AmpliSAS and AmpliHLA: web server tools for MHC typing of non-model species and human using NGS data

Alvaro Sebastian [1,2*], Magdalena Migalska [2], Aleksandra Biedrzycka [3]

[1] Sixth Researcher - www.sixthresearcher.com
[2] Evolutionary Biology Group, Faculty of Biology, Adam Mickiewicz University, ul. Umultowska 89, 61-614 Poznan, Poland (http://evobiolab.biol.amu.edu.pl)
[3] Institute of Nature Conservation, Polish Academy of Sciences, Al. Mickiewicza 33, 31-120 Kraków, Poland

[*]To whom correspondence should be addressed: sixthresearcher@gmail.com

Abstract AmpliSAS and AmpliHLA are web server tools for automatic genotyping of MHC genes from high-throughput sequencing data. AmpliSAS is designed specifically to analyze amplicon sequencing data from non-model species and it is able to perform de-novo genotyping without any previous knowledge of the reference alleles. AmpliHLA is a human specific version, it performs HLA typing by comparing sequenced variants against human reference alleles from the IMGT/HLA database. Here we describe four genotyping protocols: the first two use amplicon sequencing data to genotype the MHC genes of a passerine bird and human respectively; the third and fourth present the HLA typing of a human cell line starting from RNA and exome sequencing data respectively.

Keywords: Bioinformatics, next-generation sequencing, NGS, amplicon sequencing, RNA-Seq, WES, WXS, MHC, HLA, genotyping, alleles, haplotypes, IMGT, AmpliSAS, AmpliHLA

Running head: AmpliSAS and AmpliHLA: web server tools for MHC typing

Index

1. Introduction

The major histocompatibility complex (MHC) encodes a family of genes of central importance in vertebrate adaptive immunity. In human, the MHC is commonly named as human leukocyte antigen system (HLA). MHC molecules are responsible for binding and presenting antigens to the immune system T-cells. There are two classes of classical MHC genes involved in adaptive immunity: class I which encode molecules that present peptides from the intracellular environment (*e.g.* viruses) to T-cells; and class II that encode molecules to present peptides from the extracellular environment (*e.g.* bacteria) [1]. MHC genes are the most polymorphic genes currently characterised in vertebrates [2]. This polymorphism is believed to be primarily driven by co-evolving pathogens and through mate choice, and maintained in the populations by balancing selection [3–7]. The number of MHC genes differs greatly between species (**Figure 1**), and in some taxa the number of MHC genes differs also between individuals of the same species [8–19].

As a simplification, we can say that humans have a fix number of classical, functional major HLA loci: A, B and C for class I and DP, DQ and DR for class II (**Figure 1**) [14]. Some of the HLA loci are extremely polymorphic (**Table 1**), for example, currently 4828 alleles have been described for HLA-B locus (IMGT/HLA database 3.29 release, July 2017) [20]. In birds there is a great variability in the complexity of the MHC regions, the number of MHC genes and their variabilities (**Figure 1**). Chickens have a minimal essential MHC with two classical class I genes but only one is expressed at a high level and its diversity is considerably lower than human [17, 21, 22]. In contrast, birds of the order Passeriformes have much more complex MHC systems with dozens of genes highly duplicated and polymorphic [8, 23–25].

Table 1. Number of alleles for each HLA gene as registered at the 3.29 release of the IMGT/HLA database from the European Bioinformatics Institute (EBI) at July 2017.

Class I		Class II			
Gene	**Alleles**	**Gene**	**A Alleles**	**B Alleles**	**A x B**
HLA-A	3 968	**DR**	7	2 376	16 632
HLA-B	4 828	**DQ**	94	1 142	107 348
HLA-C	3 579	**DP**	53	894	47 382

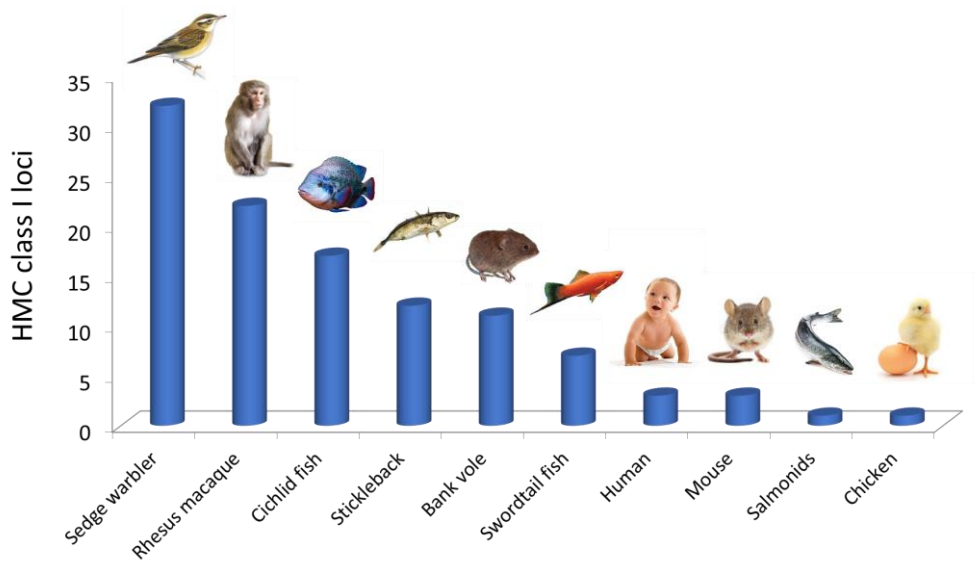


Figure 1. Maximum number of MHC class I loci extracted from the literature for some representative species up to date.

MHC genotyping is especially demanding in non-model species with highly duplicated MHC genes and where limited genomic information does not allow to design locus-specific probes or primers. In the past, expensive and time-consuming methods were required for genotyping [26], but nowadays, next-generation sequencing (NGS) of MHC amplicons has become the method of choice for non-model species [24, 27–30] and human [31–33]. Amplicon sequencing technique is based on sequencing multiple PCR products (amplicons) at once by means of NGS technologies (**Figure 2**). With a single experiment it is possible to accurately genotype hundreds of individuals with complex MHC systems [34, 35]. The **Note 1** describes briefly the technique workflow and the **Note 2** presents the benefits introduced by NGS.

However, relatively high error rates in the amplicon sequences, stemming both from intrinsic sequencing error rates of NGS technologies and PCR errors, such as chimera formation, arise new genotyping challenges (see **Note 3**). Different strategies have been proposed to detect sequence errors and correct them without altering genotypes [36, 37]. AmpliSAS is an error correction strategy that clusters real alleles with low frequency similar variants based on the particular error-rate of the NGS technology used [38]. In a recent benchmark, AmpliSAS produced reliable genotypes for the sedge warbler (*Acrocephalus schoenobaenus*), a passerine bird with a highly complex MHC system composed of dozens of loci and thousands of alleles (**Note 4**) [37]. Furthermore, AmpliSAS has been successfully validated in other non-model species as bank vole, guppy, three-spine stickleback, blue petrel or black-tailed godwit [12, 38–40].

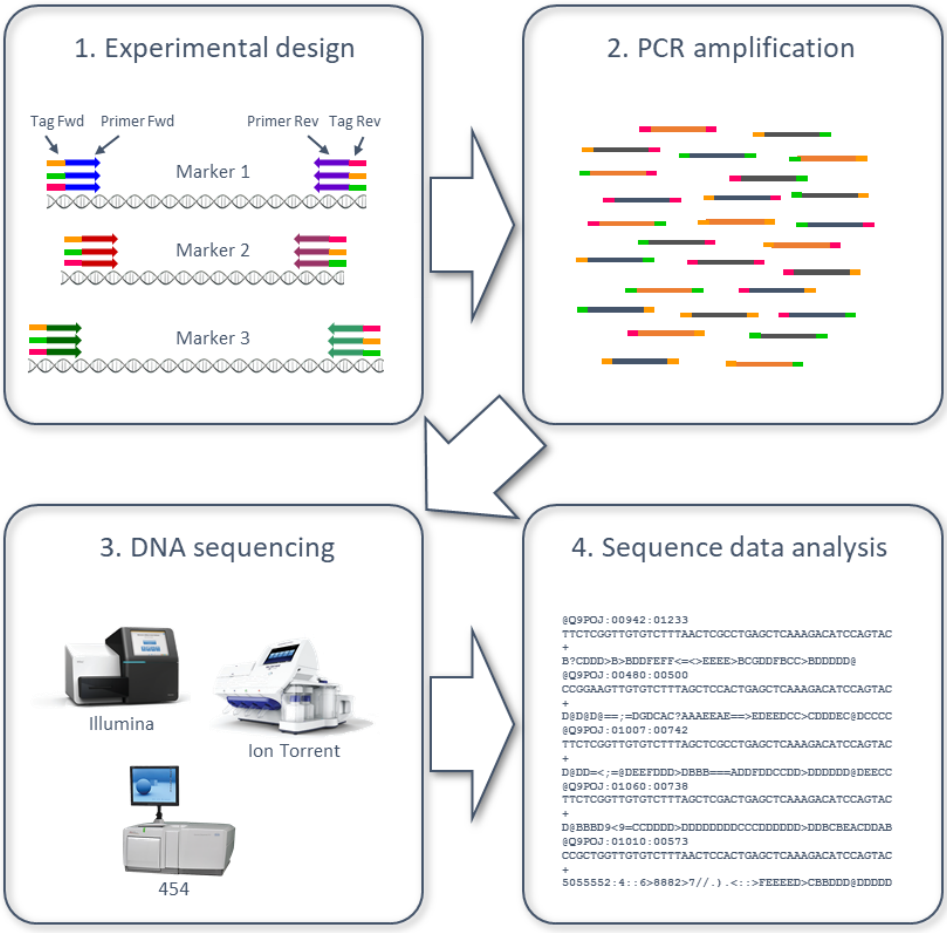


Figure 2. Amplicon sequencing workflow schema: 1) experimental design (marker regions, primers and tags), 2) PCR amplification, 3) DNA sequencing, and 4) sequence data analysis.

AmpliHLA is an adaptation of the AmpliSAS algorithm for human HLA typing, it combines the genotypes from multiple markers of the same locus and compares them with the deposited HLA alleles from the IMGT/HLA database [20]. As a result, it retrieves the HLA types with the highest possible resolution. AmpliSAS algorithm has previously been tested with human amplicon data retrieving accurate genotypes [38]. Recently, AmpliHLA has been expanded with an adaptation of the Seq2HLA algorithm to be able to analyse RNA-Seq and whole exome sequencing (WES or WXS) data [41].

Both, AmpliSAS and AmpliHLA, are available as ready-to-use web server tools at: http://evobiolab.biol.amu.edu.pl/amplisat/index.php .

Here, we present four MHC genotyping protocols. The first describes how to process amplicon data with AmpliSAS to obtain the MHC class I genotypes of five sedge warblers (passerine birds) that possess up to 56 MHC class I alleles per individual. The second presents the HLA typing with AmpliHLA of five human cell lines whose alleles were previously characterized by Sanger sequencing. The third and fourth protocols use also AmpliHLA to type of one of the previous human cell lines using RNA-Seq and WES data instead of amplicons.

2. Materials

The only resources required to replicate the analysis described in this chapter are an Internet connection and a web browser. These will suffice to learn how to use AmpliSAS, AmpliHLA and other tools from the AmpliSAT suite (Amplicon Sequencing Analysis Tools) which are presented here.

2.1.AmpliSAS

Amplicon Sequence Assignment tool (AmpliSAS) is a web server tool designed to analyse NGS amplicon data and perform automatic genotyping of complex MHC systems (http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplisas, **Figure 4A**) [38]. AmpliSAS workflow is divided into three main steps (**Figure 3**): 1) sequence de-multiplexing, 2) sequence clustering, and 3) artefact filtering. In summary, first the reads are de-replicated and classified into amplicons. Second, during clustering, variants are aligned to each other to find sequencing errors, these erroneous variants are removed and their coverages added to the true ones. Third, remaining low frequency variants are inspected to remove artefacts and chimeric PCR products. Finally, allele sequences and frequencies for each amplicon are retrieved in an Excel spreadsheet format, making them easy to interpret. Definitions of amplicon, variant and other useful terms are listed in **Table 2**. Complete details about AmpliSAS algorithm can be found in [38], and a comparison of the performance of AmpliSAS against other MHC genotyping methods in [37].

Table 2. Definitions of commonly used terms in the amplicon sequencing technique. They can slightly differ between authors.

Term	Definition
Marker	A DNA region to be amplified.
Sample	A single genetic material to be sequenced (usually from an individual of the study organism).
Tag	A unique short DNA sequence that identifies unambiguously a sample. Tags are usually ligated after PCR amplification or directly included in one or both primers.
Read	Each individual sequence retrieved by a sequencing run. A sequence run will retrieve thousands/millions of reads.
Amplicon	A set of reads derived from a single PCR (one marker, one sample); may comprise products of several co-amplifying loci.
Amplicon depth	Number of reads per amplicon.
Variant	Unique sequence retrieved by a sequencing run. Usually multiple reads correspond to one variant (= one sequence).
Variant depth / coverage	Number of reads per variant.
Per amplicon frequency	Number of reads per sequence divided by the total number of reads in a single amplicon.
True Variant / Allele	Sequence that matches a real allele or real sequence in the sample genome.
Artefact	Variant resulting from experimental/technical errors: sequencing errors, polymerase errors, non-specific amplifications (paralogs, pseudogenes), contaminants, PCR chimeras, etc.

2.2.AmpliHLA

Amplicon Sequencing HLA typing tool (AmpliHLA) is a web server tool designed to retrieve automatically HLA haplotypes from amplicon, RNA-Seq or WES data (http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplisas, **Figure 4B**). AmpliHLA implements the AmpliSAS algorithm explained in the previous section to analyze amplicon data [38], but additionally it is able to combine the information from several amplified regions of a single locus and compare their sequences with the thousands of HLA alleles annotated in the IMGT/HLA database [20].

In the analysis of RNA-Seq and WES data, AmpliHLA uses a modified version of the Seq2HLA algorithm [41]. First, the reads are mapped with BOWTIE [42] against a curated dataset of HLA variable regions (exons 2 and 3) extracted from the IMGT/HLA database [20, 41]. Then AmpliHLA analyses mapping results in a locus-specific manner:

i) the allele with the maximum number of mapped reads is selected and these reads are substracted from the remaining allele mappings, ii) step i) is successively repeated and the corrected mapped read numbers are annotated until there are no more alleles with mapped reads left, iii) one or two alleles with the highest corrected numbers of mapped reads are selected based on the drop of their values as in [29] and their HLA types are printed.

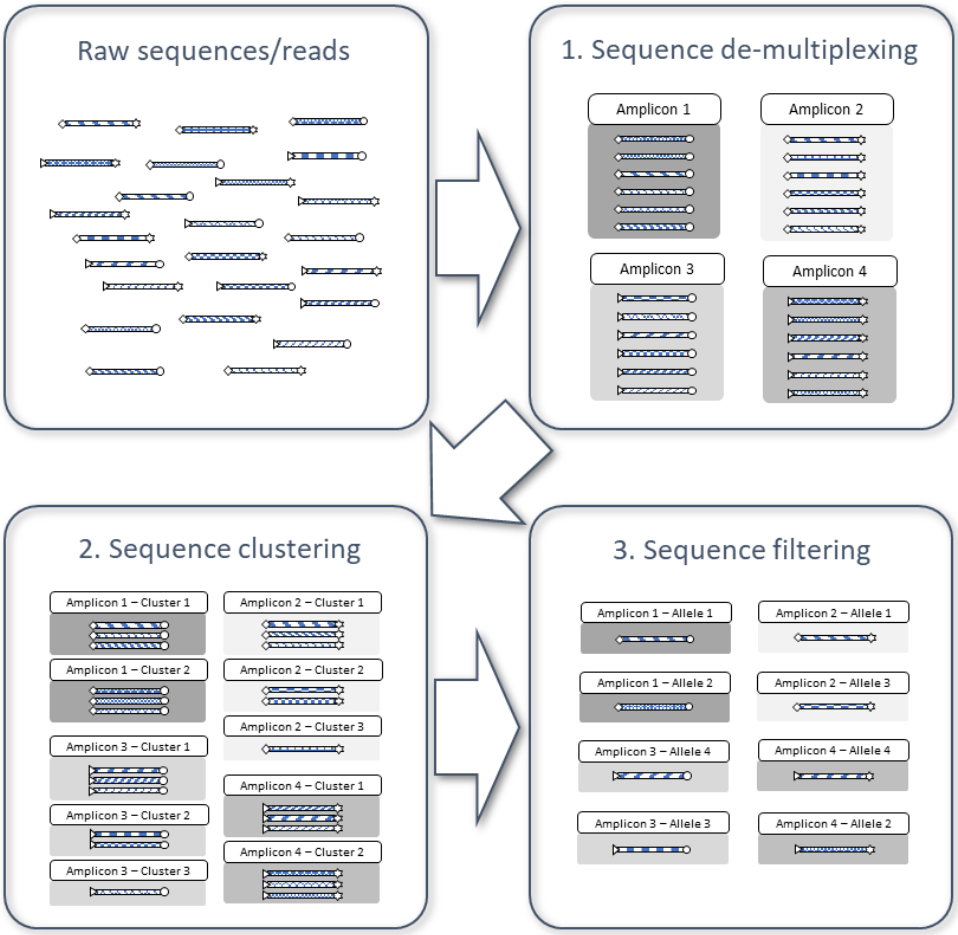


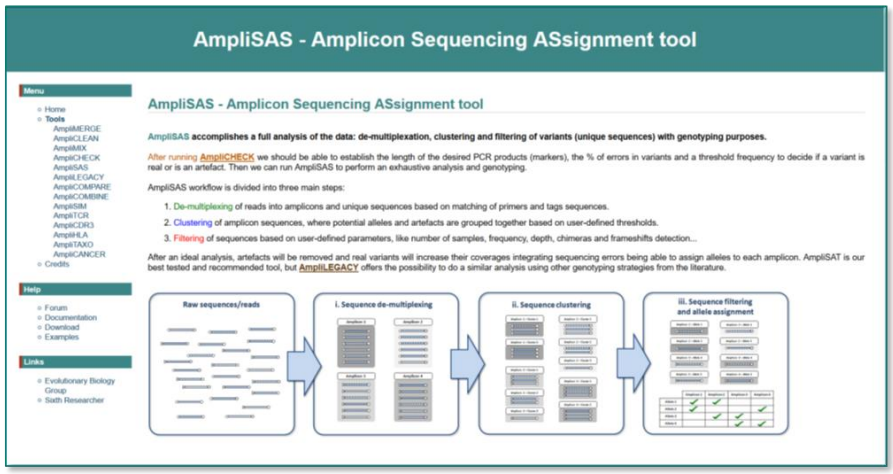
Figure 3. AmpliSAS workflow schema: 1) sequence de-multiplexing, 2) clustering, and 3) filtering and allele assignment.

2.3.AmpliCOMPARE

AmpliCOMPARE is another tool from the AmpliSAT suite. It is available at: http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplicompare

AmpliCOMPARE compares the genotyping results of two experimental or technical replicates, or between two different genotyping strategies. It accepts as input two Excel files with the same format as the AmpliSAS/AmpliHLA output one. With AmpliCOMPARE it is easy to detect genotyping discrepancies highlighted in the comparison output Excel file.

A



B

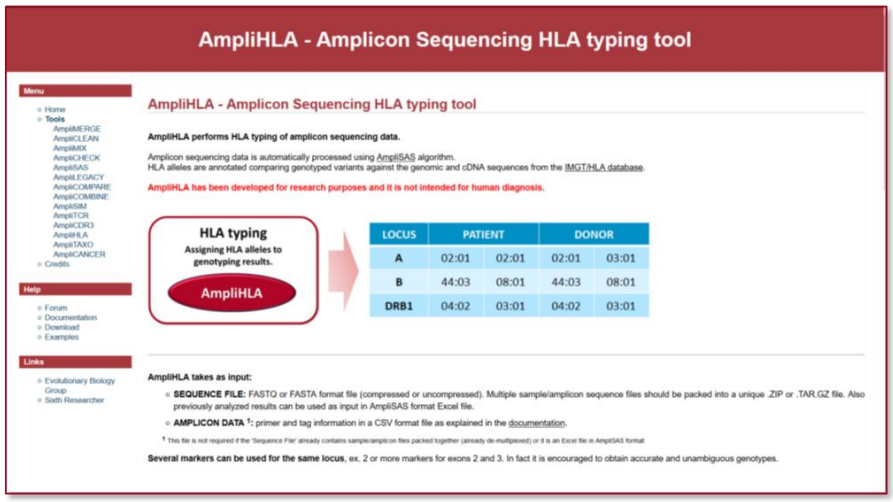


Figure 4. AmpliSAS (A) and AmpliHLA (B) web interfaces.

3. Methods

3.1. MHC class I genotyping in a passerine bird

As previously explained, AmpliSAS algorithm is designed to genotype complex MHC gene families, such as those in the sedge warbler, a passerine bird with MHC class I copy number variation and dozens of MHC class I loci in a single individual [8].

In the present protocol we will analyse data from a previous sedge warbler MHC class I genotyping study (accession PRJEB11775 at the European Nucleotide Archive - ENA). The purpose of the study was to use ultra-deep Illumina sequencing to resolve genotypes at exon 3 of MHC class I genes in the sedge warbler [37]. We will use a pre-processed and compressed FASTQ file with already merged and cleaned Illumina paired-end reads (see **Note 5**).

1) Open the AmpliSAS online submission form (**Figure 5A**):
   http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplisas
2) Enter a name for the run and, optionally, an email address if you desire to receive the results by email.
3) Copy and paste the following link into the 'Sequence file URL' field: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR113/008/ERR1136308/ERR1136308.fastq.gz

   Optionally, you can download the compressed FASTQ file in your computer and upload it with the 'Browse' button (slow). If reads have been separated into single amplicon files after sequencing, they can be packed into a single ZIP or TGZ format file and used as input (see **Note 6**).
4) Select 'Auto' from the 'Technology' options to ask the program to automatically detect the sequencing technology used.

5) Copy and paste into the 'Amplicon data' field the following information about primers used to amplify the MHC class I exon 3 region and DNA tags included at the end of the primers to identify the first 5 individuals from the experiment:

```
>marker,feature,primer_f,primer_r
MHCI,EXON3,GAGYGGGGGTCTCCACAC,TGCGMTCCAGYTCCTTCTGCCC
>sample,tag_f,tag_r
BIRD_37,ACAACC,AGCCTC
BIRD_38,ACAACC,TCGTTA
BIRD_39,ACAACC,TGTGGC
BIRD_40,ACAACC,CTCTGC
BIRD_41,ACAACC,CCTAAT
```

The information for all the individuals is available at the ENA experiment page: https://www.ebi.ac.uk/ena/data/view/ERX1215174

If the input is a compressed file containing single amplicon files for each individual, then 'Amplicon data' must be left empty (see **Note 6**).

6) Adjust the following parameters in the submission form: 'Maximum number of alleles per amplicon': 50 and 'Maximum number of reads per amplicon': 5000. Keep the rest of parameters with default values.

7) Click the 'Run' button, and after a while (first the server has to upload the reads file), it will display the message 'AmpliSAS job has been queued in the server, be patient' together with a link to access the results when the analysis will be completed (**Figure 5B**).

8) Click the link to the results page and reload it until the analysis is completed. The analysis should be finished in a few minutes, but the waiting time may depend on the current load on the server and the size of the uploaded files. Keep the results page open (**Figure 5C**), in the **Section 3.3** we will learn how to interpret them.

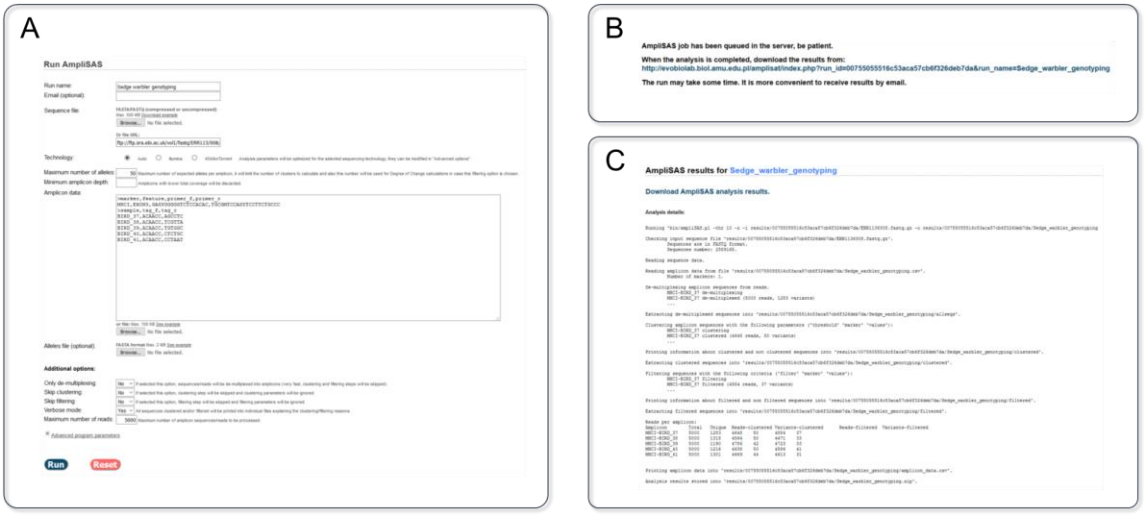


Figure 5. A: AmpliSAS input form. B: Message with the link to the results page. C: Output summary after analysis completion and link to the results ZIP file.

3.2. Customizing the MHC class I genotyping

In the previous section we performed a complex MHC class I genotyping with mostly automatic and default options. Now we will learn how to customize the analysis adjusting manually some program parameters (see **Note 7**).

1) Open a new tab in the web browser and repeat the steps 1-3 as explained on the **Section 3.1**.

4) Select 'Illumina' from the 'Technology' options. Scroll down to the 'Advanced program parameters' section, you will notice that some clustering and filtering parameters have been set to recommended values for Illumina data.

5) Copy and paste the following data in the 'amplicon data' field:

```
>marker,feature,length,primer_f,primer_r
MHCI,EXON3,235 238 241,GAGYGGGGGTCTCCACAC,TGCGMTCCAGYTCCTTCTGCCC
>sample,tag_f,tag_r
BIRD_37,ACAACC,AGCCTC
BIRD_38,ACAACC,TCGTTA
BIRD_39,ACAACC,TGTGGC
BIRD_40,ACAACC,CTCTGC
BIRD_41,ACAACC,CCTAAT
```

In the previous analysis we did not specify expected allele lengths, so the program automatically set them to 241 (see **Note 8**).

6) Adjust the same parameters as in step 6 from **Section 3.1**. Go to 'Advanced program parameters – Clustering parameters' section and set 'Exact length required': yes, 'Minimum dominant frequency': 10%. Go to 'Advanced program parameters – Filtering parameters' section and set 'Minimum amplicon frequency': 0.4%. These additional parameters were defined by the authors after manual inspection of the data in the original article (see **Note 8**) [37]. Keep the rest of parameters with the default values.
7) Repeat the steps 7-8 listed in **Section 3.1**.

3.3. Interpreting the genotyping results

1) If we have correctly followed the steps from **Sections 3.1** or **3.2**, the output of AmpliSAS should look like this (**Figure 5C**):

```
AmpliSAS results

Download AmpliSAS analysis results.

Analysis details:

Running 'bin/ampliSAS ...

Checking input sequence file ...
    Sequences are in FASTQ format.
    Sequences number: 2589165.

Reading sequence data.

Reading amplicon data from file ...
    Number of markers: 1.

De-multiplexing amplicon sequences from reads.
    MHCI-37 de-multiplexing
    MHCI-37 de-multiplexed (5000 reads, 1283 variants)
    ...

Extracting de-multiplexed sequences into ...

Checking data and setting marker lengths.
    Marker 'MHCI' lengths: 235,238,241 (manual)

Clustering amplicon sequences with the following parameters
('threshold' 'marker' 'values'):
    substitution_threshold             all   1
    indel_threshold                    all   0.001
    cluster_exact_length               all   1
```

```
    min_dominant_frequency_threshold    all   10

    MHCI-BIRD_37 clustering
    MHCI-BIRD_37 clustered (4383 reads, 47 variants)
    ...

Printing information about clustered and not clustered sequences into ...

Filtering sequences with the following criteria ('filter' 'marker' 'values'):
    min_amplicon_depth            all   100
    min_amplicon_seq_frequency    all   0.4
    min_chimera_length            all   10
    max_allele_number             all   50

    MHCI-BIRD_37 filtering
    MHCI-BIRD_37 filtered (4344 reads, 36 variants)
    ...

Printing information about filtered and non-filtered sequences into ...

Reads per amplicon:
Amplicon  Total Unique      Reads-clustered   Variants-clustered     Reads-
filtered  Variants-filtered
MHCI-BIRD_37    5000  1283  4383  47    4344  36
MHCI-BIRD_38    5000  1315  4513  50    4450  33
MHCI-BIRD_39    5000  1190  4744  35    4723  33
MHCI-BIRD_40    5000  1216  4600  47    4586  41
MHCI-BIRD_41    5000  1301  4597  38    4573  30

Printing amplicon data into ...

Analysis results stored into ...
```

2) Click the ‘Download AmpliSAS analysis results’ link to download a ZIP compressed file with the following contents:

    - ‘results.xlsx’: Excel file with the final genotyping results. See step 3.
    - ‘allseqs’, ‘clustered’ and ‘filtered’ folders: contain single amplicon FASTA files with variants recovered after every analysis step: de-multiplexing, clustering and filtering respectively. Each variant has annotated in the FASTA header its sequencing depth and frequency. All this data is also included into an Excel file per folder.
    - ‘amplicon_data.csv’: comma-separated values format file including the amplicon data and analysis parameters.
    - ‘summary.txt’: a tab-delimited file with the number of variants and associated reads retrieved after every analysis step.

3) The most informative file is ‘results.xlsx’, samples (individuals) are shown in columns and variants (alleles) in rows, the numeric values represent the variants’ depths within each amplicon (**Figure 6)**. For example, the variant MHCI-0001 is present in the five individuals, having in the BIRD_39 the maximum depth (914

reads) and the variant MHCI-0002 is present in four individuals but not in the BIRD_39. The columns at the left include additional information about the variants: DNA sequence, length, sum of the depths in all the samples, number of samples containing the variant and mean, maximum and minimum frequencies along all the samples.

	BIRD_37	BIRD_38	BIRD_39	BIRD_40	BIRD_41
MHCI-0001	381	375	914	601	411
MHCI-0002	190	525		138	180
MHCI-0003	144		170	256	392
MHCI-0006		398	298		
MHCI-0004	172	186		132	203
MHCI-0007		184	236		266
MHCI-0005	177			291	206

Figure 6. AmpliSAS Excel output file example. The numeric values show the variants' depths within each amplicon.

3.4. Comparing two genotyping result files

In the previous **Sections 3.1 and 3.2** we obtained slightly different genotyping results due to the auto vs. manual adjustment of the genotyping parameters (see **Note 8**), here we will learn how to compare them with the tool AmpliCOMPARE.

1) Open the AmpliCOMPARE online submission form: http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplicompare
2) Upload the 'results.xlsx' file obtained in the **Section 3.1** as 'First results file' and the file from the **Section 3.2** as 'Second results file'.
3) Click the 'Run' button and follow the link to the results page.
4) The output of the comparison process will be like this:

```
AmpliCOMPARE results

Download AmpliCOMPARE analysis results.

Analysis details:

Running 'bin/ampliCOMPARE.pl ...

Reading File ...

MARKER 'MHCI':
Total unique samples: 5 (file1: 5, file2: 5)
Total seqs: 117 (file1: 117, file2: 114)
Compared samples: 5 (excluded from file1: 0, from file2: 0)
Compared seqs: 117 (missing in file1: 0, in file2: 3)
Total assignments: 176 (missing in file1: 1, missing in file2: 3)

Comparison results written into ...
```

5) Click the link 'Download AmpliCOMPARE analysis results' to retrieve an Excel file with the differences between both genotypes.

6) Open the Excel file, variants (alleles) retrieved in the first analysis but not in the second are marked in cyan color and the opposite in magenta (**Figure 7**). Common variants in both analysis remain un-formatted with their depths separated by a slash (**Figure 7**). There should be three variants marked in cyan retrieved with automatic parameters (**Section 3.1**) and not retrieved with manual ones (**Section 3.2**). If we check the lengths of the different variants, they are 230 and 236 bp long, consequently they are not in-frame with the real allele lengths (235, 238 and 241) that we specified manually in the **Section 3.2** analysis. Two of three variants have low depths, so most probably they are PCR or sequencing artefacts derived from other, higher frequency alleles. The third variant could be a product of non-specific amplification, or a pseudo-gene, rather than a technical artefact. As conclusion, both genotyping strategies perform well, but the manual adjustment of AmpliSAS parameters retrieves higher quality genotypes.

	BIRD_37	BIRD_38	BIRD_39	BIRD_40	BIRD_41
MHCI-0001	364/381	366/375	887/914	577/601	396/411
MHCI-0002	177/190	511/525		130/138	168/180
MHCI-0003	137/144		161/170	250/256	383/392
...					
MHCI-0022	112/122				130/133
MHCI-0026	234				
MHCI-0028			220/228		
...					
MHCI-0048				134/136	
MHCI-0040	43/45	25		36/39	53/55
MHCI-0052			126/131		

Figure 7. AmpliCOMPARE Excel output file example. The numeric values show the variants' depth in the compared files. Cyan color marks a variant that is present in the first file and not in the second, the opposite is marked in magenta.

3.5. HLA typing with amplicon sequencing data

To show the functionality of AmpliHLA we will use a targeted amplicon sequencing dataset that consists of genomic sequences from exon 2 and exon 3 regions from HLA-A and HLA-B loci in five human cell lines sequenced with Illumina MiSeq [33] (ENA study accession: PRJEB4744). The data has been pre-processed for simplicity (see **Note 9**).

1) Open the AmpliHLA online submission form (**Figure 8A**): http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplihla
2) Enter a name for the run and, optionally, an email address if you desire to receive the results by email.
3) Choose 'Amplicons' as the analysis 'Data type'.
4) Copy and paste the following link into the 'Sequence file URL' field: http://evobiolab.biol.amu.edu.pl/amplisat/bin/examples/amplihla_example.fq.gz Optionally, you can download the compressed FASTQ file in your computer and upload it with the 'Browse' button.

5) Select ‘Auto’ from the ‘Technology’ options, the program will automatically detect the sequencing technology used to adjust the variant clustering parameters and it will calculate the optimum frequency thresholds for each marker to filter artefacts.

6) The field ‘Alleles’ shows information about the reference sequences used by AmpliHLA to identify the HLA alleles.

7) Copy and paste in the ‘Amplicon data’ field the following information about primers and tags for the 5 human cell lines sequenced (see **Note 6**):

```
>marker,length,primer_f,primer_r
HLA_A2,344,CRGGTCTCAGCCACTSCTC,CTCGGACCCGGAGACTGT
HLA_A3,353,CTYGGGGGACYGGGCTGAC,CCCAATTGTCTCCCCTCCTTG
HLA_B2,391,GGSAGGGAAATGGCCTCT,GGATGGGGAGTCGTGACCT
HLA_B3,385,GCGTTTACCCGGTTTCATT,CGGCGACCTATAGGAGATGG
>sample,tag_f
C1Rneo,GTCGTA
Daudi,AAGCGA
HEK293,TGTCTC
NCI_H929,GGTGCT
Raji,TGCGAG
```

9) Click the ‘Run’ button and wait until the analysis is completed as explained in the steps 7-8 of the **Section 3.1** (**Figure 8B**). Keep the results page open, in the next section we will learn how to interpret them.

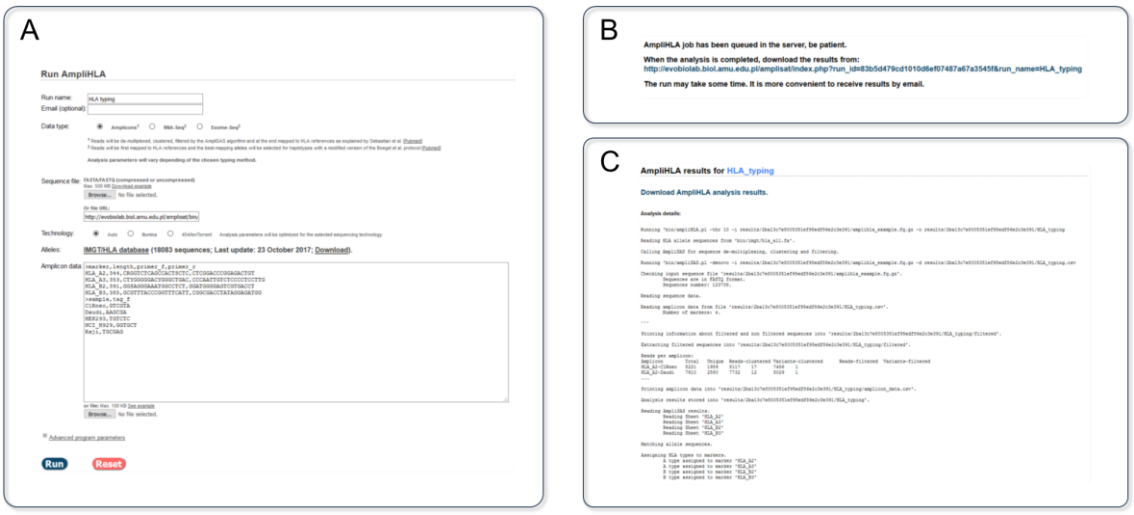


Figure 8. A: AmpliHLA input form. B: Message with the link to the results page. C: Output summary after analysis completion and link to the results ZIP file.

3.6. Interpreting the HLA typing results

1) If we have correctly followed the steps from **Section 3.5**, the AmpliHLA output should look like this (**Figure 8C**):

```
AmpliHLA results for test

Download AmpliHLA analysis results.

Analysis details:


Running 'bin/ampliHLA.pl ...

Reading HLA allele sequences ...

Calling AmpliSAS for sequence de-multiplexing, clustering and filtering.

Running 'bin/ampliSAS.pl ...

... AMPLISAS OUTPUT ...

Reading AmpliSAS results.
     Reading Sheet 'HLA_A2'
     Reading Sheet 'HLA_A3'
     Reading Sheet 'HLA_B2'
     Reading Sheet 'HLA_B3'

Matching allele sequences.

Assigning HLA types to markers.
     A type assigned to marker 'HLA_A2'
     A type assigned to marker 'HLA_A3'
     B type assigned to marker 'HLA_B2'
     B type assigned to marker 'HLA_B3'
```

2) Click the 'Download AmpliHLA analysis results' link to download a ZIP compressed file including several files and folders as explained in the Section 3.3, step 2.
3) Open the 'results.xlsx' file to check the assigned genotypes, individuals are shown in columns and alleles in rows, the numeric values represent the average frequencies of the alleles within the amplicons (**Figure 9)**. Genotypes are given with the highest resolution that can be achieved by the program (maximum 4-digits), it will depend of the number and length of HLA regions (markers) sequenced in the experiment (in the example two exonic regions per locus). Sometimes a variant shares the same identity with several alleles (e.g. Daudi A*66:01) or the type cannot be resolved with 4-digit resolution (e.g. C1Rneo

A*02), then a list of allele ambiguities is listed below the table (**Figure 9**). At the left column 'SEQUENCES' there is a list of the variant sequences that match a particular allele.

8) If we compare AmpliHLA typing RESULTS with the expected ones in the **Table 3** validated by Sanger sequencing [33], we observe that AmpliHLA has a 95% of accuracy in assigning genotypes with 2-digit resolution for both loci (the only error is the ambiguous assignment of the A*66:01 Daudi allele, **Figure 9**) and 70% of accuracy with 4-digit resolution. Nevertheless, resolution could be improved by sequencing additional regions of the loci or selecting from the ambiguities the most frequent human alleles for the studied population.

Table 3. Correspondence of human cell lines and HLA types determined by Sanger sequencing.

HLA Class I		
C1Rneo	A*02:01	B*35:03
Daudi	A*01:02 A*66:01	B*58:01 B*58:02
HEK293	A*02:01 A*03:01	B*07:02
NCI-H929	A*03:01 A*24:02	B*07:02 B*18:01
Raji	A*03:01	B*15:10

HLA Class II			
Daudi	DQA1*01:02 DQA1*01:03	DQB1*06:02 DQB1*06:04	DRB1*13:01 DRB1*13:02
Raji	DQA1*01:01 DQA1*05:01	DQB1*02:01 DQB1*05:01	DRB1*03:01 DRB1*10:01

	C1Rneo	Daudi	HEK293	NCI_H929	Raji
A*03:01			0,4	0,34	0,89
A*02	0,9				
A*01:02		0,54			
A*24:02				0,51	
A*02:01			0,44		
A*66:01 | ...		0,17			

ALLELE	AMBIGUITIES
A*02	A*02:01, A*02:03, A*02:04, A*02:07, A*02:16, A*02:17, A*02:22, A*02:24, A*02:2
A*66:01 | ...	A*25:01, A*26:01, A*26:03, A*43:01, A*66:01

Figure 9. AmpliHLA Excel output file example. Every sample has assigned one or two HLA alleles, numeric values show the amplicon frequencies of the alleles.

3.7. HLA typing with RNA-Seq data

AmpliHLA functionality is not restricted to NGS amplicon data, in the present protocol we will analyze an RNA-Seq experiment from the Daudi cell line (ENA run accession SRR387401 from the study SRP009316) [43].

1) Open the AmpliHLA online submission form (**Figure 8A**): http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplihla
2) Enter a name for the run and, optionally, an email address if you desire to receive the results by email.
3) Select 'RNA-Seq' as the 'Data type'.
4) Copy and paste the following link into the 'Sequence/reads file URL' field: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR387/SRR387401/SRR387401_1.fastq.gz and into the 'Paired-end reads file URL' field: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR387/SRR387401/SRR387401_2.fastq.gz
5) The field 'Alleles' shows information about the reference sequences used by AmpliHLA to identify the HLA alleles.
6) Click the 'Run' button and wait until the analysis is completed as explained in the steps 7-8 of the **Section 3.1**. The output should look like this:

```
AmpliHLA results

Analysis details:

Running 'bin/ampliHLA.pl ...

Retrieving allele information from BAM data.

Parsing reference file ...

Parsing alignment file ...

74488 reads mapped to 5702 reference alleles.

5702 alleles are kept after filtering artefacts.

RESULTS:

LOCI   ALLELE     SCORE  READS  DISAMBIGUATION

A      A*66:01     48.4   4038
       A*01:02     36.4   3041

B      B*58:01     57.9  16704
```

```
        B*58:02     22.4   6466

C       C*03:02     35.2   8417
        C*06:02     33.8   8074

DQA1    DQA1*01:03  94.5   3473

DQB1    DQB1*06:13  50.4   1997

DRB1    DRB1*13:02  62.3   7820
```

7) Typing results are printed in five columns with the following information: HLA locus, assigned alleles, confidence score (allele-specific reads divided by the total number of mapped reads in the locus), corrected allele coverage (see explanation in **Section 2.2**) and allele ambiguities in case the allele has not been unequivocally assigned.

8) Comparing AmpliHLA results with laboratory validated HLA types for the Daudi cell line **(Table 3**), the genotyping accuracy is of 100% for the MHC class I loci with 4-digit resolution (HLA-A, B and C). Instead, MHC class II loci (HLA-DQA1, DQB1 and DRB1) are typed at 33% of accuracy with 4-digit resolution. If we look at the 2-digit resolution genotypes, all of them have an accuracy of 100%. The difference in accuracy between the class I and class II genotypes is explained because class II genes have only one variable exon and the high similarity among allele sequences makes their genotyping more problematic.

The previous analysis can be replicated for any pair of FASTQ files from the ENA project SRP009316 (https://www.ebi.ac.uk/ena/data/view/SRP009316). For example, the RNA-Seq data from the Raji cell line (ENA run accession SRR387394):

9) Repeat steps 1-7 replacing ‘SRR387401’ with ‘SRR387394’ in the reads file URLs at step 4. Analysis results will look like this:

```
LOCI    ALLELE      SCORE  READS  DISAMBIGUATION

A       A*03:01      69.3   2024

B       B*15:10      92.5   8318

C       C*04:01      47.9   3248
        C*03:04      29.3   1983
```

```
DQA1    DQA1*05:01  63.7   1911
        DQA1*01     36.1   1083  DQA1*01:01,DQA1*01:04

DQB1    DQB1*05:01  55.4    691
        DQB1*02     44.6    556  DQB1*02:01,DQB1*02:02

DRB1    DRB1*10:01  48.6   3700
        DRB1*03:01  48.5   3687
```

10) Comparing retrieved genotypes with Raji expected ones (**Table 3**), we obtain a 100% of accuracy for the 4-digit class I genotypes and the same for the 2-digit class II ones as previously stated by the Seq2HLA authors after the analysis of the same dataset [44].

3.8. HLA typing with exome sequencing data

In this final protocol, HLA typing will be performed with whole exome sequencing (WES) data from a Daudi cell line (see **Note 10**) [45].

1) Open the AmpliHLA online submission form (**Figure 8A**): http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplihla
2) Enter a name for the run and, optionally, an email address if you desire to receive the results by email.
3) Select 'Exome-Seq' as the 'Data type'.
4) Copy and paste the following link into the 'Sequence/reads file URL' field: http://evobiolab.biol.amu.edu.pl/amplisat/examples/DAUDI_R1.fq.gz and into the 'Paired-end reads file URL' field: http://evobiolab.biol.amu.edu.pl/amplisat/examples/DAUDI_R2.fq.gz
   Optionally, you can download the compressed FASTQ file in your computer and upload it with the 'Browse' button.
5) The field 'Alleles' shows information about the reference sequences used by AmpliHLA to identify the HLA alleles.

6) Click the ‘Run’ button and wait until the analysis is completed as explained in the steps 7-8 of the **Section 3.1**. The output should look like this:

```
AmpliHLA results

Analysis details:

Running 'bin/ampliHLA.pl ...

Retrieving allele information from BAM data.

Parsing reference file ...

Parsing alignment file ...

543 reads mapped to 1111 reference alleles.

1111 alleles are kept after filtering artefacts.

RESULTS:

LOCI   ALLELE     SCORE  READS  DISAMBIGUATION

A      A*66:01     51.5     69
       A*01:02     47.8     64

B      B*58:01     96.3     78

C      C*06        51.0     49  C*06:02,C*06:04

DQA1   DQA1*01:02  89.2    149

DQB1   DQB1*06:03  90.6     29

DRB1   DRB1*13     91.7     22  DRB1*13:01,DRB1*13:02
```

7) Typing results are printed in five columns as explained in the step 7 at **Section 3.7**.

8) Comparing AmpliHLA results with laboratory validated HLA types for the Daudi cell line **(Table 3**), the genotypes for the six MHC class I and class II loci evaluated have an accuracy of 92% and 33% for 2-digit and 4-digit resolutions respectively. Accuracy is noticeably lower than in **Section 3.7** because there are far fewer WES reads mapping to HLA references (dozens) than RNA-Seq reads in the previous protocol (thousands).

4. Notes

1. Basically, there are 4 main steps in the NGS amplicon sequencing workflow (**Figure 2**):
    i. Design of the primers to amplify the desired gene regions (markers).
    ii. Library preparation by PCR amplification of the selected regions, addition of sample-specific DNA tags and of platform-specific sequencing adaptors.
    iii. NGS sequencing of the amplification products. The most commonly used platforms are: Illumina, Ion Torrent and previously 454.
    iv. Bioinformatic analysis of the sequencing data. The analysis should include: classification of reads into amplicons, sequencing error correction, filtering of spurious and contaminant reads, and final displaying of results in a human readable way, e.g. an Excel spreadsheet.

    For a list of definitions of commonly used terms in amplicon sequencing see the **Table 2**.

    In the following link you will find a video explaining the amplicon sequencing process using NGS in a metagenomics experiment:

    http://www.jove.com/video/51709/next-generation-sequencing-of-16s-ribosomal-rna-gene-amplicons

2. Before NGS technologies were available, PCR products were Sanger sequenced individually. Sanger sequencing is only able to resolve one DNA sequence (allele) per sample. In special cases, a mix of two alleles is also

possible (if they differ only by one nucleotide position (i.e. heterozygous individual at given locus). If a primer pair amplifies more than one locus (as it is often the case in MHC genotyping of non-model organisms) the only way to distinct multiple, mixed sequences was to clone sequences to bacterial vectors and further isolation, amplification and sequencing of individual clones. Nevertheless, bacterial cloning is a time-consuming and error prone approach that is only feasible with few dozens of sequences.

Fortunately, NGS techniques are able to sequence millions of sequences with individual resolution. The combination of amplicon sequencing with NGS allows us to genotype hundreds/thousands of samples in a single experiment. The only requirement is to include different DNA tags to identify the individuals/samples in the experiment. A DNA tag is a short and unique sequence of nucleotides (e.g. ACGGTA) that is either ligated to a PCR product or attached at the end of one of the PCR primers (**Figure 2**). Tags have to be unique for each sample/individual to enable assignment of the reads back to the original amplicon (individual or sample) [34, 35].

However, the NGS techniques have some limitations: the lengths of the sequences are shorter than in Sanger sequencing and frequent sequencing errors result in a high number of artefacts. To alleviate those shortcomings, long sequences can be fragmented and assembled together later by computer and increasing the depth/coverage ("reading" more times the same sequence) can correct random sequencing errors.

3. Homopolymer regions are a major issue for pyrosequencing and ion semiconductor NGS technologies (454 and Ion Torrent, respectively), where erroneous indels are introduced in high rates. Technology based on

reversible dye-terminators (Illumina) suffers from a high number of mostly random substitutions [46–51]. PCR products also incorporate polymerase substitution errors and chimeras (sequences formed from two different sequences due to incomplete primer extension) [52].

4. Four different genotyping approaches were quantitatively evaluated for removing artefacts from NGS amplicon data and assigning MHC class I alleles in a set of sedge warbler individuals [37]. Among the four methods considered, AmpliSAS retrieved accurate, repeatable genotypes requiring lower coverages than the others. Furthermore, AmpliSAS supports different NGS platforms data and it is available as a web server.
5. Usually an amplicon sequencing experiment sequenced with Illumina technology produces paired-end reads that should be cleaned and merged/overlapped before further processing. In the presented example, both steps were skipped for simplicity and reads are ready-to-use. Paired-end read overlapping and read cleaning were performed with the tools AmpliMERGE and AmpliCLEAN respectively, both are part of the AmpliSAT suite (see Materials section).
6. If reads have been separated into multiple files after sequencing, one file per amplicon, they can be packed into a single ZIP or TGZ format file and used as input. In such case the ‘amplicon data’ field should be empty, AmpliSAS will use the folder and filenames to name the markers and amplicons respectively (Example of organization of reads files into the packaged file: ./MARKER/SAMPLENAME.FASTQ).
7. Adjusting analysis parameters is important because error profiles are affected by many factors: the sequencing platform, length of the amplicon,

number of co-amplifying alleles, amplification bias introduced by each set of primers, etc. Specifically, frequency thresholds that separate genuine alleles and technical artefacts may vary between experimental setups, and should be carefully adjusted by the researcher.

8. Sedge warbler MHC class I amplicons have a major length of 241, but there are minor variants of 238 and 235 bp experimentally validated [8]. If we do not specify the lengths, AmpliSAS automatically sets the value to 241 bp and it will not detect the variants with 3 and 6 bp in-frame deletions.
9. The amplicon sequencing data from the five human cell lines that will be used in this example has been pre-processed for simplicity: paired-end reads have been merged and sample-specific DNA tags have been artificially attached to the forward primers.
10. WES data has been kindly provided by R. Siebert, A. Franke and G. Hemmrich-Stanisak from their original article [45]. To save time in the analysis, the reads have been previously aligned to HLA genomic references with BOWTIE [42] and the mapped reads extracted with 'SamToFastq' command from Picard Tools suite [53]. As a result, only few hundreds of paired-end reads from the initial 34 millions have been saved into two FASTQ files that are used as input in the AmpliHLA procotol.

5. References

1. Murphy KM, Travers P, Walport M (2007) Janeway's Immunobiology, 7 edition. Garland Science, New York
2. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE (2015) The IPD and IMGT/HLA database: Allele variant databases. Nucleic Acids Res 43:D423–D431. doi: 10.1093/nar/gku1161
3. Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F (2005) Pathogen-driven selection and worldwide HLA class I diversity. Curr Biol 15:1022–7. doi: 10.1016/j.cub.2005.04.050
4. Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc Biol Sci 277:979–88. doi: 10.1098/rspb.2009.2084
5. Sommer S (2005) The importance of immune gene variability (MHC) in evolutionary ecology and conservation. Front Zool 2:16. doi: 10.1186/1742-9994-2-16
6. Apanius V, Penn D, Slev PR, Ruff LR, Potts WK (1997) The nature of selection on the major histocompatibility complex. Crit Rev Immunol 17:179–224.
7. Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: What have we learned about natural selection in 15 years? J Evol Biol 16:363–377. doi: 10.1046/j.1420-9101.2003.00531.x
8. Biedrzycka, Aleksandra O'Connor E, Migalska M, Radwan J, Zając T, Bielański W, Solarz W, Ćmiel A, Westerdahl H (2017) Extreme MHC class I diversity in the sedge warbler (Acrocephalus schoenobaenus); selection patterns and allelic divergence suggest that different genes have different functions. BMC Evol Biol 17:159. doi: 10.1186/s12862-017-0997-9
9. Wiseman RW, Karl J a, Bohn PS, Nimityongskul F a, Starrett GJ, O'Connor DH (2013) Haplessly hoping: macaque major histocompatibility complex made easy. ILAR J 54:196–210. doi: 10.1093/ilar/ilt036
10. Sato A, Dongak R, Hao L, Takezaki N, Shintani S, Aoki T, Klein J (2006) Mhc class I genes of the cichlid fish Oreochromis niloticus. Immunogenetics 58:917–928. doi: 10.1007/s00251-006-0151-0
11. Stutz WE, Bolnick DI (2014) Stepwise Threshold Clustering: A New Method for Genotyping MHC Loci Using Next-Generation Sequencing Technology. PLoS One 9:e100587. doi:

10.1371/journal.pone.0100587

12. Migalska M, Sebastian A, Konczal M, Kotlík P, Radwan J, Kotlik P, Radwan J (2017) De novo transcriptome assembly facilitates characterisation of fast-evolving gene families, MHC class I in the bank vole (Myodes glareolus). Heredity (Edinb) 118:348–357. doi: 10.1038/hdy.2016.105
13. Figueroa F, Mayer W, Sato A, Zaleska-Rutczynska Z, Hess B, Tichy H, Klein J (2001) Mhc class I genes of swordtail fishes, Xiphophorus : variation in the number of loci and existence of ancient gene families. Immunogenetics 53:695–708. doi: 10.1007/s00251-001-0378-8
14. Mehra NK (2001) Histocompatibility Antigens. Encicl. Life Sci.
15. Trowsdale J, Campbell RD (2001) Mouse MHC Genes and Products. In: Curr. Protoc. Immunol. John Wiley & Sons, Inc., Hoboken, NJ, USA, p Appendix 1L
16. Lukacs MF, Harstad H, Grimholt U, Beetz-Sargent M, Cooper GA, Reid L, Bakke HG, Phillips RB, Miller KM, Davidson WS, Koop BF (2007) Genomic organization of duplicated major histocompatibility complex class I regions in Atlantic salmon (Salmo salar). BMC Genomics 8:251. doi: 10.1186/1471-2164-8-251
17. Kaufman J, Milne S, Göbel TW, Walker BA, Jacob JP, Auffray C, Zoorob R, Beck S (1999) The chicken B locus is a minimal essential major histocompatibility complex. Nature 401:923–925. doi: 10.1038/44856
18. Kelley J, Walter L, Trowsdale J (2005) Comparative genomics of major histocompatibility complexes. Immunogenetics 56:683–95. doi: 10.1007/s00251-004-0717-7
19. Kulski JK, Shiina T, Anzai T, Kohara S, Inoko H (2002) Comparative genomic analysis of the MHC: the evolution of class I duplication blocks, diversity and complexity from shark to man. Immunol Rev 190:95–122. doi: imr19008 [pii]
20. Robinson J, Halliwell J a, McWilliam H, Lopez R, Parham P, Marsh SGE (2013) The IMGT/HLA database. Nucleic Acids Res 41:D1222-7. doi: 10.1093/nar/gks949
21. Wallny H-J, Avila D, Hunt LG, Powell TJ, Riegert P, Salomonsen J, Skjødt K, Vainio O, Vilbois F, Wiles M V, Kaufman J (2006) Peptide motifs of the single dominantly expressed class I molecule explain the striking MHC-determined response to Rous sarcoma virus in chickens. Proc Natl Acad Sci U S A 103:1434–9. doi: 10.1073/pnas.0507386103
22. Livant EJ, Brigati JR, Ewald SJ (2004) Diversity and locus specificity of chicken MHC B class I sequences. Anim Genet 35:18–27.

23. Westerdahl H, Wittzell H, von Schantz T, Bensch S (2004) MHC class I typing in a songbird with numerous loci and high polymorphism using motif-specific PCR and DGGE. Heredity (Edinb) 92:534–42. doi: 10.1038/sj.hdy.6800450

24. Sepil I, Moghadam HK, Huchard E, Sheldon BC (2012) Characterization and 454 pyrosequencing of major histocompatibility complex class I genes in the great tit reveal complexity in a passerine system. BMC Evol Biol 12:68. doi: 10.1186/1471-2148-12-68

25. O'Connor EA, Strandh M, Hasselquist D, Nilsson J-å., Westerdahl H (2016) The evolution of highly variable immunity genes across a passerine bird radiation. Mol Ecol 25:977–989. doi: 10.1111/mec.13530

26. Babik W (2010) Methods for MHC genotyping in non-model vertebrates. Mol Ecol Resour 10:237–51. doi: 10.1111/j.1755-0998.2009.02788.x

27. Babik W, Taberlet P, Ejsmond MJ, Radwan J (2009) New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. Mol Ecol Resour 9:713–9. doi: 10.1111/j.1755-0998.2009.02622.x

28. Radwan J, Zagalska-Neubauer M, Cichoń M, Sendecka J, Kulma K, Gustafsson L, Babik W (2012) MHC diversity, malaria and lifetime reproductive success in collared flycatchers. Mol Ecol 21:2469–2479. doi: 10.1111/j.1365-294X.2012.05547.x

29. Lighten J, van Oosterhout C, Paterson IG, McMullan M, Bentzen P (2014) Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (Poecilia reticulata). Mol Ecol Resour 14:753–767. doi: 10.1111/1755-0998.12225

30. Sommer S, Courtiol A, Mazzoni CJ (2013) MHC genotyping of non-model organisms using next-generation sequencing: a new methodology to deal with artefacts and allelic dropout. BMC Genomics 14:542. doi: 10.1186/1471-2164-14-542

31. Moonsamy P V, Williams T, Bonella P, Holcomb CL, Höglund BN, Hillman G, Goodridge D, Turenchalk GS, Blake L a, Daigle D a, Simen BB, Hamilton a, May a P, Erlich H a (2013) High throughput HLA genotyping using 454 sequencing and the Fluidigm Access Array™ System for simplified amplicon library preparation. Tissue Antigens 81:141–9. doi: 10.1111/tan.12071

32. Erlich RL, Jia X, Anderson S, Banks E, Gao X, Carrington M, Gupta N, DePristo M a, Henn MR, Lennon NJ, de Bakker PIW (2011) Next-generation sequencing for HLA typing of class I loci.

BMC Genomics 12:42. doi: 10.1186/1471-2164-12-42

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

34. Binladen J, Gilbert MTP, Bollback JP, Panitz F, Bendixen C, Nielsen R, Willerslev E (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS One 2:e197. doi: 10.1371/journal.pone.0000197

35. Meyer M, Stenzel U, Myles S, Prüfer K, Hofreiter M (2007) Targeted high-throughput sequencing of tagged nucleic acid samples. Nucleic Acids Res 35:e97. doi: 10.1093/nar/gkm566

36. Lighten J, van Oosterhout C, Bentzen P (2014) Critical review of NGS analyses for de novo genotyping multigene families. Mol Ecol 23:3957–72. doi: 10.1111/mec.12843

37. Biedrzycka A, Sebastian A, Migalska M, Westerdahl H, Radwan J (2017) Testing genotyping strategies for ultra-deep sequencing of a co-amplifying gene family: MHC class I in a passerine bird. Mol Ecol Resour 17:642–655. doi: 10.1111/1755-0998.12612

38. Sebastian A, Herdegen M, Migalska M, Radwan J (2016) Amplisas: A web server for multilocus genotyping using next-generation amplicon sequencing data. Mol Ecol Resour 16:498–510. doi: 10.1111/1755-0998.12453

39. Leclaire S, Strandh M, Mardon J, Westerdahl H, Bonadonna F (2017) Odour-based discrimination of similarity at the major histocompatibility complex in birds. Proceedings Biol Sci 284:20162466. doi: 10.1098/rspb.2016.2466

40. Pardal S, Drews A, Alves JA, Ramos JA, Westerdahl H (2017) Characterization of MHC class I in a long distance migratory wader, the Icelandic black-tailed godwit. Immunogenetics 69:463–478. doi: 10.1007/s00251-017-0993-7

41. Boegel S, Löwer M, Schäfer M, Bukur T, de Graaf J, Boisguérin V, Türeci O, Diken M, Castle JC, Sahin U (2013) HLA typing from RNA-Seq sequence reads. Genome Med 4:102. doi: 10.1186/gm403

42. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. doi: 10.1186/gb-2009-10-3-r25

43. Boegel S, Löwer M, Bukur T, Sahin U, Castle JC (2014) A catalog of HLA type, HLA expression, and neo-epitope candidates in human cancer cell lines. Oncoimmunology 3:e954893. doi:

10.4161/21624011.2014.954893

44. Boegel S, Scholtalbers J, Löwer M, Sahin U, Castle JC (2015) In Silico HLA Typing Using Standard RNA-Seq Sequence Reads. In: Bugert P (ed) Mol. Typing Blood Cell Antigens, Methods Mol. Biol. Springer Science, pp 115–121

45. Kreck B, Richter J, Ammerpohl O, Barann M, Esser D, Petersen BS, Vater I, Murga Penas EM, Bormann Chung CA, Seisenberger S, Lee Boyd V, Smallwood S, Drexler HG, Macleod RAF, Hummel M, Krueger F, Häsler R, Schreiber S, Rosenstiel P, Franke A, Siebert R (2013) Base-pair resolution DNA methylome of the EBV-positive Endemic Burkitt lymphoma cell line DAUDI determined by SOLiD bisulfite-sequencing. Leukemia 27:1751–3. doi: 10.1038/leu.2013.4

46. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB (2013) Characterizing and measuring bias in sequence data. Genome Biol 14:R51. doi: 10.1186/gb-2013-14-5-r51

47. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012:251364. doi: 10.1155/2012/251364

48. Loman NJ, Misra R V, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30:434–9. doi: 10.1038/nbt.2198

49. Vandenbroucke I, Van Marck H, Verhasselt P, Thys K, Mostmans W, Dumont S, Van Eygen V, Coen K, Tuefferd M, Aerssens J (2011) Minor variant detection in amplicons using 454 massive parallel pyrosequencing: experiences and considerations for successful applications. Biotechniques 51:167–77. doi: 10.2144/000113733

50. Gilles A, Meglécz E, Pech N, Ferreira S, Malausa T, Martin J-F (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics 12:245. doi: 10.1186/1471-2164-12-245

51. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW (2013) Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS Comput Biol 9:e1003031. doi: 10.1371/journal.pcbi.1003031

52. Potapov V, Ong JL (2017) Examining Sources of Error in PCR by Single-Molecule Sequencing. PLoS One 12:e0169774. doi: 10.1371/journal.pone.0169774

53. Broad Institiute (2014) Picard tools: Java-based command-line utilities for manipulating SAM files.