frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 25 April 2023
DOI 10.3389/fimmu.2023.1146826

Prediction of HLA genotypes from single-cell transcriptome data

Benjamin D. Solomon[1]*, Hong Zheng[2,3], Laura W. Dillon[4], Jason D. Goldman[5,6,7], Christopher S. Hourigan[4], James R. Heath[8,9] and Purvesh Khatri[2,3]

[1]Department of Pediatrics, Stanford University, Palo Alto, CA, United States, [2]Institute for Immunity, Transplantation and Infection, School of Medicine, Stanford University, Stanford, CA, United States, [3]Center for Biomedical Informatics Research, Department of Medicine, School of Medicine, Stanford University, Stanford, CA, United States, [4]Laboratory of Myeloid Malignancies, National Heart Lung and Blood Institute, Bethesda, MD, United States, [5]Swedish Center for Research and Innovation, Swedish Medical Center, Seattle, WA, United States, [6]Providence St. Joseph Health, Renton, WA, United States, [7]Division of Allergy & Infectious Diseases, University of Washington, Seattle, WA, United States, [8]Institute for Systems Biology, Seattle, WA, United States, [9]Department of Bioengineering, University of Washington, Seattle, WA, United States

The human leukocyte antigen (HLA) locus plays a central role in adaptive immune function and has significant clinical implications for tissue transplant compatibility and allelic disease associations. Studies using bulk-cell RNA sequencing have demonstrated that HLA transcription may be regulated in an allele-specific manner and single-cell RNA sequencing (scRNA-seq) has the potential to better characterize these expression patterns. However, quantification of allele-specific expression (ASE) for HLA loci requires sample-specific reference genotyping due to extensive polymorphism. While genotype prediction from bulk RNA sequencing is well described, the feasibility of predicting HLA genotypes directly from single-cell data is unknown. Here we evaluate and expand upon several computational HLA genotyping tools by comparing predictions from human single-cell data to gold-standard, molecular genotyping. The highest 2-field accuracy averaged across all loci was 76% by arcasHLA and increased to 86% using a composite model of multiple genotyping tools. We also developed a highly accurate model (AUC 0.93) for predicting *HLA-DRB345* copy number in order to improve genotyping accuracy of the *HLA-DRB* locus. Genotyping accuracy improved with read depth and was reproducible at repeat sampling. Using a metanalytic approach, we also show that HLA genotypes from PHLAT and OptiType can generate ASE ratios that are highly correlated (R^2 = 0.8 and 0.94, respectively) with those derived from gold-standard genotyping.

KEYWORDS

HLA typing algorithm, next-generation sequencing data (NGS), HLA genotype, single-cell sequencing (scRNA-seq), human leukocyte antigen (HLA), major histocompatibility (MHC), allele specific expression

Check for updates

OPEN ACCESS

EDITED BY
Giovanna Del Pozzo,
National Research Council (CNR), Italy

REVIEWED BY
Sudeep Kumar Maurya,
University of Pittsburgh Medical Center,
United States
Michael Peres,
Centre Hospitalier Universitaire de Nancy,
France

*CORRESPONDENCE
Benjamin D. Solomon
solomonb@stanford.edu

SPECIALTY SECTION
This article was submitted to
Systems Immunology,
a section of the journal
Frontiers in Immunology

RECEIVED 17 January 2023
ACCEPTED 04 April 2023
PUBLISHED 25 April 2023

CITATION
Solomon BD, Zheng H, Dillon LW,
Goldman JD, Hourigan CS, Heath JR and
Khatri P (2023) Prediction of HLA
genotypes from single-cell
transcriptome data.
Front. Immunol. 14:1146826.
doi: 10.3389/fimmu.2023.1146826

Highlights

- Benchmarking of HLA genotype prediction accuracy from 5'- and 3'- based single-cell RNA-seq data compared to molecular HLA genotyping
- Quantification of transcript coverage by 5'- and 3'-based single cell sequencing methods
- Accurate prediction of complex *HLA-DRB345* copy numbers using supervised learning
- Balancing accuracy and performance through composite HLA genotyping
- Meta-analytic approach to summarizing allele-specific expression of HLA genotypes at single-cell level

Introduction

The human leukocyte antigen (HLA) locus is the most polymorphic region of the human genome. It encodes a wide range of important immune proteins, including the class I (HLA-A, -B, -C) and class II (HLA-DP, -DQ, -DR) major histocompatibility complex (MHC) proteins responsible for antigen presentation. The allelic diversity of the HLA locus contributes to the ability of MHC proteins to present a large range of possible peptides to lymphocytes and underlies the association of HLA genotypes with transplant compatibility and disease susceptibility (1).

HLA alleles are identified by four fields of increasing specificity, each with two or more digits. Together, the first two "low-resolution" fields define unique protein coding sequences, while the final two "high-resolution" fields denote synonymous -exonic and -intronic nucleic acid variation, respectively. The earliest HLA typing methods relied on low-resolution serologic assays that have largely been replaced by high-resolution molecular genotyping based on a variety of sequencing techniques. Clinical HLA genotyping typically utilizes site-specific amplicon sequencing of HLA loci, while research applications have increasingly focused on next generation sequencing (NGS) (2, 3).

HLA transcription is dynamically regulated throughout an inflammatory process and also represents a common immune escape mechanism. For example, reduced expression of MHC class 2 genes on monocytes correlates with mortality in septic shock (4) and multiple viral infections (5). Regulation of HLA expression also extends to the level of individual alleles (6). Temporal patterns in allele-specific expression (ASE) of certain *HLA-DQB1* alleles can discriminate their relative association with type 1 diabetes (7) and increased ASE imbalance of MHC class I genes was observed in colorectal (8) and other cancers (9). Single-cell sequencing allows for fine resolution of ASE and such studies have demonstrated much of the transcriptome is mono-allelically expressed at a given point in time (10–12). However, while methods for assessing single-cell ASE for HLA loci are available (13), they require high-resolution reference genotypes for accurate read mapping due to extensive HLA polymorphism, limiting their broader application to the increasingly large amount of publicly-available sequencing data.

Several computational tools exist to predict HLA genotypes from bulk RNA sequencing data and the accuracy of these methods have been robustly characterized (14). However, it is unclear if these methods can be applied to data from single-cell RNA sequencing (scRNA-seq) experiments due to differences in sequencing chemistry and transcriptome coverage. One attempt to obtain genotypes from single cells observed that very few cells express HLA genes at a high enough level to generate genotype predictions for individual HLA loci (15).

Here, we show that data from commonly used scRNA-seq platforms can be condensed into subject-specific "pseudobulk" sequence files that can be used to predict HLA genotypes. Using five *in silico* HLA genotyping tools, we compare the accuracy of genotypes derived from scRNA-seq to gold standard molecular HLA genotyping obtained from the same individuals. We further expand on these methods to better predict complex *HLA-DRB345* genotypes and obtain maximal accuracy using a composite of the tested genotyping tools. We also show that even inaccurate genotypes from several tools can result in an accurate assessment of ASE in downstream use.

Methods

HLA multiple sequence alignment and variation

HLA-allele sequences were obtained from IMGT/HLA version 3.42.0. All sequences without atypical expression suffixes (e.g. –N, -L, etc.) were included in multiple sequence alignment performed by DECIPHER/2.18.1 in R/4.0.4. Sequence variability was determined using 2-bit Shannon entropy based on nucleotide identity at each sequence position. Gap nucleotides were not factored into entropy calculations, as published HLA reference alleles often include only partial exon sequences

Samples, sequencing, and gold standard HLA genotyping

Samples used for 5' scRNA-seq have been described previously (16) as part of the ISB-Swedish INCOV study and include peripheral blood samples obtained from 157 patients at one or two time points. Samples were processed using 10X Genomics 5' Chromium Single Cell Kits and sequenced using the Illumina Novaseq platform. Samples used for 3' scRNA-seq and bulk RNA-seq have been previously described (17) as part of the NIH-HBM cohort and include bone marrow samples from 20 healthy patients. Samples were processed with using 10X Genomics Single Cell 3′ Solution Kits, TruSeq Stranded Total (bulk) RNA Sample Preparation Kits, or both. Libraries were sequenced on Illumina HiSeq 3000. For both sample cohorts, 3-field molecular HLA genotyping was performed by Scisco Genetics on peripheral blood aliquots or gDNA. As described by the manufacturer, Scisco HLA genotyping utilizes 2-stage, amplicon-based PCR amplification of genomic DNA. Four multiplex primer sets provide complete amplification of HLA-A, B, and C at exons 1 through 8; DRB1, DRB3/B4/B5, DQA1, and DQB1 at exons 1 through 6; DPB1 at exons 1 through 5; and DPA1 at exons 1 through 4. This represents complete coverage of all exons. Primers also include sufficient intron sequences to detect all known intron encoded null alleles. Sequencing is performed on the Illumina MiSeq platform using 500 cycles. The minimal read depth criteria for genotyping calling is 50 reads per amplicon.

FASTQ file preparation

For scRNA-seq samples, raw FASTQ files were demultiplexed into subject-specific FASTQ files utilizing UMI-tools/1.1.1. Quality control was performed with TrimGalore/0.6.5. Files were then mapped to the GRCh38 human reference genome with HISAT2/2.2.1 and indexed with SAMTOOLS/1.9. The `extract` function of arcasHLA/0.2.0 was used to isolate chromosome 6 and unmapped reads to be used for downstream HLA mapping and genotyping.

HLA genotype prediction

Genotyping by arcasHLA was performed with the `genotype` function. The `parameters.p` file in arcasHLA was modified to allow for genotyping of the HLA-DRB4 locus. Genotyping by PHLAT/1.1, OptiType/1.3.3, and HLAMiner/1.4 and scHLAcount/DEV were performed using default settings. When able to specify a user-defined HLA reference, IMGT/HLA version 3.42.0 was used.

Prediction validity

For a given genotyper, success compares only those loci that a genotyper generated a prediction for to the gold standard and is represented by:

$$Success = \frac{\#\,Correct\ alleles}{\#\,Correct\ alleles\ +\ \#\,Incorrect\ alleles}$$

In contrast, accuracy reflects the validity of a given genotyper across all alleles in the gold standard, even those that the genotyper failed to generate a prediction for. This is analogous to Bray-Curtis similarity and is represented by:

$$Accuracy = \frac{\#\,Correct\ alleles}{\#\,Correct\ alleles\ +\ Incorrect\ alleles\ +\#\,Absent\ alleles}$$

"Correct allele" indicates an allele prediction that matches the ground truth allele. Agreement between two predicted genotypes is a case of accuracy where "correct allele" indicates the two predicted genotypes are identical. Genotype accuracy and success were assessed independently at each HLA locus and thus only take values of 0.0, 0.5, or 1.0, reflecting 0, 1, or 2 correct alleles out of 2 possible alleles. All assessments of accuracy, success, or agreement at each field includes the full genotype up to that field. For example, in the case of a ground-truth genotype of HLA-A*02:01 and a genotyping algorithm prediction of HLA-A*01:01, field 2 accuracy represents the comparison of 02:01 and 01:01, not xx:01 *vs*. xx:01.

Read depth analysis

For read subsampling, FASTQ files generated by arcasHLA extract were subsampled to 10%, 1%, and 0.1% of their original read count using seqtk/1.3. For cell subsampling, cellular barcodes were subsampled to 10%, 1%, and 0.5% of total sample barcodes. All reads corresponding to subsampled barcodes were isolated from FASTQ files generated by arcasHLA extract using BBMAP/38.90. Subsampled files were genotyped as described above. Linear modeling of genotyping accuracy/success by read/cell depth was performed with ordinary least squares regression on log-transformed read counts using base R/4.0.4.

Statistical modeling

HLA-DRB345 allele copy numbers were predicted from a K-nearest neighbor classifier using the ratio of *HLA-DRB3, -DRB4*, or *-DRB5* mapped reads to *HLA-DRB1* mapped reads as feature and the number of *HLA-DRB3, -DRB4, -DRB5* alleles identified by molecular genotyping as ground truth. Multi-class AUC was determined using the Hand & Till method (18). Manhattan distance was used to quantify similarity between the allele count vectors {*-DRB3, -DRB4*, *-DRB5*} obtained from the kNN model and molecular genotyping. This distance was standardized to the maximum possible distance of 6, representing complete dissimilarity of the two alleles at each of the three loci. This value was then inverted to create an index where 1 represents complete similarity of all allele counts and 0 represents complete dissimilarity. Decision trees for composite HLA genotypes included presence/absence of a valid genotype from each genotyper, locus, and field level as input features and the identity of the highest accuracy genotyper for a given sample as ground truth. Both models were trained on 70% of the 5' scRNA-seq samples with 10-fold cross validation and tested on the 30% hold-out set using Tidymodels/0.1.3 in R/4.0.4.

Analysis of single-cell RNA-seq data

Gene expression analysis of 5' scRNA-seq samples including pre-processing, integration, and cellular annotation was published previously (5). Allele specific expression was determined by scHLAcount/DEV (commit 5ce7b2d) and incorporated into single cell data using Seurat.

Allele-specific expression effect size

To prevent lower-confidence ASE ratios from cells with low read counts from biasing the overall sample ASE, we determined a summary effect size using the log-odds of each cell's HLA expression ratio. For an individual cell, expected allele counts were calculated by summing the total number of reads between the two alleles, then distributing them equally between the two alleles. An odds ratio was obtained from the observed counts and this expected count, with the highest expressed allele as reference. To obtain a summary effect size across all cells, this odds ratio was converted to log-odds then weighted by the inverse variance of the log-odds. These weighted odds were summed in a random-effects model to find the effect size. Computation was performed using the R-package meta/4.19.0.

Data and code availability

All datasets, software, and algorithms used in this study are publicly available and appropriately cited as they are introduced. All code necessary to reproduce our analyses and to implement our models is available at https://github.com/BenSolomon/hla_benchmark. Any gold standard HLA genotyping not previously published is included in this code repository. For convenience, compiled HLA genotypes from all samples and genotyping methods are also included as supplementary data files.

Results

Single cell sequencing methods can produce sufficient sequence coverage to assess HLA sequence diversity

Compared to paired-end bulk RNA-seq methods, most high throughput scRNA-seq methods generate cDNA libraries that are enriched for either the 3' or 5' ends of transcripts. Given the extensive polymorphism of the HLA locus, incomplete sequence coverage could impair genotyping accuracy by excluding sequence variations outside of these enriched regions. This is particularly true for 3'-based protocols, as HLA diversity is concentrated in the 5' region of each gene (Figure 1A).

To quantify the relationship between HLA-sequence diversity and the positional bias of sequencing platforms, we compared the HLA loci coverage of 3'- and 5'-based scRNA-seq data, as well as bulk RNA-seq data. We selected two RNA-seq data sets with matched molecular HLA genotyping for this comparison and all downstream analyses. The first dataset, referred to as INCOV, included 5'-based scRNA-seq data from blood samples of 157 individuals taken at multiple time points (16). The second dataset, referred to as NIH-HBM, included both 3'-based scRNA-seq and bulk RNA-seq data from healthy bone marrow donors (17).

Interestingly, while both 3' and 5' coverage bias was evident in the respective sequencing methods, both resulted in reads that could be mapped across the full extent of HLA loci (Figure 1B). In particular, coverage from 5' based scRNA-seq data was comparable or better than that of bulk RNA-seq across the full extent of most HLA loci. This was further reflected in the overall mapping efficiency of reads in each HLA loci, as the median number of reads aligned to HLA loci from 5'-based scRNA-seq methods were an order of magnitude greater than both bulk RNA-seq and 3'-base RNA-seq data (Figure 1C).

ScRNA-seq pseudobulk data results in few HLA genotyping failures

Darby et al. previously demonstrated that individual cells do not express HLA transcripts at a sufficient level for reliable and complete genotyping at the single cell level (15). Therefore, we sought to achieve sufficient read coverage at HLA loci by pooling reads from all cells of an individual sample into a "pseudobulk" data set. As most single-cell methods generate single-end reads, we

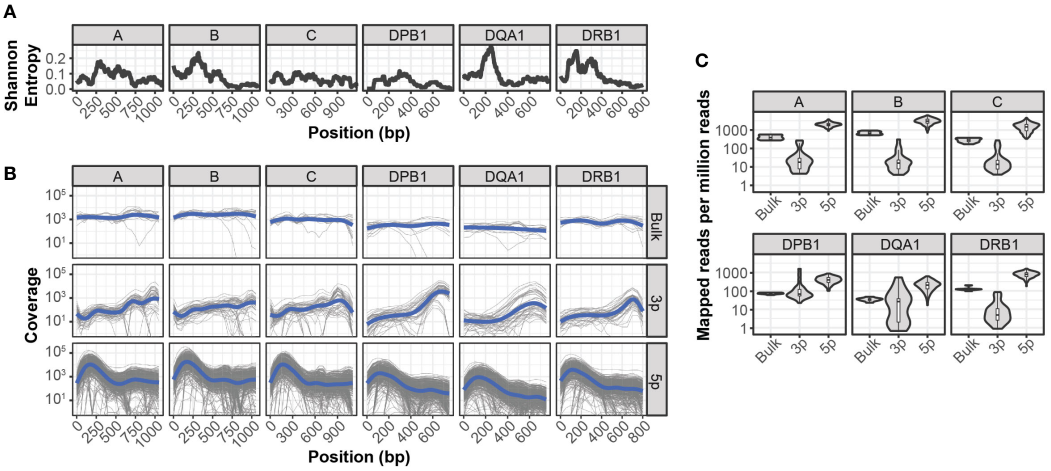

FIGURE 1
Sequencing coverage of HLA allelic diversity **(A)** Rolling (100bp) mean Shannon entropy for published allele sequences of indicated HLA loci. **(B)** HLA coverage of reads mapped from bulk RNA-seq, 3' (3p-based) scRNA-seq, and 5' (5p)-based scRNA-seq. Grey lines represent individual samples, blue lines represent loess regression. **(C)** HLA-mapped reads per million total reads from bulk RNA-seq, 3' (3p)-based scRNA-seq, and 5' (5p)-based scRNA-seq.

identified 4 compatible HLA genotyping tools: arcasHLA (19), HLAminer (20), PHLAT (21), and OptiType (22). Though primarily used to quantify ASE at HLA loci, we also compared predictions from the genotyping function of scHLAcount (13). To control and minimize the variability resulting from the different global alignment strategies incorporated into each of these tools, we first performed common global alignment with HISAT2 then isolated chromosome 6 reads (containing the HLA locus) and unmapped reads to use as a starting point for each genotyper.

Overall, most tools generated two allele predictions for the first two HLA fields, while predictions at field 3 had a higher rate of failure (Figure 2A). Notably, arcasHLA and PHLAT made complete 2-allele predictions across all loci, with the exception of *HLA-DPA1* and *HLA-DPB1* which PHLAT does not assess. scHLAcount was a notable exception, typically producing only a single allele prediction.

Genotyping accuracy varies by loci, genotyper, and sequencing direction

We compared predicted HLA genotypes to ground-truth sample-matched molecular genotyping using two metrics previously utilized by Bauer et al. (14). "Success" describes the proportional match of a predicted genotype with the ground truth genotype but ignores missing predictions. Conversely, "accuracy" assesses how likely a genotyper is to predict the complete, correct genotype by penalizing failed genotype predictions.

We focused primarily on 5'-based scRNA-seq INCOV data due to its greater transcript coverage. For class 1 genes, OptiType had the highest success at fields 1 and 2, though lower accuracy compared to PHLAT due to occasional prediction failures of OptiType (Figures 2B, C; Tables S1; S2). aracasHLA had moderate success and accuracy for class 1 genes (55% - 64% at field 2), but had substantially higher accuracy for class 2 genes (75% - 93% at field 2). Its accuracy for class 2 genes and was only surpassed by PHLAT at *HLA-DQB1*. Notably, HLAMiner had the lowest success and accuracy compared to the other genotyping tools, with an average field 2 accuracy of less than 50% across all HLA loci. The genotyping function of scHLAcount also had poor accuracy. We excluded scHLAcount from further analysis due to its low accuracy and high rate of incomplete genotype predictions. This variation in accuracy did not result from bias of particular genotypes for specific alleles or samples (Figures S2A, B).

3'-based scRNA-seq NIH-HBM data resulted in lower accuracy than 5'-based data. The average 2-field MHC class 2 accuracy for arcasHLA ranged from 32% - 95% (Figure S1A; Table S3). Interestingly, MHC class 1 accuracy appeared less affected by the reduced sequence coverage compared to MHC class 2 accuracy. By comparison, bulk RNA-sequencing frequently resulted in excellent accuracy (Figure S1B; Table S4) such as average 2-field ranging from 88%-100% across all HLA loci for arcasHLA.

Increasing accuracy of HLA-DRB345 predictions

The *HLA-DRB1* gene can occur alone or in close linkage disequilibrium with one of three functional paralogs, *HLA-DRB3*,

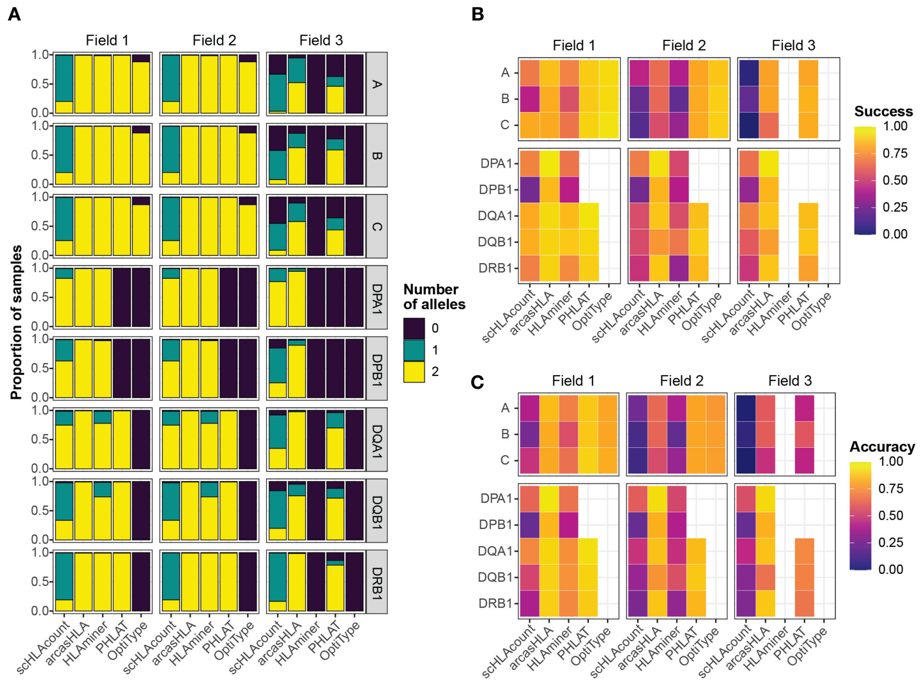

FIGURE 2
Accuracy of HLA genotype predictions **(A)** Relative proportion of samples with 0, 1, or 2 allele predictions. **(B)** Mean success and **(C)** mean accuracy of predicted genotypes compared to molecular genotyping. Empty heatmap tiles represent genotype parameters not assessed by a given genotyper.

-*DRB4*, or -*DRB5*, and pairing of *HLA-DRB1* and *HLA-DRB345* on each homologous chromosome occurs independently (23). While arcasHLA and HLAminer can predict *HLA-DRB345* identity, neither accounts for copy number, frequently resulting in biologically invalid *HLA-DRB345* predictions. While tools such as CaSpER (24) are capable of predicting copy number variants, the significant linkage disequilibrium and high sequence homology between genes of the *HLA-DRB345* locus prevents the application of conventional methods.

To address this, we sought to predict the number of *HLA-DRB3* -*DRB4*, and -*DRB5* copies prior to assessing allele accuracy. Zhang et al. previously showed that copy numbers could be predicted from targeted sequencing using the ratio of *HLA-DRB345* reads to *HLA-DRB1* reads (25). Using a similar approach, we trained a K-nearest neighbor (kNN) classifier to predict the number of *HLA-DBR3*, -*DRB4*, and –*DRB5* alleles based on their relative read abundance (Figure S3A). We chose kNN for its low complexity and minimal assumptions when applied to multiclass modeling. We used 70% of the 5'-based scRNA-seq INCOV data to train the kNN-based classifier. The model was highly accurate when applied to a holdout subset (AUROC = 0.97; Figure 3A). This model generalized well to 3'-based scRNA-seq and bulk RNA-seq data from NIH-HBM, with an AUC of 0.79 and 1.0, respectively (Figures S3D, F).

Prior to application of the kNN classifier, copy numbers associated with genotype predictions were notably discordant from those reflected in ground truth genotypes. In samples with no HLA-DRB3 copies per molecular genotyping, arcasHLA and HLAminer correctly predicted an HLA-DRB3 copy number of zero in only 30% and 3% of these samples, respectively (Figure 3B). By comparison, the kNN classifier correctly identified 93% of these zero HLA-DRB3 copy number samples. When similarity to ground truth copy numbers was summarized across all HLA-DRB345 loci, the kNN classifier had a mean standardized similarity of 0.97, compared to 0.64 and 0.74 for arcasHLA and HLAminer, respectively (Figure 3C).

Using these copy number predictions, we filtered the *HLA-DRB345* genotypes from arcasHLA and HLAminer to their top *n* alleles, where *n* is the number of copies predicted by our kNN classifier. This significantly improved *HLA-DRB345* accuracy, indicative of a high frequency of false positive predictions in unfiltered genotypes. Accuracy ranges for 2-field arcasHLA predictions increased from 35%-46% using unfiltered samples to 66%-88% with filtered samples. (Figures 3D, E; Table S5). As with other loci, arcasHLA significantly outperformed HLAminer.

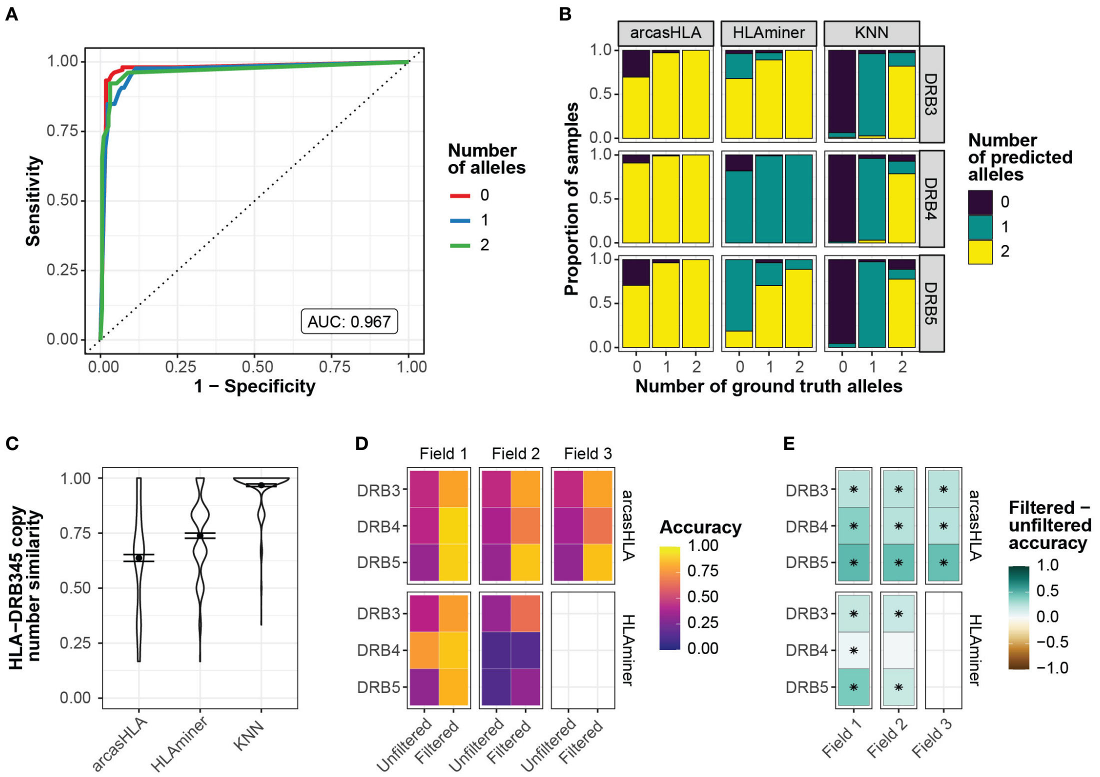

FIGURE 3
Prediction of HLA-DRB345 genotypes **(A)** ROC for HLA-DRB345 copy number classifier by kNN. Tuned on 70% test set with 10-fold cross validation. AUC represents performance on 30% hold-out test set. **(B)** Relative proportion of samples with ground truth HLA-DRB3, -DRB4, and -DRB5 copy numbers of 0, 1, or 2 with corresponding predicted copy numbers by arcasHLA, HLAminer, or the kNN classifier applied to all samples. **(C)** Similarity between the set of ground truth and predicted HLA-DRB3, -DRB4, and -DRB5 copy numbers for each sample. Similarity represents 1 - maximum-standardized Manhattan distance. Error bars represent standard error of the mean. **(D)** Mean accuracy of HLA-DRB345 genotype predictions when unfiltered or filtered by kNN classifier. **(E)** Difference in filtered and unfiltered mean accuracy. (*) FDR-adjusted p-value ≤ 0.05 by Wilcoxon rank sum test.

Genotype predictions from scRNA data are precise

While HLA genotypes are static, differences in HLA gene expression levels might affect genotype predictions when based on scRNA-seq data. The 5'-based scRNA-seq dataset contained samples from individuals obtained at two time points, allowing us to assess prediction reproducibility. Both time points had similar prediction patterns, with no significant variation in accuracy when compared to molecular HLA genotyping (Figure 4A; Table S6). In the case of arcasHLA, the difference in average 2-field accuracy between the two time points ranged from 0%-12%.

Next, we directly compared how well the genotypes from both time points agreed with one another. An expected agreement between time points was derived from the product of their respective accuracy values. For arcasHLA and OptiType there was no significant difference between expected and observed agreement at the 2-field level with the exception of arcasHLA at *HLA-DQA1*, which showed lower observed agreement (Figure 4B). Alternatively, observed 2-field agreement was either equivalent to or significantly higher than expected agreement for HLAminer and PHLAT across all HLA loci.

Increased read depth improves genotyping accuracy

Previous studies have shown conflicting results regarding the effect of read depth on the accuracy of HLA genotyping from NGS data (14) (26). We evaluated this relationship by assessing genotype accuracy from multiple levels of subsampled reads. Overall, accuracy increased with read depth for most genotypers and loci (Figure 5). OptiType accuracy was relatively constant over read depth due in part to genotyping failures at higher read counts (Figure S4). Conversely, the positive trend between read depth and accuracy was most prominent with arcasHLA, accentuated by its minimum read parameter that excludes genotype predictions for loci with a total number of mapped reads below a specified threshold.

Composite genotypes increase overall accuracy

Since no single genotyper produced the most accurate predictions across all samples and HLA loci, we sought to determine if combing multiple predictions could increase accuracy. The genotypers tested in our study have significant variation in run time ranging from a median runtime of 1.6 minutes per sample for arcasHLA to a median runtime of 183 minutes per sample for HLAMiner (Figure 6A). As such, when combining genotyper predictions, we also sought to balance any gain in accuracy with the increased processing time needed to run multiple tools.

Lee et al. previously demonstrated success with a consensus-voting approach to determine the reliability of HLA genotyping from genomic sequencing data (27). Using a related approach, we trained a decision tree classifier to determine the most accurate genotyper based on (1) the set of all genotypers with successful predictions for an individual sample, (2) the HLA locus, and (3) the field level. We trained two trees, one that incorporated arcasHLA, OptiType, and PHALT (AOP) and one with only arcasHLA and OptiType (AO) (Figures S6A, B). HLAminer was excluded entirely due to its long run time and overall low accuracy. When tested on a 30% hold out dataset, both models were highly accurate at identifying the optimal genotyper for a given sample (AOP AUC 0.84, AO AUC 0.93).

We then created a composite HLA genotype for each individual based on the genotype prediction from the optimal genotyper at each locus and allele field level. Composite accuracy was consistently higher than accuracy from individual genotypers (Figures 6B, C; Table S7). 2-field accuracy from the AOP composite ranged from 84%-86% for MHC class 1 loci and 69%-93% for MHC class 2 loci. Notably, the accuracy of predictions from the AO composite were not significantly different from those obtained by the AOP composite with the exception of *HLA-DQB1*, suggesting that a near optimal prediction can be obtained from the combination of only arcasHLA and OptiType. Moreover, the addition of arcasHLA to OptiType had a negligible impact on

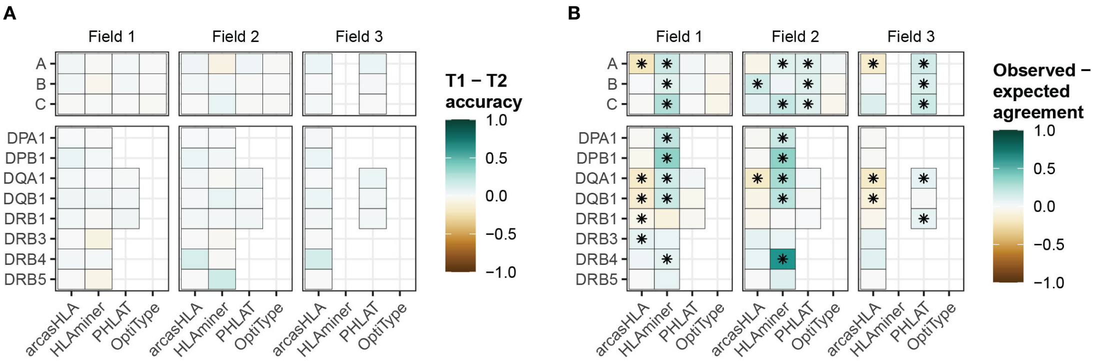

FIGURE 4
Reproducibility of predicted HLA genotypes **(A)** Difference in mean accuracy between genotype predictions at time point 1 (T1) and time point 2 (T2). **(B)** Mean difference between the expected agreement of T1 and T2 genotype predictions and their observed agreement. (*) FDR-adjusted p-value ≤ 0.05 by Wilcoxon rank sum test.

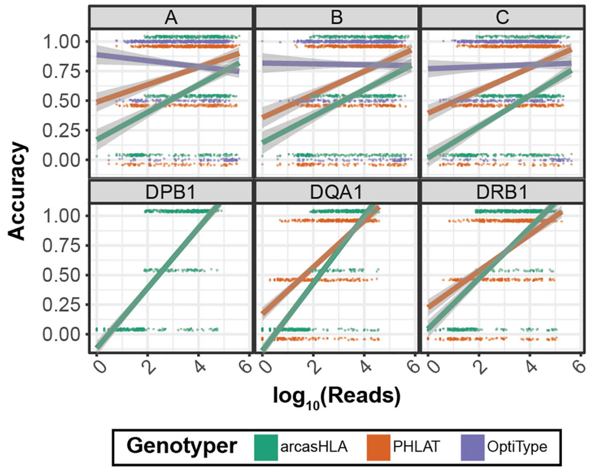

FIGURE 5
Effect of sequence sampling depth on HLA predictions 2-field accuracy based on number of reads aligned to the indicated locus. Raw sequencing data sampled to 100%, 10%, 1%, and 0.1% of original sample read total. Points represent individual samples offset from {0, 0.5, 1} to distinguish genotypers. Trend lines represent linear regression using log transformed reads with ribbon representing 95% CI.

total pipeline runtime, increasing median runtime from 73.5 minutes to 75.0 minutes, while composite of arcasHLA, OptiType, and PHLAT increased median total runtime to 154.5 minutes (Figure S6E).

Inaccurate genotype predictions can result in accurate determination of HLA allele-specific expression

Single-cell studies provide enhanced resolution of allele-specific transcriptional regulation, though analysis of ASE at HLA loci is complicated by the effect of HLA polymorphism on read mapping. This can be overcome by supplying sample-specific HLA references to tools like scHLAcount (13), though it is unclear how much allele-mistyping affects downstream determination of ASE ratios. To address this question, we evaluated how variation in accuracy from different HLA genotypers affected ASE.

In a representative sample (Figure 7A), inaccurate genotypes resulted in broadly different patterns of ASE (Figures 7B; S7A). To quantify this apparent difference in ASE across cells, we used a meta-analytic approach to summarize ASE ratios. When focusing on CD14+ monocytes from a representative sample, this approach demonstrated that the genotyping inaccuracy associated with HLAminer resulted in a notable deviation in the ASE log-odds ratio from that obtained using the ground truth genotype (Figures 7C; S7B).

When comparing ground truth- and genotyper-derived ASE log-odds ratios across all samples and cell types, PHLAT and OptiType resulted in the greatest correlation with ground truth-derived ASE, with R^2 values of 0.8 and 0.94, respectively (Figure 7D). Interestingly, this high level of correlation occurred even in samples where both allele predictions were incorrect, though not when only a single allele prediction was correct (Figure S7). This suggests that, even when inaccurate, these tools predict genotypes with sufficient sequence similarity to the ground truth genotype to allow for similar proportional mapping of reads to the two reference alleles.

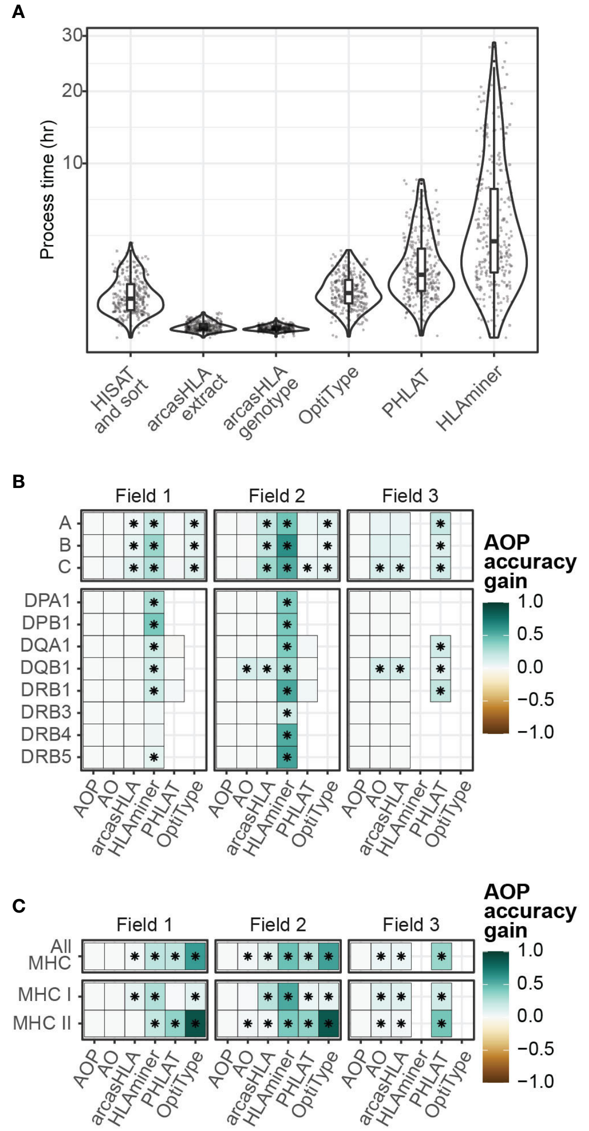

FIGURE 6
Composite HLA prediction accuracy. **(A)** Runtime of major steps in genotyping pipeline. Y-axis distributed along square root scale. **(B, C)** Difference in mean accuracy between AOP composite and indicated alternative genotyping methods across **(B)** individual loci and **(C)** loci classes. (AOP) composite of arcasHLA, OptiType, PHLAT, (AO) composite of arcasHLA, OptiType. (*) FDR-adjusted p-value ≤ 0.05 by Wilcoxon rank sum test.

Discussion

In this study, we describe the feasibility and best practices for obtaining HLA genotypes from scRNA-seq data. Using bulk RNA-

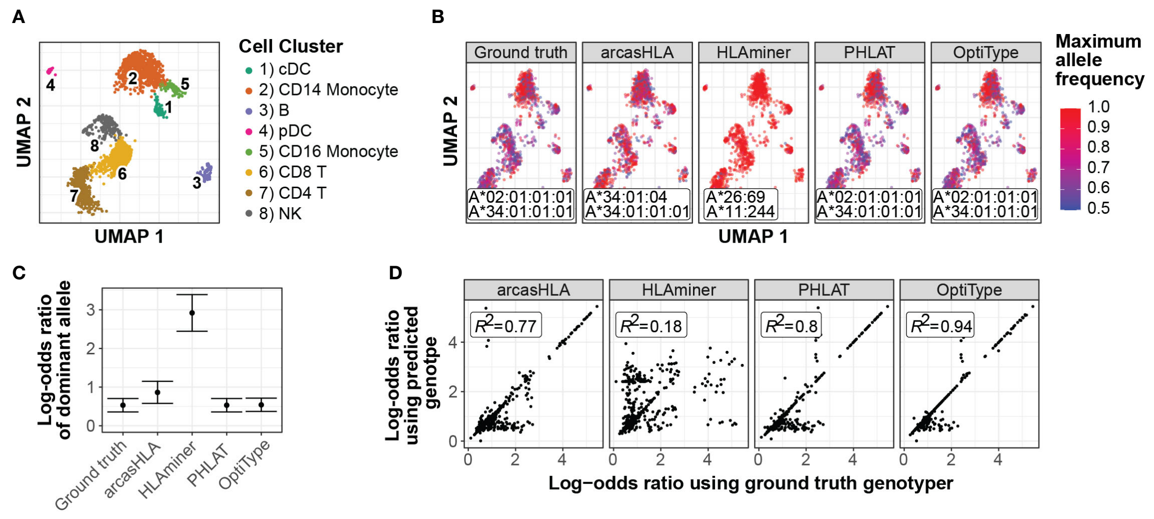

FIGURE 7
Effect of prediction accuracy on analysis of allele-specific HLA-A expression **(A-C)** Representative sample from 5′ data set. **(A)** UMAP projection of single-cell transcriptome data, annotated as previously described (4) **(B)** UMAP plots colored by cell-specific expression frequency of the most highly expressed HLA allele determined by scHLAcount. The reference genotype from each genotyper used by scHLAcount is annotated. **(C)** Summary log-odds ratios of dominant HLA-A allele from all cells in cDC cluster determined by random effects model. Error bars represent summary standard error. **(D)** Correlation of ground truth- and genotyper-derived HLA-A allele log-odds ratio from all samples and cell types.

seq data as a baseline, we found that our evaluation pipeline generated similar accuracy results to those seen in other benchmarking studies (14) (19). By comparison, sequencing from 5'-based single-end scRNA-seq data was moderately accurate with average accuracy across all loci for arcasHLA of 76% and 74% for 2-field and 3-field genotypes, respectively. This accuracy could be increased by composite predictions assembled using a decision tree model from multiple genotypers, resulting in an average accuracy of 86% and 78% for 2-field and 3-field genotypes, respectively.

In order to obtain accurate genotype predictions of highly polymorphic regions like the HLA locus, full transcript coverage is critical to ensure all underlying sequence diversity is incorporated into predictions. We observed that 5'-based scRNA-seq methods provides excellent read coverage across the entire length of HLA transcripts compared to 3'-based scRNA-seq coverage, which is more restricted and enriched for regions with lower HLA polymorphism. Both protocols utilize a template switch reaction, though the proportion of the poly-A tail binding region of reverse-transcription primer used in the 5'-based protocol is nearly double that of the primer in the 3'-based protocol, possibly contributing to the difference in coverage we observe. The observations that transcript coverage from 5'-based scRNA-seq is sufficient to genotype highly polymorphic HLA loci, suggests that this platform could also reliably quantify other areas of small sequence variation such as single nucleotide polymorphisms.

Our observation that genotyping accuracy improves with increased locus coverage contrasts with that of Bauer et al. (14). Several differences may account for this. First, the bulk RNA-seq data analyzed previously is likely more robust to positional HLA sequence variation than the single-end scRNA-seq data assessed here. As such, more read depth may be necessary to overcome the relative loss of allele-identifying sequence information in single-ended libraries compared to paired-end libraries. In addition, our subsampling approach ensured that our analysis would test genotypers at the lowest limits of starting material. Without subsampling, the more limited distribution of coverage may fall well within the optimal range of sensitivity for the genotyping tools tested.

At 86% 2-field accuracy, even optimized composite predictions for 5' scRNA-seq data are unlikely to be useful for clinical genotyping. However, the primary novel application of HLA typing to single-cell data is its research use in assessing ASE. For instance, using bulk RNA-seq, Liu et al. demonstrated that colorectal cancer samples were associated with greater skewing of HLA ASE ratios. Similar to studies that demonstrated an association between loss-of-HLA-heterozygosity with multiple tumor types and failure of therapeutic checkpoint blockade (28), this skewing away from equal biallelic HLA expression could represent an immunological escape mechanism that functions to reduce the diversity of possible tumor peptide-antigens displayed by MHC molecules. Interestingly, in our study, we show that even inaccurate genotype predictions from arcasHLA, PHLAT, and OptiType can result in ASE ratios that correlate highly with those derived from ground truth genotypes. This suggests that these inaccurate predictions still identify alleles with sufficient sequence similarity to the true alleles that proportional read mapping between the homologous chromosomes is maintained, resulting in relatively accurate assessments of ASE.

However, molecular HLA genotyping still represents the gold standard for accuracy and is readily available through commercial laboratories and kits. Yet, it is not always feasible to perform these assays for single-cell experiments. As scRNA-seq methods are limited by cost and cell number, building robust data sets often requires integration of published data, for which prior HLA genotyping or remaining biological samples may not be available. Moreover, particularly in the case of clinical studies, limited sample material may preclude the use of multiple analytical platforms. Therefore, the ability to obtain HLA genotypes directly from scRNA-seq data represents a way to maximize the utility of these methods. We believe our study demonstrates that such direct genotyping can achieve sufficient accuracy for downstream applications unique to single-cell experiments, such as precise evaluation of ASE.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: (5 prime scRNA sequencing) https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-9357, (3 prime scRNA sequencing) https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120221, (Bulk RNA sequencing) https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120446.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

BS and PK conceived of the study, interpreted results, and wrote the manuscript. BS and HZ processed raw sequencing data and BS conducted subsequent data analysis. LD and CH collected and analyzed previously described scRNA-seq data and performed additional sample preparation for molecular HLA genotyping of the NIH-HBM cohort. JG and JH collected and analyzed previously described scRNA-seq data and provided HLA genotyping of the ISB-Swedish INCOV cohort. All authors contributed to the article and approved the submitted version.

Funding

PK is funded in part by the Bill and Melinda Gates Foundation (OPP1113682); the National Institute of Allergy and Infectious Diseases (NIAID) grants 1U19AI109662, U19AI057229, U19AI167903, and 5R01AI125197; Department of Defense contracts W81XWH-18-1-0253 and W81XWH-19-1-0235; and the Ralph & Marian Falk Medical Research Trust. BS is additionally funded by the National Heart, Lung, and Blood Institute (NHLBI) grant R38HL143615. JG and JH are funded by the National Cancer Institute grant R01CA264090.

Conflict of interest

PK is a shareholder and a consultant to Inflammatix, Inc. JH is founder and board member of Isoplexis and PACT Pharma. JG declared contracted research with Gilead, Lilly, and Regeneron.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2023.1146826/full#supplementary-material

References

1. Beck S, Trowsdale J. The human major histocompatibility complex: lessons from the DNA sequence. *Annu Rev Genom Hum Genet* (2000) 1(1):117–37. doi: 10.1146/annurev.genom.1.1.117

2. Wittig M, Anmarkrud JA, Kassens JC, Koch S, Forster M, Ellinghaus E, et al. Development of a high-resolution NGS-based HLA-typing and analysis pipeline. *Nucleic Acids Res* (2015) 43(11):e70–0. doi: 10.1093/nar/gkv184

3. Edgerly CH, Weimer ET. The past, present, and future of HLA typing in transplantation. *Methods Mol Biol* (2018) 1802:1–10. doi: 10.1007/978-1-4939-8546-3_1

4. Monneret G, Lepape A, Voirin N, Bohé J, Venet F, Debard A-L, et al. Persisting low monocyte human leukocyte antigen-DR expression predicts mortality in septic shock. *Intensive Care Med* (2006) 32(8):1175–83. doi: 10.1007/s00134-006-0204-8

5. Zheng H, Rao AM, Dermadi D, Toh J, Jones LM, Donato M, et al. Multi-cohort analysis of host immune response identifies conserved protective and detrimental modules associated with severity across viruses. *Immunity* (2021) 54(4):753–768.e5. doi: 10.1016/j.immuni.2021.03.002

6. Johansson T, Yohannes DA, Koskela S, Partanen J, Saavalainen P. HLA RNA sequencing with unique molecular identifiers reveals high allele-specific variability in mRNA expression. *Front Immunol* (2021) 12:629059. doi: 10.3389/fimmu.2021.629059

7. Gutierrez-Arcelus M, Baglaenko Y, Arora J, Hannes S, Luo Y, Amariuta T, et al. Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci. *Nat Genet* (2020) 52(3):247–53. doi: 10.1038/s41588-020-0579-4

8. Liu Z, Dong X, Li Y. A genome-wide study of allele-specific expression in colorectal cancer. *Front Genet* (2018) 9:570. doi: 10.3389/fgene.2018.00570

9. Filip I, Orenbuch R, Zhao J, Manji G, de Maturana EL, Malats N, et al. HLA allele-specific expression loss in tumors can shorten survival and hinder immunotherapy. *medRxiv* (2020) 2020.09.30.20204875. doi: 10.1101/2020.09.30.20204875

10. Jiang Y, Zhang NR, Li M. SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biol* (2017) 18(1):74. doi: 10.1186/s13059-017-1200-8

11. Borel C, Ferreira PG, Santoni F, Delaneau O, Fort A, Popadin KY, et al. Biased allelic expression in human primary fibroblast single cells. *Am J Hum Genet* (2015) 96 (1):70–80. doi: 10.1016/j.ajhg.2014.12.001

12. Deng Q, Ramskold D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* (2014) 343 (6167):193–6. doi: 10.1126/science.1245316

13. Darby CA, Stubbington MJT, Marks PJ, Martínez Barrio Á., Fiddes IT. ScHLAcount: allele-specific HLA expression from single-cell gene expression data. *Bioinformatics* (2020) 36(12):3905–6. doi: 10.1093/bioinformatics/btaa264

14. Bauer DC, Zadoorian A, Wilson LOW, Thorne NP. Evaluation of computational programs to predict HLA genotypes from genomic sequencing data. *Briefings Bioinf* (2018) 19(2):179–87. doi: 10.1093/bib/bbw097

15. Tian R, Zhu H, Pang Z, Tian Y, Liang C. Extraordinary diversity of HLA class I gene expression in single cells contribute to the plasticity and adaptability of human immune system. *Immunology* (2019). doi: 10.1101/725119

16. Su Y, Chen D, Yuan D, Lausted C, Choi J, Dai CL, et al. Multi-omics resolves a sharp disease-state shift between mild and moderate COVID-19. *Cell* (2020) 183 (6):1479–1495.e20. doi: 10.1016/j.cell.2020.10.037

17. Oetjen KA, Lindblad KE, Goswami M, Gui G, Dagur PK, Lai C, et al. Human bone marrow assessment by single-cell RNA sequencing, mass cytometry, and flow cytometry. *JCI Insight* (2018) 3(23):e124928. doi: 10.1172/jci.insight.124928

18. Hand DJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning* (2001) 45:171–86. doi: 10.1023/A:1010920819831

19. Orenbuch R, Filip I, Comito D, Shaman J, Pe'er I, Rabadan R. arcasHLA: high-resolution HLA typing from RNAseq. *Bioinformatics* (2020) 36(1):33–40. doi: 10.1093/bioinformatics/btz474

20. Warren RL, Choe G, Freeman DJ, Castellarin M, Munro S, Moore R, et al. Derivation of HLA types from shotgun sequence datasets. *Genome Med* (2012) 4 (12):95. doi: 10.1186/gm396

21. Bai Y, Ni M, Cooper B, Wei Y, Fury W. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. *BMC Genomics* (2014) 15(1):1–16. doi: 10.1186/1471-2164-15-325

22. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* (2014) 30 (23):3310–6. doi: 10.1093/bioinformatics/btu548

23. Dorak MT, Lawson T, Machulla HKG, Mills KI, Burnett AK. Increased heterozygosity for MHC class II lineages in newborn males. *Genes Immun* (2002) 3 (5):263–9. doi: 10.1038/sj.gene.6363862

24. Serin Harmanci A, Harmanci AO, Zhou X. CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. *Nat Commun* (2020) 11(1):89. doi: 10.1038/s41467-019-13779-x

25. Zhang Y, Song Y, Cao H, Mo X, Yang H, Wang J, et al. Typing and copy number determination for HLA-DRB3, -DRB4 and -DRB5 from next-generation sequencing data. *HLA* (2017) 89(3):150–7. doi: 10.1111/tan.12966

26. Major E, Rigó K, Hague T, Bérces A, Juhos S. HLA typing from 1000 genomes whole genome and whole exome illumina data. *PloS One* (2013) 8(11):e78410. doi: 10.1371/journal.pone.0078410

27. Lee M, Seo JH, Song S, Song IH, Kim SY, Kim YA, et al. A new human leukocyte antigen typing algorithm combined with currently available genotyping tools based on next-generation sequencing data and guidelines to select the most likely human leukocyte antigen genotype. *Front Immunol* (2021) 12:688183. doi: 10.3389/fimmu.2021.688183

28. Zhang X, Sjöblom T. Targeting loss of heterozygosity: a novel paradigm for cancer therapy. *Pharmaceuticals* (2021) 14(1):57. doi: 10.3390/ph14010057