Supplemental Material

CBE—Life Sciences Education

Yang *et al.*

Supplemental Materials

Using Next-Generation Sequencing to Explore Genetics and Race in the High School Classroom

Xinmiao Yang,[†‡] Mark R. Hartman,[†‡] Kristin T. Harrington,[†] Candice M. Etson,[†] Matthew B. Fierman,[†] Donna K. Slonim,[§] and David R. Walt[†*]

[†]Department of Chemistry, Tufts University, Medford, MA 02155; [§]Department of Computer Science, Tufts University, Medford, MA 02155

[‡]These authors contributed equally to this work.

[*]Address correspondence to: David R. Walt (david.walt@tufts.edu)

Materials for the laboratory portion of the experiment

Equipment, consumables, and reagents used in the classroom
Hot water bath or hot plate (95° C)
Vortex mixer
Pipettes (for pipetting volumes between 20-200 µL)
Sterile pipette tips
Timer
Biohazard waste collection container
1.5 mL microcentrifuge tubes (sterile)
Catch-All Sample Collection Swab, soft pack (Epicenter, Cat# QEC89100)
Extract-N-Amp Tissue Kit (Sigma-Aldrich, Cat# XNAT2-1KT)

Reagents used in the sequencing center (for later-stagesample processing and sequencing)
Agencourt AMPure XP (Beckman Coulter, Cat# A63880)
Bioanalyzer DNA 1000 Kit (Agilent, Cat# 5067-1504)
Nextera Index Kit (Illumina, Cat# FC-121-1012)
MiSeq Nano Reagent Kit (Illumina, Cat# MS-102-2001)

Genetics of Race Module: Cost Considerations

In our implementation, we made use of our educational sequencing center staffed by a technician. The sequencing center needs to have standard equipment on hand, including thermocyclers, centrifuges, micropipettes, and pipette tips; we considered this equipment to be part of the lab setup and did not account for their cost in the discussion below.

Labor. In our implementation, students carry out the first steps of the procedure involving sample collection and cell lysis to extract DNA. Subsequent steps are performed in the sequencing center by a technician. Processing samplesconsists of the following steps: setting up the PCR reaction for targeted amplification of the regions of interest and running a second PCR reaction to attach adapters and index sequences, confirming DNA size, quality, and concentration, and running the sequencer. The technician typically dedicates approximately 40 hours to prepare 50samples, corresponding to $800. This time commitment increases incrementallywith the number of samples: an additional set of 50 samples would require an additional 4 hours, because many of the steps are performed in parallel using multiwell plates and multichannel micropipettes. This time commitment could be reduced by having students become more involved in the sample processing steps. Students could potentially complete all steps up to the final quality control and sequencing steps.This approach would reduce the technician time to approximately 16 hours but would require greater time commitment in the classroom.

Consumables and reagents. The largest expense in this category is the sequencing kit. We use the MiSeq reagent nano kit v2, 500 cycles ($340). This kit is one of the smallest and least expensive kits currently available, yet it provides much greater sequencing capacity than needed for this experiment. The need to confirm samples for quality control is also expensive: we use the Bioanalyzer system from Agilent ($820 for a kit capable of analyzing 300 samples). Additional costs include sample collection swabs ($70 for 100 swabs), DNA extraction and PCR reagents ($330, sold in kits for 100 samples), DNA quantification reagents ($85 for 100 samples), and post-PCR clean-up reagents ($320for 100 samples). Combining these costs results in a total of approximately $2,000 for a single classroom run of 50 samples (note many of these consumables could be used for multiple classroom runs). Bulk purchase or educational discounts could substantially reduce this estimate.

Materials	**Cost**
MiSeq reagent nano kit v2, 500 cycles	$340
Bioanalyzer system from Agilent (300 samples)	$820
Swabs (100)	$70
DNA extraction and PCR reagents (100 samples)	$330
DNA quantification reagents (100 samples)	$85
Post-PCR clean-up reagents (100 samples)	$320
TOTAL	$1,965

Table 1: Summary of materials and cost for NGS calculated for 100 samples, not including $800 for technician labor cost.

Cost considerations for Sanger sequencing. The cost of Sanger sequencing is fixed at approximately $4 per sample. Labor and consumable expenses for sample collection, DNA extraction, targeted PCR amplification, and post-PCR cleanup would remain the same as for NGS preparation. The main changesinclude eliminating the Bioanalyzer expense (agarose gel electrophoresis is recommended for quality control prior to submitting samples for sequencing) and eliminating the sequencing kit expense. Due to the fewer processing steps, the time commitment required by a technician is approximately 20 hours, and the labor cost is therefore estimated at $400. Increased student involvement in sample preparation could reduce or eliminate the need for a technician. With access to appropriate equipment and support, students could potentially complete all steps of this process until samples are shipped out for sequencing.

Materials	**Cost**
Sanger sequencing (100 samples)	$400
Swabs (100)	$70
DNA extraction and PCR reagents (100 samples)	$330
DNA quantification reagents (100 samples)	$85
Post-PCR clean-up reagents (100 samples)	$320
TOTAL	$1,200

Table 2: Summary of materials and cost for Sanger sequencing calculated for 100 samples, not including $400 for technician labor cost

Adapting the module to use Sanger sequencing

Our module is built upon previous educational programs that used Sanger sequencing for mitochondrial DNA sequencing and exploring race; we refer readers to these papers and websites for additional details and perspectives(Public Broadcasting Service 2003; Kosinski, Weinbrenner and Cross 2008). We also direct readers to the DNA Learning Center website, which hosts several relevant resources (Cold Spring Harbor Laboratory 2016).

Adapting the classroom teaching materials. The current classroom materials are designed to introduce students to next-generation sequencing. In particular, Lesson 3: Sequencing by Synthesis gives an overview and basic technical background on the most prevalent NGS technology. This material can be modified to focus more broadly on sequencing in general instead of NGS. In this lesson, the “modeling sequencing by synthesis” activity must be replaced with a different “modeling Sanger sequencing” activity.

Adapting the laboratory procedure. The sample collection procedure performed in the classroom (included below in the Supplement) is the same for NGS or Sanger sequencing.The subsequent steps involving PCR amplificationfollowed by post-PCR cleanup also remain the same.

After this point, the NGS and Sanger sequencing preparation procedures diverge, as indicated in the more detailed steps below.For Sanger sequencing, after the post-PCR cleanup, DNA quantification and agarose gel electrophoresis should be performed to confirm size and quantity of the resulting cleaned-up PCR product. If the quality of the DNA samples is satisfactory, they can be shipped to a Sanger sequencing facility or service.

These steps could be performed by a technician, or could be carried out by students. If students perform these steps, they will likely require at least three 50-minute classroom periods: one to set up PCR, one to perform PCR cleanup, and one for gel electrophoresis.

References.

Cold Spring Harbor Laboratory. DNA Learning Center https://dnalc.org/. 2016.

Kosinski RJ, Weinbrenner DR, Cross MG. Extraction, sequencing, and analysis of mitochondrial DNA. *Assoc Biol Lab Educ* 2008;**29**:137–66.

Public Broadcasting Service. Race - The Power of an Illusion http://www.pbs.org/race. 2003.

Genetics of Race: Sample Collection Procedure

In this procedure, students collect cheek cells from buccal swabs using the Extract-N-Amp Tissue Kit from Sigma-Aldrich. These cells will then be lysed in order to release the DNA for use in subsequent steps.

Put on your safety glasses and gloves prior to beginning the protocol.

1. Carefully open the sterile swab.
2. Collect a sample of cells from the inside of your cheek by twirling and scraping the inside of your cheek with the foam tip of the swab for 10 seconds. *It is not enough to just wet the swab with saliva; it's also important to scrape cells off of your cheek.*
3. Let the swab air-dry for 10 minutes. Make sure nothing touches it to avoid contamination.
4. Pipette 225 µL of the Extraction Solution into your microcentrifuge tube.
5. Submerge your swab into the Extraction Solution and leave there for one minute.
6. Twirl the swab in solution 10 times. Then squeeze out extra solution from the swab by pressing it against the side of the tube. Vortex for 3-4 seconds to mix it.
7. Discard the swab in the biohazard waste.
8. Close your sample collection tube and leave the sample at room temperature for 10 minutes.
9. Put your sample into a floating tube holder and place into the in the 95°C hot water bath for three minutes. The high temperature lyses the cells and exposes the DNA.
10. Add 200 µL of Neutralizing Solution and vortex the tube briefly. This makes the sample more suitable for PCR amplification.
11. Bring your sample to the collection tray. Your sample should now contain exposed DNA that is ready for PCR amplification and sequencing.

Genetics of Race: Details of Sequencing Process

Polymerase chain reaction (PCR)for targeted amplification

The following process requires extracted DNA prepared as abovefromthe student samples (following the “Genetics of Race: Sample Collection Procedure”).PCR is performed to amplify specific regions of the mitochondrial DNA (mtDNA) as desired for sequencing. Two regions are targeted for each student’s sample: hypervariable region I (HVRI) and hypervariable region II (HVRII). These regions are non-coding regions that are often used by scientists to track ancestry, migration, and relatedness.

These primers contain: targeting regions specific to the mitochondrial DNA, a variable number of “N” nucleotides that serve to increase diversity during sequencing, and additional regions to serve as linkers for subsequent steps that enable multiple samples to be sequenced at the same time. Because there are a variable number of N nucleotides (0 to 3), we used an equal mixture of four DNA sequences that correspond to each primer.

HVRI primer set (amplifies mtDNA from position 15,971 - 16,391, product length 440 bp)

HVRI forward primer

5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTAACTCCACCATTAGCACC 3'
5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNTTAACTCCACCATTAGCACC 3'
5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNTTAACTCCACCATTAGCACC 3'
5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNTTAACTCCACCATTAGCACC 3'

HVRI reverse primer

5′ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGAGGATGGTGGTCAAGGGAC 3'
5′ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNGAGGATGGTGGTCAAGGGAC 3'
5′ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNGAGGATGGTGGTCAAGGGAC 3'
5′ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNGAGGATGGTGGTCAAGGGAC 3'

HVRII primer set (amplifies mtDNA from position 8 - 430, product length 422 bp)

HVRII forward primer

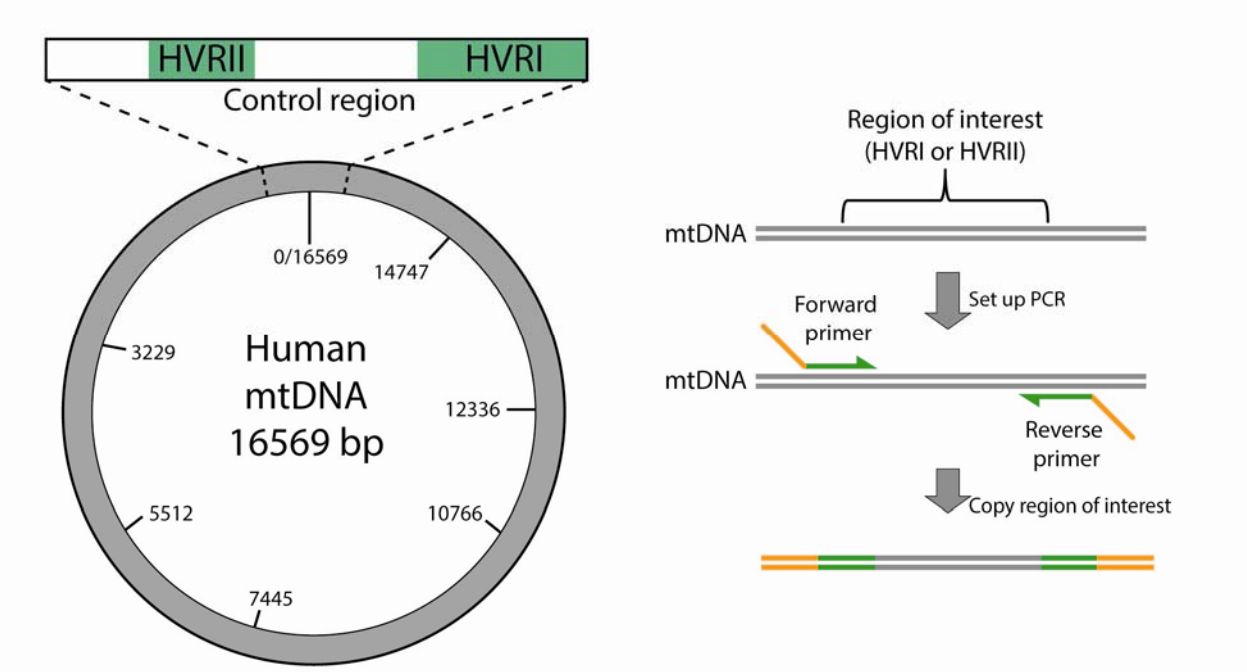
5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGTCTATCACCCTATTAACCAC 3'
5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNGGTCTATCACCCTATTAACCAC 3'
5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNGGTCTATCACCCTATTAACCAC 3'
5′ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNGGTCTATCACCCTATTAACCAC 3'

HVRII reverse primer

5′ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTGTTAAAAGTGCATACCGCCA 3'
5′ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNCTGTTAAAAGTGCATACCGCCA 3'
5′ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNCTGTTAAAAGTGCATACCGCCA 3'

5′ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNCTGTTAAAAGTGCATACCGCCA 3'

Primers for HVRI were originally designed by Kosinki et al., 2008 "Extraction, Sequencing, and Analysis of Mitochondrial DNA".Primers for HVRII were designed by Schmuczerova et al., 2010 "HVRII of mtDNA in Cord Blood Cells of Newborn Children and in Their Saliva 10 Years Later".



PCR conditions were as follows: initial denaturation 94°C for 3 minutes; 30 cycles of denaturation (94°C for 30 seconds), annealing (53°C for 30 seconds), extension (72°C for 1 minute); final extension 72°C for 10 minutes.

Preparing amplified DNA for sequencing

After amplification, the PCR product is cleaned up using magnetic beads(Agencourt AMPure XP) to isolate the amplified DNA and remove unwanted components such as residual primers and salts. Then, index regions designed for Illumina sequencing are added via a second round of PCR, and the product is again cleaned up with magnetic beads. Finally, each sample is checked for quantity, size, and quality using capillary electrophoresis (Agilent Bioanalyzer), and samples are pooled. The MiSeq sequencer is loaded according to Illumina protocols.

For Sanger sequencing, the protocol diverges immediately after PCR clean-up using the Agencourt AMPure XP beads. Instead of a second PCR amplification to attach index regions, each sample should be checked for sufficient quantity and quality, using a DNA quantification method and agarose gel electrophoresis, respectively.

Analysis to compare student sequences

The MiSeq onboard analysis uses a mitochondrial DNA reference file to call variants. These variants are used to build FASTA files corresponding to each sample. These FASTA files

undergo multiple sequence alignments(we use Clustal Omega: https://www.ebi.ac.uk/Tools/msa/clustalo/) to determine percent genetic identity between each student in the class.

BioSeq Knowledge Test

Sequencing and Ethics

1. Compared to the Sanger method of DNA sequencing, next-generation sequencing technologies ___.
    a. have made sequencing faster and cheaper.
    b. allow many DNA fragments to be sequenced simultaneously.
    c. do not require any special sample preparation.
    d. a and b
    e. b and c

2. What is meant by an "index" sequence?
    a. The PCR primer sequence
    b. One of the adapter sequences that will allow the DNA to be captured on the surface of the flow cell
    c. The unique sequence that allow us to figure out which clusters came from which person's sample
    d. none of the above

3. Prior to sequencing, a sample of DNA is quantified. What is the significance of this step?
    a. Too much DNA may reduce the quality of a read.
    b. Too little DNA will result in no readable signal.
    c. The length of the fragments of DNA that will be loaded into the sequencer must be within an appropriate range.
    d. All of the above.

4. The genome of an organism is sequenced and found to be different from the reference genome for the species. Based on this information, which statement below will also be true?
    a. We cannot say exactly what this observation means without further investigation.
    b. The organism represents a new, previously undiscovered species.
    c. The organism is likely a diseased specimen.
    d. The organism will have very different traits from other members of its species.

5. A sample of cells collected from a student's skin contains DNA sequences that match the reference genome for *Staphylococcus aureus*, the species responsible for most *Staph* infections. What does this mean?
    a. The student has an active *Staph* infection and should see a doctor immediately.
    b. The student is about to develop an active *Staph* infection.
    c. There were *Staphylococcus aureus* cells present on the student's skin.
    d. The student needs to use better hygiene practices.

Bioinformatics

1. What is one of the uses of BLAST?
    a. Find DNA sequences similar to a collected sample's sequence.
    b. Determine which molecules are poisonous to humans.
    c. Advise which chemical mixtures are explosive.
    d. Simulate protein folding.

2. Looking at the difference between the reference sequence and the sample, what is the mutation that has occurred below?

```
REFERENCE: GCTTAGGTCAGTTAGCC
READ:      GCTTAGGT--GTTAGCC
```

    a. Substitution
    b. Deletion in read
    c. Insertion in read
    d. No mutation is present.

3. What does sequence coverage mean?
    a. Coverage refers to the length of the sequencing read.
    b. Coverage is the average number of sequence reads per chromosome.
    c. Coverage is the number of times each nucleotide in the original sample is seen by the sequencer.
    d. Coverage refers to how much of an organism's genome has ever been sequenced.

4. What is the difference between *de novo* assembly and genomic resequencing?
    a. For *de novo* assembly, you have a reference sequence to which you can align sequencing reads, while for resequencing you do not.
    b. In *de novo* assembly, you do not know the underlying sequence of the sample, while in resequencing, you do have a reference sequence for the sample.
    c. *De novo* assembly refers to deep sequencing of a small sample, while resequencing refers to broad sequencing of a big sample.
    d. *De novo* assembly refers to the sequencing of a new sample, while resequencing refers to the sequencing of sample you have already sequenced before.

5. What is paired-end sequencing?
    a. The DNA and RNA from a sample are sequenced.
    b. The hydrophobic and hydrophilic parts of the DNA are sequenced.
    c. Nuclear and mitochondrial DNA are sequenced.
    d. Each end of each fragment of DNA is read, one in the forward and one in the reverse direction.

BioSeq Attitudinal Survey

Student Survey
This survey will ask you to rate yourself before participating in the Tufts BioSeq program and then again after having completed it. Please be as honest as possible. Your answers will be kept confidential and they will not affect your grades.

1. **What is your sample ID? Example: 2ASH077**

2. **Your school**

3. **Your teacher's name**

4. **How much of the BioSeq program did you attend?**Choose among the following options: I was not here for any of it; I missed one or 2 sessions; I was here for most of it; I was here for all of it.

5. **How well were you able to understand the experiment in the BioSeq program?** Choose among the following options: I did not understand it at all; I only understood it a little; I understood about half of it; I understood most of it; I understood all of it.

6. **What are the first words that come to your mind when you think about the BioSeq program?**

7. **How would you rate your knowledge of the following before participating in the BioSeq program and then again at the current time?**

   Choose among the following options: 1 – I didn't know anything about this; 2 – I had only vague ideas; 3 – I knew what this was; 4 – I knewsomething about this; 5 – I knew quite abit about this; 6 – I knew a great deal about this.

	BEFORE participating in BioSeq	AFTER participating in BioSeq
Next-generation sequencing		
Sample preparation for sequencing		
How the sequencer works		
How resequencing works		
Types of mutations		
Ethical issues associated with analyzing sequences		

8. **What for you is the most important thing you have learned about DNA sequencing?**

9. **Of all the things you have learned about DNA sequencing what has been your most valuable source of information? Please rank the following.**

NOTE: Each ranking can be used only ONCE.

	1 - The least important	2	3	4	5 - The most important
Your regular school classes	○	○	○	○	○
The BioSeq program	○	○	○	○	○
Your own independent study	○	○	○	○	○
Your friends	○	○	○	○	○
Your family	○	○	○	○	○

10. **An important feature of the BioSeq program is to help you develop a number of skills. How would you rate yourself before and after participating for the following?** Choose among the following options: 1 – I didn't know anything about this; 2 – I had only vague ideas; 3 – I knew what this was; 4 – I knew something about this; 5 – I knew quite a bit about this; 6 – I knew a great deal about this.

	BEFORE participating in BioSeq	AFTER participating in BioSeq
Determining how to build an answer to a research question		
Sample preparation for sequencing		
Understanding the process of sequence assembly		
Using BLAST to align reads		
Interpreting sequencing results		

11. **Imagine that you could take a swab from the sink in the kitchen in your home. What would a DNA analysis of the sample tell you about your family? Please provide as much detail as you can.**

12. **How would you rate yourself before and after participating in the BioSeq program for the following?** Choose among the following options: 1 – I didn't know anything about this; 2 – I knew about this but my skills were/are weak; 3 – My skills were/are just ok; 4 – My skills were/are pretty good; 5 – My skills were/are very good; 6 – My skills were/are excellent.

	BEFORE participating in BioSeq	AFTER participating in BioSeq
Your ability to explain research questions pertaining to your microbiome, genetics, or mutations to another student		
Your ability to collect DNA samples without contamination		
Your ability to explain how next-generation sequencing technologies work to another student		
Your ability to explain the bioinformatic analysis completed in this program to another student		

13. **This next set of questions will ask you about your attitudes toward research in genetics and interest in pursuing a career in a related field.**

    On a scale of 1 to 6, how would you rate the following? 1 – Not at all accurate; 6 – Extremely accurate.

	BEFORE participating in BioSeq	AFTER participating in BioSeq
I feel like I know how to develop a good research question		
I feel like I can do well in the bioinformatics field		
I feel like I can learn the concepts used in genetics		
I feel like I can be a valuable member of a research team		
I would like to pursue a career in the biological sciences		

14. **How likely are you to take a course in bioinformatics in college?**

    Not at all likely

    Maybe, but not likely

    It would depend

    I will probably take a bioinformatics course

    I will definitely take a bioinformatics course

15. **How often do you do the following?**Choose among the following options: Never; Maybe once a month; 2-3 times a month; Every week; Several times a week; Every day.

	Frequency
Talk with friends about what your experience in the BioSeq program?	
Talk with your family about your experience in the BioSeq program?	
Notice things online that are relevant to what you are learning in the BioSeq program?	

16. This next set of questions asks about how much you have told others who are NOT participating in the program about your experience in the BioSeq program.

	0	1-2	3-5	6-10	11+
How many people have you told about the BioSeq program?	○	○	○	○	○
How many other students have you encouraged to get involved with the BioSeq program?	○	○	○	○	○

17. How likely would you be to recommend the BioSeq program to another student?

Not at all likely

Maybe, but not likely

It would depend

I would probably recommend the program

I would definitely recommend the program

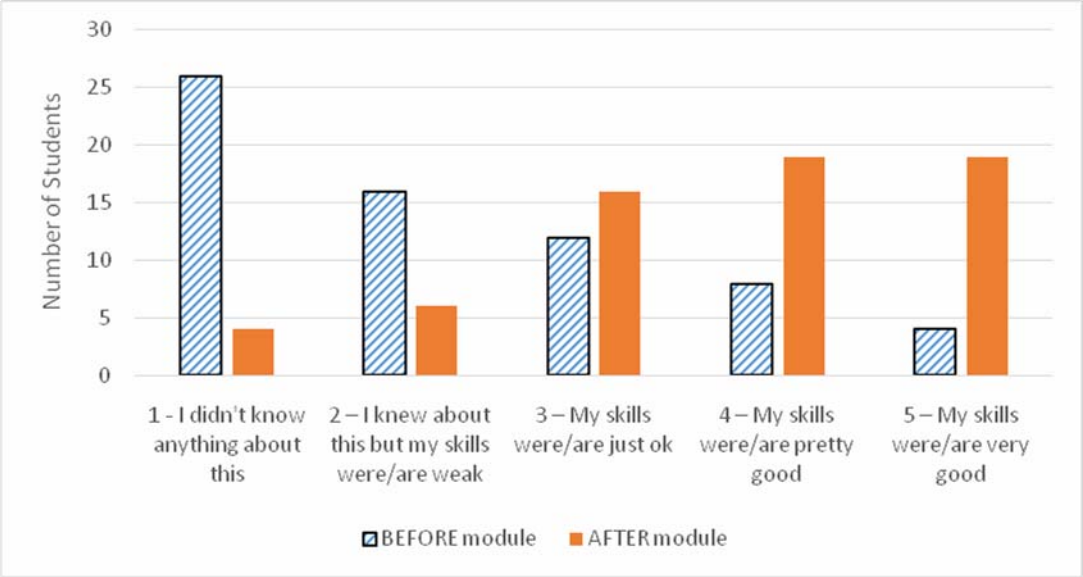
18. What would you say if another student asked you about participating in the BioSeq program?

Sampling of responses to the BioSeq Attitudinal Survey

Selected responses to Question 11,an open-ended question,from the BioSeq Attitudinal Survey: “What for you is the most important thing you have learned about DNA sequencing?” Responses are presented as given, including typos or grammatical errors.

- “Genes and Race have no exact correalation”
- …you may be genetically similar to someone who u would of never guessed.”
- “I learned that race has little to nothing to do with who you are most common to.”
- “How genetically similar we all are”
- “…I know know that even though we look different we are quite similiar in DNA.”
- “There are no scienfical evidence for the existence of human race”
- “…humans migrate all the time and might interbreed and still identify as a certain race, but that does not mean there isn't mixtures of different races in a particular person.”

Below, responses to the first part of Question 12 from the BioSeq Attitudinal Survey: “How would you rate yourself before and after participating in the BioSeq program for the following – your ability to explain research questions pertaining to your microbiome, genetics, or mutations to another student?”



Demographic information

High school students participating in this study included 68 high school students enrolled in four standard track biology classes at the same school.

Gender		
Male	**Female**	**Not Reported**
35	32	1

Ethnic categories		
Hispanic or Latino	**Not Hispanic or Latino**	**Unknown**
27	36	5

Racial categories						
American Indian/ Alaska Native	**Asian**	**Native Hawaiian/ Pacific Islander**	**African American**	**White**	**More than One Race**	**Unknown**
0	8	0	6	25	14	15

Age				
19 yrs old	**18 yrs old**	**17 yrs old**	**16 yrs old**	**15 yrs old or younger**
1	23	32	10	2