BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 3 2010, pages 403–404
doi:10.1093/bioinformatics/btp670

Sequence analysis

QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects

Emese Meglécz[1,*], Caroline Costedoat[1], Vincent Dubut[1], André Gilles[1], Thibaut Malausa[2], Nicolas Pech[1] and Jean-François Martin[3]

[1]Aix-Marseille Université, CNRS, IRD, UMR 6116 – IMEP, Equipe Evolution, Génome et Environnement, Centre Saint-Charles, Case 36, 3 Place Victor Hugo, 13331 Marseille Cedex 3, [2]Institut National de la Recherche Agronomique, UMR 1301, INRA/UNSA/CNRS, Equipe BPI, 400, route des Chappes. BP 167. 06903 Sophia-Antipolis Cedex and [3]Montpellier SupAgro, INRA, CIRAD, IRD, Centre de Biologie et de Gestion des Populations, Campus International de Baillarguet, CS30016, 34988 Montferrier-sur-Lez, France

Received on August 6, 2009; revised on November 24, 2009; accepted on December 1, 2009

Advance Access publication December 10, 2009

Associate Editor: Limsoon Wong

ABSTRACT

Summary: QDD is an open access program providing a user-friendly tool for microsatellite detection and primer design from large sets of DNA sequences. The program is designed to deal with all steps of treatment of raw sequences obtained from pyrosequencing of enriched DNA libraries, but it is also applicable to data obtained through other sequencing methods, using FASTA files as input. The following tasks are completed by QDD: tag sorting, adapter/vector removal, elimination of redundant sequences, detection of possible genomic multicopies (duplicated loci or transposable elements), stringent selection of target microsatellites and customizable primer design. It can treat up to one million sequences of a few hundred base pairs in the tag-sorting step, and up to 50 000 sequences in a single input file for the steps involving estimation of sequence similarity.

Availability: QDD is freely available under the GPL licence for Windows and Linux from the following web site: http://www.univ-provence.fr/gsite/Local/egee/dir/meglecz/QDD.html

Contact: emese.meglecz@univ-provence.fr

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

1 INTRODUCTION

Microsatellites are among the most informative and thus the most frequently used molecular markers in population biology. So far their isolation for non-model species has been dependent on time- and labour-consuming laboratory protocols providing typically a few hundred sequences. New advances in pyrosequencing now produce large amounts of sequences from any DNA sample at low cost and open new avenues for the setup of markers for species with no previous genomic information. Availability of thousands of sequences may help to optimize the setup of microsatellite markers by allowing the selection of microsatellites that are not compound or interrupted and likely follow a simple mutation model. Moreover, microsatellite amplification by polymerase chain reaction (PCR) can be seriously affected by microsatellite association with mobile element. The large number of sequences allows the detection of putative mobile elements by spotting sequence similarity groups and eliminating them. As a result, the proportion of primers allowing specific amplification at unique loci with a clear banding will be higher and will avoid time-consuming primer tests. The presence of null alleles is also a recurrent problem in microsatellite amplification. If several copies of the same loci are identified, and the sequencing was based on pooled DNA from several individuals, some of the null alleles can be detected and avoided if consensus sequence construction is stringent.

Several programs for microsatellite detection (Tandem Repeats Finder, Benson, 1999; SciRoKo, Kofler *et al.*, 2007) sequence clustering (BLASTclust, Phrap) and primer design (Primer3, Rozen and Skaletsky, 2000; FastPCR, Kalendar *et al.*, 2009) are currently available for users. A recent paper by Castoe *et al.* (2009) even provides Perl scripts for microsatellite detection, primer design and attempts to eliminate some of the sequence redundancies, but it lacks potentially necessary tag/adapter/vector/ cleaning, and their method to detect multiple copies is based on only perfect primer site matches. Hence, a complete analysis pipeline is needed for the routine treatment of thousands or millions of sequences coming from next generation sequencing.

QDD is designed to treat all bioinformatics steps from large quantities of raw sequences to the design of PCR primers for microsatellites amplification. We provide results on efficiency of the primers selected by QDD and a glossary of unusual terms in the Supplementary Materials.

2 IMPLEMENTATION

QDD is written in Perl and can be run as a standalone application on Windows or Linux systems. The Windows interface provides a user-friendly version. It is a collection of small modules that use the following freely available programs: ActivePerl (http://www.activestate.com/activeperl/), BLAST (ftp://ftp.ncbi.nih.gov/blast/executables/), CLUSTALw2 (Larkin *et al.*, 2007; ftp://ftp.ebi.ac.uk/pub/software/clustalw2/) and Primer3

*To whom correspondence should be addressed.

(Rozen and Skaletsky, 2000; http://primer3.sourceforge.net/). They should be installed in order to be able to use QDD.

QDD proceeds in three successive stages:

2.1 Sequence cleaning and microsatellite detection

The input files of QDD are the following fasta files (i) tag.fas is a file containing all user tags (optional), (ii) adapter.fas contains all user adapter/vector sequences when suitable (optional), (iii) one or several fasta files containing all raw sequence reads.

If sequences have been tagged, the program prepares one fasta file for each tag containing all sequences cleaned from the given tag, plus one file with all sequences for which no tag was detected. This step is optional.

Removal of adaptors or possible vector contaminations if needed.

Selection of sequences that are longer than a user-defined limit.

Selection of sequences that have perfect microsatellites with a minimum number of repeats defined by the user.

These sequences are written in a fasta file that will be the input for stage 2.

2.2 Sequence similarity detection

This stage is the most time consuming but essential part of the program. Removing redundancy is obviously necessary, but eliminating sequences that are part of a repetitive region of the genome is often neglected by researchers. This omission is due to the lack of genomic background information for most species. QDD detects sequence groups that potentially fall in this category. It is not designed to describe mobile elements, but to be a conservative method that eliminates many of the sequences that might be problematic for microsatellite amplifications (e.g. involving null alleles).

This stage is composed of the following steps:

'All against all' BLAST of the input file with soft-masked microsatellites.

Concatenation of 100% identical sequences.

Elimination of potential minisatellites (more than one BLAST hit between two compared sequences).

If similarity was detected by BLAST with a user-defined limit, QDD calculates pair wise identity along flanking regions. In this way, variability in the microsatellite length is not taken into account in the percentage of pair wise identity.

Establishment of contigs if pair wise identity along the flanking regions is higher than a user-defined limit (e.g. 95%).

Reconstruction of consensus sequences for each contig, where ambiguous sites are replaced by 'Ns'.

'All against all' BLAST of file containing all consensus sequences plus all original sequences that are not unique, not potential minisatellites and not included in the contigs.

Selection of contig consensus sequences that did not have hit to any other sequence in the previous BLAST.

Preparation of a file with selected consensus sequences and all original 'unique' sequences (either did not have a BLAST hit or only to 100% identical sequences). This file will be the input file for stage 3.

2.3 Microsatellite selection and primer design

The mutation model of compound, interrupted microsatellites or short repetitions ('nanosatellites') in the flanking region is complicated. Therefore, it is preferable to use perfect microsatellites with nanosatellite free flanking regions. However, since nanosatellites are abundant, most of the selected sequence regions for primer design are short. This can be problematic in multiplexing. Therefore, QDD automatically runs Primer3 to design primers several times, each time for a different PCR product size range. Stage 3 of QDD accomplishes the following steps:

Selection of sequences that contain a target microsatellite and a flanking region free from nanosatellites (users set the minimum number of repeats for the target microsatellites, the maximum number of allowed repeats for nanosatellites, the minimum length of flanking region and the minimum length of PCR product). An option is available to select compound or interrupted microsatellite as a target.

Preparation of an input file for Primer3 and a fasta file with all target microsatellites and nanosatellites printed in lower case.

Primer3 runs for all user-defined PCR product ranges. Primer3 parameters can be set by the user. Results of the different runs including all output parameters of Primer3 and information on the target microsatellites are summarized in a user-friendly table including sequence code.

ACKNOWLEDGEMENTS

We thank Frédéric Calendini for invaluable help in building the Windows interface.

Funding: University of Provence, Montpellier SupAgro, and a grant from the French Institut National de la Recherche Agronomique (INRA), AIP BioRessources EcoMicro.

Conflict of Interest: none declared.

REFERENCES

Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.

Castoe,T.A. *et al.* (2009) Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. *Mol. Ecol. Resources*, [Epub ahead of print, doi: 10.1111/j.1755-0998.2009.02750.x, Jul 30, 2009].

Kalendar,R. *et al.* (2009) Invited review: FastPCR software for PCR primer and probe design and repeat search. *Genes Genomes Genomics*, **3** (In press).

Kofler,R. *et al.* (2007) SciRoKo: a new tool for whole genome microsatellite search and investigation. *Bioinformatics*, **23**, 1683–1685.

Larkin,M.A. *et al.* (2007) ClustalW and ClustalX version 2. *Bioinformatics*, **23**, 2947–2948.

Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, 132, 365–386.