Population Genetics and a Study of Speciation Using Next-Generation Sequencing: An Educational Primer for Use with "Patterns of Transcriptome Divergence in the Male Accessory Gland of Two Closely Related Species of Field Crickets"

Patricia J. Wittkopp[1]
Department of Ecology and Evolutionary Biology, and Department of Molecular, Cellular, and Developmental Biology,
University of Michigan, Ann Arbor, Michigan 48109

SUMMARY Understanding evidence for the genetic basis of reproductive isolation is imperative for supporting students' understanding of mechanisms of speciation in courses such as Genetics and Evolutionary Biology. An article by Andrés *et al.* in the February 2013 issue of *GENETICS* illustrates how advances in DNA sequencing are accelerating studies of population genetics in species with limited genetic and genomic resources. Andrés *et al.* use the latest sequencing technologies to systematically identify and characterize sites in the DNA that vary within, and have diverged between, species to explore speciation in crickets. This primer, coupled with that article, will help instructors introduce and reinforce important concepts in genetics and evolution while simultaneously introducing modern methodology in the undergraduate classroom.

Related article in *GENETICS*: Andrés, J. A., E. L. Larson, S. M. Bogdanowicz, and R. G. Harrison, 2013 Patterns of transcriptome divergence in the male accessory gland of two closely related species of field crickets. Genetics **193:** 501–513.

Background

Some of the fastest evolving proteins in insect genomes are transferred from males to females in their seminal fluid (Findlay and Swanson 2010). Amino acid sequence divergence in these proteins can contribute to speciation by creating incompatibilities between sperm and eggs when two distantly related genotypes mate. Based on these observations, it has been proposed that divergent seminal fluid proteins might contribute to the early stages of speciation in closely related species of field crickets Andrés *et al.* (2006). Andrés *et al.* (2013) explore this hypothesis by characterizing DNA sequence variation in the coding regions of proteins expressed in accessory glands where seminal proteins are found.

Study system: Gryllus firmus and Gryllus pennsylvanicus

The senior author of the Andrés *et al.* (2013) study, Richard G. Harrison, began investigating speciation in crickets in the mid-1970s using protein electrophoresis techniques. Since then, he and his colleagues have examined properties of cricket mitochondrial DNA (Harrison *et al.* 1985; Rand and Harrison 1986, 1989), studied hybrid zones where different species of crickets meet (Willett *et al.* 1997; Ross and Harrison 2002; Maroja *et al.* 2009), and investigated whether the bacterial parasite *Wolbachia* (Mandel *et al.* 2001; Maroja *et al.* 2008) and divergent reproductive proteins (Andrés *et al.* 2006, 2008) contribute to speciation in crickets. Most recently, their studies have focused on *G. firmus* and *G. pennsylvanicus*, a pair of closely related species of field cricket found in the eastern region of North America. *G. firmus* is found on sandy soils along the east coast, and *G. pennsylvanicus* is found on firmer inland loam soils. Hybrids resulting from mating between these two species are found along the eastern edge of the Appalachian Mountains (Figure 1). When *G. pennsylvanicus* females mate

doi: 10.1534/genetics.112.148171
[1] Address for correspondence: 1061 Natural Science Bldg., 830 N. University Ave., University of Michigan, Ann Arbor, MI 48109-1048. E-mail: wittkopp@umich.edu

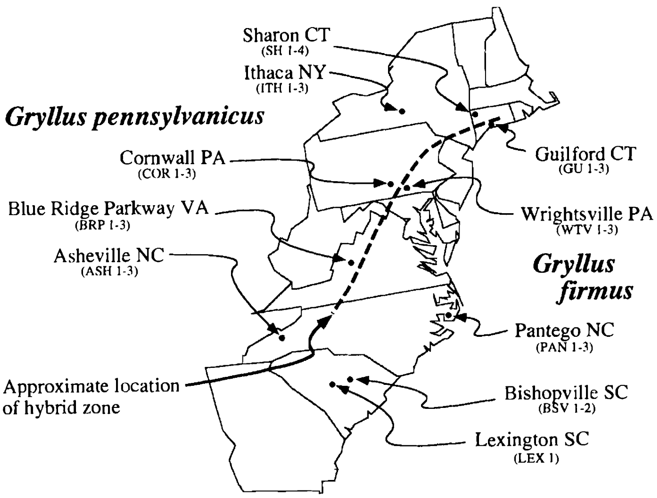

Figure 1 Collection sites (black circles) for *Gryllus pennsylvanicus* and *Gryllus firmus* are shown along with the approximate location of the hybrid zone (dotted line) between these two species. Reproduced with permission from Willett *et al.* (1997).

with *G. firmus* males, viable and fertile offspring are produced; however, when *G. pennsylvanicus* males mate with *G. firmus* females, no hybrid offspring are observed (Harrison 1985).

Using sequence variation to identify genes contributing to speciation

G. firmus and *G. pennsylvanicus* are thought to have become different species via allopatric speciation (Willett *et al.* 1997). This is a mode of speciation in which populations of the same species are geographically isolated from each other and thereby evolve genetic differences over time that reduce their ability to produce fully fertile offspring when their ranges overlap. When the isolation initially occurs, the two populations share alleles throughout the genome. Over generations, mutation, drift, and selection cause genetic differences to accumulate between the two populations. Most of these genetic differences will have no effect on morphology, physiology, or the ability of individuals from one population to interbreed with individuals from the other population—but some will affect one or more of these traits.

When hybridization between two species occurs and fertile offspring are produced, alleles are moved between species. This mixing of genetic information, or gene flow, causes introgression that reduces the genetic divergence between species. Regions of the genome contributing to reproductive isolation, however, cannot be exchanged since reproductively incompatible individuals cannot mate. This leads to the "mosaic of different evolutionary histories" across the genome described by the authors. Gene flow through a hybrid zone accentuates regions of the genome contributing to reproductive isolation by decreasing differences in DNA sequence elsewhere in the genome. Based on the idea that divergent seminal fluid proteins have played an important role in speciation, Andrés *et al.* (2013) hypothesized that they would see greater sequence divergence between *G. firmus* and *G. pennsylvanicus* in genes encoding seminal fluid proteins than in genes encoding other proteins expressed in the same tissue (the accessory gland).

Sequencing targeted regions of the genome using the transcriptome

Virtually all cells within an organism contain the same set of genes, but different subsets of genes are expressed (*i.e.*, transcribed) in each cell type. These differences in gene expression make one cell type different from another. Because Andrés *et al.* (2013) are interested in the evolution of genes that can disrupt interactions between sperm and eggs, they examined genes expressed in the male accessory gland. That is, they sequenced the accessory gland transcriptome, the subset of sequences in the genome that are transcribed to produce RNA in accessory glands. Note that the word "transcriptome" has also been used by other authors to refer to the collection of RNAs produced from all transcribed sequences.

To isolate DNA sequences coding for proteins found in the accessory gland, Andrés *et al.* (2013) dissected accessory glands from *G. firmus* and *G. pennsylvanicus* males, extracted RNA from each tissue sample, and used reverse transcriptase to synthesize DNA complementary to the RNA (termed "cDNA"). Consequently, the cDNA samples contained only DNA sequences from genes that were transcribed into RNA. But sequences from each expressed gene are not represented equally in cDNA: some genes are expressed at higher levels (*i.e.*, transcribed more times into RNA molecules) than others, and these differences in RNA abundance are reflected in the population of cDNA. If the researchers had directly sequenced this cDNA, they would have recovered sequences from highly expressed genes many times before observing sequences from genes with lower expression.

To make the representation of all expressed genes more similar in the cDNA sample, a technique called "normalization" was used. After using a DNA polymerase to turn the single-stranded cDNA molecules synthesized by reverse transcriptase into double-stranded DNA, heat is used to separate the complementary strands. As the sample is cooled, complementary DNA sequences find each other again and anneal, recreating fragments of double-stranded DNA. Transcripts that are the most abundant in the sample are most likely to find a complementary sequence quickly and form double-stranded cDNA first. By treating the sample with an enzyme that specifically degrades double-stranded DNA, sequences from the high-abundance transcripts are selectively eliminated. This creates a cDNA sample in which transcribed sequences from all expressed genes are present at similar frequencies.

Unpacking the Difficult Bits

Combining results from multiple sequencing technologies

During the past decade, advances in DNA sequencing have revolutionized the way biological research is performed.

Table 1 Samples and sequencing used by Andrés *et al*. (2013).

Biological sample(s)	Sequencing technology	Purpose
Accessory gland RNA from 10 *G. firmus* males	Sanger	Assemble transcriptome
Accessory gland RNA from a single *G. firmus* male	454/Roche	Assemble transcriptome
Accessory gland RNA from 15 *G. firmus* males	Illumina	Find SNPs; quantify allele frequencies
Accessory gland RNA from 15 *G. pennsylvanicus* males	Illumina	Find SNPs; quantify allele frequencies
16 genomic DNA samples from individual *G. firmus*	Sanger	Verify allele frequencies; infer gene genealogies
16 genomic DNA samples from individual *G. pennsylvanicus*	Sanger	Verify allele frequencies; infer gene genealogies

Instead of sequencing only one DNA fragment at a time, methods are now available to sequence billions of DNA fragments simultaneously. This makes it feasible for individual researchers to sequence the genomes of their favorite organisms as well as to survey genomic variation within and between closely related species. Andrés et al. (2013) took advantage of these advances by combining traditional single-fragment ("Sanger") sequencing with two next-generation sequencing techniques that perform massively parallel sequencing ("454" and "Illumina") (Table 1). One reaction of Sanger sequencing produces a single continuous sequence, typically ~800 bp long, whereas one run of 454 (also called Roche) sequencing produces millions of sequences each ~500 bp long and one run of Illumina sequencing produces billions of sequences each ~100 bp long. Illumina sequencing also has the ability to perform "paired-end" sequencing in which ~100-bp sequences are collected from both ends of a larger fragment, and information about the distance between these sequences is retained.

Longer reads from Sanger and 454 sequencing were used to assemble the accessory gland transcriptome of crickets *de novo* (from scratch). They generated enough of these sequencing reads to cover each site in the transcriptome an average of four times (4× coverage). Sequence assembly was accomplished by finding fragments of sequence that shared identical bases and could be overlapped (aligned) to form "contigs," stretches of uninterrupted DNA sequence longer than individual sequence reads. If all sequence reads assembled perfectly, each transcript would correspond to a single contig. In reality, problems with sequence assembly (often due to repetitive sequences that occur at multiple places in the genome) result in some transcripts being assigned to multiple contigs. The assembled set of contigs for the accessory gland transcriptome was compared to genes previously identified in crickets as well as to genes previously identified in other insect genomes and was used as a reference to assign short reads generated by Illumina sequencing to individual contigs and/or genes. Illumina sequencing reads from *G. firmus* and *G. pennsylvanicus* were aligned separately to this reference genome, and software was used to identify single nucleotide polymorphisms (SNPs), or single base changes, within each species. Sanger sequencing of genomic DNA extracted from 32 individual crickets (16 *G. firmus* and 16 *G. pennsylvanicus*) was then used to validate the frequency of SNPs detected by Illumina sequencing in samples derived from mixing equal amounts of RNA from 15 *G. firmus* or 15 *G. pennsylvanicus* male accessory glands.

Quantifying and interpreting differences in DNA sequence

Andrés *et al.* (2013) were most interested in identifying fixed differences between species that might contribute to reproductive isolation. "Fixed differences" refers to sites in the genome at which all *G. firmus* individuals have one nucleotide and all *G. pennsylvanicus* individuals have another. The authors began by identifying all sites that showed differences in the frequency of alternative alleles between species. To avoid interpreting sequencing errors as polymorphisms, they considered only sites that (1) were classified as high quality by the sequence analysis software, (2) were present in at least 20 sequencing reads per species, and (3) showed the rare allele in at least 1.0% of all reads. After compiling this list of variable sites, they calculated a statistic called "D" that captures the divergence in allele frequency between *G. firmus* and *G. pennsylvanicus*; the more divergent the two species are at a site, the larger the value of D at that site. For each contig, the average value of D was computed and used as a measure of sequence divergence for that gene. The greater the average value of D, the more evidence that the gene corresponding to that contig might be contributing to isolation between species.

Other descriptors of DNA sequence variation analogous to commonly used parameters in population genetics were also calculated, including the ratio of nonsynonymous to synonymous polymorphisms ($p\text{N}/p\text{S}$). Nonsynonymous polymorphisms are those that alter the amino acid encoded by a codon, whereas synonymous polymorphisms are those that do not. A similar statistic calculated from divergent sites, $d\text{N}/d\text{S}$, is commonly used to test coding sequence for evidence of natural selection. Values of $d\text{N}/d\text{S} > 1$ are assumed to result from adaptive evolution because the rate of fixation for mutations that change the amino acid sequence is not expected to exceed the rate of fixation for mutations that do not change the amino acid sequence unless natural selection favors divergent protein functions. It is important to note, however, that interpretations of $p\text{N}/p\text{S}$ values derived from sites that have multiple alleles within each population are not as straightforward (Kryazhimskiy and Plotkin 2008).

Andrés *et al.* (2013) also calculated the average number of nucleotide differences (π) between pairs of alleles from the same species for each contig. These measures of polymorphism describe sequence variation within *G. firmus* and *G. pennsylvanicus*. In figure 3 of Andrés *et al.* (2013), measures of π are compared to measures of divergence (average D) for each contig. Comparisons between polymorphism and divergence are often made in population genetics and molecular evolution as another way to test for evidence of adaptive evolution. In the absence of positive selection, neutral evolution should cause a correlation between polymorphism and divergence such that genes that vary more within species also vary more between species. Gene sequences with low levels of polymorphism (within a species) and high levels of divergence (between species) are considered candidates for genes that have recently been subject to strong positive selection.

Gene genealogies: inferring haplotype trees

After characterizing sequence divergence of all contigs between *G. firmus* and *G. pennsylvanicus* using Illumina sequencing, regions of the genome corresponding to 10 of the most divergent contigs were Sanger sequenced in 16 individuals from each species. Haplotypes (continuous sequences for each allele) from all 32 alleles were aligned, and variable sites within these alignments were used to generate gene trees according to the neighbor-joining method (Saitou and Nei 1987). These trees are shown in figure 4 of Andrés *et al.* (2013).

In the neighbor-joining method, as with all such DNA-based phylogenetic tree-building approaches, haplotypes with the most sequence similarity are assumed to be the most closely related, and therefore are connected by branches with the shortest lengths on the tree. Confidence in specific branch points on each tree is determined using a technique called bootstrapping, in which subsets of variable sites from each alignment are resampled (with replacement) hundreds or thousands of times and used to produce hundreds or thousands of individual trees. The bootstrap value of a branch is the percentage of times that that particular branch was seen in the entire set of these trees. Higher bootstrap values indicate higher confidence. In other words, the more often a particular branch is seen in trees generated by bootstrapping, the more likely it is that this branch is a "correct" representation of the gene genealogy. Haplotypes were color-coded white and black for *G. firmus* and *G. pennsylvanicus*, respectively, on these trees to show how DNA sequence variation is partitioned between the two species for each contig.

Connections to Genetics Concepts

This article integrates many topics commonly taught in introductory genetics or evolutionary biology courses, including transcription, mutations, population genetics, evolutionary genetics, and genomics. Consequently, it might be best discussed near the end of the course after all of these ideas have been introduced. All too frequently, population and evolutionary genetics are among the most out-of-date sections in genetics textbooks. Discussing the Andrés *et al.* (2013) article will show students how fundamental principles presented in the textbook (*e.g.*, allele frequencies, gene flow, polymorphism, divergence, speciation, phylogenetic trees) are used in contemporary research. It will also show how the mechanisms of speciation and reproductive isolation, which are notoriously difficult to study yet fundamentally important to biology, can be investigated. Finally, although the sequencing methods used in the Andrés *et al.* (2013) article might be unfamiliar to students, they are good to introduce given that these methods are now used in virtually all areas of genetics.

Approach to Classroom Use

Discussing the primary literature with undergraduate students can be challenging for both instructors and students. Students are charged with decoding complex and often unfamiliar technical language while they are still trying to learn the fundamental concepts, whereas instructors need to ensure that students are sufficiently prepared for reading the assigned article and feel comfortable discussing it with their peers. To help with both of these tasks, instructors are encouraged to provide this primer to students along with Andrés *et al.* (2013).

Instructors may wish to assign not only reading before class, but also a core set of discussion questions that students are required to answer before class. Students bring copies of their answers to class and could be required to correct and/or expand upon their preclass answers during in-class discussion. Discussion can take place in one large group, in multiple small groups, or in both depending on the class size and structure. This assignment can be graded based on effort in the preclass answers and/or correctness/completeness of the final answers. From the instructor's perspective, this strategy holds students accountable for class preparation as well as for staying engaged during the discussion. From the students' perspective, the discussion questions help them focus on key aspects of the article when preparing for class and allow them time to develop thoughtful answers that they feel confident sharing during class. Open-ended discussion questions work best with this format. Asking students to list unfamiliar vocabulary while reading and then working together as a group to define them is a great way to get initial conversations going. Some potential questions for discussion of Andrés *et al.* (2013) are provided below.

Questions for Further Exploration

1. Describe the similarities and differences between genomic DNA, RNA, and cDNA sequences for a gene with one exon. How does your answer change if the gene has more than one exon?
2. To assess the quality of their DNA sequencing, the authors compared the frequency of transitions and transversions observed among single nucleotide polymorphisms (SNPs). Why did they expect sequencing errors to show a transition:transversion ratio of 1:2? (Hint: consider the types of mutations classified as transitions and transversions.)

Assuming no sequencing errors, and given that they observed an average of 1.55 transitions for every transversion, what can you conclude about mutation and/or selection in natural populations of these field crickets?

3. The authors used Sanger and 454 sequencing of cDNA from *G. firmus* to establish a reference transcriptome and then aligned Illumina sequencing reads from both *G. firmus* and *G. pennsylvanicus* to this sequence. How might sequence differences between the two species complicate their results?
4. The authors sequenced large regions of the cricket genome for the first time. How did they figure out which of the sequences that they recovered from accessory gland RNA were likely to encode seminal fluid proteins? What assumptions underlie the methods they used?
5. What are the advantages and disadvantages for understanding cricket speciation by sequencing the transcriptome from accessory glands as opposed to whole crickets? How do you expect the transcriptomes from these two samples to differ?
6. Figure 4 in Andrés *et al.* (2013) shows haplotype networks for the 32 alleles observed by Sanger sequencing of 16 individuals each from *G. firmus* and *G. pennsylvanicus*. Which of the 10 contigs analyzed in this study shows the most evidence of completely fixed differences between *G. firmus* and *G. pennsylvanicus*? Of the two previously analyzed seminal fluid proteins shown in figure 4 of Andrés *et al.* (2013), which shows more evidence of gene flow between species? Explain your reasoning for both answers.
7. Assuming that transgenic analysis is possible in crickets (which it is not currently), design an experiment to determine whether one of the highly divergent accessory gland proteins identified in this study actually contributes to isolation between *G. firmus* and *G. pennsylvanicus*.
8. The authors state that their ultimate goal “is to understand the genetics of speciation, not simply the genetics of species differences.” Are their data sufficient to differentiate between the two? Why or why not?
9. The authors use a candidate gene strategy as well as a genome-scan approach to investigate differences between the two species. Compare and contrast these different methods, identifying the benefits and limitations of each.

Acknowledgments

I thank Brian Metzger for leading, and Anna Arias, Joseph Coolon, Wei-Chin Ho, Ling Huang, Alisha John, Chuan Li, Alexander Taylor, and Ya Yang for participating in, a discussion of Andrés *et al.* (2013) that helped shape the questions and background topics included in this Primer.

Literature Cited

Andrés, J. A., L. S. Maroja, S. M. Bogdanowicz, W. J. Swanson, and R. G. Harrison, 2006 Molecular evolution of seminal proteins in field crickets. Mol. Biol. Evol. 23: 1574–1584.

Andrés, J. A., L. S. Maroja, and R. G. Harrison, 2008 Searching for candidate speciation genes using a proteomic approach: seminal proteins in field crickets. Proc. Biol. Sci. 275: 1975–1983.

Andrés, J. A., E. L. Larson, S. M. Bogdanowicz, and R. G. Harrison, 2013 Patterns of transcriptome divergence in the male accessory gland of two closely related species of field crickets. Genetics 193: 501–513.

Findlay, G. D., and W. J. Swanson, 2010 Proteomics enhances evolutionary and functional analysis of reproductive proteins. Bioessays 32: 26–36.

Harrison, R. G., 1985 Barriers to gene exchange between closely related cricket species. II. Life cycle variation and temporal isolation. Evolution 39: 244–259.

Harrison, R. G., D. M. Rand, and W. C. Wheeler, 1985 Mitochondrial DNA size variation within individual crickets. Science 228: 1446–1448.

Kryazhimskiy, S., and J. B. Plotkin, 2008 The population genetics of dN/dS. PLoS Genet. 4: e1000304.

Mandel, M. J., C. L. Ross, and R. G. Harrison, 2001 Do *Wolbachia* infections play a role in unidirectional incompatibilities in a field cricket hybrid zone? Mol. Ecol. 10: 703–709.

Maroja, L. S., M. E. Clark, and R. G. Harrison, 2008 *Wolbachia* plays no role in the one-way reproductive incompatibility between the hybridizing field crickets *Gryllus firmus* and *G. pennsylvanicus*. Heredity 101: 435–444.

Maroja, L. S., J. A. Andrés, and R. G. Harrison, 2009 Genealogical discordance and patterns of introgression and selection across a cricket hybrid zone. Evolution 63: 2999–3015.

Rand, D. M., and R. G. Harrison, 1986 Mitochondrial DNA transmission genetics in crickets. Genetics 114: 955–970.

Rand, D. M., and R. G. Harrison, 1989 Molecular population genetics of mtDNA size variation in crickets. Genetics 121: 551–569.

Ross, C. L., and R. G. Harrison, 2002 A fine-scale spatial analysis of the mosaic hybrid zone between *Gryllus firmus* and *Gryllus pennsylvanicus*. Evolution 56: 2296–2312.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406–425.

Willett, C. S., M. J. Ford, and R. G. Harrison, 1997 Inferences about the origin of a field cricket hybrid zone from a mitochondrial DNA phylogeny. Heredity 79: 484–494.

Communicating editor: Elizabeth A. De Stasio