Title

A Pipeline and Recommendations for Population and Individual Diagnostic SNP Selection in Non-Model Species.

Journal

Molecular Ecology Resources, 25(3)

Authors

Armstrong, Ellie
Li, Chenyang
Campana, Michael
et al.

Publication Date

2025-04-01

DOI

10.1111/1755-0998.14048

Molecular Ecology Resources

WILEY

MOLECULAR ECOLOGY RESOURCES

RESOURCE ARTICLE OPEN ACCESS

A Pipeline and Recommendations for Population and Individual Diagnostic SNP Selection in Non-Model Species

Ellie E. Armstrong[1,2] | Chenyang Li[3] | Michael G. Campana[4] | Tessa Ferrari[3] | Joanna L. Kelley[5] | Dmitri A. Petrov[6,7,8] | Katherine A. Solari[6] | Jazlyn A. Mooney[3]

[1]School of Biological Sciences, Washington State University, Pullman, Washington, USA | [2]Department of Evolution, Ecology and Organismal Biology, University of California, Riverside, Riverside, California, USA | [3]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, California, USA | [4]Smithsonian's National Zoo and Conservation Biology Institute, Washington, DC, USA | [5]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, California, USA | [6]Department of Biology, Stanford University, Stanford, California, USA | [7]Chan Zuckerberg BioHub, San Francisco, California, USA | [8]Program for Conservation Genomics, Center for Computational, Evolutionary, and Human Genomics, Stanford, California, USA

Correspondence: Jazlyn A. Mooney (jazlynmo@usc.edu)

Received: 9 July 2024 | **Revised:** 16 October 2024 | **Accepted:** 6 November 2024

Handling Editor: Luke Browne

Keywords: conservation genetics | individual ID | marker selection | population assignment | SNP panel

ABSTRACT

Despite substantial reductions in the cost of sequencing over the last decade, genetic panels remain relevant due to their cost-effectiveness and flexibility across a variety of sample types. In particular, single nucleotide polymorphism (SNP) panels are increasingly favoured for conservation applications. SNP panels are often used because of their adaptability, effectiveness with low-quality samples, and cost-efficiency for population monitoring and forensics. However, the selection of diagnostic SNPs for population assignment and individual identification can be challenging. The consequences of poor SNP selection are underpowered panels, inaccurate results, and monetary loss. Here, we develop a novel and user-friendly SNP selection pipeline (mPCRselect) that can be used to select SNPs for population assignment and/or individual identification. mPCRselect allows any researcher, who has sufficient SNP-level data, to design a successful and cost-effective SNP panel for a diploid species of conservation concern.

1 | Introduction

Whole-genome sequencing (WGS) approaches have increased in feasibility and popularity as sequencing costs have declined and computational tools have improved. For many species, however, sequencing whole genomes from a large number of individuals or using WGS for continuous population monitoring is still cost-prohibitive and computationally challenging. Well-established systems that rely heavily on genomic technologies, such as human (e.g. LaFramboise 2009) or agricultural genetics (e.g. Fan et al. 2010), frequently turn to single-nucleotide polymorphism (SNP) panels, arrays or other forms of reduced-representation sequencing, such as RADseq or ddRAD for population genetic studies (Hirsch et al. 2014; Scheben, Batley, and Edwards 2017).

SNP panels have been created for numerous non-model organisms including: Iberian lynx (Kleinman-Ruiz et al. 2017), pumas (Fitak et al. 2015), lampreys (Hess et al. 2015), cichlids (Ciezarek et al. 2022), lions (Bertola et al. 2022), tigers (Khan et al. 2022; Natesh et al. 2019), bison (Wehrenberg et al. 2024), canids (Parker et al. 2022) and many others. This increase in popularity has occurred primarily because panels are cost-effective

Ellie E. Armstrong and Chenyang Li contributed equally to this study.

https://doi.org/10.1111/1755-0998.14048

and flexible (Puckett 2017). SNP panels are often designed to run on certain technologies (e.g., Fluidigm 96.96 Dynamic Array, Agena Bioscience MassARRAY), which limits them to specific facilities or ultimately requires additional validation on other machines (Carroll et al. 2018). Multiplex PCR (mPCR) is an alternative and more flexible approach, which involves creating a primer pool that amplifies many SNPs simultaneously. However, mPCR presents its own obstacles. For example, it can be challenging to avoid issues in the primer pools, such as primer dimers or cross amplification. Conversely, because it is relatively straightforward to add indexes and adapters for many individuals and adapt primers to be compatible with a variety of sequencing technologies, mPCR also has considerable benefits. Methods such as GT-seq (Campbell, Harmon, and Narum 2015) have made advances in the design and pooling of primers around specific loci, but the selection of those loci is left to the user.

The selection of loci for a targeted panel is critical to ensure that panel aims (such as population assignment or sample identity) are accurately achieved. For example, it is possible to include SNPs that are not informative, which have the potential to introduce noise into downstream analyses. Further, when creating panels for multiple populations, SNPs that represent the genetic variation in one population may not be informative in the other due to allele frequency differences (Biddanda, Rice, and Novembre 2020). This result has led the human genetics community to create multiple panels over time to better represent non-European human diversity, improve imputation accuracy, and bolster detection of variants that are associated with complex traits (Bien et al. 2016).

Panel design often includes the goals of population assignment and individual identification. The first hurdle in marker selection is that the methods used to filter and select SNPs to achieve these two aims varies widely. Methods used for marker selection and filtering include but are not limited to, estimates of pi (π) or theta (θ), Hardy–Weinberg equilibrium (HWE), differentiation (F_{ST}), minor allele frequency (MAF), linkage disequilibrium (LD), in conjunction with various quality filters such as mapping and genotype quality (see e.g., Bertola et al. 2022; Ciezarek et al. 2022; Fitak et al. 2015; Hess et al. 2015; Kleinman-Ruiz et al. 2017; Natesh et al. 2019; Wehrenberg et al. 2024). Human geneticists previously developed approaches to select optimal genetic markers for ancestry (Balding and Nichols 1994; Baye et al. 2009; Galanter et al. 2012; Kidd et al. 2006; Manel, Gaggiotti, and Waples 2005; Rosenberg 2005; Rosenberg et al. 2003) and individual identification (Balding and Nichols 1994; Kidd et al. 2006; Pakstis et al. 2007, 2010). However, these insights are not always applied in non-model species. Significantly, a majority of the theory that guides SNP selection for population assignment and individual identification assumes that markers are in linkage equilibrium and segregating at an appreciable frequency within populations, making MAF and LD filtering most relevant during the SNP filtering stage (Kidd et al. 2006). Further, in human genetics, filtering for HWE is often used to ensure genotype quality since HWE outlier loci are often caused by sequencing errors. Whereas in conservation genetics, HWE filters have been shown to remove informative loci that are variable between populations (Chen, Cole, and Grond-Ginsbach 2017; Hemstrom et al. 2024; Pearman, Urban, and Alexander 2022), primarily because conservation projects have substantially lower sample sizes or unknown population history and are potentially structured. Given the numerous approaches that exist and the potential large variance of their success in non-human species, the selection of loci remains challenging for conservation-oriented projects, but ultimately the selection of SNPs must be guided by the desired end result of the panel (individual identification, population assignment, parentage or a combination therein).

Adding to this list of considerations for marker selection, researchers must also assess marker informativeness to build robust SNP panels. There are many different statistics that measure marker informativeness from the perspective of population assignment, including Fisher Information Content (FIC; [Pfaff et al. 2004]), Shannon Information Content (SIC; [Rosenberg et al. 2003]), F statistics (in particular, F_{ST}; [Wright 1951]), Informativeness for Assignment Measure (I_{n}; [Rosenberg et al. 2003]) and the Absolute Allele Frequency Differences (delta, δ; [Rosenberg et al. 2003]). Principal Component Analysis (PCA) has additionally been used as a tool to select SNPs for structure identification and assignment (Paschou et al. 2007). Fisher Information Content (FIC) is typically used in admixed populations and quantifies a marker's informativeness of the genetic contributions of ancestral populations. In contrast, Shannon Information Content (SIC) assesses the reduction in entropy (uncertainty) provided by a marker, reflecting its overall effectiveness in distinguishing between different ancestral source populations. F statistics (here we will focus on F_{ST}) quantifies genetic differentiation between source populations. The Informativeness for Assignment Measure (I_{n}) evaluates a marker's practical utility in assigning individuals to particular populations or groups while taking into account self-reported ancestry of a sampled individual. Absolute Allele Frequency Differences (delta, δ) measures the information content of a marker through quantifying the absolute allele frequency differences between different ancestral source populations. Lastly, PCA is a dimensionality reduction technique and is often used to visualise and infer population structure in genetic data by identifying the main axes of variation in a population-level dataset. Previous work has suggested that the two best methods of estimating marker informativeness for biallelic loci are I_{n} and F_{ST} (Ding et al. 2011), with I_{n} performing better for mixed ancestry populations. Studies have also shown that selecting markers which maximise F_{ST} perform better than selecting markers using PCA (Wilkinson et al. 2011). While these approaches aim to optimise SNP selection, they do not inform on the ability of particular SNPs to assign individuals to the populations of interest.

In addition to estimating individual marker informativeness, there are methods for determining which combinations of markers will yield the most effective panel based on their ability to accurately assign individuals to the populations of interest, such as f_{ORCA} (Rosenberg 2005). Broadly, f_{ORCA} is an assignment function that computes the Optimal Rate of Correct Assignment (ORCA) across a set of markers, with the goal of assigning an individual to the source population from which the individual's genotypes are most likely to have arisen from. This approach has rarely been implemented outside of human and agriculturally relevant species, such as salmon (Storer et al. 2012), sheep (Sottile et al. 2018) and crop species

(Morrell and Clegg 2007). In rare cases, the method has been applied to non-model systems, such as in the domestic cat and European wildcat (Oliveira et al. 2015). Despite its relevance to optimising the selection of small panels (which are desirable in the conservation sector), this method has seldom been applied in non-model species.

Here, we seek to optimise diagnostic SNP selection using F_{ST} and f_{ORCA}, specifically for the purposes of population assignment. We demonstrate the utility of this method of selection and assignment evaluation in humans, tigers and domestic dogs. We also briefly explore the overlap between selecting SNPs for population assignment and individual identification. Last, we present an accompanying pipeline, mPCRselect, that when provided with a variant call file (VCF) and predesignated populations uses a greedy algorithm to provide users with diagnostic SNPs suitable for population assignment and/or individual identification in the context of a multiplex PCR assay.

2 | Materials and Methods

2.1 | Simulated Genotype Data for Two Populations

We conducted forward-in-time simulations using SLiM 4.0.1 (Haller and Messer 2023) on a high-performance computing cluster, utilising Dell node models R440 and xl170 with core speeds of 2.1 Ghz. We simulated a burn-in with neutral dynamics for a population of 10,000 individuals, each with a uniform 100 Mb genomic segment, until all lineages in the ancestral trees were fully coalesced. After the burn-in, the ancestral population was instantaneously split into two subpopulations of 10,000 individuals each. We allowed no migration between subpopulations after divergence. Simulations were conducted with a neutral mutation rate of 1e−8 and a recombination rate of 1e−8. As the subpopulations diverged, we recorded individual genotypes every 200 generations until 2000 generations after the population split. Increments of 200 generations were selected based on theoretical estimates of F_{ST} from Nicholson et al. (2002), and resulted in the populations having an F_{ST} ranging [0.01, 0.1] with increments of 0.01.

Finally, we sub-sampled 100 individuals from each of the two populations and output a VCF for classification. The VCF was converted to a plink file with PLINK 2 (version 2.00a3.7LM; [Chang et al. 2015]). The plink file was filtered to 10,000 markers in linkage equilibrium with each other. Linkage pruning was achieved with the command '--indep-pairwise 500kb 0.2'.

2.2 | Simulated Genotype Data for Three Populations

Using the same mutation rate, recombination rate and saved burn-in state as in the two-population simulations, we conducted two additional sets of simulations of three diverging populations. In the first set, the ancestral population split into three populations, (A, B, and C) after the burn-in, each consisting of 10,000 individuals. In the second, the ancestral population split into two populations of 10,000 individuals after the burn-in, populations A and B. Then after 1000 generations (when A–B F_{ST} is expected to be 0.05) a third population of 10,000, population C, split off from population A. Again, as the subpopulations diverged, we recorded individual genotypes every 200 generations until 2000 generations after the A-B population split.

Once again, we sub-sampled 100 individuals from each of the three populations and output a VCF for classification. The VCF was converted to a plink file with PLINK 2 (version 2.00a3.7LM; [Chang et al. 2015]). The plink file was filtered to 10,000 markers in linkage equilibrium with each other. Linkage pruning was conducted with the command '--indep-pairwise 500kb 0.2'.

2.3 | Empirical Data

We applied marker selection methods (δ, F_{ST}, and PCA) to genomic data from three different species. We used whole-genome sequence data from humans in the 1000 Genomes Project (Dai, CDX; Puerto Rican, PUR; Luhya, LWK; Colombian, CLM; and Afro-Caribbean, ACB) (1000 Genomes Project Consortium et al. 2010); from tigers (Amur, Bengal, and Generic, [Armstrong et al. 2024]); and dogs (Labrador retriever and Yorkshire terriers, [Mooney et al. 2023]). Sample sizes were 88 humans from each population ($N = 440$ individuals total), 13 tigers from each of the Amur and Bengal subspecies and 13 Generic ($N = 39$ individuals total), and 100 individuals from each dog breed ($N = 200$ individuals total). Unrelated individuals from the 1000 Genomes Project were identified using the ped file (20130606_g1k.ped) that is provided with the hg19 data. For the tiger data, unrelated individuals were previously identified in Armstrong et al. (2024). For the dog data, unrelated individuals were previously identified in Mooney, Yohannes, and Lohmueller (2021). We used PLINK 2 (version 2.00a3.7LM; [Chang et al. 2015]) and filtered for a minor allele frequency that was at least 5% (common in the population) and markers that were in linkage equilibrium with each other. We used the command '--maf 0.05 --indep-pairwise 500kb 0.2' to accomplish this filtering. Then, we created a subset of 10,000 markers randomly sampled from across the genome using the remaining markers in linkage equilibrium.

2.4 | Population Assignment Performance Function

We assigned individuals to populations with a performance function, f_{ORCA} (Rosenberg et al. 2003), which uses the genotype of an individual and population-level allele frequencies of the selected marker for population assignment. For a detailed description of the f_{ORCA} approach, see Rosenberg (2005). Briefly, for each sample we calculated the probability that the individual originated from a given source population (k), then we assigned the individual to the population that had the highest probability of assignment.

2.5 | Marker Selection and Population Assignment

F_{ST} (Wright 1951) was used to quantify the genetic differentiation among subpopulations. The higher the F_{ST} value, the more pronounced the genetic differentiation. Various definitions and formulations of F_{ST} exist, and in this work, we use

$$F_{\mathrm{ST}} = \frac{\mathrm{var}(p)}{\bar{p} * \bar{q}}$$

to calculate F_{ST} values (Balding 2003). We calculated F_{ST} values for all of the markers and sorted the results in descending order. A greedy algorithm was applied to select the top M markers to compose a marker panel for individual classification.

Population-level allele frequencies were computed on the basis of the sampled individuals. We used f_{ORCA} to assign N individuals a population label computed with either M random markers or the M top markers. Accuracy was measured as the proportion of empirical individuals that were classified with the correct population label. We ran 20 replicates of each classification scenario in the simulated and empirical datasets.

2.6 | mPCRselect Pipeline

We designed a novel Nextflow pipeline to select optimal SNP markers for population differentiation and individual identification. The pipeline is compatible with macOS and Linux operating systems and can run locally on a desktop/laptop computer, remotely on a computing cluster, or in the cloud. Most dependencies can be installed automatically in their own Conda environment through Nextflow (Di Tommaso et al. 2017). Only the optional dependencies (BaitsTools; [Campana 2018] and NGS-PrimerPlex; [Kechin et al. 2020]) must be installed by the end-user as no Conda recipe currently exists for these packages. We describe the pipeline briefly below. A simplified flow diagram of the pipeline is available as Figure S1. See the mPCRselect documentation (https://github.com/ellieearmstrong/mPCRselect) for a complete diagram including the various parameters, internal processes and outputs.

mPCRselect's primary input files are a variant call format (VCF) file of individual genotypes and a comma-separated value (CSV) table assigning each individual in the VCF to a designated population. Optional input includes a list of individuals to remove from the dataset, a list of chromosomes to retain in the analysis and a browser extensible data (BED) format file of genomic coordinates to remove from the dataset (e.g., regions of low mappability, repetitive regions, etc.). mPCRselect uses VCFtools (Danecek et al. 2011) to remove unwanted individuals, chromosomes and genomic regions from the dataset. VCFtools is also used to identify and remove singleton and individual-unique doubleton sites, remove sites that failed previously applied filters (e.g., those without a 'PASS' flag), exclude non-biallelic sites, and to filter the input VCF by genotype quality (GQ), site, and individual data missingness. The pipeline uses a custom Python script ('Culling.py') to remove SNPs that are within a user-specified distance of another SNP (default is 40 bp; recommended based on typical size of amplicons). This filter aims to exclude variants where primer attachment sites would contain additional SNPs and inhibit amplification. Afterwards, the sites are optionally thinned by physical distance using VCFtools and by LD using PLINK 2 (Chang et al. 2015). The resulting file then is pushed through two distinct paths to select markers for population assignment and markers for individual identification, respectively.

For population assignment, we implement the greedy algorithm described above using a custom R script ('make_fst_plots.R'). In order to account for differences in sample size between populations, we bootstrap individuals to a user-specified population size (default is 20 individuals per population). The greedy algorithm is run a user-specified number of times (default is 20 repetitions per comparison of two populations). Users are provided with output plots which correlate the number of markers with assignment accuracy for each repetition. Additionally, mPCRselect identifies the sites with the highest F_{ST} values overall using VCFtools (flag '--weir-fst-pop'). mPCRselect then uses a custom Ruby script ('get_best_snps.rb') to identify the sites that appear most frequently in the lists of highest F_{ST} sites from each of the greedy algorithm runs and from the VCFtools F_{ST} analysis.

To select markers for individual identification, the VCF file is first split into distinct populations and π is then estimated using VCFtools with the flag '--site-pi', which calculates nucleotide divergence on a per site basis using the following equation:

$$\pi = \frac{AC * (AN - AC) + (AN - AC) * AC}{AN * (AN - 1)}$$

where AC is allele count and AN is allele number. We use π to select sites for individual identification, because π will also be maximized at alleles with intermediate frequencies. For each population, we compile a list of sites with the highest π. We then use the 'get_best_snps.rb' script to identify the most frequently appearing sites across the datasets. Afterwards, we calculate the Probability of Identity (PID) of the selected SNPs using a custom R script ('RMP_calc.R'). PID is the probability that two randomly sampled individuals have identical genotypes. For biallelic loci at Hardy–Weinberg equilibrium, this probability is:

$$p^4 + 4p^2(1-p)^2 + (1-p)^4$$

After site selection, the individual identification and ancestry assignment SNP sets are combined and the user can choose to design baits for capture-based projects using BaitsTools (Campana 2018) or alternatively import the sites into NGS-PrimerPlex (Kechin et al. 2020) to design primers for multiplex SNP panels. Finally, as an *in silico* validation of the panel design, the programme performs PCA (using PLINK 2) on the biallelic sites from the unfiltered input VCF, the post-filtering VCF and the final chosen population-assignment and individual-identification sites (both separately and concatenated).

We benchmarked the mPCRselect pipeline using the 'time' command and default mPCRselect settings (as specified in the default 'nextflow.config' file for mPCRselect 0.3.1) on a 2022 Mac Studio with an Apple M1 Max chip and 64 GB of memory. Our benchmarking analysis used a sub-selected dataset of 40 tigers representing six populations (Amur, Bengal, Generic, Indochinese, Malayan and Sumatran: Armstrong et al. 2021; dataset available on dryad under doi: 10.5061/dryad.0k6djhb96).

The dataset consisted of 1,319,280 variant sites on three chromosomes (B1, F2 and D4), which the benchmark settings further restrict to two chromosomes (B1 and F2; maximum 1,020,529 sites) to test the chromosome filtration step.

3 | Results

3.1 | Population Assignment Using Realistic Simulated Data

To better understand how F_{ST} influenced the ability of f_{ORCA} to assign individuals to a population, we simulated a split-model of two populations with no migration in SLiM. We used the full data to estimate allele frequencies and compute F_{ST} between the two populations. Then, we classified subsets of (N) individuals using (M) markers for population assignment. For each set of parameters, we computed the rate of misclassification while varying N, M, and F_{ST} values. Overall, we found that when F_{ST} between populations is 0.01, individuals could be correctly assigned to a population with as few as 10 markers (Figure 1). However, the probability of incorrectly assigning individuals is quite large, the average misclassification rate was approximately 37% (Table S1). Importantly, increasing the marker set to as few as 100 markers decreased the average misclassification rate to approximately 12% (Table S2). Conversely, when $F_{ST} = 0.09$, which is on the same order of magnitude as the average F_{ST} between two human populations (Ramachandran et al. 2005; Rosenberg et al. 2005), 10 markers resulted in an average misclassification rate of 21% (Table S1). When we increased the marker set and included 80 markers, we achieved misclassification rates that were on-average $< 1\%$ (Figure 1). Increasing the number of markers improved accuracy for every tested value of F_{ST} (Figure 1). Our results demonstrate that when F_{ST} is small (0.01), it is necessary to include more markers for accurate classification (Tables S1 and S2), and as F_{ST} increases accurate results can be achieved with fewer markers and fewer individuals (Figure 1).

3.2 | Empirical Data

After testing the algorithm with simulated data, we tested the method using three empirical data sets from humans, tigers and domestic dogs. Following the simulations, we used the full data to estimate allele frequencies and compute F_{ST} between the two populations. Then, we classified subsets N individuals with M markers for population assignment. First, we compared publicly available human data from the 1000 Genomes Project (1000 Genomes Project Consortium et al. 2010). We used whole-genome sequence data from an African population with a single origin, the Luhya from Kenya, and an Asian population with a single origin, the Dai from China (Figure 2A). The F_{ST} between these population approximately 0.11. When the number of markers was the smallest, we observed the largest accuracy gain using the markers selected from the method developed here versus randomly selected markers in linkage equilibrium. For example, when the number of markers was the smallest ($M = 10$), the average accuracy across random marker sets was 0.8884 ± 0.0651, while the average accuracy across top marker sets is a perfect accuracy of 1. When the number of randomly selected markers was 20, we observed an average accuracy across groups of 0.9611 ± 0.0040, compared to the top marker set which had an average accuracy of 1. As the number of markers increased, the accuracy gain decreased. This pattern was observed across all datasets.

Given their endangered status and relatively recent efforts to sequence tiger populations, we only had access to 13 Amur individuals for performing classification (Figure 2B). The limited sample size also impacted our ability to vary the number of individuals we sampled for classification. However, because the F_{ST} value (~0.2; [Armstrong et al. 2021]) between the wild Amur and Bengal tiger populations is larger than that of human populations, even using only 10 random markers results in a classification accuracy of 0.9173 ± 0.0652 (Figure 2A,B). This accuracy was higher than what was achieved in the human

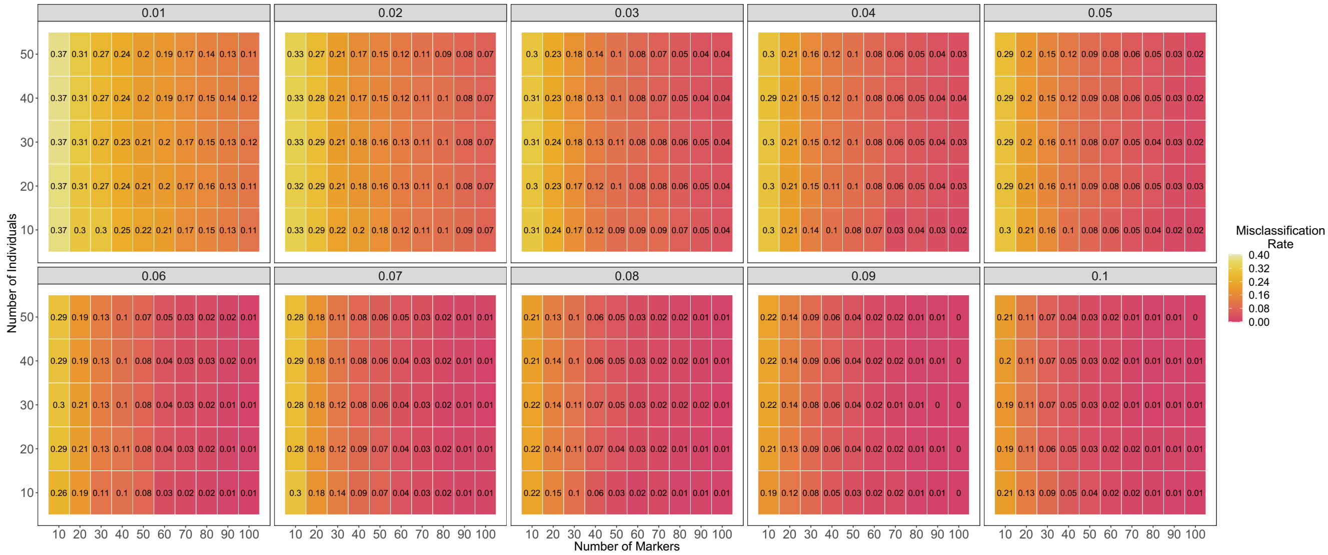

FIGURE 1 | Classification using simulated data. Two populations with a given measure of F_{ST} (as indicated in the grey bar above each panel) were classified using (M) markers (x-axis) and that consisted of (N) individuals per-population (y-axis). The mean misclassification rate over 20 simulation replicates per parameter combination ranges from 0 (pink) to 0.4 (light yellow). As the F_{ST} value and number of markers increases, the misclassification rate decreases.

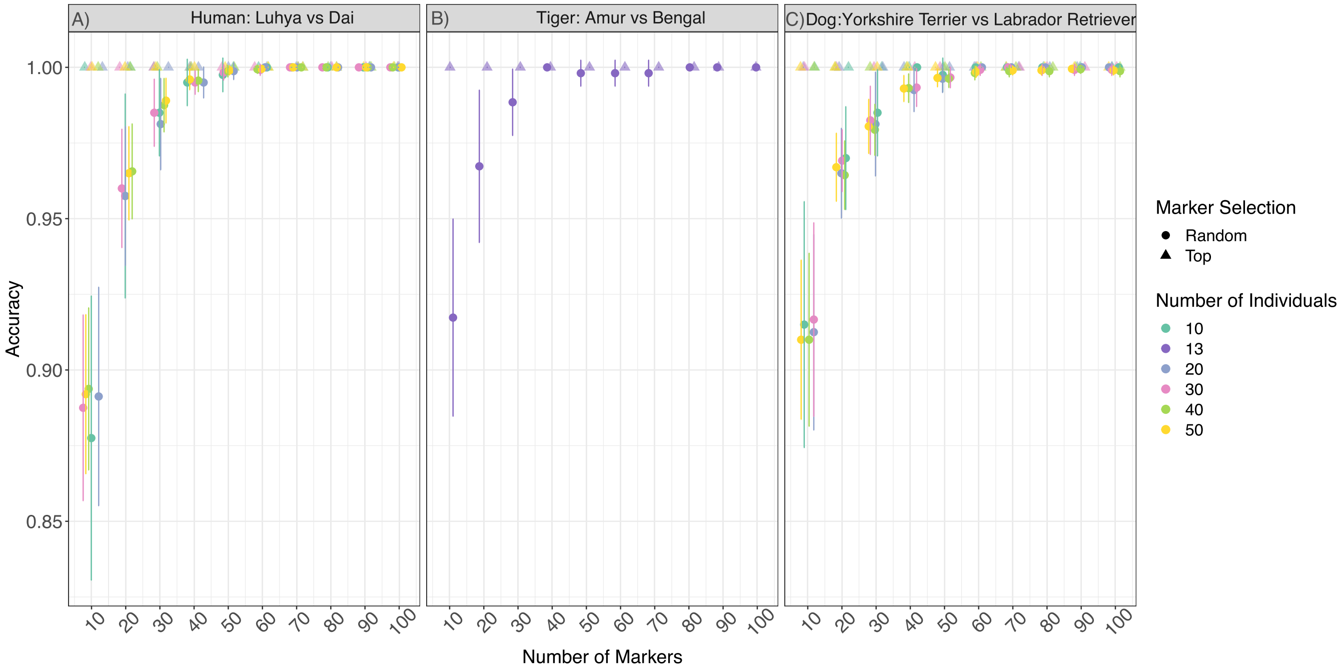

FIGURE 2 | Classification accuracy using two approaches, top F_{ST} markers (triangles) and random markers in linkage equilibrium (circles). For the random markers, each dot signifies the mean over 20 simulation replicates, error bars represent the standard deviation. The *x*-axis indicates the number of markers used for classification and the dot colour indicates the number of individuals. The accuracy of classification is shown on the *y*-axis. Here, three classifications were conducted using human populations (Luhya and Dai), tiger populations (Amur and Bengal) and breed dogs (Yorkshire Terrier and Labrador Retriever).

dataset with the same number of markers for each set of classifications. Additionally, the random marker set accuracy began to match the top markers more quickly in the tigers than humans.

Lastly, we examined breed dogs (Figure 2C), which have high homozygosity within breeds, but are quite divergent between the various clades (Parker et al. 2017). We used two breeds with the largest samples from Mooney, Yohannes, and Lohmueller (2021): Yorkshire Terrier and Labrador Retriever. Since both dog (F_{ST} ~ 0.14) and tiger populations (F_{ST} ~ 0.2) had a slightly larger F_{ST} than the human populations (F_{ST} ~ 0.1), we once again observed that the random marker set started with a high classification accuracy (0.9128 ± 0.0639) and reached the same performance as top markers faster relative to humans (Figure 2).

We also created a dataset that allowed us to explore whether the degree of admixture influenced our ability to accurately classify populations (Figures S2 and S3). In order to achieve this, we conducted classification in additional human populations, specifically Puerto Rican, Colombian and Afro-Caribbean populations, (Figure S2) as well as the captive (Generic) tiger population (Figure S3). We found that admixture decreased classification accuracy in both human and tiger populations. The drop in accuracy was most impactful when using random marker sets (Figures S2 and S3), and less severe when using f_{ORCA} in conjunction with the pipeline developed here which selected the top F_{ST} markers. It is also important to note that the degree to which there was shared ancestry between the two populations played a role in the magnitude of the decreased accuracy. For example, the Puerto Rican population represented in 1000 genomes has a more similar ancestry composition to sampled Colombian population than the sampled Afro-Caribbean population (1000 Genomes Project Consortium et al. 2010). Thus, the classification is much worse across all random marker sets in the Colombian population compared with the Afro-Caribbean. Classification also required more top F_{ST} markers to perform accurately in the admixed populations. In the tigers, we observed a similar pattern when comparing classification accuracy for differentiating the Amur and Bengal populations versus the Amur and Generic (captive) populations. On average, a random individual in the captive population could have up to approximately 39% Amur ancestry (Armstrong et al. 2024), which led to a marked (0.9173 ± 0.0652 to 0.8442 ± 0.0813) drop in our classification accuracy at the minimum marker set, $M = 10$, and a lag in the random marker classification accuracy reaching the accuracy of the top markers.

We also repeated all of the classifications while splitting the data into a training (humans $N = 30$, dog $N = 30$, and tiger $N = 5$) and test set of individuals (humans $N = 50$, dog $N = 50$, and tiger $N = 10$) and obtained results for each classification (Figure S4). Overall, our results with the split and full data were similar. We observed the worst performance when classifying the two admixed populations with the lowest F_{ST} values. When the full data were used, we observed slightly better classification results with both the top markers and random markers in linkage equilibrium. Though the full data resulted in better accuracy for classification, in the case of replicates with random markers, the standard deviations of the full data overlapped with the split data (Figure S4). The stability of accuracy across these two scenarios demonstrated that the most important component for classification is the number of individuals that were independently sampled to infer the allele frequency within the population and compute F_{ST} between populations.

Overall, our marker selection method with F_{ST} always outperformed the randomly selected marker sets. In order to accurately classify population pairs with lower F_{ST} values, we always had to use more markers. For randomly selected markers in linkage equilibrium, as more markers are incorporated into the panel, there was consistently better performance for population assignment. Our results are comparable to previous conclusions from both theory and empirical data (McVean 2009; Patterson, Price, and Reich 2006; Rosenberg 2005).

3.3 | Allele Frequency Differences of the Most Informative Markers

We next examined the allele frequency distributions of the top marker sets in humans, tigers and dogs (Figure 3). As we expected, the majority of top F_{ST} markers were concentrated at opposite allele frequencies, irrespective of the compared populations. Given that F_{ST} quantifies the magnitude of drift between two populations, this was expected. The allele frequency differences were smallest in the human populations studied (Figure 3A) and largest in tigers and dogs (Figure 3B,C), which have larger values of F_{ST} overall.

Taking the same approach as with classification, we explored whether the degree of admixture influenced the magnitude of the allele frequency gap of top F_{ST} markers between the same admixed human and captive (Generic) tiger populations (Figures S5 and S6). Indeed, admixture decreased the allele frequency gap and the degree to which there was shared ancestry affected how close the allele frequency gap was. This was expected, given that recent shared ancestry will likely result in populations having more similar allele frequencies (Ramachandran et al. 2005). Since the Puerto Rican population has a more similar ancestry composition to sampled Colombian population (F_{ST} ~ 0.005), we expected that the allele frequency gap would be smaller than both the Afro-Caribbean (F_{ST} ~ 0.054) and Luhya (F_{ST} ~ 0.065) populations, which was ultimately what we observed (Figure S5). The tigers followed this expectation as well, and the allele frequency gap between the Amur and Bengal tigers was larger than that between the Amur and Generic tigers (Figure S6).

3.4 | The Relationship Between Top Ancestry Markers and Probability of Identity

One measure of the degree to which a marker is useful in individual identification is the probability that two random individuals from the population have matching genotypes at the marker—if this probability is low, then the marker will distinguish individuals often. We will refer to this probability as the PID. PID is defined as the probability that two randomly sampled individuals have identical genotypes. For biallelic loci in Hardy–Weinberg equilibrium, this probability is:

$$p^4 + 4p^2(1-p)^2 + (1-p)^4$$

The equation above is minimised at allele frequency $p = 0.5$. In other words, markers with allele frequencies near 0.5 are the most useful for individual identification. Such markers tend to not be found in our top F_{ST} marker sets (Figure 3). In tigers and dogs, markers with high F_{ST} seldom had allele frequencies near 0.5 in either population. For human populations, though we found more alleles that existed at frequencies closer to 0.5, we

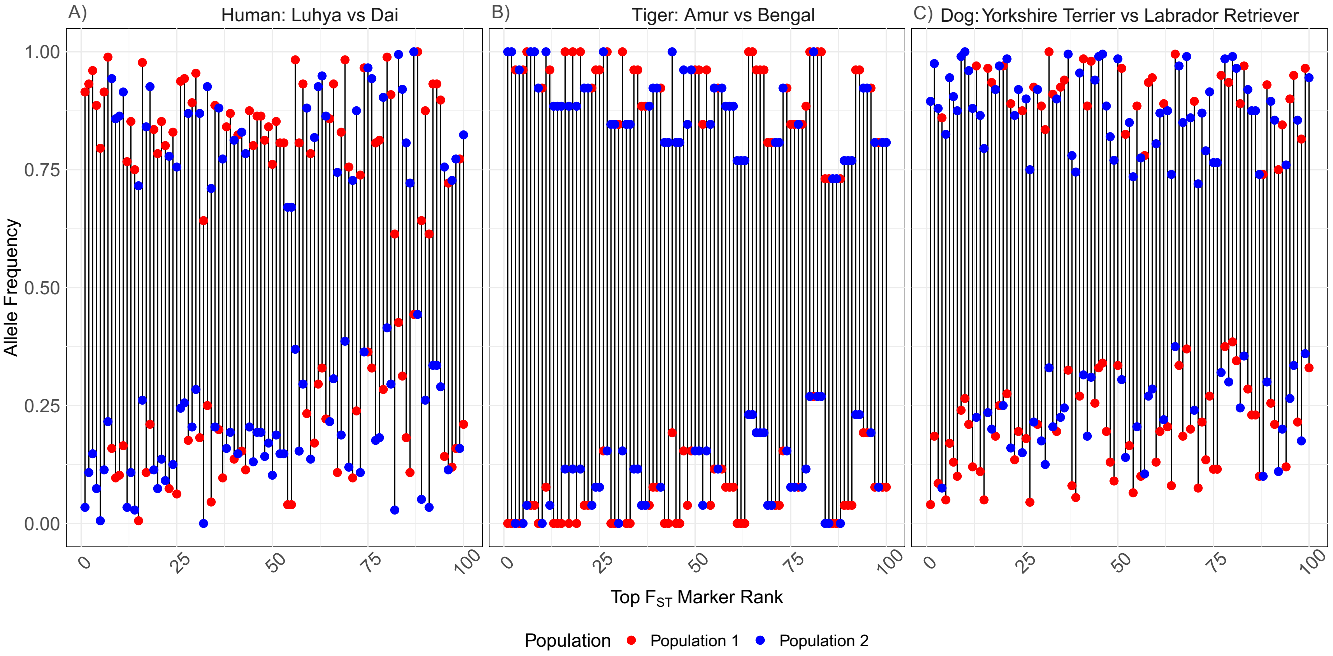

FIGURE 3 | Allele frequency of the top 100 F_{ST} markers between populations, and the frequency difference for each marker between various populations. The *x*-axis indicates the rank of F_{ST} values of the markers, and the *y*-axis indicates the markers' allele frequency in both Luhya (represented in blue) and Dai (represented in red) populations; Amur (represented in blue) and Bengal (represented in red); and Labrador Retriever (represented in blue) and Yorkshire Terrier (represented in red).

still did not observe a strong overlap between top F_{ST} and PID markers when comparing our reference population. When we identified the top 100 markers for minimising PID, we observed no overlapping markers in the Luhya population, no overlapping markers in the Amur subspecies and no overlapping markers in Yorkshire Terriers when the marker set was overlapped with the top F_{ST} markers in Figure 3. When we further explored the overlap between top F_{ST} and PID markers, we found that two markers overlapped when comparing the Puerto Rican and Luhya populations, one marker overlapped when comparing the Puerto Rican and Colombian populations, two markers overlapped when comparing the Puerto Rican and Afro-Caribbean populations and no markers overlapped when comparing the Generic and Amur populations.

3.5 | mPCRselect Performance

The mPCRselect analysis completed in 2 wall clock hours (17.4 CPU hours) using the equivalent of 8.86 CPUs. As analysis time scales approximately linearly with both the number of sites and the number of individuals, these results indicate that large (millions of SNPs and hundreds of individuals) can be processed on a current workstation in a reasonable time frame (a few days). Extremely large datasets or those that utilise a large number of populations (>6) may require parallelisation on a high-performance computing cluster or cloud instance as the number of F_{ST} comparisons scales approximately quadratically with the number of specified populations.

4 | Discussion

This work introduces the mPCRselect pipeline which is designed to provide users with a sufficient marker set to distinguish populations within their species of interest and/or a marker set to identify individuals. Markers for population assignment are selected using F_{ST} and tend to be close to fixation or loss when comparing two populations, while markers which optimise individual identification hold intermediate frequencies within populations. Implementing our marker selection method consistently reduces the number of markers required for accurate population assignment.

Our findings are consistent with previous research, where the relationship between the number of SNPs, their frequency, and the power to detect differentiation between populations has been explored in a conservation context (Morin, Martien, and Taylor 2009; Willing, Dreyer, and van Oosterhout 2012). Morin, Martien, and Taylor (2009) primarily explored this relationship from the perspective of initial study design (i.e., determining how many loci are necessary to detect population structure without having knowledge of the MAF or linkage status of a marker) and confirmed that more SNPs are necessary to detect differentiation between groups with lower values of F_{ST}. Willing, Dreyer, and van Oosterhout (2012) also echoed results from Patterson, Price, and Reich (2006) showing that even with small sample sizes, large numbers of markers can compensate to provide accurate estimates of F_{ST}. A graphical user interface with the explicit goal of helping conservation practitioners select the appropriate number of samples and markers and avoid suboptimal sampling was presented in Hoban et al. (2013), again from the perspective of initial study design. Critically, these studies emphasize that a sufficient number of individuals must be sampled in order to get an accurate estimate of allele frequencies in the population and F_{ST}.

Our method is in line with previous findings, but we approach the other end of the problem when populations have already been identified and one desires identifying optimal markers. We found that population assignment accuracy increases as more informative markers are added. We observed the lowest accuracy when we used the smallest set of random markers ($M=10$), and as the number of random markers increased, they achieved a performance similar to the top markers. This limitation on information content when using F_{ST} for marker selection was previously highlighted in several studies (Balding and Nichols 1994; Baye et al. 2009; Galanter et al. 2012; Kidd et al. 2006; Manel, Gaggiotti, and Waples 2005; Rosenberg 2005; Rosenberg et al. 2003). Our method, which identifies the most informative markers by computing F_{ST}, then conducts population assignment with f_{ORCA} consistently outperforms or does as well as random markers when a sufficient set size is achieved. However, we emphasise the findings of previous studies which show that sufficient data are necessary for detecting population structure initially (i.e., that sufficient markers and individuals are required to detect structure between populations; Patterson, Price, and Reich 2006, Morin, Martien, and Taylor 2009, Willing, Dreyer, and van Oosterhout 2012). Insufficient data at this stage would ultimately result in inaccurate allele frequencies and hinder the accurate estimation of F_{ST} and π, which are critical for optimal marker selection.

Our pipeline independently estimates the population level allele frequency and F_{ST} between populations before classification. Thus, when individuals are added or removed from assignment, we see only slight fluctuations in assignment accuracy. For example, in Figure 1, in the case where $F_{ST}=0.01$ and the number of markers is fixed at $N=20$, as we increase the number of individuals being classified, we observe a slight increase in the average misclassification rate from 0.3 to 0.31 over simulation replicates. Importantly, the average misclassification rate is stable, and the standard deviations as we increase the number of individuals overlap (Figure S7). Additionally, we observe the same slight fluctuations across F_{ST}. When we compare F_{ST} of 0.08, 0.09, and 0.1 and we select a fixed number of markers ($N=10$) and number of individuals ($N=10$) for classification, average accuracy is similar across F_{ST} values. The slight change is due to generating a new replicate and selecting new individuals for each classification replicate. However, there is no significant difference between these small fluctuations and average accuracy remains stable across the number of selected markers and individuals when the change in F_{ST} is small (Figure S7).

It is important to note that populations with higher divergence (as measured by F_{ST}) will require fewer markers for classification, whereas populations that are less divergent will require more markers. If divergence between populations is large enough, one could even use very few random markers in linkage equilibrium and accurately classify individuals to a population. For example, in Figure S3 when F_{ST} is extremely high (e.g., tigers), we have the ability to classify individuals (accuracy > 0.9) with even 20 random markers. In contrast when F_{ST} is lower, such as in Figure S2 with human populations, when we add more random markers, we do not observe as steep of a gain in assignment accuracy for classification. Our work is not the first to highlight this result while using f_{ORCA} for population assignment. For example, in Rosenberg (2005), as few as 20 randomly selected markers in linkage equilibrium could classify human populations with an accuracy of ~90%. For other organisms (carp, cat, chicken, dog, fly, grayling and maize) as few as six random markers in linkage equilibrium could be used to achieve an assignment accuracy around ~90%.

We also explored alternate methods for marker selection for classification apart from F_{ST}, since f_{ORCA} and mPCRselect can perform assignment with any set of markers the user desires. We observed that F_{ST} and δ were consistently the best methods for marker selection across species (Figure S8). Generally, PCA demonstrated the worst performance. In the case of classifying the two dog populations, PCA had a lower accuracy than even random markers. The general poor performance of PCA for population assignment in relation to F_{ST} and δ has also been observed in previous work (Wilkinson et al. 2011).

Lastly, we used simulations to explore the accuracy of classifying a different population than either of those used to select markers. In the first set, the ancestral population split into three populations, A, B and C. In the second set, the ancestral population splits into two populations A and B, and after 1000 generations (when A–B F_{ST} is expected to be 0.05), a third population, C, split off from population A. This split time results in A and C having a divergence that is half of A and B.

In the case where population C split off from the ancestral population at the same time as A and B, and we used populations A and B to compute allele frequencies and select markers, we found on average (across 20 replicates) that the misclassification rate increased when C (Figure S9) was classified instead of A (Figure 1). When F_{ST} between populations is 0.01, individuals could still be correctly assigned to a population with as few as 10 markers (Figure S9). However, the probability of incorrectly assigning individuals is quite large, the average misclassification rate increased from approximately 37% (Figure 1) to approximately 43% (Figure S9). When we increased the marker set to 100, the average misclassification rate more than doubled to approximately 28% relative to Figure 1, where the average misclassification rate was approximately 12% (Figure S9). When $F_{\mathrm{ST}} = 0.1$, we observed an average misclassification rate as high as 33% ($M = 10$) in contrast to Figure 1 where the largest average misclassification rate was approximately 21% (Figure S9).

In the simulations where population C split from A and population C maintained an equivalent F_{ST} with B as A, our approach is capable of classifying individuals from C using markers selected from populations A and B (Figure S10). However, as expected, we observed that population assignment accuracy was on average slightly lower than classifying population A, though the confidence intervals across simulation replicates overlapped (Figure S10). For example, when F_{ST} between populations A and C was 0.01, and 0.06 between A and B and B and C, if 10 markers were used the average accuracy across the 20 replicates was 0.6950 ± 0.0809 when A was classified versus 0.7075 ± 0.0950 when C was classified (Figure S10). When the number of markers increased to 100, the average accuracy across the 20 replicates was the same 0.9825 ± 0.0335 when A was classified versus 0.9825 ± 0.0245 when C was classified (Figure S10). The same pattern held as F_{ST} between populations A and C increased 0.05, and F_{ST} between A and B and B and C increased to 0.1. When we used 100 markers for classification, the average accuracy across the 20 replicates was the same 0.9950 ± 0.0154 when A was classified versus 0.9950 ± 0.0224 when C was classified (Figure S10). Through these two sets of simulations, we observed that classification of a related population is possible using markers selected from its close relative. mPCRselect will produce stable classification results when the populations being interchanged are closely related, here $F_{\mathrm{ST}} \leq 0.05$, and genetically equidistant from the other population of interest.

In the future, one natural extension to explore is optimising markers for PID and population assignment simultaneously, rather than as distinct steps. Incorporating PID will tend to increase the number of markers on the panel, because markers that are most informative for population classification (Figure 3) typically have relatively low minor allele frequencies in each subpopulation. Conversely, PID (as defined above) is minimised by alleles of intermediate frequency. Thus, for biallelic loci, the top F_{ST} markers, which are the markers most useful in population assignment, tend to be the least useful markers in the individual identification. This contrasts with the situation for multi-allelic human STRs, where a correlation has been observed between markers' usefulness for individual identification and for population classification (Algee-Hewitt et al. 2016). Another natural extension of this method could be for identification and selection of marker sets for sex identification and relatedness. If one wanted to approach creating a marker set that is balanced for both population assignment and individual identification, a greedy approach might be useful. The user begins with a SNP that is high in f_{ORCA} (or low in PID) across populations and for every subsequent SNP measure whether they are farther away from their desired f_{ORCA} or PID goals. Then, they would select a SNP that maximises the approach for f_{ORCA} or PID, conditional on the SNPs already in the set.

Though we do not discuss the practical applications of applying SNP panels here, it is broadly recommended to design and test more primer sets than what is identified as the 'minimum number' to achieve successful population assignment or individual identification, since some SNPs will not amplify as well as others, or may not work due to other factors such as primer dimerisation. Panels designed for low-quality samples (scat, hair, environmental or forensic materials) may require an increase in the overall number of SNPs being screened, as drop out will occur due to sample degradation, which will impact the ability to accurately assign or identify any one sample. This must be balanced with the expected amount of endogenous DNA in the sample because primers will not amplify well if there is too little DNA template. Similarly, we recommend designing 'excess' hybridisation capture baits as individual baits vary in their capture efficiency (e.g., due to GC content, secondary structures, synthesis efficiency, etc.). This can result in biased capture and locus drop out. While hybridisation capture works well on fragmented and low-concentration DNA samples, capture can become inefficient if the capture conditions are too non-specific and can fail if the target sequences are too divergent from the capture baits (> ~20%: Hawkins et al. 2016). To maximise recovery of specific loci or genotype low-quality samples, we recommend increased tiling density (the depth of bait coverage over a specified target) to improve the likelihood of capture (e.g., Parker et al. 2022).

Our novel pipeline mPCRselect streamlines SNP panel design for ancestry assignment and individual identification. Further, the flexibility of this software allows for straight-forward integration of novel algorithms for marker selection in the future. Similar pipelines have already been created for selecting markers from human genomic data (Chen et al. 2020). One such pipeline is the R Package AIMsetfinder, which uses a Bayesian approach to identify SNPs which are informative of population assignment. AIMsetfinder's approach is different from the approach described here because it tests every marker in the dataset, then goes backwards to create a set of markers that minimise a logarithmic loss function (Pfaffelhuber et al. 2020). Contrastingly, our pipeline maximises our assignment function, f_{ORCA}, using the most informative markers while simultaneously providing the user with a seamless connection to amplicon primer or hybridisation capture design software.

In sum, to create an effective SNP panel, one must carefully consider the minimum number of markers that should be present on the panel and the number of individuals to sample. These values should be determined by how divergent the populations of interest are and how accurate the classification needs to be. Our pipeline mPCRselect streamlines the process of selecting optimised markers for population assignment and individual identification for any user with sufficient data.

Author Contributions

E.E.A. and J.A.M. conceived of the project. C.L., J.A.M. and T.F. performed simulations and data analysis. E.E.A., C.L. and M.G.C. built the mPCRselect pipeline, with contributions from K.A.S. and J.A.M. E.E.A., C.L. and J.A.M. wrote the manuscript. M.G.C, J.L.K., D.A.P. and K.A.S. provided feedback and input on manuscript content and analyses. All authors approved of the final manuscript.

Acknowledgements

We thank Michael ‘Doc’ Edge for extremely thoughtful and helpful comments and conversations about this work. E.E.A. was supported by a Washington Research Foundation Postdoctoral Fellowship. C.L. was supported by the WiSE Summer Undergraduate Research Fellowship. J.A.M. was supported by the start-up funds from Dornsife College of Letters, Arts and Sciences through the Department of Quantitative and Computational Biology and the USC WiSE Gabilan Assistant Professorship. M.G.C. was supported by the Smithsonian Institution.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Summary files, simulation scripts, classification code and empirical data for this project is provided at GitHub repository https://github.com/ChenyangLi6/SNP-panel. mPCRSelect is available at https://github.com/ellieearmstrong/mPCRselect.

References

1000 Genomes Project Consortium, G. R. Abecasis, D. Altshuler, et al. 2010. “A Map of Human Genome Variation From Population-Scale Sequencing.” *Nature* 467, no. 7319: 1061–1073.

Algee-Hewitt, B. F. B., M. D. Edge, J. Kim, J. Z. Li, and N. A. Rosenberg. 2016. “Individual Identifiability Predicts Population Identifiability in Forensic Microsatellite Markers.” *Current Biology: CB* 26, no. 7: 935–942.

Armstrong, E. E., A. Khan, R. W. Taylor, et al. 2021. “Recent Evolutionary History of Tigers Highlights Contrasting Roles of Genetic Drift and Selection.” *Molecular Biology and Evolution* 38, no. 6: 2366–2379.

Armstrong, E. E., J. A. Mooney, K. A. Solari, et al. 2024. “Unraveling the Genomic Diversity and Admixture History of Captive Tigers in the United States.” *Proceedings National Academy of Sciences of the United States of America* 121, no. 39: e2402924121. https://doi.org/10.1073/pnas.2402924121.

Balding, D. J. 2003. “Likelihood-Based Inference for Genetic Correlation Coefficients.” *Theoretical Population Biology* 63, no. 3: 221–230.

Balding, D. J., and R. A. Nichols. 1994. “DNA Profile Match Probability Calculation: How to Allow for Population Stratification, Relatedness, Database Selection and Single Bands.” *Forensic Science International* 64, no. 2-3: 125–140.

Baye, T. M., H. K. Tiwari, D. B. Allison, and R. C. Go. 2009. “Database Mining for Selection of SNP Markers Useful in Admixture Mapping.” *Biodata Mining* 2, no. 1: 1.

Bertola, L. D., M. Vermaat, F. Lesilau, et al. 2022. “Whole Genome Sequencing and the Application of a SNP Panel Reveal Primary Evolutionary Lineages and Genomic Variation in the Lion (*Panthera leo*).” *BMC Genomics* 23, no. 1: 321.

Biddanda, A., D. P. Rice, and J. Novembre. 2020. “A Variant-Centric Perspective on Geographic Patterns of Human Allele Frequency Variation.” *eLife* 9: e60107. https://doi.org/10.7554/eLife.60107.

Bien, S. A., G. L. Wojcik, N. Zubair, et al. 2016. “Strategies for Enriching Variant Coverage in Candidate Disease Loci on a Multiethnic Genotyping Array.” *PloS One* 11, no. 12: e0167758.

Campana, M. G. 2018. “BaitsTools: Software for Hybridization Capture Bait Design.” *Molecular Ecology Resources* 18, no. 2: 356–361.

Campbell, N. R., S. A. Harmon, and S. R. Narum. 2015. “Genotyping-in-Thousands by Sequencing (GT-seq): A Cost Effective SNP Genotyping Method Based on Custom Amplicon Sequencing.” *Molecular Ecology Resources* 15, no. 4: 855–867.

Carroll, E. L., M. W. Bruford, J. A. DeWoody, et al. 2018. “Genetic and Genomic Monitoring with Minimally Invasive Sampling Methods.” *Evolutionary Applications* 11, no. 7: 1094–1119.

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. 2015. “Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets.” *GigaScience* 4: 7.

Chen, B., J. W. Cole, and C. Grond-Ginsbach. 2017. “Departure from Hardy Weinberg Equilibrium and Genotyping Error.” *Frontiers in Genetics* 8: 167.

Chen, S., S. Ghandikota, Y. Gautam, and T. B. Mersha. 2020. “MI-MAAP: Marker Informativeness for Multi-Ancestry Admixed Populations.” *BMC Bioinformatics* 21, no. 1: 131.

Ciezarek, A., A. G. P. Ford, G. J. Etherington, et al. 2022. “Whole Genome Resequencing Data Enables a Targeted SNP Panel for Conservation and Aquaculture of Oreochromis Cichlid Fishes.” *Aquaculture* 548: 737637.

Danecek, P., A. Auton, G. Abecasis, et al. 2011. "The Variant Call Format and VCFtools." *Bioinformatics* 27, no. 15: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

Di Tommaso, P., M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame. 2017. "Nextflow Enables Reproducible Computational Workflows." *Nature Biotechnology* 35, no. 4: 316–319.

Ding, L., H. Wiener, T. Abebe, et al. 2011. "Comparison of Measures of Marker Informativeness for Ancestry and Admixture Mapping." *BMC Genomics* 12: 622.

Fan, B., Z.-Q. Du, D. M. Gorbach, and M. F. Rothschild. 2010. "Development and Application of High-Density SNP Arrays in Genomic Studies of Domestic Animals." *Asian-Australasian Journal of Animal Sciences* 23, no. 7: 833–847.

Fitak, R. R., A. Naidu, R. W. Thompson, and M. Culver. 2015. "A New Panel of SNP Markers for the Individual Identification of North American Pumas." *Journal of Fish and Wildlife Management* 7, no. 1: 13–27.

Galanter, J. M., J. C. Fernandez-Lopez, C. R. Gignoux, et al. 2012. "Development of a Panel of Genome-Wide Ancestry Informative Markers to Study Admixture Throughout the Americas." *PLoS Genetics* 8, no. 3: e1002554.

Haller, B. C., and P. W. Messer. 2023. "SLiM 4: Multispecies Eco-Evolutionary Modeling." *American Naturalist* 201, no. 5: E127–E139.

Hawkins, M. T., C. A. Hofman, T. Callicrate, et al. 2016. "In-Solution Hybridization for Mammalian Mitogenome Enrichment: Pros, Cons and Challenges Associated with Multiplexing Degraded DNA." *Molecular Ecology Resources* 16, no. 5: 1173–1188.

Hemstrom, W., J. A. Grummer, G. Luikart, and M. R. Christie. 2024. "Next-Generation Data Filtering in the Genomics Era." *Nature Reviews Genetics*: 1–18. https://doi.org/10.1038/s41576-024-00738-6.

Hess, J. E., N. R. Campbell, M. F. Docker, et al. 2015. "Use of Genotyping by Sequencing Data to Develop a High-Throughput and Multifunctional SNP Panel for Conservation Applications in Pacific Lamprey." *Molecular Ecology Resources* 15, no. 1: 187–202.

Hirsch, C. D., J. Evans, C. R. Buell, and C. N. Hirsch. 2014. "Reduced Representation Approaches to Interrogate Genome Diversity in Large Repetitive Plant Genomes." *Briefings in Functional Genomics* 13, no. 4: 257–267.

Hoban, S., O. Gaggiotti, G. Bertorelle, and ConGRESS Consortium. 2013. "Sample Planning Optimization Tool for Conservation and Population Genetics (SPOTG): A Software for Choosing the Appropriate Number of Markers and Samples." *Methods in Ecology and Evolution/British Ecological Society* 4, no. 3: 299–303.

Kechin, A., V. Borobova, U. Boyarskikh, E. Khrapov, S. Subbotin, and M. Filipenko. 2020. "NGS-PrimerPlex: High-Throughput Primer Design for Multiplex Polymerase Chain Reactions." *PLoS Computational Biology* 16, no. 12: e1008468.

Khan, A., S. M. Krishna, U. Ramakrishnan, and R. Das. 2022. "Recapitulating Whole Genome Based Population Genetic Structure for Indian Wild Tigers Through an Ancestry Informative Marker Panel." *Heredity* 128, no. 2: 88–96.

Kidd, K. K., A. J. Pakstis, W. C. Speed, et al. 2006. "Developing a SNP Panel for Forensic Identification of Individuals." *Forensic Science International* 164, no. 1: 20–32.

Kleinman-Ruiz, D., B. Martínez-Cruz, L. Soriano, et al. 2017. "Novel Efficient Genome-Wide SNP Panels for the Conservation of the Highly Endangered Iberian lynx." *BMC Genomics* 18, no. 1: 556.

LaFramboise, T. 2009. "Single Nucleotide Polymorphism Arrays: A Decade of Biological, Computational and Technological Advances." *Nucleic Acids Research* 37, no. 13: 4181–4193.

Manel, S., O. E. Gaggiotti, and R. S. Waples. 2005. "Assignment Methods: Matching Biological Questions With Appropriate Techniques." *Trends in Ecology & Evolution* 20, no. 3: 136–142.

McVean, G. 2009. "A Genealogical Interpretation of Principal Components Analysis." *PLoS Genetics* 5, no. 10: e1000686.

Mooney, J. A., C. D. Marsden, A. Yohannes, R. K. Wayne, and K. E. Lohmueller. 2023. "Long-Term Small Population Size, Deleterious Variation, and Altitude Adaptation in the Ethiopian Wolf, a Severely Endangered Canid." *Molecular Biology and Evolution* 40, no. 1: msac277. https://doi.org/10.1093/molbev/msac277.

Mooney, J. A., A. Yohannes, and K. E. Lohmueller. 2021. "The Impact of Identity by Descent on Fitness and Disease in Dogs." *Proceedings of the National Academy of Sciences* 118, no. 16: e2019116118.

Morin, P. A., K. K. Martien, and B. L. Taylor. 2009. "Assessing Statistical Power of SNPs for Population Structure and Conservation Studies." *Molecular Ecology Resources* 9, no. 1: 66–73.

Morrell, P. L., and M. T. Clegg. 2007. "Genetic Evidence for a Second Domestication of Barley (*Hordeum vulgare*) East of the Fertile Crescent." *Proceedings of the National Academy of Sciences of the United States of America* 104, no. 9: 3289–3294.

Natesh, M., R. W. Taylor, N. K. Truelove, et al. 2019. "Empowering Conservation Practice With Efficient and Economical Genotyping From Poor Quality Samples." *Methods in Ecology and Evolution/British Ecological Society* 10, no. 6: 853–859.

Nicholson, G., A. V. Smith, F. Jónsson, Ó. Gústafsson, K. Stefánsson, and P. Donnelly. 2002. "Assessing Population Differentiation and Isolation from Single-Nucleotide Polymorphism Data." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 64, no. 4: 695–715.

Oliveira, R., E. Randi, F. Mattucci, J. D. Kurushima, L. A. Lyons, and P. C. Alves. 2015. "Toward a Genome-Wide Approach for Detecting Hybrids: Informative SNPs to Detect Introgression Between Domestic Cats and European Wildcats (*Felis silvestris*)." *Heredity* 115, no. 3: 195–205.

Pakstis, A. J., W. C. Speed, R. Fang, et al. 2010. "SNPs for a Universal Individual Identification Panel." *Human Genetics* 127, no. 3: 315–324.

Pakstis, A. J., W. C. Speed, J. R. Kidd, and K. K. Kidd. 2007. "Candidate SNPs for a Universal Individual Identification Panel." *Human Genetics* 121, no. 3-4: 305–317.

Parker, H. G., D. L. Dreger, M. Rimbault, et al. 2017. "Genomic Analyses Reveal the Influence of Geographic Origin, Migration, and Hybridization on Modern Dog Breed Development." *Cell Reports* 19, no. 4: 697–708.

Parker, L. D., M. G. Campana, J. D. Quinta, et al. 2022. "An Efficient Method for Simultaneous Species, Individual, and Sex Identification via In-Solution Single Nucleotide Polymorphism Capture From Low-Quality Scat Samples." *Molecular Ecology Resources* 22, no. 4: 1345–1361.

Paschou, P., E. Ziv, E. G. Burchard, et al. 2007. "PCA-correlated SNPs for structure identification in worldwide human populations." *PLoS Genetics* 3, no. 9: 1672–1686.

Patterson, N., A. L. Price, and D. Reich. 2006. "Population Structure and Eigenanalysis." *PLoS Genetics* 2, no. 12: e190.

Pearman, W. S., L. Urban, and A. Alexander. 2022. "Commonly uSed Hardy-Weinberg Equilibrium Filtering Schemes Impact Population Structure Inferences Using RADseq Data." *Molecular Ecology Resources* 22, no. 7: 2599–2613.

Pfaff, C. L., J. Barnholtz-Sloan, J. K. Wagner, and J. C. Long. 2004. "Information on Ancestry From Genetic Markers." *Genetic Epidemiology* 26, no. 4: 305–315.

Pfaffelhuber, P., F. Grundner-Culemann, V. Lipphardt, and F. Baumdicker. 2020. "How to Choose Sets of Ancestry Informative Markers: A Supervised Feature Selection Approach." *Forensic Science International. Genetics* 46: 102259.

Puckett, E. E. 2017. "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, no. 2: 289–304.

Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza. 2005. "Support From the Relationship of Genetic and Geographic Distance in Human Populations for a Serial Founder Effect Originating in Africa." *Proceedings of the National Academy of Sciences of the United States of America* 102, no. 44: 15942–15947.

Rosenberg, N. A. 2005. "Algorithms for Selecting Informative Marker Panels for Population Assignment." *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 12, no. 9: 1183–1201.

Rosenberg, N. A., L. M. Li, R. Ward, and J. K. Pritchard. 2003. "Informativeness of Genetic Markers for Inference of Ancestry." *American Journal of Human Genetics* 73, no. 6: 1402–1422.

Rosenberg, N. A., S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, and M. W. Feldman. 2005. "Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure." *PLoS Genetics* 1, no. 6: e70.

Scheben, A., J. Batley, and D. Edwards. 2017. "Genotyping-By-Sequencing Approaches to Characterize Crop Genomes: Choosing the Right Tool for the Right Application." *Plant Biotechnology Journal* 15, no. 2: 149–161.

Sottile, G., M. T. Sardina, S. Mastrangelo, et al. 2018. "Penalized Classification for Optimal Statistical Selection of Markers From High-Throughput Genotyping: Application in Sheep Breeds." *Animal: An International Journal of Animal Bioscience* 12, no. 6: 1118–1125.

Storer, C. G., C. E. Pascal, S. B. Roberts, W. D. Templin, L. W. Seeb, and J. E. Seeb. 2012. "Rank and Order: Evaluating the Performance of SNPs for Individual Assignment in a Non-Model Organism." *PLoS One* 7, no. 11: e49018.

Wehrenberg, G., M. Tokarska, B. Cocchiararo, and C. Nowak. 2024. "A Reduced SNP Panel Optimised for Non-Invasive Genetic Assessment of a Genetically Impoverished Conservation Icon, the European Bison." *Scientific Reports* 14, no. 1: 1875.

Wilkinson, S., P. Wiener, A. L. Archibald, et al. 2011. "Evaluation of Approaches for Identifying Population Informative Markers From High Density SNP Chips." *BMC Genetics* 12: 45.

Willing, E.-M., C. Dreyer, and C. van Oosterhout. 2012. "Estimates of Genetic Differentiation Measured by F_{ST} Do Not Necessarily Require Large Sample Sizes When Using Many SNP Markers." *PLoS One* 7, no. 8: e42649.

Wright, S. 1951. "The Genetical Structure of Populations." *Annals of Eugenics* 15, no. 4: 323–354.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.