Fast sequence-based microsatellite genotyping development workflow.

Olivier Lepais[1,2*], Emilie Chancerel[1], Christophe Boury[1], Franck Salin[1], Aurélie Manicki[2], Laura Taillebois[2], Cyril Dutech[1], Abdeldjalil Aissi[3], Cecile F. E. Bacles[2], Françoise Daverat[4], Sophie Launey[5], Erwan Guichoux[1].

[1] BIOGECO, INRA, Univ. Bordeaux, 33610 Cestas, France

[2] ECOBIOP, INRA, Université de Pau et Pays de l'Adour, 64310 Saint-Pée-sur-Nivelle, France

[3] LAPAPEZA, University of Batna 1 Hadj Lakhdar, 0500 Batna, Algeria

[4] EABX, IRSTEA, Cestas Cedex 33612, France

[5] ESE, Ecology and Ecosystem Health, Agrocampus Ouest, INRA, 35000 Rennes, France

[*]Corresponding author

Abstract

Application of high-throughput sequencing technologies to microsatellite genotyping (SSRseq) has been shown to remove many of the limitations of electrophoresis-based methods and to refine inference of population genetic diversity and structure. We present here a streamlined SSRseq development workflow that includes microsatellite development, multiplexed marker amplification and sequencing, and automated bioinformatics data analysis. We illustrate its application to five groups of species across phyla (fungi, plant, insect and fish) with different levels of genomic resource availability. We found that relying on previously developed microsatellite assay is not optimal and leads to a resulting low number of reliable locus being genotyped. In contrast, *de novo ad hoc* primer designs gives highly multiplexed microsatellite assays that can be sequenced to produce high quality genotypes for 20 to 40 loci. We highlight critical upfront development factors to consider for effective SSRseq setup in a wide range of situations. Sequence analysis accounting for all linked polymorphisms along the sequence, quickly generates a powerful multi-allelic haplotype-based genotypic dataset, calling to new theoretical and analytical frameworks to extract more information from multi-nucleotide polymorphism marker systems.

Keywords: sequence-based microsatellite genotyping, SSR-GBS, SSR-seq, haplotype sequence, SNPSTR, HapSTR.

Introduction

Development of high-throughput sequencing (HTS) technologies since the mid 2000's has greatly widened the scope of genetic applications in research fields such as agronomy, ecology and evolutionary biology where the use of molecular markers is common place. The easy collection of millions of DNA sequences has also facilitated the development of molecular markers such as single nucleotide polymorphism (SNP; (Delord et al., 2018; Garvin, Saitoh, & Gharrett, 2010)) or microsatellites (or single sequence repeats SSR; (Guichoux et al., 2011; Malausa et al., 2011)) for any species of interest. Beyond marker development, HTS has been largely adopted for genotyping-by-sequencing (GBS), thus taking advantage of restriction enzymes to target sequencing efforts in specific parts of the genome that are shared by conspecific individuals (Baird et al., 2008; Miller, Dunham, Amores, Cresko, & Johnson, 2007). Restriction site-associated DNA polymorphism sequencing allows simultaneous identification and genotyping of SNP in non-model species and interrogation of thousands to hundreds of thousands of loci genome-wide but remains expensive when population samples comprise hundreds of individuals. Recent development of amplicon sequencing which targets previously identified SNP makes it possible to genotype hundreds to thousands of individuals at a few hundreds of SNP at a very reasonable cost (Aykanat, Lindqvist, Pritchard, & Primmer, 2016; Campbell, Harmon, & Narum, 2015; Nguyen-Dumont, Pope, Hammet, Southey, & Park, 2013). If sequence-based genotyping of SNP is gaining momentum, application of HTS to microsatellite genotyping was until recently lagging behind. Updating microsatellite genotyping to modern technologies remains important for several reasons. Firstly, some current scientific questions in ecology or evolutionary biology can be answered using a moderate number (e.g. a dozen) of highly polymorphic multi-allelic loci such as microsatellites (Harrison, Saenz-Agudelo, Planes, Jones, & Berumen, 2013). Secondly, variation in the number of repeated oligonucleotide motif is a unique kind of polymorphism with specific mutation mechanism and rate which in itself provides a complementary picture of genetic variation to nucleotide substitutions across populations (Haasl & Payseur, 2011) and genomes (Willems et al., 2014). Thirdly, it becomes more and more obvious that microsatellite polymorphism is involved in numerous biological processes such as gene expression regulation and epigenetic mechanisms (Bagshaw, 2017; Sadd, Schaack, Zhao, Sun, & Su, 2018), and more generally in phenotypic variation (Xie et al., 2019) including human diseases (Gymrek, 2017; Hannan, 2018). Thus, while marker preference evolves through time with specific markers dominating the genotyping field over a period of time following technological advances (Schlötterer, 2004; Seeb et al., 2011), maintaining our capability to interrogate any kind of

polymorphism in the context of rapid HTS technological advances is paramount and should be prioritized.

In the age of HTS, and in comparison with SNP genotyping, microsatellite genotyping suffers from several drawbacks: homoplasy (alleles of identical size having different underlying sequence (Estoup, Jarne, & Cornuet, 2002; Viard, Franck, Dubois, Estoup, & Jarne, 1998)), time and cost consuming development and genotyping, low throughput, lack of automation and data standardization (Ellis et al., 2011; Moran, Teel, LaHood, Drake, & Kalinowski, 2006). Yet, all of these limitations are linked to the fact that currently microsatellite genotyping relies on allele discrimination based on amplicon size assessed by capillary electrophoresis (De Barba et al., 2016) and do not hold true if microsatellite genotyping transitions to sequence-based genotyping. Previous direct comparisons of capillary electrophoresis and sequence-based microsatellite genotyping (called SSRseq thereafter) validated SSRseq as a reliable method (Darby, Erickson, Hervey, & Ellis-Felege, 2016; Vartia et al., 2016). Advantages of sequence-based over capillary electrophoresis-based microsatellite genotyping are significant. Direct access to allele sequence reveals additional polymorphisms that remain hidden when using only allele size to identify variation (Darby et al., 2016; Šarhanová, Pfanzelt, Brandt, Himmelbach, & Blattner, 2018; Vartia et al., 2016). Sequence data thus reduces allele homoplasy because alleles of the same size may contain molecular variation that does not translate into size variation such as SNP, indels masking variation in repeat number, or presence of two adjacent SSR motifs with complementary size variation (Darby et al., 2016). As a result, SSRseq offers refined genetic diversity estimation and population structure inference (Bradbury et al., 2018; Darby et al., 2016; Neophytou et al., 2018; Viruel et al., 2018).

Studies published so far have explored specific technical or analytical aspects of SSRseq. Several bioinformatics approaches have been developed (Barbian et al., 2018; Hoogenboom et al., 2016; Suez et al., 2016; Zhan et al., 2017), different laboratory protocols tested (Pimentel et al., 2018; Vartia et al., 2016) and means to account for molecular variation on population genetic inference compared (Curto et al., 2019; Neophytou et al., 2018; Šarhanová et al., 2018; Viruel et al., 2018). Together, these studies explored numerous issues surrounding the technical and analytical advantages of SSRseq over traditional methods.

Here, we propose an integrative workflow for the development of a SSRseq analysis for application to non-model species. We apply this workflow to five species groups from families taken across phyla (Basidiomycota: *Armillaria ostoyae*; Angiosperms: *Quercus faginea* and *Q. canariensis*; Euarthropoda: *Melipona variegatipes*; Chordata: *Alosa alosa, A. fallax* and *Salmo*

salar) that markedly differ in the amount of genomic data already available for them. We compare a broad range of possible development scenarios including sequencing of already optimized microsatellite assay traditionally genotyped on capillary sequencers, optimizing primers around already developed microsatellites, and developing microsatellite *de novo* from a range of genomic resources when available, or from newly generated low coverage random genome sequences for species without existing genomic resources. Building on our previous experience in development of highly multiplex microsatellite genotyping protocols (Guichoux et al., 2011; Lepais & Bacles, 2011), we propose a streamlined approach with demonstrated application to groups of species which a wide range of genetic and evolutionary characteristics. We applied a microsatellite sequence data analysis pipeline to produce haplotypic data accounting for all polymorphisms detected in sequenced alleles, validated by extensive blind-repeat genotyping to estimate SSRseq error rates. We emphasize that efficient and powerful multi-polymorphism haplotype-based genotyping approaches are easy to develop and apply, calling for new theoretical and analytical development to extract more information from this new kind of molecular markers.

Materials and Methods

Studied species, SSRseq development strategies and DNA isolation

We selected a range of species from different Kingdoms among biological models studied in our laboratories (Table 1). For *S. salar* we took the most straightforward route by amplifying and sequencing microsatellites using previously developed primers. To develop a refined workflow, we chose species with different level of genomic resource availability to test alternative *de novo* microsatellite development strategies that are likely to cover a wide range of situations (Table 1).

Table 1: SSRseq development strategy and DNA characteristics of species used in this study.

Kingdom	Class	Species	SSRseq development strategy	Number of DNA extracted individuals	DNA extraction protocol	DNA quality	Reference
Animalia	Actinopterygii	*Salmo salar*	Previously developed loci	1152	Salt-chloroform (Gauthey et al., 2015)	High	(Bacles et al., 2018; Lepais et al., 2017)
Plantae	Eudicots	*Quercus faginea*, *Q.canariensis*	Re-designed primers around already developed loci using reference genome sequence of closely related species	380	Invisorb DNA Plant HTS 96 kit	High	this study
Animalia	Actinopterygii	*Alosa alosa*, *A.fallax*	*De novo* loci development based on available repeat-enriched library sequencing	382	Invitrogen PureLink Genomic DNA Mini kit	Highly degraded	(Taillebois et al., submitted)
Fungi	Agaricomycetes	*Armillaria ostoyae*	*De novo* loci development based on reference genome sequence	384	CTAB (Prospero, Lung-Escarmant, & Dutech, 2008)	Heterogeneous	this study
Animalia	Insecta	*Melipona variegatipes*	*De novo* loci development based on newly generated low coverage whole genome shotgun sequencing	91	Qiagen DNeasy 96 Blood & Tissue Kit	High	this study

SSRseq using previously-developed primers

The most straightforward approach to SSRseq microsatellite genotyping, i.e. based on sequence information from existing primers, was applied to *S. salar* using two marker selection strategies. In the first strategy, we selected 23 primers from a list of 81 microsatellites available for *S. salar* (O'Reilly and Wright 1995; Slettan *et al.* 1996; Ozaki *et al.* 2001; Rexroad *et al.* 2001; Gilbey *et al.* 2004; Paterson *et al.* 2004; King *et al.* 2005; Vasemägi *et al.* 2005; Thorsen *et al.* 2005; Yano *et al.* 2013, see details in Supporting Table S1). Selection criteria included allele size smaller than 300 bp to ensure that sequencing reads can span the entire allele length including library construction and absence of sequences complementarity between primers tested using Multiplex Manager (Holleley & Geerts, 2009). In the second strategy, we chose to sequence a set of 15 microsatellites (Supporting Table S1) that are routinely amplified in a single multiplexed PCR and genotyped using electrophoresis-based capillary sequencers (Bacles et al., 2018; Lepais et al., 2017).

SSRseq with microsatellite (re)development

Genomic resources for microsatellite (re)development

For the other species, microsatellites primers were either re-designed or developed *de novo* from various genomic resources (Table 1, Figure 1 top panel).

For *Quercus* sp., primers were re-design primers in flanking regions of existing microsatellite markers to optimize multiplex amplification and sequence interpretability while taking advantage

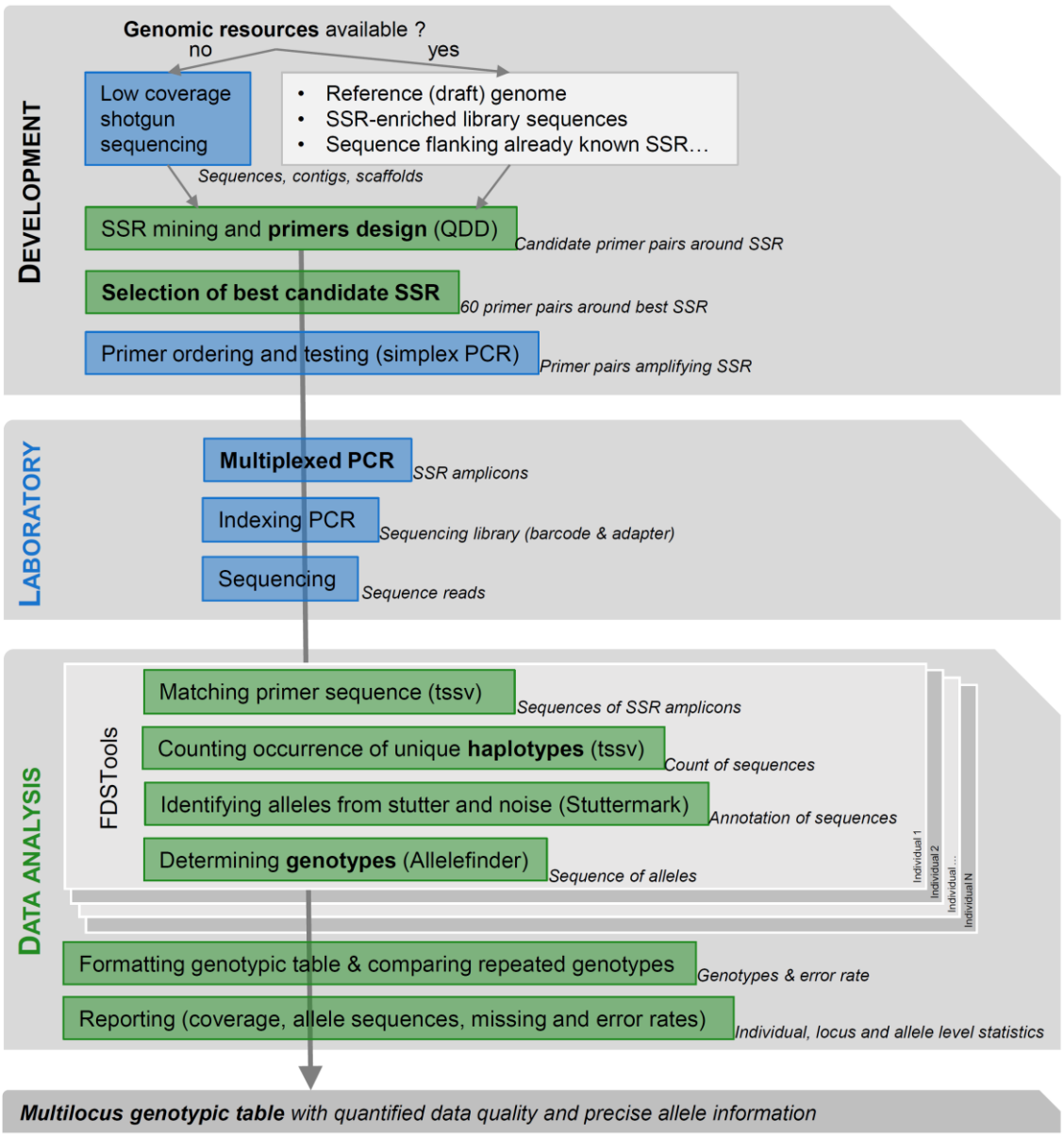
of already validated microsatellite markers. We extracted primer sequences from 259 polymorphic and mapped EST-derived (Durand et al., 2010) and 35 genomic microsatellites (Kampfer, Lexer, Glössl, & Steinkellner, 1998; Steinkellner et al., 1997)). The primer sequences were mapped on the *Q. robur* reference genome (Plomion *et al.* 2018, GenBank accession GCA_003013145.1) using bowtie 2 v2.3.4.1 (Langmead & Salzberg, 2012) and genomic sequence spanning from 200 bp downstream of the forward primer to 200 bp upstream of the reverse primer position were extracted using bedtools v2.25.0 (Quinlan, 2014) resulting in 294 sequences used as genomic resource to design new primers.

For *Alosa* sp., we used sequences obtained from a Roche 454 GS-FLX sequencing run on a microsatellite-enriched DNA library (Rougemont, Besnard, Baglinière, & Launey, 2015) following the method described in Malausa *et al.* (2011).

For *A. ostoyae*, we used the reference genome sequence as genomic resource to identify microsatellites loci (Sipos *et al.* 2017, GenBank accession GCA_900157425.1).

Finally, no genomic resources were available for *M. variegatipes*. We therefore used DNA from one individual to construct a whole-genome sequencing library using Illumina TruSeq DNA kit. The resulting library was sequenced on an Illumina MiSeq flowcell using v3 2x300 pb paired-end sequencing kit. Mothur software v1.39.5 (Schloss et al., 2009) was used to assemble paired reads and keep paired reads with a minimum overlap of 100 bp without mismatch. We randomly subsampled 500,000 reads from the resulting 6.74 million paired reads for subsequent microsatellite identification, because a few hundreds of thousands of random sequences are sufficient to identify thousands of microsatellites (Castoe et al., 2010; Curto et al., 2019; Lepais & Bacles, 2011).

Figure 1: Workflow for SSRseq markers optimization or development depending on genomic resource availability, from selection to multiplexed amplification and library preparation to bioinformatics analysis (refined approach).



De novo *microsatellite development or primer re-design*

The command line version of QDD pipeline v 3.1 (Meglécz et al., 2010, 2014) was run on either i) a reference genome sequence, ii) a set of low coverage random sequences or iii) sequence extracted around already characterised microsatellite loci (Table 2) , to detect sequences containing microsatellites, identify good quality sequence (singletons and consensus) from problematic sequences (sequences showing low complexity, minisatellites or multiple BLAST hits

with other sequences) and design primer pairs flanking the identified microsatellites (Figure 1). QDD pipeline was run with default parameters, except for the primer design step (pipe3) where parameters were stringently defined in order to improve their capacity to be amplified jointly in a single multiplexed PCR (Qiagen Multiplex kit handbook; Lepais & Bacles 2011): primer optimal size was set to 25 nucleotides (min: 21, max: 26), optimal annealing temperature to 68°C (min: 60°C, max: 75°C) with a maximal difference of 10°C between primers of a same pair and optimal percentage of cytosine and guanine of 50% (min: 40%, max: 60%). In addition, PCR product size was set between 120 and 200 bp to be compatible with a wide range of sequencing platforms and to produce robust genotyping assays that can be used to analyse degraded or low quantity DNA samples. QDD analysis results in a large number of candidate loci with designed primer pairs from which a restricted number of loci can be selected (Figure 1). At the exception of *Quercus* sp. where a restricted set of input sequences necessarily limited choice among resulting candidate microsatellites, several quality criteria were used to select 60 microsatellites among hundreds to thousands candidates for further testing. We followed recommendations of Meglécz *et al.* (2014) to prioritize primer pairs with increased likelihood of amplification success by selecting microsatellite from singletons and not from consensus sequence, with pure repeat instead of compound motifs, showing at least 20 bp between the primers and the repeat motif, and with flanking region showing high complexity (e.g. primer pairs from the design group A following QDD terminology: no minisatellite, no other microsatellite in the flanking region, no homopolymer in the flanking region or the primer). In addition, we further selected microsatellites with the highest number of repeats to increase the probability of selecting polymorphic loci, avoided motif that can form hairpin such as AT repeats and when possible included a variety of di-, tri- and tetra nucleotide repeats.

Primer modification and simplex amplification tests

For Ion Torrent sequencing (Table 2), tag A 5'-GCCTTGCCAGCCCGCTCAG-3' was add to the 5' end of each forward primer and tag B sequence 5'-GCCTCCCTCGCGCCA-3' (Blacket, Robin, Good, Lee, & Miller, 2012; Margulies et al., 2005) was added to the 5' end of each reverse primer. For Illumina sequencing (Table 2), specific tags 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3' and 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-3' were added to the 5' end of the forward and reverse primer sequences respectively. Designed primers were tested for potential primer dimer formation using Primer Pooler (Brown et al., 2017). Primer pairs showing a deltaG lower than -6 kcal/mol are likely to form dimer and result in poor amplification in a multiplexed PCR. For locus involved in significant interactions, alternative primers were selected or in absence of

alternative, another locus was selected from the candidate list. Oligonucleotides were ordered in a plate format from Integrated DNA Technologies with standard desalt purification at a final concentration of 100 µM. Primer pairs were tested using simplex amplification performed on one individuals per species using Qiagen Multiplex kit in a final volume of 10µL and with a final concentration of each primer of 0.2µM. Amplification conditions consisted of an initial denaturation step at 95°C for 15 minutes, followed by 35 cycles consisting of denaturation at 95°C for 20s, annealing at 59°C for 60s and extension at 72°C for 30s, and a final extension step for 10 minutes at 72°C. Amplicons and 1 Kb size standard were then loaded on a 3% agarose gel containing GelRed or SyberSafe dye and migrated at 100v for 15 minutes. Each locus was screened under UV light for positive amplification with a clear band at the expected size.

Table 2: Summary of the tested scenarios for SSRseq genotyping

Species	SSRseq development strategy	Candidate loci	Screened loci	Number of loci in a single multiplexed PCR	Sequencing Plateform	Total number of individuals sequenced	Number of individual analyzed for this study	Number of repeated individuals	Sequenced loci[†]	Mean (cv) sequences / loci / individual	Reliable loci genotyped[‡]	Overall success rate
S. salar	Previously developed loci	81	-	23	Ion Torrent PGM i316	960	66	66	20	161 (61%)	9	39%
				23	Illumina MiSeq - 1/2 nano PE	192	96	96	20	70 (41%)	10	43%
S. salar	Previously developed loci	15[§]	-	15	Illumina MiSeq - 1/2 nano PE	192	96	96	13	99 (65%)	7	47%
Quercus sp.	Primer redesign around previously developped loci	462	60	60	Illumina MiSeq - 1/3 V2 PE	380	46	46	53	260 (32%)	40	67%
Alosa sp.	*De novo*	2872	60	28	Illumina MiSeq - 1 nano SE	382	156	156	25	95 (58%)	21	75%
				28	Illumina MiSeq - 2 nano SE	382	156	156	26	198 (58%)	24	86%
				28	Illumina MiSeq - 3 nano SE	382	156	156	26	267 (58%)	24	86%
A. ostoyae	*De novo*	1806	60	51	Illumina MiSeq - 1/2 V2 PE	384	384	96	48	243 (83%)	38	75%
M. variegatipes	*De novo*	8937	60	54	Illumina MiSeq - 1/4 V2 PE	182	91	91	49	176 (45%)	39	72%

[†]: loci showing substantial evidence for minimum sequencing success (at least 20 sequences in at least 50% of the individuals); [‡]: reliable loci (less than 50% of missing data among individuals and less than 6% of genotyping error based on comparison of repeated genotyping); [§] routinely genotyped using optimized multiplexed PCR and capillary-based sequencer. *FDSTools* analysis using two parameter sets: *stutterfinder* -s:-1:50,+1:10 *allelefinder* -m 15 -n 20; and *stuttermark* -s:-1:70,+1:10 *allelefinder* -m 10 -n 20. For each marker, four parameter combination were used (two strategies and two parameters set) and for each strategy, the best parameter set was used for a given locus.

Multiplex microsatellite amplification and sequencing library construction

From 192 to 960 individuals were analyzed depending on the taxa considered including from 46 to 156 repeated individuals to check the reproducibility of the method (Table 2). For each of the taxonomic groups, a three-round multiplex PCR approach was used to amplify all loci simultaneously and improve amplification homogeneity and thus coverage of sequence between loci (Chen et al., 2016). In the first round, a multiplexed PCR including all selected locus primers was performed (Figure 1, Supporting Table S1 for locus characteristics including primer sequences). PCR amplification were carried out in 96-well plates in a final volume of 5µL or 10µL using Qiagen Multiplex kit, 0.05 µM of each forward and reverse tailed primers and about 40 ng of template DNA (depending on the species, 1µL of undiluted or diluted isolated DNA). PCR cycles were performed on Applied Biosystems 2720 or Verity thermocyclers and consisted of a denaturing step of 5 min at 95°C followed by 20 cycles of 95°C for 30 s, 59°C for 180 s and 72°C for 30 s (Qiagen Multiplex kit handbook; Lepais & Bacles 2011). In the second round, additional Taq polymerase added with the aim to use remaining primers completely. The PCR mixture of a final volume of 5 or 10 µL consisted of 2.5 or 5 µL of Qiagen Multiplex kit, 1.5 or 3 µL of undiluted amplicon and 1 or 2 µL of water. The PCR cycles were identical as in the first round. The third round is the indexing PCR (Figure 1) that add Ion Torrent or Illumina sequencing adaptors and barcodes used to assign each sequence to an individual. For Ion Torrent sequencing, we used 106 different barcodes resulting in a total of 960 barcode combinations. For Illumina sequencing, we used the Nextera XT index set allowing for 384 barcode combinations. The PCR mixture of a final volume of 10 µL consisted of Qiagen Multiplex kit Master Mix, 0.5 µM of sequencing platform-specific adaptor and 5 µL of undiluted amplicon resulting from the second PCR round. PCR cycles consisted of a denaturing step of 5 min at 95°C followed by 15 cycles of 95°C for 30 s, 59°C for 90 s and 72°C for 30 s and a final extension step of 68°C for 10 minutes. Amplicons from the 96 wells within a plate were pooled together in an Eppendorf tube, and purified with 1.8X Agencourt AMPure XP beads (Beckman Coulter, UK). Quality check and quantification were done using Agilent Tapestation D1000 kit and Qubit fluorometric system (Thermo Fisher Scientific), and quantified using Kapa libraryquantification kit in a Roche LightCycler 480 quantitative PCR. The resulting two to ten pools were pooled in equimolar concentration and sequenced using an Ion Torrent PGM i316 chip or Illumina MiSeq flowcell using nano or v2 2x250 bp paired-end sequencing kit (Table 2) at the Genome Transcriptome Facility of Bordeaux.

Bioinformatics data analysis

Sequence preparation

After sequence demultiplexing and adaptor trimming using a sequencer platform built-in software, quality was controlled using *FastQC* (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and reads shorter than 70 bp were removed using *cutadapt* (Martin, 2011). When paired-end sequencing was used, paired reads were assembled into contigs using *pear* (Zhang, Kobert, Flouri, & Stamatakis, 2014) with the default scoring method based on assembly score (allowing for mismatch and accounting for base quality scores), a minimum overlap of 50 bp and a maximum assembled sequence length of 450 bp. For *Alosa* sp., reverse reads quality was generally poor, therefore, only the forward read was used as its length (250 bp) encompasses the whole length of the sequenced loci (max. 200 bp). In some cases, microsatellite amplicons from several species were pooled prior to PCR indexing so that two individuals from two species shared an identical barcode combination and were sequenced in a single run. Forward primer sequences of species-specific loci were used to sort sequences belonging to different species into different fastq files using *fqgrep* tool (https://github.com/indraniel/fqgrep) allowing for one mismatch.

Converting microsatellite sequences to genotypes

We used *FDSTools* v1.1.1 pipeline (Hoogenboom et al., 2016) to identify sequences corresponding to the microsatellite alleles and call genotypes for each individual (Figure 1). This analytical tool was chosen because it accounts for any kind of polymorphism detected across the analysed sequences (including variation in the number of repeated motifs, SNP or indels) while integrating specific tools to detect true allele from stutter mutation introduced during amplification that are typical of microsatellite markers.

First, *tssv* (Anvar et al., 2014) matches primer sequences, allowing for 8% of mismatch, to identify sequences originating from each locus and count the occurrence of each unique sequence found for each locus for each individual (Figure 1). Then, *Stuttermark* uses the number of repeats of the microsatellite motif and the coverage of each unique sequence to flag unique sequences as potential allele, stutter resulting from slippage mutation during PCR and erroneous sequences (Figure 1). FInallt, *Allelefinder* calls one or two alleles among the most abundant sequences flagged as potential alleles by *Stuttermark* (Figure 1). Following the FDSTools analysis, several custom-made *bash* routines were used to format the tabulated genotypic table, compare genotypes from repeated individuals to estimate locus specific allelic error rate defined by the number of allele mismatches between replicated genotypes divided by the number of diploid alleles comparred. In addition, allele level information was extracted including allele sequences, three-digit code used for genotype

annotation, number of occurrence across individuals and allele length. Locus characteristics such as missing data rate and number of alleles are also summarized across the analysed individuals (Figure 1). All these bioinformatics steps have been embedded into a single bash script (*SSRseq_DataAnalysis_ParametersComparison.sh* available at https://doi.org/10.15454/HBXKVA) that allow to modify key analytical parameters (Supporting Materials & Methods) to evaluate their effect on the quality of the genotypic call (number of detected alleles, error and missing data rates).

Two analytical strategies were compared. In the first strategy (called *FullLength* thereafter), all variation identified between primers was considered as a haplotype irrespectively of the nature of the polymorphism because all polymorphisms are physically-linked to each other in reads that encompass the whole locus. The *FullLenghth* strategy may be too complex for some loci or species showing high levels of polymorphism. In the second strategy, the analysis was therefore restricted to the repeat motif only (strategy called *RepeatFocused* thereafter). In this case, primer and flanking sequences surrounding the repeated motif are indicated in the primer sequence field of the *FDSTools* input file. The *tssv* step then extracts the sequence corresponding to the repeat motif region (still allowing for 8% of mismatch which will accommodate flanking sequence polymorphism) to perform subsequent genotypic call with *Stuttermark* and *Allelefinder* as described above. As the analytical approach used by *FDSTools* is based on counting coverage of unique sequences, any variation identified within the repeated motif region, including variation in repeated motif number, SNP and indel within the motif region, will still be accounted for when defining alleles. While this *RepeatFocused* strategy may be more robust due to the shorter length of sequence analysed, it should identify a smaller number of alleles compared to the *FullLength* strategy.

Comparing analytical approaches

For each locus, we determined the best analytical strategy using the following criteria by order of importance: estimated allelic error, amount of missing genotypes and number of detected alleles. Loci that showed more than 6% of allelic error or more than 50% of missing data across individuals (within each species group) were flagged as failed and removed from further inspection. For each species and analytical strategy, we recorded the number of genotyped loci, mean allelic error and missing data rate and the total number of alleles (haplotypes) across loci. We then determined the best overall approach for each locus and used it to genotype each locus generating a final *Combined* genotypic dataset for each species using a specific bash script (*SSRseq_DataAnalysis_FinalGenotyping.sh* available https://doi.org/10.15454/HBXKVA). Finally, the overall development success rate was computed for each species by dividing the number of reliable loci by the number of loci included in the multiplexed PCR.

Gain from sequence information

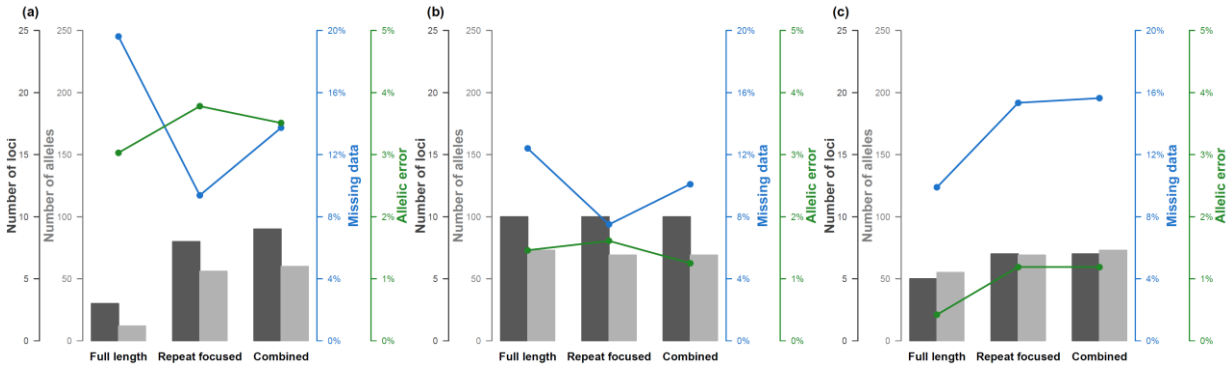
The number of identified alleles based on sequence information (haplotypes) was compared to the number of alleles differing in amplicon length only for all analysed loci to assess the gain of information obtained by using sequence data and estimate size homoplasy. We investigated further the nature of the detected polymorphism by counting, for each locus, the number of variations in the number of repeats, SNP and indels in the repeated motif and the flanking regions that differ between haplotypes.

Results

Sequence-based genotyping of previously developed microsatellites

Attempts to genotype previously developed microsatellites in *S. salar* either from a new combination of 23 loci or a routinely-used multiplex of 15 loci resulted in a low number of reliable loci genotyped (Table 2). The overall success rate (the percentage of reliable loci over the number of loci amplified in the multiplexed PCR) ranged from 39% to 47% and the number of reliable loci from 7 to 10 (Table 2). Moreover, the quality of the generated genotypic datasets is relatively low with a rate of missing data and allelic error above 10% and 1% respectively (Figure 2). The sequences produced on the PGM Ion Torrent platform resulted in the lowest genotypic data quality (Figure 2). The same genotyping protocol sequenced on the Illumina MiSeq platform produced higher genotypic quality with lower missing data and allelic error rates (Figure 2) in spite of a 2.3 times lower mean coverage per sequenced locus per individual (Table 2). The lowest performance of the Ion Torrent platform is due to the higher sequencing error rate linked to spurious insertion-deletion around homopolymer tracts. This results in a waste of sequencing reads, increasing noise (e.g. erroneous singletons: unique sequence with a coverage of one or a very few reads), at the expense of sequence exactly matching the true alleles. The Illumina MiSeq sequencing platform was thus used for subsequent analyses in other species.

Figure 2: Results of SSRsq development from previously developed microsatellites in *S. salar* for (a) a new multiplex of 23 microsatellite sequenced with Ion Torrent PGM and (b) Illumina MiSeq sequencing platforms, and (c) a routinely-used multiplex of 15 microsatellites sequenced with Illumina MiSeq sequencing platform. Number of reliable loci, total number of alleles, missing data and allelic error rates are indicated for three bioinformatics analysis strategies that focused either on all polymorphism across the sequence, on polymorphism within the repeated motif only, or a combination of the best strategy for each locus.



Sequence-based genotyping of *de novo* developed microsatellites

In contrast, overall success rate of *de novo* microsatellite development ranged from 67% to 86% (Table 2). Given the high number of candidate microsatellites typically identified from high-throughput sequencing or reference genome sequence, we were able to screen as much as 60 new loci, and combined most of them (from 28 to 60) in a single multiplexed PCR for amplification (Table 2). As a result, the final number of reliable loci was consistently high amounting to 24 for *Alosa* sp., 38 for *A. ostoyae*, 39 for *M. variegatipes* and 40 for *Quercus* sp. (Table 2). All these protocols produced high quality genotypic dataset, with low missing data (2.6% for *Quercus sp.*, 6.8% for *Alosa sp.*, 5.4% for *M. variegatipes*) at the exception of *A. ostoyae* (20.7%) and low allelic error rates as estimated based on blind-repeat genotyping (0.4% for *Quercus* sp., 0.6% for *Alosa* sp., 0.9% for *M. variegatipes* and 0.7% for *A. ostoyae*). For *Alosa* sp. and *Quercus* sp. all reliable loci were found to be transferable between species.

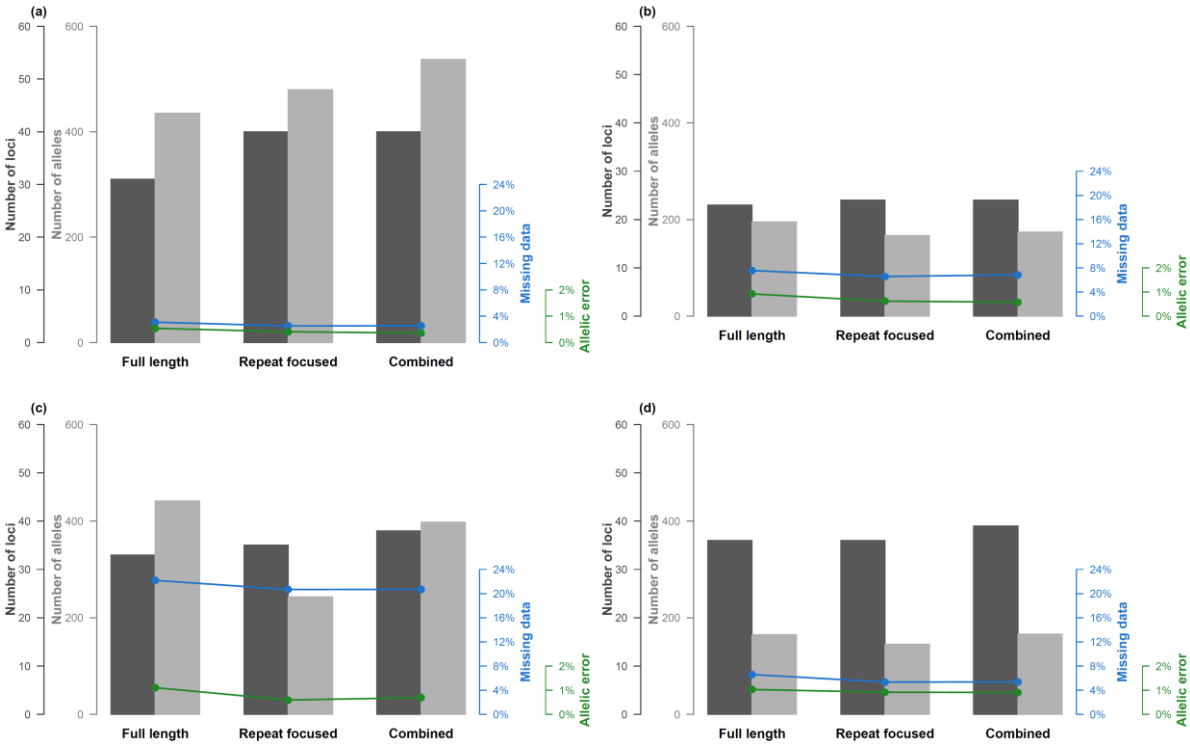
We explored the effect of sequence coverage on genotypic data quality on *Alosa* sp. by sequencing the same set of 28 microsatellites amplified in 156 individuals using one, two or three Illumina MiSeq nano flowcells (Table 2). The resulting increased coverage, from 95 to 198 and 267 sequences respectively per locus per individual (Table 2), recovers more data for those highly degraded DNA samples. First, increasing the coverage from 95 to 198 sequences by locus by individual detected three additional loci, while a further increase in coverage failed to recover additional locus (Table 2). Second, the missing data rate linearly decreases with the increase in coverage, from 16.4% to 9.0% and 6.8% with 95, 198 and 267 sequences per locus per individual respectively (Supporting Material Figure S1). It is worth noting that the allelic error rate is not affected by genome coverage, as it varied only slightly between 0.5% and 0.7% without any correlation to coverage (Supporting Material

Figure S1). A significant result is that even in conditions when only highly degraded DNA templates are available, reliable genotypic data can be obtained with moderate coverage. With the pipeline parameters used here, increasing coverage will reduce missing data but not genotyping error rate, however with more stringent pipeline parameters, increasing coverage may have more significant impact on the number of reliable loci genotyped and missing data rate.

Whole sequence VS repeated motif polymorphism analysis

Focusing the analysis on the repeated motif slightly increased the number of reliable loci and tends to produce fewer missing data and allelic errors (Figure 3). However, numerous polymorphisms that may be present in the flanking sequences are not accounted for. Indeed, analysing all polymorphism detected between the PCR primers resulted in a higher mean number of allele per locus, at the expense of slightly higher missing data and allelic error rates (Figure 3). Interestingly, 17% of the loci can be analysed reliably using either the *FullLength* or the *RepeatFocused* analytical approach. Thus combining analytical strategies by selecting the best approach for each locus resulted into an optimized dataset (Figure 3). Even for loci with reliable genotypes irrespectively of the analytical approach chosen, selecting the one that produces the best quality data (in terms of number of alleles, missing data and error rate) leads to an improved dataset quality. This combined strategy results in recovering the highest number of loci and alleles while keeping missing data and allelic error rates at the lowest (Figure 3).

Figure 3: Results of SSRseq development based on newly optimized microsatellites in (a) *Quercus* sp., (b) *Alosa* sp., (c) *A. ostoyae* and (d) *M. variegatipes* sequenced with Illumina MiSeq sequencing platform. Number of reliable loci, total number of alleles, missing data and allelic error rates are indicated for three bioinformatics analysis strategies that focused either on all polymorphism across the sequence, on polymorphism within the repeated motif only, or a combination of the best strategy for each locus.



Types of polymorphism detected across species

While most common population genetics applications do not necessitate to characterize the nature of the polymorphism differentiating alleles, the main advantage of sequence data (in addition to analysing a much higher number of loci) is to be able to identify allelic variation that does not translate into size variation, i.e. the only variation that is detected when using classical electrophoretic approaches. Across species, the proportion of alleles would have remained undetected by capillary electrophoresis (size homoplasy) ranging from 6% for *M. variegatipes*, 11% for *S. salar*, 14% for *Alosa* sp., 35% for *Quercus* sp. and 53% for *A. ostoyae* (Table 3). Conversely, the increase in the proportion of allele detected by accessing sequence data ranges from 6% for *M. variegatipes*, 13% for *S. salar*, 16% for *Alosa* sp., 56% for *Quercus* sp. and 113% for *A. ostoyae* (Table 3). Indeed, beside variation in repeat number, we identified numerous SNP and indel either in the flanking sequence or in the repeat motif itself (Figure 4, Table 3). In fact, additional polymorphism beyond variation in repeat number was the rule rather than the exception (Table 3). However, differences in the proportions of the type of the detected polymorphism were found between

species (Figure 4). While repeat number variation represented more than 80% of the polymorphism detected in *S. salar*, *Quercus* sp., *Alosa* sp. and *M. variegatipes*, SNP in the flanking sequence was the most frequent polymorphism for *A. ostoyae* representing 49.8% of the variation, much high than variation of repeat number estimated at 29.8% (Figure 4). For *A. ostoyae*, SNP in the repeat motif and indel in the flanking sequence also represent a significant proportion of the polymorphism detected (12.6% and 6.5% respectively). *Quercus* sp. were characterised by SNP both within the repeat motif (7.5%) and the flanking region (9.3%). The second most common polymorphism for *Alosa* sp. was SNP in the repeat motif (8.5%) and for *M. variegatipes* SNP in the flanking region (5.5%). For *S. salar*, variation in the number of motif was much more frequent than for other species (93.0%) compared to other polymorphisms that represent a marginal proportion of the variation (less than 3% each).

Figure 4: Proportion of detected polymorphism types within the repeat motif or in the flanking sequence for each sample per species group.

Table 3: Detected polymorphism.

						Polymorphism in the repeat motif[§]			Polymorphism in the flanking sequences[§]	
Species	**Number of loci**	**Number of alleles with sequence difference[†]**	**Number of alleles differing by amplicon size[‡]**	**% of size homoplasy**	**% of increase in alleles due to sequence**	**Repeat number variation**	**SNP**	**Indel**	**SNP**	**Indel**
S. salar[¶]	14	122	108	11%	13%	107 (14)	3 (3)	-	2 (2)	3 (3)
Quercus sp.	40	537	346	35%	55%	406 (40)	38 (25)	1 (1)	47 (18)	13 (10)
Alosa sp.	24	174	150	14%	16%	130 (23)	13 (15)	2 (2)	4 (4)	3 (3)
A. ostoyae	38	398	187	53%	113%	187 (33)	79 (26)	8 (7)	312 (23)	41 (16)
M. variegatipes	39	166	156	6%	6%	147 (39)	3 (3)	1 (1)	9 (7)	5 (5)

[†]: irrespectively of polymorphism type, computed based on the Combined analysis strategy; [‡]: simulating the number of alleles that would have been identified using traditional capillary electrophoresis, computed based on the FullLength analysis strategy and accounting for allele size only on the same locus as the Combined approach. [§]: total number of alleles (and number of loci in brackets) for each polymorphism type. [¶]: combination of two sequence based microsatellite genotyping protocols.

Discussion

While relying on previously developed microsatellite assay is far from optimal, we found that primer redesign around known locus or *de novo* microsatellite development based on strict criteria gives successful single highly multiplexed PCR amplification and sequencing for 20 to 40 loci. However, key initial factors need to be considered for efficient SSRseq setup.

Not all roads lead to Rome: navigating pitfalls when adopting HTS for microsatellite genotyping

Our preliminary attempts to apply SSRseq using previously developed microsatellite primers or capillary-based multiplexed microsatellite genotyping protocols clearly failed to produce reliable genotypic data for a sufficient number of loci. In the best case scenario, sequencing of 15 microsatellites in *Salmo* led to the reliable genotyping of 7 loci with 15.6% of missing data and 1.19% of allelic error, a result far worse than the high quality dataset obtained for the same multiplex using traditional capillary-electrophoresis of 14 loci with 0.5% missing data and 0.35% allelic error (Bacles et al., 2018). Similar trends were observed in previous studies that validated the use of HTS to genotype microsatellites and relied on previously developed primers: the generated datasets were characterised by high levels of missing data (up to 45%, (Vartia *et al.* 2016)) or low number of genotyped loci (7 to 8, (Suez *et al.* 2016; Barbian *et al.* 2018)). Two characteristics are problematic for SSRseq from microsatellites initially developed for capillary electrophoresis-based genotyping. First, high variability in locus length and primer characteristics leads to heterogeneous amplification intensity and sequencing coverage across loci. Indeed, efficient multiplexed PCR necessitates careful primer design using strict criteria (Guichoux et al., 2011; Lepais & Bacles, 2011). In addition, the need for variable locus length for optimal multiplexing without allele size range overlap in capillary-based electrophoresis becomes unnecessary for sequencing, because same size loci can be reliably identified simply based on primer sequences. Secondly, starting from a limited number of loci (e.g. 10-20 typical of capillary electrophoresis-based microsatellite genotyping approaches) results in a handful of reliable loci that may be too low for downstream applications. This conclusion agrees well with a previous study developing SSRseq in *S. salar*, where only one out of six (17%) of previously developed loci was successfully integrated into the final panel, compared with a 26% success rate for newly developed primers (Bradbury et al., 2018). Adapting previously developed loci in the same species resulted in a success rate of about 45% in our case (7 out of 15 and 10 out of 23 reliable loci), but *de novo* development in other species was much more successful with a success rate of 75% on

average. Not relying of previous primer design is thus of primary importance for the success of reliable SSRseq protocol development.

Workflow for efficient SSRseq development in non-model species

We propose here a workflow for developing new SSRseq approaches and we demonstrated its efficiency for a range of species with different level of genomic resources availability. Starting from a reasonable number of candidate loci that were identified using readily available or newly generated genomic resources, resulted in an average of 75% of the loci included in the PCR multiplex generating reliable genotypic data. This result compares favourably with previous studies where success rates were lower (47% (Farrell, Carlsson, Carlsson, & Farrell, 2016), 53% (Neophytou et al., 2018)) or similar (78% (Tibihika et al., 2018)) when developing about 20 to 40 loci from a moderate number of candidate loci. However, extensive screening of numerous markers resulted in a much lower success rate: 101 validated loci from 385 tested (26% success rate (Bradbury et al., 2018)) or 43 validated loci from 448 locus tested (10% (Zhan et al., 2017)). In this study, we propose a balanced approach for effective development that relies on simple amplification tests of a limited number of carefully selected candidate loci. Locus validation is then made when sequencing the final set of loci in a way that the validation step, based on the analyse of blind-repeat of at least 48 individuals, jointly generates useful genotypic data (for additional individuals included in the sequencing run, Table 2). In addition, most of the ordered primers will be validated and integrated in the final set of reliable loci which greatly minimise development costs. The workflow is also flexible in terms of number of loci chosen for analysis: if a higher number of loci is required; additional sets of 60 loci could be selected from the candidate locus list and amplified in separated PCR multiplexes that can be pooled together before PCR indexing leading to an effective way to genotype additional loci (Bradbury et al., 2018). We did not test multiplexing more than 60 loci in a single PCR, but we did not see difficulties in doing so (Campbell et al., 2015). In such a situation, careful primer design with strict criteria and control for primer interactions will be key to increase multiplexed amplification success.

The proposed workflow for SSRseq development minimized laboratory steps and analytical optimizations. We chose to start from a moderate number of loci, designed to maximize their compatibility and sequence interpretability, and remove any locus that failed to amplify, produce interpretable or repeatable genotypes. This strategy avoids tedious optimization and reduce the number and complexity of laboratory steps necessary to produce genotypes. Admittedly, additional DNA or amplicon clean up or normalisation might improve sequence quality and coverage across individuals and loci. However, we chose to keep the laboratory procedure as simple as possible and compensate the increase in amplification heterogeneity by additional sequence output resulting in

sufficient coverage (220 reads/loci/individuals on average) to obtain a nearly complete genotypic dataset (generally 5% of missing data excluding the atypical case of *Armilaria* species). Moreover, increasing the number of laboratory steps inflates the risk of handling error and subsequently of genotyping error (Vartia et al., 2016). In this respect, highly multiplexed PCR are a useful technique to reduce laboratory steps and potential associated errors in addition to saving time and cutting costs. Additional tests not presented here showed that the two-stage multiplex amplification prior to PCR indexing that was used to increase amplification homogeneity across loci is not necessary: high coverage is still efficient to compensate the increase in amplification heterogeneity when using a single multiplex PCR step for locus amplification as performed in previous studies (Bradbury et al., 2018; De Barba et al., 2016; Tibihika et al., 2018; Zhan et al., 2017).

We chose to rely on the set of 384 Illumina barcodes combinations because we found it to fit well with the output of the MiSeq sequencing platform when analysing from 20 loci to 300 loci depending on the type of flow cell used. However, studying more than 384 individuals from a single species necessitates either several MiSeq runs (Bradbury et al., 2018) or custom made dual-indexing strategies, as was successfully performed for the Ion Torrent PGM run (960 barcode combinations used) or in previous studies using the MiSeq platform (960 and 1024 barcode combinations (Farrell et al., 2016; Zhan et al., 2017)).

Finally, including repeated individuals is of paramount importance to assess the reliability of the produced genotypic data for each locus. This procedure aims to be best practice in capillary electrophoresis-based microsatellite genotyping (Guichoux et al., 2011; Hoffman & Amos, 2005; Pompanon, Bonin, Bellemain, & Taberlet, 2005) but is even more important in SSRseq. Indeed, not all sequenced loci produced reliable genotypic data. It is thus necessary to be able to identify and exclude loci producing high genotyping errors. At the bioinformatics analysis stage, selecting the best analytical strategy for each locus is an easy task that improve the number of reliable loci and decrease the genotyping error rate. However, a few loci will still show high genotyping error rate due to low coverage or complex polymorphism patterns and should be excluded from the final genotypic dataset. Exploratory analyses testing a much wider number of parameters resulted in limited success in increasing the final number of reliable locus. Extensive parameter testing is a tedious task requiring high computation time with only minor improvement for the final genotypic dataset quality and is not worth the extra effort as long as a sufficient number of loci have been included in the genotyping panel. Furthermore, additional analyses using alternative microsatellite sequences analysis tools such as Megasat (Zhan et al., 2017) and MicNeSs (Suez et al., 2016) also lead to the conclusion that all loci cannot be analysed reliably using a single set of parameters. Thus, we stress the importance to include a significant number of repeated individuals (at least 48) in the first

analysis of a new SSRseq panel, irrespective of the bioinformatics data analysis strategy used, to (1) coarsely optimize analysis parameters for reliable loci, (2) quantify genotyping error rate and (3) identify and exclude unreliable loci. Such procedure has been implemented only in a limited number of previous studies describing SSRseq methods (6 out of 14 published studies thus far (Barbian et al., 2018; Bradbury et al., 2018; De Barba et al., 2016; Šarhanová et al., 2018; Viruel et al., 2018; Zhan et al., 2017)) but should be generalized.

Implications of haplotype based genotyping

We took advantage of the fact that reads span across entire loci to analyse all linked and phased polymorphisms encountered using the FDSTools pipeline (Hoogenboom et al., 2016). This haplotype approach differs from the methods implemented in other sequence-based microsatellite genotyping software such as Megasat (Zhan et al., 2017) which focuses on amplicon length or Micness (Suez et al., 2016) which estimates the number of repeated motifs while accounting for up to one substitution within the microsatellites motif. The haplotype approach has several advantages. Firstly, it is relatively insensitive to sequencing error because the analysis focused on unique sequence with high coverage and thus does not consider the noise generated by sequencing error which produce numerous unique low coverage sequences that are removed. Secondly, by analysing unique sequences, it accounts for any kind of polymorphisms, while at the same time includes an algorithm to identify stutters resulting from slippage mutation due to microsatellite instability during PCR amplification.

Previous studies have demonstrated the added benefit of analysing different types of polymorphism within sequences compared to using amplicon size only (as in traditional capillary electrophoresis-based genotyping) to differentiate alleles (Barbian et al., 2018; Curto et al., 2019; Darby et al., 2016; Neophytou et al., 2018; Tibihika et al., 2018). Size homoplasy, due to alleles identical by size but not by sequence, ranges from to 32% and 64% (Barbian et al., 2018; Darby et al., 2016; Šarhanová et al., 2018; Vartia et al., 2016). Here, we found high variability in size homoplasy ranging from 6% to 53% between the studied species in direct correlation to the different types of variation observed across species. SNP in the repeat motif or in the flanking region was the main source of size homoplasy which showed high prevalence in *A. ostoyae* (SNP represented 62.4% of the detected variation and 53% of size homoplasy) and to a lesser extent in *Quercus* sp. (SNP represented 16.8% of the detected variation and 35% of apparent homoplasy). Even for species with lower apparent polymorphism levels, such as *M. variegatipes* (6%), *S. salar* (11%) and *Alosa* sp. (16%), the increase in the observed number of alleles will substantially improve genotypic resolution power (Darby et al., 2016; De Barba

et al., 2016). Haplotype-based analysis that accounts for all linked variations across the whole sequence will make the most of the information available from sequence data (Barthe et al., 2012).

Finally, the ability to detect different sources of variation originating from several mutation mechanisms occurring at different rates provides renewed opportunities to study ecological and evolutionary events that occur at different timescales (Barthe et al., 2012; Ramakrishnan & Mountain, 2004). Combining information on linked microsatellites and SNP (into a system called SNPSTR (Mountain et al., 2002) or HapSTR (Hey, Won, Sivasundar, Nielsen, & Markert, 2004; Sorenson & Dacosta, 2011)) was demonstrated to be a promising approach thanks to the increased phylogenetic resolution offered by explicitly considering complementary mutation properties of the markers. While theoretical and analytical implications of these approaches have been derived (Hey et al., 2004; Payseur & Cutter, 2006; Ramakrishnan & Mountain, 2004), empirical applications remain scarce and restricted to a very small number of systems due to the previous difficulties encountered to generate such empirical data (Hey et al., 2004; Mountain et al., 2002). This early limitation does not hold anymore with the generalisation of sequence-based microsatellite genotyping, as proposed herein, and the new ability to analyse linked microsatellites and SNP as haplotype. In addition, the flexibility of coalescent programs to simulate linked loci of different types (e.g. fastsimcoal2 (Excoffier, Dupanloup, Huerta-Sánchez, Sousa, & Foll, 2013)) will authorize far more realistic simulation of mutation mechanisms specifically tailored to each marker system. Such improvement, especially when applied to tens of loci, would make simulation-based inference (Beaumont, Zhang, & Balding, 2002) more accurate over an extended timescale range even for complex evolutionary history scenarios (Mountain et al., 2002).

In conclusion, this study proposes an integrated approach to expedite the development of SSRseq protocol for non-model species and provides several recommendations to improve development efficiency. The two most important advices are to optimize marker selection and primer design for effective multiplex PCR amplification and sequence interpretability, and to use repeated individuals to assess the quality of the generated genotypic data. The ability of SSRseq to characterize SNP and indel present along the sequences, in addition to the targeted microsatellite, represents a new opportunity to produce empirical data to apply existing theoretical and statistical frameworks that integrate linked polymorphism with different mutation characteristics (Payseur & Cutter, 2006). Finally, the ease of parallel development for multiple species make these approach convenient to develop powerful multilocus datasets for comparative population and community genetics studies (Crutsinger, 2016), and to further investigate the functional implications (Bagshaw, 2017) and adaptive potential of microsatellite variation among natural populations (Xie et al., 2019).

Acknowledgments

This research was supported by grants to OL of the Agence de l'Eau Adour-Garonne (project 2017/3273), the Région Nouvelle-Aquitaine (project 2016-1R20602-00007239), the Agence Française pour la Biodiversité (project 2016–18 A13) and INRA. A. A. benefited from a travel grant from the Department of Agronomy ISVSA of the University Batna 1 Hadja Lakhdar. We thank François Meurgey for the sampling of *Melipona sp.* individuals. Technical developments and sequencing were performed at the Genome Transcriptome Facility of Bordeaux (Grants from Investissements d'Avenir, Convention attributive d'aide EquipEx Xyloforest ANR-10-EQPX-16-01).

References:

Anvar, S. Y., Van Der Gaag, K. J., Van Der Heijden, J. W. F., Veltrop, M. H. A. M., Vossen, R. H. A. M., De Leeuw, R. H., … Laros, J. F. J. (2014). TSSV: A tool for characterization of complex allelic variants in pure and mixed genomes. *Bioinformatics*, *30*(12), 1651–1659. doi:10.1093/bioinformatics/btu068

Aykanat, T., Lindqvist, M., Pritchard, V. L., & Primmer, C. R. (2016). From population genomics to conservation and management: A workflow for targeted analysis of markers identified using genome-wide approaches in Atlantic salmon Salmo salar. *Journal of Fish Biology*, 1–22. doi:10.1111/jfb.13149

Bacles, C. F. E., Bouchard, C., Lange, F., Manicki, A., Tentelier, C., & Lepais, O. (2018). Estimating the effective number of breeders from single parr samples for conservation monitoring of wild populations of Atlantic salmon Salmo salar. *Journal of Fish Biology*, *92*(3), 699–726. doi:10.1111/jfb.13537

Bagshaw, A. T. M. (2017). Functional mechanisms of microsatellite DNA in eukaryotic genomes. *Genome Biology and Evolution*, *9*(9), 2428–2443. doi:10.1093/gbe/evx164

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., … Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, *3*(10), e3376. doi:10.1371/journal.pone.0003376

Barbian, H. J., Connell, A. J., Avitto, A. N., Russell, R. M., Smith, A. G., Gundlapally, M. S., … Hahn, B. H. (2018). CHIIMP: An automated high-throughput microsatellite genotyping platform reveals greater allelic diversity in wild chimpanzees. *Ecology and Evolution*, *8*, 7946–7963. doi:10.1002/ece3.4302

Barthe, S., Gugerli, F., Barkley, N. A., Maggia, L., Cardi, C., & Scotti, I. (2012). Always look on both sides: phylogenetic information conveyed by simple sequence repeat allele sequences. *PLoS ONE*, *7*(7), e40699. doi:10.1371/journal.pone.0040699

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate bayesian computation in population genetics. *Genetics*, *162*(4), 2025–2035.

Blacket, M. J., Robin, C., Good, R. T., Lee, S. F., & Miller, A. D. (2012). Universal primers for fluorescent labelling of PCR fragments – an efficient and cost-effective approach to genotyping by fluorescence. *Molecular Ecology*, *12*(3), 456–463. doi:10.1111/j.1755-0998.2011.03104.x

Bradbury, I. R., Wringe, B. F., Watson, B., Paterson, I., Horne, J., Beiko, R., … Bentzen, P. (2018). Genotyping-by-sequencing of genome-wide microsatellite loci reveals fine-scale harvest composition in a coastal Atlantic salmon fishery. *Evolutionary Applications*, *11*(6), 918–930. doi:10.1111/eva.12606

Brown, S. S., Chen, Y.-W., Wang, M., Clipson, A., Ochoa, E., & Du, M.-Q. (2017). PrimerPooler: automated primer pooling to prepare library for targeted sequencing. *Biology Methods and Protocols*, *2*(1), 1–10. doi:10.1093/biomethods/bpx006

Campbell, N. R., Harmon, S., & Narum, S. R. (2015). Genotyping-in-Thousands by sequencing (GT-

seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources*, *15*, 855–867. doi:10.1111/1755-0998.12357

Castoe, T. A., Poole, A. W., Gu, W., Jason de Koning, A. P., Daza, J. M., Smith, E. N., & Pollock, D. D. (2010). Rapid identification of thousands of copperhead snake (Agkistrodon contortrix) microsatellite loci from modest amounts of 454 shotgun genome sequence. *Molecular Ecology Resources*, *10*(2), 341–347. doi:10.1111/j.1755-0998.2009.02750.x

Chen, K., Zhou, Y. xun, Li, K., Qi, L. xin, Zhang, Q. fei, Wang, M. chun, & Xiao, J. hua. (2016). A novel three-round multiplex PCR for SNP genotyping with next generation sequencing. *Analytical and Bioanalytical Chemistry*, 1–7. doi:10.1007/s00216-016-9536-6

Crutsinger, G. M. (2016). A community genetics perspective: Opportunities for the coming decade. *New Phytologist*, *210*(1), 65–70. doi:10.1111/nph.13537

Curto, M., Winter, S., Seiter, A., Schmid, L., Scheicher, K., Barthel, L. M. F., … Meimberg, H. (2019). Application of a SSR-GBS marker system on investigation of European Hedgehog species and their hybrid zone dynamics. *Ecology and Evolution*, *9*(5), 2814–2832. doi:10.1002/ece3.4960

Darby, B. J., Erickson, S. F., Hervey, S. D., & Ellis-Felege, S. N. (2016). Digital fragment analysis of short tandem repeats by high-throughput amplicon sequencing. *Ecology and Evolution*, *6*(13), 4502–4512. doi:10.1002/ece3.2221

De Barba, M., Miquel, C., Lobréaux, S., Quenette, P. Y., Swenson, J. E., & Taberlet, P. (2016). High-throughput microsatellite genotyping in ecology: Improved accuracy, efficiency, standardization and success with low-quantity and degraded DNA. *Molecular Ecology Resources*, *17*(3), 492–507. doi:10.1111/1755-0998.12594

Delord, C., Lassalle, G., Oger, A., Barloy, D., Coutellec, M.-A., Delcamp, A., … Petit, E. J. (2018). A cost-and-time effective procedure to develop SNP markers for multiple species: A support for community genetics. *Methods in Ecology and Evolution*, *9*(9), 1959–1974. doi:10.1111/2041-210X.13034

Durand, J., Bodenes, C., Chancerel, E., Frigerio, J. M., Vendramin, G., Sebastiani, F., … Plomion, C. (2010). A fast and cost-effective approach to develop and map EST-SSR markers: oak as a case study. *BMC Genomics*, *11*(1). doi:10.1186/1471-2164-11-570

Ellis, J. S., Gilbey, J., Armstrong, A., Balstad, T., Cauwelier, E., Cherbonnel, C., … Stevens, J. R. (2011). Microsatellite standardization and evaluation of genotyping error in a large multi-partner research programme for conservation of Atlantic salmon (Salmo salar L.). *Genetica*, *139*(3), 353–367. doi:10.1007/s10709-011-9554-4

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*(9), 1591–1604. doi:10.1046/j.1365-294X.2002.01576.x

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, *9*(10), e1003905. doi:10.1371/journal.pgen.1003905

Farrell, E. D., Carlsson, J. E. L., Carlsson, J., & Farrell, E. D. (2016). Next Gen Pop Gen : implementing a high-throughput approach to population genetics in boarfish (Capros aper). *Royal Society Open Science*, *3*(12), 160651. doi:http://dx.doi.org/10.1098/rsos.160651

Garvin, M. R., Saitoh, K., & Gharrett, A. J. (2010). Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources*, *10*(6), 915–934. doi:10.1111/j.1755-0998.2010.02891.x

Gauthey, Z., Freychet, M., Manicki, A., Herman, A., Lepais, O., Panserat, S., … Labonne, J. (2015). The concentration of plasma metabolites varies throughout reproduction and affects offspring number in wild brown trout (Salmo trutta). *Comparative Biochemistry and Physiology. Part A, Molecular & Integrative Physiology*, *184*, 90–96. doi:10.1016/j.cbpa.2015.01.025

Gilbey, J., Verspoor, E., McLay, A., & Houlihan, D. (2004). A microsatellite linkage map for Atlantic salmon (*Salmo salar*). *Animal Genetics*, *35*(2), 98–105. doi:10.1111/j.1365-2052.2004.01091.x

Guichoux, E., Lagache, L., Wagner, S., Chaumeil, P., Léger, P., Lepais, O., … Petit, R. J. (2011). Current trends in microsatellite genotyping. *Molecular Ecology Resources*, *11*(4), 591–611.

doi:10.1111/j.1755-0998.2011.03014.x
Gymrek, M. (2017). A genomic view of short tandem repeats. *Current Opinion in Genetics & Development*, *44*, 9–16. doi:10.1016/j.gde.2017.01.012
Haasl, R. J., & Payseur, B. A. (2011). Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. *Heredity*, *106*(1), 158–171. doi:10.1038/hdy.2010.21
Hannan, A. J. (2018). Tandem repeats mediating genetic plasticity in health and disease. *Nature Reviews Genetics*, *19*(5), 286–298. doi:10.1038/nrg.2017.115
Harrison, H. B., Saenz-Agudelo, P., Planes, S., Jones, G. P., & Berumen, M. L. (2013). Relative accuracy of three common methods of parentage analysis in natural populations. *Molecular Ecology*, *22*(4), 1158–1170. doi:10.1111/mec.12138
Hey, J., Won, Y.-J., Sivasundar, A., Nielsen, R., & Markert, J. A. (2004). Using nuclear haplotypes with microsatellites to study gene flow between recently separated Cichlid species. *Molecular Ecology*, *13*(4), 909–919. doi:10.1046/j.1365-294X.2003.02031.x
Hoffman, J. I., & Amos, W. (2005). Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology*, *14*(2), 599–612. doi:10.1111/j.1365-294X.2004.02419.x
Holleley, C. E., & Geerts, P. G. (2009). Multiplex Manager 1.0: a cross-platform computer program that plans and optimizes multiplex PCR. *BioTechniques*, *46*(7), 511–517. doi:10.2144/000113156
Hoogenboom, J., de Knijff, P., Laros, J. F. J., de Leeuw, R. H., van der Gaag, K. J., & Sijen, T. (2016). FDSTools: A software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise. *Forensic Science International: Genetics*, *27*, 27–40. doi:10.1016/j.fsigen.2016.11.007
Kampfer, S., Lexer, C., Glössl, J., & Steinkellner, H. (1998). Characterization of (GA)n microsatellite loci from Quercus robur. *Hereditas*, *129*(2), 183–186. doi:10.1111/j.1601-5223.1998.00183.x
King, T. L., Eackles, M. S., & Letcher, B. H. (2005). Microsatellite DNA markers for the study of Atlantic salmon (Salmo salar) kinship, population structure, and mixed-fishery analyses. *Molecular Ecology Notes*, *5*(1), 130–132. doi:10.1111/j.1471-8286.2005.00860.x
Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. doi:10.1038/nmeth.1923
Lepais, O., & Bacles, C. F. E. (2011). Comparison of random and SSR-enriched shotgun pyrosequencing for microsatellite discovery and single multiplex PCR optimization in Acacia harpophylla F. Muell. Ex Benth. *Molecular Ecology Resources*, 711–724. doi:10.1111/j.1755-0998.2011.03002.x
Lepais, O., Manicki, A., Glise, S., Buoro, M., & Bardonnet, A. (2017). Genetic architecture of threshold reaction norms for male alternative reproductive tactics in Atlantic salmon (Salmo salar L.). *Scientific Reports*, *7*(January), 43552. doi:10.1038/srep43552
Malausa, T., Gilles, A., Meglécz, E., Blanquart, H., Duthoy, S., Costedoat, C., … Martin, J.-F. (2011). High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries. *Molecular Ecology Resources*, *11*(4), 638–644. doi:10.1111/j.1755-0998.2011.02992.x
Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380. doi:10.1038/nature03959
Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10. doi:10.14806/ej.17.1.200
Meglécz, E., Costedoat, C., Dubut, V., Gilles, A., Malausa, T., Pech, N., & Martin, J.-F. (2010). QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects. *Bioinformatics*, *26*(3), 403–4. doi:10.1093/bioinformatics/btp670
Meglécz, E., Pech, N., Gilles, A., Dubut, V., Hingamp, P., Trilles, A., … Martin, J.-F. (2014). QDD version 3.1: a user-friendly computer program for microsatellite selection and primer design revisited: experimental validation of variables determining genotyping success rate. *Molecular Ecology*

Resources, *14*(6), 1302–1313. doi:10.1111/1755-0998.12271
Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, *17*(2), 240–8. doi:10.1101/gr.5681207
Moran, P., Teel, D. J., LaHood, E. S., Drake, J., & Kalinowski, S. (2006). Standardising multi-laboratory microsatellite data in Pacific salmon: an historical view of the future. *Ecology of Freshwater Fish*, *15*(4), 597–605. doi:10.1111/j.1600-0633.2006.00201.x
Mountain, J. L., Knight, A., Jobin, M., Gignoux, C., Miller, A., Lin, A. A., & Underhill, P. A. (2002). SNPSTRs: Empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes. *Genome Research*, *12*(11), 1766–1772. doi:10.1101/gr.238602
Neophytou, C., Torutaeva, E., Winter, S., Meimberg, H., Hasenauer, H., & Curto, M. (2018). Analysis of microsatellite loci in tree of heaven (Ailanthus altissima (Mill.) Swingle) using SSR-GBS. *Tree Genetics and Genomes*, *14*(6). doi:10.1007/s11295-018-1295-4
Nguyen-Dumont, T., Pope, B. J., Hammet, F., Southey, M. C., & Park, D. J. (2013). A high-plex PCR approach for massively parallel sequencing. *BioTechniques*, *55*(2), 69–74. doi:10.2144/000114052
O'Reilly, P., & Wright, J. M. (1995). The evolving technology of DNA fingerprinting and its application to fisheries and aquaculture. *Journal of Fish Biology*, *47*(sa), 29–55. doi:10.1111/j.1095-8649.1995.tb06042.x
Ozaki, A., Sakamoto, T., Khoo, S., Nakamura, K., Coimbra, M. R. M., Akutsu, T., & Okamoto, N. (2001). Quantitative trait loci (QTLs) associated with resistance/susceptibility to infectious pancreatic necrosis virus (IPNV) in rainbow trout (Oncorhynchus mykiss). *Molecular Genetics and Genomics*, *265*(1), 23–31. doi:10.1007/s004380000392
Paterson, S., Piertney, S. B., Knox, D., Gilbey, J., & Verspoor, E. (2004). Characterization and PCR multiplexing of novel highly variable tetranucleotide Atlantic salmon (Salmo salar L.) microsatellites. *Molecular Ecology Notes*, *4*(2), 160–162. doi:10.1111/j.1471-8286.2004.00598.x
Payseur, B. A., & Cutter, A. D. (2006). Integrating patterns of polymorphism at SNPs and STRs. *Trends in Genetics*, *22*(8), 424–429. doi:10.1016/j.tig.2006.06.009
Pimentel, J. S. M., Carmo, A. O., Rosse, I. C., Martins, A. P. V., Ludwig, S., Facchin, S., … Kalapothakis, E. (2018). High-throughput sequencing strategy for microsatellite genotyping using neotropical fish as a model. *Frontiers in Genetics*, *9*, 73. doi:10.3389/fgene.2018.00073
Plomion, C., Aury, J. M., Amselem, J., Leroy, T., Murat, F., Duplessis, S., … Salse, J. (2018). Oak genome reveals facets of long lifespan. *Nature Plants*, *4*(7), 440–452. doi:10.1038/s41477-018-0172-3
Pompanon, F., Bonin, A., Bellemain, E., & Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. *Nature Reviews. Genetics*, *6*(11), 847–59. doi:10.1038/nrg1707
Prospero, S., Lung-Escarmant, B., & Dutech, C. (2008). Genetic structure of an expanding Armillaria root rot fungus (Armillaria ostoyae) population in a managed pine forest in southwestern France. *Molecular Ecology*, *17*(14), 3366–3378. doi:10.1111/j.1365-294X.2007.03829.x
Quinlan, A. R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current Protocols in Bioinformatics*, *47*(1), 11.12.1-11.12.34. doi:10.1002/0471250953.bi1112s47
Ramakrishnan, U., & Mountain, J. L. (2004). Precision and accuracy of divergence time estimates from STR and SNPSTR variation. *Molecular Biology and Evolution*, *21*(10), 1960–1971. doi:10.1093/molbev/msh212
Rexroad, C. E., Coleman, R. L., Martin, A. M., Hershberger, W. K., & Killefer, J. (2001). Thirty-five polymorphic microsatellite markers for rainbow trout (Oncorhynchus mykiss). *Animal Genetics*, *32*(5), 317–319. doi:10.1046/j.1365-2052.2001.0730b.x
Rougemont, Q., Besnard, A. L., Baglinière, J. L., & Launey, S. (2015). Characterization of thirteen new microsatellite markers for allis shad (Alosa alosa) and twaite shad (Alosa fallax). *Conservation Genetics Resources*, *7*(1), 259–261. doi:10.1007/s12686-014-0352-z
Sadd, B. M., Schaack, S., Zhao, X., Sun, C., & Su, L. (2018). Tandem repeats contribute to coding

sequence variation in bumblebees (Hymenoptera: Apidae). *Genome Biology and Evolution*, *10*(12), 3176–3187. doi:10.1093/gbe/evy244

Šarhanová, P., Pfanzelt, S., Brandt, R., Himmelbach, A., & Blattner, F. R. (2018). SSR-seq: Genotyping of microsatellites using next-generation sequencing reveals higher level of polymorphism as compared to traditional fragment size scoring. *Ecology and Evolution*, *8*(22), 10817–10833. doi:10.1002/ece3.4533

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–41. doi:10.1128/AEM.01541-09

Schlötterer, C. (2004). The evolution of molecular markers — just a matter of fashion? *Nature Reviews Genetics*, *5*(1), 63–69. doi:10.1038/nrg1249

Seeb, J. E., Carvalho, G., Hauser, L., Naish, K., Roberts, S., & Seeb, L. W. (2011). Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources*, *11*, 1–8. doi:10.1111/j.1755-0998.2010.02979.x

Sipos, G., Prasanna, A. N., Walter, M. C., O'Connor, E., Bálint, B., Krizsán, K., … Nagy, L. G. (2017). Genome expansion and lineage-specific genetic innovations in the forest pathogenic fungi Armillaria. *Nature Ecology & Evolution*, *1*(12), 1931–1941. doi:10.1038/s41559-017-0347-8

Slettan, A., Olsaker, I., & Lie, Ø. (1996). Polymorphic Atlantic salmon, Salmo salar L., microsatellites at the SSOSL438, SSOSL439 and SSOSL444 loci. *Animal Genetics*, *27*(1), 57–58. doi:10.1111/j.1365-2052.1996.tb01180.x

Sorenson, M. D., & Dacosta, J. M. (2011). Genotyping HapSTR loci: phase determination from direct sequencing of PCR products. *Molecular Ecology Resources*, *11*(6), 1068–1075. doi:10.1111/j.1755-0998.2011.03036.x

Steinkellner, H., Fluch, S., Turetschek, E., Lexer, C., Streiff, R., Kremer, A., … Glössl, J. (1997). Identification and characterization of (GA/CT)n microsatellite loci from Quercus petraea. *Plant Molecular Biology*, *33*(6), 1093–1096. doi:10.1023/A:1005736722794

Suez, M., Behdenna, A., Brouillet, S., Graça, P., Higuet, D., & Achaz, G. (2016). MicNeSs: Genotyping microsatellite loci from a collection of (NGS) reads. *Molecular Ecology Resources*, *16*(2), 524–533. doi:10.1111/1755-0998.12467

Thorsen, J., Zhu, B., Frengen, E., Osoegawa, K., de Jong, P. J., Koop, B. F., … Høyheim, B. (2005). A highly redundant BAC library of Atlantic salmon (Salmo salar): an important tool for salmon projects. *BMC Genomics*, *6*(1), 50. doi:10.1186/1471-2164-6-50

Tibihika, P. D., Curto, M., Dornstauder-Schrammel, E., Winter, S., Alemayehu, E., Waidbacher, H., & Meimberg, H. (2018). Application of microsatellite genotyping by sequencing (SSR-GBS) to measure genetic diversity of the East African Oreochromis niloticus. *Conservation Genetics*, 1–16. doi:10.1007/s10592-018-1136-x

Vartia, S., Villanueva-Cañas, J. L., Finarelli, J., Farrell, E. D., Collins, P. C., Hughes, G. M., … Carlsson, J. (2016). A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding. *Royal Society Open Science*, *3*(1), 150565. doi:10.1098/rsos.150565

Vasemägi, A., Nilsson, J., & Primmer, C. R. (2005). Expressed sequence tag-linked microsatellites as a source of gene-associated polymorphisms for detecting signatures of divergent selection in Atlantic salmon (Salmo salar L.). *Molecular Biology and Evolution*, *22*(4), 1067–76. doi:10.1093/molbev/msi093

Viard, F., Franck, P., Dubois, M. P., Estoup, A., & Jarne, P. (1998). Variation of microsatellite size homoplasy across electromorphs, loci, and populations in three invertebrate species. *Journal of Molecular Evolution*, *47*(1), 42–51.

Viruel, J., Haguenauer, A., Juin, M., Mirleau, F., Bouteiller, D., Boudagher-Kharrat, M., … Baumel, A. (2018). Advances in genotyping microsatellite markers through sequencing and consequences of scoring methods for Ceratonia siliqua (Leguminosae). *Applications in Plant Sciences*, *6*(12), e01201. doi:10.1002/aps3.1201

Willems, T., Gymrek, M., Highnam, G., 1000 Genomes Project Consortium, T. 1000 G. P., Mittelman,

D., & Erlich, Y. (2014). The landscape of human STR variation. *Genome Research*, *24*(11), 1894–904. doi:10.1101/gr.177774.114

Xie, K. T., Wang, G., Thompson, A. C., Wucherpfennig, J. I., Reimchen, T. E., MacColl, A. D. C., … Kingsley, D. M. (2019). DNA fragility in the parallel evolution of pelvic reduction in stickleback fish. *Science*, *363*(6422), 81–84. doi:10.1126/science.aan1425

Yano, A., Nicol, B., Jouanno, E., Quillet, E., Fostier, A., Guyomard, R., & Guigen, Y. (2013). The sexually dimorphic on the Y-chromosome gene (sdY) is a conserved male-specific Y-chromosome sequence in many salmonids. *Evolutionary Applications*, *6*(3), 486–496. doi:10.1111/eva.12032

Zhan, L., Paterson, I. G., Fraser, B. A., Watson, B., Bradbury, I. R., Nadukkalam Ravindran, P., … Bentzen, P. (2017). megasat: Automated inference of microsatellite genotypes from sequence data. *Molecular Ecology Resources*, *17*(2), 247–256. doi:10.1111/1755-0998.12561

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*(5), 614–620. doi:10.1093/bioinformatics/btt593

Data Accessibility Statement

- DNA sequences: sequence reads from SSRseq run were submitted to the European Nucleotide Archive SRA under the accession number PRJEB31908 (*Alosa* sp.), PRJEB31909 (*A. ostoyae*), PRJEB31910 (*M. variegatipes*), PRJEB31913 (*Quercus* sp.) and PRJEB31914 (*S. salar*). Random shotgun sequencing reads from one *M. variegatipes* individual for microsatellite discovery was submitted under the accession number ERR3255838.
- Output files from the genotyping pipeline for each species will be submitted to Dryad upon acceptance.
- Pipeline scripts for automated sequence-based microsatellite genotyping with documentation and example files are available at Data Inra https://doi.org/10.15454/HBXKVA.

Authors Contributions

OL designed the study and developed the data analysis pipeline. OL, EC, CB, FS and EG developed the laboratory protocols. AM, LT, AA and CFEB contributed to laboratory analyses. EC, LT and CFEB contributed to data analysis. CD, FD and SL contributed to sampling and preliminary data collection. OL wrote the manuscript which was revised critically and approved by all authors.

Supporting / Supplemental information

Table S1: Full details about the loci used or developed in this study (including species, locus name, primer sequences, characteristics, and reference)

Figure S1: Effect of sequence coverage on SSRseq data quality for *Alosa* species.

Supporting Materials and Methods: detailed explanation of FDSTools parameters used in this study.