jMHC

Michał Stuglik, Jacek Radwan & Wiesław Babik

michal.stuglik@uj.edu.pl

The software is used for analyzing and visualization of the results of deep amplicon sequencing, as currently performed using 454 technology, but as the software operates on FASTA files, output from any technology can be used in principle. We developed this software for genotyping of major histocompatibility complex but we can imagine that it may be useful for analyzing amplicons derived from other multigene families or for genotyping other polymorphic systems.

The major difference between this kind of data and amplicons generated for example from environmental samples for biodiversity studies is that standard clustering methods will not work well here. The reason is that both very similar and more divergent alleles are expected to be present in a single amplicon, therefore simple clustering based on sequence similarity (i.e. sequences above some defined similarity threshold are clustered together and considered as a single unit) is not appropriate solution. Here, we analyze all unique sequence variants, which enables distinguishing true alleles differing even by a single nucleotide, if both are present in a substantial number of reads in the amplicon. The software is particularly useful for short amplicons, which can be fully sequenced in one sequencing read (currently ca 400 bp. With 454); exons 2 and 3 of MHC genes are examples of such amplicons.

Software performs three major tasks:

1. For a given amplicon type, as defined by amplification primers, the target sequences are extracted from all reads containing the complete tag sequence and imported to SQLite database together with corresponding tags (Fig. 1). The tag length is user-defined and either one or both amplification primers may contain tags. There is also an option of relying only on the sequence of a single primer for amplicon identification and extracting the fixed number of bases following the primer (Fig. 1B). Additional information may be imported to the database at this stage: names and sequences of already known alleles and names of amplicons (individuals) defined by particular tag sequences.

2. The user can generate from the imported data a table in tab-delimited text format, which contains all (or a subset of) sequence variants and the number of reads by which a given variant was represented in a given amplicon. This output may be analyzed further using a spreadsheet software, alternatively, for more specific tasks, the database may be queried with SQL.

3. For each amplicon (or a selected subset of amplicons), the program generates FASTA files containing all sequence variants (or a subset of them) ordered according to the number of reads in the amplicon. Such files are extremely helpful in genotyping, facilitating for example quick and easy identification of common artifacts – recombinants (chimeras) between sequences of true alleles which may be present in multiple copies in amplicons sequenced to a high coverage. Optionally, user can perform alignment of sequence variants in each file with MUSCLE.

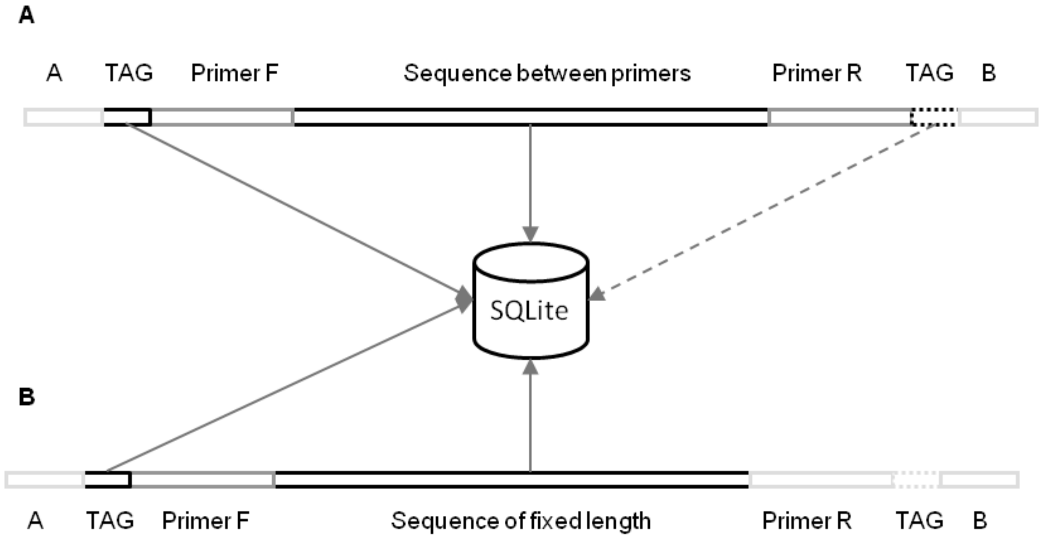

Figure 1. Schematic representation of reads and parts of the read extracted to the database: A) both primer sequences are used for identification of reads and the full sequence between primers is extracted; either one or both primers may be tagged, B) sequence of only one primer is identified and a fixed number of bases following the primer is extracted.

File->Project

"**New database**" - creates new database with a defined table structure where the sequences and additional provided data will be loaded; after clicking **“Connect”**, location and name of the new database can be selected; connection to the newly created database is established automatically.

Alternatively, an existing (created by this program) database can be used by selecting "**Open database**" and selecting an appropriate file using **“Connect”**. Again, connection to database is established after opening the file.

[optional] "**Alleles**" A FASTA file containing the names and sequences of already known alleles. Sequences can be unaligned or aligned; no ambiguity codes are allowed.

[optional] "**Tags**"

A tab delimited text file containing three columns:

1) tag sequence

If both primers were tagged the concatenated sequence of both tags (each in 5’->3’ direction should be provided tak to było z sekwencją?)

2) name of the multi-FASTA file containing the reads

3) amplicon designation

Example below:

AGTGTT	GEJFWGJ01.fna	1000ROS
CCGGAA	GEJFWGJ01.fna	101ROS1
CCGCTG	GEJFWGJ02.fna	101ROS2
AACCGA	GEJFWGJ02.fna	102ROS

Information in this file is used to automatically assign names to tag/file combinations.

If a given tag is not on the list, the amplicon designation in the database and ouput file have ”NO_TAG” value.

[optional] “**MUSCLE binary**” a path to Muscle executables – needed if generation of aligned multi-FASTA files with sequence variants present in amplicons is to be selected. File should be located on local disk and path cannot contain spaces or special characters.

After the database is connected and other options are selected, the “**Project**” window may be closed and the user may proceed to the “**Extractor**” menu.

Extractor->Import/Extracting

This window consists of two panels:

1. **“Import to database”** defines which sequences will be imported to the SQLite database. Once sequences are imported to the database, output files may be generated multiple times using the **“Generate output files”** panel (see below).

The files containing reads are loaded using the “**open**” button. Multiple files are selected with <Ctrl> or <Ctrl-Shift>.

“**Primer F**” **-** forward primer sequence in 5’->3’ direction. IUPAC nucleotide ambiguity codes as well as any combination of small and capital letters are allowed.

“**Primer R**” **-** reverse primer sequence in 5’->3’ direction. IUPAC nucleotide ambiguity codes as well as any combination of small and capital letters are allowed.

“**Fixed length**” option may be selected, but it works only in case of a single tagged primer. The fixed number of bases following the F primer is extracted from each read containing the complete tag and the F primer sequence.

“**Cutoff**” option may be selected, but it too works only in case of a single tagged primer. The program assumes that, if only one primer is tagged it is always the F primer. Cutoff value is the number of bases from the 3’ end of the R primer which must match for the sequence to be selected. This option may be useful to maximize the yield of sequences when the length of an amplicon is limiting or when, simply, a part of the R primer sequence is sufficient to identify the primer without ambiguity. "**TAG length**" the length of the tag (bp); the tag is the x (where x is tag length) bases preceding the 5' end of primer F (if "two-sided tags" option is NOT selected), or x bases preceding the 5' end of primer F and x bases preceding the 5' end of primer R.

"**Forward**" tells the program to search for reads in forward orientation

"**Reverse**" tells the program to search for reads in reverse complement orientation

Normally, if bidirectional sequencing was performed both Forward and Reverse should be selected. If the user knows that sequencing was performed only in one direction, an appropriate option may be selected and should speed up searching, although clicking both options will produce identical results.

"**Extract only starting with seq**" only partial sequences starting with required sequence motif will be extracted; the motif may be anywhere within the sequence (inside the limits set by primers). Useful in certain situations. Ambiguity codes, capital and small letters supported.

Clicking **"Start"** starts extraction of sequences to the database

The sequence between primers (but without primer sequences) is imported into the database if all the following conditions are fulfilled:

1) Contains the complete F primer and either complete R primer sequence or the number of bases from the 3' end specified in the cutoff field. The cutoff works only in case a single primer is tagged. If "**2-sided TAGs**" option is selected both primers must be complete

2) Does not contain any Ns (or other ambiguous characters) between primers

3) Does contain a complete tag sequence (or complete sequences of both tags if "**2-sided TAGs**" option is selected). The tag sequence cannot contain any ambiguities (Ns)

4) Does contain the desired sequence motif, if "Extract only starting with seq" option is selected. The reported sequence is a substring of the original one, starting with (and including) the desired motif.

2. **"Generate output files"** this panels generates output files from the current database in the spreadsheet form

"Do not output erroneous tags" suppresses output of the amplicons with tags which are not expected in the output on the basis of the "Tags" file. These are erroneous tags generated by PCR/sequencing errors.

"Output only variants with the number of reads >" suppresses output of sequence variants with the number of reads in the database smaller or equal to the given number

Both these options may be useful if the dataset is very large, as it substantially reduces the size of the output table.

"**Save**" – selects the location and the generic name of the output file. The output (tab-delimited text) generated in the output file with suffix "1" has the structure:

1st row: tag sequences; if “two-sided tags” option is selected sequences of both tags will be concatenated in 5’->3’ direction.

2nd row: name of the file with reads.

3rd row: individual/amplicon name if a file with names was provided, otherwise will be empty.

4th and following rows have identical structure: in the first column there is a sequence of a unique variant and following columns contain counts of the variant for a given tag/file combination. Variants which were present in the FASTA file with known alleles will be assigned respective names, other variants will be given generic names in the format seq00000.

The output may be easily processed further with Excel or other spreadsheet application but note that the matrix may be very large; in such cases suppressing output of singletons and/or erroneous tags will result in considerably smaller matrix.

Example output below:

tag	AACCGA	AACGCG	AAGACA	AGAATT	AGATAT
file	769_819_01.fna	769_819_01.fna	769_819_01.fna	769_819_01.fna	769_819_01.fna
ind	1000ROS	1001ROS	1002ROS	1003ROS	1004ROS
seq1	0	0	7	0	0
seq2	0	0	2	0	0
seq3	0	0	6	0	0
seq4	0	0	3	1	0
seq5	0	0	1	0	0
seq6	0	0	2	0	0

The output file generated with suffix “report” contains simple statistics concerning sequences in output file.

The output file with suffix “2” will probably be of no general interest for most users.

Extractor->Export/Alignment

This window also consists of two panels:

1. „**export to FASTA**” - for each amplicon (or a subset of amplicons selected with options described below) a FASTA file containing all sequence variants (or a subset of variants selected with options below) is created and named accordingly. Variants are sorted according to the number of reads present in a given amplicon, each variant is named with the allele name (if its sequence is identical to any of the provided alleles) or a unique sequence identifier in the format seq00000, sequence length and sequence count in a given amplicon.

Options for naming FASTA files:

[optional] **“prefix”** – all FASTA files may start with the user defined string.

“file name options” – file name may contain up to four fields, these may be: i) amplicon name as defined in the Tags file (**“amplicon”**), ii) tag sequence (**“tag”**), iii) name of the multi-FASTA file from which a given amplicon was extracted (**“file”**), iv) coverage of the amplicon (total number of reads obtained for the amplicon) (**“coverage”**). These field may be present in any combination and any order.

Options for exporting sequences

“**min. length**” – minimum length of sequences exported to FASTA files.

“**max. length**” – maximum length of sequences exported to FASTA files.

“**min. coverage**” – FASTA files will be generated only for amplicons (as defined by the tag sequence and multi-FASTA file name) with at least that number of reads.

“**muscle options**” - – user can aptionally select additional alignment options .

- anchors : use anchor optimization in tree dependent refinement iterations.
- brenner: use Steven Brenner's method for computing the root alignment.
- dimer: use dimer approximation for the SP score (faster, slightly less accurate).
- diags: use diagonal optimizations. Faster, especially for closely related sequences, but may be less accurate.

More details at http://www.drive5.com/muscle/muscle_userguide3.8.html

“**min. variant count**” – only variants with at least that number of reads in a given amplicon will be exported to FASTA files.

2. **“alignment”**

“**alignment**” perform alignment using MUSCLE – aligned FASTA file has “_align” added to the file name.

“**sort results in original order**” – the current version of MUSCLE (3.8), group sequences by similarity, this option restores the original order in the aligned file; sorted alignment has “_ordered” added to the file name.

After selecting the **“target folder”** and clicking “Start”, FASTA files will be generated, and if appropriate options were selected, aligned.