INTRODUCTION

The scale of genomic sequencing data and the complexity of bioinformatic algorithms make it difficult for students to develop a concrete understanding of assembling complete genomes from millions of short DNA sequences. The majority of genome sequencing is performed using Illumina’s sequencing-by-synthesis (SBS) technology (1). Instead of reading the genome as a single sequence of nucleotides, SBS instruments produce millions of small, 50- to 250-base pair (bp), DNA fragments called “short reads.” The short reads must be subsequently assembled into a complete genome. A single 3.2 × 10⁹ bp human genome, at the standard 30-fold coverage, requires over 190 million short reads. When learning genome sequencing, students can struggle to imagine the challenges created by so many reads. Students might believe that the challenges stem from simply the number of reads. In practice it is often the complexity, rather than the size, of genomic datasets that complicates genome assembly. Pressing bioinformatics challenges like identifying sequence variants and repetitive regions cannot be solved with only faster computers. Instead, solutions will require novel insights from current and future bioinformaticists.

This manuscript presents a problem-based lesson for discovering the challenges in genome assembly. The lesson uses physical paper reads and assumes no background in computer science or mathematics, only an introductory understanding of DNA and genomes. Topics highlighted during the lesson include overlap identification, reference sequences, and the challenges arising from sequencing errors, low-frequency mutations, and repetitive regions. Sample materials provide reads and solutions for assembling clinically relevant regions of the *S. gordonii* penicillin binding protein and the human *HTT* gene. An online tool allows instructors to generate custom read sets from other DNA sequences. We believe these minimal prerequisites will reduce barriers to engaging with computational biology.

PROCEDURE

To make the concepts of genome assembly more concrete, teams of students were given paper copies of short reads generated from segments of microbial and mammalian genomes and asked to assemble them into complete sequences (Fig. 1). The reads used in the following activities (and the original sequences) are included in Appendix 1, along with detailed instructions for recreating the activities. To aid instructors who wish to generate reads from their own sequences, we created an online interface to our read-generating software (available at http://jensenlab.net/tools). Instructors can enter custom sequences and modify the read length, coverage, and error rate. The online tool generates and randomizes reads and produces a reference sequence. The reads can be copied into a word processor and printed for lessons. Lower-case, fixed-width serif fonts with different colors for each base offered the best contrast and easiest assemblies in our experience.
Activity 1: De novo assembly from error-free short reads

Students were given 24 short reads covering a 60 bp DNA sequence, clear tape, and scissors. Students were instructed that all of the reads assemble into a single contiguous piece of DNA (a contig), and the reads are free from errors. Each group of three students received an identical set of reads. All groups completed their assembly within twelve minutes, and six of the eight groups formed correct assemblies. The other two groups both used small overlaps (as small as a single base) in parts of their assemblies, which led to incorrect sequences. The students were given a “key” —a long strip of paper containing the correct sequence to compare with their assemblies.

Discussion points included: How much overlap is required before you can be confident in your assembly? Can we quantify the probability that each overlap is correct? Answers may include \((1/4)^n\), where \(n\) is the number of overlapping bases.

Activity 2: Scaffolded assembly against a reference sequence with errors and mutations

The group discussed how errors in the sequences would confound their assembly efforts. The students repeated the exercise using reads from the same sequence but containing errors. Students were instructed that the errors were “rare,” and that no read contained multiple errors. After five minutes of assembly, groups were allowed to use the key from the previous exercise as a reference sequence, and all groups completed the assembly within five additional minutes.

Two topics were discussed. First, the students reflected on how having a reference sequence simplifies the assembly process, even in the presence of errors. The instructor commented on how high-quality reference genomes, such as the result of the human genome project, are still valuable despite being produced by long-outdated instruments. Second, students were asked how they can distinguish between rare mutations and sequencing errors. The read set contained two single nucleotide differences from the reference sequence. One was a sequencing “error,” but the second was a naturally occurring, low-level variant. The instructor revealed that reads for activities 1 and 2 were generated from a segment of the penicillin binding protein gene in Streptococcus gordonii. The polymorphism at position 45 confers 250-fold resistance to penicillin. Tracking abundance of this mutation is important for surveilling antimicrobial resistance (2).

Activity 3: De novo assembly of a trinucleotide repeat region

One drawback of using millions of short reads is difficulty assembling regions of genomes with repetitive structures. If a repeated region is longer than the read length, it is not possible to unambiguously assemble the region. To demonstrate the difficulties posed by repeats, the students were asked to assemble a third set of reads generated from the human Huntingtin (HTT) gene. The HTT gene contains a region of repeated CAGs, encoding the amino acid glutamine. The exact number of CAGs varies by individual. Too many CAGs cause a glutamine “knot,” making the protein insoluble. Because HTT is expressed in neurons, these protein precipitates cause the neurodegenerative disease known as Huntington’s (3). The short reads for this activity were generated from an HTT sequence with seven CAGs, but the reads can be assembled into contigs with as few as four and as many as 23 repeats.

After assembly, students were asked how they could diagnose genetic conditions like Huntington’s that contain repetitive sequences. (Possible answer: Many genetic conditions are diagnosed by targeted Sanger sequencing of only the region of interest. These assays are lower throughput but produce longer, more accurate reads.) Also, it was noted that a reference sequence is of no help for this problem. Assembling against a reference will result in the same number of repeats as the reference, not the number of repeats in the patient’s genome.

CONCLUSION

As the cost of DNA sequencing decreases, genomics becomes increasingly ingrained in biology. Bioinformatics is rapidly transforming from a distinct sub-discipline into a necessary part of a biologist’s toolkit (4). Like other quantitative disciplines, bioinformatics and computational biology build on the fundamentals of calculus, statistics, computer science, and physics. Viewing math and science courses as the gateway to computational biology places significant barriers to the field for students lacking confidence in these areas, especially women and underrepresented minorities (5). We hope that simple activities like the one presented here will allow more students to become interested in bioinformatics and develop the skills needed for the next generation of biological research.

SUPPLEMENTAL MATERIALS

Appendix I: Materials and instructions

ACKNOWLEDGMENTS

We thank Megan Griebel and Susan Flannegan for preparing materials and Brian San Francisco for testing early versions of the activity. Financial support for this work was provided by the WYSE Summer Camp and the University of Illinois at Urbana-Champaign College of Engineering. The author has no conflicts of interest to declare.
REFERENCES

1. Illumina, Inc. 2016. An introduction to next-generation sequencing technology. San Diego, CA. Available at: http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf. Retrieved January 1, 2017.

2. Haenni M, Moreillon P. 2006. Mutations in penicillin-binding protein (PBP) genes and in non-PBP genes during selection of penicillin-resistant Streptococcus gordonii. Antimicrob Agents Chemother 50(12):4053–4061.

3. Walker FO. 2007. Huntington’s disease. Lancet 369(9557):218–228.

4. Markowetz F. 2017. All biology is computational biology. PLOS Biol 15(3): e2002050.

5. Ellis J, Fosdick BK, Rasmussen C. 2016. Women 1.5 times more likely to leave STEM pipeline after calculus compared to men: lack of mathematical confidence a potential culprit. PLOS One 11(7): e0157447.