A Rigorous Benchmarking of alignment-based HLA callers for RNA-seq data

Abstract

Accurate identification of human leukocyte antigen (HLA) alleles is essential for various clinical and research applications, such as transplant matching and drug sensitivities. Recent advances in RNA-seq technology have made it possible to impute HLA types from high throughput sequencing data, spurring the development of a large number of computational HLA typing tools. However, the relative performance of these tools is unknown, limiting the ability for clinical and biomedical research to make informed choices regarding which tools to use. Here, we rigorously compare the performance of 9 HLA callers on 652 RNA-seq samples across 6 datasets with molecularly defined gold standard. We find that OptiType has the highest accuracy at both low and high resolution with an accuracy above 99%, followed by arcasHLA and seq2HLA with accuracies above 96%. Despite OptiType's high accuracy, it is only capable of Class I predictions, thereby limiting its application to clinical procedures like transplantation requiring Class II predictions. Furthermore, our findings reveal significant variation in accuracy across each HLA locus, with HLA-A exhibiting the highest accuracy and HLA-DRB1 exhibiting the lowest accuracy. We also find that class II genes are generally more challenging to impute than class I genes, with most typing algorithms capable of making Class I predictions to >97% accuracy whereas the best Class II tool predicts with 94.2% accuracy. Moreover, we identify notable differences in the computational resources necessary to run each tool. We find that the most computationally expensive tools are OptiType and HLA-HD which require 10^5 and 10^2 times greater RAM and CPU, respectively, than the least computationally expensive tools, seq2HLA and RNA2HLA. Furthermore, all tools have decreased accuracy for African samples with respect to European samples at four digit resolution. We conclude that RNA-Seq HLA callers are capable of returning high-quality results, but the tools that offer a good balance between accuracy, consistency, and computational expensiveness are yet to be developed.

Introduction

The Human Leukocyte Antigen (HLA) complex, otherwise known as the human Major Histocompatibility Complex (MHC), is a 3.6 Mb segment of the human genome which encodes for cell surface glycoproteins involved in cell-mediated immunity and self/nonself recognition. There are two classes of the HLA system typically implicated in immunity: Class I, which includes the loci HLA-A, HLA-B, and HLA-C and Class II,

which includes the loci HLA-DRB1 and HLA-DQB1 (Fig 1a). These five loci comprise the Classical HLA loci and are the most well documented in the entire region.

The HLA region is the most polymorphic region of the human genome with over 25,000 HLA alleles currently available in the Immuno Polymorphism Database ImMunoGeneTics (IPD-IMGT) HLA reference database[1,2]. The high polymorphism within the locus has the biological function of allowing each individual to produce a dramatically distinct immune response. Therefore, accurate identification of HLA haplotypes is extremely relevant to immunology and the biomedical sciences. For instance, HLA typing is widely used in vaccinology[3] and transplant medicine including organ and hematopoietic stem cell transplantations[4]. Furthermore, certain HLA variants are associated with pathogenesis of both infectious and autoimmune diseases[5,6], adverse reactions to certain drugs[7], and cancer development[8]. Recent studies have also found that some HLA variations are associated with susceptibility, progression, and severity of SARS-CoV-2 infection[9]. US standards require an 8/8 match for all alleles within the HLA-A, HLA-B, HLA-C, and HLA-DQB1 for transplant donor selection, and most European standards require a more stringent 10/10 match for HLA-A, HLA-B, HLA-C, HLA-DRB1, and HLA-DQB1 alleles. Whether a donor's HLA type constitutes an "HLA match" depends on the resolution of interest, with the most frequently used resolutions being either two-digit resolution (low resolution) or four-digit resolution (high resolution). High resolution typing is recommended by the National Marrow Donors Program and hence is of considerable clinical significance[10].

Established HLA typing methods involve laboratory-based techniques such as amplification and quantification of HLA genes via sequence-specific oligonucleotide probes (SSOP) or sequence-specific primers (SSP), or via PCR followed by Sanger sequencing and comparison against an HLA reference database (SBT). SBT is the current gold standard method of typing, both it is still cost prohibitive and laborious to be effective for large-scale clinical and research applications[9]. More recently, developments in accessibility and throughput of NGS technology have spurred the development of HLA callers that impute HLA types from short read data. They use algorithms that can be divided into two general categories, alignment-based methods and assembly methods. Alignment-based methods align reads to an HLA reference containing a set of known allele sequences, then use that information to either build a probabilistic or optimization model to impute the HLA type. The less common assembly method links fragments into contigs which are then aligned to the reference genome to make the imputation of HLA type.

Despite the plethora of recently developed HLA callers, there is a lack of comprehensive and systematic benchmarking of RNA-seq-based HLA callers using large-scale and realist gold standards, limiting the ability for clinical and biomedical research to make informed choices regarding which tools to use. Furthermore, most existing benchmarking papers focus solely on whole-genome (WGS) or whole-exome

(WES) data[11–14], with the only existing RNA-seq study published seven years ago[15]. Since the previous benchmarking study, seven new HLA callers have been developed, and the accuracy of HLA callers on varying read length and ancestry is yet to be evaluated. To address these limitations, we evaluated the performance of 9 HLA callers across 652 RNA-seq samples with available gold standard HLA alleles. The tools we included in this analysis are HLAminer[16], seq2HLA[17], HLAforest[18], PHLAT[19], OptiType[20], HLA-VBseq[21], HLA-HD[22], arcasHLA[23], and RNA2HLA[24], of which 4 are novel and not considered in the previous benchmarking study. All tools are capable of making predictions at all Class I and Class II classical HLA loci, with the exception of OptiType[20], which only makes Class I predictions. In each case, we produced evaluation metrics of accuracy that is the percentage of correctly predicted alleles to determine the best performing tool. We also conducted comparisons of the performance of each HLA caller across the various HLA genes to detect the genes that are consistently miscalled by HLA calling methods. Furthermore, we examined the impact of sequencing parameters on the performance of each HLA caller and thoroughly assessed the computational resources required for each software. Additionally, we compared the performance of HLA methods across different ancestry groups and identified tools and trends in prediction accuracy and quality across samples of European and African ancestry. In summary, we conducted an extensive evaluation of all publicly accessible HLA callers, revealing valuable insights for researchers and clinicians to make an informed decision selecting the most appropriate RNA-Seq based HLA caller, offering a balance of consistency, performance, and scalability for their clinical and scientific needs.

Methods

RNA-Seq Data

We retrieved data from 6 publicly available RNA-seq datasets (ERP001942[25], SRP298704[26], ERP000101[27], SRP099568[28], SRP198497[29], SRP096329[30]) to assess the performance of the tools, as detailed in Table 1. From the datasets we selected only the samples containing ground truth HLA types to be used for this study, amounting to a total of 652 samples collected from lymphoblastoid cell lines (LCLs) (n=552) and peripheral blood mononuclear cells (PMBCs) (n=100). Altogether, these datasets represent a diverse set of samples spanning 5 1000-genomes project populations (Northern Europeans in Utah, Finnish, British in England and Scotland, Tuscan in Italy, Yoruba), 4 sequencing platforms (Illumina HiSeq 2000, Illumina HiSeq 2500, Illumina Genome Analyzer II, Illumina NextSeq 550), read counts between 24.5 million and 337.8 million reads, and read lengths ranging from 36bp to 126bp.

We used the SRA-Toolkit v2.11.0 to download all samples as fastq files from the NCBI SRA archive, using the prefetch and fasterq-dump commands:

1. prefetch --option-file accession_list.txt, to download all samples in SRA format, where 'accession_list.txt' contains a list of the accession numbers of all samples separated by newlines (\n)
2. fasterq-dump --split-files sample.sra" on each sample ('sample.sra'), to convert from sra to FASTQ format

We used STAR aligner v2.7.0e to align reads to the GRCh38 reference using the command ""STAR --runThreadN 8 --genomeDir /path/to/genomeIndicesFile --readFilesIn path/to/sample1.fastq path/to/sample2.fastq --outSAMtype BAM SortedByCoordinate --outFileNamePrefix /path/to/outDir/prefix". The '--genomeDir' argument specifies the location of the file containing the genome indices, and '--outSAMtype BAM SortedByCoordinate' indicates the output files to be sorted BAM files. From the aligned BAM files, we extracted the chromosome 6 regions using Samtools v1.19 commands, with default parameters unless otherwise specified. We first used the Samtools index function to produce indexed BAI files from the BAM-format samples for fast random access. Then we then invoked Samtools view with default parameters to extract the reads aligning to chromosome 6 into output files. Finally, we invoked 'samtools view' with -b argument on each of the output files to convert them from SAM to BAM format. Finally, Bedtools v2.27.1 bamtofastq command was used to convert the BAM files to FASTQ files. After these steps we produced a pair of FASTQ samples for each of the 652 samples. These data served as the final inputs for all HLA callers.

Gold Standard HLA types

The gold standard HLA types of samples were determined via laboratory based methods summarized in Table 1. For dataset 1, HLA types were found via PCR amplification of the HLA-associated genes followed by Sanger sequencing-based typing of the exons (SBT method). Dataset 2 and 4's HLA types were found using PCR-based sequence-specific primers (SSP method) which amplifies the specific regions of DNA using primers that are designed specifically to bind to and amplify particular alleles of HLA genes. Dataset 3's HLA types were found using PCR-based sequence-specific oligonucleotide probes (SSOP method) which uses PCR to amplify a specific region of the HLA gene, and then uses sequence-specific oligonucleotide probes that will detect and differentiate specific alleles or variants in that region based on their complementary binding. Dataset 5 and 6 were created from cells from the HLA

Class I-deficient B721.221 cell line transduced with one or two Class I HLA allele-expressing cDNA vectors.

We formatted the gold standard HLA types for each dataset into a *standard format csv* where rows represent the samples and columns represent the loci(s) of interest (HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQB1) for which gold standard allele data are available. The csv files have been made available at https://github.com/ydottie/hlaproject.

HLA definition and nomenclature

To specify each HLA allele type, we used the HLA nomenclature system defined by the World Health Organization Committee for Factors of the HLA System in 2010 (Fig 1b). This system uses a four-part naming system: the HLA Prefix, the gene, the field groups, and the suffix letter as shown in Figure 1b. Every allele begins with the HLA Prefix, "HLA." Following the HLA prefix is the gene identifier, which denotes the gene or locus of the allele. Then lies the field groups which are divided into 4 fields of increasing resolution, separated by colons. Field 1 represents the allele group, or specifically what serological antigen type is present for this particular allele. Field 2, which is typically 2 or 3 digits, represents the specific HLA protein. Field 3 represents allele variants with synonymous or intronic differences. Field 4 designates whether the allele is a single or multiple nucleotide polymorphism in the non-coding region of the gene. Finally, the suffix letter denotes differences in expression of the allele as described in Table S3. Two-digit resolution is denoted by the HLA prefix, the locus, and the first field groups (i.e., HLA-A*02). Four-digit resolution is the clinically relevant resolution and is denoted by the HLA prefix, the locus, and the first 2 field groups (i.e., HLA-A*02:10). Two-digit resolution may also be referred to as low resolution, and four-digit resolution may be referred to as high resolution or clinically relevant resolution. For the purposes of this benchmarking study, we assessed performance at the two-digit and four-digit resolutions.

HLA caller choice and installation

We benchmarked all existing alignment-based HLA callers capable of imputing HLA types from RNA-seq data. With these prerequisites, we identified 12 tools to be included in this study, namely HLAminer[16], seq2HLA[17], HLAforest[18], PHLAT[19], OptiType[20], HLA-VBseq[21], HLA-HD[22], arcasHLA[23], RNA2HLA[24], HLApers[a31], HLAProfiler[32], and HISAT-genotype[18] (Table 2). We downloaded each tool from their respective repository onto the USC CARC Discovery Cluster, closely following all available

documentation and recommendations for each tool. If testing suites were provided, we ran the tools on the tests and manually verified that the tools successfully completed the tests. Furthermore, we emailed the developers of each tool to verify our process of installation and for recommendations on best practices for installation and running. During the installation process we faced unresolvable issues in 3 tools (HLApers[a31], HLAProfiler[32], and HISAT-genotype[18]), and hence excluded these tools from the study. The errors can be found in Table S1.

Run HLA profiling methods on RNA-Seq samples

After installation, we ran each HLA caller by submitting bash scripted jobs to the SLURM job scheduler on the USC CARC Discovery cluster. The commands used to run each tool are summarized in Table 3. Although some tools are capable of imputation at greater resolutions than 4 digits, we considered only the 2-digit (low) and 4-digit (high) resolution due to the limitation that our datasets contain only gold standard types up to 4-digit resolution. We parsed each tool's custom output files into the *standard csv format* which parallels the format of the gold standard datasets, with each row representing a different sample and each column representing a different loci (see "Gold Standard HLA types" subsection). We considered lack of prediction as an incorrect prediction.

Evaluating HLA predictions for accuracy

We evaluated the accuracy of each HLA tool in comparison to the gold standard using the following equation:

$$\text{□□□□□□□□□} = \frac{\text{□□□□□□ □□ □□□□□□□ □□□□□□□□□ □□□□□□□□□}}{\text{□□□□□ □□□□□□ □□ □□□□□□□ □□ □□□□ □□□□□□□□}} \times 100\%$$

We assessed accuracy at 2 digit and 4 digit resolutions, with 2-digit resolution considering only the section of the HLA allele up to the first field and 4-digit resolution considering the section of the allele up to the second field (Figure 1). The numerator "number of alleles correctly predicted" is determined via a locus-by-locus comparison of the prediction with the gold standard. Because each locus contains 2 independent haplotypes, our algorithm assesses the 2 and 4 digit accuracy of both a parallel and cross-wise comparison between the 2 alleles of the gold standard and the prediction. Between the parallel and

cross-wise comparison, we selected the option that maximizes both the 2 and 4 digit accuracies. We treated no-calls as missed-calls, with the exception of OptiType, in which all Class II loci were ignored. Furthermore, we stratified accuracy by class, locus and allele to examine for variations. We also examined the presence of specific allele haplotypes that are commonly mispredicted as a specific allele, followed by pairwise alignment of the aligned sequences with the hypothesis that these “frequent pairs” are the result of high sequence similarity.

Computational modification of RNA-Seq sample properties

We are currently working on using Seqtk v1.3 to produce simulated RNA-seq samples with truncated read lengths. We will perform this computational modification for 10 samples from each of datasets 1-4, and all samples of datasets 5-6 because these datasets contain less than 10 samples, for a total of 52 samples. So far, we have produced simulated samples for 12 samples only, but plan to complete the rest of the samples in the future. We first converted each sample from FASTQ to FASTA format using the command seqtk seq -a in.fq > out.fa. For each sample, we then iteratively invoked the seqtk trimfq -b 5 in.fa > out.fa, such that each iteration truncates 5 base pairs from the right of every read, stopping when reads become shorter than 36 bp. Each original sample is hence used as a template for multiple simulated samples, each with read lengths ranging from 36bp to the original read length, with a step size of 5bp. The final files were converted from fa back to fq format using the seqtk seq -a in.fa > out.fq command. All simulated samples were ran on the HLA caller tools with default and optimal parameters for accuracy analysis.

For analysis of coverage effect on accuracy (in progress), we will computationally modify the coverage via randomly selecting a subset of reads from our original RNA-seq samples to create new, lower-coverage samples. We will quantify coverage as a percentage of the original read count, subsampling from a coverage of 20% of the original read count to 100% of the read count with a step size of 5% of the read count for each sample. At each percentage level, we will use the seqtk subsample function to produce simulated lower coverage samples, ensuring that we use the same seed for both files from each paired-end sample. All samples will then be run on the HLA caller tools for accuracy analysis.

Assessing the effect of ancestral diversity on prediction quality

We used all samples in Dataset 1 (n=490) in order to evaluate differences in performances between European and African ancestry groups. For each caller, we determined the accuracy at each HLA locus separately for the European subset (n=423) and the African subset (n=67) of dataset 1, both at two and

four-digit resolutions. After calculating accuracy, we used a chi-square test of independence to determine the statistical significance of the differences in accuracy across both ancestral groups. We also isolated the alleles to investigate the misclassification rate for each allele when looking at European versus African samples.

Assessing computational resources required to run HLA callers

For each tool, CPU time and RAM metrics were recorded to determine the computational performance. Each tool was run with 10 samples, 10 times each on the USC Center for Advanced Research Computing (USC CARC) with 1 node, 1 task and 4 CPUs per task. Metrics were collected using an inbuilt command from the cluster's Slurm Workload Manager, called upon by adding the following line of code, '#SBATCH --mail-user=<e-mail address>' to the sbatch script header, would email the user a full record of the execution of the script, containing information on the CPU time and the RAM usage. We averaged the 10 runs for each sample to produce 10 final values for each tool's CPU and RAM, in order to minimize biases from resource availability in the shared cluster.

Results

ArcasHLA has the greatest overall accuracy across Class I and Class II loci

At two digit resolution, the accuracy varies from 14.9% to 99.4% for Class I predictions, and 7.4% to 98.4% accuracy for Class II predictions. At four digit resolution, the accuracy is lower, with class I accuracy ranging from 6.4% to 98.3% accuracy, while Class II predictions ranging from 0% to 93.4% accuracy (Fig 2a).

At both 2 digit and 4 digit resolutions, for Class I alleles only, OptiType has the best performance in terms of accuracy for Class I alleles with scores of 99.4% and 98.3% respectively. For Class II alleles, ArcasHLA has the best accuracy performance with scores of 98.4% and 93.4%. Notably, ArcasHLA is able to make both Class I and Class II predictions with consistently high accuracy, whereas OptiType is only able to make class I predictions. HLAMiner and HLAForest both struggle to make accurate predictions and have the worst performance for both Class I and Class II. These tools have a respective accuracy of 18.3% and 56.4% at two digit resolution, and 6.4% and 0% at four digit accuracy. Exact accuracy values of each tool from 0 (fully inaccurate) to 1 (fully accurate) are reported in Table S2a-c.

The HLA-B and HLA-DRB1 loci are most commonly mispredicted

We assessed each tool's performance on each locus (A, B, C, DRB1, DQB1) to determine differences in accuracy across loci (Fig 2b). We find that Class II underperform compared to Class I, with the least accurate loci being HLA-DQB1 (70.1%) and HLA-DRB1 (75.5%) at four digits of resolution. Of the Class I loci, HLA-B performs the worst, with a 75.6% accuracy, whereas HLA-A and HLA-C both perform fairly well with accuracies of 89.3% and 88.8% respectively at four digits of significance.

We additionally evaluated the percentages at which each individual allele in the gold standard datasets is mispredicted (Fig 2c-d). Our results show that the Class II loci, particularly HLA-DQB1, is mispredicted at the highest ratio in comparison with alleles from other loci. Notably, we find that HLA-VBseq and HLAMiner are responsible for a majority of the Class II mispredictions, as removing these two tools from analysis decreased the significance of the difference in accuracies between Class I and Class II genes.

Computational expense varies greatly between HLA callers

We measured the CPU time and RAM usage of each caller for two samples per dataset to assess computational expensiveness (Fig 3a-b). We find great variation in the computational resources necessary to run the tools, with RAM ranging from the order of 100 kB to 100 GB, and CPU time ranging from approximately 10 minutes to 3 hours. While there is often a trade off between accuracy and computational resources required by each tool, we observe that RNA2HLA yielded the best balance between high accuracy (93.3%) at low computational expense (average 0.36 GB RAM amd 6.9 min CPU time), making it suitable for HLA typing with large scale studies. In contrast, the best accuracy tool, OptiType, requires an average of 14.2 GB RAM and 45.3 min CPU time per sample.

Read lengths do not exhibit large effect on prediction quality

From our preliminary analysis of 12 samples across Datasets 2 and 3, we find that increasing read length is associated with an increase in accuracy (Fig 3c) for all tools except HLAMiner. A read length of 36 has an average accuracy of 63%, a read length of 51 has an average accuracy of 67% and a read length of 76 has an average accuracy of 77%. In the future we will perform more in-depth analysis with 52 samples and smaller step sizes between the read lengths.

All HLA callers have higher accuracy on European than African samples

We used all samples from Dataset 1 (n=490) to evaluate differences in performance between Europe and African ancestry groups. For each caller, we determined the accuracy at each HLA locus separately for European (n=423) and African (n=67) subsets. We observe the greatest disparity in accuracy in HLAforest[6] (p=0.019) and PHLAT[7] (p=0.017). Furthermore, we find that differences in accuracy are more pronounced at four-digit resolution than they are at two-digit resolution (Fig 4a). At four digits of resolution, every single tool performed better on European than African ancestry groups. Although the disparity in some tools is not pronounced enough to be statistically significant, a binomial test demonstrates that the overall lower accuracy in African ancestry groups across all tools is significant (p=0.0039).

When we stratify the accuracy of each ancestry group by locus (Fig 4b), we find that there is some variability in whether European or African subsets have greater accuracy for Class I alleles, but a vast majority of tools seem to be better at making accurate predictions for the European group for Class II alleles. The greatest discrepancy in accuracy out of the Class II genes is HLA-DQB1 in which the African subset accuracy is 2.7% lower than the European subset. Less extreme but still significant, between the Class I genes the greatest discrepancies in accuracy are observed in HLA-B where the average African subset accuracy is 1.3% lower than the European subset. Furthermore, for HLA-A the difference in average accuracy is 1.3% and for HLA-DRB1 it is 1.0%. HLA-C is the only loci where the African subset outperformed the European subset on accuracy, with a difference of 0.3 percentage points. We then isolated alleles and plotted the Europe against Africa misclassification rate for each allele, finding a slight skew towards higher misclassification rates for the Africa samples. We also notice a significantly greater misclassification rate for Class II alleles for both Europe and Africa alleles, which disappears with the removal of the tools HLA-VBseq and HLAMiner, again verifying that these two tools are insufficient for making accurate Class II imputations.

Discussion

Our rigorous assessment of HLA callers highlights the advantages and limitations of computational HLA typing algorithms across several Class I and Class II loci and various sequencing parameters. Altogether, this study offers crucial, up-to-date information for researchers regarding appropriate choice of tool for an

HLA typing from RNA-seq data. Furthermore, our study demonstrates the potential for RNA-seq based HLA typing tools to outperform traditional methods as a scalable and cost-effective method for imputing HLA typing. Crucially, our comparison of the performance of HLA callers across different ancestry groups highlights the pressing need to develop improved HLA typing algorithms and databases that can accurately account for the distinct genetic variations present in non-European populations.

In comparison to the previous RNA-seq benchmarking study[15], we found highly consistent results with the same tools performing well in their study, OptiType and arcasHLA performing similarly well and HLA-VBseq struggling to make accurate imputations. We also found that more recently developed tools perform similarly well to prior tools, without many significant improvements to accuracy in the past few years. For instance, Optitype remains the most accurate tool for Class I genotypes. All tools performed much better on Class I than on Class II genes, with Class II accuracy being much lower compared to the current gold standard HLA typing methods.

Our study is the first to evaluate the ability for HLA callers to impute HLA types from specifically RNA-seq data. We find that all tools, except HLA-VBseq and HLAMiner, are capable of making Class I predictions to 4 digit (clinical) resolution to 82.6% - 98.3% accuracy when compared with the gold standard method. We find that there are few tools that are able to make consistently accurate predictions across all loci and classes, and we believe that arcasHLA is the overall best tool as it performs with a consistently high 93.4% accuracy on Class I and Class II. In addition to arcasHLA, we also recommend RNA2HLA as a tool for high quality predictions at relatively low computational expense. Our study is also the first to explore the effect of read length and ancestry on the imputation quality of the callers. We found that read length does not have a substantial effect on accuracy. The tools are more accurate at making predictions for individuals of European ancestry than those of African ancestry.

In the future, we plan to complete the read length analysis and data collection for the CPU and RAM usage of the tools as we are currently missing some data. We also plan to determine the effect of coverage on the accuracy of the HLA typing algorithms. We also plan to determine the concordance in predictions across various samples, to evaluate whether the algorithms have the tendency to predict certain alleles. Once identifying patterns of misclassifications across tools, we also hope to analyze the sequence similarity of the allele with the misclassification to establish a hypothesis for why these errors occur. One final step we hope to complete is running tools with various custom parameters and arguments to determine the best set of parameters for each tool, rather than simply the default parameters like we are currently doing.

We envision that future HLA callers will accurately impute HLA genotypes for both Class I and Class II genes to 4-digit resolution, thereby offering clinicians and researchers a well performing and scalable

option for HLA typing with WGS data. Current methods show an inability to make accurate Class II predictions, and discrepancies in prediction quality across different ancestral groups, and addressing these challenges will make HLA callers a viable choice for HLA typing studies. One potential limitation is that our CPU/RAM was collected from jobs submitted to a remote high performance computing cluster, and hence may not be completely controlled because it might be dependent on resource availability of the cluster at the time we ran each sample. We mitigated this limitation by running each sample 10 times and taking the average to be used for our analysis. We hope this study offers crucial information for researchers regarding appropriate choices of methods for conducting HLA typing with RNA-seq data.

Data Availability

The RNA-seq datasets used for this analysis are available under the following accession numbers on the NCBI SRA archive: ERP001942[25], SRP298704[26], ERP000101[27], SRP099568[28], SRP198497[29], SRP096329[30]. All data required to produce the figures and analysis performed in this paper are freely available at https://github.com/ydottie/hlaproject.

Code Availability

All code required to produce the figures and analysis performed in this paper is under the Massachusetts Institute of Technology (MIT) license and is available at https://github.com/ydottie/hlaproject.

Acknowledgments

Thank you to Dr. Serghei Mangul (USC School of Pharmacy) for mentoring this project. Thank you to Ram Ayyala (PhD student in USC QCB department) for working jointly with me on the data collection for this project.

References

1. Robinson, J. *et al.* The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* **43**, D423–D431 (2015).

2. Robinson, J. *et al.* IPD-IMGT/HLA Database. *Nucleic Acids Res.* gkz950 (2019) doi:10.1093/nar/gkz950.
3. Haralambieva, I. H. *et al.* The genetic basis for interindividual immune response variation to measles vaccine: new understanding and new vaccine approaches. *Expert Rev. Vaccines* **12**, 57–70 (2013).
4. Montgomery, R. A., Tatapudi, V. S., Leffell, M. S. & Zachary, A. A. HLA in transplantation. *Nat. Rev. Nephrol.* **14**, 558–570 (2018).
5. De Santis, D. *et al.* 16 th IHIW□: Review of HLA typing by NGS. *Int. J. Immunogenet.* **40**, 72–76 (2013).
6. Thorsby, E. & Lie, B. A. HLA associated genetic predisposition to autoimmune diseases: Genes involved and possible mechanisms. *Transpl. Immunol.* **14**, 175–182 (2005).
7. Fan, W.-L. *et al.* HLA Association with Drug-Induced Adverse Reactions. *J. Immunol. Res.* **2017**, 1–10 (2017).
8. Naranbhai, V. *et al.* HLA-A*03 and response to immune checkpoint blockade in cancer: an epidemiological biomarker study. *Lancet Oncol.* **23**, 172–184 (2022).
9. Migliorini, F. *et al.* Association between HLA genotypes and COVID-19 susceptibility, severity and progression: a comprehensive review of the literature. *Eur. J. Med. Res.* **26**, 84 (2021).
10. Horowitz, M. M. High-resolution typing for unrelated donor transplantation: How far do we go? *Best Pract. Res. Clin. Haematol.* **22**, 537–541 (2009).
11. Liu, P. *et al.* Benchmarking the Human Leukocyte Antigen Typing Performance of Three Assays and Seven Next-Generation Sequencing-Based Algorithms. *Front. Immunol.* **12**, 652258 (2021).
12. Thuesen, N. H., Klausen, M. S., Gopalakrishnan, S., Trolle, T. & Renaud, G. Benchmarking freely available HLA typing algorithms across varying genes, coverages and typing resolutions. *Front. Immunol.* **13**, 987655 (2022).
13. Xin, H. *et al.* Benchmarking of 5 algorithms for high-resolution genotyping of human leukocyte antigen class I genes from blood and tissue samples. *Ann. Transl. Med.* **10**, 633–633 (2022).
14. Danish Pan Genome Consortium, Matey-Hernandez, M. L., Brunak, S. & Izarzugaza, J. M. G. Benchmarking the HLA typing performance of Polysolver and Optitype in 50 Danish parental trios. *BMC Bioinformatics* **19**, 239 (2018).
15. Bauer, D. C., Zadoorian, A., Wilson, L. O. W., Melbourne Genomics Health Alliance & Thorne, N. P. Evaluation of computational programs to predict HLA genotypes from

genomic sequencing data. *Brief. Bioinform.* bbw097 (2016) doi:10.1093/bib/bbw097.

16. Warren, R. L. *et al.* Derivation of HLA types from shotgun sequence datasets. *Genome Med.* **4**, 95 (2012).
17. Boegel, S. *et al.* HLA typing from RNA-Seq sequence reads. *Genome Med.* **4**, 102 (2012).
18. Kim, H. J. & Pourmand, N. HLA Haplotyping from RNA-seq Data Using Hierarchical Read Weighting. *PLoS ONE* **8**, e67885 (2013).
19. 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* **15**, 325 (2014).
20. Szolek, A. *et al.* OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **30**, 3310–3316 (2014).
21. Nariai, N. *et al.* HLA-VBSeq: accurate HLA typing at full resolution from whole-genome sequencing data. *BMC Genomics* **16**, S7 (2015).
22. Kawaguchi, S., Higasa, K., Shimizu, M., Yamada, R. & Matsuda, F. HLA□HD: An accurate HLA typing algorithm for next□generation sequencing data. *Hum. Mutat.* **38**, 788–797 (2017).
23. Orenbuch, R. *et al.* arcasHLA: high-resolution HLA typing from RNAseq. *Bioinformatics* **36**, 33–40 (2020).
24. Chelysheva, I., Pollard, A. J. & O'Connor, D. RNA2HLA: HLA-based quality control of RNA-seq datasets. *Brief. Bioinform.* **22**, bbab055 (2021).
25. Abi-Rached, L. *et al.* Immune diversity sheds light on missing variation in worldwide genetic diversity panels. *PLOS ONE* **13**, e0206512 (2018).
26. Houtman, M. *et al.* Haplotype-Specific Expression Analysis of MHC Class II Genes in Healthy Individuals and Rheumatoid Arthritis Patients. *Front. Immunol.* **12**, 707217 (2021).
27. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
28. Chemin, K. *et al.* EOMES-positive CD4+ T cells are increased in PTPN22 (1858T) risk allele carriers. *Eur. J. Immunol.* **48**, 655–669 (2018).
29. Sarkizova, S. *et al.* A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* **38**, 199–209 (2020).
30. Abelin, J. G. *et al.* Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity* **46**, 315–326 (2017).
31. Aguiar, V. R. C., Masotti, C., Camargo, A. A. & Meyer, D. HLApers: HLA Typing and Quantification of Expression with Personalized Index. in *Bioinformatics for Cancer Immunotherapy* (ed. Boegel, S.) vol. 2120 101–112 (Springer US, 2020).

32. Buchkovich, M. L. *et al.* HLAProfiler utilizes k-mer profiles to improve HLA calling accuracy for rare and common alleles in RNA-seq data. *Genome Med.* **9**, 86 (2017).

Figures

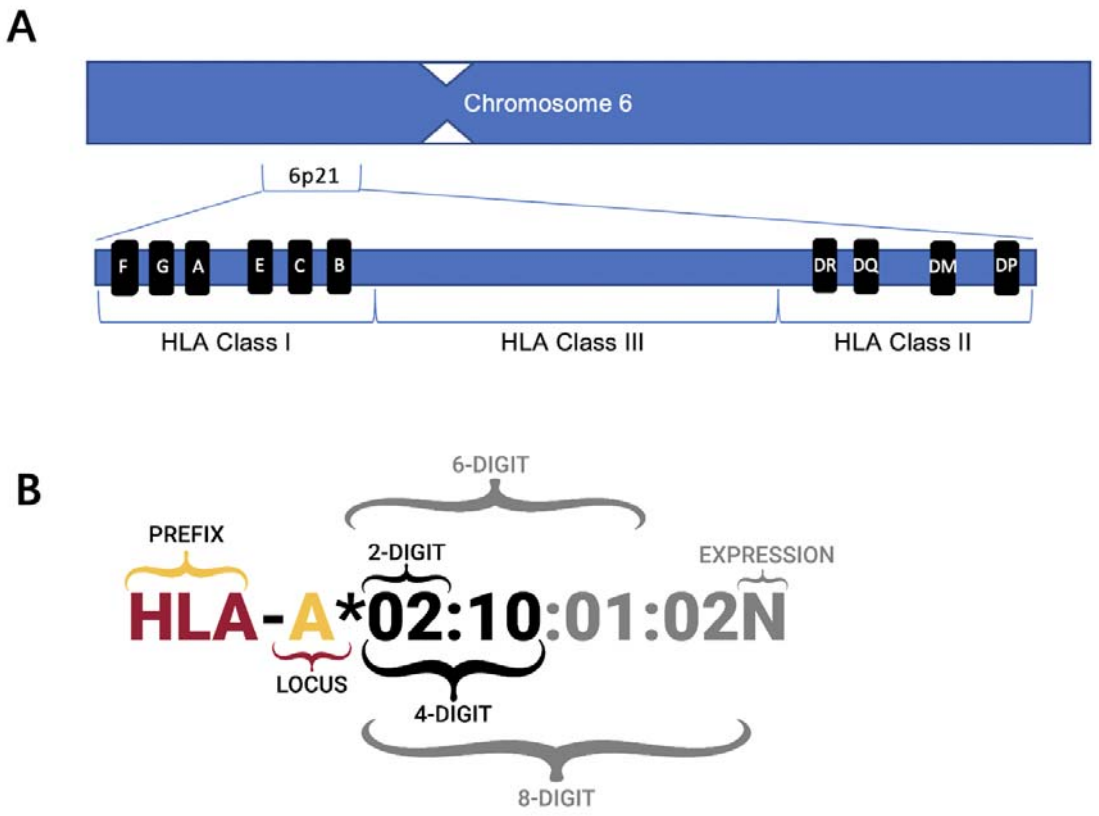

Figure 1: A: Schematic representation of HLA locus, with classical genes in Class I and Class II labeled. B: The nomenclature for specification of HLA allele types. 2-digit resolution indicates the allele group and is the lowest resolution for typing. 4-digit resolution indicates the specific protein produced and is considered high resolution and the clinically significant resolution. Higher resolution levels indicate synonymous DNA substitution in the coding region, changes in the non-coding areas of the genome, and expression level.

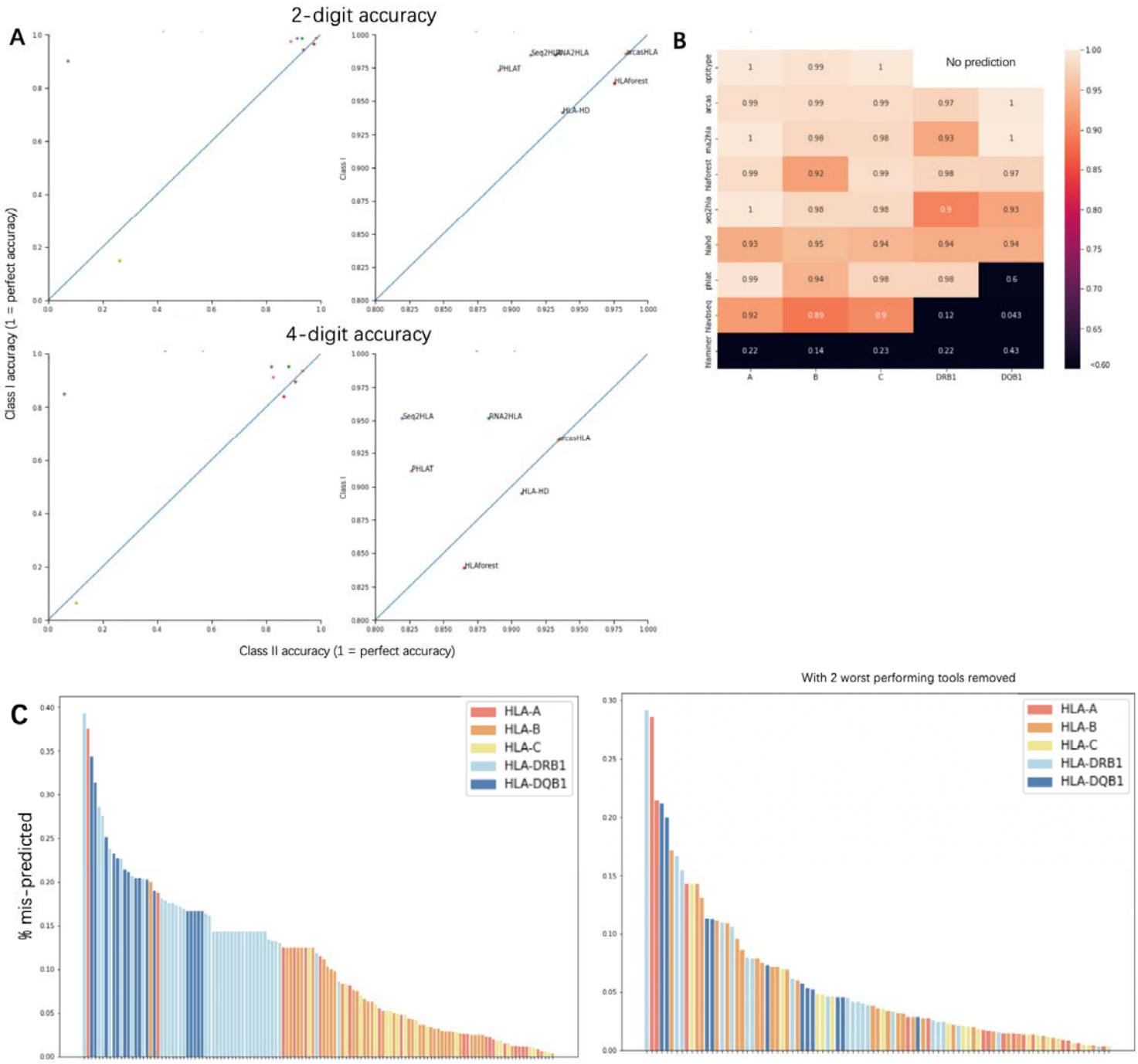

Figure 2: A: Accuracy of each tool to 2 and 4 digit accuracy, plotted with Class I accuracy against Class II accuracy. B: Accuracy of each tool for each locus, with loci on the horizontal axis and tool on the vertical axis. C: all alleles with height representing the % of mispredictions. Class I alleles are in orange whereas Class II alleles are in blue. On the right is the same plot except with the 2 lowest accuracy tools (HLA-VBseq and HLAMiner) removed.

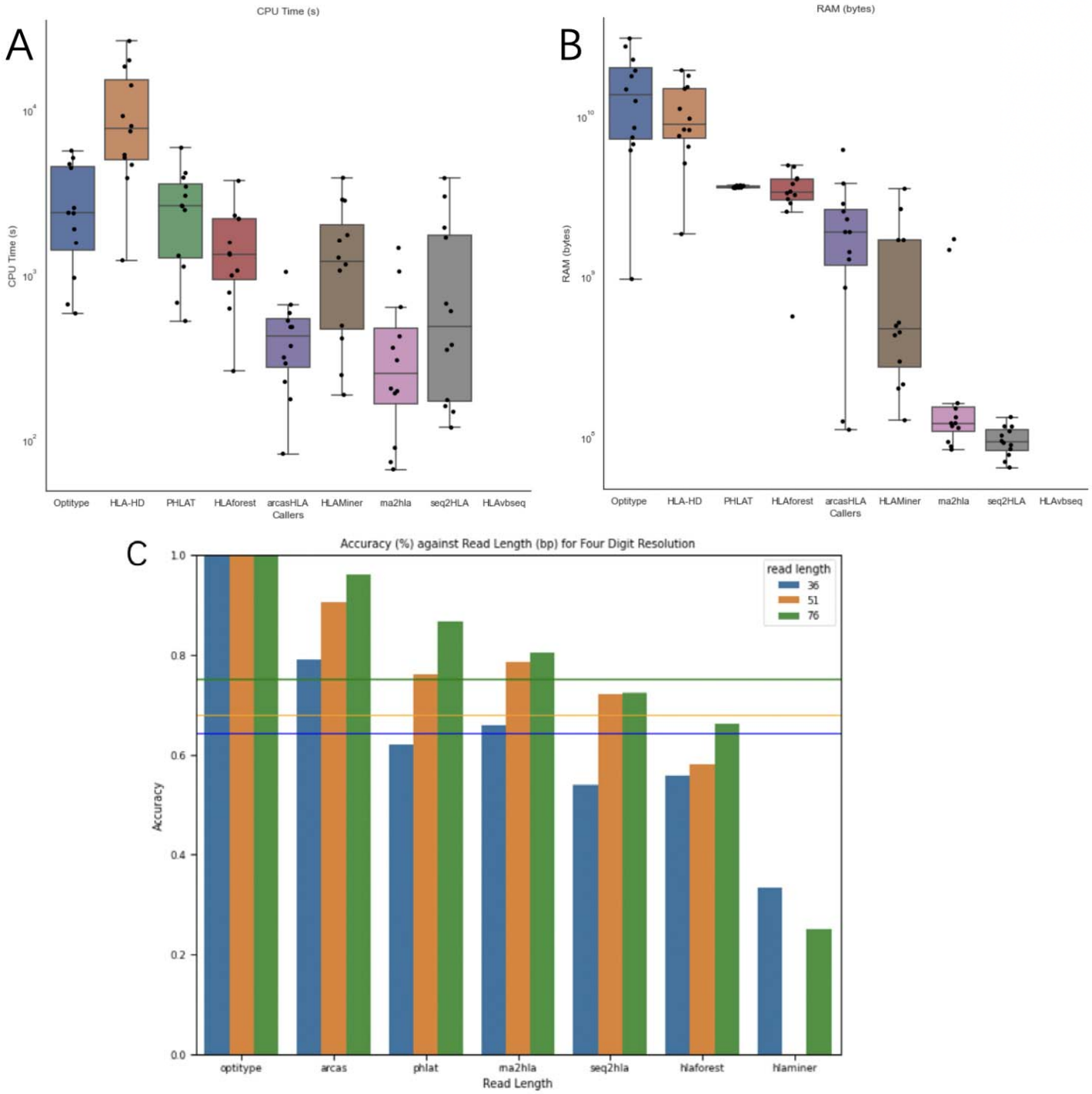

Figure 3: A: CPU time of each caller. B: RAM usage of each caller. C: Accuracy at read lengths of 36, 51, and 76 for each tool. Some data is missing from tools we did not collect data for yet.

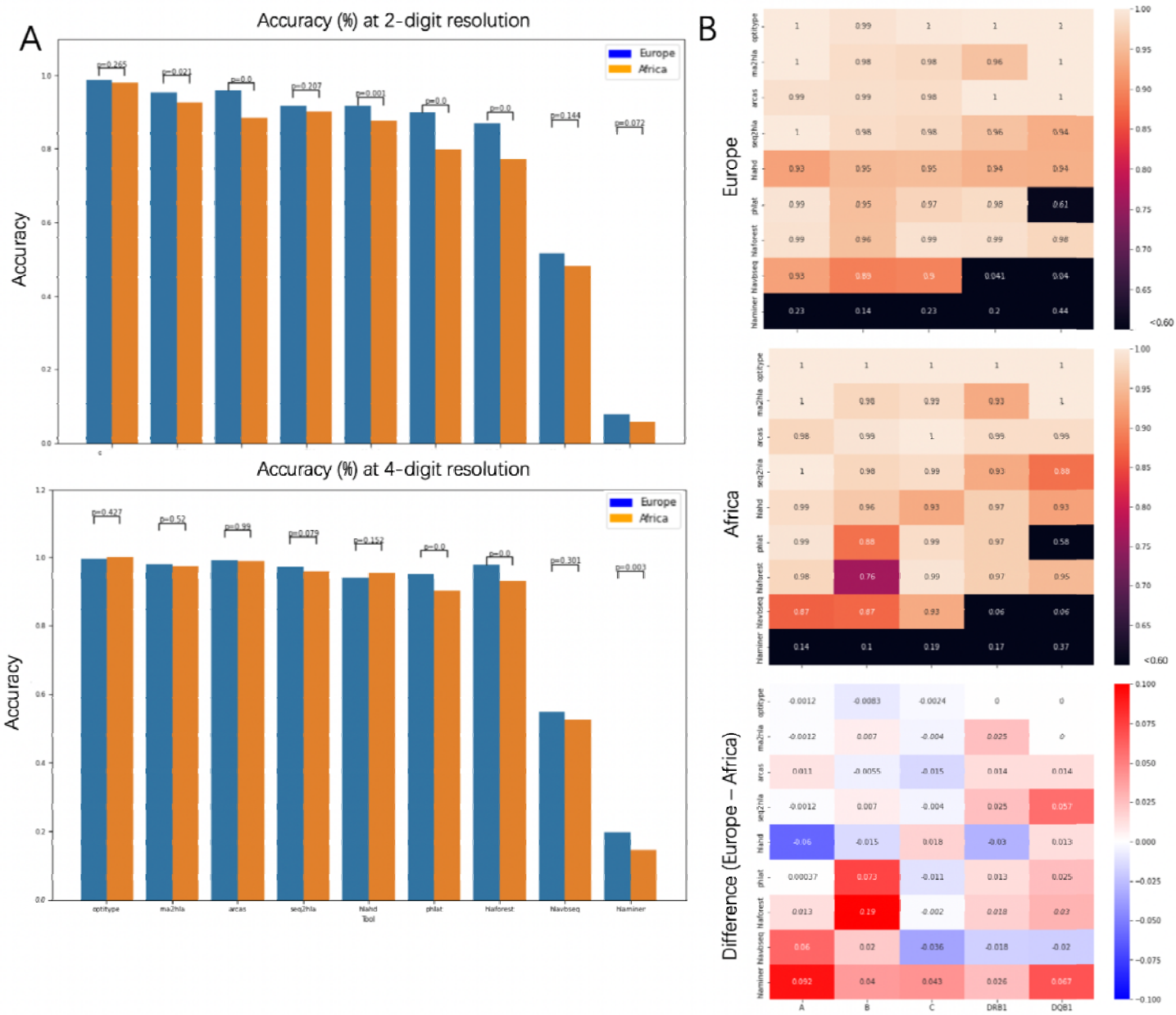

Figure 4: A: Accuracy at 2 and 4 digit resolution for the subset of samples from European ancestry (blue) and African ancestry (orange). B: Accuracy for Africa (top) and Europe (middle) subsets, stratified by locus. At the bottom is the difference, with loci at which Europe samples perform better in blue and loci at which Africa samples perform better in red.

Tables

Table 1: Overview of the gold standard datasets. “NCBI SRA Accession Number” indicates the unique identifier code for the complete dataset in the NCBI Gene Expression Omnibus record. “# total samples” indicates the total number of samples in the dataset which were used for this study. “”Sequencing platform” indicates the platform used to generate the RNA-seq reads from each sample. “Genes with Gold Standard HLA Types” indicates all classes and loci of the HLA complex that contained gold standard HLA types. “Length of RNA-seq reads” indicates the average read length of all samples in each dataset, with ‘x2’ suffix following each read count indicating that the samples are paired-end. “Tissue or cell type” indicates the type of cell or tissue that RNA was extracted from for RNA-seq. “HLA typing method” indicates the typing method used to determine the ground truth HLA type, to establish the gold standard datasets. “Read count before extraction” indicates the mean and range number of reads prior to extraction, and “Read count after extraction” indicates the mean and range number of reads after extraction of chromosome 6 mapping reads.

Data set #	NCBI SRA Accession number / NCBI GEO (ArrayExpress) accession number	Number of samples	Sequencing platform	Genes with Gold Standard HLA Types	Length of RNA-seq Reads	Tissue or cell type	HLA typing method	Read count before extraction (mean, range)	Read count after extraction (mean, range)
1	ERP001942 E-GEUV-1	490	Illumina HiSeq 2000	HLA class I (HLA-A, -B, -C) HLA class II (DRB1 and DQB1)	76bp x2	LCLs	PCR amplification followed by Sanger sequencing of the exons (SBT)	*29.9 million 8.6–83 million*	8.0 mil, 2.7-24 mil

2	SRP298704 GSE163605	86	Illumina HiSeq 2500	HLA class II (DRB1)	126bp x2	peripheral blood mononuclear cells (PMBCs), CD4+ T cells, CD8+ T cells, and CD14+ monocytes from 32 healthy individuals	PCR-based sequence-specific primers (SSP)	41 million, 24.5 million and 60.5 million	11.8 mil, 5.9 mil-20.0 mil
3	ERP000101 E-MTAB-197	50	Illumina Genome Analyzer II	HLA class I (HLA-A, -B, -C)	37bp x2	LCLs	PCR-based sequence-specific oligonucleotide probes (SSOP)	53.5 million, 35.0 million - 337.8 million	2.7mil, 0.9mil - 4.4mil
4	SRP099568 GSE94859	14	Illumina HiSeq 2500	HLA class II (DRB1)	126bp x2	Naïve helper T cells from peripheral blood	PCR-based sequence-specific primers (SSP)	54.4 mil 24.3 mil - 83.7 mil	8.0 mil, 3.4mil - 11.7mil

5	SRP198497 GSE131267	8	Illumina NextSeq 550	HLA class I (HLA-C)	37bp x2	Human B Lymphoblastoid cells (B721.221 cells)	mono-allelic B cells generated by transduction of B721.221 cells with a retroviral vector coding a single class I HLA allele	77.7 mil, 63.4 mil - 88.9 mil	7.5 mil, 5.8-8.2 mil
6	SRP096329 GSE93315	4	Illumina HiSeq 2500	HLA class I (HLA-A)	37bp x2	Human B Lymphoblastoid cells (B721.221 cells)	mono-allelic B cells generated by transduction of B721.221 cells with a retroviral vector coding a single class I HLA allele	157.6 mil, 170.0 mil - 191.6mil	28.2 mil, 17.1 mil - 29.2 mil

Table 2: Overview of HLA callers included in this study. “Tool” indicates the names of each of the 12 HLA caller tools benchmarked. “Version” is the software version of the tool which was downloaded and installed for benchmarking. “Types of reads accepted”: SR indicates Single-Read, PE indicates Paired-End. “Alignment algorithm” indicates the algorithm used for alignment of reads to the reference genome. “Programming language” indicates the programming language in which most code for the HLA caller was written in. “Published year” is the year in which the tool and corresponding paper were published. “Website” indicates the GitHub repository link from which the software was downloaded.

	Tool	Version	Types of reads accepted	Alignment algorithm	Algorithm for prediction	Programming language	Published year	Website
1	HLAminer	1.4	SR/PE	BLAST	Hybrid (reference	Perl	2012	https://github.com/bcgsc/

					based and de novo assembly algorithms)			HLAminer
2	seq2HLA	2.3	PE	Bowtie	Neural Networks	Python	2012	https://github.com/TRON-Bioinformatics/seq2HLA
3	HLAforest	1	PE	Bowtie	Random Forest	Perl	2013	https://github.com/FNaveed786/hlaforest
4	PHLAT	1.1	PE	Bowtie2	Hybrid (reference based and de novo assembly algorithms)	Python	2014	https://sites.google.com/site/phlatfortype/download
5	OptiType	1.3.3	SR/PE	RazerS3	Reference Based Probabilistic Model	Python	2014	https://github.com/FRED-2/OptiType
6	HLA-VBseq	2	SR/PE	BWA-MEM	Hybrid (reference based and de novo assembly algorithms)	Java	2015	http://nagasakilab.csml.org/hla/
7	HLA-HD	1.5.0	PE	Bowtie2	Hybrid (reference based and support vector machine (SVM))	C++	2017	https://www.genome.med.kyoto-u.ac.jp/HLA-HD/
8	arcasHLA	0.2.0	SR/PE	Kallisto	Hybrid (reference	Python	2019	https://github.com/Rabad

					based and de novo assembly algorithms)			anLab/arcasHLA
9	RNA2HLA	1.0	SR/PE	Bowtie	Neural Network	Python	2021	https://github.com/Chelysheva/RNA2HLA
10	HLApers	1.1	PE	Kallisto, Salmon, STAR	Reference Based	Python	2020	https://github.com/genevol-usp/HLApers
11	HLAProfiler	1.1.0	PE	Bowtie2, OptiType	Hybrid (reference based and de novo assembly algorithms)	Perl	2017	https://github.com/ExpressionAnalysis/HLAProfiler
12	HISAT-genotype	1.3.2	SE/PE	HISAT2	Reference based	Python	2019	https://daehwankimlab.github.io/hisat-genotype/

Table 3: Commands used to execute the HLA typing tools to obtain HLA prediction results. “IPD-IMGT/HLA default Reference version” indicates the version of the software that is the default version. “Example execution command” indicates the structure of the command we used to run each software, with all parameters set to default unless specified in the execution command.

	Tool	IPD-IMGT/HLA Default reference version	Example execution command
1	HLAminer (alignment/HPRA)	Release 3.3.0 and 3.4.0	bwa mem HLA-I_II_CDS.fasta 1.fq 2.fq > file.sam HLAminer.pl -a file.sam -h HLA-I_II_CDS.fasta -p hla_nom_p.txt -s 500 -q 20 -i 90 -z 100 -l > file.csv

2	seq2HLA	Class I: Release 3.6.0 Class II: Release 3.8.0	python seq2HLA.py -1 path_to_fastq1 -2 path_to_fastq2 -r path_to_results_dir
3	HLAforest	Release 3.10.0	scripts/CallHaplotypesPE.sh path_to_results path_to_fastq1 path_to_fastq2
4	PHLAT	Release 3.8.0	Python -O PHLAT_sm2.py -1 fastq.1.fq -2 fastq.2.fq -index phlat/b2folder -b2url bowtie2 -orientation --fr -tag sample_name -e phlat -o output/ -tmp 1
5	OPTITYPE	Release 3.14.0	Data preprocessing (recommended by developers): razers3 -i 95 -m 1 -dr 0 -o sample_1.bam hla_reference_rna.fasta path_to_fastq1 razers3 -i 95 -m 1 -dr 0 -o sample_2.bam hla_reference_rna.fasta path_to_fastq2 samtools bam2fq sample1.bam > fastq_1_finished.fastq samtools bam2fq sample2.bam > fastq_2_finished.fastq Running tool: python OptiType/OptiTypePipeline.py -i fastq_1_finished.fastq fastq_2_finished.fastq --rna -v -o output -c OptiType/config.ini
6	HLA-VBseq	Release 3.15.0	bwa mem -t 8 -P -L 10000 -a hla_all_v2.fasta.1.fq 2.fq> file.sam java -jar HLAVBSeq.jar hla_all_v2.fasta file.sam file_result.txt --alpha_zero 0.01 --is_paired
7	HLA-HD	Release 3.30.0	sh hlahd.sh -t 8 -m 10 -c 0.95 -f freq_data 1.fastq 2.fastq HLA_gene.split.3.32.0.txt dictionary sample_name .
8	arcasHLA	NA	./arcasHLA genotype 1.fastq 2.fastq -g A,B,C,DPB1,DQA1,DRB1,DMA,DMB,DPA1,DPA2,DRA,E,F,G,H,J,K,TAP1,TAP2,MICA,MICB -o results/ -t 8 -v
9	RNA2HLA	Class I: Release 3.6.0 Class II: Release 3.8.0	python RNA2HLA.py -f [path_to_dir_with_both_fastq] -r [path_to_results_dir] -g 6

10	HLApers	NA	Step 1: build transcriptome supplemented with HLA sequences hlapers prepare-ref -t file1.fq file2..fq -a gencode.v37.annotation.gtf.gz -i IMGTHLA -o hladb Step 2: Create an index for read alignment hlapers index -t hladb/transcripts_MHC_HLAsupp.fa -p 4 -o index Step 3: HLA genotyping hlapers bam2fq -b HG00096.bam -m ./hladb/mhc_coords.txt -o HG00096
11	HLAProfiler	NA	perl HLAProfiler.pl predict -fastq1 file1.fastq -fastq2 file2.fastq -database_name hla_database -database_dir test/outputs/HLAProfiler-1.0.0-db_only -reference path/to/HLAProfiler/test/outputs/HLAProfiler-1.0.0-db_only/hla_database/data/reference/hla.ref.merged.fa -if -l runlogs/filename.HLAProfiler.log
12	HISAT-genotype	Release 3.52.0	hisatgenotype -x [GENOME] --base [GENE_GROUP] -z [INDEX_DIR] [OPTIONS] -1 [FASTQ_PAIR1] -2 [FASTQ_PAIR2]

Supplementary tables

Table S1: Existing Methods for HLA typing which we failed to install or run. We opened an issue on each tool's GitHub if available. If no GitHub was publicly available, then we emailed the corresponding authors of the paper for each tool. "In progress" denotes we did not have a chance to complete this section yet, and will update the table as we make progress.

	Tool	Github Issue/Email	Log Error

1	HLAProfiler	https://github.com/ExpressionAnalysis/HLAProfiler/issues/6	Log file (ERR009106.HLAProfiler.log) shows typically ~80% alleles classified, ~20% unclassified. In prediction file however, all alleles show "not enough reads to make call".
2	HISAT-genotyope	In progress	In progress
4	HLApers	In progress	In progress

Table S2: The raw accuracy scores for each tool, with '% inaccurate' representing the percentage of alleles incorrectly predicted, '% 2 dig' representing the percentage of alleles accurate to 2 but not 4 digits, and '% 4 dig' representing the number of alleles accurate to 4 digits. The 3 categories are mutually exclusive and sum up to 1 for each tool. A: Accuracy on Class I alleles only. B: Accuracy on Class II alleles only. C: Overall accuracy on all Class I and Class II loci.

A

Tool	% inaccurate	% 2 dig	% 4 dig
OptiType	0.006173	0.011111	0.982716
arcasHLA	0.014763	0.050385	0.934852
RNA2HLA	0.015625	0.032813	0.951562
HLAforest	0.036814	0.124163	0.839023
Seq2HLA	0.015625	0.032813	0.951562
HLA-HD	0.058768	0.046512	0.89472
PHLAT	0.027177	0.061231	0.911591
HLA-VBseq	0.099871	0.05139	0.848739
HLAMiner	0.850141	0.085684	0.064175

B

Tool	% inaccurate	% 2 dig	% 4 dig
OptiType	0	0	0

arcasHLA	0.016144	0.049383	0.934473
RNA2HLA	0.067753	0.048885	0.883362
HLAforest	0.024332	0.11021	0.865458
Seq2HLA	0.085835	0.094278	0.819887
HLA-HD	0.062383	0.030019	0.907598
PHLAT	0.109276	0.064168	0.826557
HLA-VBseq	0.926346	0.013692	0.059962
HLAMiner	0.763615	0.236385	0

C

Tool	% inaccurate	% 2 dig	% 4 dig
OptiType	0.006173	0.011111	0.982716
arcasHLA	0.01532	0.049981	0.934699
RNA2HLA	0.029546	0.037105	0.933349
HLAforest	0.031668	0.118411	0.849921
Seq2HLA	0.043698	0.057389	0.898912
HLA-HD	0.060218	0.039895	0.899887
PHLAT	0.055099	0.06223	0.882671
HLA-VBseq	0.435725	0.036071	0.528204
HLAMiner	0.817405	0.142701	0.039895