Multi-individual Microsatellite identification: a multiple genome approach to microsatellite design (MiMi)

Fox, G, Preziosi, RF, Antwis, RE, Benavides-Serrato, M, Combe, FJ, Edwin Harris, W, Hartley, IR, Kitchener, AC, de Kort, SR, Nekaris, A and Rowntree, JK

http://dx.doi.org/10.1111/1755-0998.13065

MR. GRAEME FOX (Orcid ID : 0000-0001-7980-6944)
DR. RACHAEL ELLEN ANTWIS (Orcid ID : 0000-0002-8849-8194)
DR. MILENA B SERRATO (Orcid ID : 0000-0002-1644-8673)
DR. JENNIFER K ROWNTREE (Orcid ID : 0000-0001-8249-8057)

Article type : Resource Article

Multi-individual Microsatellite identification: a multiple genome approach to microsatellite design (MiMi).

Graeme Fox[1], Richard F. Preziosi[1], Rachael E. Antwis[2], Milena Benavides-Serrato[1, 3], Fraser J. Combe[1, 4], W. Edwin Harris[1, 5], Ian R. Hartley[6], Andrew C. Kitchener[7], Selvino R. de Kort[1], Anne-Isola Nekaris[8], Jennifer K. Rowntree[1]*

[1] Ecology and Environment Research Centre, Department of Natural Sciences, Manchester Metropolitan University, Manchester, M1 5GD, UK.

[2] School of Environment and Life Sciences, University of Salford, Salford, M5 4WT, UK.

[3] Universidad Nacional de Colombia, Sede Caribe-CECIMAR Calle 25 #2-55, Playa Salguero, Colombia.

[4] Kansas State University, Division of Biology, Manhattan, KS, United States.

[5] Crop and Environment Sciences, Harper Adams University, Newport, TF10 8NB, UK.

[6] Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK.

[7] Department of Natural Sciences, National Museums Scotland, Chambers Street, Edinburgh, EH1 1JF, UK.

[8] Department of Social Sciences, Faculty of Humanities and Social Sciences, Oxford Brookes University, Oxford, OX3 0BP, UK.

*Correspondence: j.rowntree@mmu.ac.uk

Abstract

Bespoke microsatellite marker panels are increasingly affordable and tractable to researchers and conservationists. The rate of microsatellite discovery is very high within a shotgun genomic dataset, but extensive laboratory testing of markers is required for confirmation of amplification and polymorphism. By incorporating shotgun next-generation sequencing datasets from multiple individuals of the same species, we have developed a new method for the optimal design of microsatellite markers. This new tool allows us to increase the rate at which suitable candidate markers are selected by 58% in direct comparisons and facilitate an estimated 16% reduction in costs associated with producing a novel microsatellite panel. Our method enables the visualisation of each microsatellite locus in a multiple sequence alignment allowing several important quality checks to be made. Polymorphic loci can be identified and prioritised. Loci containing fragment-length-altering mutations in the flanking regions, which may invalidate assumptions regarding the model of evolution underlying variation at the microsatellite, can be avoided. Priming regions containing point mutations can be detected and avoided, helping to reduce sample-site-marker specificity arising from genetic isolation, and the likelihood of null alleles occurring. We demonstrate the utility of this new approach in two species: an echinoderm and a bird. Our method makes a valuable contribution towards minimising genotyping errors and reducing costs associated with developing a novel marker panel. The Python script to perform our method of multi-individual microsatellite identification (MiMi) is freely available from GitHub (https://github.com/graemefox/mimi).

Keywords Microsatellite design, High-throughput sequencing, Short Tandem Repeat (STR), in silico quality control, Polymorphic loci detection, Cost-effective marker development.

Introduction

Microsatellites, short tandem repeats (STRs) or short simple repeats (SSRs), are exceptionally polymorphic repetitive regions of DNA found throughout the genomes of both eukaryotic and prokaryotic species (Bhargava & Fuentes, 2010; Rose & Falush, 1998). High rates of polymorphism, along with co-dominance and Mendelian inheritance, make them ideal markers for use in studies of population genetics (Abdul-Muneer, 2014; Goldstein & Pollock, 1997). Microsatellites have been the most popular choice of genetic

marker for several decades in ecology, conservation and evolutionary research, and are extensively used in contemporary studies of population genetics, parentage and kinship identification, evolutionary processes and genetic mapping (Vieira, Santini, Diniz, & de Munhoz, 2016; Ribout et al., 2019). Although single nucleotide polymorphism (SNP) markers have become increasingly popular markers for population genetics, microsatellites remain a common choice due to well-documented methodologies, ease of application, low equipment demands and well-developed statistical analyses. Furthermore, there remain scenarios where SNPs are not practical for use, or microsatellites are preferred (Zhan et al. 2016). For example, the management of captive populations has benefited enormously by the inclusion of genetic information (Fox et al., 2018; Witzenberger & Hochkirch, 2011), which must be continually updated as small numbers of new individuals are added to collections or produced through mating. In these cases, it is impractical to perform repeated SNP analyses on small numbers of samples due to the expense associated with next-generation sequencing (NGS) to acquire high coverage SNPs. Conversely, once a microsatellite panel has been developed, additional individuals can be genotyped using the existing markers very quickly, and at very low cost (Puckett, 2016). Where non-invasive sampling methods are required, for example because a species is of conservation concern (e.g. Fox et al., 2018), it may prove to be impossible to acquire sufficient high molecular weight DNA to perform NGS for SNP genotyping. In contrast, microsatellite analysis is forgiving of low DNA template input, and many contaminants that may disrupt NGS library preparation can simply be diluted out prior to amplification. A simple literature search in Google Scholar indicated the publication of approximately 2000 new microsatellite marker panels in 2018, suggesting that microsatellites are still very popular genetic markers, and we predict they will continue to be used extensively in conservation and ecology well into the future.

Ecological and conservation studies are often focused upon non-model species for which genetic markers are not available. The combination of affordable NGS and freely available bioinformatics tools can be used to identify tens of thousands of potential markers in a matter of days. Where probes were once used to target repeat regions of genetic code (Bloor, Barker, Watts, Noyes, & Kemp, 2001), shotgun genome sequencing does not require any prior knowledge of the genome, and is considered a non-targeted approach (Davey et al., 2011). Instead, random fragments of genomic DNA are sequenced, a fraction of which include SSRs within the length of the sequencing read. Free, open source software packages are available to detect SSRs and design suitable PCR primers to amplify the appropriate region of the genome; often referred to as the "seq-to-SSR" approach (Castoe et al., 2015; Griffiths et al., 2016). These developments, and the increasing availability of NGS technology globally, brings microsatellite marker discovery within the reach of ever more research

laboratories as the cost-per-base of NGS continues to decrease (Koboldt, Steinberg, Larson, Wilson, & Mardis, 2013; McPherson, 2014), even for applied, species-focused conservation research with limited funding. Thus, the development of bespoke microsatellite marker panels has become commonplace.

The use of microsatellite markers is reliant upon variation in PCR product fragment length, and therefore microsatellites must be amplifiable by PCR, and must contain fragment length altering polymorphisms within the repetitive stretch of SSR sequence. Despite improvements delivered by NGS, the optimisation of a bespoke microsatellite panel remains a time consuming and costly process, largely because the primer pair for each potential marker still requires manual laboratory confirmation of both successful amplification and the presence of multiple alleles at each locus (Bloor et al., 2001). Typically, the development of a microsatellite marker is performed through the discovery of a microsatellite locus in a single individual, followed by analysis of the locus in several more individuals to test for consistent amplification and variation in PCR fragment size (Abdelkrim, Robertson, Stanton & Gemmell, 2009). The main contributors to the cost of developing a panel of microsatellite markers are the NGS reagents, PCR reagents, PCR oligos, capillary electrophoresis, size standards and staff time. Improvements that enable reductions in cost or time associated with marker development will contribute to microsatellite markers becoming more widely available to ecological and conservation researchers.

Here we present a new conceptual approach to microsatellite marker design, demonstrated with a new bioinformatics technique applied to seq-to-SSR workflows. This technique is designed to improve the rate at which loci that are identified can be successfully amplified by PCR and produce informative genotype data. The innovation in our approach is the incorporation of information from the genomes of multiple individuals. This allows the *in silico* detection of polymorphic loci and the detection of several other important characteristics of a putative microsatellite marker, which are only detectable through multiple genome analysis. We demonstrate that this method reduces the number of markers that must be tested for polymorphism in the laboratory, and achieves an improved rate of successful marker development. Furthermore, our methods also minimise factors known to increase allelic dropout and invalidate genotyping results based upon molecular weight of PCR fragments. We refer to this technique as Multi-individual Microsatellite identification (MiMi). Here, we develop microsatellite markers using MiMi in two species: the green sea urchin (*Psammechinus miliaris*) and the Eurasian blue tit (*Cyanistes caeruleus*). For comparison, we also present the success rates of microsatellite development in *P. miliaris* and *C. caeruleus*, and in two other species (*Tragelaphus eurycerus isaaci* and *Nycticebus pygmaeus)*, which were designed using a traditional microsatellite design method (Castoe et al. 2015; Griffiths et al. 2016). The results from the successful development of each panel of markers, combined

with our refined bioinformatics method, provide a strong case for the utility of the MiMi concept and the value to microsatellite marker development.

Materials and Methods

DNA Extraction and Sequencing

Prior to DNA extraction, all samples (Table S1) were stored in 100% ethanol and stored at 4°C. Genomic DNA was extracted from samples using the DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) or the E.Z.N.A. Mollusc DNA Kit (Omega bio-tek, Georgia, USA) (Table S2). High quality and high molecular weight genomic DNA (determined by gel electrophoresis) was diluted to 2.5ng/µL and sequenced on an Illumina MiSeq (Illumina, San Diego, USA), using the Illumina Nextera XT library preparation reagents (Illumina, San Diego, USA). Paired-end, shotgun genomic DNA sequencing was performed using the Illumina MiSeq Reagent Kit v2/v3. MiMi analysis was conducted on eight individuals of each species (*P. miliaris* and *C. caeruleus*) which were indexed, pooled and sequenced on a flowcell, per species. For traditional microsatellite detection, single samples of each species (*T. eurycerus isaaci* and *N. pygmaeus*) were individually indexed, pooled and sequenced along with other species not used in this study (Table S2). Both methods were not tested for all species, due to these microsatellite markers being designed for active research projects that progressed beyond marker development as the MiMi method was being developed and iterated upon.

MiMi Microsatellite Detection Methodology

Microsatellite markers were initially designed in data from each sample using the pal_finder (Castoe et al. 2015) workflow of Griffiths et al. (2016); a traditional design method using the data of a single individual. A novel quality control procedure was developed for those datasets in which multiple individuals of the same species were sequenced (two species) with the aim of identifying polymorphic loci, filtering out primer pairs containing point mutations within the priming regions, and avoiding other potential issues with a locus including non-specific primer binding and insertion/deletion mutations in the flanking regions. Eight individuals per species were sequenced and the data pertaining to each individual were first passed separately through the traditional design method. The eight individual output files then become the input for the novel method: Multi-individual Microsatellite identification (MiMi). MiMi takes the primer sequences developed in each individual and checks for their presence in the data of every other individual. Primer pairs for which the forward primer appeared in more than 33% of the individuals were selected and all reads containing the exact primer sequence

compiled into an MSA file with the FASTA format. The MSA files were aligned using the MUSCLE alignment algorithm (Edgar, 2004) and putative loci automatically filtered to remove monomorphic loci, low quality 'gapped' alignments and loci containing sequence mutations within the primer binding sites. Loci passing all filters are retained as high quality loci and loci passing some filters but lacking enough information to confidently pass all filters are retained as good quality loci. Both high quality and good quality loci are each ranked by the size range in alleles detected. A log file is produced detailing loci which have been removed by each filter. A Python script implementing the MiMi tool is available to download and run from https://github.com/graemefox/mimi.

Optimisation of Potential Markers

Primer pairs developed under either design method were tested in 5µL reactions using the Type-it Microsatellite PCR Kit (Qiagen, Hilden, Germany) using the standard protocol and thermal cycling parameters (5 mins at 95℃, 25-28*{30s at 95℃, 90s at 60℃, 30s at 72℃}, 30 mins at 60℃). Only a single annealing temperature (60°C) was tested, as Primer3 (Koressaar & Remm, 2007; Untergasser et al. 2012) which is used during the traditional marker design process (Castoe et al. 2015; Griffiths et al. 2016), had been configured specifically for these PCR reagents and a primary goal of this method was to avoid time consuming annealing temperature optimisation. A marker was given successful amplification status if clean PCR products were clearly visible on a 2% agarose gel in the 100-1000bp range for six or more individuals out of eight tested. PCR products were analysed using a 2% agarose electrophoresis gel. Fluorescent dyes (6-FAM, TAMRA, HEX, PET) were added to PCR products using a universal tail technique (Blacket, Robin, Good, Lee & Miller, 2012). Fragment length was determined using an ABI 3730 DNA Analyzer capillary sequencer (ThermoFisher Scientific, Carlsbad, USA) with GeneScan 500 LIZ dye Size Standard (ThermoFisher Scientific, Carlsbad, USA) and analysed using Genemapper 5.0 software (ThermoFisher Scientific, Carlsbad, USA). We define an informative marker as one that produces clearly interpretable electropherogram traces after capillary electrophoresis and is polymorphic in terms of PCR fragment length between multiple individuals.

Results

Of the markers which passed each set of quality controls, we were able to optimise amplifiable and informative markers at a rate of 47.9% using the traditional design method, and 86.6% using MiMi. Comparisons between average rates of successful amplification and production of informative loci for each marker design method demonstrated a marked increase in both measures when MiMi was applied. In *P. miliaris* and *C. caeruleus*, markers were designed using both the traditional methodology and the MiMi methodology. A direct comparison between these two methods shows a very notable increase in both the rate of amplification success and the rate of development of informative markers (Figure 1). In two further species, (*T. eurycerus isaaci* and *N. pygmaeus*), markers were designed using only the traditional methodology. Rates of success for these species are presented as further evidence of a baseline of microsatellite design against which the MiMi method can be compared (Table 1). Unsuitable markers were removed at each filtering stage, reducing hundreds of thousands of possible markers designed by pal_finder, to a fewer than a hundred identified as high- or good-quality using MiMi (Table 2). Where MiMi was applied, the number of individuals sharing each common primer sequence ranged from three to seven (Figure 2). In the two example MiMi datasets presented here, 5% of potential loci were detected in sufficient individuals to allow further analysis by MiMi.

Automatic analysis of MSA files allowed the identification and removal of loci with mutations within the primer binding sites (Figure S1a and Figure S1b) and loci showing very low alignment quality. Low alignment quality is indicative of a locus potentially containing fragment length altering polymorphisms (insertions/deletions) between the primer binding sites but outside the microsatellite locus itself (Figure S1c) or non-specific primer binding. Monomorphic loci were also removed (Figure S1d and Figure S1e). Of the markers which MiMi detected in multiple individuals, we were able to discount 79.3% of potential loci as unsuitable for microsatellite analysis (Table 3). High quality loci (those which exclusively showed evidence of positive characteristics) were detected at a rate of 4.5%, and good quality loci (those which did not show any evidence of negative characteristics, but did not have enough data to confidently pass all filters) were detected at an average rate of 16.1%.

Whilst the full MiMi method requires more data than the traditional approach detailed here (we recommend a minimum of eight individuals to be sequenced using the capacity of an entire MiSeq flowcell, although fewer samples are possible), the reduction in time spent in the lab, and associated savings, justifies the larger outlay in initial sequencing costs. A recent Illumina MiSeq run cost approximately $2330, and using MiMi we recorded that 90% of the primer pairs chosen to be tested were successfully developed as informative

microsatellite markers (Table 1, dataset #2). Using the traditional method, sequencing costs were less, as only a fraction (12.5%) of the capacity of a MiSeq sequencing flowcell was required, but only 38% of primer pairs tested were ultimately found to be informative markers (Table 1, dataset #5). The reduction in time and laboratory expense associated with investing in "failed markers" (inconsistent amplification / non-polymorphic loci) ultimately results in a net saving when using MiMi. Based on our estimated rate of successful marker development, a project to develop a panel of 20 optimised markers over a two-week period using the MiMi methodology would cost less than using the traditional methodology over a four week period (16% reduction in total cost, 50% reduction in staff costs only, 19% increase in reagent costs only; see Table S3 and Table S4). The most significant savings will be in researcher time spent screening loci, which was approximately 50% less using MiMi.

Description of Output Files

The outputs from the MiMi method are two tab separated tables containing details of the loci that have passed the quality control processes, a log file detailing which loci were removed under which quality-control conditions, and a per-locus MSA file in the FASTA format. The output tables each give the following information for each locus: forward primer sequence; reverse primer sequence; number of alleles at the locus; number of individuals in which the locus was sequenced in the dataset; a description of the alleles found (the repeat motif and the number of repeats), and the predicted size range of amplicons produced using the PCR primers. The file "MiMi_output_all_loci.txt" gives details of every loci which MiMi was able to detect in multiple individuals (above the user-defined threshold) and "MiMi_output_filtered_loci.txt" gives just those loci which were able to pass all quality control filters as either high- or good quality. The log file details which loci were removed under which quality control conditions. Examples of the "MiMi_output_filtered_loci.txt" files resulting from the the MiMi analysis of *C. caeruleus* (dataset #1) and *P. miliaris* (dataset #2) are presented in tables S5(a) and S5(b) respectively. Three MSA files per locus are created: one containing the raw sequences from the input data that were found to contain the locus within the length of the read (ending .fastq); one containing these reads after alignment by MUSCLE (ending .aln) and one containing aligned reads trimmed to the position of the forward primer (ending .trimmed). The main section of the MSA file name is the forward primer sequence of the locus.

Discussion

MiMi has proved to be a fast, cost effective approach to identification and characterisation of microsatellite markers using genomic sequence data from multiple individuals. The application of a microsatellite-picking tool such as pal_finder typically results in tens of thousands of potential loci, and therefore it makes logical sense to attempt to apply *in silico* marker optimisation methods over laboratory optimisation, to increase the efficiency in identifying informative loci. MiMi is the first tool, to our knowledge, that allows this range of important characteristics to be observed at the marker design stage (but see Nichols, Conroy, Kasinadhuni, Lamont & Ogbourne, 2018). In a direct comparison between the traditional and MiMi methods, we show that the application of MiMi resulted in a 58% increase in the rate of identification of informative microsatellite markers, facilitating a 16% reduction in costs associated with the development of a microsatellite marker panel. To provide a baseline value of microsatellite design success, we also provide success rates for two species which only used the traditional methodology. Although not a true comparison, it appears that MiMi can be expected to produce amplifiable, informative markers at a consistently higher rate than the traditional methodology, facilitating an increase from ~57-60% (datasets #3 and #4) to ~80-90% (datasets #1 and #2). We feel certain that an increase of this order of magnitude, and the reduction in costs associated with the testing of markers which ultimately fail, fully justify the slight increase in sequencing costs associated with MiMi.

The incorporation of multiple genomes and construction of an MSA for each microsatellite locus allows several important quality checks to be made of each locus and facilitates notable increases in both the rate of successful amplification by PCR, and the development of informative markers. Nucleotide polymorphisms and INDEL mutations within the forward or reverse primer binding site can cause issues with inconsistent or failed PCR amplification, potentially resulting in allelic dropout (Silva, Torrezan, Brianese, Stabellini & Carraro, 2017), and can also lead to an increase in the frequency of null alleles (Rico et al., 2017). Allelic dropout can present a significant problem during microsatellite analysis, causing decreased estimates of observed heterozygosity and increased estimates of inbreeding in the population (Wang, Schroeder & Rosenberg, 2012). Two main causes of allelic dropout have been shown: sequence variation at a primer binding site (Silva et al., 2017) and PCR product size (particularly problematic for markers with large repeat counts), (Sefc, Payne & Sorenson, 2003). Through the construction of each MSA we were able to use MiMi to automatically confirm that primer-binding sites show strong sequence conservation, albeit in only a small subset of samples, thus minimising the likelihood that a putative marker would exhibit an elevated rate of allelic

dropout caused by mis-priming. Confirmation of sequence conservation in at least one primer-binding site improved the rate at which we were able to amplify loci successfully. If possible, genomes of individuals from a range of putative populations should be included in the MiMi analysis to minimise null allele bias towards a particular sub population (Oosterhout, Weetman & Hutchinson, 2005). Analysis of each microsatellite locus in an MSA also allows visualisation of the number of motif repeats, and automatic prioritisation of loci where variation is seen among samples. Rejecting monomorphic loci through MiMi produced an increase in the rate at which we were able to develop informative markers, compared to our own previous experience using other methods, and rates stated in the literature (Zhan et al. 2016). Additionally, MiMi automatically assesses the likelihood of the presence of multiple primer binding sites in the host genome by collating all sequences containing a common primer sequence. Where sequences containing the primer sequence produce low-overlap alignments, it is indicative that the corresponding primer binding site occurs in multiple locations across the genome, and thus that particular primer pair should be avoided to reduce cross-amplification.

Statistical models based upon a particular model of evolution at the microsatellite locus (the stepwise mutation model, for example) rely upon the assumption that the source of variation in fragment size is polymorphism in the number of repeats in the SSR (Dieringer and Schlötterer, 2003). The presence of other fragment length altering mutations between the primer binding sites (excluding the microsatellite itself) is indistinguishable by capillary electrophoresis from 'true' variation at the microsatellite locus (Angers & Bernatchez, 1997; Grimaldi & Crouau-Roy, 1997; Stágel et al., 2009). Markers with fragment-length-altering mutations outside the microsatellite locus, potentially invalidate the assumptions of a number of models of microsatellite evolution, and are therefore avoided in our protocol.

Whilst MiMi does not allow one to state with certainty that a putative marker will not exhibit any of the negative characteristics described (allelic dropout, null alleles arising from population differentiation, non-variable microsatellite loci, cross amplification or invalidation of assumptions of evolutionary model) when comprehensively characterised in a much larger number of samples, the opportunity to identify loci that do exhibit them, and subsequently remove them from analyses, is nevertheless valuable.

Variation in the rate at which loci were removed under each quality control category shows the importance of making each check, and that marker development in different taxa may perform differently from one another. In both examples of the application of MiMi here, we were able to remove undesirable loci, that failed at least one quality check. Considering the total markers designed and filtered in both species, we were able to pass many loci (mean: 20.7%) that did not show evidence of these negative characteristics in the eight tested samples.

The success of MiMi is dependent upon the sequence coverage achieved in each sequencing run. Very low sequence coverage would likely result in relatively little overlap in the sequences of each individual, and therefore few loci passing the MiMi filter. The development of a new marker panel is very often performed in non-model species of specialised interest and it is likely that the genome size will be unknown and sequence coverage incalculable (Shikano, Ramadevi, Shimada & Merilä, 2010). MiMi was successfully implemented in the two species tested here (with estimated coverage of 0.57X and 1.20X), suggesting that the method is suitable for genomic datasets with relatively low sequence coverage (Ekblom and Wolf, 2014). The proportion of individuals in which a primer must be detected is user definable, with a minimum of two individuals required for MiMi to provide useful information. Where loci were successfully detected in multiple individuals, we found a negative correlation between the number of potential markers and the frequency at which loci were found in multiple datasets. These frequencies are dependent upon the genome size, and the microsatellite richness of the genome, of the species of interest. Where estimates of genome coverage are approximately 1X or below, removal of duplicate primers/loci from the dataset of each individual is recommended (implemented automatically in the Griffiths et al. (2016) workflow) as coverage of >1X of a locus in a single individual does not contribute any additional information to the MiMi process. However, where estimated coverage is significantly >1X, their removal may result in the dismissal of an increased frequency of otherwise useful loci that appear multiple times in the sequence data as a result of the random nature of shotgun sequencing (Bouck, Miller, Gorrell, Muzny and Gibbs, 1998). In the event of a low number of markers ultimately being returned, the filter that removes loci appearing more than once in the data can easily be disabled at the web interface of the Griffiths et al. (2016) tool. In this case, multiple reads containing the primer sequence from the same biological sample will appear alongside each other in the output MSA, allowing the user to assess the reads as “shotgun duplicates” (ie. multiple sequence reads covering the same genomic region of an individual, by chance).

MiMi makes several important assumptions of the characteristics of microsatellite loci investigated in a small number of samples, and infers these are representative of the loci in the wider population. However, this is not always expected to be true (Goldstein, Linares, Cavalli-Sforza & Feldman, 1995) and the removal of otherwise useful markers, under the limiting assumptions of the MiMi quality control process, is likely to happen. For example, SSRs that do not show any variation in number of repeats in the sequence data are removed, but these loci may show variation in the wider population. The ethos behind the MiMi method is to select markers for which we have the most information, rather than seeking to discover as many markers as possible. Given the large numbers of potential markers we derived from the MiMi process, we do not consider the removal of potentially useful markers as a major disadvantage, and these markers can always be added back if needed.

Loci that do show allelic variation are ranked by the range size of the microsatellite repeat number (Goldstein & Schlötterer, 1999), with the assumption that the loci with the largest differences are most likely to be informative markers. A large range in the number of repeats implies that the variation seen at the locus is less likely to be the result of an amplification or sequencing error (Hosseinzadeh-Colagar, Haghighatnia, Amiri, Mohadjerani & Tafrihi, 2016) but rather is representative of a true, variable microsatellite locus. We conclude that under the assumptions we identify here, the rate and efficiency of informative microsatellite discovery are greatly increased using high-throughput sequencing data in comparison to traditional microsatellite library discovery methods, but the robustness of MiMi should be tested in additional species.

We recommend that eight unrelated individuals are sequenced for MiMi processing for optimal capture of markers exhibiting multiple alleles at microsatellite loci. Whilst it is impossible to state an optimum figure for universal use, due to varying allelic richness in species and populations (Bashalkhanov, Pandey & Rajora, 2009), in our experience, eight samples represents an acceptable balance between depth of sequencing coverage and allele rarefaction (Hale, Burg & Steeves, 2012). In species where it is not feasible to source eight samples, related or not, due to their extreme scarcity, MiMi is still applicable. MiMi will function beneficially on any number of samples >1, whether related or unrelated. Furthermore, species with extremely large genomes may not perform well due to the limitations of sequencer capacity and the requirement for approximately 1X genome sequence coverage to be achieved. Our method has been tested on Illumina MiSeq data only, but will function on paired-end data, in the .fastq format, from any sequencing platform, should additional depth of coverage be required. It is important to note that we are not attempting to detect all, or even most alleles present at a locus. Detecting the presence of multiple alleles (>1) is sufficient to enable MiMi processing. Other influencing

factors, such as the sampling of related individuals or populations experiencing low genetic diversity due to historical population bottlenecks, may impact the allelic richness of the samples and therefore the ability of MiMi to detect multiple alleles (Price & Hadfield, 2014).

Methods of genotyping microsatellites by high-throughput sequencing are a promising development and avoid many of the ambiguities inherent in genotyping by capillary electrophoresis (Zhan et al. 2016; Shin et al. 2017). Determination of accurate genotypes by these methods enables many of the additional tests required of a microsatellite marker (tests for linkage disequilibrium, frequency of null alleles, for example) to be carried out using NGS data alone. We envisage that large scale microsatellite studies be performed using two NGS runs: the first using MiMi to discover potentially informative microsatellites; and a second using a high-throughput genotyping method to genotype all experimental samples in one go (De Barba et al. 2016).

Acknowledgements

With thanks to the Genomic Technologies Core Facility of the University of Manchester for their expertise and services. *P. miliaris* samples were collected by Simon Exley of Queen's University Belfast. Funding for this PhD research comes from Manchester Metropolitan University. ACK thanks the Negaunee Foundation for their generous support of a curatorial preparator who extracted the samples of *N. pygmaeus*.

References

Abdelkrim, J., Robertson, B. C., Stanton, J.L., and Gemmell, N. J. (2009) Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. *Biotechniques,* 46(3), 185-92. doi: 10.2144/000113084

Abdul-Muneer, P.M. (2014) Application of microsatellite markers in conservation genetics and fisheries management: recent advances in population structure analysis and conservation strategies. *Genetics Research International,* 2014, 691759. doi: 10.1155/2014/691759

Angers, B., and Bernatchez, L. (1997) Complex evolution of a salmonid microsatellite locus and its consequences in inferring allelic divergence from size information. *Molecular Biology and Evolution,* 14(3), 230-8. doi: 10.1093/oxfordjournals.molbev.a025759

Bashalkhanov, S., Pandey, M. and Rajora, O.P. (2009) A simple method for estimating genetic diversity in large populations from finite sample sizes. *BMC Genetics*. 10(84). doi: 10.1186/1471-2156-10-84

Bhargava, A. and Fuentes, F.F. (2010) Mutational dynamics of microsatellites. *Molecular Biotechnol*ogy, 44(3), 250-66. doi: 10.1007/s12033-009-9230-4.

Blacket, M.J., Robin, C., Good, R.T., Lee, S.F. and Miller, A.D. (2012) Universal primers for fluorescent labelling of PCR fragments--an efficient and cost-effective approach to genotyping by fluorescence. *Molecular Ecology Resources,* 12(3), 456-63. doi: 10.1111/j.1755-0998.2011.03104.x.

Bloor, P.A., Barker, F.S., Watts, P.C., Noyes, H.A., and Kemp, S.J. (2001) Microsatellite libraries by enrichment. [online] Available from: http://www.genomics.liv.ac.uk/animal/MICROSAT.PDF [Accessed: 11/10/2018]

Bolger, A.M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 1;30(15), 2114-20. doi: 10.1093/bioinformatics/btu170.

Bouck, J., Miller, W., Gorrell, J.H., Muzny, D. and Gibbs, R.A. (1998) Analysis of the quality and utility of random shotgun sequencing at low redundancies. *Genome Research.* 8(10), 1074-1084. doi: 10.1101/gr.8.10.1074.

Castoe, T.A., Poole, A.W., de Koning, A.P.J., Jones, K.L., Tomback, D.F., Oyler-McCance, S.J., Fike, J.A., Lance, S.L., Streicher, J.W., Smith, E.N., and Pollock, D.D. (2015) Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. *PLoS One,* 2015;10(8):e0136465.

Combe, F.J., Taylor-Cox, E., Fox, G., Sandri, T., Davis, N., Jones, M.J., Cain, B., Mallon, D. and Harris, E.W. (2018) Rapid isolation and characterization of microsatellites in the critically endangered Mountain Bongo (*Tragelaphus eurycerus isaaci*). *Journal of Genetics,* 97(2), 549-553. doi: https://doi.org/10.1007/s12041-018-0922-z

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M. and Blaxter, M.L. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics.* 12, 499-510.

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

Dieringer, D. and Schlötterer, C. (2003) Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome Resources,* 13(10), 2242-2251. doi: 10.1101/gr.1416703

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research,* 19;32(5), 1792-7. doi: 10.1093/nar/gkh340

Ekblom, R. and Wolf, J.B.W. (2014) A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications* 7(9), 1026-42. doi: 10.1111/eva.12178.

Fox, G., Darolti, I., Hibbitt, JD., Preziosi, R.F., Fitzpatrick, J.L., and Rowntree, J.K. (2018) Genetic assessment of ex situ populations to aid species conservation and maintain heterozygosity in non-model species. *Journal of Zoo and Aquarium Research,* 6(2), 50-56. doi: https://doi.org/10.19227/jzar.v6i2.299

Goldstein, D.B., Linares, A.R., Cavalli-Sforza, L.L. and Feldman, M.W. (1995). *Genetics,* 139(1), 463 – 471.

Goldstein, D.B., and Pollock, D.D. (1997) Launching microsatellites: a review of mutation processes and methods of phylogenetic inference. *Journal of Heredity,* 88(5), 335-342. doi: https://doi.org/10.1093/oxfordjournals.jhered.a023114

Goldstein, D.B., and Schlötterer, C. (1999) Microsatellites: evolution & applications. Oxford, United Kingdom. Oxford University Press.

Griffiths, S.M., Fox, G., Briggs, P.J., Donaldson, I.J., Hood, S., Richardson, P., Leaver, G.W., Truelove, N.K., and Preziosi, R.F. (2016) A Galaxy-based bioinformatics pipeline for optimised, streamlined microsatellite development from Illumina next-generation sequencing data. *Conservation Genetics Resources,* 8(4), 481-486.

Grimaldi, M.C., and Crouau-Roy, B. (1997) Microsatellite allelic homoplasy due to variable flanking sequences. *Journal of Molecular Evolution,* 44(3), 336-40.

Hale, M.L., Burg, T.M. and Steeves, T.E. (2012) Sampling for Microsatellite-Based Population Genetic Studies: 25 to 30 Individuals per Population Is Enough to Accurately Estimate Allele Frequencies. *PLoS ONE.* doi: https://doi.org/10.1371/journal.pone.0045170

Hosseinzadeh-Colagar, A., Haghighatnia, M.J., Amiri, Z., Mohadjerani, M. and Tafrihi, M. (2016) Microsatellite (SSR) amplification by PCR usually led to polymorphic bands: Evidence which shows replication slippage occurs in extend or nascent DNA strands. *Molecular Biology Research Communications,* 5(3), 167-174.

Koboldt, D.C., Steinberg, K.M., Larson, D.E., Wilson, R.K., and Mardis, E.R. (2013) The next-generation sequencing revolution and its impact on genomics. *Cell,* 155(1), 27-38. doi: 10.1016/j.cell.2013.09.006.

Koressaar, T., and Remm, M. (2007) Enhancements and modifications of primer design program Primer3. *Bioinformatics,* 15;23(10), 1289-91. doi: 10.1093/bioinformatics/btm091

Larsson, A. (2014) AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics.* 30(22), 3276-8. doi: https://doi.org/10.1093/bioinformatics/btu531

McPherson, J.D. (2014) A defining decade in DNA sequencing. *Nature Methods,* 11, 1003-5.

Nichols, J., Conroy, G.C., Kasinadhuni, N., Lomont, R.W. and Ogbourne, S.M. (2018) In silico detection of polymorphic microsatellites in the endangered Isis tamarind, Alectryon ramiflorus (Sapindaceae). *Applications in Plant Sciences.* 6(11). doi: https://doi.org/10.1002/aps3.1196

Oosterhout, C.V., Weetman, D. and Hutchinson, W.F. (2005) Estimation and adjustment of microsatellite null alleles in nonequilibrium populations. *Molecular Ecology Notes.* 6(1). doi: https://doi.org/10.1111/j.1471-8286.2005.01082.x

Price M.R. and Hadfield M.G. (2014) Population genetics and the effects of a severe bottleneck in an ex situ population of critically endangered Hawaiian tree snails. *PLoS ONE* 9: e114377. doi:10.1371/journal.pone.0114377.

Puckett, E.E. (2016) Variability in total project and per sample genotyping costs under varying study designs including with microsatellites or SNPs to answer conservation genetic questions. *Conservation Genetics Resources.* 9(2), 289-304.

Ribout, C., Villers, A., Ruault, S., Bretagnolle, V., Picard, D., Monceau, K. and Gauffre, B. (2019) Fine-scale genetic structure in a high dispersal capacity raptor, the Montagu's harrier (*Circus pygargus*), revealed by a set

of novel microsatellite loci. *Genetica.* 147(1), 69-78. *d*oi: https://doi.org/10.1007/s10709-019-00053-7

Rico, C., Cuesta, J.A., Drake, P., Macpherson, E., Bernatchez, L., and Marie, A.D. (2017) Null alleles are ubiquitous at microsatellite loci in the Wedge Clam (*Donax trunculus). PeerJ,* 5: e3188. doi: 10.7717/peerj.3188

Rose, O., and Falush, D. (1998) A threshold size for microsatellite expansion. *Molecular Biology and Evolution,* 15(5), 613-5. doi: 10.1093/oxfordjournals.molbev.a025964

Sefc, K.M., Payne, R.B. and Sorenson, M.D. (2003) Microsatellite Amplification from Museum Feather Samples: Effects of Fragment Size and Template Concentration on Genotyping Errors. *The Auk.* 120(4), 982-989

Silva, F.C., Torrezan, G.T., Brianese, R.C., Stabellini, R. and Carraro, D.M. (2017) Pitfalls in genetic testing: a case of a SNP in primer-annealing region leading to allele dropout in *BRCA1*. *Molecular Genetics and Genomic Medicine,* 5(4), 443-447. doi: 10.1002/mgg3.29

Shikano, T., Ramadevi, J., Shimada, Y. and Merilä, J. (2010) Utility of sequenced genomes for microsatellite marker development in non-model organisms: a case study of functionally important genes in nine-spined sticklebacks (*Pungitius pungitius*). *BMC Genomics.*11(334). doi: 10.1186/1471-2164-11-334

Shin G., Grimes, S.M., Lee, H.J., Lau, B.T., Xia, L.C. and Ji, H.P. (2017) CRISPR-Cas9-targeted fragmentation and selective sequencing enable massively parallel microsatellite analysis. *Nature Communications*. 8 (14291).

Stágel, A., Gyurján, I., Sasvári, Z., Lanteri, S., Ganal, M., and Nagy, I. (2009) Patterns of molecular evolution of microsatellite loci in pepper (*Capsicum spp.)* revealed by allele sequencing. *Plant Systematics and Evolution,* 281(1-4), 251-254

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., and Rozen, S.G. (2012) Primer3--new capabilities and interfaces. *Nucleic Acids Research,* 40(15):e115

Vieira, M.L.C., Santini, L., Diniz, A.L., and Munhoz, C.deF. (2016) Microsatellite markers: what they mean and why they are so useful. *Genetics and Molecular Biology,* 39(3), 312-328. doi: 10.1590/1678-4685-GMB-2016-0027

Wang, C., Schroeder, K.B., and Rosenberg, N.A. (2012) A Maximum-Likelihood Method to Correct for Allelic Dropout in Microsatellite Data with No Replicate Genotypes. *Genetics.* 192(2), 651-69. doi: 10.1534/genetics.112.139519

Witzenberger, K.A., and Hochkirch, A. (2011) Ex situ conservation genetics: a review of molecular studies on the genetic consequences of captive breeding programmes for endangered animal species. *Biodiversity and Conservation,* 20(9), 1843-1861. doi: 10.1007/s10531-011-0074-4.

Zhan, L., Paterson, I.G., Fraser, B.A., Watson, B., Bradbury, I.R., Ravindran, P.N., Reznick, D., Beiko, R.G. and Bentzen, P. (2016) MEGASAT: automated inference if microsatellite genotypes from sequence data. *Molecular Ecology Resources,* 17(2), 247-256. doi: https://doi.org/10.1111/1755-0998.12561.

Data Accessibility

The MiMi quality processing procedure is performed by an open-source Python script, freely available from https://github.com/graemefox/mimi. A small subset of example data is included at the repository. The pal_finder and pal_filter process required prior to MiMi is easily run and accessed via an online service hosted by the University of Manchester https://palfinder.ls.manchester.ac.uk/. Raw sequence reads are available from the N.C.B.I. BioProject and Sequence Read Archive (*C. caeruleus*: Accession PRJNA507250; *P. miliaris*: Accession PRJNA510714; *T. eurycerus isaaci*: Accession PRJNA509530; *N. pygamaeus*: Accession PRJNA509330).

Author Contributions

GF, RFP & JKR conceived the concept. GF wrote and tested the programme and performed the marker optimisation in *C. caeruleus*. GF, RFP & JKR verified the methods and the interpretation of the results. GF, RFP & JKR discussed the results and drafted the manuscript, with helpful comments and contributions from remaining authors to the final manuscript. RA assisted with DNA sequencing at the University of Salford. FC and WEH provided the *T. eurycerus isaaci* and *C. caeruleus* samples and associated sequence data, FC performed the marker optimisation in these species and FC ran the capillary sequencer. ACK and AN provided the *N. pygamaeus* sample. MBS provided the *P. miliaris* samples and sequence data, and performed the marker optimisation in this species. IH and SrdK provided the *C. cyanistes* samples.

Tables and Figures

Table 1. A summary of the design methods used in each species, including the dataset number (ID), species, treatment (T_x), number of individuals sequenced (N), number of PCR primers tested (Pp), number of PCR primers tested successfully amplifying in 75% of samples tested (Amp.), number of amplifiable PCR primers producing informative data after capillary electrophoresis (easily interpretable and polymorphic) (Inf), percentage of amplifiable primers which were informative (Inf / Amp), percentage of total primers tested which were informative (Inf / Pp) genome size estimate (C-val), raw sequence reads per sample (Reads), (mean and SD given where MiMi applied), estimated sequence coverage (Cov), literature reference and/or accession numbers of NGS data (REF / SRA) where applicable. All genome sizes were retrieved from the Animal Genome Size Database (www.genomesize.com) with the closest related species used. Panels of markers were developed in *P. miliaris* and *C. caeruleus* using both the traditional method (Castoe et al., 2015; Griffiths et al., 2016) and MiMi methods. The application of the MiMi quality control process produces higher rates of both amplification and production of informative markers in both these instances.

ID	Species	T_x	N	Pp	Amp	Inf	Inf / Amp	Inf / Pp	C val	Reads	Cov	REF / SRA
1	*C. caeruleus*	MiMi	8	10	10	8	80%	80%	1.47	8* 2,901,027, (STDEV +/- 878, 838)	1.20X	SRX5066864 to SRX5066869
2	*P. miliaris*	MiMi	8	20	19	18	95%	90%	1.30	8* 1,482,736, (STDEV +/- 280, 686)	0.57X	SRX5162614 to SRX5162621
3	*T. eurycerus isaaci*	Trad.	1	30	21	18	86%	60%	3.94	8,980,510	1.10X	Combe et al., 2018 / SRX5116712
4	*N. pygamaeus*	Trad.	1	30	26	17	65%	57%	3.58	5,309,686	0.74X	SRX5112421
5	*C. caeruleus*	Trad.	1	10	4	1	25%	10%	1.47	3,913,299	1.60X	SRX5066867
6	*P. miliaris*	Trad.	1	24	13	9	69%	38%	1.30	1,359,615	0.52X	SRX5162614

Table 2. The total number of potential microsatellite loci discovered using the traditional design methodology, retained after filtering with the Griffiths et al. (2016) method and retained after MiMi quality control processing.

Species	pal_finder loci	Griffiths et al. loci	MiMi loci
Cyanistes caeruleus	158,147	4,513 (2.9%)	302 (0.19%)
Psammechinus miliaris	469,047	5,657 (1.2%)	250 (0.05%)

Table 3. Potential loci are automatically filtered by the MiMi script. Loci are removed under the following conditions: Low quality alignments = loci rejected due to not meeting a minimum requirement for overall quality of alignment. This is indicative of multiple primer binding occurring in the host genome, and of size-altering INDEL mutations occurring in the flanking regions. Primer mutations = loci rejected due to SNP or INDEL mutations detected within the primer binding sites. Non-variable = loci rejected due to multiple reads spanning the microsatellite but no motif number variation present. High Quality = loci passed due to consistent forward and reverse primer sequences seen in multiple individuals, multiple reads spanning the microsatellite and variable motif number observed, no evidence of INDEL or multiple binding sites, Good Quality = identical criteria as 'High Quality', but alignment provided no information afforded relating to consistent reverse PCR primer or INDEL mutations,

ID	Species	Total	Low Quality Alignments	Primer mutations	Non-variable	High Quality	Good Quality
1	*Cyanistes caeruleus*	302	14 (4.6%)	7 (2.3%)	205 (67.9%)	13 (4.3%)	63 (20.9%)
2	*Psammechinus miliaris*	250	102 (40.8%)	9 (3.6%)	101 (40.4%)	12 (4.8%)	26 (10.4%)

Figure 1. Summary statistics showing the rate at which potential microsatellite markers were successfully amplified in the laboratory, and the rate at which they were discovered to be informative. Markers were designed using both methodologies in *P. miliaris* and *C. caeruleus*. Stated values are the average for each design method, in each measure of success (amplification rate and informative loci rate). Dashed lines show the standard deviations. The use of MiMi results in both an increase in the rate at which markers amplify and are informative, and also a reduction in the variability at each of these measures compared to the traditional workflow.

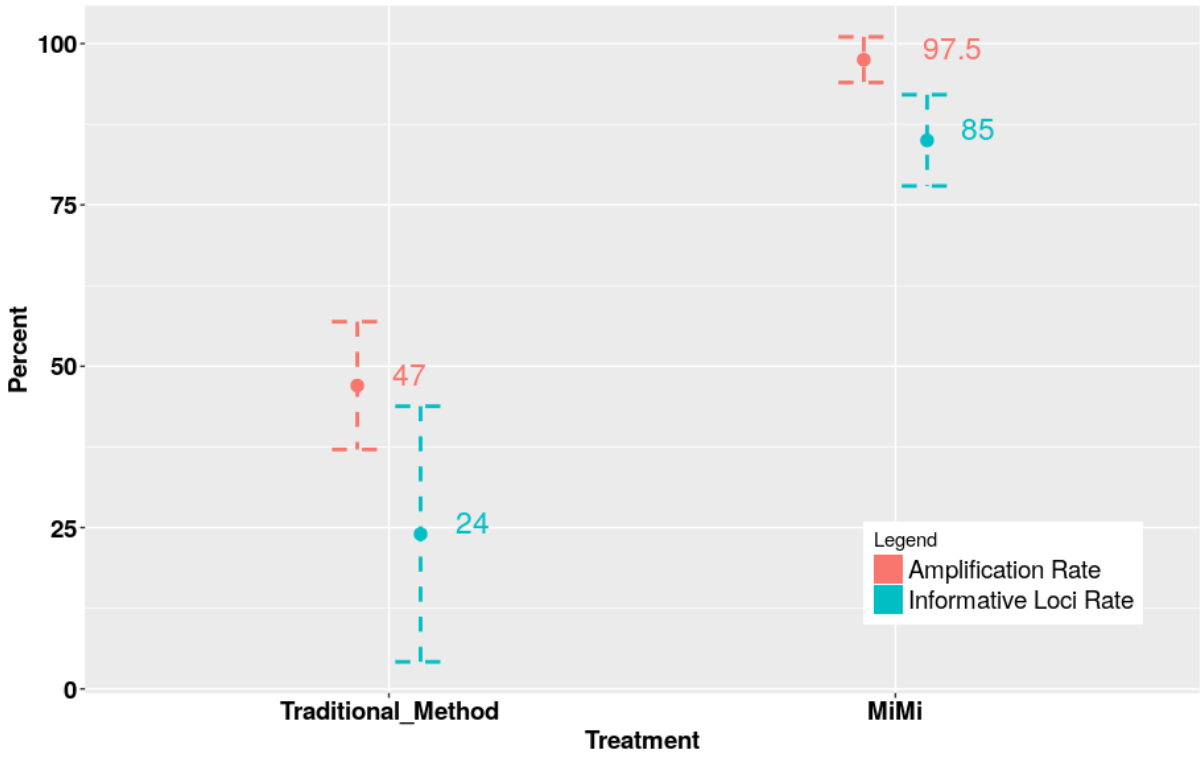

Figure 2. The MiMi tool was used to analyse 5,657 potential microsatellite loci discovered in *P. miliaris* sequence data and 4,513 discovered in *C. caeruleus.* Loci were filtered to just those which appeared in the sequence data of three or more individuals. The total number of loci which were successfully detected in multiple individuals, and in how many individuals they were detected is shown below. The bar labels are the absolute number of loci that were detected in each category (number of individuals).