A short course in analyzing next-generation sequencing data

September 24, 2010

Keywords: bioinformatics, next generation sequencing, genomics.

1 Project Summary

Modern biomedical research is increasingly making use of genome-scale from next-generation sequencing platforms, including Roche 454, Illumina GA2, and ABI SOLiD. 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 generate large data sets, biomedical researchers 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 computational resources necessary to run these analyses.

Our specific aims are to help bridge this gap in a short, two-week 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. All specific aims will be accompanied by in-depth hands-on practical training in the relevant techniques. Our experience is that this practical training leads to a substantial improvement in the basic computational sophistication of participants.

This short course will 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. Our end goal is increase the efficiency and sophistication with which biomedical researchers make use of novel sequencing technologies.

2 Specific Aims

The vast increase in sequencing capacity available with the introduction of the Roche 454, Illumina GA2, ABI SOLiD, and other platforms, 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. Most biologists lack training in computational data analysis, and combined with the vast increase in data-gathering abilities, 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), advanced level course to disseminate computational sequence analysis skills specifically designed for analysis of next-generation sequencing data.** Our objective for this course is to provide biomedical 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, to enable the researchers to apply greater resources in the future as needed. **We have already run this course once, with positive results.**

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 Roche, Illumina, ABI, 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 (e.g. Myrna, QuEST, breseq), 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 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.

3 Research Education Program Plan

3.1 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. More recently, 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 the first version of this proposed course but also for the BEACON Center for the Study of Evolution in Action, where he is teaching a course entitled "Computational Science for Evolutionary Biologists". Dr. Brown is an active researcher in many aspects of genomics and next-generation sequence analysis.

3.2 Program Faculty/Staff

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 Assistant 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. 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://biostar.stackexchange.com/

3.3 Proposed Research Education Program

The vast increase in sequencing capacity available with the introduction of the Roche 454, Illumina GA2, ABI SOLiD, and other platforms, 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 of gut bacteria. 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. While traditional biomedical education paths have lacked training in computational data analysis, this omission is now exacerbated by the speed with which both computational science is moving and the increase in data gathering capabilities available to biologists. Moreover, most researchers and many institutions do not possess sufficient computational infrastructure to perform large-scale sequence analysis in a timely manner. **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 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), advanced level course to disseminate computational sequence analysis skills specifically designed for analysis of next-generation sequencing data.** Our objective for this course is to provide biomedical 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, to enable the researchers to apply greater resources in the future as needed. An additional key component of the course is the opportunity for researchers to bring personal data sets with them for individualized instruction. 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 once, with positive results.**

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 Roche, Illumina, ABI, 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 (e.g. Myrna, QuEST, breseq), 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 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 experience with this overall approach suggests 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 of substantial compute resources without a large up-front infrastructure investment; and, because we use normal Linux systems on Amazon EC2, the skills transfer to the most commonly available compute systems, e.g. at their home institution. Finally, by making our notes openly available (see http://ged.msu.edu/angus/), and linking them to the Software Carpentry computational science tutorials, we provide long-term resources for self-training and growth in computational science.

3.4 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 is rapidly changing, with the plethora of existing production-stage technologies such as Roche 454, Helicos, Illumina GA2, and ABI SOLiD, moving to maturity, even while new technologies (such as Illumina HiSeq and Pacific Biosciences) are emerging. 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.

- 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. 454 vs Illumina vs SOLiD). This has reached the point where sequencing centers suggest using multiple technologies for each resequencing project! [Harismendy et al., 2009, Gomez-Alvarez et al., 2009, Barrick and Lenski, 2009].

- 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 transcriptomic. (The answer depends critically on the quality of the reference genome or transcriptome; generally the answer is yes, but many people balk at the additional cost because they don't understand why it is important.)

- 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 quick speed of technological change in this area makes rote learning an ineffective approach.

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 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.

The two most focused and integrated courses available for *computational training* are a course (the precursor to this proposed course) that we ran at the Kellogg Biological Station (KBS) at Michigan State University (MSU) in May 2010 [1], and a course at UC Davis that is currently running [2]. 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.

Another course, offered at Leiden in the Netherlands, is considerably shorter than either the MSU or UC Davis course and offers a correspondingly more focused look at specific technologies[3]. The Broad Institute has also posted materials from a similar one-day workshop in early 2010 that primarily addressed resequencing analysis[4].

[1] http://bioinformatics.msu.edu/ngs-summer-course-2010

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

[3] http://www.expertiseteam-separationsciences.nl/avans-training-2542.htm

[4] http://www.broadinstitute.org/science/programs/medical-and-population-genetics/slides-next-generation-

Cornell runs a 6 week in-term workshop on practical bioinformatics training for people in the area[5].

Other courses, such as those offered by the Wellcome Trust [6] and Cold Spring Harbor Labs[7], 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[8].

3.5 Preliminary Results

Driven by the need to train students and postdocs in analyzing data, Michigan State University (the PI's host institution) funded a first version of this course in May 2010. The course was held at the Kellogg Biological Station, about 90 minutes west of the Lansing airport, and attended by 20 students from biology graduate and postdoc programs (including MSU, U. Chicago, UC Irvine, and Yale), along with three additional industry scientists.

For this first course we implemented a hands-on educational strategy (described below in more detail), in which we gave lectures, provided students with detailed recipes in using computational tools, and discussed strategies on how to use those tools to accomplish specific scientific tasks. We then encouraged them to explore data sets on their own and with the help of TAs. (See Research Strategy and Educational Strategy sections, below, for details.)

Feedback from the course, obtained through anonymous surveys and an open "post-mortem", was extremely positive. A number of students continue to use the online materials and several students have already incorporated additional sequence analysis approaches into their research.

From this first course we produced a Web site, http://ged.msu.edu/angus/, containing all of the lectures and computational recipes used in the course. Since the course, several journal clubs and ad hoc working groups at other institutions have formed to explore these materials, indicating both the utility of our course and the timeliness of the subject.

3.6 Proposed course

We propose to run an evolved version of our 2010 course for two weeks during each summer (between May and July) from 2011 to 2013. Our proposed course will be held at the Kellogg Biological Station and follow much the same outline as the 2010 course, with updates for new technology and some modifications suggested by the students from the 2010 course.

Note: the majority of the tutorials below are already available in draft form at http://ged.msu.edu/angus, under tutorials/.

Specific Aim 1: Teach practical remote use of UNIX systems on the Amazon cloud computing platform The first two days of the course will be spent introducing the UNIX command line as a concept, and learning setting up Amazon EC2 machines on which to perform analysis. Specific

sequencing-workshop

[5] http://cbsu.tc.cornell.edu/nextgenworkshop2010.aspx

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

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

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

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 will immediately appreciate as a critically useful tool.

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

Specific Aim 3: Use sample data sets to tackle biological problems After a Sunday break, we will spend the next three days introducing mRNAseq analysis with Myrna, resequencing analysis with breseq, and ChIP-seq analysis with QuEST [Langmead et al., 2010, Barrick et al., 2009, Valouev et al., 2008]. A key part of this section will be 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.

mRNAseq In the first course, we used a home-grown mRNAseq analysis pipeline to map, count, normalize, and analyze lamprey mRNAseq data from multiple tissues. Now that Myrna is available, we propose to refactor the course to use it instead [Langmead et al., 2010]. We will also address assembly of mRNAseq data with Velvet and splicing and isoform analysis with Cufflinks [Trapnell et al., 2010] and TopHat [Trapnell et al., 2009].

breseq The breseq pipeline for resequencing analysis in *E. coli* is a pipeline developed at Michigan State University to analyze resequencing data from 454 and Illumina technologies [Barrick et al., 2009, Barrick and Lenski, 2009]. breseq offers a wide variety of options, including the ability to generate your own mutations and simulated resequencing data. However, it is also complex to install (requiring R, Perl, BioPerl, SSAHA, and a number of other packages), run, and configure. It therefore offers an excellent introduction to the complexity of modern pipelines while also providing a full environment for learning about resequencing.

QuEST and MEME The QuEST tool is a high-quality tool for analyzing ChIP-seq data [Valouev et al., 2008]. We will introduce its use on locally available *M. xanthus* and *G. gallus* data sets. An important part of ChIP-seq is downstream motif analysis, and we will demonstrate the use of the MEME and MAST packages for this purpose [Bailey et al., 2006].

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 tools, students become very interested in applying their knowledge 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.

One important component of this section is to provide students with a variety of data sets, so that they can compare e.g. ABI SOLiD data and processing tools with Illumina and 454 sequence, or paired-end vs non-paired end mapping and sequence assembly.

We have relationships with a number of sequencing companies, including Roche, Illumina, and Pacific Biosciences, to supply us with “latest generation” data in March of 2011, just prior to the next course. This will help give our students a chance to work with recent data.

3.7 Educational strategy

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 propose to 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 will 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 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[9] and BioStar[10] Web sites. We will also provide relevant links to the Software Carpentry site[11] 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. One goal of our evaluation and assessment efforts is to verify that this approach works (see Challenges, and Evaluation, below).

[9] SeqAnswers.com

[10] http://biostar.stackexchange.com

[11] http://software-carpentry.org/blog/

Bridging to biology We will 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. If we can demonstrate this through assessment, this will have a significant impact in our ability to cross-train computational scientists in the future.

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, we expect the new Pacific Biosciences technology (with fast, inexpensive long reads) to quickly revolutionize resequencing approaches, but to have little immediate impact on e.g. transcriptome 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 agreed to serve as advisors to the course. Their role will be to ensure that we are covering the latest sequencing technologies and approaches, and to help promote the course. In particular, they will be consulted on each year's syllabus, will help provide interesting data sets, will help advertise the course, and may provide TAs.

- **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.

3.8 Responsible Conduct of Research

While not required for short courses, we will focus considerable attention on the question of replication and reproducibility of computational analyses, including a section on source code control systems and automation.

3.9 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 will be assumed.

Course location and cost We have reserved space for 24 students and up to 8 TAs and faculty at the Kellogg Biological Station (KBS) in summer 2011. 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 for two weeks is projected to be under $700 per student.

3.10 Student selection criteria

We will select students from our applicant pool based primarily 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.

3.11 Diversity Recruitment

We will make a special effort to 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.

3.12 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 hope that other courses will 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 NGS course (see Preliminary Results, this proposal) 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 access to 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 both the ABySS assembler and the breseq bacterial resequencing pipeline than exists anywhere else. We have received a number of e-mails about this material, and several thousand unique visitors have viewed material since its initial posting.

- The popular BioStar bioinformatics Q&A site [12] is run by Dr. Albert, a course instructor.

- 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[13], 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 world.

We propose 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 continue to encourage students to contact us personally after the course. We will also encourage students to correspond on a class-derived (but open) mailing list for those who want to discuss their research further. Finally, we will provide materials in a format suitable for efforts like the Hacker Within[14], a "student-run, skill-sharing interest group for scientific software development" at the University of Wisconsin.

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

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

[14] http://hackerwithin.org/

As early evidence of our success at disseminating the original class notes, several students at Rockefeller have already self-organized to run through the notes in a group in summer/fall 2010.

3.13 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;
- 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;

In order to assess students' progress in meeting these learning objectives, we have engaged evaluation services from the Center for Engineering Education Research (CEER) at Michigan State University[15]. The plan of assessment involves two main activities.

First, CEER personnel will conduct student interviews at the end of the course to determine the degree to which students feel they have mastered course material. Educational literature reports a positive, mild correlation between student perceived "self-efficacy" and actual student performance (see, for example, [Lent et al., 1984]). This quantitative method will help to set the broad parameters of the success of the course.

Second, CEER personnel will develop rubrics for assessing the degree to which students meet each of the learning objectives. Moreover, CEER personnel will train the course TAs in applying the rubrics to each student, for each learning outcome. The TAs will mark the rubrics in situ during their interaction with students as they apply the course material to problems. CEER personnel will collect and judge the development of students based on the applied rubrics. The course instructor will not be involved in scoring any of the rubrics, and to further assure fair scoring, at least two TAs will score rubrics for each student, thus enabling triangulation between multiple observers and the ultimate generation of global judgements for the class as a whole.

As part of a broader attempt to evaluate our educational strategy, which is being applied in related areas (most notably, Dr. Brown is teaching a "Computational Science for Evolutionary Biologists" course as part of the BEACON NSF STC grant at MSU), we will leverage multiple additional sources of funding, including matching funds for MSU for assessment (see attached letter from Dr.

[15] http://ceer.egr.msu.edu

Arnosti) and BEACON internal funding. This funding will be used for followup interviews and a more longitudinal study on the application of computational skills in actual research. We will submit our results to the ACM journal “Transactions of Computing Education.”