

**Student writing reveals their heterogeneous thinking about the origin of genetic variation in populations**

Luanna B Prevost<sup>1</sup>, Jennifer K Knight<sup>2</sup>, Michelle K Smith<sup>3</sup>, and Mark Urban-Lurain<sup>1</sup>

<sup>1</sup> Center for Engineering Education Research, Michigan State University, East Lansing MI 48824

<sup>2</sup> Department of Molecular, Cellular and Developmental Biology and the Science Education Initiative, University of Colorado, Boulder, CO 80309

<sup>3</sup> School of Biology and Ecology and the Center for Research in Science, Technology, Engineering, and Math Education, University of Maine, Orono, ME 04469

This material is based upon work supported by the National Science Foundation (DUE 0717768). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

Correspondence should be sent to Luanna Prevost. Contact: [prevostl@msu.edu](mailto:prevostl@msu.edu)

**Reference:** Prevost, L. B., Knight, J. K., Smith, M. K., Urban-Lurain, M. (2013). Student writing reveals their heterogeneous thinking about the origin of genetic variation in populations. In Proceedings of the National Association for Research in Science Teaching (NARST) annual conference. Rio Grande, Puerto Rico.

### **Abstract**

Variation is a core concept of genetics which helps set the foundation for evolutionary thinking for undergraduate biology students. We investigated whether students rely on surface features or a deeper understanding of genetic variation by analyzing students' written explanations. Constructed response assessments allow students to demonstrate their thinking in their own words. Because constructed response assessments can be more difficult to analyze than multiple choice assessments, we employed computerized lexical analysis, which has been shown to reveal student thinking about complex biology concepts. For this study, we developed a pair of constructed response items asking students to explain the origin of new alleles in animal and bacterial populations. Our results show that some students relied on surface features and did not recognize origins of new alleles common to both populations. Fewer than five percent of the responses addressed mechanisms unique to bacterial populations. We also observed that students incorrectly identified natural selection as the predominant mechanism by which new alleles arise. Coupling constructed response assessment with lexical analysis revealed that students rely on surface level features to explain genetic variation and hold heterogeneous ideas (both correct and incorrect) about variation.

## **Student writing reveals their heterogeneous thinking about the origin of genetic variation in populations**

### **Introduction**

Well-constructed assessments can reveal the depth of understanding that students have about a scientific concept. Novice students often rely on surface features of an assessment to determine the principles they should apply in completing the assessment. For example, unlike physics experts, novices were not able to see past surface features to identify underlying scientific ideas (Chi, Feltovich, & Glaser, 1981). Similarly in biology, item features such as the trait under selection and the type of organism have been shown to influence students' explanations of natural selection (Heredia, Furtak, & Morrison, 2012; R. H. Nehm & Ha, 2011). In these studies, students were able to apply key concepts to natural selection in familiar organisms, but failed to apply them with unfamiliar organisms.

Ideally, assessments should also provide feedback to both instructors and students about student learning (Pellegrino, Chudowsky, & Glaser, 2001). Although multiple choice questions are easy to grade, constructed response questions provide instructors with greater insight into student understanding, as they give students the opportunity to represent their thinking in their own words (Birenbaum & Tatsuoka, 1987). In order to analyze the heterogeneity of ideas that students hold about scientific concepts when they answer constructed response questions, researchers have begun using computerized lexical analysis. These techniques use various algorithms to extract and categorize words and phrases within a larger block of text. Two recent examples of such an approach demonstrated students can hold mixed models about acid-base chemistry (Haudek, Prevost, Moscarella, & Merrill, 2012), and that students can incorporate both important constructs and misconceptions in their thinking about natural selection (Ha & Nehm, 2012).

We are exploring student understanding of genetic variation using constructed response items to allow students to articulate their thinking in their own words. Genetic variation describes the diversity of alleles within or among populations. In genetics, it is important that undergraduate students understand the concept of variation, as it is a core concept of natural selection (Mayr, 1963) and for evolutionary thinking. Furthermore, understanding this diversity is important in our changing global environment, as selection on genetic variation drives population adaptation (Orr, 2005). By varying the surface features of these items, we can examine the depth of student understanding of the origins of genetic variation. Therefore, we developed a constructed response item based on a question from the Genetics Concept Assessment (GCA), a multiple-choice assessment that examines key areas of student understanding of genetics, including student understanding of the origins of genetic variation (Smith, Wood, & Knight, 2008). We explore whether students identify origins of genetic variation by asking questions on the origin of new gene variants or alleles in populations. We also examine whether a surface feature – organism type – influences students' explanations.

### **Methods**

#### **Data collection**

We investigated undergraduate student thinking about genetic variation using two question forms (Figure 1). One form of the question asked how new alleles arise in an animal population; the other form asked how new alleles arise in a bacterial population. In Fall 2010, a

total of 348 responses to both question were collected from 233 students in an introductory genetics class at one large research university and 114 students in an introductory biology class at a second large research university.

Question 1. Explain how new alleles (gene versions or variants) arise in populations of animals  
 Question 2. Explain how new alleles (gene versions or variants) arise in a bacterial colony

Figure 1. Questions on the source of allele variation administered to students. Questions were administered in the order presented above.

### Computerized lexical analysis

We use the Text Mining node in IBM SPSS Modeler 14.2 (*IBM SPSS Modeler 14.2*, 2011) to analyze student responses. The software identifies *terms* from custom-built *libraries* (Haudek et al., 2012; Moscarella et al., 2008), similar to dictionaries. Terms are classified into *categories* by predefined computer algorithms that are subsequently modified by the researcher. For example the category *mutation* includes the terms *missense mutation*, *deletions* and *SNP* as well as the term *mutation*. Each response can contain multiple terms, with each term belonging to one or more categories.

### Statistical analysis

For each lexical category we compared responses to the two questions using McNemar's nonparametric related-samples tests for binary data (Sheskin, 2004). For each lexical category, there are two possible conditions: the category is either present in or absent from a given student response. We used a Bonferroni correction to adjust the alpha level to 0.0016 for the 30 comparisons performed.

### Results

We created 30 lexical categories to capture relevant ideas identified in students' responses (Figure 2). As shown in Figure 2, 18 of the 30 categories show significantly different frequencies for question 1 compared to question 2, ( $p < 0.0016$  indicated by an asterisk). We expected some differences in student responses since bacterial populations and animal populations do not share all processes by which new alleles arise. However, there were several categories that describe processes that apply to both populations for which the distributions were different in the students' answers: *mutation*, *random mutation*, *population*, *advantage*, *passed on to offspring*, and *gene*. This result may indicate that students do not recognize underlying concepts common to both types of organisms, but use surface features to construct their explanations.

We further examined the individual categories to identify differences in students' responses to the two forms of the question. The category *mutation* contained the most responses. Students identified various types of mutations including silent mutations, missense mutations, deletions and insertions. Students mentioned these types of mutations significantly more often when referring to animal populations than to bacterial populations ( $\bar{x}_{\text{animal}} = 0.64$ ,  $\bar{x}_{\text{bacteria}} = 0.52$ ,  $p < 0.001$ ). Responses assigned to the *mutation* category also were frequently assigned to the *passed on to offspring* category, as students often explained that mutations could be passed on to

the next generation. An example of such a response to the animal population question is shown below:

“Mutations within the genes of an individual's gametes can cause him or her to produce progeny with new alleles, which may then be spread throughout the population in subsequent generations, as the new allele is passed down from parent(s) to offspring. “

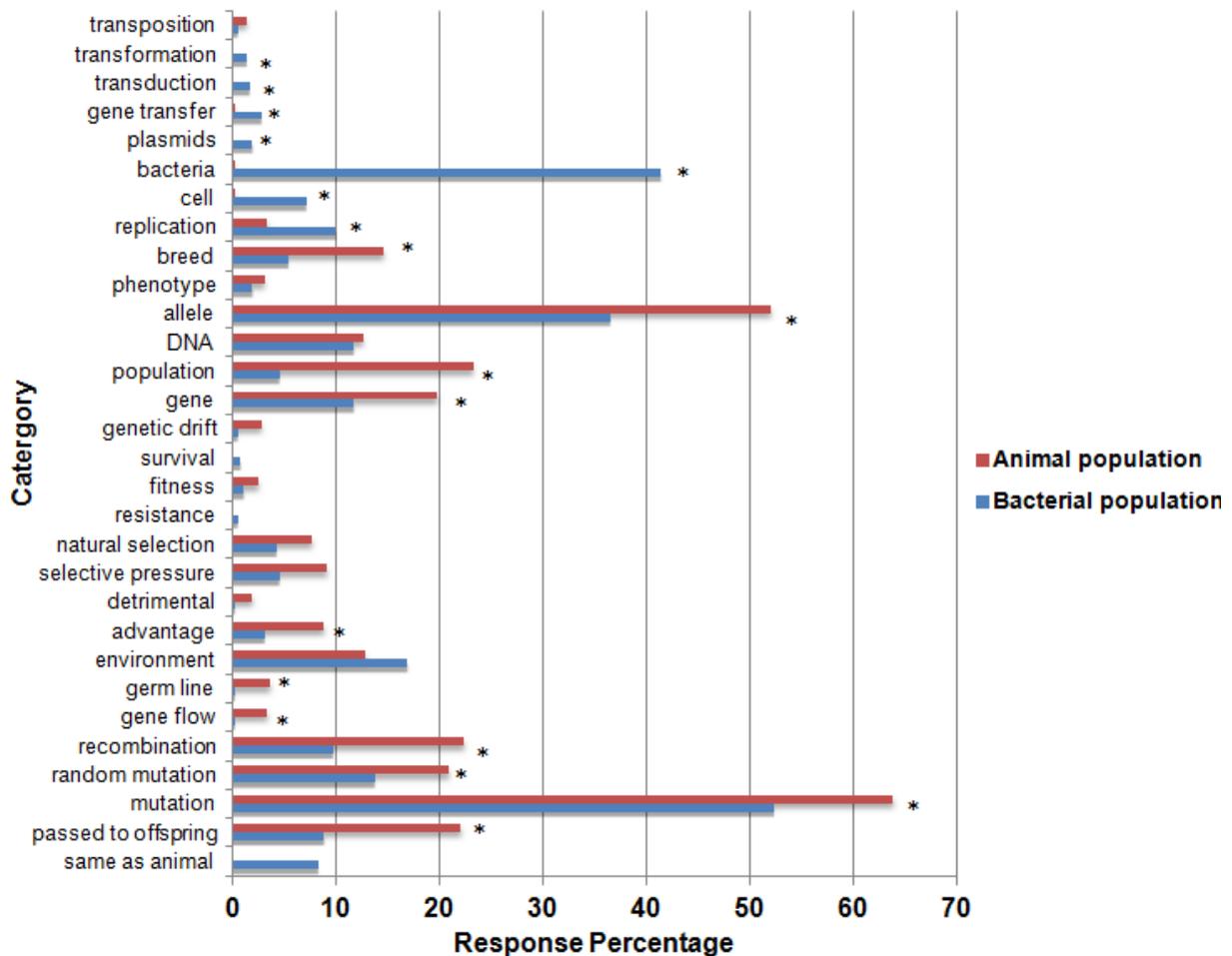


Figure 2. Distribution of responses among lexical categories.  
 Note: \* difference between Animal and Bacterial significant  $p < 0.0016$

Most students using the terms ‘random mutation’ or ‘spontaneous mutation’ identified mutation as the source of new alleles but did not focus on passing the mutation to subsequent generations. Therefore, these responses, placed in the category *random mutation*, were less likely to be classified in the *passed on to offspring* category. For example, one student responded:

“New alleles (or variants) arise in populations of animals due to spontaneous or induced mutations. Therefore, a random mutation occurs in the genome of an animal, and this mutation leads to the creation of a different protein. Once this occurs, the animal may have new phenotype based on the functionality of this

new type of protein. New alleles are NOT a function of the animal trying to adapt to its environment.”

Few students identified gene flow, migration or immigration as means of introducing new alleles to the population. The category *gene flow*, which comprises these terms, contained few responses. In addition, the *gene flow* category was more frequently assigned to responses for animal than bacterial population ( $\bar{x}_{\text{animal}} = 0.034$ ,  $\bar{x}_{\text{bacteria}} = 0.003$ ,  $p < 0.001$ ). A similar pattern was observed with the category *recombination*: this category was more frequently used to explain how new alleles arise in animal populations ( $\bar{x}_{\text{animal}} = 0.224$ ,  $\bar{x}_{\text{bacteria}} = 0.098$ ,  $p < 0.001$ ).

We were interested in determining whether students identified similar ways in which mutations arise in the two populations. Eight percent of students did so by explaining that new alleles arose in the bacterial population “in the same way as animal populations”. These responses were classified in the *same as animal* category, as they contained no other ideas to categorize.

Very few students identified mechanisms by which new alleles arise that were specific to bacterial populations. Less than 5% of the responses were classified in each of the following categories that are unique to bacteria: *gene transfer*, *plasmid*, *transduction* and *transformation*.

Some students also displayed some misunderstandings about how new alleles arise. They indicated *natural selection*, *selective pressure*, *fitness*, and *survival* as a source of new alleles rather than as processes that can change frequencies of already present alleles. Students were equally likely to use these terms to describe new alleles arising in bacterial and animal populations ( $p$  values range from 0.006 to 0.083). Student referred to mutations that have an advantage or benefit (*advantage* category) more frequently with animal than bacterial populations ( $\bar{x}_{\text{animal}} = 0.089$ ,  $\bar{x}_{\text{bacteria}} = 0.032$ ,  $p < 0.001$ ).

## Discussion

This study supports previous research demonstrating that students struggle with genetics concepts (Marbach-Ad, 2001; Smith & Knight, 2012) and that students may respond to surface features of assessment rather than address the underlying concepts (Chi et al., 1981; R. H. Nehm & Ha, 2011). In this study we demonstrate that students also have difficulty with surface features in genetic variation questions. For example, we would expect students to identify a mutation as a source of new alleles in *both* animal and bacterial populations. However, students defined mutation in this way significantly more often for animal populations than for bacterial populations, supporting the idea that students who have demonstrated understanding of a concept with one type of organism may fail to apply that concept when given an unfamiliar organism (Ross H. Nehm, Beggrow, Opfer, & Ha, 2012).

Students hold a variety of ideas about how new alleles arise, and how genetic variation arises in a population. We were able to demonstrate the heterogeneity of students’ ideas using lexical analysis of student answers to two different questions, one specifying animals, and the other specifying bacteria. Some students recognized that genetic variation arises in animal populations and bacterial populations through similar mechanisms such as mutation and gene flow. However, few students identified the mechanisms that were unique to bacterial populations, such as transduction, and conjugative plasmid transfer. Lexical analysis of student responses also revealed that students hold incorrect ideas about how new alleles arise in a population, including identifying natural selection and selective pressures as predominant

mechanisms. Our results suggest that students should be presented with more opportunities to examine the similarities and differences between sources or new alleles in bacterial and animal populations to help them develop a deeper understand of genetic variation concepts rather than relying on surface features. Overall we observe that computerized lexical analysis of student writing can help instructors gain a deeper understanding of how students think about genetic variation than traditional multiple choice questions.

#### **Future directions**

This work builds on existing studies of student writing about genetics concepts (Prevost, Knight, et al., 2012) and presents the first step in automating the analysis of student writing on the origin of allelic variation in population. In the next phase of analysis we will compare human coding of students responses with computerized coding to develop statistical models that illustrate relationship between lexical categories and human coding. We have created such models for other biology content (Haudek et al., 2012; Prevost, Haudek, Merrill, & Urban-Lurain, 2012), and have demonstrated that these models can be used to automatically rate student responses with inter-rater reliability (IRR) with expert human coding on par with expert-to-expert IRR. These models can facilitate the use and assessment of writing, particular in large enrollment courses.

### References

- Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats—it does make a difference for diagnostic purposes. *Applied Psychological Measurement*, 11(4), 385–395. doi:10.1177/014662168701100404
- Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive science*, 5(2), 121–152.
- Ha, M., & Nehm, R. H. (2012). Using machine-learning methods to detect key concepts and misconceptions of evolution in students' written explanations. In Proceedings of the National Association for Research in Science Teaching (NARST) annual conference. Indianapolis, IN.
- Haudek, K. C., Prevost, L. B., Moscarella, R. A., & Merrill, J. E. (2012). What are they thinking? Automated analysis of student writing about acid/base chemistry in introductory biology. *CBE Life Sciences Education*, 11(3), 283–293. doi: 10.1187/cbe.11-08-0084
- Heredia, S., Furtak, E. M., & Morrison, D. (2012). Item context: how organisms used to frame natural selection items influence student response choices. In Proceedings of the National Association for Research in Science Teaching (NARST) annual conference. Indianapolis, IN.
- IBM SPSS Modeler 14.2 User's Guide. 2011. IBM Corporation.
- Marbach-Ad, G. (2001). Attempting to break the code in student comprehension of genetic concepts. *Journal of Biological Education*, 35(4), 183–189.
- Mayr, E. 1963. *Animal species and evolution*. Cambridge, MA: Belknap Press of Harvard University Press.
- Moscarella, R. A., Urban-Lurain, M., Merritt, B. W., Long, T., Richmond, G., Merrill, J. E., Parker, J. M., Patterson R., & Wilson, C. (2008). Understanding undergraduate students' conceptions in science: using lexical analysis software to analyze students' constructed responses in biology. In Proceedings of the National Association for Research in Science Teaching (NARST) annual conference. Baltimore, MD.
- Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching*, 48(3), 237–256.
- Nehm, Ross H., Beggrow, E. P., Opfer, J. E., & Ha, M. (2012). Reasoning about natural selection: diagnosing contextual competency using the ACORNS instrument. *The American Biology Teacher*, 74(2), 92–98.
- Orr, H. (2005). Theories of adaptation: what they do and don't say. *Genetics of Adaptation*, 3–13.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: the science and design of educational assessment*. National Academies Press.
- Prevost, L. B., Haudek, K. C., Merrill, J. E., & Urban-Lurain, M. (2012). Examining student constructed explanations of thermodynamics using lexical analysis. In Proceeding of the Frontiers in Education Conference. Seattle, Washington.
- Prevost, L. B., Knight, J. K., Lai, E., Smith, M. K., Haudek, K. C., Merrill, J. E., & Urban-Lurain, M. (2012). Using lexical analysis to explore students' written responses to genetics concept assessment derived items. Presented at the Society for the Advancement of Biology Education Research National Meeting, Minneapolis, MN.
- Sheskin, D. (2004). *Handbook of parametric and nonparametric statistical procedures*. CRC Press.
- Smith, M. K., & Knight, J. K. (2012). Using the genetics concept assessment to document persistent conceptual difficulties in undergraduate genetics courses. *Genetics*, 191(1), 21–32.

Smith, M. K., Wood, W. B., & Knight, J. K. (2008). The Genetics Concept Assessment: a new concept inventory for gauging student understanding of genetics. *CBE Life Sciences Education*, 7(4), 422–430. doi:10.1187/cbe.08-08-0045