FecalSeq: methylation-based enrichment for noninvasive population genomics from feces

Kenneth L. Chiou[1,*] and Christina M. Bergey[2,3,4,*]

[1]Department of Anthropology, Washington University, St. Louis, MO 63130, USA
[2]Department of Anthropology, New York University, New York, NY 10003, USA
[3]New York Consortium in Evolutionary Primatology, New York, NY, USA
[4]Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556, USA

November 25, 2015

Abstract

We developed an inexpensive methylation-based capture method for enriching host DNA from noninvasively obtained fecal samples. We demonstrate that the enrichment is robust, efficient, and compatible with downstream library preparation methods, including complexity-reduction approaches for massively parallel sequencing. Because feces are widely available and convenient to collect, our method empowers researchers to explore genomic-scale population-level questions in organisms for which invasive sampling is challenging or undesirable.

The past decade has witnessed a rapid transformation of biological studies with the continuing development and application of massively parallel sequencing technology. This genomic-scale revolution, however, has had a relatively muted effect on studies of free-ranging nonmodel organisms due largely to the difficulty of obtaining high-quality samples. Feces provide a convenient and widely available source of host genetic information[1], but a preponderance of microbial DNA has thus far hindered the adoption of massively parallel sequencing for noninvasive studies.

Prior efforts have employed targeted sequence capture methods for genomic-scale enrichment of host DNA (hDNA) from fecal DNA (fDNA)[2,3]. In this approach, DNA or RNA baits are selectively hybridized to target DNA fragments based on sequence complementarity, then isolated by magnet[4] or on an array[5,6]. This approach, however,

* These authors contributed equally to this work
Correspondence should be addressed to K.L.C. (kenneth.chiou@wustl.edu) and C.M.B. (cbergey@nd.edu).

has distinct drawbacks. Oligonucleotide baits are costly to synthesize. Custom transcription of RNA baits[3,7,8] reduces costs, but is labor-intensive, time-consuming, and requires high-quality genomic DNA that is consumed by the procedure. Capture biases are introduced by sequence differences among samples and, because the composition of transcribed baits inevitably varies across bait sets, particularly when different genomic DNA templates are used, hybridization may be inconsistent, promoting allelic dropout among samples.

In this study, we introduce FecalSeq, a sequence-independent method for enriching hDNA from feces with subsequent library preparation for massively parallel sequencing. In a modification of a previous application[9], FecalSeq makes use of a methyl-binding-domain (MBD) protein to selectively bind DNA fragments with a relatively high proportion of methylated CpG dinucleotides. Because microbes generally have low CpG methylation relative to animals[10], this enables partitioning of hDNA from microbial DNA.

To evaluate our approach, we enriched and prepared double-digest RADseq (ddRADseq)[11] libraries from the feces of six captive and eight wild baboons. We also prepared ddRADseq libraries from blood-derived genomic DNA of the six captive baboons to facilitate intra-individual comparisons of blood and fecal libraries. All libraries were sequenced on the Illumina MiSeq platform.

Quantitative PCR estimates of starting hDNA proportions in fDNA extracts ranged widely, but were generally lower in samples obtained from the wild (captive samples: mean 5.9%, range <0.01%-17.4%; wild samples: mean 1.6%, range <0.01%-4.9%; **Supplementary Table 1**).

Based on two pilot libraries constructed from MBD-enriched fDNA, we found that there was large variation in the proportion of reads mapping to the baboon reference genome (mean 24.2%, range 0.7%-79.3%; **Supplementary Fig. 1** and **Supplementary Table 1**), with the read mapping proportion correlating highly with starting hDNA proportions (library A: R^2=0.9778, library B: R^2=0.9074; **Supplementary Fig. 2**). Following a series of protocol optimization experiments (**Supplementary Table 2**), we created and sequenced a third library from MBD-enriched fDNA, incorporating several protocol improvements (**Supplementary Protocol 1**).

Despite having similar or even lower starting hDNA proportions, read mapping proportions in the third library were substantially higher than the prior two (mean 48.2%, range 8.7%-73.9%; **Supplementary Fig. 1** and **Supplementary Table 1**).

MBD binding may in principle select for genomic regions with relatively high methylation density, leading to dropout of other loci. Cumulative coverage plots for feces- and blood-derived libraries, however, exhibited no discernible differences (**Supplementary Fig. 3**), indicating similar recovery rates of restriction-site-flanked fragments ("RADtags") in MBD-enriched libraries.

Assessment of the concordance between blood- and feces-derived reads from the same individual was complicated by the correlation in ddRADseq between total reads and expected RADtags recovered and thereby SNPs discovered: a given RADtag is sequenced at a frequency inversely proportional to the deviation of its length from the mean of the size selection. Thus, we were tasked with discerning between dropout due to coverage-related stochasticity inherent in ddRADseq[11] and that due to MBD enrichment. To perform this comparison, we computed distance, measured in differing alleles, between blood- and feces-derived reads from the same individual, with samples

downsampled as necessary to equalize total coverage among same-individual samples. Allelic dropout due to MBD enrichment would result in a higher proportion of alleles unique to blood-derived libraries relative to feces-derived libraries. We did not find a significant discrepancy (multi-sample-called SNPs: mean proportion unique alleles in blood = 3.2%%, mean proportion unique alleles in feces = 3.2%; Wilcoxon signed rank test, $p = 0.92$; **Supplementary Fig. 7**).

Dropout of entire RADtags is easily detectable given a reference genome or sufficient samples for comparison; dropout of a single allele at heterozygous sites is a more insidious potential bias. Allelic dropout due to MBD enrichment would result in a decrease in heterozygosity in MBD-enriched fecal libraries. Inbreeding coefficients from same-individual blood- and feces-derived libraries, however, exhibited no significant difference (Wilcoxon rank sum test, $p = 0.60$, **Supplementary Fig. 8**), indicating low allelic dropout attributable to the MBD enrichment.

Stringent filtration of SNP sets, as would be implemented in a standard population genetic study, reduced the apparent biases attributable to fecal enrichment, measured both as total SNPs with a significant association with sample type (unfiltered: 12,935 out of 336,295, or 3.8%; filtered: 33 out of 4,099, or 0.8%) as well as total SNPs with significant missingness assessed via a chi-square test (unfiltered: 40,850 out of 323,840, or 12.6%; filtered: 0 out of 2,672, or 0%). Though more work is needed to carefully quantify the extent and causal factors that lead to missingness, many population genetic analyses are robust to the low level of dropout our analyses reveal in addition to that which is inherent in the RADseq family of techniques[12].

Feces are among the most readily accessible sources of information on wild animals[1], and are particularly useful for population-level studies or studies of endangered or elusive species for which obtaining high-quality samples is difficult or undesirable. By exploiting methylation differences rather than sequence differences between host and bacterial DNA, FecalSeq is a *de novo* enrichment strategy that requires neither prior genome sequence knowledge nor the use of high-quality DNA for preparation of capture baits. This results in enrichment which is both inexpensive—we estimate a per-sample enrichment cost of $0.70 (**Supplementary Note 1**)—and replicable. Importantly, FecalSeq is to our knowledge the first genomic-scale fecal hDNA enrichment method that is compatible with most downstream library preparation methods for massively parallel sequencing. Through our use of ddRADseq, we demonstrate that our method facilitates low-cost high-capacity genotyping of wild populations without introducing significant bias. Further, because ddRADseq is customizable[11], there is substantial flexibility for researchers to optimize the number of samples and the fraction of the genome sequenced for particular research questions.

We robustly found that sequencing efficiency of MBD-enriched fDNA libraries correlates strongly with starting proportions of hDNA. Future attention should therefore be directed towards fecal sample collection, storage, and extraction methods that maximize the selective recovery of host nuclear DNA[13]. We have preliminarily found, however, that incorporating a second, sequential enrichment procedure may substantially improve the overall enrichment of samples with low proportions of hDNA (**Supplementary Fig. 9**).

Because MBD enrichment partitions DNA based on CpG-methylation density, FecalSeq does not enrich hypomethylated host mitochondrial DNA[14] but may co-enrich nuclear DNA from exogenous eukaryotes, such as from plant or animal digesta. Care

should therefore be taken to minimize the presence of exogenous eukaryotic tissues or cells, although the degree to which this is a problem in practice is currently unknown.

Since PCR amplification of DNA from feces was first achieved in the 1990s[15–17], noninvasive genetic studies have revolutionized our understanding of the evolution, ecology, and behavior of nonmodel organisms. By facilitating low-cost genomic-scale sequencing from feces, our method connects a community of field researchers with the benefits of massively parallel sequencing, ushering noninvasive organism studies into the genomic age.

Methods

Samples. Blood and fecal samples were collected from six captive baboons (genus *Papio*) housed at the Southwest National Primate Research Center (SNPRC) at the Texas Biomedical Research Institute. The individuals were of either *P. anubis* or hybrid ancestry (**Supplementary Table 1**). All six baboons were fed a diet manufactured by Purina LabDiet ("Monkey Diet 15%") containing 15% minimum crude protein, 4% minimum crude fat, and 10% maximum crude fiber. In separate sedation events, blood and feces were collected from the same individual who was isolated for the duration of the sedation. Following centrifugation, the buffy coat was isolated from blood samples and stored at -80°C. 2 ml of feces were also collected into 8 ml tubes containing 4 ml of RNALater (Ambion). All procedures were conducted under the Texas Biomedical Research Institute IACUC protocol #1403 PC 0. Sedation and blood draws were performed under the supervision of a veterinarian and animals were returned immediately to their enclosures following recovery.

In addition, we collected or obtained fecal samples from eight wild baboons in Zambia. Four samples were collected in 2006 from the Luangwa Valley, Zambia. Four samples were collected in 2014 from Kafue National Park, Zambia. The Luangwa Valley baboons are of hybrid *P kindae* x *P. cynocephalus* ancestry while the Kafue National Park baboons are of hybrid *P. kindae* x *P. ursinus* ancestry. As with the SNPRC samples, 2 ml of feces were collected into 8 ml tubes containing 4 ml of RNALater. In contrast to the SNPRC samples, however, these samples were collected noninvasively from unhabituated animals in remote field conditions. Samples therefore could not be attributed to particular animals, although these eight samples were chosen from distinct groups and locations to avoid duplication. Following collection, samples were stored without refrigeration for 1-6 months before being frozen at -20°C for long-term storage.

Buffy coat extractions were performed using the QIAamp DNA Blood Mini Kit (Qiagen), following manufacturer's instructions. Fecal extractions were performed using the QIAamp DNA Stool Mini Kit (Qiagen) following manufacturer's instructions for optimizing host DNA yields. DNA concentration and yield were measured using a Qubit dsDNA BR Assay (Life Technologies).

We estimated the proportion of host DNA (hDNA) for each fecal DNA (fDNA) sample using quantitative PCR (qPCR) by comparing estimates of hDNA concentration obtained by qPCR to estimates of total fDNA concentration obtained by Qubit. Amplification was conducted using universal mammalian *c-myc* primers[18] and evaluated against a standard curve constructed from the liver DNA of an individual baboon. Samples and standards were run in duplicate alongside positive and negative controls (see **Supplementary Protocol 1**: Auxiliary protocol B for full details).

DNA Enrichment. DNA was enriched using the NEBNext Microbiome DNA Enrichment Kit (New England BioLabs). This enrichment procedure[9] captures eukaryotic DNA using a methylated CpG-specific binding domain protein fused to the Fc fragment of human IgG (MBD2-Fc) to selectively target sequences with high CpG methylation density.

MBD2-Fc-bound magnetic beads were prepared according to manufacturer instructions in batches ranging from 40 to 160 µl. For each n µl batch, we prebound 0.1 x n µl MBD2-Fc protein to n µl protein A magnetic beads by incubating the mixture with rotation for 10 min at room temperature. The bound MBD2-Fc magnetic beads were then collected by magnet and washed twice with 1 ml ice-cold 1x bind/wash buffer before being resuspended in n µl ice-cold 1x bind/wash buffer.

As a pilot experiment, we prepared two successive libraries, library A and library B, following manufacturer's instructions for capturing methylated hDNA, with minor protocol modifications incorporated for the second pilot library (library B). Library A included MBD-enriched fDNA from 4 SNPRC baboons and 2 Luangwa Valley baboons, as well as blood DNA from the same SNPRC baboons. Library B included MBD-enriched fDNA from 4 SNPRC baboons (with two repeats from library A), 4 Kafue National Park baboons, and 2 Luangwa Valley baboons, as well as blood DNA from 2 SNPRC baboons. For each fDNA sample, we combined 1-2 µg of extracted fDNA with 160 µl of prepared protein-bound beads and a variable volume of ice-cold 5x bind/wash buffer for maintaining 1x concentration of bind/wash buffer. After combining beads and DNA, we incubated the mixture at room temperature with rotation for 15 min. DNA and MBD2-Fc-bound magnetic beads were then collected by magnet and the supernatant removed. For library A, we washed the collected beads with 1 ml of ice-cold 1x bind/wash buffer. For library B, we conducted three expanded wash steps to maximize the removal of unbound DNA. For each wash in library B, we added 1 ml of ice-cold 1x bind/wash buffer and mixed the beads on a rotating mixer for three minutes at room temperature before collecting the beads by magnet and removing the supernatant. Following the final wash, we resuspended and incubated the beads at 65°C with 150 µL of 1x TE buffer and 15 µL of Proteinase K for 20 min with occasional mixing. The eluted DNA was then separated by magnet, purified with 1.5x homemade SPRI beads[19], and quantified using a Qubit dsDNA HS Assay (Life Technologies).

Our pilot sequencing results from libraries A and B revealed large variation in the percentage of reads mapping to the baboon genome, with mapping percentages ranging from 1.1% to 79.3%, with much of the variation correlating with the proportion of hDNA in the unenriched fDNA sample (**Supplementary Fig. 2**). To expand the utility of the enrichment protocol to all fDNA samples, we conducted a series of capture experiments designed to optimize the enrichment of hDNA from "low-quality" samples (i.e., samples with low proportions of hDNA). For these experiments, we artificially simulated fDNA by combining high-quality baboon liver or blood genomic DNA (liver: SNPRC ID #19334; blood: SNPRC ID# 14068 or 25567) with *E. coli* DNA (K12 or ATCC 11303 strains) at controlled proportions. The resulting post-enrichment proportion of baboon and *E. coli* DNA was evaluated by qPCR in two analyses using (1) universal mammalian *c-myc*[18] and (2) universal bacterial *16s*[20] primers along with standards created from the same respective organisms (experiments and results are described in detail in **Supplementary Table 2**).

We prepared a final library, library C, incorporating modifications (**Supplementary**

Protocol 1) based on results from our capture optimization experiments. For these captures, we added a much smaller volume of prepared MBD2-Fc-bound magnetic beads (1-22 µl) based on the estimated proportion of starting hDNA, kept the capture reaction volume consistent at a relatively low 40 µl (concentrating samples as needed using a SPRI bead cleanup), added an extra wash step in which samples were resuspended in 100 µl of 1x bind/wash buffer then incubated at room temperature for 3 minutes with rotation, and eluted samples in 100 µl 2 M NaCl. For four fDNA samples, we serially enriched the samples by repeating the capture reaction with 30 µl of MBD-enriched DNA (post SPRI-bead cleanup). Library C included fDNA from 5 SNPRC baboons, 2 Kafue National Park baboons, and 1 Luangwa Valley baboon. At least one fDNA sample from all three locations was both singly and doubly enriched to facilitate comparison. The composition of libraries A-C are described in detail in **Supplementary Table 1**.

Library Preparation and Sequencing. Library preparation followed standard double-digest restriction site-associated DNA sequencing (ddRADseq) procedures[11] with modifications to accommodate low input as described below.

For all samples, including blood DNA and MBD-enriched fDNA, we digested DNA with *Sph*I and *Mlu*CI. Enzymes were diluted using compatible diluents to facilitate pipetting of small quantities. As the total amount of post-enrichment fDNA was low, we adjusted adapter concentrations in the ligation reaction to ~0.1 µM for barcoded P1 and ~3 µM for P2, which correspond to excesses of adapters between 1-2 orders of magnitude. Since adapter-ligated samples are multiplexed into pools in equimolar amounts, we made efforts to combine samples with similar concentrations and enrichment when known. We used the BluePippin (Sage Science) with a 1.5% agarose gel cassette for automated size selection of pooled individuals, with a target of 300 bp (including adapters) and extraction of a "tight" collection range. For PCR amplification, we ran all reactions in quadruplicate to minimize PCR biases and attempted to limit the number of PCR cycles. As the concentration of post-size-selection pools was below the limits of detection without loss of a considerable fraction of the sample, estimation of the required number of PCR cycles was difficult. We therefore iteratively quantified products post-PCR and added cycles as necessary. The total number of PCR cycles per pool is reported in **Supplementary Table 1**, but was usually 24. Finally, libraries were sequenced in single runs of the Illumina MiSeq using 2x150 cycles and 30% spike-in of PhiX control DNA.

Analysis. We demultiplexed reads by sample and mapped them to the baboon reference genome (papAnu2; Baylor College of Medicine Human Genome Sequencing Center) using BWA with default parameters[21]. For every pair of blood and fecal samples from the same individual, we downsampled mapped reads to create new pairs with equal coverage in order to control for biases due to differences in sequencing depth. After realignment around indels, we identified variants using GATK's Unified Genotyper[22], in parallel analyses (1) calling variants in all samples at once and (2) processing each sample in isolation to avoid biasing variant calls from other samples at the expense of accuracy. Homozygous sites matching the reference genome were listed as missing when variants were inferred in single individuals. Variants were filtered with GATK (VariantFiltration: QD < 2.0, MQ < 40.0, FS > 60.0, HaplotypeScore > 13.0, MQRankSum $<$ -12.5, ReadPosRankSum $<$ -8.0) and indels were excluded.

We digested the baboon reference genome *in silico*, tallied reads within each predicted RADtag, and gathered the following information about each region: length, GC percentage, and CpG count in region ± 5 kb. We also calculated read depth in these simulated RADtags. Coverage plots (**Supplementary Fig. 3**) and distributions of blood and fecal RADtags' length, GC percentage, and local CpG density (**Supplementary Fig. 10** and **Supplementary Fig. 11**) were visually inspected for gross distortion due to widespread dropout.

If the fecal enrichment procedure caused widespread allelic dropout, the proportion of alleles unique to the blood samples would be higher than that to the fecal sample. We tallied these unique alleles with VCFtools[23] and tested for an excess with a Wilcoxon signed rank test.

To quantify loss of heterozygosity due to allelic dropout, we computed the inbreeding coefficient, F for all blood-feces pairs with equalized coverage, using both the individually called and multi-sample called SNP sets. The presence of dropout is expected to inflate F. We tested for differences in paired samples' estimates of F via a Wilcoxon rank sum test. The dataset is not filtered for missingness, so sequencing errors inferred to be true variants may inflate heterozygosity estimates, thus deflating F.

To create a stringently filtered dataset with high genotyping rate, we filtered the multi-sample called SNPs in PLINK[24], retaining only those genotyped in at least 90% of samples and removing samples with genotypes at fewer than 10% of sites. This filtered set was further pruned for linkage disequilibrium by sliding a window of 50 SNPs across the chromosome and removing one random SNP in each pair with $r^2 > 0.5$. Using all samples, we performed multidimensional scaling to visualize identity by state (IBS). Using just the samples that were part of the same-individual blood-feces pairs, we then performed an association test and missingness chi-square test to detect allele frequencies or missingness that correlated with sample type. We did the same with the unfiltered dataset as well. Though we had few pairs of fecal samples from the same individual, we computed distance between pairs of samples from the same individual using the stringently filtered dataset (**Supplementary Fig. 6**) to compare distance between and within sample types via a Wilcoxon rank sum test.

All code generated for this project can be accessed at https://github.com/bergeycm/RAD-faex.

Acknowledgments

We thank C. Jolly and J. Phillips-Conroy for providing fecal samples from the Luangwa Valley, Zambia. We thank the Zambia Wildlife Authority and the University of Zambia for granting permission and providing support for fieldwork. We thank E. Yigit, A. Burrell, and T. Disotell for helpful conversations. This study was funded by the National Science Foundation (BCS 1341018, BCS 1260816, BCS 1029302, SMA 1338524), the Leakey Foundation, the Wenner-Gren Foundation, and the NYU University Research Challenge Fund. The Genome Technology Center at NYU is supported by NIH/NCATS UL1 TR00038 and NIH/NCI P30 CA016087. K.L.C. is supported by NSF fellowship DGE 1143954.

Author Contributions

K.L.C. and C.M.B. conceived the project and wrote the paper. K.L.C. collected samples and led the labwork. C.M.B. led the analysis.

Competing Interests Statement

New England Biolabs, the commercial vendor of the enrichment reagents, supplied three enrichment kits (retail value of $210 each) and 30 µg of purified K-12 *E. coli* genomic DNA at no cost for use in this project. K.L.C. and C.M.B. are not affiliated with New England Biolabs and declare no other competing interests.

References

1. Kohn, M. H. & Wayne, R. K. *Trends Ecol Evol* **12,** 223–227 (1997).

2. Perry, G. H., Marioni, J. C., Melsted, P. & Gilad, Y. *Mol Ecol* **19,** 5332–5344 (2010).

3. Snyder-Mackler, N. *et al.* Preprint at http://biorxiv.org/content/early/2015/10/21/029520 (2015).

4. Gnirke, A. *et al. Nat Biotechnol* **27,** 182–189 (2009).

5. Okou, D. T. *et al. Nat Methods* **4,** 907–909 (2007).

6. Albert, T. J. *et al. Nat Methods* **4,** 903–905 (2007).

7. Melnikov, A. *et al. Genome Biol* **12,** R73 (2011).

8. Carpenter, M. L. *et al. Am J Hum Genet* **93,** 852–864 (2013).

9. Feehery, G. R. *et al. PLoS One* **8,** e76096 (2013).

10. Zemach, A., McDaniel, I. E., Silva, P. & Zilberman, D. *Science* **328,** 916–919 (2010).

11. Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S. & Hoekstra, H. E. *PLoS One* **7,** e37135 (2012).

12. Gautier, M. *et al. Mol Ecol* **22,** 3165–3178 (2013).

13. Ramón-Laca, A., Soriano, L., Gleeson, D. & Godoy, J. A. *Wildlife Biology* **21,** 195–203 (2015).

14. Rebelo, A. P., Williams, S. L. & Moraes, C. T. *Nucleic Acids Res* **37,** 6701–6715 (2009).

15. Höss, M., Kohn, M., Pääbo, S., Knauer, F. & Schröder, W. *Nature* **359,** 199 (1992).

16. Constable, J. J., Packer, C., Collins, D. A. & Pusey, A. E. *Nature* **373,** 393–393 (1995).

17. Gerloff, U. *et al. Mol Ecol* **4,** 515–518 (1995).

18. Morin, P. A., Chambers, K. E., Boesch, C. & Vigilant, L. *Mol Ecol* **10,** 1835–1844 (2001).

19. Rohland, N. & Reich, D. *Genome Res* **22,** 939–946 (2012).

20. Corless, C. E. *et al. J Clin Microbiol* **38,** 1747–1752 (2000).

21. Li, H. & Durbin, R. *Bioinformatics* **25,** 1754–1760 (2009).

22. DePristo, M. A. *et al. Nat Genet* **43,** 491–498 (2011).

23. Danecek, P. *et al. Bioinformatics* **27,** 2156–2158 (2011).

24. Purcell, S. *et al. Am J Hum Genet* **81,** 559–575 (2007).