
https://doi.org/10.1111/1755-0998.13065

http://repository.nms.ac.uk/2396

Deposited on: 27 September 2019
Multi-individual microsatellite identification: A multiple genome approach to microsatellite design (MiMi)

Graeme Fox1 | Richard F. Preziosi1 | Rachael E. Antwis2 | Milena Benavides-Serrato1,3 | Fraser J. Combe1,4 | W. Edwin Harris1,5 | Ian R. Hartley6 | Andrew C. Kitchener7 | Selvino R. de Kort1 | Anne-Isola Nekaris8 | Jennifer K. Rowntree1

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 data set, but extensive laboratory testing of markers is required for confirmation of amplification and polymorphism. By incorporating shotgun next-generation sequencing data sets 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
cost-effective marker development, high-throughput sequencing, in silico quality control, microsatellite design, polymorphic loci detection, short tandem repeat (STR)
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 codominance 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 (Ribout et al., 2019; Vieira, Santini, Diniz, & de Munhoz, 2016). 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 2,000 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 nontargeted 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
For traditional microsatellite detection, single samples of each species (T. eurycerus isaci 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.

2.2 | 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 data sets 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 nonspecific 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.

2.3 | Optimisation of potential markers

Primer pairs developed under either design method were tested in 5 μl reactions using the Type-it Microsatellite PCR Kit (Qiagen) using the standard protocol and thermal cycling parameters (5 min at 95°C, 25–28*30 s at 95°C, 90 s at 60°C, 30 s at 72°C, 30 min at 60°C). 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.
A summary of the design methods used in each species, including the data set number (ID), 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 (Inf), percentage of amplifiable primers which were informative (Inf/Pp), genome size estimate (C-val), raw sequence reads per sample (Reads), and estimated sequence coverage (Cov) for each species used. All genome sizes were retrieved from the Animal Genome Size Database (www.genomessize.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 method against which the MiMi methodology can be compared (Table 1).

<table>
<thead>
<tr>
<th>ID</th>
<th>Species</th>
<th>Treatment</th>
<th>N</th>
<th>Pp</th>
<th>Amp</th>
<th>Inf</th>
<th>Inf/Pp</th>
<th>C-val</th>
<th>Reads</th>
<th>Cov</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>C. caeruleus</em></td>
<td>MiMi</td>
<td>8</td>
<td>10</td>
<td>18</td>
<td>100</td>
<td>80%</td>
<td>2,901,027</td>
<td>8.2</td>
<td>1.20X</td>
</tr>
<tr>
<td>2</td>
<td><em>T. eurycerus isaaci</em></td>
<td>Trad.</td>
<td>1</td>
<td>30</td>
<td>21</td>
<td>71</td>
<td>3.94</td>
<td>8,980,510</td>
<td>1.10X</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td><em>T. p. eurycerus</em></td>
<td>Trad.</td>
<td>1</td>
<td>10</td>
<td>4</td>
<td>1</td>
<td>1.47</td>
<td>3,913,299</td>
<td>0.57X</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td><em>N. pygmaeus</em></td>
<td>Trad.</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>10</td>
<td>3.58</td>
<td>5,309,686</td>
<td>0.74X</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td><em>C. caeruleus</em></td>
<td>Trad.</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>100</td>
<td>3.91</td>
<td>3,913,299</td>
<td>1.60X</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td><em>P. miliaris</em></td>
<td>Trad.</td>
<td>1</td>
<td>24</td>
<td>13</td>
<td>55</td>
<td>1.30</td>
<td>1,339,515</td>
<td>0.32X</td>
<td></td>
</tr>
</tbody>
</table>

**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 data sets 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 (Figures S1a,b) 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 nonspecific primer binding. Monomorphic loci were also removed (Figures S1d,e). 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%. 

---

**TABLE 1** A summary of the design methods used in each species, including the data set number (ID), 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 (Inf), percentage of amplifiable primers which were informative (Inf/Pp), genome size estimate (C-val), raw sequence reads per sample (Reads), and estimated sequence coverage (Cov) for each species used. All genome sizes were retrieved from the Animal Genome Size Database (www.genomessize.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 method against which the MiMi methodology can be compared (Table 1).
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.

<table>
<thead>
<tr>
<th>Species</th>
<th>pal_finder loci</th>
<th>Griffiths et al. (2016) loci</th>
<th>MiMi loci</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyanistes caeruleus</td>
<td>158,147</td>
<td>4,513 (2.9%)</td>
<td>302 (0.19%)</td>
</tr>
<tr>
<td>Psammechinus miliaris</td>
<td>469,047</td>
<td>5,657 (1.2%)</td>
<td>250 (0.05%)</td>
</tr>
</tbody>
</table>

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 laboratory, and associated savings, justifies the larger outlay in initial sequencing costs. A recent Illumina MiSeq run cost approximately $2,330, and using MiMi we recorded that 90% of the primer pairs chosen to be tested were successfully developed as informative microsatellite markers (Table 1, data set No.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, data set No.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 Tables S3 and S4). The most significant savings will be in researcher time spent screening loci, which was approximately 50% less using MiMi.

3.1 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; 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 (data set No. 1) and P. miliaris (data set No. 2) are presented in Tables S5a,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.

4 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 amenable, informative markers at a consistently higher rate than the traditional methodology, facilitating an increase from ~57%–60% (data sets Nos. 3 and 4) to ~80%–90% (data sets Nos. 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 subpopulation (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 & 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; Stågel 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, nonvariable 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.

### TABLE 3

<table>
<thead>
<tr>
<th>ID</th>
<th>Species</th>
<th>Total</th>
<th>Low quality alignments</th>
<th>Primer mutations</th>
<th>Nonvariable</th>
<th>High quality</th>
<th>Good quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>Cyanistes caeruleus</em></td>
<td>302</td>
<td>14 (4.6%)</td>
<td>7 (2.3%)</td>
<td>205 (67.9%)</td>
<td>13 (4.3%)</td>
<td>63 (20.9%)</td>
</tr>
<tr>
<td>2</td>
<td><em>Psammechinus miliaris</em></td>
<td>250</td>
<td>102 (40.8%)</td>
<td>9 (3.6%)</td>
<td>101 (40.4%)</td>
<td>12 (4.8%)</td>
<td>26 (10.4%)</td>
</tr>
</tbody>
</table>
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 probably 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 data sets with relatively low sequence coverage (Ekblom & 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 data sets. 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 data set 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, & 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” (i.e., 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 (Hosseinizadeh-Colagar, Haghighatnia, Amiri, Mohadjerani, & Tafrhi, 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 (Shin et al., 2017; Zhan et al., 2016). 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.

AUTHOR CONTRIBUTIONS
G.F., R.F.P., & J.K.R. conceived the concept. G.F. wrote and tested the programme and performed the marker optimisation in C. caeruleus. G.F., R.F.P., & J.K.R. verified the methods and the interpretation of the results. G.F., R.F.P., & J.K.R. discussed the results and drafted the manuscript, with helpful comments and contributions from remaining authors to the final manuscript. R.A. assisted with DNA sequencing at the University of Salford. F.C., and W.E.H. provided the T. eurycerus isaaci and C. caeruleus samples and associated sequence data, F.C. performed the marker optimisation in these species and F.C. ran the capillary sequencer. A.C.K., and A.N. provided the N. pygmaeus sample. M.B.S. provided the P. miliaris samples and sequence data, and performed the marker optimisation in this species. I.H. and S.R.D.K. provided the C. cyanistes samples.

DATA AVAILABILITY STATEMENT
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.man.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. pygmaeus: Accession PRJNA509330).

REFERENCES


**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of the article.