MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) **11**, 739–742 doi: 10.1111/j.1755-0998.2011.02997.x

jMHC: software assistant for multilocus genotyping of gene families using next-generation amplicon sequencing

MICHAŁ T. STUGLIK, JACEK RADWAN and WIESŁAW BABIK
Institute of Environmental Sciences, Jagiellonian University, ul. Gronostajowa 7, Krakow 30-387, Poland

Abstract

Genotyping of multilocus gene families, such as the major histocompatibility complex (MHC), may be challenging because of problems with assigning alleles to loci and copy number variation among individuals. Simultaneous amplification and genotyping of multiple loci may be necessary, and in such cases, next-generation deep amplicon sequencing offers a great promise as a genotyping method of choice. Here, we describe jMHC, a computer program developed for analysing and assisting in the visualization of deep amplicon sequencing data. Software operates on FASTA files; therefore, output from any sequencing technology may be used. jMHC was designed specifically for MHC studies but it may be useful for analysing amplicons derived from other multigene families or for genotyping other polymorphic systems. The program is written in Java with user-friendly graphical interface (GUI) and can be run on Microsoft Windows, Linux OS and Mac OS.

Keywords: 454, bioinformatics, deep amplicon sequencing, genotyping, MHC, next generation sequencing

Received 26 November 2010; revision received 6 January 2011; accepted 13 January 2011

Background

Genotyping multilocus gene families such as the major histocompatibility complex may be a formidable task. This is because individuals commonly differ in the number of loci and interlocus recombination may blur orthologous relationships among alleles within loci (Kumanovics *et al.* 2003; Ellis *et al.* 2005; Westerdahl 2007). In such cases, simultaneous amplification and genotyping of multiple loci may be necessary. Difficulties with assigning alleles to loci, variation in the number of alleles per individual and the presence of both divergent and highly similar alleles may complicate matters further (reviewed in Babik 2010). Therefore, next-generation deep amplicon sequencing has been widely regarded as a technology offering a great promise for genotyping major histocompatibility complex (MHC) systems of any complexity (Babik *et al.* 2009; Galan *et al.* 2010; Kloch *et al.* 2010). For the purpose of multilocus genotyping, deep amplicon sequencing is equivalent to cloning amplicons in a cell-free system and sequencing a high number of clones (Babik 2010). Of course, this technology generates artefacts, during both PCR (substitutions and recombinants—chimeras) and sequencing (substitutions and indels), so there remains a challenge of distinguishing true alleles from artefacts. Although approaches that minimize the frequency of PCR artefacts exist (Lenz & Becker 2008; Babik 2010), complete elimination of artefacts is unlikely and these approaches may be cumbersome and hard to implement in large-scale studies for logistic reasons. Thus, screening for artefacts and their elimination must be performed a posteriori, as a part of the analysis of the results of a sequencing run. Simple clustering methods used in biodiversity studies (Roesch *et al.* 2007; Christen 2008) are of limited use here, because true alleles may be extremely similar to each other or very divergent. However, artefacts should be relatively rare compared with true sequences, and thus, not only sequence divergence but also abundance of sequence variants in individual amplicons should be taken into account to perform reliable genotyping. We aimed at developing a universally applicable software tool jMHC that would assist in genotyping through extraction and tabulation of variants, reporting their frequencies, as well as producing alignments of sequence variants present in individual amplicons. jMHC was developed for analysing and assisting in the visualization of 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. jMHC was designed specifically for MHC studies but we believe it may be useful for analysing amplicons derived from other multigene families

Correspondence: Michał T. Stuglik, fax: (012) 664 69 12;
E-mail : michal.stuglik@uj.edu.pl

or for genotyping other polymorphic systems. The software is particularly useful for short amplicons, which can be fully sequenced in one sequencing read (currently ca 400 bp for 454 Titanium technology); exons 2 and 3 of MHC genes are examples of such amplicons. Recently, another database-oriented tool for performing similar tasks has appeared (Meglécz *et al.* 2010). SESAME is a web-based application requiring a multistep installation and thus appears better suited to a multiuser environment. SESAME provides extensive graphical interface for data visualization and assistance in genotyping. As both jMHC and SESAME possess unique features, both programs should be useful for analysing amplicons from multigene families and genotyping highly polymorphic systems such as MHC.

Software usage and description

A single sequencing run will contain hundreds of thousands to millions of reads and may consist of a mix of PCR products amplified with different primer sets (e.g. different loci or the same loci from different organisms) obtained from different individuals. To identify individuals, sequence tags of various length are typically used (Fig. 1) (e.g. Galan *et al.* 2010; Kloch *et al.* 2010). Both primers may be tagged, which enables the analysis of a large number of samples with a reasonable number of tagged primers.

The application performs three major tasks (Fig. 2). First, for a given amplicon type, as defined by amplification primers, the target sequences are extracted

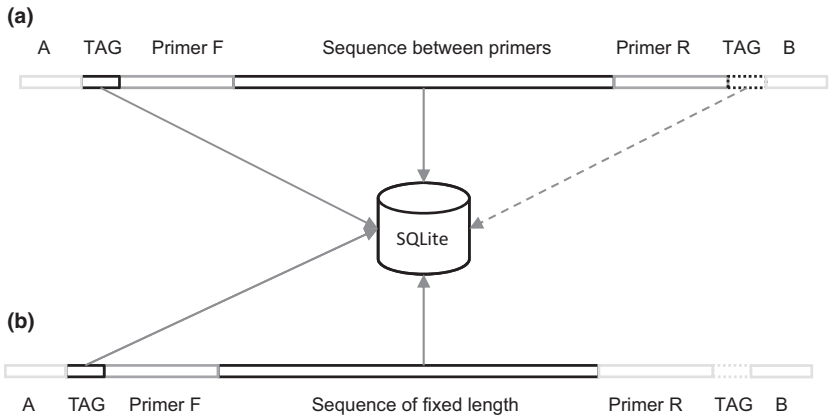

Fig. 1 Schematic representation of reads and parts of the read extracted to the database: (a) both primer sequences are used for the identification of reads and the full sequence between primers is extracted; either one or both primers may be tagged and (b) sequence of only one primer is identified and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

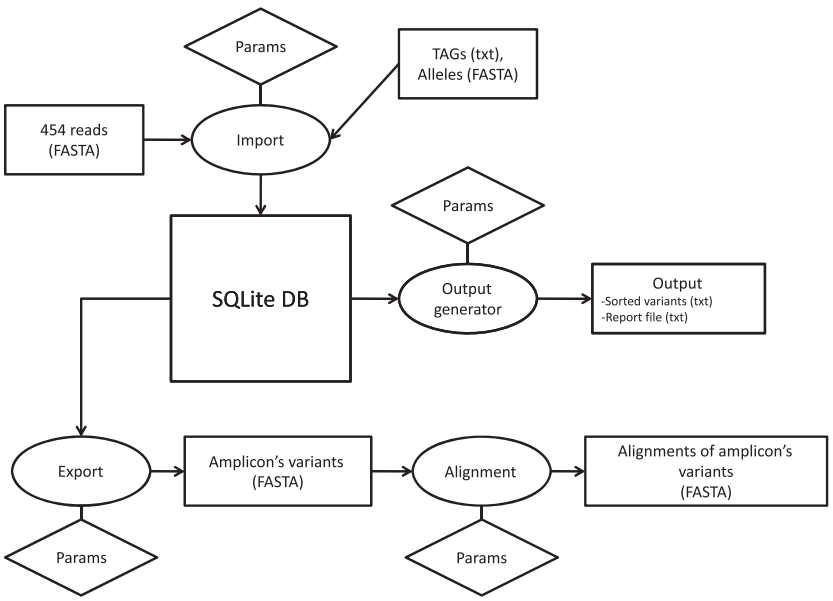

Fig. 2 jMHC workflow diagram illustrating major tasks, parameters as well as input and output data types.

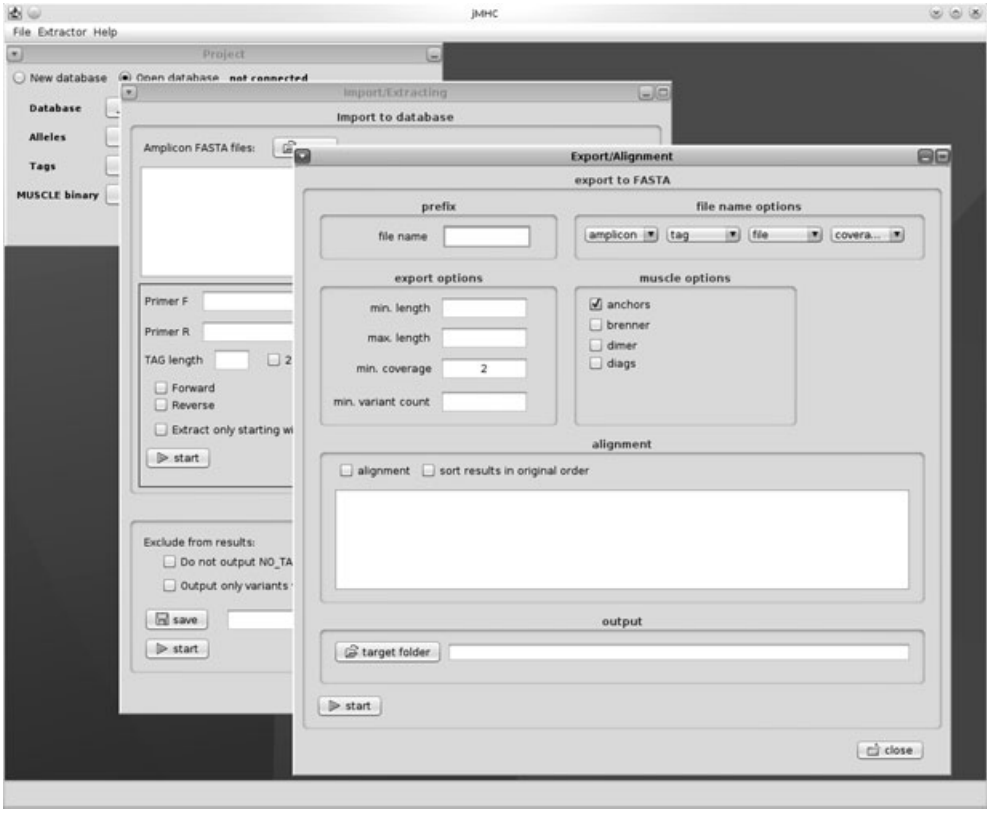

Fig. 3 jMHC graphical user interface.

from all reads containing the complete tag sequence and imported to SQLite (http://sqlite.org/) database together with corresponding tags (Fig. 1). No errors are allowed in primer and tag sequences. 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.

Second, 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 analysed further using a spreadsheet software; alternatively, for more specific tasks, the database may be queried with SQL.

Third, 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 artefacts—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 (Edgar 2004), which may be useful in further data processing/automation. In all steps, output can be modified on specific user demand.

jMHC is written in Java programming language using the Swing Graphical User Interface (GUI) (Fig. 3) toolkit, BioJava module (Holland *et al.* 2008) as bioinformatics core library and SQLite database for storing data. The application runs on Microsoft Windows, Linux OS and Mac OS and requires recent Java Runtime Environment (http://www.java.com/en/download/index.jsp).

Current version: binaries, source code, sample data and user manual are available at http://code.google.com/p/jmhc/. Program is published under the terms of the GNU General Public License v3.

Acknowledgements

This work was supported by the Foundation for Polish Science, professor subsidy 9/2008 to JR and the Jagiellonian University (DS/WBINOZ/INOŚ/762/10).

References

Babik W (2010) Methods for MHC genotyping in non-model vertebrates. *Molecular Ecology Resources*, **10**, 237–251.

Babik W, Taberlet P, Ejsmond MJ, Radwan J (2009) New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. *Molecular Ecology Resources*, **9**, 713–719.

Christen R (2008) Global sequencing: a review of current molecular data and new methods available to assess microbial diversity. *Microbes and Environments*, **23**, 253–268.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

Ellis SA, Morrison WI, MacHugh ND *et al.* (2005) Serological and molecular diversity in the cattle MHC class I region. *Immunogenetics*, **57**, 601–606.

Galan M, Guivier E, Caraux G, Charbonnel N, Cosson JF (2010) A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics*, **11**, 296.

Holland RCG, Down TA, Pocock M *et al.* (2008) BioJava: an open-source framework for bioinformatics. *Bioinformatics*, **24**, 2096–2097.

Kloch A, Babik W, Bajer A, Siński E, Radwan J (2010) Effects of an MHC-DRB genotype and allele number on the load of gut parasites in the bank vole *Myodes glareolus*. *Molecular Ecology*, **19**, 255–265.

Kumanovics A, Takada T, Lindahl KF (2003) Genomic organization of the mammalian MHC. *Annual Review of Immunology*, **21**, 629–657.

Lenz TL, Becker S (2008) Simple approach to reduce PCR artefact formation leads to reliable genotyping of MHC and other highly polymorphic loci - Implications for evolutionary analysis. *Gene*, **427**, 117–123.

Meglécz E, Piry S, Desmarais E *et al.* (2010) SESAME (SEquence Sorter & AMplicon Explorer): genotyping based on high-throughput multiplex amplicon sequencing. *Bioinformatics*, doi: 10.1093/bioinformatics/btq641; November 16, 2010.

Roesch LFW, Fulthorpe RR, Riva A *et al.* (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME Journal*, **1**, 283–290.

Westerdahl H (2007) Passerine MHC; genetic variation and disease resistance in the wild. *Journal of Ornithology*, **148**, S469–S477.