RADseq Works in Primates, Dammit.

Christina M. Bergey Luca Pozzi Todd R. Disotell
Andrew S. Burrell

December 12, 2012

Abstract

. . . (This is just a bit of the blurb from the email.) Our paper is an introduction to a 2nd generation sequencing technique for typing thousands of genome-wide markers from non-model organisms. We demonstrate it in five primates and discuss its promise for doing mutlilocus phylogenetics and population genetics in primates. . . .

1 Introduction

During the first decades of molecular primatology, a formidable impediment to researchers was the need to develop and type polymorphic markers in a taxon of interest. The markers that resulted from this tedious and expensive task were often uninformative when applied outside of the population used in their design, necessitating further rounds of primer design or microsatellite assays, for example (Davey et al., 2011). Due to the bottleneck of marker discovery, many population genetic or phylogenetic studies in molecular primatology have been based on one or few loci (Ting and Sterner, published yet?). Such inferences can reliably give the evolutionary history of those particular regions of the genome, but they fail to adequately capture the complete complex history of the population given the mosaic nature of genomic evolution (some gene-tree species tree review ref). Adequate resolution depends on high marker density, and until recently that goal has been out of reach for many primate researchers.

The rapidly decreasing costs of DNA sequencing technology have promised revolutionary gains for primatology (Enard and Paabo 2004, maybe? Good-

man et al 2005?). A primate researcher benefits from the many nearby sequenced and assembled reference genomes in the order, but genomic studies of non-model organisms nevertheless remain difficult. Though the cost of low coverage whole genome sequencing has fallen to a level feasible for many researchers' budgets, sequencing whole genomes for the tens or hundreds of individuals desired in a typical population genetic study is often prohibitively expensive and quite possibly superfluous.

Fortunately, researchers have recently developed techniques that reduce the complexity of the genome and allow discovery and typing of thousands or tens of thousands of genome-wide makers in many individuals in a single step. RAD-seq is one such simple, inexpensive reduced representation technique which allows for the sequencing of small fragments of the genome adjacent to restriction enzyme cut sites (Baird et al., 2008). These RAD tags, or restriction-site association DNA tags, were originally developed for use in microarray hybridization typing (Miller et al., 2007), but an updated protocol substitutes second-generation DNA sequencing to rapidly discover and type SNPs (Baird et al., 2008). The lack of reliance on a reference genome and applicability to datasets of many individuals make it a promising technique for phylogenetic or population genetic studies in non-model organisms, such as many primates.

In the present study, we summarize the RAD-seq method in brief and note the many and varied applications since its development. We go on to discuss reasons that make RAD-seq promising for primatologists, and then demonstrate the technique in 5 primates: a lemur, New World monkey, Old World monkey, and two apes.

1.1 The RAD-seq Technique

The following is a summary of the RAD tag library preparation of Etter et al., 2011 (Fig. 1). The RAD-seq library preparation begins when genomic DNA from multiple individuals is digested with a restriction enzyme, such as *EcoRI* or *PspXI*. The P1 adapter is then ligated to the fragments, connected to the sticky end at the restriction enzyme cut site. The P1 adapter contains an amplification site for PCR, an Illumina sequencing priming site, and an individual-specific barcode of five basepairs. Once the barcode has been added, fragments from multiple individuals can be pooled, and the DNA is randomly sheared with a sonicator to have a length distribution under 1 kilobase. To select for reads that are suitable for sequencing on the Illumina

platform, the sheared samples are size selected via agarose gel electrophoresis, extracting fragments between 300 and 500 bp in length. The second adapter, P2, is a Y adapter meaning its two halves are complementary for only part of their length. It is ligated to the fragments and then the fragments are amplified via PCR. Because the second adapter has divergent ends, the reverse amplification primer is unable to bind until after the forward amplification primer has filled in its complementary sequence. This ensures that only RAD tags ligated to P1 are able to amplify. After few cycles of PCR to minimize the risk of introducing PCR artifacts or biases, the library is ready for final clean-up, quality control, and sequencing.

1.2 Previous RAD-seq Studies

[Previous RADseq studies. Focus on pop gen, phylogeny studies]

1.3 Prospective Primate Applications

[Why it would be good in primates.]

Changing the restriction enzyme to a more or less frequent cutter allows fine grained control over the number of loci, number of individuals multiplexed, and depth of coverage, allowing researchers to tailor marker discovery to their hypothesis.

2 Methods

Library Preparation and Sequencing We digested genomic DNA from 5 primates with *PspXI* (New England Biolabs) and used it to create a multiplexed RAD tag library. Our library preparation method followed that of Etter et al, 2011 with the following modifications: the P1 adapter top(?) oligonucleotide was modified to have an overhang corresponding to the cut site of *PspXI*, and a longer P2 adapter suitable for paired end sequencing was used (P2_top: 5'-SEQUENCEHERE-3'; P2_bottom: 5'-SEQUENCEHERE-3'). Individual-specific barcodes contained in the P1 adapter differed by at least three nucleotides. We chose *PspXI* based on the results of *in silico* digestion of the human, rhesus macaque, and baboon reference genomes using custom Perl scripts (refs). We sequenced the prepared library as one

150-cycle paired-end run of an Illumina MiSeq at the NYU Langone Medical Center's Genome Technology Center using a spike-in of 30% PhiX DNA to control for low diversity in the library at the barcode and restriction sites. Other individuals were sequenced alongside those of the present study. Sequences are available to download from the NCBI Short Read Archive (accession number SRAXXXXXX.X).

Sequence Analysis - Clustering and SNP Discovery We demultiplexed, or separated by barcode, sequence reads and excluded from the analysis reads without an expected barcode or an intact restriction enzyme cut site. We also removed reads with any quality scores below 10. Using the program Stacks, we clustered all reads into matching sets and compared closely related sets to detect orthologous loci and SNPs using a maximum likelihood approach (Stacks ref). We tallied orthologous SNPs using VCFtools (ref).

[Phylogeny?]

Sequence Analysis - Assess RAD Tag Coverage To assess the RAD tag coverage, we mapped human and chimpanzee reads to the human reference genome. Again, we excluded reads without an expected barcode or an intact restriction enzyme cut site. We aligned reads to the human reference genome (hg19, ref) using BWA with default parameters. We removed reads that were unmapped or that had low mapping quality after alignment using Picard (ref) and BamTools (ref).

Cut this paragraph? After performing local realignment around indels with GATK (ref), we identified SNPs and short indels using SAMtools mpileup and BCFtools (ref). We required a minimum coverage of 3 reads and a maximum of 100 to call a SNP or an indel at a given location. We tallied orthologous SNPs using VCFtools (ref).

To assess how many restriction sites were successfully sequenced and to analyze the degree of overlap between multiplexed individual's datasets, we first found all possible *PspXI* cut sites in the human genome using the oligoMatch utility in the USCS Genome Browser program (ref). This allowed us to calculate the coverage of these restriction site-associated regions using BEDtools' multiBamCov program (ref).

3 Results

___ million sequencing reads with an intact barcode and restriction enzyme cut site could be assigned confidently to one of the five primates in the present study. Roughly ___ million of those reads passed quality control filtration and ___ could be assigned to a stack. In these stacks our study identified ___ SNPs among all samples. Information for each individual is summarized in Table 1.

Table 1: Clustered Reads Data

Taxon	Source	# Reads	# Filtered Reads	# Stacks	Mean Coverage	# SNPs
Microcebus murinus?	Unknown					
Cebus sp.?	Unknown					
Theropithecus gelada	Unknown					
Pan troglodytes	Unknown					
Homo sapiens	Unknown					

In the human genome, we found ___ possible cut sites for *PspXI* and ___ possible sequencing sites (two per cut site, one upstream and one downstream). Of those, ___ sites (___%) were covered by at least one read in human and ___ sites (%) in chimpanzee. For ___ sites (__%), both chimpanzee and human had at least one read. When we restrict the analysis to sites with at least three reads, ___ sites (___%) were covered in human, ___ sites (___%) were covered in chimpanzee, and ___ sites (___%) in both.

4 Discussion

- RADseq is viable tool for researcher interested in primate phylogenetics, pop. gen.

Table 2: Mapped Reads Data

Taxon	# Reads	# Filtered Reads	# Loci ≥ 1 Read	# Loci ≥ 3 Reads	# SNPs
Pan troglodytes					
Homo sapiens					

  - Cheap and easy to create libraries. Competent lab can do it in two days. (Ha!)
  - Multiplexing is great for pop gen studies.
  - Enzyme choice allows control over coverage, number of individuals, number of loci.
  - Baird: "different marker densities can be attained by choice of restriction enzyme"

- Potential problems with RADseq method
- ddRADseq
- Promise for primatology

5 Acknowledgements

- NYU Med School folk
- Grants?