1 **Title:** Efficient genome-wide sequencing and low coverage pedigree analysis from non-
2 invasively collected samples
3
4 **Authors**
5 Noah Snyder-Mackler[1], William H. Majoros[2], Michael L. Yuan[1], Amanda O. Shaver[1],
6 Jacob B. Gordon[3], Gisela H. Kopp[4], Stephen A. Schlebusch[5], Jeffrey D. Wall[6], Susan C.
7 Alberts[3,7], Sayan Mukherjee[8,9,10], Xiang Zhou[*,11,12], Jenny Tung[*,13,7,13]
8 **Author Affiliations**
9 [1] Department of Evolutionary Anthropology, Duke University, Durham, NC
10 [2] Graduate Program in Computational Biology and Bioinformatics, Duke University,
11 Durham, NC
12 [3] Department of Biology, Duke University, Durham, NC
13 [4] Cognitive Ethology Laboratory, German Primate Center, Leibniz Institute for Primate
14 Research, Göttingen, Germany
15 [5] Department of Molecular and Cell Biology, University of Cape Town, Cape Town,
16 South Africa
17 [6] Institute for Human Genetics, University of California San Francisco, San Francisco,
18 CA
19 [7] Institute of Primate Research, National Museums of Kenya, Nairobi, Kenya
20 [8] Department of Statistical Science, Duke University, Durham, NC
21 [9] Department of Mathematics, Duke University, Durham, NC
22 [10] Department of Computer Science, Duke University, Durham, NC
23 [11] Department of Biostatistics, University of Michigan, Ann Arbor, MI

[12] Center for Statistical Genetics, University of Michigan, Ann Arbor, MI

[13] Duke University Population Research Institute, Duke University, Durham, NC 27708

Keywords: Capture-based enrichment, non-invasive samples, baboons, paternity analysis, pedigree, genome resequencing

***Authors for Correspondence**

Jenny Tung

Department of Evolutionary Anthropology, Duke University

104 Biological Sciences Building, Box 90383

Durham, NC 27708-9976

Fax: 919-660-7348; Phone: 919-684-3910

E-mail: jt5@duke.edu

Xiang Zhou

Department of Biostatistics, University of Michigan

1415 Washington Heights #4623

Ann Arbor, MI 48109

Phone: 734-764-7067

E-mail: xzhousph@umich.edu

Running title: Genomic enrichment of non-invasive samples

ABSTRACT:

Research on the genetics of natural populations was revolutionized in the 1990's by methods for genotyping non-invasively 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 non-invasively 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 non-invasive samples are available. The lab protocol and software ("*WHODAD*") are freely available at www.tung-lab.org/protocols and www.xzlab.org/software, respectively.

The capacity to generate genetic data from low-quality or non-invasively collected samples, first developed in the 1990's[1,2], 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[3–5]; helped define the role of admixture in mammalian evolution[6–8]; and enabled theoretical expectations about paternal investment, kin recognition, and reproductive skew to be empirically tested, sometimes for the first time[9–12]. 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[13–17]. Non-invasive genetic analysis has thus changed the ways we study population, ecological, and conservation genetics. Indeed, these fields would look very different today—and we would know far less about many species—without it.

However, techniques for non-invasive genetic analysis have changed little in the past twenty years. Collection of genetic data from non-invasively collected tissues (e.g., feces, hair, urine) continues to be labor-intensive, time-intensive, and vulnerable to technical artifacts such as allelic dropout and cross-contamination[18,19]. Further, current methods ultimately yield very small amounts of data by today's standards. Typical studies genotype only a dozen 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, such as identifying signatures of natural selection,

reconstructing population history and demography, and testing for genetic associations with phenotypic variation[20–22]. Similarly, analyses that require local (i.e., gene- or region-specific) information on genetic diversity, structure, or ancestry instead of genome-wide averages cannot be conducted[23–27]. Finally, because non-invasively collected genotype data are most often based on microsatellites, they cannot take advantage of new tools designed specifically for single nucleotide variants[28–30].

Generating genome-scale data sets from non-invasive 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 non-host DNA. For example, only about 1% of DNA extracted from fecal-derived samples is endogenous to the donor animal (most is microbial)[31]. Sequence capture methods, in which synthesized baits are used to enrich for pre-specified target sequences from a larger DNA pool[32], 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 and colleagues[31] were able to target and sequence 1.5 megabases of the chimpanzee genome from fecal-derived DNA, using a modified version of sequence capture, with very low genotyping error rates relative to blood-derived DNA. More recently, Carpenter et al.[33] reported a method for performing genome-wide sequence capture from low-quality ancient DNA samples, which recapitulate many of the challenges posed by non-invasive 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 post-capture enrichment associated with

currently available protocols are not cost-effective for population studies of non-invasive samples. Substantially higher rates of enrichment, particularly in non-repetitive 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[34–36] 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 non-invasively 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 lab methods and computational tools using non-invasively collected samples from 54 members of an intensively studied wild baboon population in the Amboseli basin of Kenya[37]. We also demonstrate the generalizability of our methods to non-invasive 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, non-repetitive elements, 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 = 0.49x 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 valuable 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 non-invasive samples. Similar to Carpenter et al[33], 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 Gnirke et al.[32] and Perry et al.[31]), but can also produce a bait set that includes a large proportion of low complexity, repetitive regions. Consequently, reads generated from captured DNA cannot be uniquely mapped, lowering the protocol's efficiency relative to using a more diverse bait set. To address this concern, we incorporated a novel duplex specific nuclease (DSN) digestion in the bait construction step (Fig. 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 hour incubation of sheared DNA at 68°C followed by a 20 minute DSN digestion in the presence of human Cot-1 greatly improved the efficiency of capture, producing 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%; Fig. 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; Fig. 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[9,37,38]. 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; Table S2 provides detailed 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 high-quality genomic DNA samples (gDNA) extracted from blood for the same individuals.

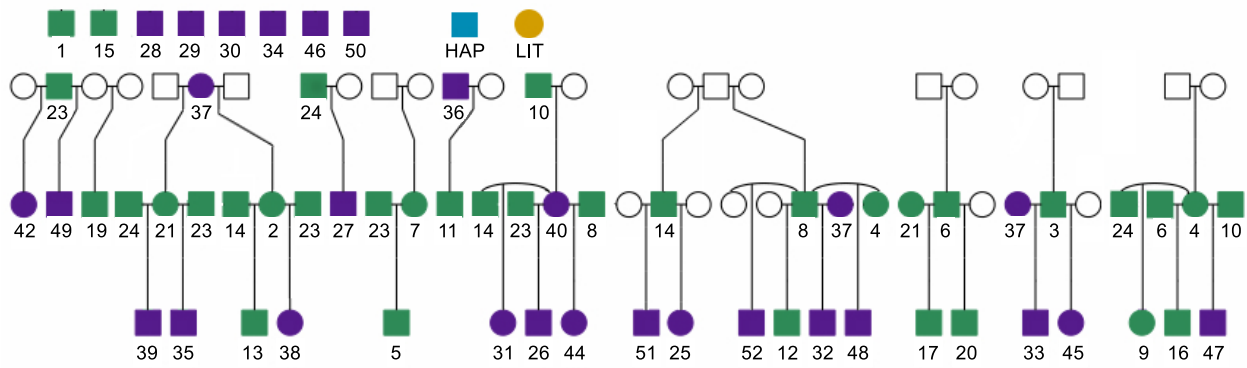

Figure 1. Pedigree of a subset of baboons monitored by the Amboseli Baboon Research Project. Samples from both males (squares) and females (circles) were enriched in Capture 1 (green) or Capture 2 (purple). Unfilled circles/squares represent baboons that connect individuals in our pedigree, but who were not sequenced as part of this study. Each sequenced individual is represented by a unique number (below the circle/squares); note that some individuals are repeated in the figure 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.

Our protocol (Fig. S1) resulted in substantial enrichment of baboon DNA in the post-capture versus pre-capture samples (see Table S2 for sample-specific details). A mean of 44.56% (range: 10.28-83.17%) of post-capture fragments mapped to the baboon genome, despite starting with pre-capture samples that contained a mean of only 2.04% endogenous baboon DNA, as estimated by qPCR (range 0.19-8.37%). However, in Capture 1 a large proportion of the mapped fragments were identified as PCR duplicates ($mean_{capture1}$=71.97% of mapped fragments, $range_{capture1}$: 51.43-88.46%; Fig. 2A). After removing PCR duplicates, a mean of 9.16% of the post-capture reads in Capture 1 were non-duplicate mappable fragments ($range_{capture1}$=2.23%-23.75%), producing a mean coverage of 0.20x per sample relative to the mappable baboon genome (mean sequencing depth of 5.8 Gb per sample; $range_{capture1}$ = 0.04-0.49x; Fig. 2B). These numbers translated to an overall mean fold enrichment of 39.8-

188 fold for mapped reads ($range_{capture1}$: 8.0-111.8-fold, s.d.=25.2), and 9.6x enrichment of
189 non-PCR duplicate mapped reads ($range_{capture1}$:3.9-22.4-fold, s.d.=5.0; Fig. 2C).

190 Based on our results for Capture 1, we made multiple protocol improvements
191 prior to conducting Capture 2 (Table S1). The improved protocol was twice as effective
192 on average, resulting in a mean 18-fold enrichment of high quality, analysis-ready reads
193 and a maximum fold enrichment of close to 40-fold ($range_{capture2}$ = 8.0-39.2-fold; Fig. 2C;
194 by comparison, methods optimized for ancient DNA achieved a mean of 5.5-fold
195 enrichment of non-PCR duplicate fragments[33]; Fig. 2A). Specifically, the protocol
196 changes improved the proportion of non-duplicate mapped fragments by more than
197 four-fold, from a mean proportion of 9.16% in Capture 1 to a mean proportion of 37.74%
198 in Capture 2 ($range_{capture2}$=6.16-68.61%) and reduced the proportion of PCR duplicates
199 among mapped reads two-fold (from 71.97% in Capture 1 to 36.97% in Capture 2). This
200 improvement translated to an increase in overall genomic coverage from a mean of
201 0.20x in Capture 1 to 0.73x in Capture 2 (mean total sequencing of 5.7Gb per sample;
202 $range_{capture2}$ = 0.19-1.24x; Fig. 2B). This improvement in coverage was not explained by
203 increased sequencing depth in Capture 2 (Table S2). Thus, while we would need to
204 sequence a pre-capture fDNA sample 50-100 times as deeply as a blood or tissue-
205 derived sample to produce the same level of coverage, our capture method reduces this
206 difference to approximately 2 times the sequencing effort. Importantly, our method was
207 also successful in enriching fDNA samples (n=8) from independent samples collected
208 from Guinea baboons (*P. papio*; Fig. 2A, Table S2), suggesting that our results are
209 highly generalizable across different species and storage and extraction methods.

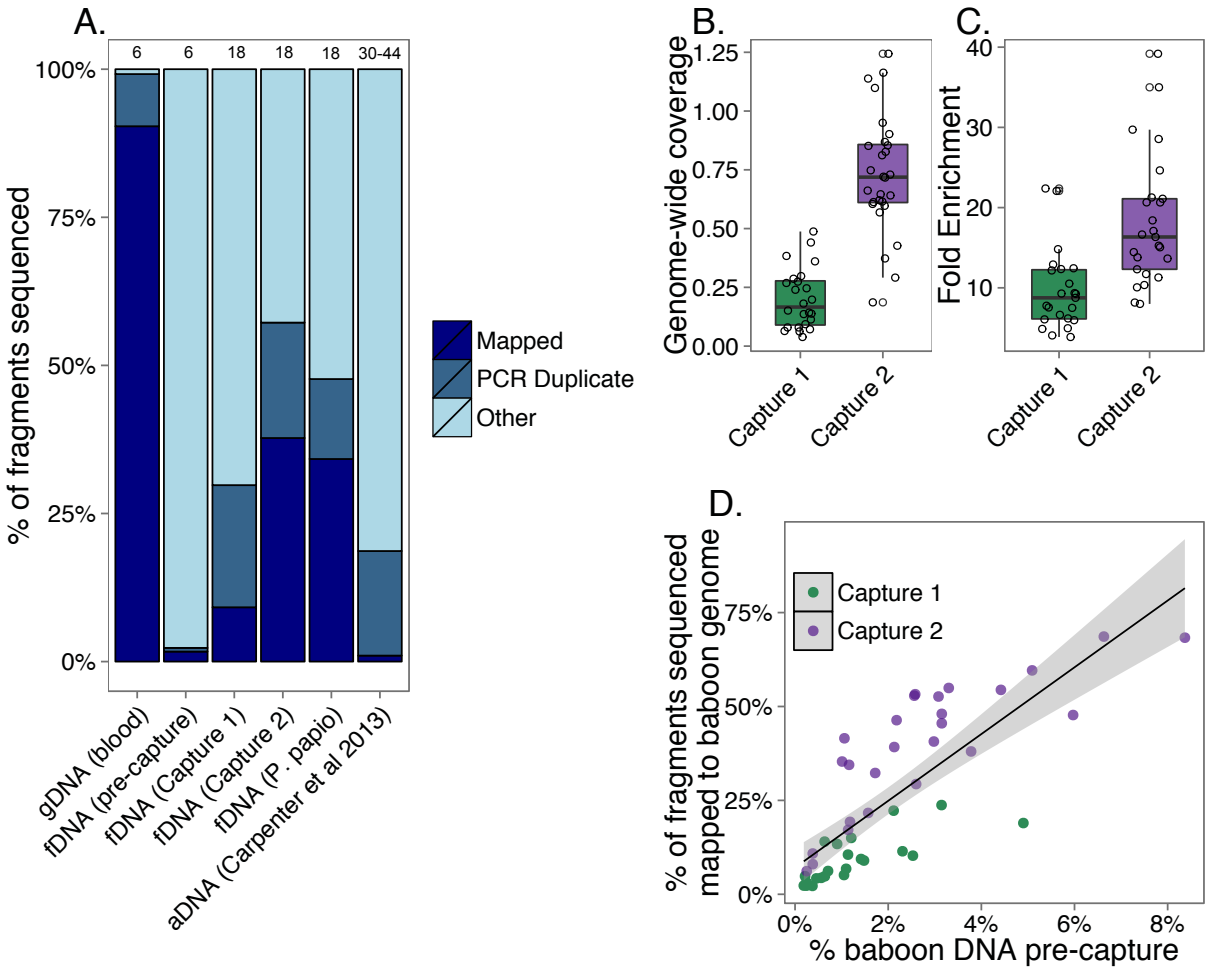

Figure 2. **fDNA enrichment results**. (A) Percent of sequencing reads that mapped to the baboon genome and were not PCR duplicates ("Mapped:" dark blue); mapped and were PCR duplicates ("PCR Duplicate:" blue); or did not map and likely represent environmental or bacterial DNA in the case of fDNA/aDNA 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[33]. 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 similar number of overall reads generated per sample (two-sample t-test, T=9.7, $p=3.0x10^{-12}$). On average in Capture 2, we obtained ~0.73x coverage of the genome with 5.76Gb of sequencing. If all 5.76Gb mapped to the baboon genome as non-PCR duplicates, we would have produced ~2.2x genome-wide coverage. (C) Capture 2 also produced significantly greater fold enrichment of baboon DNA (fold enrichment is measured as % non-duplicate baboon DNA post-capture divided by % baboon DNA pre-capture: two-sample t-test, T=4.4, $p=7.3x10^{-5}$). (D) The amount of baboon DNA in the sample pre-capture (% baboon DNA pre-capture, based on qPCR of the single copy *c-myc* gene[39]) is strongly correlated with the percentage of baboon fragments obtained in post-enrichment sequencing (Pearson's r=0.80, $p=1.0x10^{-11}$). However, even samples with low amounts of endogenous DNA (<2%) exhibit substantial fold enrichment using our protocol

(mean$_{capture1}$=10.60x, mean$_{capture2}$=24.82x).

Sample attributes influencing capture efficiency

The amount of baboon DNA in the pre-capture fDNA sample was the strongest predictor of enrichment success. Specifically, the percent of baboon DNA pre-capture, as assessed via qPCR, was positively correlated with the percentage of non-duplicate fragments mapped post-capture (Fig. 2D; T=6.88, p=1.72x10^{-8}). Samples from Capture 2 had more pre-capture 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 Capture 1. This pattern is observable whether assessed using the percent of baboon DNA fragments sequenced post-capture ($T_{capture2}$=10.00, p=6.76x10^{-13}) or fold enrichment relative to pre-capture amounts ($T_{capture2}$=6.89, p=1.69x10^{-8}), and could not be explained by differences in the length of sequence fragments or overall sequencing depth (Fig. S3; Table S2). The amount of fDNA library used in the capture reaction was also weakly positively correlated with the percent of baboon DNA fragments sequenced post-capture, after controlling for the amount of baboon DNA in the pre-capture sample ($T_{ng\_fDNA\_library}$=2.09, p=0.042; Table S2).

Library complexity, distribution of reads, and GC content

231 The post-capture libraries included a higher proportion of PCR duplicates relative
232 to reads generated from high-quality genomic DNA samples, for which fewer rounds of
233 PCR amplification were required (PCR duplicate proportion: $mean_{fDNA\_capture1}$=69.6%,
234 $mean_{fDNA\_capture2}$=36.8%, $mean_{gDNA}$=11.3% of mapped reads; 18 rounds of PCR in the
235 capture protocol versus 6 for the high-quality samples). For comparison, this proportion
236 is much lower than reported for aDNA samples, which go through more rounds of PCR
237 amplification ($mean_{aDNA}$=94.6%; Fig. 2A and Fig. S4[33]). Despite increases in clonality,
238 the number of non-duplicate reads continued to increase with increasing sequencing
239 depth, with the slope of this relationship especially favorable for Capture 2 (Fig. 3).
240 Thus, deeper sequencing of post-capture libraries should continue to increase genome-
241 wide coverage, albeit not as efficiently as sequencing blood-derived gDNA samples.

242

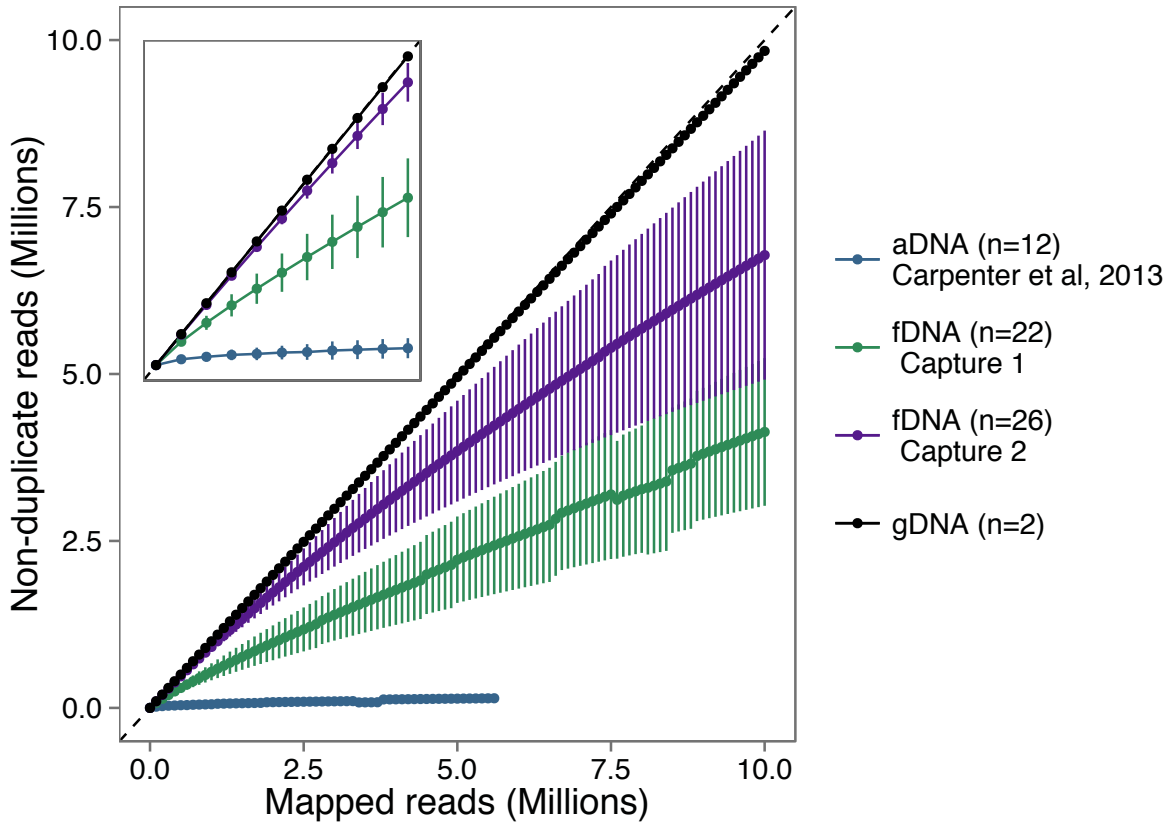

Figure 3. Increased sequencing effort produces increased numbers of non-duplicate reads. The number of mapped reads plotted against the number of non-duplicate reads mapped (mean ± SD; plotted using the program "preseq"[40]). More complex libraries (i.e., those containing more non-duplicate fragments) 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 ten million mapped reads for each sample. The inset shows the same plot for the first million mapped reads.

As with other capture-based methods[33,41], 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 ($mean_{capture1}$=6.55%; $mean_{capture2}$=6.73%; Fig. S5A). However, Capture 2 resulted in significantly more unambiguously non-duplicate mtDNA-mapped reads than Capture 1, largely due to the paired-end sequencing used in Capture 2 ($mean_{capture1}$=0.47% of all mapped reads; $mean_{capture2}$=6.46%; Fig. S5B). The higher number of non-duplicate

mtDNA reads in Capture 2 thus produced much deeper overall coverage of the mitochondrial genome (Fig. S5C), despite the fact that the ratio of mtDNA to nuclear DNA mapped reads was comparable between the two captures (Fig. S5D). Finally, the distributions of read GC content for post-capture reads using our protocol, the DNA template for the RNA baits, and aDNA libraries were highly similar (Fig. S6). This observation suggests that any GC bias relative to the genome appears during bait construction and/or sequencing, not during the hybridization step.

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

To assess the accuracy of genotypes called from post-capture fDNA libraries, we compared genotype data from paired blood-derived gDNA (without capture) and post-capture 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%, or 270,724 of 312,739 sites for the LIT paired samples; 77%, or 30,948 of 40,132 sites for the HAP paired samples; note that we had lower coverage for the HAP fecal-derived sample than for the LIT 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% of discordant LIT_{gDNA} heterozygous sites; 74.4% of discordant HAP_{gDNA} heterozygous sites). Importantly, the fDNA genotype captured at least one of the alleles from the gDNA genotype in 99.8% (LIT) and 99.6% (HAP) of these

discordant sites. 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 post-capture fDNA libraries were more similar to the genotypes called from their high-quality gDNA counterparts than they were to other post-capture fDNA libraries. Specifically, *k0* values from *lcMLkin*[42], 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} versus other fDNA samples = 0.996 – 1.000; Z=849.2, $p < 10^{-20}$; *k0* range HAP_{fdna} versus other fDNA samples = 0.786 – 0.999; Z=10.6, $p < 10^{-20}$; Fig. 4A).

For the 48 extended pedigree individuals (Fig. 1, including 8 Amboseli baboons with no known relatives in the pedigree), we then tested if estimated relatedness values from *lcMLkin*[42] in the post-capture data were correlated with 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 for filtering parameters), we found a strong correlation between the two measures (Pearson's r=0.73, $p<10^{-16}$; Fig. 4B). This correlation improved further if we imposed thresholds for the minimum number of sites genotyped in both individuals ("shared sites") in a dyad (Fig. S7). For example, if we removed all dyads with fewer than 2,000 shared sites (84 of 1,128 dyads, or 7.4%), the correlation between pedigree relatedness and genotype similarity reached r=0.86 ($p<10^{-16}$).

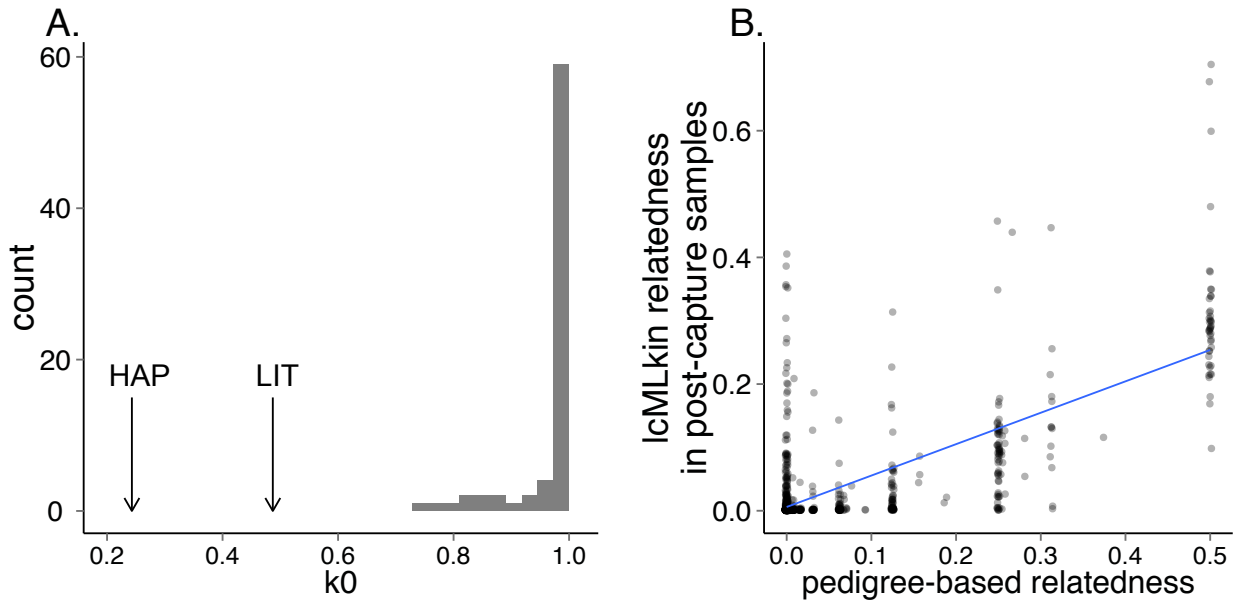

Figure 4. Post-capture genotype data are consistent with individual identity and pedigree relationships. (A) The *k0* values for the HAP and LIT fDNA-gDNA paired samples (arrows) were significantly lower than the range of *k0* values for LIT_{fDNA} and HAP_{fDNA} versus any other fDNA sample (gray distribution). Lower *k0* values reflect increased relatedness (i.e., decreased probability of no IBD sharing). (B) Estimated dyadic relatedness values were correlated with independently obtained pedigree relatedness values calculated using the R package *kinship2* (Sinnwell et al. 2014; $r=0.73$, $p<10^{-16}$). Both *k0* and the estimated relatedness values were calculated with *lcMLkin*[42].

296

297 *Paternity inference using WHODAD*

298 Current methods for assigning paternity (e.g., CERVUS[34,35] and exclusion[36])
299 assume genotype certainty, such that individuals are assigned a deterministic genotype
300 at each locus (i.e., 0, 1, or 2, or a microsatellite repeat number; while a low level of
301 measurement error, i.e., due to lab handling, can be modeled, this error rate is held
302 constant across genotype calls). This assumption is violated in low coverage
303 sequencing data, in which genotypes are not known with certainty and this uncertainty
304 varies across genotype calls. However, the relative *probabilities* of each genotype can
305 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 Fig. S8).

Using *WHODAD*, we assigned paternity to all father-offspring pairs (n = 27) represented in the independently established extended pedigree in Fig. 1. Note that our approach represents a particularly conservative test 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 Fig. 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 three of these cases, sequencing coverage was very low for either the pedigree-identified father or offspring: 0.04-0.17x). For the remaining 12 offspring, we did not generate genotype data using our enrichment protocol for their mothers (i.e., their mothers were not among the samples run in Capture 1 or 2). To test all 27 father-offspring dyads together, we therefore re-ran *WHODAD* excluding maternal genotype information. In this setting, *WHODAD*'s paternity assignments agreed with the pedigree data in 22 of 27 (81%) cases (Fig. 5).

329 Notably, when the pedigree-identified father was included in the data set, *WHODAD*
330 never assigned paternity to a different male, whether or not maternal genotype data
331 were available. Because our method is highly robust to exclusion of maternal genotype
332 data, we therefore performed all subsequent analyses assuming maternal genotype
333 data were *not* available. This approach allowed us to evaluate all father-offspring dyads,
334 and also captures a scenario that may often occur in studies of natural populations.

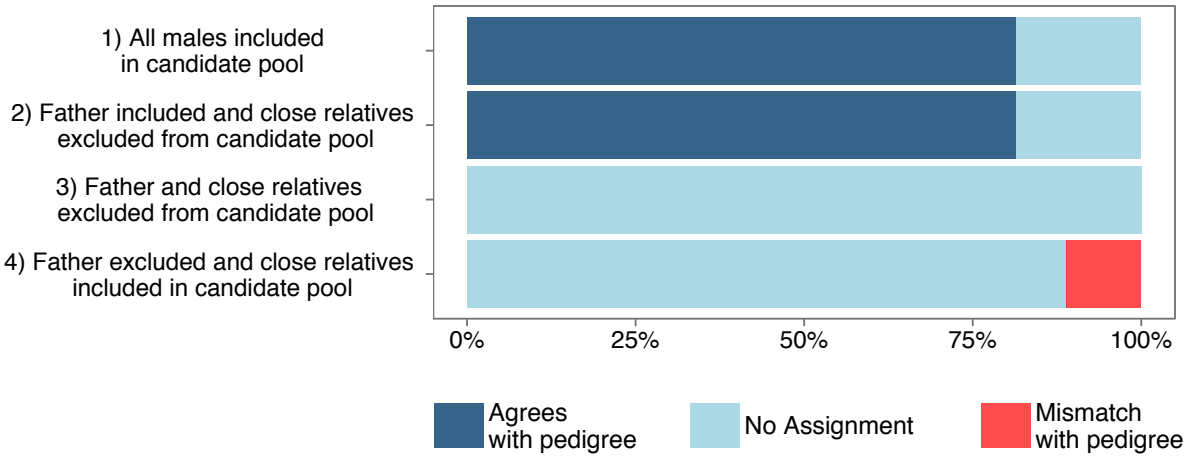

Figure 5. Paternity inference with *WHODAD* using low coverage genotype calls. 1) When all males (n=34) were included in the pool of candidate fathers (top bar), *WHODAD* assigned paternity to the same father identified in the pedigree for 22 of 27 (81%) of offspring (see assignment criterion in Methods; dark blue). The remaining offspring were not assigned a father based on *WHODAD*'s assignment criteria (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).

335 The presence of close relatives, such as full or half-siblings, can influence the
336 accuracy of paternity assignment if these close relatives are also included as candidate
337 fathers[35,44–46]. 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 (Fig. 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 (Fig. 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 (Fig. S9). For example, when all males were included in the candidate pool, the combined approach resulted in an 81% true positive rate and a 0% false positive rate, while just using the *k0* values in a mixture model resulted in the same true positive rate (81%), but an additional 11% false positive rate (Fig. S9).

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

361 enrichment-based capture methods[33,47,48] for non-invasively collected samples, which
362 represent a major resource for behavioral, conservation, and evolutionary genetic
363 studies in natural populations. Importantly, our protocol increases efficiency and lowers
364 cost by reducing the input requirements and number of PCR cycles relative to previous
365 methods[31] and, in our final protocol, achieves up to 40-fold enrichment of post-capture
366 endogenous DNA relative to pre-capture levels. We also show, for the first time since
367 Perry et al[31], that capture libraries from low-quality samples produce genotype data that
368 are highly concordant with genotype data derived from high-quality, non-captured
369 samples from the same individuals.

370 We anticipate that data generated through this protocol could be leveraged for a
371 wide variety of applications. To illustrate this point for paternity analysis, one of the most
372 central components of genetic studies in natural populations, we present an
373 accompanying method, *WHODAD*, that produces results in near-perfect concordance
374 with an independently constructed pedigree, using low-coverage data generated with
375 our enrichment protocol (note that the few cases in which assignments could not be
376 confidently made could readily be addressed with slightly deeper sequencing coverage,
377 similar to typing more markers in conventional microsatellite analysis). By incorporating
378 demographic and behavioral data often used to constrain pedigree reconstruction, as
379 well as prior information about other pedigree links, its performance would be improved
380 even further. For instance, in reconstructing pedigree links in the Amboseli population,
381 we generally include only plausible candidates (e.g., we exclude males who were
382 immature or not yet born at the offspring's conception), not all males with genotype
383 data, as we did here.

384 Together, these results provide valuable, accessible wet lab and computational
385 tools for moving studies of difficult-to-sample natural populations forward into the
386 genomics era. Importantly, our methods can be generalized to produce low complexity
387 DNA-depleted RNA baits for any species in which at least one high-quality DNA sample
388 is available (or potentially a closely related species[48]).
389
390 *Costs of performing the protocol*
391 At the time of publication, using the same reagents as we used here and sourced
392 from the same locations, the cost of generating these data is ~$60/sample (including
393 sequencing costs). Importantly, our method does not require the commercial synthesis
394 of targeted capture probes, which is a relatively expensive step for many capture-based
395 approaches[31,32]. Thus, the majority of the costs are accounted for by the streptavidin-
396 coated Dynalbeads ($11/prep), RNA baits ($5/sample) and High Sensitivity Bioanalyzer
397 chips for quality control ($9/sample). Replacing Ampure XP beads with homemade
398 SPRI beads would reduce the per-sample costs considerably, as would pooling
399 adapter-ligated fDNA samples prior to hybridization (instead of post-hybridization, as
400 reported here). For a multiplexed pool of 10 samples, we estimate that using these two
401 strategies would result in a per-sample cost of ~$29. Indeed, we have verified that
402 multiplexing samples prior to hybridization does not result in loss of capture efficiency,
403 and actually resulted in improved yield of mapped, non-PCR duplicate reads (~61% of
404 reads; mean of 117-fold enrichment, range = 54.8 – 257.2-fold; Fig. S10A), although it
405 did result in more uneven coverage of samples sequenced within a pool (Fig. S10B).
406 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 1x genome for baboon (~2.9 Gb) would be about $200 (based on paired-end, 125-bp sequencing at $2,000 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[49]—but provides substantially more genetic information. These estimates will drop further as the cost of high-throughput sequencing continues to fall, making application of our approach to whole populations increasingly feasible. Notably, our finding that useful sequencing reads do not asymptote with deeper sequencing (Fig. 3) also suggests the feasibility of producing a high-quality, high-coverage genome from such samples if one were to sequence more deeply than required for the analyses reported here.

Finally, to make the current protocol as cost-effective as possible, we recommend that researchers use qPCR quantification to choose DNA samples with the highest proportion of host DNA possible—the strongest predictor of the foldchange enrichment in endogenous DNA post- versus pre-capture (Fig. 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 lab 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[34,35,45,46]. Importantly, while analyses of pedigree structure using previously available methods are greatly aided by prior knowledge of mother-offspring relationships[34], maternal links do not appear to be necessary for WHODAD analyses, which performs well even when no maternal information is available (Fig. 5; Fig. S8).

Conclusions

High-throughput sequencing approaches solve one problem of working with low-quality, non-invasive samples: the sheared nature of the original samples. Capture approaches have demonstrated great promise for solving the second major problem—large proportions of non-endogenous DNA—since the results published by Perry et al (2010). Motivated by parallel work on ancient DNA, our results help to fulfill this promise by providing methods to perform cost-effective scaling of sequence capture from non-invasive samples on a genome-wide scale, coupled with analytical methods to deal with the resulting data. Our protocols add an important tool to the range of available options for genetic data generation from such samples. Notably, for questions in which

453 investigators are specifically interested in variants in *a priori*-defined subsets of the
454 genome (e.g., the exome[50,51]), targeted capture with synthesized baits, followed by
455 much deeper sequencing, may still be the best option. However, for the many types of
456 analyses that use genome-scale data (e.g., local ancestry analysis; genome-wide scans
457 for selection, including in non-coding regions; reconstruction of population demographic
458 history[20–27,30]), our approach will be more useful, especially as the costs of high-
459 throughput sequencing continue to fall.

460 Here, we focused specifically on DNA obtained from fecal samples, which are one
461 of the most commonly collected types of non-invasive samples: they contain information
462 not only about host genetics, but also about endocrinological parameters[52], gut
463 microbiota[53], parasite burdens[54], and, as recently demonstrated for human infants, gene
464 expression levels[55]. The sample banks already available for many natural populations
465 thus open the door to population and evolutionary genomic studies in species in which
466 such analyses were previously impossible. As the costs of data generation continue to
467 fall, and the limiting factor for many studies becomes high quality phenotypic data, we
468 envision that such studies will rapidly move far beyond the simple analyses of paternity
469 and pedigree structure reported here.

470

Methods

Bait generation

Similar to Carpenter et al[33], we use a cost-effective *in vitro* synthesis method based on T7 RNA polymerase amplification of sheared DNA from a high-quality sample (Fig. S1A). We extracted genomic DNA from a blood sample collected from an olive baboon (*Papio anubis*) who 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 1.8x ratio of AMPure beads to sample volume.

We annealed custom adapters to the A-tailed library by incubating the following reagents for 15 minutes at 20 °C: 10 µL 5x 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 50µl total volume. The custom adapters we used (EcoOT7dTV: Fwd 5'-GGAAGGAAGGAAGAGATAATACGACTCACTATAGGGCCTGGT, EcoOT7dTV: Rev 5'-/5Phos/CCAGGCCCTATAGTGAGTCGTATTATCTCTTCCTTCCTTCC) differ from those used in other protocols[33,47,48]. Specifically, they contained: 1) a T7 RNA polymerase recognition site; 2) flanking sequence that improves T7 transcription efficiency[56]; 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.[33]).

We then digested the purified, adapter-ligated DNA with duplex-specific nuclease (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 (Fig. S2)[57]. We performed DSN digestion in fifteen 2 µL aliquots, each mixed with 1 µL 4x 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°C for 3 minutes, held the reaction at 68°C for 4 hours, and then added 4 µl H_2O, 1 mL 10x DSN Buffer, and 1 µl DSN (1 U/µL) to the reaction. After 20 minutes of digestion, we stopped the reaction by adding 5 µL 2x DSN Stop Solution (10 mM EDTA) and purified it with 2.4x AMPure beads.

Next, we used Klenow DNA polymerase to blunt end the non-digested DNA, size-selected for 200 - 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 2x 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 1.8x 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°C for 4 hours in the following reaction: 2 µL T7 10x 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

518 total volume. We then digested the DNA template by adding 1 µL TURBO DNase (Life
519 Technologies) to the reaction and incubating at 37°C for 15 minutes. We purified the
520 resulting reaction with the MEGAClear Transcription Clean-Up Kit (Life Technologies)
521 and eluted in a final volume of 70 µL. To cleave off the adapter sequence, we digested
522 the RNA baits with the *Eco*O1091 enzyme (NEB). Lastly, the baits were again purified
523 with the MEGAclear Clean-Up Kit, eluted in 70 µL, and quantified on a Bioanalyzer RNA
524 6000, Eukaryote Total RNA chip (Agilent Technologies).
525
526 *Samples, DNA extraction, and qPCR quantification*
527 Baboon samples were stored in 95% ethanol and fDNA was extracted using the
528 QIAamp DNA Stool Mini Kit (Qiagen; with slight modifications as described in Alberts et
529 al.[38]), or using the QIAxtractor (protocol available here:
530 http://amboselibaboons.nd.edu/assets/84050/alberts_fecal_genotyping_protocol_sca.do
531 cx)[38]. The majority of the sampled individuals (48 of 54) were either members of a single
532 extended pedigree or were unrelated males living in the same study population that
533 were genotyped for inclusion in pedigree building/paternity testing for members of that
534 pedigree (Fig. 1). For LIT and HAP, gDNA was extracted from blood samples using the
535 Qiagen Maxi Kit (Qiagen).
536 To assess our protocol's generalizability to samples collected and stored using
537 different methods, we also extracted fDNA samples from 8 unhabituated Guinea
538 baboons (*P. papio*) sampled in West Africa. These samples were stored in either 90%
539 ethanol or soaked in 90% ethanol and then dried using silica beads (i.e., the "two-step"

method[58,59]). They were then extracted using either the Qiagen DNA Stool Mini Kit or the Gen-ial First DNA All Tissue Kit (Table S2).

We assessed the proportion of endogenous DNA in each fDNA sample using qPCR against the *c-myc* gene, as described in Morin et al.[39].

Library preparation

All samples were fragmented to the desired size (200 or 400 base pairs: 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 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 except where explicitly noted.

Capture-based enrichment

We modified the capture methods from Gnirke et al[32] and Perry et al[31] (Fig. 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

562 reagent ("IBR", 50 μM). This mixture was incubated for 5 minutes at 95°C followed by
563 12 minutes at 65°C. Next, we added 13 μL of hybridization buffer (10x SSPE, 10x
564 Denhardt's solution, 10 mM EDTA, 0.2% SDS, preheated to 65°C), 7 μL hybridization
565 bait mixture (1 μL SUPERase-In, 750 ng RNA baits, and water up to a total volume of 7
566 μL, preheated to 65°C) to the fDNA mixture, and incubated the complete mixture at
567 65°C for 48 hours (see Fig. S11 for comparison of alternative bait concentrations and
568 incubation times).

569 After incubation, we purified the enriched fDNA sample using 50 μL Dynal
570 MyOne Streptavidin T1 beads (Invitrogen). To do so, the beads were washed a total of
571 three times with 200 μL binding buffer (1 M NaCl, 10 mM Tris-HCl [pH 7.5], 1 mM
572 EDTA) and resuspended in 200 μL of binding buffer. Next, the entire fDNA/RNA
573 hybridization mix was added to the 200 μL Dynal MyOne Streptavidin T1 bead and
574 binding buffer slurry. We incubated this mixture at room temperature for 30 minutes on
575 an Eppendorf Thermomixer at 700 rpm. The mixture was placed on a magnetic rack,
576 the supernatant was discarded, and the beads were washed once with 500 μL low
577 stringency wash buffer (1x SSC, 0.1% SDS) followed by a 15-minute incubation at room
578 temperature. The beads were then washed three times with 500 μL high stringency
579 wash buffer (0.1x SSC, 0.1% SDS) with a 10 minute room temperature incubation
580 between each wash. After the final wash, the enriched fDNA fraction was eluted from
581 the beads with 50 μL elution buffer (0.1 M NaOH), transferred to a new tube containing
582 70 μL "neutralization buffer" (1 M Tris-HCl, pH 7.5), purified with 1.8x AMPure beads,
583 and eluted in a 30 μL volume. A final PCR was carried out in a 50 μL reaction volume
584 using 23 μL of the post-hybridization fDNA and either: 1) 25 μL 2x KAPA High Fidelity

585 master mix and 2 µL TruSeq universal primer (Capture 1); or 2) 25 µL 2x NEBNext High
586 Fidelity PCR master mix, 1 µL universal PCR primer, and 1 µL NEB indexing primer
587 (Capture 2). After 12 PCR cycles the final reaction was purified with 1x AMPure beads,
588 eluted in 20 µL H_2O, and visualized on a Bioanalyzer High Sensitivity DNA chip.

589

590 *Sequencing and alignment*

591 All high-throughput sequence generation was conducted on the Illumina HiSeq
592 platform (see Table S1 for sequencing details). The resulting sequencing reads were
593 mapped to a *de novo* assembly of the *Papio cynocephalus* genome (alignment available
594 at https://abrp-genomics.biology.duke.edu/index.php?title=Other-downloads/Pcyn1.0)
595 using the default settings of the *bwa mem* alignment algorithm v0.7.4-r385[60]. Duplicate
596 reads were marked and discarded in subsequent analyses using the "MarkDuplicates"
597 function in PicardTools (http://picard.sourceforge.net). To facilitate comparison across
598 samples of differing coverage, and because coverage of the gDNA samples was much
599 higher (~30X) than for the fDNA samples for LIT and HAP (1.4 and 0.27 respectively),
600 we downsampled the gDNA libraries to 0.73x coverage (the median coverage of
601 samples in Capture 2) using "DownsampleSam" in PicardTools.

602

603 *Comparison sequencing data sets*

604 In several analyses, we compared our capture-based enrichment results to two
605 independent datasets: i) a previously published capture-based enrichment of ancient
606 DNA (aDNA) samples (NCBI SRA accession: SRP042225)[33], and ii) shotgun

607 sequencing from six Capture 1 fDNA samples prior to hybridization ("pre-capture"; Table
608 1). The aDNA samples were aligned to the human genome (*hg38*) and the pre-capture
609 fDNA samples were mapped to the *de novo P. cynocephalus* genome assembly.

610

611 *Library complexity, distribution of reads, and GC content*

612 We calculated the complexity of each library using two methods. First, we used
613 the ENCODE Project's PCR Bottleneck Coefficient (PBC), which calculates the percent
614 of non-duplicate mapped reads out of the total number of mapped reads[61,62].The PBC
615 ranges from 0 to 1, where more complex libraries have higher numbers. Second, we
616 used the function "c_curve" from the program *preseq* (v1.0.2) to plot the number of non-
617 duplicate fragments mapped vs. the number of total mapped fragments[40]. More complex
618 libraries (i.e., those with fewer duplicate fragments) have a c_curve slope closer to 1,
619 meaning that increasing sequencing depth continues to provide novel information. Less
620 complex libraries have a shallower slope and asymptote at smaller values. Lastly, we
621 evaluated the GC bias for each sequencing library using Picard Tools'
622 "CollectGCBiasMetrics" (http://picard.sourceforge.net).

623

624 *Sample attributes influencing capture efficiency*

625 To determine the sample attributes that predicted the success of our capture
626 protocol, we first modeled the relationship between the proportion of non-duplicate
627 reads that mapped to the baboon genome after capture (our primary measure of
628 protocol success) and (i) the percent of endogenous baboon DNA in the pre-capture

629 samples; (ii) the amount of fDNA library (ng) that went into the capture; and (iii) whether
630 the sample was captured using our initial protocol or the second version of the protocol
631 (i.e., in “Capture 1” or “Capture 2”). Second, we investigated the relationship between
632 the same three variables and a secondary measure of protocol success, the fold-
633 change enrichment of baboon DNA in the sample pre- versus post-capture. Pre-capture
634 concentrations of endogenous DNA in fDNA samples were measured as the
635 concentration of baboon DNA estimated using qPCR, relative to the concentration of
636 total DNA estimated using the Qubit High Sensitivity fluorometer (Life Technologies). To
637 ensure that our qPCR-based measures were well calibrated, we confirmed the
638 relationship between qPCR-based estimates and pre-capture sequence-based
639 estimates of endogenous DNA in 6 samples for which both values were available (R^2=
640 0.92; Fig. S12). All statistical analyses were carried out in R[63].

641

642 *Variant calling*

643 We used two different approaches to call variants and genotypes in our sample:
644 SAMTOOLS[64,65] and the Genome Analysis Toolkit (GATK)[66–68]. In downstream
645 analyses, we only retained variants that were identified by both methods, a strategy that
646 produces a higher ratio of true positives to false positives than variants identified by a
647 single method alone[69]. Duplicate-marked alignments were used as input for both
648 methods. SAMTOOLS variant calling was carried out using *mpileup* and *bcftools*, with a
649 maximum allowed read depth (-D) of 100. GATK variant calling was carried out
650 following the GATK Best Practices for GATK v3.0, for variant calling from DNA-seq. To
651 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 re-called 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 ≥100; QD < 2.0; MQ < 35.0; FS > 60.0; HaplotypeScore >13.0; MQRankSum < −12.5; and ReadPosRankSum < −8.0 (as described in Tung, Zhou, et al.[70]). 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 3 additional rounds. Finally, we identified the intersection set between the variants called from GATK and SAMTOOLS, respectively, using the *bcftools* function *isec*[64]. 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*[71].

Estimating 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*[42]. *lcMLkin* uses the genotype likelihoods generated by

675 GATK for each genotype call to calculate two measures: (i) *k*0, the probability that two
676 individuals share no alleles that are identical by descent, and (ii) *r*, the coefficient of
677 relatedness[42]. Several other methods have been developed[72,73] to estimate relatedness
678 from thousands of SNPs, but *lcMLkin* yielded the best match to pedigree-based
679 estimates in our data set (Fig. S13).

680 We also compared genotype calls for the matched fecal and blood-derived
681 samples using GATK's GenotypeConcordance function[68]. This tool allowed us to
682 determine concordance rates between data sets for different classes of variants (e.g., 0,
683 1, or 2). For the majority of variant sites, we expected that the genotypes would be
684 completely concordant (i.e., the same genotype called in the fDNA and gDNA samples
685 from the same individual). However, for calls reported as discordant, we expected that
686 most errors would reflect cases in which the low coverage sample was called as
687 homozygous and the high coverage sample was called as heterozygous, as low read
688 depth makes observation of both alleles at a heterozygous site less likely.

689

690 *WHODAD: Paternity inference and pedigree reconstruction*

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

696 Our approach can be broken into three steps (Fig. S8). First, we estimate, for
697 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 (ii) and (iii) perform subtly different functions in our analysis: (ii) tests that the top candidate is significantly more related to the offspring than any other candidate, whereas (iii) tests that the dyadic similarity between the candidate and the offspring look 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 (Fig. S8).

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 i^{th} 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 i^{th} potential candidate is the father is:

$$P(F_i|M,O) = P(F_i,M,O)/(\sum_{k=1}^{n} P(F_k,M,O)) \quad (1)$$

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:

$$P(F_i|M,O) \approx P(F_i,M,O)^{1/L_i} / (\sum_{k=1}^{n} P(F_k,M,O)^{1/L_k}) \quad (1a)$$

Each joint probability can be calculated in turn as:

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

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:

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

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.,

738 Marshall et al., 1998). Due to genotype uncertainty in low coverage data, the values of
739 $P(F_i|M,O)$ are small. However, the highest value is usually assigned to the most likely
740 father (based on comparison to the pedigree; see Results) and we can directly assess
741 the strength of the relative evidence for the top candidate versus other candidates in
742 Step 2 by calibrating these values against permuted data.

743 *Step 2: calculating resampling-based p-values.* To compute p-values for each
744 paternity assignment, candidates are ranked based on their conditional probability
745 $P(F_i|M,O)$ of being the true father. The log ratio of conditional probabilities between the
746 highest probability father and the second best candidate is the test statistic:

$$r = \log\left(\frac{P(F_{best}|M,O)}{P(F_{second}|M,O)}\right) \quad (4)$$

747 To assess significance for *r*, we then simulate genotype data for a set of *n*
748 unrelated candidate fathers based on allele frequency information for each locus in the
749 analysis and sequence coverage information for the real candidates, at each of the loci
750 for which they were genotyped in the true data set. Specifically, for each locus-
751 simulated unrelated candidate combination, f_{ij}, where *i* indexes a (real) candidate male
752 and *j* indexes the locus, we simulate a vector of genotype probabilities for the candidate
753 father, $(f_{ij0}, f_{ij1}, f_{ij2})$, which sum to 1. The number of probability vectors simulated for
754 each candidate is based on the number and identity of the loci observed in the real
755 data. For example, if the top candidate in the real data was evaluated based on 10,000
756 sites, we would simulate an unrelated male with genotype vector probabilities simulated
757 for each of those 10,000 sites; if the second best candidate was evaluated at 9,000
758 sites, we would simulate an unrelated male with genotype vector probabilities simulated

759 for each of those 9,000 sites; and so on. The variant sets for different simulated
760 candidates need not be identical, and are in fact highly unlikely to be so in practice.

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

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

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

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

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

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

where the expectations are based on the allele frequencies (as above) across all individuals and loci, and across all 3 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 *r* (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[34,35]), 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 non-trivial with low-coverage data, and by itself can also yield incorrect inferences (Fig. S9).

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, 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 high pedigree relatedness (other measures of dyadic relatedness could be substituted, but the *k0* values produced the best correlation with known pedigree-based measures of relatedness in our sample: Fig. S13). 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 (fit using the *mixtools* function in R, with a default value of K=5). 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 from either 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:

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

where for the *i*th 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 fit 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 if the i^{th} 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

843 assignments. To infer $P(z_{bi}=1)$, we use an expectation-maximization algorithm (see
844 Supplementary Methods for detailed information about the EM steps). WHODAD
845 considers a male as the likely true father of a focal offspring if he was (i) the candidate
846 with the highest conditional probability of paternity; (ii) assigned a p-value from our
847 simulations < 0.05; and (iii) $P(z_{bi}=1) > 0.9$.
848
849 *Testing the accuracy of paternity assignment using WHODAD*
850 We assigned paternity using the methods detailed above for all previously
851 identified father-offspring pairs (n = 27) in the Amboseli pedigree (Fig. 1). This pedigree
852 was constructed using a combination of observational life history data on female
853 pregnancies and infant care (to infer maternal-offspring dyads), demographic data to
854 identify possible candidate fathers, and microsatellite genotyping data analyzed in the
855 program CERVUS (with confidence >95%; see Alberts et al.[38] for additional detail).
856 Our data set contained maternal genotype information derived from the fecal
857 enrichment protocol for 15 of these individuals (56%). We first used *WHODAD* to assign
858 paternity for these 15 offspring while incorporating the genotype data from their
859 mothers. To assess the accuracy of *WHODAD* in the absence of maternal genotype
860 data, we then repeated the paternity analysis for the same 15 offspring without including
861 the mother's genotype. For this analysis, we were also able to include the 12 additional
862 offspring for whom we did not have genotype data from the mother, but had genotype
863 data from the known father (n=27).
864 To examine how the presence of close male kin influenced the accuracy and
865 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, half-sibling or full-sibling brothers) from the pool of potential fathers. Second, to test if *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 less than 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 $P(z_{bi}=1) < 0.90$.

880

881 **ACKNOWLEDGEMENTS**

882 We would like to thank the Kenya Wildlife Service, Institute of Primate Research,
883 National Museums of Kenya, National Council for Science and Technology, members of
884 the Amboseli-Longido pastoralist communities, Tortilis Camp, and Ker & Downey
885 Safaris for their assistance in Kenya. We also thank Jeanne Altmann and Elizabeth
886 Archie for their generous support and access to the Amboseli Baboon Research Project
887 data set and samples; Raphael Mututua, Serah Sayialel, Kinyua Warutere, Mercy
888 Akinyi, Tim Wango, and Vivian Oudu for invaluable assistance with the Amboseli
889 baboon sample collection; Emily McLean for assistance in identifying samples from the
890 extended pedigree; and Tauras Vilgalys for assistance in drawing the pedigree. For
891 access to the Guinea baboon samples, we thank Julia Fischer, Dietmar Zinner and José
892 Carlos Brito; the Wild Chimpanzee Foundation for logistical support in Guinea; and the
893 Ministère de l'Environnemnt et de la Protection de la Nature and the Direction des Parcs
894 Nationaux in Senegal, the Opération du Parc National de la Boucle du Baoulé and the
895 Ministère de l'Environnement et de l'Assainissement in Mali, the Office Guinéen de la
896 Diversité Biologique et des Aires Protégées and the Ministère de L'Environnement, des
897 Eaux et Forêts in Guinea, and the Ministère Délégue auprès du Premier Ministre,
898 Chargé de l'Environnement et du Développement Durable in Mauritania. Finally, we
899 thank PJ Perry, Luis Barreiro, Greg Crawford, Tim Reddy, and members of the Alberts
900 and Tung labs for their feedback on earlier versions of this work. This work was
901 supported by National Science Foundation grants DEB-1405308 (to JT) and SMA-
902 1306134 (to JT and NSM). GHK was supported by the German Academic Exchange

903 Service (DAAD), the Christiane-Nüsslein-Volhard Foundation, The Leakey Foundation,
904 and the German Primate Center. XZ was supported by a grant from the Foundation for
905 the National Institutes of Health through the Accelerating Medicines Partnership
906 BOEH15AMP.

907

DATA ACCESSIBILITY:

All resequencing data sets reported in this manuscript will be deposited in the NCBI Short Read Archive (SRA) upon acceptance. A reviewer URL for the metadata and run info for these data sets is available at ftp://ftp-trace.ncbi.nlm.nih.gov/sra/review/SRP064514_20151007_093529_a3f2a910685f5b07f5f45a5fc1fdb389.

AUTHOR CONTRIBUTIONS:

JT, NSM, SM, and XZ conceived and designed the research. MLY, AOS, JBG, GHK, and NSM performed all laboratory experiments. SAS, JT, and JDW provided the genome assembly. WHM, SM and XZ developed the computational methods. WHM and XZ implemented the software. NSM, JT, and XZ analyzed the data. SCA and GHK provided samples, reagents, and logistical support. NSM, JT, and XZ wrote the manuscript with input from all of the coauthors.

COMPETING FINANCIAL INTERESTS:

The authors declare no competing financial interests.

References Cited

1. Höss, M., Kohn, M., Pääbo, S., Knauer, F. & Schröder, W. Excrement analysis by PCR. *Nature* **359,** 199–199 (1992).

2. Ashley, M. & Dow, B. The use of microsatellite analysis in population biology: background, methods and potential applications. *Mol. Ecol. Evol. approaches Appl.* **69,** 185–201 (1994).

3. Clifford, S. L. *et al.* Mitochondrial DNA phylogeography of western lowland gorillas (Gorilla gorilla gorilla). *Mol. Ecol.* **13,** 1551–65, 1567 (2004).

4. Nyakaana, S., Arctander, P. & Siegismund, H. R. Population structure of the African savannah elephant inferred from mitochondrial control region sequences and nuclear microsatellite loci. *Heredity (Edinb).* **89,** 90–8 (2002).

5. Kohn, M. & Knauer, F. Phylogeography of Brown Bears in Europe and Excremental PCR: The New Tool in the Genetic Analysis of Animals in the Wild. *Ursus* **10,** Phylogeography of brown bears in Europe and excrem (1998).

6. Charpentier, M. J. E. *et al.* Genetic structure in a dynamic baboon hybrid zone corroborates behavioural observations in a hybrid population. *Mol. Ecol.* **21,** 715–731 (2012).

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

8. Pérez, T. *et al.* Evidence for improved connectivity between Cantabrian brown bear subpopulations. *Ursus* **21,** 104–108 (2010).

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

10. Smith, K., Alberts, S. C. & Altmann, J. Wild female baboons bias their social behaviour towards paternal half-sisters. *Proc. R. Soc. B Biol. Sci.* **270,** 503 (2003).

11. Archie, E. A. *et al.* Behavioural inbreeding avoidance in wild African elephants. *Mol. Ecol.* **16,** 4138–48 (2007).

12. Gottelli, D., Wang, J., Bashir, S. & Durant, S. M. Genetic analysis reveals promiscuity among female cheetahs. *Proc. R. Soc. B Biol. Sci.* **274,** 1993–2001 (2007).

13. Mondol, S. *et al.* Evaluation of non-invasive genetic sampling methods for estimating tiger population size. *Biol. Conserv.* **142,** 2350–2360 (2009).

14. Nagata, J., Aramilev, V. V., Belozor, A., Sugimoto, T. & McCullough, D. R. 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,** 863–866 (2005).

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

16. Rudnick, J. A., Katzner, T. E., Bragin, E. A. & DeWoody, J. A. 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,** 667–676 (2007).

17. Valière, N. *et al.* Long-distance wolf recolonization of France and Switzerland inferred from non-invasive genetic sampling over a period of 10 years. *Anim. Conserv.* **6,** 83–92 (2003).

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

19. Gagneux, P., Boesch, C. & Woodruff, D. S. Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. *Mol. Ecol.* **6,** 861–8 (1997).

20. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* (2011).

21. Sabeti, P., Reich, D. & Higgins, J. Detecting recent positive selection in the human genome from haplotype structure. *Nature* (2002).

22. Price, A., Tandon, A., Patterson, N. & Barnes, K. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* (2009).

23. Sankararaman, S. & Sridhar, S. Estimating local ancestry in admixed populations. *Am. J. …* (2008).

24. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43,** 519–25 (2011).

25. Ma, Y. *et al.* Accurate inference of local phased ancestry of modern admixed populations. *Sci. Rep.* **4,** 5800 (2014).

26. Huang, B., Amos, C. & Lin, D. Detecting haplotype effects in genomewide association studies. *Genet. Epidemiol.* (2007).

27. Li, Y., Sung, W. & Liu, J. Association mapping via regularized regression analysis of single-nucleotide–polymorphism haplotypes in variable-sized sliding windows. *Am. J. Hum. Genet.* (2007).

28. Purcell, S. *et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses*. *Am. J. Hum. Genet.* **81,** (2007).

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

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

1034 31. Perry, G. H., Marioni, J. C., Melsted, P. & Gilad, Y. Genomic-scale capture and
1035 sequencing of endogenous DNA from feces. *Mol. Ecol.* **19,** 5332–44 (2010).
1036

1037 32. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for
1038 massively parallel targeted sequencing. *Nat. Biotechnol.* **27,** 182–189 (2009).
1039

1040 33. Carpenter, M. L. *et al.* Pulling out the 1%: whole-genome capture for the targeted
1041 enrichment of ancient DNA sequencing libraries. *Am. J. Hum. Genet.* **93,** 852–64
1042 (2013).
1043

1044 34. Kalinowski, S. T., Taper, M. L. & Marshall, T. C. Revising how the computer
1045 program CERVUS accommodates genotyping error increases success in
1046 paternity assignment. *Mol. Ecol.* **16,** 1099–1106 (2007).
1047

1048 35. Marshall, T. C., Slate, J., B., K. L. E. & Pemberton, J. M. Statistical confidence for
1049 likelihood-based paternity inference in natural populations. *Mol. Ecol.* **7,** 639–655
1050 (1998).
1051

1052 36. Chakraborty, R., Shaw, M. & Schull, W. J. Exclusion of paternity: the current state
1053 of the art. *Am. J. Hum. Genet.* **26,** 477–88 (1974).
1054

1055 37. Alberts, S. C. & Altmann, J. in *Long-term F. Stud. primates* (Kappeler, P. M. &
1056 Watts, D. P.) 261–287 (Springer, 2012).
1057

1058 38. Alberts, S. C., Buchan, J. C. & Altmann, J. Sexual selection in wild baboons: from
1059 mating opportunities to paternity success. *Anim. Behav.* **72,** 1177–1196 (2006).
1060

1061 39. Morin, P. A., Chambers, K. E. K., Boesch, C. & Vigilant, L. Quantitative
1062 polymerase chain reaction analysis of DNA from noninvasive samples for
1063 accurate microsatellite genotyping of wild chimpanzees (Pan troglodytes verus).
1064 *Mol. Ecol.* **10,** 1835–1844 (2001).
1065

1066 40. Daley, T. & Smith, A. D. A. Predicting the molecular complexity of sequencing
1067 libraries. *Nat. Methods* **10,** 325–329 (2013).
1068

1069 41. Samuels, D. C. *et al.* Finding the lost treasures in exome sequencing data. *Trends*

Genet. **29,** 593–9 (2013).

42. Lipatov, M., Sanjeev, K., Patro, R. & Veeramah, K. *Maximum Likelihood Estimation of Biological Relatedness from Low Coverage Sequencing Data*. *bioRxiv* (Cold Spring Harbor Labs Journals, 2015). doi:10.1101/023374

43. Sinnwell, J. P., Therneau, T. M. & Schaid, D. J. The kinship2 R package for pedigree data. *Hum. Hered.* **78,** 91–3 (2014).

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

45. Olsen, J. B., Busack, C., Britt, J. & Bentzen, P. The aunt and uncle effect: an empirical evaluation of the confounding influence of full sibs of parents on pedigree reconstruction. *J. Hered.* **92,** 243–7 (2001).

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

47. Ávila-Arcos, M. C. *et al.* Comparative Performance of Two Whole Genome Capture Methodologies on Ancient DNA Illumina Libraries. *Methods Ecol. Evol.* n/a–n/a (2015). doi:10.1111/2041-210X.12353

48. Enk, J. M. *et al.* Ancient whole genome enrichment using baits built from modern DNA. *Mol. Biol. Evol.* **31,** 1292–4 (2014).

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

50. George, R. D. *et al.* Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection. *Genome Res.* **21,** 1686–94 (2011).

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

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

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

54. Gillespie, T. R. Noninvasive Assessment of Gastrointestinal Parasite Infections in Free-Ranging Primates. *Int. J. Primatol.* **27,** 1129–1143 (2006).

55. Knight, J. M. *et al.* Non-invasive analysis of intestinal development in preterm and term infants using RNA-Sequencing. *Sci. Rep.* **4,** 5453 (2014).

56. Moll, P. R., Duschl, J. & Richter, K. Optimized RNA amplification using T7-RNA-polymerase based in vitro transcription. *Anal. Biochem.* **334,** 164–74 (2004).

57. Shagina, I. *et al.* Normalization of genomic DNA using duplex-specific nuclease. *Biotechniques* **48,** 455–9 (2010).

58. Roeder, A. & Archer, F. A novel method for collection and preservation of faeces for genetic studies. *Mol. Ecol. …* **4,** 761–764 (2004).

59. Nsubuga, A. M. *et al.* Factors affecting the amount of genomic DNA extracted from ape faeces and the identification of an improved sample storage method. *Mol. Ecol.* **13,** 2089–94 (2004).

60. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).

61. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22,** 1813–31 (2012).

62. Kharchenko, P. V, Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* **26,** 1351–9 (2008).

63. R Development Core Team. R: A Language and Environment for Statistical Computing. (2015).

64. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–9 (2009).

65. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–93 (2011).

66. Van der Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr. Protoc. Bioinforma.* 11.10.1–11.10.33 (2013). doi:10.1002/0471250953.bi1110s43

67. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–303 (2010).

68. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).

69. O'Rawe, J. *et al.* Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.* **5,** 28 (2013).

70. Tung, J., Zhou, X., Alberts, S. C., Stephens, M. & Gilad, Y. The genetic architecture of gene expression levels in wild baboons. *Elife* **4,** e04729 (2015).

71. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–8 (2011).

72. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42,** 565–9 (2010).

73. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26,** 2867–73 (2010).