AmpliSAT Manual

Amplicon Sequencing Analysis Tools

http://evobiolab.biol.amu.edu.pl/amplisat

v. 8 (24/06/18)

Alvaro Sebastian

Evolutionary Biology Group
Adam Mickiewicz University

CONTENTS

1. What is AmpliSAT?

AmpliSAT are a set of online tools that make easy the analysis of Amplicon Sequencing experiments.

Amplicon Sequencing (AS) is a useful technique in the genotyping task of complex gene families, such as MHC, with multiple loci and copy number variation among individuals (Babik 2010).

AS is widely used for taxonomic classification by amplifying a variety of marker genes: cytochrome c oxidase subunit 1 (CO1), rRNA genes (16S/18S/28S), plant specific genes (rbcL, matK, and trnH-psbA) and nuclear internal transcribed spacers (ITSs) (Sogin *et al.* 2006; Joly *et al.* 2014; Kress *et al.* 2014).

Another novel and promising application of AS is **the study of antibody and T cell receptor (TCR) repertoires** in mammals and birds (Baum *et al.* 2012; Georgiou *et al.* 2014; Ruggiero *et al.* 2015; Robinson 2015).

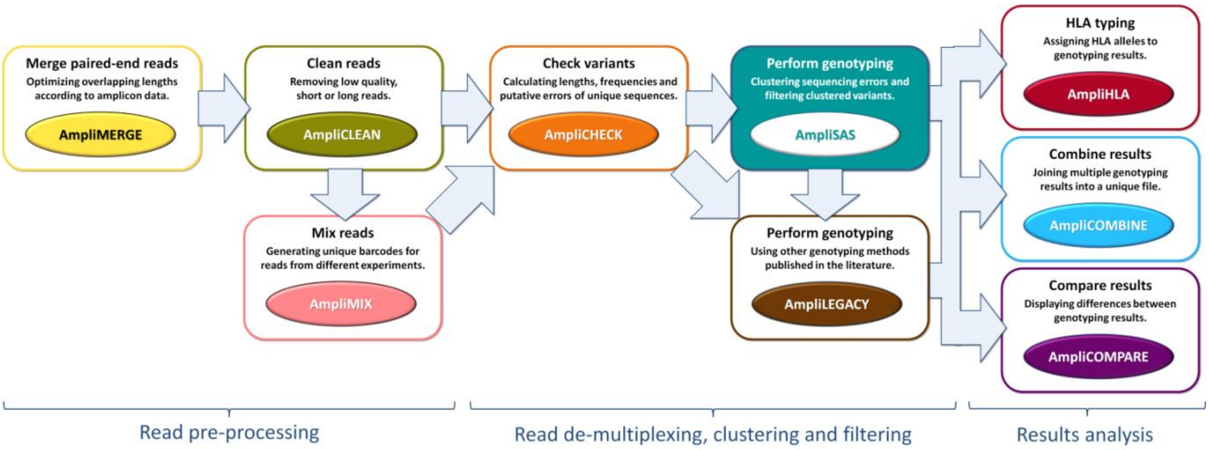
Current tools included in AmpliSAT are:

- AmpliMERGE - Amplicon Sequencing MERGing tool
- AmpliCLEAN - Amplicon Sequencing CLEANing tool
- AmpliMIX - Amplicon Sequencing MIXing tool
- AmpliCHECK - Amplicon Sequencing CHECKing tool
- AmpliSAS - Amplicon Sequencing ASsignment tool
- AmpliLEGACY - Amplicon Sequencing LEGACY genotyping methods
- AmpliCOMPARE - Amplicon Sequencing COMPARE results tool
- AmpliCOMBINE - Amplicon Sequencing COMBINing tool
- AmpliHLA - Amplicon Sequencing HLA typing tool
- AmpliTCR - Amplicon Sequencing T-Cell Receptor segments discovery tool
- AmpliCDR3 – Amplicon Sequencing T-cell CDR3 repertoire analysis tool

Tool descriptions:

- AmpliMERGE: merges paired-end reads, optimizing their overlapping lengths according to amplicon data.
- AmpliCLEAN: removes low quality reads and anomalous short or long ones.
- AmpliMIX: mix reads from different experiments into a unique file unifying barcodes.
- AmpliCHECK: calculates lengths, frequencies and putative errors for the most frequent variants.
- AmpliSAS: helps in genotyping task by clustering sequencing errors with real alleles and filtering PCR artefacts.
- AmpliLEGACY: implements other genotyping methods published in the literature.
- AmpliCOMPARE: displays differences between genotyping results (eg. from replicates or different genotyping parameters).
- AmpliCOMBINE: combines multiple genotyping results.
- AmpliHLA: assigns HLA alleles to human genotyping results.
- AmpliTCR: analyzes a set of genomic or transcriptomic TCR sequences recognizing and extracting their Variable, Joining, Diversity, CDR3 and/or Constant regions.
- AmpliCDR3 – extracts and analyses T-cell receptor (TCR) CDR3 region sequences and provides basic statistics such as number of CDR3 variants retrieved per sample, CDR3 length distribution, repertoire sharing between samples and V-J segment usage.

The following schema shows a typical AS data analysis workflow:


Merge paired-end reads
Optimizing overlapping lengths according to amplicon data.
AmpliMERGE
Clean reads
Removing low quality, short or long reads.
AmpliCLEAN
Mix reads
Generating unique barcodes for reads from different experiments.
AmpliMIX
Check variants
Calculating lengths, frequencies and putative errors of unique sequences.
AmpliCHECK
Perform genotyping
Clustering sequencing errors and filtering clustered variants.
AmpliSAS
Perform genotyping
Using other genotyping methods published in the literature.
AmpliLEGACY
HLA typing
Assigning HLA alleles to genotyping results.
AmpliHLA
Combine results
Joining multiple genotyping results into a unique file.
AmpliCOMBINE
Compare results
Displaying differences between genotyping results.
AmpliCOMPARE
Read pre-processing
Read de-multiplexing, clustering and filtering
Results analysis

First stage in AS is the experimental design: identify the target regions of the desired genes to analyze (markers) and design primers to amplify them. Unique short DNA sequences (barcodes) are added to the primers before or after PCR amplification to unambiguously identify each sample. An amplicon will be a set of sequences derived from a single PCR (one marker and one sample).

Second, amplicons obtained in the wet-lab experiments are sequenced using any of the available technologies. Next generation sequencing (NGS) technologies retrieve thousands to several millions of sequences (reads) from a single AS experiment. The main advantage of NGS is its ability to differentiate different co-amplifying alleles from one or more loci as individual DNA molecules are sequenced separately. The main disadvantages are that retrieved reads are usually short (<300bp) and contain a variable percentage of erroneous nucleotides.

Some NGS technologies can read amplicon sequences from both sides (covering partially or totally the amplified region) retrieving paired-end reads that overlap (e.g. Illumina Mi-Seq). These paired-end reads must be merged into a unique sequence before continuing the data analysis. The merging process will correct errors in low quality sequenced positions and will increase the length of the sequence if it was not fully covered by any of the paired-end reads. **AmpliMERGE** is the suite tool that optimizes overlapping parameters and merges paired-end reads.

Sequenced data (merged or not) may contain reads from other experiments (and sequenced together, usually to decrease costs), **and low quality or abnormally short/long ones** that it is recommended to remove before continuing further analysis to save disk space and time. For this purpose, **AmpliCLEAN** will separate experiment -specific reads complying the desired length and quality parameters.

If different experiments have been carried out and their sequenced data need to be combined and analyzed together, the tool **AmpliMIX** will unify the reads into a unique dataset and will generate new unique barcode sequences in case of redundancy.

Before optimizing parameters for accurate genotyping, we should **know the length, coverage and frequency of the most abundant variants in each amplicon** and the potential erroneous ones (artefacts derived from sequencing or PCR errors). Because usually multiple reads correspond to the same sequence, a variant is defined as a unique sequence. **AmpliCHECK** de-multiplexes reads and performs a fast analysis of the higher frequency variants in each individual amplicon retrieving the previously mentioned information.

After running AmpliCHECK we should be able to establish the length of the desired PCR products (markers), the error rate of the sequencing technique and a threshold frequency to decide if a variant is real or is an artefact. Then we can run **AmpliSAS** to **perform an exhaustive analysis and genotyping**. AmpliSAS workflow is divided into three main steps that will be explained in detail in Chapter **Error! Reference source not found.**: 1) sequence de-multiplexing; 2) sequence clustering; and 3) artefact filtering. **After an ideal analysis, artefacts will be removed and real variants will increase their coverages integrating sequencing errors being able to assign alleles to each amplicon.** AmpliSAT is our best tested and recommended tool, but **AmpliLEGACY offers the possibility to do a similar analysis using other genotyping strategies from the literature**.

If we are interested in genotyping human data, **AmpliHLA will retrieve the human alleles and their frequencies** for each individual taking into account the information from several amplified fragments of the same gene (when possible). As human alleles can be described with several levels of resolution (Nunes *et al.* 2011), the highest one will be given when unequivocal allele assignment is possible, if not the maximum one together with the multiple allele ambiguous assignments.

Finally, it is possible to **compare or combine multiple genotyping results using the tools AmpliCOMPARE and AmpliCOMBINE respectively.** AmpliCOMPARE annotates the differences between genotyping results from two experimental or technical replicates, or using different clustering/filtering parameters. AmpliCOMBINE allows to merge multiple genotyping results unifying the names and statistics of the retrieved alleles.

2. Useful concepts and definitions

Here is a list of the most common terms used in AS analysis:

Table 1. Definitions of commonly used terms in amplicon sequencing and genotyping studies. 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).
Barcode / Molecular Identifier Tag (MID)	A unique short DNA sequence that identifies unambiguously a sample. Barcodes 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 of 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.
Variant frequency or Per-Amplicon Frequency (PAF)	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 / Artefactual sequence	Variant resulting from experimental/technical errors: sequencing errors, polymerase errors, non-specific amplifications (paralogs, pseudogenes), contaminants, PCR chimeras, etc.
Read de-multiplexing	Classification of reads into amplicons based on predesignated primer and barcode sequences and assignment of reads to variants annotating their coverages.
Variant clustering	Removal of artefacts by grouping them with the true variant from which they are derived increasing the coverage of the variant.
Variant filtering	Removal of artefacts by excluding variants from the amplicon based on different criteria: low frequency, low coverage, wrong length, frameshifts, chimeras, etc.

3. AMPLICON DATA FORMAT

Several AmpliSAT tools require as input the amplicon data, which is the information about the primers used to generate the PCR products and the barcode sequences used to differentiate the samples.

Here is an example of amplicon data format accepted by AmpliSAT:

```
>marker,length,primer_f,primer_r,gene,feature,species
MHC2_RET,214-220,GTTGTGTCTTTARCTCSHCTG,ATCGGCTCACCTGATHTA,MHCII,exon2,P. reticulata
MHC2_SUS,200 203,GAGTGTCATTTCTCCAACGGGA,TCACCTCTCCGCTCCACAGTGAA,DRB MHC2,exon 2,S.suslicus
MHC2_SUS,200 203,GAGTGTCATTTCTCCAACGGGA,TCACCTCTCCTCTCCACAGTGAA,DRB MHC2,exon 2,S.suslicus

>sample,barcode_f,barcode_r
RET01,TTCTCG,AACCGA
RET02,AGTGTT,AACGCG
SUS01,AACGCG
SUS02,TCACTC
SUS03,,CTTGGT
SUS04,,CGTCAC
```

Legend:

```
primer_f: Forward primer      barcode_f: Forward barcode
primer_r: Reverse primer      barcode_r: Reverse barcode
```

The format must follow these rules:

- Field names are on the first line after the character ‘>’ and on the next rows are their values in the same order.
- Field names and values must be separated by commas (comma-separated values, CSV format).
- Primer and barcode sequences must be in 5'->3' sense. Primer sequences can contain IUPAC ambiguity codes (e.g. R=A/G), also several primer pairs can be specified for a unique marker .
- The marker length is the length of the PCR product excluding barcodes and primers. Several lengths (200 203) or a range of lengths (214-220) can be specified. AmpliCHECK tool can be used to retrieve these values.
- Sample names must be unique (1st column).
- Single barcodes or pairs of barcodes (forward and reverse) can be specified in the correct column.

4. AmpliSAT data example

In the present section, we will explore the basic functionalities of AmpliSAT with a real case (data kindly provided by M. Herdegen). **The example is a genotyping experiment of a wild population of speckled ground squirrel** (*Spermophilus suslicus*). A 203bp long (excluding primers) exon 2 fragment of the MHC class II DRB locus was amplified in 5 individuals and sequenced by Illumina MiSeq obtaining 150bp paired-end reads. Forward primer: SusL1 (5'-GAGTGTCATTTCTCCAACGGGA-3'); reverse primer: SusR2 (5'-TCACCTCTCCKCTCCACAGTGAA-3'); 6bp long barcodes were added to both primer sequences to differentiate the individuals.

Few ground **squirrel allele sequences** were previously retrieved by single strand conformational polymorphism (SSCP) isolation and Sanger sequencing (GenBank accession nos. EF569186–EF569201, download), also primers and PCR conditions are described in the literature (Biedrzycka & Radwan 2008; Biedrzycka *et al.* 2011).

The two compressed FASTQ files used in the present example can be downloaded here: file 1, file 2. They contain a representative subset of 10000 paired-end reads from the original experiment

5. Merging reads with AmpliMERGE

Considering that the Illumina reads are 150bp long and the DRB gene amplified fragment 203bp, we will need to merge the paired-end reads to obtain the full amplified product. If we add the lengths of the primers (22+23bp) and barcodes (6+6bp) to the fragment length (203bp), a 260bp fragment must be reconstructed between the two paired-end reads. The total length covered by 2 paired-end reads is 300bp, so the theoretical overlap between them should be of 40bp (300-260bp). See the example in Figure 1.

```
Read 1:

TCACTCGAGTGTCATTTCTCCAACGGGACGGAGCGGGTGCGGTTCCTGGAGAGACACTTCTACAACCGGGAGGCGAACGTGCGCTTC
GACAGCGACGTGGGGGAGTTCCGCGCGGTGAGCGAGCTGGGGCGGCCGGACGCCGAGTACTGG

Read 2 (reverse complementary):

CGCGGTGAGCGAGCTGGGGCGGCCGGACGCCGAGTACTGGAACAGCCAGAAGGACTTCCTGGAGGGGAGGCGGGCCGCGGTGGACAA
CTACTGCCGACACAACTACGGGGTTGGTGAGAGCTTCACTGTGGAGAGGAGAGGTGAGTGACG

Merged:

TCACTCGAGTGTCATTTCTCCAACGGGACGGAGCGGGTGCGGTTCCTGGAGAGACACTTCTACAACCGGGAGGCGAACGTGCGCTTC
GACAGCGACGTGGGGGAGTTCCGCGCGGTGAGCGAGCTGGGGCGGCCGGACGCCGAGTACTGGAACAGCCAGAAGGACTTCCTGGAG
GGGAGGCGGGCCGCGGTGGACAACTACTGCCGACACAACTACGGGGTTGGTGAGAGCTTCACTGTGGAGAGGAGAGGTGAGTGACG


Barcode Fwd (6bp): TCACTC      Primer Fwd (22bp): GAGTGTCATTTCTCCAACGGGA
Barcode Rev (6bp): GTGACG      Primer Rev (23bp): TTCACTGTGGAGAGGAGAGGTGA
Overlapping region (40bp): CGCGGTGAGCGAGCTGGGGCGGCCGGACGCCGAGTACTGG
```

Figure 1. Example of paired-end reads before and after merging. Barcodes, primers and overlapping region are colored.

To merge the paired-end reads contained in the two FASTQ files follow these steps:

1. Go to AmpliMERGE web form (Figure 2A): http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplimerge

2. Give a name to the run and write your email address (optional) if you want to receive a link to the results by email.

3. Give the location of the two files with the paired-end reads to merge clicking on the 'Browse…' button (download file 1, file 2).

4. Introduce the following information into the 'Amplicon data' textbox:

   ```
   >marker,length,primer_f,primer_r,gene,feature,species
   MHC2,203,GAGTGTCATTTCTCCAACGGGA,TCACCTCTCCKCTCCACAGTGAA,DRB MHC2,exon 2,Spermophilus
   suslicus
   ```

   Remember to write the field names on the first line after the character '>' and on the next rows write the values in the same order than the field names. Field names and values must be comma-separated.

5. Click on the 'Run' button at the bottom of the page to start the merging process.

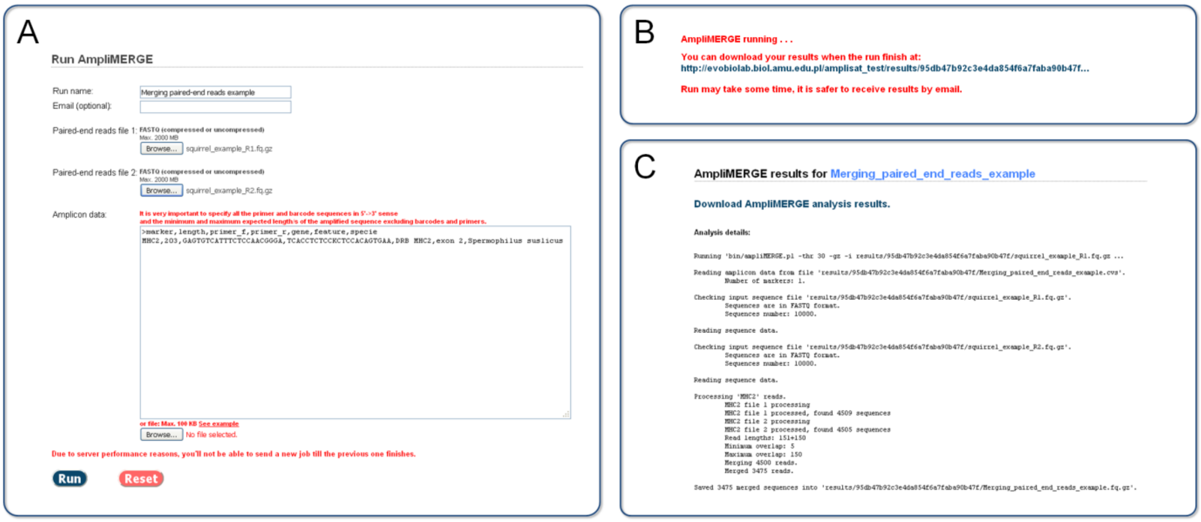
6. A link to the results will appear (Figure 2B), copy it to download the results later or wait for the run to finish and the results will appear directly on the screen (Figure 2C). If you wrote your email address, you will receive a notification when the process will finish with the link to download the results.

The output of the merging process will be something like this (Figure 2C):

```
Processing 'MHC2' reads.
  MHC2 file 1 processing
  MHC2 file 1 processed, found 4509 sequences
  MHC2 file 2 processing
  MHC2 file 2 processed, found 4505 sequences
  Read lengths: 151+150
  Minimum overlap: 5
  Maximum overlap: 150
  Merging 4500 reads.
  Merged 3475 reads.

Saved 3475 merged sequences into...
```

That indicates that AmpliMERGE found 4509 reads matching a primer sequence in the first file and 4505 in the second. From them, only 3475 could be merged. Each file contained 10000 reads, which means that more than half of the sequences are from other experiments. Furthermore, not all the reads matching a primer in one of the reads contain the other primer in the paired one (usually because second read is only partially sequenced or contains sequencing errors).

Following the 'Download AmpliMERGE analysis results' link, we will obtain a GZIP compressed FASTQ file with the 3475 merged reads.



Run AmpliMERGE

Run name: Merging paired-end reads example

Email (optional):

Paired-end reads file 1: FASTQ (compressed or uncompressed)
Max. 2000 MB
squirrel_example_R1.fq.gz

Paired-end reads file 2: FASTQ (compressed or uncompressed)
Max. 2000 MB
squirrel_example_R2.fq.gz

Amplicon data: It is very important to specify all the primer and barcode sequences in 5'->3' sense
and the minimum and maximum expected length/s of the amplified sequence excluding barcodes and primers.

```
>marker,length,primer_f,primer_r,gene,feature,specie
MHC2,203,GAGTGTCATTTCTCCAACGGGA,TCACCTCTCCKCTCCACAGTGAA,DRB MHC2,exon 2,Spermophilus suslicus
```

or file: Max. 100 KB See example
No file selected.

Due to server performance reasons, you'll not be able to send a new job till the previous one finishes.

AmpliMERGE running . . .

You can download your results when the run finish at:
http://evobiolab.biol.amu.edu.pl/amplisat_test/results/95db47b92c3e4da854f6a7faba90b47f...

Run may take some time, it is safer to receive results by email.

AmpliMERGE results for Merging_paired_end_reads_example

Download AmpliMERGE analysis results.

Analysis details:

```
Running 'bin/ampliMERGE.pl -thr 30 -gz -i results/95db47b92c3e4da854f6a7faba90b47f/squirrel_example_R1.fq.gz ...

Reading amplicon data from file 'results/95db47b92c3e4da854f6a7faba90b47f/Merging_paired_end_reads_example.cvs'.
        Number of markers: 1.

Checking input sequence file 'results/95db47b92c3e4da854f6a7faba90b47f/squirrel_example_R1.fq.gz'.
        Sequences are in FASTQ format.
        Sequences number: 10000.

Reading sequence data.

Checking input sequence file 'results/95db47b92c3e4da854f6a7faba90b47f/squirrel_example_R2.fq.gz'.
        Sequences are in FASTQ format.
        Sequences number: 10000.

Reading sequence data.

Processing 'MHC2' reads.
        MHC2 file 1 processing
        MHC2 file 1 processed, found 4509 sequences
        MHC2 file 2 processing
        MHC2 file 2 processed, found 4505 sequences
        Read lengths: 151+150
        Minimum overlap: 5
        Maximum overlap: 150
        Merging 4500 reads.
        Merged 3475 reads.

Saved 3475 merged sequences into 'results/95db47b92c3e4da854f6a7faba90b47f/Merging_paired_end_reads_example.fq.gz'.
```

Figure 2. A: AmpliMERGE input form. B: Message with the link to download results. C: Output after merging process completion.

6. Cleaning reads with AmpliCLEAN

In the present example, the merging process has already cleaned the original set of sequences by removing reads from other experiments and reads not containing the PCR primers. But if we work with single-end reads, AmpliCLEAN will be the only cleaning option. In our example, we will additionally clean the previously merged reads removing all of them that do not contain any barcode sequences from the five individuals to genotype. Also we will remove reads with lower average Phred quality score than 30.

To clean the reads contained in a FASTQ file let's do the following steps:

1. Go to AmpliCLEAN web form (Figure 3A):
   http://evobiolab.biol.amu.edu.pl/amplisat/index.php?ampliclean

2. Give a name to the run and write your email (optional) if you want to receive a link to the results by email.

3. Select the 'Sequences file' with the merged reads (download) clicking on the 'Browse…' button.

4. Introduce the following information into the 'Amplicon data' textbox:

   ```
   >marker,length,primer_f,primer_r,gene,feature,species
   MHC2,203,GAGTGTCATTTCTCCAACGGGA,TCACCTCTCCKCTCCACAGTGAA,DRB MHC2,exon 2,Spermophilus
   suslicus
   >sample,barcode_f,barcode_r
   S1,AACGCG,AAGACA
   S2,TCACTC,CGTCAC
   S3,CTTGGT,TTGAGT
   S4,TGGAAC,TAACAT
   S5,CGAATC,GGTCGA
   ```

   Remember to write the field names on the first line after the character '>' and on the next rows write the values in the same order than the field names. Field names and values must be separated by commas.

5. Give the value 30 to the 'Minimum Phred quality score' filtering parameter.

6. Click on the 'Run' button at the bottom of the page to start the cleaning process.

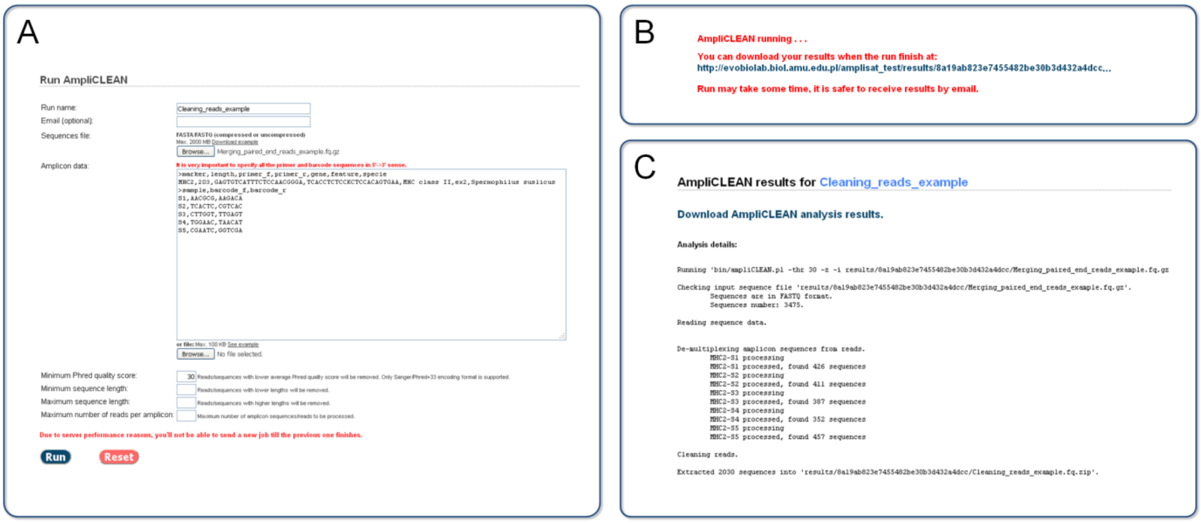
7. A link to the results will appear (Figure 3B), copy it to download the results later or wait for the run to finish and the results will appear directly on the screen (Figure 3C). If you wrote your email address, you will receive a notification when the process will finish with the link to download the results.

The output of the cleaning process will be something like this (Figure 3C):

```
De-multiplexing amplicon sequences from reads.
  MHC2-S1 processing
  MHC2-S1 processed, found 426 sequences
  MHC2-S2 processing
  MHC2-S2 processed, found 411 sequences
  MHC2-S3 processing
  MHC2-S3 processed, found 387 sequences
  MHC2-S4 processing
  MHC2-S4 processed, found 352 sequences
  MHC2-S5 processing
  MHC2-S5 processed, found 457 sequences

Cleaning reads.

Extracted 2030 sequences into...
```

From the initial number of 3475 reads, AmpliCLEAN retrieved 2030 reads that correspond to our five individuals (samples) and fulfil our quality criteria.

Following the ‘Download AmpliCLEAN analysis results’ link, we will obtain a GZIP compressed FASTQ file with the 2030 reads.



Run AmpliCLEAN

Run name: Cleaning_reads_example
Email (optional):
Sequences file: FASTA FASTQ (compressed or uncompressed)
Max. 2000 MB Download example
Merging_paired_end_reads_example.fq.gz
Amplicon data: It is very important to specify all the primer and barcode sequences in 5'->3' sense.

```
>marker,length,primer_f,primer_r,gene,feature,specie
MHC2,203,GAGTGTCATTTCTCCAACGGGA,TCACCTCTCCKCTCCACAGTGAA,MHC class II,ex2,Spermophilus suslicus
>sample,barcode_f,barcode_r
S1,AACGCG,AAGACA
S2,TCACTC,CGTCAC
S3,CTTGGT,TTGAGT
S4,TGGAAC,TAACAT
S5,CGAATC,GGTCGA
```

or file: Max. 100 KB See example
No file selected.

Minimum Phred quality score: 30 Reads/sequences with lower average Phred quality score will be removed. Only Sanger/Phred+33 encoding format is supported.
Minimum sequence length: Reads/sequences with lower lengths will be removed.
Maximum sequence length: Reads/sequences with higher lengths will be removed.
Maximum number of reads per amplicon: Maximum number of amplicon sequences/reads to be processed.

Due to server performance reasons, you'll not be able to send a new job till the previous one finishes.

AmpliCLEAN running . . .

You can download your results when the run finish at:
http://evobiolab.biol.amu.edu.pl/amplisat_test/results/8a19ab823e7455482be30b3d432a4dcc...

Run may take some time, it is safer to receive results by email.

AmpliCLEAN results for Cleaning_reads_example

Download AmpliCLEAN analysis results.

Analysis details:

```
Running 'bin/ampliCLEAN.pl -thr 30 -z -i results/8a19ab823e7455482be30b3d432a4dcc/Merging_paired_end_reads_example.fq.gz

Checking input sequence file 'results/8a19ab823e7455482be30b3d432a4dcc/Merging_paired_end_reads_example.fq.gz'.
        Sequences are in FASTQ format.
        Sequences number: 3475.

Reading sequence data.


De-multiplexing amplicon sequences from reads.
        MHC2-S1 processing
        MHC2-S1 processed, found 426 sequences
        MHC2-S2 processing
        MHC2-S2 processed, found 411 sequences
        MHC2-S3 processing
        MHC2-S3 processed, found 387 sequences
        MHC2-S4 processing
        MHC2-S4 processed, found 352 sequences
        MHC2-S5 processing
        MHC2-S5 processed, found 457 sequences

Cleaning reads.

Extracted 2030 sequences into 'results/8a19ab823e7455482be30b3d432a4dcc/Cleaning_reads_example.fq.zip'.
```

Figure 3. A: AmpliCLEAN input form. B: Message with the link to download results. C: Output after cleaning process completion.

7. Mixing reads with AmpliMIX

We have run a second experiment with the same squirrel individuals plus three new ones and we want to analyze the data from both experiments together (data is already merged). For this purpose, the tool AmpliMIX will generate a unique FASTQ file including the reads from the new individuals and increasing the coverage of the previous 5 ones. The barcodes of the 5 duplicated samples will be unified with the previous ones (only 1 barcode combination per individual) and if any new sample has barcodes identical to the first experiment ones, they will be replaced for new and unique ones. If we do not desire to mix the duplicated individuals with the previous ones it will be enough to give them a different name and AmpliMIX will not combine their reads (eg. one sample can be 'S1' and the other 'S1dup')

To mix the reads contained in two FASTQ files let's complete the following steps:

1. Go to AmpliMIX web form (Figure 4A):
   http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplimix

2. Give a name to the run and write your email address (optional) if you want to receive a link to the results by email.

3. Fill the 'Experiment A data' section with the merged reads file (download) and the amplicon data of the first experiment:

   ```
   >marker,length,primer_f,primer_r,gene,feature,species
   MHC2,203,GAGTGTCATTTCTCCAACGGGA,TCACCTCTCCKCTCCACAGTGAA,MHC class II,ex2,Spermophilus
   suslicus
   >sample,barcode_f,barcode_r
   S1,AACGCG,AAGACA
   S2,TCACTC,CGTCAC
   S3,CTTGGT,TTGAGT
   S4,TGGAAC,TAACAT
   S5,CGAATC,GGTCGA
   ```

   Remember to write the field names on the first line after the character '>' and on the next rows write the values in the same order than the field names. Field names and values must be separated by commas.

4. Fill the 'Experiment B data' section with the merged reads file (download) and the amplicon data of the second experiment:

   ```
   >marker,length,primer_f,primer_r,gene,feature,species
   MHC2,203,GAGTGTCATTTCTCCAACGGGA,TCACCTCTCCKCTCCACAGTGAA,DRB MHC2,exon 2,Spermophilus
   suslicus
   >sample,barcode_f,barcode_r
   S1,AGTGTT,CCGTCC
   S2,CCGGAA,CAATCG
   S3,AGTGTT,CAATCG
   S4,AACCGA,GAACTA
   S5,AACCGA,CACAGT
   S6,CCGCTG,CAATCG
   S7,AACCGA,CCGTCC
   S8,CCGGAA,CCGTCC
   ```

5. Click on the 'Run' button at the bottom of the page to start the mixing process.

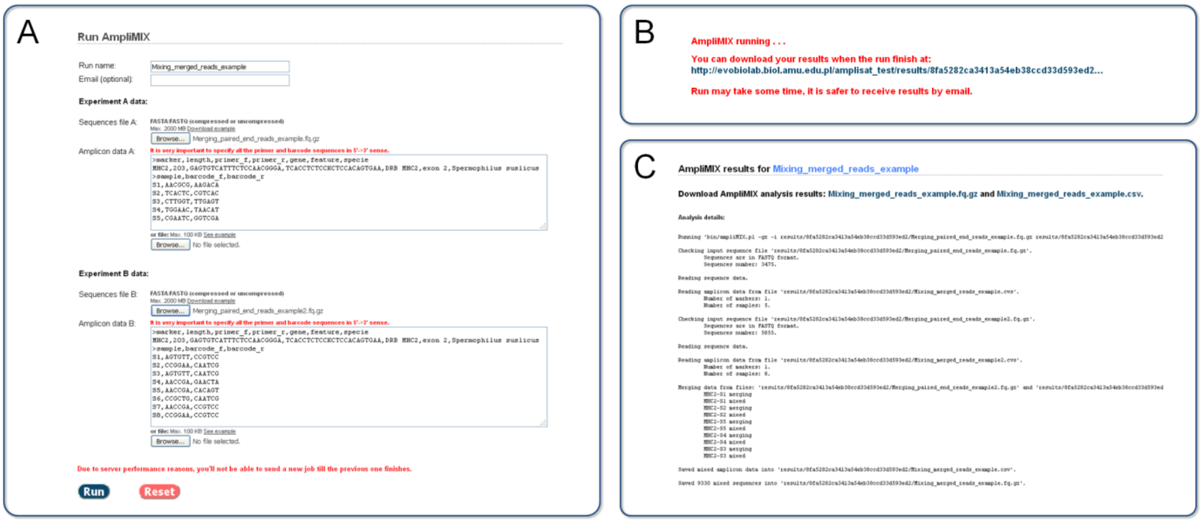
6. A link to the results will appear (Figure 4B), copy it to download the results later or wait for the run to finish and the results will appear directly on the screen (Figure 4C). If you wrote your email address, you will receive a notification when the process will finish with the link to download the results.

The output of the mixing process will be something like this (Figure 4C):

```
Merging data from files...
```

```
   MHC2-S1 merging
   MHC2-S1 mixed
   MHC2-S2 merging
   MHC2-S2 mixed
   MHC2-S5 merging
   MHC2-S5 mixed
   MHC2-S4 merging
   MHC2-S4 mixed
   MHC2-S3 merging
   MHC2-S3 mixed

 Saved mixed amplicon data into...

 Saved 9330 mixed sequences into...
```

AmpliMIX will retrieve two files: a compressed FASTQ file with the reads from the 2 experiments mixed together and a text file in comma-separated values format with the information and sequences of the primers and barcodes from both experiments.



Run AmpliMIX

Run name: Mixing_merged_reads_example
Email (optional):

Experiment A data:

Sequences file A: FASTA/FASTQ (compressed or uncompressed)
Max. 2000 MB Download example
Merging_paired_end_reads_example.fq.gz

Amplicon data A: It is very important to specify all the primer and barcode sequences in 5'->3' sense.

```
>marker,length,primer_f,primer_r,gene,feature,specie
MHC2,203,GAGTGTCATTTCTCCAACGGGA,TCACCTCTCCKCTCCACAGTGAA,DRB MHC2,exon 2,Spermophilus suslicus
>sample,barcode_f,barcode_r
S1,AACGCG,AAGACA
S2,TCACTC,CGTCAC
S3,CTTGGT,TTGAGT
S4,TGGAAC,TAACAT
S5,CGAATC,GGTCGA
```

or file: Max. 100 KB See example
No file selected.

Experiment B data:

Sequences file B: FASTA/FASTQ (compressed or uncompressed)
Max. 2000 MB Download example
Merging_paired_end_reads_example2.fq.gz

Amplicon data B: It is very important to specify all the primer and barcode sequences in 5'->3' sense.

```
>marker,length,primer_f,primer_r,gene,feature,specie
MHC2,203,GAGTGTCATTTCTCCAACGGGA,TCACCTCTCCKCTCCACAGTGAA,DRB MHC2,exon 2,Spermophilus suslicus
>sample,barcode_f,barcode_r
S1,AGTGTT,CCGTCC
S2,CCGGAA,CAATCG
S3,AGTGTT,CAATCG
S4,AACCGA,GAACTA
S5,AACCGA,CACAGT
S6,CCGCTG,CAATCG
S7,AACCGA,CCGTCC
S8,CCGGAA,CCGTCC
```

or file: Max. 100 KB See example
No file selected.

Due to server performance reasons, you'll not be able to send a new job till the previous one finishes.

AmpliMIX running . . .

You can download your results when the run finish at:
http://evobiolab.biol.amu.edu.pl/amplisat_test/results/8fa5282ca3413a54eb38ccd33d593ed2...

Run may take some time, it is safer to receive results by email.

AmpliMIX results for Mixing_merged_reads_example

Download AmpliMIX analysis results: Mixing_merged_reads_example.fq.gz **and** Mixing_merged_reads_example.csv.

Analysis details:

```
Running 'bin/ampliMIX.pl -gz -i results/8fa5282ca3413a54eb38ccd33d593ed2/Merging_paired_end_reads_example.fq.gz results/8fa5282ca3413a54eb38ccd33d593ed2

Checking input sequence file 'results/8fa5282ca3413a54eb38ccd33d593ed2/Merging_paired_end_reads_example.fq.gz'.
        Sequences are in FASTQ format.
        Sequences number: 3475.

Reading sequence data.

Reading amplicon data from file 'results/8fa5282ca3413a54eb38ccd33d593ed2/Mixing_merged_reads_example.cvs'.
        Number of markers: 1.
        Number of samples: 5.

Checking input sequence file 'results/8fa5282ca3413a54eb38ccd33d593ed2/Merging_paired_end_reads_example2.fq.gz'.
        Sequences are in FASTQ format.
        Sequences number: 5855.

Reading sequence data.

Reading amplicon data from file 'results/8fa5282ca3413a54eb38ccd33d593ed2/Mixing_merged_reads_example2.cvs'.
        Number of markers: 1.
        Number of samples: 8.

Merging data from files: 'results/8fa5282ca3413a54eb38ccd33d593ed2/Merging_paired_end_reads_example2.fq.gz' and 'results/8fa5282ca3413a54eb38ccd33d593ed
        MHC2-S1 merging
        MHC2-S1 mixed
        MHC2-S2 merging
        MHC2-S2 mixed
        MHC2-S5 merging
        MHC2-S5 mixed
        MHC2-S4 merging
        MHC2-S4 mixed
        MHC2-S3 merging
        MHC2-S3 mixed

Saved mixed amplicon data into 'results/8fa5282ca3413a54eb38ccd33d593ed2/Mixing_merged_reads_example.csv'.

Saved 9330 mixed sequences into 'results/8fa5282ca3413a54eb38ccd33d593ed2/Mixing_merged_reads_example.fq.gz'.
```

Figure 4. A: AmpliMIX input form. B: Message with the link to download results. C: Output after mixing process completion.

8. Checking variants with AmpliCHECK

At this stage we should have a unique FASTQ file with our experiment reads after merging, cleaning and/or mixing (see previous sections). Before doing any more advanced analysis, we should use AmpliCHECK for a fast de-multiplexing of our reads and a variant/artifact frequency assessment. That means that the reads will be separated into amplicons, in our case, each amplicon will include all the MHC II DRB sequences of a particular individual. All the identical reads will be added to the coverage of a unique variant and variant frequency will be calculated. Additionally, all variants with potential sequencing and PCR errors will be annotated. Definitions of amplicon, variant and other useful terms are in Table 1.

To check the reads contained in a FASTQ file let's do the following steps:

1. Go to AmpliCHECK web form (Figure 6A):
   http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplicheck

2. Give a name to the run and write your email (optional) if you want to receive a link to the results by email.

3. Select the 'Sequences file' with the merged reads (download) using the 'Browse…' button.

4. Select 'Illumina' in 'Technology' check box.

5. Introduce the following information into the 'Amplicon data' textbox:

   ```
   >marker,length,primer_f,primer_r,gene,feature,species
   MHC2,203,GAGTGTCATTTCTCCAACGGGA,TCACCTCTCCKCTCCACAGTGAA,DRB MHC2,exon 2,Spermophilus
   suslicus
   >sample,barcode_f,barcode_r
   S1,AACGCG,AAGACA
   S2,TCACTC,CGTCAC
   S3,CTTGGT,TTGAGT
   S4,TGGAAC,TAACAT
   S5,CGAATC,GGTCGA
   ```

   Remember to write the field names on the first line after the character '>' and on the next rows write the values in the same order than the field names. Field names and values must be separated by commas.

6. We can include a FASTA file with the Genbank sequences of the previously known speckled squirrel alleles to rename the variants with the allele names included in the FASTA file (download).

7. Click on the 'Run' button at the bottom of the page to start the checking process.

8. A link to the results will appear (Figure 6B), copy it to download the results later or wait for the run to finish and the results will appear directly on the screen (Figure 6C). If you wrote your email address, you will receive a notification when the process will finish with the link to download the results.

The output of the checking process will be something like this (Figure 6C):

```
Running 'bin/ampliCHECK.pl ...

Checking input sequence file ...
  Sequences are in FASTQ format.
  Sequences number: 3475.

Reading sequence data.

Reading amplicon data from file ...
  Number of markers: 1.
  Number of samples: 5.

Reading allele sequences from ...
```

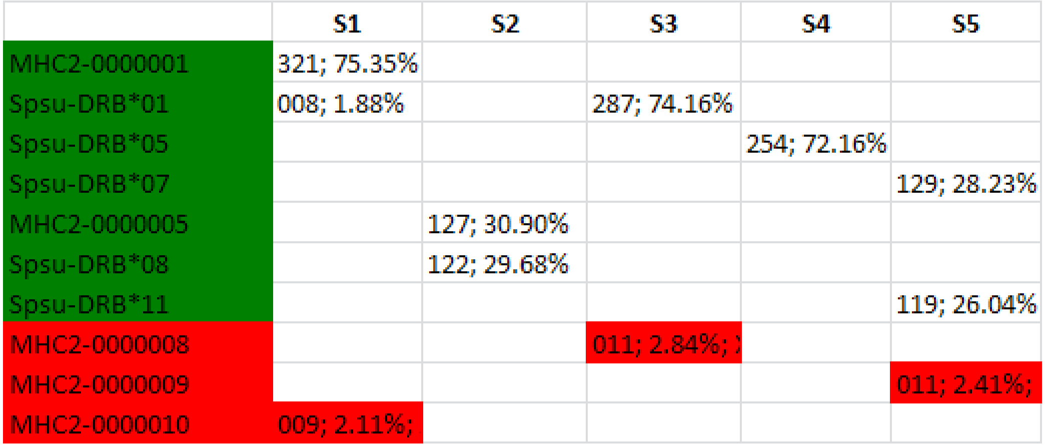
```
Printing amplicon data into ...

De-multiplexing amplicon sequences from reads.
  MHC2-S1 de-multiplexing
  MHC2-S1 de-multiplexed (426 sequences, 67 unique)
  ...

Matching allele sequences.

Extracting de-multiplexed sequences into ...

Filtering sequences with the following criteria ('filter' 'marker' 'values'):
  min_amplicon_seq_frequency    all    1

  MHC2-S1 filtering
  MHC2-S1 filtered (338 sequences, 3 unique)
  ...

Comparing amplicon sequences that passed the filters with the following parameters
('threshold' 'marker' 'values'):
  substitution_threshold all     1
  indel_threshold        all     0.001

  MHC2-S1 pairwise comparing sequences
  MHC2-S1 pairwise compared 3 sequences
  ...

Sequences per amplicon:
Amplicon Total  Unique
MHC2-S1  426    67
MHC2-S2  411    124
MHC2-S3  387    78
MHC2-S4  352    81
MHC2-S5  457    140


Analysis results stored into...
```

Following the ‘Download AmpliCHECK analysis results’ link, we will obtain a ZIP compressed file including:

- ‘results.xlsx’ Excel file: annotations of variant depths, frequencies and possible errors.
- ‘allseqs’ folder: contains FASTA files with de-multiplexed variants for each individual amplicon.
- ‘amplicon_data.csv’: a comma-separated values format file including the amplicon data and analysis parameters.

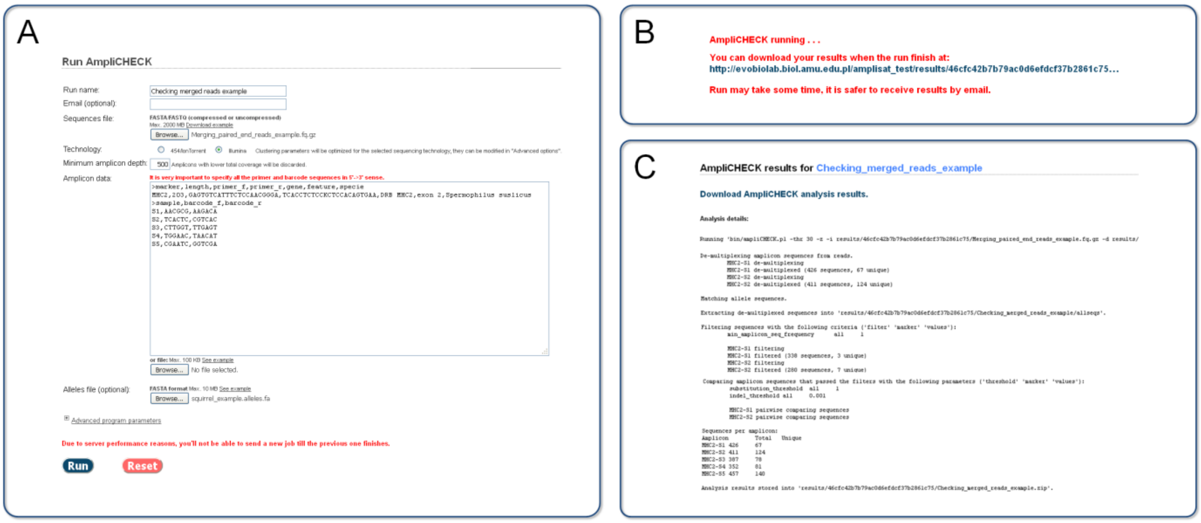
The most informative file is ‘results.xlsx’, in the Figure 5 we can see how it looks like for the most abundant variants. Individuals/samples are shown in columns and variants in rows. The numeric values show is the variant depths and frequencies into the amplicons. In green color are shown the names of the most frequent variants without known errors. The variants that can be explained as sequencing errors or chimeras of the most abundant ones are colored in red. For example, the variant ‘MHC2-0000010’ differs in one substitution (186 G/T) from ‘MHC2-0000001’, so it is probably a sequencing artefact. Variants whose sequences are identical to the Genbank ones included in the alleles FASTA file are renamed with the name of the allele, eg. ‘Spsu-DRB*01’.

In the results from the Figure 5 we observe that 2 variants (MHC2-0000001 and MHC2-0000005) are not present in Genbank data and they are probably novel alleles. In the sample ‘S1’ we notice a possible contamination from another individual (probably S3) because the allele ‘Spsu-DRB*01’ appears in very low frequency. Analyzing the frequencies from the putative real alleles (green) and artefacts (red) we can conclude that using a frequency threshold of 25% in further analysis we will filter most of the artefacts that obscure the genotyping (in fact, we can use a threshold between 3 and 25%).



	S1	S2	S3	S4	S5
MHC2-0000001	321; 75.35%				
Spsu-DRB*01	008; 1.88%		287; 74.16%		
Spsu-DRB*05				254; 72.16%	
Spsu-DRB*07					129; 28.23%
MHC2-0000005		127; 30.90%			
Spsu-DRB*08		122; 29.68%			
Spsu-DRB*11					119; 26.04%
MHC2-0000008			011; 2.84%;		
MHC2-0000009					011; 2.41%;
MHC2-0000010	009; 2.11%;				

Figure 5. AmpliCHECK 'results.xlsx' Excel file example. The numeric values show is the variant depths and frequencies into the amplicons. In green color are shown the names of the most frequent variants without known errors. The variants that can be explained as sequencing errors or chimeras of the most abundant ones are colored in red.



Run AmpliCHECK

Run name: Checking merged reads example

Email (optional):

Sequences file: FASTA/FASTQ (compressed or uncompressed)
Max. 2000 MB Download example
Browse... Merging_paired_end_reads_example.fq.gz

Technology: ○ 454/IonTorrent ● Illumina Clustering parameters will be optimized for the selected sequencing technology, they can be modified in "Advanced options".

Minimum amplicon depth: 500 Amplicons with lower total coverage will be discarded.

Amplicon data: It is very important to specify all the primer and barcode sequences in 5'->3' sense.

```
>marker,length,primer_f,primer_r,gene,feature,specie
MHC2,203,GAGTGTCATTTCTCCAACGGGA,TCACCTCTCCKCTCCACAGTGAA,DRB MHC2,exon 2,Spermophilus suslicus
>sample,barcode_f,barcode_r
S1,AACGCG,AAGACA
S2,TCACTC,CGTCAC
S3,CTTGGT,TTGAGT
S4,TGGAAC,TAACAT
S5,CGAATC,GGTCGA
```

or file: Max. 100 KB See example
Browse... No file selected.

Alleles file (optional): FASTA format Max. 10 MB See example
Browse... squirrel_example.alleles.fa

⊞ Advanced program parameters

Due to server performance reasons, you'll not be able to send a new job till the previous one finishes.

Run Reset

AmpliCHECK running . . .

You can download your results when the run finish at:
http://evobiolab.biol.amu.edu.pl/amplisat_test/results/46cfc42b7b79ac0d6efdcf37b2861c75...

Run may take some time, it is safer to receive results by email.

AmpliCHECK results for Checking_merged_reads_example

Download AmpliCHECK analysis results.

Analysis details:

```
Running 'bin/ampliCHECK.pl -thr 30 -z -i results/46cfc42b7b79ac0d6efdcf37b2861c75/Merging_paired_end_reads_example.fq.gz -d results/

De-multiplexing amplicon sequences from reads.
        MHC2-S1 de-multiplexing
        MHC2-S1 de-multiplexed (426 sequences, 67 unique)
        MHC2-S2 de-multiplexing
        MHC2-S2 de-multiplexed (411 sequences, 124 unique)

Matching allele sequences.

Extracting de-multiplexed sequences into 'results/46cfc42b7b79ac0d6efdcf37b2861c75/Checking_merged_reads_example/allseqs'.

Filtering sequences with the following criteria ('filter' 'marker' 'values'):
        min_amplicon_seq_frequency      all     1

        MHC2-S1 filtering
        MHC2-S1 filtered (338 sequences, 3 unique)
        MHC2-S2 filtering
        MHC2-S2 filtered (280 sequences, 7 unique)
 Comparing amplicon sequences that passed the filters with the following parameters ('threshold' 'marker' 'values'):
         substitution_threshold  all     1
         indel_threshold all     0.001

         MHC2-S1 pairwise comparing sequences
         MHC2-S2 pairwise comparing sequences

Sequences per amplicon:
Amplicon        Total   Unique
MHC2-S1 426     67
MHC2-S2 411     124
MHC2-S3 387     78
MHC2-S4 352     81
MHC2-S5 457     140

Analysis results stored into 'results/46cfc42b7b79ac0d6efdcf37b2861c75/Checking_merged_reads_example.zip'.
```

Figure 6. A: AmpliCHECK input form. B: Message with the link to download results. C: Output after cleaning process completion.

9. Genotyping variants with AmpliSAS

At this stage we should have a unique FASTQ file with our experiment reads after merging, cleaning and/or mixing (see previous sections). Also, before running AmpliSAS it is highly recommended to use AmpliCHECK to have a better knowledge of the quality of the data, allele frequencies and errors. AmpliSAS can take several hours to run while AmpliCHECK should take few minutes.

The main difference between AmpliSAS and AmpliCHECK is that AmpliSAS performs a clustering step (slow) that increase the coverage of the alleles by incorporating to them sequencing errors. Also AmpliSAS allows a large set of filtering options including like chimera removal that can remove artefacts or small contaminations that can escape the clustering. For the present example we could perform the genotyping by manually correcting AmpliCHECK results, but more complex data or with lower quality could require AmpliSAS (eg. 454 or Ion Torrent data that contain a high number of indels, transcriptomic data where some alleles can be present at low frequencies, etc.).

The AmpliSAS workflow is divided into three main steps: 1) sequence de-multiplexing; 2) sequence clustering; and 3) artefact filtering. In summary, the reads are de-multiplexed into amplicons, in our case, each amplicon will include all the MHC II DRB sequences of a particular individual. All the identical reads will be added to the coverage of a unique variant and variant frequency will be calculated. During clustering, variants will be aligned to each other to find sequencing errors, these erroneous variants will be removed and their coverages will be added to the true ones. Definitions of amplicon, variant and other useful terms are in Table 1. For more details about AmpliSAS algorithm check Sebastian et al. publication (Sebastian *et al.* 2016).

We will perform AmpliSAS analysis with almost all default parameters, advanced options will be explained in further sections. Let's run AmpliSAS following thsese steps:

1. Go to AmpliSAS web form (**Figure 8**A):
   http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplisas

2. Give a name to the run and write your email (optional) if you want to receive a link to the results by email.

3. Select the 'Sequences file' with the merged reads (download) using the 'Browse…' button.

4. Select 'Illumina' in 'Technology' check box.

5. Keep in 10 the 'Maximum number of alleles per amplicon'.

6. Decrease 'Minimum amplicon depth' to 100, this is an easy genotyping example and amplicon depths are small.

7. Introduce the following information into the 'Amplicon data' textbox:

```
>marker,length,primer_f,primer_r,gene,feature,species
MHC2,203,GAGTGTCATTTCTCCAACGGGA,TCACCTCTCCKCTCCACAGTGAA,DRB MHC2,exon 2,Spermophilus
suslicus
>sample,barcode_f,barcode_r
S1,AACGCG,AAGACA
S2,TCACTC,CGTCAC
S3,CTTGGT,TTGAGT
S4,TGGAAC,TAACAT
S5,CGAATC,GGTCGA
```

   Remember to write the field names on the first line after the character '>' and on the next rows write the values in the same order than the field names. Field names and values must be separated by commas.

8. We can include a FASTA file with the Genbank sequences of the previously known speckled squirrel alleles to rename the variants with the allele names included in the FASTA file (download).

9. Click on the 'Run' button at the bottom of the page to start the genotyping process.

10. A link to the results will appear (**Figure 8**B), copy it to download the results later or wait for the run to finish and the results will appear directly on the screen (**Figure 8**C). If you wrote your email address, you will receive a notification when the process will finish with the link to download the results.

The output of AmpliSAS will be something like this (**Figure 8**C):

```
Running 'bin/ampliSAS.pl ...

Checking input sequence file ...
  Sequences are in FASTQ format.
  Sequences number: 3475.

Reading sequence data.

Reading amplicon data from file ...
  Number of markers: 1.
  Number of samples: 5.

Reading allele sequences from ...

Printing amplicon data into ...

De-multiplexing amplicon sequences from reads.
  MHC2-S1 de-multiplexing
  MHC2-S1 de-multiplexed (426 sequences, 67 unique)
  ...

Matching allele sequences.

Extracting de-multiplexed sequences into ...

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

  MHC2-S1 clustering
  MHC2-S1 clustered (420 sequences, 4 unique)
  ...

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

Matching allele sequences.

Extracting clustered sequences into...

Filtering sequences with the following criteria ('filter' 'marker' 'values'):
  min_amplicon_depth     all     100
  min_amplicon_seq_frequency     all     3
  min_chimera_length     all     10
  max_allele_number      all     10

  MHC2-S1 filtering
  MHC2-S1 filtered (402 sequences, 1 unique)
  ...

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

Extracting filtered sequences into...

Sequences per amplicon:
Amplicon Total   Unique  Total-clustered       Unique-clustered      Total-filtered
  Unique-filtered
MHC2-S1   426     67      420     4       402     1
MHC2-S2   411     124     398     6       351     2
MHC2-S3   387     78      384     1       384     1
MHC2-S4   352     81      352     1       352     1
```

```
MHC2-S5   457     140     440     6       385     2

Analysis results stored into...
```

Following the ‘Download AmpliSAS analysis results’ link, we will obtain a ZIP compressed file including:

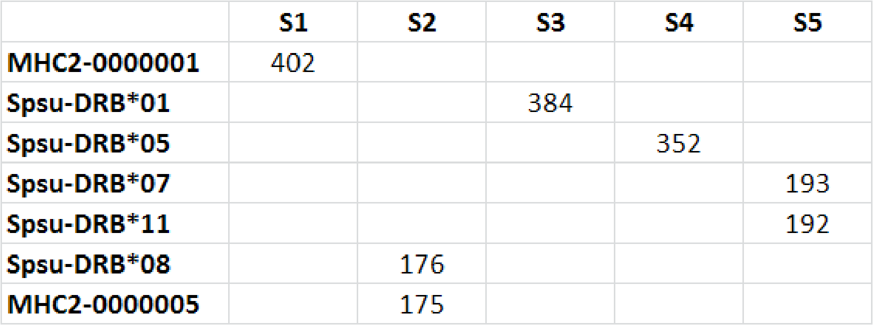
- ‘results.xlsx’ Excel file: final genotyping results, after de-multiplexing, clustering and filtering steps.
- ‘allseqs’ folder: contains FASTA files with de-multiplexed variants for each individual amplicon and also an Excel file with de-multiplexed variant depths.
- ‘clustered’ folder: contains FASTA files with clustered variants for each individual amplicon and also an Excel file with clustered variant depths.
- ‘filtered’ folder: contains FASTA files with filtered variants for each individual amplicon and also an Excel file with filtered variant depths.
- ‘amplicon_data.csv’: a comma-separated values format file including the amplicon data and analysis parameters.

The most informative file is ‘results.xlsx’, in the

Figure 7 we can see how it looks like. Individuals/samples are shown in columns and variants in rows, the numeric values show is the variant depths into the amplicons. Variants whose sequences are identical to the Genbank ones included in the alleles FASTA file are renamed with the name of the allele, eg. ‘Spsu-DRB*01’.

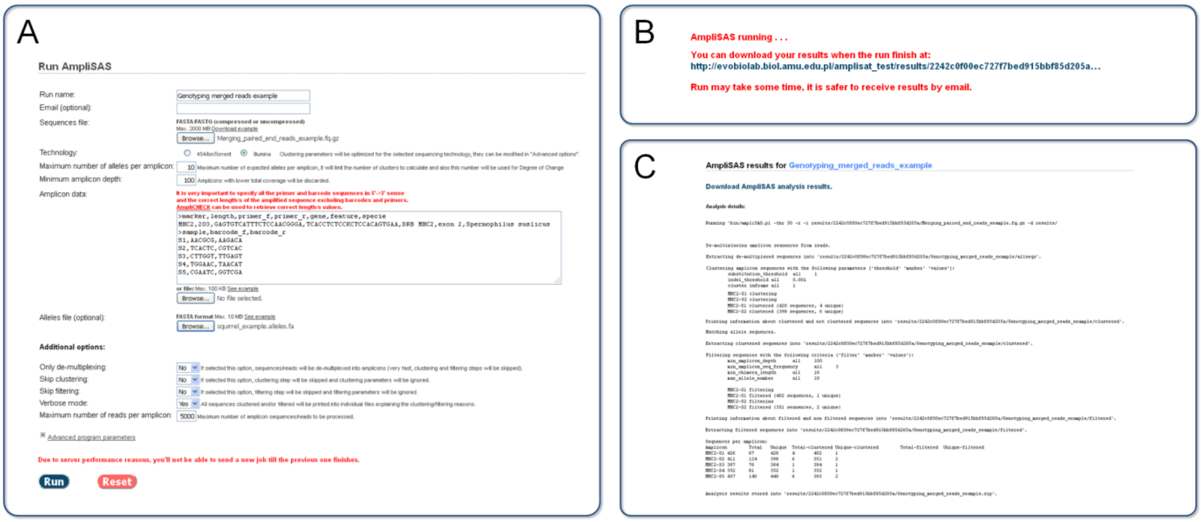
In the results from the

Figure 7 we observe that 2 variants (MHC2-0000001 and MHC2-0000005) are not present in Genbank data and they are probably novel alleles. The low possible variant contamination in the sample ‘S1’ that we noticed in AmpliCHECK analysis (see Section 8) has been filtered. Also we observe how the variant coverages have increased after the clustering step by removing the sequencing errors and incorporating their depths. For example, the unique allele from the first individual has increased a 25% its depth from 321 to 402 reads. The artefactual variants that AmpliCHECK annotated as putative errors have been removed and most of them incorporated in their original alleles. Later (Section 11) we will automatically compare AmpliCHECK against AmpliSAS results.



	S1	S2	S3	S4	S5
MHC2-0000001	402				
Spsu-DRB*01			384		
Spsu-DRB*05				352	
Spsu-DRB*07					193
Spsu-DRB*11					192
Spsu-DRB*08		176			
MHC2-0000005		175			

Figure 7. AmpliSAS ‘results.xlsx’ Excel file example. The numeric values show the variant depths after clustering and filtering.



Run AmpliSAS

Run name: Genotyping merged reads example

Email (optional):

Sequences file: FASTA FASTQ (compressed or uncompressed) Max. 2000 MB Download example
Merging_paired_end_reads_example.fq.gz

Technology: ○ 454/IonTorrent ● Illumina Clustering parameters will be optimized for the selected sequencing technology, they can be modified in "Advanced options".

Maximum number of alleles per amplicon: 10 Maximum number of expected alleles per amplicon, it will limit the number of clusters to calculate and also this number will be used for Degree of Change

Minimum amplicon depth: 100 Amplicons with lower total coverage will be discarded.

Amplicon data: It is very important to specify all the primer and barcode sequences in 5'->3' sense
and the correct length/s of the amplified sequence excluding barcodes and primers.
AmpliCHECK can be used to retrieve correct length/s values.

```
>marker,length,primer_f,primer_r,gene,feature,specie
MHC2,203,GAGTGTCATTTCTCCAACGGGA,TCACCTCTCCKCTCCACAGTGAA,DRB MHC2,exon 2,Spermophilus suslicus
>sample,barcode_f,barcode_r
S1,AACGCG,AAGACA
S2,TCACTC,CGTCAC
S3,CTTGGT,TTGAGT
S4,TGGAAC,TAACAT
S5,CGAATC,GGTCGA
```

or file: Max. 100 KB See example
No file selected.

Alleles file (optional): FASTA format Max. 10 MB See example
squirrel_example.alleles.fa

Additional options:

Only de-multiplexing: No If selected this option, sequences/reads will be de-multiplexed into amplicons (very fast, clustering and filtering steps will be skipped).

Skip clustering: No If selected this option, clustering step will be skipped and clustering parameters will be ignored.

Skip filtering: No If selected this option, filtering step will be skipped and filtering parameters will be ignored.

Verbose mode: Yes All sequences clustered and/or filtered will be printed into individual files explaining the clustering/filtering reasons.

Maximum number of reads per amplicon: 5000 Maximum number of amplicon sequences/reads to be processed.

⊞ Advanced program parameters

Due to server performance reasons, you'll not be able to send a new job till the previous one finishes.

AmpliSAS running . . .

You can download your results when the run finish at:
http://evobiolab.biol.amu.edu.pl/amplisat_test/results/2242c0f00ec727f7bed915bbf85d205a...

Run may take some time, it is safer to receive results by email.

AmpliSAS results for Genotyping_merged_reads_example

Download AmpliSAS analysis results.

Analysis details:

```
Running 'bin/ampliSAS.pl -thr 30 -z -i results/2242c0f00ec727f7bed915bbf85d205a/Merging_paired_end_reads_example.fq.gz -d results/


De-multiplexing amplicon sequences from reads.

Extracting de-multiplexed sequences into 'results/2242c0f00ec727f7bed915bbf85d205a/Genotyping_merged_reads_example/allseqs'.

Clustering amplicon sequences with the following parameters ('threshold' 'marker' 'values'):
        substitution_threshold  all     1
        indel_threshold all     0.001
        cluster inframe all     1
        MHC2-S1 clustering
        MHC2-S2 clustering
        MHC2-S1 clustered (420 sequences, 4 unique)
        MHC2-S2 clustered (398 sequences, 6 unique)
Printing information about clustered and not clustered sequences into 'results/2242c0f00ec727f7bed915bbf85d205a/Genotyping_merged_reads_example/clustered'.

Matching allele sequences.

Extracting clustered sequences into 'results/2242c0f00ec727f7bed915bbf85d205a/Genotyping_merged_reads_example/clustered'.

Filtering sequences with the following criteria ('filter' 'marker' 'values'):
        min_amplicon_depth      all     100
        min_amplicon_seq_frequency      all     3
        min_chimera_length      all     10
        max_allele_number       all     10

        MHC2-S1 filtering
        MHC2-S1 filtered (402 sequences, 1 unique)
        MHC2-S2 filtering
        MHC2-S2 filtered (351 sequences, 2 unique)

Printing information about filtered and non filtered sequences into 'results/2242c0f00ec727f7bed915bbf85d205a/Genotyping_merged_reads_example/filtered'.

Extracting filtered sequences into 'results/2242c0f00ec727f7bed915bbf85d205a/Genotyping_merged_reads_example/filtered'.

Sequences per amplicon:
Amplicon        Total   Unique  Total-clustered Unique-clustered        Total-filtered  Unique-filtered
MHC2-S1 426     67      420     4       402     1
MHC2-S2 411     124     398     6       351     2
MHC2-S3 387     78      384     1       384     1
MHC2-S4 352     81      352     1       352     1
MHC2-S5 457     140     440     6       385     2


Analysis results stored into 'results/2242c0f00ec727f7bed915bbf85d205a/Genotyping_merged_reads_example.zip'.
```

Figure 8. A: AmpliSAS input form. B: Message with the link to download results. C: Output after cleaning process completion.

10. Genotyping variants with AmpliLEGACY

AmpliSAT is our best tested and recommended tool (see previous section), but AmpliLEGACY offers the possibility to do an equivalent genotyping using other strategies from the literature. The three alternative analysis methods available in AmpliLEGACY are:

1. An approach based in comparing variants from two replicates to discard erroneous ones and contaminations (Sommer *et al.* 2013).
2. Degree of Change (DOC) method that looks for a drop in cumulative frequency after clustering variants to differentiate between real ones and artefacts (Lighten *et al.* 2014).
3. A comparison among rare variants and common ones to determine if they can be explained as sequencing arterfacts or PCR chimeras of the more common ones (Radwan *et al.* 2012; Herdegen *et al.* 2014).

Subsequently, other published genotyping protocols have been incorporated (Sommer *et al.* 2013; Herdegen *et al.* 2014; Lighten *et al.* 2014). AmpliSAT users can decide which method fits better to their data or try several and later compare the retrieved genotypes by the different approaches with AmpliCOMPARE (Section 11).

We will genotype with the DOC method (Lighten *et al.* 2014) our example data (with default parameters) following these steps:

1. Go to AmpliLEGACY web form (Figure 10A):
   http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplilegacy

2. Give a name to the run and write your email (optional) if you want to receive a link to the results by email.

3. Select the 'Sequences file' with the merged reads (download) using the 'Browse…' button.

4. Choose 'Lighten et al.' as 'Genotyping method'.

5. Decrease 'Minimum amplicon depth' to 100, this is an easy genotyping example and amplicon depths are small.

6. Introduce the following information into the 'Amplicon data' textbox:

   ```
   >marker,length,primer_f,primer_r,gene,feature,species
   MHC2,203,GAGTGTCATTTCTCCAACGGGA,TCACCTCTCCKCTCCACAGTGAA,DRB MHC2,exon 2,Spermophilus
   suslicus
   >sample,barcode_f,barcode_r
   S1,AACGCG,AAGACA
   S2,TCACTC,CGTCAC
   S3,CTTGGT,TTGAGT
   S4,TGGAAC,TAACAT
   S5,CGAATC,GGTCGA
   ```

   Remember to write the field names on the first line after the character '>' and on the next rows write the values in the same order than the field names. Field names and values must be separated by commas.

7. We can include a FASTA file with the Genbank sequences of the previously known speckled squirrel alleles to rename the variants with the allele names included in the FASTA file (download).

8. Click on the 'Run' button at the bottom of the page to start the genotyping process.

9. A link to the results will appear (Figure 10B), copy it to download the results later or wait for the run to finish and the results will appear directly on the screen (Figure 10C). If you wrote your email address, you will receive a notification when the process will finish with the link to download the results.

The output of AmpliLEGACY will be something like this (Figure 10C):

```
Running 'bin/ampliLEGACY.pl ...

Checking input sequence file ...
  Sequences are in FASTQ format.
  Sequences number: 3475.

Reading sequence data.

Reading amplicon data from file ...
  Number of markers: 1.
  Number of samples: 5.

Reading allele sequences from ...

Printing amplicon data into ...

De-multiplexing amplicon sequences from reads.
  MHC2-S1 de-multiplexing
  MHC2-S1 de-multiplexed (426 sequences, 67 unique)
  MHC2-S2 de-multiplexing
  MHC2-S2 de-multiplexed (411 sequences, 124 unique)
  ...

Matching allele sequences.

Extracting de-multiplexed sequences into ...

Genotyping sequences with 'Lighten' method and the following criteria
('parameter' 'marker' 'values'):
  cluster_inframe       all     1
  min_amplicon_depth    all     100
  max_allele_number     all     10
  min_dominant_frequency_threshold     all     2
  error_threshold       all     3

  MHC2-S1 genotyping
  MHC2-S2 genotyping
  MHC2-S1 genotyped (395 sequences, 1 unique)
  MHC2-S2 genotyped (349 sequences, 2 unique)
  ...

Printing verbose information about alleles and artifacts into ...

Extracting putative allele sequences into ...


Sequences per amplicon:
Amplicon Total  Unique  Total-lighten  Alleles-lighten
MHC2-S1  426    67      395     1
MHC2-S2  411    124     349     2
MHC2-S3  387    78      376     1
MHC2-S4  352    81      345     1
MHC2-S5  457    140     353     2


Analysis results stored into ...
```

Following the ‘Download AmpliLEGACY analysis results’ link, we will obtain a ZIP compressed file including:

- ‘results.xlsx’ Excel file: final genotyping results, after de-multiplexing, clustering and filtering steps.
- ‘allseqs’ folder: contains FASTA files with de-multiplexed variants for each individual amplicon and also an Excel file with de-multiplexed variant depths.

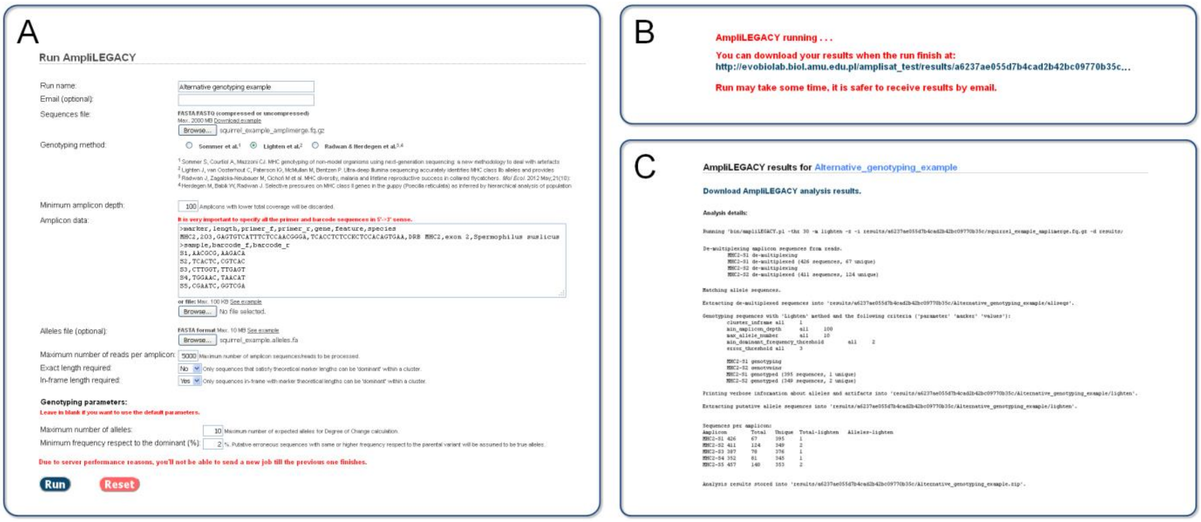
- 'genotyping' folder: contains FASTA files with clustered variants and genotyping details for each individual amplicon and also an Excel file with clustered variant depths.
- 'amplicon_data.csv': a comma-separated values format file including the amplicon data and analysis parameters.

Comparing AmpliLEGACY genotyping results in Figure 9 against AmpliSAS (

Figure 7) we observe that both methods retrieve identical genotypes for this example, only variant depths are slightly different.

	S1	S2	S3	S4	S5
MHC2-0000001	395				
Spsu-DRB*01			376		
Spsu-DRB*05				345	
Spsu-DRB*07					181
Spsu-DRB*08		176			
MHC2-0000005		173			
Spsu-DRB*11					172

Figure 9. AmpliLEGACY 'results.xlsx' Excel file example. The numeric values show the variant depths after clustering and filtering by the DOC method.



Run AmpliLEGACY

Run name: Alternative genotyping example
Email (optional):
Sequences file: FASTA/FASTQ (compressed or uncompressed) Max. 2000 MB Download example
squirrel_example_amplimerge.fq.gz
Genotyping method: ○ Sommer et al.[1] ● Lighten et al.[2] ○ Radwan & Herdegen et al.[3,4]

[1] Sommer S, Courtiol A, Mazzoni CJ. MHC genotyping of non-model organisms using next-generation sequencing: a new methodology to deal with artefacts
[2] Lighten J, van Oosterhout C, Paterson IG, McMullan M, Bentzen P. Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides
[3] Radwan J, Zagalska-Neubauer M, Cichoń M et al. MHC diversity, malaria and lifetime reproductive success in collared flycatchers. *Mol Ecol.* 2012 May;21(10):
[4] Herdegen M, Babik W, Radwan J. Selective pressures on MHC class II genes in the guppy (Poecilia reticulata) as inferred by hierarchical analysis of population

Minimum amplicon depth: 100 Amplicons with lower total coverage will be discarded.
Amplicon data: It is very important to specify all the primer and barcode sequences in 5'->3' sense.

```
>marker,length,primer_f,primer_r,gene,feature,species
MHC2,203,GAGTGTCATTTCTCCAACGGGA,TCACCTCTCCKCTCCACAGTGAA,DRB MHC2,exon 2,Spermophilus suslicus
>sample,barcode_f,barcode_r
S1,AACGCG,AAGACA
S2,TCACTC,CGTCAC
S3,CTTGGT,TTGAGT
S4,TGGAAC,TAACAT
S5,CGAATC,GGTCGA
```

or file: Max. 100 KB See example
No file selected.
Alleles file (optional): FASTA format Max. 10 MB See example
squirrel_example.alleles.fa
Maximum number of reads per amplicon: 5000 Maximum number of amplicon sequences/reads to be processed.
Exact length required: No Only sequences that satisfy theoretical marker lengths can be 'dominant' within a cluster.
In-frame length required: Yes Only sequences in-frame with marker theoretical lengths can be 'dominant' within a cluster.

Genotyping parameters:
Leave in blank if you want to use the default parameters.

Maximum number of alleles: 10 Maximum number of expected alleles for Degree of Change calculation.
Minimum frequency respect to the dominant (%): 2 %. Putative erroneous sequences with same or higher frequency respect to the parental variant will be assumed to be true alleles.

Due to server performance reasons, you'll not be able to send a new job till the previous one finishes.

AmpliLEGACY running . . .

You can download your results when the run finish at:
http://evobiolab.biol.amu.edu.pl/amplisat_test/results/a6237ae055d7b4cad2b42bc09770b35c...

Run may take some time, it is safer to receive results by email.

AmpliLEGACY results for Alternative_genotyping_example

Download AmpliLEGACY analysis results.

Analysis details:

```
Running 'bin/ampliLEGACY.pl -thr 30 -m lighten -z -i results/a6237ae055d7b4cad2b42bc09770b35c/squirrel_example_amplimerge.fq.gz -d results/

De-multiplexing amplicon sequences from reads.
        MHC2-S1 de-multiplexing
        MHC2-S1 de-multiplexed (426 sequences, 67 unique)
        MHC2-S2 de-multiplexing
        MHC2-S2 de-multiplexed (411 sequences, 124 unique)

Matching allele sequences.

Extracting de-multiplexed sequences into 'results/a6237ae055d7b4cad2b42bc09770b35c/Alternative_genotyping_example/allseqs'.

Genotyping sequences with 'Lighten' method and the following criteria ('parameter' 'marker' 'values'):
        cluster_inframe all     1
        min_amplicon_depth      all     100
        max_allele_number       all     10
        min_dominant_frequency_threshold        all     2
        error_threshold all     3

        MHC2-S1 genotyping
        MHC2-S2 genotyping
        MHC2-S1 genotyped (395 sequences, 1 unique)
        MHC2-S2 genotyped (349 sequences, 2 unique)

Printing verbose information about alleles and artifacts into 'results/a6237ae055d7b4cad2b42bc09770b35c/Alternative_genotyping_example/lighten'.

Extracting putative allele sequences into 'results/a6237ae055d7b4cad2b42bc09770b35c/Alternative_genotyping_example/lighten'.

Sequences per amplicon:
Amplicon        Total   Unique  Total-lighten   Alleles-lighten
MHC2-S1 426     67      395     1
MHC2-S2 411     124     349     2
MHC2-S3 387     78      376     1
MHC2-S4 352     81      345     1
MHC2-S5 457     140     353     2

Analysis results stored into 'results/a6237ae055d7b4cad2b42bc09770b35c/Alternative_genotyping_example.zip'.
```

Figure 10. A: AmpliLEGACY input form. B: Message with the link to download results. C: Output after comparison completion.

11. Compare results with AmpliCOMPARE

To show the capabilities of the AmpliCOMPARE tool we will compare the results obtained previously in AmpliCHECK and AmpliSAS runs (see Sections 8 and 9). As noticed before (read Section 9), low frequency variants that include sequencing errors and contaminations have been removed after clustering and filtering with AmpliSAS, obtaining clean genotyping results. AmpliCOMPARE can be also used to compare the genotyping results between experimental replicates, or to mark the changes after using different genotyping strategies.

To compare two Excel result files obtained in previous analyses let's do the following steps:

1. Go to AmpliCOMPARE form (Figure 12A):
   http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplicompare

2. Give a name to the run and write your email address (optional) if you want to receive a link to the results by email.

3. Give the location of the Excel files with genotyping results to compare clicking on the 'Browse…' button (download file 1 & file 2).

   IMPORTANT: spreadsheet names associated with the markers must be the same in both files.

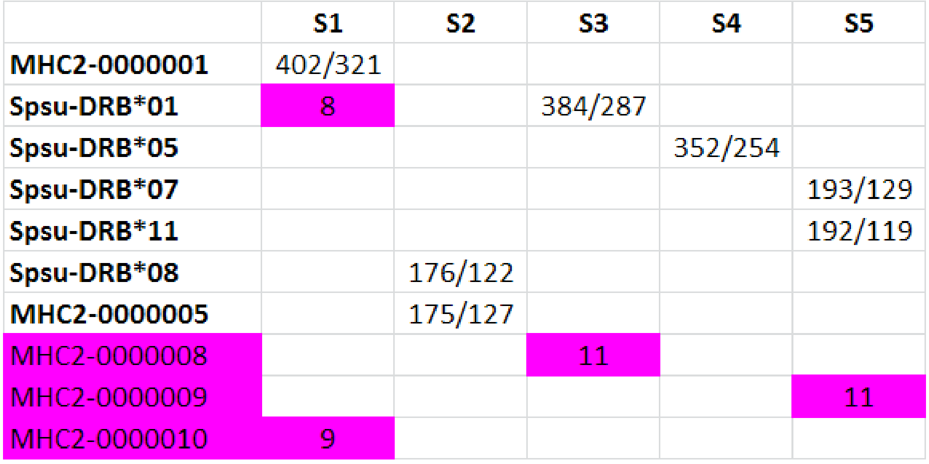
4. Click on the 'Run' button at the bottom of the page to start the merging process.

5. A link to the results will appear (Figure 12B), copy it to download the results later or wait for the run to finish and the results will appear directly on the screen (Figure 12C). If you wrote your email address, you will receive a notification when the process will finish with the link to download the results.

The output of the comparison process will be something like this (Figure 12C):

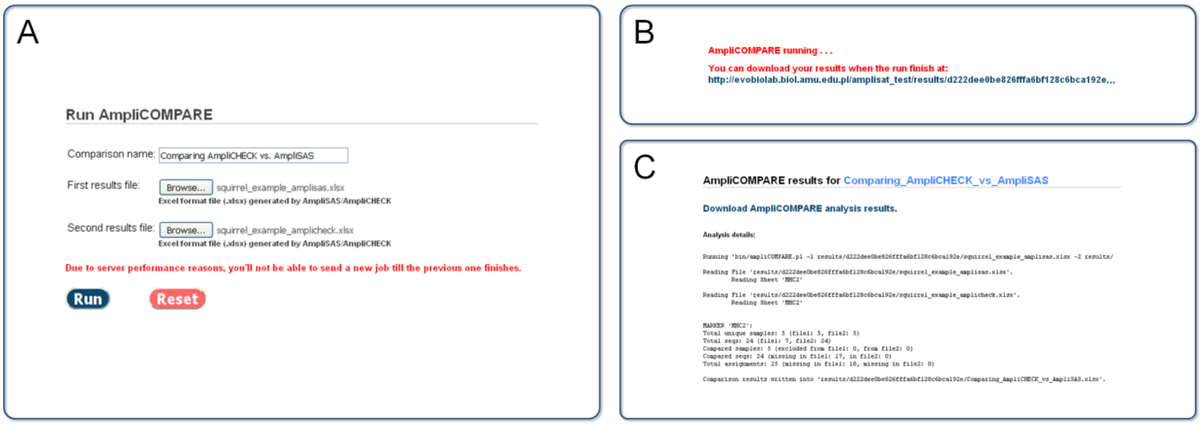
```
Running 'bin/ampliCOMPARE.pl ...

Reading File  ...

MARKER 'MHC2':
Total unique samples: 5 (file1: 5, file2: 5)
Total seqs: 24 (file1: 7, file2: 24)
Compared samples: 5 (excluded from file1: 0, from file2: 0)
Compared seqs: 24 (missing in file1: 17, in file2: 0)
Total assignments: 25 (missing in file1: 18, missing in file2: 0)

Comparison results written into  ...
```

In the Figure 11 we can check the comparison results for this example. Individuals/samples are shown in columns and variants in rows, the numeric values show is the variant depths into the amplicons. The numeric values show the depth of the variants shared by both compared files. Magenta color marks variants that are present in the second file and not in the first. Variants that are in the first and not in the second file will be marked in cyan (not happening in this example). The 'Spsu-DRB*01' 8 reads contamination detected with AmpliCHECK in the sample 'S1', latterly filtered by AmpliSAS (see Sections 8 and 9, Figure 5 and Figure 7) is magenta colored. Also other low frequency sequencing errors corrected by AmpliSAS are colored (MHC2-0000008-MHC2-0000010). The rest of the variants agree in both experiments and their depths are shown separated by a dash ("/").



	S1	S2	S3	S4	S5
MHC2-0000001	402/321				
Spsu-DRB*01	8		384/287		
Spsu-DRB*05				352/254	
Spsu-DRB*07					193/129
Spsu-DRB*11					192/119
Spsu-DRB*08		176/122			
MHC2-0000005		175/127			
MHC2-0000008			11		
MHC2-0000009					11
MHC2-0000010	9				

Figure 11. AmpliCOMPARE Excel output file example. The numeric values show the depth of the variants shared by both compared files. Magenta color marks variants that are present in the second file and not in the first. Variants that are in the first and not in the second file will be marked in cyan (not happening in this example).



Run AmpliCOMPARE

Comparison name: Comparing AmpliCHECK vs. AmpliSAS

First results file: squirrel_example_amplisas.xlsx
Excel format file (.xlsx) generated by AmpliSAS/AmpliCHECK

Second results file: squirrel_example_amplicheck.xlsx
Excel format file (.xlsx) generated by AmpliSAS/AmpliCHECK

Due to server performance reasons, you'll not be able to send a new job till the previous one finishes.

AmpliCOMPARE running . . .

You can download your results when the run finish at:
http://evobiolab.biol.amu.edu.pl/amplisat_test/results/d222dee0be826fffa6bf128c6bca192e...

AmpliCOMPARE results for Comparing_AmpliCHECK_vs_AmpliSAS

Download AmpliCOMPARE analysis results.

Analysis details:

```
Running 'bin/ampliCOMPARE.pl -1 results/d222dee0be826fffa6bf128c6bca192e/squirrel_example_amplisas.xlsx -2 results/

Reading File 'results/d222dee0be826fffa6bf128c6bca192e/squirrel_example_amplisas.xlsx'.
        Reading Sheet 'MHC2'

Reading File 'results/d222dee0be826fffa6bf128c6bca192e/squirrel_example_amplicheck.xlsx'.
        Reading Sheet 'MHC2'


MARKER 'MHC2':
Total unique samples: 5 (file1: 5, file2: 5)
Total seqs: 24 (file1: 7, file2: 24)
Compared samples: 5 (excluded from file1: 0, from file2: 0)
Compared seqs: 24 (missing in file1: 17, in file2: 0)
Total assignments: 25 (missing in file1: 18, missing in file2: 0)

Comparison results written into 'results/d222dee0be826fffa6bf128c6bca192e/Comparing_AmpliCHECK_vs_AmpliSAS.xlsx'.
```

Figure 12. A: AmpliCOMPARE input form. B: Message with the link to download results. C: Output after comparison completion.

12. Combining results with AmpliCOMBINE

We have a second experiment with the same squirrel individuals sequenced in different runs to confirm the genotyping results of the first one (replicate). Now, we are going to combine the genotyping results obtained from both experiments into a unique Excel file with the tool AmpliCOMBINE.

The paired-end reads of the second experiment have been merged previously as shown in Section 1.4 and the individual genotypes have been retrieved with AmpliSAS as explained in Section 1.8.

To combine two Excel result files obtained in previous analyses let's do the following steps:

1. Go to AmpliCOMBINE web form (Figure 14A): http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplicombine

2. Give a name to the run and write your email address (optional) if you want to receive a link to the results by email.

6. Give the location of the Excel files with genotyping results from both experiments by clicking on the 'Browse…' button (download file 1 & file 2).

   IMPORTANT: spreadsheet names associated with the markers must be the same in both files.

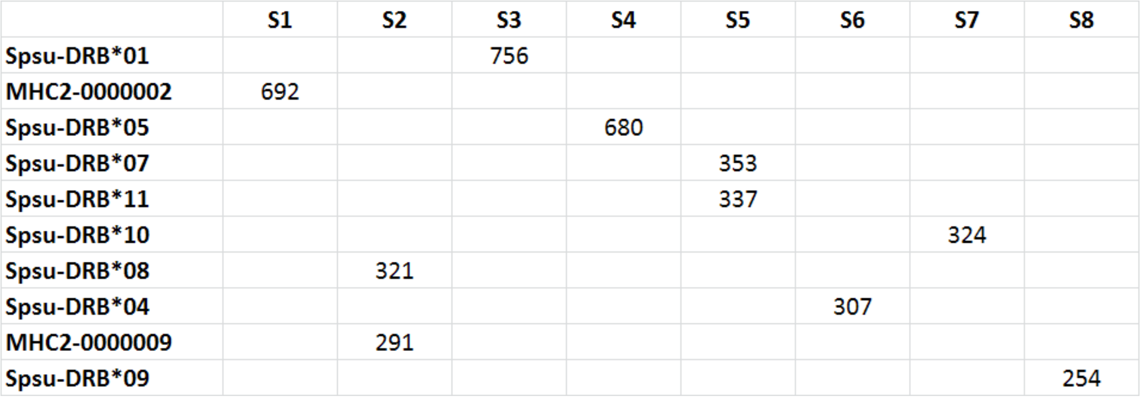
3. Click on the 'Run' button at the bottom of the page to start the merging process.

4. A link to the results will appear (Figure 14B), copy it to download the results later or wait for the run to finish and the results will appear directly on the screen (Figure 14C). If you wrote your email address, you will receive a notification when the process will finish with the link to download the results.

The output of the combination process will be something like this (Figure 14C):

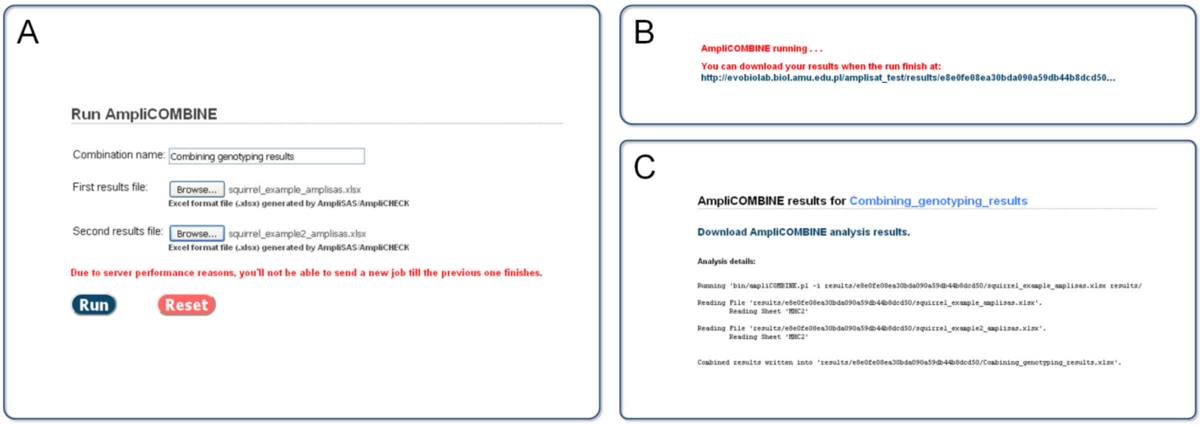
```
Running 'bin/ampliCOMBINE.pl ...

Reading File ...
  Reading Sheet 'MHC2'

Reading File ...
  Reading Sheet 'MHC2'


Combined results written into ...
```

In the Figure 13 we can check the combination results for this example. Individuals/samples are shown in columns and variants in rows, the numeric values show the variant depths into the amplicons. The three additional samples from the second experiment have been added to the results (S6, S7 and S8). When a variant is shared by both files their depths are summed together (e.g. variant 'Spsu-DRB*01' depth in sample 'S3' is 756, the sum of both files depths: 384+372). When variants have the same name in both files, they are kept, if not the variants are renamed (e.g. variant 'MHC2-0000002' was named 'MHC2-0000001' in the first file and 'MHC2-0000004' in the second).



	S1	S2	S3	S4	S5	S6	S7	S8
Spsu-DRB*01			756					
MHC2-0000002	692							
Spsu-DRB*05				680				
Spsu-DRB*07					353			
Spsu-DRB*11					337			
Spsu-DRB*10							324	
Spsu-DRB*08		321						
Spsu-DRB*04						307		
MHC2-0000009		291						
Spsu-DRB*09								254

Figure 13. AmpliCOMBINE Excel output file example. Numeric values show the depths of the variants, when variants are shared by both files their depths are summed together.



Run AmpliCOMBINE

Combination name: Combining genotyping results

First results file: squirrel_example_amplisas.xlsx
Excel format file (.xlsx) generated by AmpliSAS/AmpliCHECK

Second results file: squirrel_example2_amplisas.xlsx
Excel format file (.xlsx) generated by AmpliSAS/AmpliCHECK

Due to server performance reasons, you'll not be able to send a new job till the previous one finishes.

AmpliCOMBINE running . . .

You can download your results when the run finish at:
http://evobiolab.biol.amu.edu.pl/amplisat_test/results/e8e0fe08ea30bda090a59db44b8dcd50...

AmpliCOMBINE results for Combining_genotyping_results

Download AmpliCOMBINE analysis results.

Analysis details:

```
Running 'bin/ampliCOMBINE.pl -i results/e8e0fe08ea30bda090a59db44b8dcd50/squirrel_example_amplisas.xlsx results/

Reading File 'results/e8e0fe08ea30bda090a59db44b8dcd50/squirrel_example_amplisas.xlsx'.
        Reading Sheet 'MHC2'

Reading File 'results/e8e0fe08ea30bda090a59db44b8dcd50/squirrel_example2_amplisas.xlsx'.
        Reading Sheet 'MHC2'


Combined results written into 'results/e8e0fe08ea30bda090a59db44b8dcd50/Combining_genotyping_results.xlsx'.
```

Figure 14. A: AmpliCOMBINE input form. B: Message with the link to download results. C: Output after comparison completion.

13. HLA typing with AmpliHLA

AmpliHLA is designed to retrieve human genotypes by identifying amplicon variants among the thousands of HLA alleles annotated in the IMGT/HLA reference database. Several regions from each HLA locus should be amplified to obtain enough number variants and assign genotypes with good accuracy.

An example data is provided in AmpliSAT examples section, it consists in genomic sequences from exon 2 and exon 3 regions from class I HLA-A and HLA-B loci in five human cell lines sequenced with Illumina MiSeq. The first step will be to run AmpliSAS (or AmpliLEGACY) to obtain an Excel file with the variants assigned to each individual. Then we will follow these steps:

1. Go to AmpliHLA web form (Figure 16A):
   http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplihla

2. Give a name to the run.

3. Provide the location of the Excel file with genotyping results by clicking on the 'Browse…' button (download).

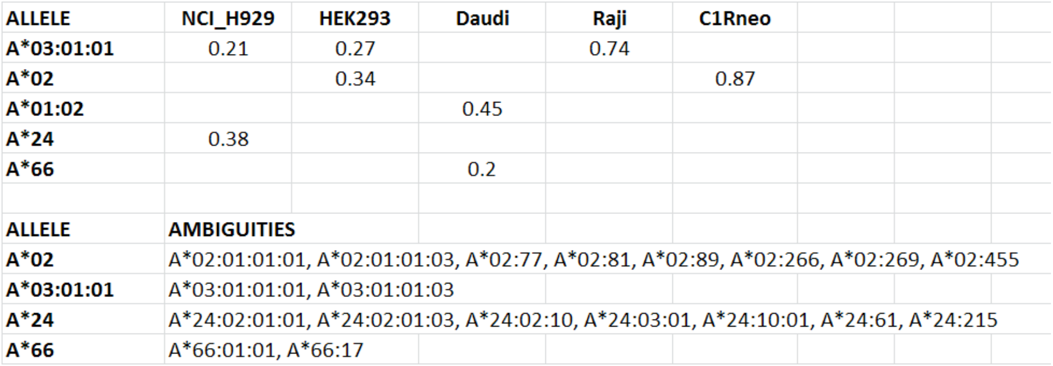
4. Click on the 'Run' button at the bottom of the page to start the analysis.

5. A link to the results will appear (Figure 16B), copy it to download the results later or wait for the run to finish and the results will appear directly on the screen (Figure 16C). If you wrote your email address, you will receive a notification when the process will finish with the link to download the results.

The output of the combination process will be something like this (Figure 16C):

```
Running 'bin/ampliHLA.pl ...

Reading HLA allele sequences from 'bin/imgt/201509_hla.fa.gz'.

Reading File ...
  Reading Sheet 'HLA_A2'
  Reading Sheet 'HLA_A3'
  Reading Sheet 'HLA_B2'
  Reading Sheet 'HLA_B3'

Matching allele sequences.

HLA typing results written into ....
```

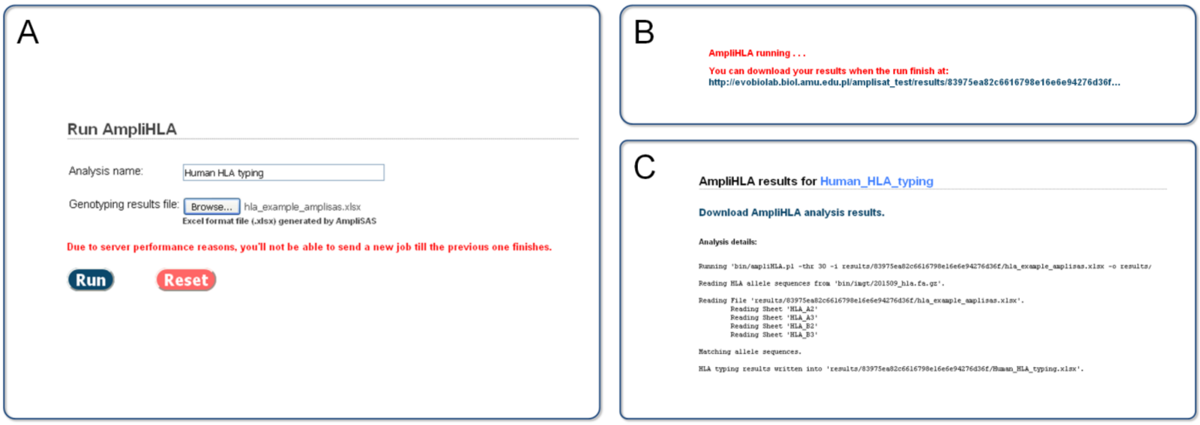
In the Figure 15 are displayed the HLA typing results. Individuals/samples are shown in columns and variants in rows, the numeric values show the allele frequencies in each amplicon. Genotypes are given with the highest resolution that can be achieved by the markers used in the experiment. In this example only 2 exonic regions are amplified so most of the genotypes are low-resolution. The full set of alleles that match individual variants are listed as ambiguities.



ALLELE	NCI_H929	HEK293	Daudi	Raji	C1Rneo
A*03:01:01	0.21	0.27		0.74	
A*02		0.34			0.87
A*01:02			0.45		
A*24	0.38				
A*66			0.2		

ALLELE	AMBIGUITIES
A*02	A*02:01:01:01, A*02:01:01:03, A*02:77, A*02:81, A*02:89, A*02:266, A*02:269, A*02:455
A*03:01:01	A*03:01:01:01, A*03:01:01:03
A*24	A*24:02:01:01, A*24:02:01:03, A*24:02:10, A*24:03:01, A*24:10:01, A*24:61, A*24:215
A*66	A*66:01:01, A*66:17

Figure 15. AmpliHLA Excel output file example. Numeric values show the amplicon frequencies of the alleles. Genotypes are given with the highest resolution that can be achieved by the markers used in the experiment, in the example only 2 exonic regions are amplified. The full set of alleles that match individual variants are listed as ambiguities.



Run AmpliHLA

Analysis name: Human HLA typing

Genotyping results file: Browse... hla_example_amplisas.xlsx
Excel format file (.xlsx) generated by AmpliSAS

Due to server performance reasons, you'll not be able to send a new job till the previous one finishes.

AmpliHLA running . . .

You can download your results when the run finish at:
http://evobiolab.biol.amu.edu.pl/amplisat_test/results/83975ea82c6616798e16e6e94276d36f...

AmpliHLA results for Human_HLA_typing

Download AmpliHLA analysis results.

Analysis details:

```
Running 'bin/ampliHLA.pl -thr 30 -i results/83975ea82c6616798e16e6e94276d36f/hla_example_amplisas.xlsx -o results/

Reading HLA allele sequences from 'bin/imgt/201509_hla.fa.gz'.

Reading File 'results/83975ea82c6616798e16e6e94276d36f/hla_example_amplisas.xlsx'.
        Reading Sheet 'HLA_A2'
        Reading Sheet 'HLA_A3'
        Reading Sheet 'HLA_B2'
        Reading Sheet 'HLA_B3'

Matching allele sequences.

HLA typing results written into 'results/83975ea82c6616798e16e6e94276d36f/Human_HLA_typing.xlsx'.
```

Figure 16. A: AmpliHLA input form. B: Message with the link to download results. C: Output after comparison completion.

14. TCR ANALYSIS WITH AMPLITCR

AmpliTCR was designed to enable *de novo* identification of V, J and D segments from high-throughput sequencing (HTS) reads that cover the entire Variable domain of the receptor (e.g. 2×300bp paired-end Illumina sequencing). AmpliTCR analyzes a set of genomic or transcriptomic TCR sequences through recognition and extraction of their Variable (V), Joining (J), Constant (C) and Complementarity Determining 3 (CDR3) regions. Input sequences are automatically translated to **recognize the TCR regions based only on patterns of highly conserved residues** (e.g. the cysteines that form disulfide bonds to link the TCR chains), and do not require prior knowledge about the V/D/J segments in the species of interest.

The extracted TCR region sequences are clustered based on similarity, to discard errors and retrieve only allelic variants. In this step sequencing artefacts (i.e. low-frequency variants highly similar to a higher frequency variants) are detected and corrected. In the case of the CDR3 region, the sequences are not clustered and the full sequence repertoire is extracted. However, for the purpose of a detailed analysis of CDR3 region, the AmpliCDR3 tool is recommended (see next Section), as it implements a more robust error-correction pipeline that can integrate unique molecular identifiers (UMIs). A comprehensive example of *de novo* TCR repertoire analysis in a non-model species with AmpliTCR and AmpliCDR3 is available in Migalska et al. 2017 (preprint version at bioRxiv).

An example of TCR data from this publication is provided here: it consists of TCRβ transcriptomic data from a bank vole (*Myodes glareolus*), obtained by the 5'RACE-based HTS library preparation method (more information about the technique: Mamedov *et al.* 2013; Migalska *et al.* 2017).

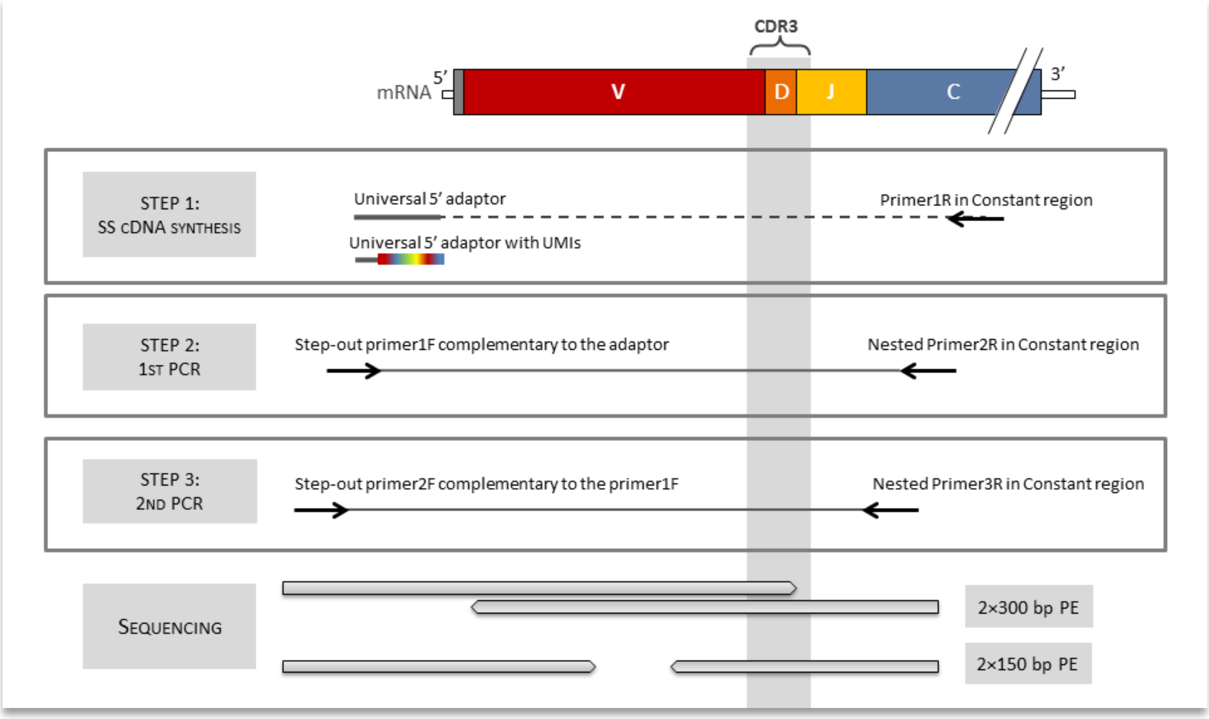

CDR3
mRNA
5'
V
D
J
C
3'
STEP 1:
SS cDNA SYNTHESIS
Universal 5' adaptor
Primer1R in Constant region
Universal 5' adaptor with UMIs
STEP 2:
1ST PCR
Step-out primer1F complementary to the adaptor
Nested Primer2R in Constant region
STEP 3:
2ND PCR
Step-out primer2F complementary to the primer1F
Nested Primer3R in Constant region
SEQUENCING
2×300 bp PE
2×150 bp PE

Figure 17. Simplified overview of a library preparation protocol from RNA with 5'RACE for deep HTS profiling. Based on Mamedov *et al.* 2013; Migalska *et al.* 2017**.**

To facilitate the analysis, the paired-end sequencing reads have been already merged with AmpliMERGE. To analyze this data we will do the following steps:

1. Go to AmpliTCR web form (Figure 18A):
   http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplitcr

2. Give a name to the run.

3. Select the 'Sequences file' with the merged reads (download) clicking on the 'Browse…' button.

4. Provide the desired protein patterns to extract the TCRβ regions in PROSITE syntax.
   - Variable region pattern: Qx[PS]x(14)Cx(10,11)WYx(39,42)[LM]x(14)C
   - Joining region pattern: GxGx(2)Lx[VI]
   - Constant region pattern: EDL*

   The provided patterns are also valid for human TCRβ, these conserved residues should be common in most of the mammals.

5. Click on the 'Run' button at the bottom of the page to start the analysis.

6. A link to the results will appear (Figure 18B), copy it to download the results later or wait for the run to finish and the results will appear directly on the screen (Figure 18C). If you wrote your email address, you will receive a notification when the process will finish with the link to download the results.

The output of the combination process will be something like this (Figure 18C):

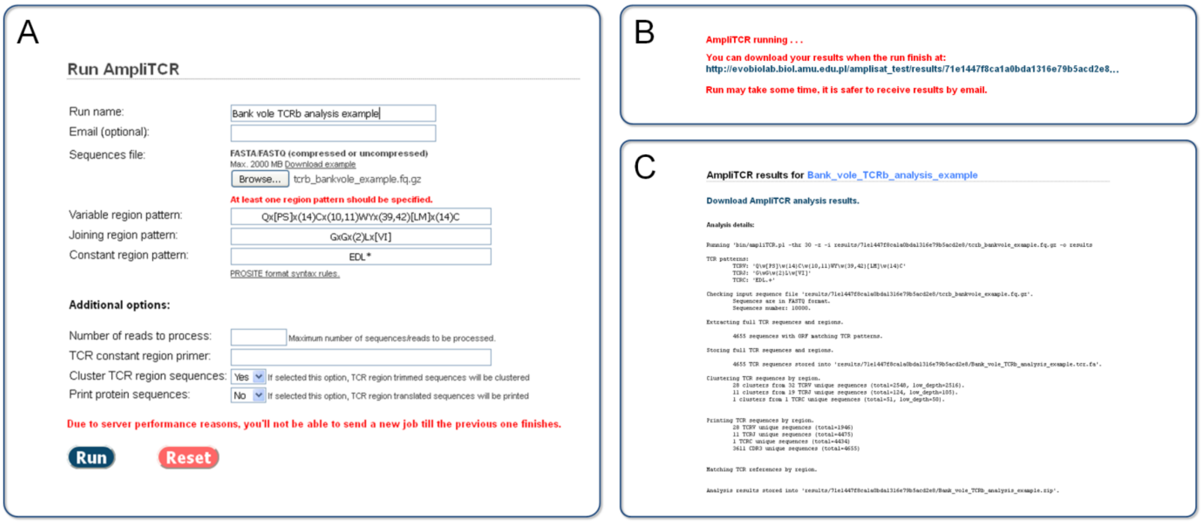
```
Running 'bin/ampliTCR.pl ...

TCR patterns:
  TCRV: 'Q\w[PS]\w{14}C\w{10,11}WY\w{39,42}[LM]\w{14}C'
  TCRJ: 'G\wG\w{2}L\w[VI]'
  TCRC: 'EDL.+'

Checking input sequence file ...
  Sequences are in FASTQ format.
  Sequences number: 10000.

Extracting full TCR sequences and regions.

  4655 sequences with ORF matching TCR patterns.

Storing full TCR sequences and regions.

  4655 TCR sequences stored into ...

Clustering TCR sequences by region.
  28 clusters from 32 TCRV unique sequences (total=2548, low_depth=2516).
  11 clusters from 19 TCRJ unique sequences (total=124, low_depth=105).
  1 clusters from 1 TCRC unique sequences (total=51, low_depth=50).

Printing TCR sequences by region.
  28 TCRV unique sequences (total=1946) stored into ...
  11 TCRJ unique sequences (total=4475) stored into ...
  1 TCRC unique sequences (total=4434) stored into ...
  3611 CDR3 unique sequences (total=4655) stored into ...

Matching TCR references by region.

Analysis results stored into ...
```

The compressed results will contain individual FASTA files with the allelic variants of the TCRβ regions and an additional file with the CDR3 repertoire (as stated before, this repertoire is not error-corrected, use AmpliCDR3 for a more accurate analysis).

There are additional parameters that can be tuned:

- *Number of reads to process:* limitation will result in random sub-sampling of the number of reads; this option is useful to control the sequencing depth between amplicons.

- *TCR constant segment primer*: when the primer sequence is provided reads will be filtered before the analysis – all non-specific amplification products will be removed, therefore increasing speed and accuracy of the overall analysis.
- *Cluster TCR segment sequencing errors with the following algorithm*: if this option is enabled the trimmed sequences of particular V/D/J segments will be clustered based on similarity and frequency thresholds using either AmpliSAS or CD-HIT algorithms. This step allows grouping of putative allelic variants with their respective sequencing artefacts. Further manual inspection and curation of retrieved alleles is however advised.
- *Print protein sequences*: If this option is enabled, translated TCR allelic variant sequences will be printed in additional FASTA files.



Run AmpliTCR

Run name: Bank vole TCRb analysis example
Email (optional):
Sequences file: FASTA/FASTQ (compressed or uncompressed)
Max. 2000 MB Download example
Browse... tcrb_bankvole_example.fq.gz
At least one region pattern should be specified.
Variable region pattern: Qx[PS]x(14)Cx(10,11)WYx(39,42)[LM]x(14)C
Joining region pattern: GxGx(2)Lx[VI]
Constant region pattern: EDL*
PROSITE format syntax rules.

Additional options:

Number of reads to process: Maximum number of sequences/reads to be processed.
TCR constant region primer:
Cluster TCR region sequences: Yes If selected this option, TCR region trimmed sequences will be clustered
Print protein sequences: No If selected this option, TCR region translated sequences will be printed

Due to server performance reasons, you'll not be able to send a new job till the previous one finishes.

AmpliTCR running . . .

You can download your results when the run finish at:
http://evobiolab.biol.amu.edu.pl/amplisat_test/results/71e1447f8ca1a0bda1316e79b5acd2e8...

Run may take some time, it is safer to receive results by email.

AmpliTCR results for Bank_vole_TCRb_analysis_example

Download AmpliTCR analysis results.

Analysis details:

```
Running 'bin/ampliTCR.pl -thr 30 -z -i results/71e1447f8ca1a0bda1316e79b5acd2e8/tcrb_bankvole_example.fq.gz -o results

TCR patterns:
        TCRV: 'Q\w[PS]\w(14)C\w(10,11)WY\w(39,42)[LM]\w(14)C'
        TCRJ: 'G\wG\w(2)L\w[VI]'
        TCRC: 'EDL.+'

Checking input sequence file 'results/71e1447f8ca1a0bda1316e79b5acd2e8/tcrb_bankvole_example.fq.gz'.
        Sequences are in FASTQ format.
        Sequences number: 10000.

Extracting full TCR sequences and regions.

        4655 sequences with ORF matching TCR patterns.

Storing full TCR sequences and regions.

        4655 TCR sequences stored into 'results/71e1447f8ca1a0bda1316e79b5acd2e8/Bank_vole_TCRb_analysis_example.tcr.fa'.

Clustering TCR sequences by region.
        28 clusters from 32 TCRV unique sequences (total=2548, low_depth=2516).
        11 clusters from 19 TCRJ unique sequences (total=124, low_depth=105).
        1 clusters from 1 TCRC unique sequences (total=51, low_depth=50).


Printing TCR sequences by region.
        28 TCRV unique sequences (total=1946)
        11 TCRJ unique sequences (total=4475)
        1 TCRC unique sequences (total=4434)
        3611 CDR3 unique sequences (total=4655)


Matching TCR references by region.


Analysis results stored into 'results/71e1447f8ca1a0bda1316e79b5acd2e8/Bank_vole_TCRb_analysis_example.zip'.
```

Figure 18. A: AmpliTCR input form. B: Message with the link to download results. C: Output after comparison completion.

15. TCR CDR3 REGION ANALYSIS WITH AMPLICDR3

AmpliCDR3 recognizes and extracts CDR3 region sequences from targeted TCR amplicon sequencing with HTS. Reads can cover the full Variable region of the receptor (e.g. paired-end 2×300bp Illumina sequencing), or partial length of the Variable region (e.g. paired end 2×125/2×150 bp Illumina sequencing). CDR3 region boundaries are recognized by the presence of DNA motifs that mark the beginning and the end of the CDR3 region. AmpliCDR3 uses as input DNA motifs in the standard REGEX format used in script programming. For human and mouse alpha, beta, gamma and delta TCR chains, and the beta chain in bank vole, these motifs are defined by default in AmpliCDR3 (Table 2),. For other non-model species, it is possible to create user-defined motifs and provide them in a field: '*CDR3 region pattern*'.

Table 2. AmpliCDR3 default DNA motifs in REGEX format to extract CDR3 regions in TCR alpha, beta, gamma and beta chains from human, mouse and bank vole (only beta).

<table>
<tr><th>Species</th><th>TCR chain</th><th>DNA motif</th></tr>
<tr><td rowspan="4">Human</td><td>Alpha</td><td>TA[TC]\w[TA][CT]TG[TC](\w+?)\w{34}ATATCCAGAA</td></tr>
<tr><td>Beta</td><td>TA[TC]\w{3}[TC][GA][TC]([AG][GC]\w+?)\w{31}AGGACCTGA</td></tr>
<tr><td>Gamma</td><td>[TA]A[TC][TC]ACTG[TC](\w+?)\w{34}AACAACTTGA</td></tr>
<tr><td>Delta</td><td>TACT[AT][CT]TGT(GC\w+?)\w{33}AGAAGTCAG</td></tr>
<tr><td rowspan="4">Mouse</td><td>Alpha</td><td>T\w[TC]\wT\w[TC]T\wTG[TC]([GCA][CG]\w+?)(TTT|TTC|CTG)GG\w{4}GG</td></tr>
<tr><td>Beta</td><td>T[AT]\w[TC][TA][TCG][TG]G[TCG]([GAT][CG]\w+?)\w{31}AGGATCTGA</td></tr>
<tr><td>Gamma</td><td>TA[TC]TACTGT(\w+?)\w{34}(AAAAGCCAG|AAAGGCTTG|ACAAAGCTC)</td></tr>
<tr><td>Delta</td><td>TA[TC][TC][AT]CTGT(G\w+?)\w{33}AAAAGCCAG</td></tr>
<tr><td>Vole</td><td>Beta</td><td>TG[TC]([GA][CG]\w+)\w{31}AGGA[CT]CTGA</td></tr>
</table>

In principle, AmpliCDR3 can analyze data from any vertebrate species, without reference sequences of V and J segments. When possible, and to improve analysis efficiency, **specify the distance from the last nucleotide in the CDR3 region to the first one in the constant region of the receptor, or more practically – to the primer complementary to constant region of the transcript** (parameter 'primer_r_dist' within the 'amplicon data' CSV input file, see also Figure 19).

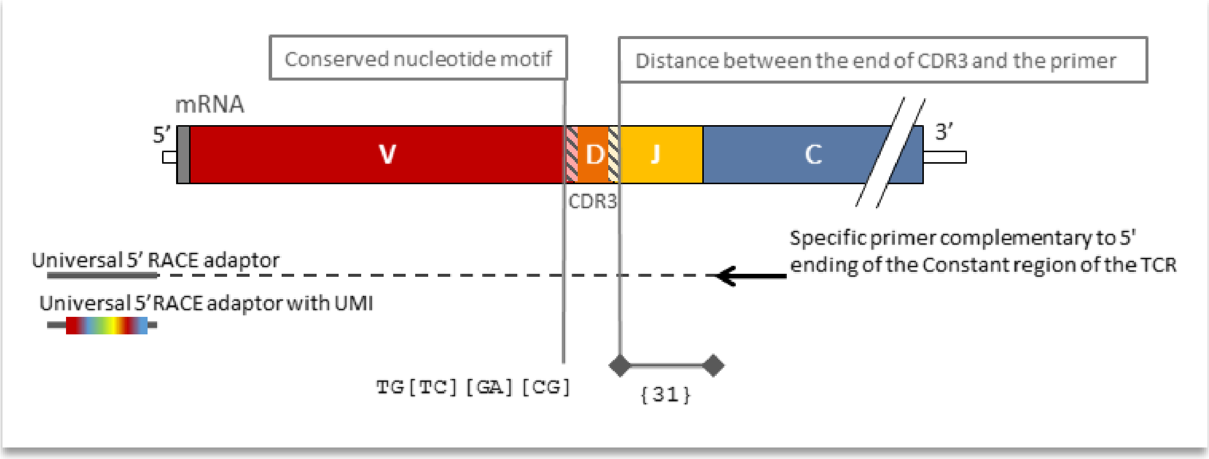

Conserved nucleotide motif
Distance between the end of CDR3 and the primer
mRNA
5'
V
D
J
C
3'
CDR3
Specific primer complementary to 5' ending of the Constant region of the TCR
Universal 5' RACE adaptor
Universal 5'RACE adaptor with UMI
TG[TC][GA][CG]
{31}

Figure 19. Schematic of the key AmpliCDR3 parameters that allow extraction of the CDR3 region. Values correspond to the parameters used for bank vole TCRβ analysis (Migalska et al. 2017).

Extracted CDR3 sequences are de-multiplexed amplicon by amplicon into variants (information about the sequencing depth of each variant is preserved), and filtered, to remove variants abnormally short or long (<15 and >63 bp), not in-frame and/or containing stop-codons.

The last step allows error correction. If an option '*Cluster CDR3 errors*' is enabled, CDR3 variants are clustered based on their similarity to remove artefacts such as sequencing errors, PCR errors or chimeras. However - **use this option carefully because clusters may include several real CDR3 variants!** Currently, the most efficient way to tackle HTS (and PCR) errors in immune repertoire profiling is to implement **molecular barcoding with Unique Molecular Identifiers** (UMIs, (Shugay *et al.* 2014) at the library preparation step. AmpliCDR3 can process reads with UMIs (Figure 20): their presence must be specified in the 'amplicon data' CVS file as multiple Ns, and an option '*Cluster CDR3 within UMIs*' should be enabled. Error-correction algorithm implemented in AmpliCDR3 tool is based on the principles of the MIGEC (molecular identifier groups–based error correction) strategy developed by Shugay and colleagues (2014), but it is simplified to increase speed and modified to account for lower sequencing depths. To infer the correct sequence of a CDR3, AmpliCDR3 groups reads with identical UMIs into clusters, or, to follow nomenclature developed by Shugay *et al.* 2014: molecular identifier groups (MIGs). If more than 50% of reads in an UMI cluster (MIG) differ by more than 2bp from the major sequence, the consensus cannot be inferred and the cluster is discarded. Such ambiguous/mixed UMI clusters (MIGs) can result from e.g. presence of PCR chimeras or early PCR errors. With insufficient per amplicon sequencing depth, several UMI clusters (MIGs) will contain only one read (singleton UMIs), which also precludes reliable error correction. Such sequences can be discarded by enabling option: '*Remove singletons*'.

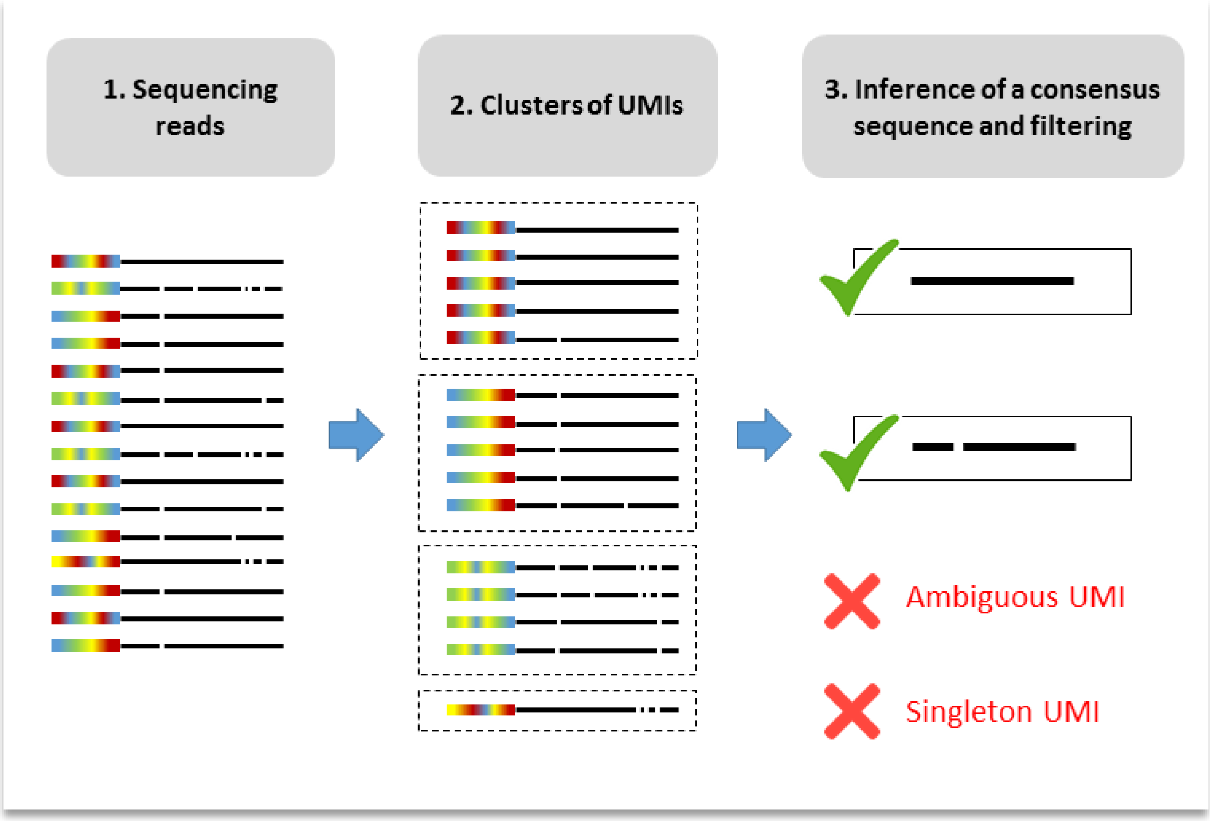


Figure 20. UMI-based error correction scheme. UMIs are shown as rainbow-colored rectangles, CDR3 reads are presented as black lines. Differences between sequences are depicted as white marks.

An example of TCR data from Migalska *et al.* 2017 is provided here, it consists of TCRβ transcriptomic data from a bank vole (*Myodes glareolus*) individual obtained by the 5'RACE-based library preparation method (more information about the technique: see Mamedov *et al.* 2013; Migalska *et al.* 2017). To facilitate the analysis, the paired-end sequencing reads have been already merged with AmpliMERGE. To analyze this data we will do the following steps:

1. Go to AmpliCDR3 web form (Figure 21A):
   http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplicdr3

2. Give a name to the run.

3. Select the 'Sequences file' with the merged reads (download) clicking on the 'Browse…' button. If reads have been already de-multiplexed into separate files (one file per sample), you can pack them into a single .zip or .tar.gz format file and use it as input.

4. Choose 'Vole' in the list of available species or specify manually the DNA motif to be used to extract the CDR3 sequences. If your species of interest is not on the list, choose 'Any', and modify the recognition pattern in the field: '*CDR3 region pattern*'.

5. Introduce the following information into the 'Amplicon data' textbox:

   ```
   >marker,primer_f,primer_f_id
   TCRB,CAACGCAGAG(NNNNNNNNNNNNNNNN),FWD_RACE
   >marker,primer_r,primer_r_id,primer_r_dist
   TCRB,CTCAGATCCT,REV_RACE,31
   >sample
   s092_A
   s092_B
   s092_C
   ```

   Important remarks:
   - Remember to write the field names on the first line after the character '>' and on the next rows write the values in the same order as the field names.
   - Field names and values must be separated by commas.
   - It is recommended to specify the name of the chain in the field marker.
   - If technical replicates for the same individual/sample are present, each replicate's ID should start with identical name followed by '_' or '-' and unique designation (see the example input).
   - **The parameter 'primer_r_dist'** specifies the distance from the last nucleotide in the CDR3 region to the first one in the primer complementary to the constant region of the TCR.
   - **Unique Molecular Identifier sequences (UMIs) must be indicated between parenthesis.** The number of Ns must match the length of the UMI.
   - Shortening primer sequences to 7-9 nts can increase the number of retrieved sequences (eg. GAGTGTCAT instead of GAGTGTCATTTCTCCAACGGGA).

6. If you want to retrieve V-J usage statistics provide a suitable reference in the filed '*Alleles file (optional)*'. Here, we could add a FASTA file with the bank vole Variable and Joining region beta chain gene alleles (download an example of bank vole TCRβ segments retrieved by Migalska et al. 2017), however, a suitable reference will be automatically included in the analysis after selecting 'Vole' from the Species list. While analyzing other species (that are not included in our software), prepare a suitable reference list of V and J segment genes. For a number of model species the reference sequences can be found in immune databases (e.g., THE INTERNATIONAL IMMUNOGENETICS INFORMATION SYSTEM®, http://www.imgt.org/). If you work with non-model species that lack references, you can use AmpliTCR (see previous section) to identify these genes.

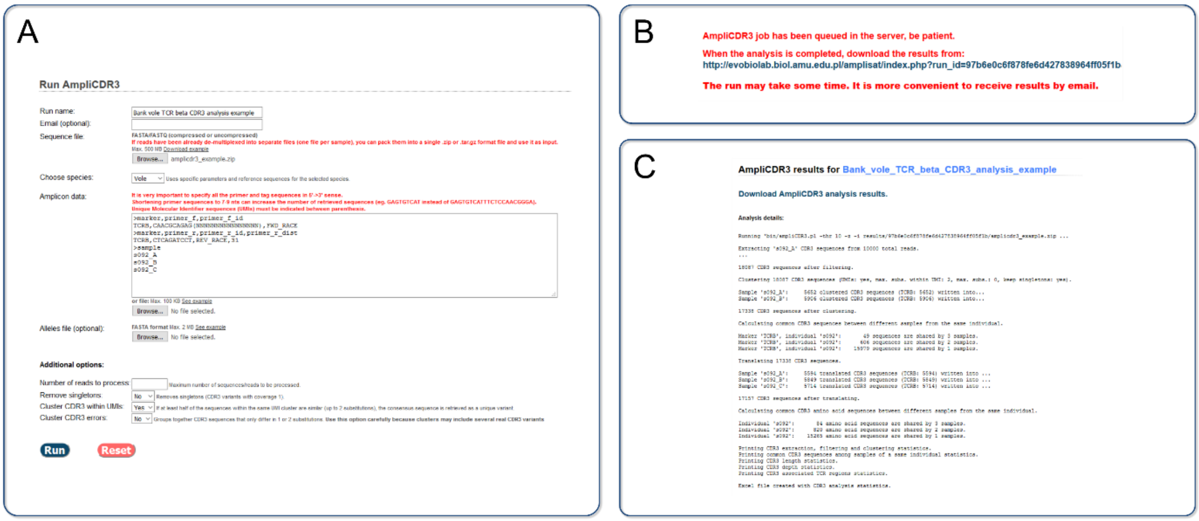
7. Additional options may be specified. In this case, it is particularly important to select '*Cluster CDR3 within UMIs*' to activate the UMI-based error correction and clustering. In this tutorial we do not discard singletons - because the provided file with sample reads contains a small subset of the original amplicon, most of the variants are covered by just one read. While analyzing real, high-depth data (especially with UMIs), we advise enabling this option.

8. Click on the 'Run' button at the bottom of the page to start the analysis.

9. A link to the results will appear (Figure 21B), copy it to download the results later or wait for the run to finish and the results will appear directly on the screen (Figure 21C). If you wrote your email address, you will receive a notification when the process will finish with the link to download the results.

The output of the combination process will be something like this (Figure 21C):

```
Running 'bin/ampliTCR.pl ...

TCR patterns:
  TCRV: 'Q\w[PS]\w{14}C\w{10,11}WY\w{39,42}[LM]\w{14}C'
  TCRJ: 'G\wG\w{2}L\w[VI]'
  TCRC: 'EDL.+'

Checking input sequence file ...
  Sequences are in FASTQ format.
  Sequences number: 10000.

Extracting full TCR sequences and regions.

  4655 sequences with ORF matching TCR patterns.

Storing full TCR sequences and regions.

  4655 TCR sequences stored into ...

Clustering TCR sequences by region.
  28 clusters from 32 TCRV unique sequences (total=2548, low_depth=2516).
  11 clusters from 19 TCRJ unique sequences (total=124, low_depth=105).
  1 clusters from 1 TCRC unique sequences (total=51, low_depth=50).

Printing TCR sequences by region.
  28 TCRV unique sequences (total=1946) stored into ...
  11 TCRJ unique sequences (total=4475) stored into ...
  1 TCRC unique sequences (total=4434) stored into ...
  3611 CDR3 unique sequences (total=4655) stored into ...

Matching TCR references by region.

Analysis results stored into ...
```



Run AmpliCDR3

Run name: Bank vole TCR beta CDR3 analysis example
Email (optional):
Sequence file: FASTA/FASTQ (compressed or uncompressed)
If reads have been already de-multiplexed into separate files (one file per sample), you can pack them into a single .zip or .tar.gz format file and use it as input.
Max. 500 MB Download example
Browse... amplicdr3_example.zip

Choose species: Vole Uses specific parameters and reference sequences for the selected species.

Amplicon data: It is very important to specify all the primer and tag sequences in 5'->3' sense.
Shortening primer sequences to 7-9 nts can increase the number of retrieved sequences (eg. GAGTGTCAT instead of GAGTGTCATTTCTCCAACGGGA).
Unique Molecular Identifier sequences (UMIs) must be indicated between parenthesis.

```
>marker,primer_f,primer_f_id
TCRB,CAACGCAGAG(NNNNNNNNNNNNNNNN),FWD_RACE
>marker,primer_r,primer_r_id,primer_r_dist
TCRB,CTCAGATCCT,REV_RACE,31
>sample
s092_A
s092_B
s092_C
```

or file: Max. 100 KB See example
Browse... No file selected.

Alleles file (optional): FASTA format Max. 2 MB See example
Browse... No file selected.

Additional options:

Number of reads to process: Maximum number of sequences/reads to be processed.
Remove singletons: No Removes singletons (CDR3 variants with coverage 1).
Cluster CDR3 within UMIs: Yes If at least half of the sequences within the same UMI cluster are similar (up to 2 substitutions), the consensus sequence is retrieved as a unique variant.
Cluster CDR3 errors: No Groups together CDR3 sequences that only differ in 1 or 2 substitutions. Use this option carefully because clusters may include several real CDR3 variants

Run Reset

AmpliCDR3 job has been queued in the server, be patient.

When the analysis is completed, download the results from:
http://evobiolab.biol.amu.edu.pl/amplisat/index.php?run_id=97b6e0c6f878fe6d427838964ff05f1b

The run may take some time. It is more convenient to receive results by email.

AmpliCDR3 results for Bank_vole_TCR_beta_CDR3_analysis_example

Download AmpliCDR3 analysis results.

Analysis details:

```
Running 'bin/ampliCDR3.pl -thr 10 -z -i results/97b6e0c6f878fe6d427838964ff05f1b/amplicdr3_example.zip ...

Extracting 's092_A' CDR3 sequences from 10000 total reads.
...

18087 CDR3 sequences after filtering.

Clustering 18087 CDR3 sequences (UMIs: yes, max. subs. within UMI: 2, max. subs.: 0, keep singletons: yes).

Sample 's092_A':     5652 clustered CDR3 sequences (TCRB: 5652) written into...
Sample 's092_B':     5906 clustered CDR3 sequences (TCRB: 5906) written into...

17338 CDR3 sequences after clustering.

Calculating common CDR3 sequences between different samples from the same individual.

Marker 'TCRB', individual 's092':       49 sequences are shared by 3 samples.
Marker 'TCRB', individual 's092':      606 sequences are shared by 2 samples.
Marker 'TCRB', individual 's092':    15979 sequences are shared by 1 samples.

Translating 17338 CDR3 sequences.

Sample 's092_A':     5594 translated CDR3 sequences (TCRB: 5594) written into ...
Sample 's092_B':     5849 translated CDR3 sequences (TCRB: 5849) written into ...
Sample 's092_C':     5714 translated CDR3 sequences (TCRB: 5714) written into ...

17157 CDR3 sequences after translating.

Calculating common CDR3 amino acid sequences between different samples from the same individual.

Individual 's092':       84 amino acid sequences are shared by 3 samples.
Individual 's092':      820 amino acid sequences are shared by 2 samples.
Individual 's092':    15265 amino acid sequences are shared by 1 samples.

Printing CDR3 extraction, filtering and clustering statistics.
Printing common CDR3 sequences among samples of a same individual statistics.
Printing CDR3 length statistics.
Printing CDR3 depth statistics.
Printing CDR3 associated TCR regions statistics.

Excel file created with CDR3 analysis statistics.
```

Figure 21. A: AmpliCDR3 input form. B: Message with the link to download results. C: Output after comparison completion.

Results can be downloaded as a compressed folder, that contains (for each individual analyzed): several FASTA files, text files and an Excel file with summary statistics.

FASTA files are generated for each amplicon separately, with the **CDR3 variants obtained after each step of the analysis** (CDR3 extraction, filtering, clustering and translation).

Text files are tab-separated files generated for each amplicon separately. **In rows are unique CDR3 sequences** recovered during analysis. Each sequence is assumed to represent one T cell clonotype (hence the name of the file). Each file contains the following columns:

- **cloneId –** ID number (arbitrarily assigned, starting from the highest-coverage sequences) of each unique CDR3 sequence.
- **cloneCount -** if UMI-correction had been enabled: number of different UMIs tagging identical CDR3 sequence. This should be equal to the number of unique cDNA templates containing given CDR3 sequence.
- **CloneFraction -** if UMI-correction had been enabled: fraction of the UMIs representing given CDR3 sequence.
- **totalReads –** number of reads representing given CDR3 sequence (if UMI-correction had been enabled: this value will be a sum of reads across all clusters of UMIs/MIGs representing given CDR3).
- **readFraction -** fraction of reads representing given CDR3 sequence.
- **nSeqCDR3 –** nucleotide sequence of CDR3.
- **aaSeqCDR3 –** amino acid sequence of CDR3.
- **allVHitsWithScore –** all V segments associated with given CDR3 (reported if reference V segment sequences had been provided). Note that identical CDR3 sequences may be generated as a results of a convergent recombination of different V segments.
- **allJHitsWithScore -** all j segments associated with given CDR3 (reported if reference J segment sequences had been provided). Note that identical CDR3 sequences may be generated as a results of a convergent recombination of different J segments.

Excel file **is generated for each individual (all amplicons from single individual). It summarizes basic statistics for all replicates available for this individual** (e.g., number of CDR3 variants retrieved per sample, CDR3 length distribution, repertoire sharing between replicates) and several statistics of V-J segment usage (if reference V and J segment sequences had been provided or reference is available for analyzed species at the website – currently only human, mice and vole). The following sheets should be present in each Excel file:

- **Unique -** TCR CDR3 extraction, filtering and clustering statistics for each amplicon.
- **Common -** Common CDR3 variants among technical replicates from the same individual/sample (this data can be used to assess repeatability or estimate total repertoire size using incidence-based richness estimators, such as Chao2 – see Migalska et al. 2017 for implementation).
- **Lengths** - Distribution of the CDR3 lengths – *in silico* spectratyping.
- **Depths** - Distribution of depths of CDR3 sequences (clustered or corrected with UMIs).
- **Regions -** Distribution of TCRJ and TCRV genes (per amplicon).
- **Regions2 -** J-V gene usage (per amplicon).
- **Regions3 -** CDR3 lengths vs. TCRJ and TCRV gene usage (per amplicon).
- **Clones –** the same results as in the text files (see above), limited to top 10000 positions.

APPENDIX

1. AMPLICON SEQUENCING TECHNIQUE

Amplicon sequencing (AS) basically consists in sequencing at once the amplification products of multiple PCRs, see Figure 22. Using new generation sequencing (NGS) technologies together with combinations of primers and barcodes it is possible to sequence targeted gen regions with deep coverage for hundreds, even thousands of individuals in a single experiment. The utility of AS is only limited by the intrinsic sequencing error rates of NGS technologies and other error sources like polymerase amplification or chimeras.

One of the AS pioneer applications was to assess the microbial diversity in deep sea water samples, discovering thousands of low-abundance populations not retrieved in previous studies (Sogin *et al.* 2006). The technique has been improved including barcorde sequences, also called molecular identifier tags (MIDs), to PCR products to be able to separate reads into samples after sequencing (Binladen *et al.* 2007; Meyer *et al.* 2007). Many successful studies have been carried ever since, and **AS is nowadays a widely used technique in metagenomics, ecology, population genetics and evolutionary biology** (Swenson 2012; Di Bella *et al.* 2013; Joly *et al.* 2014) (Di Bella et al. 2013; Swenson 2012; Joly et al. 2014), also it is getting increasing interest in clinical diagnostics and therapeutics (Erlich 2012; Chang & Li 2013). **Major NGS platforms have AS commercial solutions in their catalogues**: Illumina TruSeq Custom Amplicon, Roche 454 Fluidigm Access Array or Life Technologies Ion Torrent Ion AmpliSeq. As a new technique, different authors use their own definitions for common technical terms, in Table 1 are listed the most commonly used terms in amplicon sequencing and genotyping studies

In the field of evolutionary genetics AS has a main role in the genotyping of polymorphic multilocus systems such as Major Histocompatibility Complex (MHC), replacing the traditional cloning and Sanger sequencing method. MHC class I and class II gene families encode cell surface receptors that present antigens to immune cells and they are the most polymorphic genes among vertebrates (Mehra 2001; Penn 2002). The best studied example is human MHC, also called Human Leukocyte Antigen (HLA), because it plays a key role in organ transplantation compatibility. Thousands of HLA alleles have been identified, most of them by classical cloning and Sanger sequencing, their names and sequences are annotated in the IMGT/HLA database (Robinson *et al.* 2013). Classical HLA typing methods are: hybridization with sequence-specific oligonucleotide (SSO) probes and sequence-specific priming (SSP), but nowadays NGS technology, including AS, is starting to be used for high-resolution genotyping accuracy (Nunes *et al.* 2011; Wang *et al.* 2012; Erlich 2012; Boegel *et al.* 2013).

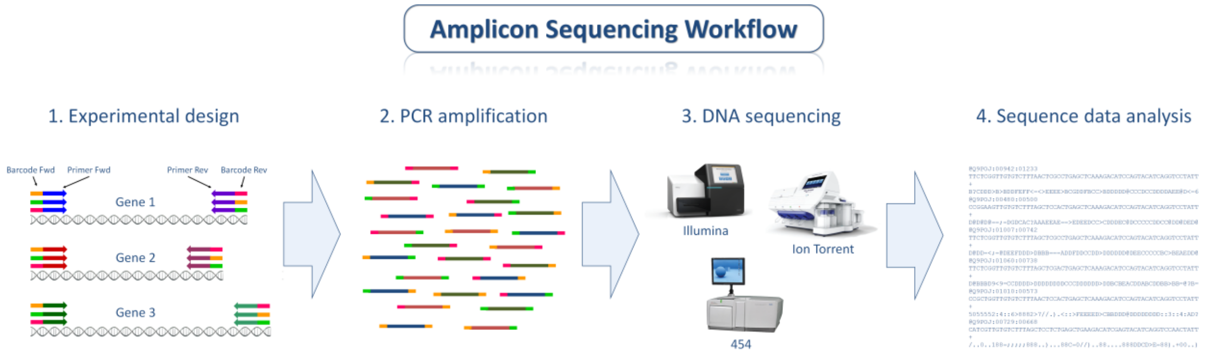


Figure 22. Amplicon sequencing typical workflow.

2. Sequencing errors

Relatively high error rates associated with AS, stemming both from intrinsic sequencing error rate of high-throughput technologies and PCR errors (Table 3), such as chimera formation, make genotyping using NGS challenging. For example, homopolymer regions are a major issue for pyrosequencing and ion semiconductor technologies (454 or Ion Torrent), where erroneous indels are introduced in high rates, whereas technology based on reversible dye-terminators (Illumina) suffers from a high number of not necessarily random substitutions (Gilles *et al.* 2011; Vandenbroucke *et al.* 2011; Liu *et al.* 2012; Loman *et al.* 2012; Bragg *et al.* 2013; Ross *et al.* 2013)

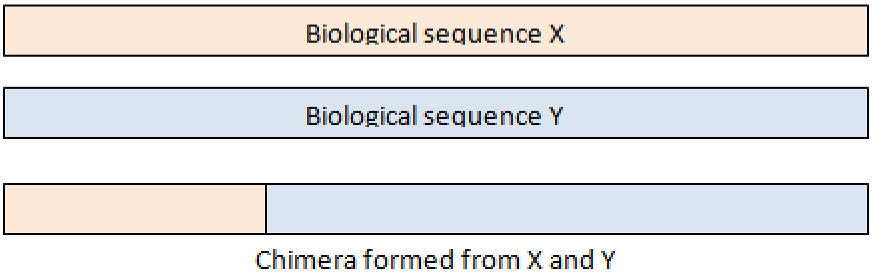
Table 3. Error rate (%, per 100 nts) comparison among several NGS technologies and sources. A. Ross *et al.* 2013**. B.** Liu *et al.* 2012**. C.** Loman *et al.* 2012**. D.** Vandenbroucke *et al.* 2011**. *HiSeq 2000 Illumina.**

	MiSeq Illumina			PGM Ion Torrent			454 Life Sciences	
	A	**B***	**C**	**A**	**B**	**C**	**C**	**D**
Mismatches	0.36	1.00	?	0.18	0.34	?	?	0.07
Insertions	<0.01	<0.01	<0.01	0.44	0.69	0.73	0.28	0.14
Deletions	<0.01	<0.01	<0.01	0.53	0.96	0.76	0.09	0.08
Total	0.37	1.01	?	1.15	1.99	>1.50	?	0.29

3. PCR CHIMERAS DETECTION

Source: http://drive5.com/usearch/manual/chimera_formation.html

Chimeras are sequences formed from two or more biological sequences joined together. Amplicons with chimeric sequences can form during PCR. Chimeras are rare with shotgun sequencing, but are common in amplicon sequencing where closely related sequences are amplified (Smyth *et al.* 2010; Schloss *et al.* 2011). Although chimeras can be formed by a number of mechanisms, the majority of chimeras are believed to arise from incomplete extension. During subsequent cycles of PCR, a partially extended strand can bind to a template derived from a different but similar sequence. This then acts as a primer that is extended to form a chimeric sequence. A chimeric template is created during one round and then amplified by subsequent rounds to produce chimeric amplicons. In amplicon sequencing, we typically find that only a small fraction of amplicon reads is chimeric, usually less than 1%. Chimeras from more than two sequences are very rare.



Though there is specific software available for chimera detection (Edgar *et al.* 2011), AmpliSAT has implemented AS-specific chimera detection rules:

- A chimeric variant is composed only by higher frequency ones not classified previously as artefacts (parental sequences).
- Each parental sequence must be present with a minimum length of 10bp in the chimera.
- Any parental sequence cannot have higher sequence identity with the chimera than the sequencing error identity threshold established for clustering. E.g. for a indel threshold of 0.001% and a substitution threshold of 1%, a parental sequence of 100bp must differ more than 1bp from the chimeric one. This is done to avoid erroneous chimera classifications when alleles are highly similar.
- The chimera detection algorithm compares all the combinations of higher frequency variants, e.g. A,B,C... against the lower frequency one Q:
    1. Compares A with Q to annotate how many nucleotides have in common in both sides.
    2. If A matches Q in the one side (>10bp), then looks for another sequence matching the opposite side the remaining length not matched by A.
    3. If another parental sequence B is found to match the opposite side, then the Q sequence is annotated as chimera of A+B.

BIBLIOGRAPHY

Babik W (2010) Methods for MHC genotyping in non-model vertebrates. *Molecular ecology resources*, **10**, 237–51.

Baum PD, Venturi V, Price D a (2012) Wrestling with the repertoire: the promise and perils of next generation sequencing for antigen receptors. *European journal of immunology*, **42**, 2834–9.

Di Bella JM, Bao Y, Gloor GB, Burton JP, Reid G (2013) High throughput sequencing methods and analysis for microbiome research. *Journal of microbiological methods*, **95**, 401–14.

Biedrzycka A, Kloch A, Buczek M, Radwan J (2011) Major histocompatibility complex DRB genes and blood parasite loads in fragmented populations of the spotted suslik Spermophilus suslicus. *Mammalian Biology - Zeitschrift für Säugetierkunde*, **76**, 672–677.

Biedrzycka A, Radwan J (2008) Population fragmentation and major histocompatibility complex variation in the spotted suslik, Spermophilus suslicus. *Molecular ecology*, **17**, 4801–11.

Binladen J, Gilbert MTP, Bollback JP *et al.* (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PloS one*, **2**, e197.

Boegel S, Löwer M, Schäfer M *et al.* (2013) HLA typing from RNA-Seq sequence reads. *Genome medicine*, **4**, 102.

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 computational biology*, **9**, e1003031.

Chang F, Li MM (2013) Clinical application of amplicon-based next-generation sequencing in cancer. *Cancer genetics*, **206**, 413–9.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics (Oxford, England)*, **27**, 2194–200.

Erlich H (2012) HLA DNA typing: past, present, and future. *Tissue antigens*, **80**, 1–11.

Georgiou G, Ippolito GC, Beausang J *et al.* (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature biotechnology*, **32**, 158–68.

Gilles A, Meglécz E, Pech N *et al.* (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC genomics*, **12**, 245.

Herdegen M, Babik W, Radwan J (2014) Selective pressures on MHC class II genes in the guppy (Poecilia reticulata) as inferred by hierarchical analysis of population structure. *Journal of Evolutionary Biology*, **27**, 2347–2359.

Joly S, Davies TJ, Archambault A *et al.* (2014) Ecology in the age of DNA barcoding: the resource, the promise and the challenges ahead. *Molecular ecology resources*, **14**, 221–32.

Kress WJ, García-Robledo C, Uriarte M, Erickson DL (2014) DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology & Evolution*, **30**, 25–35.

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). *Molecular ecology resources*, **14**, 753–767.

Liu L, Li Y, Li S *et al.* (2012) Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*, **2012**, 251364.

Loman NJ, Misra R V, Dallman TJ *et al.* (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, **30**, 434–9.

Mamedov IZ, Britanova O V, Zvyagin I V *et al.* (2013) Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling. *Frontiers in immunology*, **4**, 456.

Mehra NK (2001) Histocompatibility Antigens. *Enciclopedia of Life Sciences*.

Meyer M, Stenzel U, Myles S, Prüfer K, Hofreiter M (2007) Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic acids research*, **35**, e97.

Migalska M, Sebastian A, Radwan J (2017) Comprehensive profiling of TCRβ repertoire in a non-model species (the bank vole) using high-throughput sequencing. *bioRxiv*, 217653.

Nunes E, Heslop H, Fernandez-Vina M *et al.* (2011) Definitions of histocompatibility typing terms. *Blood*, **118**, e180-3.

Penn DJ (2002) Major Histocompatibility. *Enciclopedia of Life Sciences*.

Radwan J, Zagalska-Neubauer M, Cichoń M *et al.* (2012) MHC diversity, malaria and lifetime reproductive success in collared flycatchers. *Molecular Ecology*, **21**, 2469–2479.

Robinson WH (2015) Sequencing the functional antibody repertoire--diagnostic and therapeutic discovery. *Nature reviews. Rheumatology*, **11**, 171–82.

Robinson J, Halliwell J a, McWilliam H *et al.* (2013) The IMGT/HLA database. *Nucleic acids research*, **41**, D1222-7.

Ross MG, Russ C, Costello M *et al.* (2013) Characterizing and measuring bias in sequence data. *Genome biology*, **14**, R51.

Ruggiero E, Nicolay JP, Fronza R *et al.* (2015) High-resolution analysis of the human T-cell receptor repertoire. *Nature communications*, **6**, 8081.

Schloss PD, Gevers D, Westcott SL (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PloS one*, **6**, e27310.

Sebastian A, Herdegen M, Migalska M, Radwan J (2016) Amplisas: A web server for multilocus genotyping using next-generation amplicon sequencing data. *Molecular Ecology Resources*, **16**, 498–510.

Shugay M, Britanova O V, Merzlyak EM *et al.* (2014) Towards error-free profiling of immune repertoires. *Nature methods*, **11**, 653–5.

Smyth RP, Schlub TE, Grimm A *et al.* (2010) Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene*, **469**, 45–51.

Sogin ML, Morrison HG, Huber JA *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 12115–20.

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.

Swenson NG (2012) Phylogenetic analyses of ecological communities using DNA barcode data. *Methods in molecular biology (Clifton, N.J.)*, **858**, 409–19.

Vandenbroucke I, Van Marck H, Verhasselt P *et al.* (2011) Minor variant detection in amplicons using 454 massive parallel pyrosequencing: experiences and considerations for successful applications. *BioTechniques*, **51**, 167–77.

Wang C, Krishnakumar S, Wilhelmy J *et al.* (2012) High-throughput, high-fidelity HLA genotyping with deep sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 8676–81.