Genetics Society of America
PREP
Peer-Reviewed Education Portal

Finding Selection in All the Right Places: A College Genetics Laboratory Inquiry-Based Learning Exercise

Juliet K. F. Noor* and Mohamed A. F. Noor
Duke University Department of Biology
Box 90338
Durham, NC 27708

**Corresponding Author: jkfnoor@duke.edu*

Accepted for publication February 7, 2013

Citation:

> Noor, J. K. F. and Noor, M. A. F. (2013). Finding Selection in All the Right Places: A College Genetics Laboratory Inquiry-Based Learning Exercise. *Genetics Society of America Peer-Reviewed Education Portal (GSA PREP)*: 2013.001; doi: 10.1534/gsaprep.2013.001

Introduction:

Many college faculty are transitioning their course laboratory exercises from "cookbook" activities to inquiry-based learning exercises, where students are actively engaged in primary research (e.g., CHEN *et al.* 2005; SHAFFER *et al.* 2010). This change brings many advantages, including for example, a) students do not know *a priori* what results are expected, leading to greater interest, b) inquiry-based exercises frequently force students to deal with uncertainty, leading to decisions on application and interpretation rather than simple acquisition of results, and c) students contribute to the broader scientific enterprise, adding to their experience and benefiting the community. However, faculty sometimes struggle with developing inquiry-based exercises that are useful yet can be applied over multiple iterations.

We have developed an inquiry-based evolutionary genetics laboratory exercise that leverages the recent availability of population genomic sequence datasets. Many recent studies have published multiple aligned genome sequences from within single species and their close relatives (e.g., LANGLEY *et al.* 2012; LITI *et al.* 2009; MCGAUGH *et al.* 2012). These tremendous datasets are typically computationally aligned, annotated, and analyzed, and population genetic tests are then applied on the genome as a whole and/ or individual genes to test for natural selection. While these computational approaches provide an excellent "first look," manual curation and analysis sometimes identifies misalignments or misannotations that can alter interpretations of the evolutionary forces operating on individual genes. One of the most common sources of error with computational curation is the exact placement of alignment gaps associated with indels relative to a reference genome. Also, computationally annotated predicted genes from a reference can sometimes be falsified when additional sequences are obtained from other individuals bearing frequent or abundant internal stop codons.

With our activity, students assess the observed alignments and annotations of multiple genes from population genomic datasets. They repair alignment errors, and they then apply population genetic tests for natural selection (e.g., MCDONALD and KREITMAN 1991). They interpret the results, and, if possible, hypothesize about the nature of forces that may be operating in the context of the function information available for those genes. With our study organism, *Drosophila pseudoobscura* and *D. miranda,* most students obtain some statistically significant results, and almost half in each class identify what may be a signature of positive selection. The results can be collated and disseminated in many ways: written publication, oral presentation at conferences, development of an online database, etc. A subset of students at research-oriented universities may even find opportunities to follow up on their findings with an independent study project, but this is not required for the activity to meet the learning objectives.

We developed this activity for our Genetics and Evolution course. It is one of the first two biology courses that our biology majors are required to take and so has no pre-requisites. However, since it focuses on just transmission genetics and evolution, it does cover some advanced as well as basic concepts in these areas. The lecture portion of the course introduces the students to the concepts that the students focus on in lab, as well as the more basic ones that are building blocks for these concepts.

The particular learning objectives are for all the participating students to be able to:

1. Explain the need for, and process of, DNA sequence alignment before analysis of the sequence.
2. Use statistical tests on DNA sequence data to interpret the nature of past evolutionary events, such as natural selection favoring rapid evolutionary change.
3. Form hypotheses for the observed presence or absence of evidence of natural selection on individual genes based on results of objective number 2 above.

Below, we apply this approach to assembled, aligned, and annotated genome sequences of *Drosophila pseudoobscura* and its sister species *D. miranda* (McGaugh et al. 2012; McGaugh and Noor 2012).

Approach:

Step 1: Analyze the alignment of the DNA sequences.
Step 2: Perform two versions of the McDonald-Kreitman test on the sequences and interpret the results.
Step 3: Research two of the genes to form hypotheses to explain the results of the tests.

As homework before the lab period, students complete an assignment looking up some basic information in an organism-specific database (FlyBase: McQuilton *et al.* 2012, in our recent iterations) and interpreting McDonald-Kreitman (1991) results presented in the format they will see in lab. This familiarizes the students with the interfaces they will use in lab.

The lab activity starts with a short exercise illustrating the importance of DNA sequence alignment. Specifically, students witness how alignment allows a researcher to identify homologous sites across individuals in order to document base changes, and they see how insertions and deletions (indels, collectively) present a challenge to this process.

After completing this introductory activity, the students work in pairs to analyze the alignment of five unique genes they have been assigned using the free software Mega (Tamura *et al.* 2011). They confirm that the computational alignment was successful, and occasionally perform manual corrections if possible without introducing additional gaps. The students also pay particular attention to DNA sequences that have an excessive number of unidentified nucleotides, premature stop codons, or indels that result in a frameshift of the predicted amino acid sequence. The McDonald-Kreitman test has to ignore codon positions in which even one strain has an unidentified nucleotide, so an excessive number of these in any one strain's sequence decreases the power of the test. Similarly, a premature stop codon or indel that results in the frameshift of the protein indicates either a sequencing error, in which case the sequence is unreliable, or that the gene is potentially non-functional in that particular strain. In either case, these sequences arguably should not be included in the McDonald-Kreitman test, though they may be logged for other sorts of investigations (e.g., Hoehn *et al.* 2012).

After analyzing and re-aligning the sequences for all five of their genes, the students input the sequences into the Standard & Generalized McDonald-Kreitman Test website at http://mkt.uab.es . The students then generate results from two different versions of the McDonald-Kreitman test. The first is the standard McDonald-Kreitman test that compares the ratio of nonsynonymous (resulting in a different amino acid) to synonymous (any nucleotide substitution resulting in the same amino acid) nucleotide substitutions in different strains within a species to the same ratio between different species. The second test is a modified and more conservative version of the test that compares the same ratios but only considers four-fold degenerate sites (where any nucleotide substitution will result in no change to the amino acid) as synonymous. This corrects for any possible selective biases associated with two- and six-fold degenerate codons. Alternatives to this modified test that we are currently exploring are having the students calculate the dN/dS ratio (Hughes and Nei 1988), or carry out the Fay and Wu test (Fay and Wu 2000), or the McDonald-Kreitman test considering only abundant variation within species (Charlesworth and Eyre-Walker 2008).

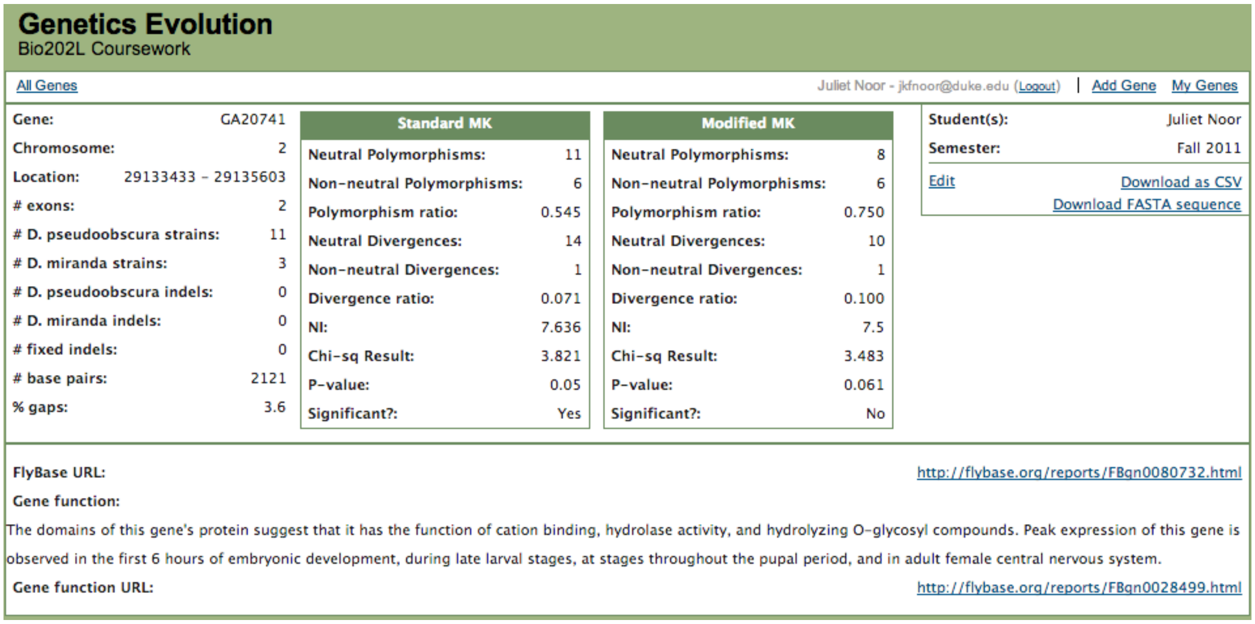
Once the students have collated their results, they interpret them to determine whether there is evidence that selection was acting on their genes. If there is statistically significant evidence of natural selection, they interpret whether selection seems to have been acting to maintain the same amino acid sequence (negative selection) or whether it has been driving a change in the amino acid sequence (positive selection) causing divergence between the species. Students enter all of their results in a publicly accessible class database.

Finally, students pick two of their five genes, usually two with interesting McDonald-Kreitman results, to research further. They look up the function of these two genes in the organism-specific database (FlyBase) and then interpret their McDonald-Kreitman results further in light of the predicted protein function. Specifically, why might there not be evidence of selection acting on this gene, why might selection be acting to maintain the amino acid sequence, or why might selection be acting to cause divergence in this amino acid sequence? They are also asked to design an experiment to test this hypothesis. Given that we have employed this exercise at a research university and using model organisms such as flies and yeast, they are encouraged to seek an independent research opportunity to potentially pursue their hypothesis directly.

Sample Results:

We have employed versions of this activity with over 1100 students in our introductory genetics and evolution course thus far, spanning five semesters. Figure 1 shows a sample of the results from our class. Pictured is the detailed results page for a single gene from our database (http://geneticsevolution.biology.duke.edu). Note that the individual student (or students) who contributed the data are named so that they are credited with, and held accountable for, their contribution. This database is publicly accessible, and so functions as a form of publication of the data. While others adopting this exercise would not be able to contribute to this particular database, their results could be collated and disseminated in many ways: written publication, oral presentation at conferences, development of their own online database, etc.

In our most recent iterations, approximately 11% of genes had some need to realign, so, on average, about 45% of students had at least one gene that required realignment and thus contributed beyond what using computational approaches alone provided. Approximately 33% of genes analyzed by our students had statistically significant deviations from neutrality, so over 99% of the students had at least one gene with a statistically significant signature of natural selection. In addition, about 10% of the genes exhibited a statistically significant result indicating positive selection, so 40% of students on average had at least one gene showing this very interesting result. Of course, these results are very specific to the dataset and so will vary greatly when other taxa are used for the analysis.



Genetics Evolution
Bio202L Coursework

All Genes

Juliet Noor - jkfnoor@duke.edu (Logout) | Add Gene My Genes

Gene:	GA20741
Chromosome:	2
Location:	29133433 - 29135603
# exons:	2
# D. pseudoobscura strains:	11
# D. miranda strains:	3
# D. pseudoobscura indels:	0
# D. miranda indels:	0
# fixed indels:	0
# base pairs:	2121
% gaps:	3.6

Standard MK	
Neutral Polymorphisms:	11
Non-neutral Polymorphisms:	6
Polymorphism ratio:	0.545
Neutral Divergences:	14
Non-neutral Divergences:	1
Divergence ratio:	0.071
NI:	7.636
Chi-sq Result:	3.821
P-value:	0.05
Significant?:	Yes

Modified MK	
Neutral Polymorphisms:	8
Non-neutral Polymorphisms:	6
Polymorphism ratio:	0.750
Neutral Divergences:	10
Non-neutral Divergences:	1
Divergence ratio:	0.100
NI:	7.5
Chi-sq Result:	3.483
P-value:	0.061
Significant?:	No

Student(s):	Juliet Noor
Semester:	Fall 2011

Edit Download as CSV
Download FASTA sequence

FlyBase URL: http://flybase.org/reports/FBgn0080732.html

Gene function:

The domains of this gene's protein suggest that it has the function of cation binding, hydrolase activity, and hydrolyzing O-glycosyl compounds. Peak expression of this gene is observed in the first 6 hours of embryonic development, during late larval stages, at stages throughout the pupal period, and in adult female central nervous system.

Gene function URL: http://flybase.org/reports/FBgn0028499.html

Figure 1. The detailed results page for a single gene (*GA20741* in this case) from our database.

Challenges:

Initially, we used fairly poor-quality aligned sequences, and the students were frustrated with the length of time and effort required to produce a reasonable alignment. When we switched to the higher-quality *Drosophila pseudoobscura* sequences, the activity was received far better by the students, but they still appreciated the need for small manual refinements of the computation alignments. Hence, we encourage anyone interested in this approach to focus on high-quality sequences and have high read coverage. Specifically, we suggest a minimum of 20X sequence coverage over all the genome. For the sequences themselves, we suggest using datasets that excluded bases with read coverage less than four or with Phred-scale consensus mapping score of less than 30.

A subset of students each semester adopt the mistaken impression that they're being "used" by a research laboratory in doing this activity – as though they're basically being forced to do our research for us. It's been essential to stress that this activity is designed ***primarily*** for their education and has secondary value as research that may be publishable in the future. Given the recent push for inquiry-based research, we did not anticipate this sort of backlash from students, but the few students who adopted this conspiracy-theory mentality were generally not students who were at the top of the class or who valued research endeavors.

Assessment of Student Learning & Interest:

Initially, students did not appreciate the somewhat tedious process of analyzing the sequences for problems, but we've refined this activity over 5 semesters, and the average response to the question “on a scale of 1-5, with 1 being not so much and 5 being very much, to what extent did this lab exercise help you to understand the course subject material and concepts?” rose from 3.19 to 3.77. For reference, the highest any of our lab exercises has scored on this question is 4.10. Importantly, 47-64% of students rated this exercise as the "most difficult or challenging" one of the semester. This observation suggests the students really did need more practice with these concepts. Although we do not have direct pre-/post-testing results, we are confident that the activity increased their understanding of the use and interpretation of this particular statistical test for evidence of selection at the molecular level.

Synopsis:

We've been quite satisfied with our implementations this exercise, and the potential to apply it more broadly will grow exponentially with the growth of genome sequencing efforts. Many research laboratories have already assembled databases of coding sequence polymorphism and divergence in taxa (from both model- and non-model organisms), and having introductory undergraduate students vet this data and analyze it independently for evidence of natural selection provides added pedagogical and research value to these endeavors.

References:

CHARLESWORTH, J., and A. EYRE-WALKER, 2008 The McDonald-Kreitman Test and Slightly Deleterious Mutations. Mol Biol Evol **25:** 1007-1015.

CHEN, J., G. B. CALL, E. BEYER, C. BUI, A. CESPEDES *et al.*, 2005 Discovery-based science education: functional genomic dissection in Drosophila by undergraduate researchers. PLoS Biol **3:** e59.

FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405-1413.

HOEHN, K. B., S. E. MCGAUGH and M. A. F. NOOR, 2012 Effects of Premature Termination Codon Polymorphisms in the Drosophila pseudoobscura Subclade. J Mol Evol **75:** 141-150.

HUGHES, A. L., and M. NEI, 1988 Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335:** 167-170.

LANGLEY, C. H., K. STEVENS, C. CARDENO, Y. C. G. LEE, D. R. SCHRIDER *et al.*, 2012 Genomic variation in natural populations of Drosophila melanogaster. Genetics **192:** 533-598.

LITI, G., D. M. CARTER, A. M. MOSES, J. WARRINGER, L. PARTS *et al.*, 2009 Population genomics of domestic and wild yeasts. Nature **458:** 337-341.

MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the Adh locus in Drosophila. Nature **351:** 652-654.

MCGAUGH, S. E., C. S. HEIL, B. MANZANO-WINKLER, L. LOEWE, S. GOLDSTEIN *et al.*, 2012 Recombination modulates how selection affects linked sites in Drosophila. PLoS Biol **10:** e1001422.

MCGAUGH, S. E., and M. A. F. NOOR, 2012 Genomic impacts of chromosomal inversions in parapatric Drosophila species. Philos Trans R Soc Lond B Biol Sci **367:** 422-429.

MCQUILTON, P., S. E. ST PIERRE, J. THURMOND and F. CONSORTIUM, 2012 FlyBase 101--the basics of navigating FlyBase. Nucleic Acids Res **40:** D706-D714.

SHAFFER, C. D., C. ALVAREZ, C. BAILEY, D. BARNARD, S. BHALLA *et al.*, 2010 The genomics education partnership: successful integration of research into laboratory classes at a diverse group of undergraduate institutions. CBE Life Sci Educ **9:** 55-69.

TAMURA, K., D. PETERSON, N. PETERSON, G. STECHER, M. NEI *et al.*, 2011 MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol **28:** 2731-2739.