A 2b-RAD parentage analysis pipeline for complex and mixed DNA samples

Word count: abstract: 232 words; main text: 6857 words
Tables: 0 [Supplemental: 3]
Figures: 4 [Supplemental: 9]

Key words: parentage analysis, paternity testing, next-generation sequencing, 2b-RAD sequencing, identity-by-state matrix, combined paternity index

ABSTRACT

Next-generation sequencing technology has revolutionized genotyping in many fields of study, yet parentage analysis often still relies on microsatellite markers that are costly to generate and are currently available only for a limited number of species. 2b-RAD sequencing (2b-RAD) is a DNA sequencing technique developed for ecological population genomics that utilizes type IIB restriction enzymes to generate consistent, uniform fragments across samples. This technology is inexpensive, effective with low DNA inputs, and robust to DNA degradation. Here, we developed a probabilistic genotyping-by-sequencing genetic testing pipeline for parentage analysis by using 2b-RAD for inferring familial relationships from mixed DNA samples and populations. Our approach to partial paternity assignment utilizes a novel weighted outlier paternity index (WOPI) adapted for next-generation sequencing data and an identity-by-state (IBS) matrix-based clustering method for pedigree reconstruction. The combination of these two parentage assignment methods overcomes two major obstacles faced by other genetic testing methods: 1) It allows detection of parentage when closely related or inbred individuals are in the alleged parent population (e.g., in laboratory strains); and 2) it resolves mixed DNA samples. We successfully demonstrate this novel approach by correctly inferring paternity for samples pooled from multiple offspring (i.e., entire clutches) in a highly inbred population of an East African cichlid fish. The unique advantages of 2b-RAD in combination with our bioinformatics pipeline enable straightforward and cost-effective parentage analysis in any species regardless of genomic resources available.

INTRODUCTION

Genetic testing is fundamental to both ecology and forensic science for inferring relationships among individuals without direct historical knowledge (Flanagan, 2018). Its success is based on the insight that knowledge of variation in a relatively small number of Mendelian loci is sufficient to infer the structure and history of a population or to identify familial relationships (Thompson, 1976; Thompson & Meagher, 1987). For decades, such analyses have relied on short tandem repeats (STRs, often referred to as microsatellite markers), which take considerable time to develop and validate (Jones, 2010). Due to the large initial cost of establishing and validating microsatellites, their use has been limited to relatively few species, to outbred populations with numerous polymorphic loci, and to studies with relatively small sample sizes. Additionally, the requirement for human curation of microsatellite data can be considered more of an 'art form' than quantitative approach, with difficulty transferring criteria between laboratories (Flanagan, 2018). In addition, mixed samples (i.e., samples containing DNA from multiple individuals) results can be difficult to ascertain with STRs, especially when there are more than three contributors or any DNA degradation (Yang, 2019). Lastly, microsatellite-based approaches are ill-suited to automation of bioinformatic analysis pipelines (Hodel, 2016).

In its simplest form, parentage analysis is based on diploid offspring receiving one allele per locus from each parent. If the offspring and a putative parent share no alleles, then this individual can be excluded. However, **parentage analysis by exclusion** assumes that there are no errors for biological (e.g., mutations during meiosis) or technical (e.g. genotyping error) reasons (Chakraborty, 1974). Because exclusion testing only relies on homozygous sites, thus discarding most of the genetic information, this approach is rarely used anymore (Flanagan et al.,

2018; Kalinowski, 2007; Marshall, 1998). Instead, maximum-likelihood methods were developed to identify parent-offspring pairs in natural populations (Meagher, 1986). **Categorical allocation**, the most common parentage analysis used within this framework, calculates the relative likelihood of different hypotheses about putative triadic relationship being true. The likelihood is the probability of observing the genotypes given the proposed relationship, which can then be calculated through Mendelian inheritance rules (Kalinowski et al., 2007; Marshall et al., 1998). Instead of using absolute likelihood, a log-likelihood ratio is calculated by dividing the proposed triad likelihood by the likelihood that the members of a given triad are unrelated (Marshall et al., 1998). A positive log-likelihood ratio indicates that the triad is likely related but is difficult to interpret statistically. Therefore, parentage confidence is assessed by the difference between the highest log-likelihood ratio and the second highest log-likelihood ratio score. This in turn is compared to a critical value generated by simulation that uses observed allele frequencies and considers number of alleged fathers, proportion of potential fathers sampled, completeness of genetic data, and the genotyping error rate. Importantly, the reliability of the categorical allocation procedure critically depends on marker quality, the number of candidate fathers, and that the mother's genotype is known (Marshall, 1998). Another popular method for parentage analysis is **partial paternity testing**, which uses a Bayesian posterior probability to partially assign offspring to candidate parents, with the highest posterior probability indicating likely parentage (Devlin, 1988). Additionally, a prior for parentage can be assigned using known ecological or behavioral variables instead of assuming that mating is random, though this is generally not done, as it would confound the testing of those variables. This method outperforms categorical likelihood models as it avoids systematic biases such as over-assigning paternity to males with a relatively higher number of homozygous loci (Devlin, 1988). Partial paternity

testing fell out of favor and is underutilized in the study of paternity since in most cases it is impractical to consider fractions of paternity (Flanagan, 2018).

The advent of next-generation sequencing (NGS) has made it possible to efficiently identify thousands or even millions of single nucleotide polymorphisms (SNPs) in a population at low cost, which has revolutionized population genetics (Pool, 2010). **Genotyping-by-sequencing** approaches have eliminated the need for expensive and labor-intensive development and validation of microsatellite markers, as SNPs are much more abundant, have lower mutation rates, and can be genotyped with lower error rates (Anderson 2006). In fact, depending on the frequency of minor and null alleles, degree of linkage disequilibrium, and number of parental pairs, as few as 60-200 SNP markers, or ~500 if minor allele frequencies were low, outperform any microsatellite-based approaches (Dussault, 2018; Premachandra, 2019; Andrews, 2018; Anderson, 2006; Flanagan, 2018; Fernández, 2013). SNPs are particularly attractive when a population has low polymorphism (e.g., due to inbreeding) or when samples are mixed or contaminated with other sources of DNA (e.g., in forensic settings) (Flanagan et al., 2018; Hodel, 2016). Importantly, SNP-based approaches lend themselves to automation, which further increases efficiency and decreases cost. Given these numerous benefits, it is not surprising that the potential of genotyping-by-sequencing to dramatically advance our genotyping abilities for parentage analysis was recognized early (Glaubitz, 2003), yet to date remarkably few studies have utilized SNPs for parentage analysis (Flanagan, 2018).

One common NGS method in population genomics is Restriction-site-associated DNA sequencing (RAD-seq), which requires as little as 10-100 ng of DNA as input (Andrews, 2016) and uses short-read sequencing of a large library of DNA fragments to generate genotypes across millions of loci (Baird, 2008). Because RAD-seq methods do not require a reference genome,

this approach is ideally suited for species with limited genomic resources. The type IIB restriction endonucleases RAD-seq (2b-RAD, Wang, 2012) method produces smaller uniform fragment sizes with greater efficiency and lower cost than other RAD-seq methods, while still providing large numbers of SNP markers to assess paternity (Puritz, 2014; Andrews, 2016). The target fragment size in 2b-RAD is small and uniform (36 bp), which makes this method robust to DNA degradation and thus well suited for forensic applications if the degraded fragment sizes remain above ~50b (Barbanti, 2020).

While the use of NGS in parentage analysis has been growing, the effectiveness of this approach for more challenging applications, such as closely related individuals or mixed samples, has yet to be established. Current bioinformatic analysis pipelines for genotyping-by-sequencing usually rely on either categorical allocation or sibship reconstruction (Flanagan, 2018). Using multiple full- or half-siblings and one parent's full multi-locus genotype it is possible to reconstruct the genotype of an unknown relative with parental sibship reconstruction (Wang, 2004). A pedigree reconstruction method is required when related individuals may be present in the pool of alleged parents, although this approach requires testing more individuals than those of interest. Parentage analysis is particularly challenging in populations with high *Fis*, or high inbreeding, due to the reduction in informative distinctive loci when heterozygosity is low. However, RAD-seq approaches provide sufficient coverage for genome-wide analyses with only a few hundred SNP loci required (Andrews, 2016; Kardos, 2015). The use of marker-based approaches is encouraged for highly inbred populations, particularly when using non-model organisms as individuals are more homozygous across sites due to a greater degree of loci being 'identical by descent' (IBD) (Kardos, 2015).

A powerful method to measure relatedness in populations is clustering of an identity-by-state matrix (IBS), which is optimized for heterogeneous populations but is still capable of distinguishing closely related individuals (Stevens, 2011; Jin, 2017). IBS evaluates genetic similarity between pairs of samples by calculating the average degree of matching across all loci. However, clustering of an IBS matrix does not consider known data, such as pedigree data or maternal information, and therefore can greatly benefit from combination with techniques that do (Luan, 2012). A study in Pacific and European oysters combined both categorical allocation and identity-by-state clustering to successfully identify closely related individuals by grouping with multidimensional scaling (Gutierrez, 2017).

Nevertheless, there is an urgent need to develop efficient and robust parentage analysis pipelines for RAD-seq methods, especially *de novo* methods such as 2b-RAD, that can overcome real-world challenges such as complex population structure, inbred families, and mixed or contaminated DNA samples. The field of forensic genetics has set out guidelines for handling DNA mixtures, typically constrained with the inclusion of closely related individuals, that requires estimating relative contribution from each individual (Gill, 2006; Gill, 2015). Crucially, any approach of pooling more than two individuals requires a SNP based approach with many sites (Yang, 2019).

Here, we systematically investigated several 2b-RAD-based parentage analysis methods in the African cichlid fish, *Astatotilapia burtoni*, a model system in social neuroscience (Hofmann, 2006; Weitekamp & Hofmann, 2014). This species forms highly complex and dynamic social communities that can be readily studied and manipulated in the laboratory (Hofmann, 1999; Maruska, 2015). *A. burtoni* males of this species attract females to territorial bowers for mating, after which females incubate their offspring in their mouth for two weeks

(Fernald & Hirata, 1977). However, even though a female may spend considerable time with any given male, the time spent in or near a male's bower is no reliable indicator of successful mating (Kidd, 2006). In fact, females can mate with multiple males and thus incubate clutches with multiple paternity (Theis, 2012). Assigning paternity based on behavior alone is thus unreliable. Laboratory populations of *A. burtoni* are, however, highly inbred (Salzburger, 2018), which has foiled prior attempts to establish genotyping based on microsatellite markers (unpublished observations; for *A. burtoni* microsatellites see: Sanetra, 2009). These challenging characteristics make this species an ideal model system for systematically testing the performance of various 2b-RAD parentage analysis methods with genetically homogeneous and/or mixed samples. In the present study, we first validate the use of novel parentage analysis technique in triads of known paternity (Fig 1). We then demonstrate the potential of this approach in naturalistic communities.

METHODS

Behavioral Experiments

All animals used in this study were obtained from a laboratory population descended for about 60 generations from a wild-caught stock of 400 individuals (Fernald & Hirata, 1977). All work was done in compliance with the Institutional Animal Care and Use Committee (IACUC) at The University of Texas at Austin.

In the first experiment we established n=12 **triads with known paternity** consisting of one male (known father), a female incubating his offspring, and the offspring themselves by placing one male each (standard length SL 5.5 – 6.5 cm) into a compartment equivalent to one third of a 120 L hexagonal aquarium (i.e., four aquaria in total), along with three reproductive,

non-brooding females (SL 4.0 – 4.9 cm). Clear dividers between the compartments allowed for social interactions between all inhabitants of a given aquarium, while preventing any matings to take place across compartments, thus ensuring known paternity of any resulting offspring. To allow females to go through at least one full 28-day reproductive cycle (Kidd et al., 2013), we maintained these communities for two months. Eight males fathered at least one brood from 13 females, resulting in a total of 15 broods collected. There were 2 cases in which the same father and mother pair had multiple broods together resulting in biological replicates. Two males, one with biological replicate broods and another with two broods from two different females, were selected to be technically replicated and sequenced in duplicate. Any females that incubated fry more than once served as a biological replicate for the parentage analysis. A further five broods and one mother were randomly selected for technical replicates as well, resulting in a total of 20 broods with replicates.

In a second experiment, we established n=6 **naturalistic communities** of *A. burtoni* in 120 L aquaria, each consisting of 8 males (SL 5.0 – 6.6 cm) and 8 females (SL 4.0 – 5.5 cm), which ensured that multiple males in each community could establish a territory and seek out mating opportunities, while at the same time affording females the opportunity to have eggs fertilized by more than one male in a single mating bout, thus potentially creating broods with multiple paternity. For each community, we monitored social behavior, male social status, and space uses three times a week at 15:00 hours for 10 minutes each using a digital video system, while also measuring body mass and standard length every other week (data not shown). Over the 12-week observation period we collected 25 broods from 23 mothers (1 – 6 broods per community), with two females incubating two broods each. Two males and two broods from different communities served as technical replicates.

Throughout either experiment, broods were collected from females' buccal cavity approximately one week after fertilization and stored in 70% ethanol at 4° C. At that stage, fry are large enough to be easily separated from any remaining yolk and to yield abundant DNA. A razor and slide were used to separate any yolk and cut individuals in half. The bottom and top halves for all the fry in each brood were then pooled and stored separately. This allowed for each brood pool to consist of approximately equal proportions of each offspring. At the end of each experiment, we collected fin-clips collected from all adults and stored them in 70% ethanol at 4° C until DNA extraction.

Broods are named by the 3-letter tank code, the color-tag of their mother, and the date collected. Females are named by their color-tag followed by their 3-letter tank code, males are named in a similar fashion. Any name that ends in an underscore by a letter (i.e. '_A' or '_B') indicates a technical replicate. Therefore, a mother and brood will share both the unique tank ID and color, while the brood will also indicate a date. In the known triads, with only one male per tank, the unique tank id can identify the correct father for any given brood. In the naturalistic communities, only real mothers can be identified by unique tank id. In the known triad, the alleged father pool consisted of all adult males used in triads. In the unknown community, the alleged father pool was limited to males within each tank.

Library preparation and sequencing

DNA was extracted from fin clips and fry using Maxwell 16 Tissue DNA Purification kit (Promega, USA) and then purified using Zymo DNA Clean & Concentrator kit (Zymo Research, USA) according to the manufacturer's instructions. We then prepared sequencing libraries according to Wang (2012) (we used version "2bRAD_protocol_may15_2017_nnrw", the most

up-to-date detailed protocol is available at https://github.com/z0on/2bRAD_denovo). Briefly, a type IIB restriction enzyme BcgI (New England Biolabs) was used to digest DNA into uniform 36 base pair fragments. Adaptors with unique molecular identifiers (UMI) ligated to the fragments barcoded only on the 3' end before being stored overnight at 4°C. The ligase was then heat-inactivated with a 10-minute incubation at 65°C. Samples were then pooled with 12 different 3' barcodes and amplified before a final purification step of the pooled libraries for the band at 160-180 base pairs using the Pippin Prep (Sage Science, USA) protocol. Libraries were sequenced on the Illumina HiSeq 2500 platform (Illumina, USA) at UT Austin's Genomic Sequencing and Analysis Facility generating 418 million reads (2.9 million reads per sample on average).

Bioinformatic analyses

Processing of raw reads and quality control

The 2bRAD sequencing reads were de-multiplexed, trimmed, and de-duplicated using the custom script accommodating the 2bRAD-specific triple-barcoding scheme and degenerate ligated tags to identify PCR duplicates (https://github.com/z0on/2bRAD_denovo). The SNP profiles were generated by 2bRAD sequencing using the 2b-RAD pipeline from Wang, 2012 and mapped to the reference *A. burtoni* genome (RefSeq assembly version GCF_000239415.1 AstBur1.0; Brawand, 2015). The resulting mapped to the *A. burtoni* genome with 81% efficiency, and to closely related Nile Tilapia genome with 55% efficiency. ANGSD (Korneliussen, 2014) with SAMtools (Li, 2009) model produced genotype likelihoods for each individual across all 1.7 million loci. Two males from one of the naturalistic communities (G2) were removed at this stage, as they only had sequence coverage for less than 1% of these sites

while all the remaining fish had >60% coverage. Having less than 1% coverage of sites not only indicates a likely technical issue with sequencing on those samples but also does not provide enough sites to establish paternity.

To avoid sampling each egg individually or be limited to only a small portion of brood as is common, pooling brood DNA and using a read depth of around 50X enabled an assessment of the proportion of paternity attributable to each male. This level of coverage is not needed for adults, instead 20X coverage was used to sufficiently resolve heterozygous SNPs. Quality control from the bam files for all adults (supplemental figure 1) and all broods (supplemental figure 2) indicate good quality and sequencing depth. Of note, is the variation in coverage among adults which would result in differential rates of confident base calls among males. Therefore, males sequenced at higher depth have more sites to match with broods which could skew paternity testing towards highly sequenced males.

Parentage analysis techniques

CERVUS

We applied the popular paternity analysis software CERVUS version 3.0.7 to the known triad dataset (Kalinowski, 2007). This program uses allele frequencies and individual genotype calls to calculate a likelihood score for each potential parent and the combination of a known parent and an alleged parent as represented by the log-likelihood ratio. A log-likelihood ratio, or the delta log-likelihood ratio score for comparing to the next most likely parent, above 0 is considered a likely paternity match. CERVUS utilizes a simulation of the observed allele frequencies to determine the predicted likelihood difference of the real parent compared to a

random individual in the population. Additionally, CERVUS has an option to incorporate an estimate of genotyping error provided by the user.

We assigned genotype calls to the known triad samples by assigning genotype probabilities above 0.75 as the correct genotype for that site in an individual. Next, SNPs were filtered by the minor allele frequency (MAF) to reduce the number of total sites using six different cutoffs: 0.4, 0.3, 0.2, 0.1, 0.05, 0.03 (Andrews, 2018). Paternity testing was run twice for each MAF cutoff with CERVUS calculating allele frequency once using just adults and once using all samples. Simulation was therefore done 12 times for 100,000 offspring, 12 potential males, 95% proportion of fathers sampled, 50% proportion of typed alleles missing data, estimated genotyping error rate of 1%, and minimum typed loci of 50% total loci per analysis (Crain, 2020; Andrews, 2018)

Relative Combined Paternity Index

After bam files were generated using SAMtools and referenced to *A. burtoni* genome, ANGSD was used to filter out SNPs and assign genotype likelihoods at the remaining site for each individual and brood. A custom R-script was used to filter out sites based on adult population genotype frequency using all adults in known triads and unknown communities respectively to avoid unwanted biases (Flanagan, 2018). Using the function 'paternityIndex' from the R package 'paternity' (Rosyara, 2014), each pair of mother and brood was used to calculate paternity index for every alleged father at the filtered sites.

Paternity index is a ratio of the likelihood of the offspring's genotype conditional on the mother and alleged father's genotype over the likelihood of the offspring's genotype given the mother's genotype. This means that increase in paternity index can be considered an increase in

paternity probability and is standard method of partial paternity allocation (Baur, 1986). The paternity index from the R package paternity uses a set of equations that utilize population allele frequency to calculate the paternity index for a given locus given the genotype of offspring, mother, and alleged father at that site (Elston, 1986).

A combined paternity index (CPI) for a given alleged father is then calculated by taking the product of the paternity index for every site. This method, developed for microsatellites, requires genotypes to be assigned and drops down to zero if there are any exclusion sites. We attempted to replicate this method using our sequencing data by setting a genotyping threshold. This method failed as every male including the fathers had a CPI score of zero, and we had limited success when we excluded exclusion sites altogether. While null alleles can be easily identified or ignored, allelic dropouts are particularly challenging for parentage analysis as this type of sequencing error can create false exclusions between parents and offspring, although false alleles can also pose challenges (Wang, 2010). Since this approach requires a genotyping call across the mother, brood, and alleged father, allele sites that were not present in at least one individual of the triad being tested were removed from the analysis on a per triad basis. Similarly, sites that would indicate an exclusion for an alleged father were removed to ensure that no possible sequencing or genotyping errors altered paternity, as a single exclusion site would result in a paternity index of zero. Taken together, only sites that would add paternity information were included with the goal that the most likely father would maintain the highest relative CPI score of all alleged fathers. With microsatellite data, the probability of paternity is traditionally determined as the CPI divided by one plus the CPI, assuming a uniform prior, is commonly used. This method of probability of paternity does not work with the large number of sites used as most alleged fathers would end up with probability of paternity well above 99 %.

Therefore, a relative CPI was calculated by dividing paternal CPI by the sum of all the CPI scores for every alleged father of a mother and brood pair. This novel relative CPI approach mirrors the use of delta log-likelihood ratio score in categorical allocation parentage analysis, in which the top two highest scoring males are compared (Marshall et al., 1998). The false-positive rate threshold determined in the known triads was used to filter CPI for unknown paternity. The relative CPI percentage and false-positive threshold were used to assign likelihood of paternity and identify cases in which paternity could not be assigned, respectively. For relative CPI, genotypes were assigned to loci with a genotype likelihood above 0.6 resulting in ~8,000 sites used in parentage analysis.

Weighted Outlier Paternity Index (WOPI)

After bam files were generated after mapping reads to *A. burtoni* genome using bowtie2, ANGSD was to assign genotype likelihoods at the remaining sites for each individual and brood. A custom script was used to filter out non-variable sites by selecting sites with at least two samples having an alternative allele with a read count of 2 or greater. Each pair of mother and brood was used to calculate paternity index for every alleged father at the filtered sites. Novel to this approach, the genotype probabilities assigned by ANGSD were used directly without applying a threshold to assign genotypes. This produces an output beagle format file, which is a standardized table that includes genotype likelihood for each individual at every locus with a single SNP for all genotype combinations: homozygous major allele, homozygous minor allele, and heterozygous. Here we used the reference state to assign the reference and alternative alleles, although these can be determined *de novo* per population. Importantly, sites with no data for an individual are given equal probability for all three possible genotypes. This allows for the

incorporation of the sequencing error-correction inherent in the ANGSD output directly into the paternity calculation, since no one genotype at any loci can have an absolute genotype probability.

We developed a weighted paternity index to incorporate genotype probabilities directly with CPI. For each site, we calculated every paternity index value for all possible genotype combinations across the alleged father, mother, and brood. We then multiplied each paternity index value for a given set of genotypes at a specific site by the genotype probabilities that the individuals have those genotypes at that site. The weighted paternity index for a site is the sum of all these paternity index values that have been weighted by the probability that the individuals have that specific genotype combination (Fig 2a). To achieve this a custom R function was developed, taking inspiration from the R function 'paternityIndex' from the package 'paternity' (Rosyara, 2014). Importantly, weighted paternity index maintains exclusion sites, either real or from sequencing error, as they no longer have a value of zero instead assigning a value based on the probability that it is an exclusion site.

An information score criterion was developed to filter out sites that had no read coverage, in which case an individual had equal probability of all three genotypes. The information score was calculated by taking the difference of the highest and lowest genotype probability for a given site in an individual. An information score of zero would therefore indicate that the site had been assigned an equal probability (e.g., 0.33) for all three genotypes. Implementing this filter reduces random noise due to variation in coverage, as sites with no data are assigned equal probability to all three genotypes.

Performing a standard CPI does not work with sequencing data as multiplying that many values below one will result in a number too small to compute. The theoretical distribution of

paternity index values has a mean around one, as we would predict most sites would not be informative regarding paternity, with any exclusion sites having a value of zero. Importantly, the lowest possible paternity index value for a non-exclusion site is 0.5. That means any weighted paternity index below 0.5 indicates either a likely exclusion site or that the alleged father is unlikely to be the father compared with the population. Likewise, any value on the other side of the distribution above 1.5 indicates that the alleged father is more likely to be the father. Therefore, we can limit the number of sites by focusing on the outlier tails and taking the combined product of the values above 1.5 and below 0.5, termed weighted outlier paternity index (WOPI). A father was assigned paternity with a WOPI score well above the distribution of WOPI scores for all other males. Determining the degree of separation from other alleged fathers was done by generating a mean and standard deviation of the WOPI for a specific brood and mother pair across the pool of alleged father, excluding the alleged father with the highest WOPI score. Then this mean and standard deviation was used to calculate a z-score for each alleged father and paternity assigned if the highest scoring male passed a z-score threshold. This threshold was determined by selecting a value that correctly identified all correct fathers from the known triad experiment. If no male scores above this z-score threshold than paternity could not be determined. Since this method also depends on a well sequenced mother, broods were filtered out that did not cluster with their mother (see '*Identity-by-State (IBS) Matrix'*).

Identity-by-State (IBS)

Genotyping data for adult samples, with technical replicates removed, were processed through ANGSD to create a table of ~23,000 adult sites present in at least 10 individuals at a read depth of at least 2, which was then indexed through ANGSD. This indexed site file was

used to filter sites for the ANGSD command that generated the IBS matrix. An IBS matrix was generated for the broods with the females and males separately for the known triads and each naturalistic community, respectively. The dendrogram of the IBS matrix was generated with a custom R script using the function pvclust (Suzuki, 2006) to generate hierarchical clustering, with agglomeration method UPGMA and euclidean distance, providing both an approximately unbiased (AU) p-value and edge height for each dendrogram.

To assign paternity or maternity, the first internal node from the offspring had to be above an AU p-value threshold and only contain one other individual. The AU p-value threshold was determined by selecting a value that successfully identified correct fathers in the known triads. Offspring that did not properly cluster with known mother after IBS matrix clustering were removed. Paternity assignment was determined by finding the closest node to a brood with a putative father and assessing the AU p-value. If that first node had multiple fathers than paternity could not be determined.

Population heterozygosity

The original fish population was allowed to breed freely, rendering it too inbred for microsatellite analysis (Pauquet et al., 2018). In addition, the individuals used here were selected based on size and other attributes, not their relatedness status, to set up functional social groups. We determined individual global heterozygosity and inbreeding coefficient for all adults using 242,308 high quality sites present in 99% of adults. Heterozygosity was calculated as the site frequency spectrum (SFS) estimation for a single sample using ANGSD and realSFS to get the proportion of heterozygous genotypes. Finally, we performed a test for Hardy-Weinberg

equilibrium (HWE) based on genotype likelihoods using ANGSD to determine the inbreeding coefficient.

RESULTS

Experiment 1: Known triads

CERVUS

With a total of 2,400 paternity tests run, 12 for each of the 20 broods, only 4 of the 20 broods resulted in a trio log-likelihood ratio and trio delta score above 0 for any of the allele frequencies (see supplemental table 1). All 4 of these broods, with at least one positive trio log-likelihood ratio across all the parameters, did identify the correct father indicating no false positives but a low success rate. Additionally, we found that it took longer to run compared to the other parentage analysis methods, due to both the number of simulations run and the fact that it uses a GUI instead of an R script. Overall, this method did not identify any false-positives while only assigning paternity to 4 out of 20 known triad broods.

Relative Combined Paternity Index

For the known triads, a CPI threshold was set to eliminate any false-positives, and while it added three false-negatives, this stringent filter can confidently assign both paternity and identify cases in which it is unknown. A realistic father threshold of e^{27} was sufficient to eliminate any false positives from the known paternity triads, as such is used as the threshold under which a male is not considered the likely father. This means that any brood that does not have a male above this threshold is considered to have unknown paternity. Using known paternity triads, 16 out of 20 broods the male with the highest relative CPI was the correct father,

including brood technical replicates, but three of these fell below CPI threshold (see supplemental table 2).

Weighted Outlier Paternity Index (WOPI)

For the known triads, the WOPI approach correctly assigned paternity for all 15/15 broods and all technical replicates (see supplemental figure 3). A z-score threshold of e^{20} was selected empirically as the lowest value that clearly distinguished correct fathers from all the other alleged fathers (see supplemental figure 4). This threshold prevented false positives when testing WOPI by removing the true father (see supplemental figure 5). This method outperformed paternity testing via CERVUS and a relative CPI approach in correctly assigning paternity (see supplemental table 2).

Identity-by-State (IBS) Matrix

Each technical replicate paired with its appropriate counterpart at an AU p-value of at least 80, which empirically served as the threshold for a successful node. For the known triads, in the IBS matrix of just females and broods, every brood shared the closest node with the correct mother with an AU p-value above 80 (see supplemental figure 8). Hierarchical clustering of the male and brood IBS matrix correctly identified the father at the first node for 13/15 broods (Fig 4a). Of the remaining two broods, one did pair with the correct father at the first node, but the AU p-value was below threshold at 54. The second brood, which was also technically replicated, had both the correct father and another male at the first node that had an AU p-value of 68 (see supplemental figure 8). Both of these males, ‘Black.I4A’ and ‘Grey.I5B’, had additional broods that clustered correctly indicating that the issue may be with the brood not the males.

442

443 Experiment II: Naturalistic communities

444 *Weighted Outlier Paternity Index (WOPI)*

445 For the naturalistic communities, the WOPI approach assigned paternity above the z-
446 score threshold for 11 of 15 collected broods after removing broods that failed maternal pairing
447 in the IBS matrix approach (Fig 4b). The four broods that did not assign a father also did not
448 have a father assigned by the IBS approach. Additionally, of the 11 broods that did not cluster
449 with their mother in the IBS approach only 2 had a father assigned by WOPI (see supplemental
450 table 3). Together, these results indicate that the WOPI method is conservative when calling
451 paternity. Each brood shows a distinct range of the WPI outlier tails across alleged fathers (see
452 supplemental figure 6). Putative fathers are easily distinguished when applying the z-score
453 threshold determined in the Known Triad experiment (see supplemental figure 7).

454

455 *Identity-by-State (IBS)*

456 The naturalistic communities had variable success with the IBS approach, probably due
457 to the much higher incidence of closely related individuals present in each community. After
458 filtering out cases in which mothers and broods did not match, fathers were successfully assigned
459 to 10/15 of these broods, all assignments agreeing with the WOPI results (Fig 4b). All technical
460 replicates were easily identified and appropriately paired (see supplemental figure 9). Two
461 additional broods that failed maternal pairing also had a father assigned (see supplemental table
462 3). One brood that had no father assigned by the IBS approach had it assigned by the WOPI
463 approach.

464

Population heterozygosity

We determined individual global heterozygosity for adults by calculating the SFS estimation for a single sample using ANGSD to get the proportion of heterozygous genotypes across 242,308 sites present in 99% of adults. We found low levels of heterozygosity across sites in the adult population with a mean of 0.00246 (s = 0.00017) (see supplemental figure 10). Additionally, we performed a test for Hardy-Weinberg equilibrium (HWE) based on genotype likelihoods across these same sites in adults. We found that the interindividual F statistic, inbreeding coefficient, was mostly positive across sites with a mean of 0.638 (s = 0.142). This average positive value indicates a high degree of inbreeding, as the sites are heterozygous deficient compared with HWE expectations (see supplemental figure 10).‘

DISCUSSION

We used 2b-RAD to develop a parentage analysis method that uses a combination of the novel WOPI approach and IBS clustering (Fig 1). Together, these approaches are specifically designed to deal with mixed samples and genetically homogeneous populations. The WOPI approach accounts for genotyping uncertainty and integrates data from both parents. IBS clustering is crucial in identifying cases in which the mother and offspring do not cluster together, indicating a potential issue with sequencing or presence of a close maternal relative in the dataset. Therefore, it serves as an appropriate filter for the WOPI approach, which is dependent on both maternal and paternal data. Additionally, IBS clustering of potential fathers provides insight into the population structure to identify problematic closely related males. Together, these methods outperformed traditional methods of paternity such as CPI and CERVUS (see supplemental table 2).

The WOPI approach was able to correctly identify paternity for pooled broods in all 15 known triads, while also determining a z-score threshold that prevented false-positives when the correct father was not present. IBS matrix clustering correctly identified paternity for pooled broods in 13 of the 15 known triads, with one being correct but below threshold and another not assigning any father. The combination of these two techniques identified paternity for 100% of the known triad broods (Fig 4a).

Application of WOPI and IBS matrix clustering to naturalistic communities resulted in 11 of 15 broods having paternity assigned by at least one method, 10 of which had concordant assignments by both methods (Fig 4b). The failure of maternal clustering provides an appropriate filter for WOPI as this method relies on both quality paternal and maternal data. Compared to the known triads, in which every mother-offspring pair was correctly identified by IBS clustering, the naturalistic communities only have correct identification of mother-offspring pairs in 15 broods, with 12 broods failing to have a mother identified (see supplemental table 3). Both techniques appear robust to false positives, as evidenced by the high concordance of cases which both methods did not assign parentage.

Limitations

Samples comprised of DNA mixtures pose difficulties when determining how related individuals and genotypes are represented in the mixture (Gill, 2015). One goal of the present study was to understand the effects of pooling all the offspring within a brood on parentage assignment. The possibility to pool offspring could considerably lower the cost of brood parentage analysis. Had individuals within a brood been sequenced separately, a maximum-likelihood algorithm could be used to generate the full set of possible parents with parental

sibship reconstruction (Wang, 2004). We treated each pooled brood as a population with genotype probabilities reflecting brood allele frequencies. Therefore, we had to use techniques that did not rely on genotype calls but considered relative probabilities of every possible genotype.

Broods with multiple paternity are common in *A. burtoni* (Kellog, 1995; Theis, 2012). When using pooled broods, partial paternity testing is the only method capable of detecting multiple paternity. An advantage of partial analysis is that uncertainty from parentage analysis is incorporated as the uncertainty in the final estimate, whereas categorical allocation typically discards uncertainty early in the analysis. We predict that a multiple paternity brood would result in an instance where one male could not be identified as a father, as the multiple fathers would be equally likely. Future work will examine the effect multiple paternity broods have on these parentage analyses and whether true partial paternity can be resolved.

In some naturalistic communities neither the WOPI nor IBS approach reliably identified a father. While this is to be expected if maintaining a low false positive rate is a goal, the thresholds determined in our study will not be universal. Specifically, we assigned empirical thresholds from the Known triad experiment that maximized the number of correct father calls while allowing for no false positives. Future research will need to focus on expanding these approaches to other systems and datasets to ascertain the exact relationship between number of sites, sample size, and population allele frequency with appropriate thresholds.

Inbred populations

A majority of parentage testing techniques, such as categorical allocation, work under the assumption that parents are unrelated, and the population of putative parents contain no close

relatives, as this can lead to instances in which full-siblings can be incorrectly assigned parentage over actual parents (Thompson, 1976; Marshall et al., 1998; Thompson & Meagher, 1987). Inbred populations pose a problem to both microsatellite and SNP assays due to low levels of variation among individuals (Fisher, 2009). Nevertheless, an analysis with fewer than 100 SNPs can outperform the use of microsatellites in homogenous populations (Fisher, 2009; Tokarska, 2009). The most informative SNP loci are ones with high minor allele frequency and low likelihood of allelic dropouts, with more loci required with lower allelic diversity (Flanagan, 2018). Parentage analysis can be skewed when closely related males (e.g., brothers) are present in the sample as they will cluster together and can result in a set of related putative fathers (Double, 1997). Therefore, our success in developing a parentage analysis pipeline even in a highly inbred, homozygous population demonstrates the overall effectiveness of this approach (supplemental figure 10). If close relatives are suspected to be in the sample, we recommend including broader pedigree analysis such as IBS clustering (Flanagan, 2018). The combination of WOPI and IBS testing allows detection of parentage in sample populations from closely related individuals.

Use of RAD-seq for parentage analysis

Few studies have employed a next-generation sequencing for parentage analysis, possibly due to the perception that this approach is expensive, involves intensive molecular biology skills, or requires advanced bioinformatics expertise (Flanagan, 2018; Palaiokostas, 2020; Crain, 2020.). However, with the widespread adoption of bioinformatics training, the introduction of more user-friendly analysis pipelines, it is only a matter of time before NGS becomes the preferred method of parentage analysis. Financial obstacles have diminished over time with

RAD-seq analysis becoming more accessible and affordable, particular with cost-effective approaches such as 2b-RAD (Hodel, 2016; Puritz, 2014). The cost per sample can be further decreased by reducing sequencing depth or utilizing reduced-representation adapters, which decreases the number of sites sequenced by 4- or 16-fold. Importantly, techniques such as 2b-RAD are highly amenable for use with lower-quality, slightly degraded DNA samples from non-model species (Barbanti, 2020). Additionally, 2b-RAD provides an excellent tool for analysis beyond parentage and is well-established in the field of molecular-ecology (Puritz, 2014; Wang, 2012).

With NGS it is highly unlikely that any data produced will be error free, especially with large numbers of samples and/or markers. Most current parentage analysis techniques incorporate some form of error rate correction that the user provides. Generally, these are based on expectations for microsatellites and may not account for sequencing error and allelic dropouts of PCR bias that arise from NGS techniques. (Kalinowski, 2007; Flanagan, 2018). Therefore, we recommend using sequencing methods that incorporate some form of PCR duplicate discrimination, such as 2b-RAD, and analysis pipelines that can calculate genotyping probability, such as ANGSD.

CONCLUSIONS

2b-RAD is a cost-effective sequencing-based method capable of handling complex biological samples with limited genomic resources. In the present study, we combined two approaches to parentage analysis: WOPI, a novel partial paternity allocation, and IBS clustering, a pedigree reconstruction analysis. Together, these techniques can confirm paternity cases while accounting for genotyping uncertainty. We expect this novel approach to have broad applications

in public health, forensics, crop and life stock breeding, conservation management, and evolutionary ecology studies.

ACKNOWLEDGEMENTS

We thank Dr. Becca Young and members of the Hofmann lab for guidance and discussion. This work was supported by a US Department of Justice graduate fellowship to IMC, an EEB start-up grant to IMC, and NSF-IOS grant 1326187 to HAH.

DATA AND CODE AVAILABILITY

The filtered deduplicated reads have been deposited to the NCBI Short Read Archive (SRA), bioproject PRJNA754415. The code used in this paper together with documentation can be accessed as the GitHub repository, https://github.com/imillercrews/ParentageAnalysis.

CONFLICT OF INTEREST

The authors have no competing interests to declare.

REFERENCES

Anderson, E. C., & Garza, J. C. (2006). The power of single-nucleotide polymorphisms for large-scale parentage inference. Genetics, 172(4), 2567–2582. https://doi.org/10.1534/genetics.105.048074

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Publishing Group, 17. https://doi.org/10.1038/nrg.2015.28

Andrews, K. R., Adams, J. R., Cassirer, E. F., Plowright, R. K., Gardner, C., Dwire, M., Hohenlohe, P.A., Waits, L. P. (2018). A bioinformatic pipeline for identifying informative SNP panels for parentage assignment from RADseq data. Molecular Ecology Resources, 18(6), 1263–1281. https://doi.org/10.1111/1755-0998.12910

Barbanti, A., Torrado, H., Macpherson, E., Bargelloni, L., Franch, R., Carreras, C., & Pascual, M. (2020). Helping decision making for reliable and cost-effective 2b-RAD sequencing and genotyping analyses in non-model species. Molecular Ecology Resources, 20(3), 795–806. https://doi.org/10.1111/1755-0998.13144

Baur, M. P., Elston, R. C., Gürtler, H., Henningsen, K., Hummel, K., Matsumoto, H., Mayr, W., Moris, J. W., Niejenhuis, L., Polesky, H., Salmon, D., Valentin, J., Walkers, R. (1986). No fallacies in the formulation of the paternity index. American Journal of Human Genetics, 39(4), 528. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1683973/?page=2

Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., … Di Palma, F. (2015). The genomic substrate for adaptive radiation in African cichlid fish. Nature, 513(7518), 375–381. https://doi.org/10.1038/nature13726

617 Chakraborty, R., Shaw, M., & Schull, W. J. (1974). Exclusion of paternity: the current state of
618 the art. American Journal of Human Genetics, 26(4), 477–488. Retrieved from
619 http://www.ncbi.nlm.nih.gov/pubmed/4841637
620 Crain, J., Larson, S., Dorn, K., Hagedorn, T., DeHaan, L., & Poland, J. (2020). Sequenced-based
621 paternity analysis to improve breeding and identify self-incompatibility loci in
622 intermediate wheatgrass (*Thinopyrum intermedium*). Theoretical and Applied Genetics,
623 133(11), 3217–3233. https://doi.org/10.1007/s00122-020-03666-1
624 Devlin, B., Roeder, K., & Ellstrand, N. C. (1988). Fractional paternity assignment: theoretical
625 development and comparison to other methods. Theoretical and Applied Genetics, 76(3),
626 369–380. https://doi.org/10.1007/BF00265336
627 DeWoody, J. A., & Avise, J. C. (2000). Microsatellite variation in marine, freshwater and
628 anadromous fishes compared with other animals. Journal of Fish Biology. John Wiley &
629 Sons, Ltd (10.1111). https://doi.org/10.1006/jfbi.1999.1210
630 Double, M. C., Cockburn, A., Barry, S. C., & Smouse, P. E. (1997). Exclusion probabilities for
631 single-locus paternity analysis when related males compete for matings. Molecular
632 Ecology, 6(12), 1155–1166. https://doi.org/10.1046/j.1365-294X.1997.00291.x
633 Dussault, F. M., & Boulding, E. G. (2018). Effect of minor allele frequency on the number of
634 single nucleotide polymorphisms needed for accurate parentage assignment: A
635 methodology illustrated using Atlantic salmon. Aquaculture Research, 49(3), 1368–1372.
636 https://doi.org/10.1111/are.13566
637 Elston, R. C. (1986). Probability and paternity testing. American Journal of Human Genetics,
638 39(1), 112–122. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3752078

639 Fernald, R. D., & Hirata, N. R. (1977). Field study of Haplochromis burtoni: Quantitative
640 behavioral observations. Animal Behaviour, 25, 964–975.
641 Fernández, M. E., Goszczynski, D. E., Lirón, J. P., Villegas-Castagnasso, E. E., Carino, M. H.,
642 Ripoli, M. V., … Giovambattista, G. (2013). Comparison of the effectiveness of
643 microsatellites and SNP panels for genetic identification, traceability and assessment of
644 parentage in an inbred Angus herd. Genetics and Molecular Biology, 36(2), 185–191.
645 https://doi.org/10.1590/S1415-47572013000200008
646 Fisher, P. J., Malthus, B., Walker, M. C., Corbett, G., & Spelman, R. J. (2009). The number of
647 single nucleotide polymorphisms and on-farm data required for whole-herd parentage
648 testing in dairy cattle herds. Journal of Dairy Science, 92(1), 369–374.
649 https://doi.org/10.3168/jds.2008-1086
650 Flanagan, S. P., & Jones, A. G. (2018). The future of parentage analysis: From microsatellites to
651 SNPs and beyond. Molecular Ecology, mec.14988. https://doi.org/10.1111/mec.14988
652 Gill, P., Brenner, C. H., Buckleton, J. S., Carracedo, A., Krawczak, M., Mayr, W. R., … Weir, B.
653 S. (2006). DNA commission of the International Society of Forensic Genetics:
654 Recommendations on the interpretation of mixtures. Forensic Science International,
655 160(2–3), 90–101. https://doi.org/10.1016/j.forsciint.2006.04.009
656 Gill, P., Haned, H., Bleka, O., Hansson, O., Dørum, G., & Egeland, T. (2015). Genotyping and
657 interpretation of STR-DNA: Low-template, mixtures and database matches - Twenty
658 years of research and development. Forensic Science International: Genetics, 18, 100–
659 117. https://doi.org/10.1016/j.fsigen.2015.03.014

660 Glaubitz, J. C., Rhodes, O. E., & Dewoody, J. A. (2003). Prospects for inferring pairwise
661 relationships with single nucleotide polymorphisms. Molecular Ecology, 12(4), 1039–
662 1047. https://doi.org/10.1046/j.1365-294X.2003.01790.x
663 Gutierrez, A. P., Turner, F., Gharbi, K., Talbot, R., Lowe, N. R., Peñaloza, C., … Houston, R. D.
664 (2017). Development of a Medium Density Combined-Species SNP Array for Pacific and
665 European Oysters (*Crassostrea gigas* and *Ostrea edulis*). G3 (Bethesda, Md.), 7(7),
666 2209–2218. https://doi.org/10.1534/g3.117.041780
667 Hadfield, J. D., Richardson, D. S., & Burke, T. (2006). Towards unbiased parentage assignment:
668 Combining genetic, behavioural and spatial data in a Bayesian framework. Molecular
669 Ecology, 15(12), 3715–3730. https://doi.org/10.1111/j.1365-294X.2006.03050.x
670 Hodel, R. G. J., Segovia-Salcedo, M. C., Landis, J. B., Crowl, A. A., Sun, M., Liu, X., … Soltis,
671 P. S. (2016). The Report of My Death was an Exaggeration: A Review for Researchers
672 Using Microsatellites in the 21st Century. Applications in Plant Sciences, 4(6), 1600025.
673 https://doi.org/10.3732/apps.1600025
674 Hofmann, H. a, Benson, M. E., & Fernald, R. D. (1999). Social status regulates growth rate:
675 consequences for life-history strategies. Proceedings of the National Academy of
676 Sciences of the United States of America, 96(24), 14171–14176.
677 https://doi.org/10.1073/pnas.96.24.14171
678 Jin, Y., Schäeffer, A. A., Sherry, S. T., & Feolo, M. (2017). Quickly identifying identical and
679 closely related subjects in large databases using genotype data. PLoS ONE, 12(6).
680 https://doi.org/10.1371/journal.pone.0179106

681 Jones, A. G., & Ardren, W. R. (2003, October 1). Methods of parentage analysis in natural
682 populations. Molecular Ecology. John Wiley & Sons, Ltd (10.1111).
683 https://doi.org/10.1046/j.1365-294X.2003.01928.x
684 Jones, A. G., Small, C. M., Paczolt, K. A., & Ratterman, N. L. (2010, January 1). A practical
685 guide to methods of parentage analysis. Molecular Ecology Resources. John Wiley &
686 Sons, Ltd (10.1111). https://doi.org/10.1111/j.1755-0998.2009.02778.x
687 Kalinowski, S. T., Taper, M. L., & Marshall, T. C. (2007). Revising how the computer program
688 CERVUS accommodates genotyping error increases success in paternity assignment.
689 Molecular Ecology, 16(5), 1099–1106. https://doi.org/10.1111/j.1365-
690 294X.2007.03089.x
691 Kardos, M., Luikart, G., & Allendorf, F. W. (2015). Measuring individual inbreeding in the age
692 of genomics: Marker-based measures are better than pedigrees. Heredity, 115(1), 63–72.
693 https://doi.org/10.1038/hdy.2015.17
694 Kellogg, K. A., Markert, J. A., Stauffer, J. R., & Kocher, T. D. (1995). Microsatellite variation
695 demonstrates multiple paternity in lekking cichlid fishes from Lake Malawi, Africa.
696 Proceedings of the Royal Society B: Biological Sciences, 260(1357), 79–84.
697 https://doi.org/10.1098/rspb.1995.0062
698 Kidd, M. R., Danley, P. D., & Kocher, T. D. (2006). A direct assay of female choice in cichlids:
699 all the eggs in one basket. Journal of Fish Biology, 68(2), 373–384.
700 https://doi.org/10.1111/j.0022-1112.2006.00896.x
701 Kidd, M. R., O'Connell, L. A., Kidd, C. E., Chen, C. W., Fontenot, M. R., Williams, S. J., &
702 Hofmann, H. A. (2013). Female preference for males depends on reproductive

physiology in the African cichlid fish Astatotilapia burtoni. General and Comparative Endocrinology, 180(1), 56–63. https://doi.org/10.1016/j.ygcen.2012.10.014

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics, 15(1), 356. https://doi.org/10.1186/s12859-014-0356-4

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Luan, T., Woolliams, J. A., Degård, J., Dolezal, M., Roman-Ponce, S. I., Bagnato, A., & Meuwissen, T. H. E. (2012). The importance of identity-by-state information for the accuracy of genomic selection. Genetics Selection Evolution, 44(1), 28. https://doi.org/10.1186/1297-9686-44-28

Marshall, T. C., Slate, J., Kruuk, L. E. B., & Pemberton, J. M. (1998). Statistical confidence for likelihood-based paternity inference in natural populations. Molecular Ecology, 7(5), 639–655. https://doi.org/10.1046/j.1365-294x.1998.00374.x

Maruska, K. P. (2015). Social Transitions Cause Rapid Behavioral and Neuroendocrine Changes. Integrative and Comparative Biology, 55(2), 294–306. https://doi.org/10.1093/icb/icv057

Meagher, T. R., & Thompson, E. (1986). The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. Theoretical Population Biology, 29(1), 87–106. https://doi.org/10.1016/0040-5809(86)90006-7

Palaiokostas, C., Clarke, S. M., Jeuthe, H., Brauning, R., Bilton, T. P., Dodds, K. G., … de Koning, D. J. (2020). Application of low coverage genotyping by sequencing in

selectively bred arctic charr (*Salvelinus alpinus*). G3: Genes, Genomes, Genetics, 10(6), 2069–2078. https://doi.org/10.1534/g3.120.401295

Pauquet, G., Salzburger, W., & Egger, B. (2018). The puzzling phylogeography of the haplochromine cichlid fish Astatotilapia burtoni. Ecology and Evolution, 8(11), 5637–5648. https://doi.org/10.1002/ece3.4092

Pool, J. E., Hellmann, I., Jensen, J. D., & Nielsen, R. (2010, March 1). Population genetic inference from genomic sequence variation. Genome Research. Cold Spring Harbor Laboratory Press. https://doi.org/10.1101/gr.079509.108

Premachandra, H. K. A., Nguyen, N. H., & Knibb, W. (2019). Effectiveness of SNPs for parentage and sibship assessment in polygamous yellowtail kingfish *Seriola lalandi*. Aquaculture, 499, 24–31. https://doi.org/10.1016/j.aquaculture.2018.09.022

Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., Bolnick, D. I., & Bird, C. E. (2014). Demystifying the RAD fad. Molecular Ecology, 23(24), 5937–5942. https://doi.org/10.1111/mec.12965

Rosyara, U. (2014). paternity: Paternity tests using SNPs. Retrieved from https://r-forge.r-project.org/projects/familyassoc

Salzburger, W. (2018). Understanding explosive diversification through cichlid fish genomics. Nature Reviews Genetics. Nature Publishing Group. https://doi.org/10.1038/s41576-018-0043-9

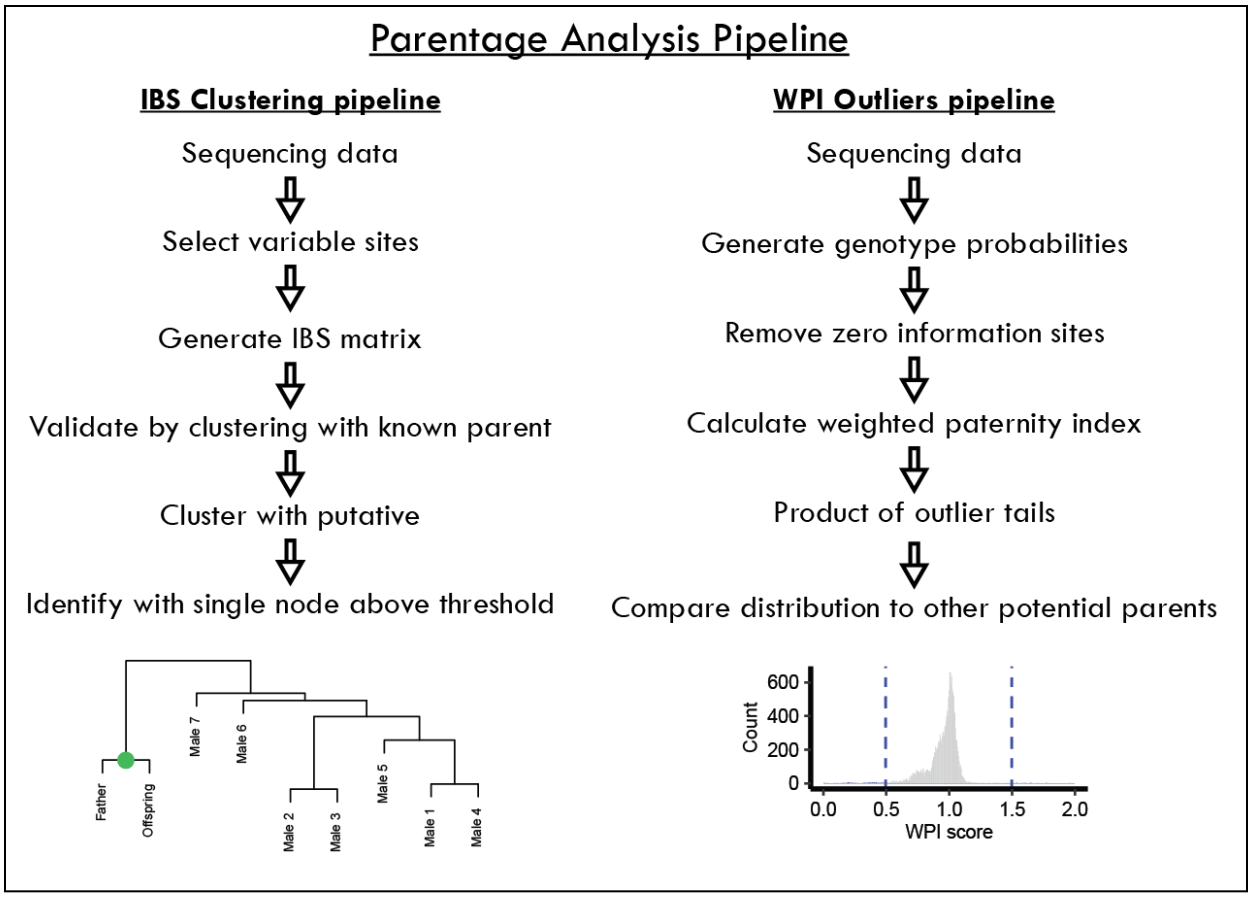
Sanetra, M., Henning, F., Fukamachi, S., & Meyer, A. (2009). A microsatellite-based genetic linkage map of the cichlid fish, Astatotilapia burtoni (Teleostei): A comparison of genomic architectures among rapidly speciating cichlids. Genetics. https://doi.org/10.1534/genetics.108.089367

748 Stevens, E. L., Heckenberg, G., Roberson, E. D. O., Baugher, J. D., Downey, T. J., & Pevsner, J.
749 (2011). Inference of Relationships in Population Data Using Identity-by-Descent and
750 Identity-by-State. PLoS Genetics, 7(9), e1002287.
751 https://doi.org/10.1371/journal.pgen.1002287
752 Suzuki, R., & Shimodaira, H. (2006). Pvclust: An R package for assessing the uncertainty in
753 hierarchical clustering. Bioinformatics, 22(12), 1540–1542.
754 https://doi.org/10.1093/bioinformatics/btl117
755 Theis, A., Salzburger, W., Egger, B., & Steinke, D. (2012). The Function of Anal Fin Egg-Spots
756 in the Cichlid Fish *Astatotilapia burtoni*. PLoS ONE, 7(1). https://doi.org/10.1371/
757 Thompson, E. A. (1976). A paradox of genealogical inference. Advances in Applied Probability,
758 8(04), 648–650. https://doi.org/10.2307/1425927
759 Thompson, E. A., & Meagher, T. R. (1987). Parental and Sib Likelihoods in Genealogy
760 Reconstruction. Biometrics, 43(3), 585. https://doi.org/10.2307/2531997
761 Tokarska, M., Marshall, T., Kowalczyk, R., Wójcik, J. M., Pertoldi, C., Kristensen, T. N., …
762 Bendixen, C. (2009). Effectiveness of microsatellite and SNP markers for parentage and
763 identity analysis in species with low genetic diversity: The case of European bison.
764 Heredity, 103(4), 326–332. https://doi.org/10.1038/hdy.2009.73
765 Wang, J. (2010). Effects of genotyping errors on parentage exclusion analysis. Molecular
766 Ecology, 19(22), 5061–5078. https://doi.org/10.1111/j.1365-294X.2010.04865.x
767 Wang, J. (2004). Sibship Reconstruction from Genetic Data with Typing Errors. Genetics,
768 166(4), 1963–1979. https://doi.org/10.1534/genetics.166.4.1963
769 Wang, S., Meyer, E., Mckay, J. K., & Matz, M. V. (2012). 2b-rad: a simple and flexible method
770 for genome-wide genotyping. https://doi.org/10.1038/nmeth.2023

771 Yang, J., Lin, D., Deng, C., Li, Z., Pu, Y., Yu, Y., … Chen, F. (2019, August 1). The advances in
772 DNA mixture interpretation. Forensic Science International. Elsevier Ireland Ltd.
773 https://doi.org/10.1016/j.forsciint.2019.05.024
774
775

776 **FIGURES**

777



778

779 **Figure 1:** Parentage analysis pipelines developed for complementary methods from next-
780 generation sequencing data utilizing IBS matrix clustering, which is responsive to relatedness
781 among samples, and WOPI outliers, directly incorporates maternal data.

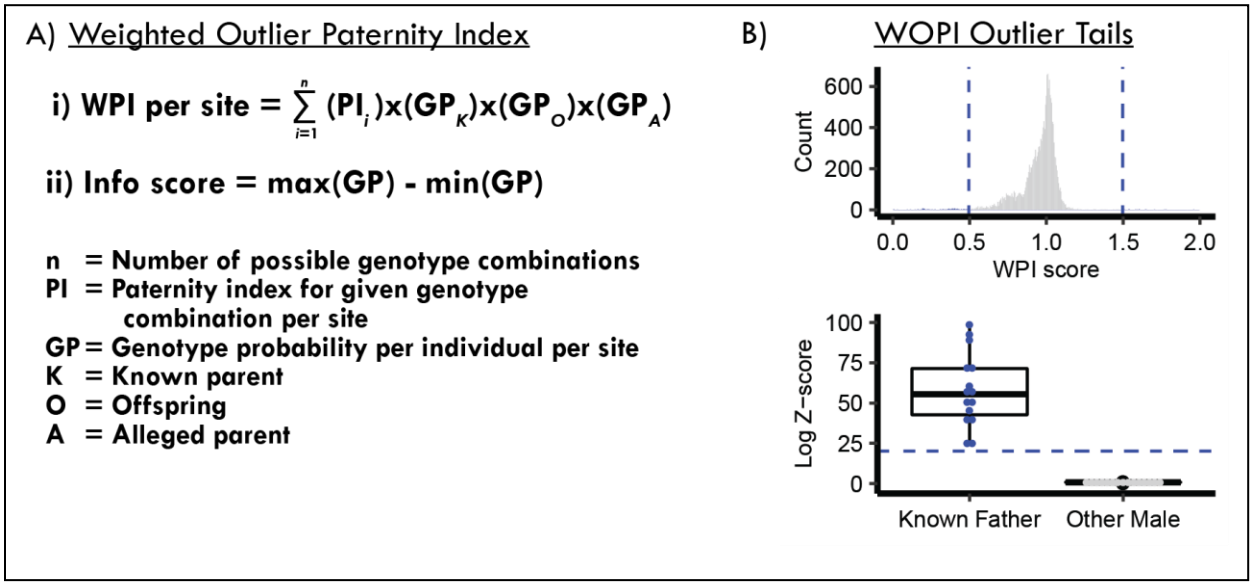


Figure 2: The weighted outlier paternity index (WOPI) method adapts a Bayesian approach to parentage analysis for next-generation sequencing to identify fathers from a pool of samples and requires the mother being known. A) Two equations used in the WOPI pipeline. (i) Weighted paternity index values are calculated for each site taking the traditional paternity index weighted by the probability of a specific genotype combination summed across all possible genotype combinations. (ii) The information score is calculated to filter out sites with no genotyping information as a technique to reduce noise. The info score for a site in an individual is calculated as the difference between the maximum and minimum genotype probability of the three possible genotypes. An info score of zero indicates that there is an equal probability (i.e. 0.33) chance that at that site an individual is any genotype and is therefore filtered out. B) A histogram of weighted paternity index (WPI) scores from a sample for which the correct father is known with dashed lines denoting outlier cutoffs (Top). The WOPI outlier score is the product of the tails of the distribution outside the theoretical outlier cutoffs. Known fathers have WOPI outlier scores above a z-score threshold when compared to the distribution of the other alleged parents (Bottom).

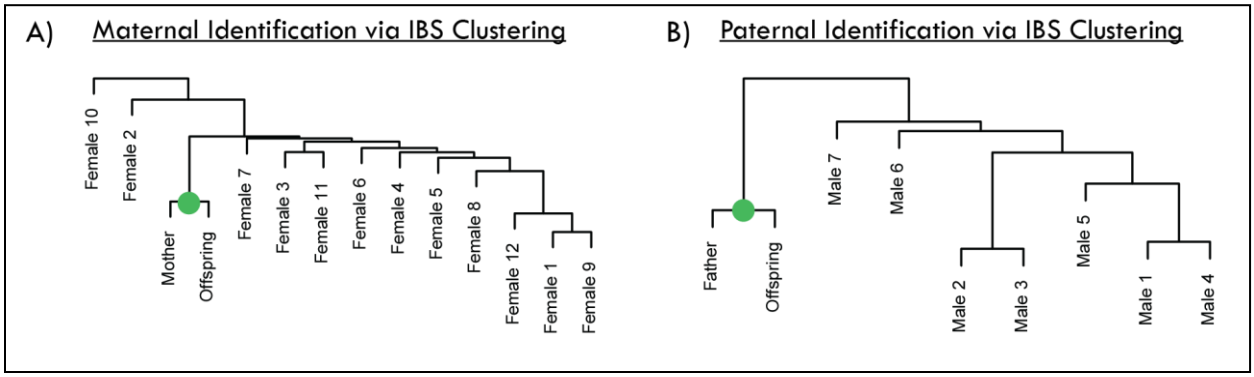
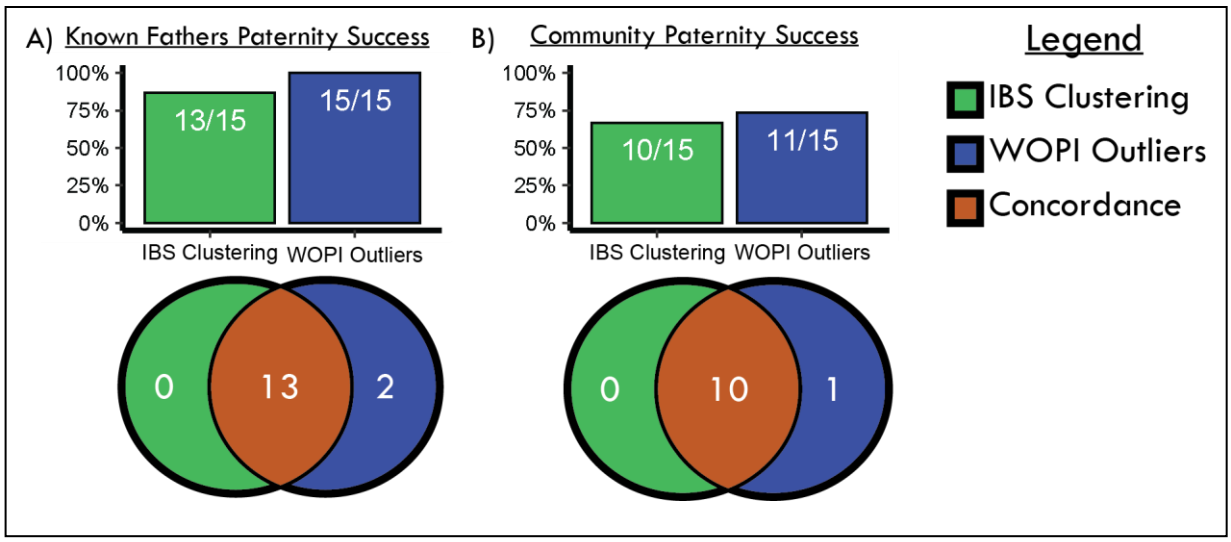


Figure 3: Hierarchal clustering of IBS distance matrix from known triad experiment. Limiting the samples in the matrix to potential parents of one sex and offspring creates a dendrogram wherein the offspring pair with their parent at the first internal node (green dot). Paternity or maternity was assigned if the first internal node from the offspring connected to a single individual and was above AU p-value threshold. A) A sample offspring, with known maternity and paternity (same as Fig 2.), clustered with a pool of all the females from the known triad experiment to check appropriate pairing of known mother at first node. B) Dendrogram generated using a pool of males from the known triad experiment and same sample offspring correctly identifies known father.



808

809 **Figure 4:** Complementary methods successfully identify paternity with concordance between
810 methods. Bar charts represent the number of paternity calls made by each technique, respectively.
811 Venn diagrams show the overlap between the samples that received paternity assignments. A)
812 Across both methods, all 15/15 broods were assigned the correct father when using triads with
813 known paternity. B) Across six naturalistic communities, conservative paternity methods assigned
814 paternity to 11/15 broods, after filtering the offspring that failed to appropriately cluster with their
815 respective mother via IBS matrix clustering.

816