RADseq Works in Primates, Dammit.

Christina M. Bergey Andrew S. Burrell
Luca S.J. Pozzi Todd, I suppose

December 8, 2012

Abstract

... Blah, blah, blah, RADseq, blah, blah, Cercopithecoidea. ...

1 Introduction

- Next-gen sequencing revolution promises gains in primatology
- Still expensive
- Many genomes, but still tough doing genomics on non-model organisms
- What is RADseq?
- Previous RADseq studies
- Why would it be good for primates
- PRESENT STUDY
  - We did RADseq on 6 Cercopithecoids
  - Assessed how well it worked
  - Show it has promise for primates

2 Methods

Library Preparation and Sequencing Genomic DNA from 6 primates was digested with *PspXI* (New England Biolabs) and used 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 Sequence reads were demultiplexed, or separated by barcode, and reads without an expected barcode or an intact restriction enzyme cut site were excluded from the analysis. Reads were then aligned to the rhesus macaque reference genome (v.1.0, Mmul_051212/rheMac2, ref) using BWA with default parameters. Reads that were unmapped, unpaired, duplicates, or that had low mapping quality were removed after alignment using Picard (ref) and BamTools (ref).

After performing local realignment around indels with GATK (ref), SNPs and short indels were identified using SAMtools mpileup and BCFtools (ref). A minimum coverage of 3 reads and a maximum of 100 was required to call a SNP or an indel at a given location. Orthologous SNPs were tallied 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 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

6.1 million sequencing reads with an intact barcode and restriction enzyme cut site could be assigned confidently to one of the six Old World monkeys in the present study. Roughly 2.0 million of those reads were successfully mapped to the rhesus macaque genome and passed all quality control filtration steps. Relative to the reference genome, our study identified 531,175 SNPs and 24,260 small indels among all samples. Information for each individual is summarized in Table 1.

In the rhesus macaque genome, we found 108,727 possible cut sites for *PspXI*.

Table 1: Individual Sample Information

Taxon	Source	# Reads	# Filtered Reads	# Loci ≥ 1 Read	# Loci ≥ 3 Reads	# SNPs
Langur?	Unknown	1,367,050	306,248			363,501
Allenopithecus nigroviridis	Unknown	1,325,558	447,279			356,750
Macaca mulatta	Unknown	481,376	234,659			279,350
Cercocebus torquatus atys	Unknown	297,290	124,700			319,766
Papio anubis?	Unknown	743,556	313,676			316,118
Theropithecus gelada	Unknown	1,911,030	603,024			366,097

- Number of loci hit
- Number of loci hit with coverage $\geq N$
- SNP info from merged analysis
- SNP Venn diagram?
- Calculate coverage of restriction site-associated regions
- Total number possible targets in rhesus genome (compare to human too?)

- Total possible target BP
- How many targets did we hit?
- BEDtools multiBamCoverage for this job
- Number and percentage of targets with coverage ≥ 1
- Number and percentage of targets with coverage $\geq N$
- Count orthologous SNPs shared between individuals
- VCFtools vcf-compare for this job

4 Discussion

- RADseq is viable tool for researcher interested in primate phylogenetics, pop. gen.
- Enzyme choice allows control over coverage, number of individuals, number of loci.
- Potential problems with RADseq method
- Promise for primatology

5 Acknowledgements

Acknowledgements