Using experimental evolution and next-generation sequencing to teach bench and bioinformatic skills

Alexander S. Mikheyev[1*] and Jigyasa Arora[1]

[1]Ecology and Evolution Unit

Okinawa Institute of Science and Technology

1919-1 Tancha

Onna-son, Kunigami-gun

Okinawa, Japan, 904-0412

[*]Corresponding author

email: sasha@homologo.us

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1356v1 | CC-BY 4.0 Open Access | rec: 9 Sep 2015, publ: 9 Sep 2015

Abstract

Advances in sequencing technology have exponentially increased data-generating capabilities, and data analysis has now become the major hurdle in many research programs. As sequencing tools become more accessible and automated, experimental design and data analysis skills become the key factors in determining the success of a study. However, proper bioinformatic analysis also relies on a deep understanding of laboratory workflow, in order to prevent biases in the data. This is particularly true if commercial kits are used, as proprietary reagents frequently obfuscate underlying reactions and their conditions. Here we present a training module that seamlessly combines laboratory components (experimental evolution of T5 bacteriophage resistance by *Escherichia coli*, and library preparation), with bioinformatic analysis of the resulting data. Students conduct a simple genetic variant discovery experiment in the course of about a week. The module uses mature Illumina chemistry for both library preparation and sequencing, though it can be modified for use with any sequencing platform. Because most students do not use Linux, the bioinformatic pipeline is available inside a cross-platform virtual machine, simplifying software installation, and providing a non-threatening introduction to the command line. The analysis, which is made simpler by the fact that most resistance mutations occur in one gene, making them easier to find, emphasizes the potential pitfalls of using short-read data for mutational analysis, and explores biases inherent to the methodology. This module can fill an existing training gap in advanced undergraduate, or early graduate education, allowing student to experience first-hand design, execution, and analysis of next-generation sequencing experiments.

Keywords: bioinformatics, education, experimental evolution, microbiology, phages, sequencing

Introduction

The next-generation sequencing revolution has created a demand for researchers with an understanding of both laboratory techniques and bioinformatic analysis. However, the two skill sets are still typically taught separately, though there are increasing efforts to integrate them (Furge *et al.* 2009; Lopatto *et al.* 2008; Robertson & Phillips 2008; Temple *et al.* 2010; Weisman 2010). Lack of bioinformatic skills by laboratory scientists can lead to poor experimental design, for example as a result of inadequate replication (Lynn *et al.* 2003). Similarly, an inadequate knowledge of laboratory techniques can make bioinformaticians oblivious to potential sources of data bias. Here we present an integrative educational module that examines the outcome of an evolutionary interaction, and covers all parts of a scientific investigation from bench work to genetic variant discovery and analysis. The goal of this exercise is to allow students to complete an entire study, very similar in design to experiments resulting in publishable data, with emphasis on proper microbiological, molecular, and bioinformatic technique throughout.

This set of exercises builds on those developed by Hyman (2014) for introductory biology students, modifying them for advanced undergraduate or early graduate students by incorporating next-generation sequencing and bioinformatic analysis. The module does not require advanced molecular biology skills, and can be performed in a basic molecular biology laboratory, though some steps, such as DNA quantification may rely on equipment not found in regular teaching laboratories. However, the instructor can easily complete this step between classes. The in-class bioinformatic analyses can be performed without programming, focusing instead on the Linux command line, and on interactive data exploration. Additional homework problems do require some elementary programming skills, at a level that can be self-taught in about a week, though we strongly encourage a lecture component introducing

basic programming and statistical analysis.

Bacteria and phages as research and teaching tools. Bacteria and phages offer exciting possibilities for laboratory and bioinformatic instruction. In the mid-20th century, the relative simplicity of the bacteriophage was crucial to understanding the genetic code and the fine structure of genes (Benzer 1961; Cairns *et al.* 2007; Crick *et al.* 1961). The reproducibility and elegance of these experiments, as well as the relative simplicity of phage genomes, make them powerful teaching tools (Allen & Gyure 2013; Caruso *et al.* 2009; Hatfull *et al.* 2006; Hyman 2014; Jordan *et al.* 2014; Temple *et al.* 2010). A commercially available 'canned' laboratory based on Benzer's gene substructure mapping work also exists (Carolina Biological #124550).

The small size of the *Escherichia coli* genome made it among the first to be completely sequenced (Blattner *et al.* 1997). Although that was a major undertaking at the time, we can now re-sequence many *E. coli* isolates at high coverage using a bench top sequencer. Microbial genome analysis is also increasingly being used in a teaching environments (Ditty *et al.* 2010). Streamlined library preparation protocols also make it possible to quickly and reliably prepare samples for sequencing using commercial kits. Currently, it is possible to go from experiment to genome-wide sequence data in the course of a few days, a turnaround time that allows the entire experimental process to be replicated in a teaching lab.

T5 phage system. The T5 phage is a member of the Family Siphoviridae, which is characterized by a long, flexible, non-contractile tail and an isometric, icosahedral capsid, containing the double-stranded DNA genome (Effantin *et al.* 2006). It has a large 121,752 bp genome with composition that is typical of members of this genus (Wang *et al.* 2005). T5 infects *E. coli* by binding to the bacterial ferrichrome transporter, encoded by the FhuA gene, and specifically to its gating loop (Killmann *et al.* 1995). In the course of several terms using

this lab for undergraduate instruction, Hyman (2014) found that resistance typically evolves by removing the receptor target for T5 via a wide variety of knockout mutants. The reproducibility of the evolutionary outcome in this relatively simple and well-characterized system makes it ideal for an introduction to bioinformatic analysis. The obvious candidate gene focuses the analysis of *E. coli* mutants resistant to T5, on one small region of the bacterial genome, greatly minimizing the complexity inherent to next-generation sequencing data sets. The candidate gene approach also allows the introduction of more sophisticated concepts, such as the role of sequencing alignment software, and why some types of mutations are not easily detected by short-read re-sequencing approaches.

Laboratory Exercises

Students should ideally perform the laboratory exercises alone, in order to gain experience with the full range of techniques. We designed the module so that each student can sequence a single mutant, at a total cost of under $100 per student (Table 1). The major cost is sequencing reagents, which do not decrease in price for classes of fewer than a couple dozen students, so there are few advantages to grouping students.

Freeze-dried cultures of both the T5 phage and the host *E. coli* B strain can be obtained from ATCC (catalog numbers 11303-B5 and 11303). The phage and bacteria can be easily propagated and stored for use between classes, offsetting the initial investment. The *E. coli* B strain genome has been sequenced (Jeong *et al.* 2009), and can be used in bioinformatic analysis.

Day 1: Evolving resistance. The experimental evolution component of the lab takes place quickly when an overnight bacterial culture is exposed to a high titer of phage. The actual class exercise should require less than half an hour, since it merely involves mixing the bacteria

and phage, and plating them for overnight incubation. It can be paired with a lecture.

Day 2: Isolating mutants. Resistant mutants grow overnight, and should be streaked to isolate single colonies. Again this does not take long, although it is best to have students streak excessive numbers of colonies to learn this skill. They will not see the results of their technique until the next day, so this is a possible failure point. If there are few students, we suggest that the instructor also make several so as to have reserve sets of properly streaked plates. This activity should take about 30 minutes.

Day 3: Confirming resistance. Isolated colonies should be checked to verify that they are indeed *E. coli* and that they are resistant to the phage. We use a PCR-based method for species verification (Chen & Griffiths 1998). This provides the students with an opportunity to practice reaction set-up for later library preparation. Since the reactions of the entire class will most likely be incubated on one thermocycler, we recommend using a hot-start DNA polymerase in order to minimize differences in reaction start time between students. We optimized the PCR, so that it takes under an hour, and the next phase of the lab can take place during this time. In practice, verification of resistance takes somewhat less time than that required for PCR, creating an opportunity for another activity. With loading, running, and visualizing results on a gel, this exercise takes under two hours.

The same colony can be used for species verification, and for confirming resistance. Mixing some of the putative mutant with T5 phage stock, and incubating overnight on a Petri plate confirms resistance. A positive control should be included, to test for any decrease in phage titer. In essence, this is the same experiment as on Day 1, but with a different outcome for mutant cells.

Day 4: DNA extraction. Overnight cultures prepared on Day 3 from resistant *E. coli* mutants can be extracted using commercial kits (Qiagen DNeasy Blood and Tissue). This is a simple procedure, but we found that DNA may need to be subjected to an additional purification step, or the tagmentation reaction may be inhibited during library preparation. Consequently, we include an additional de-salting step using centrifugal filter devices (Amicon Ultra-0.5). The entire exercise should take about an hour.

Day 5: Illumina Nextera library preparation. This is technically the most challenging part of the entire experiment (and the most expensive), so it should be approached with care. The actual set of reactions is not difficult, but attention must be paid that they are conducted following the manufacturer's instructions. In order to minimize the cost of library preparation, we halve the volume of all reagents. This exercise should take about an hour and a half. There will be an hour-long interval between the start of PCR and running of the gel, and it should probably be filled with a lecture or some other activity.

We also include a PCR purification and size selection step following Tin *et al.* (2015), in the instructor manual (see Data Accessibility), which generates an optimal range of fragments for sequencing. This step can be omitted in favor of a kit-based PCR purification step performed by the students, *e.g.*, using a commercial kit, but sequencing performance will likely suffer. However, the experiment is designed with redundant sequencing coverage in mind, so it should be reasonably robust.

Bioinformatic analysis

One of the major obstacles to conducting sophisticated bioinformatic analysis is the wide range of tools that can be used, and achieving computing environment consistency for each student (Cummings & Temple 2010). Each of these tools has platform-specific installation instructions

and requirements. Although compiling and running software is an important part of bioinformatics, it is not necessarily instructional. Since the vast majority of computational clusters use some form of the Linux operating system, and most students tend to use Apple or Windows machines, installation skills may be non-transferable. Another obstacle to bioinformatic analysis is the slow speed of computation, particularly on a laptop computer, even with relatively small genome-scale data sets. To circumvent both these problems, we prepared the complete pipeline and intermediate files produced by all time-intensive data analysis steps as part of a virtual machine (VM) implemented in cross-platform VirtualBox software (Oracle). The VM provides a 'plug-and-play' computational environment that works together with source code and instructions stored in a `git` repository that guides the students through a realistic analysis of their data. Rather than using a VM, other approaches are possible, such as installing the required software cloud services such (e.g., Amazon's E3, or iPlant's Atmosphere), or an institutional computational cluster.

In-class exercises. Before the students are allowed to analyze the data, they will need a brief introduction to the Linux operating system, and possibly a quick introduction to computer programming (both are beyond the scope of this lab). After downloading the VM and data before class, the students go through an interactive exploration of the re-sequenced bacterial data. They are provided with an alignment of reads from the resistant mutants and the original strain, and the output of a variant call analysis pipeline (Danecek *et al.* 2011; Garrison & Marth 2012; Langmead & Salzberg 2012). Results are visualized using IGV (Thorvaldsdóttir *et al.* 2012), which allows simultaneous display of annotations, aligned reads, and variant calls (Figure 1). The in-class laboratory component allows students to interact with their instructor regarding operation of various software components and interpretation of the data. No programming background is necessary for the in-class exercises, although it requires fairly close

attention to detail. We find that this activity takes well over an hour for most students, and is even capable of filling a 2-3 hour lab period.

Homework programming exercises. A more challenging set of exercises is available for students who have some programming and data analysis skills. These exercises combine higher-level data analyses similar to those that would be performed by bioinformaticians in practice. They depart from 'point-and-click' analyses of the data to address important issues, such as the presence of possible false positives and bias in the data. They also allow students to transition between merely using available tools to writing their own, which is a key component of bioinformatics education (Counsell 2003). We see them as a critical component of the laboratory module and recommend that some basics of programming and some elementary statistical concepts, such as correlation, be taught alongside the lab in order that the students perform them independently. We also feel that it is important that the students figure out as much of the data analysis as possible on their own, with solutions discussed in a subsequent class.

Discussion

At the completion of this set of exercises, students will have been exposed to many of the skills necessary to conduct a molecular ecology investigation. More importantly, they will understand molecular and bioinformatic components that go into such an investigation, allowing them to plan their own studies, building on the skills learned in this module. At the very minimum, this module provides students an opportunity to witness an evolutionary event and to functionally characterize it in a remarkably short period of time, illustrating the power of modern technology to provide fundamental biological insights.

The current reagent cost for this teaching module is less than $100, assuming one mutant per

student (Table 1). In principle, this cost can be halved by simply omitting the actual sequencing and using the provided data instead, or reduced even further by eliminating the library preparation step. Although this has the potential to minimize the impact of the exercise, the analysis would be functionally the same. Because all the bacterial isolates are individually indexed during library preparation, it may be possible to reduce sequencing costs by running them together with other projects, although the timing of other experiments would need to be carefully arranged. We intend to add additional data to this exercise over time, possibly using other sequencing platforms, which will provide additional material.

The pace of sequencing platform development has been breathtaking, with fundamentally novel platforms being released regularly (Mikheyev & Tin 2014). This ensures that many of the laboratory techniques described here will be obsolete in coming years. However, the basic evolutionary interaction used in this teaching module will be suitable for analysis using the next generation of sequencing tools, and the teaching module can be easily adapted to work with them also. Indeed, many of the mutations involve structural variations such as large deletions and transposon insertions (Hyman 2014) (Figure 1), which are more difficult to analyze using short-read data (Alkan *et al.* 2011; Medvedev *et al.* 2009), and which the longer reads of new sequencing platforms should detect with greater ease.

This module may be adapted in a large number of ways that could even result in publishable, short-term, descriptive studies. For instance, one could combine novel *E. coli* phage isolation from sewage (Bisen 2014; Luciano *et al.* 2002) to identify novel phages, and what receptor sites they bind. Alternatively, one could compare biases and outcomes from different library preparation methods, or of bioinformatic techniques, such as reference-based mapping *vs. de novo* assembly. One could imagine even more next-generation sequencing based exercises that

use the elegance of the bacteria/phage interaction to explore the microbial world, and the ecological and evolutionary forces that shape it.

On a final note, the corresponding author would like to hear if anyone carries out this exercise to exchange notes. Although the laboratory includes a specific implementation of the exercise and analysis, numerous modifications and improvements are possible. The corresponding author would be happy to curate additional data sets, and maintain a page with different versions of the protocols. If this teaching tool proves useful, it can evolve into a more sophisticated and comprehensive community-driven resource.

Acknowledgements. We thank Paul Hyman for extensive advice on his laboratory and for helpful feedback on the manuscript. We thank OIST class B02 (Introduction to Biology), who served as the guinea pigs for this lab, for their feedback. We are grateful to Mandy Tin and Qiu Liu for laboratory assistance. We are grateful to Steven D. Aird for editing the manuscript.

Data accessibility

The student and instructor manuals are available at: http://t5-lab.homologo.us/

The virtual machine and data will be available on DataDryad after manuscript acceptance, but links to them are available on the above site for review.

References

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. *Nature Reviews Genetics* **12**, 363-376.

Allen ME, Gyure RA (2013) An undergraduate laboratory activity demonstrating bacteriophage specificity. *Journal of Microbiology & Biology Education* **14**, 84-92.

Benzer S (1961) On the topography of the genetic fine structure. *Proceedings of the National Academy of Sciences of the United States of America* **47**, 403-415.

Bisen PS (2014) *Laboratory Protocols in Applied Life Sciences* CRC Press.

Blattner FR, Plunkett G, Bloch CA, *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453-1462.

Cairns J, Stent GS, Watson JD (2007) Phage and the Origins of Molecular Biology, The Centennial Edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Caruso SM, Sandoz J, Kelsey J (2009) Non-STEM undergraduates become enthusiastic phage-hunters. *CBE-Life Sciences Education* **8**, 278-282.

Chen J, Griffiths M (1998) PCR differentiation of *Escherichia coli* from other Gram‐negative bacteria using primers derived from the nucleotide sequences flanking the gene encoding the universal stress protein. *Letters in Applied Microbiology* **27**, 369-371.

Counsell D (2003) A review of bioinformatics education in the UK. *Briefings in Bioinformatics* **4**, 7-21.

Crick FH, Barnett L, Brenner S, Watts-Tobin RJ (1961) General nature of the genetic code for proteins. *Nature* **192**, 1227-1232.

Cummings MP, Temple GG (2010) Broader incorporation of bioinformatics in education: opportunities and challenges. *Briefings in Bioinformatics*.

Danecek P, Auton A, Abecasis G, *et al.* (2011) The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158.

Ditty JL, Kvaal CA, Goodner B, *et al.* (2010) Incorporating genomics and bioinformatics across the life sciences curriculum. *PLoS Biology* **8**.

Effantin G, Boulanger P, Neumann E, Letellier L, Conway J (2006) Bacteriophage T5 structure reveals similarities with HK97 and T4 suggesting evolutionary relationships. *Journal of Molecular Biology* **361**, 993-1002.

Furge LL, Stevens‐Truss R, Moore DB, Langeland JA (2009) Vertical and horizontal integration of bioinformatics education. *Biochemistry and Molecular Biology Education* **37**, 26-36.

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.

Hatfull GF, Pedulla ML, Jacobs-Sera D, *et al.* (2006) Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. *PLoS Genetics* **2**.

Hyman P (2014) Bacteriophage as instructional organisms in introductory biology labs. *Bacteriophage* **4**.

Jeong H, Barbe V, Lee CH, *et al.* (2009) Genome sequences of *Escherichia coli* B strains REL606 and BL21 (DE3). *Journal of Molecular Biology* **394**, 644-652.

Jordan TC, Burnett SH, Carson S, *et al.* (2014) A broadly implementable research course in phage discovery and genomics for first-year undergraduate students. *MBio* **5**, e01051-01013.

Killmann H, Videnov G, Jung G, Schwarz H, Braun V (1995) Identification of receptor binding sites by competitive peptide mapping: phages T1, T5, and phi 80 and colicin M bind to the gating loop of FhuA. *Journal of Bacteriology* **177**, 694-698.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-359.

Lopatto D, Alvarez C, Barnard D, *et al.* (2008) UNDER GRADUATE RESEARCH: Genomics Education Partnership. *Science* **322**, 684-685.

Luciano CS, Young MW, Patterson RR (2002) Bacteriophage: a model system for active learning. *Microbiology Education* **3**, 1.

Lynn DJ, Lloyd AT, O'Farrelly C (2003) Bioinformatics: implications for medical research and clinical practice. *Clinical and Investigative Medicine* **26**, 70-74.

Medvedev P, Stanciu M, Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods* **6**, S13-S20.

Mikheyev AS, Tin MM (2014) A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources* **14**, 1097-1102.

Robertson AL, Phillips AR (2008) Integrating PCR Theory and Bioinformatics into a Research-oriented Primer Design Exercise. *CBE-Life Sciences Education* **7**, 89-95.

Temple L, Cresawn SG, Monroe JD (2010) Genomics and bioinformatics in undergraduate curricula: Contexts for hybrid laboratory/lecture courses for entering and advanced science students. *Biochemistry and Molecular Biology Education* **38**, 23-28.

Thorvaldsdóttir H, Robinson JT, Mesirov JP (2012) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, bbs017.

Tin MM, Rheindt FE, Cros E, Mikheyev AS (2015) Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data

improve genotype calling accuracy. *Molecular Ecology Resources* **15**, 329-336.

Wang J, Jiang Y, Vincent M*, et al.* (2005) Complete genome sequence of bacteriophage T5. *Virology* **332**, 45-65.

Weisman D (2010) Incorporating a collaborative web‐based virtual laboratory in an undergraduate bioinformatics course. *Biochemistry and Molecular Biology Education* **38**, 4-9.

Figure 1. Visualization of a segment of the *E. coli* genome, including the FhuA gene, with three different mutation classes: deletion (top), transposon insertion (center), and an amber nonsense mutation mutation (bottom). Each panel has the coverage histogram on top, and individual reads on the bottom. In the top panel there are no mapped reads, suggesting that the reference sequence is missing, as would be expected for a deletion. In the middle panel, there is a gap in coverage and apparent short insertions. Inspection of the rest of the read (clipped during alignment) would show an IS insertion site. In the bottom panel a single nucleotide substitution (orange bars) introduced a stop codon. Lines with various other colors interspersed through the data are sequencing errors. Structural mutations, such as the transposon insertion and deletion are difficult to detect using short read mapping to a reference, and they serve as good examples of the dangers of relying blindly on software output.

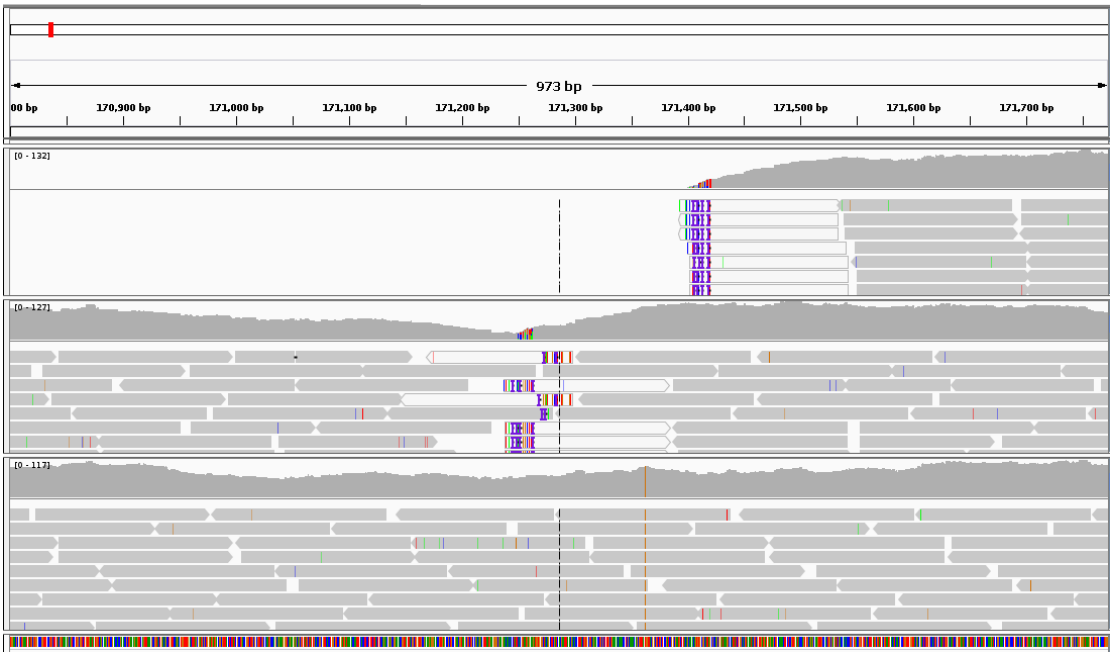

Table 1. Approximate costs associated with the teaching module. The sequencing cost assumes that the entire MiSeq flow cell is used to achieve 30x coverage of the bacterial genome.

Item	**List price (US$)**	**Reactions**	**Cost per mutant**
Amicon Ultra	44	8	5.5
DNeasy Blood & Tissue Kit	158	50	3.2
Nextera® XT DNA Sample Preparation Kit	775	48	16
Nextera XT Index Kit	230	96	2
MiSeq Reagent Kit v3 (150 cycle)	850	24	35
PCR reagents, agarose gel, media, *etc.*			10
Total:			**73**