Human Immunology 79 (2018) 678–684

Contents lists available at ScienceDirect

Human Immunology

journal homepage: www.elsevier.com/locate/humimm

Hla-mapper: An application to optimize the mapping of HLA sequences produced by massively parallel sequencing procedures

Erick C. Castelli[a,b,*], Michelle A. Paz[a], Andréia S. Souza[a], Jaqueline Ramalho[a], Celso Teixeira Mendes-Junior[c]

[a] São Paulo State University (UNESP), Molecular Genetics and Bioinformatics Laboratory, Experimental Research Unit (UNIPEX), School of Medicine, Botucatu, State of São Paulo, Brazil
[b] São Paulo State University (UNESP), Pathology Department, School of Medicine, Botucatu, State of São Paulo, Brazil
[c] Departamento de Química, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, 14040-901 Ribeirão Preto, SP, Brazil

ARTICLE INFO

Keywords:
MHC
HLA
Next Generation Sequencing (NGS)
Second Generation Sequencing
Variability
Polymorphisms
Typing
Aligners
Mapping tool

ABSTRACT

A challenging task when more than one HLA gene is evaluated together by second-generation sequencing is to achieve a reliable read mapping. The polymorphic and repetitive nature of HLA genes might bias the read mapping process, usually underestimating variability at very polymorphic segments, or overestimating variability at some segments. To overcome this issue we developed *hla-mapper*, which takes into account HLA sequences derived from the IPD-IMGT/HLA database and unpublished HLA sequences to apply a scoring system. This comprehends the evaluation of each read pair, addressing them to the most likely HLA gene they were derived from. *Hla-mapper* provides a reliable map of HLA sequences, allowing accurate downstream analysis such as variant calling, haplotype inference, and allele typing. Moreover, *hla-mapper* supports whole genome, exome, and targeted sequencing data. To assess the software performance in comparison with traditional mapping algorithms, we used three different simulated datasets to compare the results obtained with *hla-mapper*, BWA MEM, and Bowtie2. Overall, *hla-mapper* presented a superior performance, mainly for the classical HLA class I genes, minimizing wrong mapping and cross-mapping that are typically observed when using BWA MEM or Bowtie2 with a single reference genome.

1. Introduction

The human leukocyte antigen (HLA) complex comprehends the most variable segment of the human genome. This complex plays a central role in the immune response since HLA molecules are key features for antigen presentation and immune response modulation [1]. HLA genes compatibility between recipient and donor influences graft acceptance, therefore, HLA variability is frequently evaluated for clinical purposes. Moreover, different HLA variants might be related to differential immune responses against pathogens, susceptibility to autoimmune diseases, or even to specific tumors [2]. The classical HLA class I genes, *HLA-A*, *HLA-B*, and *HLA-C*, are highly polymorphic loci and encode a key molecule for antigen presentation to T CD8+ lymphocytes. The non-classical HLA class I genes, *HLA-G*, *HLA-E*, and *HLA-F*, are conserved at the DNA and protein level and encode immunemodulatory molecules [3–5].

Much effort has been made to characterize HLA polymorphisms and complete sequences in worldwide populations. Most of the studies are restricted to the coding region, or even to exon segments only. However, with the advent of next-generation sequencing (NGS), or massively parallel sequencing, information regarding introns and regulatory segments has been obtained, as well as the characterization of many new allele variants, providing deeper insights regarding HLA worldwide variability and extended haplotypes. This is particularly evident for non-classical HLA class I genes [6–8]. In addition, NGS strategies do allow the evaluation of several genes all at once, and also allow the phase definition (haplotypes) among part of the variants detected. Nevertheless, when two or more HLA genes are sequenced at the same time using NGS methods, a challenging task is to achieve a reliable read mapping. Because of the polymorphic and repetitive nature of most of the HLA genes due to their paralogous origins, the following issues may arise when raw data (the reads) are mapped to a single reference genome.

First, the classical HLA class I genes are among the most variable genes in the human genome. The International Immunogenetics Database (IPD-IMGT/HLA database, version 3.31.0) describes more

* Corresponding author at: São Paulo State University (UNESP), Pathology Department, School of Medicine, Botucatu, State of São Paulo, CEP 18618-687, Brazil.
E-mail address: erick.castelli@unesp.br (E.C. Castelli).

https://doi.org/10.1016/j.humimm.2018.06.010
Received 23 November 2017; Received in revised form 15 June 2018; Accepted 29 June 2018
Available online 03 July 2018

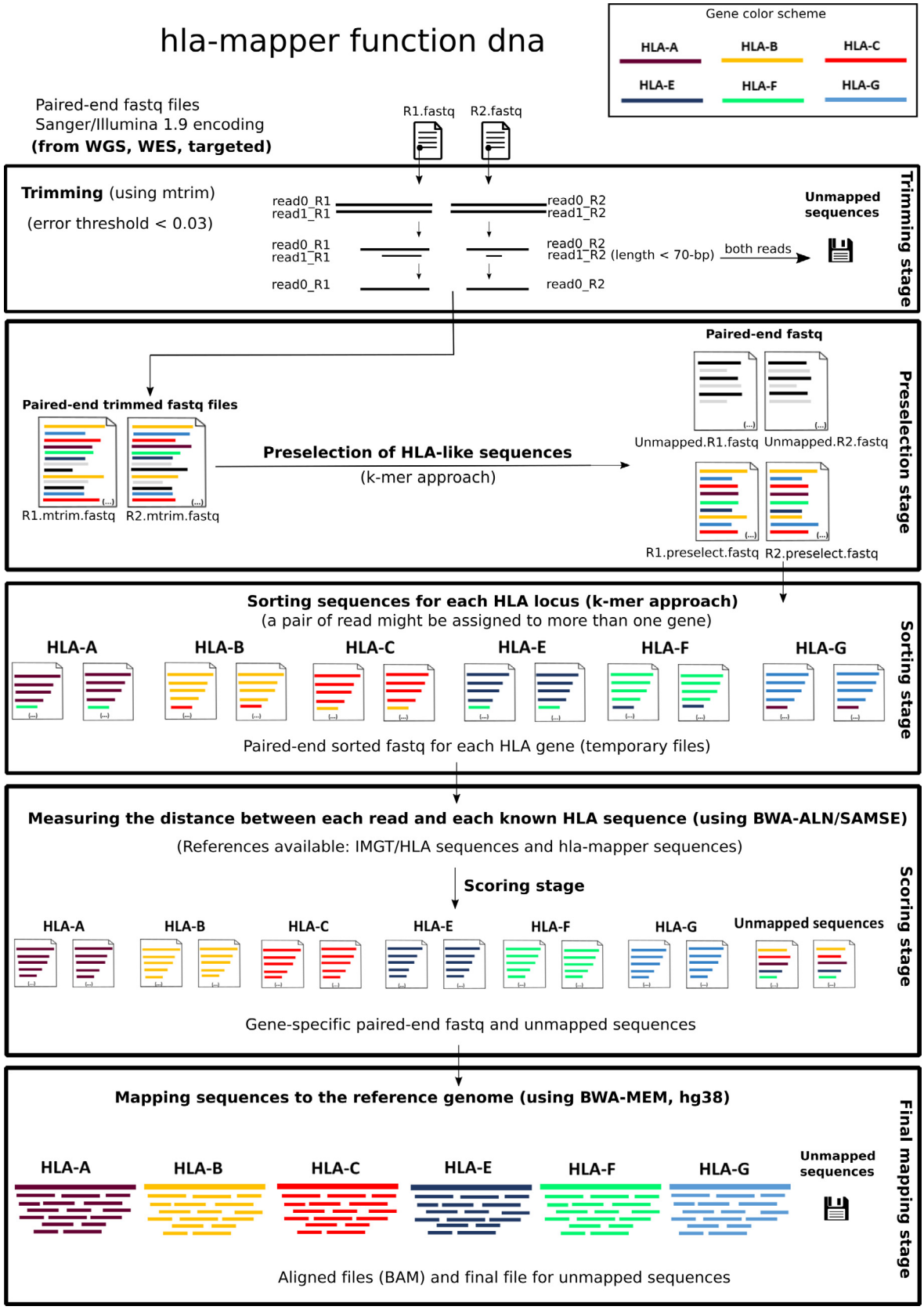

Fig. 1. *Hla-mapper* software workflow, from the input FASTQ to the outputted BAM files.

than 12,893 class I alleles [9]. Because of the polymorphic nature of classical HLA genes, the sequences (or reads) obtained by NGS methods usually present too many nucleotide differences when compared with the reference human genome. Thus, well-established aligners such as BWA (the Burrows-Wheeler Aligner) and Bowtie2 [10,11] frequently do not map classical HLA class I sequences correctly. Usually, when these algorithms are used with default parameters, because of the high polymorphism, many reads do not find a match in the reference genome, leading to a mapping bias that underestimates HLA variability and overestimates reference allele frequencies [12–14]. Moreover, when a higher tolerance for mismatches is defined, many reads are incorrectly mapped as explained below.

Second, mostly due to the high sequence similarity among HLA genes, polymorphisms may lead a sequence to be more compatible with the reference sequence of another HLA gene than with the reference of the original gene. Thus, when the tolerance for mismatches is increased to avoid the aforementioned mapping bias, this second issue leads to a large number of reads mapping to more than one HLA gene into the reference genome, or simply mapping to the wrong gene. In this scenario, it is expected an overestimation of genetic diversity in segments presenting a large number of incorrectly mapped reads. Third, depending on the NGS method used, a large number of very short reads is produced and they could exacerbate the issues already presented.

The BWA developers do acknowledge the issues presented above. They created the bwa.kit to improve HLA read mapping based on alternative contigs and known HLA sequences from the IPD-IMGT/HLA. However, its use is not straightforward and the data obtained might bias downstream genotyping procedures as discussed later. In addition, other attempts to evaluate the efficiency of many mappers and variant call methods in complex regions have been made, but mapping improvement depends on the mapper and variant caller combination [15,16].

All the issues introduced above could be circumvented if the reads of each HLA gene were tagged with different indices, but this strategy is cost ineffective. Usually, the aim is to sequence all HLA genes from a given individual in a single sequencing run, tagging each individual with specific indices, thus allowing the evaluation of many samples at the same time. Additionally, research groups could be interested in other non-HLA genes that may be sequenced together with HLA, or even in evaluating HLA from whole genome sequencing.

Although many companies have introduced different NGS HLA-typing kits and specific applications to call HLA alleles, these products are mainly focused on reporting pre-defined alleles from the IPD-IMGT/HLA database, for clinical purposes. In this context, they are usually restricted to the segments tracked by the aforementioned database, which does not include the complete upstream regulatory regions and the complete 3′ untranslated segments. In addition, these applications usually produce good results when the HLA typing kits from the same manufacturer are used. However, they may not be suitable when alternative approaches to characterize HLA genes are applied (e.g., when other non-HLA genes are included in the sequencing). These applications are usually not freely available.

Many publicly available typing tools have recently been developed to call HLA alleles from NGS data [14,17], including HLAminer [18], HLA-VBseq [19], HLAreporter [20], OptiType [21], and others. In general, these tools report only the HLA typing and do not properly handle new HLA alleles. Moreover, they usually do not output aligned (BAM/SAM) files that can be further processed (for instance, to infer SNP genotypes).

To achieve a better evaluation of HLA genes when several HLA genes are sequenced together or when other genes outside the HLA complex are also included, we developed *hla-mapper* to optimize read mapping for HLA class I genes. *hla-mapper* uses a scoring system in order to address each read pair to the most likely gene, providing a reliable map of those sequences, as described in the methods section. Here we present the *hla-mapper* application, comparing its mapping accuracy with other aligners such as BWA and Bowtie2.

2. Methods

2.1. The hla-mapper software

The input for *hla-mapper* is composed of paired-end FASTQ files from amplicon, exon, or whole-genome sequencing. The software has a trimming algorithm to remove short reads (the default value is 70 nucleotides) and low-quality segments, identifying the largest sequence fragment for each read in which all nucleotides present a quality value higher than 97% ($Q \geq 15$). This process assures that only high-quality sequences pass forward to the scoring stage (Fig. 1).

After the trimming process, *hla-mapper* identifies all read pairs presenting sequence resemblance with any class I HLA gene sequence, by using a k-mer approach. In this step, all possible 15-mer motifs are computed, considering every HLA sequence available in the *hla-mapper* database. Subsequently, every read pair with both sequences presenting at least one of these motifs is considered as a possible HLA sequence. This approach is particularly useful when dealing with whole genome, whole exome sequencing, or when other non-HLA sequences are present. This feature may be turned off when the input includes only HLA class I gene sequences. After the initial read pair selection, *hla-mapper* sorts the preselected pairs into subgroups according to the similarities with each HLA class I gene. At this point, a read pair might be preselected for more than one gene (Fig. 1).

At the scoring stage, *hla-mapper* uses the BWA ALN/SAMSE algorithms [10] to align each read against a database of known HLA sequences from IPD-IMGT/HLA [9] and from curated HLA sequences provided by the hla-mapper database. The parameters used for BWA ALN are “maximum number or fraction of gap opens = 2” and the “max #diff (int) or missing prob under 0.02 err rate = 10”. The BWA SAMSE algorithm is used with default parameters. For each read, *hla-mapper* measures the number of mismatches between this read and each HLA class I sequence in the database, registering the lowest number observed and with which reference it was associated. A mismatch score is calculated for each read by summing the number of soft-clipped bases (calculated from the CIGAR string) and the number of mismatches at the aligned segment (calculated by using the NM:i field at the SAM file). The algorithm allows a maximum of 6 mismatches for any read. Reads not attending this criterion are excluded. Then, for each read pair, a divergence score is calculated for each class I gene by summing the mismatch score of the forward and reverse sequences observed for each gene. If any sequence of a read pair failed to align against a given gene reference, this pair is no longer considered for that specific gene. When a given read pair presents the same divergence score for more than one gene, the pair is excluded by default and placed within the unmapped sequence files. This feature may be turned off, forcing these pairs to map with MQ = 0. After scoring each read pair, they are assigned to the HLA class I gene presenting the lowest divergence score. Then, gene-specific paired-end FASTQ files are created (Fig. 1). All read pairs not assigned to a class I gene are placed at FASTQ files containing the unmapped sequences.

For some HLA alleles, the IPD-IMGT/HLA database sequence comprehends the segment between the proximal promoter (around position -300) up to a partially characterized 3′ untranslated region (3′UTR). However, there is no information regarding complete promoters, introns, or 3′UTR segments for most of the known alleles. This is particularly evident for *HLA-A*, *HLA-B,* and *HLA-C*. Nonetheless, this database comprehends known HLA sequences that were properly cloned and confirmed by Sanger sequencing.

To circumvent the lack of reference sequences regarding the 5′ upstream, 3′UTR, and intron segments, the *hla-mapper* database presents curated and manually analyzed sequences that have not been described at the IPD-IMGT/HLA database (mainly for the regulatory segments) or that have partially been characterized at the

aforementioned database (a list of references regarding these sequences is available in the software manual). In addition, the user may also add locally known HLA sequences to improve the *hla-mapper* scoring system.

After the scoring stage, each gene-specific paired-end FASTQ file is then mapped using the hg19 or hg38 sequences as references, with a less restrictive approach and allowing a higher mismatch rate. At this point, *hla-mapper* uses the BWA MEM algorithm [10] to map, for instance, *HLA-A* specific FASTQ files against an adapted version of the reference genome that only includes the *HLA-A* sequence as a reference. This procedure is repeated for each class I gene. The BWA MEM algorithm is used with default parameters, except for the "penalty for a mismatch = 2", minimizing mapping failure of highly divergent sequences. The *hla-mapper* BAM files are then corrected regarding mapping positions, making them compatible with the GRCh37 (hg19) or GRCh38 (hg38) genome references (the default is hg38). Finally, gene-specific BAM files are created. These BAM files can be further processed to infer genotypes and haplotypes following any suitable method for the user, such as the GATK package (Genome Analysis Toolkit) [22,23].

All steps described above are automatically performed by a single program and using a single command. *hla-mapper* supports HLA class I genes (*HLA-A*, *HLA-B*, *HLA-C*, *HLA-E*, *HLA-F,* and *HLA-G*) and it is compatible with Linux and macOS. *hla-mapper* is freely available for download at www.castelli-lab.net/apps/hla-mapper.

2.2. Assessing hla-mapper performance by comparison with other mapping algorithms

To assess the *hla-mapper* performance (version 2.0, with database version 2.1), we compared its results with the ones obtained using two different well-established aligners, BWA MEM (version 0.7.16a) and Bowtie2 (version 2.3.3). BWA MEM and Bowtie2 were used with default parameters. We compared the performance of these three methods using three different datasets, each one simulating a targeted sequencing of 1000 virtual samples.

In order to simulate the targeted sequencing, for each virtual sample, we selected two allele sequences of each HLA class I gene (*HLA-A*, *HLA-B*, *HLA-C*, *HLA-E*, *HLA-F,* and *HLA-G*). These allele sequences were retrieved from the IPD-IMGT/HLA database version 3.31 based on the frequencies already reported in the Brazilian population. Brazilians were selected because (a) *HLA-E, HLA-F,* and *HLA-G* complete variability has already been evaluated for the same sample [6–8]; (b) *HLA-A*, *HLA-B,* and *HLA-C* complete variability has already been evaluated (data not published) for this same sample; (c) Brazilians are highly admixed and a great number of different alleles is usually observed. Consequently, we used this data to obtain allele frequencies from each HLA class I gene in a real population sample. The algorithm for allele sequence selection considered the reported frequency of each allele in a way that, when considering the entire dataset, all alleles presented a frequency that resembles the ones observed in a real population sample.

For each sample, a set of 700 random fragments of approximately 600 nucleotides was produced for each allele sequence. Each fragment was transformed into two reads, a forward (the 5′ end of the fragment) and a reverse (the 3′ end of the fragment, complemented and inverted), to simulate 2 × 150 or 2 × 250 paired-end sequencing. Additionally, we inserted random mutations to simulate sequencing errors, with a maximum of 5 mutations per read. To track the allele sequence used to produce each read pair, reads were identified according to the name of the allele used to generate them.

As previously stated, we used 3 different datasets to evaluate the *hla-mapper* performance. The first dataset (dataset 1) consisted of 1000 virtual samples with *HLA-A*, *HLA-B*, *HLA-C*, *HLA-E*, *HLA-F,* and *HLA-G* sequences, simulating a 2 × 250b sequencing procedure. The allele frequencies resemble the ones reported for a Brazilian population sample. The second dataset (dataset 2) also consisted of 1000 virtual samples with *HLA-A*, *HLA-B*, *HLA-C*, *HLA-E*, *HLA-F,* and *HLA-G* sequences, simulating 2 × 250b sequencing. However, the HLA alleles were mutated in order to generate new sequences, by introducing up to 5 random mutations for each original allele sequence. These mutations consisted of nucleotide exchanges or single nucleotide deletions. Thus, this dataset simulates the sequencing of samples with only new HLA alleles, besides the sequencing errors introduced for each read. The third dataset (dataset 3) also consisted of 1000 virtual samples, but, in order to simulate a more complex mapping task, we added four HLA class I pseudogenes (*HLA-H*, *HLA-J*, *HLA-K,* and *HLA-L*). The read sizes were reduced to 150b, simulating a 2 × 150b sequencing. All datasets were further processed by *hla-mapper*, and also by BWA MEM and Bowtie2 using chromosome 6 from hg38 as a reference.

Mapping performance was evaluated by five metrics: (a) *apparent mapping failure*, indicating the rate of reads that failed to map or were mapped more than once; (b) *real mapping failure*, indicating the rate of reads that failed to map or reads that were mapped to the wrong gene; (c) *gene-specific unmapped sequences*, indicating the rate of reads from a specific gene that were not mapped at all; (d) *gene-specific mapping failure*, indicating the rate of reads from a given gene that failed to map to the right gene; and (e) *pairwise cross-mapping rates*, indicating the rate of reads from one gene mapped to another.

2.3. Assessing HLA typing tools performance when data is pre-processed with hla-mapper

In order to evaluate whether pre-processing data with *hla-mapper* would improve the performance of HLA typing, we have tested the following tools using dataset 3: HLAminer [18], HLA-VBseq [19], OptiType [21], with updated databases (IPD-IMGT/HLA version 3.31), and the trial version of NGSEngine version 2.8.1 (from GenDX) using the complete class I database (that includes the non-classical loci, IMGT/HLA version 3.29). All these typing tools were used with default parameters. We compared the results when these tools were used directly (the raw data as input) and when the *hla-mapper* FASTQ files (pre-processed with *hla-mapper*) were used as input.

3. Results

3.1. Hla-mapper performance at datasets 1 and 2

hla-mapper performance was assessed in three datasets, as described earlier. The results for dataset 1 and 2 (that do not include HLA pseudogenes) are shown in Tables 1 and 2. Here we assessed the *hla-mapper* performance when dealing with simulated samples that resemble an actual population sample with known alleles (dataset 1, Table 1) and when dealing with only new HLA sequences (dataset 2, Table 2). In these datasets, there were 8.4 million read pairs (2 alleles × 700 fragments × 6 genes × 1000 samples).

Hla-mapper presented the highest rate of the *apparent mapping failure* metric, in which 0.1% of the reads failed to map for both datasets. When the *real mapping failure* metric is taken into account, we observed higher values when compared with the *apparent mapping failure*. This is related to the presence of many reads mapped to the wrong gene. In this scenario, for both datasets, higher rates were detected for Bowtie2, followed by BWA MEM. Thus, although these two algorithms present a low proportion of reads that fail to map, they present a higher mapping error rate. The best performance for this metric was achieved by *hla-mapper*, since almost all reads were successfully mapped to the right gene.

When the proportion of *gene-specific unmapped sequences* is evaluated, BWA MEM and Bowtie2 usually do not fail to map sequences (not necessarily in a correct fashion, as revealed by the next metric), while the proportion of unmapped reads using *hla-mapper* is slightly higher (around 0.1%) for both datasets (Tables 1 and 2).

Regarding mapping accuracy of individual reads (i.e., the *gene-specific mapping failure* metric), it can be noticed that mapping accuracy

Table 1
Rates, in percentage, for apparent mapping failure, real mapping failure, gene-specific unmapped sequences, gene-specific mapping failure and pairwise cross-mappings, for dataset 1.

Parameters	BWA MEM	Bowtie2	*Hla-mapper*
Apparent mapping failure	0.0001%	0.0142%	0.1031%
Real mapping failure	1.0077%	2.0769%	0.2027%
Gene-specific unmapped sequences			
HLA-A	0.0003%	0.0198%	0.1005%
HLA-B	0.0001%	0.0388%	0.1030%
HLA-C	0.0001%	0.0249%	0.1029%
HLA-E	0.0001%	0.0003%	0.1075%
HLA-F	0.0000%	0.0004%	0.1033%
HLA-G	0.0000%	0.0011%	0.1017%
Gene-specific mapping failure			
HLA-A	3.4591%	4.4307%	0.1788%
HLA-B	0.8176%	2.8071%	0.2064%
HLA-C	1.7678%	3.9884%	0.2057%
HLA-E	0.0006%	0.1980%	0.2151%
HLA-F	0.0003%	0.1877%	0.2066%
HLA-G	0.0006%	0.8494%	0.2034%
Pairwise cross-mapping rates[*]			
HLA-A to HLA-B	0.0006%	0.0912%	0.0000%
HLA-A to HLA-H	3.4774%	4.0757%	–
HLA-A to HLA-J	0.0021%	0.0781%	–
HLA-A to HLA-K	0.0000%	0.0610%	–
HLA-B to HLA-C	0.8153%	2.6557%	0.0003%
HLA-B to HLA-H	0.0022%	0.0790%	–
HLA-C to HLA-B	0.0011%	3.8652%	0.0001%
HLA-E to HLA-A	0.0000%	0.0619%	0.0001%
HLA-G to HLA-A	0.0000%	0.1165%	0.0000%
HLA-G to HLA-H	0.0001%	0.0557%	–
HLA-G to HLA-J	0.0000%	0.5580%	–

– *Hla-mapper* does not align sequences against these genes. Instead, sequences from these genes are placed into FASTQ files containing unmapped sequences.
* Pairs of genes presenting rates higher than 0.05% for at least one algorithm.

Table 2
Rates, in percentage, for apparent mapping failure, real mapping failure, gene-specific unmapped sequences, gene-specific mapping failure and pairwise cross-mapping, for dataset 2.

Parameters	BWA MEM	Bowtie2	*Hla-mapper*
Apparent mapping failure	0.0002%	0.0142%	0.1033%
Real mapping failure	0.9815%	2.0512%	0.2032%
Gene-specific unmapped sequences			
HLA-A	0.0003%	0.0171%	0.1060%
HLA-B	0.0004%	0.0389%	0.1026%
HLA-C	0.0003%	0.0259%	0.1048%
HLA-E	0.0001%	0.0008%	0.1040%
HLA-F	0.0000%	0.0008%	0.1036%
HLA-G	0.0000%	0.0018%	0.0991%
Gene-specific mapping failure			
HLA-A	3.3280%	4.2824%	0.1910%
HLA-B	0.8048%	2.7889%	0.2052%
HLA-C	1.7555%	3.9808%	0.2092%
HLA-E	0.0004%	0.2038%	0.2081%
HLA-F	0.0000%	0.1859%	0.2072%
HLA-G	0.0004%	0.8651%	0.1983%
Pairwise cross-mapping rates[*]			
HLA-A to HLA-B	0.0003%	0.0958%	0.0000%
HLA-A to HLA-H	3.3460%	3.9308%	–
HLA-A to HLA-J	0.0018%	0.0770%	–
HLA-A to HLA-K	0.0004%	0.0600%	–
HLA-B to HLA-C	0.8032%	2.6478%	0.0001%
HLA-B to HLA-H	0.0022%	0.0771%	–
HLA-C to HLA-B	1.7536%	3.8601%	0.0000%
HLA-E to HLA-A	0.0001%	0.0636%	0.0001%
HLA-E to HLA-B	0.0000%	0.0528%	0.0000%
HLA-E to HLA-H	0.0000%	0.0468%	–
HLA-F to HLA-A	0.0000%	0.0508%	0.0000%
HLA-G to HLA-A	0.0001%	0.1159%	0.0000%
HLA-G to HLA-H	0.0000%	0.0599%	–
HLA-G to HLA-J	0.0000%	0.5719%	–
HLA-G to HLA-K	0.0000%	0.0435%	–

– *Hla-mapper* does not align sequences against these genes. Instead, sequences from these genes are placed into FASTQ files containing unmapped sequences.
* Pairs of genes presenting rates higher than 0.05% for at least one algorithm.

varies according to the method used and the HLA class I gene. The highest error rate was observed for classical genes. The worst performance was observed for Bowtie2 for classical class I genes. Overall, the best performance was achieved by *hla-mapper* when considering the classical class I genes individually, or all genes together. Although dataset 2 comprehends only new HLA sequences (that are not present at the *hla-mapper* database), *hla-mapper* performance for dataset 2 is practically equal to dataset 1. Regarding non-classical HLA class I sequences, all methods reach good results in this metric, probably due to the lower extent of sequence diversity of these genes.

When the *pairwise cross-mapping rate* is considered, two major issues emerge. The first one is the cross-mapping between *HLA-A* and *HLA-H*, in which many *HLA-A* reads (around 4%) are addressed to *HLA-H* when using BWA MEM or Bowtie2 (Tables 1 and 2). The second one is the cross-mapping between *HLA-B* and *HLA-C*, in which a high proportion of reads from one gene is mapped to the other when using BWA MEM or Bowtie2. Both issues are circumvented when using *hla-mapper*.

3.2. Hla-mapper performance at dataset 3

The results for dataset 3 (that include HLA pseudogenes and smaller read sizes) are shown in Table 3. For this dataset, there are 14 million read pairs (2 alleles × 700 fragments × 10 genes × 1000 samples). Here we opted to not compute the *apparent mapping failure* because *hla-mapper* would not map at least 40% of the reads as these are derived from HLA pseudogenes. The *real mapping failure* rates were similar to the ones observed for datasets 1 and 2, with Bowtie2 presenting the highest rate (worst performance), followed by BWA MEM. In general, a higher *gene-specific unmapped sequences* metric is observed for Bowtie2 considering classical genes, and for *hla-mapper* considering non-classical ones. When *gene-specific mapping failure* is taken into account, Bowtie2 presented the worst performance, followed by BWA MEM. Overall, the best performance was achieved by *hla-mapper*, in which reads from all genes presented a mapping failure below 0.63%.

Cross-mappings involving *HLA-A* and *HLA-H* were observed when using any of the algorithms except *hla-mapper*. Since dataset 3 presents HLA pseudogene sequences, it can be noticed that a high proportion of *HLA-H* reads also map to *HLA-A* when using BWA MEM or Bowtie2, mainly the last one. Finally, as dataset 3 presents smaller read sizes (150b), a higher rate of cross-mapping between *HLA-B* and *HLA-C* can be noticed for all algorithms, except for *hla-mapper*.

3.3. Typing tools performance after hla-mapper pre-processing

We used dataset 3 (with pseudogenes and smaller read sizes) to evaluate the HLA typing performance of the following tools: HLAminer, HLA-VBseq, OptiType, and NGSEngine from GenDX. For HLAminer, the overall accuracy (both alleles called correctly) was increased from 21.8% (no pre-processing) to 49.5% (after *hla-mapper*). It should be mentioned that HLAminer calls only 2-digit alleles (different HLA proteins). For HLA-VBseq and OptiType, the same results were obtained with and without *hla-mapper* pre-processing, with an overall accuracy of 98% at the 4-digit level. This high OptiType accuracy is compatible with previous reports [14].

For NGSEngine, the overall accuracy increased from 45.1% to 97.3% after *hla-mapper* pre-processing. However, it should be mentioned that the majority of the correct calls before *hla-mapper* pre-

Table 3
Rates, in percentage, for real mapping failure, gene-specific unmapped sequences, gene-specific mapping failure and pairwise cross-mapping, for dataset 3.

Parameters	BWA MEM	Bowtie2	*Hla-mapper*
Real mapping failure	1.0033%	3.1307%	0.6000%
Gene-specific unmapped sequences			
HLA-A	0.0005%	0.3157%	0.2967%
HLA-B	0.0004%	1.3763%	0.2928%
HLA-C	0.0004%	0.3408%	0.2977%
HLA-E	0.0000%	0.0005%	0.3112%
HLA-F	0.0000%	0.0009%	0.3063%
HLA-G	0.0000%	0.0079%	0.2975%
Gene-specific mapping failure			
HLA-A	2.5060%	5.6479%	0.5885%
HLA-B	1.0800%	5.4614%	0.5861%
HLA-C	2.4326%	5.8561%	0.5954%
HLA-E	0.0006%	0.5126%	0.6226%
HLA-F	0.0001%	0.5388%	0.6125%
HLA-G	0.0002%	0.7675%	0.5951%
Pairwise cross-mapping rates[*]			
HLA-A to HLA-B	0.0037%	0.2824%	0.0000%
HLA-A to HLA-G	0.0015%	0.0957%	0.0000%
HLA-A to HLA-H	2.5281%	4.2857%	–
HLA-A to HLA-J	0.0352%	0.4799%	–
HLA-A to HLA-K	0.0006%	0.1015%	–
HLA-B to HLA-A	0.0281%	0.1274%	0.0000%
HLA-B to HLA-C	1.0025%	3.5877%	0.0006%
HLA-B to HLA-H	0.1112%	0.3398%	–
HLA-C to HLA-A	0.0345%	0.1336%	0.0000%
HLA-C to HLA-B	2.3970%	5.2476%	0.0002%
HLA-E to HLA-A	0.0001%	0.1824%	0.0001%
HLA-E to HLA-B	0.0000%	0.0671%	0.0000%
HLA-E to HLA-L	0.0000%	0.0591%	–
HLA-F to HLA-A	0.0000%	0.1018%	0.0000%
HLA-F to HLA-G	0.0000%	0.0824%	0.0000%
HLA-F to HLA-H	0.0000%	0.1891%	–
HLA-F to HLA-L	0.0000%	0.0637%	–
HLA-G to HLA-H	0.0001%	0.2544%	–
HLA-G to HLA-J	0.0000%	0.3325%	–
HLA-H to HLA-A	0.1981%	3.5702%	0.0022%
HLA-H to HLA-K	0.0017%	0.0997%	–
HLA-J to HLA-A	0.0000%	0.0950%	0.0000%
HLA-J to HLA-G	0.0002%	0.1632%	0.0000%
HLA-J to HLA-H	0.0000%	0.0685%	0.0000%
HLA-K to HLA-A	0.0001%	0.1983%	0.0000%
HLA-K to HLA-G	0.0001%	0.0528%	0.0000%
HLA-K to HLA-H	0.0006%	0.3444%	–

– *Hla-mapper* does not align sequences against these genes. Instead, sequences from these genes are placed into FASTQ files containing unmapped sequences.

* Pairs of genes presenting rates higher than 0.05% for at least one algorithm.

processing were related to non-classical genes (*HLA-G*, *HLA-E,* and *HLA-F*), and the accuracy for the classical genes was no higher than 4%. In contrast, after *hla-mapper* pre-processing, the accuracy for all genes increased and reached 100% for all classical genes. These results do indicate that, if other genes are included in the sequencing (in this case, some HLA pseudogenes), *hla-mapper* should be used to pre-process data before typing class I genes with NGSEngine. This issue is probably related with the NGSEngine database used, which does not include HLA pseudogenes. Finally, we are not sure whether NGSEngine supports datasets with HLA genes that are not included in the software database or datasets containing non-HLA genes. The HLA alleles of each virtual sample from dataset 3 are available at Table S1.

3.4. Comparing hla-mapper and BWA.kit performances

As previously mentioned, BWA developers present a set of scripts (bwa.kit) for HLA-related reads mapping improvement. These scripts use alternative contigs and known HLA sequences from the IPD-IMGT/HLA to re-estimate mapping quality of reads with ambiguous mappings. After applying this post-processing tool in some samples from dataset 3, we noticed that it depleted the sequencing depth within certain gene segments (mainly exons). This phenomenon was particularly evident when reads with low mapping quality (MQ) were removed. For instance, the mean sequencing depth observed in the middle of *HLA-B* exon 3 (at dataset 3) was 140 when using *hla-mapper*, but 12 when using bwa.kit and 8 when reads with MQ = 0 are removed. Although the remaining reads after the bwa.kit post-processing are usually correctly mapped, the low sequencing depth at these important segments might bias SNP genotyping inference, downgrading the usability of this tool.

4. Discussion

Much effort has been made in the development of typing tools to call HLA alleles from NGS data. Many commercial or freely available typing tools have recently been developed. Among the public ones, we can find HLAminer, HLAreporter, HLA-VBseq, OptiType, ATHLATES, PHLAT, HLAforest, and others. The major goal of these tools is HLA typing, i.e., the definition of the two HLA alleles at each locus. Among the commercial ones, we may cite NGSEngine (from GenDX) and HLA Twin (from Omixon). Minor attention has been devoted to the development of mechanisms to minimize the read mapping bias addressed earlier. Moreover, many research groups focus on haplotype structure, Linkage Disequilibrium levels, natural selection and evolution, case-control association studies at the SNP and haplotype levels, and other goals, in which an accurate BAM/SAM file is more important than the allele typing. Besides, the above-mentioned tools usually do not evaluate the promoter segment. To circumvent all these issues, we developed *hla-mapper*.

hla-mapper was firstly introduced in the evaluation of the *HLA-E* variability in two African population samples [24] and it successfully assigned *HLA-E* sequences when several class I genes were sequenced together. Later, this approach was used to address *HLA-E* and *HLA-F* variability in Brazilian samples [7,8] and *HLA-G* variability in two geographically distinct populations, Brazil and Cyprus [6]. The comparison among the results obtained with *hla-mapper*, BWA MEM, and Bowtie2 demonstrated that BAM files produced by *hla-mapper* are more accurate than the ones produced when using a single genome draft as a reference. The *hla-mapper* scoring system minimizes cross-mappings and wrong mappings. This scoring system relies on a database of published sequences and also includes some unpublished ones, allowing the scoring of introns and regulatory sequences, mainly the promoter segment. These unpublished sequences were manually curated and obtained from either homozygous local samples or samples whose HLA genes were sequenced independently. Together, the database sequences cover around 1500 nucleotides upstream the first translated ATG up to 500 nucleotides downstream the 3′UTR segment of each HLA class I gene. With this database, *hla-mapper* enhances the accuracy of the scoring and mapping procedure at exons, introns and regulatory segments, thus reducing the incidence of false-positive and false-negative mapping rates. The database comprises references for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-E*, *HLA-F*, *and HLA-G*, but *hla-mapper* is compatible with data containing non-HLA genes (e.g., whole-genome sequencing) as it preselects sequences before the scoring process.

Here we detected two major issues when mapping HLA class I gene sequences directly with BWA or Bowtie2: cross-mappings between *HLA-A* and *HLA-H*, and between *HLA-B* and *HLA-C*. *HLA-H* is commonly observed as a low-frequency extra amplification when using *HLA-A* specific primers [25]. In addition, *HLA-H* is also present in whole-genome data. Therefore, a high genotyping bias for *HLA-A*, *HLA-B,* and *HLA-C* (and also *HLA-H*) is expected when using BWA MEM or Bowtie2 directly. These errors are actually expected since *HLA-A/HLA-H* and *HLA-B/HLA-C* pairs share more sequence motifs when compared with others (Table S2). For instance, considering frequent alleles and a 70-

mer motif, which is the minimum read size considered by *hla-mapper* when using default parameters, the pairs *HLA-A/HLA-H* and *HLA-B/HLA-C* share 1.4% and 1.9% of all possible motifs, respectively. However, when using a 20-mer motif (the typical size of a primer), the pairs *HLA-A/HLA-H* and *HLA-B/HLA-C* share 12.1% and 14.2% of all possible motifs, respectively (Table S2). The scoring and mapping strategy proposed by *hla-mapper* (Fig. 1) circumvents these issues and generates accurate read mappings for HLA class I genes, even when only new HLA alleles are considered (dataset 2). Moreover, *hla-mapper* performance was superior even when smaller read sizes and HLA pseudogenes were present (dataset 3).

A downside of *hla-mapper* is that the scoring system uses known HLA sequences as a reference. Although mismatches are allowed in the scoring process, minimizing the impact of new variable sites, *hla-mapper* could exclude sequences carrying new large indels or a large number of new point mutations. Thus, it is possible that reads from samples presenting new alleles with several divergences in comparison with known HLA sequences would not be properly mapped. Nevertheless, as observed in dataset 2, the strategy used by *hla-mapper* is straightforward and suitable to map HLA sequences to their proper reference even when the dataset includes only new HLA alleles (Table 2).

We evaluated the performance of some of the above-mentioned typing tools when data was pre-processed with *hla-mapper*. While hla-mapper pre-processing did not influence some tools performance, others were much improved. For instance, we noticed a better performance for HLAminer and NGSEngine. At least for the latter, this improvement was probably due to the database used by NGSEngine, since there were no HLA pseudogenes included. NGSEngine performance reached 100% for classical class I genes when data was pre-processed with *hla-mapper*, mostly because the processed data no longer contains HLA pseudogenes sequences. Since both HLA-VBseq and OptiType were updated with complete HLA databases, that include pseudogenes, typing accuracy was the same with or without *hla-mapper* pre-processing. Moreover, many typing tools are not compatible with whole-genome data or they work quite slow when dealing with it. Since *hla-mapper* preselects HLA sequences and creates gene-specific FASTQ and BAM files, mapping accuracy would be greatly improved by pre-processing the data with *hla-mapper*, as demonstrated for NGSEngine and HLAminer, even for non-classical genes.

In conclusion, we hereby present a strategy and an application to handle HLA NGS data in order to achieve an accurate sequence mapping of HLA sequences to the human reference genome (hg19 or hg38). Many different genotyping, haplotyping, and allele calling methods might be applied afterward using the *hla-mapper* outputs. *hla-mapper* is freely available at www.castelli-lab.net/apps/hla-mapper and is compatible with most of the UNIX-based systems.

Acknowledgements

This work was supported by FAPESP/Brazil (Grant# 2013/17084-2). E.C.C and C.T.M.J. are supported by CNPq/Brazil (Grants# 304471/2013-5 and 309572/2014-2).

Conflict of interests

There is no financial conflict of interest.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.humimm.2018.06.010.

References

[1] J. Klein, A. Sato, The HLA system. First of two parts, N. Engl. J. Med. 343 (2000) 702.
[2] V. Matzaraki, V. Kumar, C. Wijmenga, A. Zhernakova, The MHC locus and genetic susceptibility to autoimmune and infectious diseases, Genome Biol. 18 (2017) 76.
[3] T.V. Hviid, HLA-G in human reproduction: aspects of genetics, function and pregnancy complications, Hum. Reprod. Update 12 (2006) 209.
[4] E.A. Donadi, E.C. Castelli, A. Arnaiz-Villena, M. Roger, D. Rey, P. Moreau, Implications of the polymorphism of HLA-G on its function, regulation, evolution and disease association, Cell Mol. Life Sci. 68 (2011) 369.
[5] B.K. Kaiser, J.C. Pizarro, J. Kerns, R.K. Strong, Structural basis for NKG2A/CD94 recognition of HLA-E, Proc. Natl. Acad. Sci. USA 105 (2008) 6696.
[6] E.C. Castelli, P. Gerasimou, M.A. Paz, J. Ramalho, I.O. Porto, T.H. Lima, et al., HLA-G variability and haplotypes detected by massively parallel sequencing procedures in the geographicaly distinct population samples of Brazil and Cyprus, Mol. Immunol. 83 (2017) 115.
[7] T.H. Lima, R.V. Buttura, E.A. Donadi, L.C. Veiga-Castelli, C.T. Mendes-Junior, E.C. Castelli, HLA-F coding and regulatory segments variability determined by massively parallel sequencing procedures in a Brazilian population sample, Hum. Immunol. 77 (2016) 841.
[8] J. Ramalho, L.C. Veiga-Castelli, E.A. Donadi, C.T. Mendes-Junior, E.C. Castelli, HLA-E regulatory and coding region variability and haplotypes in a Brazilian population sample, Mol. Immunol. 91 (2017) 173.
[9] J. Robinson, J.A. Halliwell, J.D. Hayhurst, P. Flicek, P. Parham, S.G. Marsh, The IPD and IMGT/HLA database: allele variant databases, Nucl. Acids Res. 43 (2015) D423.
[10] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (2009) 1754.
[11] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (2012) 357.
[12] D.Y. Brandt, V.R. Aguiar, B.D. Bitarello, K. Nunes, J. Goudet, D. Meyer, Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes Project Phase I Data, G3 (Bethesda) (5) (2015) 931.
[13] J.F. Degner, J.C. Marioni, A.A. Pai, J.K. Pickrell, E. Nkadori, Y. Gilad, et al., Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data, Bioinformatics 25 (2009) 3207.
[14] K. Kiyotani, T.H. Mai, Y. Nakamura, Comparison of exome-based HLA class I genotyping tools: identification of platform-specific genotyping errors, J. Hum. Genet. 62 (2017) 397.
[15] S. Tian, H. Yan, M. Kalmbach, S.L. Slager, Impact of post-alignment processing in variant discovery from whole exome data, BMC Bioinf. 17 (2016) 403.
[16] S. Tian, H. Yan, C. Neuhauser, S.L. Slager, An analytical workflow for accurate variant discovery in highly divergent regions, BMC Genom. 17 (2016) 703.
[17] D.C. Bauer, A. Zadoorian, L.O. Wilson, Melbourne Genomics Health A, Thorne NP: Evaluation of computational programs to predict HLA genotypes from genomic sequencing data, Brief Bioinform. (2016).
[18] R.L. Warren, G. Choe, D.J. Freeman, M. Castellarin, S. Munro, R. Moore, et al., Derivation of HLA types from shotgun sequence datasets, Genome Med. 4 (2012) 95.
[19] N. Nariai, K. Kojima, S. Saito, T. Mimori, Y. Sato, Y. Kawai, et al., HLA-VBSeq: accurate HLA typing at full resolution from whole-genome sequencing data, BMC Genom. 16 (Suppl 2) (2015) S7.
[20] Y. Huang, J. Yang, D. Ying, Y. Zhang, V. Shotelersuk, N. Hirankarn, et al., HLA reporter: a tool for HLA typing from next generation sequencing data, Genome Med. 7 (2015) 25.
[21] A. Szolek, B. Schubert, C. Mohr, M. Sturm, M. Feldhahn, O. Kohlbacher, OptiType: precision HLA typing from next-generation sequencing data, Bioinformatics 30 (2014) 3310.
[22] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, et al., The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010) 1297.
[23] M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, et al., A framework for variation discovery and genotyping using next-generation DNA sequencing data, Nat. Genet. 43 (2011) 491.
[24] E.C. Castelli, C.T. Mendes-Junior, A. Sabbagh, I.O. Porto, A. Garcia, J. Ramalho, et al., HLA-E coding and 3' untranslated region variability determined by next-generation sequencing in two West-African population samples, Hum. Immunol. 76 (2015) 945.
[25] K. Hosomichi, T.A. Jinam, S. Mitsunaga, H. Nakaoka, I. Inoue, Phase-defined complete sequencing of the HLA genes by next-generation sequencing, BMC Genom. 14 (2013) 355.