Chapter 4

In Silico Fingerprinting (ISIF): A User-Friendly In Silico AFLP Program

Margot Paris and Laurence Després

Abstract

The Amplified fragment Length Polymorphism (AFLP) is one of the cost-effective and useful fingerprinting techniques to study non-model species. One crucial AFLP step in the AFLP procedure is the choice of restriction enzymes and selective bases providing good-quality AFLP profiles. Here, we present a user-friendly program (ISIF) that allows carrying out in silico AFLPs on species for which whole genome sequences are available. Carrying out in silico analyses as preliminary tests can help to optimize the experimental work by allowing a rapid screening of candidate restriction enzymes and the combinations of selective bases to be used. Furthermore, using in silico AFLPs is of great interest to limit homoplasy and amplification of repetitive elements to target genomic regions of interest or to optimize complex and costly high-throughput genomic experiments.

Key words: AFLP fingerprint, In silico genotyping, Whole genome data, Homoplasy, Restriction enzymes

1. Introduction

The Amplified fragment Length Polymorphism (AFLP (1)) is one of the cost-effective and useful fingerprinting techniques to study non-model species. AFLP is based on the selective polymerase chain reaction (PCR) amplification of subsets of genomic restriction fragments. Genomic DNA is digested in thousands of fragments using restriction enzymes, and a subset of fragments is amplified by PCR using primers with one to four selective bases, thereby reducing the number of fragments on the profile. Fragments are separated by their length using electrophoresis, and discrete peaks can be visualized on a typical AFLP profile. Each discrete peak position is scored and characterized as a dominant biallelic locus (coded 0/1) in a 50–500-bp range.

François Pompanon and Aurélie Bonin (eds.), *Data Production and Analysis in Population Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 888, DOI 10.1007/978-1-61779-870-2_4,

Recently, many authors have focused on improving the reliability and the accuracy of the AFLP technique, from the molecular steps to the data analysis. First, the AFLP protocol has to be carefully chosen depending on the study species; the initial AFLP protocol described for plants by Vos et al. (1) has been already successfully modified for the study of more challenging organisms like vertebrates (2) or insects (3). To control the quality of the AFLP procedure (contaminations, reliability of the method, or genotyping errors (4, 5)), negative controls and sample replicates are now included in most experiments (6–14). Then, several marker selection algorithms have been developed to optimize the challenging step of AFLP marker scoring by discarding biases due to subjective and unreliable personal procedures (15–17). Finally, statistical analyses appropriate for dominant markers have to be applied and many methods are now available to assess genetic diversity and population structure from AFLP data sets and to detect AFLP markers linked to selection (see ref. 18 for a review). More recently, a Bayesian method taking into account the distribution of band intensities in populations has been developed to allow the analyses of AFLPs as codominant markers (19). This method improves considerably the estimation of population structure and inbreeding coefficients from AFLP data sets and allows reaching a precision for these estimates very close to that obtained with SNPs (19).

Another crucial step in the AFLP procedure consists in the choice of restriction enzymes and/or selective bases that will generate AFLP profiles with an adequate number of peaks (typically between 20 and 100) with homogeneous length distribution and homogeneous fluorescence. Indeed, one of the major flaws of AFLPs is the presence of homoplasious peaks in the profiles that are due to co-migrating fragments of the same length (20–23). Here, we present the user-friendly program ISIF (22) that allows carrying out in silico AFLPs on species for which whole genome sequences are available. ISIF program is freely available at http://www-leca.ujf-grenoble.fr/logiciels.htm. It works in a Windows® environment and requires The Microsoft .NET Framework version 2.0 (freely available at http://www.microsoft.com/downloads). The program performs in silico AFLPs from any sequences by simulating the AFLP procedure step by step. First, it identifies all the restriction sites along the sequence and produces the pool of all possible restriction fragments. From those, it selects the final set of fragments that exhibit the selective bases used for the amplification. Finally, it determines the length of all the peaks of the AFLP profile, with the adaptor and primer lengths added when specified by the user. ISIF can provide the sequences of the virtual fragments for any known sequence, and for any restriction enzyme and selective bases combinations. Furthermore, it provides for each AFLP fragment the position along the genome. It, therefore,

allows quickly detecting homoplasious fragments. ISIF program is also very useful for a rapid screening of candidate restriction enzymes and of the combinations of selective bases to be used in order to optimize the experimental work. Indeed, testing many primer combinations before the genotyping can help:

- Selecting enzymes and selective bases providing AFLP profiles with the appropriate number of peaks
- Choosing primer combinations that provide AFLP profiles with homogeneous length distribution
- Choosing primer combinations with low homoplasy rate
- Detecting and discarding primer combinations amplifying repetitive elements in the genome, such as transposable elements
- Combining primer pairs in order to maximize the distribution of the AFLP fragments throughout the genome
- Targeting genomic regions of interest by using primer pairs generating AFLP fragments in these regions
- Optimizing complex and costly high-throughput genomic experiments, such as Diversity Arrays Technology (DArT (24, 25)), pyrosequencing of AFLPs (26, 27), or Restriction-site Associated DNA (RAD (28, 29))

2. Program Usage

2.1. Reference Sequences Import

The program performs in silico AFLPs from all sequences written in capital or small letters saved as plain text without line numbers and spaces, such as text files. Import reference sequence files using the "+" button in the middle of the user interface of ISIF program (Fig. 1). The names of the imported files are indicated in the white square on the left side of the user interface. For genomes divided in several chromosomes or contigs, one separate file per chromosome/contig should be imported. Use the "–" button to remove the selected files.

2.2. Restriction Sites' Specification

ISIF can perform in silico AFLPs with any classical restriction enzymes (i.e., enzymes that cleave only once, and inside the recognition site). Restriction sites of restriction enzymes have to be specified in the "Left Cut" and in the "Right Cut" columns, on the right side of the user interface. Each line corresponds to one restriction enzyme/site. "Left Cut" column corresponds to the part of the sequence in 5′ of the enzyme cleavage location, and the "Right Cut" column corresponds to the 3′ part of the sequence after the cleavage location. For example, for the EcoRI enzyme restriction site 5′G↓AATTC3′ (↓ indicates the cleavage location), "G" corresponds to the "Left Cut" and "AATTC" corresponds to the "Right Cut" (see Note 1).

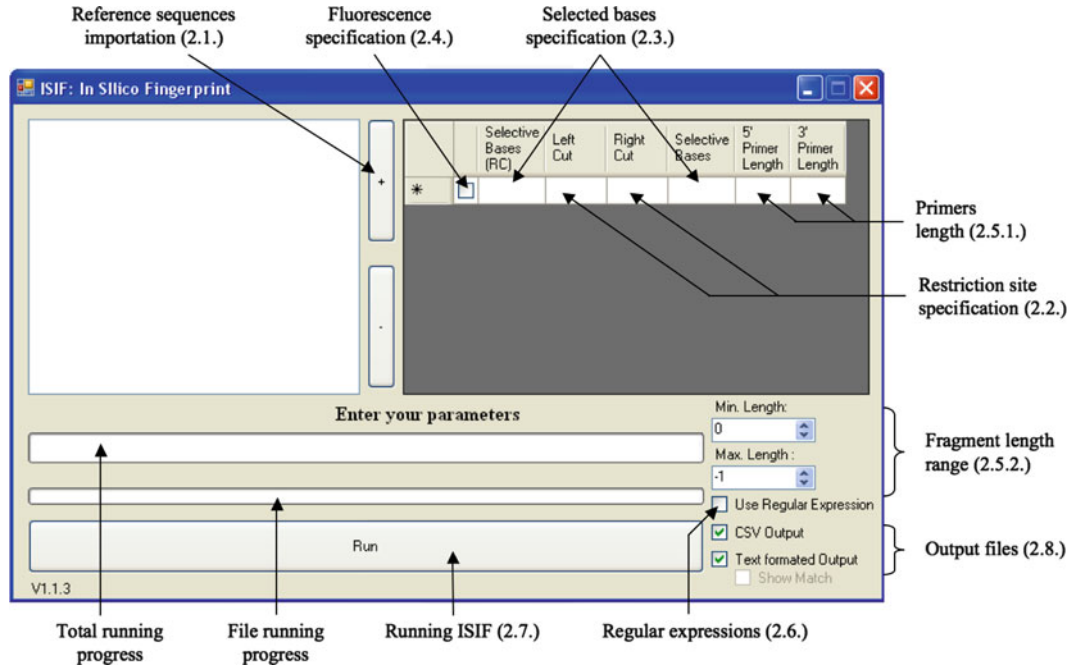

Fig. 1. ISIF user interface and parameters. For each parameter, the corresponding chapter number is indicated in *bracket*.

2.3. Selective Bases' Specification

Selective bases used in combination with a specific enzyme have to be specified in the same line of this enzyme restriction site, in the "Selective Bases" column (see Note 1). In silico AFLPs are performed on one 5′–3′ DNA strand of the reference genome (Fig. 2a, b). However, AFLP restriction sites are palindromic and both sides of the cutting sites are ligated with AFLP adaptors and potentially amplified. Therefore, to properly simulate AFLP procedure, the reverse complement sequences of the selected bases have to be specified in the "Selective Bases (RC)" column. They correspond to the selective bases sequences at the 3′ extremity of the AFLP restriction fragments of the reference genome (Fig. 2).

2.4. Fluorescent Enzymes' Specification

Two restriction enzymes are used in classical AFLP protocols and only one is favored during the amplification and the detection steps. This is achieved by using during the amplification step of the AFLP procedure a fluorescent primer in excess, which is associated with the enzyme restriction site that is favored. The favored enzyme has to be indicated in the enzyme line by checking the appropriate box (see Fig. 1). Only fragments cleaved at least in one extremity by this favored enzyme are presented in the ISIF output files.

2.5. Fragment Length

ISIF calculates and provides the fragment length of restriction fragments, from the cleavage site in 5′ to the cleavage site in 3′ (Fig. 2b). During the AFLP procedure, adaptors specific of each enzyme restriction site are ligated to the restriction fragments. After this ligation step, the fragments are amplified using primers,

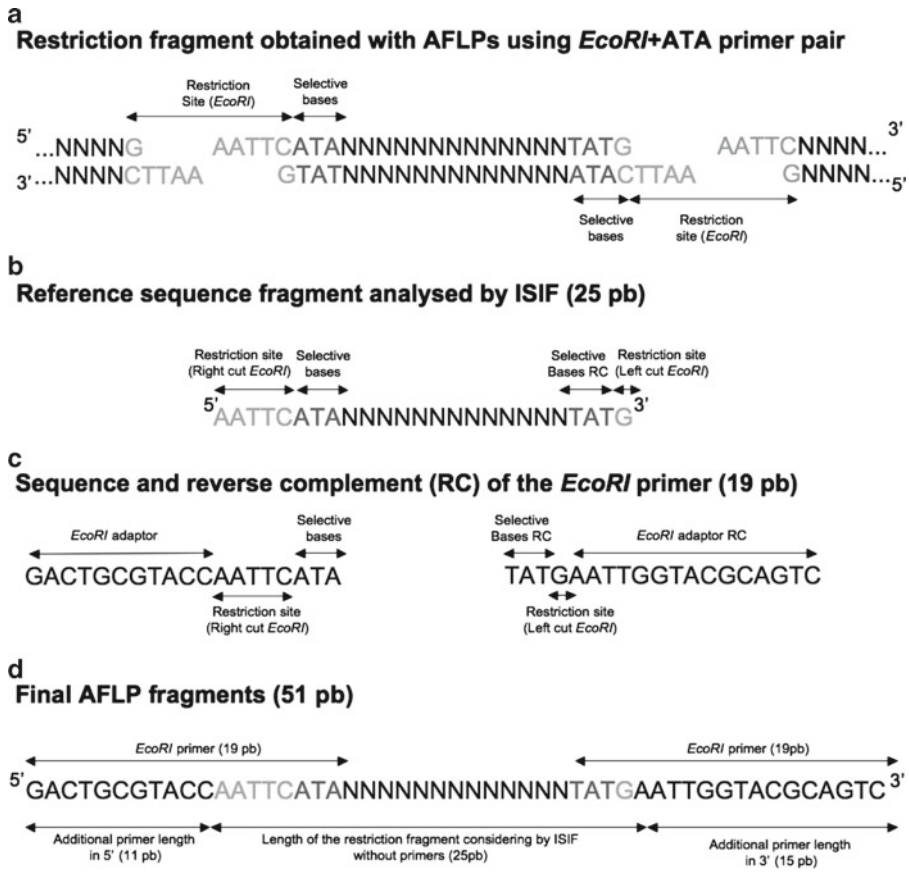

Fig. 2. Example of ISIF procedure and fragment length calculation for an AFLP restriction fragment amplified by the primer *EcoRI*+ATA. "RC" abbreviation corresponds to "Reverse Complement".

the sequence of which corresponds to the adaptor, plus the restriction sites and some supplementary selective bases (see Fig. 2c). ISIF can calculate and provide the length of final AFLP fragments by adding the primer length to restriction fragments (Fig. 2d, see Note 2).

2.5.1. Adding the Primer Length to Obtain the Final Length of the AFLP Fragments

The additional lengths of the primers have to be indicated in the "5′ primer length" and "3′ primer length" columns (Fig. 1). For each restriction enzyme, the additional lengths due to the primers have to be calculated as follows (Fig. 2d):

(a) 5′ primer length = total primer length − enzyme right cut length − selective bases length

(b) 3′ primer length = total primer length − enzyme left cut length − selective bases length

2.5.2. Selecting Fragment Length Range

By default, all fragments are presented in ISIF output files. Furthermore, as AFLP method focuses generally on fragments ranging between 50 pb and 500 pb, minimum and maximum

lengths can be added in ISIF parameters using the "Min. Length:" and the "Max. Length:" options (see Fig. 1). If no primer length is indicated (see Subheading 2.5.1), ISIF output files provide only restriction fragments in the selected length range. If primer lengths are indicated, ISIF output files provide only final AFLP fragments in the selected length range.

2.6. Use of Regular Expressions

ISIF does not take into account the IUPAC nucleotide code for unknown degenerated bases, such as N, R, or H. However, it is possible to specify these degenerated bases using regular expressions (see Note 3). ISIF allows the use of regular expressions by checking the box "Use regular expression." First, this option can be useful to perform in silico AFLPs on reference sequences containing genetic polymorphism (see Note 4). Indeed, heterozygosity is important when using dominant markers such as AFLPs because both homozygote and heterozygote status lead to AFLP peaks. Second, by using regular expressions, in silico AFLPs can also be performed with restriction enzymes containing degenerated bases (see Note 5). Such restriction enzymes can be used in classical AFLPs or in other restriction-based genotyping methods, such as DArT (24, 25).

2.7. Running ISIF

When all parameters are specified, press the "Run" button to start ISIF. The running progress is indicated for the total analyses, as well as for each of the reference file analyses (Fig. 1). For indication, performing in silico APFLs on a computer with an Intel® Pentium® D CPU 2.80 GHz and 2.00 Go of RAM, the running time is about 2 min for the *Arabidopsis thaliana* genome (genome size of 120 Mb) and 10 min for the *Aedes aegypti* genome (genome size of 1,310 Mb).

2.8. Program Output

ISIF provides two different output files for each of the reference sequences: a "CSV reference-file-name" and a "Text reference-file-name" file. Uncheck the box "csv Output" or "text formatted Output" in the user interface (Fig. 1) when an output file is not wanted. The Text-formatted file provides the following parameters for each restriction fragment (Fig. 4):

`SEQUENCE No.`

`Starting Cut:`	5′ restriction site (and corresponding selective bases)
`Ending Cut:`	3′ restriction site (and corresponding selective bases)
`Start Index:`	Starting position in the reference genome (in pb)
`End Index:`	Ending position in the reference genome (in pb)
`Length:`	Total fragment length (restriction fragment length)
`Fragment:`	Restriction fragment sequence

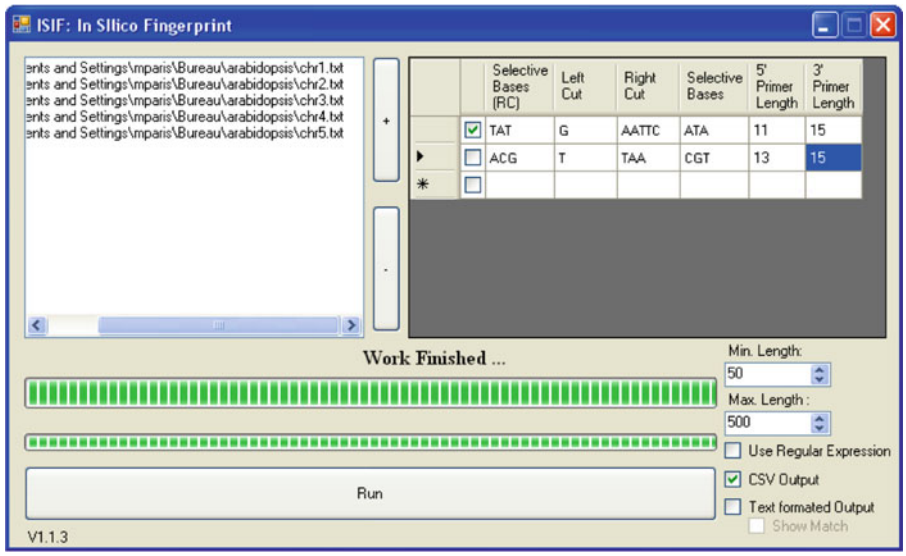

Fig. 3. Example of ISIF parameters to perform in silico AFLPs on the 5 *Arabidopsis thaliana* chromosomes using the primer pair *EcoRI*+ATA/*MseI*+CGT.

The CSV output file provides one line per restriction fragment using ";" as separator, and is compatible with spreadsheet editors or R program (30) for analyses (Fig. 5). The following parameters are presented in this order: starting cut, ending cut, start index, end index, total fragment length, restriction fragment length, and restriction fragment sequence.

3. Example

Figure 3 presents an example of program parameters for performing in silico AFLPs with the restriction enzyme and selective bases pair *EcoRI*+ATA/*MseI*+CGT. The first and the second lines correspond to the *EcoRI* and *MseI* enzyme parameters, respectively. The "fluorescent enzyme box" is checked only for *EcoRI*, and only fragments cleaved at least in one extremity by this enzyme are presented in the ISIF output files. The 19-pb primers GACTGCGTACCAATTCATA and GATGAGTCCTGAGTAACGT were used in this example to amplify *EcoRI* and *MseI* fragments, respectively. Therefore, the additional length of primers were 11 pb in 5′ and 15 pb in 3′ for the *EcoRI* enzyme (Fig. 2d) and 13 pb in 5′ and 15 pb in 3′ for the *MseI* enzyme. In this example, final AFLP fragments range from 50 to 500 pb. Figures 4 and 5 present, respectively, the "Text" and the "CSV" output files obtained for the chromosome 1 of *Arabidopsis thaliana*. Using this primer pair, two AFLP fragments of 118 and 76 pb were obtained for this chromosome (Figs. 4 and 5).

Fig. 4. "Text" output file provided by ISIF for in silico AFLPs on the chromosome 1 of *Arabidopsis thaliana* using the primer pair *EcoRI*+ATA/*MseI*+CGT.

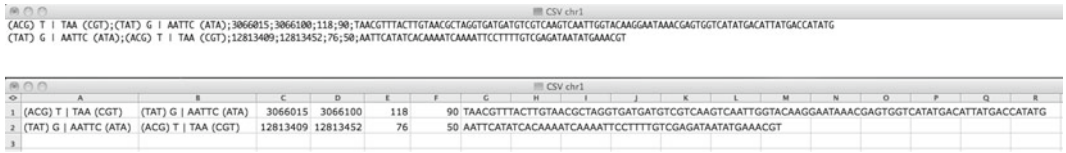

Fig. 5. "CSV" output file provided by ISIF for in silico AFLPs on the chromosome 1 of *Arabidopsis thaliana* using the primer pair *EcoRI*+ATA/*MseI*+CGT (screen shots from both text and spreadsheet editors).

4. Notes

1. ISIF distinguishes between capital and small letters. Therefore, restriction site sequences and selective bases have to be written in capital or small letters in order to match the reference genome format.
2. If no additional primer lengths are specified in ISIF parameters, the total fragment length corresponds to the restriction fragment length in output files.
3. Using regular expression, the special character "." denotes any single character and corresponds to the nucleotide code N in sequence data sets (IUPAC nucleotide code). The bracket expression "[ ]" matches a single character that is contained within the brackets. For example, the bracket expression "[AG]" denotes "A" or "G" and corresponds to the IUPAC nucleotide code R; and the bracket expression "[ACT]" denotes "A," "C," or "T" and corresponds to the IUPAC nucleotide code H.
4. To perform in silico AFLPs on reference sequences containing degenerate bases (IUPAC nucleotide code), these possible

bases have to be included in the restriction site and the selective bases' specifications using regular expressions. For example, the regular expression "[GSKBDVN]" represents all the possibilities to get a G in the reference sequence. On diploid species, avoiding the use of degenerated bases coding for more than two bases (B, D, H, V, and N) can help to limit the biases due to errors or uncertainties in sequences and to focus on polymorphisms. In this case, the specifications for the restriction enzyme EcoRI are "[GRSK]"for the "Left Cut" and "[ARWM][ARWM][TYWK][TYWK][CYSM]" for the "Right Cut" instead of "G" and "AATTC" (see Fig. 3).

5. In silico AFLPs can be performed with restriction enzymes containing degenerated bases using regular expressions. For example, the restriction enzyme Bsp1286I with the restriction site 5′GDGCH↓C3′ was used for the genotyping of *Aedes aegypti* mosquito strains (24, 25). Considering reference sequences containing no degenerated bases, "G[AGT]GC[ACT]" corresponds to the "Left Cut" and "C" corresponds to the "Right Cut" of the restriction site specifications for this enzyme.

Acknowledgments

This work was supported by a grant from the French Rhône-Alpes region (grant 501545401) and by the French National Research Agency (project ANR-08-CES-006-01 DIBBECO).

References

1. Vos P, Hogers R, Bleeker M et al (1995) AFLP – a new technique for DNA-fingerprinting. Nucleic Acids Res 23:4407–4414
2. Bonin A, Pompanon F, Taberlet P (2005) Use of amplified fragment length polymorphism (AFLP) markers in surveys of vertebrate diversity. Mol Evol 395:145–161
3. Paris M, Boyer S, Bonin A et al (2010) Genome scan in the mosquito *Aedes rusticus*: population structure and detection of positive selection after insecticide treatment. Mol Ecol 19: 325–337
4. Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. Nat Rev Genet 6:847–859
5. Bonin A, Bellemain E, Eidesen PB et al (2004) How to track and assess genotyping errors in population genetics studies. Mol Ecol 13: 3261–3273
6. Conord C, Lemperiere G, Taberlet P, Despres L (2006) Genetic structure of the forest pest *Hylobius abietis* on conifer plantations at different spatial scales in Europe. Heredity 97: 46–55
7. Fink S, Fischer MC, Excoffier L, Heckel G (2010) Genomic scans support repetitive continental colonization events during the rapid radiation of voles (Rodentia: Microtus): the utility of AFLPs versus mitochondrial and nuclear sequence markers. Syst Biol 59: 548–572
8. Karrenberg S, Favre A (2008) Genetic and ecological differentiation in the hybridizing campions *Silene dioica* and *S. latifolia*. Evolution 62:763–773
9. Meyer CL, Vitalis R, Saumitou-Laprade P, Castric V (2009) Genomic pattern of adaptive divergence in *Arabidopsis halleri*, a model

species for tolerance to heavy metal. Mol Ecol 18:2050–2062

10. Mraz P, Gaudeul M, Rioux D et al (2007) Genetic structure of *Hypochaeris uniflora* (Asteraceae) suggests vicariance in the Carpathians and rapid post-glacial colonization of the Alps from an eastern Alpine refugium. J Biogeogr 34:2100–2114
11. Nosil P, Egan SP, Funk DJ (2008) Heterogeneous genomic differentiation between walking-stick ecotypes: "isolation by adaptation" and multiple roles for divergent selection. Evolution 62: 316–336
12. Poncet BN, Herrmann D, Gugerli F et al (2010) Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*. Mol Ecol 19:2896–2907
13. Puscas M, Choler P, Tribsch A et al (2008) Post-glacial history of the dominant alpine sedge *Carex curvula* in the European Alpine System inferred from nuclear and chloroplast markers. Mol Ecol 17:2417–2429
14. Tilquin M, Paris M, Reynaud S et al (2008) Long lasting persistence of *Bacillus thuringiensis subsp israelensis* (*Bti*) in mosquito natural habitats. PLoS One 3:10
15. Arrigo N, Holderegger R, Alvarez N (2012) Automated scoring of AFLPs using RawGeno v 20, a free R CRAN library. In: Bonin A, Pompanon F (eds) Data Production and Analysis in Population Genomics, Methods in Mol Biol Series, Humana Press
16. Arrigo N, Tuszynski JW, Ehrich D et al (2009) Evaluating the impact of scoring parameters on the structure of intra-specific genetic variation using RawGeno, an R package for automating AFLP scoring. BMC Bioinformatics 10:33
17. Herrmann D, Poncet BN, Manel S et al (2010) Selection criteria for scoring amplified fragment length polymorphisms (AFLPs) positively affect the reliability of population genetic parameter estimates. Genome 53:302–310
18. Bonin A, Ehrich D, Manel S (2007) Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. Mol Ecol 16:3737–3758
19. Foll M, Fischer MC, Heckel G, Excoffier L (2010) Estimating population structure from AFLP amplification intensity. Mol Ecol 19:4638–4647
20. Caballero A, Quesada H (2010) Homoplasy and distribution of AFLP fragments: an analysis in silico of the genome of different species. Mol Biol Evol 27:1139–1151
21. Caballero A, Quesada H, Rolan-Alvarez E (2008) Impact of amplified fragment length polymorphism size homoplasy on the estimation of population genetic diversity and the detection of selective loci. Genetics 179:539–554
22. Paris M, Bonnes B, Ficetola GF et al (2010) Amplified fragment length homoplasy: in silico analysis for model and non-model species. BMC Genomics 11:287
23. Vekemans X, Beauwens T, Lemaire M, Roldan-Ruiz I (2002) Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. Mol Ecol 11: 139–151
24. Bonin A, Paris M, Despres L et al (2008) A MITE-based genotyping method to reveal hundreds of DNA polymorphisms in an animal genome after a few generations of artificial selection. BMC Genomics 9:459
25. Bonin A, Paris M, Tetreau G et al (2009) Candidate genes revealed by a genome scan for mosquito resistance to a bacterial insecticide: sequence and gene expression variations. BMC Genomics 10:551
26. Paris M, Meyer C, Blassiau C et al (2012) Two methods to easily obtain nucleotide sequences from AFLP loci of interest. In: Bonin A, Pompanon F (eds) Data Production and Analysis in Population Genomics. Methods in Mol Biol Series, Humana Press
27. Van Orsouw NJ, Hogers RJ, Janssen A et al (2007) Complexity reduction of polymorphic sequences (CRoPS (TM)): a novel approach for large-scale polymorphism discovery in complex genomes. PLoS One 2:11
28. Baird NA, Etter PD, Atwood TS et al (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:10
29. Hohenlohe PA, Bassham S, Etter PD et al (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet 6:2
30. R Development Core Team (2005) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org, ISBN 3-900051-900007-900050