A new method for genome-wide marker development and typing holds great promise for primate population genetics

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

December 13, 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

In molecular primatology, as in all fields of molecular biology, the development of variable genetic markers is essential for the study of organisms at different levels, from population genetic to phylogeographic to phylogenetic research. During the first decades of the discipline, an impediment to researchers was the need to develop and type polymorphic markers in a taxon of interest. The markers that resulted from this time-consuming 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 caused by the inefficiency of marker discovery, many population genetic or phylogenetic studies in molecular primatology have been based on one or few loci, usually mitochondrial DNA or a few microsatellites (Ting and Sterner, 2012). 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 (Degnan and Rosenberg, 2009; Maddison, 1997; Edwards, 2009; Maddison and Knowles, 2006). Adequate resolution depends on high marker density, and until recently that goal has been out of reach for many primate researchers (Edwards, 2009).

The rapidly decreasing costs of DNA sequencing technology have promised revolutionary gains for primatology (Enard and Paabo 2004; Goodman et al 2005; Ting and Sterner, 2012). 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 whole genome sequencing has fallen to a level feasible for many researchers' budgets (e.g., Perry et al., 2012), 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 (McCormack, et al. 2012). 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 (Davey et al., 2011; McCormack, et al. 2012). 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; Etter et al., 2011). 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 demonstrate the technique in 5 primates: a lemur, New World monkey, Old World monkey, and two apes, and discuss the features that make RAD-seq promising for primatologists.

1.1 The RAD-seq Technique

The following is a summary of the RAD tag library preparation protocol of Etter et al., 2011 (Fig. 1). The RAD-seq library preparation begins when genomic DNA is digested with a restriction enzyme, such as *EcoRI* or *PspXI* (1A). 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 (1B). Once the barcode has been added, fragments from multiple individuals can be pooled (1C), and the DNA is randomly sheared with a sonicator to have a length distribution under 1 kilobase (1D). 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 (1E). It is ligated to the fragments and then the fragments are amplified via PCR (1F). 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

Restriction site-associated DNA Sequencing (RAD-Seq) is an economical and efficient method for SNP discovery and genotyping. Since its first application by Baird and colleagues (2008) on two model organisms the fungus *Neurospora crassa* and the three-spined stickleback, *Gasterosteus aculeatus* RAD-seq has been successfully applied to several organisms for which reference genome information was not available.

The ability of RAD-seq technology to identify thousands of orthologous SNPs across multiple individuals at both intra- and interspecific level makes this technique extremely promising for the study of population structure (Hohenlohe et al., 2010; Emerson et al., 2010; Keller et al., 2012), gene flow and hybridization (Hohenlohe et al., 2011; Keller et al., 2012), phylogeography (Emerson et al., 2010) and phylogeny (Wagner et al., 2012; Rubin et al., 2012). The RAD-tag sequencing approach has been particularly used to generate SNP data to address questions in population genomics. For example, a series of studies conducted by Hohenlohe and colleagues investigated parallel adaptation and hybridization in several species of fish (Hohenlohe et al. 2010, 2011, 2012), while Emerson et al (2010) identified more than 3,700 SNPs for pitcher plant mosquitoes in eastern North America to provide the first phylogeographic study using RAD sequence data.

Although RAD sequencing is more effective in addressing questions at or below the level of a single species, a few recent studies have used this technique in the analysis of phylogenetic questions. Rubin et al (2012) provided a simulation study in which they investigated the accuracy of RAD-seq data to reconstruct phylogenies in organisms with different population sizes and clade ages (*Drosophila*, mammals, and yeasts). In their study the authors supported the efficiency of RAD-seq data in inferring phylogenies, but they also caution about how this approach achieves the best results in younger clades where more orthologous restriction sites are likely to be retained across species. This simulation analysis was confirmed in two recent empirical studies in which a RAD-tag sequencing approach was successfully used to reconstruct phylogenetic relationships in two recent but speciose radiation, such as the African cyclids (Wagner et al., 2012) and the *Heliconius* butteries (Nadeau et al., 2012).

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. We sequenced the prepared library as one 150-cycle paired-end run and one 150-cycle single-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 As input for the clustering analysis, we combined the first read of the paired-end run and the single-end run reads. 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 sets that differed by no more than two basepairs and compared closely related sets to detect orthologous loci and SNPs using a maximum likelihood approach (Catchen et al., 2011). We tallied orthologous SNPs using VCFtools (Danecek et al., 2011).

Sequence Analysis - Assess RAD Tag Coverage To assess the RAD tag coverage, we mapped human and chimpanzee reads to the highest quality primate reference genome, that of humans. Again, we excluded reads without an expected barcode or an intact restriction enzyme cut site. We aligned reads to the human reference genome (GRCh37/hg19, Human Genome Consortium 2001) using BWA with default parameters (Li and Durbin, 2009). We separately mapped the single-end and paired-end data and then combined the resultant files after alignment. We removed reads that were unmapped or that had low mapping quality using Picard (http://picard.sourceforge.net) and BamTools (Barnett et al., 2011).

After performing local realignment around indels with GATK (DePristo et al., 2011), we identified SNPs and short indels using SAMtools mpileup and BCFtools (Li et al., 2009). 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 (Danecek et al., 2011).

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 and created a BED file of all regions 1000 basepairs upstream and downstream. (Meyer et al., 2012). This allowed us to calculate the coverage of these restriction site-associated regions using BEDtools' multiBamCov program (Quinlan and Hall 2010).

3 Results

12.3 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 9.1 million of those reads passed quality control filtration and were clustered into stacks. By comparing these stacks and including only SNPs that were present in multiple individuals, our study identified 7,910 SNPs among all samples. Information for each individual is summarized in Table 1.

Table 1: Clustered Reads Data

Taxon	# Reads	# Filtered Reads	# Ind. Stacks	Mean Coverage (SD)	# Shared SNPs
Microcebus sp.	2,830,832	2,025,103	248,324	7.66 (20.71)	13
Cebus sp.	1,946,096	1,427,413	107,829	12.64 (139.08)	56
Theropithecus gelada	1,918,425	1,392,709	136,657	9.70 (17.34)	212
Pan troglodytes	2,616,062	1,910,560	157,775	11.50 (42.27)	5,886
Homo sapiens	3,032,823	2,374,733	131,544	17.37 (25.30)	5,786

In the human genome, we found 58,172 possible cut sites for *PspXI* and 116,344 possible sequencing sites (two per cut site, one upstream and one downstream). Of those possible location, 111,686 sites (96.00%) had at least one mapped read present in human and 91,646 sites (78.77%) in chimpanzee. For 90,022 sites (77.38%), both chimpanzee and human had at least one read. When we restrict the analysis to sites with at least three reads, 109,098 sites (93.77%) had sequences in human, 89,628 sites (77.04%) in chimpanzee, and 86,604 sites (74.44%) in both. From this data, we found 9,275 SNPs relative to the human reference genome that were present in both chimpanzee and human datasets.

Table 2: Mapped Reads Data

Taxon	# Reads	# Filtered Reads	# Loci ≥ 1 Read	# Loci ≥ 3 Reads	# SNPs
Pan troglodytes	3,917,046	2,826,643	91,646	89,628	309,703
Homo sapiens	4,542,978	3,784,192	111,686	109,098	35,651

4 Discussion

[Why it would be good in primates.]

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

- 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"

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.

Potential problems with RADseq method including the phylogenetic distance and the need of high quality DNA (at least for now)

RADseq is shitty when phylogenetic distances are great. Works much better within a genus or species level. We managed to infer the accepted primate phylogeny from our SNP data using maximum likelihood approach, but the support values were crap

ddRADseq

5 Acknowledgements

The authors would like to thank the NYU Langone Medical Center's Genome Technology Center for assistance with library preparation and sequencing.

- NYU Med School folk
- Grants?