HGG Advances

ARTICLE

Evaluating genetic ancestry inference from single-cell transcriptomic datasets

Jianing Yao[1,2,3] and Steven Gazal[1,2,4,5,*]

Summary

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 10 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.

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 impact 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 10 HCA datasets[19–28] (including 5 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.

[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
[5]Lead contact
*Correspondence: gazal@usc.edu
https://doi.org/10.1016/j.xhgg.2026.100564.

Material and methods

Common SNP calling in single-cell transcriptomic data

We identified genetic variants from single-cell transcriptomic data using the Genome Analysis Toolkit (GATK)[29,30] following the best-practice pipeline recommended by Schnepp et al.,[31] and 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 RNA-seq short variant discovery workflow as in the following. First, processed sequencing reads were aligned to the human hg19 reference genome using STARsolo version 2.7.11a in the two-pass mode to achieve better alignments around 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 that were supported only by multi-mapped reads or with a unique read count <2. 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 untranslated regions (UTRs), defined using UCSC Genome Browser reference files and extended by 150 base pairs upstream and downstream.[32] We used PLINK to exclude SNPs with genotype missingness >10% and donors with genotype missingness >10%.

We have released an open-source pipeline implementing our framework (see data and code availability).

Single-cell transcriptomics datasets

To evaluate the genotype error rate of our pipeline, we leveraged 25 snRNA-seq samples from Genotype-Tissue Expression (GTEx), obtained from 16 donors across 8 different tissues.[33] We compared genotypes inferred by our pipeline with those obtained by GTEx via whole-genome sequencing (WGS), which we considered to be ground truth. We also analyzed 12 individuals with scRNA-seq data from both peripheral blood mononuclear cells (PBMCs) and SNP-array data,[34] considering the SNP-array genotypes to be ground truth.

We next investigated the 50 HCA projects with the largest number of donors (as of July 2025). We report their sample size, sequencing technology, availability of FASTQ files on the HCA website, tissues, and reported ancestry information in Table S1. We identified 15 projects with no or limited information about ancestry and inferred genetic ancestry for the five of these with available FASTQ files (datasets labeled airway epithelium, breast, lung, nose, and skin, see Table 1). To enable a comprehensive evaluation of genetic ancestry inference across varying combinations of genes expressed and technical platforms, we also selected five additional datasets with available FASTQ files that encompass diverse tissues and cell types, employ different sequencing technologies (i.e., Smart-seq, Drop-seq, Microwell-seq, and 10x Genomics 3′ sequencing), and include donors of different ancestries (i.e., African American, East Asian, and European) (Table 1). Three datasets had ancestry information available (datasets labeled CD45$^+$, cell landscape, and heart), while two datasets were collected from European countries (United Kingdom and Sweden), where donor ancestry can be reasonably inferred from population demographics (datasets labeled bone marrow and embryo). One dataset included both scRNA-seq and snRNA-seq data; individuals profiled with these different technologies were analyzed separately. After QC, these 10 datasets comprised 401 individuals.

Genetic ancestry inference approaches

We considered the harmonized HGDP+1KGP dataset[17] as a reference population dataset for genetic ancestry inference and selected six genetic ancestry groups (Africa, America, Europe, Middle East, East Asia, and South Asia; note that the Finnish population was included within the European group). Several QC steps were applied before downstream analysis. Specifically, we excluded individuals who were outliers in principal-component analyses (PCAs) within their reported genetic ancestry groups to obtain homogeneous reference populations. The final dataset consisted of 3,481 individuals from 67 populations (Table S2). Further analyses were restricted to common SNPs (MAF >5%) in this dataset.

We considered three commonly used approaches to infer the genetic ancestry of study samples (i.e., donors from single-cell studies) using a reference dataset (i.e., HGDP+1KGP). Two approaches relied on PCA performed on the HGDP+1KGP dataset, with projection of donors onto the first five PCs. First, we assigned each donor to the genetic ancestry group whose centroid had the smallest Euclidean distance to that donor (the method labeled PCA-Distance). Second, we used a random forest classifier trained on the five PCs to predict genetic ancestry probabilities for each study sample, as previously described by Koenig et al.[17] (method labeled PCA-RandomForest). Finally, we applied ADMIXTURE[18] in supervised mode, using the HGDP+1KGP genetic ancestry groups as reference (e.g., K = 6) to estimate the proportions of ancestral populations for each donor. All three approaches were run after linkage disequilibrium (LD) pruning performed with PLINK[35] (using the –indep-pairwise option, with a window size of 50 SNPs, a step size of 10 SNPs, and an r^2 threshold of 0.1) on the HGDP+1KGP dataset restricted to the investigated SNPs derived from HCA single-cell datasets (between 2,968 and 17,544 SNPs after LD pruning; see Table 1). We also performed additional analyses relaxing the MAF cutoff (1% and 2%) and the r^2 threshold (0.15 and 0.2).

Other methods exist to infer genetic ancestry, but we did not investigate them in this study. For instance, STRUCTURE[36] employs a probabilistic model similar to that of ADMIXTURE, but it is substantially less computationally efficient for large datasets. Furthermore, methods such as RFmix[37] infer local genetic ancestry, but they rely on haplotype information and require phasing of genetic data; we opted against these methods because phasing genetic variants inferred from single-cell data presents a considerable

Table 1. HCA single-cell transcriptomic datasets analyzed in this study

Dataset	Technology	Tissue	No. of donors before/after QC	No. of cells	No. of QCed SNPs before/after pruning	Reported ancestry	Reference
Airway epithelium	10 × 3′ v2/v3, Drop-seq	lung epithelium, submucosal gland	50/43	not reported	9,100/4,515	not reported	Carraro et al.[19]
Breast	10 × 3′ v2/v3	basal, epithelial, luminal cells	35/33	103,000	26,603/8,757	not reported	Murrow et al.[20]
Lung	10 × 3′ v1/v2/v3, 10 × 5′ v1	blood, lung	24/24	362,000	22,416/8,058	not reported	Leader et al.[21]
Nose	10 × 3′ v3	nasal cavity mucosa	39/34	not reported	13,106/5,846	not reported	Winkley et al.[22]
Skin	10 × 3′ v2	diverse skin regions	22/21	233K	65,252/13,594	not reported	Ganier et al.[23]
Bone marrow	Smart-seq2	bone marrow	42/37	2.3K	6,589/2,968	European[a]	Giustacchini et al.[24]
Embryo	Smart-seq2	embryonic cell, inner cell mass cell	88/81	1.5K	14,039/5,014	European[a]	Petropoulos et al.[25]
$CD45^+$	10 × 3′ v2	$CD45^+$ cells from diverse tissues	37/35	126K	21,895/7,614	European	Cillo et al.[26]
Cell landscape	Microwell-Seq	major human organs	58/53	703K	21,049/7,504	Asian, European	Han et al.[27]
Heart	10 × 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	Koenig et al.[28]

These 10 datasets provided 11 sets of sc-SNPs, as scRNA-seq and snRNA-seq data from the heart dataset were analyzed separately.
[a]Ancestry not provided but expected from demographics.

technical challenge due to data sparsity and may compromise accuracy. Finally, we acknowledge the utility of recent tools that work directly with single-cell measurements for variant calling and genetic ancestry inference, such as Monopogen,[13] which leverages LD information from external reference panels to call germline and somatic SNVs from sparse single-cell profiles and support global and local genetic ancestry mapping, and scAI-SNP,[14] which is trained on ancestry-informative SNPs derived from the 1KGP to estimate ancestry contributions for each donor. Nonetheless, in this work, we restrict our evaluations to variants called using a standard GATK-based workflow and to global genetic ancestry inference approaches based on PCA and ADMIXTURE, as this combination is widely used in practice and provides a simple baseline that can be readily adopted in existing population-genetic and single-cell analysis pipelines.

Evaluating genetic ancestry inference using HGDP+1KGP

We evaluated the three genetic ancestry inference approaches using a leave-one-population-out strategy on HGDP+1KGP. Specifically, for each of the 67 HGDP+1KGP populations, we removed the corresponding individuals from the reference dataset and inferred their genetic ancestry using each approach to assess the robustness of the inference methods when prior information about the excluded population was unavailable. We repeated this procedure for every population. For PCA-RandomForest and ADMIXTURE, we assigned each individual from the removed population to the genetic ancestry category with the highest predicted proportion. Finally, we compared the inferred genetic ancestry to the labels provided in the HGDP+1KGP dataset and evaluated the inference error rate both within each genetic ancestry group and across all individuals.

We evaluated each approach using different sets of genotypes. First, to define a gold standard, we considered genotypes from all common SNPs in the HGDP+1KGP dataset, performed LD pruning, and retained 295,434 SNPs; we labeled this genotype dataset "all-SNPs." Second, to evaluate genetic ancestry inference from SNPs identified in single-cell datasets, we considered 11 distinct sets of SNPs corresponding to the LD-pruned SNPs identified in the 10 HCA datasets; one dataset had two sets of SNPs from scRNA-seq and snRNA-seq data, leading to 11 sets of sc-SNPs. We labeled these genotype datasets "sc-SNPs." Finally, to account for genotype detection error using single-cell data, we leveraged the error rate estimated from the GTEx and PBMCs datasets (8%; see results) and simulated an 8% genotype error rate within each sc-SNPs dataset for individuals in the removed population; we labeled these genotype datasets "sc-SNPs-8%error." Genotype error was simulated by randomly selecting 8% of the genotypes of each individual and changing each selected genotype to an alternative value. We also explored additional analyses using a range of simulated genotype error rates (see results). We note that all the investigated methods have negligible computation time when using sc-SNPs (<1 min for PCA and PCA-RandomForest, <5 min for ADMIXTURE). In comparison, ADMIXTURE is more computationally intensive when using all-SNPs (∼1 h).

Finally, we evaluated the estimation of genetic admixture proportions with ADMIXTURE in four admixed populations from HGDP+1KGP. First, we considered African Americans from the US Southwest (ASW; 68 individuals) and African Caribbeans in Barbados (ACB; 112 individuals) and focused on their proportions of African and European genetic ancestry.[38] Second, we considered Mexican Ancestry individuals from Los Angeles (MXL; 96 individuals), and focused on their proportions of African, American, European, and Middle Eastern genetic ancestry. Finally, we considered Balochi individuals from HGDP (21

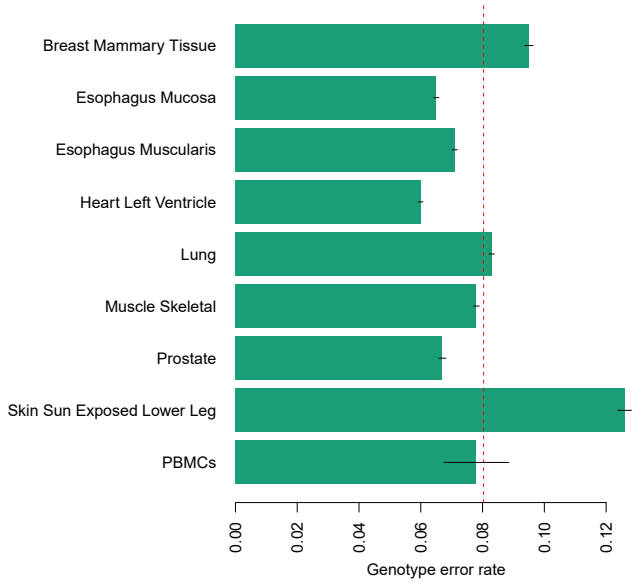

Figure 1. Genotype error rate across tissues
We report genotype error rates across eight tissues from GTEx and PBMCs. Genotypes identified in the GTEx datasets were compared to whole-genome sequencing genotypes; genotypes identified in the PBMCs dataset were compared to SNP-array genotypes. The dashed vertical red line represents the mean and was used to simulate genotype error in downstream analyses. Error bars represent 95% confidence intervals. Numerical results are reported in Table S3.

individuals), and focused on their proportions of African, European, Middle Eastern, and South Asian genetic ancestry. We first ran ADMIXTURE using the set of 295,434 LD-pruned SNPs (all-SNPs) and treated the resulting genetic ancestry proportions as the gold standard. We then reran ADMIXTURE on these individuals using genotypes from each set of sc-SNPs, as well as on genotypes in which we simulated an 8% genotype error rate (sc-SNPs-8%error).

Results

Estimating genotype error rate

We first assessed the precision of SNP detection by GATK in single-cell transcriptomic datasets by comparing genotypes called in GTEx snRNA-seq data from eight tissues and in scRNA-seq data from PBMCs to those obtained via WGS and SNP-array, respectively. Across these nine datasets, we observed an average genotype error rate of 8.02% (Figure 1; Table S3). The genotype error rate varied across tissue types, ranging from 5.98% in heart left ventricle samples to 12.64% in skin sun-exposed lower leg samples from GTEx, while the error rate remained similar across polymorphism types (e.g., SNPs with A and C alleles) or across alleles (e.g., A-to-C error) (Figure S1). We also note that restricting SNPs to coding exons and UTRs reduced the genotype error rate from 12.05% to 8.02% (Figure S1), confirming that this QC step helps remove potential false-positive genotypes. Based on these results, we used a genotype error rate of 8% for the main analyses and varied the genotype error rate from 1% to 13% in additional analyses.

Detecting common genetic polymorphisms in HCA single-cell transcriptomic datasets

We next ran our pipeline on 10 datasets from the HCA encompassing diverse tissues and cell types, employing different sequencing technologies, and including donors with different ancestries. After QC, we detected between 6,589 and 115,931 SNPs that were also present in the HGDP+1KGP dataset (average: 34,682), which were further reduced to between 2,968 and 17,544 common SNPs (average: 8,609) after LD pruning (Table 1). To assess the number of ancestry-informative SNPs derived from each HCA dataset, we examined SNPs that were significantly associated with the first five PCs computed using all-SNPs. We observed that about 6,000 SNPs were associated with the first and second PCs, about 3,000 SNPs were associated with the third and fourth PCs, and about 300 SNPs were associated with the fifth PC; on average, this represents a ~30-fold decrease in ancestry-informative SNPs compared with using all-SNPs (Figure S2).

Quantifying genetic ancestry inference accuracy

We estimated genetic ancestry inference error using a leave-one-population-out approach on the HGDP+1KGP dataset; results within the six genetic ancestry groups, along with the overall average, are reported in Figure 2 and Table S4. Using the all-SNPs dataset as the gold standard, we observed that ADMIXTURE provided a very low error rate (0.11% across the groups), lower than that obtained with PCA-Distance (0.59%) and PCA-RandomForest (4.84%). When restricting these analyses to sc-SNPs, ADMIXTURE still yielded an extremely low error rate (0.21% across the 11 sets of sc-SNPs), more than six times lower than those of PCA-Distance and PCA-RandomForest (1.37% and 5.05%, respectively). After simulating genotype error in these datasets (sc-SNPs-8%error), the error rate for ADMIXTURE remained lower than that for PCA-Distance and PCA-RandomForest (1.42%, 2.24%, and 4.30%, respectively). The higher error rate for PCA-RandomForest was driven by the misclassification of individuals from the Druze and Mozabite populations in the Middle Eastern group. The increased error rate for ADMIXTURE from sc-SNPs to sc-SNPs-8%error was driven by the misclassification of individuals from the Adygei and Toscani populations, who were inferred as having admixture of European and Middle Eastern genetic ancestries (Table S5). We observed similar trends across the 11 sets of sc-SNPs (Figure S3).

We next investigated the impact of genotype error rate on genetic ancestry inference and found that the ADMIXTURE error rate increased exponentially, whereas the error rate for PCA-based methods remained stable (PCA-RandomForest) or increased linearly (PCA-Distance) (Figure 3; Table S6). The elevated ADMIXTURE inference error under high simulated genotype error was driven by

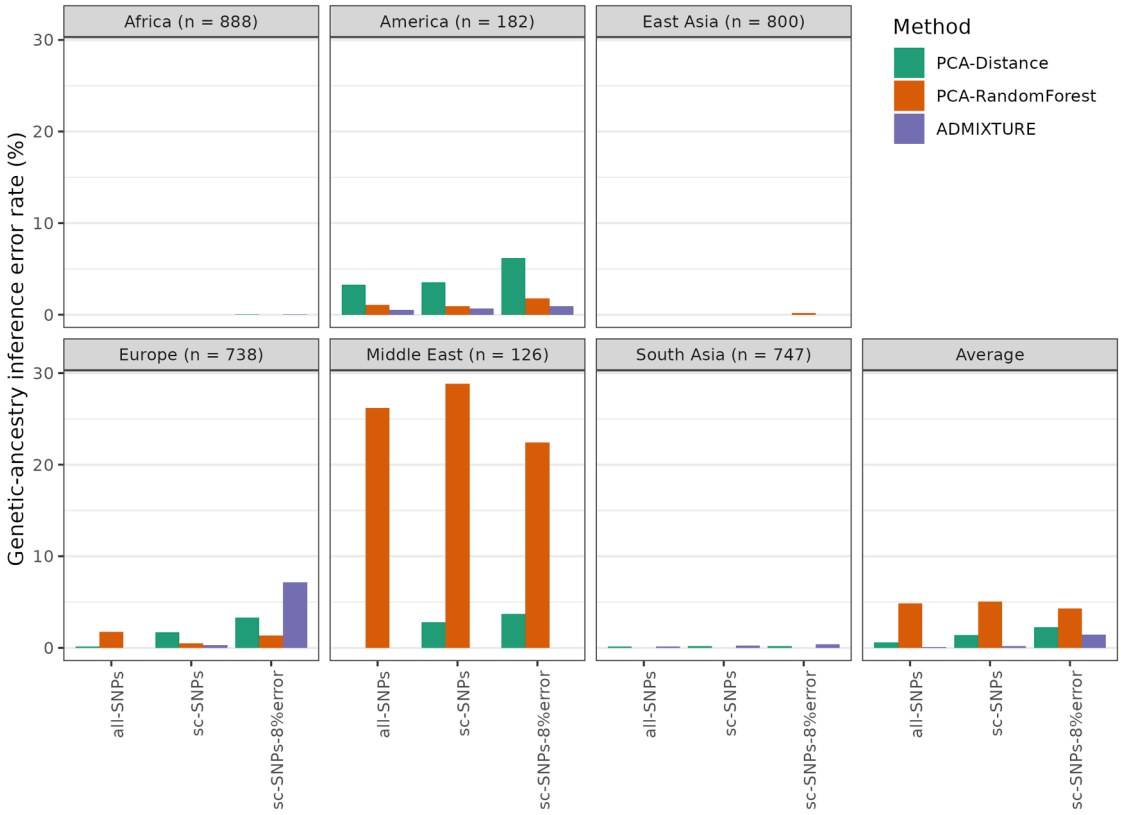

Figure 2. Genetic ancestry inference error rate in HGDP+1KGP

We report genetic ancestry inference error rates estimated using 3,481 HGDP+1KGP individuals from 67 populations spanning 6 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 11 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 Figure S3. Numerical results are reported in Table S4.

the misclassification of individuals from European populations (Adygei, Iberian, and Toscani; Table S5).

To further assess the robustness and limitations of our approach, we performed two supplementary analyses. First, we investigated whether including SNPs with lower MAF and relaxing the LD pruning thresholds would improve genetic ancestry inference. We found that using less-stringent values for these parameters did not affect the error rate when using all-SNPs and sc-SNPs; however, including SNPs with MAF <5% increased the error rate of all methods when using sc-SNPs-8%error (Figure S4). Second, we investigated whether SNPs identified in single-cell datasets allow for more fine-grained genetic ancestry inference. Specifically, we tested whether subpopulations within an ancestry group could be inferred (e.g., identifying Japanese rather than simply East Asian genetic ancestry).[39] We observed that although ADMIXTURE can yield an acceptable error rate within some ancestry groups when using sc-SNPs, all inference approaches produced high error rates when using sc-SNPs-8%error (Figures S5 and S6).

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

Quantifying genetic admixture inference accuracy

Donors from single-cell datasets are likely to be genetically admixed. We therefore evaluated genetic admixture inference with ADMIXTURE by leveraging admixed individuals from the ASW, ACB, MXL, and Balochi populations from HGDP+1kGP (Figure 4; Table S7). Results obtained with sc-SNPs were nearly identical to those obtained with all-SNPs (used here as the gold standard). Results obtained with sc-SNPs-8%error were highly correlated with those from all-SNPs, although they slightly underestimated some admixture proportions in certain cases. For example, European genetic admixture was well estimated in ASW

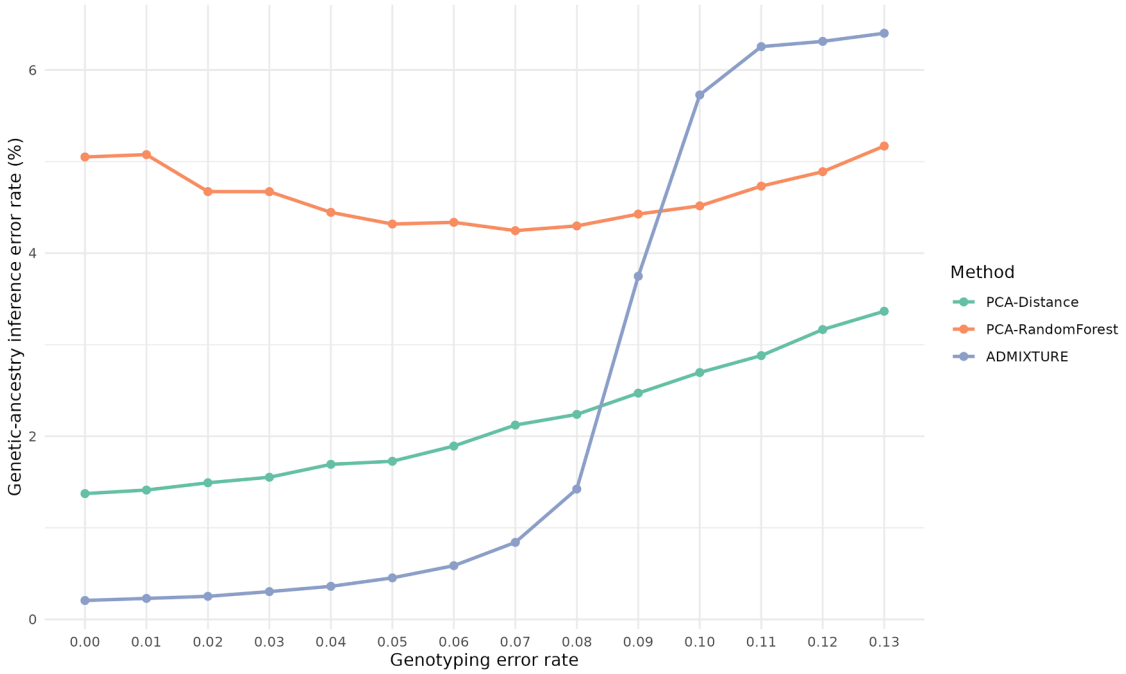

Figure 3. Genetic ancestry inference error rate as a function of genotyping error rate
For various genotyping error rates, we simulated errors within each sc-SNPs dataset for individuals in the removed population and reported the corresponding genetic ancestry inference error rate. Numerical results are reported in Table S6.

and ACB populations (regression slopes = 1.01 and 1.14, respectively) but underestimated in the MXL population (regression slope = 0.79). We reached similar conclusions across the 11 sets of sc-SNPs (Figure S7; Table S8). Overall, these results suggest that ADMIXTURE can also provide robust estimates of genetic admixture proportions.

Characterizing the ancestry of donors from the HCA

Finally, we aimed to characterize the ancestry of donors represented in the HCA by investigating the 50 HCA projects with the largest numbers of donors (Table S1). Ancestry information for donors was explicitly available for 20 datasets and could be deduced for 3 additional datasets in which the corresponding publications reported a single ancestry origin. These 23 datasets contained 3,505 unique donors, including 2,283 (65.1%) of European ancestry, 806 of Asian ancestry (467 East Asian, 111 South Asian, 228 unreported/other; 23.0%), and 158 African American donors (4.5%); ancestry for the remaining individuals was mostly missing (5.1%). An additional 12 datasets were expected to include donors from European or East Asian countries based on sample demographics. Assuming that these individuals have European or East Asian ancestries, the numbers of European and Asian donors would increase to 2,741 (65.2%) and 1,046 (24.9%), respectively, across 4,203 donors. Overall, these results indicate that donors contributing to HCA projects are overwhelmingly of European and Asian ancestry. Among the 15 remaining projects with no or limited ancestry information, 5 had FASTQ files available on the HCA website, allowing us to infer genetic ancestry.

To benchmark our pipeline on real data, we first inferred genetic ancestry for three datasets in which ancestry information was available for 10 donors of African ancestry, 52 donors of East Asian ancestry, and 66 donors of European ancestry (Figure S8; Table S9). ADMIXTURE results were largely consistent with reported ancestry, with a mean African genetic admixture proportion of 73.9% for donors of African ancestry, a mean East Asian genetic admixture proportion of 99.2% for donors of East Asian ancestry, and a mean European genetic admixture proportion of 91.0% for donors of European ancestry. Only three individuals from the same dataset (heart) had genetic ancestry that did not match the reported ancestry. One donor of African ancestry was inferred as European and one donor of European ancestry was inferred as African, suggesting either sample mix-up or mislabeled ancestry. Another donor of European ancestry was inferred as having American and European genetic admixture (62.9% and 27.0%, respectively). Similar conclusions were obtained with PCA-Distance and PCA-RandomForest.

We next applied ADMIXTURE to infer genetic ancestry for 118 individuals from two datasets generated in European countries (United Kingdom and Sweden), for which European genetic ancestry is expected (Figure S8; Table S9). Although the inferred European ancestry proportion was high across samples (mean = 90.3%), six individuals had <50% European genetic ancestry. Two individuals had primarily African or South Asian ancestry, and one individual had a genetic admixture of European and East Asian ancestry (57.7% and 34.3%, respectively). The three remaining individuals had European and

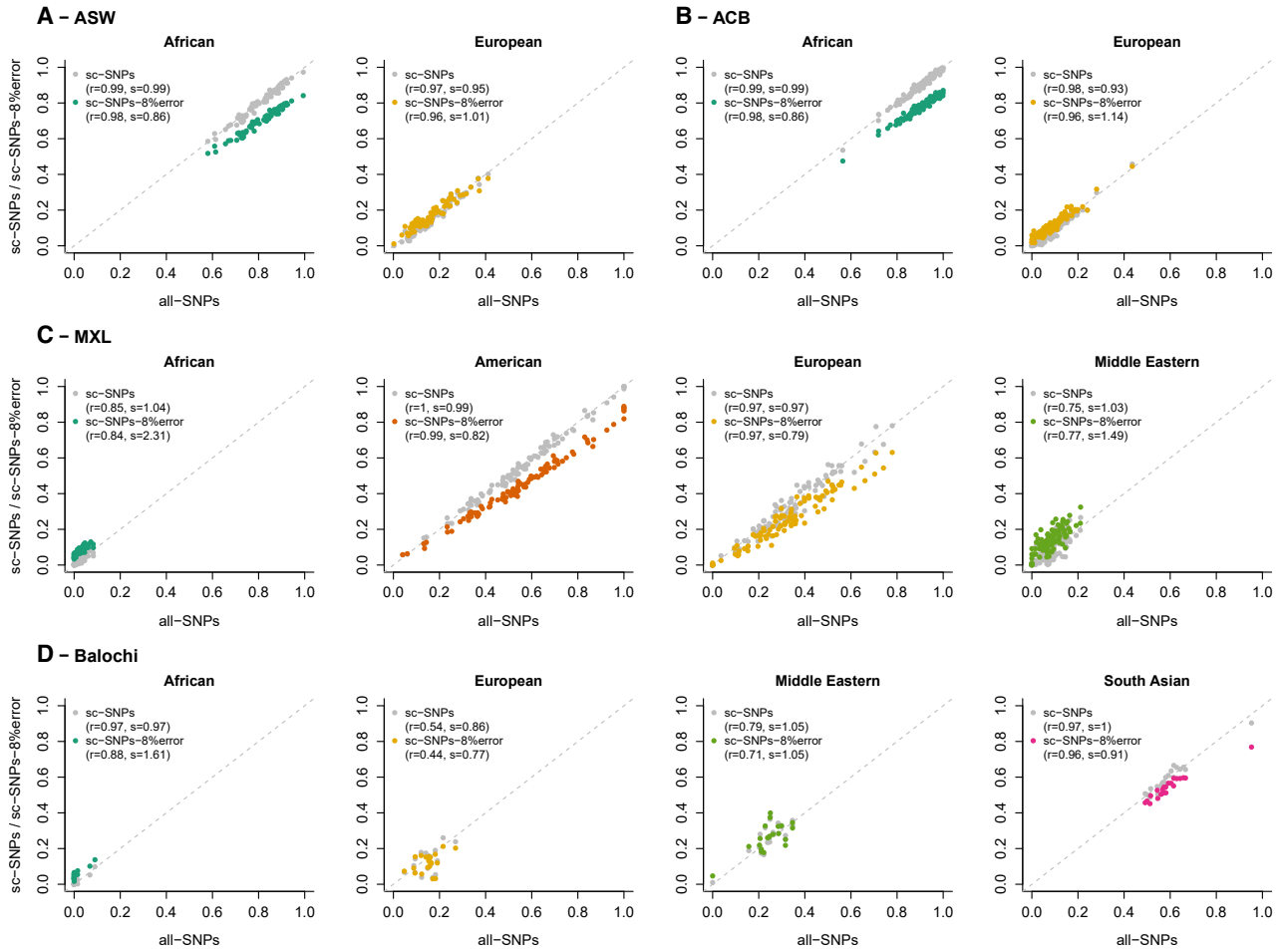

Figure 4. ADMIXTURE estimates among four admixed populations

We report ADMIXTURE estimates for (A) African Americans from the US Southwest (ASW), (B) African Caribbeans in Barbados (ACB), (C) Mexican ancestry individuals from Los Angeles (MXL), and (D) Balochi individuals from HGDP. Genetic admixture was estimated using genotypes from all-SNPs (*x* axis), sc-SNPs (*y* axis, gray dots), and sc-SNPs-8%error (*y* axis, blue dots). Proportions estimated using sc-SNPs and sc-SNPs-8%error were averaged across 11 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 Table S7; estimates using SNPs from each single-cell dataset are reported in Table S8.

Middle Eastern genetic admixture, but we caution against overinterpreting the genetic ancestry of these donors based on our previous results. We obtained consistent conclusions with PCA-Distance and PCA-RandomForest.

Finally, we applied ADMIXTURE to infer genetic ancestry for 155 individuals from 5 projects where ancestry was unreported but FASTQ files were available (Figure 5; Table S9). 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 5 datasets, 16 individuals had predominantly African genetic admixture, 2 had American, 1 had East Asian, and 2 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.

Discussion

Characterizing the ancestry of donors in single-cell transcriptomic studies is essential to ensure genetic homogeneity of the dataset and to identify ancestry-specific regulatory mechanisms linked to disease risk. Here, we propose a framework to evaluate methods for inferring genetic ancestry from single-cell data. We considered 11 realistic sets of sc-SNPs that capture the variability in SNP detection across different single-cell technologies and human tissues, and we simulated genotype errors to reflect imperfect variant calling in single-cell data. We demonstrated that ADMIXTURE provides accurate

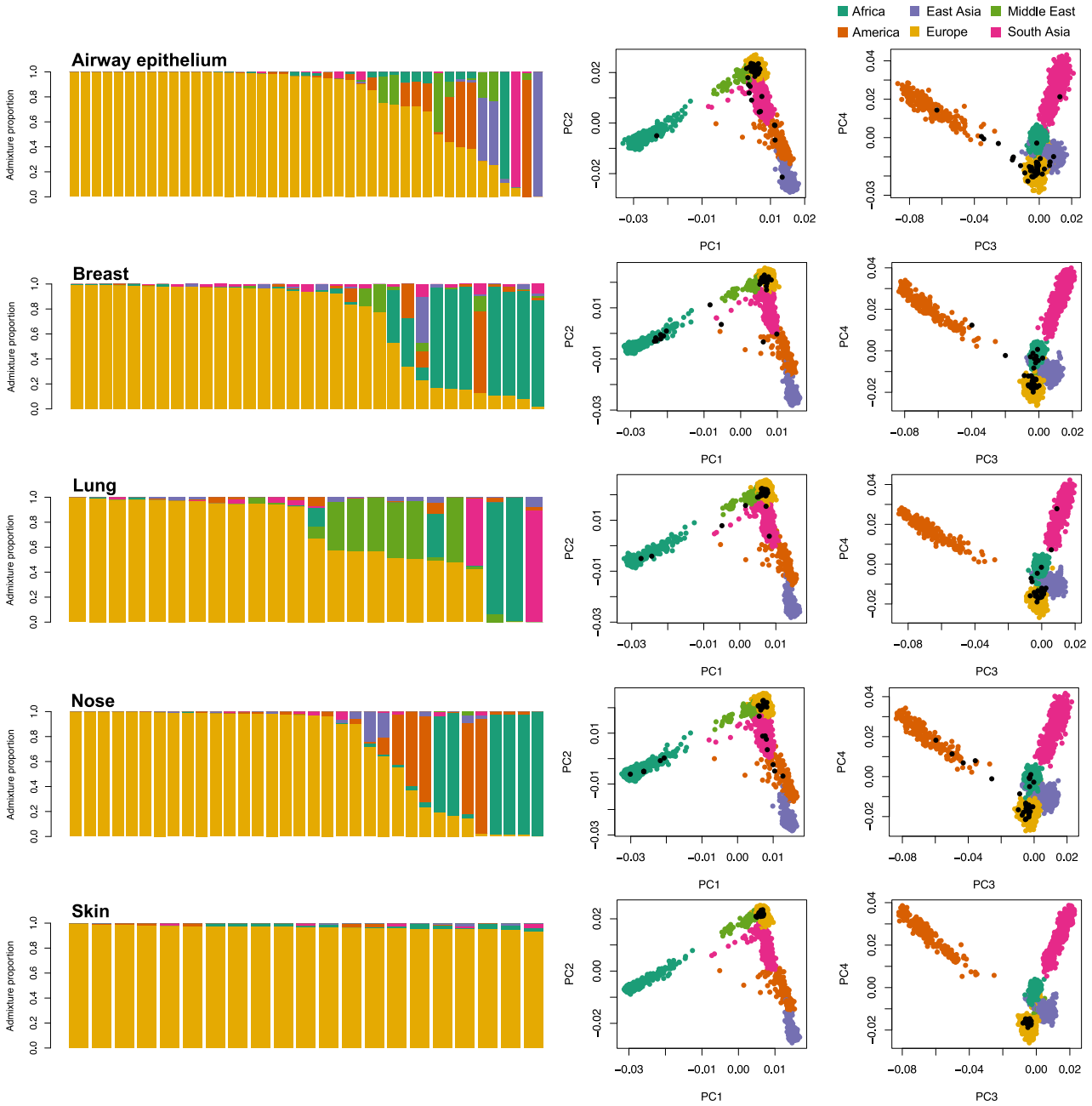

Figure 5. ADMIXTURE and PCA analyses of five HCA single-cell datasets with unreported ancestry
(Left) Admixture proportions for HCA donors from each dataset. (Right) PCA plots of HGDP+1KGP individuals (colored dots) with projected HCA donors (black dots). Results for five additional HCA datasets are reported in Figure S8. Numerical results are reported in Table S9.

inference of genetic ancestry and genetic admixture, despite the limited number of genetic polymorphisms and the presence of variant-calling errors in single-cell sequencing reads. However, we recommend careful interpretation of individuals inferred as a genetic admixture of European and Middle Eastern ancestries. We anticipate that our easy-to-use pipeline (see data and code availability) will enable researchers to infer and document the genetic ancestry of their samples and facilitate the characterization of genetic ancestry across publicly available single-cell datasets.

We observed an overrepresentation of donors with European genetic ancestry in HCA datasets, consistent with observations in RNA-seq datasets[40] and GWASs.[41] Although recent studies have generated large single-cell datasets using hundreds of Asian donors,[34,42] there remains a need to generate datasets in more genetically diverse populations to ensure that existing datasets are representative of human genetic diversity. As recent methods that integrate GWAS results with single-cell data from a few donors have identified precise cellular processes impacting disease,[43,44] generating single-cell

data for even a small number of non-European donors will substantially improve the interpretation of GWASs in 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.[45] 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 with 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 10 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 (~24,000 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.

Data and code availability

Code used in this study is available at: https://github.com/JianingYao/scRNA-seq-genetic-ancestry. Public datasets analyzed in this study are available from the following repositories: Human Cell Atlas (HCA) datasets: https://data.humancellatlas.org; Genotype-Tissue Expression (GTEx) v9: https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_GTEx_V9_hg38; HGDP+1KGP dataset: https://gnomad.broadinstitute.org/downloads.

Acknowledgments

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

Declaration of interests

S.G. reports consulting fees from Eleven Therapeutics that are unrelated to the present work.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2026.100564.

Received: March 27, 2025
Accepted: January 5, 2026

References

1. Nédélec, Y., Sanz, J., Baharian, G., Szpiech, Z.A., Pacis, A., Dumaine, A., Grenier, J.C., Freiman, A., Sams, A.J., Hebert, S., et al. (2016). Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. Cell *167*, 657–669.e21. https://doi.org/10.1016/j.cell.2016.09.025.
2. Quach, H., Rotival, M., Pothlichet, J., Loh, Y.H.E., Dannemann, M., Zidane, N., Laval, G., Patin, E., Harmant, C., Lopez, M., et al. (2016). Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. Cell *167*, 643–656.e17. https://doi.org/10.1016/j.cell.2016.09.024.
3. Idaghdour, Y., Czika, W., Shianna, K.V., Lee, S.H., Visscher, P.M., Martin, H.C., Miclaus, K., Jadallah, S.J., Goldstein, D.B., Wolfinger, R.D., and Gibson, G. (2010). Geographical genomics of human leukocyte gene expression variation in southern Morocco. Nat. Genet. *42*, 62–67.
4. Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.
5. Martin, A.R., Costa, H.A., Lappalainen, T., Henn, B.M., Kidd, J.M., Yee, M.C., Grubert, F., Cann, H.M., Snyder, M., Montgomery, S.B., and Bustamante, C.D. (2014). Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture. PLoS Genet. *10*, e1004549.
6. Hughes, D.A., Kircher, M., He, Z., Guo, S., Fairbrother, G.L., Moreno, C.S., Khaitovich, P., and Stoneking, M. (2015). Evaluating intra- and inter-individual variation in the human placental transcriptome. Genome Biol. *16*, 54.
7. Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., Young, T.R., Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J., et al. (2015). Human genomics. The human transcriptome across tissues and individuals. Science *348*, 660–665.
8. Randolph, H.E., Fiege, J.K., Thielen, B.K., Mickelson, C.K., Shiratori, M., Barroso-Batista, J., Langlois, R.A., and Barreiro, L.B. (2021). Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. Science *374*, 1127–1133.
9. Aquino, Y., Bisiaux, A., Li, Z., O'Neill, M., Mendoza-Revilla, J., Merkling, S.H., Kerner, G., Hasan, M., Libri, V., Bondet, V., et al. (2023). Dissecting human population variation in single-cell responses to SARS-CoV-2. Nature *621*, 120–128.
10. Wang, J., Zhang, Z., Lu, Z., Mancuso, N., and Gazal, S. (2024). Genes with differential expression across ancestries are enriched in ancestry-specific disease effects likely due to gene-by-environment interactions. Am. J. Hum. Genet. *111*, 2117–2128.
11. Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). Science Forum: The Human Cell Atlas. eLife *6*, e27041. https://doi.org/10.7554/eLife.27041.
12. HCA Data Portal. https://data.humancellatlas.org/.

13. Dou, J., Tan, Y., Kock, K.H., Wang, J., Cheng, X., Tan, L.M., Han, K.Y., Hon, C.C., Park, W.Y., Shin, J.W., et al. (2024). Single-nucleotide variant calling in single-cell sequencing data with Monopogen. Nat. Biotechnol. *42*, 803–812.
14. Hong, S.C., Muyas, F., Cortés-Ciriano, I., and Hormoz, S. (2024). scAI-SNP: a method for inferring ancestry from single-cell data. Preprint at bioRxiv. https://doi.org/10.1101/2024.05.14.594208.
15. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., and Myers, R.M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science (New York, N.Y.) *319*, 1100–1104.
16. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.
17. Koenig, Z., Yohannes, M.T., Nkambule, L.L., Zhao, X., Goodrich, J.K., Kim, H.A., Wilson, M.W., Tiao, G., Hao, S.P., Sahakian, N., et al. (2024). A harmonized public resource of deeply sequenced diverse human genomes. Genome Res. *34*, 796–809.
18. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.
19. Carraro, G., Langerman, J., Sabri, S., Lorenzana, Z., Purkayastha, A., Zhang, G., Konda, B., Aros, C.J., Calvert, B.A., Szymaniak, A., et al. (2021). Transcriptional analysis of cystic fibrosis airways at single-cell resolution reveals altered epithelial cell states and composition. Nat. Med. *27*, 806–814.
20. Murrow, L.M., Weber, R.J., Caruso, J.A., McGinnis, C.S., Phong, K., Gascard, P., Rabadam, G., Borowsky, A.D., Desai, T.A., Thomson, M., et al. (2022). Mapping hormone-regulated cell-cell interaction networks in the human breast at single-cell resolution. Cell Syst. *13*, 644–664.e8.
21. Leader, A.M., Grout, J.A., Maier, B.B., Nabet, B.Y., Park, M.D., Tabachnikova, A., Chang, C., Walker, L., Lansky, A., Le Berichel, J., et al. (2021). Single-cell analysis of human non-small cell lung cancer lesions refines tumor classification and patient stratification. Cancer Cell *39*, 1594–1609.e12.
22. Winkley, K., Banerjee, D., Bradley, T., Koseva, B., Cheung, W.A., Selvarangan, R., Pastinen, T., and Grundberg, E. (2021). Immune cell residency in the nasal mucosa may partially explain respiratory disease severity across the age range. Sci. Rep. *11*, 15927.
23. Ganier, C., Mazin, P., Herrera-Oropeza, G., Du-Harpur, X., Blakeley, M., Gabriel, J., Predeus, A.V., Cakir, B., Prete, M., Harun, N., et al. (2024). Multiscale spatial mapping of cell populations across anatomical sites in healthy human skin and basal cell carcinoma. Proc. Natl. Acad. Sci. USA *121*, e2313326120.
24. Giustacchini, A., Thongjuea, S., Barkas, N., Woll, P.S., Povinelli, B.J., Booth, C.A.G., Sopp, P., Norfo, R., Rodriguez-Meira, A., Ashley, N., et al. (2017). Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. Nat. Med. *23*, 692–702.
25. Petropoulos, S., Edsgärd, D., Reinius, B., Deng, Q., Panula, S.P., Codeluppi, S., Reyes, A.P., Linnarsson, S., Sandberg, R., and Lanner, F. (2016). Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell *167*, 285.
26. Cillo, A.R., Kürten, C.H.L., Tabib, T., Qi, Z., Onkar, S., Wang, T., Liu, A., Duvvuri, U., Kim, S., Soose, R.J., et al. (2020). Immune Landscape of Viral- and Carcinogen-Driven Head and Neck Cancer. Immunity *52*, 183–199.e9.
27. Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., Chen, H., Wang, J., Tang, H., Ge, W., et al. (2020). Construction of a human cell landscape at single-cell level. Nature *581*, 303–309.
28. Koenig, A.L., Shchukina, I., Amrute, J., Andhey, P.S., Zaitsev, K., Lai, L., Bajpai, G., Bredemeyer, A., Smith, G., Jones, C., et al. (2022). Single-cell transcriptomics reveals cell-type-specific diversification in human heart failure. Nat. Cardiovasc. Res. *1*, 263–280.
29. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.
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., and Zhou, X. (2019). SNV identification from single-cell RNA sequencing data. Hum. Mol. Genet. *28*, 3569–3583.
32. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., and Zahler, A.M. (2002). Haussler D.The human genome browser at UCSC. Genome Res *12*, 996–1006.
33. Eraslan, G., Drokhlyansky, E., Anand, S., Fiskin, E., Subramanian, A., Slyper, M., Wang, J., Van Wittenberghe, N., Rouhana, J.M., Waldman, J., et al. (2022). Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. Science *376*, eabl4290.
34. Perez, R.K., Gordon, M.G., Subramaniam, M., Kim, M.C., Hartoularos, G.C., Targ, S., Sun, Y., Ogorodnikov, A., Bueno, R., Lu, A., et al. (2022). Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. Science *376*, eabf1970.
35. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
36. Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.
37. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288.
38. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2020). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am. J. Hum. Genet. *107*, 788–789.
39. Bollas, A.E., Rajkovic, A., Ceyhan, D., Gaither, J.B., Mardis, E.R., and White, P. (2024). SNVstory: inferring genetic

ancestry from genome sequencing data. BMC Bioinf. *25*, 76.

40. Romero, I.G., Rodenberg, G., Arner, A.M., Li, L., Beasley, I.J., Rossow, R., Ryan, N., Wang, S., and Lea, A.J. (2024). Public RNA-seq data are not representative of global human diversity. Preprint at bioRxiv. https://doi.org/10.1101/2024.10.11.617967.
41. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591.
42. Tian, C., Zhang, Y., Tong, Y., Kock, K.H., Sim, D.Y., Liu, F., Dong, J., Jing, Z., Wang, W., Gao, J., et al. (2024). Single-cell RNA sequencing of peripheral blood links cell-type-specific regulation of splicing to autoimmune and inflammatory diseases. Nat. Genet. *56*, 2739–2752.
43. Zhang, M.J., Hou, K., Dey, K.K., Sakaue, S., Jagadeesh, K.A., Weinand, K., Taychameekiatchai, A., Rao, P., Pisco, A.O., Zou, J., et al. (2022). Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. Nat. Genet. *54*, 1572–1580.
44. Jagadeesh, K.A., Dey, K.K., Montoro, D.T., Mohan, R., Gazal, S., Engreitz, J.M., Xavier, R.J., Price, A.L., and Regev, A. (2022). Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-sequencing and human genetics. Nat. Genet. *54*, 1479–1492.
45. Lewis, A.C.F., Molina, S.J., Appelbaum, P.S., Dauda, B., Di Rienzo, A., Fuentes, A., Fullerton, S.M., Garrison, N.A., Ghosh, N., Hammonds, E.M., et al. (2022). Getting genetic ancestry right for science and society. Science *376*, 250–252. https://doi.org/10.1126/science.abm7530.