A short course in analyzing next-generation sequencing data

September 25, 2012

Keywords: bioinformatics, next generation sequencing, genomics.

Project Narrative

Many biomedical researchers are not trained in computational tools that would help them make use of genomic and other bioinformatics data. We propose to continue teaching a two-week short course for advanced researchers that will help train them to take advantage of sequence data in their research.

1 Project Summary

Modern biomedical research is increasingly making use of genome-scale data from next-generation sequencing platforms, including Illumina HiSeq and MiSeq machines and Pacific Biosciences SMRT. These platforms make it possible for individual labs to quickly and cheaply generate vast amounts of genomic and transcriptomic data from *de novo* sequencing, resequencing, ChIP-seq, mRNA-seq, and allelotyping experiments.

Despite this ability to quickly generate large data sets, **biologists are rarely trained in the computational and statistical techniques necessary to make sense of this data**. Thus, many researchers must rely on others – often computational scientists with little biological training – to design and implement appropriate data reduction and data mining techniques. Moreover, most institutions do not have access to the substantial computational resources necessary to run these analyses.

We will continue to help bridge this gap with a short, two-week intensive summer course, by teaching biomedical researchers to (1) run analyses on remote UNIX servers hosted in the Amazon Web Services "cloud"; (2) perform mapping and assembly on large short-read data sets; (3) tackle specific biological problems with existing short-read data; and (4) design computational pipelines capable of addressing their own research questions. This will be accomplished by in-depth hands-on practical training in the relevant techniques. Our experience, confirmed by assessment, is that this practical training leads to a substantial improvement in the basic computational sophistication of participants. We believe that in the long term our cadre and those of other courses will contribute to a significant improvement in the general area of data-driven biology.

This short course will continue to help train the current and next generation of independent biomedical researchers in basic computational thinking and procedure, as well as teaching them how to make use of scalable Internet computing resources for their own research. Moreover, we will continue to develop and extend our extensive online materials, which are freely available online and widely used. Our end goal is increase the efficiency and sophistication with which biomedical researchers make use of novel sequencing technologies.

For this renewal, we propose to continue offering the course at a low cost; expand our RNAseq discussion; address student needs by expanding the available materials for learning programming and UNIX; and increase our statistics component significantly.

2 Specific Aims

The vast increase in sequencing capacity over the last few decade presents many opportunities for biologists. However, this democratization of sequencing capacity has also led to “data overload”, in which individual scientists produce large sequence datasets but cannot effectively analyze them. In large part this is because most biologists lack training in computational data analysis. Combined with the vast increase in data-gathering abilities and a general lack of substantial biological cyberinfrastructure, researchers are now often blocked from analyzing their own data.

Our long-term goal is to systematically increase the capabilities and sophistication of biomedical researchers in thinking about and analyzing large sequence data sets. In support of this goal, **we propose a short (two week), intermediate course to disseminate computational sequence analysis skills specifically designed for analysis of next-generation sequencing data.** Our objective for this course is to provide biological researchers with explicit, portable, hands-on technical training in applying cutting-edge computational tools to existing data sets, and to do this training with remote “cloud computing” resources, which enables the researchers to apply greater resources in the future as needed. **We have run this course every summer since 2010, and are funded for the 2013 course.** Our course is designed to achieve the following Specific Aims:

Specific Aim 1: Teach practical remote use of UNIX systems on the Amazon cloud computing platform Most sequence analysis software is developed for and runs on UNIX systems. We use the Amazon Elastic Cloud Computing (EC2) system to provide compute resources during the class. EC2 provides a range of machine types on a rental basis, including large-memory, multicore, and cluster systems, together with a range of operating systems. We introduce students to basic configuration and installation of UNIX software on a Debian system through copy-paste “recipes”, as well as showing them UNIX navigation, download of remote data, running remote programs, and automation through simple scripting. For the 2012 workshop, we also started using the IPython Notebook to provide interactive data exploration abilities [Pérez and Granger, 2007].

Specific Aim 2: Introduce short-read mapping and assembly techniques Analysis of genome and transcriptome sequence data generally starts with either de novo assembly or mapping of that data to a reference genome. We introduce Bowtie and BWA, several common mapping tools, as well as Velvet and SGA, two common and freely available de Bruijn graph assemblers. We show how to install and run them, how to apply them to existing data sets from Illumina and PacBio systems, and discuss parameter choice and analysis evaluation.

Specific Aim 3: Use sample data sets to tackle biological problems We provide students with a variety of existing data sets from genome sequencing, genome resequencing, mRNA-seq, ChIP-seq, and allelotyping projects. (Increasingly, students are coming with their own data sets.) We use these data sets to demonstrate a number of software packages built for these kinds of analysis address issues of up-front experimental design and their downstream consequences, and use them as foils to illustrate the various features and biases present in each type of data set.

Specific Aim 4: Work with individual students to develop research-oriented computational approaches By the end of the first week (after learning how to do mapping and assembly), students are eager to work with their own data; when they do not yet have their own data, we provide similar data sets from publications or collaborations. We then work with students to develop computational approaches and simple pipelines to help them appropriately analyze this data.

3 Research Education Program Plan

3.1 Proposed Research Education Program

The vast sequencing capacity now available presents many opportunities for biologists to study genome content, large-scale population heterogeneity, whole-transcriptome response to perturbation, evolution of microbial pathogenesis and drug resistance, and even entire microbial communities. The resulting "data overload" in biology challenges the expertise of many biologists. The omission of computation in traditional biological and biomedical is now exacerbated by the speed with which computational biology is moving. Many researchers and institutions do not possess sufficient computational infrastructure to perform large-scale sequence analysis, either. **This combined lack of training and resources has led to a "perfect storm" of sequence analysis, in which researchers are blocked from analyzing existing data sets relevant to their research.**

Our proposed training program addresses this in a short, intensive, two week period by, first, introducing biologists to general computational practice, including UNIX and scripting; second, making use of standardized "cloud" computing machines that allow scaling of compute resources on demand; third, providing the underlying theory behind mapping and assembly; and fourth, giving practical, hands-on tutorials in applying cutting-edge research software to large data sets.

To teach this course, we have assembled a team with strong backgrounds in sequence analysis, statistics, molecular biology, genomics, and population genetics, as well as substantial expertise in computational data analysis and software development. **We have already run this course three times.**

Our course is designed to achieve the following Specific Aims:

Specific Aim 1: Teach practical remote use of UNIX systems on the Amazon cloud computing platform Most sequence analysis software is developed for and runs on UNIX systems. We will use the Amazon Elastic Cloud Computing (EC2) system to provide compute resources during the class. EC2 provides a range of machine types on a rental basis, including large-memory, multicore, and cluster systems, together with a range of operating systems. We will introduce students to basic configuration and installation of UNIX software on a Debian system through copy-paste "recipes", as well as showing them UNIX navigation, download of remote data, running remote programs, and automation through simple scripting.

Specific Aim 2: Introduce short-read mapping and assembly techniques Analysis of genome and transcriptome sequence data generally starts with either de novo assembly or mapping of that data to a reference genome. We will introduce Bowtie and BWA, several common mapping tools, as well as Velvet and ABySS, two common and freely available De Bruijn graph assemblers. We will show how to install and run them, how to apply them to existing data sets from Illumina, Ion Torrent, and PacBio systems, and how to choose parameters and evaluate parameter choice.

Specific Aim 3: Use sample data sets to tackle biological problems We will provide students with a variety of existing data sets from genome sequencing, genome resequencing, mRNA-seq, ChIP-seq, and allelotyping projects. We will use these data sets to demonstrate a number of software packages built for these kinds of analysis, address issues of up-front experimental design and their downstream consequences, and use them as foils to illustrate the various features and biases present in each type of data set.

Specific Aim 4: Work with individual students to develop research-oriented computational approaches By the end of the first week (after learning how to do mapping and assembly), our experience is that students are eager to apply their skills to their own data. We encourage students to bring personal data sets, and if none are available we can often provide similar data sets; this allows students to work with data sets that apply to their research. We will work with students to develop computational approaches and simple pipelines to help them appropriately analyze this data.

Our evaluation of this overall approach demonstrates that we markedly increase the computational capabilities and sophistication of students, help them develop the mental outlook necessary for computational data analysis, and empower them to tackle future research. By using Amazon EC2, we also show them how to use substantial compute resources without a large up-front infrastructure investment; and, because we use normal Linux systems on Amazon EC2, often these skills transfer to commonly available compute systems at their home institution. Finally, by making our notes openly available (see http://ged.msu.edu/angus/), and linking them to several external sites, including the BioStar Q&A Web site and the Software Carpentry computational science tutorials, we provide long-term and open resources for self-training and growth in computational science.

3.2 Background and Significance

Biology is faced with an ever-increasing deluge of genomic and transcriptomic data from next-generation sequencers. Yet most biologists lack the computational experience and expertise to analyze this data, extract hypotheses for later biological validation, and validate biological hypotheses with computational data. In addition to this "expertise gap", the landscape of sequencing technologies continues to rapidly change, with technologies such as Illumina HiSeq moving to maturity, while the Ion Torrent, Illumina MiSeq, and PacBio SMRT platforms begin production use.

This shifting landscape seems unlikely to become static anytime soon, which presents a number of problems for researchers:

- First, the tools and approaches that work for yesterday's technology do not work for today's; witness, for example, the replacement of overlap-layout-consensus assemblers in favor of De Bruijn graph assemblers such as Velvet, which scale much better with large amounts of data. Even newer assemblers, including ALLPATHS-LG and SGA, are now in common use [Gnerre et al., 2011, Simpson and Durbin, 2012].

- Second, the biases and errors present in one technology often do not apply to another technology, which requires radically different approaches to handling data (e.g. Illumina vs Torrent vs PacBio).

- Third, it is not easy to match the appropriate technology to a given problem - e.g. we are commonly asked if paired-end sequencing from Illumina is important for transcriptomics, or if PacBio will replace mRNAseq.

- Fourth, standard tools in common use simply do not apply to new types of sequences, but the reasons why are not always clear. For example, some researchers are still using BLAST to map large quantities of Illumina reads to reference genomes, despite the inappropriate default gapping model and poor scaling of BLAST for this purpose.

- Fifth, the different tools that do exist are only usable from the command line, are often difficult to build and install, and usually possess little or no documentation. This renders them useless to most biologists.
- Sixth, each tool provides a plethora of command line options, with tradeoffs that are inappropriate for various tasks because of the heuristics. For example, the two basic alignment modes of the commonly used bowtie mapping tool can return radically different answers, of widely variant utility for resequencing and transcriptome analyses.
- Seventh, the volume of data that is produced by the sequencers overwhelms the compute infrastructure available to and accessible by most biologists.

This collection of problems is further compounded by the continued divide in training between the more numerical sciences (physics, math, and chemistry) and the biomedical sciences, leading to a very small intersection between those who both think computationally and also understand biology. Were next-generation sequence analysis simply a rote exercise, this lack of training would be less of a problem; however, the rate of technological change in this area makes rote learning ineffective.

This "perfect storm" of inadequate training, rapidly changing sequencing technologies, tools and algorithms, and insufficient infrastructure, means that many biomedical researchers are incapable of taking advantage of the new sequencing technologies. Those that *are* capable often struggle to connect the dots between the biology and the computation, or reach incorrect conclusions due to bias, error, and incorrect tool use.

Internet resources There are a number of mailing lists, Web forums, and tutorial sites available to help scientists make use of new sequencing technologies and tools - for example, the 'velvet-users' mailing list discusses sequence assembly tips and tricks, BioStar is a question and answer site for both novice and experienced bioinformaticians, and SeqAnswers.com is a general Web forum for discussing next-generation sequencing. However, most of these fora assume a minimum level of computational facility with basic UNIX commands and simple file formats, as well as short shell scripts and Perl or Python programs. Thus they are not accessible to most biology researchers without any prior training.

Workshops and courses One potentially effective way to train current biology researchers in next-generation sequence analysis is to provide training through short, bootcamp-style workshops or short courses. There are a number of courses that take this general approach.

Two of the most focused and integrated courses available for *computational training* are our course, run every May since 2010 at the Kellogg Biological Station (KBS) Michigan State University (MSU); and a course at UC Davis that is also run every summer[1]. Both courses focus on training biologists in basic command line UNIX and a broad range of the most common tools for mapping, assembly, and further processing of data. The MSU course integrates cloud computing into its curriculum, doing all analyses on the Amazon cloud, while the UC Davis short course offers a separate two-day cloud computing workshop. The MSU course is also residential, with a two week stay at KBS required in order to take the course.

[1] http://bsc2010.bioinformatics.ucdavis.edu/

Other courses, such as those offered by the Wellcome Trust [2] and Cold Spring Harbor Labs[3], are centered on *experimental* training and sample handling, and do not focus on computational data handling or analysis. CSHL does offer an advanced computational course in comparative genome analysis, but it is designed for students who already have some background in computational biology[4].

Regardless, we believe that the significant demand evidenced by our receipt of over 300 applications in two years demonstrates that our course is neither redundant nor unnecessary.

3.3 Progress Report

Our course was initially funded in 2010 by Michigan State University, which saw a need to train students and postdocs in analyzing NGS data. The course was held at the Kellogg Biological Station, about 90 minutes west of the Lansing airport. We received 33 applications for the course and admitted a total of 24 students – 20 students from graduate and postdoc programs (including MSU, U. Chicago, UC Irvine, and Yale), along with three additional industry scientists.

We ran the course again in 2011 (with MSU funding) and received 133 applications, from students in 23 states and 16 countries. We admitted 24 of these students (see Figure 1). The majority of the course attendees were graduate students, with 5 postdocs and 3 faculty attending.

We received NIH funding to run the course in 2012 and 2013, and for 2012 we received over 169 applications, from 19 countries and 34 states. We again admitted 24 students (see Figure 1). These 24 attendees consisted of 8 graduate students, 9 postdocs, 4 tenure-line faculty (including two full professors), and 3 industry researchers.

Online materials The course materials that we created and now maintain for this course are freely available and broadly useful for biologists.

The course materials have been hosted at http://ged.msu.edu/angus/ for all three years of the course. In 2011 we instrumented the site with Google Analytics to study the viewer statistics. Google Analytics reports that within the last year (August to August), the materials received 43,000 unique visitors in 80,000 visits, with approximately 204,000 page views.

In 2012, we added disqus commenting to the course Web site. This allows students and remote viewers to annotate the material and ask questions. Since July we have received several dozen comments.

The biostars.org Web site, partially supported by the NIH through our previous grant, has over 52,000 posts from over 5,000 users. It is one of the two central bioinformatics question and answer sites.

All of these materials are widely used. Moreover, they have been extremely valuable for other courses such as the STAMPS microbial ecology course at MBL, where Dr. Brown has extended them to provide tutorials in metagenomic assembly and shotgun analysis. They have also been reused in an NIH-funded Human Microbiome Project workshop on cloud computing.

[2] http://www.wellcome.ac.uk/Education-resources/Courses-and-conferences/Advanced-Courses/Courses/WTX056918.htm

[3] http://meetings.cshl.edu/courses/c-seqtech10.shtml

[4] http://meetings.cshl.edu/courses/c-ecg10.shtml

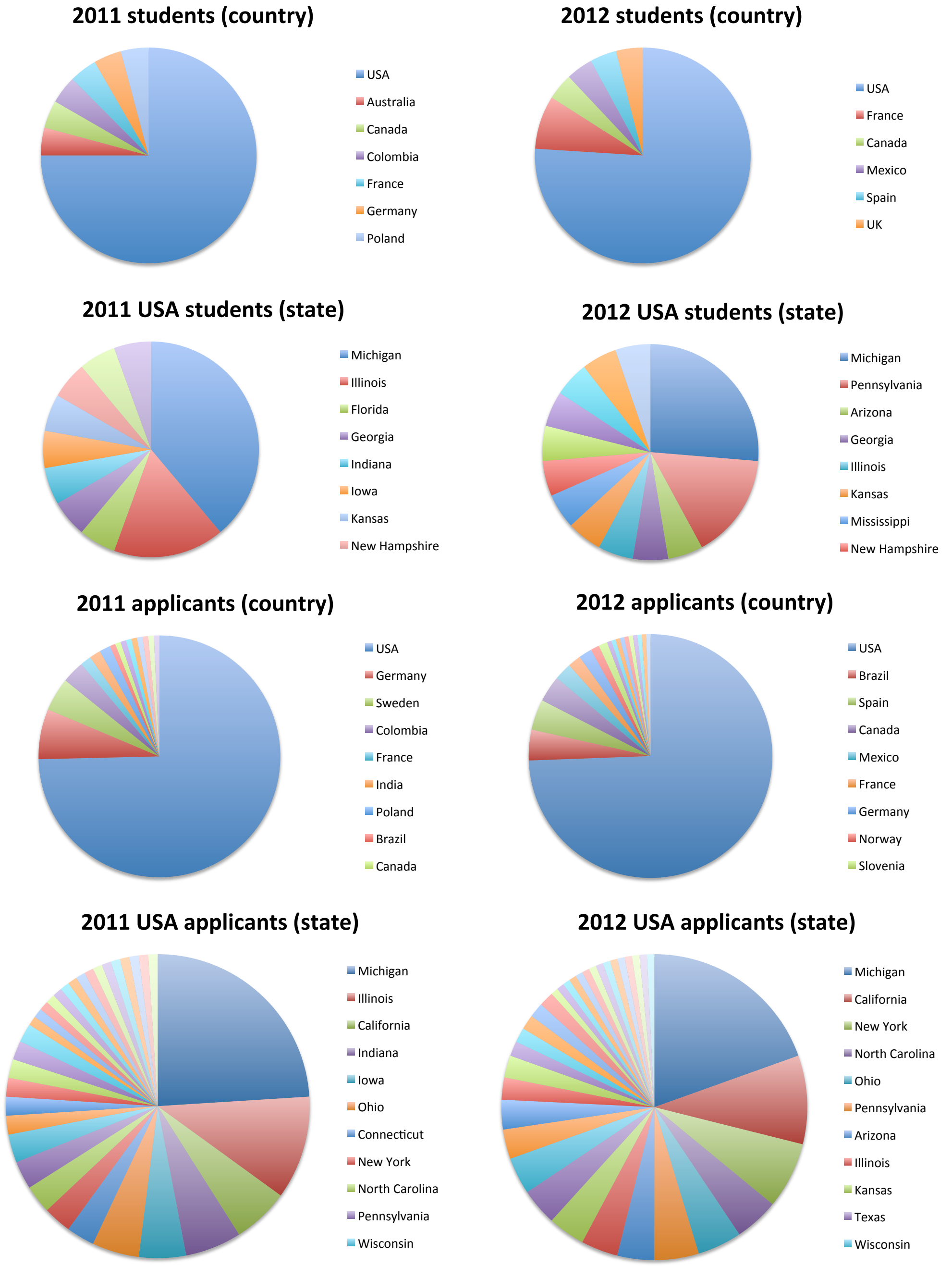

Figure 1: **Applicant and student statistics by country/state of origin.** In 2011 and 2012 we had over 302 applicants, combined; we admitted 24 each year. The large majority of admitted students came from the United States (top row), from across 10 states (2011) and 12 states (2012). These 48 students were selected from a far larger applicant pool of students from 16 countries and 23 states in 2011, and 19 countries and 34 states in 2012. This amazing breadth of student applications illustrates the strong continued need for this training program.

Evaluation Summary The executive summary of the evaluation performed by StemEd LLC for the 2012 summer course follows; the entire evaluation report is attached as an appendix.

"A pre-workshop evaluation of the NGS Summer Workshop 2012 - Analyzing Next-Generation Sequencing Data was conducted on June 4, with a post-workshop evaluation occurring June 15, 2012. Observations were also conducted at the start, middle, and end of the workshop. In all, 25 participants completed the pre-survey and 23 completed both the pre- and post-surveys.

In summary, we found that:

1. Scores on the Perception of Computational Ability scale were calculated for both the pre- and post-workshop surveys. Results from the Wilcoxon Signed Ranks Test indicate that pre- and post-workshop results are statistically different (Z = -4.116, $p < 0.001$), with higher post-workshop scores. **This indicates that participants perceived greater computational ability after engagement in the workshop.**

2. Scores on the Computational Understanding - Sequencing Data scale were calculated for both the pre- and post-workshop surveys. Results from the Wilcoxon Signed Ranks Test indicate that pre- and post-workshop results are statistically different (Z = -4.111, $p < 0.001$), with higher post-workshop scores. **This indicates that participants perceived greater understanding after engagement in the workshop.**

3. Scores on the Python Coding Ability scale were calculated for both the pre- and post-workshop surveys. Results from the Wilcoxon Signed Ranks Test indicate that pre- and post-workshop results are statistically different (Z = -4.374, $p < 0.001$), with higher post-workshop scores. **This indicates that participants perceived greater coding ability after engagement in the workshop.**

4. **Participants were generally very satisfied with the workshop.** On average, participants rated the workshop components as Good-Very Good.

5. **Participants generally felt the workshop met their needs and would overwhelmingly recommend it to others.**

6. **Participants were generally positive about the workshop in their open-ended comments.** Suggestions for improvement include: more time on RNA sequencing and differential expression/data, less focus on why tools are not good, more focus on basics of programming, scripts and/or UNIX early on, and more details about daily activities."

In response to these concerns we will be adding more screencasts about programming, UNIX, and scripting, and we will provide a detailed daily schedule in advance of each year's course. The RNAseq comments are being addressed by splitting mRNAseq analysis into reference-based transcriptome analysis (with e.g. Cufflinks) and de novo transcriptome analysis (with e.g. Trinity) and then presenting quantification as a third topic at the beginning of the second week.

With reference to the comment about "less focus on why tools are not good", each year there are several students who simply want us to teach a tool without providing associated caveats. We disagree! For each tool we present we provide a detailed description of the assumptions made by the underlying algorithm and provide examples of where it can go awry. We believe this to be an essential component of the underlying science.

We also received feedback that more and better information on statistical approaches and assumptions were necessary; we therefore plan to devote a whole day of lectures and exercises to discuss appropriate statistical inference and experimental design.

Outcomes We now have over 70 course alumni, including graduate students, postdocs, faculty, and core directors. Many of them have gone on to expand the computational component of their work; while we are still working on longitudinal surveys and interviews from our first year of NIH funding, we have received many letters of support (see attachments) detailing the impact on their research and careers. Many of our alumni are now teaching similar workshops and courses at their own institution.

3.4 Proposed course

We propose to continue running this extremely successful course in 2014, 2015, and 2016. While we plan to continue focusing on the same specific aims as before (see below), we will update our technology focus each year based on the advice of our Advisory Committee and our Visiting Faculty. It is increasingly challenging to predict even the near future of sequencing technology, but the next two years will undoubtedly include data from Illumina MiSeq, PacBio SMRT, and Ion Torrent, as well as perhaps Oxford Nanopore data. In addition to updating existing course material and developing new course material as needed, we will also extend our educational offerings into scripting and data visualization, and continue to run yearly evaluations to focus on areas that need improvement. Finally, we plan to extend the site with screencast videos focusing on specific tasks, e.g. mRNAseq assembly and quantification, as we develop new tutorials.

Our course addresses the following specific aims:

Specific Aim 1: Teach practical remote use of UNIX systems on the Amazon cloud computing platform We introduce the UNIX command line and Amazon EC2 machines in the first two days of the course. Specific topics covered include logging in, program installation via package managers, download of files and editing of text files, and executing long-running processes.

Our primary example on the second day will be command-line BLAST, which all biologists are familiar with and immediately appreciate as a critically useful tool.

Specific Aim 2: Introduce short-read mapping and assembly techniques The third and fourth days are spent introducing short read mapping and assembly programs, through bowtie, BWA, Velvet, and SGA [Langmead et al., 2009] [Langmead et al., 2009] [Simpson et al., 2009] [Zerbino and Birney, 2008]. These are the building blocks upon which most later analyses rest.

Specific Aim 3: Use sample data sets to tackle biological problems After a Sunday break, we spend the next three days introducing mRNAseq analysis with TopHat and Cufflinks, resequencing analysis with GATK, and ChIP-seq analysis with QuEST [Langmead et al., 2010, Barrick et al., 2009, Valouev et al., 2008]. A key part of this section is discussing appropriate statistical issues in experiment design, both by examining how specific packages (e.g. DEGseq and Myrna) handle data interdependencies, and by constructing hypothetical situations.

Specific Aim 4: Work with individual students to explore and develop research-oriented computational approaches Our experience has been that as soon as they learn to run the mapping and assembly tools, students immediately become very interested in applying mapping and assembly to their own data sets – even in advance of the specific topics to follow. The loose structure of the course, with plenty of time between tutorials and late evening sessions with TAs and faculty present, is designed to allow students sufficient time to explore their own data sets, or data sets that we provide up front.

We combine these specific aims with a specific educational strategy designed to give students a very in-depth and hands-on experience.

3.5 Educational approach

Our goals are to bridge the gap between the students' area of expertise in biological science and the computational skills required to apply that expertise to large data sets. Specifically, we couple strongly guided tutorial-based education in computational science with an inquiry-based discussion of biomedical research.

Computational education strategy Inquiry-based strategies appear not to be effective ways to introduce students to new material, largely because students struggle to recall and apply recently learned facts from short-term memory [Kirschner et al., 2006]. Based on this research, we have developed a strongly guided approach. Each day in the course consists of a steady progression through several stages:

- First, a **lecture providing an overview** of the area. For example, for mapping, we discuss the basic computational challenges and algorithmic solutions, as well as their assumptions and consequent drawbacks in the face of real data.

- Next, **a guided tutorial**, consisting of a detailed run-through of copy/paste instructions for performing a particular task on an Amazon EC2 instance.

- Next, **a period of exploration**, during which students can repeat the tutorial on their own, but in the presence of TAs. Students are also encouraged to "play" with parameters to see the effect. For example, in mapping with bowtie, we encourage students to explore the difference between the "-n" and "-v" alignment modes, and to examine the resulting disparity in mappings as well as the sensitivity of mapping to various additional parameters. In later days this is tied into application to mRNAseq, genome resequencing, ChIP-seq. etc.

- Fourth, we individually discuss **application of the tools to their own data or problems**, with the relevant faculty and TAs. Students who have data are encouraged to begin analysis, with the full help of TAs and faculty; students without their own data are provided with data sets applicable to their "home" research problems.

- Fifth and finally, **links to additional information**. We provide Web links to additional software and publications that bear on each question, as well as connecting to relevant questions and answers on the SeqAnswers[5] and BioStar[6] Web sites. We also provide relevant links

[5] SeqAnswers.com

[6] http://biostars.org

to the Software Carpentry site[7] for beginning and continuing education in computational science. All of these materials are available before, during, and after the course (see Dissemination).

We believe that this strongly guided approach helps maximize the utility of the course, the utility of the highly available TAs and faculty, and the learning of the students. Our evaluation and assessment efforts confirm this.

A typical day Now that we have run the course three times, we have settled on the following daily schedule for the 10 days of the course (Mon-Sat week 1, Mon-Th week 2).

1. Breakfast until 9am.
2. Opening lecture from 9am-10:30am, with questions encouraged.
3. Initial tutorial (e.g. "running mapping and examining results") until noon.
4. Lunch from noon until 1:30pm.
5. Second tutorial from 1:30pm-3pm (e.g. "visualizing mapping and adjusting parameters").
6. Free work/play time (3pm-5pm); TA present in room.
7. Dinner from 5:30-7pm
8. Tutorial, lecture, or panel from 7pm-8pm
9. Free work/play time (8pm-midnight); TA present in room until 11pm.

The specific daily topics have varied during the second week, but the first week is typically structured to introduce all the basic concepts necessary for starting independent work. At the beginning of the second week we have started rearranging the tables so that students are grouped by topic (e.g. mRNAseq, ChIP-seq, etc.) and the course shifts to morning lectures with optional tutorials, and primarily focuses on letting students work with their own data. (This shift was requested by students in 2011 and carried out for 2012.)

1. Amazon EC2 configuration and logging in
2. UNIX command line, BLAST, and scripting.
3. Mapping, visualization, and more scripting.
4. Assembly, mRNAseq assembly, and more scripting.
5. Statistics, plotting, and data set evaluation.
6. Quality control and analysis of data

[7] http://software-carpentry.org/

7. mRNAseq quantification and differential expression
8. ChIP-Seq analysis
9. Resequencing analysis
10. RAD-tag analysis

One concern raised by reviewers in our first application was that long compute times might interfere with the daily schedule and yield too much "downtime" for students. In practice, we have avoided this by (1) choosing small data sets as exemplars, (2) providing intermediate results as appropriate, and (3) showing students how to set up long jobs overnight. As a consequence "downtime" has not been a practical concern; we find that computational exhaustion on the parts of students is a more pressing concern!

Bridging to biology We couple this strongly guided approach with in-depth discussions of how to evaluate computational strategies and software, intersect those evaluations with your research needs at the moment, and think about the research from a computational perspective. In our personal experience, researchers trained in biology are extremely quick to key in on this perspective once they have run programs and seen the effect of parameter variation on their results.

Challenges A principle challenge in teaching non-computational scientists to effectively use computers is that many of them are apprehensive about the technical knowledge and requirements involved. Our experience in several courses has been that a gentle initial introduction in the first day or two, coupled with a few very relevant examples (e.g. BLAST), and the generally strong motivation of biological scientists to learn in this area, yields great rewards in terms of interest, engagement, and ultimately learning. This has been anecdotally confirmed by students and is supported by our first year's assessment. If we can demonstrate this through multi-year assessment, this will have a significant impact in our ability to cross-train computational scientists in the future.

Teaching assistants We have had 4-5 teaching assistants each year, and have had no problem recruiting qualified candidates. The TAs are selected from the labs and courses of the instructors; for 2013, for example, we expect to have two TAs from Dr. Brown's lab, one from Dr. Albert's lab, and one from Dr. Corbin Jones' sequencing core. Their travel and lodging is paid for through this grant.

In our new proposal, we request one summer RA to update the course material before and after the course: before the course, the RA will update software version numbers and screenshots; after the course, the RA will update the material in response to questions and critiques that arise during the course.

Advisory committee and curriculum change As technology progresses, new biological applications open up; as biomedical research advances, existing technology can be applied in new ways. Keeping our course up to date with both the fast changing fields of genomics and large-scale sequencing is extremely important, and not always straightforward; for example, the new Pacific Biosciences technology (with fast, inexpensive long reads) has begun to impact de novo assembly, but its high error rate and low sampling has limited its utility in resequencing and mRNAseq quantification.

To ensure that we keep the course abreast of the latest advances in biotechnology and genomics, we have recruited an advisorial board of internationally recognized leading-edge genomic scientists. The following scientists have served as advisors to the course over the past two years. Their role is to ensure that we are covering the latest sequencing technologies and approaches, and to help promote the course. In particular, they are consulted on each year's syllabus, provide interesting data sets, help advertise the course, and provide TAs and students.

- **Human genetics and genomics:** Kevin White, James and Karen Frank Family Professor at University of Chicago.
- **Animal genetics and genomics:** Paul W. Sternberg, Thomas Hunt Morgan Professor of Biology at the California Institute of Technology; member of the National Academy of Sciences.
- **Microbial evolution and resequencing:** Richard E. Lenski, Hannah Distinguished Professor at Michigan State University; member of the National Academy of Sciences.
- **Plant genetics and genomics:** Robin Buell, Associate Professor at Michigan State University.
- **Microbial population sequencing and metagenomics:** James M. Tiedje, Professor, Michigan State University; member of the National Academy of Sciences.
- **Bioinformatics and genomics:** Lincoln Stein, Platform Leader, Informatics and Bio-Computing, Ontario Institute for Cancer Research; Professor, Cold Spring Harbor Laboratory.

3.6 Program Director/Principle Investigator

Dr. Brown has almost 20 years of experience in computational science, including digital life and climate modeling. His undergraduate degree is in Math and he has a substantial amount of experience in practical software engineering, including several open source bioinformatics toolkits. Dr. Brown received his PhD in Developmental Biology from Caltech in Dr. Eric Davidson's lab, where he was trained in genomics and regulatory genomics. After brief post-doctoral work with Dr. Marianne Bronner-Fraser at Caltech, he took a faculty position split between two departments, Computer Science and Engineering and Microbiology and Molecular Genetics at Michigan State University. Since then he has continued to work in genomics, metagenomics, next-gen sequencing data analysis, and software development. He has also devoted considerable effort to interdisciplinary training, not only for this course but also for the BEACON Center for the Study of Evolution in Action, where he has developed and taught a course entitled "Computational Science for Evolutionary Biologists".

Dr. Brown is an active researcher in many aspects of genomics and next-generation sequence analysis, and has published in bacterial resequencing analysis and *de novo* metagenome assembly [Jerome et al., 2011, Pell et al., 2012]. He has also developed approaches for efficient streaming data reduction in biological data sets [Brown et al., 2012].

Dr. Brown's role in the course is to advertise the course, manage admissions, organize the schedule, recruit visiting faculty, and coordinate materials development. He is in attendance at the course during the entire two weeks.

3.7 Program Faculty/Staff

The program faculty are responsible for coordinating lectures and tutorials within their areas of expertise.

Ian Dworkin holds a Ph.D. in Evolutionary Genetics from the University of Toronto, and has worked on Quantitative and statistical genetics, Evolutionary biology, genomics and developmental biology. He is currently an Associate Professor in Zoology, and in the Program in Ecology, Evolutionary Biology and Behavior, at Michigan State University, where his lab works on evolutionary genomics of development, morphology and behavior, in addition to the development of new statistical tools. Dr. Dworkin runs the resequencing and statistics components of the course.

Dr. Istvan Albert holds a Ph.D. in Physics from the University of Notre Dame and has worked in statistical physics, data mining, bioinformatics and genomics. He is currently an Associate Professor in Biochemistry and Molecular Biology at the Pennsylvania State University where his group works on developing novel data analysis and visualization methods in the fields of bioinformatics and medical informatics. He is also the maintainer of BioStar, a question and answer site for bioinformatics research: http://biostars.org/. Dr. Albert runs the ChIP-seq and scripting components of the course.

3.8 Visiting Lecturers

Each year we invite several visiting lecturers to focus on topics not represented amongst the core faculty. The three visiting lecturers from 2012 were Corbin Jones, Erich Schwarz, and Julian Catchen.

Dr. Corbin Jones is an Associate Professor at UNC, where he runs the sequencing core. In 2012 he gave lectures on sequencing technologies and and mRNAseq.

Dr. Erich Schwarz is a Research Scientist at Cornell with an adjunct position at Caltech in Wormbase. In 2012 he lectured on nematode genome assembly and genome annotation.

Dr. Julian Catchen is a postdoctoral fellow in computational biology at the University of Oregon, where he works on population genetics analysis of model and non-model organisms. He ran two workshops on the Stacks software in 2012.

3.9 Responsible Conduct of Research

We find that many biologists have not been exposed to sound computational practice and are unfamiliar with data archiving policies and basic approaches in reproducible research. We therefore include many lecture components on the question of replication and reproducibility of computational analyses, including a section on source code control systems and analysis automation.

3.10 Program Participants

The intended participants for this course are advanced graduate students, postdocs, and junior faculty trained and working in biomedical research areas. No computational background of any kind is required.

Course location and cost We provide space for 24 students and up to 8 TAs and faculty at the Kellogg Biological Station (KBS) in early May. KBS is located approximately 90 minutes from the Lansing and Grand Rapids International Airports, as well as 30 minutes from the Kalamazoo

Regional Airport. The total cost for room and board is approximately $500 per student, which makes this one of the least expensive courses available.

3.11 Student selection criteria

We will select students from our applicant pool based on career stage, research focus, recommendation letters, and diversity considerations.

Specifically,

- Early-stage research faculty, postdoctoral fellows, and advanced graduate students will receive priority over early career graduate students.
- Students self-identifying as members of under-represented groups (e.g. early-career women faculty and post-docs) will receive priority.
- Students with specific plans to use next-generation sequencing, in their research, or who already have data to analyze, will receive priority over other students.

We will also provide "free ride" funding for several attendees, based on maximizing the ability of members of under-represented groups to attend. (See Budget.) We will also advertise the course at SACNAS and more regional conferences for underrepresented groups in science.

3.12 Diversity Recruitment

We will recruit members of under-represented groups into the course by contacting diversity program officers directly at multiple institutions, and by advertising the course at SACNAS and other conferences for underrepresented groups in science. We will also partner with the Sloan Engineering Program at Michigan State University, a program designed to recruit, mentor, and graduate underrepresented minorities with doctoral degrees, to identify pools of potential advanced graduate students from underrepresented groups.

3.13 Dissemination

Properly constructed, the material used for "boot-camp" courses can be immensely valuable, both for students during the course and also for later perusal. In particular, we find that students during the course are willing to explore the relevant "dark corners" of the material that may not be presented in depth during the course, if they are particularly interested in the subject. We also find that students pay more attention during the course if they know that the materials are available openly after the course, and if the instructors indicate a willingness to answer questions via e-mail. We also think that there is a great opportunity for open course notes to be "Google bait", i.e. if course notes can be found by Web search and linked to by others, then they act as a nexus of information and serve as a force multiplier for the entire field. We also expect that other courses will be able to make use of our material as an adjunct source of information.

Most similar courses do not make their course material openly available or reusable.

Our team has demonstrated an outstanding devotion and a continuing commitment to online education and dissemination of research and teaching materials. Specifically,

- The entire course contents for the 2010, 2011, and 2012 NGS courses are freely and publicly available, without registration, for both perusal and modification/redistribution under a Creative Commons - Share Alike 2.0 license (CC-BY-SA). This maximizes availability and utility of the lectures and course materials.

  This material is useful far beyond the course: for example, our Web site has more detailed documentation on installing and running many assemblers than exists anywhere else. We have received a number of e-mails about this material, and several tens of thousands of unique visitors have viewed material since its initial posting.

- The popular BioStar bioinformatics Q&A site [8] is run by Dr. Albert, a course instructor, and has been supported by this course since 2012.

- Michigan State University, under the direction of Dr. Brown (the PI), has contributed to funding of the Software Carpentry materials, specifically to aid biological scientists in confronting the same challenges addressed in this course. Dr. Gregory V. Wilson, the creator and ongoing editor of the Software Carpentry materials, was course faculty on our NGS course in 2010.

- Dr. Brown practices "open science", blogging regularly about (e.g.) assembly work on his personal blog[9], as well as using Twitter and Facebook; these posts are connected into the course material when relevant.

- Both Dr. Albert and Dr. Brown are regular contributors to open source frameworks within the bioinformatics community.

We will continue to provide up-to-date and leading-edge tutorials and techniques through the course Web site, through BioStar, and through our personal blogs. We will also continue to encourage students to remain in contact with us personally after the course. Finally, we will provide materials in a format suitable for efforts like the Hacker Within[10], a "student-run, skill-sharing interest group for scientific software development" at the University of Wisconsin.

3.14 Evaluation plan

At the top level, our learning objectives focus on student ability to ascend to at least the middle upper levels of the Bloom hierarchy of learning objectives applied to the disciplinary domain of sequence analysis and computational science [Bloom et al., 1984]. Specifically, students will demonstrate their abilities in five areas:

- Ability to manipulate large data sets;

- Ability to apply approaches they do not understand in detail by utilizing black box computational approaches;

- Ability to place controls on these black box computational approaches;

[8] see: http://biostar.stackexchange.com

[9] http://ivory.idyll.org/blog/

[10] http://hackerwithin.org/

- Ability to integrate multiple sources of data to make biological inferences based on computational approaches;
- Ability to make a principled and defensible choice of a statistical approach given dependencies in a target data set or sets;
- Ability to accurately describe the limitations and biases resulting from applying a given computational approach to a particular data set;

We have engaged StemEd LLC to develop assessment materials, apply them independently to the students, and provide evaluation reports each year. Three primary modes of assessment were applied in 2012. First, a pre- and post- written questionnaire was given to the students to evaluate their baseline knowledge and gains in knowledge during the course. Second, several lectures were observed by personnel from StemEd LLC. And third, the post-mortem session was attended by evaluation personnel. This culminated in an assessment report that concluded that some relatively minor refocusing was needed in the background of an overwhelmingly successful course (see Appendix for full evaluation).

Future assessment will continue as above, with written instruments and observations of class sessions. However, in an effort to better document the impact of our course and focus it on students' continuing needs, StemEd LLC will also conduct post-course interviews of both admitted students and rejected students. The primary goals of this component of the assessment will be to determine how accepted students are making use of their knowledge, and whether and how rejected students are learning material on their own. We also hope to develop a better understanding of how advanced professionals learn computational practice.

We are already coordinating this broader evaluation with the BEACON NSF Science and Technology Center at MSU, which is keenly interested in how to teach modeling and data analysis skills to biologists. As part of this broader effort Dr. Brown has received a $200k supplement to work across NSF BIO Centers in this area, which we will be able to leverage to improve this course.

4 Resources and Environment

The Kellogg Biological Station is a full research and education campus external to the main Michigan State campus. KBS has a 100 Mbit Internet connection, WiFi throughout the campus, and a dedicated IT staff on-site.

The classroom allocated to the proposed course is a 2000 sq ft classroom with a projector, conference table, individual student tables, a lounge area, and a small kitchen area. It is within walking distance of dorms and cafeteria.

KBS campus contains room for over 5 dozen faculty and students, including 36 units of dorm housing that have been allocated to our two week course. Each dorm room hosts two people, has a shared bathroom, and WiFi throughout.

KBS is located 90 minutes west of Michigan State University main campus and easily accessible from multiple airports, including Lansing, Detroit, Grand Rapids, Flint, and Kalamazoo Regional.

References

[Barrick et al., 2009] Barrick, J., Yu, D., Yoon, S., Jeong, H., Oh, T., Schneider, D., Lenski, R. and Kim, J. (2009). Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature *461*, 1243–7.

[Bloom et al., 1984] Bloom, B., Krathwohl, D. and Masia, B. (1984). Taxonomy of educational objectives: the classification of educational goals. New York: Longman.

[Brown et al., 2012] Brown, C., Howe, A., Zhang, Q., Pyrkosz, A. and Brom, T. (2012). A reference-free algorithm for computational normalization of shotgun sequencing data. In review at PLoS One, Aug 2012; Preprint at http://arxiv.org/abs/1203.4802.

[Gnerre et al., 2011] Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F., Burton, J., Walker, B., Sharpe, T., Hall, G., Shea, T., Sykes, S., Berlin, A., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E. and Jaffe, D. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A *108*, 1513–8.

[Jerome et al., 2011] Jerome, J., Bell, J., Plovanich-Jones, A., Barrick, J., Brown, C. and Mansfield, L. (2011). Standing genetic variation in contingency loci drives the rapid adaptation of Campylobacter jejuni to a novel host. PLoS One *6*, e16399.

[Kirschner et al., 2006] Kirschner, P., Sweller, J. and Clark, R. (2006). Educational Psychologist *41*, 75–86.

[Langmead et al., 2010] Langmead, B., Hansen, K. and Leek, J. (2010). Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol *11*, R83.

[Langmead et al., 2009] Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

[Pell et al., 2012] Pell, J., Hintze, A., Canino-Koning, R., Howe, A., Tiedje, J. and Brown, C. (2012). Scaling metagenome sequence assembly with probabilistic de bruijn graphs. Accepted at PNAS, July 2012; Preprint at http://arxiv.org/abs/1112.4193.

[Pérez and Granger, 2007] Pérez, F. and Granger, B. E. (2007). IPython: a System for Interactive Scientific Computing. Comput. Sci. Eng. *9*, 21–29.

[Simpson and Durbin, 2012] Simpson, J. and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. Genome Res *22*, 549–56.

[Simpson et al., 2009] Simpson, J., Wong, K., Jackman, S., Schein, J., Jones, S. and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. Genome Res *19*, 1117–23.

[Valouev et al., 2008] Valouev, A., Johnson, D., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods *5*, 829–34.

[Zerbino and Birney, 2008] Zerbino, D. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res *18*, 821–9.