HOW TO DO IT

Illuminating the Black Box of Genome Sequence Assembly: A Free Online Tool to Introduce Students to Bioinformatics

RECOMMENDED FOR *AP Biology*

D. LELAND TAYLOR,
A. MALCOLM CAMPBELL,
LAURIE J. HEYER

Abstract

Next-generation sequencing technologies have greatly reduced the cost of sequencing genomes. With the current sequencing technology, a genome is broken into fragments and sequenced, producing millions of "reads." A computer algorithm pieces these reads together in the genome assembly process. PHAST is a set of online modules (http://gcat.davidson.edu/phast) designed to teach advanced high school and college students the genome assembly process. PHAST allows users to assemble phage genomes in real time and includes tutorials detailing the complexities of genome assembly. With PHAST, students learn concepts behind genome assembly and understand how mathematics solves biological problems such as genome assembly.

Key Words: *Genome assembly; bioinformatics; computational biology; teaching tool.*

Genome-scale DNA sequencing has transformed biological research. Scientists can sequence whole genomes of microbes, plants, animals, and humans (Fleischmann et al., 1995; Blattner et al., 1997; *C. elegans* Sequencing Consortium, 1998; *Arabidopsis* Genome Initiative, 2000; International Human Genome Sequencing Consortium, 2001; Mouse Genome Sequencing Consortium, 2002). Investigators are sequencing the genomes of individual cells within tumors to determine cellular DNA variance (Navin et al., 2011). Metagenomic data sets are composed of DNA from mixed populations of microbial species, and the composition of species is deduced from assembled genomes (Tyson et al., 2004; Venter et al., 2004; Tringe et al., 2005). All of these research advances rely on genome assembly algorithms. As sequencing costs continue to plummet, the sequencing of whole genomes will become progressively commonplace (Wetterstrand, 2012). With large data sets, the field of biology has become increasingly quantitative and interdisciplinary, drawing from the fields of mathematics and computer science.

Genome-scale DNA sequencing has transformed biological research.

The rapid transformation in biological research has made it difficult for biology curricula to keep pace with the demands of high-throughput biology. In 2003, the National Research Council's (NRC) *Bio2010* report emphasized the need to update traditional biology curricula to reflect the interdisciplinary nature of modern biological research (NRC, 2003). Similarly, in 2009, the Association of American Medical Colleges (AAMC) and Howard Hughes Medical Institute (HHMI) released a report stressing the importance of mathematics in undergraduate biology education for future physicians (AAMC & HHMI, 2009). A recent report, Vision and Change (AAAS, 2011), reiterates the need for biology students to learn the most recent interdisciplinary research methods. Finally, the redesigned AP Biology curriculum requires high school students to learn to use computational tools (College Board, 2012). To better equip students for the requirements of modern science, our students need to make connections among biology, mathematics, and computer science.

This article describes a free Web resource that allows students to explore genome assembly through hands-on experimentation. The website is tailored for students participating in the HHMI-sponsored Phage Hunters program (Hatfull et al., 2006; Pope et al., 2011) but can be used by any college or advanced high school student to understand how genomes are assembled. Typically, the genome sequencing and assembly process is a "black box" for most students who annotate and analyze genomes. For example, Phage Hunter students isolate mycobacteriophages (bacterial viruses), send them to be sequenced, and annotate the fully assembled genome sequence. PHAST helps students understand genome sequencing and assembly.

○ Objectives

PHAST (Phage Assembly Suite and Tutorial; http://gcat.davidson.edu/phast) is appropriate for students who understand DNA structure and how DNA constitutes an organism's genome. Through PHAST's tutorials, students will learn (1) the tradeoffs in sequencing technologies

The American Biology Teacher, Vol. 75, No. 8, pages 572–577. ISSN 0002-7685, electronic ISSN 1938-4211. ©2013 by National Association of Biology Teachers. All rights reserved. Request permission to photocopy or reproduce article content at the University of California Press's Rights and Permissions Web site at *www.ucpressjournals.com/reprintinfo.asp.*
DOI: 10.1525/abt.2013.75.8.9

and how reads are stitched together into larger sequences called "contigs" (contiguous sequences of DNA) or "scaffolds" (ordered contigs); (2) the application of basic graph theory in popular genome assembly methods; and (3) the computational complexities of the genome assembly problem, such as the difficulty in identifying a correct assembly. In addition, PHAST's real-time assembly feature enables students to generate their own assembled genomes. Combined, these features provide students with the tools necessary to discover the utility of mathematics and computer science to enrich their understanding of biology.

○ Materials

This exercise requires a computer connected to the Internet with at least Safari 5, Chrome 16, or Firefox 9 installed. Internet Explorer does not work well with PHAST, so users are encouraged to use one of the free browsers listed above. Instructors do not need to install any specialized software, because all of the computation is performed on the PHAST server.

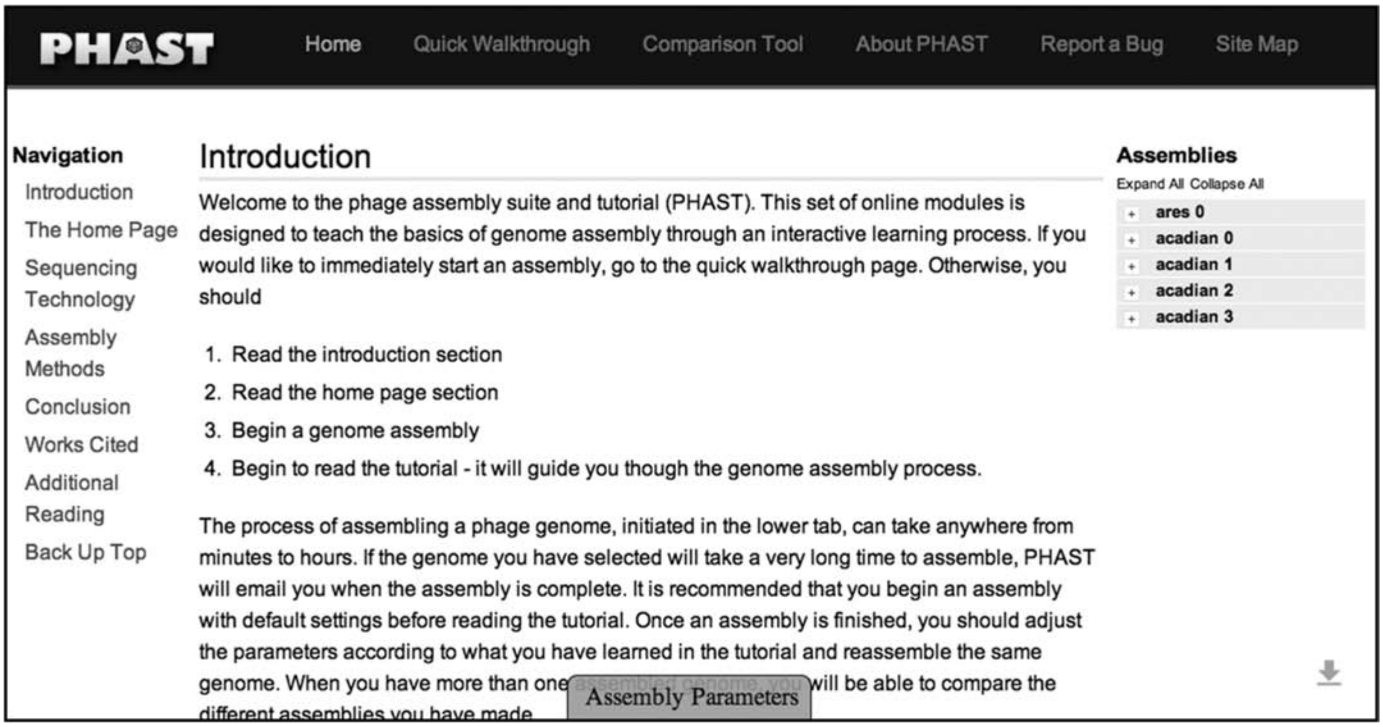
○ Engagement Strategies

PHAST engages students through an interactive discovery of how genomes are assembled from large numbers of raw reads. Learning activities take place entirely online through the website. PHAST is a "one-stop" genome-assembly teaching tool, consisting of several modules. As students work through the modules, they are encouraged to generate their own, real-time assemblies from the supplied phage reads.

PHAST contains a prepopulated database of raw reads from several phages sequenced through the HHMI PHAGES (Phage Hunters Advancing Genomics and Evolutionary Science; http://phagesdb.org/) program. Embedded within PHAST is a genome assembly program used by bioinformaticians at sequencing centers, called MIRA (Mimicking Intelligent Read Assembly; Chevreux et al., 1999). PHAST simplifies the assembly process to three adjustable parameters that highlight key genome assembly concepts described in the modules (see "Student Procedures"). The interactive format of PHAST encourages students to develop and test their own hypotheses as they generate novel phage genome assemblies using real data. After generating several assemblies, students can compare the structure of their assembled genomes through an integrated dotplot tool called "gepard" (GEnome PAir – Rapid Dotter; Krumsiek et al., 2007). In this way, students discover the consequence of assembly parameter choices on the structure of the assembled genome.

○ Teacher Procedures

To familiarize yourself with the PHAST website, first read the overview provided by the "Quick Walkthrough" (http://gcat.davidson.edu/phast/walk.html). The walkthrough guides you through a sample assembly, which illustrates the different features of PHAST and the assembly workflow. Next, we recommend that you read the main tutorial, located on the "Home" page (Figure 1). PHAST assumes your students have a basic understanding of the nucleotide structure of DNA, with minimal knowledge of DNA sequencing. The main tutorial introduces Sanger sequencing, followed by next-generation sequencing technologies. PHAST explains how next-generation



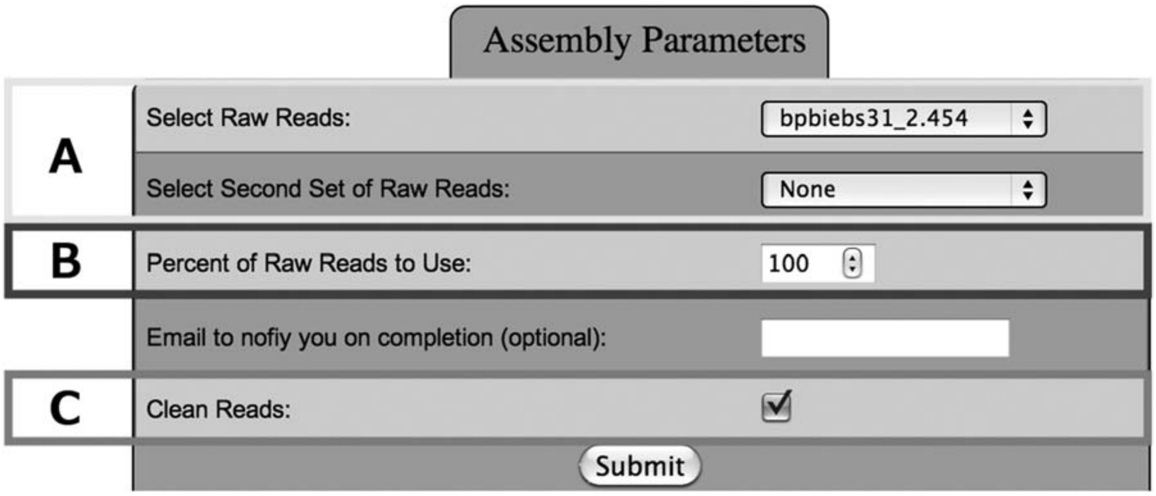
Introduction

Welcome to the phage assembly suite and tutorial (PHAST). This set of online modules is designed to teach the basics of genome assembly through an interactive learning process. If you would like to immediately start an assembly, go to the quick walkthrough page. Otherwise, you should

1. Read the introduction section
2. Read the home page section
3. Begin a genome assembly
4. Begin to read the tutorial - it will guide you though the genome assembly process.

The process of assembling a phage genome, initiated in the lower tab, can take anywhere from minutes to hours. If the genome you have selected will take a very long time to assemble, PHAST will email you when the assembly is complete. It is recommended that you begin an assembly with default settings before reading the tutorial. Once an assembly is finished, you should adjust the parameters according to what you have learned in the tutorial and reassemble the same genome. When you have more than one assembled genome, you will be able to compare the different assemblies you have made

Figure 1. The PHAST homepage. Assemblies are generated using options within the bottom tab labeled "Assembly Parameters." Completed assemblies appear in the right-hand column entitled "Assemblies." Clicking on an assembly will display the parameters used to generate the assembly along with statistics such as the average read coverage, the number of contigs, and the assembled genome size. The "Assembly Parameters" tab and the "Assemblies" column appear on every page. The left-hand "Navigation" column enables skipping through the sections of each page.



Assembly Parameters

Select Raw Reads: bpbiebs31_2.454

Select Second Set of Raw Reads: None

Percent of Raw Reads to Use: 100

Email to nofiy you on completion (optional):

Clean Reads: ☑

Figure 2. Required Assembly Parameters. (A) The raw phage reads used to generate the assembly. Users can choose to generate an assembly using one or two sets of raw reads. *bpbeibs31* is the name of a phage isolated through the phage hunters program. *bpbeibs31_2* is the second set of raw reads, where *bpbeibs31_1* is the first set. The number *454* indicates that the 454 sequencing platform was used to generate the reads. (B) The percentage of raw reads the user has chosen for assembly. (C) The option to turn on or off the removal DNA adaptor sequences from each read.

technologies suit whole genome sequencing, which leads to the genome assembly problem. The "Assembly Methods" section of the main tutorial provides an overview of genome assembly techniques and contains links to other modules that discuss specific assembly techniques in detail. The "Additional Reading" section located at the bottom of each module contains useful resources related to the topic on each page.

Student Procedures

First, familiarize yourself with the PHAST workflow by generating an assembly as directed by the "Quick Walkthrough." After initiating an assembly, proceed to the "Home" page and read the tutorials while your genome is being assembled. PHAST reduces the assembly process to three adjustable parameters (Figure 2) that will enable you to explore genome assembly concepts described in the tutorials. For example, you can decrease genome coverage (the average number of times any base was sequenced) by reducing the percentage of raw reads (Figure 2B). As coverage decreases, the assembler's ability to accurately identify sequencing errors and predict the correct genome arrangement also decreases. You can choose not to "clean" 454 reads of adaptor sequences that are ligated to genomic DNA during DNA preparation of 454 sequencing (Figure 2C). Not cleaning 454 reads complicates the assembly by generating false nucleotide overlaps between reads.

For a more advanced exploration, try to assemble two sets of reads, either from the same or a different phage (Figure 2A). Sequencing facilities may sequence the genome of an organism more than once in order to fill gaps or resolve areas of low read coverage. Assembling two sets of reads for a single genome will allow you to assemble a phage genome using all available reads from two distinct sequencing runs. You will soon discover that having more reads (up to a point) increases the ability of assemblers to stitch together a complete genome and to detect sequencing errors. As an advanced analysis, you could select reads from two separate genomes to see whether the program can simultaneously assemble each genome separately. If the sequence structures of two genomes share few similarities (e.g., acadian and timshel), a two-genome assembly will produce a few, large contigs, roughly corresponding to the two different genomes. However, assembling two similar genomes (e.g., bpbiebs31 and euphoria) will produce an extremely fragmented assembly. All four of these phages can be used separately to generate assemblies on PHAST.

After performing two or more assemblies, you can compare your results using gepard, a dotplot tool integrated into PHAST (accessible on "Comparison Tool" in the top menu bar). With this tool, you can select two of your own assemblies from a dropdown menu to generate a dotplot comparing the two sequences. Dotplots are scatter plots that visualize the similarities shared between two sequences on the x-axis and y-axis. In this graph, a "dot" is added at every nucleotide position where the two sequences match. Using this method, two identical sequences are displayed as a diagonal line whereas insertions or deletions appear as lines slightly shifted off the main diagonal. Through dotplots, you can easily visualize the relationship of one assembly to another and assess the effect of the different parameters used for each assembly. For example, you could compare two assemblies of the acadian phage, one using 100% of the reads and one using 10% of the reads. In the resulting dotplot, you can quickly identify rearranged, fragmented areas of the 10% assembly, caused by a decrease in genome coverage. A series of assemblies using different percentages of the available reads will allow you to determine a minimum threshold necessary for complete phage assembly. Finally, you could use gepard to identify similar genomic regions between two different phages, just as PHAGES researchers do to classify unknown phage genomes (Pope et al., 2011).

Because assemblies are generated in real time, you should submit only one assembly job at a time, so as not to overconsume server resources. Assemblies with large amounts of data may take a long time. PHAST will automatically prompt you to enter an e-mail address for jobs that may take longer than a few minutes. After the job is complete, PHAST will e-mail you with instructions for accessing the results.

Guided Discovery Questions

Questions	Instructions
What is the minimum read coverage necessary to assemble a genome?	The minimum coverage necessary to assemble a genome will vary for each genome. To find the minimal coverage necessary to assemble acadian: 1. Generate three assemblies with the raw reads labeled **acadian.454** using 100%, 30%, and 20% of the reads. 2. Under the "Assemblies" column, click *expand all.* Locate the assemblies you generated by using the *parameters* section. In the *statistics* section, you will find the number of contigs and average read coverage for each assembly. 3. Using 100% and 30% of the reads, you generate a single contig, representing a complete genome. However, when you use 20% of the reads, the assembler is not able to stitch the reads together into a single contig. 4. Additionally, you see that the average coverage decreases as you use fewer reads. By comparing the average coverage, you can determine the minimum read coverage necessary to assemble acadian. 5. You can view the assembled raw reads by clicking *Mira Consensus*, but ignore the "Tag legend" and "Statistics" sections. Mira consensus is generated by the assembler, MIRA, and is not controlled by PHAST.
What happens if you forget to clean adaptor sequences from raw reads?	By not cleaning adaptor sequences, you introduce many false overlaps in the assembly graph. To illustrate the effect of false overlaps, follow this example using acadian. 1. Generate two assemblies with the raw reads labeled **acadian.454** using 100% of the reads. For one assembly check the box labeled *Clean Reads*, and for the other assembly uncheck the same box. 2. Under the "Assemblies" column, click *expand all.* Locate the assemblies you generated by using the *parameters* section. Look at the number of contigs under the *statistics* section. 3. You can visualize the arrangement of the two assemblies. Click on *Comparison Tool*. Select your two assemblies from the dropdown menu and click *Submit*. The resulting graph shows there are many genomic rearrangements and fragments in the uncleaned assembly.
Is it possible to assemble two separate genomes simultaneously?	Highly similar genomes will share many nodes and edges within the assembly graph. These similarities make it difficult for the assembler to tease apart two separate genomes. Let's look at an example. 1. Perform a two-genome assembly by selecting **bpbiebs31_1.454** for your first set of reads and **euphoria_1.454** for your second set of reads. 2. While this assembly is running, visit http://phagesdb.org/ and search for bpbiebs31 (http://phagesdb.org/phages/BPBiebs31/) and euphoria (http://phagesdb.org/phages/Euphoria/). 3. In the "Characterization" section of PhagesDB, you will find the cluster and subcluster of these phages. 4. Generate a second assembly using the **acadian.454** and **timshel.454** reads. 5. Return to PhagesDB and determine the cluster and subcluster for acadian (http://phagesdb.org/phages/Acadian/) and timshel (http://phagesdb.org/phages/Timshel/). 6. Armed with cluster information, predict which two-genome assembly will produce the two least fragmented genomes. Test your hypothesis by determining the number of contigs in each assembly. 7. Use the *Comparison Tool* to visualize and further evaluate the quality of the acadian and timshel assembly. Generate two separate assemblies for both acadian and timshel using the default parameters. Compare the acadian assembly to the acadian+timshel assembly. In the resulting plot, you can see that the acadian genome is rearranged but segregated from the timshel assembly in the acadian+timshel assembly. If you compare the timshel assembly to the acadian+timshel assembly, you will see a similar trend, but with more rearrangements and fragmentation. On the basis of your contig investigation, predict the results of a similar analysis for bpbiebs31+euphoria.

Teaching with PHAST

You could require students to read the PHAST tutorials outside of class and then use PHAST during class or lab to reinforce what they read. Alternatively, students could explore genome assembly with minimal assistance, as a dry lab exercise or as an exploratory question on an exam. The estimated time for students working completely on their own to read the main tutorial and explore several assemblies is roughly an hour. With directions from an instructor, students can generate assembled genome sequences in less than 5 minutes.

Assemblies generated are unique to each user and cannot be accessed by another user. Therefore, students must do their own work. Students could be asked to answer specific questions by generating assemblies and taking screen shots to support their conclusions. Below are several questions that students could answer with PHAST. For a detailed explanation of concepts, refer to the "Home" page.

Discuss the importance of genome coverage. Support your conclusions with an example.

"Genome coverage" refers to the ratio between the cumulative size, in nucleotides, of a set of reads and the size of the genome. Increased coverage gives assemblers more information to stitch together a complete, continuous genome. A student could generate two assemblies using 10% and 100% of the available reads. By clicking on the statistics of each assembly, a student could show that the average coverage increases with more reads and the number of contigs decreases, indicating a more complete, continuous genome assembly. These results could be further confirmed by generating a dotplot. An excellent example can quickly be generated using acadian.

How are read overlaps used in genome assemblies?

Overlaps between reads allow genome assemblers to stitch reads together into the assembled genome. To demonstrate basic comprehension, students could discuss how an overlap between two sequences, AGT and GTT, can be used to infer a larger sequence, AGTT. An advanced student might discuss how overlaps between reads or subsequences within reads form an assembly graph that is compressed into contigs. This level of sophistication would require reading the tutorials that detail specific assembly methods (http://gcat.davidson.edu/phast/olc.html and http://gcat.davidson.edu/phast/debruijn.html).

What factors complicate genome assembly?

There are many factors that can complicate genome assembly, and PHAST touches on a few. Students might discuss how sequencing errors must be identified and resolved when computing the consensus sequence (which ties back to the importance of genome coverage). Advanced answers might link concepts back to the assembly graph. For example, it is important to identify and clean adaptor sequences because they lead to false overlaps that obscure the assembly graph and contig generation. Students could also mention how repetitive sequences complicate assembly by creating alternative assembly paths through an assembly graph. Although not discussed in detail, sequencing errors also complicate the assembly graph, especially in de Bruijn assemblers (http://gcat.davidson.edu/phast/debruijn.html), by introducing new subsequences.

○ Conclusion

Biology is benefiting from a data explosion due to high throughput technologies. With massive amounts of data, the study of biology has become extremely interdisciplinary, often requiring high-performance computational approaches to make analysis headway. Modern biology demands that today's students be able to think across disciplines and use tools from mathematics and computer science in order to succeed in their professional lives (NRC, 2003; AAMC & HHMI, 2009; AAAS, 2011). With PHAST, students can interact with real data from the highly interdisciplinary field of genomics. PHAST bundles powerful bioinformatics tools in a convenient, user-friendly Web interface, allowing students to manipulate raw data sets and appreciate the mathematical concepts and computer science techniques used to process such information. Teachers can use this free Web tool to enhance their existing curricula so that students can learn about the important issue of genome assembly through interactive experimentation in real time.

○ Acknowledgments

Funding was provided by the Davidson Research Initiative, the Mimms Summer Research Fellowship, and the James G. Martin Genomics Program. We are grateful for support from those involved with the HHMI PHAGES program: Dr. Graham Hatfull, Dan Russell, Dr. Tuajuanda Jordan, Melvina Lewis, and Dr. Lucia Barker. Special thanks to Bill Hatfield and Paul Brantley for help configuring the PHAST server.

References

AAAS. (2011). *Vision and Change in Undergraduate Education: A Call to Action*. Washington, DC: American Association for the Advancement of Science.

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, *408*, 796–815.

Association of American Medical Colleges & Howard Hughes Medical Institute. (2009). *Scientific Foundations for Future Physicians: Report of the AAMC-HHMI Committee*. Washington, DC: AAMC.

Blattner, F.R., Plunkett, G., III, Bloch, C.A., Perna, N.T., Burland, V., Riley, M. & others. (1997). The complete genome sequence of *Escherichia coli* K–12. *Science*, *277*, 1453–1462.

C. elegans Sequencing Consortium. (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, *282*, 2012–2018.

Chevreux, B., Wetter, T. & Suhai, S. (1999). Genome sequence assembly using trace signals and additional sequence information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics*, *99*, 45–56.

College Board. (2012). *AP Biology Course and Exam Description.* Available online at http://media.collegeboard.com/digitalServices/pdf/ap/2012advances/AP-Biology_CED_Fall2012.pdf.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R. & others. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, *269*, 496–512.

Hatfull, G.F., Pedulla, M.L., Jacobs-Sera, D., Cichon, P.M., Foley, A., Ford, M.E. & others. (2006). Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. *PLoS Genetics*, *2*(6), e92.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921.

Krumsiek, J., Arnold, R. & Rattei, T. (2007). Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, *23*, 1026–1028.

Mouse Genome Sequencing Consortium. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*, 520–562.

National Research Council. (2003). *Bio2010: Transforming Undergraduate Education for Future Research Biologists.* Washington, DC: National Academies Press.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J. & others. (2011). Tumour evolution inferred by single-cell sequencing. *Nature*, *472*, 90–94.

Pope, W.H., Jacobs-Sera, D., Russell, D.A., Peebles, C.L., Al-Atrache, Z., Alcoser, T.A. & others. (2011). Expanding the diversity of mycobacteriophages: insights into genome architecture and evolution. *PLoS ONE*, *6*(1), e16329.

Tringe, S.G., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W. & others. (2005). Comparative metagenomics of microbial communities. *Science*, *308*, 554–557.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S. & Banfield, J.F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*, 37–43.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W. & others. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, *304*, 66–74.

Wetterstrand, K.A. (2012). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available online at http://www.genome.gov/sequencingcosts.

D. LELAND TAYLOR is a Computational Biology major at the Center for Interdisciplinary Studies, Davidson College, Box 7118, Davidson, NC 28035; e-mail: leland.taylor@nih.gov. A. MALCOLM CAMPBELL is Professor of Biology, Davidson College, Box 7118, Davidson, NC 28035; e-mail: macampbell@davidson.edu. LAURIE J. HEYER is Professor of Mathematics, Davidson College, Box 6959, Davidson, NC 28035; e-mail: laheyer@davidson.edu.