BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 2 2011, pages 277–278
doi:10.1093/bioinformatics/btq641

Sequence analysis

Advance Access publication November 16, 2010

SESAME (SEquence Sorter & AMplicon Explorer): genotyping based on high-throughput multiplex amplicon sequencing

Emese Meglécz[1,*,†], Sylvain Piry[2,†], Erick Desmarais[3], Maxime Galan[2], André Gilles[1], Emmanuel Guivier[2], Nicolas Pech[1] and Jean-François Martin[2]

[1]Aix-Marseille Université, CNRS, IRD, UMR 6116 – IMEP, Equipe Evolution, Génome et Environnement, 13331 Marseille Cedex 3, [2]UMR CBGP (INRA/IRD/Cirad/Montpellier SupAgro), Campus international de Baillarguet, CS 30016, F-34988 Montferrier-sur-Lez cedex and [3]Université Montpellier 2, CNRS, UMR 5554 – Institut des Sciences de l'Evolution – Montpellier – CC 065, 34095 Montpellier Cedex 05, France

Associate Editor: Martin Bishop

ABSTRACT

Summary: Characterizing genetic diversity through genotyping short amplicons is central to evolutionary biology. Next-generation sequencing (NGS) technologies changed the scale at which these type of data are acquired. SESAME is a web application package that assists genotyping of multiplexed individuals for several markers based on NGS amplicon sequencing. It automatically assigns reads to loci and individuals, corrects reads if standard samples are available and provides an intuitive graphical user interface (GUI) for allele validation based on the sequences and associated decision-making tools. The aim of SESAME is to help allele identification among a large number of sequences.

Availability: SESAME and its documentation are freely available under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported Licence for Windows and Linux from http://www1.montpellier.inra.fr/CBGP/NGS/ or http://tinyurl.com/ngs-sesame.

Contact: emese.meglecz@univ-provence.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 6, 2010; revised on October 20, 2010; accepted on November 8, 2010

1 INTRODUCTION

Since the breakthrough of PCR in the 1980s, researchers have been using various DNA markers and approaches to characterize the genetic diversity such as microsatellites, single strand conformation polymorphism (SSCP) and amplified fragment length polymorphism (AFLP). However, direct sequencing of DNA loci from a large number of individuals was costly and labor intensive. This was even further an issue for nuclear markers where cloning must precede Sanger sequencing to disentangle alleles. Technical advances involved in next-generation sequencing (NGS) technologies, including clonal amplification (454®, Roche) or single molecule sequencing (SMRT®, Pacific Bioscience), open new horizons where gigabases of DNA sequences can be obtained easily. One can take advantage from these innovations to directly and massively sequence amplicons. However, this approach necessitates tagging individual amplicons and sorting out real alleles from reads with sequencing errors.

We have created SESAME as a user-friendly web application package for analyzing amplicon sequences obtained through NGS, where the amplicon is completely covered by a read. It includes automatic sequence assignation to multiple markers and individuals via oligonucleotide tags. It provides an intuitive point-and-click interface to validate sequences as allele from individual-based variant alignments. The results are exported as genotypes or sequences of alleles. To our knowledge, no similar software has been published that accomplishes all these tasks.

2 IMPLEMENTATION

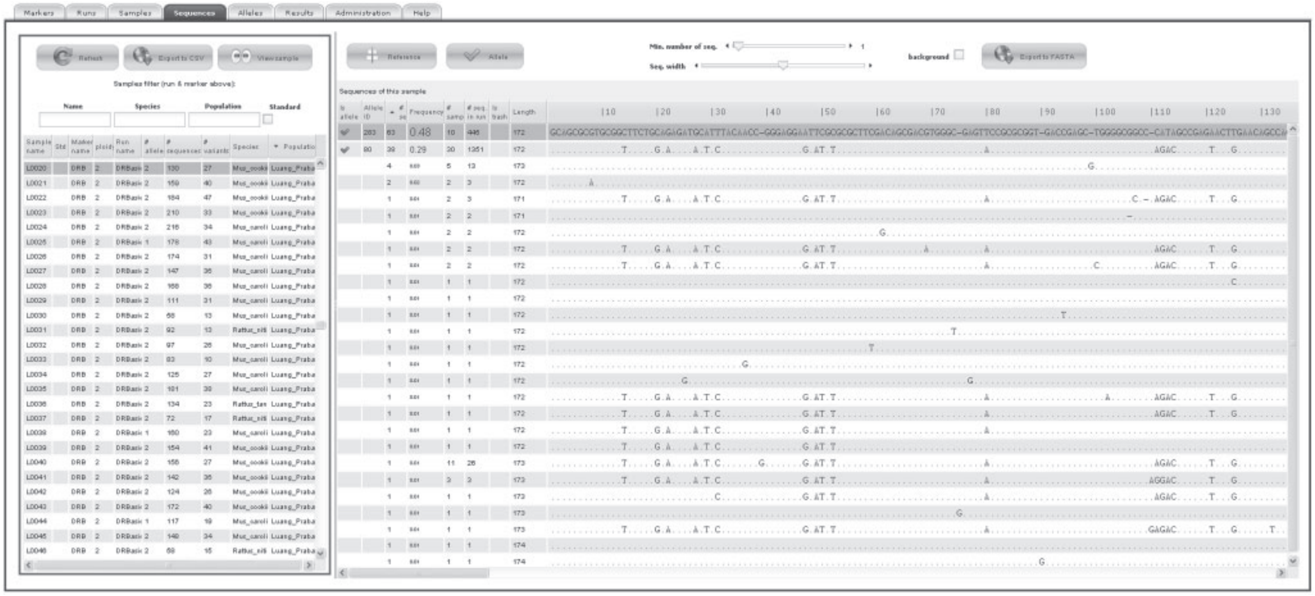
SESAME is a database oriented web application with an easy to use graphical user interface (GUI). The database can be accessed simultaneously by several users sharing their data within a project. An assistant guides the user for input data upload and through the sequence analysis steps (Supplementary Materials). All data are stored in a relational database. The user interacts with the database through an intuitive interface during the allele validation procedure.

2.1 Input data and automatic sequence analysis

Amplicons are individually identified by oligonucleotide tags linked as described in Galan *et al.* (2010). Although only one tag is strictly needed for assignation, using both forward and reverse tags increases the number of possible combinations. Furthermore, SESAME allows simultaneous analysis of several loci multiplexed in a single run. Sample (individual-locus combination) information is read from a comma-separated values (CSVs) file providing sample name, marker name, expected ploidy level, primer and tag sequences, status (sample or standard), population name and species (step 1). DNA sequences are read from one single FASTA file (step 2). Furthermore, users are asked to provide the list of DNA markers with corresponding reference sequences that will be used for assignation of the reads to loci and for trimming them from primers and tags. More precisely, DNA sequences are BLASTed (with user-defined *E*-value) against the reference sequences to assign each read to a locus and determine read orientation (step 3) then BLASTed against the list of tag sequences concatenated with primer sequences for sample assignation (step 4). Only reads with perfect match

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.



Samples filter (run & marker above):

Name □ Species □ Population □ Standard □

Sample name	Std	Marker name	ploidy	Run name	# allele	# sequences	# variants	Species	▼ Population
L0020		DRB	2	DRBasi	2	130	27	Mus_cooki	Luang_Praba
L0021		DRB	2	DRBasi	2	159	40	Mus_cooki	Luang_Praba
L0022		DRB	2	DRBasi	2	184	47	Mus_cooki	Luang_Praba
L0023		DRB	2	DRBasi	2	210	33	Mus_cooki	Luang_Praba
L0024		DRB	2	DRBasi	2	216	34	Mus_caroli	Luang_Praba
L0025		DRB	2	DRBasi	1	178	43	Mus_caroli	Luang_Praba
L0026		DRB	2	DRBasi	2	174	31	Mus_caroli	Luang_Praba
L0027		DRB	2	DRBasi	2	147	36	Mus_caroli	Luang_Praba
L0028		DRB	2	DRBasi	2	168	36	Mus_caroli	Luang_Praba
L0029		DRB	2	DRBasi	2	111	31	Mus_caroli	Luang_Praba
L0030		DRB	2	DRBasi	2	68	13	Mus_caroli	Luang_Praba
L0031		DRB	2	DRBasi	2	92	13	Rattus_niti	Luang_Praba
L0032		DRB	2	DRBasi	2	97	26	Mus_caroli	Luang_Praba
L0033		DRB	2	DRBasi	2	83	10	Mus_caroli	Luang_Praba
L0034		DRB	2	DRBasi	2	125	27	Mus_caroli	Luang_Praba
L0035		DRB	2	DRBasi	2	181	30	Mus_caroli	Luang_Praba
L0036		DRB	2	DRBasi	2	134	23	Rattus_tan	Luang_Praba
L0037		DRB	2	DRBasi	2	72	17	Rattus_niti	Luang_Praba
L0038		DRB	2	DRBasi	1	160	23	Mus_caroli	Luang_Praba
L0039		DRB	2	DRBasi	2	154	41	Mus_cooki	Luang_Praba
L0040		DRB	2	DRBasi	2	156	27	Mus_cooki	Luang_Praba
L0041		DRB	2	DRBasi	2	142	36	Mus_cooki	Luang_Praba
L0042		DRB	2	DRBasi	2	124	26	Mus_cooki	Luang_Praba
L0043		DRB	2	DRBasi	2	172	40	Mus_cooki	Luang_Praba
L0044		DRB	2	DRBasi	1	117	19	Mus_caroli	Luang_Praba
L0045		DRB	2	DRBasi	2	140	34	Mus_caroli	Luang_Praba
L0046		DRB	2	DRBasi	2	69	15	Rattus_niti	Luang_Praba

Min. number of seq. 1

Seq. width

background □

Sequences of this sample

Is allele	Allele ID	▲ # seq	Frequency	# samp	# seq. in run	Is trash	Length
✓	263	63	0.48	10	446		172
✓	80	38	0.29	30	1351		172
		4	0.03	5	13		173
		2	0.02	2	3		172
		1	0.01	2	3		171
		1	0.01	2	2		171
		1	0.01	2	2		172
		1	0.01	2	2		172
		1	0.01	2	2		172
		1	0.01	1	1		172
		1	0.01	1	1		172
		1	0.01	1	1		172
		1	0.01	1	1		172
		1	0.01	1	1		172
		1	0.01	1	1		172
		1	0.01	1	1		172
		1	0.01	1	1		172
		1	0.01	1	1		172
		1	0.01	1	1		172
		1	0.01	1	1		172
		1	0.01	11	26		173
		1	0.01	3	3		173
		1	0.01	1	1		173
		1	0.01	1	1		173
		1	0.01	1	1		173
		1	0.01	1	1		174
		1	0.01	1	1		174

```
|10       |20       |30       |40       |50       |60       |70       |80       |90       |100      |110      |120      |130
GCAGCGCGTGCGGCTTCTGCAGAGATGCATTTACAACC–GGGAGGAATTCGCGCGCTTCGACAGCGACGTGGGC–GAGTTCCGCGCGGT–GACCGAGC–TGGGGCGGCC–CATAGCCGAGAACTTGAACAGCCA
```

Fig. 1. Screenshot of the allele validation step. For illustration, the selected sample L0020 (left frame) is a diploid species *Mus cookii* from the population Luang Prabang and has 130 sequences representing 27 variants of *DRB* gene with two validated alleles. These two validated alleles are in high frequency in the sample (right frame), respectively, 0.48 (63 sequences) and 0.29 (38 sequences), compared with other variants (frequency < 0.03). Analysis of the sequence alignment shows that these variants in low frequency display errors (indels, substitutions) or are chimeric recombination of the two validated alleles.

of both tags are assigned to sample, and thus all assigned reads cover the whole amplicon. All of the assigned reads are kept for further analyses. For each sample, all reads are aligned by MUSCLE (Edgar, 2004) or MAFFT (Katoh *et al.*, 2005) according to the administrator's choice. Tags and primers are trimmed off (step 5). An optional position-specific correction of reads (step 6) can be performed if standard samples are provided by the user (typically a single cloned allele amplified and sequenced along with the samples; for details see Supplementary Materials).

2.2 Assisted allele validation

Users visualize the alignments of reads for each individual amplicon with an intuitive GUI. For a selected locus, all samples appear together with descriptive statistics, such as the number of sequences and variants assigned to the sample or the number of user-validated alleles. The number of sequences is displayed in red if too low to insure high-quality genotyping given the expected ploidy level (see User guide for details). Identical reads of a sample are pooled into variants and their numbers and frequencies are provided. Selecting a sample displays the alignment of all corresponding variants. Each variant is characterized by its length, its total occurrence in the whole run and the number of samples in which it is found (Fig. 1). Filters can be applied to help genotyping (e.g. the minimum number of reads for a variant; see User guide for details). Users can validate a variant as an allele by a simple click. A unique allele identifier is attributed allowing comparisons of genetic diversity among multiple runs within a project. After allele validation, the genotype table or the FASTA file of alleles can be exported. Filtering options are available to display or export only selected subsets (e.g. run, populations, species).

Information on a test dataset is available in the Supplementary Materials.

3 DISCUSSION

Obtaining high number of sequences for amplicons from nuclear markers brings new challenges with regard to deciphering alleles from each other and distinguishing weakly amplified alleles from sequencing errors. With this respect, bioinformatics tools did not cope yet with the analysis of amplicons sequenced in this high-throughput environment. Moreover, user-friendly tools are needed for sorting and analyzing this type of data, as end users are interested in the biological information and do not necessarily have programming or sharp bioinformatics skills. SESAME fills this gap and makes routine sequencing as DNA barcoding of a large number of amplicons at hand. SESAME is resolutely oriented toward biologists with regard to user-friendly interfacing and easy installation. Furthermore, most functions to support analyses are unique and provide a strong and productive framework for high-throughput amplicon sequencing.

Conflict of Interest: none declared.

REFERENCES

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

Galan,M. *et al.* (2010) A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics*, **11**, 269.

Katoh,K. *et al.* (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.