Technical Report
1177

Bioinformatics Challenge Days

P.A. Carr
D.O. Ricke
A. Shcherbina

30 December 2013

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

Prepared for the Defense Threat Reduction Agency Joint Science and Technology Office (DTRA/JSTO) under Air Force Contract FA8721-05-C-0002.

Approved for public release; distribution is unlimited.

This report is based on studies performed at Lincoln Laboratory, a federally funded research and development center operated by Massachusetts Institute of Technology. This work was sponsored by the Defense Threat Reduction Agency Joint Science and Technology Office (DTRA/JSTO), under Air Force Contract FA8721-05-C-0002.

This report may be reproduced to satisfy needs of U.S. Government agencies.

DTRA/JSTO has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Gary Tutungian
Administrative Contracting Officer
Enterprise Acquisition Division

Non-Lincoln Recipients

PLEASE DO NOT RETURN

Permission has been given to destroy this document when it is no longer needed.

Massachusetts Institute of Technology
Lincoln Laboratory

Bioinformatics Challenge Days

P.A. Carr
D.O. Ricke
A. Shcherbina
Group 48

Technical Report 1177

30 December 2013

Approved for public release; distribution is unlimited.

Lexington Massachusetts

This page intentionally left blank.

ABSTRACT

The Bioinformatics Challenge Days were two one-day events sponsored by the Defense Threat Reduction Agency (DTRA) and carried out by MIT Lincoln Laboratory in cooperation with Edgewood Chemical Biological Center (ECBC). These events explored the utility of a short-term "hack day" format for educating interested individuals in bioinformatics issues relevant to the Department of Defense. They also explored to what extent these events could spark innovation in a diverse technical community that includes computer scientists, engineers, and biologists.

This page intentionally left blank.

TABLE OF CONTENTS

1. OVERVIEW 1

2. BIOINFORMATICS CHALLENGE DAY 1 3

2.1 Preparation for the Event 3
2.2 The Event 4
2.3 Post-Event Analysis 5

3. BIOINFORMATICS CHALLENGE DAY 2 9

3.1 Preparation for the Event 9
3.2 The Event 10
3.3 Post-Event Analysis 14

4. CONCLUSION 19

APPENDIX A DATASET SUMMARIES 21

Time Course Challenge (BCD1) 21
From Metagenomic Sample to Useful Visual (BCD1, 2) 21
De Novo Genetic Assembly (BCD1, 2) 22
Identifying Markers of Genetic Engineering (BCD2) 22

APPENDIX B CHALLENGE DAY SCHEDULES 23

APPENDIX C PARTICIPANT SURVEY (BOTH CHALLENGE DAYS) 25

APPENDIX D TOPICAL PRIMERS FOR BIOINFORMATICS CHALLENGE DAY 2 27

Basic DNA Biology 27
Next-Generation Sequencing Technology 27
Bioinformatics 27
Genetic Engineering Techniques 28
Genetic Engineering Designs 28

This page intentionally left blank.

1. OVERVIEW

The Bioinformatics Challenge Days were conceived as an experiment applying a short "hack day" format to bioinformatics problems of interest to DTRA. Participants of diverse technical backgrounds formed teams to tackle challenges selected by the organizers. The planning for these events was shaped by these motivations:

> *Aggregate*: attract talented individuals with a broad range of skills and experiences; build connections between participants and with organizers
>
> *Educate*: give participants an understanding of bioinformatics challenges of interest to the Department of Defense; fill gaps in their knowledge of bioinformatics in general
>
> *Innovate*: produce ideas, projects, and results that progress toward serving unmet needs in the field of bioinformatics
>
> *Investigate*: explore the successfulness of this one-day format in tackling these challenges; consider how to optimize the effectiveness of such events for future challenges in bioinformatics and other fields

The plan for the Bioinformatics Challenge Days was to conduct two such events, with the second event several months after the first. The second event would provide an opportunity to learn from the first and refine the approach, updating the choice of specific challenges.

Organizing Team

Nancy Burgess, PhD, Defense Threat Reduction Agency, Program Manager

Peter Carr, PhD, Synthetic Biology lead at MIT Lincoln Laboratory

Darrell Ricke, PhD, Bioinformatics lead at MIT Lincoln Laboratory

Anna Shcherbina, MIT Lincoln Laboratory

Nicole Rosenzweig, PhD, Edgewood Chemical Biological Center

Calvin Chue, PhD, Edgewood Chemical Biological Center

This page intentionally left blank.

2. BIOINFORMATICS CHALLENGE DAY 1

The first Bioinformatics Challenge Day (BCD1) took place on 10 September 2012, hosted in coordination with the IEEE High Performance Extreme Computing ("Big Data") conference at the Westin Hotel in Waltham, MA.

2.1 PREPARATION FOR THE EVENT

The organizing team worked together through a series of meetings and conference calls over a period of three months leading up to BCD1. During this time, various possible challenges were considered, with final selections detailed in Appendix A and summarized below:

1. Environmental Time Course: How to analyze and visualize environmental samples for a location sampled on multiple days?

2. Metagenomic Visual: Developing visualization methods to facilitate analysis of metagenomic data with unknown numbers of genomes at varying concentrations

3. Genome Assembly for the Clinic: Performing de novo assembly from clinical samples with an emphasis on pathogen identification

In considering the format for the first event, the organizers took note of different types of previous hack day formats that had been very open-ended (i.e., gave participants a collection of hardware or software, to see what they can do with it) or more narrowly defined (with a very specific problem all participants should focus on). For BCD1 the organizing team chose to give participants broadly worded challenges with relatively open-ended metrics of success. Participants would be encouraged to self-organize into teams as they saw fit.

The event lasted approximately 10 hours, from 9 AM to 7 PM. The first ~2 hours would be spent with presentations from the organizers to orient the participants and introduce the challenges. After a brief break and time for organizing into groups, the participants worked on the challenges throughout the rest of the day with support from the advisors. (See schedule in Appendix B.)

Datasets and available software were made accessible to participants in two ways: through the software distribution website SourceForge and on 32 GB USB sticks the day of the event. These files included precomputed BLAST alignment results for the datasets. To provide access to additional computing power beyond the laptops of individual participant laptops, a server was physically brought to the event locale and made available.

The date (10 September 2012) and location (Westin Hotel, Waltham, MA) were chosen to coincide with the IEEE High Performance Extreme Computing ("Big Data") conference being held there, starting the day afterward. In this way, BCD1 would benefit by drawing on the conference while recruiting potential participants. Coordination with the IEEE Boston Chapter organizing the conference would also facilitate event logisitics.

Outreach to potential BCD1 participants was primarily through the IEEE conference website, which was also used for BCD1 registration. Event registration was free, and participants were not required to register for the adjoining IEEE conference. An advertisement for the event also was sent to e-mail lists for MIT's Electrical Engineering department and MIT CSAIL (Computer Science and Artificial Intelligence Laboratory). A target range of 15–50 participants was prepared for, with an ideal of 30, based on a preferred participant to advisor ratio no greater than 10:1. It was felt that more than this number would limit the ability of the advisors to provide needed feedback throughout the event. By the day of the event, 29 participants were registered.

2.2 THE EVENT

The number of registrants for BCD1 (29) was within the targeted range, but the no-show rate was significantly higher than expected. Eight individuals took part in the event from start to finish (not including organizers), with two more participants more joining later in the day.

While the number of participants was lower than expected, participant engagement was very high. After the introductory lectures, participants broke up into groups according to which challenge they wished to focus on. Each challenge advisor (Ricke, Shcherbina, Rosenzweig) became the focal point of one or more of these groups, helping to answer numerous questions and fill in knowledge gaps. In general, the advisors encountered a broad range of experience in electrical engineering and computing, but less expertise in biology and bioinformatics. This led to the teams spending a great deal of time throughout the day learning important core biology concepts in order to apply their computational skills.

As teams worked throughout the afternoon, it became clear near 6 PM that teams had accomplished what they wished to for this Challenge Day, and effort shifted toward preparing short summary presentations for the close of the event. Some of these employed PowerPoint while others were given using the flip-chart pads that had been provided. Team projects and presentations are summarized below.

Team Gerardo: 1 person, Genome Assembly Challenge

Analyzed oral cavity data to produce a population/diversity analysis bases on Bayesian inference.

Team Awesome: 4 people, Time Course Challenge

Focused on comparing DNA sequence word lengths (e.g., 1–4 DNA bases) for analyzing short sequence reads.

Team Lightning Assembly: 3 people, Genome Assembly Challenge

This team evaluated various available assembly tools, finding Velvet Assembly to be particularly useful.

Team NK: 2 people, Metagenomic Visual Challenge

This team experimented with visual means of displaying differentials between datasets derived from healthy and diseased samples, in particular the different read depths observed for bacteria.

Figure 1. Bioinformatics Challenge Day 1, 10 September 2012, IEEE Big Data Conference, Waltham, MA.

2.3 POST-EVENT ANALYSIS

The organizing team assessed the success of BCD1 in follow-up discussions, both the day of the event and in a separate meeting two days afterward. Qualitative assessment considered the motivations described previously:

Aggregate: attract talented individuals with a broad range of skills and experiences; build connections between participants and with organizers

- The number of participants was small, but it was a motivated group with varied backgrounds across computing, electrical engineering, and biology.
- Biology-oriented backgrounds were less represented than engineering backgrounds.
- Participants and advisors interacted closely throughout the day.

Educate: give participants an understanding of bioinformatics challenges of interest to the Department of Defense; fill gaps in their knowledge of bioinformatics in general

- Participants expressed a great deal of enthusiasm for what they were learning, and appreciation for being able to take part in the event. Several participants described themselves as understanding a great deal more about bioinformatics at the end of the day than when they arrived, including the larger challenges facing the field.
- The small numbers of participants per advisor was in fact felt to be critical to the success of the event, given the need for educating participants in the relevant biology. There was consensus that had the event actually reached the target of 30 people, there would not have been enough advisor time to go around.

Innovate: produce ideas, projects, and results that progress toward serving unmet needs in the field of bioinformatics

- It was felt that innovation during this first event did not reach its full potential.
- A blend of technical backgrounds that included more biology perspective would have been more optimal.
- The format gravitated toward dependence on the advisors. Making advisors less central to the team organization was suggested for the next event.

Investigate: explore the successfulness of this one-day format in tackling these challenges; consider how to optimize effectiveness of such events for future challenges in bioinformatics and other fields

- This initial Challenge Day was thought to be most successful in terms of educating participants and stimulating enthusiasm for the field of bioinformatics.
- Several ways to improve the events were apparent over the course of the day.

The majority of participants also filled out and returned a survey designed to gauge what had been the best part of the experience, and what could improve the event. (See Appendix 3.) The six responses

are summarized below, including numerical ranking of various aspects of BCD1, as well as specific comments. (Not all comments are included—a representative set has been selected.)

How did you hear about the Challenge Day? Conference website (3); e-mail announcement (2), other (1, "IEEE")

How would you rate the…	**(scale of 1–5, with 5 the best)**	**mean**	**+/– sd**
schedule		4.3	0.5
introductory talks		4.8	0.4
opportunities for teaming/networking		5.0	0.0
choice of challenges		4.5	0.5
interactions with organizers/mentors		4.8	0.4
interactions with other participants		4.8	0.4
overall experience		4.8	0.4
How likely are you to…			
continue working on these challenges after today?		4.2	0.4
continue working with your team?		3.7	0.8

What did you learn from the Challenge Day?

- It is possible to make some semblance of progress in even a short time
- Current problems in genomic analyses
- Other perspectives of analyzing data
- That the techniques I use in signal processing can be extended to analysis of next generation DNA sequencing
- It's great to learn overview about bioinformatics in general and some of the tools that are publically available

What kinds of challenges would you like to see in the future?

- Maybe have some that are specifically directed toward a solution
- Wetlab techniques
- Computational science, generally

Other ways Challenge Day could be improved?

- Homework ahead of time
- More advertisement to colleges and graduate departments
- Could use longer time to work on project, but understandable

Additional comments?

- Excellent event, this is what the field needs
- The bioinformatics was confusing to me. I thought it would involve biometrics such as iris identification rather than DNA sequencing
- Great job

In looking ahead to the second Bioinformatics Challenge Day, the organizers considered which strategic changes could be made to improve on the first event. Some of these were modest logistical considerations—for example, the server which had been brought to the event and set up on site had not provided much extra benefit, so it was decided for BCD2 that participant laptops and on-site wireless access would be sufficient. It was especially felt that the low turnout needed to be improved, and multiple factors were discussed that may have contributed. For example, another conference adjacent to the event had been scheduled for the same day, and it had not been clear to participants that their schedules were exclusive (many had signed up for both). Thus, several individuals realized at the last minute that they needed to choose one or the other. The organizing team also considered how the event could be better advertised, including to individuals with more biology background, and how greater commitment could be encouraged among potential applicants (discussed further under Bioinformatic Challenge Day 2).

3. BIOINFORMATICS CHALLENGE DAY 2

The second Bioinformatics Challenge Day (BCD2) took place on 2 February 2013, on the campus of the Massachusetts Institute of Technology at the MIT Media Lab in Cambridge, MA.

3.1 PREPARATION FOR THE EVENT

In planning for BCD2, lessons learned from BCD1 were taken closely into account. Several steps were taken to reach a broader pool of potential participants, and to foster greater commitment from the participants.

Outreach:

- An announcement was made at the annual symposium for the Synthetic Biology Center at MIT (SBC@MIT).
- Fliers advertising BCD2 were posted in several locations around the MIT campus, including the MIT Media Lab, MIT Stata Center, and MIT Infinite Corridor.
- E-mail announcements for BCD2 were sent to MIT CSAIL and MIT Electrical Engineering as before, and also to the MIT Media Lab, Synthetic Biology Center at MIT, Boston Area Synthetic Biology Working Group (mainly the synthetic biology communities at Boston University, Harvard, and MIT), the MIT Bioinformatics seminar series, and the IEEE Boston Bioinformatics Chapter.

Fostering commitment:

- Announcements made clear that the number of participant slots was limited, and that interested participants were required to apply, giving a moderate amount of information about themselves and describe their interest in the event.
- Applicants were asked to confirm in their applications that they would be available to participate for the full day.
- Organizers pursued more communication with participants in advance of the event, including communicating participant acceptance to BCD2, and answering questions from potential participants.
- Access was given in advance for participants to examine event presentations, available software tools, and primers/tutorials on biology and bioinformatics.

As before, during self-organzing/teaming, participants gathered in different areas of the room, according to which challenge they wanted to tackle, but without the advisors seated with them. Advisors instead floated from group to group, offering guidance and perspective as requested. This change from BCD1 was implemented in order to foster team-building and initiative on the part of the participants, and less dependence on the advisors.

The schedule for BCD2 was quite similar to BCD1. (See schedules in Appendix B.) After the introductory lectures, participants were encouraged to spend time getting to know each other's technical backgrounds and ideas, and to form teams of complementing abilities. The campus of the Massachusetts Institute of Technology (MIT) was selected as the location for BCD2, specifically the meeting facilities on the sixth floor of the MIT Media Lab (MIT Building E14). The flexible open floor plan of this space would allow teams to spread out in different parts of the room. It was also felt that holding the event at a well-known academic institution would help recruit the broad range of desired participants. MIT is centrally located in the Boston area and easily accessible via public transportation.

For BCD2, four challenges were planned, the original three plus an additional challenge centered on genetic engineering. This last challenge was intended to motivate participants to consider how one might identify a genetically engineered organism present in a mixture with naturally occurring organisms. However, due to U.S. budget sequestration, organizing team members from ECBC were unable to attend. With three advisors remaining, it was felt best to move ahead with no more than three challenges, allowing one advisor to serve as point person for each challenge.

1. Environmental Time Course: How to analyze and visualize environmental samples for a location sampled on multiple days?

2. Metagenomic Visual: Developing visualization methods to facilitate analysis of metagenomic data with unknown numbers of genomes at varying concentrations

3. Genome Assembly for the Clinic: Performing de novo assembly from clinical samples with an emphasis on pathogen identification

4. Genetic Engineering: Identifying which components in a sample may have been engineering, how they have been engineered, and what they have been engineered to do

The new Genetic Engineering challenge was also intended to provide a more goal-oriented challenge, in response to participant feedback from BCD1. Success in this challenge would be determined by how many genetically engineered features the teams could correctly identify, and how much of their function could be interpreted.

3.2 THE EVENT

As the morning of BCD2 began, the organizers were optimistic that the attendance rate would be much higher than for BCD1, with an expectation that roughly 80% of the 30 registered and accepted particpants would arrive. In fact, 100% of these participants arrived, and were eager to get started. In addition, several other interested individuals who had not been able to commit to the full day visited to see how the event was progressing. Two additional MIT undergraduates representing the Executive Committee of the MIT Biology Undergraduate Student Association (BUSA) were also allowed to attend as visitors, as BUSA had expressed interest in hosting a future similar event. Furthermore, two BCD2 participant expressed strong interest in organizing a similar BCD event on their own campuses, and a third requested use of some BCD2 materials in teaching a course.

The majority of participants chose to focus their attention on the new BCD2 challenge, Genetic Engineering. Many responded (see survey results below) that they liked the more defined nature of this challenge, as there was a clearer objective.

The organizers agreed that for this event the sense of energy and enthusiasm in the room was high. The new strategy of making sure advisers were not central to the team structure was also deemed successful, facilitating more initiative on the part of the participants in working together to solve problems. This choice also allowed the advisors to support a larger number of participants than would have been possible for BCD1. As teams approached the 6 PM break point (when they were asked to shift focus to preparing their presentations), multiple participants expressed the desire to keep working on their challenge. One team was so intent on their work that it required multiple friendly encouragements from the organizers to bring them back together with the larger group at the time of presentations.

Participants were encouraged to think of the presentations not as final reports, but as progress reports. With the understanding that only so much could be accomplished in the one day format, the efforts invested in their initial work could be built upon post-BCD2. Most team presentations employed PowerPoint while one was given using the flip-chart pads that had been provided. Team projects and presentations are summarized below.

Team Disease Easy: 7 people, Metagenomic Visual Challenge

Focused their visual output on Swann plots (a type of annotated scatter plot), linking data on sequencing reads to potential pathogen matches, pulling metagenomic data from the PATRIC database of pathogenic organisms.

Team Sajjad: 1 person, Metagenomic Visual Challenge

Mapped genetic differences between samples graphically, representing these differences as "distances," employing an existing tool (D3).

Team CDX: 6 people, Genetic Engineering Challenge

Focused their efforts on how to identify common DNA sequence elements employed in genetic engineering, such as "scars" left behind by methods using restriction enyzmes, as well as frequently used functional DNA sequences (certain promoters, double terminators, fluorescent proteins).

Team Zach & George: 2 people, Genetic Engineering Challenge

Employed a series of available tools to identify the species present in the dataset. They then went on to identify 25 genes that had been deleted from the modified E. coli strain, as well as several SNPs.

Team Eddie: 1 person, Genetic Engineering Challenge

Laid out a framework for "Threat Reduction Thinking" with several stages of analysis, ranging from quick ID of genetic markers, to more in-depth analysis of mutations (kinds and numbers), likelihood of who would have access to the specific engineering technology, and detailed sequence pattern analysis.

Team RESTPRSMJ: 9 people, Genetic Engineering Challenge

Used the available dataset to build contiguous stretches of DNA sequence ("contigs") before searching for matches to specific outside DNA sequence databases. In doing so, they correctly identified two of the engineered elements in the dataset, phase lambda, and the specific plasmid.

(A few individuals also opted not to present, or needed to leave before the presentation time.)

Figure 2. Bioinformatics Challenge Day 2, February 2, 2013, MIT Media Lab, Cambridge, MA.

3.3 POST-EVENT ANALYSIS

The organizing team assessed the success of BCD2 in follow-up discussions, both during and following the event. This assessment considered the motivations described previously:

Aggregate: attract talented individuals with a broad range of skills and experiences; build connections between participants and with organizers

- BCD2 participants had backgrounds sampling the desired range of experience, including biology, bioinformatics, and computer science.
- Ages ranged from high school (1 participant) to late career professionals.
- A large number of participants were present from both academia and industry.
- BCD2 was very successful in recruiting the desired number of participants (30). Of the 30 who had been accepted for the event, 100% attended.
- Two additional MIT undergraduates representing the Executive Committee of the MIT Biology Undergraduate Student Association (BUSA) were allowed to attend as visitors, as BUSA had expressed interest in hosting a future similar event.

Educate: give participants an understanding of bioinformatics challenges of interest to the Department of Defense; fill gaps in their knowledge of bioinformatics in general

- Many participants expressed their gratitude for being able to take part in BCD2, that the day had been very worthwhile, and that they had learned a great deal.
- Participants in BCD2 also learned a great deal from each other, not only from the advisors.
- (See participant comments below for more details.)

Innovate: produce ideas, projects, and results that progress toward serving unmet needs in the field of bioinformatics

- While there were more innovative ideas presented than at BCD1, maximizing innovation in this context is difficult, especially with the time required to educate the new practitioners in the field.
- Innovation/productivity might be furthered by selecting challenges that are more highly defined, with very specific goals.
- A number of the most motivated and intense participants indicated that they would have preferred more time to work on the challenges.

Investigate: explore the successfulness of this one-day format in tackling these challenges; consider how to optimize the effectiveness of such events for future challenges in bioinformatics and other fields

- It was generally agreed by the organizers that the lessons learned from BCD1 and implemented in BCD2 helped make the second event much more successful overall.
- With this short one-day time frame, this format can be quite useful for aggregating and educating groups of talented, motivated, diverse participants.
- Fostering true innovation in this context seems quite difficult, though producing seeds of projects to be completed later has significant potential.

More than two-thirds of participants also filled out and returned a survey designed to gauge what had been the best part of the experience, and what could improve the event. (This survey was the same for BCD1 and BCD2; see Appendix 3.) The 22 responses are summarized below, including numerical ranking of various aspects of BCD1, as well as specific comments. (Not all comments are included—a representative set has been selected.)

How did you hear about the event? E-mail announcement (13); Friend/colleague (6); Flier (3)

How would you rate the… (scale of 1-5, with 5 the best)	**mean**	**+/– sd**
schedule	4.4	0.7
introductory talks	4.2	1.0
opportunities for teaming/networking	4.4	0.7
choice of challenges	4.2	0.9
interactions with organizers/mentors	4.6	0.5
interactions with other participants	4.5	0.6
overall experience	4.4	0.7
How likely are you to…		
continue working on these challenges after today?	4.0	1.0
continue working with your team?	3.0	1.1

What did you learn from the Challenge Day?

- How the bioinformatics field is diverse and rich
- Bioinformatics basics—tools and thinking and future directions
- Challenges of big data visualization
- Collection of tools and got a little experience using them

- A lot about sequencing tools and genetic engineering
- Bioinformatics is plagued by annotation problems and limited information, but surprisingly much progress can be made to decipher complicated information, even in one day
- Different ways people are thinking about organism-type inference
- Many tools and file formats for bioinformatics
- Processing multi-organism raw data (metagenomic) with BLAST
- Potential applications of metagenomic data
- Teamwork

What kinds of challenges would you like to see in the future?

- Synthetic biology-related, i.e., design, compiler, ways to get more people involved
- Bioinformatics algorithms improvement/discover
- Machine learning for pattern searching
- I'd like to continue working on reverse engineering organisms
- More challenges with focused goals like the genetic design challenge
- I liked the genetic engineering challenge, especially because it had very clear questions to pursue
- More focus on nonsimulated datasets
- Having a puzzle to figure out for each challenge would be great
- Challenges with more specific goals. The Assembly challenge was ambiguous. Also more datasets (replicates)
- Medical data; ENCODE
- Mobile app development; provide APIs

Other ways Challenge Day could be improved?

- Scheduling more time to work
- More details ahead about the challenges, recommended lecture to get a better picture the day of the challenge
- More topics to choose

- You could probably get a guest speaker to tell about some methods or tools
- I think many people gravitated toward the genetic engineering challenge because it was clearly defined, so definitely having clear goals is helpful
- Primer on pipeline that processed raw data
- More in-depth lectures
- Less intro talks. Get prizes → huge incentive. More time to work on projects (overnight event?)
- Less introductory talks. More time to code.
- 24-hour competition with prizes

Additional comments?

- Awesome experience, thanks for putting this together!
- I love this challenge day. Fantastic!
- It was great to be part of [this] innovative group of folks
- Excellent! Highly enjoyable and informative. I would definitely participate again, if given the opportunity.
- There's a gap between what biologists and programmers can do

---

This page intentionally left blank.

4. CONCLUSION

During both Challenge Days there was a strong consensus among participants that it had been a worthwhile, interesting, and educational experience. A large proportion of participants expressed a great deal of appreciation for being included in the event and for the efforts of the organizers. There was also interest in future events of this nature, and several participants indicated a desire to organize their own similar event at their campus or place of work.

Several factors contributing to the success of the Challenge Days have been discussed above, as well as potential for improvement. Especially important among these was reaching out to participants well in advance of the event, through several different channels, and emphasizing to potential applicants the value of taking part. Also of great value was having challenges clearly defined so that participants could work toward well-articulated goals.

To further maximize the innovative output of these events, there would very likely need to be more time for project work, including interaction within teams. For this purpose one might consider providing a website that facilitates teaming, as well as online discussions of challenges. Teams could form and begin work on some aspects of the challenges before the event day. (Some very energetic participants also indicated interest in a 24-hour version of the event.) A further challenge (for the organizers) would be to devise means to foster team development and continued work after the event as well.

This page intentionally left blank.

APPENDIX A
DATASET SUMMARIES

TIME COURSE CHALLENGE (BCD1)*

Metagenomic datasets (FASTQ format) from environmental samples were taken at three timepoints. The environmental samples were collected once a day for three days. Each day, biological material and particulates were captured in a buffer that is *not* DNA-free. Particulates (inorganic or plant material) were removed by centrifugation, DNA was extracted, and the genetic material was sequenced on Illumina HiSeq 2000.

FROM METAGENOMIC SAMPLE TO USEFUL VISUAL (BCD1, 2)

One publicly available FASTQ dataset was obtained from the Metagenomics Analysis Server (MG-RAST). It consisted of two supragingival dental plaque samples from the project Oral Metagenome in Health and Disease with IDs 4447943.3 and 4447192.3. Sample 4447943.3 was taken from a healthy subject, and sample 44471923 was extracted from a subject that suffered from periodontal disease. The rationale for including these two samples was to identify the differences and similarities between the metagenomes in the healthy and diseased states. 454 pyrosequencing was used to generate both samples.

Four publicly available FASTQ datasets were obtained from the NCBI Sequence Read Archive (SRA).

- Sample SRS084836: RNA was extracted from a nose/throat swab from a Nicaraguan child with acute respiratory illness. cDNA was randomly amplified using the Illumina Genome Analyzer II. The dataset contained one lane of Illumina paired-end sequences, 65 bases in each direction. The sample contained 2 Gbp of data.

- Sample SRX001682: 16S rRNA was obtained from a human hand skin sample. The goal of the study was to determine the influence of sex, handedness and washing on the diversity of hand surface bacteria. The 454 GS FLX instrument was used to perform pyrosequencing on the sample. The sample contained 133 Mbp of data.

- Sample DRX000972: A human abscess sample of unknown etiology was used for a study on comprehensive detection of possible bioterrorism agents, Francisella sp, from clinical specimens using next-generation direct DNA sequencing. The sample contained 42 Mbp of data. From the study page:

  > We performed unbiased direct sequencing using a next-generation DNA sequencer Illumina GAIIx to detect potential pathogens in a abscess sample of unknown etiology. The direct deep sequencing identified the potential pathogen Francisella tularensis which

---

* Provided by Edgewood Chemical Biological Center (ECBC), Dr. C. Nicole Rosenzweig.

is a category A select agent. Genomic single nucleotide variations on the Francisella spp. genomes, the case was associated with Francisella tularensis subsp. holarctica (biovar japonica) infection, but not highly virulent Type A or other subsp. holarctica. This case was concluded as a sporadic infection in Japan, instead of bioterrorism (http://sra.dnanexus.com/experiments/DRX000972/studies).

- Sample SRX116574: The soil metagenome sample was obtained via Illumina sequencing of a Kansas cultivated corn soil metagenome reference library. The Illumina Genome Analyzer II was used. The sample contained 1.4 Gbp of data.

DE NOVO GENETIC ASSEMBLY (BCD1, 2)

The same datasets were used for this challenge as for the Metagenomic Sample to Useful Visual Challenge (see above).

IDENTIFYING MARKERS OF GENETIC ENGINEERING (BCD2)

The background dataset consisted of a human blood metagenome sample sequenced on an IonTorrent single-end sequencer.[†] The dataset was roughly 92.6 MB in size. Analysis determined that the primary organisms included in the background were *Escherichia coli* (8 reads) and Enterobacteria phage T4 (25 reads).

The background dataset was then spiked in silico with codon-substituted Enterobacteria phage lambda (gi|215104|gb|J02459.1|LAMCG). Codon substitution was performed to remove/recode TTA, TTG, TAG, AGT, AGC, AGA, AGG, representing a reduced genetic code. The phage genome was combined with Escherichia coli reads to mimic the error/mutation rate of the background profile, and the combined phage/*E. coli* reads were spiked in at 20× relative to background.

A plasmid (EU496103 with xis-AAV BioBrick cloning vector pSB3C5-I52001) was spiked in at 200× relative to the background.

Overall, a total of 46% of the dataset consisted of *in silico*–generated reads from *E. coli*, Enterobacteria phage lambda, and plasmid EU496103.

[†] All data used for the Challenge Days was in FASTQ format.

APPENDIX B
CHALLENGE DAY SCHEDULES

Bioinformatics Challenge Day 1

IEEE-HPEC Conference

The Westin Hotel, Waltham, MA

September 10, 2012

8:00 AM Breakfast/check-in

9:00 AM Welcome (Nancy Burgess, DTRA)

9:15 AM Challenge Day overview and logistics (Peter Carr, MIT LL)

9:45 AM The Challenges

1. Environmental Time Course (Nicole Rosenzwieg, ECBC)
2. Metagenomic Visual (Anna Shcherbina, MIT LL)
3. Genome Assembly for the Clinic (Darrell Ricke, MIT LL)

10:45 AM Coffee/break into project groups/teaming

12:30 PM Lunch served (groups can continue to work)

3:30 PM Snack (groups can continue to work)

7:00 PM Results submitted by dinnertime

7:00 PM Dinner

8:00 PM Final results announced

Bioinformatics Challenge Day 2

MIT Media Lab, Cambridge, MA

February 2, 2013

8:00 AM Breakfast/check-in

9:00 AM Welcome (Peter Carr, MIT LL)

9:15 AM Overview and logistics (Peter Carr, MIT LL)

9:45 AM The Challenges:

1. Metagenomic Visual (Anna Shcherbina, MIT LL)
2. Genome Assembly for the Clinic (Darrell Ricke, MIT LL)
3. Genetic Engineering (Peter Carr, MIT LL)

10:45 AM Coffee/Break into project groups

12:30 PM Lunch served (groups can continue to work)

3:30 PM Snack (groups can continue to work)

6:30 PM Progress updates ready by dinnertime

6:30 PM Dinner and progress reports

8:00 PM+ Groups can continue to work

APPENDIX C
PARTICIPANT SURVEY (BOTH CHALLENGE DAYS)

Name (optional):_____

Bioinformatics Challenge Day

Survey/Feedback

Thank you for taking part in the Bioinformatics Challenge Day! One of the major outcomes we are hoping for is to learn from this experience. Your feedback will make a huge difference toward putting together future events that can be even better.

How did you hear about the Challenge Day? (circle one)

conference website friend/colleague e-mail announcement other:_____

Suggestions for other places to advertise?

How would you rate the…	**poor**				**excellent**
schedule	1	2	3	4	5
introductory talks	1	2	3	4	5
opportunities for teaming/networking	1	2	3	4	5
choice of challenges	1	2	3	4	5
interactions with organizers/mentors	1	2	3	4	5
interactions with other participants	1	2	3	4	5
overall experience	1	2	3	4	5
How likely are you to…	**unlikely**				**very likely**
continue working on these challenges after today?	1	2	3	4	5
continue working with your team?	1	2	3	4	5

What did you learn from the Challenge Day?

What kinds of challenges would you like to see in the future?

Other ways Challenge Day could be improved?

Additional comments?

This page intentionally left blank.

APPENDIX D
TOPICAL PRIMERS FOR BIOINFORMATICS CHALLENGE DAY 2

BASIC DNA BIOLOGY

DNA structure and replication:

http://www.hartnell.edu/tutorials/biology/dnareplication.html

Transcription:

http://www.hartnell.edu/tutorials/biology/transcription.html

Translation:

http://www.hartnell.edu/tutorials/biology/translation.html

NEXT-GENERATION SEQUENCING TECHNOLOGY

The Wikipedia entry is quite good for a basic intro:

http://en.wikipedia.org/wiki/DNA_sequencing#Next-generation_methods

A recent analysis comparing three next-gen sequencing platforms:

http://www.biomedcentral.com/1471-2164/13/341

Special issue of the journal *Bioinformatics*, focus on next-gen sequencing:

http://www.oxfordjournals.org/our_journals/bioinformatics/nextgenerationsequencing.html

Clinical potential of sequencing:

http://www.genengnews.com/gen-articles/dna-sequencing-the-clinical-potential/4684/

BIOINFORMATICS

Bioinformatics primer:

http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html

Wikipedia entry (especially “Software and tools”):

http://en.wikipedia.org/wiki/Bioinformatics

Overview of many types of bioinformatics:

http://www.bioinformatics.nl/webportal/background/techniques.html

More specifics on techniques:

http://lectures.molgen.mpg.de/online_lectures.html

GENETIC ENGINEERING TECHNIQUES

A very simple glossary of key terms:

http://theagricos.com/biotechnology/genetic-engineering/genetic-engg-techniques/

Restriction enzymes:

http://en.wikipedia.org/wiki/Restriction_enzyme

DNA assembly, with links to traditional, Gibson, and Golden Gate methods:

http://j5.jbei.org/j5manual/pages/76.html

MAGE: Multiplex Automatable Genome Engineering:

http://www.nature.com/nature/journal/v460/n7257/full/nature08187.html

Registry of Standard Biological Parts:

http://partsregistry.org/Main_Page

Assembly methods such as Biobrick standard assembly:

http://partsregistry.org/Help:Assembly

GENETIC ENGINEERING DESIGNS

The Elowitz oscillator:

http://www.ncbi.nlm.nih.gov/pubmed/10659856

The Collins toggle:

http://www.ncbi.nlm.nih.gov/pubmed/10659857

Weiss yeast cell-cell communication:

http://www.nature.com/nbt/journal/v23/n12/abs/nbt1162.html

Weiss, Benenson cancer cell classifier circuit:

http://www.sciencemag.org/content/333/6047/1307

Re-organizing a simple genome:

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1681472/

Toward a new genetic code (uses MAGE, introduces CAGE):

http://www.sciencemag.org/content/333/6040/348

The Registry contains many genetic circuit designs from the iGEM competition. Some work, some don't:

http://partsregistry.org/Main_Page

This page intentionally left blank.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
30 December 2013	Technical Report	

4. TITLE AND SUBTITLE

Bioinformatics Challenge Days

5a. CONTRACT NUMBER
FA8721-05-C-0002

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

P.A. Carr, D.O. Ricke, and A. Shcherbina

5d. PROJECT NUMBER

5e. TASK NUMBER
2317-11

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420-9108

8. PERFORMING ORGANIZATION REPORT NUMBER

TR-1177

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Nancy Burgess, Defense Threat Reduction Agency
8725 John J Kingman Rd Stop 6201
Fort Belvoir, VA 22060-6201

10. SPONSOR/MONITOR'S ACRONYM(S)
DTRA/JSTO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution Statement A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The Bioinformatics Challenge Days were two one-day events sponsored by the Defense Threat Reduction Agency (DTRA) and carried out by MIT Lincoln Laboratory in cooperation with Edgewood Chemical Biological Center (ECBC). These events explored the utility of a short-term "hack day" format for educating interested individuals in bioinformatics issues relevant to the Department of Defense. They also explored to what extent these events could spark innovation in a diverse technical community that includes computer scientists, engineers, and biologists.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*
Unclassified	Unclassified	Unclassified	Same as report	40	

This page intentionally left blank.