hla-mapper

version 4.1, May 2021

www.castelli-lab.net/apps/hla-mapper

Contact: Erick C. Castelli (erick.castelli@unesp.br)

Molecular Genetics and
Bioinformatics Laboratory
www.castelli-lab.net

DOCUMENTATION ROADMAP

OVERVIEW

The polymorphic and repetitive nature of HLA and KIR genes leads to alignment failures or read mapping bias, leading to genotyping and haplotyping errors.

We have designed hla-mapper to minimize these errors. hla-mapper uses a database of known sequences to apply an algorithm to decide where each read should be aligned. The final output is a BAM file with reads mapped to the human reference genome (version hg38), allowing any downstream analysis that suites the user. The resulting BAM file is far more reliable than those generated using BWA or Bowtie2.

hla-mapper supports single-end or paired-end sequencing data (Illumina, Ion) from targeted sequencing, whole-genome sequencing, and whole-exome sequencing. It supports the following genes:

HLA-A, *HLA-B*, *HLA-C*, *HLA-G*, *HLA-E*, *HLA-F* (validated).

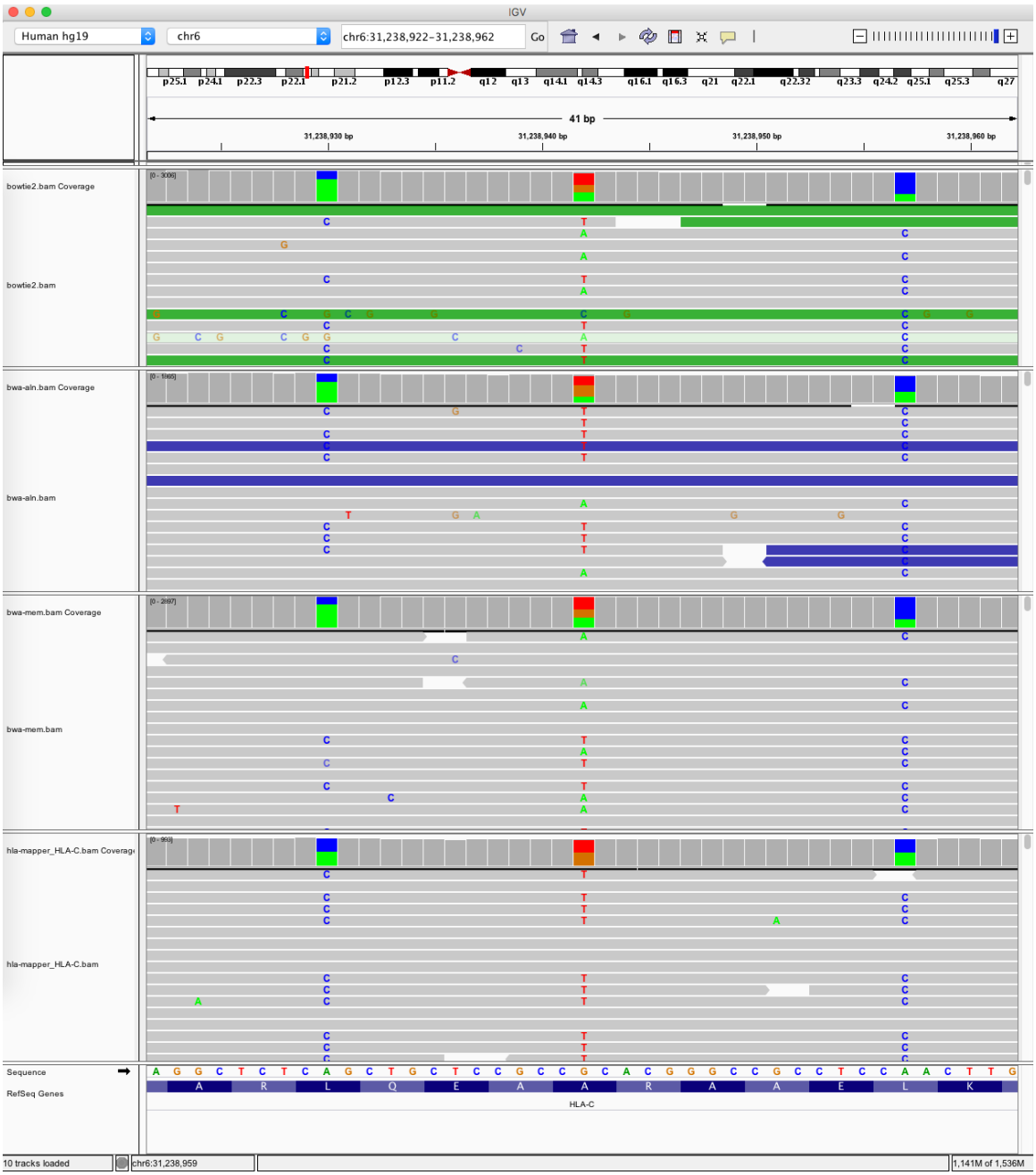
TAP1, TAP2, MICA, MICB, HLA-DOA, HLA-DOB, and HLA class II genes (available, but under beta test)

KIR genes (available, but under beta test)

LILRB1, *LILRB2*, *LAIR1*, and *LAIR2* (available, but under beta test)

hla-mapper selects relevant sequences before the scoring process using a k-mer approach. It also trims all reads considering base quality and size before the scoring algorithm, but it map the original read sequence in the final BAM. Each read pair is compared and scored against the database, registering the number of mismatches observed for each gene. The scoring system compares each gene's scores, addressing the read pair to the locus with the lowest mismatch rate. Read pairs with more than one possible destination are not removed but mapped as secondary in the final BAM file. Finally, hla-mapper produces a BAM file with all alignments.

The following image is an example of the hla-mapper output, comparing the alignments obtained with BWA ALN, BWA MEM, Bowtie2, and hla-mapper. It can be noticed a significant number of misaligned reads when using BWA ALN, BWA MEM, or Bowtie2, with the presence of unbalanced variable sites and one variable site presenting three alleles, all of them prone to genotyping errors.



LEGAL ISSUES

You cannot modify or distribute this software.

Permission is hereby granted to use this software for research purposes, without fee, subject to this condition: (1) Permission for use of this software is granted only if the user accepts full responsibility for any undesirable consequences; the authors accept NO LIABILITY for damages of any kind. (2) Permission for use of this software is granted only if the user accepts to not modify any part of the software. (3) Permission for use of this software is granted only if the user accepts to not use it for clinical purposes or to not use as the basis of commercial products. These conditions apply to any software derived from or based on *hla-mapper*. If you use our work, you ought to acknowledge us:

Castelli EC et al. Hla-mapper: An application to optimize the mapping of HLA sequences produced by massively parallel sequencing procedures. Hum Immunol. 2018 Sep;79(9):678-684

About the hla-mapper database:

hla-mapper databases were built based on sequences available at the IPD-IMGT/HLA, the IPD-IMGT/KIR, GENBANK, and curated sequences from the developer laboratory.

About BWA, SAMTOOLS, MIRA:

These are third party software that must be installed before using hla-mapper. Please follow the instructions on their websites to install them properly. MIRA is optional.

HOW TO INSTALL AND CONFIGURE HLA-MAPPER

First, you need boost installed in your system. For MacOS, we recommend homebrew (brew.sh), by typing “brew install boost”. For Linux, we recommend type “sudo apt-get install libboost-all-dev”.

For now, the hla-mapper source code is not provided. However, binaries for MacOS and Linux 64bits are available. You need to download the distribution file at www.castelli-lab.net/apps/hla-mapper and unzip it. You also need to download the last database version as well. You may use the application directly from the extracted folder (select the correct version at the /bin folder), or you may move the hla-mapper program to another location.

To use hla-mapper, you need a working copy of both BWA and SAMTOOLS. Please follow the instructions at the BWA and SAMTOOLS websites to install them properly. We also recommend the use of IGV (https://www.broadinstitute.org/igv/) to visualize the BAM files generated by hla-mapper.

Attention: some BWA and Samtools versions are not compatible with hla-mapper.

BWA compatible versions: 0.7.15, 0.7.16a, 0.7.17

SAMTOOLS compatible versions: 1.8, 1.9, 1.10, 1.11, 1.12

The next step is optional, but it will save you time when using hla-mapper. Using any text editor (we suggest nano), you need to create a file named *.hla-mapper* at your home folder using a command such as:

sudo nano ~/.hla-mapper

Add the following information and save it. Do not use blank spaces!

db=path_to_the_hla-mapper_database [example: db=/home/user/hla-mapper/hla-mapper_db_003.1/]

bwa= path_to_bwa [not necessary if bwa is available by just typing *bwa* in the terminal]

samtools=path_to_samtools [not necessary if samtools is available by just typing *samtools* in the terminal]

This will set the path do the hla-mapper database and you won't need to type it every time you use hla-mapper.

USING HLA-MAPPER DNA

You need to download an hla-mapper database at www.castelli-lab.net/apps/hla-mapper

There are two main hla-mapper functions, dna and select. By running the hla-mapper software using the Linux/Mac terminal, you will see both these functions. Function ***dna*** is used to map DNA sequences (WGS, WES, amplicons, etc) and produce aligned "BAM format" files. Function ***select*** is used to select relevant sequences (from WGS, for instance).

Here is an example of a simple hla-mapper run using the example data provided (at folder /sample). At the Linux/Mac terminal, assuming that you have moved hla-mapper to /usr/local/bin and the configuration file was properly set, type:

hla-mapper dna (this will show you all the options for this algorithm)

You have three options to indicate the data to be processed:

Using r1 and r2 for raw paired-end sequencing data [.fastq, .fq, or .gz]

Using r0 for raw single-end sequencing data [.fastq, .fq, or .gz]

Using bam for aligned data. Reads must be aligned using the hg38 reference genome.

You also need to indicate a sample name: sample=sample_name

Usage samples:

hla-mapper dna r1=R1.fastq.gz r2=R2.fastq.gz sample=Test

hla-mapper dna r0=R1.fastq.gz sample=Test

hla-mapper dna bam=hg38.bam sample=Test

The above example will produce BAM files for all supported genes, using the human hg38 genome version as reference. There are some adjustments you can make when running hla-mapper. To get a list of these adjustments/options, just type "hla-mapper dna" at the terminal. Bellow is a list of them, with their full descriptions.

db

The path to the hla-mapper database. There is one for the MHC and another for KIR genes. E.g.: db=/home/user/hla-mapper/db/

output

The folder where all files will be placed. By default, hla-mapper creates a folder named "hla-mapper" next to the R0/R1/BAM file. Do not use blank spaces. E.g.: output=/home/user/output/hla-mapper

Attention: all previous files in the output folder are deleted!!! Be careful!

threads

Number of threads to be used. By default, hla-mapper uses half the computer capacity. E.g.: threads=4

bed

A BED file with coordinates to be used by the algorithm. Use this only if you know what you are doing, because the hla-mapper database already provide this file. E.g.: bed=/home/user/file.bed

buffer

Number of sequences processed per thread. Increasing this value may speedy up the analysis but needs more memory. You should not change this unless you know what you are doing. E.g.: buffer=5000

error

This is the error threshold for read trimming. By default, hla-mapper do extract the largest possible fragment in which all nucleotides present an error probability lower than the one fixed here. The default value is 0.08. E.g: error=0.08

Attention: for now, hla-mapper supports only Sanger / Illumina 1.9 encoding. Undesired behavior will occur when using other encodings.

tolerance

How many mismatches are allowed to assign a read to any locus or to be dropped out. The default value is 0.05, or 5% of the read size. E.g. tolerance=0.05

downsample

Downsampling for the adjustment procedure for. E.g.: downsample=30. Increase this number will increase speed exponentially. This value is set for WGS, WES, and amplicon sequencing.

bwa

If BWA is not available by only typing *bwa* at terminal, the complete path to the bwa binary (including the binary name) should be indicated here. E.g.: bwa=/home/user/bwa-aligner/bwa

Tip: you can set this path at the configuration file. If the BWA aligner is already installed and available by only typing *bwa*, hla-mapper will automatically detect it.

samtools

If SAMTOOLS is not available by only typing *samtools* at terminal, the complete path to SAMTOOLS (including the binary name) should be indicated here. E.g.: samtools=/home/user/samtools

Tip: you can set this path at the configuration file. If SAMTOOLS is already installed and available by only typing *samtools*, hla-mapper will automatically detect it.

--skip-unmapped

Ignore all unmapped reads from BAM files (not recommended).

--skip-adjust

Do not proceed with the typing and adjustment phase (not recommended).

--low-mem

Only use the low memory mode for read selection. This should be used only when necessary. It will slow down the entire pipeline.

--quiet

Do not display any warnings or messages. Quiet mode.

UNDERSTANDING THE HLA-MAPPER OUTPUTS

In the output folder, all files start with the sample name.

General files

_addressing_table.txt

The distance (mismatches) observed among this read and each gene.

_selected.R1.fast, _selected.R2.fastq, _selected.R0.fastq

The FASTQ files with the sequences considered for the scoring system.

.hla-mapper.log

A log file with the hla-mapper configuration.

.adjusted.bam

This is the BAM file with all reads mapped to the hg38 reference genome. **You should use this one for genotyping. Open this file in IGV!**

Gene-specific files

_R1.fastq, _R2.fastq, _R0.fastq

Gene-specific fastq files. Some reads may be present in more than one loci and it is ambiguous. **You should use this files for genotyping.**

_.log

Log file with reads that have been mapped as primary, secondary, and other parameters.

HOW TO CITE HLA-MAPPER

Hla-mapper: an application to optimize the mapping of hla sequences produced by massively parallel sequencing procedures. Human Immunology 2018, https://doi.org/10.1016/j.humimm.2018.06.010

Since hla-mapper uses BWA and Samtools, these are their proper references. Please cite these references if you use hla-mapper.

The Sequence Alignment/Map format and SAMtools. Li H et al. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352.

Fast and accurate short read alignment with Burrows-Wheeler transform. Li H, Durbin R. Bioinformatics. 2009 Jul 15;25(14):1754-60. doi: 10.1093/bioinformatics/btp324.