1. BACKGROUND – Biodiversity is predicted to decline for the foreseeable future[1], and recent studies have documented this in diverse ecosystems (e.g.,[2,3]). This underscores the proximate need to understand the evolutionary processes that create biodiversity through speciation. Recently, the genomics revolution has led to a near exponential growth of archived genomics data since 2006, now with petabytes of data from a large variety of study systems[4]. This revolution has ushered in unprecedented opportunities to bring new quantitative insights to long standing problems relating biogeographic context with speciation in the emerging field of speciation genomics[5]. In addition, it has only recently been computationally feasible to identify complex genetic variants, and they have been found to be important to the speciation process (e.g., reviewed in Bohne et al.[6]). Further, using online biological collections provides a cost effective and robust way to generate novel insights while using freely available curated biological data[7]. Using these data, I propose to investigate the relationship of speciation genomics with biogeography using genetic variants. First, I will identify complex genomic variants from *selected species groups* to investigate their importance as gene flow isolating mechanisms (**Aim 1**). Further, I will identify regions that show a pattern of reinforcement to determine the relative contribution to reinforcement of reproductive versus ecological character displacement from gene functional enrichment analysis (**Aim 2**).

There are two main biogeographic contexts important to the speciation process. First, allopatric speciation results from interrupting a continuous species range (e.g., mountain ranges[8]), which decreases migration and subsequent gene flow, and, given ample time, results in two incipient species. In contrast, sympatric speciation results without geographic interruption, proceeds with gene flow, and results in two incipient species[9]. Though empirical support for the latter has been difficult to find (reviewed in Feder et al.[10]), various mechanisms have been posited as to how species can arise in a sympatric context with selection on enhanced prezygotic isolation (i.e., reinforcement) being empirically important[11]. Reinforcement, selection completing the process of speciation in previously allopatric populations, has been found to be relatively common (reviewed in Coyne and Orr[12]), and an important route to sympatric speciation (reviewed in Abbott et al.[13]). Reinforcement can isolate sympatric species causing a barrier to gene flow in two ways: 1) reproductive character displacement, or 2) ecological character displacement (reviewed in Servedio and Noor[14]). Noor[15] has provided a test for reinforcement with the expectation that a given species will be more divergent from a closely related sympatric species than a closely related allopatric species. While this test was formulated for reproductive isolation, the logic of this test can be applied to other characters that might isolate species such as genetic variants. Thus, I will use Noor's[15] test for reinforcement to investigate the relationship of geographic context with four broad classes of genetic variants: 1) transposable elements, 2) short tandem repeats, 3) structural variants, and 4) single nucleotide polymorphisms (SNPs).

Isolation in a genomic context can arise through a variety of mechanisms. While previously only simple genetic variants have been considered (e.g. SNPs), novel bioinformatics software has been developed to computationally identify complex genetic variants. Therefore, these complex genetic variants represent an understudied, but likely important, component of the genome that might transform our understanding of how speciation occurs in a biogeographic context. First, **transposable elements** are a broad class of genetic variant that can make up a significant proportion of eukaryotic genomes (e.g., 40% in Humans[6]). Transposable elements self replicate and insert themselves into other parts of the genome, and have been recently identified as important to speciation (reviewed in Bohne et al.[6]). Second, **short tandem repeats** might be important to the speciation process because they have been recently shown to affect gene expression in diverse groups including humans[16,17,18]. Third, **structural variants** have been identified as an important source of genomic variability[19]. There are four different types of structural variants that are important to adaptation and speciation: 1) inversions[20], 2) copy number variants[21], 3) insertions and deletions[22], and 4) translocations[23]. Finally, SNPs are an important source of genomic variability, and have been implicated in environmental adaptation and subsequent speciation[24]. While complex genetic variants have been shown to be important to the speciation process, to my knowledge, this study will be the first that 1) investigates the relationship between complex genetic variants and biogeographic context, and 2) accesses functional aspects of genetic differences in a biogeographic context related to the speciation process across a diverse set of taxa.

To investigate the relationship of biogeography and speciation genetics, I propose a study using genomic biological collections (i.e., National Center for Biotechnology Information [NCBI] sequence read archive). I will investigate two related hypotheses that will be important to understanding how genetic variants (discussed above) and associated functions of genes that contain genetic variants contribute to speciation in contrasting biogeographic distributions. **1) I hypothesize that complex genomic variants will be more often different between sympatric than allopatric species highlighting their importance as a gene flow isolating mechanism** and **2) I hypothesize that polymorphic sites (SNPs) that are more often different between sympatric than allopatric species will be associated with functions indicative of ecological character displacement.**

2. PROPOSED RESEARCH – The primary goal of my research is to **investigate the relationship of speciation genomics and biogeography.** I will achieve this by using ***electronic biological collections*** to retrieve raw genomic data for six replicate biogeographic 4-species groups from four continents selected from the literature including insects, fish, mammals, and birds (**Table 1**). Further, I will add more species groups from the primary literature when they are made available. My work will uncover general patterns related to speciation genomics, and the functional genetic differences that underlie these divergent genomic regions. **In Aim 1**, I will investigate the relationship of biogeographic structure and complex genetic elements to understand the speciation process. **In Aim 2**, I will identify single nucleotide polymorphisms (SNPs) indicative of reinforcement within a biogeographic context. I will use a gene ontology approach to investigate potential functions of the underlying genes to understand why these might have led to speciation. Already, I have successfully conducted most of the genomic analysis techniques proposed in the Stevison lab.

***Aim 1: Identify complex genomic variants from selected species group to investigate their importance as gene flow isolating mechanisms*.** Because this investigation relies on publicly available whole genome sequence collections contained in NCBI's sequence read archive (SRA), I will download raw paired-end reads for species groups that have been pre-selected from the literature (**Table 1**) having the biogeographic and phylogenetic relationships represented in **Figure 1**. To identify candidate groups, *I conducted a literature search* for species groups which have BOTH whole genome sequencing data deposited in NCBI's SRA and the appropriate biogeographic and phylogenetic structure to investigate my hypothesis (**Table 1**). In order to analyze these data, I will use the significant computational resources of two high performance computers (HPCs) available to me: 1) Auburn University Hopper HPC with 80 cores dedicated to our lab, and 2) the computing resources of the Alabama Supercomputer Authority.

Table 1: Proposed organisms and accession numbers in NCBI's SRA. Each **Organism** represents four species in the appropriate geographic and phylogenetic context (see text for details). Data is from the National Center for Biotechnology Information (NCBI) and the European Nucleotide Archive (ENA). NCBI and ENA accession numbers can be used to access the associated data in NCBI's SRA.

Organism	Accession Numbers
Mosquitoes (Anopheles)[25]	NCBI: PRJNA6751 and PRJNA254046.
Horses (Equus)[26]	ENA: PRJEB7446
Butterflies (Heliconius)[27]	ENA: ERP002440
Flycatchers (Ficula)[28]	ENA: PRJEB7359
Dogs (Canis)[29]	Authors Contacted
Cichlids[30]	NCBI: PRJNA78915, PRJNA60369, PRJNA60363, and PRJNA78185

To analyze these data, I will download the data from NCBI's SRA, align to the reference genome, and use the GATK best practices pipeline[31]. For Aim 1, I will only use the GATK pipeline through the base quality recalibration score step (i.e., not calling variants), which will provide high quality alignment data to discover complex genetic variants. I will use three different classes of genetic variants for Aim 1: 1) transposable elements, 2) small tandem repeats, and 3) structural variants. In order to detect both transposable elements and short tandem repeats I will use RED[33]. Because RED requires fasta format files to discover genetic variants, I will convert the high quality alignment files derived above to fasta format before running RED. I will use metaSV[32] software with the alignment data to detect structural variants such as insertions, deletions, inversions, and translocations that have been shown to be important to the speciation process (see introduction). MetaSV uses the consensus from four independent software packages, producing a high quality set of structural variants. These genetic variants will be used to investigate biogeographic speciation.

Using the identified complex genetic variants, I will apply a 4-species test (modified from previous 4-species tests[34,35]) in order to determine if complex genetic elements are divergent between the test group (A) and the sympatric (B) and allopatric species (C), respectively, while limiting the search for complex genetic variants derived after the outgroup (O) diverged from the most recent common ancestor (**Figure 1**). A simple expectation of reinforcement following Noor[15] is that species in sympatry will be more genetically isolated than those species in allopatry. Specifically, I will use the 4-species test to compute a modified D statistic based on biogeographic relationships in **Figure 1**. I modified the D statistic to be calculated as $\frac{\sum_i^n Sym_i - Allo_i}{\sum_i^n Sym_i + Allo_i}$ (1) where *Sym* and *Allo* are indicator variables that take the value 1 (else 0) when the test species (A) has a different genetic variant than the sympatric (B) and allopatric (C) species, respectively.

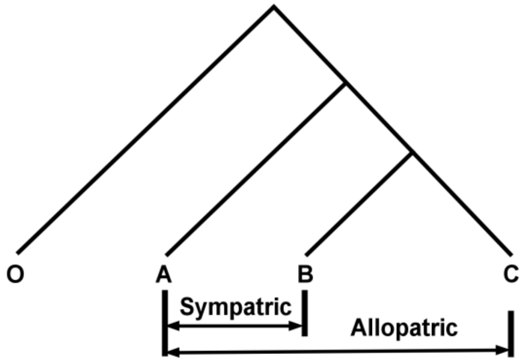

Figure 1. Schematic for the geographic and phylogenetic relationships of organisms used in test for reinforcement (below).

The specific predictions for this test are that if D=0; the genetic variant is equally shared, D>0; the genetic variant is less often shared with the sympatric species (i.e., reinforcement), and D<0; the genetic variant is less often shared with the allopatric species. In order to assess whether there is a significant deviation from 0, I will use a **novel permutation based method**. Briefly, I will shuffle species labels to break the link between species and genetic variants, and recalculate D ≥ 1000 times resulting in a p-value equal to the sum of the count of D greater than or less than 0 divided by number of permutations.

I hypothesize that complex genomic variants will be more often different between sympatric than allopatric species (e.g., D>0). By conducting this analysis with multiple types of complex genetic variants, my work will highlight which is most likely to have a role in contributing to speciation. Additionally, by conducting this analysis in a variety of taxa, my work will also determine if the same gene flow isolating mechanisms are important across various taxa.

***Aim 2.* Identify regions that show a pattern of reinforcement (D>0) to determine the relative contribution to reinforcement of reproductive versus ecological character displacement from gene functional enrichment analysis.** In contrast to Aim 1 where D will be calculated at the whole genome scale, in Aim 2, I will use a sliding window approach to identify regions that show the pattern of reinforcement (D>0). Because the window size will likely be taxon specific, the sliding window size will be determined to maximize the signal to noise ratio[36] based on SNP density[37] and/or genome size[35]. Specifically, I will identify SNPs through GATK's variant calling pipeline[38], using the base quality recalibrated alignment files from Aim 1. I will further investigate gene function in genomic regions showing the pattern of reinforcement, targeting genes containing identified SNPs in regions with D>0. I will use a gene ontology (GO) approach to determine gene functional enrichment[39]. In two recent studies, GO has been used successfully to show reproductive character displacement (RCD)[40] and ecological character displacement (ECD)[41]. Specifically, GO terms related to RCD should be relatively consistent across taxa. For instance, I would consider genes related to recombination and sperm development (among others) to be related to RCD[40]. In contrast, genes related to ECD will, likely, be more variable, but indicative of ecological factors influencing selection (e.g., habitat type). For instance, genes related to habitat type indicative of ECD were found to be important in mole rats[41]. For the GO analysis, I will use the recently updated DAVID bioinformatics resources 6.8[42], specifically the RDAVIDWebService R package[43], to test for significant gene enrichment (or depletion) for both RCD and ECD categories with a Fisher's exact test corrected for the false discovery rate[44].

Expected challenges and potential solutions: As pointed out in reviews of GO analyses[45,46], one possible challenge of using GO with non-model organisms is that their genomes may not be thoroughly annotated. Ensemble[47] has well annotated reference genomes for all organisms in **Table 1** except Heliconius. The Heliconius genome has been recently assembled and updated[48], which may result in incomplete annotation. In order to better annotate this genome (and potentially others), I will use Blast2GO which combines a homology search and GO annotation into one automated step[49,50] with fasta sequences containing SNPs indicative of reinforcement identified in Aim 2. Blast2GO has been used successfully in other non-model organisms[22]. Finally, new species groups will likely be published that fit the biogeographic and phylogenetic relationships outlined in **Figure 1,** and I will add these to the analyses as these data become available.

Significance: This study will investigate the genetics of speciation in a biogeographic context, and proposes to use a wide array of organisms. The proposed research should be able to uncover general patterns from robust inferences about how biogeographic context affects the speciation process because of the taxonomic breadth investigated. To my knowledge, a study of this taxonomic breadth has not been conducted to investigate the relationship of biogeographic context and speciation, and could potentially transform our understanding of the relationship of biogeography and speciation.

3. BROADER IMPACTS – Because the proposed work uses electronic collections, I am proposing three major broader impact activities, complementary to the proposed research, that focus on interactions between the public and traditional collections, and bring traditional and online collections together. **Activities A** and **B** will be targeted at 7/8th graders, which is the year of (or after) life science is taught in Alabama schools[51], and I will use the guidance of a recent project to teach biology to Alabama state standards[52]. This will build off of my previous experience in developing and leading one day 7/8th grade field based, hypothesis driven stream ecology workshops. These were developed in collaboration with a 7/8th grade teacher, and a South Carolina Department of Natural Resources scientist. In addition, these workshops were held for 6 years in South Carolina and served an under-represented, low-income school (data for Liberty Middle School from the US Department of Education[53]). **Activity C** will serve to aid integration between electronic and traditional collections. These will be developed and implemented in collaboration with the Auburn University Museum of Natural History (AUMNH), and the College of Math and Sciences outreach office.

A. Speciation at the junior curator camp: First, AUMNH holds an annual week long curator camp for 7/8th graders. This camp consists of curators from the museum leading field expeditions and then teaching students how to curate museum specimens. This camp covers a broad range of organisms (from insects to vertebrates). Since speciation is a process that occurs across all organismal groups, I propose to develop a 4 hour speciation lesson in order to synthesize the week's curator camp activities that will act as a capstone for the week's activities.

B. Online bioinformatics lesson**:** Second, modern bioinformatics came into existence with the introduction of high throughput sequencing (HTS) in 2006[54]. Even before HTS, bioinformatics was seen as having a strong future with an increasing importance[55]. As a result, it is important to introduce children to this very important STEM career early in their academic careers. I propose to develop a 7th grade web based, interactive bioinformatics module incorporating knowledge gained from successful grade level lessons [56] that will be initially deployed to schools in the Auburn area, and, importantly, to the Wehle center that serves under-represented groups in STEM located in rural Bullock County, Alabama. This investigation will be hypothesis driven, will be tailored to the grade level by collaborating with an education graduate student through a class offered at Auburn University (BIOL 7960), and will require input of simplified commands at a web-based command line that will produce graphical outputs to introduce students to bioinformatics data analysis. I will use the R language for statistical computing[57,58] with the shineyR web interface builder to build the data analysis part of this lesson, and I will use my previous website development experience to develop the remainder of the lesson.

C. Bringing online and traditional collections together: Third, I have met with AUMNH curator staff, David Werneke (Fish Collections Manager) and Dr. Brian Helms (Assistant Research Professor and Invertebrate Collections Manager), and we identified a need to connect online collections with traditional collections. The museum does loan tissue for sequencing, and these sequencing data are accessioned on NCBI, but accession numbers are not always reported to the museum. In order to connect the AUMNH with NCBI accession numbers, I will write a python program that will access the SQL database backend (sqlite3 python package) for the museum's collections database. Following this, the script will take the AUMNH accession numbers and use them to query the NCBI database with biopython[59]. Following this, a comma separated value file of changes will be written to the server for manual curation, and an email will be sent to the collection managers informing them that the script has run. ***I will release this code on github as it may be useful for other collections across the country.***

4. TRAINING OBJECTIVES AND CAREER DEVELOPMENT – As the proposed work represents a large departure from the work I conducted for my PhD, I believe that this fellowship represents a unique opportunity to learn and apply bioinformatics approaches to evolutionary biological questions. This will substantially expand my research abilities, opening to me a completely new area of research, that will not only be important for the proposed research, but by adding genomic expertise will allow for my future research to bridge genomics and traditional ecology. Because of this change in focus, I am using an **Individual Development Plan (IDP)**[60] to guide my exploration of careers and attainment of career goals. Since this is an iterative process, Dr. Stevison and I will discuss my career goals and progress at regular meetings. The results of the **IDP** indicated that my interests, skills, and values would be a good fit for research focused careers (academic or government). Additionally, a career in industry was identified as a potential good career path. I am exploring this novel (to me) career path by researching this potential career, and using Dr. Stevison's contacts in industry to develop contacts and gain knowledge (e.g. Monsanto, 23andMe, Ancestry.com, etc.). In addition, I have set areas of improvement that should benefit my career development: 1) improve writing grants and papers, 2) learn a new programming language (java or C), and 3) further refine and develop undergraduate mentoring skills. Upon completion of this fellowship (if funded) and implementing my **IDP**, I will possess a unique and highly competitive skill set for pursuing a career in academia, government, or industry.

5. SPONSORING SCIENTIST AND HOST INSTITUTION – *Sponsoring Scientist:* Dr. Stevison will be a superb mentor for both my proposed research and professional development. With regards to my proposed research, Dr. Stevison is building a highly collaborative and active computational biology lab, and is actively advising students in related areas of inquiry. This project will require significant computational resources, and poses unique programming challenges. Dr. Stevison has expertise in both speciation genetics and computational biology, making her an excellent sponsor for the proposed project. In addition, her lab works as a team to solve complex problems in computational biology, and I have already developed a working rapport with the lab. Finally, Dr. Stevison is significantly committed to the broader impact activities proposed here, and has proposed similar activities in a recent NSF proposal. Thus, the Stevison lab is an ideal place for me to conduct the proposed research, and continue to develop as a scientist. In addition, this postdoc represents a significant departure from my PhD training in quantitative Stream Ecology, but Dr. Stevison has been and will continue to be an excellent mentor for my new area of inquiry.

Host Institution: Auburn University will provide an excellent place to conduct this fellowship, and the biological sciences department offers a wide range of expertise with a core area being in evolutionary genetics. Additionally, due to its location in the Southern US, Auburn offers a unique location for my broader impact goals to reach targeted underrepresented groups in STEM. Although I received my PhD from Auburn University in Stream Ecology, I have chosen Auburn as the host institution for the proposed fellowship because I have greatly changed my area of inquiry from my PhD studies. Dr. Stevison started her lab group in the final year of my PhD and her work offers an excellent opportunity to **significantly broaden my biological training** from classical community ecology (PhD advisor: Dr. Jack Feminella) to speciation genomics. As a result, I believe that my choice of mentor and host institution are the best choice for my career development and proposed research.

6. PROJECTED TIME TABLE

		Y1	**Y 2**
Aim 1	Compile/Download Sequence Data	X	
	Develop Data Processing Pipeline, and write custom software	X	
	Calculate D statistic whole genome	X	
Aim 2	Call Variants		X
	Calculate D statistics in sliding windows		X
	GO analysis		X
BI	Speciation at the junior curator camp	X	X
	Online bioinformatics lesson		X
	Bringing online and traditional collections together – python script	X	X
	Submit Manuscripts		X