Benchmark of tools for *in silico* prediction of MHC class I and class II genotypes from NGS data

Authors and affiliations

Arne Claeys[1,2], Jasper Staut[1], Peter Merseburger[1], Kathleen Marchal[2,3,4] and Jimmy Van den Eynden[1,2*]

1. Department of Human Structure and Repair, Anatomy and Embryology Unit, Ghent University, Ghent, Belgium
2. Cancer Research Institute Ghent, Ghent, Belgium
3. Department of Information Technology, IDLab, Ghent University, Ghent, Belgium
4. Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium

* Corresponding author:

E-mail: jimmy.vandeneynden@ugent.be

Abstract

The Human Leukocyte Antigen (HLA) genes are a group of highly polymorphic genes that are located in the Major Histocompatibility Complex (MHC) region on chromosome 6. The HLA genotype affects the presentability of tumour antigens to the immune system. While knowledge of these genotypes is of utmost importance to study differences in immune responses between cancer patients, gold standard, PCR-derived genotypes are rarely available in large Next Generation Sequencing (NGS) datasets. Therefore, a variety of methods for *in silico* NGS-based HLA genotyping have been developed, bypassing the need to determine these genotypes with separate experiments. However, there is currently no consensus on the best performing tool. Here, we compiled a list of 13 HLA callers and evaluated their accuracy on three different datasets. Based on these results, best-practice

guidelines were constructed, and consensus HLA allele predictions were made for DNA and RNA samples from The Cancer Genome Atlas (TCGA).

Introduction

The human Major Histocompatibility Complex (MHC) is a gene complex located on the p-arm of chromosome 6 that contains two large clusters of genes with antigen processing and presentation functions: the MHC class I and MHC class II regions [1–3].

MHC class I molecules are involved in the presentation of endogenous antigens to cytotoxic T-cells and consist of a heavy chain encoded by one of the Human Leukocyte Antigen Class I genes (*HLA-A*, *HLA-B* or *HLA-C)*, and a light β_2 microglobulin chain [4–6]. Their role in tumour immunity has been established for a long time. Indeed, they can present neoantigens, small mutated peptides, to CD8+ T cells, resulting in an immune response and cancer cell death.

The most frequently studied MHC class II genes include *HLA-DPA1, HLA-DPB1*, *HLA-DQA1, HLA-DQB1*, *HLA-DRA1* and *HLA-DRB1*. They encode alpha and beta heterodimers that form the MHC class II protein complex. The role of these genes in anti-tumour immunity is emerging [7–9]. MHC class II mediated tumour-immune interaction occurs either via an indirect or a direct mechanism. First, cancer cells can secrete neoantigens that are subsequently taken up and presented on the MHC-II of antigen presenting cells infiltrating the tumour [7,10,11]. Additionally, some tumours express MHC-II themselves and can directly interact with CD4+ T-cells [7,11].

The peptide-binding region of HLA molecules is highly polymorphic and specific HLA alleles determine neoantigen binding and presentation to the immune system. Genotype dependent differences in HLA binding affinity could lead to differential responses to immunotherapy, as illustrated by the association that has been described between MHC-I genotypes (e.g., *HLA-B62*) and survival in immune checkpoint blockade (ICB)-treated advanced melanoma patients [12]. It is currently unclear whether MHC-II genotypes also determine responses to immunotherapy.

48 Such association studies require knowledge of the HLA genotype. PCR methods are currently the gold
49 standard for this genotyping but datasets with PCR-based HLA calls are rarely available [13–15]. HLA
50 genotyping can also be performed on Next Generation Sequencing (NGS) data. A plethora of tools
51 has been developed for this task. *Polysolver* and *Optitype* are often recommended as the best
52 performing tools for MHC-I genotyping [16]. For MHC-II genotyping there is currently no consensus
53 about the best method. Several benchmarks have been performed previously [14,16–23], but these
54 were either not applied to MHC class II or did not include some recently published tools.

55 In this study, we compiled a list of 13 tools that predict HLA genotypes from NGS data and
56 benchmarked their performance on both the 1000 genomes dataset and on an independent cell line
57 dataset. Subsequently we assessed their performance on 9162 DNA and 9761 RNA sequencing files
58 from The Cancer Genome Atlas (TCGA) by comparing the predicted allele frequencies with reference
59 population allele frequencies. Based on these findings, we give recommendations on which tool to
60 use for a given data type and how the outputs of multiple tools can be combined into a consensus
61 prediction.

62 Results

63 Selection of 13 HLA genotyping tools with variable computational 64 resource requirements

65 We identified 22 available HLA genotyping tools from literature (Table 1). Thirteen tools that were
66 free for academic use, applicable on Whole Exome Sequencing (WES), Whole Genome Sequencing
67 (WGS) or RNA-Seq data and ran on Ubuntu 20.04 were included in this study: *arcasHLA*, *HLA-HD*,
68 *HLA-VBSeq*, *HLA*LA*, *HLAforest*, *HLAminer*, *HLAscan*, *Kourami*, *Optitype*, *PHLAT*, *Polysolver*, *seq2HLA*
69 and *xHLA (*Table S1*)*. These HLA typing algorithms mainly differ in the method to align reads to the
70 HLA allele reference sequences and in the subsequent allele prioritization approach (Table S2). All 13
71 tools can make allele predictions for the three MHC class I genes (*HLA-A*, *HLA-B* and *HLA-C*) and 9

tools support additional calling of the MHC class II genes *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1* and *HLA-DRB1*. Two methods support only a subset of the MHC class II genes: *xHLA* does not support calling *HLA-DPA1* and *HLA-DQA1*, while *PHLAT* does not support *HLA-DPA1* and *HLA-DPB1*. The tools also differ in which data types they support: 6 of them require DNA data, 3 tools require RNA data and the 4 remaining tools support both data types (Table 1).

Firstly, the computing time and memory usage of the thirteen selected tools were measured on a random subset of 10 DNA and 10 RNA sequencing files from the TCGA project (Fig 1).

Among the 10 DNA-supporting methods *Optitype* (median 2.48 hours) and *HLA*LA* (median 1.84 hours) require the largest computing time. The remaining DNA tools take less than 1 hour per file, with *HLAminer*, *Kourami* and *PHLAT* being the fastest (97s, 225s and 253s respectively). Apart from being computationally intensive, *HLA*LA* is also the most memory demanding DNA tool (median 36.3 GiB per file). Other DNA tools with a median memory consumption higher than 5 GiB are *xHLA* (median 22.9 GiB), *Kourami* (median 9.3 GiB) and *HLA-HD* (median 6.7 GiB). The relatively low memory usage of *Polysolver* makes it feasible to compensate for its long running time by processing multiple samples in parallel.

Among the 7 RNA-supporting methods, *HLA-HD* has the longest computing time per sample (median 15.0 hours). At the other end of the spectrum, the sole pseudoalignment-based tool *arcasHLA* takes only 38s per file. The most memory intensive tool is *HLA-HD* (median memory peaks of 103.1 GiB), followed by *Optitype* (median 34.1 GiB). The other RNA tools have a memory usage lower than 10 GiB. Remarkably, *HLAminer*, *PHLAT* and *HLA-HD*, which are compatible with both DNA and RNA data take a longer time on RNA data (median computing time per sample is 29.4, 8.9, 6.8 times longer for *HLA-HD*, *PHLAT* and *HLAminer* respectively).

		Data type		Input filetype		HLA loci								Version
		DNA	RNA	BAM	FASTQ	A	B	C	DPA1	DPB1	DQA1	DQB1	DRB1	
Included	arcasHLA [15]	✗	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	0.2.0
	HLA-HD [24]	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	1.3.0
	HLA-VBSeq [25]	✓	✗	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	2
	HLA*LA [26]	✓	✗	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	1.0.1
	HLAforest [27]	✗	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	1
	HLAminer [28]	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	1.4
	HLAscan [29]	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	2.1.4
	Kourami [30]	✓	✗	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	0.9.6
	Optitype [13]	✓	✓	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	1.3.5
	PHLAT [31]	✓	✓	✗	✓	✓	✓	✓	✗	✗	✓	✓	✓	1.1
	Polysolver [32]	✓	✗	✓	✗	✓	✓	✓	✗	✗	✗	✗	✗	4
	seq2HLA [33]	✗	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	2.3
	xHLA [34]	✓	✗	✓	✗	✓	✓	✓	✗	✓	✗	✓	✓	0.0.0
Not included	ALPHLARD-NT [35]	✓	✗	?	?	✓	✓	✓	✓	✓	✓	✓	✓	
	ATHLATES [36]	✓	✗	✗	✓	✓	✓	✓	✗	✗	✗	✓	✓	
	HLAProfiler [37]	✗	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	
	HLAreporter [38]	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	
	HLAssign [39]	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	
	OncoHLA [40]	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	
	PolyPheMe [41]	✓	✗	✗	✓	✓	✓	✓	✗	✗	✗	✓	✓	
	SNP2HLA [42]	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	
	SOAP-HLA [43]	✓	✗	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	

Table 1. Overview of evaluated tools for HLA genotyping. List of tools for NGS-based HLA genotyping. Checkmarks and crosses indicate which NGS methods (DNA-Seq and/or RNA-Seq) and input file types (FASTQ and/or BAM) are supported and for which genes predictions can be made. The tools in the upper part of the table are benchmarked in this study. Tools in the lower part of the table did not fulfil our inclusion criteria and were not further considered.

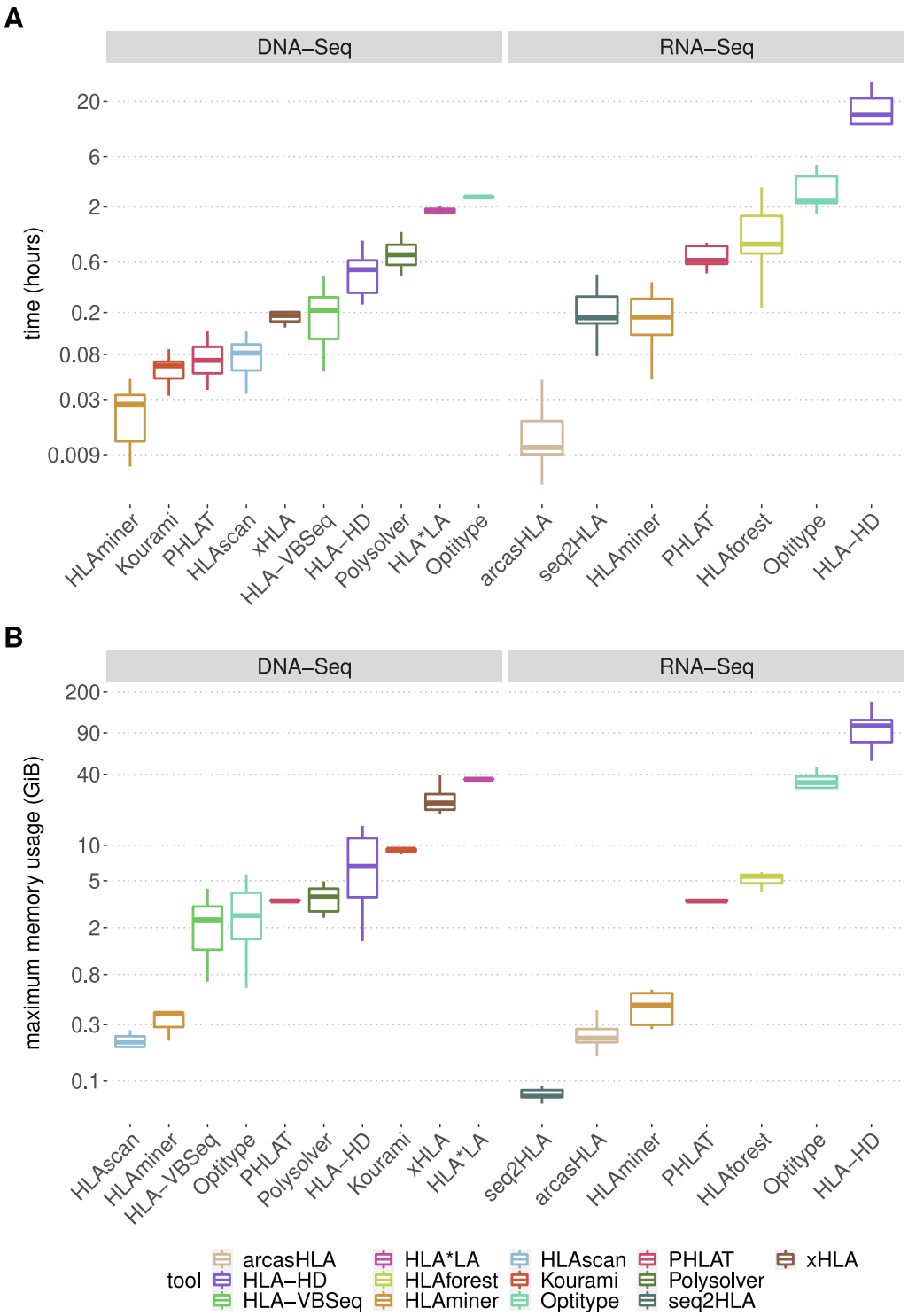

Figure 1. Computational resource consumption of the 13 selected tools. (**A-B**) Boxplots compare the resources needed by the different tools to analyse one sequencing file on a system with a single CPU core. Each tool was applied on DNA and/or RNA sequencing files *(n=10)*, as indicated at the top of the figure. Different tools are represented with a different colour of the boxplot, as indicated in the legend on the right. (**A**) Time consumption per sample. (**B**) Maximal memory consumption per sample.

106 # HLA*LA and HLA-HD are the best performing MHC class II

107 # genotyping tools on DNA data

108 The 10 selected algorithms that are compatible with DNA sequencing data were benchmarked using
109 WES data from the *1000 Genomes project* [44] (Fig 2). Predictions were made for *HLA-A* (n = 1012),
110 *HLA-B* (n = 1011), *HLA-C* (n = 1010), *HLA-DQB1* (n = 1008), *HLA-DRB1* (n = 1000) and *HLA-DQA1* (n =
111 68). *HLA-DPA1* and *HLA-DPB1* were not benchmarked due to the lack of available gold standard calls.
112 For MHC-I genes (*HLA-A*, *HLA-B*, *HLA-C*), the best accuracy was obtained with *Optitype* (98.0%),
113 followed by *Polysolver* and *HLA*LA* (94.9% and 94.4% respectively). For MHC-II genes (*HLA-DQA1*,
114 *HLA-DQB1* and *HLA-DRB1*), the best allele predictions were made using *HLA-HD* and *HLA*LA* (96.2%
115 and 95.7% accuracy respectively). These were the only two methods to reach an accuracy of 90% on
116 all tested MHC-II genes. *HLAscan* (74.2%), *HLA-VBSeq* (60.2%) and *HLAminer* (53.8%) performed
117 considerably worse than the other tools.

118 We observed large variabilities in calling accuracies between MHC class II genes (Fig 2). Overall, *HLA-*
119 *DQB1* was the hardest MHC-II gene to call. Except for *PHLAT*, all tools obtained their worst MHC-II
120 call accuracy on this gene. *HLA-DQA1*, on the other hand, was the gene with the highest calling
121 accuracy for all tools that support it, except for *HLAminer* and *Kourami*.

122 Incorrect calls are either caused by wrong allele calls or a failure to make an allele call. When
123 *HLAscan* and *Kourami* were able to make a call, their predictions were most often reliable (S1 Fig),
124 but these tools regularly produced no output at all (S2 Fig). *HLA-VBSeq* and *HLAminer* had both a high
125 rate of incorrect and failed calls (S1-S2 Figs).

126 Subsequently, we performed an independent benchmark using the smaller NCI-60 cell line dataset
127 (n=58), which largely confirmed our results (S3 Fig). Additionally, this analysis indicated that the best
128 performing MHC class II supporting tools also performed well on *HLA-DPB1*.

129 **HLA-HD, PHLAT and arcasHLA are the best performing MHC class II**

130 **genotyping tools on RNA data**

131 We then evaluated the 7 selected methods that support HLA calling on RNA sequencing data from
132 the *1000 genomes project* [45] (Fig 2). Predictions were made for *HLA-A* (n = 373), *HLA-B* (n = 372),
133 *HLA-C* (n = 372), *HLA-DQB1* (n = 371), *HLA-DRB1* (n = 362) and *HLA-DQA1* (n = 53).

134 A*rcasHLA* and *Optitype* had the best MHC-I allele predictions (99.4% and 99.2% accuracy,
135 respectively), followed by *HLA-HD* (98.0%), *seq2HLA* (95.9%) and *PHLAT* (95.4%). Similar accuracies
136 were found for MHC-II allele predictions, with *HLA-HD*, *PHLAT* and *arcasHLA* performing the best
137 (99.4%, 98.9% and 98.1%, respectively). Contrary to its good prediction of MHC class I alleles,
138 *seq2HLA* has a lower accuracy for MHC class II (87.8%).

139 The high MHC-I accuracies of *arcasHLA* and *Optitype* were confirmed on the independent NCI-60
140 dataset (91.8% and 90.0%, respectively; n=58). The accuracy of *HLA-HD*, *PHLAT* and *seq2HLA* was
141 worse on the cell lines than in the benchmark on the 1000 genomes data (86.6%, 83.3% and 82.3%,
142 respectively). As MHC-II is generally not expressed in cell lines, this benchmark was not performed
143 for those genes.

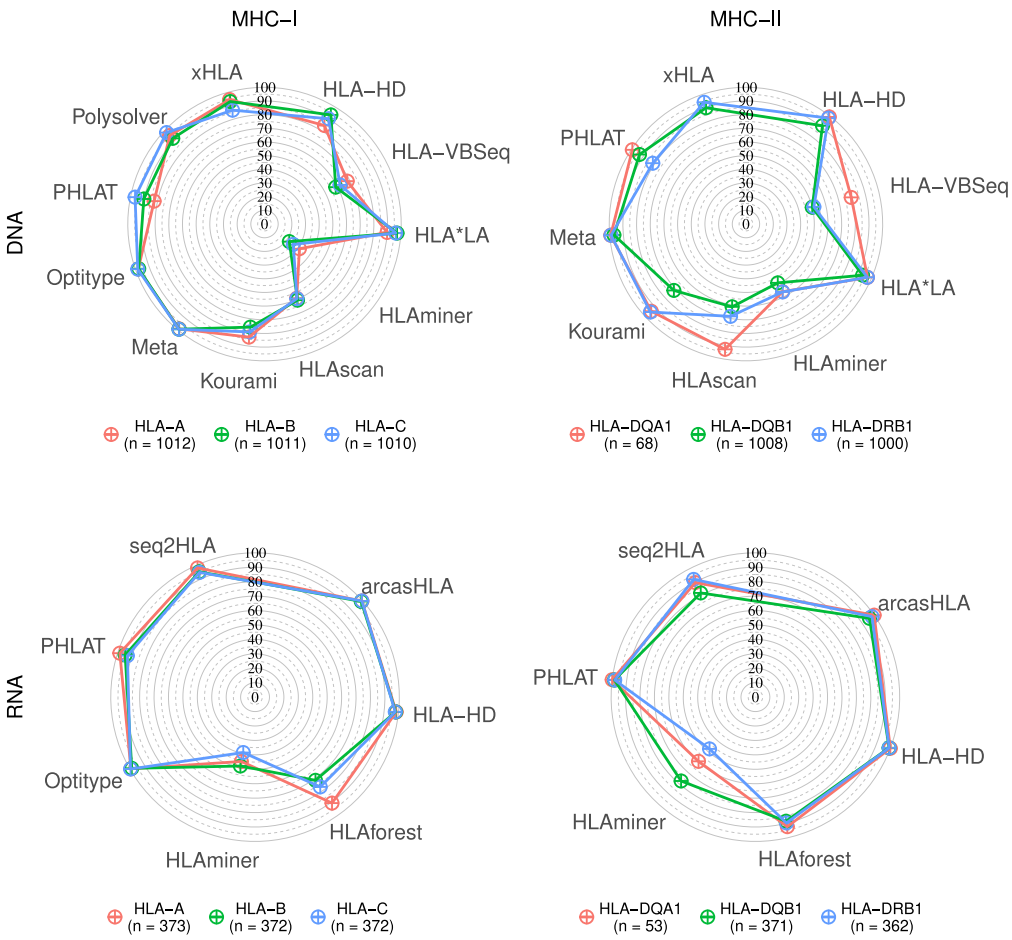

Figure 2. HLA allele prediction accuracies. Radar plots of HLA allele prediction accuracies on samples from the 1000 Genomes Project. Coloured lines represent different genes, as indicated in the legend below the plots. Corners of the radar plots correspond to the tools that were evaluated for that data type. The Meta tool correspond to the 4-tool consensus metaclassifier.

Correlation and concordance analyses on large independent datasets confirm the benchmarking results.

Being one of the few large sequencing datasets for which gold standard HLA genotypes for both MHC classes are available, many algorithms included in our benchmark were developed, optimized and validated using files from the *1000 genomes* project, introducing a potential bias. Additionally, no evaluation was possible for *HLA-DPA1* and *HLA-DPB1*, due to the lack of gold standard HLA calls. Therefore, we performed an indirect and independent evaluation on a large NGS dataset obtained from TCGA (n=9162 and n=9761 for DNA and RNA respectively).

157 We first compared the observed allele frequencies for each tool with the expected population
158 frequencies. We calculated how often each of the alleles was predicted by a certain tool to obtain an
159 observed allele frequency, stratifying for Caucasian American (n= 7935) and African American
160 (n=938) ethnicities. By comparing these frequencies to the expected allele frequencies, as derived
161 from *Allele Frequency Net* [46], strong correlations (i.e., Pearson's *r* 0.94 or higher for all genes and
162 both populations) were found for the DNA-based tools *HLA-HD*, *HLA*LA*, *Optitype*, *Polysolver* and
163 *xHLA* and for the RNA-based tools *Optitype*, *arcasHLA* and *PHLAT*. The correlations were considerably
164 worse for *HLA-VBSeq*, *HLAminer* and *HLAforest* than for the other tools (Fig 3). These findings largely
165 confirm the results of the benchmark on the 1000 genomes data. Notably, among the well
166 performing tools, *arcasHLA* had a worse correlation for *HLA-DRB1*, which is mainly due to the
167 discrepancy between the observed and predicted frequency of *HLA-DRB1*14:02* in this population
168 (S4 Fig).

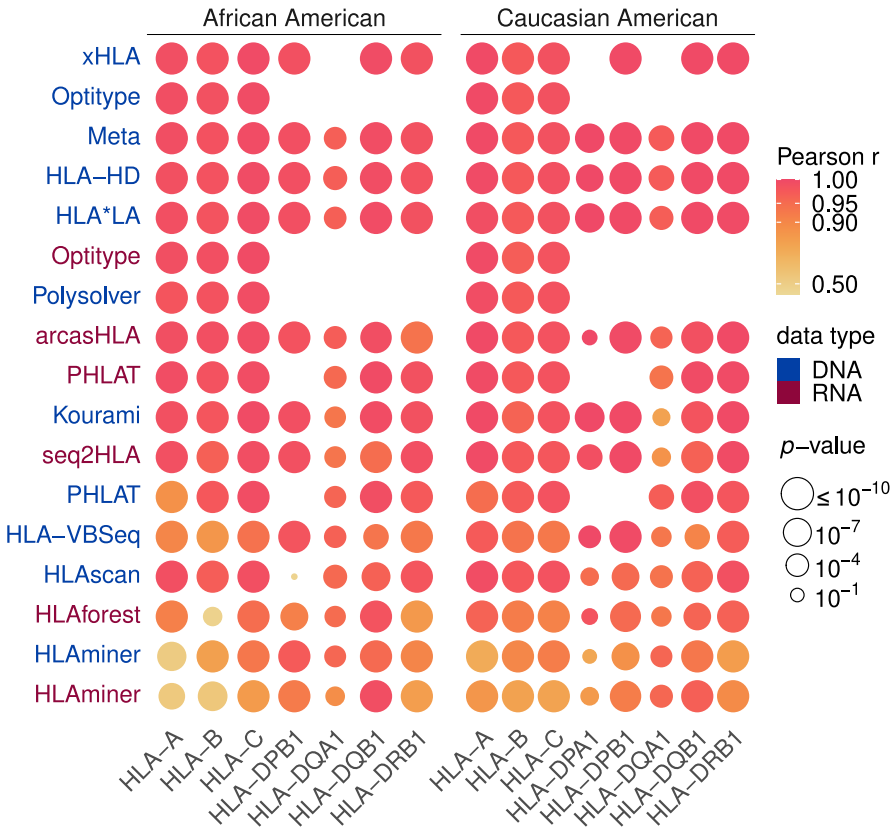

Figure 3. Correlations between observed and expected allele frequencies. Heatmap of correlations between observed allele frequencies and frequencies expected in an African American and in a Caucasian American population. Vertical axis indicates the tools, with different colours representing the data type (DNA or RNA) on which the tool was applied. Rows were sorted according to the mean correlation of the tool. Size of the circles indicates the *p*-value of the correlation test as indicated in legend. Absent circles indicate that the tool could not be evaluated on that gene.

We then calculated for each pair of tools how often their predictions are concordant (S5-S8 Figs). Tools that performed poorly in the previous analyses (e.g., *HLAminer*, *HLA-VBSeq* and *HLAforest*) consistently have a low concordance with all other tools. In contrary, tools that scored high in the previous analyses (such as *Optitype, HLA*LA, arcasHLA* and *HLA-HD*) made predictions that are consistent with each other. Noteworthy, this is also the case for *HLA-DPA1* and *HLA-DPB1*, two genes for which no gold standard data was available, suggesting that predictions for these genes are reliable as well.

A consensus metaclassifier improves HLA predictions for DNA data

We noted that only for a very small fraction of the samples the genotypes are wrongly typed by all tools simultaneously (median 0.79% for DNA and 0.68% for RNA; Figures S9-S10). This

186 complementarity of the tools' allele predictions opens the possibility to combine predictions of
187 different HLA callers into a consensus prediction. We first applied a majority voting algorithm to the
188 output of all tools, with the predicted allele pair being the one with most votes. On the DNA data,
189 this approach outperforms the predictions of each individual tool for all genes. This is best illustrated
190 by the *HLA-DQB1* gene, where the accuracies increased from 93.2% with the best performing tool
191 (*HLA*LA*) to 96.3% when the voting metaclassifier was used. On RNA data, where the best tools
192 already attain accuracies over 99% by themselves, only minor improvements were made by
193 combining the results (S11 Fig).

194 Based on these results, we determined the minimal number of tools that must be included in the
195 DNA-based metaclassifier to produce reliable results (Fig 4). For the DNA data, including 4 tools in
196 the model led to a considerable improvement for all genes for both MHC classes. The best accuracies
197 were observed when *Optitype*, *HLA*LA*, *Kourami* and *Polysolver* were combined for MHC-I
198 predictions (99.0% accuracy) and with *HLA*LA*, *HLA-HD*, *PHLAT* and *xHLA* for MHC-II predictions
199 (98.4% accuracy). Raising the number of tools further only resulted in marginal gains. Strikingly, the
200 accuracy of the *HLA-DQB1* allele predictions even decreases when more tools were included in the
201 model. Therefore, we suggest combining the output of 4 tools for both MHC classes.

202 To evaluate whether the good performance of this approach is generalizable to other datasets, we
203 assessed the correlation between the expected allele frequencies and the allele frequencies
204 observed using the 4-tool DNA consensus predictions on the TCGA dataset and compared the results
205 with our previous findings. The allele frequencies predicted by the metaclassifier correlated better
206 with the expected allele frequencies (Fig 3) than was the case for the individual tools that supported
207 all genes of interest.

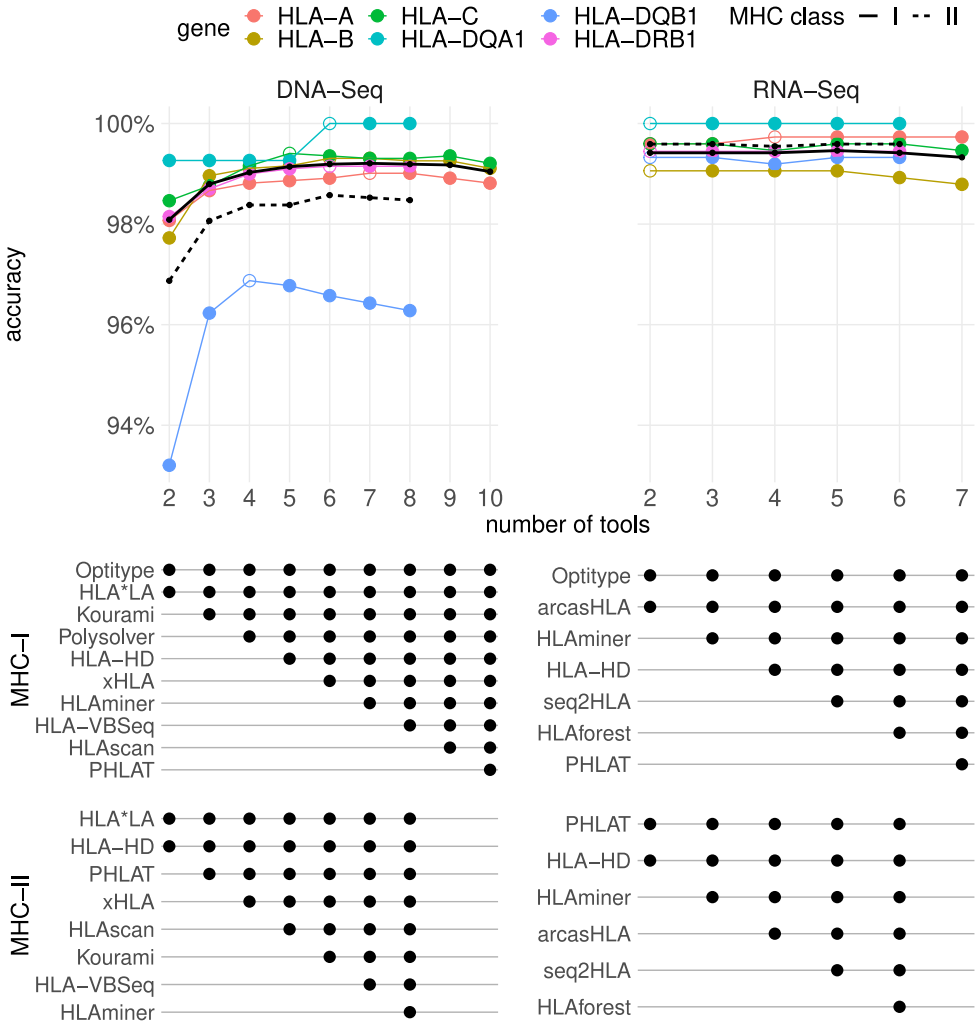

Figure 4. Accuracies of meta-prediction models with an increasing number of included tools. Tools were added one by one to the consensus metaclassifier model. At each step, the prediction accuracies of the best performing metaclassifier model for a given number of tools were plotted at the top of the figure. Unfilled markers are placed at the smallest number of tools where the maximal accuracy was obtained for that gene. Black lines indicate the average accuracy of the consensus predictions for the two MHC classes (averaged over all genes of that class). The table below the plot indicates which tools were selected in each model for a given number of tools.

Discussion

Rapid technological advancements in NGS have resulted in the generation of numerous publicly available WES and RNA-Seq datasets. These data have been critical for understanding the genomic basis of human carcinogenesis [47]. In the field of immuno-oncology, genomic data have also been used to study immune selection [48,49] and, additionally, the availability of corresponding clinical data opens possibilities for studying HLA-dependent cancer susceptibility or even differences in clinical ICB responses between cancer patients [12,50–52]. However, this requires that the HLA

223 genotype for each subject can be accurately determined. An ever-increasing number of NGS-based
224 HLA typing software applications have been developed. In this study, we benchmarked the
225 performance of 13 publicly available tools. To our knowledge, this is the most extensive benchmark
226 of MHC genotyping tools that has been performed so far (Table S4).

227 First, we evaluated the tools by comparing their output to genotypes derived from a PCR-based
228 approach. While PCR methods are the gold standard for HLA typing, they have limitations that could
229 lead to ambiguous typing results [53]. Furthermore, inconsistencies have been reported across PCR-
230 based HLA typing datasets that are available for the 1000 genomes samples [54] which could have
231 affected our benchmarking results. Therefore, we also used 2 other, indirect approaches to assess
232 the performance of the different tools.

233 Both a concordance analysis between the tools' predictions and a correlation analysis between
234 predicted and expected allele frequencies confirmed our benchmarking results. To avoid biasing the
235 results of this correlation analysis, we disabled ethnicity-specific allele frequencies for the algorithms
236 that support this (i.e., *arcasHLA* and *Polysolver*). However, in the case of *arcasHLA*, when no specific
237 ethnicity is specified, it uses prior frequencies that depend on the prevalence of the alleles in the
238 entire human population, possibly hindering its ability to call alleles that are uncommon in the
239 specified population. This is illustrated by the worse correlation between observed and expected
240 allele frequencies of *arcasHLA* for *HLA-DRB1* in the African American population, due to an
241 overestimation of the frequency of the rare *HLA-DRB1*14:02* allele.

242 We found that *Optitype, Polysolver, HLA-HD, HLA*LA* and *xHLA* are all solid choices for DNA-based
243 MHC genotyping, while *Optitype*, *HLA-HD*, *arcasHLA* and *PHLAT* are the better performing tools for
244 RNA data. On the other hand, *HLAminer*, *HLA-VBSeq* and *HLAScan* performed rather poorly in our
245 benchmark. Similar trends were observed in previous independent benchmarking studies [14,17,19–
246 23] that focused on a subset of tools and/or genes (Table S4), with the exception of *xHLA* where we
247 obtained considerably higher accuracies on WES data than reported in a study by Chen et al. [19].

248 The optimal strategy for HLA genotyping depends on a few factors: the availability of DNA or RNA
249 data, the size of the dataset that needs to be analysed and the available computational resources.
250 Additionally, MHC class II typing based on RNA data is only feasible on sequencing data derived from
251 MHC-II expressing cells. For DNA data, *Optitype* and *HLA-HD* are the best performing individual tools
252 for MHC class I and MHC class II typing, respectively. For RNA data, the same tools are recommended
253 when sufficient computational resources are available. However, the large resource and time
254 consumption of *HLA-HD* on RNA data makes its usage rather impractical on large datasets. As an
255 alternative, *arcasHLA* is recommended, which is both the fastest and more accurate tool for RNA that
256 supports all 5 MHC class II genes. Finally, we have demonstrated that the accuracy of the DNA-based
257 HLA genotype predictions can be improved further by combining the output of *Optitype*, *HLA*LA*,
258 *Kourami* and *Polysolver* for MHC-I typing and combining *HLA*LA*, *HLA-HD*, *PHLAT* and *xHLA* for MHC-
259 II typing using a majority voting rule. For RNA data a similar approach did not lead to a further
260 improvement of the prediction accuracies.

261 Materials and Methods

262 Selection of tools

263 A list of existing HLA genotyping tools for NGS data was compiled from literature between October
264 and December 2020. The tools that fulfilled the following criteria were selected for further analysis:
265 the tool should be free for academic use, support DNA and/or RNA sequencing data, should not
266 require enrichment of the HLA region before sequencing and should be a Linux command line tool
267 that we could successfully run on our system. When the authors provided instructions on how to
268 update the IPD-IMGT/HLA database used by their tool, this database was updated to version 3.43.
269 This was the case for three tools: *HLA-HD*, *HLAminer* and *Kourami*.

Next-generation sequencing datasets for benchmark

Slices of the 1012 CRAM files of WES data from the *1000 Genomes on GRCh38* dataset [44] that were used for the benchmark on DNA data were obtained from the *International Genome Sample Resource* using the *samtools view* command (version 1.12). The following contigs were included in the download: the MHC region on the primary assembly (chr6:28509970-33480727), all 525 contigs starting with *HLA*- and all unmapped reads. The sliced BAM files for the RNA benchmark were obtained from the *Geuvadis* [45] RNA-Seq dataset (part of the *1000 genomes* project) via *ArrayExpress* (accession number *E-GEUV-1*). All reads mapped to the MHC region and the unmapped reads were included in the download. Sequencing data from NCI-60 cell lines [24] were obtained from the *Sequence Read Archive* with accession numbers *SRP150855* (WES) [55] and *SRP133178* (RNA) [56]. The NCI-60 sequencing data was realigned according to the same alignment pipeline used by the *1000 Genomes on GRCh38* dataset [44]: reads were aligned to the complete GRCh38 reference genome, including ALT contigs and HLA sequences, using an alternative scaffold-aware version of BWA-MEM. As done in the same 1000 genomes alignment pipeline, PCR-introduced duplicates were marked using the *markduplicates* function in BioBamBam (version 2.0.182). Aligned sequences of Whole Exome Sequencing (WES) and RNA sequencing experiments from *The Cancer Genome Atlas (TCGA)* were downloaded in BAM format from the *Genomic Data Commons (GDC)* portal. All 9162 available BAM files of Blood Derived normal WES samples were selected. For RNA-Seq, all 9762 RNA-Seq samples that were derived from primary tumours and were aligned using the "STAR 2-Pass" workflow, were selected. Reads mapped to the MHC region of chromosome 6 (chr6:28509970-33480727) and unmapped reads were extracted from the BAM files and downloaded following the instructions that are described in the GDC API. For the RNA-Seq samples one file failed to download after multiple attempts. The resulting dataset consists of 9162 blood-derived normal WES samples and 9761 primary tumour RNA-Seq samples from 33 available cancer types. The most resource intensive RNA tools were applied on a subset of the TCGA dataset. *Optitype*

295 was applied on 2226 RNA files , *HLAforest* on 2900 files and *HLA-HD* was not applied on the TCGA
296 data.

297 Gold standard HLA typing data

298 Gold standard PCR-based HLA calls for the samples from the *1000 genomes on GRCh38* dataset were
299 provided by three earlier studies [26–28]. The HLA genotypes from these datasets were merged.
300 Where the calls did not agree, the calls by Gourraud et al. [57] were preferred. For the NCI-60 cell
301 lines, PCR-based HLA genotypes were provided in a study by Adams et al [29]. For both reference
302 datasets alleles were mapped to the corresponding G-groups, as defined by IPD-IMGT
303 (http://hla.alleles.org/alleles/g_groups.html), and trimmed to the second-field resolution.

304 HLA allele predictions

305 All 13 selected tools were run on the sliced BAM files following the guidelines of the authors. For
306 tools requiring FASTQ input files, a FASTQ file was extracted from the sliced BAM files using *samtools*
307 *fastq.* For *HLAScan*, which supports input files in either file format, the input was provided in BAM
308 format. For tools that allowed to specify a list of loci that should be called: *HLA-A, HLA-B, HLA-C, HLA-*
309 *DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1* and *HLA-DRB1* were chosen. *Kourami* was run with the *-a*
310 (additional loci) parameter to call the *HLA-DPA1* and *HLA-DPB1* genes. In rare cases, this led to a
311 crash of the tool and *Kourami* was run again without the *-a* parameter. For *HLAminer* only the HPRA
312 mode was evaluated. *xHLA*, *Polysolver* and *HLA-VBSeq* were not compatible with BAM files that are
313 aligned to a reference genome build that includes alternative contigs. For these tools, an additional
314 realignment step was performed before the tool was executed. Input data for *xHLA* and *Polysolver*
315 were realigned to a GRCh38 build that excludes alternative contigs. The input data for *HLA-VBSeq*
316 was realigned to GRCh37. All allele predictions were mapped to the corresponding G-groups and
317 trimmed at second-field resolution.

318 ## Measuring the resource consumption

319 The running time and memory consumption required by the tools were measured for a random
320 subset of 10 DNA and 10 RNA sequencing files from the TCGA project. Each tool was executed in a
321 separate Docker container (version 19.03.3) that was allocated a single CPU core. When the package
322 provided a parameter to specify the number of threads, this was set to 1. Per file, the memory usage
323 of the Docker container was monitored using the *docker stats* command. The running time was
324 calculated as the time interval between the start and the end of the tool, excluding the time to start
325 the Docker container. Pre-processing steps related to realignment to a different genome build (as
326 required for *xHLA*, *Polysolver* and *HLA-VBSeq*) were not included in the resource consumption
327 assessment. For HLA-HD the analysis of a single sample did not complete successfully as the required
328 amount of memory exceeded what we have available on our system.

329 ## Performance metric

330 For each sample, two allele predictions were made. An allele prediction was labelled "correct" when
331 it was listed as one of the two alleles in the gold standard for that patient. When a tool made a
332 homozygous prediction, while the gold standard was heterozygous, at most one of the two
333 predictions was labelled "correct" for that sample. The accuracy of the predictions is then defined as
334 the proportion of all correctly predicted alleles divided by twice the number of samples. Samples
335 where the gold standard was missing for a particular gene were ignored for that gene.

336 ## Population frequency data

337 Lists of expected HLA allele frequencies for an African American and for a Caucasian American
338 population were constructed based on 18 different studies in the *Allele Frequency Net* [46] database
339 (Table S3). The studies were selected based on the following criteria. First, we required that the
340 study was conducted on a *Black* or *Caucasoid* population from the United States. This was not
341 possible for *HLA-DPA1* where no HLA allele frequencies were available for these ethnicities. As a

342 substitute, the allele frequencies of three European populations (French, Swedish and Basques) were
343 used to approximate the allele frequencies for this gene in Caucasian Americans. As a second
344 requirement, the HLA calls should be determined by a PCR-based method. Thirdly, the *Allele*
345 *Frequency Net* database should have assigned a gold label (i.e., allele frequency sums to 1, sample
346 size of study > 50, and at least 2-field resolution) to the study for the gene of interest. Lastly, it was
347 required that the subjects included in the selected studies were healthy subjects (i.e., selected for an
348 anthropological study, blood donors, bone marrow registry or controls for a disease study). Allele
349 frequencies from different studies were combined by taking the average frequency, weighted
350 according to the study's sample size. All alleles were mapped to the corresponding G-groups and
351 trimmed at second-field resolution.

352 Correlation between expected and observed allele frequencies

353 For all tools and for each supported data type, the number of times that each allele was called was
354 counted. This count was divided by the total number of samples to obtain the "observed allele
355 frequency". The Pearson correlation was calculated between observed allele frequencies and the
356 allele frequencies that were expected based on the *Allele Frequency Net* database.

357 Concordance of predictions among different tools

358 Per gene, the concordance of the predictions between each pair of tools was assessed by counting
359 the number of allele pair predictions made by the first tool that were also made by the second tool
360 (for the same sample and gene). Samples where one of both tools did not make a prediction were
361 not considered. This analysis was performed on the 1000 genomes and TCGA dataset.

362 Consensus HLA predictions

363 A majority voting rule was used to determine the most likely HLA genotype for each sample. For each
364 gene of interest, we selected the pair of alleles that has been predicted the most frequently for that
365 sample (i.e., outputted by the highest number of tools). When ties occurred (i.e., multiple allele pairs

had equal numbers of predictions), priority was given to the allele pair that was predicted by the tool with the best individual performance for that gene.

Selecting a minimum number of tools to make consensus HLA predictions

The minimal set of tools that must be included in the majority voting scheme to make reliable consensus predictions was determined using an iterative procedure. Initially, two tools were selected for the model: the tool that performed the best in the benchmark on the 1000 genomes data and the one that best complements that tool. The latter tool was defined as the tool that most often made a correct prediction (for both alleles) on the samples that were wrongly predicted by the best performing tool. Tools were added one by one to this initial model. At each step, the consensus predictions were made and evaluated using the gold standard HLA calls. The tool that led to the greatest increase in accuracy was added to the model. This procedure was repeated until all tools were selected.

Hardware and software environment

Analyses were performed on Ubuntu 20.04 on a Dell EMC PowerEdge R940xa server with 4 Intel Xeon Gold 6240 CPUs (2.60 GHz), each with 18 physical CPU cores, and 376 GiB RAM installed.

Data processing and statistical analysis

Data processing and statistical analyses were performed using R (version 4.0).

Supporting information

S1 Fig. Fraction of correct allele predictions (1000 genomes). Radar plots depicting the fraction of correct allele predictions relative to the total number of alleles for which the algorithm was able to make a prediction on the 1000 genomes dataset. Coloured lines represent different genes, as

388 indicated in the legend below the plots. Corners of the radar plots correspond to the tools that were
389 evaluated for that data type. The Meta tools correspond to the 4-tools metaclassifiers.

390 **S2 Fig. Fraction of successful allele predictions (1000 genomes)**. Radar plots depicting the fraction of
391 alleles for which the tool was able to make a prediction on the 1000 genomes dataset. Coloured lines
392 represent different genes, as indicated in the legend below the plots. Corners of the radar plots
393 correspond to the tools that were evaluated for that data type. The Meta tools correspond to the 4-
394 tools metaclassifiers.

395 **S3 Fig. HLA allele prediction accuracies on NCI-60 cell lines**. Radar plots of HLA allele prediction
396 accuracies on data from NCI-60 cell lines. Coloured lines represent different genes, as indicated in the
397 legend below the plots. Corners of the radar plots correspond to the tools that were evaluated for
398 that data type.

399 **S4 Fig. Expected frequency of HLA-DRB1 alleles in an African American population vs frequencies**
400 **predicted by arcasHLA**. Scatter plot that compares the allele frequency as predicted by arcasHLA (x-
401 axis) with the expected allele frequencies based on data from *Allele Frequency Net* (y-axis).

402 **S5 Fig. Concordance of HLA calls between each pair of tools on DNA data (1000 genomes)**.
403 Heatmaps representing the concordance of the HLA calls between each pair of tools, applied on the
404 1000 genomes DNA data. Hierarchical clustering was applied on the tools. The Meta tool corresponds
405 to the 4-tool consensus metaclassifier.

406 **S6 Fig. Concordance of HLA calls between each pair of tools on RNA data (1000 genomes)**.
407 Heatmaps representing the concordance of the HLA calls between each pair of tools, applied on the
408 1000 genomes RNA data. Hierarchical clustering was applied on the tools.

409 **S7 Fig. Concordance of HLA calls between each pair of tools on DNA data (TCGA)**. Heatmaps
410 representing the concordance of the HLA calls between each pair of tools, applied on the TCGA DNA

411 data. Hierarchical clustering was applied on the tools. The Meta tool corresponds to the 4-tool
412 consensus metaclassifier.

413 **S8 Fig. Concordance of HLA calls between each pair of tools on RNA data (TCGA)**. Heatmaps
414 representing the concordance of the HLA calls between each pair of tools, applied on the TCGA RNA
415 data. Hierarchical clustering was applied on the tools.

416 **S9 Fig. Correctness of predictions on DNA data**. Heatmap indicating correctness of predictions on
417 DNA data for each sample (rows) and tool (columns). Hierarchical clustering was applied on tools and
418 samples. Dendrogram for the tools is shown on top of the plots. Dendrogram for the samples is
419 shown right of the plots.

420 **S10 Fig. Correctness of predictions on RNA data**. Heatmap indicating correctness of predictions on
421 RNA data for each sample (rows) and tool (columns). Hierarchical clustering was applied on tools and
422 samples. Dendrogram for the tools is shown on top of the plots. Dendrogram for the samples is
423 shown right of the plots.

424 **S11 Fig. Comparison of accuracies of all-tool metaclassifier with best performing individual tool per**
425 **gene.** Barplots comparing the accuracy of the best tool for each gene and data type to the accuracy
426 of a classifier that chooses an HLA genotype from the output of all tools that support that data type
427 and gene based on a majority voting rule. Bars in a red correspond to the accuracies of the voting
428 classifier. Bars in blue correspond to the accuracies of the best individual tool for that gene.

429 **S1 Table. Motivation for exclusion of 9 tools from the study.** Overview of tools that were not
430 benchmarked in our study and the reason for their exclusion. Excluded tools were either not *freely*
431 *available for academic use*, do not use *FASTQ or BAM input files from WGS, WES and/or RNA-Seq*
432 experiments (e.g., enrichment of the HLA region prior to sequencing is needed) or we were not able
433 to get them *running on Ubuntu 20.04*.

434 **S2 Table. Algorithmic characteristics of the 13 selected tools.** Table that compares the 13 selected
435 HLA genotyping algorithms. Methods differ in the method to align reads to the HLA allele reference
436 sequences (alignment method) and in the subsequent allele prioritization approach (allele pair
437 prioritization). The column *tool-specific steps* describes additional particularities of the tools.

438 **S3 Table. Overview of AFN datasets used for expected allele frequencies.** Overview of populations
439 (studies) from the *Allele Frequency Net database* which were used to compile the list of expected
440 HLA allele frequencies.

441 **S4 Table. Comparison of our results with 7 other independent benchmark studies.** The allele
442 prediction accuracies obtained in our benchmark on the 1000 genomes data are compared with 7
443 other independent benchmark papers. Each score was assigned a colour code: red under 50%, yellow
444 between 50% and 80% and green above 80%. A grey box indicates that a tool was not evaluated for
445 the corresponding gene in that study. The second row of the table indicates the gene or MHC class,
446 where *class I + II* corresponds to the overall (combined) accuracy for MHC class I and class II. For the
447 Lee, Kiyotani, Bauer and Yu studies the tools were evaluated on data from the 1000 genomes project.
448 The studies by Liu, Yi and Chen used a different in-house dataset. DNA data always refers to WES
449 data.

450 Code and Data Availability

451 The code used to produce the results described in this manuscript are available at
452 https://github.com/CCGGlab/mhc_genotyping.

453 Acknowledgments

454 The results shown here are in whole or part based upon data generated by the TCGA Research
455 Network: https://www.cancer.gov/tcga.

456 # References

457 1. Trowsdale J. Genomic structure and function in the MHC. Trends in Genetics. 1993;9: 117–
458 122. doi:10.1016/0168-9525(93)90205-V

459 2. Beck S, Geraghty D, Inoko H, Rowen L, Aguado B, Bahram S, et al. Complete sequence and
460 gene map of a human major histocompatibility complex. Nature 1999 401:6756. 1999;401:
461 921–923. doi:10.1038/44853

462 3. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, et al. Gene map of the
463 extended human MHC. Nature Reviews Genetics 2004 5:12. 2004;5: 889–899.
464 doi:10.1038/nrg1489

465 4. Halenius A, Gerke C, Hengel H. Classical and non-classical MHC I molecule manipulation by
466 human cytomegalovirus: so many targets—but how many arrows in the quiver? Cellular &
467 Molecular Immunology 2015 12:2. 2014;12: 139–153. doi:10.1038/cmi.2014.105

468 5. Allen RL, Hogan L. Non-Classical MHC Class I Molecules (MHC-Ib). eLS. 2013 [cited 18 Mar
469 2022]. doi:10.1002/9780470015902.A0024246

470 6. Hewitt EW. The MHC class I antigen presentation pathway: strategies for viral immune
471 evasion. Immunology. 2003;110: 163. doi:10.1046/J.1365-2567.2003.01738.X

472 7. Axelrod ML, Cook RS, Johnson DB, Balko JM. Biological consequences of MHC-II expression by
473 tumor cells in cancer. Clinical Cancer Research. American Association for Cancer Research Inc.;
474 2019. pp. 2392–2402. doi:10.1158/1078-0432.CCR-18-3200

475 8. Alspach E, Lussier DM, Miceli AP, Kizhvatov I, DuPage M, Luoma AM, et al. MHC-II neoantigens
476 shape tumour immunity and response to immunotherapy. Nature. 2019;574: 696–701.
477 doi:10.1038/s41586-019-1671-8

478 9. Sun Z, Chen F, Meng F, Wei J, Liu B. MHC class II restricted neoantigen: A promising target in
479 tumor immunotherapy. Cancer Letters. 2017;392: 17–25. doi:10.1016/J.CANLET.2016.12.039

480 10. Corthay A, Skovseth DK, Lundin KU, Røsjø E, Omholt H, Hofgaard PO, et al. Primary antitumor
481 immune response mediated by CD4+ T cells. Immunity. 2005;22: 371–383.
482 doi:10.1016/J.IMMUNI.2005.02.003

483 11. Haabeth OAW, Fauskanger M, Manzke M, Lundin KU, Corthay A, Bogen B, et al. CD4+ T-cell–
484 Mediated Rejection of MHC Class II–Positive Tumor Cells Is Dependent on Antigen Secretion
485 and Indirect Presentation on Host APCs. Cancer Research. 2018;78: 4573–4585.
486 doi:10.1158/0008-5472.CAN-17-2426

487 12. Chowell D, Morris LGT, Grigg CM, Weber JK, Samstein RM, Makarov V, et al. Patient HLA class
488 I genotype influences cancer response to checkpoint blockade immunotherapy. Science
489 (1979). 2018;359: 582–587.
490 doi:10.1126/SCIENCE.AAO4572/SUPPL_FILE/AAO4572_CHOWELL_SM-TABLE-S9.XLSX

491 13. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA
492 typing from next-generation sequencing data. Bioinformatics. 2014;30: 3310–3316.
493 doi:10.1093/bioinformatics/btu548

494 14. Bauer DC, Zadoorian A, Wilson LOW, Alliance MGH, Thorne NP. Evaluation of computational
495 programs to predict HLA genotypes from genomic sequencing data. Briefings in
496 Bioinformatics. 2018;19: 179–187. doi:10.1093/BIB/BBW097

497 15. Orenbuch R, Filip I, Comito D, Shaman J, Pe'Er I, Rabadan R. arcasHLA: high-resolution HLA
498 typing from RNAseq. Bioinformatics. 2020;36: 33–40. doi:10.1093/BIOINFORMATICS/BTZ474

499 16. Matey-Hernandez ML, Brunak S, Izarzugaza JMG. Benchmarking the HLA typing performance
500 of Polysolver and Optitype in 50 Danish parental trios. BMC Bioinformatics. 2018;19: 1–12.
501 doi:10.1186/S12859-018-2239-6

502 17. Lee M, Seo JH, Song S, Song IH, Kim SY, Kim YA, et al. A New Human Leukocyte Antigen Typing
503 Algorithm Combined With Currently Available Genotyping Tools Based on Next-Generation
504 Sequencing Data and Guidelines to Select the Most Likely Human Leukocyte Antigen
505 Genotype. Frontiers in Immunology. 2021;12: 4080. doi:10.3389/FIMMU.2021.688183

506 18. Li X, Zhou C, Chen K, Huang B, Liu Q, Ye H. Benchmarking HLA genotyping and clarifying HLA
507 impact on survival in tumor immunotherapy. Molecular Oncology. 2021;15: 1764–1782.
508 doi:10.1002/1878-0261.12895

509 19. Chen J, Madireddi S, Nagarkar D, Migdal M, vander Heiden J, Chang D, et al. In silico tools for
510 accurate HLA and KIR inference from clinical sequencing data empower immunogenetics on
511 individual-patient and population scales. Briefings in Bioinformatics. 2021;22: 1–11.
512 doi:10.1093/BIB/BBAA223

513 20. Yu Y, Wang K, Fahira A, Yang Q, Sun R, Li Z, et al. Systematic comparative study of
514 computational methods for HLA typing from next-generation sequencing. HLA. 2021;97: 481–
515 492. doi:10.1111/TAN.14244

516 21. Kiyotani K, Mai TH, Nakamura Y. Comparison of exome-based HLA class I genotyping tools:
517 identification of platform-specific genotyping errors. Journal of Human Genetics 2017 62:3.
518 2016;62: 397–405. doi:10.1038/jhg.2016.141

519 22. Liu P, Yao M, Gong Y, Song Y, Chen Y, Ye Y, et al. Benchmarking the Human Leukocyte Antigen
520 Typing Performance of Three Assays and Seven Next-Generation Sequencing-Based
521 Algorithms. Frontiers in Immunology. 2021;12: 840. doi:10.3389/FIMMU.2021.652258/BIBTEX

522 23. Yi J, Chen L, Xiao Y, Zhao Z, Su X. Investigations of sequencing data and sample type on HLA
523 class Ia typing with different computational tools. Briefings in Bioinformatics. 2021;22: 1–6.
524 doi:10.1093/BIB/BBAA143

525 24. Kawaguchi S, Higasa K, Shimizu M, Yamada R, Matsuda F. HLA-HD: An accurate HLA typing
526 algorithm for next-generation sequencing data. Human Mutation. 2017;38: 788–797.
527 doi:10.1002/HUMU.23230

528 25. Wang YY, Mimori T, Khor SS, Gervais O, Kawai Y, Hitomi Y, et al. HLA-VBSeq v2: improved HLA
529 calling accuracy with full-length Japanese class-I panel. Human Genome Variation 2019 6:1.
530 2019;6: 1–5. doi:10.1038/s41439-019-0061-y

531 26. Dilthey AT, Mentzer AJ, Carapito R, Cutland C, Cereb N, Madhi SA, et al. HLA*LA—HLA typing
532 from linearly projected graph alignments. Bioinformatics. 2019;35: 4394–4396.
533 doi:10.1093/BIOINFORMATICS/BTZ235

534 27. Kim HJ, Pourmand N. HLA Haplotyping from RNA-seq Data Using Hierarchical Read Weighting.
535 PLOS ONE. 2013;8: e67885. doi:10.1371/JOURNAL.PONE.0067885

536 28. Warren RL, Choe G, Freeman DJ, Castellarin M, Munro S, Moore R, et al. Derivation of HLA
537 types from shotgun sequence datasets. Genome Medicine. 2012;4: 1–8. doi:10.1186/GM396

538 29. Ka S, Lee S, Hong J, Cho Y, Sung J, Kim HN, et al. HLAscan: Genotyping of the HLA region using
539 next-generation sequencing data. BMC Bioinformatics. 2017;18: 1–11. doi:10.1186/S12859-
540 017-1671-3

541 30. Lee H, Kingsford C. Kourami: Graph-guided assembly for novel human leukocyte antigen allele
542 discovery. Genome Biology. 2018;19: 1–16. doi:10.1186/S13059-018-1388-2

543 31. Bai Y, Ni M, Cooper B, Wei Y, Fury W. Inference of high resolution HLA types using genome-
544 wide RNA or DNA sequencing reads. BMC Genomics. 2014;15: 1–16. doi:10.1186/1471-2164-
545 15-325/FIGURES/3

546 32. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of
547 tumors associated with local immune cytolytic activity. Cell. 2015;160: 48–61.
548 doi:10.1016/j.cell.2014.12.033

549 33. Boegel S, Löwer M, Schäfer M, Bukur T, de Graaf J, Boisguérin V, et al. HLA typing from RNA-
550 Seq sequence reads. Genome Medicine. 2012;4: 1–12. doi:10.1186/GM403

551 34. Xie C, Yeo ZX, Wong M, Piper J, Long T, Kirkness EF, et al. Fast and accurate HLA typing from
552 short-read next-generation sequence data with xHLA. Proceedings of the National Academy of
553 Sciences. 2017;114: 8059–8064. doi:10.1073/PNAS.1707945114

554 35. Hayashi S, Moriyama T, Yamaguchi R, Mizuno S, Komura M, Miyano S, et al. ALPHLARD-NT:
555 Bayesian Method for Human Leukocyte Antigen Genotyping and Mutation Calling through
556 Simultaneous Analysis of Normal and Tumor Whole-Genome Sequence Data. Journal of
557 Computational Biology. 2019;26: 923–937. doi:10.1089/cmb.2018.0224

558 36. Liu C, Yang X, Duffy B, Mohanakumar T, Mitra RD, Zody MC, et al. ATHLATES: accurate typing
559 of human leukocyte antigen through exome sequencing. Nucleic Acids Research. 2013;41:
560 e142–e142. doi:10.1093/NAR/GKT481

561 37. Buchkovich ML, Brown CC, Robasky K, Chai S, Westfall S, Vincent BG, et al. HLAProfiler utilizes
562 k-mer profiles to improve HLA calling accuracy for rare and common alleles in RNA-seq data.
563 Genome Medicine. 2017;9: 1–15. doi:10.1186/S13073-017-0473-6

564 38. Huang Y, Yang J, Ying D, Zhang Y, Shotelersuk V, Hirankarn N, et al. HLAreporter: A tool for
565 HLA typing from next generation sequencing data. Genome Medicine. 2015;7: 1–12.
566 doi:10.1186/S13073-015-0145-3

567 39. Wittig M, Anmarkrud JA, Kässens JC, Koch S, Forster M, Ellinghaus E, et al. Development of a
568 high-resolution NGS-based HLA-typing and analysis pipeline. Nucleic Acids Research. 2015;43:
569 e70. doi:10.1093/NAR/GKV184

570 40. Sverchkova A, Anzar I, Stratford R, Clancy T. Improved HLA typing of Class I and Class II alleles
571 from next-generation sequencing data. HLA. 2019;94: 504–513. doi:10.1111/TAN.13685

572 41. Abi-Rached L, Gouret P, Yeh JH, Cristofaro J di, Pontarotti P, Picard C, et al. Immune diversity
573 sheds light on missing variation in worldwide genetic diversity panels. PLOS ONE. 2018;13:
574 e0206512. doi:10.1371/JOURNAL.PONE.0206512

575 42. Jia X, Han B, Onengut-Gumuscu S, Chen WM, Concannon PJ, Rich SS, et al. Imputing Amino
576 Acid Polymorphisms in Human Leukocyte Antigens. PLOS ONE. 2013;8: e64683.
577 doi:10.1371/JOURNAL.PONE.0064683

578 43. Cao H, Wu J, Wang Y, Jiang H, Zhang T, Liu X, et al. An Integrated Tool to Study MHC Region:
579 Accurate SNV Detection and HLA Genes Typing in Human MHC Region Using Targeted High-
580 Throughput Sequencing. PLoS ONE. 2013;8. doi:10.1371/JOURNAL.PONE.0069388

581 44. Zheng-Bradley X, Streeter I, Fairley S, Richardson D, Clarke L, Flicek P, et al. Alignment of 1000
582 Genomes Project reads to reference assembly GRCh38. Gigascience. 2017;6: 1.
583 doi:10.1093/GIGASCIENCE/GIX038

584 45. Lappalainen T, Sammeth M, Friedländer MR, ’T Hoen PAC, Monlong J, Rivas MA, et al.
585 Transcriptome and genome sequencing uncovers functional variation in humans. Nature 2013
586 501:7468. 2013;501: 506–511. doi:10.1038/nature12531

587 46. Gonzalez-Galarza FF, McCabe A, Santos EJM dos, Jones J, Takeshita L, Ortega-Rivera ND, et al.
588 Allele frequency net database (AFND) 2020 update: gold-standard data classification, open
589 access genotype data and new query tools. Nucleic Acids Res. 2020;48: D783–D788.
590 doi:10.1093/NAR/GKZ1029

591 47. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome
592 landscapes. Science (1979). 2013;340: 1546–1558. doi:10.1126/science.1235122

593 48. Claeys A, Luijts T, Marchal K, van den Eynden J. Low immunogenicity of common cancer hot
594 spot mutations resulting in false immunogenic selection signals. Thomas PG, editor. PLOS
595 Genetics. 2021;17: e1009368. doi:10.1371/journal.pgen.1009368

596 49. van den Eynden J, Jiménez-Sánchez A, Miller ML, Larsson E. Lack of detectable neoantigen
597 depletion signals in the untreated cancer genome. Nature Genetics. 2019;51: 1741–1748.
598 doi:10.1038/s41588-019-0532-6

599 50. Riaz N, Havel JJ, Makarov V, Desrichard A, Urba WJ, Sims JS, et al. Tumor and
600 Microenvironment Evolution during Immunotherapy with Nivolumab. Cell. 2017;171: 934-
601 949.e16. doi:10.1016/J.CELL.2017.09.028

602 51. Liu D, Schilling B, Liu D, Sucker A, Livingstone E, Jerby-Amon L, et al. Integrative molecular and
603 clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma.
604 Nature Medicine 2019 25:12. 2019;25: 1916–1927. doi:10.1038/s41591-019-0654-5

605 52. Naranbhai V, Viard M, Dean M, Groha S, Braun DA, Labaki C, et al. HLA-A*03 and response to
606 immune checkpoint blockade in cancer: an epidemiological biomarker study. Lancet Oncol.
607 2022;23: 172–184. doi:10.1016/S1470-2045(21)00582-9

608 53. Adams SD, Barracchini KC, Chen D, Robbins F, Wang L, Larsen P, et al. Ambiguous allele
609 combinations in HLA Class I and Class II sequence-based typing: when precise nucleotide
610 sequencing leads to imprecise allele identification. 2004 [cited 29 Oct 2021].
611 doi:10.1186/1479-5876-2-30

612 54. Bauer-Mehren A, Bundschus M, Rautschka M, Mayer MA, Sanz F, Furlong LI. Gene-disease
613 network analysis reveals functional modules in mendelian, complex and environmental
614 diseases. PLoS ONE. 2011;6: e20284. doi:10.1371/journal.pone.0020284

615 55. Abaan OD, Polley EC, Davis SR, Zhu YJ, Bilke S, Walker RL, et al. The exomes of the NCI-60
616 panel: a genomic resource for cancer biology and systems pharmacology. Cancer Res.
617 2013;73: 4372–4382. doi:10.1158/0008-5472.CAN-12-3342

618 56. Reinhold WC, Varma S, Sunshine M, Elloumi F, Ofori-Atta K, Lee S, et al. RNA Sequencing of
619 the NCI-60: Integration into CellMiner and CellMiner CDB. Cancer Res. 2019;79: 3514–3524.
620 doi:10.1158/0008-5472.CAN-18-2047

621 57. Gourraud PA, Khankhanian P, Cereb N, Yang SY, Feolo M, Maiers M, et al. HLA Diversity in the
622 1000 Genomes Dataset. PLOS ONE. 2014;9: e97282. doi:10.1371/JOURNAL.PONE.0097282

623 58. Adams S, Robbins FM, Chen D, Wagage D, Holbeck SL, Morse HC, et al. HLA class I and II
624 genotype of the NCI-60 cell lines. Journal of Translational Medicine. 2005;3: 11.
625 doi:10.1186/1479-5876-3-11

626