QDD version 3.1: A user-friendly computer program for microsatellite selection and primer design revisited: Experimental validation of variables determining genotyping success rate

Article *in* Molecular Ecology Resources · April 2014

DOI: 10.1111/1755-0998.12271

CITATIONS
94

READS
607

8 authors, including:

Emese Meglécz
Aix-Marseille Université
64 PUBLICATIONS **2,749** CITATIONS
SEE PROFILE

André Gilles
Aix-Marseille Université
255 PUBLICATIONS **4,148** CITATIONS
SEE PROFILE

Vincent Dubut
Aix-Marseille Université
67 PUBLICATIONS **1,310** CITATIONS
SEE PROFILE

Pascal Hingamp
Aix-Marseille Université
116 PUBLICATIONS **8,963** CITATIONS
SEE PROFILE

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2014) **14**, 1302–1313 doi: 10.1111/1755-0998.12271

QDD version 3.1: a user-friendly computer program for microsatellite selection and primer design revisited: experimental validation of variables determining genotyping success rate

EMESE MEGLÉCZ,* NICOLAS PECH,* ANDRÉ GILLES,* VINCENT DUBUT,* PASCAL HINGAMP,† AURÉLIE TRILLES,* RÉMI GRENIER* and JEAN-FRANÇOIS MARTIN‡
*Aix-Marseille Université, CNRS, IRD, Univ. Avignon, UMR 7263 – IMBE, Equipe EGE, Centre Saint-Charles, Case 36, 3 Place Victor Hugo, 13331 Marseille Cedex 3, France, †Aix-Marseille Université, CNRS, UMR7256 – Information Génétique et Structurale, 163 Avenue de Luminy Case 934, 13288 Marseille Cedex 9, France, ‡Montpellier SupAgro, UMR CBGP, 34988 Montferrier-sur-Lez, France

Abstract

Microsatellite marker development has been greatly simplified by the use of high-throughput sequencing followed by *in silico* microsatellite detection and primer design. However, the selection of markers designed by the existing pipelines depends either on arbitrary criteria, or older studies on PCR success. Based on wet laboratory experiments, we have identified the following factors that are most likely to influence genotyping success rate: alignment score between the primers and the amplicon; the distance between primers and microsatellites; the length of the PCR product; target region complexity and the number of reads underlying the sequence. The QDD pipeline has been modified to include these most pertinent factors in the output to help the selection of markers. Furthermore, new features are also included in the present version: (i) not only raw sequencing reads are accepted as input, but also contigs, allowing the analysis of assembled high-coverage data; (ii) input data can be both in fasta and fastq format to facilitate the use of Illumina and IonTorrent reads; (iii) A comparison to known transposable elements allows their detection; (iv) A contamination check can be carried out by BLASTing potential markers against the nucleotide (nt) database of NCBI; (v) QDD3 is now also available imbedded into a virtual machine making installation easier and operating system independent. It can be used both on command-line version as well as integrated into a Galaxy server, providing a user-friendly interface, as well as the possibility to utilize a large variety of NGS tools.

Keywords: *Chondrostoma*, microsatellite, next-generation sequencing, primer design, quality scores, transposable elements

Received 27 February 2014; revision received 21 April 2014; accepted 25 April 2014

Introduction

Microsatellites or simple sequence repeats (SSR) have been widely used in molecular ecology for the last 20 years as they are highly polymorphic DNA markers and are omnipresent in all eukaryotes (Tóth *et al.* 2000; Meglécz *et al.* 2012a). In spite of the recent and rapid spread of SNP markers (Seeb *et al.* 2011), the use of microsatellites remains pertinent and widespread for several good reasons (Guichoux *et al.* 2011). Firstly, microsatellites have advantages over other molecular markers in revealing fine-scale population structure, selective sweeps and for relatedness estimation (Vignal *et al.* 2002; Wiehe 2007; Narum *et al.* 2008; Hess *et al.* 2011; DeFaveri *et al.* 2013). Secondly, microsatellites provide comparable and reproducible results for different samples of the same species, while some of the techniques of single-nucleotide polymorphism (SNP) genotyping, such as RAD (restriction-site-associated DNA) sequencing, only provide a one-shot result. RAD sequencing is highly dependent on the experimental conditions, thus can be hardly replicated. Therefore, including further samples post-data acquisition and following populations over time is not possible. Finally, cost and time of microsatellite development and genotyping has been considerably reduced due to high-throughput sequencing (Santana *et al.* 2009; Gardner *et al.* 2011; Malausa *et al.*

Correspondence: Emese Meglécz, Fax: +33-413550786;
E-mail: emese.meglecz@imbe.fr

2011) and by multiplexing markers (Guichoux *et al.* 2011) vastly increasing the potential number of markers utilized.

As megabases of sequences are easily and inexpensively obtained for non-model organisms, the analysis of large data sets requires dedicated computer programs that automatize most of the bioinformatics steps of data analyses from raw sequence reads till primer design. QDD (Meglécz *et al.* 2010) was among the first software dedicated to this task. The original version of QDD was designed to incorporate stringent marker selection to maximize laboratory genotyping success rate and to avoid markers with interrupted or compound microsatellites with complex mutation models (Estoup *et al.* 2002; Selkoe & Toonen 2006). This first version proved to be an important tool to help ecologists in microsatellite marker development at the beginning of next-generation sequencing (NGS) era. However, there is a continuous need for implementing further options to improve genotyping success rate and increase the number of potential markers.

QDD3 therefore integrates several new options. Firstly, as sequence contamination can be a non-negligible problem in high-throughput sequencing (Leese *et al.* 2012), information on the best hit to the NCBI nucleotide (nt) database is given for each of the selected sequences in the QDD3 output. This option helps in identifying serious sequence contamination or errors due to incorrect labelling of samples from different species. Secondly, the original QDD pipeline intends to eliminate reads from potential transposable elements by avoiding clusters of reads that are similar to each other but unlikely to come from the same locus. QDD3 can also compare sequences to a bank of known transposable elements, making the identification of potential transposable elements more efficient. These two approaches are complementary, as the first depends on genome coverage but works for any species without previous knowledge on transposons, while the second works only for species where transposable elements have been identified for related species, but does not depend on genome coverage. Thirdly, QDD was originally designed to analyse 454 sequences, as this was the only available NGS platform that produced reads long enough for marker design. Recently, both Ion Torrent and paired-end Illumina sequences have been shown to be suitable for microsatellite marker design (Castoe *et al.* 2012; Elliott *et al.* 2014). As the fastq is the standard output format of these last two technologies, QDD3 can also accept this format as an input to facilitate the use of the pipeline. Fourthly, obtaining more than two to three-fold genome coverage data became affordable, and reads can be assembled into contigs or scaffolds. QDD3 can now also take assembled sequences as an input, thus taking advantages of high-coverage data and the use of dedicated *de novo* assembly software.

The necessity for more potential markers stems from several sources. While it was common to use <10 markers in the pre-NGS era (microsatellite isolation being labour intensive), this is now considered too few to reliably answer many questions in ecological genetics (e.g. Nikolic *et al.* 2009). Genotyping time and cost can be significantly reduced by multiplexing microsatellites, but this can be achieved only if a balanced number of markers of different size ranges are available (Guichoux *et al.* 2011). Finally, for taxa where the genotyping success rate is low (e.g. frequent problem of null alleles), a high number of potential markers are needed for obtaining a sufficiently high number of useable markers (Sinama *et al.* 2011). To increase the number of potential markers, we have decreased the stringency on the selected marker types, and QDD3 provides ample information on microsatellite type and potential problems in the amplicons to facilitate the choice of markers and primer pairs from the pool designed by the pipeline. As a result, users are faced with a large output file, with multiple primer pairs for each sequence and many parameters potentially relevant to increase genotyping success rate. To help users to choose a set of primers, we performed wet laboratory test to determine the most important factors influencing genotyping success.

Genotyping success rate depends on several factors such as the number of priming sites in the genome (the fewer, the better; Andreson *et al.* 2008), the GC content of the primer and the amplicon (closer to 50, the better; Andreson *et al.* 2008; Mallona *et al.* 2011), primer length (the longer, the better; Innis *et al.* 1999) and the primer melting temperature (preferred range: 50–80 °C; Chavali *et al.* 2005). However, earlier studies did not take into account the presence of repetitive elements in the flanking regions (ranging from very short tandem repetitions, like nanosatellites or homopolymers, to potential transposable elements) and the advantages/disadvantages of using consensus sequences in primer design. To pinpoint the main factors influencing genotyping success of microsatellites, we tested 250 primer pairs designed for two Cyprinid fish species *Chondrostoma nasus* (Linnaeus, 1758) and *Parachondrostoma toxostoma* (Vallot, 1837). This step is crucial, as the aim of QDD is to provide a selection of primer pairs that are likely to amplify. Thus, the pertinence of *in silico* analyses by QDD can only be validated by wet laboratory tests. All factors that were significant are included in the output file of QDD3. Furthermore, based on our results, an automatic selection of one primer pair per locus is implemented in QDD3, which helps users to choose a primer pair for each sequence that is more likely to amplify among the

many ones designed. Furthermore, the results also provide guidelines for the choice of the loci among all the potentially available.

Materials and methods

Implementation

QDD version 3 is an open-source program written in PERL and available in two different forms: (i) as QDD-VM, where QDD3 and all its dependencies are packed into a virtual machine (VM). This version is operating system independent. The QDD-VM version can be accessed in two ways, either run as a command line, or integrated into preconfigured local Galaxy server (Giardine *et al.* 2005; Blankenberg *et al.* 2010; Goecks *et al.* 2010) which provides a user-friendly interface; (ii) QDD3 scripts can also be downloaded directly (without the VM), and either installed into an existing Galaxy server or used as a command-line version. This option has the advantage of being a light-weight version, and it is particularly adapted for regular users with multiple data sets. The command-line version can also de-multiplex sequences (i.e. sort sequences according to sequence tags) if relevant. However, third party programmes used by QDD3 should be installed before running the scripts. The command-line version can run under Linux and Windows operating systems. All versions are freely available under the Creative Commons Attribution-NonCommercial-sharealike 3.0 Unported License from http://www.imbe.fr/~emeglecz/qdd.

QDD3 uses the following freely available software: PERL (http://www.perl.org/get.html), BIOPERL (http://www.bioperl.org/), BLAST+ (BASIC LOCAL ALIGNMENT SEARCH TOOL; ftp://ftp.ncbi.nih.gov/blast/executables/blast+/), CLUSTALW2 (Larkin *et al.* 2007) available at ftp://ftp.ebi.ac.uk/pub/software/clustalw2/, PRIMER3 (Rozen & Skaletsky 2000) available at http://primer3.sourceforge.net/. They should be installed if using the command-line version of QDD, but they are already installed in the virtual machine. The installation of REPEATMASKER Open-4.0 (http://www.repeatmasker.org) is optional and needed only for comparing selected sequences against known transposable elements. REPEATMASKER runs under Linux, but it is already installed in the VM, thus becomes system independent. However, it requires users to download the REPBASE Database from http://www.girinst.org. As this database is freely available only for academic users, it is not included in the VM. As this database is regularly updated, users are encouraged to update their database before running the analyses.

QDD3 is designed in four steps that form a pipeline of sequence analysis from raw sequences to primer design (Fig 1).

Step 1: Sequence preparation and microsatellite detection. The input sequences can be either assembled (contigs, scaffolds or chromosomes) or non-assembled sequences. In case of non-assembled sequences, the ideal read length range is 150–500 bp, as shorter reads are less likely to produce functional markers and longer reads have a higher chance of being eliminated in step 2 due to partial similarity to other reads. QDD is not adapted for high-coverage (>one to three-fold) data. These datasets should be first assembled by dedicated assembly software, followed by the analysis of contigs by QDD3.

From non-assembled reads, sequencing adapters can be removed (if relevant), short reads (<80 bp by default) are eliminated and only sequences with microsatellites are kept for further analysis. From assembled sequences, microsatellites are extracted with flanking region length defined by the user.

Step 2: Sequence similarity detection. Sequences are soft masked (i.e. microsatellites are in lower case, while the remaining regions are in upper case letters) and compared using an all-against-all BLAST (-task blastn -evalue 1e-40 -lcase_masking -soft_masking true). This ensures that BLAST searches for alignment seed only in the unmasked regions of the sequences, but then the alignment can be extended into the masked region. Based on the results of the BLAST, sequences are sorted according to their similarities into the following categories: (i) unique sequences are singletons that had BLAST hit to themselves only (autohit); (ii) Nohit CSS (CSS are Cryptically Simple Sequences) are low-complexity sequences that are detected by the lack of an autohit in BLAST (i.e. most of the sequence is masked, thus BLAST does not find a seed for sequence alignment); (iii) Multihit sequences are putative minisatellites, detected by more than one hit between a pair of sequences; (iv) Consensus sequences are based on the alignment of reads, where the pairwise identity was at least 95% (default parameter value) on the overlapping region. Consensus sequences are marked as polymorphic when polymorphism in the microsatellite length is detected among the aligned reads. (v) Grouped sequences had BLAST hits to other sequences, with <95% identity of the overlapping region. Regions covered by BLAST hits are masked by lower case letters. The input for step 3 contains all unique singleton (i) and consensus sequences (iv).

If the sequences were extracted from assemblies, making consensus sequences does not make sense. In this case, step 2 input sequences are still compared by an all-against-all BLAST, but only singletons are kept for primer design. In this way, potential repetitive elements such as transposons or minisatellites are eliminated before primer design.

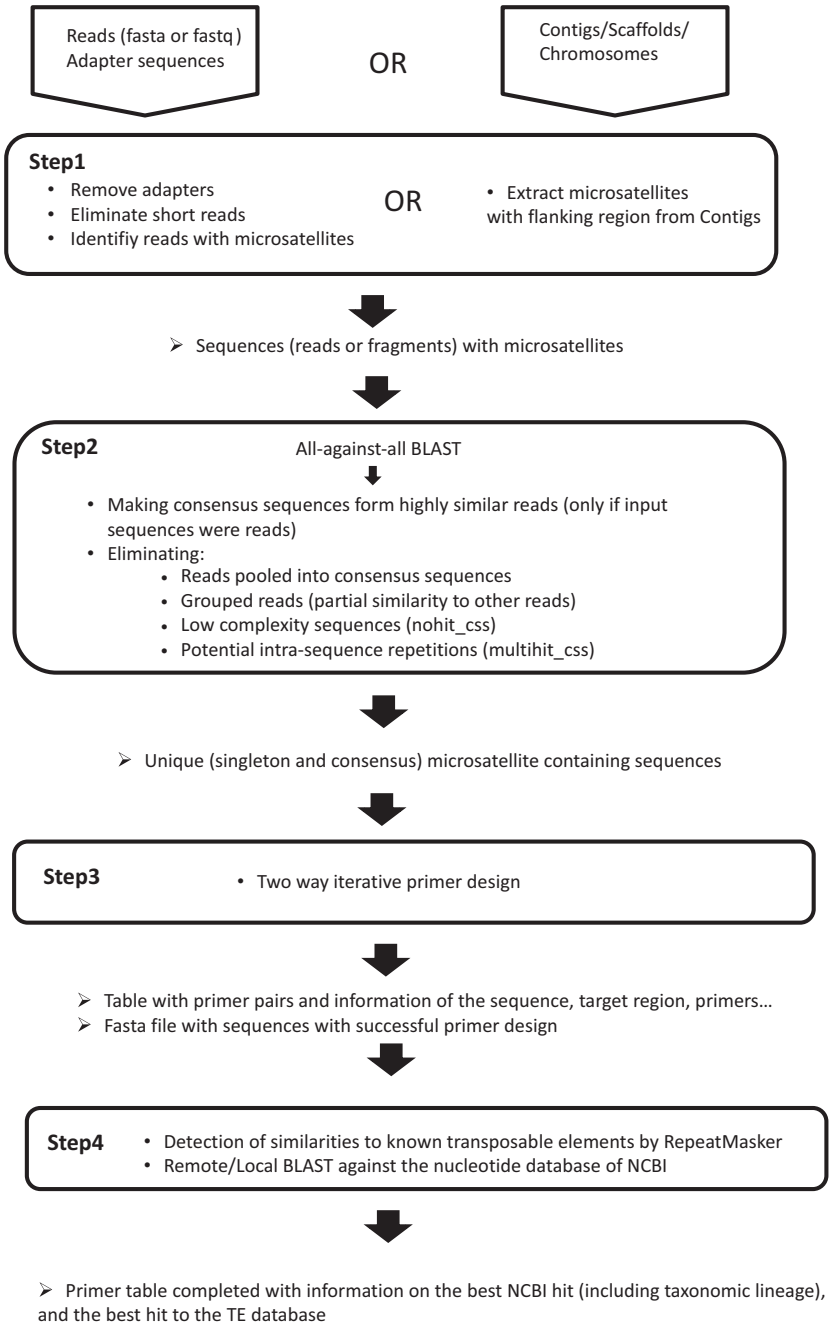

Fig 1. Flow chart of the QDD3 pipeline.

Step 3: Primer design. Primers are designed in a twofold iterative way for each sequence: (i) by increasing the size range of the PCR products, to force the design of primers with varied PCR product size; and (ii) from best to worst scenario concerning the amplicon (including the primer annealing sites) composition. In the best scenario, the amplicon includes only one pure microsatellite with no nanosatellites (three to four repetitions of two to six base-pair motif) or homopolymers (at least five repeats of a single base), while in the worst scenario multiple compound microsatellites, nanosatellites and homopolymers are allowed (Appendix S1, Table S1, Supporting information).

Step 4 (optional): contamination check and comparison to a collection of transposable elements An optional contamination check can be carried out by BLASTing all sequences with successful primer design against the nt database of NCBI and checking the taxonomic linage of the best hit. It is particularly useful for species of generally well-represented taxonomic groups, but it is also useful for most of the species to pinpoint serious contamination from,

for example, bacterial or viral sequences or from an error such as the unintended mix-up of samples from different species. This step can be done either by a local or by remote BLAST. The first option requires the download of the nt database from NCBI (15 Gb) but is faster than the second. The second option relies on a high-speed Internet connection.

Sequences with successful primer design can also be compared to known transposable elements by running REPEATMASKER from within QDD3. This step is optional as only academic users can freely access the RepBase database required for RepeatMasking.

Output. Main output of QDD3 is a primer table which is formatted as a tab delimited text file, easily read by any spreadsheet software (e.g. MS Excel). Each line corresponds to a primer pair, and several primer pairs are designed for each sequence that makes it through to this stage of the pipeline. For each primer pair, information is given on the sequence, the target region (number, type, motif, length of microsatellite) and the primers (e.g. position, length, annealing temperature). If the appropriate options are selected, information on the best hit against NCBI nt database (accession, e-value, score, taxonomy) and the best hit to a known transposable element are also given. Two columns help the easy selection of the 'best' primer pair per read/target region. The choice of the best primer pair within each locus is based on the results of our wet laboratory tests. For each target region, primer pairs are first ordered by the alignment score to the amplicon (excluding the primer annealing sites; the lower, the better), then secondly, by the distance from the target microsatellites (the greater, the better), and lastly by the length of the PCR product (the shorter, the better). The choice between different target regions for the same sequence is based on the complexity of the target region (number and type of microsatellites in the target region) and microsatellite length; preference is given to amplicons with a single pure microsatellite and microsatellites with greater numbers of repeats. Users, however, do not have to rely on our selection of the 'best' primer pairs, as most of the factors studied in our wet laboratory test can be directly accessed by a dedicated column in the primer table. Thus, users can define their own priorities for primer pair selection. A detailed description of the output files is found at http://www.imbe.fr/~emeglecz/qdd_output, and examples of the output files can be downloaded from http://www.imbe.fr/~emeglecz/qdd_download.

Selection of primers, PCR and genotyping

Chondrostoma nasus and *P. toxostoma* microsatellite-enriched partial genomic libraries were sequenced by 454 FLX Titanium technology (Martin *et al.* 2010; Malausa *et al.* 2011). We obtained 4.2 and 5.3 Mb of 454 sequences, containing 29 951 and 40 201 reads, respectively. All reads were deposited to Sequence Read Archives (SRA) of NCBI under the run accessions SRR1137232 for *P. toxostoma* and SRR1137121 for *C. nasus*. Reads were analysed by QDD (version 2.3; default parameters) to select microsatellites of singletons or consensus sequences and design primers. Of the 1696 and 2580 unique sequences containing microsatellite, primers could be designed for 383 and 475 in *C. nasus* and *P. toxostoma*, respectively, and 125 of them were selected from each of the two species for testing in the laboratory. For choosing the 250 sequences and one primer pair for each sequence among the several designed by QDD, seven classes of stringency (Appendix S1, Table S1, Supporting information) were used as a proxy for obtaining the set of primer pairs. This procedure allowed us to cover a wide range of values for the variables tested for the genotyping success rate analyses. The complete list of 46 variables and their values for each of the primers pairs can be found in Appendix S2 (Supporting information).

Chondrostoma nasus and *P. toxotoma* are two cyprinid species that hybridise in nature (Gilles *et al.* 1998; Costedoat *et al.* 2007; Sinama *et al.* 2013). We aimed to establish markers that were useable in both species. All 250 PCR primer pairs were tested on four individuals of each species. First, we used unlabelled primers for PCR amplification, and the PCR products were tested on agarose gels. PCR protocol and conditions as described in Grenier *et al.* (2013). If the amplification of a primer pair produced a clear band of the expected size for all eight individuals, the forward primer was labelled with a fluorescent dye (6-FAM, VIC, NED or PET; Life Technologies), and PCRs were conducted again for eight individuals. Genotyping was conducted as in Grenier *et al.* (2013), and secondary rounds of validation were conducted based on genotype profiles and polymorphism. Genotyping was considered as successful when all eight tested individuals provided unambiguously interpretable profile for genotyping.

If a primer pair failed to amplify in the first stage (unlabelled primers), a new pair of primers was picked for the same sequence (secondary primer pairs). Whenever possible, both primers were replaced, and all three remaining combinations of the two forward and two reverse primers were also tested in the above-described two-stage procedure. For some of the sequences, we did not find a second forward or reverse primer sufficiently different from the first one (i.e. 3′ end of the priming site is more than two bases away from the first primer). In these cases, only the new primer was tested with the originally selected one.

Statistical identification of factors influencing genotyping success

We used a multivariate logistic model to identify factors influencing genotyping success (binary variable). Genotyping success can be influenced by several different factors, such as (i) the presence of nanosatellites/homopolymers in the amplicon, (ii) the complementarity between primers/PCR product, (iii) whether the sequence is a consensus of more than one read or a singleton, (iv) the GC content and length of the primers and amplicons, (v) the distance between the priming sites and the microsatellite, (vi) the penalty score of the primer pair provided by PRIMER3, (vii) the number of hits to a *Danio rerio* (Hamilton, 1822) genome, which is the closest species with a published draft genome, (viii) the microsatellite type and length, (ix) the similarity to a known transposable element and (x) the quality scores of the reads in the priming sites.

Specifying a meaningful variable however is complex, as most of the above-cited factors can be expressed in different ways. For example, variables describing primers can be expressed for each primer individually or for the primer pair. In this latter case, one can take the minimum, maximum or the average value of the primers. Furthermore, values such as GC content or quality scores can be calculated for the whole primer, or only the 3′end, but one also has to define the length of the 3′end. This leads to a plethora of correlated variables. To overcome this problem, the statistical analyses were carried out in two stages. First, variables were pooled into groups according to the factor they describe (Table 1), and for each group with at least two variables, a logistic model was constructed to determine which variables best explain the genotyping success (within the category). In this way, a subset of variables was selected using a stepwise procedure based on Analysis Information Criterion (Akaike 1992). The variables selected at this stage were gathered with all single variables (where only one variable is available for a factor, e.g. GC content of amplicon) to define a reduced set of explanatory variables. This set was subjected to the same selection procedure as described above. To ensure the independence of the analysed primer pairs, these analyses were conducted only on the initial primer pair of each sequence. The selected model was evaluated by considering the rate of correct assignation, using cross-validation.

This global model – based on the first pair of primers for each locus – was used to predict a probability of success of each secondary primer pair (i.e. the pair designed for the loci when PCR was unsuccessful in the first amplification with the unlabelled original primers). Comparison of primer success probabilities was carried out by Wilcoxon signed rank test when the original primer pair of a sequence failed and the second amplified successfully. A *t*-test was conducted when the second primer pair failed, as the number of observations was larger in this latter case.

The effect on polymorphism (one variable per species and a pooled variable over species) of microsatellite motif length, microsatellite type (pure, compound or multiple) and microsatellite repeat number was also tested by univariate logistic models, and the interaction between the significant variables was tested by multivariate logistic models. All statistical analyses were conducted in R (R Development Core Team 2011).

Results

Testing the first primer pair for each of the 250 locus

Of the 250 original pairs of primers tested, 95 (38%; 44 from *C. nasus* and 51 from *P. toxostoma* sequences) gave an unambiguous, interpretable pattern with fluorescently labelled markers for all eight tested individuals (four individuals per species). A total of 70 (28%; 31 and 39 from *C. nasus* and *P. toxostoma*, respectively) of the primer pairs were validated as being polymorphic for at least one of the two species and seven (2.8%; three and four from *C. nasus* and *P. toxostoma*, respectively) were monomorphic for both species but for different alleles, thus discriminating the two species. Primers and PCR conditions on each validated polymorphic or discriminant primer pairs are found in Appendix S3 (Supporting information).

The first stage of the variable selection procedure – conducted for each of the six variable groups (factors) separately – allowed us to select variables that are pertinent for explaining genotyping success and to eliminate the redundancy between variables of each group (Table 1). None of the variables were retained from the variable group describing primer length and GC content, indicating that none of these factors significantly affected the genotyping success. From each of the other five variable groups, one or two variables were selected (Table 1). A further nine 'single' variables each described a different factor, thus a within-group logistic model could not be done in these cases. These 'single' variables were used in the global test together with the seven variables selected from the groups to construct a global logistic model. According to the results of the global model, the maximum alignment score between any of the primers and the amplicon without the primer annealing sites (primer_pcr_alignscore_max variable from 'Primer-amplicon match' group of variables; the lower the better) had the most significant effect of genotyping success (P = 7.5e-10, Odds ratio: OR = 0.73; Table 1). Further variables that explained a significant part of the deviance

Table 1 Variables tested for their effect on genotyping success

Variable	*P* group	*P* global	Variable definition
Variable group 1: Nanosatellite or homopolymer in amplicon			
Overlap_nanosat_primer			[0,1,2] Number of primers that overlaps with nanosatellites
Max_nano_repn_primer			Maximum number of repeats in primer
Nano_primer3	0.040		[0,1,2] Number of primers where the 3′ end overlaps with nanosatellites
nano_rep_target	0.067		Total number of repeats in nanosatellites present in the amplicon
nano_rep_target_star			[0,1] Absence/presence of nanosatellite in amplicon
homo_rep_star			Absence/presence of homopolymer in amplicon
homo_rep			Total length of homopolymers in the amplicon (sum of the length of all homopolymers)
homo_max			Length of the longest stretch of homopolymer in the amplicon
Variable group 2: Primer-amplicon match			
primer_pcr_alignscore_max	3.6e-12	7.5e-10	Maximum alignment score between any of the primers and the amplicon
primer_pcr_match_max			Maximum length of uninterrupted match between any of the primers and the amplicon
primer_pcr_alignscore_mean			Mean of alignment score between each of the primers and the amplicon
primer_pcr_match_mean			Mean of the max. length of uninterrupted match between two primers and the amplicon
Variable group 3: Primer duplex			
fw_rev_alignscore	0.144		Alignment score between the two primers
fw_rev_match			Maximum length of uninterrupted match between the two primers
Variable group 4: Consensus vs. singleton			
cons_n_star			1 = singleton, 2 = consensus
cons_n	0.010	0.059	Number of reads underlying the sequence (1 for singleton, >1 for consensus)
Variable group 5: Primer length and GC content			
primer_length_min			Length of the shorter primer
primer_length_max			Length of the longer primer
primer_length_mean			Average length of the FW and REV primer
gc_primer_min			GC proportion: minimum of the two primers
gc_primer_max			GC proportion: maximum of the two primers
gc_primer_mean			GC proportion: average of the two primers
gc_primer_5_min			GC proportion: minimum of five bases at the 3′ end of the two primers
gc_primer_5_max			GC proportion: maximum of five bases at the 3′ end of the two primers
gc_primer_5_mean			GC proportion: average of five bases at the 3′ end of the two primers
gc_primer_3_min			GC proportion: minimum of three bases at the 3′ end of the two primers
gc_primer_3_max			GC proportion: maximum of three bases at the 3′ end of the two primers
gc_primer_3_mean			GC proportion: average of three bases at the 3′ end of the two primers
gc_primer_half_min			GC proportion: minimum of the 3′ half of the two primers
gc_primer_half_max			GC proportion: maximum of the 3′ half of the two primers
gc_primer_half_mean			GC proportion: average of the 3′ half of the two primers
Variable group 6: Quality scores of the reads in the primer annealing sites			
primer_score_mean			Average over all bases of both primers
primer_score_min			Minimum over all bases of both primers
mean_3prime_3bases			Average over three bases et the 3′ end of the primers
min_3prime_3bases	0.076		Minimum over three bases et the 3′ end of the primers
mean_3prime_5bases	0.155		Average over five bases et the 3′ end of the primers
min_3prime_5bases			Minimum over five bases et the 3′ end of the primers
Factors described by a single variable			
GC content of amplicon			
gc_pcr_product	NA		GC proportion of the amplicon
Length of the amplicon			
PCR_length	NA	0.028	Length of the PCR product
Distance between primer 3′ and microsatellite			
Min_primer_target_dist	NA	0.032	Minimum distance of primers and the microsatellite
Primer pair penalty			
primer_pair_pen	NA		Primer pair penalty calculated by Primer3

Table 1 (Continued)

Variable	*P* group	*P* global	Variable definition
Number of hits to the assembled Danio rerio Genome			
danio_hit_star	NA		0, 1 or multiple BLAST hit to the Danio rerio genome
Microsatellites Type			
target_ms_type	NA	0.047	Pure, compound or multiple microsatellites
Length of the microsatellite motif			
target_motif_length	NA		Length of the microsatellite motif (2–6 bases)
Microsatellite length			
target_max_rep	NA		Number of repeats of the longest uninterrupted stretch of microsatellite
Hit to a known transposable element			
RM_hit	NA		Presence (1) or absence (0) of hit by REPEATMASKER

P group, *P* values of the selected variables from each variable group. Factors that can affect genotyping success rate can be expressed by many different ways leading to different but probably redundant variables. From each group of variables, 0–2 variables were selected based on a within-group logistic model; *P* global, *P* values of the variables retained in the final model; When a variable is not selected, no *P* value is indicated in the *P* group and *P* global columns; nanosatellite, 3–4 repeats of 2–6 bp motifs; homopolymer, tandem repeat of a single base at least five times; Amplicon, amplified DNA fragment excluding primers; NA, when only one variable describes a factor, a within-group logistic model cannot be done, but the variable is used in the global test.

of the genotyping success were the number of reads that underlay the sequence of the locus (cons_n, the lower the better; OR = 0.74): the length of the PCR product (PCR_length; the shorter the better OR = 0.99); the minimum distance between the primers and the microsatellite (Min_primer_target_dist; the longer the better, OR = 1.02); the target region complexity (target_ms_type). Target region complexity refers to microsatellite type (pure or compound) if there was only one microsatellite in the target region and to the number of microsatellites if there were more than one. Primer pairs targeting more than two microsatellites never amplified (OR < 0.5), while one pure targeted microsatellite amplified better than all other categories (Appendix SI, Fig. S1, Supporting information). The error rate of assignation via cross-validation is equal to 0.304, which means that one expects a correct assignation in ca. 70% of the cases using the previously selected four variables as predictors.

The effect of factors influencing polymorphism was studied both separately for the two species (toxostoma_polymorphism and nasus_polymorphism) and for a pooled variable of the two species. Target region complexity did not have a significant effect on the polymorphism in any of the tests ($P = 0.358$ for *C. nasus*, $P = 0.996$ for *P. toxostoma* and $P = 0.094$ for the pooled variable). Univariate logistic models indicated that polymorphism increased from di- to tetra-nucleotide motifs ($P = 0.008$ for *C. nasus*, $P < 1e\text{-}6$ for *P. toxostoma* and $P = 7e\text{-}4$ for the two species together), and with the number of repeats of the microsatellites ($P < 1e\text{-}6$ for *C. nasus*, $P = 3e{-}4$ for *P. toxostoma* and $P < 1e\text{-}6$ for the two species together). Multivariate logistic models were constructed to study the interaction between the microsatellite motif length and the number of repeats in the microsatellites. For numerical reasons, a model with interaction between these two explanatory variables could be constructed only for polymorphism in *P. toxostoma*, but the interaction was not significant ($P = 0.712$). Additive models indicated a significant effect of the number of repeats at fixed repeat motif length values on polymorphism ($P = 0.015$ for *C. nasus*, $P = 0.015$ for *P. toxostoma* and $P = 0.005$ for the two species together), but the effect of motif length at fixed repeat number values was significant only for *P. toxostoma* ($P = 0.013$) and not for *C. nasus* ($P = 0.226$) or for the pooled variable over species ($P = 0.088$).

Primer redesign

After testing the first primer pair for each of the 250 loci with unlabelled primers, 134 of them were rejected, as they either did not give clear bands of the expected size, or some of the eight tested individuals did not amplify. For these loci, in 29 cases (21.6%), no new primer pair could be picked, for 51 loci (38.1%) either a new forward or reverse primer could be selected (that permitted one new PCR) and for 54 loci (40.3%), a new pair of primers was chosen (that permitted three new PCRs). From the 105 loci (54 from *C. nasus* and 51 from *P. toxostoma*, respectively) with at least one new primer, 20 (19.0%) were accepted based on the genotyping pattern of the marked primers. Fourteen of them (13.3%; eight from *C. nasus* and six from *P. toxostoma*, respectively) were validated as polymorphic and two (1.9%; one from each species) as monomorphic but discriminating between species.

Based on the model constructed from the first pair of primers for each locus, a probability of success was predicted for each primer pair, both for the original and for

the redesigned pairs. When comparing the probability of success between the original primer pairs that did not work and the secondary pairs, which worked, secondary primers had a slightly but non-significantly higher probability of success (Wilcoxon signed rank test; $V = 147$, $P = 0.123$, $n = 20$). A paired *t*-test did not indicate differences when comparing unsuccessful primer pairs for both the original and secondary design ($t = -0.993$, d.f. = 210, $P = 0.322$).

Discussion

Factors affecting genotyping success rate

Although a considerable amount of work has been done on identifying parameters that influence PCR or genotyping success (Innis *et al.* 1999; Chavali *et al.* 2005; Andreson *et al.* 2008; Mallona *et al.* 2011), our analyses highlighted some further parameters that can be helpful in predicting the genotyping success before conducting wet laboratory tests. The most important factor coming out of our analyses was, not surprisingly, the alignment score between primers and the amplified fragment (excluding primer annealing sites). Primers designed in low-complexity and/or semirepetitive regions (CSS as Cryptically Simple Sequence) are likely to partially align to other sites as well. This factor could be particularly relevant for microsatellite amplification, as in many microsatellite-containing sequences, flanking regions contain low-complexity and/or CSS regions. Although one can set parameters in PRIMER3 to only pick primers with close to 50% GC content to avoid primer design in low-complexity regions, this can be very restricting and could considerably reduce the number of sequences with primers. Furthermore, PRIMER3 cannot eliminate primers with CSS regions. Although at the second step of QDD, reads composed mainly of CSS regions are eliminated, reads with short stretches of CSS will pass this filtering. It is therefore possible that primers for which the 3′ end overlaps with a low-complexity/CSS regions could align to several sites of the read thus failing to amplify a single fragment. Following the same line of logic, it is also easy to understand the importance of the distance between primers and microsatellites. Our analyses showed that primers further away from the target microsatellite amplify better than the ones close to it, which is likely the consequence of the observation that very often the pure stretches of microsatellite repetitions are surrounded by CSS sequences.

Target region complexity was also found to be a significant parameter affecting genotyping success. Fragments with one pure microsatellite in the target region amplified better than multiple or compound microsatellites. The presence of multiple microsatellites may be a sign that a locus is part of a repetitive region of the genome. Primer pairs producing shorter PCR products were found to be less prone to PCR failure than the ones with longer PCR product length. This unfortunately leads to a trade-off, as for efficient multiplexing, long PCR products are also needed.

A very interesting and counter-intuitive result was that primers based on consensus sequences are more prone for genotyping failure than primers based on singletons. *A priori*, we supposed that consensus sequences can give better results, as sequencing errors are corrected and potential polymorphic sites are masked by making consensus sequences from reads of a same locus. This, however, supposes that highly similar reads correspond to the same locus (i.e. different alleles of a polymorphic locus and not paralogous sequences). However, we found that the genotyping success decreased with increasing consensus read number (the number of reads underlying a consensus sequence). As we used microsatellite-enriched DNA for sequencing, it is difficult to estimate the coverage of the targeted genomic regions. By extracting microsatellites from the assembled *Danio rerio* genome with 150 bp flanking region on both sides, the total length of the extracted fragments was 222 Mb, which gives an approximate coverage of 0.02× for the 4–5 Mb of data. This coverage level is in fact very low. As it was found that there is no substantial bias in genome coverage in 454 sequencing of enriched (Martin *et al.* 2010) or shotgun libraries (Meglécz *et al.* 2012b), we would expect few cases where more than one read covers the same locus. It thus seems likely that the reads pooled into consensus sequences are rather reads of different loci of the same transposable elements (TE), which explains the increasing PCR failure rate with increasing number of reads building the consensus (consensus read number). An alternative way of detecting transposable elements is to compare reads to a library of known TEs, as it is done by REPEATMASKER (http://www.repeatmasker.org/). Fernandez-Silva *et al.* (2013) found that stringent filtering based on the quality score of 454 reads, and similarity to known transposable elements, increased the genotyping success rate in the wet laboratory. We have also included the presence or absence of similarity to a known transposable element as detected by REPEATMASKER (RM_hit). Although, there was a correlation between the variables RM_hit and cons_n (Spearman correlation test; $S = 642\,206.1$, $P < 2.2e\text{-}16$), according to the global model used here, the number of reads underlying the consensus sequences is a better parameter to indicate genotyping success. However, this result can be model dependent. Comparing rodent or primate sequences to a database of known TEs is likely to reveal most of the TE elements in the data set, as TEs are well described in these groups. On the other hand, for other groups that

have been less studied for TE content, the consensus read number can be more informative. One should note, however, that the consensus read number should not be taken into account in absolute terms. In low-coverage data set, consensus read number as low as two or three can be a sign of transposable elements, while for 1× or higher-genome coverage data, it is expected that many loci are covered by more than one reads.

Following the results of Fernandez-Silva *et al.* (2013), who found a higher primer-to-marker conversion rate when applying quality control of the reads and TE screening than using the QDD pipeline alone, we have also tested the effect of the quality scores in the primer binding sites. None of the quality variables had a significant effect on genotyping success in our analyses. This implies either that most of the improvement in their study is likely to come from the TE screening or that sequence quality matters below a certain quality score threshold. In this last case, one would not expect an effect of sequence quality scores when most of the sequences are good quality, but the effect of sequence quality could become important when errors become frequent. Therefore, as a common sense practice, users should not neglect quality filtering of sequences as a first step of any analyses.

New features of QDD3

Several programs have been published for microsatellite detection and primer design (Fukuoka *et al.* 2005; Dereeper *et al.* 2007; Kraemer *et al.* 2009; Martins *et al.* 2009; Meglécz *et al.* 2010; Churbanov *et al.* 2012), and QDD is one of the most widely used. The major disadvantage of earlier programmes is that they were not designed to handle tens of thousands of sequences that are typical of NGS sequencing (Fukuoka *et al.* 2005; Dereeper *et al.* 2007; Martins *et al.* 2009). Programmes that can deal with NGS data, however, are not necessarily easy to install and use by researchers with basic informatics skills. HighSSR (Churbanov *et al.* 2012) is a recent program designed for NGS data and deals with the microsatellite detection, elimination of redundancy and primer design. The data are stored in a ProstgreSQL database, which provides the potential of exploiting the data for researchers with bioinformatics skills, but not necessarily easy for others. The main output of the program consists of a list of primers and not primer pairs, and it is not organized into a table, which would be easy to handle by any user. STAMP is an extension of the STADEN package (Kraemer *et al.* 2009) for microsatellites detection and primer design from contigs assembled by Staden, but it is not suited for low-coverage NGS data. QDD3 can handle both raw reads and assembled sequences, and available in different forms to suit the need of researches with different informatics background. The command-line version is very light, and batch submission of many files is possible and can use the full capacity of the computer. Its disadvantage is the necessity to install third party programs. QDD galaxy packed into a virtual machine (running QDD3 scripts) provides an easy installation, where all the third party programs are set up, and galaxy provides a graphical interface. The drawbacks of this version are the size of package, and the VM cannot use the full capacity of the computer. The main output of the QDD3 (run in command line or Galaxy) is an easy to use synthetic table, where each row represents a primer pair with information on primers, target region, GenBank and REPEATMASKER hit organized in columns. Two columns are designed for the automatic selection of only one primer pair per sequence/target region, where the selection of the 'best pair' is based on our wet laboratory results. All the others columns with diverse information on the primer pairs, target region of sequences type can be used for refining the selection of the markers to be tested based on the preferences of the users.

QDD3 has been tested mostly on 454 data, as the read length of 454 reads (300–700 bp) is compatible with the most frequently used microsatellite marker sizes. However, other platforms, such as Illumina and IonTorrent, can produce a much higher amount of data for a lower cost; thus, it is very tempting to use them. Although there is no theoretical reason to exclude these types of sequences for QDD3 analyses, there are two potential problems. The typical read length of 100–250 bp of these platforms (especially for earlier versions) would provide only short markers – making multiplexing less efficient – and the chance of finding a sufficient flanking region of both sides of the microsatellites is lower than for long reads (Elliott *et al.* 2014). Furthermore, it is more likely that primers are designed close to the target microsatellites and thus increasing the risk of PCR failure. However, these problems are likely to be resolved soon with the improvement of the read length of these technologies. Using the advantage of paired-end reads can also overcome of this problem, and this is nicely implemented in the PAL_FINDER program (Castoe *et al.* 2012).

The other potential problem is related to the genome coverage of the NGS data. QDD was designed for low-coverage data, where overlapping reads are rare, and can be easily combined into a consensus. It cannot replace assembly programs which are designed to produce contigs from high-coverage data. Gigabases of sequences come at a reasonable cost with Illumina and Ion Torrent sequencing, leading to a very wide coverage range of different species. For low-coverage data (<0.2–0.5×), QDD3 can be used, although run time can be considerable due to an all-against-all sequence comparison. For high-coverage data (more than 1–3×), the best way is

to assemble the reads by a dedicated assembler [e.g. CAP3 (Huang & Madan 1999), Velvet (Zerbino & Birney 2008), MIRA (Chevreux *et al.* 2004)], and use the contigs in QDD3 for microsatellite detection and primer design. For intermediate coverage, however, run time might be excessively long and consensus sequences can be meaningless. It is also important to note that the coverage would also influence the results of PAL_FINDER software (Castoe *et al.* 2012), where eliminating the redundancy is based on the number of exact matches of the designed primers to the whole data set. High-coverage data are likely to produce fewer PALs (potentially amplifiable SSR loci), while for low coverage, the chance of classifying reads to high quality PALs that are in fact repetitive is increased.

Due to the presence of different sequencing platforms, the differences in sequencing cost, their throughput and read length, it is increasingly difficult to suggest one single program that fits to all kind of data and users. QDD3 is suitable for analysing assemblies coming from high-coverage data, and raw reads from low-coverage data. We have provided guidelines for marker selection based on wet laboratory tests, rather than choosing markers based on theoretical considerations only. Furthermore, QDD3 is adapted to be used by researchers with different bioinformatics backgrounds (from novice to expert) and can run under Linux and Windows operating system. Moreover, the program is continuously under development to follow the evolution of data types and the suggestions of the users.

Acknowledgements

We are grateful for users who provided feedback on earlier version of QDD and Cyrille Lepoivre who tested QDD3. We thank Michael G. Gardner for useful comments on the manuscript. Data used in this work were partly produced through the technical facilities of the Centre Méditerranéen Environnement Biodiversité.

References

Akaike H (1992) Information theory and an extension of the maximum likelihood principle. In: *Breakthroughs in Statistics Springer Series in Statistics* (eds Kotz S, Johnson NL), pp. 610–624. Springer, New York.

Andreson R, Möls T, Remm M (2008) Predicting failure rate of PCR in large genomes. *Nucleic Acids Research*, **36**, e66.

Blankenberg D, Kuster GV, Coraor N *et al.* (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology*, **89**, 19.10.1–19.10.21.

Castoe TA, Poole AW, de Koning APJ *et al.* (2012) Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. *PLoS ONE*, **7**, e30953.

Chavali S, Mahajan A, Tabassum R, Maiti S, Bharadwaj D (2005) Oligonucleotide properties determination and primer designing: a critical examination of predictions. *Bioinformatics*, **21**, 3918–3925.

Chevreux B, Pfisterer T, Drescher B *et al.* (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research*, **14**, 1147–1159.

Churbanov A, Ryan R, Hasan N *et al.* (2012) HighSSR: high-throughput SSR characterization and locus development from next-gen sequencing data. *Bioinformatics*, **28**, 2797–2803.

Costedoat C, Pech N, Chappaz R, Gilles A (2007) Novelties in hybrid zones: crossroads between population genomic and ecological approaches. *PLoS ONE*, **2**, e357.

DeFaveri J, Viitaniemi H, Leder E, Merila J (2013) Characterizing genic and nongenic molecular markers: comparison of microsatellites and SNPs. *Molecular Ecology Resources*, **13**, 377–392.

Dereeper A, Argout X, Billot C, Rami J-F, Ruiz M (2007) SAT, a flexible and optimized Web application for SSR marker development. *BMC Bioinformatics*, **8**, 465.

Elliott CP, Enright NJ, Allcock RJN *et al.* (2014) Microsatellite markers from the Ion Torrent: a multi-species contrast to 454 shotgun sequencing. *Molecular Ecology Resources*, **4**, 554–568.

Estoup A, Jarne P, Cornuet J-M (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular ecology*, **11**, 1591–1604.

Fernandez-Silva I, Whitney J, Wainwright B *et al.* (2013) Microsatellites for next-generation ecologists: a post-sequencing bioinformatics pipeline. *PLoS ONE*, **8**, e55990.

Fukuoka H, Nunome T, Minamiyama Y *et al.* (2005) Read2Marker: a data processing tool for microsatellite marker development from a large data set. *BioTechniques*, **29**, 472,474–476.

Gardner MG, Fitch AJ, Bertozzi T, Lowe AJ (2011) Rise of the machines - recommendations for ecologists when using next generation sequencing for microsatellite development. *Molecular Ecology Resources*, **11**, 1093–1101.

Giardine B, Riemer C, Hardison RC *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Research*, **15**, 1451–1455.

Gilles A, Lecointre G, Faure E, Chappaz R, Brun G (1998) Mitochondrial phylogeny of the European cyprinids: implications for their systematics, reticulate evolution, and colonization time. *Molecular Phylogenetics and Evolution*, **10**, 132–143.

Goecks J, Nekrutenko A, Taylor J, Team Galaxy (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, **11**, R86.

Grenier R, Costedoat C, Chappaz R, Dubut V (2013) Two multiplexed sets of 21 and 18 microsatellites for *Phoxinus phoxinus* (L.) and *Gobio gobio* (L.) developed by cross-species amplification. *European Journal of Wildlife Research*, **59**, 291–297.

Guichoux E, Lagache L, Wagner S *et al.* (2011) Current trends in microsatellite genotyping. *Molecular Ecology Resources*, **11**, 591–611.

Hess JE, Matala AP, Narum SR (2011) Comparison of SNPs and microsatellites for fine-scale application of genetic stock identification of Chinook salmon in the Columbia River Basin. *Molecular Ecology Resources*, **11**(Suppl. 1), 137–149.

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. *Genome research*, **9**, 868–877.

Innis MA, Gelfand DH, Sninsky JJ (1999) *PCR Applications: Protocols for Functional Genomics*. Academic Press, San Diego.

Kraemer L, Beszteri B, Gäbler-Schwarz S *et al.* (2009) STAMP: extensions to the STADEN sequence analysis package for high throughput interactive microsatellite marker design. *BMC Bioinformatics*, **10**, 41.

Larkin MA, Blackshields G, Brown NP *et al.* (2007) Clustal W and clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Leese F, Brand P, Rozenberg A *et al.* (2012) Exploring Pandora's box: potential and pitfalls of low coverage genome surveys for evolutionary biology. *PLoS ONE*, **7**, e49202.

Malausa T, Gilles A, Meglécz E *et al.* (2011) High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries. *Molecular Ecology Resources*, **11**, 638–644.

Mallona I, Weiss J, Egea-Cortines M (2011) pcrEfficiency: a Web tool for PCR amplification efficiency prediction. *BMC Bioinformatics*, **12**, 404.
Martin J-F, Pech N, Meglécz E *et al.* (2010) Representativeness of microsatellite distributions in genomes, as revealed by 454 GS-FLX Titanium pyrosequencing. *BMC Genomics*, **11**, 560.
Martins WS, Soares Lucas DC, de Souza Neves KF, Bertioli DJ (2009) WebSat - A web software for microsatellite marker development. *Bioinformation*, **3**, 282–283.
Meglécz E, Costedoat C, Dubut V *et al.* (2010) QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects. *Bioinformatics*, **26**, 403–404.
Meglécz E, Nève G, Biffin E, Gardner MG (2012a) Breakdown of phylogenetic signal: a survey of microsatellite densities in 454 shotgun sequences from 154 non model Eukaryote species. *PLoS ONE*, **7**, e40861.
Meglécz E, Pech N, Gilles A, Martin J-F, Gardner MG (2012b) A shot in the genome: how accurately do shotgun 454 sequences represent a genome? *BMC Research Notes*, **5**, 259.
Narum SR, Banks M, Beacham TD *et al.* (2008) Differentiating salmon populations at broad and fine geographical scales with microsatellites and single nucleotide polymorphisms. *Molecular Ecology*, **17**, 3464–3477.
Nikolic N, Fève K, Chevalet C, Høyheim B, Riquet J (2009) A set of 37 microsatellite DNA markers for genetic diversity and structure analysis of Atlantic salmon *Salmo salar* populations. *Journal of Fish Biology*, **74**, 458–466.
R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology*, **132**, 365–386.
Santana QC, Coetzee MPA, Steenkamp ET *et al.* (2009) Microsatellite discovery by deep sequencing of enriched genomic libraries. *BioTechniques*, **46**, 217–223.
Seeb JE, Carvalho G, Hauser L *et al.* (2011) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources*, **11**, 1–8.
Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters*, **9**, 615–629.
Sinama M, Dubut V, Costedoat C *et al.* (2011) Challenge of microsatellite development in Lepidoptera: *Euphydryas aurinia* (Nymphalidae) as a case study. *European Journal of Entomology*, **108**, 261–266.
Sinama M, Gilles A, Costedoat C *et al.* (2013) Non-homogeneous combination of two porous genomes induces complex body shape trajectories in cyprinid hybrids. *Frontiers in Zoology*, **10**, 22.
Tóth G, Gáspári Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Research*, **10**, 967–981.
Vignal A, Milan D, SanCristobal M, Eggen A (2002) A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution*, **34**, 275–305.
Wiehe T (2007) Identification of selective sweeps using a dynamically adjusted number of linked microsatellites. *Genetics*, **175**, 207–218.
Zerbino DR, Birney E (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

E.M. wrote QDD, participated in experimental design and analysis and wrote the manuscript. N.P. did the statistical analyses, wrote the manuscript. V.D., A.G. and J.F.M. participated in experimental design, tested QDD3 and wrote the manuscript. P.H. participated in the scripting and tested QDD3. A.T. and R.G. did the laboratory work.

Data Accessibility

QDD is freely available under the Creative Commons Attribution-NonCommercial- SHAREALIKE 3.0 Unported License from http://www.imbe.fr/~emeglecz/qdd. The website includes the executable, full documentation, and sample files.

The raw 454 reads used of *C. nasus* and *P. toxostoma* are deposited to the SRA database of NCBI (study accessions SRP035552 and SRP035548). Sequences of validated markers are deposited in GenBank with accessions KJ169457–KJ169549.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 Stringency levels used by QDD in primer design.

Fig. S1 PCR amplification success in function of target region complexity.

Appendix S1 Supplementary figures and Tables.

Appendix S2 Data for all tested primer pairs and all variables (xlsx file).

Appendix S3 Table of all validated polymorphic or discriminant primer pairs with PCR product size range, number of alleles and GenBank accessions.