Who are you? A framework to identify and report genetic sample mix-ups

Running title: A call to check for individual sample mix-ups

Laura Duntsch[1], Patricia Brekke[2], John G. Ewen[2], Anna W. Santure[1]

E-mail address of main author: ldun612@aucklanduni.ac.nz

[1] Centre for Biodiversity and Biosecurity, School of Biological Sciences, University of Auckland, Auckland, NZ
[2] Institute of Zoology, Zoological Society of London, Regents Park, London, UK

Abstract

Sample mix-ups occur when samples have accidentally been duplicated, mislabelled or swapped. When samples are subsequently genotyped or sequenced, this can lead to individual IDs being incorrectly linked to genetic data, resulting in incorrect or biased research results, or reduced power to detect true biological patterns. We surveyed the community and found that almost 80% of responding researchers have encountered sample mix-ups. However, many recent studies in the field of molecular ecology do not appear to systematically report individual assignment checks as part of their publications. Although checks may be done, lack of consistent reporting means that it is difficult to assess whether sample mix-ups have occurred or been detected. Here, we present an easy-to-follow sample verification framework that can utilise existing metadata, including species, population structure, sex and pedigree information. We demonstrate its application to a dataset representing individuals of a threatened Aotearoa New Zealand bird species, the hihi, genotyped on a 50K SNP array. We detected numerous incorrect genotype-ID associations when comparing observed and genetic sex or comparing to relationships in a verified microsatellite pedigree. The framework proposed here helped to confirm 488 individuals (39%), correct another 20 bird-genotype links, and detect hundreds of incorrect sample IDs, emphasizing the value of routinely checking genetic and genomic datasets for their accuracy. We therefore promote the implementation and reporting of this simple yet effective sample verification framework as a standardized quality control step for studies in the field of molecular ecology.

Keywords

Sample mix-up; SNP array data; pedigree verification; duplicate check; data QC; framework

Introduction

Modern sequencing and genotyping technologies allow for high-quality processing and relatively cost-effective evaluation of biological data. At the same time, standardized laboratory handling protocols and quality checks should ensure sample identification – in theory. However, even the most experienced laboratory is not safe from the occasional sample mix-up, often resulting in the affected sample being discarded if it is detected (Have et al., 2014; Wang et al., 2019). Sample mix-ups can happen at any stage in a research project: during data collection, labelling, transport and storage, or handling and processing in wet and dry laboratories (Figure 1), with laboratory mix-ups appearing to be particularly common (McClure et al., 2018). For example, label switches during lab work and sample contamination were detected in a recent avian genome sequencing project of hundreds of genomes (Feng et al., 2020), pipetting error was concluded as the likely cause for sample mix-ups in a mouse gene expression study (Broman et al., 2019), and multiple samples were found to be cross-contaminated during lab work for a mouse microbiome study (Lobo et al., 2019).

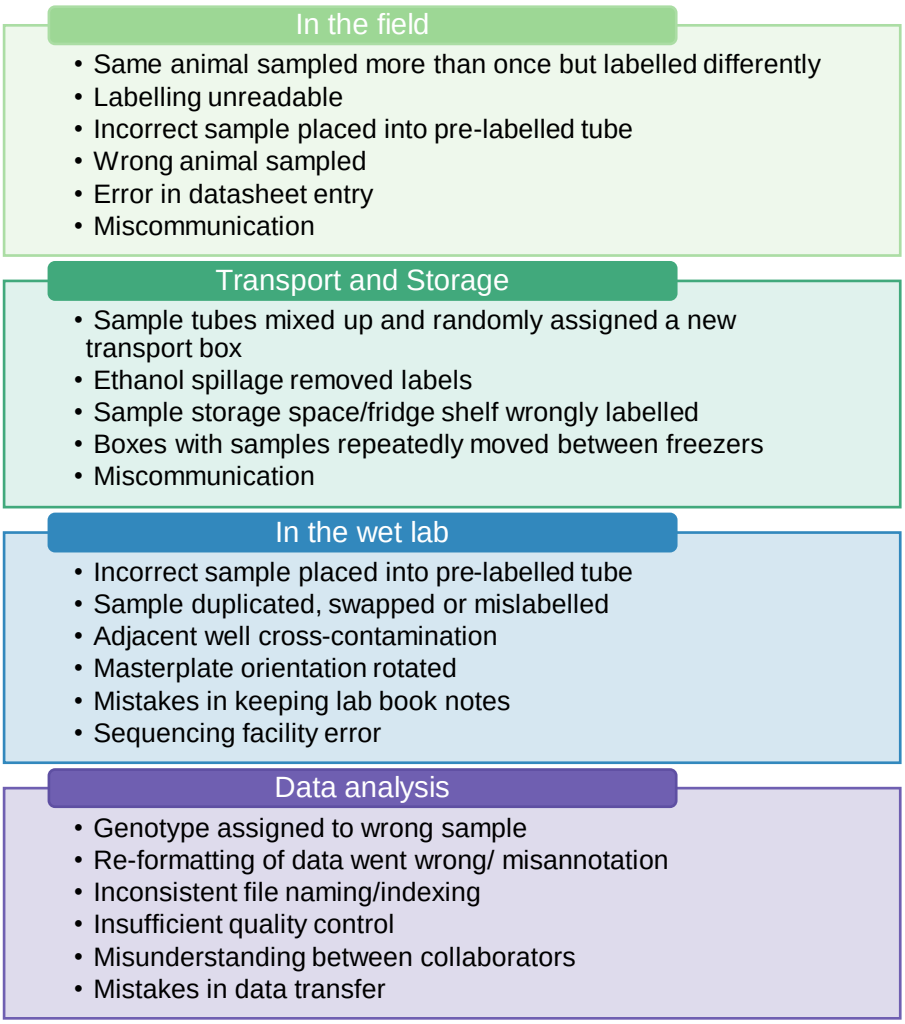

Figure 1: Examples of points in a research project where sample mix-ups could potentially occur. While sample errors are most likely to be detected at the end of a project from examining sequence or genotype data, sample mix-ups can happen at any stage, and can dramatically influence downstream conclusions.

Although a few undetected mix-ups in a large-scale genomics study are unlikely to bias downstream analyses, large numbers of mix-ups bear significant costs. If detected, these mix-ups represent substantial monetary loss and the ethical cost of sampling individuals that cannot be utilised. If undetected, sample mix-ups may result in incorrect research conclusions and suboptimal downstream decisions in applied contexts such as conservation management (Huang et al., 2013; Lohr et al., 2015). Despite these costs, and the increasing number of research projects that are generating large-scale individual genetic or genomic data, it appears common to assume that the ID of an individual is correctly associated with the right genetic information. Yet, even though some publications acknowledge that anomalies in their results

could be due to sample swaps (Li et al., 2020), many genomic studies do not appear to have implemented and/or reported a standardised approach for verifying sample identification, unless it was the main objective of the paper (Broman et al., 2015; Pedersen & Quinlan, 2017; Lobo et al., 2019).

Beyond more general genomic data quality control such as sample duplicate checks, verifying sample identity will require the utilisation of existing metadata associated with the samples in a genomic dataset. For example, for many species, morphological or behavioural information can be used to infer the sex of an individual, and if sex markers can be identified from genomic data this can provide an initial check of the minimum number of sample misidentifications. In some cases, samples may be sequenced or genotyped on multiple platforms (for example, low coverage whole genome sequencing and transcriptome sequencing), providing opportunity to identify data from shared genomic regions and check for genotype consistency between datasets. Sample verification is also greatly facilitated by a pre-existing pedigree from previously generated genetic data (for example, a panel of microsatellite markers), as is the case in many long-term monitored populations (Dugdale et al., 2007; Walling et al., 2010; Nielsen et al., 2012; Chen et al., 2016; Johnston et al., 2016; Malenfant et al., 2016; de Villemereuil et al., 2019; Fitzpatrick et al., 2019; Niskanen et al., 2020). Samples can then be checked for Mendelian consistencies between previously identified close relatives. Further, genetic or genomic data has enabled family relationships to be inferred for thousands of additional shorter-term studies (see, for example, Flanagan and Jones (2019) and references therein). Inferred genetic relationships can then be compared with data recorded at collection (for example, fledglings sampled from the same nest) and checked for compatibility.

Previous individual-based ecological studies have utilised some, although not necessarily all, available sample metadata to verify that genotyped or sequenced samples are correctly identified (see, for example, Sardell et al., 2010; Van Oers et al., 2014; Santure et al., 2015; Husby et al., 2015; Nietlisbach et al., 2015; Johnston et al., 2016; Huisman et al., 2016;

Lundregan et al., 2018; Duntsch et al., 2020; Feng et al., 2020; Cockburn et al., 2021; Debes et al., 2021; Grueber et al., 2021). Checks in these studies have included testing the consistency of genetic and morphological sex, detecting (unintended) sample duplicates, checking consistency with previously generated genomic data from the same loci, and generating additional targeted sequence data to confirm the presumed species. For populations where it is possible to infer pedigree relationships, Mendelian inheritance checks are also commonly reported. However, when we reviewed more than 200 recent publications in the field of molecular ecology (see Supplementary Material 1), we found that few individual-based studies mentioned sample checks. The most commonly employed and reported measures to mitigate sample errors were the inclusion of positive and negative controls or a duplicate check (found in 30% of the publications). However, less than ten percent of the examined studies documented at least one individual sample-ID check in their main manuscript and none of the studies reported following a standardized protocol (Supplementary Material 1).

This is, to our knowledge, because there is no sample-verification guideline available, neither for individual based ecological genomic data, nor in other applications such as eDNA or human studies where sample mix-ups have been reported (Have et al., 2014; Nicholson et al., 2020). Further, it does not appear to be standard practice to systematically report the validation of genetic, genomic or transcriptomic data and sample-ID assignments in ecological studies. This suggests that there may be numerous peer-reviewed studies that could have been impacted by sample mis-annotations, or that there are a significant amount of sample checks that go unmentioned.

Given the costs of sample mix-ups, the challenge now is to move beyond inconsistent implementation and reporting of quality control steps and to put an intuitive and systematic framework in place. Admittedly, if very little observational and genetic metadata is available, it may not be possible to infer genotypes or sequences that have been incorrectly assigned to

112 sample identifiers. However, the majority of individual-level studies in molecular ecology are
113 likely to be able to identify and also, in some cases, correct, sample mix-ups. With this in mind,
114 we have developed a framework to serve as a guideline for ecologists to quality check their
115 data and confidently identify, quantify and potentially resolve sample mix-ups. The proposed
116 standard process for ecological data checking should be universally applicable to any
117 individual-based dataset, including those that may contain morphological, location, additional
118 genetic, relationship / pedigree or other metadata.

119 Here, we present a novel sample verification framework for molecular ecologists and
120 demonstrate its application to a single nucleotide polymorphism (SNP) array dataset that was
121 intended to include genotypes of 1,256 hihi (stitchbird; *Notiomystis cincta*), a threatened bird
122 species of Aotearoa New Zealand. In early 2019, a routine quality control check comparing
123 recorded morphological sex with SNP array-inferred sex of the genotyped individuals revealed
124 a large number of discrepancies. This incident motivated the development of a framework that
125 would help researchers detect and occasionally resolve sample errors before they perform
126 downstream analyses that require individual-level data. As far as we are aware, this is the first
127 sample verification framework to provide a step-by-step guide, detailed examples and
128 additional notes on data handling pre and post analysis and we recommend implementing this
129 easy-to-follow routine to anyone dealing with individual genetic or genomic data.

130 A survey among molecular ecologists

131 In 2020, we designed a short questionnaire for researchers working with genetic data, asking
132 whether they had encountered sample mix-ups, and about their sample protocols and how
133 they detect and deal with erroneous samples. The survey was designed in *Qualtrics* and
134 included four questions, all of which allowed an optional free-text response. Participants did
135 not need to answer all questions, and for some questions, multiple options could be chosen
136 (Supplementary Material 2). We distributed the survey via email invitations and Twitter.

Participants were not limited to those working in molecular ecology, but were likely to be the majority of respondents given our networks and contacts. We had the survey open online for one month and received 303 responses, 285 of which answered and met the eligibility criteria of (i) 18 or older and (ii) frequently or occasionally working with genetic data.

Our survey results indicate that sample mix-ups occur regularly in laboratories around the world, with 79.55% (214/269 that answered this question) of respondents agreeing that they had encountered a sample mix-up in their lab. For those that have protocols in place to verify sample identity, checking population structure and sex were one of the most commonly used methods, along with checks for Mendelian consistency based on known family relationships. Participants indicated that they believe most mistakes happen in the wet lab, with tube mislabelling, inconsistent sample indexing and pipetting mistakes on genotyping plates provided as common errors made in sample processing. Further, the survey indicated that once a sample mix-up was detected, samples were usually discarded, and other samples checked. Eighty nine percent (195/219) of the participants stated that they would welcome protocols for an extra quality control step that ensures sample identity (Supplementary Material 2). Whilst our survey was open to all who met our eligibility criteria, we caution that the group of voluntary participants may be biased toward researchers motivated to report their sample mix-up experiences.

Overview of the framework

Together with our own experience of hihi sample mix-ups, our survey findings motivated this manuscript and the construction of a sample verification framework (Figure 2) to serve as a resource for the wider community of molecular ecologists at the post-data collection stage. As detailed below, we propose a framework for genomic data to identify potential sample mix-ups. We recommend initially assuming all samples are **unvalidated** and following a set of steps to identify and **flag** those whose metadata and genetic information do not agree. Samples should

ultimately be classified as **confirmed**, **corrected**, **rejected** or remaining **unvalidated**, and a decision made as to the degree of uncertainty that is acceptable in taking these samples forward for analysis. While we predominantly focus on SNP data generated from, for example, whole genome sequencing, reduced-representation resequencing or targeted SNP genotyping, the presented framework is equally applicable to transcriptome data and other molecular genetic markers such as microsatellites. We assume that all studies have already followed standard workflows to produce a high quality genomic dataset (see e.g. O'Leary et al., 2018). Sample checks that everyone can then perform and report include checking positive and negative controls, removing duplicates, identifying mixed samples and an initial analysis to infer structure in the genetic or gene expression data. Moreover, additional metadata (including species or subspecies designation, location, observed sex, cohort, year of birth and experimental control treatment) can be used to help check and cross-validate sample identity. If relationship information is available, a check for Mendelian errors and a comparison of pedigree and genomic relatedness can identify further mix-ups. Parentage assignment programs may be used to further confirm sample identities and additional genomic data may help verify genotype-ID associations. Given that all genetic and genomic datasets are unique we do not provide recommended software for each of these checks, although we provide our own choices for hihi in Supplementary Material 3.

Control, duplicate and mixture check

Where experiments have included positive and negative controls during the extraction process and genotyping step, initial checks should identify whether genetic data has been produced from a well that should technically be free from it (negative control) or if the desired target sequence has indeed been generated (positive control). Unexpected data from these wells is likely to indicate a plating error, and we recommend **flagging** but keeping this sample in the dataset in the hope its identity can be resolved. Identity by state (IBS) allele sharing can be calculated between all pairs of individuals to identify expected (to quantify the genotyping error

rate) and unexpected sample duplicates, where two genetic samples have very high levels of identity. The sample with the highest quality genotype or sequence data and correct ID can be maintained in the dataset while their duplicate can be removed. Finally, a check on per-sample heterozygosity by plotting the distribution of heterozygosities could identify samples with unusually high or low heterozygosity. High heterozygosity is likely to indicate a mixed sample (cross-contamination; e.g. due to spill-over across wells) and should in most cases be **rejected**. Individuals with very low heterozygosity may point to issues with coverage and in most cases will have been identified from the genomic quality control preceding this framework.

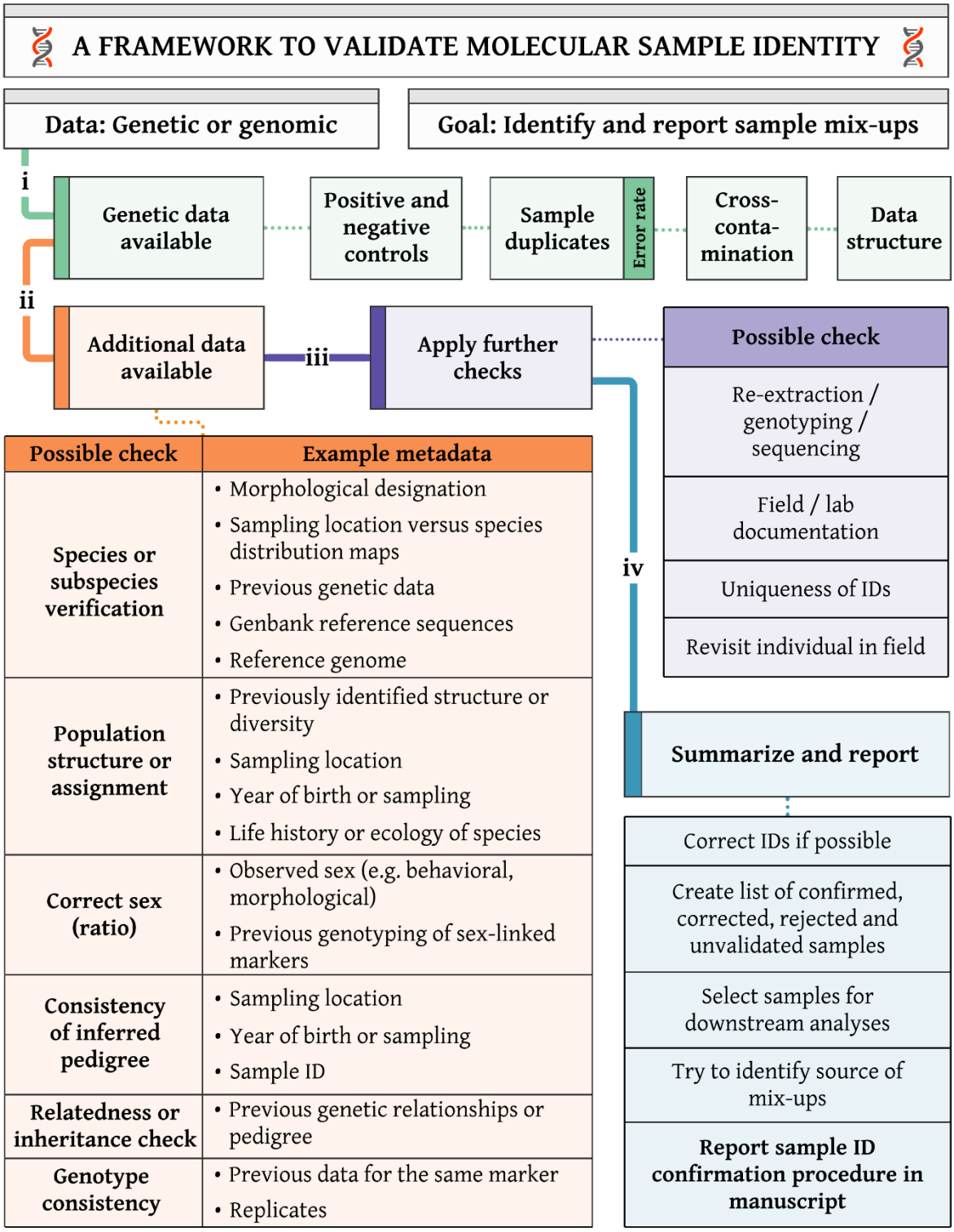

Figure 2: A molecular ecology framework to help detect genomic sample mix-ups. The framework presents common data checks (positive and negative controls, duplicates) and an analysis of data structure as universally applicable steps (i. green). The orange pathway describes sample checks if additional metadata (such as phenotypes, birth year, plate information and field notes) is available. Some studies can also draw information from previously established pedigrees or phylogenies (ii. orange) and sometimes additional genetic data (iii. purple). The goal of this framework is to make lists that contain the confirmed, corrected, rejected and unvalidated samples for future data analyses and management (iv. blue). Figure created with Lucidchart.com.

Data structure analysis

We suggest inferring genetic or transcriptomic structure, for example by running a principal component analyses (PCA) on all samples. This will enable an initial check to determine whether individuals fall into clearly defined clusters as might be expected between different treatment groups in a transcriptomic study. In the case that genetically differentiated individuals or groups are identified, and in the absence of *any* other sample information and/or expectation of genetic structuring, this analysis might suggest that e.g. individuals from cryptic species have unintentionally been sampled, or the sample has been contaminated. These samples should be **flagged** and treated carefully in further analyses.

Sample checks with additional metadata

In many cases, additional metadata can be leveraged to check sample identities. Observations from many wild populations include documentation of the (presumed) sex of an individual, whether it is an adult or juvenile, its location and a date of sampling. Other metadata including morphological, life-history, relationship and previous genetic data can also be utilised to check and confirm sample identity. We note that discrepancies between genetic data and the metadata they relate to may in fact reveal inaccuracies in the metadata or in the assumptions underlying that data, e.g. it may be assumed a species is monogamous but genomic data suggests instances of extra pair paternity, or a species presumed not to be migratory appears outside its range. We therefore recommend also treating metadata with some caution, and assessing whether there is enough evidence to firmly reject an ID-genetic sample association if it is discordant with this data.

Species or subspecies verification

In many cases, individuals will have been identified to species or subspecies level during collection, based on e.g. morphology or occurrence within known species ranges. Population structure and assignment plots can help verify whether individuals group consistently

compared to expectation. Where sequence data from the present experiment is available, sequences can be compared to previous genetic data from the same species, available genome assemblies or reference databases such as Genbank. Low coverage whole genome resequencing, sequence capture and even reduced representation sequencing may capture mitochondrial genome sequences (Stobie et al., 2019; Allio et al., 2020), with confirmation that this mitochondrial sequence matches the expected species being particularly useful in phylogenomic studies, where the samples are distantly related. The quality and proportion of reads mapping to an existing reference genome can also confirm species identity. If multiple species have been sampled, a sample mix-up would be easily identifiable if it appears in a very different clade within a phylogenetic tree and should be **rejected**, although in some cases it may be possible to **correct** these samples. There is the option to confirm morphological identifications by using BLAST or to genotype additional genomic loci to identify misidentifications within the sampled pool. When individuals have been sampled across various geographic locations, population clustering can be verified based on the genetic data with published distribution maps, and if samples cluster unusually, especially if migration between locations is not possible, it can be **flagged** as a potential sample misidentification. Additionally, datasheets containing all measured traits and metadata should be checked for recording errors and consistent data entry.

Population structure analysis

In almost all cases, sampling location will be recorded or available, and/or there will be some previous knowledge of the ecology or genetic structure of the species. In this case, the inference of population structure (see above) or genetic assignment analysis will enable a check to determine whether individuals fall into defined populations or groups as might be expected from previous work, such as from mitochondrial haplotype networks or spatial or temporal structure inferred from microsatellite genotyping. Previous genetic knowledge about sub-populations of interest can also help **confirm** sample IDs. For instance, summary statistics

such as the relative genetic diversity of different cohorts, ages or locations can be compared across new and old datasets. Strong structure may also be expected based on the life history or ecology of the species. For example, in the case where there is an expectation that genotyped populations have little or no gene flow, individuals from one sampling site clustering with another can reveal sample mix-ups across locations, and is an indication that within-population mix-ups are also likely to have occurred. Population structure across time may also be expected for e.g. species with sweepstakes reproductive success, with a lack of expected structure similarly indicating that sample mix-ups may have occurred. Individuals that clearly fall in the wrong cluster can be **flagged** (and **rejected** if there is no known migration or other process that would explain this).

Sex check

For species with some evidence of genetic sex-determination, sex-linked markers may be known based on previous work. Heterozygosity at these markers can be used to distinguish the homogametic (very low or no heterozygosity) and heterogametic (higher heterozygosity) sex. Putative sex-linked markers can sometimes be identified de novo based on unusual genotype frequencies or alignment of sequencing reads to a reference genome of the species or a closely related species where the sex chromosome has been identified. In some species morphological or observational data provides unambiguous sex for an individual that the genetic sample can be checked against. Individuals where there is confidence in the recorded sex, and the recorded and newly assigned genetic sex differ, should in most cases be **rejected**, although in some cases where it may be possible to re-visit an individual in the field and check their sex (e.g. a banded bird), the observational sex can be corrected. In addition, if the proportion of male and female individuals inferred from the genomic data is significantly different from the proportions that were expected when selecting samples for genotyping, this may also indicate that sample mix-ups have occurred.

Metadata consistency with inferred pedigree

For populations without intensive monitoring or where samples are anonymous (e.g. faecal sampling), pedigrees or relationships are unlikely to be known or are not previously genetically verified. In these cases, pedigree construction from the current genomic data is a useful approach to help validate sample identities, particularly when knowledge of the spatial or temporal sampling of individuals can exclude the possibility of first-degree relationships between pairs or groups of individuals. For example, a parent-offspring relationship between two individuals sampled ten years apart is unlikely in a short-lived species and one or both of these individuals should be **rejected**. For many species, co-occurrence of individuals, for example offspring in a nest or seedlings surrounding a plant, can indicate potential first-degree relationships, for example between siblings or mother-offspring, often indicated by consecutive numbering of samples. Inconsistency with these observed putative relationships may not necessarily indicate that the sample needs to be rejected, but consistency with this spatial or temporal metadata such as sample location, year of birth and sampling and sample naming can help **confirm** sample IDs.

Relatedness and inheritance check

For some populations, particularly those for which individuals are tracked (e.g. through marking or banding), robust information about relationships may be known from previous genetic work. Where a verified genetic pedigree is available, detection of errors requires that the new genomic data and the original DNA (that was used to build the pedigree in the first place) do not come from the same, potentially erroneous, individual sample. Once the new genetic or genomic data has been obtained, genotyped individuals can be checked for genetic compatibility with their parents by counting the number of Mendelian inheritance errors for autosomal SNPs. When numerous genotypic mismatches are observed between the offspring and only one previously genetically verified parent, the parent sample should be **rejected**, while individuals that show similarly high numbers of mismatches with each parent should be

rejected. We note that inconsistencies with a previous pedigree may reflect lower power of a previous genetic dataset (e.g. a set of microsatellite markers) to correctly infer relationships. In this case, additional metadata may help resolve the true relationships.

Building a custom matrix

Where a multi-generational genetic pedigree is available, we propose to build a custom matrix, based on pairwise genomic and verified pedigree relatedness (Table 1). Rows represent all the genotyped individuals. The columns represent each individuals' genotyped parents, full-sibs and offspring based on the verified pedigree. Values in each cell are calculated based on the pedigree and genomic relatedness between the focal individual (row name) and each of their first-degree relatives' ID. The genomic relatedness estimate can be chosen to be broadly consistent with the range of pedigree relatedness values, or standardised to similar values. Standardising the range of genomic relatedness to pedigree relatedness enables inbreeding to be taken into account, as both expected (pedigree) and realised (genomic) relatedness between first-degree relatives can exceed 0.5 (Hedrick & Lacy, 2015). Relatedness thresholds to designate related versus unrelated can be chosen based on the distribution of the difference between pedigree and (standardised) genomic relatedness estimates for verified parent-offspring and siblings from the Mendelian error check when both parents are genotyped. Relatedness values for parents and offspring will have smaller variance than for full sibs, but in most cases and with sufficient numbers of informative markers the distribution of first-degree relatives is relatively distinct from that of second or higher degree relatives (Städele & Vigilant, 2016; Galla et al., 2020). Note that this approach will not detect sample mix-ups among full-siblings if they themselves do not have offspring and their genetic and recorded sex are concordant, but mix-ups in most other cases should be identifiable and these individuals **rejected**.

Table 1: Three example rows from a matrix with pairwise genomic relatedness values, and the difference between expected pedigree and genomic relatedness, between focal individuals A, B and C and their first-degree relatives (e.g. with F = father). In the case of individual A, a very low relatedness value with their mother (M) but relatedness

consistency with siblings (S) and offspring (O) suggest that the mother is a sample mix-up. For individual B, relatedness inconsistencies with all genotyped relatives suggest that individual B is a mix-up. All available pedigree relatedness values for individuals A and B are 0.5 (i.e. there is no inbreeding). For individual C, despite very high relatedness values reflecting extensive inbreeding in their pedigree, consistency between pedigree and genomic relatedness indicates no mix-up. NA designates ungenotyped relatives. Note: The pedigree and genomic relatedness values are taken from the worked hihi example as mentioned in the Supplementary Materials.

Individual	Pedigree relatedness						Genomic relatedness						Pedigree – genomic relatedness					
	M	F	S1	S2	O1	O2	M	F	S1	S2	O1	O2	M	F	S1	S2	O1	O2
A	0.5	0.5	0.5	NA	NA	NA	0.06	0.45	0.49	NA	NA	NA	0.44	0.05	0.01	NA	NA	NA
B	0.5	0.5	0.5	0.5	0.5	0.5	-0.01	-0.04	-0.02	0.02	-0.05	0.06	0.51	0.54	0.52	0.48	0.55	0.44
C	0.84	0.78	0.81	NA	NA	NA	0.78	0.76	0.78	NA	NA	NA	-0.14	-0.02	0.03	NA	NA	NA

The matrix may also be extended to include self-self relatedness, as highly inbred individuals should be expected to have both high pedigree and genomic relatedness values, and a large discordance between these values may indicate a sample mix-up. In addition, more distant relatives could be included in the matrix, although we caution that the variance in the difference between pedigree and genomic relatedness for these relationships may be too high to confidently confirm sample identity. If time is of essence, plotting genomic versus pedigree relatedness between all individuals (including individuals with themselves) can reveal individuals with pairwise genomic relatedness values that are very high or very low compared to their expected pedigree relatedness with others. These individuals can then be identified, **flagged** or **rejected** from the dataset. Note that given the large variance in relatedness estimates from a small number of markers, such as a microsatellite panel (Santure et al., 2010; Galla et al., 2020), we do not recommend directly comparing genomic and microsatellite relatedness to validate sample identity.

Cross-validation

As a last step for populations where a genetic pedigree is available, we suggest to cross-validate that the **confirmed** samples are indeed assigned to the correct genotypes. For example, between all confirmed individuals and each class of first-degree relatives, we recommend creating a scatterplot to compare the distribution of pedigree-based relatedness

to that of the genomic relatedness and visually inspect for any outliers. This procedure is also an additional way of double-checking that all **rejected** ID-genotype associations for certain individuals are indeed false, or at least different from the verified pedigree relationships.

A final step to cross-validate individuals whose relatedness and inheritance checks support a correct ID-genotype association is to use the new genomic data to reconstruct the pedigree. Doing so can serve three purposes. First, as noted above, it can be used to validate the pedigree relationships, and hence sample IDs, that may have been assigned based on using fewer markers e.g. using a microsatellite dataset. This again implies that the different datatypes do not come from the same DNA extraction, in the case that the sample has been mixed up early on. Second, pedigree reconstruction may enable the identification of further sample mix-ups that were not apparent because individuals did not previously have any close pedigree relatives genotyped. In particular, if an individual is assigned as a parent, offspring or sibling to another individual in the dataset, but none of the original ID's observed relatives were included in the genotyping, it is possible this individual has been mixed-up and hence should be **rejected**. Further, newly constructed pedigrees may be able to assign individuals new **corrected** sample IDs. For example, transposition of digits or numbers in IDs can easily occur in the field or lab. If individual P1009, identified in a matrix (such as Table 1) as a sample mix-up, is assigned as a parent to an individual with pedigree father P1090, further checks (correct sex, relatedness consistency with other genotyped relatives) can be done to determine whether the sample ID has been incorrectly recorded and can be confidently **corrected** to P1090. Individuals passing all the above quality control steps are suitable for all downstream analyses, while **rejected** IDs could become subject to revision of their pedigree, re-sampling, or re-sequencing where possible.

Genotype consistency

Previous or additional genomic data including reduced representation, low-coverage or even high-coverage whole-genome sequencing all provide a means of **confirming** the sample ID

for the current genotypes or sequences under investigation. Where there is substantial overlap between genotyped regions (for example, a SNP array was designed using variation identified from low coverage whole genome sequencing, or individuals have been previously sequenced for a mitochondrial region that can also be identified from the current genomic data), genotypes from identical loci can be compared across platforms to check IBS sharing of an individual with itself or close relatives. Moreover, sequence data is often delivered as multiple sequencing runs per individual, hence replicates between flow cells or sequencing lanes can help to check for genotype consistency.

Additional checks

As outlined above, generating additional sequence data can be one way to help confirm the identity of samples. In addition, patterns during sample collection and processing, recorded in lab or field documentation, can also serve as an initial sample check (e.g. when certain labels have only been assigned on certain days to certain individuals). Lab and field notes can also be carefully checked to ensure that a putative sample mix-up is not due to e.g. individual IDs being reused across seasons or sites. If sex allocations are uncertain, and the sample population can be re-visited, additional observations can help re-assign or confirm a sex, at least if the individual is identifiable in the population, and the species shows sexual dimorphism. The individual can also be resampled and resequenced or regenotyped. Finally, for all analyses we recommend the comparison of multiple software outputs (such as clustering algorithms, inference of sex or parentage checks) in order to give further confidence that confirmed samples are in fact reliable.

Summarise and report

A final step is to prepare a summary table of **confirmed**, **corrected**, **unvalidated** and **rejected** ID-genetic data associations and a comprehensive summary of why these samples were confirmed, corrected, unvalidated or rejected. This information will help tailor datasets for

410 further analyses. For example, ‘unvalidated’ individuals that cluster in their expected cohort
411 may still be useful for inference of population structure, and if individuals in a phylogenetic
412 study have only been mixed up within species, they will still cluster as one group that is
413 separated from the other clades. On the other hand, as soon as pedigrees or family
414 relationships are being investigated or correlated, or genotype - phenotype analyses such as
415 association mapping are planned, correct individual genotype-ID associations are paramount.
416 Overall, it depends on the particular study question whether it is important to correctly identify
417 to the species, the population or the individual level.

418 If sources of error have been identified, appropriate measures should be put into place in order
419 to minimize the risk of re-occurring sample mix-up incidences in the future. Re-sampling and
420 re-sequencing where (ethically) possible are, of course, an alternative way of making sure the
421 correct individuals are being genotyped or sequenced. In order to standardize field and
422 laboratory protocols, we strongly encourage molecular ecologists to use this framework as a
423 mandatory process any time data is being analysed, and to report all performed sample checks
424 in the methods or supplementary material of their publication.

425 Identifying sources of error

426 As outlined above, a number of the steps in the framework can identify likely sample mix-ups.
427 In some cases it may be possible to identify when these errors occurred, such as transcription
428 errors in the field or lab. In a bird population, for example, sample mix-ups may be more likely
429 to happen (and are harder to detect) between relatives if individual bird samples are collected
430 within a nest setting. On the other hand, individual sampling by mist netting migrants or culling,
431 may mean that sampling occurs more randomly and as a result, the error will be less biased
432 by relationship. In both cases, the sampled individuals will most likely have sequential sample
433 numbers, defined by those temporal and spatial factors. Additionally, one can try to trace back
434 the sample mix-ups to specific extraction, genotyping or sequencing plates used in the wet lab
435 in order to narrow down the number and source of samples that were possibly affected, which

436 can also help identify the cause for the sample mix-up (Broman et al., 2015). For example,
437 errors arising from a small number of genotyping plates may indicate a systematic problem
438 with one cohort of samples or one single step in the laboratory procedure. In some cases, it
439 might even be feasible to generate duplicate genetic or genomic data for a small number of
440 focal individuals to check whether data is consistent with previous genotyping. Consistent data
441 indicates that the original samples may be mislabelled (perhaps due to a field error), while
442 inconsistent data may point to an error during or after the most recent wet lab process (perhaps
443 due to sample mis-plating).

444 Implementation: the hihi project

445 A total of 1,536 hihi individuals from five populations were genotyped on a custom 50K SNP
446 array (Lee et al., 2021). To demonstrate the implementation of our framework, we focus on the
447 verification of the 1,256 individuals from the reintroduced population of Tiritiri Mātangi
448 (36°36'8"S, 174°53'13"E) presumed to be included on the array. These individuals had
449 extensive metadata available, including multi-generational pedigree information previously
450 verified using microsatellite markers (Brekke et al., 2013; de Villemereuil et al., 2019).
451 Following individual and marker data control, we removed six unexpected duplicate samples.
452 No positive or negative controls were included and could not be checked. Principal component
453 analyses of individuals from all five populations indicated that population structure was weak,
454 but did seem to indicate that Tiritiri Mātangi individuals were clustering as expected by
455 population. We checked for and detected 126 inconsistencies between recorded and genetic
456 sex in the Tiritiri Mātangi individuals.

457 **Table 2:** Summary table of all the confirmed and rejected sample-genotype associations after following the steps
458 of the suggested framework with the 1,256 hihi genotypes from Tiritiri Mātangi. 488 IDs were 'confirmed' through
459 parentage assignment and having the correct sex and relatedness with other close relatives. 42 samples had two
460 parentage assignment softwares agreeing on a different parental pair than the validated pedigree. Based on these
461 assignments, 20 of these samples could unambiguously be assigned a new ID and are shown as 'corrected' while
462 the other 22 were 'rejected' The remaining 'rejected' samples were duplicates, had a different parental pair in all

pedigrees or were the wrong sex. Unvalidated samples were the correct sex but did not have enough additional information (i.e., few or no genotyped close relatives) available to be categorized in any way.

Sample status	*Confirmed*	*Corrected*	*Rejected*	*Unvalidated*
Number of samples	488	20	256	492

Comparisons between genomic and validated microsatellite pedigree relatedness revealed more than one hundred hihi individuals with near-zero relatedness towards all their (expected) relatives. We used two pedigree reconstruction softwares to reconstruct and check pedigree relationships for all Tiritiri Mātangi individuals. A total of 508 samples could be confirmed or corrected, although a further 256 are clearly incorrectly labelled and 492 hihi could not be validated (Table 2; detailed methods in Supplementary Material 3). Even though hundreds of animals remained unvalidated, the implementation of our framework allowed for more than one third (39%) of the Tiritiri Mātangi hihi in this genotyping project to be confirmed (Figure 3). These confirmed individuals have been reliably used for downstream population analyses for this threatened species (Duntsch et al., 2020; Duntsch et al., 2021).

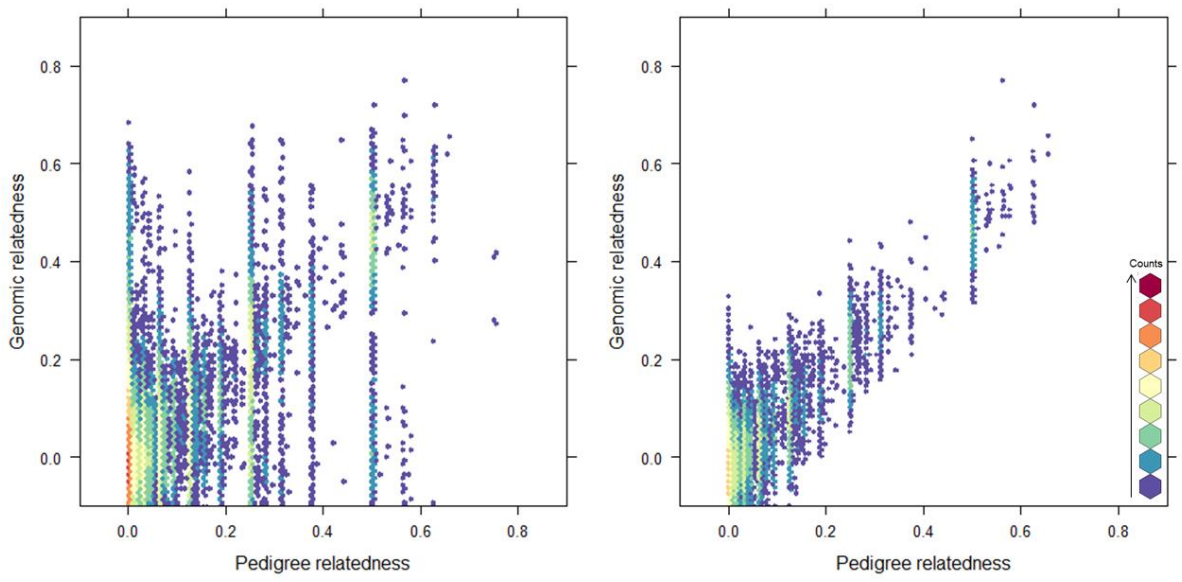

Figure 3: The association between genomic and pedigree relatedness of the hihi on Tiritiri Mātangi for all unique samples (right panel, N=1,250) and only the confirmed samples (right panel, N=488). The pedigree-based relationship matrix was calculated using the R package *kinship2* (Sinnwell et al., 2014), the genomic relationships were calculated with the tool *GCTA* (Yang et al., 2011). The warmer the colour, the more pairs show a specific relatedness, with most pairs being unrelated. In the left panel, some individuals show high pedigree yet no genomic

relatedness, or low pedigree with high genomic relatedness, an indication of sample error. These erroneous links have disappeared after sample checking (right panel).

After re-tracing the entire sampling and sequencing process, we were able to narrow down a potential origin of the sample mix-ups of the hihi genotypes. Because the same set of samples had previously been extracted and microsatellite genotyped (Brekke et al., 2013; de Villemereuil et al., 2019) and in this previous work the vast majority of observational and genotyped mother-offspring relationships were in agreement with each other, sample errors are unlikely to have occurred in the field. Therefore, we suspect that most errors in our SNP datasets occurred in the wet lab during the re-extraction of samples for the SNP array genotyping.

Discussion

Here, we present a framework for the use of genetic data and additional metadata to check sample IDs, and apply it to validate the sample identity of over 500 hihi individuals genotyped on a SNP array. The framework is designed to guide and encourage researchers to routinely implement and report an additional quality control step into their data processing routine. Incorrect ID-genetic data links lower the robustness and power of a study and can possibly corrupt many of the underlying statistics and assumptions (Huang et al., 2013; Lohr et al., 2015). Hence, it is important to standardize protocols for data sampling and handling and ensure detailed documentation (e.g. online data sheets, shared drives, data sharing platforms such as the Genomics Observatory Metadatabase (GEOME; https://geome-db.org; Riginos et al., 2020)). For individual-based research in particular, such as genotype-phenotype analyses or inbreeding depression studies, it is crucial to be able to correctly match the phenotype of individuals with their genotype. This is particularly important when re-genotyping or re-sampling is difficult, especially in small laboratories where funding is scarce, or when the raw

samples are no longer available to be re-analysed (e.g., the original sample has been used up).

With more and more data being generated in laboratories all over the world, now is the time to develop standardised protocols as a resource for the wider science community. Our recent survey showed that sample errors have occurred in most laboratories (80%), and nearly 90% of the participants stated that they would welcome protocols for an extra quality control step that ensures sample identity. We encourage researchers to consistently document the results of their sample quality control in publications, in the same way that sequence and marker quality control is routinely reported. This will avoid the same problems being tackled by different researchers independently, reveal common and significant mistakes, improve the exchange of novel applications and methods, and finally contribute to more transparent research and reliable publications.

The sample verification framework

Although genomic data quality checks are relatively standard in molecular ecology research, we found very few studies that consistently report sample data checks. Most ecological studies will contain useful metadata that can also be leveraged to check sample identity. In addition, when some relationships are known from field observation or previous genetic data, there is the option to compare pedigree-based and genomic relatedness of the individuals in order to identify erroneous samples. Unfortunately, we were unable to identify a publicly available tool that reliably (and in a straightforward manner) checks for Mendelian errors across all close relatives. As this is an intuitive method when wanting to detect pedigree errors, we propose the construction of a custom-built relatedness matrix to check for inconsistencies between datasets, until a more appropriate tool becomes available. Approaches that estimate identity by descent sharing to classify more distant relatives are also likely to be helpful, for example when wanting to distinguish full sibs from parent-offspring (Waples et al., 2019).

After designing a framework to confirm sample identity with existing genomic data by comparing with our recorded metadata (including sex, location and pedigree relationships), we applied this protocol to 1,256 genotyped hihi samples. We were able to resolve more than a third of the samples and can therefore be confident about those individuals in our analyses. Our hihi dataset presents one of those scenarios where samples are scarce and precious, and re-genotyping of the individuals is simply not feasible from a financial perspective. If the sample mix-up had remained undetected, any downstream analysis would be biased and not representative of the true nature of evolutionary processes, such as inbreeding depression, in the examined population. Application of this framework has enabled us to create a smaller, yet reliable genomic dataset that has been used to quantify the adaptive potential of this threatened species (Duntsch et al., 2020; Duntsch et al., 2021).

Where do sample mix-ups happen?

An additional step when investigating sample mix-ups is to determine where in the data processing pipeline the mix-up may have occurred in order to prevent them from happening in the future. This can turn out to be a difficult task if the study system is lacking additional information such as a verified pedigree and phenotypic information or if the data handling procedures are not well documented and the genotyping technology unexplored (Have et al., 2014). Sampling errors can happen from the moment the sample was taken in the field, during any stage of transportation and storage, in any step of the wet lab procedures and up to the moment when the bioinformatics processing commences (Figure 1). Our survey shows that researchers believe the majority of mistakes happen in the wet lab, meaning that human errors such as tube mislabelling, inconsistent sample indexing and pipetting on genotyping or sequencing plates may be common errors made in the processing of genetic data.

Avoiding sample mix-ups

Ideally, of course, sample mix-ups are avoided throughout the entire process from data collection to research publication. In the field and before transport, it helps to regularly scratch sample ID and sampling date onto the sampling tubes with a pin or needle to avoid poor legibility or an accidental removal of labels. In the wet lab, one could move Eppendorf tubes from which or into which a sample was pipetted into a new rack to avoid pipetting from it (or into it) more than once. To avoid mix-ups within a genotyping plate, one could use aluminium plate covers when pipetting between plates and use the tip to puncture the plate cover, as well as make use of multi-channel pipettes to transfer samples between plates. Plate orientation should be carefully checked at each step of a protocol and when transferring between plates. Generally, it is advisable to consistently include positive and negative controls throughout the entire sample preparation and sequencing process, particularly when genetic work is being outsourced to external sequencing facilities, to minimise the number of sample handlers and to be careful with data entry, sorting and transfer. Further, it may be helpful to check samples received from collaborators or that have been in storage for some time – for example, amplifying a barcode sequence to confirm species identity, or amplifying previously identified sex markers to confirm individual sex.

At a time where genomic data generation is not the limitation anymore, it is becoming more and more important to ensure thorough documentation and to standardise and share as many of the common lab practises as possible, to allow the early detection of sample errors. Interestingly, the scientific community has recently become more aware of the benefits of standardised processes across research groups, universities and countries. Biological meta-databases like GEOME (Riginos et al., 2020) and *Ira Moana* (https://sites.massey.ac.nz/iramoana/; Liggins et al., 2021) collect commonly employed research methodology, promote data reusability and study reproducibility while providing templates and recommendations for study design and execution. What is more, there are

open-source platforms such as *Galaxy* (https://usegalaxy.org/), which aim to provide tools for researchers who are working with genomic data.

Conclusion

This project highlights the potential for samples to be resolved, but most importantly demonstrates the potential to detect the sample errors that inevitably can happen. While we can never fully avoid human error, we can certainly employ methods in order to make sure that those sample mix-ups, mislabelling and plating errors are identified, corrected and accounted for. In this paper, we developed a framework for working with individual genomic data samples, have explored the properties of a dataset that has undergone a major sample mix-up and demonstrated the potential of detecting (or neglecting) sample errors with regard to downstream analyses. We strongly recommend that a sample verification step is implemented into any data quality control routine of laboratories around the world, and identified errors routinely reported. Taking this extra measure of caution early in the sample handling process will prove to be crucial not only to adjust for human error and consequently reduce data processing costs, but also to be able to correctly inform ecological studies, inform conservation management and other applied outcomes and further contribute to a transparent science community.

Acknowledgements & Funding

We acknowledge Ngāti Manuhiri as Mana Whenua and Kaitiaki of Te Hauturu-o-Toi and its taonga, including hihi. We extend many thanks to the volunteers, past students, and Department of Conservation staff who have contributed to monitoring the Tiritiri Mātangi hihi population. We thank the Hihi Recovery Group, Department of Conservation, and Supporters of Tiritiri Mātangi for maintaining such a long-term vision in monitoring and management of this population, with a special mention of the current hihi conservation officer Mhairi McCready.

Thanks to Johanna Nielsen and Kang-Wook Kim for help with microsatellite genotyping. We thank Kate Lee for developing the hihi SNP array and early discussions about resolving sample mix-ups. We acknowledge the use of New Zealand eScience Infrastructure (NeSI) high-performance computing facilities and thank GEOME for the inspiration to standardize protocols. We would like to mention the extraordinary responsiveness of some of the creators of the software employed, and the helpfulness of the science community on *stackoverflow* and *researchgate*. The survey has been approved by the University of Auckland Human Participants Ethics Committee (UAHPEC2528), and we extend many thanks to all survey participants for kindly taking the time to share their experiences. We particularly thank Steve Goldstein and Christine Couldrey who saw the survey and took time to contact us directly and speak to us of their experiences with genomic samples. A Marsden Grant (UOA1408) awarded to A.W.S. from the New Zealand Royal Society Te Aparangi supported A.W.S., P.B., and J.G.E. The High Quality Genomes and Population Genomics project of Genomics Aotearoa and a New Zealand National Science Challenge Biological Heritage Project Grant, Project 1.4, also supported A.W.S. P.B and J.G.E are supported by Research England. A University of Auckland Doctoral Scholarship and Centre for Biodiversity and Biosecurity writing stipends supported L.D. Finally, we sincerely thank three reviewers for their helpful and insightful suggestions that significantly improved both the manuscript and the framework.

References

Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., & Delsuc, F. (2020). MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Molecular Ecology Resources, 20*(4), 892-905. doi:10.1111/1755-0998.13160

Brekke, P., Cassey, P., Ariani, C., & Ewen, J. G. (2013). Evolution of extreme-mating behaviour: patterns of extrapair paternity in a species with forced extrapair copulation. *Behavioral Ecology and Sociobiology, 67*(6), 963-972. doi:10.1007/s00265-013-1522-9

Broman, K. W., Gatti, D. M., Svenson, K. L., Sen, Ś., & Churchill, G. A. (2019). Cleaning Genotype Data from Diversity Outbred Mice. *G3: Genes|Genomes|Genetics, 9*(5), 1571. doi:10.1534/g3.119.400165

Broman, K. W., Keller, M. P., Broman, A. T., Kendziorski, C., Yandell, B. S., Sen, Ś., & Attie, A. D. (2015). Identification and Correction of Sample Mix-Ups in Expression Genetic Data: A Case Study. *G3: Genes|Genomes|Genetics, 5*(10), 2177-2186. doi:10.1534/g3.115.019778

Chen, N., Cosgrove, E J., Bowman, R., Fitzpatrick, J W., & Clark, A G. (2016). Genomic Consequences of Population Decline in the Endangered Florida Scrub-Jay. *Current Biology, 26*(21), 2974-2979. doi:10.1016/j.cub.2016.08.062

Cockburn, A., Peñalba, J. V., Jaccoud, D., Kilian, A., Brouwer, L., Double, M. C., . . . van de Pol, M. (2021). hiphop: Improved paternity assignment among close relatives using a simple exclusion method for biallelic markers. *Molecular Ecology Resources, 21*(6), 1850-1865. doi:10.1111/1755-0998.13389

de Villemereuil, P., Rutschmann, A., Lee, K. D., Ewen, J. G., Brekke, P., & Santure, A. W. (2019). Little Adaptive Potential in a Threatened Passerine Bird. *Current Biology, 29*(5), 889-894.e883. doi:10.1016/j.cub.2019.01.072

Dugdale, H. L., Macdonald, D. W., Pope, L. C., & Burke, T. (2007). Polygynandry, extra-group paternity and multiple-paternity litters in European badger (*Meles meles*) social groups. *Molecular Ecology, 16*(24), 5294-5306. doi:10.1111/j.1365-294X.2007.03571.x

Duntsch, L., Tomotani, B. M., de Villemereuil, P., Brekke, P., Lee, K. D., Ewen, J. G., & Santure, A. W. (2020). Polygenic basis for adaptive morphological variation in a threatened Aotearoa| New Zealand bird, the hihi (*Notiomystis cincta*). *Proceedings of the Royal Society B, 287*(1933), 20200948. doi:10.1098/rspb.2020.0948

Duntsch, L., Whibley, A., Brekke, P., Ewen, J. G., & Santure, A. W. (2021). Genomic data of different resolutions reveal consistent inbreeding estimates but contrasting homozygosity landscapes for the threatened Aotearoa New Zealand hihi. *Molecular Ecology, 30*(23), 6006-6020. doi:10.1111/mec.16068

Feng, S., Stiller, J., Deng, Y., Armstrong, J., Fang, Q., Reeve, A. H., . . . Zhang, G. (2020). Dense sampling of bird diversity increases power of comparative genomics. *Nature, 587*(7833), 252-257. doi:10.1038/s41586-020-2873-9

668 Fitzpatrick, S. W., Bradburd, G. S., Kremer, C. T., Salerno, P. E., Angeloni, L. M.,
669 & Funk, W. C. (2019). Genetic rescue without genomic swamping in wild
670 populations. *bioRxiv*, 701706. doi:10.1101/701706
671 Flanagan, S. P., & Jones, A. G. (2019). The future of parentage analysis: From
672 microsatellites to SNPs and beyond. *Molecular Ecology, 28*(3), 544-567.
673 doi:10.1111/mec.14988
674 Galla, S. J., Moraga, R., Brown, L., Cleland, S., Hoeppner, M. P., Maloney, R. F., .
675 . . Steeves, T. E. (2020). A comparison of pedigree, genetic and genomic
676 estimates of relatedness for informing pairing decisions in two critically
677 endangered birds: Implications for conservation breeding programmes
678 worldwide. *Evolutionary Applications, 13*(5), 991-1008.
679 doi:10.1111/eva.12916
680 Grueber, C. E., Farquharson, K. A., Wright, B. R., Wallis, G. P., Hogg, C. J., &
681 Belov, K. (2021). First evidence of deviation from Mendelian proportions in
682 a conservation programme. *Molecular Ecology, 30*(15), 3703-3715.
683 doi:10.1111/mec.16004
684 Have, C. T., Appel, E. V. R., Grarup, N., Hansen, T., & Bork-Jensen, J. (2014).
685 Identification of Mislabeled Samples and Sample Mix-ups in Genotype Data
686 using Barcode Genotypes. *International Journal of Bioscience,*
687 *Biochemistry and Bioinformatics, 4*(5), 355-360.
688 doi:10.7763/IJBBB.2014.V4.370
689 Hedrick, P. W., & Lacy, R. C. (2015). Measuring relatedness between inbred
690 individuals. *J Hered, 106*(1), 20-25. doi:10.1093/jhered/esu072
691 Huang, J., Chen, J., Lathrop, M., & Liang, L. (2013). A tool for RNA sequencing
692 sample identity check. *Bioinformatics, 29*(11), 1463-1464.
693 doi:10.1093/bioinformatics/btt155
694 Huisman, J., Kruuk, L. E., Ellis, P. A., Clutton-Brock, T., & Pemberton, J. M.
695 (2016). Inbreeding depression across the lifespan in a wild mammal
696 population. *Proceedings of the National Academy of Sciences, 113*(13),
697 3585-3590. doi:10.1073/pnas.1518046113
698 Husby, A., Kawakami, T., Ronnegard, L., Smeds, L., Ellegren, H., & Qvarnstrom,
699 A. (2015). Genome-wide association mapping in a wild avian population
700 identifies a link between genetic and phenotypic variation in a life-history
701 trait. *Proceedings of the Royal Society B, 282*(1806), 20150156.
702 doi:10.1098/rspb.2015.0156
703 Johnston, S. E., Berenos, C., Slate, J., & Pemberton, J. M. (2016). Conserved
704 Genetic Architecture Underlying Individual Recombination Rate Variation in
705 a Wild Population of Soay Sheep (*Ovis aries*). *Genetics, 203*(1), 583-598.
706 doi:10.1534/genetics.115.185553
707 Lee, K. D., Millar, C. D., Brekke, P., Whibley, A., Ewen, J. G., Hingston, M., . . .
708 Santure, A. W. (2021). The design and application of a 50 K SNP chip for a
709 threatened Aotearoa New Zealand passerine, the hihi. *Molecular Ecology*
710 *Resources, n/a*(n/a). doi:10.1111/1755-0998.13480
711 Li, J., Lee, M., Davis, B. W., Lamichhaney, S., Dorshorst, B. J., Siegel, P. B., &
712 Andersson*, L. (2020). Mutations upstream of the TBX5 and PITX1
713 transcription factor genes are associated with feathered legs in the
714 domestic chicken. *Molecular Biology and Evolution*.
715 doi:10.1093/molbev/msaa093

Liggins, L., Noble, C., & , T. I. M. N. (2021). The Ira Moana Project: A Genetic Observatory for Aotearoa's Marine Biodiversity. *8*(1713). doi:10.3389/fmars.2021.740953

Lobo, A. K., Traeger, L. L., Keller, M. P., Attie, A. D., Rey, F. E., & Broman, K. W. (2019). Identification of sample mix-ups and mixtures in microbiome data in Diversity Outbred mice. *bioRxiv*, 529040. doi:10.1101/529040

Lohr, M., Hellwig, B., Edlund, K., Mattsson, J. S. M., Botling, J., Schmidt, M., . . . Rahnenführer, J. (2015). Identification of sample annotation errors in gene expression datasets. *Archives of toxicology, 89*(12), 2265-2272. doi:10.1007/s00204-015-1632-4

Lundregan, S. L., Hagen, I. J., Gohli, J., Niskanen, A. K., Kemppainen, P., Ringsby, T. H., . . . Jensen, H. (2018). Inferences of genetic architecture of bill morphology in house sparrow using a high-density SNP array point to a polygenic basis. *Molecular Ecology, 27*(17), 3498-3514. doi:10.1111/mec.14811

Malenfant, R. M., Coltman, D. W., Richardson, E. S., Lunn, N. J., Stirling, I., Adamowicz, E., & Davis, C. S. (2016). Evidence of adoption, monozygotic twinning, and low inbreeding rates in a large genetic pedigree of polar bears. *Polar Biology, 39*(8), 1455-1465. doi:10.1007/s00300-015-1871-0

McClure, M. C., McCarthy, J., Flynn, P., McClure, J. C., Dair, E., O'Connell, D. K., & Kearney, J. F. (2018). SNP Data Quality Control in a National Beef and Dairy Cattle System and Highly Accurate SNP Based Parentage Verification and Identification. *Frontiers in Genetics, 9*(84). doi:10.3389/fgene.2018.00084

Nicholson, A., McIsaac, D., MacDonald, C., Gec, P., Mason, B. E., Rein, W., . . . Hanner, R. H. (2020). An analysis of metadata reporting in freshwater environmental DNA research calls for the development of best practice guidelines. *Environmental DNA, 2*(3), 343-349. doi:10.1002/edn3.81

Nielsen, J. F., English, S., Goodall-Copestake, W. P., Wang, J., Walling, C. A., Bateman, A. W., . . . Pemberton, J. M. (2012). Inbreeding and inbreeding depression of early life traits in a cooperative mammal. *Molecular Ecology, 21*(11), 2788-2804. doi:10.1111/j.1365-294X.2012.05565.x

Nietlisbach, P., Camenisch, G., Bucher, T., Slate, J., Keller, L. F., & Postma, E. (2015). A microsatellite-based linkage map for song sparrows (Melospiza melodia). *Molecular Ecology Resources, 15*(6), 1486-1496. doi:10.1111/1755-0998.12414

Niskanen, A. K., Billing, A. M., Holand, H., Hagen, I. J., Araya-Ajoy, Y. G., Husby, A., . . . Jensen, H. (2020). Consistent scaling of inbreeding depression in space and time in a house sparrow metapopulation. *Proceedings of the National Academy of Sciences, 117*(25), 14584-14592. doi:10.1073/pnas.1909599117

O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology, 27*(16), 3193-3206. doi:10.1111/mec.14792

Pedersen, B. S., & Quinlan, A. R. (2017). Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. *American Journal of Human Genetics, 100*(3), 406-413. doi:10.1016/j.ajhg.2017.01.017

765 Riginos, C., Crandall, E. D., Liggins, L., Gaither, M. R., Ewing, R. B., Meyer, C., .
766 . . Deck, J. (2020). Building a global genomics observatory: Using GEOME
767 (the Genomic Observatories Metadatabase) to expedite and improve
768 deposition and retrieval of genetic data and metadata for biodiversity
769 research. *Molecular Ecology Resources, 20*(6), 1458-1469.
770 doi:10.1111/1755-0998.13269
771 Santure, A. W., Poissant, J., De Cauwer, I., Van Oers, K., Robinson, M. R.,
772 Quinn, J. L., . . . Slate, J. (2015). Replicated analysis of the genetic
773 architecture of quantitative traits in two wild great tit populations.
774 *Molecular Ecology, 24*(24), 6148-6162. doi:10.1111/mec.13452
775 Santure, A. W., Stapley, J., Ball, A. D., Birkhead, T. R., Burke, T., & Slate, J.
776 (2010). On the use of large marker panels to estimate inbreeding and
777 relatedness: Empirical and simulation studies of a pedigreed zebra finch
778 population typed at 771 SNPs. *Molecular Ecology, 19*(7), 1439-1451.
779 doi:10.1111/j.1365-294X.2010.04554.x
780 Sardell, R. J., Keller, L. F., Arcese, P., Bucher, T., & Reid, J. M. (2010).
781 Comprehensive paternity assignment: genotype, spatial location and social
782 status in song sparrows, Melospiza Melodia. *Molecular Ecology, 19*(19),
783 4352-4364. doi:10.1111/j.1365-294X.2010.04805.x
784 Sinnwell, J. P., Therneau, T. M., & Schaid, D. J. (2014). The kinship2 R Package
785 for Pedigree Data. *Human Heredity, 78*(2), 91-93.
786 doi:10.1159/000363105
787 Städele, V., & Vigilant, L. (2016). Strategies for determining kinship in wild
788 populations using genetic data. *Ecology and Evolution, 6*(17), 6107-6120.
789 doi:10.1002/ece3.2346
790 Stobie, C. S., Cunningham, M. J., Oosthuizen, C. J., & Bloomer, P. (2019).
791 Finding stories in noise: Mitochondrial portraits from RAD data. *Molecular*
792 *Ecology Resources, 19*(1), 191-205. doi:10.1111/1755-0998.12953
793 Van Oers, K., Santure, A. W., De Cauwer, I., Van Bers, N. E. M., Crooijmans, R.
794 P. M. A., Sheldon, B. C., . . . Groenen, M. A. M. (2014). Replicated high-
795 density genetic maps of two great tit populations reveal fine-scale genomic
796 departures from sex-equal recombination rates. *Heredity, 112*(3), 307-
797 316. doi:10.1038/hdy.2013.107
798 Walling, C. A., Pemberton, J. M., Hadfield, J. D., & Kruuk, L. E. (2010).
799 Comparing parentage inference software: reanalysis of a red deer
800 pedigree. *Molecular Ecology, 19*(9), 1914-1928. doi:10.1111/j.1365-
801 294X.2010.04604.x
802 Wang, M. H., Cordell, H. J., & Van Steen, K. (2019). Statistical methods for
803 genome-wide association studies. *Seminars in Cancer Biology, 55*, 53-60.
804 doi:10.1016/j.semcancer.2018.04.008
805 Waples, R. K., Albrechtsen, A., & Moltke, I. (2019). Allele frequency-free
806 inference of close familial relationships from genotypes or low-depth
807 sequencing data. *Molecular Ecology, 28*(1), 35-48.
808 doi:10.1111/mec.14954
809 Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A Tool for
810 Genome-wide Complex Trait Analysis. *The American Journal of Human*
811 *Genetics, 88*(1), 76-82. doi:10.1016/j.ajhg.2010.11.011

812

Data Accessibility

Supporting methods, results figures and tables are provided in the Supplementary Material. Supplementary Material 1 provides methods and results of a literature review of recent publications in the field of molecular ecology; Supplementary Material 2 provides a summary of the survey responses while Supplementary Material 3 describes the process of checking hihi SNP array data using the above described framework. Hihi are of cultural significance to the indigenous people of Aotearoa New Zealand, the Māori, and are considered a taonga (treasured) species whose whakapapa (genealogy) is intricately tied to that of Māori. For this reason, the genotypes and associated metadata for hihi will be made available by request on the recommendation of Ngāti Manuhiri, the iwi (tribe) that affiliates as kaitiaki (guardians) for hihi. To obtain contact details for the iwi, please contact Dr Anna Santure: a.santure@auckland.ac.nz.

Author Contributions

L.D. and A.W.S. designed the research, and L.D. processed and analysed the data and performed the research. J.G.E. developed the demographic dataset and supervised the data collection. P.B. developed the microsatellite dataset, supervised the genotyping, and performed the pedigree reconstruction. L.D. led the writing of the paper, with input from A.W.S. and feedback from P.B. and J.G.E. All authors read and approved the final manuscript.

Tables and Figures (with captions)

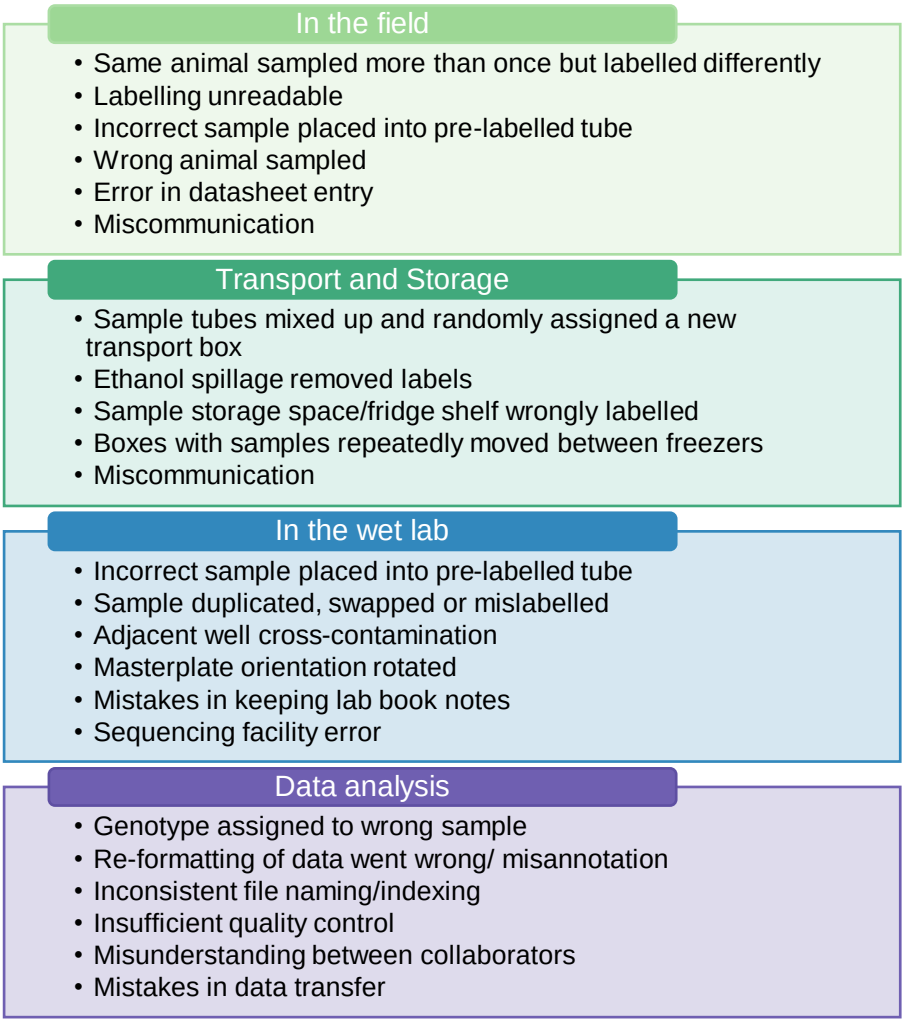

Figure 4: Examples of points in a research project where sample mix-ups could potentially occur. While sample errors are most likely to be detected at the end of a project from examining sequence or genotype data, sample mix-ups can happen at any stage, and can dramatically influence downstream conclusions.

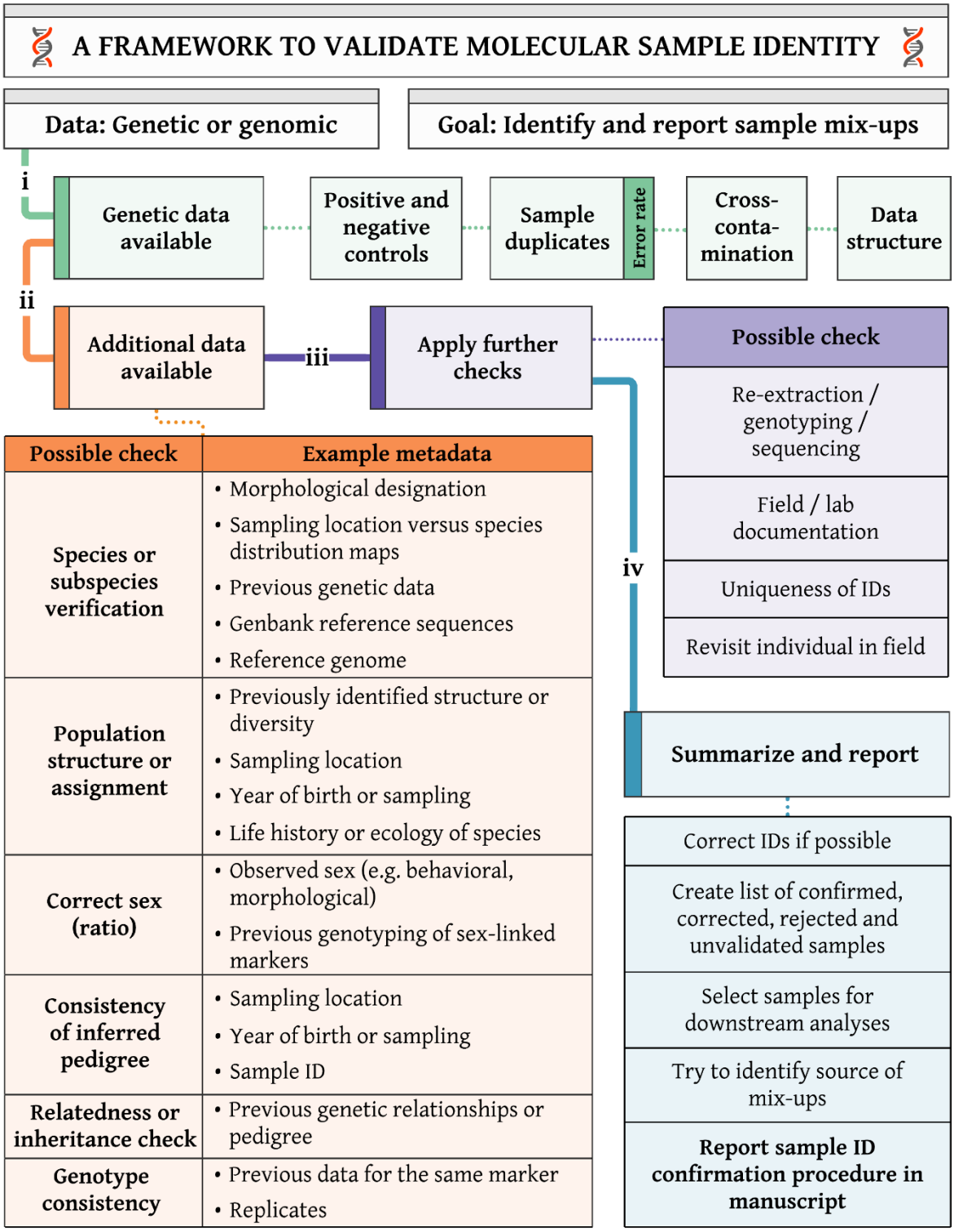

Figure 5: A molecular ecology framework to help detect genomic sample mix-ups. The framework presents common data checks (positive and negative controls, duplicates) and an analysis of data structure as universally applicable steps (i. green). The orange pathway describes sample checks if additional metadata (such as phenotypes, birth year, plate information and field notes) is available. Some studies can also draw information from previously established pedigrees or phylogenies (ii. orange) and sometimes additional genetic data (iii. purple). The goal of this framework is to make lists that contain the confirmed, corrected, rejected and unvalidated samples for future data analyses and management (iv. blue). Figure created with Lucidchart.com.

Table 1: Three example rows from a matrix with pairwise genomic relatedness values, and the difference between expected pedigree and genomic relatedness, between focal individuals A, B and C and their first-degree relatives (e.g. with F = father). In the case of individual A, a very low relatedness value with their mother (M) but relatedness consistency with siblings (S) and offspring (O) suggest that the mother is a sample mix-up. For individual B, relatedness inconsistencies with all genotyped relatives suggest that individual B is a mix-up. All available pedigree relatedness values for individuals A and B are 0.5 (i.e. there is no inbreeding). For individual C, despite very high relatedness values reflecting extensive inbreeding in their pedigree, consistency between pedigree and genomic relatedness indicates no mix-up. NA designates ungenotyped relatives. Note: The pedigree and genomic relatedness values are taken from the worked hihi example as mentioned in the Supplementary Materials.

Individual	Pedigree relatedness						Genomic relatedness						Pedigree – genomic relatedness					
	M	F	S1	S2	O1	O2	M	F	S1	S2	O1	O2	M	F	S1	S2	O1	O2
A	0.5	0.5	0.5	NA	NA	NA	0.06	0.45	0.49	NA	NA	NA	0.44	0.05	0.01	NA	NA	NA
B	0.5	0.5	0.5	0.5	0.5	0.5	-0.01	-0.04	-0.02	0.02	-0.05	0.06	0.51	0.54	0.52	0.48	0.55	0.44
C	0.84	0.78	0.81	NA	NA	NA	0.78	0.76	0.78	NA	NA	NA	-0.14	-0.02	0.03	NA	NA	NA

Table 2: Summary table of all the confirmed and rejected sample-genotype associations after following the steps of the suggested framework with the 1,256 hihi genotypes from Tiritiri Mātangi. 488 IDs were 'confirmed' through parentage assignment and having the correct sex and relatedness with other close relatives. 42 samples had two parentage assignment softwares agreeing on a different parental pair than the validated pedigree. Based on these assignments, 20 of these samples could unambiguously be assigned a new ID and are shown as 'corrected' while the other 22 were 'rejected' The remaining 'rejected' samples were duplicates, had a different parental pair in all pedigrees or were the wrong sex. Unvalidated samples were the correct sex but did not have enough additional information (i.e., few or no genotyped close relatives) available to be categorized in any way.

Sample status	*Confirmed*	*Corrected*	*Rejected*	*Unvalidated*
Number of samples	488	20	256	492

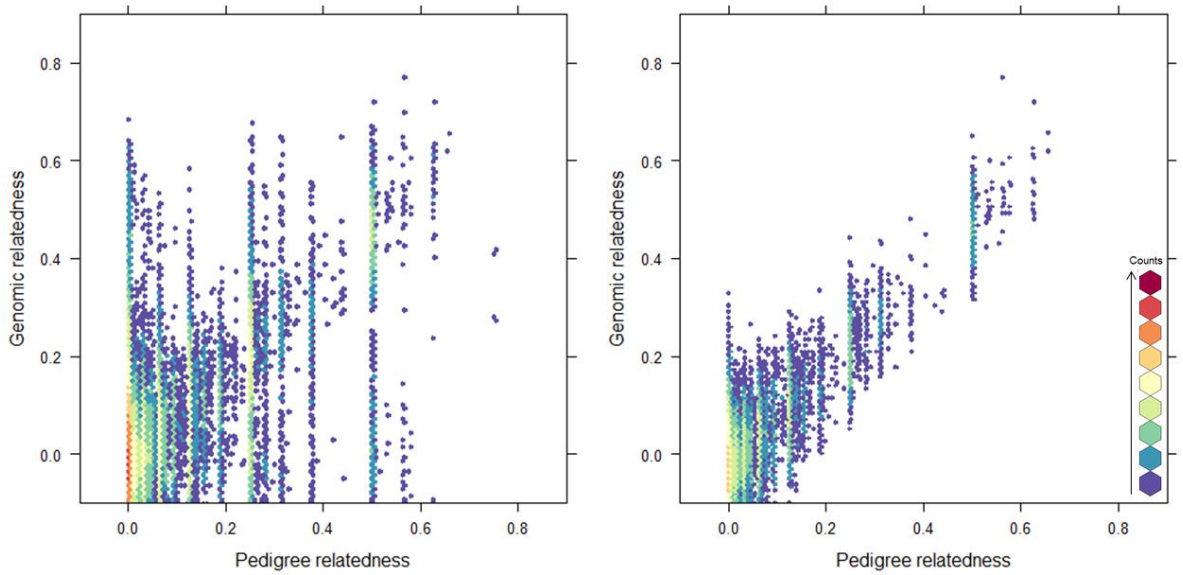

Figure 6: The association between genomic and pedigree relatedness of the hihi on Tiritiri Mātangi for all unique samples (right panel, N=1,250) and only the confirmed samples (right panel, N=488). The pedigree-based relationship matrix was calculated using the R package *kinship2* (Sinnwell et al., 2014), the genomic relationships were calculated with the tool *GCTA* (Yang et al., 2011). The warmer the colour, the more pairs show a specific relatedness, with most pairs being unrelated. In the left panel, some individuals show high pedigree yet no genomic relatedness, or low pedigree with high genomic relatedness, an indication of sample error. These erroneous links have disappeared after sample checking (right panel).