTURE error rate in European populations of HGDP+1kGP.**
548 We report the genetic-ancestry inference error rate of ADMIXTURE for European populations as
549 a function of genotype error rate.

550

551 **Supplementary Table 6. Genetic-ancestry inference error rate as a function of genotyping**
552 **error rate.** For various genotyping error rates, we simulated errors within each sc-SNPs dataset
553 for individuals in the removed population and reported the corresponding genetic-ancestry
554 inference error rates.

555

556 **Supplementary Table 7. ADMIXTURE estimates among four admixed populations.** We
557 report ADMIXTURE estimates for ASW, ACB, MXL, and Balochi individuals. Genetic-admixture
558 was estimated using genotypes from all-SNPs, sc-SNPs, and sc-SNPs-8%error. Proportions
559 estimated using sc-SNPs and sc-SNPs-8%error were averaged across eleven sets of sc-SNPs.

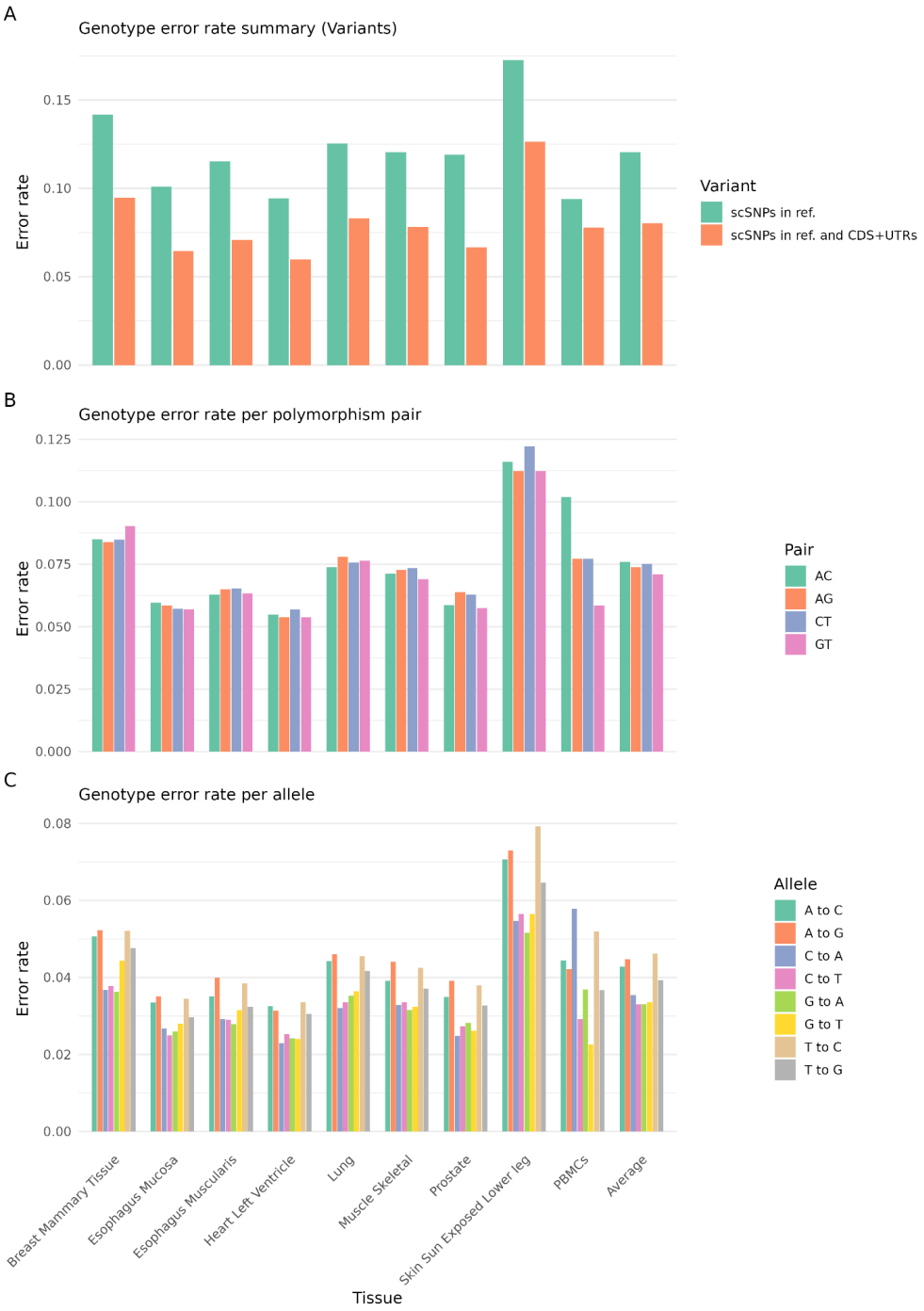
560

561 **Supplementary Table 8. ADMIXTURE estimates among four admixed populations across**
562 **eleven sets of sc-SNPs.** We report ADMIXTURE estimates for ASW, ACB, MXL, and Balochi
563 individuals. Genetic-admixture was estimated using genotypes from sc-SNPs and sc-SNPs-
564 8%error across eleven sets of sc-SNPs.

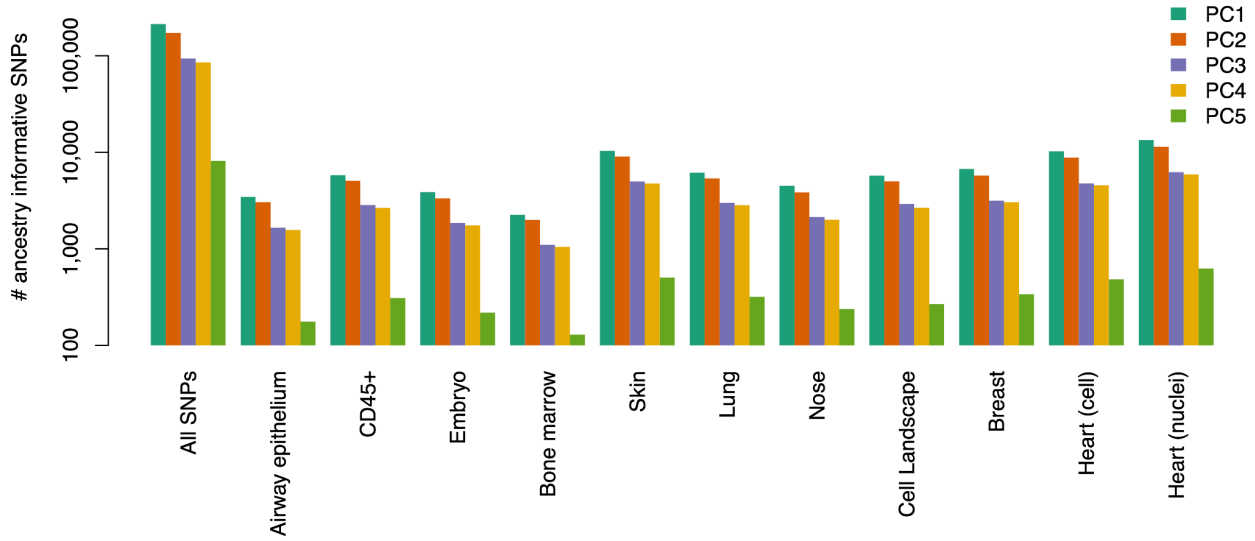
565

566 **Supplementary Table 9. ADMIXTURE estimates in the ten HCA single-cell datasets.** We
567 report genetic-ancestry inference results obtained using three approaches, along with admixture
568 proportions and random forest probabilities for HCA donors from each dataset.

Supplementary Figures

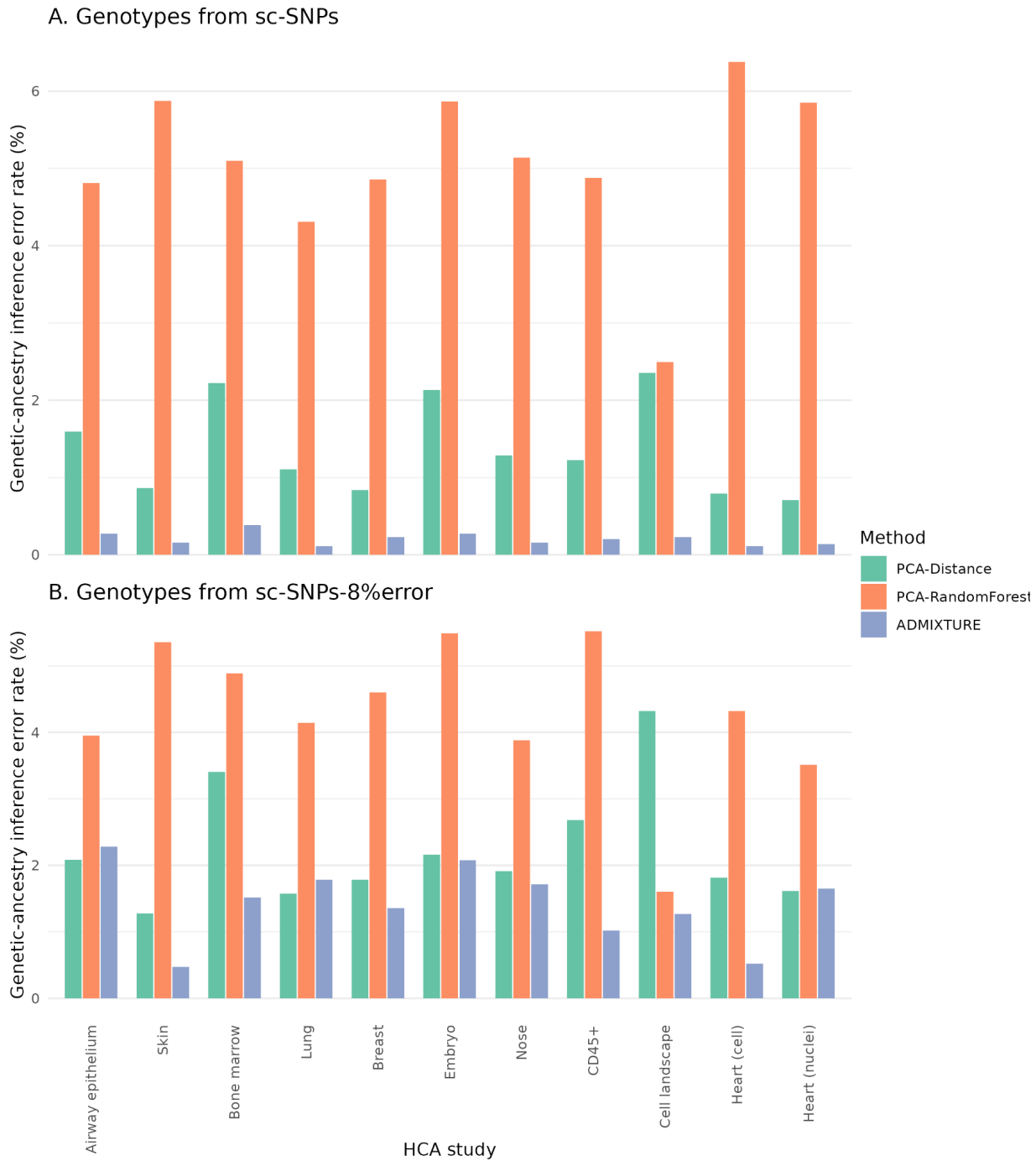


Supplementary Figure 1. Genotype error rate across tissues. (A) We report genotype error rates before and after filtering variants in coding exons and UTRs. **(B)** We report genotype error rates across polymorphism types (e.g., SNPs with A and C alleles). **(C)** We report genotype error rates across alleles (e.g., A-to-C error). Numerical results are reported in **Supp. Table 3**.

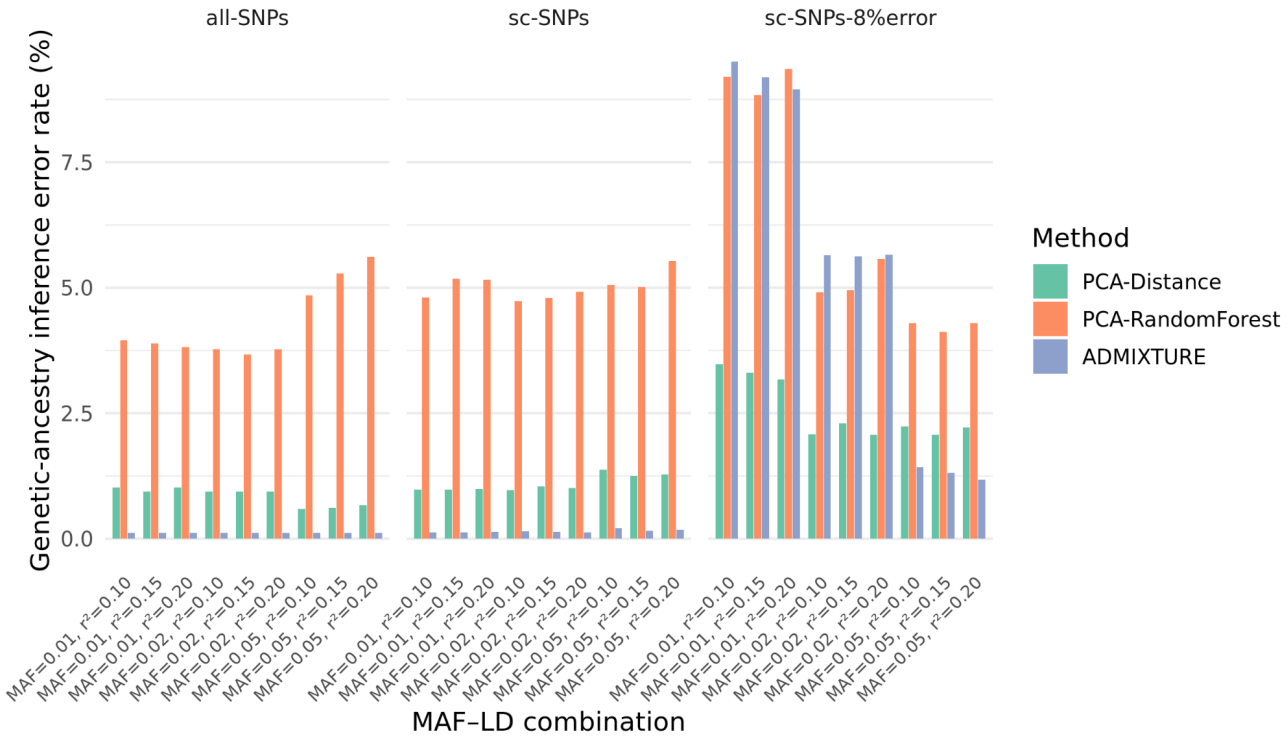


575

576 **Supplementary Figure 2. Ancestry-informative markers stratified by single-cell dataset.** For
577 each single-cell dataset, we report the number of LD-pruned SNPs associated with the first five
578 PCs computed using all-SNPs (after LD pruning). An ancestry-informative SNP is defined as
579 having a *P* value < 5 x10^{-8}.

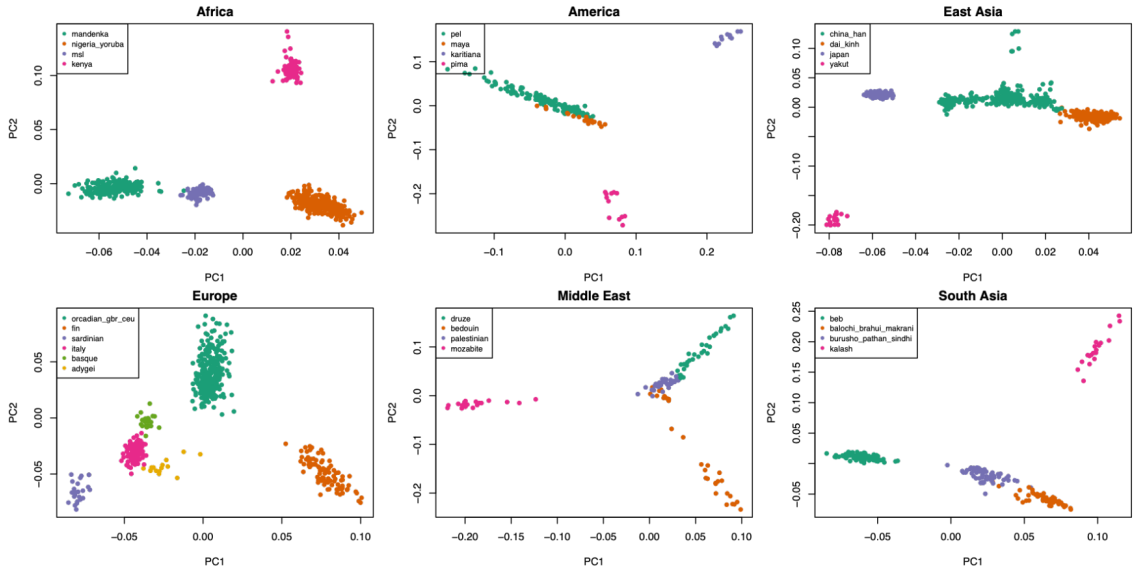


Supplementary Figure 3. Genetic-ancestry inference error rates stratified by single-cell dataset. We report results for genotypes from sc-SNPs (top panel) and sc-SNPs-8%error (bottom panel), stratified by single-cell dataset.

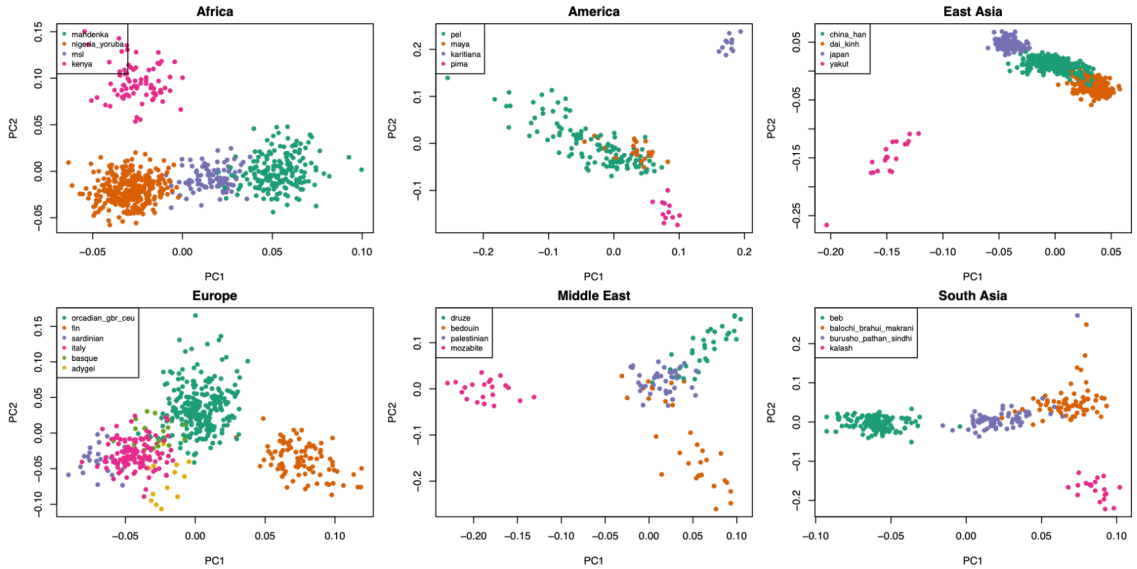


Supplementary Figure 4. Genetic-ancestry inference error rate when relaxing MAF and LD pruning thresholds. We report genetic-ancestry inference error rates when relaxing the MAF cutoff (1% and 2%, default 5%) and the LD r^2 threshold (0.15 and 0.2, default 0.1).

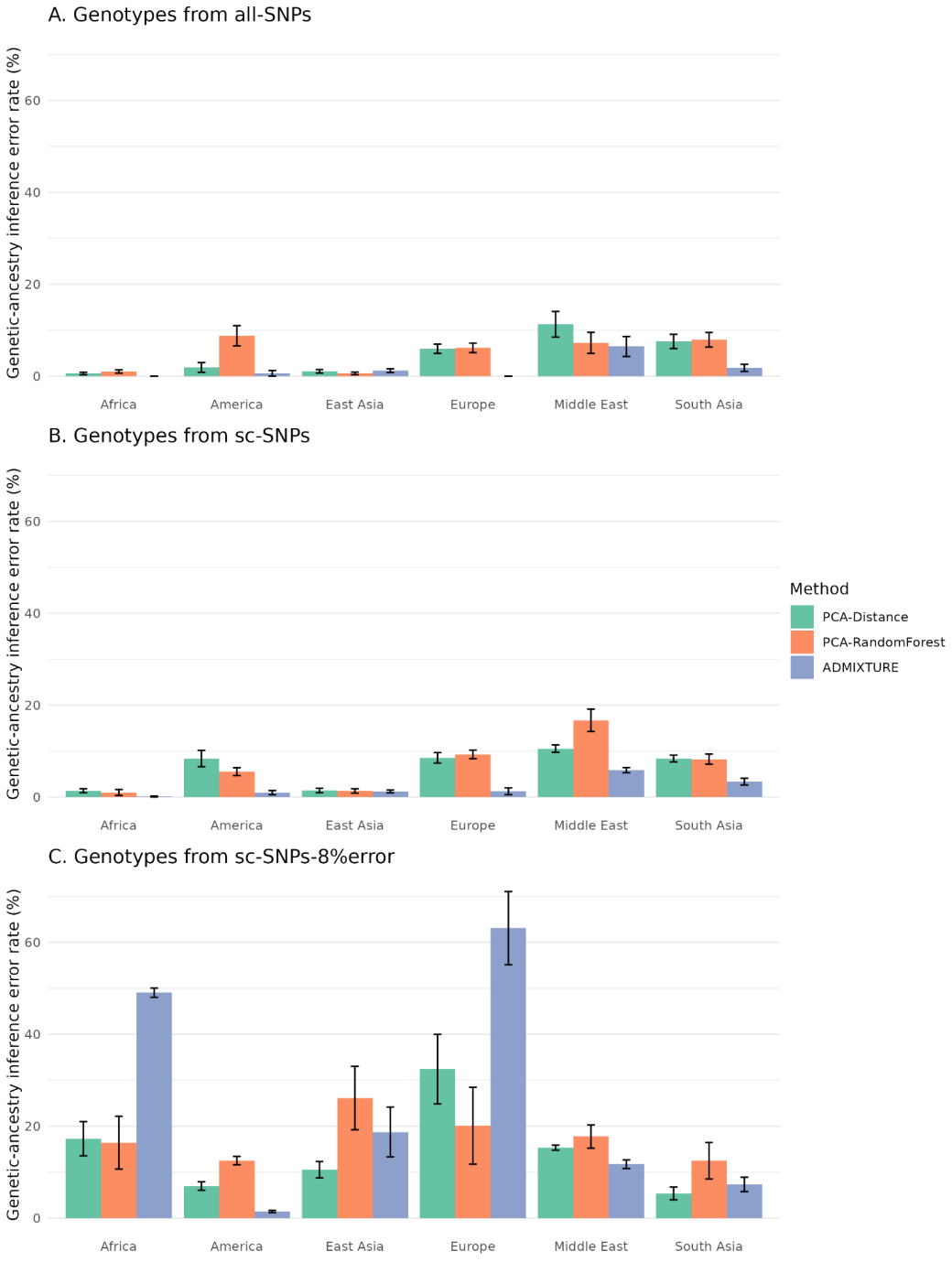
A. Genotypes from all-SNPs



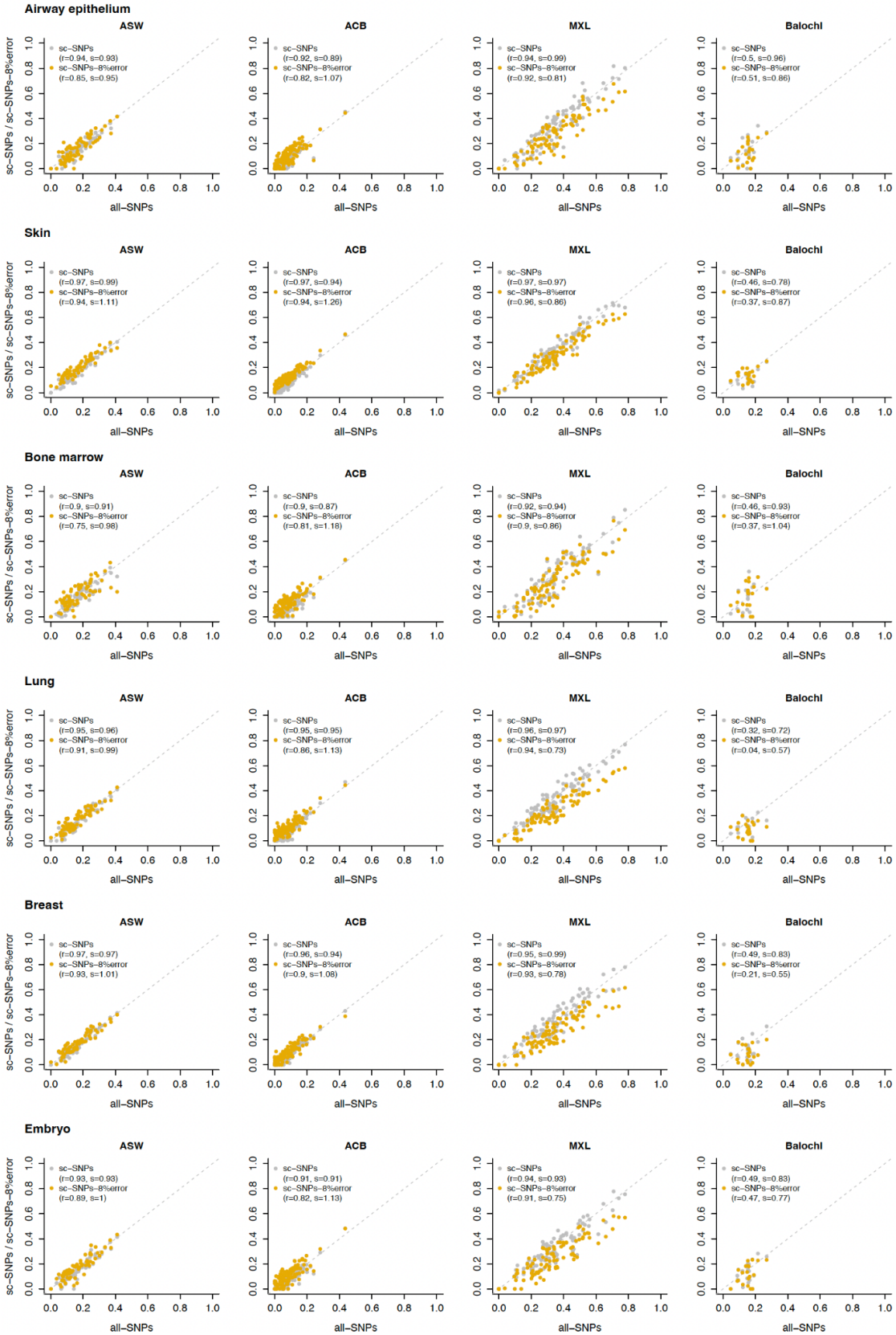
B. Genotypes from sc-SNPs from Giustacchini et al. (bone marrow) dataset

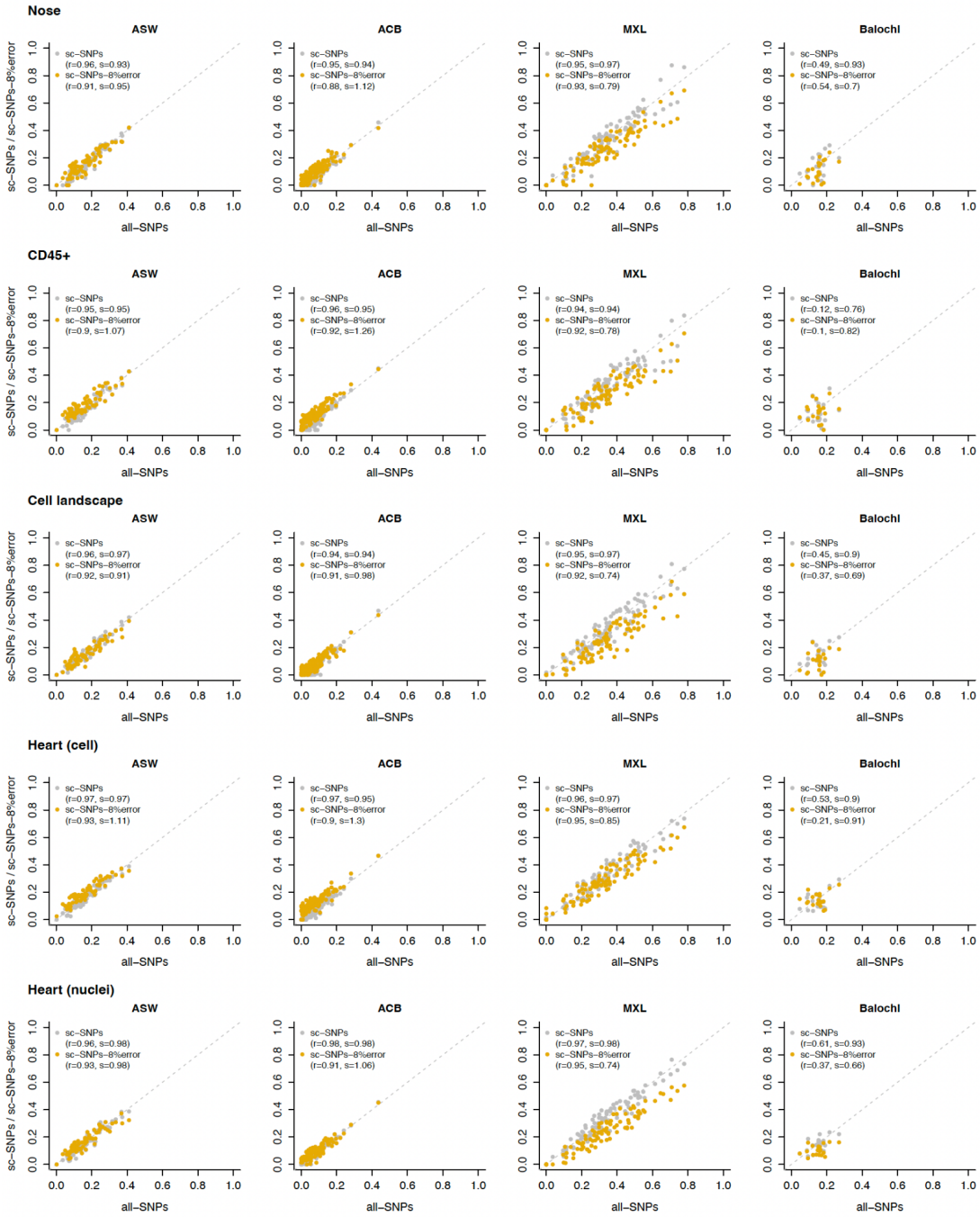


Supplementary Figure 5. Principal component analysis within each ancestry group. For each ancestry group, we performed PCA using genotypes from (**A**) all-SNPs and (**B**) sc-SNPs from the Giustacchini et al. (bone marrow) dataset. HGDP+1kGP subpopulations were grouped into homogeneous clusters. This figure illustrates the challenge of inferring subpopulation structure within an ancestry group using sc-SNPs.

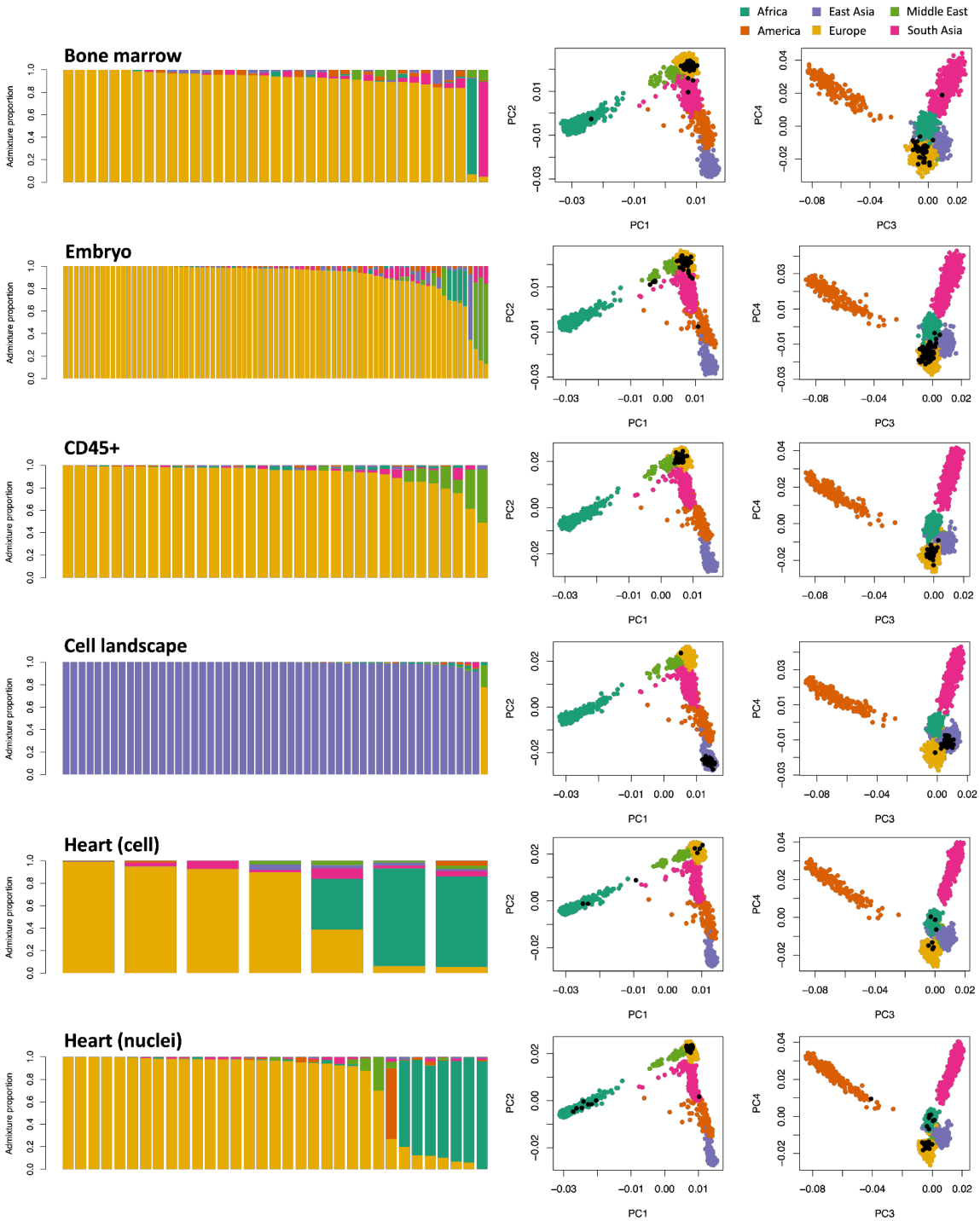


Supplementary Figure 6. Genetic-ancestry inference error rate within each ancestry group. For each ancestry group, we considered homogeneous populations (see **Supp. Figure 5**), divided the dataset into ten stratified subsets with balanced representation from each population, and performed genetic-ancestry inference on one subset at a time using the remaining nine subsets as the reference.


Airway epithelium
ASW
sc–SNPs
(r=0.94, s=0.93)
sc–SNPs–8%error
(r=0.85, s=0.95)
ACB
sc–SNPs
(r=0.92, s=0.89)
sc–SNPs–8%error
(r=0.82, s=1.07)
MXL
sc–SNPs
(r=0.94, s=0.99)
sc–SNPs–8%error
(r=0.92, s=0.81)
Balochl
sc–SNPs
(r=0.5, s=0.96)
sc–SNPs–8%error
(r=0.51, s=0.86)
Skin
ASW
sc–SNPs
(r=0.97, s=0.99)
sc–SNPs–8%error
(r=0.94, s=1.11)
ACB
sc–SNPs
(r=0.97, s=0.94)
sc–SNPs–8%error
(r=0.94, s=1.26)
MXL
sc–SNPs
(r=0.97, s=0.97)
sc–SNPs–8%error
(r=0.96, s=0.86)
Balochl
sc–SNPs
(r=0.46, s=0.78)
sc–SNPs–8%error
(r=0.37, s=0.87)
Bone marrow
ASW
sc–SNPs
(r=0.9, s=0.91)
sc–SNPs–8%error
(r=0.75, s=0.98)
ACB
sc–SNPs
(r=0.9, s=0.87)
sc–SNPs–8%error
(r=0.81, s=1.18)
MXL
sc–SNPs
(r=0.92, s=0.94)
sc–SNPs–8%error
(r=0.9, s=0.86)
Balochl
sc–SNPs
(r=0.46, s=0.93)
sc–SNPs–8%error
(r=0.37, s=1.04)
Lung
ASW
sc–SNPs
(r=0.95, s=0.96)
sc–SNPs–8%error
(r=0.91, s=0.99)
ACB
sc–SNPs
(r=0.95, s=0.95)
sc–SNPs–8%error
(r=0.86, s=1.13)
MXL
sc–SNPs
(r=0.96, s=0.97)
sc–SNPs–8%error
(r=0.94, s=0.73)
Balochl
sc–SNPs
(r=0.32, s=0.72)
sc–SNPs–8%error
(r=0.04, s=0.57)
Breast
ASW
sc–SNPs
(r=0.97, s=0.97)
sc–SNPs–8%error
(r=0.93, s=1.01)
ACB
sc–SNPs
(r=0.96, s=0.94)
sc–SNPs–8%error
(r=0.9, s=1.08)
MXL
sc–SNPs
(r=0.95, s=0.99)
sc–SNPs–8%error
(r=0.93, s=0.78)
Balochl
sc–SNPs
(r=0.49, s=0.83)
sc–SNPs–8%error
(r=0.21, s=0.55)
Embryo
ASW
sc–SNPs
(r=0.93, s=0.93)
sc–SNPs–8%error
(r=0.89, s=1)
ACB
sc–SNPs
(r=0.91, s=0.91)
sc–SNPs–8%error
(r=0.82, s=1.13)
MXL
sc–SNPs
(r=0.94, s=0.93)
sc–SNPs–8%error
(r=0.91, s=0.75)
Balochl
sc–SNPs
(r=0.49, s=0.83)
sc–SNPs–8%error
(r=0.47, s=0.77)
sc–SNPs / sc–SNPs–8%error
all–SNPs
0.0 0.2 0.4 0.6 0.8 1.0



Supplementary Figure 7. ADMIXTURE estimates of European genetic-admixture proportions among ASW, ACB, MXL, and Balochi individuals, stratified by single-cell dataset. We report the results from **Figure 4**, stratified by single-cell dataset. Numerical values, as well as genetic-admixture proportions for other ancestries, are reported in **Supp. Table 8**.



608

609 **Supplementary Figure 8. ADMIXTURE and PCA analyses of five HCA single-cell datasets**
610 **with reported or expected ancestry information.** The left panels show admixture proportions
611 for HCA donors from each dataset. The right panels display PCA plots of HGDP+1kGP individuals
612 (colored dots) with projected HCA donors (black dots). Numerical results are reported in **Supp.**
613 **Table 9**.