Title

Efficient Genome-Wide Sequencing and Low-Coverage Pedigree Analysis from Noninvasively Collected Samples.

Journal

Genetics, 203(2)

ISSN

0016-6731

Authors

Snyder-Mackler, Noah
Majoros, William H
Yuan, Michael L
et al.

Publication Date

2016-06-01

DOI

10.1534/genetics.116.187492

Peer reviewed

Efficient Genome-Wide Sequencing and Low-Coverage Pedigree Analysis from Noninvasively Collected Samples

Noah Snyder-Mackler,* William H. Majoros,† Michael L. Yuan,* Amanda O. Shaver,* Jacob B. Gordon,‡ Gisela H. Kopp,§,,†† Stephen A. Schlebusch,‡‡ Jeffrey D. Wall,§§ Susan C. Alberts,*,‡,*** Sayan Mukherjee,†††,‡‡‡,§§§ Xiang Zhou,****,††††,1 and Jenny Tung*,‡,***,††††,1**

*Department of Evolutionary Anthropology, †Graduate Program in Computational Biology and Bioinformatics, ‡Department of Biology, †††Department of Statistical Science, ‡‡‡Department of Mathematics, §§§Department of Computer Science, and ††††Duke University Population Research Institute, Duke University, Durham, North Carolina 27708, §Cognitive Ethology Laboratory, German Primate Center, Leibniz Institute for Primate Research, 37077 Göttingen, Germany, **Department of Biology, University of Konstanz, 78457 Konstanz, Germany, ††Department of Migration and Immuno-Ecology, Max Planck Institute for Ornithology, 82319 Radolfzell, Germany, ‡‡Department of Molecular and Cell Biology, University of Cape Town, 7700 Cape Town, South Africa, §§Institute for Human Genetics, University of California, San Francisco, California 94143, ***Institute of Primate Research, National Museums of Kenya, Nairobi, Kenya, and ****Department of Biostatistics and ††††Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan 48109

ORCID IDs: 0000-0003-3026-6160 (N.S.-M.); 0000-0003-3026-6160 (W.H.M.); 0000-0003-0416-2958 (J.T.)

ABSTRACT Research on the genetics of natural populations was revolutionized in the 1990s by methods for genotyping noninvasively collected samples. However, these methods have remained largely unchanged for the past 20 years and lag far behind the genomics era. To close this gap, here we report an optimized laboratory protocol for genome-wide capture of endogenous DNA from noninvasively collected samples, coupled with a novel computational approach to reconstruct pedigree links from the resulting low-coverage data. We validated both methods using fecal samples from 62 wild baboons, including 48 from an independently constructed extended pedigree. We enriched fecal-derived DNA samples up to 40-fold for endogenous baboon DNA and reconstructed near-perfect pedigree relationships even with extremely low-coverage sequencing. We anticipate that these methods will be broadly applicable to the many research systems for which only noninvasive samples are available. The lab protocol and software ("WHODAD") are freely available at www.tung-lab.org/protocols-and-software.html and www.xzlab.org/software.html, respectively.

KEYWORDS capture-based enrichment; noninvasive samples; baboons; paternity analysis; pedigree; genome resequencing

THE capacity to generate genetic data from low-quality or noninvasively collected samples, first developed in the 1990s, revolutionized the study of genetics, evolution, behavior, and ecology in natural populations. These methodological advances facilitated phylogenetic and phylogeographic analyses of difficult-to-sample taxa; helped define the role of admixture in mammalian evolution (Pérez *et al.* 2010; Sacks *et al.* 2011; Charpentier *et al.* 2012); and enabled theoretical expectations about paternal investment, kin recognition, and reproductive skew to be empirically tested, sometimes for the first time (Buchan *et al.* 2003; Smith *et al.* 2003; Archie *et al.* 2007; Gottelli *et al.* 2007). They also yielded important insights into the genetic viability and future prospects of threatened or endangered populations from which invasive samples are impossible to obtain (Idaghdour *et al.* 2003; Valière *et al.* 2003; Nagata *et al.* 2005; Rudnick *et al.* 2007; Mondol *et al.* 2009). Noninvasive genetic analysis has thus changed the ways

doi: 10.1534/genetics.116.187492
Manuscript received January 25, 2016; accepted for publication April 18, 2016; published Early Online April 19, 2016.
Available freely online through the author-supported open access option.
Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.187492/-/DC1.
[1]Corresponding authors: Department of Evolutionary Anthropology, Duke University, 104 Biological Sciences Bldg., Box 90383, Durham, NC 27708-9976. E-mail: jt5@duke.edu; and Department of Biostatistics, University of Michigan, 1415 Washington Heights no. 4623, Ann Arbor, MI 48109. E-mail: xzhousph@umich.edu

we study population, ecological, and conservation genetics, and we would know far less about many species without it.

However, techniques for noninvasive genetic analysis have changed little in the past 20 years. Collection of genetic data from noninvasively collected tissues (*e.g.*, feces, hair, and urine) continues to be labor intensive, time intensive, and vulnerable to technical artifacts such as allelic dropout and cross-contamination (Gagneux *et al.* 1997; Taberlet *et al.* 1999). Further, current methods ultimately yield very small amounts of data by today's standards. Typical studies genotype up to several dozen microsatellite loci per individual—a trivial amount compared to the data sets now routinely generated using standard high-throughput sequencing approaches. Thus, while existing methods are sufficient for basic pedigree construction and estimating some population genetic parameters (although usually with substantial uncertainty), they are severely underpowered for many other types of analyses (Sabeti *et al.* 2002; Price *et al.* 2009; Li and Durbin 2011), including any that require local (*i.e.*, gene- or region-specific) information instead of genome-wide averages (Huang *et al.* 2007; Li *et al.* 2007; Sankararaman and Sridhar 2008; Yang *et al.* 2011; Ma *et al.* 2014). Further, because noninvasively collected genotype data are most often based on microsatellites, they cannot take advantage of new tools designed specifically for single-nucleotide variants (Purcell *et al.* 2007; Visscher 2009; Durand *et al.* 2011).

Generating genome-scale data sets from noninvasive samples is challenging for two reasons. First, in many cases, the DNA extracted from these samples is low quality and highly fragmented. Second, it contains large proportions of nonhost DNA. For example, only ~1% of DNA extracted from fecal-derived samples is endogenous to the donor animal [most is microbial (Perry *et al.* 2010)]. Sequence capture methods, in which synthesized baits are used to enrich for prespecified target sequences from a larger DNA pool (Gnirke *et al.* 2009), present a potential solution to both of these problems. Because shearing is a required step in library preparation, the problem of working with highly fragmented samples is obviated. Indeed, Perry *et al.* (2010) were able to use a modified version of sequence capture to target and sequence 1.5 Mb of the chimpanzee genome from fecal-derived DNA, with very low genotyping error rates relative to blood-derived DNA. More recently, Carpenter *et al.* (2013) reported a method for performing genome-wide sequence capture from low-quality ancient DNA samples, which recapitulate many of the challenges posed by noninvasive samples (*e.g.*, highly fragmented DNA and low proportions of endogenous DNA).

However, while considerable investment in single samples often makes sense in ancient DNA studies, the low levels of postcapture enrichment associated with currently available protocols are not cost effective for population studies of noninvasive samples. Substantially higher rates of enrichment, particularly in nonrepetitive regions of the genome, will be essential to overcome this limitation. In addition, computational methods for analyzing the resulting data are also required, especially given that genome-scale sequencing efforts for such samples are likely to produce low-coverage data. For example, current paternity assignment approaches (Chakraborty *et al.* 1974; Marshall *et al.* 1998; Kalinowski *et al.* 2007) were not designed to deal with uncertain genotypes, an inevitable component of analyzing low-coverage sequencing data. Thus, for capture-based methods to become broadly accessible, the development of appropriate new computational approaches is also essential.

Here, we report an optimized laboratory protocol for genome-wide capture of endogenous DNA from noninvasively collected samples, combined with a novel computational approach to reconstruct pedigree links from the resulting data (implemented in the program WHODAD). We validate both our laboratory methods and computational tools, using noninvasively collected samples from 54 members of an intensively studied wild baboon population in the Amboseli basin of Kenya (Alberts and Altmann 2012). We also demonstrate the generalizability of our methods to noninvasive samples collected using different methods from a different baboon species from West Africa. Our protocol is cost effective, has manageable sample input requirements, yields good capture efficiency for high complexity, nonrepetitive DNA, and minimizes the need for extensive PCR amplification. Importantly, we find that genotype data generated from fecal samples closely match data from high-quality blood-derived DNA samples from the same individuals, and provide near-perfect information on pedigree relationships even with extremely low per-sample sequencing coverage (mean = 0.49× genome coverage). Together, these methods will enable population, conservation, and ecological genetic analyses of natural populations to again take a major leap forward, into the genomic era. At the same time, they will also introduce new systems to the genomics community.

Results

DSN digestion during bait construction increases library complexity

Our protocol relies on *in vitro* transcription of biotinylated RNA baits to capture host-specific DNA from the mixed pool of host, environmental, and microbial DNA extracted from noninvasive samples. Similar to Carpenter *et al.* (2013), RNA baits are generated from DNA templates obtained from a high-quality DNA sample (here, DNA extracted from blood). This approach avoids the high cost of custom bait synthesis (as in Perry *et al.* 2010 and Gnirke *et al.* 2009), but can also produce a bait set that includes a large proportion of low-complexity, repetitive regions. Consequently, many reads generated from captured DNA cannot be uniquely mapped, lowering the protocol's efficiency. To address this concern, we incorporated a novel duplex specific nuclease (DSN) digestion in the bait construction step (Supplemental Material, Figure S1A; see *Methods*). Sequencing the DNA bait templates prior to *in vitro* amplification demonstrates that including the digestion step reduces the percentage of baits

synthesized from low-complexity/highly duplicated regions. Specifically, a 4-hr incubation of sheared DNA at 68° followed by a 20-min DSN digestion in the presence of human Cot-1 produced the highest-complexity bait library of the five conditions we tested. Compared to DNA templates from a non-DSN-digested library, bait templates produced using these conditions reduced the number of reads mapping to multiple locations by 2.6-fold (from 19.2% to 7.5%; Figure S2).

Capture-based enrichment

We validated our full capture protocol (bait construction followed by capture of endogenous DNA and sequencing of captured fragments), using fecal-derived DNA (fDNA) samples collected from 54 individually recognized yellow baboons (36 males and 18 females; Figure 1) from the Amboseli baboon population, an intensively studied population in which maternal and paternal pedigree relationships are known for a large set of individuals (Buchan *et al.* 2003; Alberts *et al.* 2006; Alberts and Altmann 2012). We produced data for 52 of the samples in two successive capture efforts: "capture 1" was conducted on fDNA from 24 baboons, and "capture 2" was conducted on fDNA from 28 additional baboons after making multiple improvements to our initial protocol (changes to the protocol between capture efforts are described in detail in Table S1 and File S1; see Table S2 for information on sequencing coverage and mapping statistics). Data from the remaining two individuals, "LIT" and "HAP," were generated to compare the captured fDNA sample with data derived from sequencing blood-derived genomic DNA (gDNA) samples from the same individuals.

Our protocol (Figure S1) resulted in substantial enrichment of baboon DNA in the postcapture *vs.* precapture samples (see Table S2 for sample-specific details). A mean of 44.56% (range: 10.28–83.17%) of postcapture fragments mapped to the yellow baboon genome (*Pcyn1.0*), despite starting with precapture samples that contained a mean of only 2.04% endogenous baboon DNA, as estimated by quantitative PCR (qPCR) (range 0.19–8.37%). However, in capture 1 a large proportion of the mapped fragments were identified as PCR duplicates ($\text{mean}_{\text{capture1}}$ = 71.97% of mapped fragments, $\text{range}_{\text{capture1}}$ = 51.43–88.46%; Figure 2A). After removing PCR duplicates, a mean of 9.16% of the postcapture reads in capture 1 were nonduplicate mappable fragments ($\text{range}_{\text{capture1}}$ = 2.23–23.75%), producing a mean coverage of 0.20× per sample relative to the mappable baboon genome (mean sequencing depth of 5.8 Gb per sample; $\text{range}_{\text{capture1}}$ = 0.04–0.49×; Figure 2B). These numbers translated to an overall mean fold enrichment of 39.8× for mapped reads ($\text{range}_{\text{capture1}}$ = 8.0–111.8-fold, SD = 25.2) and 9.6× enrichment of non-PCR duplicate mapped reads ($\text{range}_{\text{capture1}}$ = 3.9–22.4-fold, SD = 5.0; Figure 2C).

Based on our results for capture 1, we made multiple protocol improvements prior to conducting capture 2 (Table S1 and File S1). The improved protocol was twice as effective on average, resulting in a mean 18-fold enrichment of high-quality, analysis-ready reads and a maximum fold enrichment of close to 40-fold [$\text{range}_{\text{capture2}}$ = 8.0–39.2-fold, Figure 2C; by comparison, methods optimized for ancient DNA achieved a mean of 5.5-fold enrichment of non-PCR duplicate fragments (Carpenter *et al.* 2013), Figure 2A]. Specifically, the protocol changes improved the proportion of nonduplicate mapped fragments by >4-fold, from a mean proportion of 9.16% in capture 1 to a mean proportion of 37.74% in capture 2 ($\text{range}_{\text{capture2}}$ = 6.16–68.61%), and reduced the proportion of PCR duplicates among mapped reads 2-fold (from 71.97% in capture 1 to 36.97% in capture 2). This improvement translated to an increase in overall genomic coverage from a mean of 0.20× in capture 1 to 0.73× in capture 2 (mean total sequencing of 5.7 Gb per sample; $\text{range}_{\text{capture2}}$ = 0.19–1.24×; Figure 2B). This improvement in coverage was not explained by increased sequencing depth in capture 2 (Table S2). Thus, while we would need to sequence a precapture fDNA sample 50–100 times as deeply as a blood- or tissue-derived sample to produce the same level of coverage, our capture method reduces this difference to ∼2 times the sequencing effort. Importantly, our method was also successful in enriching fDNA samples (n = 8) from independent samples collected from Guinea baboons (*Papio papio*; Figure 2A, Table S2), suggesting that our results are highly generalizable across different species and storage and extraction methods.

Sample attributes influencing capture efficiency

The amount of baboon DNA in the precapture fDNA sample was the strongest predictor of enrichment success. Specifically, the percentage of baboon DNA precapture, as assessed via qPCR, was positively correlated with the percentage of nonduplicate fragments mapped postcapture (Figure 2D; T = 6.88, $P = 1.72 \times 10^{-8}$). Samples from capture 2 had more precapture baboon DNA than samples used in capture 1 because we attempted to optimize the input samples based on our initial analyses in capture 1 (capture 1 mean = 1.21%, range = 0.19–4.90%; capture 2 mean = 2.80%, range = 0.25–8.37%). However, even when controlling for this difference, enrichment of samples from capture 2 was improved over that of capture 1. This pattern is observable whether assessed using the percentage of baboon DNA fragments sequenced postcapture ($T_{\text{capture2}} = 10.00, P = 6.76 \times 10^{-13}$) or assessed using fold enrichment relative to precapture amounts ($T_{\text{capture2}} = 6.89$, $P = 1.69 \times 10^{-8}$) and could not be explained by differences in the length of sequence fragments or overall sequencing depth (Figure S3, Table S2). The amount of fDNA library used in the capture reaction was also weakly positively correlated with the percentage of baboon DNA fragments sequenced postcapture, after controlling for the amount of baboon DNA in the precapture sample ($T_{\text{ng\_fDNA\_library}} = 2.09$, $P = 0.042$; Table S2).

Library complexity, distribution of reads, and GC content

The postcapture libraries included a higher proportion of PCR duplicates relative to reads generated from high-quality

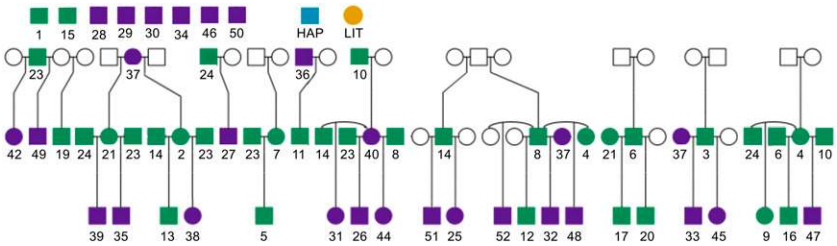


Figure 1 Pedigree of a subset of baboons monitored by the Amboseli Baboon Research Project. Samples from both males (squares) and females (circles) were enriched in capture 1 (green) or capture 2 (purple). Open circles and squares represent baboons that connect individuals in our pedigree, but who were not sequenced as part of this study. Each sequenced individual is represented by a unique number (below the circles or squares), with some individuals repeated because baboons often produce offspring with multiple mates. The paired fDNA and gDNA samples came from two individuals, HAP (blue) and LIT (orange), who were members of the study population but are not connected to this pedigree.

genomic DNA samples, for which fewer rounds of PCR amplification were required (PCR duplicate proportion: $\text{mean}_{\text{fDNA\_capture1}}$ = 69.6%, $\text{mean}_{\text{fDNA\_capture2}}$ = 36.8%, $\text{mean}_{\text{gDNA}}$ = 11.3% of mapped reads; 18 rounds of PCR in the capture protocol *vs.* 6 rounds for the high-quality samples). For comparison, this proportion is much lower than reported for ancient DNA (aDNA) samples, which go through more rounds of PCR amplification ($\text{mean}_{\text{aDNA}}$ = 94.6%; Figure 2A and Figure S4; Carpenter *et al.* 2013). Despite increases in clonality, the number of nonduplicate reads continued to increase with increasing sequencing depth, with the slope of this relationship especially favorable for capture 2 (Figure 3). Thus, deeper sequencing of postcapture libraries should continue to increase genome-wide coverage, albeit not as efficiently as sequencing blood-derived gDNA samples.

As with other capture-based methods (Carpenter *et al.* 2013; Samuels *et al.* 2013), a modest fraction of the mapped fragments mapped to the mitochondrial genome (mtDNA). When we included all mapped reads, this fraction was similar in libraries from capture 1 and capture 2 ($\text{mean}_{\text{capture1}}$ = 6.55%; $\text{mean}_{\text{capture2}}$ = 6.73%; Figure S5A). However, capture 2 resulted in significantly more unambiguously nonduplicate mtDNA-mapped reads than capture 1, largely due to the paired-end sequencing used in capture 2 ($\text{mean}_{\text{capture1}}$ = 0.47% of all mapped reads; $\text{mean}_{\text{capture2}}$ = 6.46%; Figure S5B). The higher number of nonduplicate mtDNA reads in capture 2 thus produced much deeper overall coverage of the mitochondrial genome (Figure S5C), despite the fact that the ratio of mtDNA to nuclear DNA mapped reads was comparable between the two captures (Figure S5D). Finally, the distributions of read GC content for postcapture reads using our protocol, the DNA template for the RNA baits, and aDNA libraries were highly similar (Figure S6). This observation suggests that any GC bias relative to the genome appears during bait construction and/or sequencing, not during the hybridization step.

Postcapture fDNA-derived genotype data are consistent with individual identity and independently established pedigree relationships

To assess the accuracy of genotypes called from postcapture fDNA libraries, we compared genotype data from paired blood-derived gDNA (without capture) and postcapture fDNA libraries for two individuals, LIT and HAP. Using genotypes for sites that were called with a genotype quality (GQ) > 20 in both the fDNA and gDNA data sets for either LIT or HAP, we found that the majority of the genotypes called in both data sets were concordant (86.5% of 312,739 sites for the LIT paired samples; 77% of 40,132 sites for the HAP paired samples, for whom we had much lower coverage for the fecal-derived sample). As expected, the majority of the discordant sites occurred when the low-coverage fDNA sample was called as homozygous and the high-coverage gDNA sample was called as heterozygous (77.7% and 74.4% of discordant sites in LIT and HAP, respectively; Figure S7). Further, among all sites, the fDNA genotype captured at least one of the alleles from the gDNA genotype in 99.8% (LIT) and 99.6% (HAP) of cases (Figure S7). Thus, even when genotypes called in fDNA and gDNA samples from the same individual were discordant, they were almost always compatible.

Further, we found that genotypes called from the postcapture fDNA libraries were more similar to the genotypes called from their high-quality gDNA counterparts than they were to those from other postcapture fDNA libraries. Specifically, $k0$ values from *lcMLkin* (Lipatov *et al.* 2015), which estimate the probability that two samples share no alleles that are identical by descent, were much smaller for the LIT_{fDNA}–LIT_{gDNA} paired samples (0.487) and HAP_{fDNA}–HAP_{gDNA} paired samples (0.243) than for $k0$ values calculated for the two blood-derived samples when compared to any other fDNA sample ($k0$ range LIT_{fdna} *vs.* other fDNA samples = 0.996–1.000; $Z = 849.2, P < 10^{-20}$; $k0$ range HAP_{fDNA} *vs.* other fDNA samples = 0.786–0.999; $Z = 10.6, P < 10^{-20}$; Figure 4A).

For the 48 extended-pedigree individuals (Figure 1, including 8 Amboseli baboons with no known relatives in the pedigree), we then tested whether the estimated coefficient of relatedness values, r, from *lcMLkin* (Lipatov *et al.* 2015) in the postcapture data (range: 0–1, or 2× the kinship coefficient) were correlated with coefficient of relatedness values obtained from the independently constructed pedigree (based on known mother–offspring relationships and microsatellite-based paternity assignments: see *Methods*). Using a filtered set of 127,654 single-nucleotide variants (see *Methods*), we found a strong correlation between the two

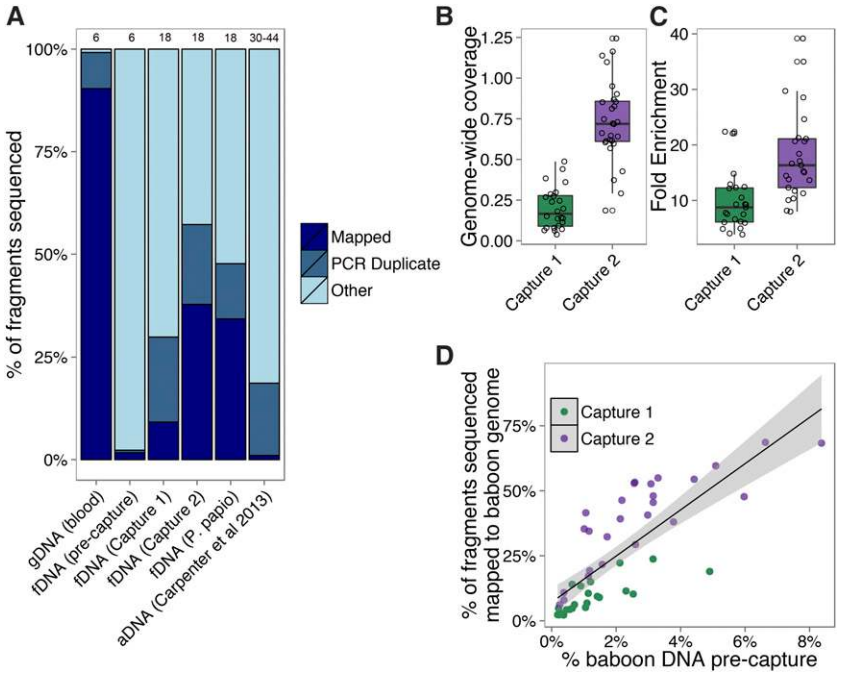


Figure 2 fDNA enrichment results. (A) Percentage of sequencing reads that mapped to the baboon genome and were not PCR duplicates ("Mapped," dark blue), mapped and were PCR duplicates ("PCR Duplicate," blue), or did not map and likely represent environmental or bacterial DNA in the case of fDNA/aDNA and unmappable fragments in the case of gDNA ("Other," light blue). "gDNA" represents genomic DNA derived from the blood samples for LIT and HAP; "aDNA" represents ancient DNA data from capture-based enrichment reported in Carpenter *et al.* (2013). Numbers above each bar show the total number of PCR cycles used in each protocol. (B) Capture 2 produced significantly greater genome coverage than capture 1, despite a similar number of reads generated per sample (two-sample *t*-test, $T = 9.7$, $P = 3.0 \times 10^{-12}$). On average in capture 2, we obtained ~0.73× coverage of the genome with 5.76 Gb of sequencing. If all 5.76 Gb mapped to the baboon genome as non-PCR duplicates, we would have produced ~2.2× genome-wide coverage. (C) Capture 2 also produced significantly greater fold enrichment of baboon DNA (fold enrichment is measured as percentage of nonduplicate baboon DNA postcapture divided by percentage of baboon DNA precapture: two-sample *t*-test, $T = 4.4$, $P = 7.3 \times 10^{-5}$). (D) The amount of baboon DNA in the sample precapture [percentage of baboon DNA precapture, based on qPCR of the single-copy *c-myc* gene (Morin *et al.* 2001)] is strongly correlated with the percentage of baboon fragments obtained in postenrichment sequencing (Pearson's $r = 0.80$, $P = 1.0 \times 10^{-11}$). However, even samples with low amounts of endogenous DNA (<2%) exhibit substantial fold enrichment using our protocol ($\text{mean}_{\text{capture1}} = 10.60\times$, $\text{mean}_{\text{capture2}} = 24.82\times$).

measures (Pearson's $r = 0.73$, $P < 10^{-16}$; Figure 4B). This correlation improved further if we imposed thresholds for the minimum number of sites genotyped in both individuals ("shared sites") in a dyad (Figure S8). For example, if we removed all dyads with <2000 shared sites (84 of 1128 dyads or 7.4%), the correlation between pedigree relatedness and genotype similarity reached Pearson's $r = 0.86$ ($P < 10^{-16}$). Notably, for one individual we prepared and sequenced capture libraries from two independently collected fecal samples (libraries AMB_018 and AMB_040). For these biological replicates, the pairwise relatedness value was 0.774, more than twice as high as for any other pair of relatives (range of estimates for parent–offspring and full-sib pairs typed at ≥2000 sites: 0.10–0.38). Thus, our methods readily distinguish replicate samples (which can be inadvertently collected, especially in unhabituated populations) from those collected from distinct individuals, even close relatives.

Paternity inference using WHODAD

Current methods for assigning paternity [*e.g.*, CERVUS (Marshall *et al.* 1998; Kalinowski *et al.* 2007) and exclusion (Chakraborty *et al.* 1974)] assume genotype certainty, such that individuals are assigned a deterministic genotype at each locus (*i.e.*, 0, 1, or 2 or a microsatellite repeat number; while a low level of measurement error due to sample mishandling can be modeled, this error rate is held constant across genotype calls). This assumption is violated in low-coverage sequencing data, in which genotypes are not known with certainty and this uncertainty varies across genotype calls. However, the relative *probabilities* of each genotype can be estimated, given estimated population allele frequencies and sequencing coverage information. To conduct paternity inference and pedigree reconstruction in this context, we therefore developed a novel approach to integrate information across low-coverage sites, implemented in the program WHODAD. Our method has two components. The first component identifies a top candidate male and tests whether he is significantly more related to the offspring than any other candidate male, using a *P*-value criterion. The second component tests whether the dyadic similarity between the top candidate and offspring is consistent with a parent–offspring dyad, using posterior probabilities obtained from a mixture model (see *Methods* and Figure S9).

Using WHODAD, we assigned paternity to all father–offspring pairs ($n = 27$) represented in the independently established extended pedigree in Figure 1. This approach is conservative because it departs from the usual practice of first identifying a likely set of candidate fathers based on demographic and prior pedigree information (the approach used in producing the pedigree in Figure 1). For 15 of the 27 offspring, we produced genotype data from the known mother with our enrichment protocol. WHODAD identified the same father as shown in the pedigree in 12 of these 15 trios (80%); in the other 3 trios (20%), no candidate male satisfied WHODAD's paternity assignment criteria (in all 3 of these cases, sequencing coverage was very low for either the pedigree-identified father or offspring: 0.04–0.17×). For the remaining 12 offspring, we did not generate genotype data using our enrichment protocol for their mothers. To test all 27

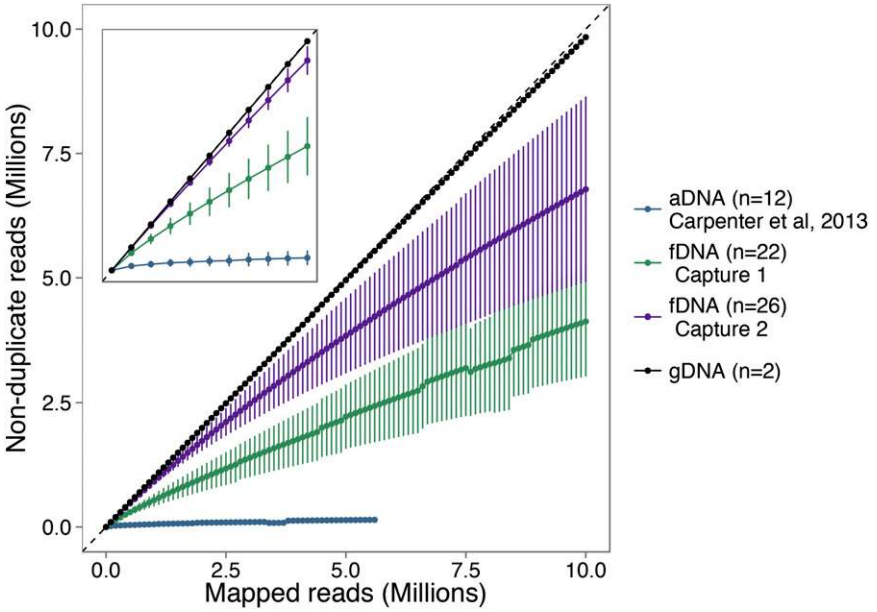


Figure 3 Increased sequencing effort produces increased numbers of nonduplicate reads. Shown is the number of mapped reads plotted against the number of nonduplicate reads mapped [mean ± SD; plotted using the program "preseq" (Daley and Smith 2013)]. More complex libraries (*i.e.*, those containing more nonduplicate fragments) have a slope closer to 1 (as in the case of the gDNA libraries), while less complex libraries have a shallower slope and asymptote at a smaller value. The main plot shows the first 10 million mapped reads for each sample. The inset shows the same plot for the first 1 million mapped reads.

father–offspring dyads together, we therefore reran WHODAD, excluding maternal genotype information. In this setting, WHODAD's paternity assignments agreed with the pedigree data in 22 of 27 (81%) cases (Figure 5). Notably, when the pedigree-identified father was included in the data set, WHODAD never assigned paternity to a different male, whether or not maternal genotype data were available. Because our method is highly robust to exclusion of maternal genotype data, we therefore performed all subsequent analyses assuming maternal genotype data were *not* available, a scenario that may often occur in studies of natural populations.

The presence of close relatives, such as full- or half-siblings, can influence the accuracy of paternity assignment if these close relatives are also included as candidate fathers (Thompson and Meagher 1987; Marshall *et al.* 1998; Olsen *et al.* 2001; Ford and Williamson 2010). Thus, to examine how the presence of close male kin influenced the accuracy and confidence of WHODAD's paternity assignments, we conducted three additional analyses. First, when all close male kin were removed from the candidate list of potential fathers ($r \geq 0.25$), but the father was retained, our method performed equivalently to the case when both father and close relatives were in the candidate pool. Second, when we removed all close male kin *including* the father, none of the best candidate fathers from the conditional probability analysis (0%) were assigned as fathers based on WHODAD's assignment criteria (Figure 5). Third, when we removed the father from the pool of candidate fathers, but included close male kin, 11% of the best remaining candidates (3 of 27 cases) were incorrectly assigned as fathers, based on comparison to the pedigree (Figure 5). All 3 of these false positives were close male relatives: in two cases WHODAD assigned the half-brother of the offspring as the likely father, and in one case WHODAD assigned the son of the offspring as the likely father. The best balance between maximizing the number of true positives while minimizing the number of false positives was achieved by combining both the *P*-value and mixture model criteria (see *Methods*). This approach outperformed either component used alone (Figure S10). For example, when all males were included in the candidate pool, the

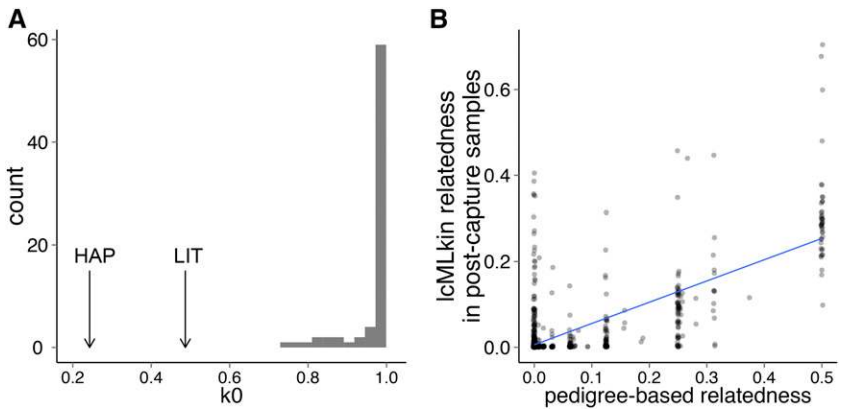


Figure 4 Postcapture genotype data are consistent with individual identity and pedigree relationships. (A) The *k*0 values for the HAP and LIT fDNA–gDNA paired samples (arrows) were significantly lower than the range of *k*0 values for LIT_{fDNA} and HAP_{fDNA} *vs.* any other fDNA sample (gray distribution). Lower *k*0 values reflect increased relatedness (*i.e.*, decreased probability of no IBD sharing). (B) Estimated dyadic coefficient of relatedness values (range: 0–1) were correlated with independently obtained pedigree relatedness values calculated using the R package *kinship2* (Sinnwell *et al.* 2014) (Pearson's $r = 0.73$, $P < 10^{-16}$). The blue line shows the best-fit slope and intercept from the linear model. Both *k*0 and the estimated relatedness values were calculated with *lcMLkin* (Lipatov *et al.* 2015).

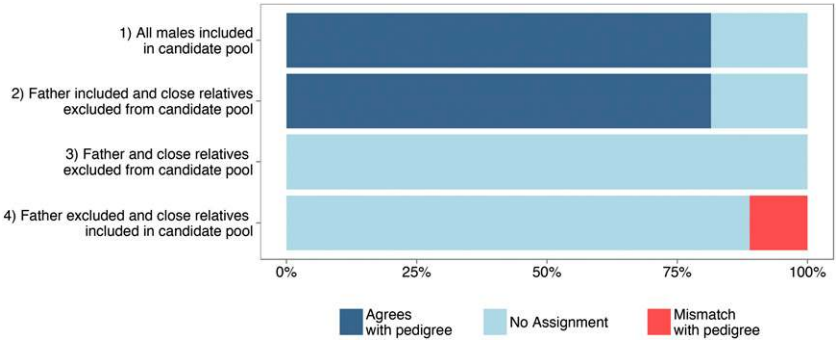


Figure 5 Paternity inference with WHODAD using low-coverage genotype calls. (1) When all males (n = 34) were included in the pool of candidate fathers (top bar), WHODAD assigned paternity to the same father identified in the pedigree for 22 of 27 (81%) of offspring (dark blue; see assignment criterion in *Methods*). The remaining offspring were not assigned a father based on WHODAD's assignment criteria, most likely due to low sequencing coverage (5 of 27; light blue). (2) WHODAD's accuracy was identical when we removed all close male relatives of the offspring ($r \geq 0.25$) from the pool of candidate fathers. (3) When we removed all close relatives, including fathers, from the candidate pool, no fathers were assigned, as expected. (4) Finally, when we removed the father from the candidate pool but retained close relatives, our method incorrectly assigned paternity to 11% of offspring (3 of 27; bottom bar). All three incorrectly assigned fathers were closely related to the offspring (in two cases the assigned father was the half-brother of the offspring and in one case the assigned father was the son of the offspring).

combined approach resulted in an 81% true positive rate and a 0% false positive rate, while using just the $k0$ values in a mixture model resulted in the same true positive rate (81%), but an additional 11% false positive rate (Figure S10).

Discussion

Our capture-based method strongly enriches the proportion of host DNA in low-quality DNA extracted from feces (fDNA). Our method is the first use of genome-wide enrichment-based capture methods (Carpenter *et al.* 2013; Enk *et al.* 2014; Ávila-Arcos *et al.* 2015) for noninvasively collected samples, which represent a major resource for behavioral, conservation, and evolutionary genetic studies in natural populations. Importantly, our protocol increases efficiency and lowers cost by reducing the input requirements (<1 μg) and number of PCR cycles relative to previous methods (Perry *et al.* 2010) and, in our final protocol, achieves up to 40-fold enrichment of postcapture endogenous DNA relative to precapture levels. We also show, for the first time since Perry *et al.* (2010), that capture libraries from low-quality samples produce genotype data that are highly concordant with genotype data derived from high-quality, noncaptured samples from the same individuals.

We anticipate that data generated through this protocol could be leveraged for a wide variety of applications. To illustrate this point for paternity analysis, we present an accompanying method, WHODAD, that produces results in near-perfect concordance with an independently constructed pedigree, using low-coverage data generated with our enrichment protocol. By incorporating prior information about pedigree links or other demographic and behavioral data, or by sequencing very low-coverage samples to additional depth (similar to typing more markers in conventional microsatellite analysis), its performance would be improved even further. For instance, in reconstructing pedigree links in the Amboseli population, we generally include only plausible candidates (*e.g.*, we exclude males who were immature or not yet born at the offspring's conception), not all males with genotype data, as we did here.

Together, these results provide valuable, accessible wet laboratory and computational tools for moving studies of difficult-to-sample natural populations forward into the genomics era. Importantly, our methods can be generalized to produce low-complexity DNA-depleted RNA baits for any species in which at least one high-quality DNA sample is available [or potentially a closely related species (Enk *et al.* 2014)]. Further, our results show that WHODAD is highly accurate for pedigree reconstruction even when the reference genome is not a high-quality chromosomal assembly (here, we used 33,120 contigs from *Pcyn1.0*) or, based on exploratory analyses, even from the same species. Specifically, when mapping to the reference genome for the rhesus macaque [*MacaM* (Zimin *et al.* 2014)] instead of baboon, which diverged from baboons 6–8 MYA (Steiper and Young 2006), WHODAD produced similarly accurate paternity assignments (21 of 27 fathers were correctly assigned using our recommended statistical thresholds compared to 22 of 27 when mapping to *Pcyn1.0*; there were no false positive assignments in either case).

Costs of performing the protocol

At the time of publication, using the same reagents as we used here and sourced from the same locations, the costs of generating these data are ~$60 per sample (excluding sequencing costs). Because our method does not require the commercial synthesis of targeted capture probes, the majority of the costs are accounted for by the streptavidin-coated Dynalbeads ($11 per preparation), RNA baits ($5 per sample) and high-sensitivity Bioanalyzer chips for quality control ($9 per sample). Replacing Ampure XP beads with homemade SPRI beads would reduce the per-sample costs considerably, as would pooling adapter-ligated fDNA samples prior to hybridization (instead of posthybridization, as reported here). For a multiplexed pool of 10 samples, we estimate that using these two strategies would result in a per-sample cost of ~$29. Indeed, we have verified that multiplexing samples prior to hybridization does not result in loss of capture efficiency and actually resulted in improved yield of mapped, non-PCR duplicate reads (~61% of reads; mean of 117-fold

enrichment, range = 54.8–257.2-fold; Figure S11A), although it did result in more uneven coverage of samples sequenced within a pool (Figure S11B) and raises the possibility of barcode swapping [which can be managed using dual barcoding approaches (Kircher *et al.* 2012)]. Multiplexing also has the advantage of reducing the amounts of input DNA per sample and the number of PCR cycles required for the initial library preparation step. We are currently pursuing improvements to the protocol along these lines.

Based on achieving 40% non-PCR duplicate, mapped reads after capture (the mean result for capture 2 samples), we estimate that the sequencing costs of a 1× genome for baboon (~2.9 Gb) would be ~$200 (based on paired-end, 125-bp sequencing at $2000 per lane and exclusion of PCR duplicates). This cost per sample is approximately twice the cost of genotyping 14 microsatellites from the same fDNA sample—the previous strategy for the main study population, the Amboseli baboons (Van Horn *et al.* 2008)—but provides substantially more genetic information. These estimates will drop farther as the cost of high-throughput sequencing continues to fall, making application of our approach to whole populations increasingly feasible. Our finding that useful sequencing reads do not asymptote with deeper sequencing (Figure 3) also suggests the feasibility of producing a high-quality, high-coverage genome from such samples if one were to sequence more deeply. This approach would alleviate cases in which both alleles at a truly heterozygous site were not observed due to low sequencing depth (for example, with 1× coverage, only one of the two alleles can possibly be observed). Notably, however, it would not fix "allelic dropout" problems in which an allele was not represented in the pool of sequenceable fragments (Pompanon *et al.* 2005). Analogous to the solution in noninvasive microsatellite typing, multiple, independent PCR reactions could be used to solve this problem.

Finally, to make the current protocol as cost effective as possible, we recommend that researchers use qPCR to choose DNA samples with the highest proportion of host DNA possible—the strongest predictor of the fold-change enrichment in endogenous DNA postcapture *vs.* precapture (Figure 2D).

Assigning paternity using WHODAD

The lack of available tools for working with low-coverage genomic data—realistically, one of the most likely data types to be produced for studies of natural populations—represents a major barrier to moving from low-throughput marker genotyping to genome-scale analyses. The pedigree structure of a study population is fundamental to understanding its genetic structure and social organization. However, current methods for pedigree reconstruction are unable to cope with high levels of genotype uncertainty. The approach we have implemented in WHODAD takes this uncertainty into account, suggesting one simple application for the wet laboratory methods presented here. Indeed, our method performed well when compared to an independently constructed extended pedigree, with its major challenges—differentiating between close relatives in a candidate pool—comparable to those reported for existing software (Marshall *et al.* 1998; Olsen *et al.* 2001; Kalinowski *et al.* 2007; Ford and Williamson 2010). Importantly, while analyses of pedigree structure using previously available methods are greatly aided by prior knowledge of mother–offspring relationships (Kalinowski *et al.* 2007), maternal links do not appear to be necessary for WHODAD analyses, which perform well even when no maternal information is available (Figure 5, Figure S9).

Conclusions

High-throughput sequencing approaches solve one problem of working with low-quality, noninvasive samples: the sheared nature of the original samples. Capture approaches have demonstrated great promise for solving the second major problem—large proportions of nonendogenous DNA—since the results published by Perry *et al.* (2010). Our results help to fulfill this promise by providing methods to perform cost-effective sequence capture from noninvasive samples on a genome-wide scale, coupled with analytical methods to deal with the resulting data (we note that our protocols could also be explored for broader application to aDNA samples). For questions in which investigators are specifically interested in variants in *a priori*-defined subsets of the genome [*e.g.*, the exome (Vallender 2011; George *et al.* 2011)], targeted capture with synthesized baits may still be the best option. However, for the many types of analyses that use genome-scale data [*e.g.*, local ancestry analysis, genome-wide scans for selection, and reconstruction of population demographic history (Sabeti *et al.* 2002; Huang *et al.* 2007; Li *et al.* 2007; Sankararaman and Sridhar 2008; Price *et al.* 2009; Durand *et al.* 2011; Li and Durbin 2011; Yang *et al.* 2011; Ma *et al.* 2014)], our approach will be more useful, especially as the costs of high-throughput sequencing continue to fall.

Here, we focused specifically on DNA obtained from fecal samples, which are one of the most commonly collected types of noninvasive samples: they contain information not only about host genetics, but also about endocrinological parameters (Palme 2005), gut microbiota (Ley *et al.* 2008), parasite burdens (Gillespie 2006), and gene expression levels (Knight *et al.* 2014). The sample banks already available for many natural populations thus open the door to population and evolutionary genomic studies in species in which such analyses were previously impossible. As the costs of data generation continue to fall, and the limiting factor for many studies becomes high-quality phenotypic data, we envision that such studies will rapidly move far beyond the simple analyses of paternity and pedigree structure reported here.

Methods

Bait generation

Similar to Carpenter *et al.* (2013), we use a cost-effective *in vitro* synthesis method based on T7 RNA polymerase amplification of sheared DNA from a high-quality sample (Figure S1A). We extracted genomic DNA from a blood sample

collected from an olive baboon (*P. anubis*) that was unrelated to any of the individuals in the samples we wished to enrich. To generate baits, we sheared 5 μg of purified DNA to a mean fragment size of 150 bp and then end repaired and A-tailed the fragments, using the KAPA DNA Library Preparation Kit for Illumina Sequencing. We purified the resulting reaction, using a 1.8× ratio of AMPure beads to sample volume.

We annealed custom adapters to the A-tailed library by incubating the following reagents for 15 min at 20°: 10 μl 5× ligation buffer (KAPA Biosystems), 5 μl DNA Ligase (KAPA Biosystems), 1 μl 25 μM custom adapter, ≤34 μl of A-tailed DNA, and H_2O up to 50 μl total volume. The custom adapters we used (EcoOT7dTV, Fwd 5′-GGAAGGAAGGAAGAGATAATACGACTCACTATAGGGCCTGGT; EcoOT7dTV, Rev 5′-/5Phos/CCAGGCCCTATAGTGAGTCGTATTATCTCTTCCTTCCTTCC) differ from those used in other protocols (Carpenter *et al.* 2013; Enk *et al.* 2014; Ávila-Arcos *et al.* 2015). Specifically, they contained (1) a T7 RNA polymerase recognition site, (2) flanking sequence that improves T7 transcription efficiency (Moll *et al.* 2004), and (3) an EcoO109I restriction enzyme cut site that allowed us to later cleave off the adapter sequence from T7 amplified RNAs (rather than blocking it, as in Carpenter *et al.* 2013).

We then digested the purified, adapter-ligated DNA with DSN (Axxora). DSN is a Kamchatka crab-derived enzyme that specifically degrades double-stranded DNA but not single-stranded DNA, allowing us to take advantage of DNA reassociation kinetics to reduce the representation of repetitive regions in the bait set (Figure S2; Shagina *et al.* 2010). We performed DSN digestion in 15 2-μl aliquots, each mixed with 1 μl 4× hybridization buffer [200 mM HEPES (pH 7.5), 2 M NaCl, 0.8 mM EDTA] and 1 μl human Cot-1 DNA (1 μg/μl). We denatured the DNA by heating to 98° for 3 min; held the reaction at 68° for 4 hr; and then added 4 μl H_2O, 1 ml 10× DSN buffer, and 1 μl DSN (1 unit/μl) to the reaction. After 20 min of digestion, we stopped the reaction by adding 5 μl 2× DSN Stop Solution (10 mM EDTA) and purified it with 2.4× AMPure beads.

Next, we used Klenow DNA polymerase to blunt end the nondigested DNA, size selected for 200- to 300-bp fragments on a 2% agarose gel, and purified the size-selected fraction using the Zymoclean Gel DNA Recovery Kit (Zymo Research). After purification the aliquots were PCR amplified for 16 cycles, using 25 μl 2× HiFi Hot Start ReadyMix (KAPA Biosystems) and 1 μl each of 25 μM primers EcoOT7_PCR1 (5′-GGAAGGAAGGAAGAGATAATACGACTCACT) and EcoOT7_PCR2 (5′-TACGACTCACTATAGGGCCTGGT). Following amplification the bait DNA libraries were purified using 1.8× AMPure beads and the resulting product was visualized on a Bioanalyzer DNA 1000 chip (Agilent Technologies).

Finally, we *in vitro* transcribed the DNA libraries to construct biotin-tagged RNA baits, using the MEGA Shortscript Kit (Life Technologies) and Biotin-UTP (Illumina). Briefly, 125–150 nM of DNA baits were incubated at 37° for 4 hr in the following reaction: 2 μl T7 10× reaction buffer; 2 μl each of T7 ATP, GTP, CTP, and UTP solutions (75 mM); 1 μl Biotin-UTP (50 mM); 2 μl T7 enzyme mix; and water to 20 μl total volume. We then digested the DNA template by adding 1 μl TURBO DNase (Life Technologies) to the reaction and incubating it at 37° for 15 min. We purified the resulting reaction with the MEGAClear Transcription Clean-Up Kit (Life Technologies) and eluted it in a final volume of 70 μl. To cleave off the adapter sequence, we digested the RNA baits with the *Eco*O1091 enzyme (NEB). Finally, the baits were again purified with the MEGAclear Clean-Up Kit, eluted in 70 μl, and quantified on a Bioanalyzer RNA 6000, Eukaryote Total RNA chip (Agilent Technologies).

Samples, DNA extraction, and qPCR quantification

Baboon samples from Amboseli (the main study population) or West Africa (8 unhabituated Guinea baboons) were collected, stored, and extracted as detailed in Table S2. For LIT and HAP, gDNA was extracted from blood samples, using the QIAGEN (Valencia, CA) Maxi Kit. The majority of the sampled Amboseli individuals (48 of 54) were either members of a single extended pedigree or unrelated males living in the same study population (Figure 1). We assessed the proportion of endogenous DNA in each fDNA sample, using qPCR against the *c-myc* gene, as described in Morin *et al.* (2001).

Library preparation

All samples were fragmented to the desired size (200 or 400 bp; see Table S1), using a Bioruptor instrument (Diagenode). Illumina sequencing libraries were then generated from the fragmented DNA, using either the KAPA DNA library kits for Illumina (capture 1) or the NEBNext DNA Ultra library kit (capture 2; see Table S1). Libraries were amplified for 6 PCR cycles prior to capture-based enrichment. Sample-specific details of library preparation and sequencing results are described in Table S1. Note that we changed several steps between capture 1 and capture 2 based on interim improvements in the protocol (also detailed in Table S1). Because the methods used in capture 2 were ultimately more effective, the updated capture 2 protocol is described in the *Methods* section except where explicitly noted.

Capture-based enrichment

We modified the capture methods from Gnirke *et al.* (2009) and Perry *et al.* (2010) (Figure S1B). For each capture, we hybridized 121–626 ng of the fDNA libraries generated as described above to the RNA baits. First, we mixed each fDNA library with 2.5 μl human Cot-1 DNA (1 mg/ml), 2.5 μl salmon sperm DNA (1 mg/ml), and 0.6 μl index-blocking reagent ("IBR") (50 μM). This mixture was incubated for 5 min at 95° followed by 12 min at 65°. Next, we added 13 μl of hybridization buffer (10× SSPE, 10× Denhardt's solution, 10 mM EDTA, 0.2% SDS, preheated to 65°), 7 μl hybridization bait mixture (1 μl SUPERase-In, 750 ng RNA baits, and water up to a total volume of 7 μl, preheated to 65°) to the fDNA mixture and incubated the complete mixture at 65° for 48 hr (see Figure S12 for comparison of alternative bait concentrations and incubation times).

After incubation, we purified the enriched fDNA sample, using 50 μl Dynal MyOne Streptavidin T1 beads (Invitrogen, Carlsbad, CA). To do so, the beads were washed a total of three times with 200 μl binding buffer [1 M NaCl, 10 mM Tris-HCl (pH 7.5), 1 mM EDTA] and resuspended in 200 μl binding buffer. Next, the entire fDNA/RNA hybridization mix was added to the 200-μl Dynal MyOne Streptavidin T1 bead and binding buffer slurry. We incubated this mixture at room temperature for 30 min on an Eppendorf Thermomixer at 700 rpm. The mixture was placed on a magnetic rack, the supernatant was discarded, and the beads were washed once with 500 μl low-stringency wash buffer (1× SSC, 0.1% SDS) followed by a 15-min incubation at room temperature. The beads were then washed three times with 500 μl high-stringency wash buffer (0.1× SSC, 0.1% SDS) with a 10-min room temperature incubation between each wash. After the final wash, the enriched fDNA fraction was eluted from the beads with 50 μl elution buffer (0.1 M NaOH), transferred to a new tube containing 70 μl "neutralization buffer" (1 M Tris-HCl, pH 7.5), purified with 1.8× AMPure beads, and eluted in a 30-μl volume. A final PCR was carried out in a 50-μl reaction volume, using 23 μl of the posthybridization fDNA and either (1) 25 μl 2× KAPA High Fidelity master mix and 2 μl TruSeq universal primer (capture 1) or (2) 25 μl 2× NEBNext High Fidelity PCR master mix, 1 μl universal PCR primer, and 1 μl NEB indexing primer (capture 2). After 12 PCR cycles the final reaction was purified with 1× AMPure beads, eluted in 20 μL H_2O, and visualized on a Bioanalyzer High Sensitivity DNA chip.

Sequencing and alignment

All high-throughput sequence generation was conducted on the Illumina HiSeq platform (see Table S1 for sequencing details). The resulting sequencing reads were mapped to a *de novo* assembly of the *P. cynocephalus* genome (Wall *et al.* 2016) (alignment available at https://abrp-genomics.biology.duke.edu/index.php?title=Other-downloads/Pcyn1.0), using the default settings of the *bwa mem* alignment algorithm v0.7.4-r385 (Li 2013). Reads that mapped to scaffold 10204 of Pcyn1 were assigned to mitochondrial DNA due to scaffold 10204's similarity (97% sequence similarity) to a published *P. anubis* mitochondrial genome (NCBI GenBank accession no. KC757406.1). Duplicate reads were marked and discarded in subsequent analyses, using the "MarkDuplicates" function in PicardTools (http://picard.sourceforge.net). To facilitate comparison across samples of differing coverage, and because coverage of the gDNA samples was much higher (∼30×) than for the fDNA samples for LIT and HAP (1.4 and 0.27, respectively), we downsampled the gDNA libraries to 0.73× coverage (the median coverage of samples in capture 2), using "DownsampleSam" in PicardTools.

Comparison of sequencing data sets

In several analyses, we compared our capture-based enrichment results to two independent data sets: (i) a previously published capture-based enrichment of aDNA samples [NCBI SRA accession no. SRP042225 (Carpenter *et al.* 2013)] and (ii) shotgun sequencing from six capture 1 fDNA samples prior to hybridization ("precapture"; Table S1). The aDNA samples were aligned to the human genome (*hg38*) and the precapture fDNA samples were mapped to the *de novo Pcyn1.0* genome assembly.

Library complexity, distribution of reads, and GC content

We calculated the complexity of each library, using two methods. First, we used the ENCODE Project's PCR bottleneck coefficient (PBC), which calculates the percentage of nonduplicate mapped reads from the total number of mapped reads (Kharchenko *et al.* 2008; Landt *et al.* 2012).The PBC ranges from 0 to 1, where more complex libraries have higher numbers. Second, we used the function "c_curve" from the program *preseq* (v1.0.2) to plot the number of nonduplicate fragments mapped *vs.* the number of total mapped fragments (Daley and Smith 2013). More complex libraries (*i.e.*, those with fewer duplicate fragments) have a c_curve slope closer to 1, meaning that increasing sequencing depth continues to provide novel information. Less complex libraries have a shallower slope and asymptote at smaller values. Finally, we evaluated the GC bias for each sequencing library, using Picard Tools' "CollectGCBiasMetrics" (http://picard.sourceforge.net).

Sample attributes influencing capture efficiency

To determine the sample attributes that predicted the success of our capture protocol, we first modeled the relationship between the proportion of nonduplicate reads that mapped to the baboon genome after capture (our primary measure of protocol success) and (i) the percentage of endogenous baboon DNA in the precapture samples, (ii) the amount of fDNA library (nanograms) that went into the capture, and (iii) whether the sample was captured using our initial protocol or the second version of the protocol (*i.e.*, in capture 1 or capture 2). Second, we investigated the relationship between the same three variables and a secondary measure of protocol success, the fold-change enrichment of baboon DNA in the sample precapture *vs.* postcapture. Precapture concentrations of endogenous DNA in fDNA samples were measured as the concentration of baboon DNA estimated using qPCR, relative to the concentration of total DNA estimated using the Qubit High Sensitivity fluorometer (Life Technologies). To ensure that our qPCR-based measures were well calibrated, we confirmed the relationship between qPCR-based estimates and precapture sequence-based estimates of endogenous DNA in six samples for which both values were available ($R^2 = 0.92$; Figure S13). All statistical analyses were carried out in R (R Development Core Team 2015).

Variant calling

We used two different approaches to call variants and genotypes in our sample: SAMTOOLS (Li *et al.* 2009; Li 2011) and the Genome Analysis Toolkit (GATK) (McKenna *et al.* 2010; DePristo *et al.* 2011; Van der Auwera *et al.* 2013). In

downstream analyses, we retained only variants that were identified by both methods, a strategy that produces a higher ratio of true positives to false positives than variants identified by a single method alone (O'Rawe *et al.* 2013). Duplicate-marked alignments were used as input for both methods. SAMTOOLS multisample variant calling was carried out using *mpileup* and *bcftools*, with a maximum allowed read depth (-D) of 100. GATK variant calling was carried out using HaplotypeCaller following the GATK v3.0 Best Practices for variant calling from DNA-seq. To minimize potential batch effects introduced by the two capture efforts, we used the following strategy. First, we called genotypes using reads from each capture independently. Second, we recalled genotypes using reads from both captures together. Third, we extracted the union set of variants called in steps 1 and 2 for downstream analysis.

Because no reference set of genetic variants is currently publicly available for baboons, we used a bootstrapping procedure for base quality score recalibration. Briefly, we performed an initial round of variant calling on read alignments without quality score recalibration. From this variant call set, we extracted a set of high-confidence variants that passed the following hard filters: quality score $\geq$100; QD $<$ 2.0; MQ $<$ 35.0; FS $>$ 60.0; HaplotypeScore $>$13.0; MQRankSum $<$ $-$12.5; and ReadPosRankSum $<$ $-$8.0 (as described in Tung *et al.* 2015). We then recalibrated the base quality scores for each alignment, using this high-confidence set as the database of "known variants," and repeated the same variant-calling and filtering procedure for three additional rounds. Finally, we identified the intersection set between the variants called from GATK and SAMTOOLS, respectively, using the *bcftools* function *isec* (Li *et al.* 2009). To produce our final call set, we removed all sites that were genotyped in only one of the capture efforts, had a minor allele frequency of $<$0.05, or were within 10 kb of one another, using *vcftools* (Danecek *et al.* 2011). For comparisons between the paired fDNA and gDNA samples, we used the above variant-calling pipeline to jointly genotype all samples sequenced in the study.

Estimating the coefficient of relatedness

To produce an estimate of relatedness between samples in our pedigree and to test for concordance between fecal and blood-derived samples for the same individuals, we used the program *lcMLkin* (Lipatov *et al.* 2015). *lcMLkin* uses the genotype likelihoods generated by GATK for each genotype call to calculate two measures: (i) $k0$, the probability that two individuals share no alleles that are identical by descent, and (ii) r, the coefficient of relatedness (Lipatov *et al.* 2015) (*i.e.*, twice the kinship coefficient). Several other methods have been developed (Manichaikul *et al.* 2010; Yang *et al.* 2010) to estimate relatedness from thousands of SNPs, but *lcMLkin* yielded the best match to pedigree-based estimates in our data set (Figure S14).

We also compared genotype calls for the matched fecal and blood-derived samples, using GATK's GenotypeConcordance function (DePristo *et al.* 2011). This tool allowed us to determine concordance rates between data sets for different classes of variants (*e.g.*, 0, 1, or 2).

WHODAD: paternity inference and pedigree reconstruction

Our paternity prediction model is based on a naive Bayes classifier that takes advantage of the rules of Mendelian segregation within pedigrees. Using data from all sites genotyped in a potential father–mother–offspring trio or, when the mother is not genotyped, all sites genotyped in a potential father–offspring dyad, it estimates the posterior probability that a potential candidate is the true father of a given offspring.

Our approach can be broken into three steps (Figure S9). First, we estimate, for each candidate male, the conditional probability that he is the true father of a given offspring, given the genotype data for the candidate, offspring, and mother, if known (below we show the case in which genotype information is available for the mother, but the model is similar when maternal genotype information is missing). Second, we assign a *P*-value for the top candidate male from the first step, for the null hypothesis that he is *not* more related to the focal offspring than the other candidates tested. Third, we calculate the probability that the genotype data for the top candidate and offspring are consistent with a true parent–offspring relationship, using a mixture model. Steps 2 and 3 perform subtly different functions in our analysis: step 2 tests that the top candidate is significantly more related to the offspring than any other candidate, whereas step 3 tests that the dyadic similarity between the candidate and the offspring looks as expected for parent–offspring dyads. We have found that combining both approaches is key to detecting true positive fathers while minimizing false positive calls that can occur when true fathers are not in the pool of genotyped candidates (Figure S10).

Step 1: estimating conditional probabilities for each trio: For a given offspring or mother–offspring dyad, our goal is to infer the true genetic father from a pool of n candidates. For the ith candidate, we use data for the L_i variants for which we have genotype information for the known mother–offspring dyad and for the candidate father. Assuming the true father is present in the candidate pool (*i.e.*, he has been genotyped), the probability that the ith potential candidate is the father is

$$P(F_i|M,O) = P(F_i,M,O) \Big/ \left(\sum_{k=1}^{n} P(F_k,M,O) \right), \qquad (1a)$$

where $P(F_i|M,O)$ denotes the probability that the candidate is the father, conditional on the (known) mother–offspring dyad; $P(F_i,M,O)$ denotes the joint probability of the whole trio; and $\sum_{k=1}^{n} P(F_k,M,O)$ is the sum of the joint probabilities of all possible trios evaluated in the analysis. In practice, we normalize these conditional probabilities to take into account differences in the number of variants evaluated for each trio by taking the L_i^{th} root:

$$P(F_i|M,O) \approx P(F_i,M,O)^{1/L_i} \Big/ \left(\sum_{k=1}^{n} P(F_k,M,O)^{1/L_k}\right). \tag{1b}$$

Each joint probability can be calculated in turn as

$$\begin{aligned} P(F_i,M,O) &= \sum_{f,m,o} P(F_i,M,O,f,m,o) \\ &= \sum_{f,m,o} \prod_{j=1}^{L_i} P(F_i,M,O,f_{ij},m_j,o_j), \end{aligned} \tag{2}$$

where m_j, f_{ij}, and o_j represent the genotype data for the jth variant of the mother, the candidate father, and the offspring, respectively. Genotypes take values in {0, 1, 2} (*i.e.*, the number of copies of the reference allele at each individual–site combination). Importantly, although Equation 2 unrealistically assumes independence across loci, this assumption does not change the relative order of trio joint probabilities.

The probability $P(F_i,M,O,f_{ij},m_j,o_j)$ for each locus can be further decomposed as

$$P(F_i,M,O,f_{ij},m_j,o_j) \propto P(o_j|m_j,f_{ij}) \frac{P(f_{ij}|F_i)P(m_j|M)P(o_j|O)}{P(o_j)}, \tag{3}$$

where we take genotype uncertainty into account by using GATK's genotype probabilities to calculate the conditional genotype probabilities for $P(f_{ij}|F_i)$, $P(m_j|M)$, and $P(o_j|O)$ over all possible genotype values at each site–individual combination (*i.e.*, the probabilities that each genotype is 0, 1, or 2, which sum to 1). We also ignore the scaling constant $P(F_i)P(M)P(O)$ because it cancels out in the numerator and denominator of (1). The marginal probability of the offspring's genotype, $P(o_j)$, is calculated from the minor allele frequency of the variant in the population. Finally, the conditional probability $P(o_j|m_j,f_{ij})$ is based on the rules of Mendelian transmission (*e.g.*, Marshall *et al.* 1998). Due to genotype uncertainty in low-coverage data, the values of $P(F_i|M,O)$ are small. However, the highest value is usually assigned to the most likely father (based on comparison to the pedigree; see *Results*) and we can directly assess the strength of the relative evidence for the top candidate *vs.* other candidates in step 2 by calibrating these values against permuted data.

Step 2: calculating resampling-based P-values: To compute *P*-values for each paternity assignment, candidates are ranked based on their conditional probability $P(F_i|M,O)$ of being the true father. The log ratio of conditional probabilities between the highest-probability father and the second best candidate is the test statistic

$$v = \log\left(\frac{P(F_{\text{best}}|M,O)}{P(F_{\text{second}}|M,O)}\right). \tag{4}$$

To assess significance for v, we then simulate genotype data for a set of n unrelated candidate fathers based on allele frequency information for each locus in the analysis and sequence coverage information for the real candidates, at each of the loci for which they were genotyped in the true data set. Specifically, for each locus-simulated unrelated candidate combination, f_{ij}, where i indexes a (real) candidate male and j indexes the locus, we simulate a vector of genotype probabilities for the candidate father, $(f_{ij0},f_{ij1},f_{ij2})$, which sum to 1. The number of probability vectors simulated for each candidate is based on the number and identity of the loci observed in the real data. For example, if the top candidate in the real data were evaluated based on 10,000 sites, we would simulate an unrelated male with genotype vector probabilities simulated for each of those 10,000 sites; if the second-best candidate was evaluated at 9000 sites, we would simulate an unrelated male with genotype vector probabilities simulated for each of those 9000 sites, and so on. The variant sets for different simulated candidates need not be identical and are in fact highly unlikely to be so in practice.

To simulate each vector, we draw values from a Dirichlet distribution (*i.e.*, a distribution on probability vectors that sum to one). In principle, the Dirichlet distribution for each biallelic site could be parameterized by the genotype frequencies for each of the three potential genotype values, $\text{Dir}(\pi_{j0},\pi_{j1},\pi_{j2})$, with genotype frequencies equal to the Hardy–Weinberg expected values based on the allele frequency of the reference allele [*i.e.*, p^2, $2p(1-p)$, $(1-p)^2$, with p estimated from the data]. However, the low coverage in our data introduces additional noise into this sampling problem, so we instead draw values from the following Dirichlet distribution,

$$(f_{ij0},f_{ij1},f_{ij2}) \sim \text{Dir}(\kappa c_{ij}(\pi_{j0},\pi_{j1},\pi_{j2})), \tag{5}$$

where c_{ij} is the read depth (coverage) for the site in (true) candidate father i, and κ is a concentration parameter common to all sites and candidate fathers, estimated from the real data using the method of moments. κ can be thought of as a scaling factor for the effect of coverage on variance in $(f_{ij0},f_{ij1},f_{ij2})$. To make the simulations as realistic as possible, all parameters are estimated from the real data as

$$\pi_{jl} = E(f_{ijl}), \tag{6}$$

where the expectation is based on the allele frequencies for the reference allele estimated across all individuals, for each locus j and genotype l combination, and

$$\kappa = \frac{E(f_{ijl}) - E(f_{ijl}^2)}{E(c_{ij}f_{ijl}^2) - E^2(c_{ij}f_{ijl})}, \tag{7}$$

where the expectations are based on the allele frequencies (as above) across all individuals and loci and across all three possible genotype values (0, 1, and 2) for each locus–individual combination. Our estimates for π_{ij} and κ are based on the

observed average values from the data, which approximate the expected value.

After simulating genotype data for each candidate male as if he were unrelated to the focal offspring, we can obtain a new value of v (Equation 4) from the simulated data. By repeating this procedure s times, we can compute a P-value for the hypothesis that the best candidate in the true data is no more related to the focal offspring than any other candidate in the data set. This P-value is equal to the proportion of times the simulated test statistics exceed the observed test statistic. It intuitively corresponds to the probability of seeing a gap as large as the true gap between the conditional probabilities for the best and second-best candidates, if all candidates were in fact unrelated (or equally related) to the focal offspring.

Step 3: estimating the posterior probability of paternity: WHODAD's inference method, like other paternity inference methods [*e.g.*, CERVUS (Marshall *et al.* 1998; Kalinowski *et al.* 2007)], can falsely assign paternity to a close relative if the true father is not included in the pool of potential fathers. Such false positives arise because these methods do not actually test the hypothesis that the assigned father is the true father, but rather whether the assigned father is significantly more closely related to the focal offspring than other candidates in the pool. A more direct method would be to test the probability of observing the data for a father–offspring dyad (or father–mother–offspring trio) under the *alternative* hypothesis that the assigned father is the true father. Testing the alternative hypothesis is nontrivial with low-coverage data and by itself can also yield incorrect inferences (Figure S10). However, in combination with the resampling-based P-values described above, it can improve paternity assignments.

To estimate the probability of the data given the best candidate–offspring dyad, we take advantage of the fact that dyadic measures of genotype similarity or relatedness or other estimates of identity-by-descent should differ for true parent–offspring pairs compared to all other dyads (except for full sibs). By utilizing the many dyadic values in a data set of mothers, offspring, and candidate fathers, we should therefore be able to distinguish father–offspring dyads from dyads involving other relatives or unrelated pairs. Notably, this method allows us to use dyadic values for mother–offspring pairs to maximum effect.

We use a normal mixture clustering approach and the $k0$ value from the R package *lcMLkin*, where low $k0$ values predict a low probability of sharing 0 alleles. We denote y_b as the vector of logit-transformed $k0$ measurements for the best candidate–offspring dyads for all tested father–offspring dyads; y_1 as the vector of logit ($k0$) measurements for all known mother–offspring dyads, if any are present (y_1 can be an empty vector if no mother–offspring dyads were sampled); and y_0 as the vector of logit ($k0$) measurements for all other dyads. Thus, y_0 captures the distribution of logit ($k0$) values for non-parent–offspring dyads; y_1 captures the distribution of logit ($k0$) values for known parent–offspring dyads; and y_b contains a mixture of logit ($k0$) values for both true parent–offspring dyads and non-parent–offspring dyads.

We first work only with y_0 and use a mixture model approach to assign the logit ($k0$) value for each dyad i into one of K component normal distributions (fitted using the *mixtools* function in R, with a default value of $K = 5$; note that our analyses are robust to reasonable choices of K, see File S1). Components with lower mean values for $k0$ can be thought of as capturing the distribution of logit ($k0$) values for highly related dyads (*e.g.*, half-siblings), whereas components with high mean values capture distantly related or unrelated dyads (if relatedness coefficients were used instead of $k0$, this direction would be reversed: low values would correspond to distantly related dyads instead). For y_1, all dyads are from the same relatedness category (mother–offspring), so logit ($k0$) values in y_1 can be modeled by a single distribution parameterized by a mean and a variance. Finally, for y_b, values of logit ($k0$) can be assumed to be drawn either from the distribution on y_1 or from one of the distributions (likely one with a low mean value) in the mixture model for y_0,

$$y_{bi} \sim \pi N\left(\mu, \sigma^2\right) + (1-\pi)N\left(\mu_i, \sigma_i^2\right), \tag{8}$$

where for the ith individual in y_b, μ_i and σ_i^2 are the mean and variance for one of the distributions in the mixture model of y_0; μ and σ^2 are the mean and variance for the distribution on y_1; and π is the probability that a value in y_b belongs to the parent–offspring distribution or one of the distributions fitted in the mixture model for other dyads. To infer these parameters, for each dyad in y_b, we assign μ_i, σ_i^2 to the mean and variance of the mostly likely normal component by evaluating the likelihood under all K components. We then combine y_1 and y_b to jointly infer π, μ, σ^2 in Equation 8.

Finally, we introduce a latent indicator variable z_{bi} for each dyad to indicate whether the ith dyad in y_b is a true father–offspring dyad. The probability of being a true father–offspring dyad, or $P(z_{bi} = 1)$, becomes the final statistic used to assess our paternity assignments. To infer $P(z_{bi} = 1)$, we use an expectation-maximization algorithm (see File S1 for detailed information about the EM steps). WHODAD considers a male as the likely true father of a focal offspring if he was (i) the candidate with the highest conditional probability of paternity, (ii) assigned a P-value from our simulations < 0.05, and (iii) $P(z_{bi} = 1) > 0.9$.

Testing the accuracy of paternity assignment using WHODAD

We assigned paternity using the methods detailed above for all previously identified father–offspring pairs ($n = 27$) in the Amboseli pedigree (Figure 1). This pedigree was constructed using a combination of observational life history data on female pregnancies and infant care (to infer maternal–offspring dyads), demographic data to identify possible candidate fathers, and microsatellite genotyping data analyzed in the program CERVUS (with confidence >95%; see Alberts *et al.* 2006 for additional detail).

Our data set contained maternal genotype information derived from the fecal enrichment protocol for 15 of these individuals (56%). We first used WHODAD to assign paternity for these 15 offspring while incorporating the genotype data from their mothers. To assess the accuracy of WHODAD in the absence of maternal genotype data, we then repeated the paternity analysis for the same 15 offspring without including the mother's genotype. For this analysis, we were also able to include the 12 additional offspring for whom we did not have genotype data from the mother, but had genotype data from the known father ($n = 27$).

To examine how the presence of close male kin influenced the accuracy and confidence of WHODAD's paternity assignments, we conducted three additional analyses. First, to assess the accuracy of WHODAD when the pedigree-assigned father is the only close male relative present, we removed all close relatives of the offspring except the father ($r \geq 0.25$, *e.g.*, grandfathers and half-sibling or full-sibling brothers) from the pool of potential fathers. Second, to test whether WHODAD assigned a father with high confidence even when no close relatives were present, we removed all close male relatives, including the pedigree-assigned father, from the pool of candidate males. Third, to assess the risk of confidently (but erroneously) assigning a close male relative as the likely father when the pedigree-assigned father was not genotyped, we removed the father from the pool of potential fathers. For all WHODAD analyses we report assignment accuracy based on whether the father was identified by WHODAD with a P-value <0.05 and a P ($z_{bi} = 1$) > 0.90. Offspring were not assigned a father ("no assignment") when the best candidate male was identified with a P-value >0.05 or a P ($z_{bi} = 1$) < 0.90.

Data availability

All sequencing data sets reported in this article have been deposited in the NCBI Short Read Archive (SRA), accession no. SRP064514. The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article.

Acknowledgments

We thank the Kenya Wildlife Service, the Institute of Primate Research, National Museums of Kenya, the National Council for Science and Technology, members of the Amboseli–Longido pastoralist communities, Tortilis Camp, and Ker & Downey Safaris for their assistance in Kenya. We also thank Jeanne Altmann and Elizabeth Archie for their generous support and access to the Amboseli Baboon Research Project data set and samples; Raphael Mututua, Serah Sayialel, Kinyua Warutere, Mercy Akinyi, Tim Wango, and Vivian Oudu for invaluable assistance with the Amboseli baboon sample collection; Emily McLean for assistance in identifying samples from the extended pedigree; and Tauras Vilgalys for assistance in drawing the pedigree. For access to the Guinea baboon samples, we thank Julia Fischer, Dietmar Zinner and José Carlos Brito; the Wild Chimpanzee Foundation for logistical support in Guinea; and the Ministère de l'Environnement et de la Protection de la Nature and the Direction des Parcs Nationaux in Senegal; the Opération du Parc National de la Boucle du Baoulé and the Ministère de l'Environnement et de l'Assainissement in Mali; the Office Guinéen de la Diversité Biologique et des Aires Protégées and the Ministère de L'Environnement, des Eaux et Forêts in Guinea; and the Ministère Délégue auprès du Premier Ministre, Chargé de l'Environnement et du Développement Durable in Mauritania. Finally, we thank P. J. Perry, Luis Barreiro, Greg Crawford, Tim Reddy, members of the Alberts and Tung laboratories, and two anonymous reviewers for their feedback on earlier versions of this work. This work was supported by National Science Foundation grants DEB-1405308 (to J.T.) and SMA-1306134 (to J.T. and N.S.-M.). G.H.K. was supported by the German Academic Exchange Service [Deutscher Akademischer Austauschdienst (DAAD)], the Christiane-Nüsslein-Volhard Foundation, The Leakey Foundation, and the German Primate Center. X.Z. was supported by a grant from the Foundation for the National Institutes of Health through the Accelerating Medicines Partnership BOEH15AMP. The authors declare no competing financial interests.

Author contributions: J.T., N.S.-M., S.M., and X.Z. conceived and designed the research. M.L.Y., A.O.S., J.B.G., G.H.K., and N.S.M. performed all laboratory experiments. S.A.S., J.T., and J.D.W. provided the genome assembly. W.H.M., S.M., J.T., and X.Z. developed the computational methods. W.H.M. and X.Z. implemented the software. N.S.-M., J.T., and X.Z. analyzed the data. S.C.A. and G.H.K. provided samples, reagents, and logistical support. N.S.-M., J.T., and X.Z. wrote the manuscript with input from all of the coauthors.

Literature Cited

Alberts, S. C., and J. Altmann, 2012 The Amboseli Baboon Research Project: 40 years of continuity and change, pp. 261–287 in *Long-Term Field Studies of Primates*, edited by P. M. Kappeler, and D. P. Watts. Springer-Verlag, Berlin/Heidelberg, Germany.

Alberts, S. C., J. C. Buchan, and J. Altmann, 2006 Sexual selection in wild baboons: from mating opportunities to paternity success. Anim. Behav. 72(5): 1177–1196.

Archie, E. A., J. A. Hollister-Smith, J. H. Poole, P. C. Lee, C. J. Moss *et al.*, 2007 Behavioural inbreeding avoidance in wild African elephants. Mol. Ecol. 16(19): 4138–4148.

Ávila-Arcos, M. C., M. Sandoval-Velasco, H. Schroeder, M. L. Carpenter, A.-S. Malaspinas *et al.*, 2015 Comparative performance of two whole genome capture methodologies on ancient DNA Illumina libraries. Methods Ecol. Evol. 6(6): 725–734.

Buchan, J. C., S. C. Alberts, J. B. Silk, and J. Altmann, 2003 True paternal care in a multi-male primate society. Nature 425 (6954): 179–181.

Carpenter, M. L., J. D. Buenrostro, C. Valdiosera, H. Schroeder, M. E. Allentoft *et al.*, 2013 Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. Am. J. Hum. Genet. 93(5): 852–864.

Chakraborty, R., M. Shaw, and W. J. Schull, 1974 Exclusion of paternity: the current state of the art. Am. J. Hum. Genet. 26(4): 477–488.

Charpentier, M. J. E., M. C. Fontaine, E. Cherel, J. P. Renoult, T. Jenkins *et al.*, 2012 Genetic structure in a dynamic baboon hybrid zone corroborates behavioural observations in a hybrid population. Mol. Ecol. 21(3): 715–731.

Daley, T., and A. D. A. Smith, 2013 Predicting the molecular complexity of sequencing libraries. Nat. Methods 10(4): 325–329.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27(15): 2156–2158.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43(5): 491–498.

Durand, E. Y., N. Patterson, D. Reich, and M. Slatkin, 2011 Testing for ancient admixture between closely related populations. Mol. Biol. Evol. 28(8): 2239–2252.

Enk, J. M., A. M. Devault, M. Kuch, Y. E. Murgha, J.-M. Rouillard *et al.*, 2014 Ancient whole genome enrichment using baits built from modern DNA. Mol. Biol. Evol. 31(5): 1292–1294.

Ford, M. J., and K. S. Williamson, 2010 The aunt and uncle effect revisited - the effect of biased parentage assignment on fitness estimation in a supplemented salmon population. J. Hered. 101 (1): 33–41.

Gagneux, P., C. Boesch, and D. S. Woodruff, 1997 Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. Mol. Ecol. 6(9): 861–868.

George, R. D., G. McVicker, R. Diederich, S. B. Ng, A. P. MacKenzie *et al.*, 2011 Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection. Genome Res. 21(10): 1686–1694.

Gillespie, T. R., 2006 Noninvasive assessment of gastrointestinal parasite infections in free-ranging primates. Int. J. Primatol. 27(4): 1129–1143.

Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust *et al.*, 2009 Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 27(2): 182–189.

Gottelli, D., J. Wang, S. Bashir, and S. M. Durant, 2007 Genetic analysis reveals promiscuity among female cheetahs. Proc. Biol. Sci. 274(1621): 1993–2001.

Huang, B., C. Amos, and D. Lin, 2007 Detecting haplotype effects in genomewide association studies. Genet. Epidemiol. 31: 803–812.

Idaghdour, Y., D. Broderick, and A. Korrida, 2003 Faeces as a source of DNA for molecular studies in a threatened population of great bustards. Conserv. Genet. 4(6): 789–792.

Kalinowski, S. T., M. L. Taper, and T. C. Marshall, 2007 Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol. Ecol. 16(5): 1099–1106.

Kharchenko, P. V., M. Y. Tolstorukov, and P. J. Park, 2008 Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat. Biotechnol. 26(12): 1351–1359.

Kircher, M., S. Sawyer, and M. Meyer, 2012 Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 40(1): e3.

Knight, J. M., L. A. Davidson, D. Herman, C. R. Martin, J. S. Goldsby *et al.*, 2014 Non-invasive analysis of intestinal development in preterm and term infants using RNA-sequencing. Sci. Rep. 4: 5453.

Landt, S. G., G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli *et al.*, 2012 ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 22(9): 1813–1831.

Ley, R. E., C. A. Lozupone, M. Hamady, R. Knight, and J. I. Gordon, 2008 Worlds within worlds: evolution of the vertebrate gut microbiota. Nat. Rev. Microbiol. 6(10): 776–788.

Li, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27(21): 2987–2993.

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv: 1303.3997.

Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature 475: 493–496.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16): 2078–2079.

Li, Y., W. Sung, and J. Liu, 2007 Association mapping via regularized regression analysis of single-nucleotide–polymorphism haplotypes in variable-sized sliding windows. Am. J. Hum. Genet. 80: 705–715.

Lipatov, M., K. Sanjeev, R. Patro, and K. Veeramah, 2015 Maximum likelihood estimation of biological relatedness from low coverage sequencing data. bioRxiv: 023374.

Ma, Y., J. Zhao, J.-S. Wong, L. Ma, W. Li *et al.*, 2014 Accurate inference of local phased ancestry of modern admixed populations. Sci. Rep. 4: 5800.

Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale *et al.*, 2010 Robust relationship inference in genome-wide association studies. Bioinformatics 26(22): 2867–2873.

Marshall, T. C., and J. Slate, L. E. Kruuk, and J. M. Pemberton, 1998 Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. 7(5): 639–655.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20(9): 1297–1303.

Moll, P. R., J. Duschl, and K. Richter, 2004 Optimized RNA amplification using T7-RNA-polymerase based in vitro transcription. Anal. Biochem. 334(1): 164–174.

Mondol, S., K. Ullas Karanth, N. Samba Kumar, A. M. Gopalaswamy, A. Andheria *et al.*, 2009 Evaluation of non-invasive genetic sampling methods for estimating tiger population size. Biol. Conserv. 142(10): 2350–2360.

Morin, P. A., K. E. K. Chambers, C. Boesch, and L. Vigilant, 2001 Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (Pan troglodytes verus). Mol. Ecol. 10(7): 1835–1844.

Nagata, J., V. V. Aramilev, A. Belozor, T. Sugimoto, and D. R. McCullough, 2005 Fecal genetic analysis using PCR-RFLP of cytochrome b to identify sympatric carnivores, the tiger Panthera tigris and the leopard Panthera pardus, in far eastern Russia. Conserv. Genet. 6(5): 863–866.

O'Rawe, J., T. Jiang, G. Sun, Y. Wu, W. Wang *et al.*, 2013 Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. 5(3): 28.

Olsen, J. B., C. Busack, J. Britt, and P. Bentzen, 2001 The aunt and uncle effect: an empirical evaluation of the confounding influence of full sibs of parents on pedigree reconstruction. J. Hered. 92(3): 243–247.

Palme, R., 2005 Measuring fecal steroids: guidelines for practical application. Ann. N. Y. Acad. Sci. 1046: 75–80.

Pérez, T., J. Naves, J. F. Vázquez, J. Seijas, A. Corao *et al.*, 2010 Evidence for improved connectivity between Cantabrian brown bear subpopulations. Ursus 21(1): 104–108.

Perry, G. H., J. C. Marioni, P. Melsted, and Y. Gilad, 2010 Genomic-scale capture and sequencing of endogenous DNA from feces. Mol. Ecol. 19(24): 5332–5344.

Pompanon, F., A. Bonin, E. Bellemain, and P. Taberlet, 2005 Genotyping errors: causes, consequences and solutions. Nat. Rev. Genet. 6(11): 847–859.

Price, A., A. Tandon, N. Patterson, and K. Barnes, 2009 Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 5: e1000519.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–575.

R Development Core Team, 2015 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Rudnick, J. A., T. E. Katzner, E. A. Bragin, and J. A. DeWoody, 2007 A non-invasive genetic evaluation of population size, natal philopatry, and roosting behavior of non-breeding eastern imperial eagles (Aquila heliaca) in central Asia. Conserv. Genet. 9(3): 667–676.

Sabeti, P., D. Reich, and J. Higgins, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.

Sacks, B. N., M. Moore, M. J. Statham, and H. U. Wittmer, 2011 A restricted hybrid zone between native and introduced red fox (Vulpes vulpes) populations suggests reproductive barriers and competitive exclusion. Mol. Ecol. 20(2): 326–341.

Samuels, D. C., L. Han, J. Li, S. Quanghu, T. A. Clark *et al.*, 2013 Finding the lost treasures in exome sequencing data. Trends Genet. 29(10): 593–599.

Sankararaman, S., and S. Sridhar, 2008 Estimating local ancestry in admixed populations. Am. J. Hum. Genet. 82: 290–303.

Shagina, I., E. Bogdanova, I. Z. Mamedov, Y. Lebedev, S. Lukyanov *et al.*, 2010 Normalization of genomic DNA using duplex-specific nuclease. Biotechniques 48(6): 455–459.

Sinnwell, J. P., T. M. Therneau, and D. J. Schaid, 2014 The kinship2 R package for pedigree data. Hum. Hered. 78(2): 91–93.

Smith, K., S. C. Alberts, and J. Altmann, 2003 Wild female baboons bias their social behaviour towards paternal half-sisters. Proc. Biol. Sci. 270(1514): 503.

Steiper, M. E., and N. M. Young, 2006 Primate molecular divergence dates. Mol. Phylogenet. Evol. 41(2): 384–394.

Taberlet, P., L. Waits, and G. Luikart, 1999 Noninvasive genetic sampling: look before you leap. Trends Ecol. Evol. 14(8): 323–327.

Thompson, E. A., and T. R. Meagher, 1987 Parental and sib likelihoods in genealogy reconstruction. Biometrics 43(3): 585–600.

Tung, J., X. Zhou, S. C. Alberts, M. Stephens, and Y. Gilad, 2015 The genetic architecture of gene expression levels in wild baboons. eLife 4: e04729.

Valière, N., L. Fumagalli, L. Gielly, C. Miquel, B. Lequette *et al.*, 2003 Long-distance wolf recolonization of France and Switzerland inferred from non-invasive genetic sampling over a period of 10 years. Anim. Conserv. 6(1): 83–92.

Vallender, E. J., 2011 Expanding whole exome resequencing into non-human primates. Genome Biol. 12(9): R87.

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel *et al*., 2013 From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics 43: 11.10.1–11.10.33.

Van Horn, R. C., J. Altmann, and S. C. Alberts, 2008 Can't get there from here: inferring kinship from pairwise genetic relatedness. Anim. Behav. 75(3): 1173–1180.

Visscher, P. M., 2009 Whole genome approaches to quantitative genetics. Genetica 136(2): 351–358.

Wall, J. D., S. A. Schlebusch, S. C. Alberts, and L. Cox, N. Snyder-Mackler *et al.*, 2016 Genome-wide ancestry and divergence patterns from low-coverage sequencing data reveal a complex history of admixture in wild baboons. Mol. Ecol. DOI: 10.1111/mec.13684.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42(7): 565–569.

Yang, J., T. A. Manolio, L. R. Pasquale, E. Boerwinkle, N. Caporaso *et al.*, 2011 Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. 43(6): 519–525.

Zimin, A. V., A. S. Cornish, M. D. Maudhoo, R. M. Gibbs, X. Zhang *et al.*, 2014 A new rhesus macaque assembly and annotation for next-generation sequencing analyses. Biol. Direct 9(1): 20.

Communicating editor: J. Shendure

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.187492/-/DC1

Efficient Genome-Wide Sequencing and Low-Coverage Pedigree Analysis from Noninvasively Collected Samples

Noah Snyder-Mackler, William H. Majoros, Michael L. Yuan, Amanda O. Shaver, Jacob B. Gordon, Gisela H. Kopp, Stephen A. Schlebusch, Jeffrey D. Wall, Susan C. Alberts, Sayan Mukherjee, Xiang Zhou, and Jenny Tung

DOI: 10.1534/genetics.116.187492

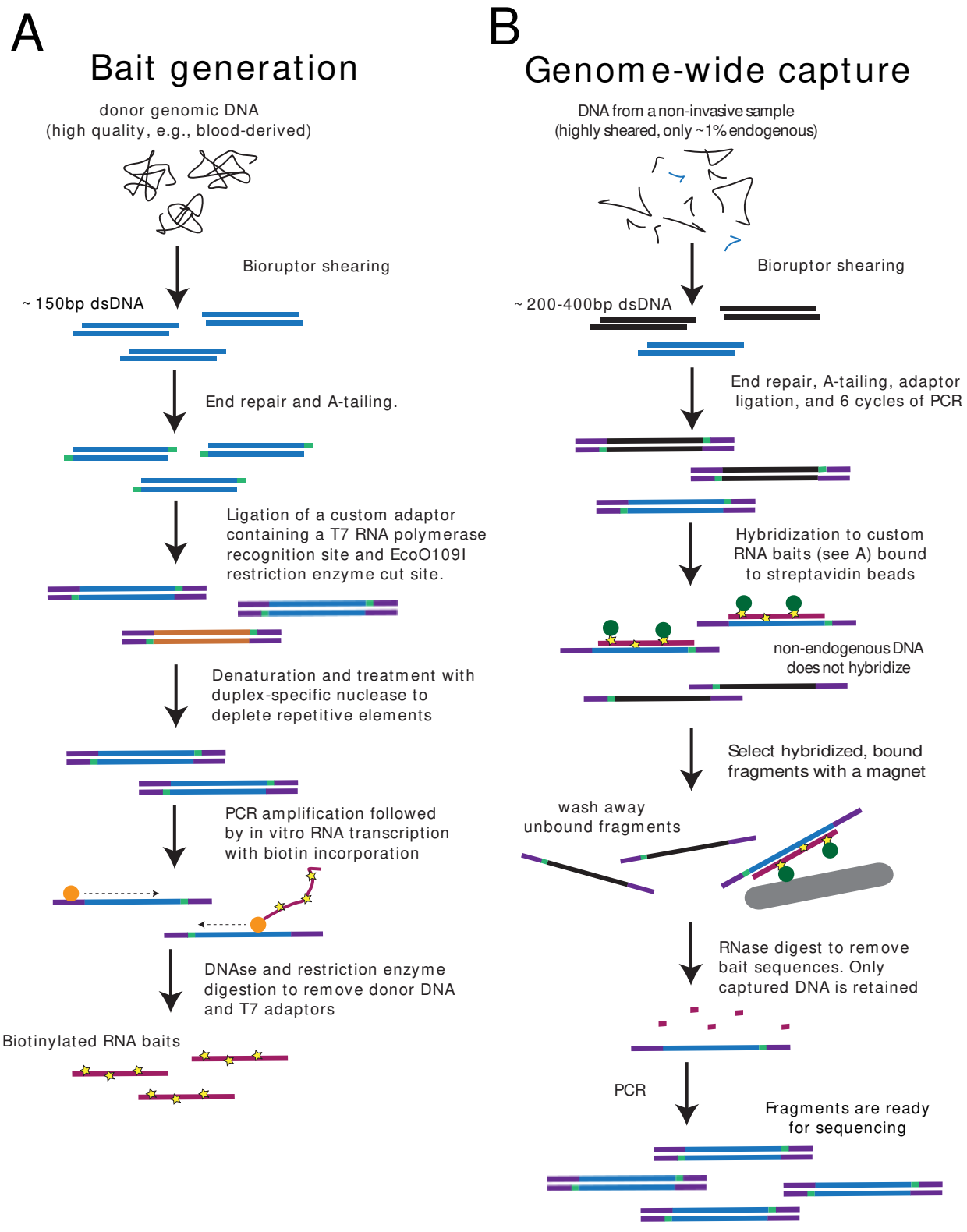

A
Bait generation
donor genomic DNA
(high quality, e.g., blood-derived)
Bioruptor shearing
~ 150bp dsDNA
End repair and A-tailing.
Ligation of a custom adaptor containing a T7 RNA polymerase recognition site and EcoO109I restriction enzyme cut site.
Denaturation and treatment with duplex-specific nuclease to deplete repetitive elements
PCR amplification followed by in vitro RNA transcription with biotin incorporation
DNAse and restriction enzyme digestion to remove donor DNA and T7 adaptors
Biotinylated RNA baits
B
Genome-wide capture
DNA from a non-invasive sample
(highly sheared, only ~1% endogenous)
Bioruptor shearing
~ 200-400bp dsDNA
End repair, A-tailing, adaptor ligation, and 6 cycles of PCR
Hybridization to custom RNA baits (see A) bound to streptavidin beads
non-endogenous DNA does not hybridize
Select hybridized, bound fragments with a magnet
wash away unbound fragments
RNase digest to remove bait sequences. Only captured DNA is retained
PCR
Fragments are ready for sequencing

Figure S1. Schematic of RNA bait generation and hybridization reaction. (A) Schematic of RNA bait generation. High-quality genomic DNA from a baboon is fragmented to 150 bp (blue fragments). Custom adaptors (purple fragments) with a T7 RNA polymerase site and a restriction enzyme cut site are then ligated to the fragmented DNA. DSN treatment is used to reduce the representation of repetitive elements (orange fragments). Finally, the library is PCR amplified, biotinylated, and transcribed into RNA baits. (B) Schematic of hybridization and capture. A genomic library is generated from fecal DNA (fDNA) sheared to 200-400 bp fragments. This fragment pool originally contains ~1% endogenous DNA (blue fragments) and ~99% environmental/microbial DNA (black fragments). Next, the fDNA library is incubated with 750 ng biotinylated RNA baits. RNA bait-bound DNA (enriched for endogenous DNA fragments) is then separated from the supernatant with a magnet. The RNA baits are digested leaving only the enriched fDNA sample, which can be PCR amplified for high-throughput sequencing.

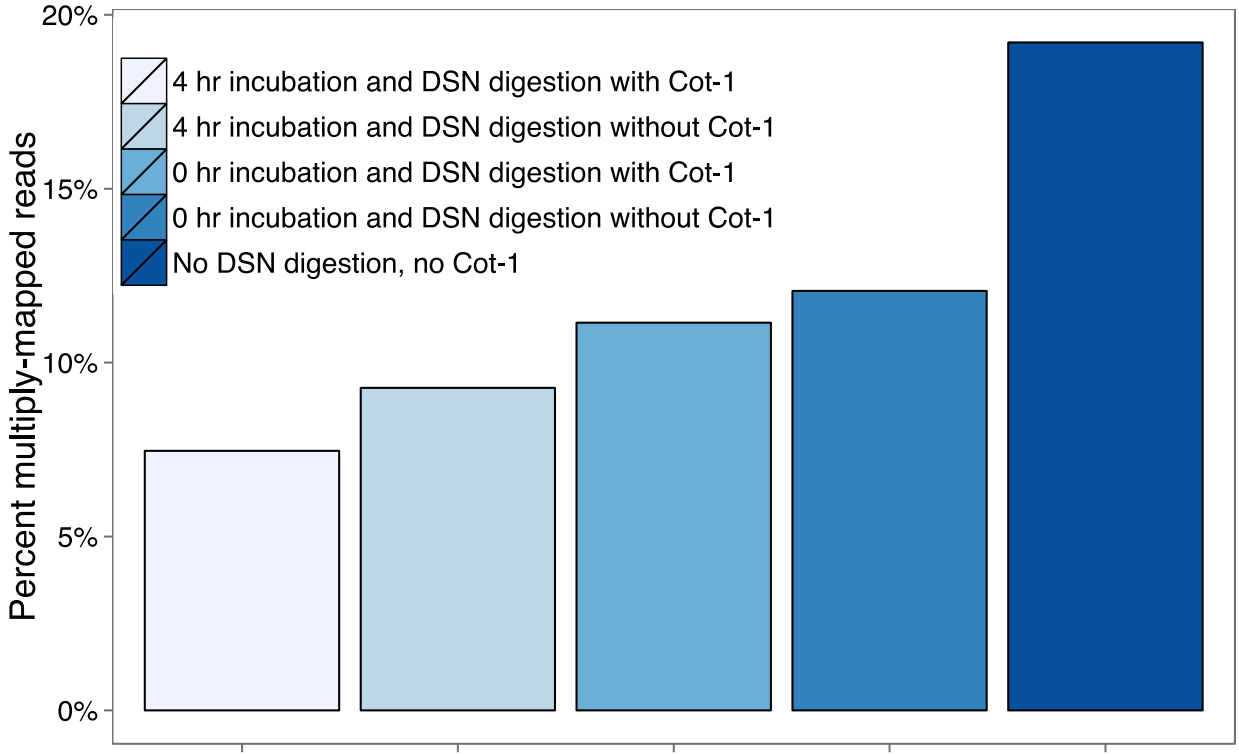


Figure S2. Depletion of multiply-mapped regions in DNA used to make RNA baits. Post DSN-treated DNA libraries (generated following the protocol used to make RNA baits, but with Illumina sequencing adapters ligated to the ends) were sequenced to assess DSN-mediated depletion of bait templates that do not uniquely map in the genome (i) with or without the presence of Cot-1, and (ii) at 0 or 4 hour incubation times. Bars show the percent of non-PCR duplicate mapped reads that mapped to multiple locations ("multiply mapped"). The four-hour incubation followed by a 20-minute DSN digestion in the presence of human Cot-1 (lightest blue bar) provided the strongest depletion of multiply mapped reads, reducing the proportion of these reads 2.6-fold (from 19.2% to 7.4%).

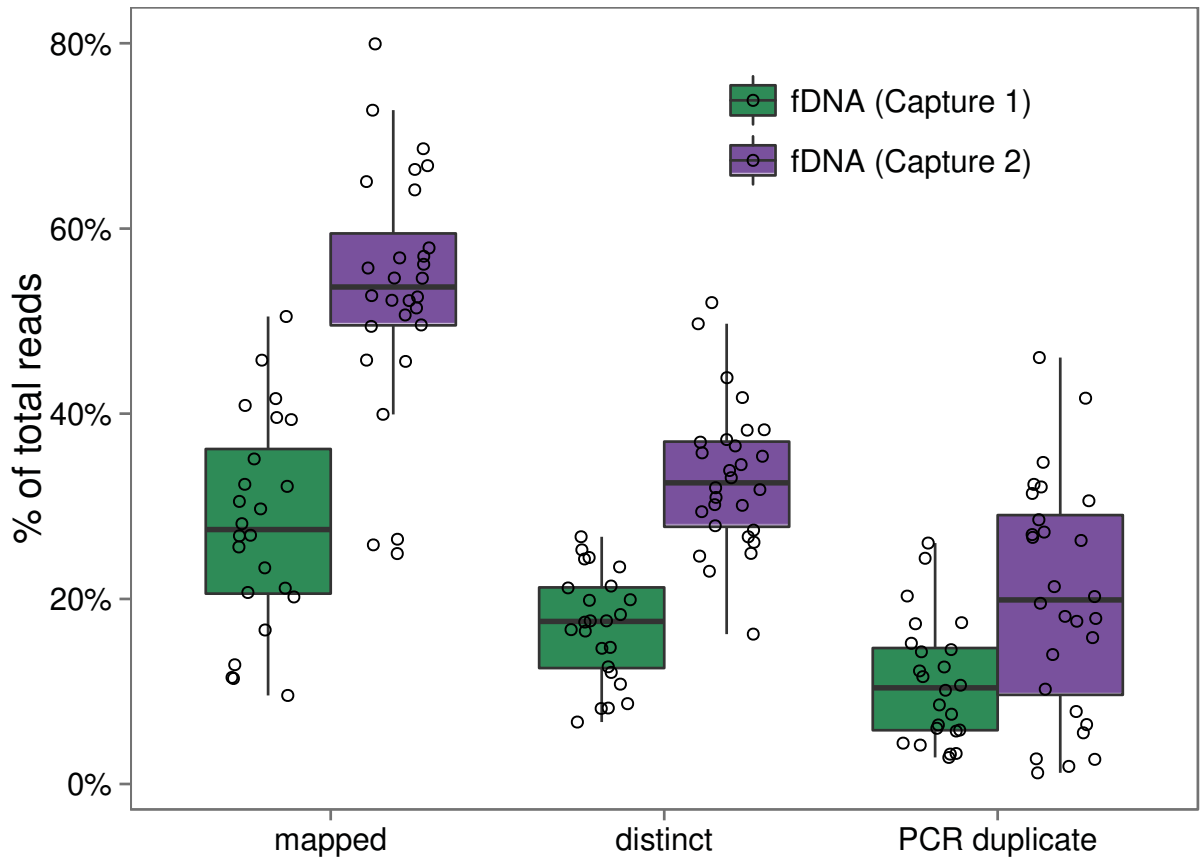


Figure S3. Comparison of mapped reads in Capture 1 and Capture 2. Capture 2 data were generated using paired-end sequencing; to compare across capture efforts, we truncated the Capture 2 reads to recapitulate the single-end, 100 bp reads generated in Capture 1. Capture 2 significantly improved over Capture 1 when considering both the percent of total reads that mapped ("mapped"; two-sample t-test, $T=7.50$, $p=1.0\times10^{-9}$) or the percent of reads that were uniquely mapped and could be distinguished from PCR duplicates ("distinct"; two-sample t-test, $T=3.41$, $p=1.4\times10^{-3}$).

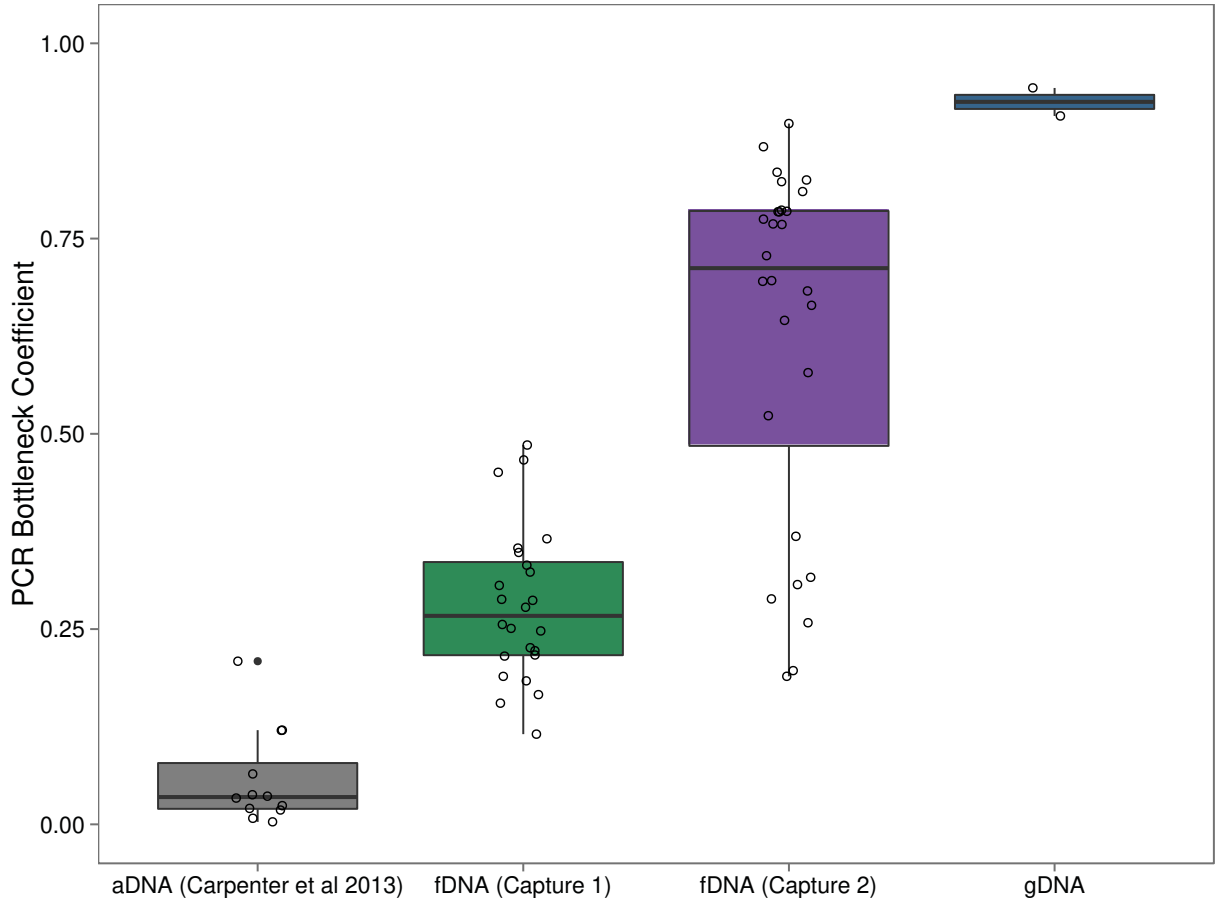


Figure S4. Library complexity for different sample preps. Library complexity measured using the ENCODE PCR Bottleneck Coefficient, which is calculated as the number of mapped, non-PCR duplicate reads divided by the number of mapped reads. Numbers closer to 1 are more complex; numbers closer to 0 are less complex.

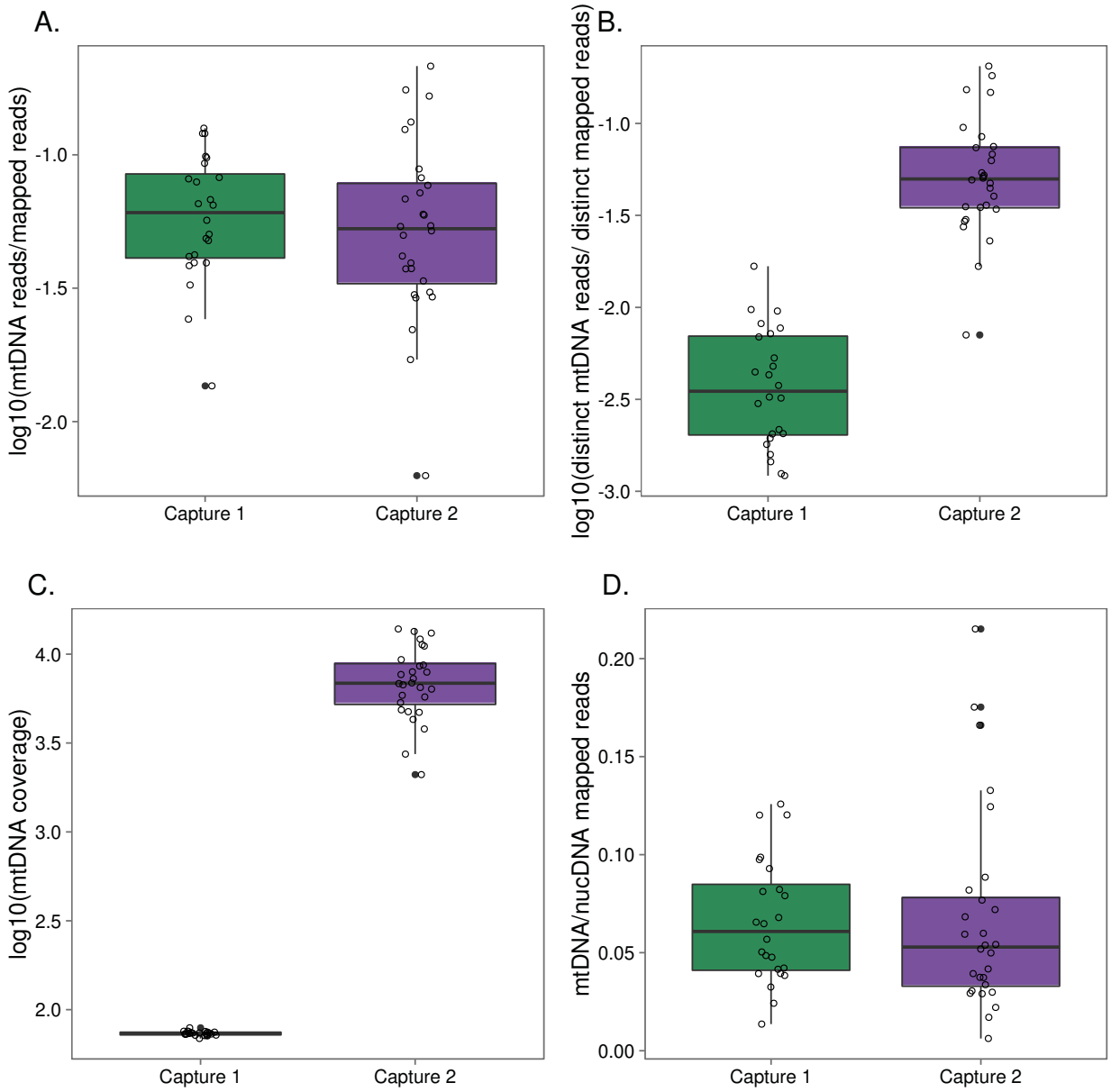


Figure S5. **Post-capture coverage of mitochondrial genome (mtDNA).** (A) Capture 2 had a similar proportion of fragments that mapped to mtDNA as Capture 1. (B) However, Capture 2 had proportionally more distinct (non-PCR duplicate) mapped reads than Capture 1. Because data from Capture 1 were generated using single-end sequencing while data from Capture 2 were generated using paired-end sequencing, many of the mtDNA mapped reads identified as PCR duplicates in Capture 1 probably represent distinct fragments that are indistinguishable from PCR duplicates. With a single end sequencing strategy, only ~16,000 reads can be identified as “distinct” reads (because mtDNA is ~16 kb in length), whereas with paired end sequencing and variable insert lengths, many more distinct reads can be identified. (C) This difference in sequencing resulted in much deeper coverage of the mtDNA genome in Capture 2 when we removed reads that could not be excluded as PCR duplicates. However, mean coverage for Capture 1 samples was still high, at ~74x. (D) There was no difference in the ratio of mtDNA to nuclear DNA (nucDNA) mapped reads between the two captures.

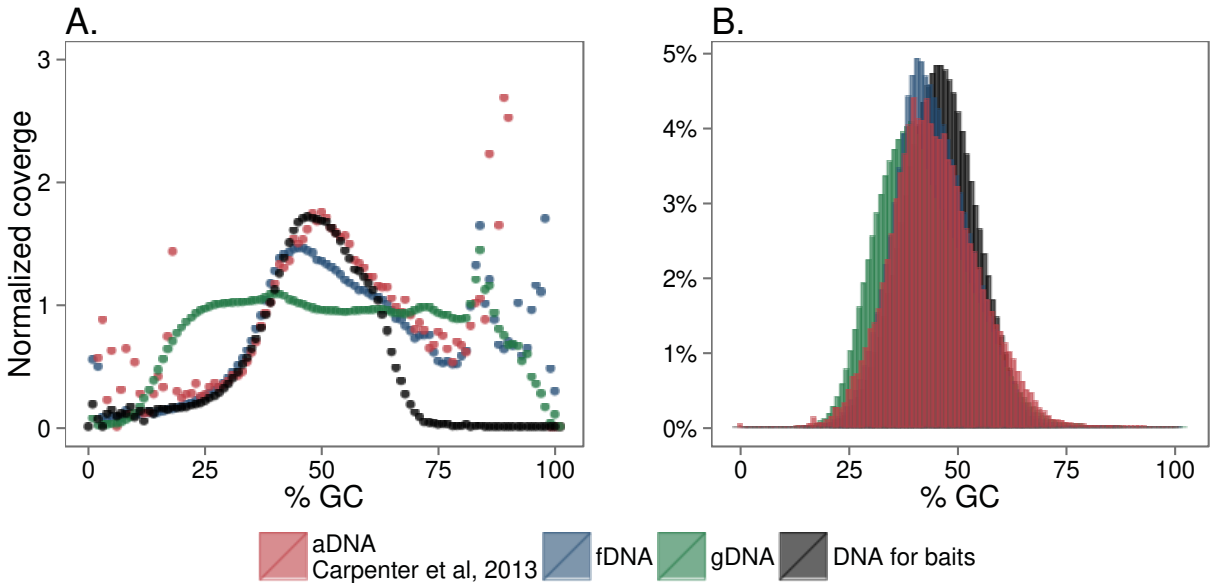
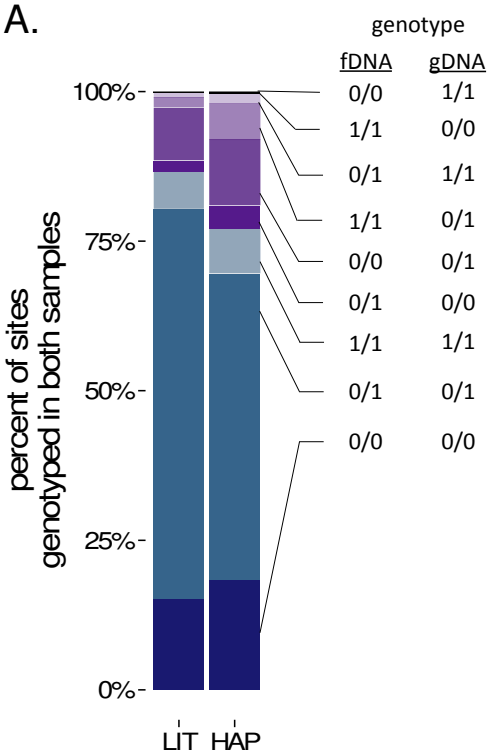


Figure S6. GC content of different library preps. fDNA capture libraries exhibit similar GC bias properties as the aDNA libraries sequenced in Carpenter et al, 2013, and the DNA fragments used to make RNA baits ("DNA for baits"), but slightly different GC properties compared to the blood-derived libraries from HAP and LIT ("gDNA"). (A) GC content vs. normalized coverage (the number of fragments per GC content window divided by the average number of fragments per window across the whole genome). (B) Distribution of GC content of fragments.

A.



B.

gDNA genotype

fDNA genotype	LIT 0/0	LIT 0/1	LIT 1/1
0/0	47559- (15.2%)	27397- (8.8%)	357- (0.1%)
0/1	5776- (1.8%)	203834- (65.2%)	3118- (1.0%)
1/1	90- (0.03%)	5277- (1.7%)	19331- (6.2%)

gDNA genotype

fDNA genotype	HAP 0/0	HAP 0/1	HAP 1/1
0/0	7403- (18.4%)	4410- (11.0%)	56- (0.1%)
0/1	1524- (3.8%)	20477- (51.0%)	681- (1.7%)
1/1	89- (0.2%)	2424- (6.0%)	3068- (7.6%)

Figure S7 Genotype concordance between paired fDNA and gDNA samples. (A) The fDNA samples captured at least one allele (blue and purple bars) from the paired gDNA samples in 99.8% (LIT) and 99.6% (HAP) of sites genotyped. Paired samples therefore had completely discordant genotypes (0/0 and 1/1 or 1/1 and 0/0; black and dark grey bars) in less than 0.5% of genotyped sites. The 9 possible combinations of genotype calls in paired fDNA (first column) and gDNA samples (2nd column) are shown to the right of the stacked barcharts; in all cases, the reference allele is represented as 0 and the alternate allele as 1. (B) Table of genotype concordance/discordance rates between sites genotyped in paired fDNA and gDNA libraries. The top table shows the results for LIT fDNA-gDNA and the bottom table for HAP fDNA-gDNA.

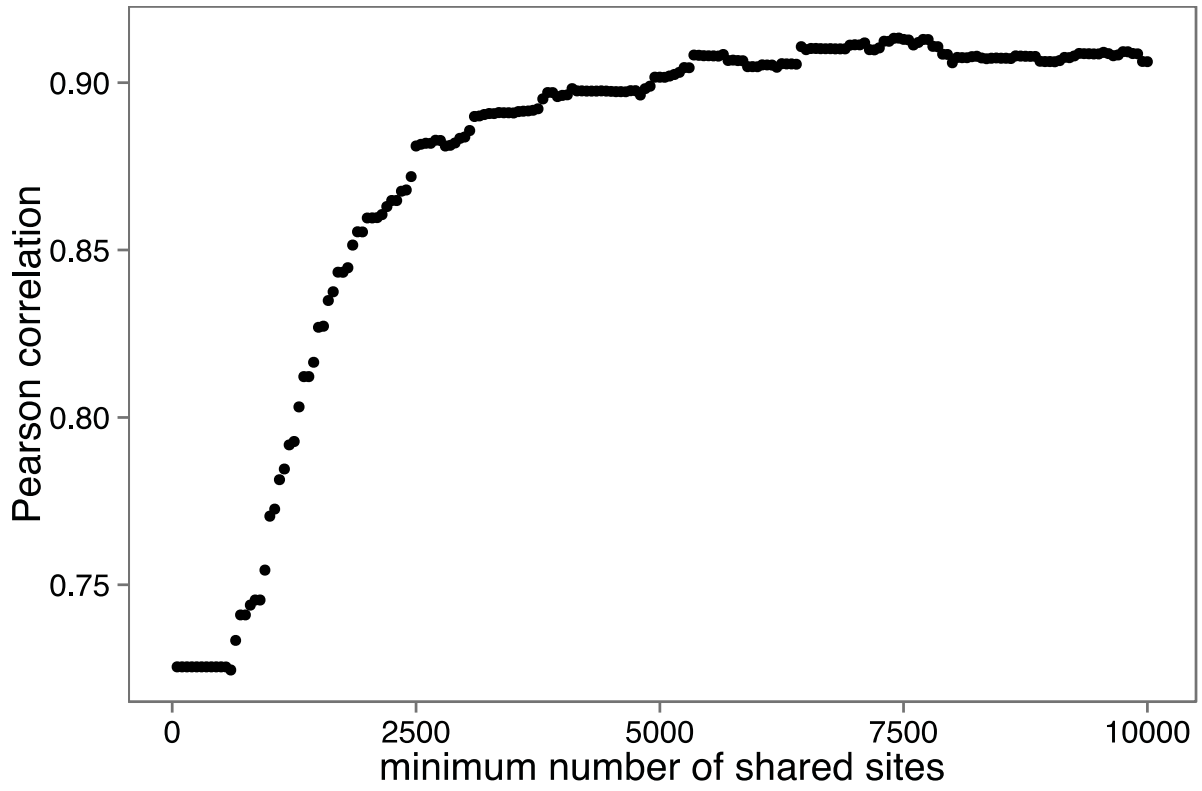


Figure S8 Correlation between independently established pedigree-based relatedness and genotype similarity, by minimum number of shared sites. The value of the correlation increased when we only included dyads with a minimum number of shared sites (x-axis). 93% of all dyads in our study had over 2,000 shared sites. The correlation coefficient asymptotes at $r \approx 0.91$.

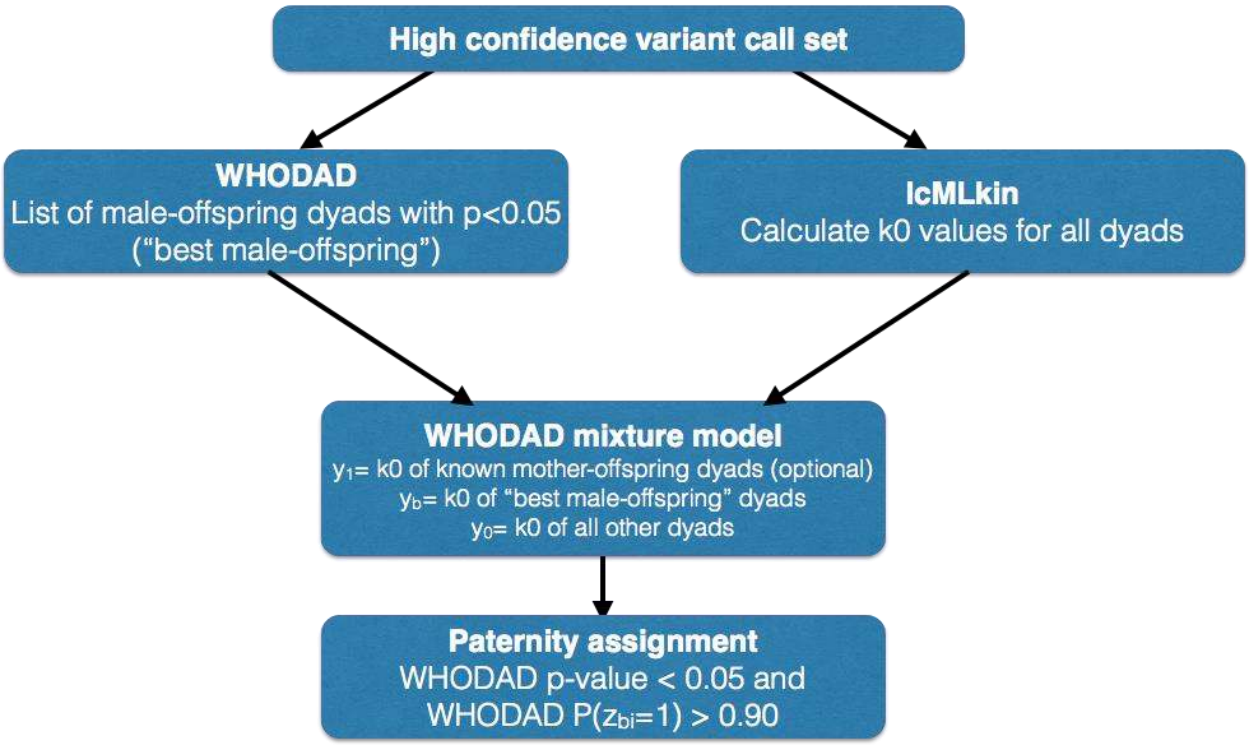


Figure S9 Flowchart for paternity inference using WHODAD. A filtered variant call set is used as input for both WHODAD and *lcMLkin* (any other relatedness estimation program can also be used). *k0* values from *lcMLkin* are then used to parameterize a mixture model that fits different normal distributions on *k0* (corresponding intuitively to different levels of relatedness). By integrating the estimates from the mixture model with the best candidate father from the genotype simulations (used to assign p-values to best candidate fathers), WHODAD then calculates $P(z_{bi}=1)$, the posterior probability that the best male-offspring dyad is drawn from the distribution of *k0* values for parent-offspring pairs. We recommend assigning paternity to candidate father-offspring dyads with a WHODAD p-value < 0.05 and $P(z_{bi}=1) > 0.90$. See methods for details.

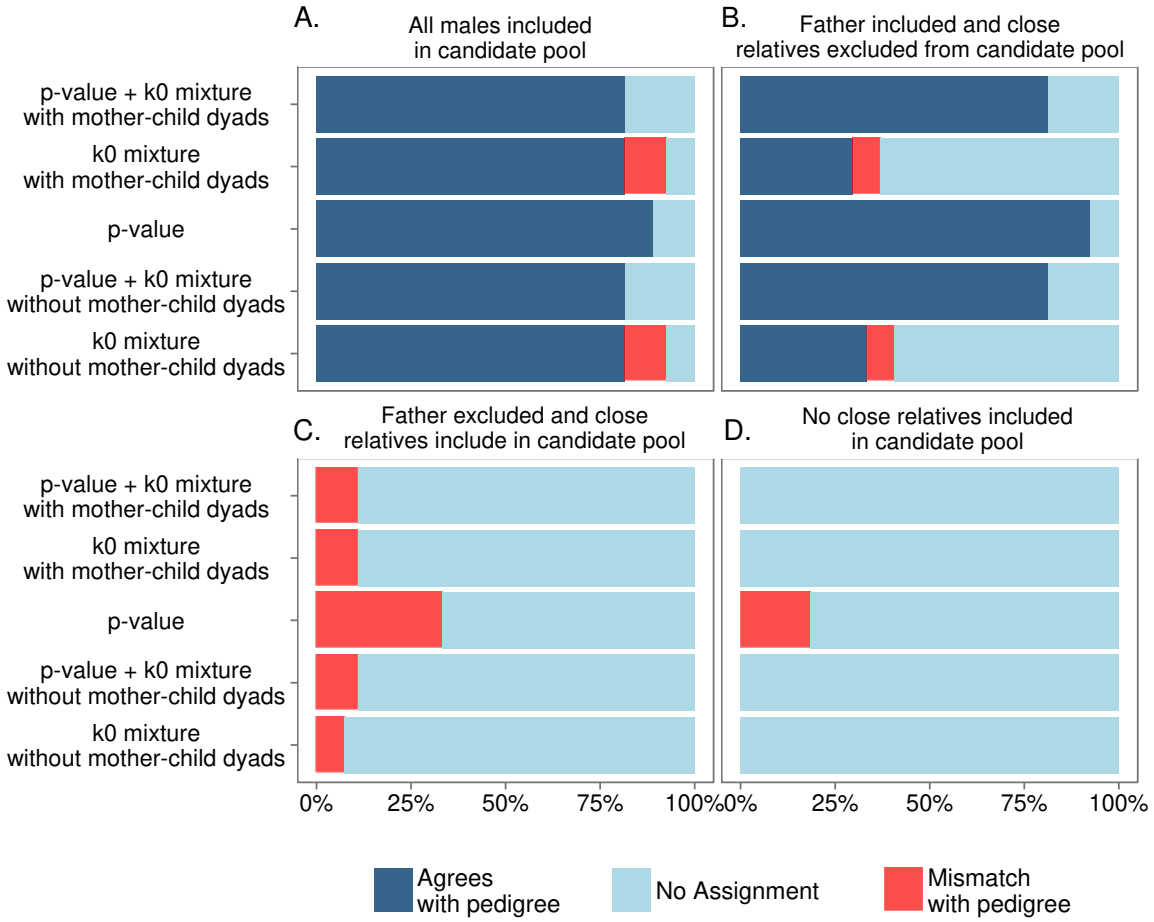


Figure S10 Comparison of paternity assignment accuracy. Using both a p-value cutoff of 0.05 and a $P(z_{bi}=1)$ cutoff of 0.90 maximized the number of correct paternity assignments while minimizing the number of incorrect assignments ("p-value plus k0 mixture, with mother-child dyads:" top bar in all four panels). Correct assignments (dark blue) reflect cases in which the pedigree-assigned father was also assigned as the father using these criteria (with known mother-offspring dyads used to inform the mixture model). "k0 mixture with mother-child dyads": $P(z_{bi}=1)$ cutoff of 0.90 and no p-value criterion, while using known mother-offspring dyads to inform the mixture model. "p-value": p-value cutoff of 0.05 and no $P(z_{bi}=1)$ criterion. "p-value + k0 mixture without mom-child dyads": p-value cutoff of 0.05 and $P(z_{bi}=1)$ cutoff of 0.90, without using known mother-offspring dyads to inform the mixture model. "k0 mixture without mother-child dyads": $P(z_{bi}=1)$ cutoff of 0.90 and no p-value criterion, without using known mother-offspring dyads to inform the mixture model. "Incorrect" assignments (red) were males incorrectly assigned as the father (compared to the previous pedigree-assigned father). All other offspring were not assigned a father ("No assignment"; light blue). (A) All males included in the candidate pool. (B) Father included in the candidate pool, but all other close male relatives ($r \geq 0.25$) excluded. (C) Pedigree-assigned father excluded from the candidate pool, but other close male relatives retained. (D) Father and all close relatives ($r \geq 0.25$) excluded from the candidate pool.

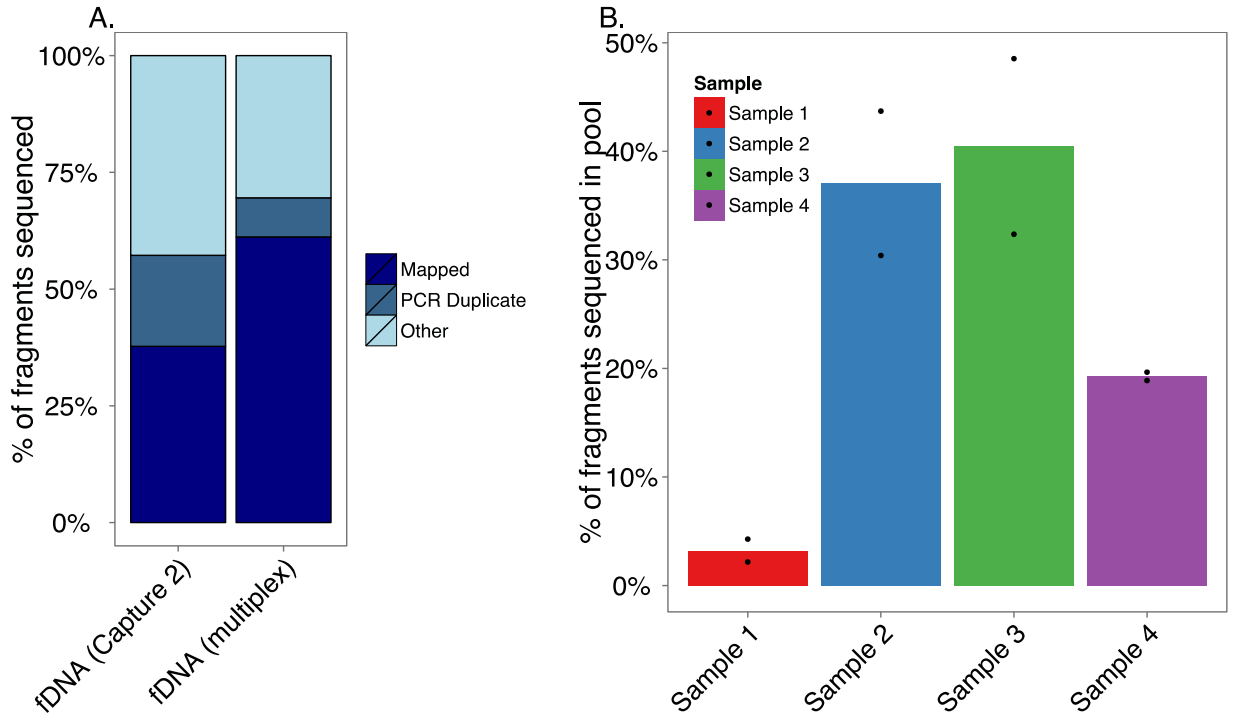


Figure S11 Post-capture mapping of multiplexed libraries. (A) Multiplexing samples prior to the hybridization resulted in similar or even better enrichment than single-plexed libraries that were pooled for sequencing after hybridization. (B) Pools multiplexed prior to capture (n=2 pools) contained 4 samples in each pool (the two raw values for each sample, one from each pool, are shown as black dots). Although the samples were pooled equally prior to hybridization, the resulting data produced uneven coverage per sample, primarily driven by poor sequencing of Sample 1.

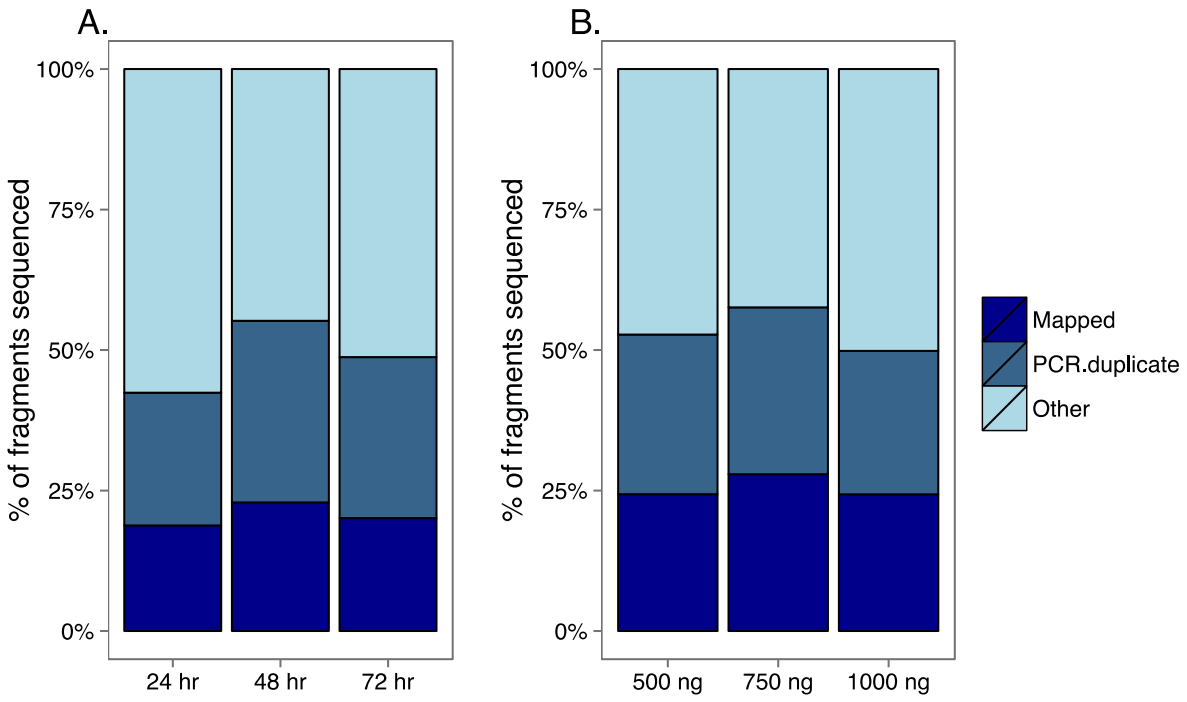


Figure S12 Optimization of probe concentration and hybridization length. (A) Capture efficiency is influenced by the duration of hybridization and (B) the amount of bait added to the capture reaction. The strongest enrichment of both mapped reads and non-duplicate fragments was observed with a 48-hour incubation and 750 ng of bait, so we used these values in the capture protocol.

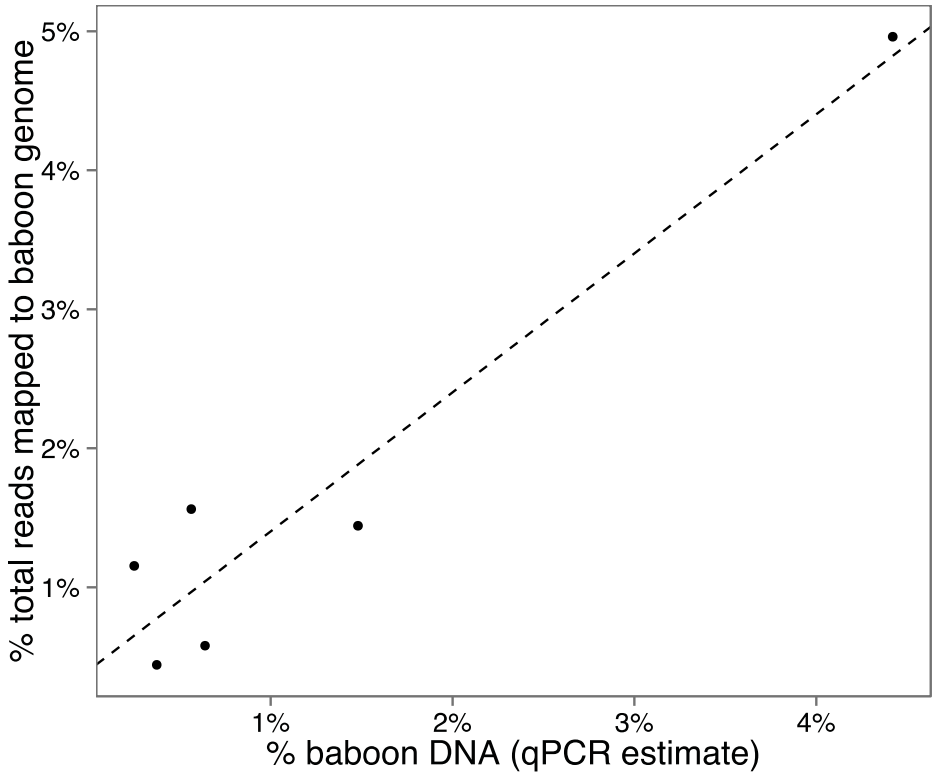


Figure S13 Accuracy of qPCR estimates of endogenous baboon DNA prior to capture. The qPCR-based measure of endogenous DNA predicted the proportion of non-duplicate mapped reads in pre-capture libraries constructed from the same fDNA samples (β=1.00, T=6.66, p=2.6x10^{-3}).

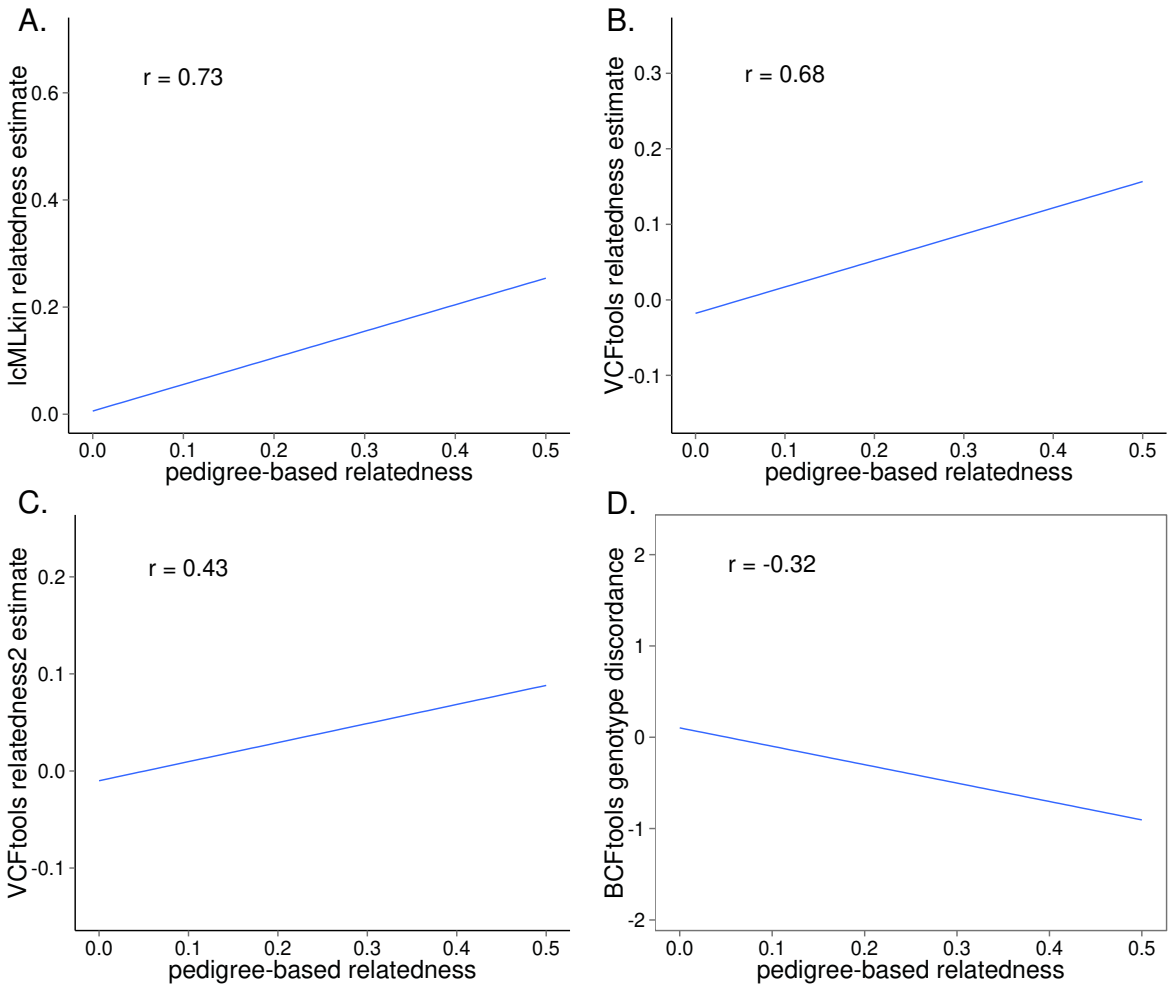


Figure S14 Relationship between pedigree relatedness and four measures of genotype similarity. (A) The correlation was the strongest between the independently established pedigree relatedness and the *lcMLkin* measure of relatedness. Other measures of genotype similarity are also correlated with independently established pedigree relatedness, although not as strongly as the *lcMLkin* measure. (B) Relatedness estimated with *vcftools* using the method of Yang et al. (2010). (C) Relatedness estimated with *vcftools* using the method of Manichaikul et al. (2010). (D) Genotype discordance per site genotyped as measured using the function "gtcheck" in the program *bcftools* (Li et al., 2009). The y-axes in panels B-D are the residual relatedness (B and C) or dissimilarity (D) scores after controlling for the batch effect of capture effort (Capture 1 or Capture 2) using a linear model with one categorical predictor variable where each dyad was categorized as either both from capture 1 ("within capture 1"), both from capture 2 ("within capture 2"), or one member of the dyad from each capture effects ("between capture"). There was no detectable effect of capture effort on the *lcMLkin* relatedness measure, so the y-axis in (A) shows the raw value of *r* estimated from *lcMLkin*. Blue lines show the fit from a linear model.

Table S1. Details of library preparation, DNA capture, and sequencing. (.xlsx, 43 KB)

Available for download as a .csv file at
www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.187492/-/DC1/TableS1.xlsx

Table S2. Sample-specific collection, extraction, and sequencing data. (.xlsx, 59 KB)

Available for download as a .csv file at
www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.187492/-/DC1/TableS2.xlsx

File S1. SUPPLEMENTARY METHODS

Expectation-maximization (EM) Algorithm for the WHODAD mixture model

As explained in the main text methods, we denote y_b as the vector of logit(*k0*) measurements for the best candidate father-offspring dyads for all tested offspring; y_1 as the vector of logit(*k0*) measurements for all known mother-offspring dyads, if any are present (y_1 can be an empty vector if no mother-offspring dyads were sampled); and y_0 as the vector of logit(*k0*) measurements for all other dyads.

We model the elements in y_0 as a mixture of K normal distributions, where different distributions capture different degrees of relatedness:

$$y_{0i} \sim \sum_{k=1}^{K} w_k N(\mu_k, \sigma_k^2) \quad (1)$$

We model the elements in y_1 as a single normal distribution:

$$y_{1i} \sim N(\mu, \sigma^2) \quad (2)$$

and we model the elements in y_b as a mixture of two normal distributions:

$$y_{bi} \sim \pi N(\mu, \sigma^2) + (1-\pi) N(\mu_i, \sigma_i^2) \quad (3)$$

where

$$\mu_i \in (\mu_1, \cdots, \mu_K), \sigma_i^2 \in (\sigma_1^2, \cdots, \sigma_K^2)$$

We fit equation (1) using the *mixtools* function in R, with a default value of K=5. We assign μ_i and σ_i^2 in equation (3) to the mean and variance of the mostly likely normal component inferred for equation (1) by evaluating the likelihood of y_{1i} under all K components. We then combine y_1 and y_b to jointly infer the hyper-parameters π, μ, σ^2 in equations (2) and (3).

To fit equations (2) and (3), we introduce a latent indicator variable z_{bi} for each dyad to indicate if the *i*th dyad in y_b belongs to the first normal component $N(\mu, \sigma^2)$ (the component that presumably captures true parent-offspring relationships). The probability of being in the first component, or $P(z_{bi}=1)$, effectively captures how similar the *i*th dyad is to the mother-offspring dyads (or, if no maternal information is available, distinct from the distribution for the next most closely related set of dyads) and thus how likely it is to be a true father-offspring dyad. We use this probability as the final statistic to assess our paternity assignments.

We use an expectation-maximization (EM) algorithm to infer both $P(z_{bi}=1)$ and the hyper-parameters π, μ, σ^2. Our EM algorithm iterates through the following steps:

Step 1: $\pi_t = \sum_i P_{t-1}(z_{bi} = 1) / n_b$

Step 2: $\mu_t = (\sum_i P_{t-1}(z_{bi} = 1) y_{bi} + \sum_j y_{1j}) / (\sum_i P_{t-1}(z_{bi} = 1) + n_1)$

$$\sigma_t^2 = (\sum_i P_{t-1}(z_{bi}=1)(y_{bi}-\mu_t)^2 + \sum_j (y_{1j}-\mu_t)^2)/(\sum_i P_{t-1}(z_{bi}=1)+n_1)$$

Step 3: $P_t(z_{bi}=1) = \pi_t N(y_{bi};\mu_t,\sigma_t^2)/(\pi_t N(y_{bi};\mu_t,\sigma_t^2)+(1-\pi_t)N(y_{bi};\mu_i,\sigma_i^2))$

where n_b is the number of dyads in y_b, n_1 is the number of dyads in y_1 and t indicates the iteration number.

Step 1 updates π (the probability that a dyad in y_b belongs in the parent-offspring distribution) by using the current average estimate of $P(z_{bi}=1)$ across all dyads in y_b. Step 2 updates μ and σ^2 by using the weighted estimates from y_1 and y_b, with the weights for y_1 (known mother-offspring dyads) equal to 1 and the weights for y_b equal to the current estimates of $P(z_{bi}=1)$; and Step 3 updates $P(z_{bi}=1)$ conditional on the updated estimates of π, μ, σ^2 (from Steps 1 and 2).

For all our analyses, we initialized the EM algorithm with values $\pi = 0.1, \mu = 0, \sigma^2 = 1$ and performed 50 iterations. The results were robust with respect to different initial values. In addition, we used K=5 components to fit equation 1. We could choose an optimal K based Akaike information criterion (AIC) or Bayesian information criterion (BIC), but we found that our results were robust to reasonable choices of K. Finally, we note that in principle we could develop an EM algorithm to fit equations 1-3 jointly. In practice, however, we found that using the above two-step procedure (i.e. fitting equation 1 first and then fitting equations 2 and 3 second) produces more stable results with respect to initial values. This presumably is because the joint model has more parameters compared with each of the two separate models; thus, fitting the joint model can be more sensitive to initial values than the two-step approach we took. With our two-step procedure described above, with any reasonable initial values, our algorithm converged to the same results often within a few iterations.

Protocol improvements

We made a number of changes to the protocol between Capture 1 and Capture 2, which are detailed in Table S1. These changes were made to increase the flexibility and decrease the cost of the protocol. Specifically: (i) we increased the library insert size from ~200bp to ~400bp by changing the Bioruptor shearing settings, which allowed us to generate longer, paired-end reads, reduce the per-base pair sequencing cost, and increase read mappability; (ii) we switched from gel-based size selection to a SPRI bead-based size selection, which decreased the amount of time needed to perform the protocol; (iii) we switched from KAPA and Illumina reagents to to NEBNext library preparation reagents (New England Biolabs) because they reduced protocol costs.

Another improvement to the protocol that was not implemented in Capture 2, but is a promising direction for the future is pooling samples for multiplexing prior to hybridization (instead of post-capture, prior to sequencing). Results from a preliminary test of this approach are shown in Fig. S11 and discussed briefly in the main text: they suggest that multiplexing not only reduces input and reagent requirements, but may also increase the overall fold-enrichment of endogenous DNA in the resulting library. However, bulk amplification of post-capture multiplexed samples may increase the possibility of barcode switching. Use of dual-indexing approaches to discriminate samples within a pool, and direct assessment of barcode switching rates (e.g., using mitochondrial DNA reads) could help alleviate this problem (although generating data from pedigrees may mean that many individuals share mtDNA haplotypes through maternal descent). Meanwhile, complementary improvements to the computational pipeline could explicitly incorporate information on haploid/sex-linked chromosomes (i.e., mtDNA and Y), and include statistical approaches that account for the uncertainty introduced by allelic dropout. We note that, in this analysis, we did not remove sex-linked loci because they are not annotated in *Pcyn1.0* (although we did remove mtDNA contigs). Our results thus suggest that *WHODAD*'s performance is robust to the presence of combined autosomal and sex-linked reads, and that high quality assemblies are not essential for accurate performance.

SUPPLEMENTARY REFERENCES:

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*(15), 2156–8. doi:10.1093/bioinformatics/btr330

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–9. doi:10.1093/bioinformatics/btp352

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)*, *26*(22), 2867–73. doi:10.1093/bioinformatics/btq559

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., … Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–9. doi:10.1038/ng.608