

SARAH M. HIRD

[SHIRD1@TIGERS.LSU.EDU]

119 FOSTER HALL; MUSEUM OF NATURAL SCIENCE; LOUISIANA STATE UNIVERSITY; BATON ROUGE, LA 70803

**1. EDUCATION**

- 2013 Ph.D. (Biology) Louisiana State University (LSU). Advisors: Drs. Robb Brumfield & Bryan Carstens.  
 2008 M.Sci. (Biology) University of Idaho (UI). Advisor: Dr. Jack Sullivan.  
 2005 B.S. (Biology) University of Idaho (UI). Mathematics minor.

**2. PEER REVIEWED PUBLICATIONS**

- McCormack JE, Maley JM, **Hird SM**, Derryberry EP, Graves GR and RT Brumfield. 2011. Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. *Molecular Ecology*, in review.
- Hird SM**, Brumfield RT and BC Carstens. 2011. PRGMATIC: an efficient pipeline for collating genome-enriched second generation sequencing data using a "provisional-reference genome". *Molecular Ecology Resources*, in press [DOI: 10.1111/j.1755-0998.2011.03005.x].
- Hird S**, Kubatko L and B Carstens. 2010. Replicated subsampling enables accurate species tree estimation in empirical systems. *Molecular Phylogenetics and Evolution*, 57(2): 888-898.
- Hird S**, Reid N, Demboski J and J Sullivan. 2010. Introgression at differentially aged contact zones in red-tailed chipmunks (*Tamias ruficaudus*). *Genetica*, 138(8): 869-883.
- Reid N, **Hird S**, Schulte-Hostedde A and J Sullivan. 2010. Examination of nuclear markers at a zone of mitochondrial introgression in two chipmunk species (*Tamias amoenus* and *Tamias ruficaudus*). *Journal of Mammalogy*, 91(6): 1389-1400.
- Koopman M, Fuselier D, **Hird S** and B Carstens. 2010. The carnivorous Pale Pitcher Plant harbors diverse, distinct and temporally dependent bacterial communities. *Applied and Environmental Microbiology*, 76: 1851-1860.
- Hird S** and J Sullivan. 2009. Assessing gene flow across a hybrid zone in red-tailed chipmunks (*Tamias ruficaudus*). *Molecular Ecology*, 18: 3097-3109.
- Good J, **Hird S**, Reid N, Demboski J, Steppan S, Martin-Nims T and J Sullivan. 2008. Ancient Hybridization and Mitochondrial Capture between Two Distantly Related Species of Chipmunks (*Tamias*: Rodentia). *Molecular Ecology*, 17 (5): 1313-1327.

**3. RESEARCH AND EDUCATION FUNDING**

2010. BioGrads Research Award, LSU. (\$300)  
 2009. Graduate Student Enhancement, LSU. (\$20,000)

**4. SELECT PRESENTATIONS**

- Hird SM**, Koopman MM, Ence DD, Brumfield RT and BC Carstens. 2010. (Poster) *An Efficient Pipeline for the Preparation of Phylogeographic Data Collected Via Next-Generation Sequencing*; Society for the Study of Evolution, Portland, OR.
- Hird S** and J Sullivan. 2009. (Talk) Nuclear markers at red-tailed chipmunk hybrid zones. LSU, Museum of Natural Science Seminar Series.
- Hird S** and J Sullivan. 2006. (Poster) *Preliminary Microsatellite Analysis of a Chipmunk Hybrid Zone*; Society for the Study of Evolution, Stony Brook, NY.

**5. PROFESSIONAL EXPERIENCE**

- 2009 – present; Graduate Research Assistant, LSU.  
 2008-2009; Research Associate, Carstens Lab, LSU.  
 2005-2008; Graduate Teaching Assistant, UI.

**6. PROFESSIONAL ORGANIZATIONS**

- American Association for the Advancement of Science (AAAS); American Society for Microbiology (ASM); Society for the Study of Evolution (SSE); Society of Systematic Biologists (SSB)

**7. REVIEWS PROVIDED**

- Molecular Ecology, Journal of Heredity

## Distinguishing paralogy from variation and error in next-generation sequences

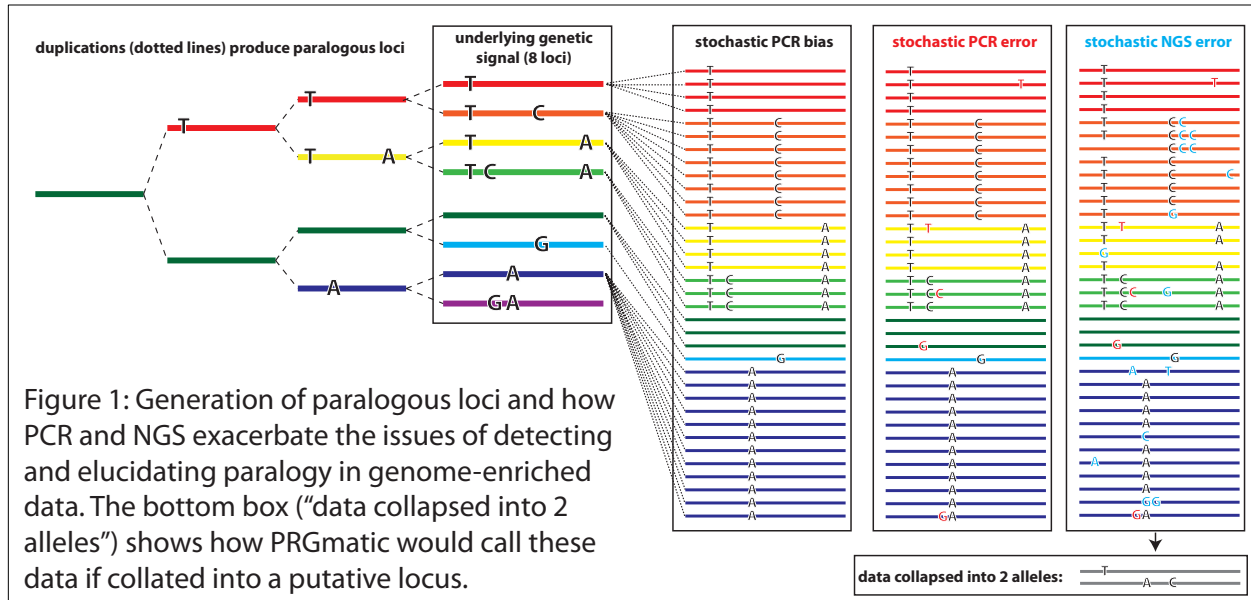
Sarah M. Hird [SHIRD1@TIGERS.LSU.EDU]

*Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803*

Funds are requested to verify and improve the performance of a software pipeline intended to allow any lab without bioinformatics capabilities to gather and analyze next-generation sequencing data. The requested funds (\$2000.00) will allow me to Sanger sequence 768 clones from 8 putatively paralogous loci (as inferred from the pipeline). These data will be used to empirically obtain underlying copy number of the loci and as a guide for how to computationally deal with highly similar, yet not homologous sequences – a current and unresolved issue in genome-enriched next-generation sequencing experiments.

**Introduction** Next-generation sequencing (NGS) refers to the wave of new sequencing technologies that allow researchers to generate gigabases of data in a single run (e.g., Illumina's Genome Analyzer, Roche 454's Genome Sequencer). Although most frequently utilized for genome sequencing, NGS technology is being increasingly applied to systematics and locus-based research through genome-enrichment techniques (e.g., (1) and for a review, see (2)). Many disciplines of evolutionary biology (e.g., phylogeography, population biology, speciation, ecology, etc.) are interested in sampling loci from across the genome because inference about population and species level questions improves as independent loci are added (3). Such techniques allow a targeted and reduced subset of the genome to be amplified and when coupled with multiplexed sequence tags, allow these loci to be sequenced in up to hundreds of individuals (1, 4). Since restriction enzymes are often used to fragment the genome and isolate homologous loci across individuals (e.g., RAD tags (5) or modified AFLP protocols (4, 6, 7)), I wrote a bioinformatics pipeline (called PRGMATIC (8)) that collates the NGS reads into putative loci based on percent identity of the sequences. Briefly, the pipeline clusters sequences within an individual into provisional alleles then clusters high-coverage provisional alleles across individuals into provisional loci. The provisional loci are concatenated into a provisional-reference genome (or PRG) that can be used in conjunction with a variety of programs that require a reference genome. The pipeline then aligns all the reads from each individual to the PRG and calls two alleles/individual, based on the reads that align to each provisional locus. The end product of the pipeline is a folder of multi-FASTA files that correspond to the provisional loci, with two alleles called for each individual. This method is fast, flexible (user can set various parameter values) and free and generates meaningful biological signal (6, 7).

While writing the pipeline and analyzing preliminary data, we noticed some loci contain too much variation to be generated from two alleles and error alone, the expectation for diploid organisms (Figure 2). We ascribed this variation to the reads being generated from copy number variants (CNVs) across the genome (i.e., multiple loc, Figure 1). Here, CNVs refer to small insertion/deletions (or INDELS, 2-1,000 base pairs), structural variants (>1,000 base pairs), pseudogenes and tandem repeats and these collectively represent most of the variation in the human genome (versus single nucleotide polymorphisms, or SNPs, (9)). In fact, CNV variability and prominence may represent the most informative population level polymorphisms (10) yet are relatively under studied (11). Being able to use these markers, instead of discard them, will greatly increase knowledge and power regarding species level questions, and potentially higher and lower taxonomic levels as well.

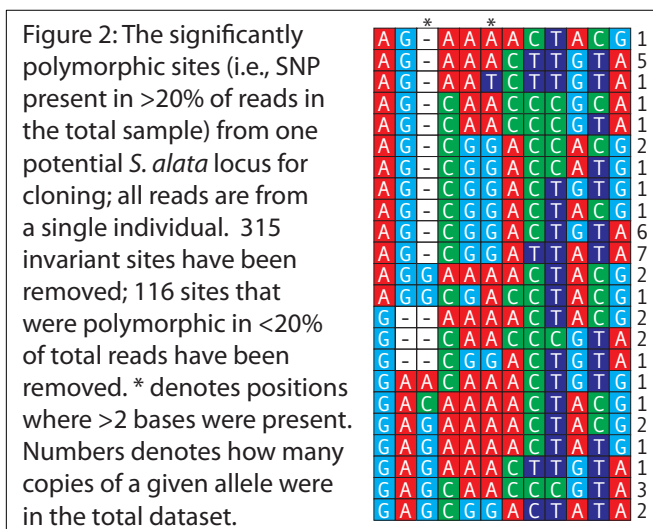


In addition to unknown copy number, error in the PCR and NGS sequence generation further complicates understanding conflicting signal in NGS datasets. Figure 1 shows how PRGMATIC would call 2 alleles from sequences within 90% identity but from 8 underlying loci; stochastic error incorporation makes deciphering true signal difficult. However, there are patterns that can be inferred from the phylogeny of the reads; the generation of cloned, phased, low error sequences will be used to ascertain copy number and understand how CNV sequences manifest in actual data. Read depth can predict copy number in high coverage genomes (12); I will use the Sanger sequence derived copy number and read depth in NGS data to write an algorithm for detecting CNVs in genome-enriched data. Singly unique nucleotides (or SUNs, private SNPs that distinguish unique alleles) will be detected with this method as well. Computationally detecting SUNs is another way to infer underlying copy number in paralogous sequences. I will also use known error rates and the correct sequences to recognize and discard error in the CNVs; incorporating error rate is another goal for improving PRGMATIC.

**Objectives** (1) Determine copy number in 8 highly polymorphic provisional loci called by the program PRGMATIC using PCR, cloning and Sanger sequencing. (2) Decipher patterns that CNVs produce in genome-enriched NGS datasets. (3) Incorporate the information obtained about copy number, allele sequence and error into PRGMATIC.

**Methods** I will select the 8 most likely paralogous loci from the *Sarracenia alata* 454 NGS dataset collected by the Carstens Lab at LSU (7) by calculating number of variable sites/site and visually checking for greater than the 2 possible alleles for a locus (for example, see Figure 2). I will then design primers for these loci based on the sequences using Primer3 (13) and amplify these loci in the 5 most variable individuals/locus. I will then use a Qiagen PCR Clone Kit to clone these PCR reactions. I will select 18-20 clones per individual per locus at random for sequencing. PCR cleanup, cycle sequencing, cycle sequencing cleanup and sequencing will be outsourced to Beckman-Coulter (Brea, CA), a cost effective means of generating Sanger sequencing.

**Significance** All comparative genetic studies rely on homology of the sequences. Next-generation sequencing provides a significant increase in cost and time efficiency and allows researchers to generate orders of magnitude more data for equivalent prices to Sanger sequencing. Applying this technology to population level questions is integral to the progression of the field and requires genome-enrichment techniques. PRGMATIC is free of software that attempts to bring the sequencing capacity to primary researchers by collating anonymously generated NGS fragments into loci. However, paralogy of sequences undermines efficiency because copy number variants represent a significant and poorly understood portion of the genome. Collecting empirical copy number variant validation will allow me to improve PRGMATIC by accounting for paralogy in multi-locus, multi-individual datasets, utilizing their unique variation instead of discarding it.



**Schedule** We have NGS data for over one thousand loci in *Sarracenia alata*; I have already identified two-dozen potential loci to design primers from, which would commence as soon as funds were available (Fall 2011). The lab work (PCR and cloning) would occur over the subsequent 2 months and clones would be sent for sequencing within 3 months. Analyzing the results and incorporating them into PRGMATIC would take another several months; I estimate I would begin writing the manuscript in mid-Spring 2012 and submit it to a journal in Summer 2012.

1. P. A. Hohenlohe *et al.*, *PLoS Genetics* **6**, 23 (2010).
2. L. Mamanova *et al.*, *Nature Methods* **7**, 111 (2009).
3. R. Brumfield, L. Liu, D. Lum, S. Edwards, *Systematic Biology* **57**, 719 (2008).
4. Z. Gompert *et al.*, *Molecular Ecology* **19**, 2455 (2010).
5. N. Baird *et al.*, *PLoS One* **3**, 3376 (2008).
6. J. McCormack *et al.*, *Molecular Ecology*, in review, (2011).
7. A. J. Zellmer, M. M. Koopman, S. M. Hird, B. C. Carstens, in prep, (2011).
8. S. M. Hird, R. T. Brumfield, B. C. Carstens, *Molecular Ecology Resources*, (2011).
9. D. F. Conrad *et al.*, *Nature* **464**, 704 (May 01, 2010).
10. R. Xi, T. Kim, P. Park, *Briefings in Functional Genomics* **9**, 405 (2010).
11. R. Mills, K. Walter, C. Stewart, R. Handsaker, *Nature*, (2011).
12. P. H. Sudmant *et al.*, *Science* **330**, 641 (2010).
13. S. Rozen, H. J. Skaletsky, in *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, S. Krawetz, S. Misener, Eds. (Humana Press, Totowa, NJ, 2000), pp. 365-386.

### **Budget**

Expense	Price(\$)/ unit	Quantity	Total(\$)
Platinum® <i>Taq</i> DNA Polymerase High Fidelity (100 reactions)	158	0.5	79
Qiagen PCR Cloning Kit (40 reactions):	806	1	806
Beckman-Coulter Sequencing (96 sequences):	144	8	1152
		Total Expenses:	2037.00
		<b>Total Requested:</b>	<b>2000.00</b>

### **Budget Justification**

In order to distinguish between paralogy and error in closely related NGS data, I will need to clone PCR products and sequence them. I will use Platinum® *Taq* DNA Polymerase (High Fidelity) to minimize PCR error prior to the cloning reaction (40 reactions plus 10 extra to account for optimization of reactions). The Qiagen PCR Cloning Kit is an all-inclusive kit that will perform 40 cloning reactions (8 loci, 5 individuals per locus). The Beckman-Coulter sequencing will produce 768 sequences, 18-20 clones sequenced per individual per locus (or 96 clones per locus). This should be ample data to begin to investigate the number of copies of a locus across individuals within a species. I am requesting these funds from SSB because systematics relies heavily on DNA sequence data and with NGS advancements and genome-enrichment techniques, it will be most useful to systematically address paralogy in these datasets thus improving gene tree inference, species delimitation, cost efficiency, etc.. The primers and other PCR reagents for generating the putatively paralogous fragments will be purchased with funds from another grant.