




OPINION ARTICLE

# Introduction to Genomic Analysis Workshop: A catalyst for engaging life-science researchers in high throughput analysis [version 1; peer review: 2 approved]

Phillip A. Richmond <sup>1,2</sup>, Wyeth W. Wasserman<sup>1,2</sup>

<sup>1</sup>Centre for Molecular Medicine and Therapeutics, BC Children's Hospital Research Institute, Vancouver, British Columbia, V5Z 4H4, Canada  
<sup>2</sup>Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada

**v1** **First published:** 30 Jul 2019, 8:1221 (<https://doi.org/10.12688/f1000research.19320.1>)  
**Latest published:** 30 Jul 2019, 8:1221 (<https://doi.org/10.12688/f1000research.19320.1>)

**Abstract**

Researchers in the life sciences are increasingly faced with the task of obtaining compute resources and training to analyze large, high-throughput technology generated datasets. As demand for compute resources has grown, high performance computing (HPC) systems have been implemented by research organizations and international consortiums to support academic researchers. However, life science researchers lack effective time-of-need training resources for utilization of these systems. Current training options have drawbacks that inhibit the effective training of researchers without experience in computational analysis. We identified the need for flexible, centrally-organized, easily accessible, interactive, and compute resource specific training for academic HPC use. In our delivery of a modular workshop series, we provided foundational training to a group of researchers in a coordinated manner, allowing them to further pursue additional training and analysis on compute resources available to them. Efficacy measures indicate that the material was effectively delivered to a broad audience in a short time period, including both virtual and on-site students. The practical approach to catalyze academic HPC use is amenable to diverse systems worldwide.

**Keywords**



genomics, education, genome analysis, high throughput computing, hpc, life sciences




This article is included in the **Bioinformatics Education and Training Collection** collection.

**Open Peer Review**

**Reviewer Status** 

	Invited Reviewers	
	1	2
<b>version 1</b> published 30 Jul 2019	 report	 report

- 1 **Richard Fitzpatrick**, University of Edinburgh, Edinburgh, UK
- Melanie I. Stefan** , University of Edinburgh, Edinburgh, UK
- 2 **Michael Springer**, Harvard Medical School, Boston, USA

Any reports and responses or comments on the article can be found at the end of the article.

**Corresponding author:** Wyeth W. Wasserman ([wyeth@cmmt.ubc.ca](mailto:wyeth@cmmt.ubc.ca))

**Author roles:** **Richmond PA:** Data Curation, Formal Analysis, Investigation, Methodology, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Wasserman WW:** Investigation, Supervision, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** The work was supported by NSERC Discovery Grant (RGPIN-2017-06824) acquired by WWW, and PAR was supported by a BC Children's Hospital Research Institute Graduate Studentship award during this work.

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2019 Richmond PA and Wasserman WW. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Richmond PA and Wasserman WW. **Introduction to Genomic Analysis Workshop: A catalyst for engaging life-science researchers in high throughput analysis [version 1; peer review: 2 approved]** F1000Research 2019, 8:1221 (<https://doi.org/10.12688/f1000research.19320.1>)

**First published:** 30 Jul 2019, 8:1221 (<https://doi.org/10.12688/f1000research.19320.1>)

## Introduction

The era of genomics and DNA sequencing is being rapidly incorporated into life science research fields, spanning fields as diverse as population-scale human genetics, model organism studies and patient-focused precision medicine. Researchers have harnessed the wealth of data produced to unveil previously unattainable insights into their research questions. Although researchers across different fields are beginning to utilize the power of genomics, two major roadblocks of accessible compute resources and lack of training prevent them from being able to effectively analyze their own data<sup>1</sup>. An emerging solution to deliver high performance computing (HPC) to researchers worldwide comes through centralization of compute resources, generally in the form of grid or cluster compute servers. Examples include Compute Canada which delivers coordinated grid computing to Canadian academic institutions, the National Computational Infrastructure that coordinates the super-computer system Raijin for Australian researchers, XCEDE which coordinates compute power for academic researchers in science and engineering across the United States, and numerous existing European academic HPC solutions such as Partnership for Advanced Computing in Europe (PRACE). Although such academic HPC systems have been historically used by researchers in physics and chemistry—fields dominated by high throughput calculations and big data computations—the systems are being increasingly used by life scientists. Even as funding for academic HPC systems grows, a lack of facilitation to introduce life scientists to their use remains a roadblock for many potential users.

The increasingly digital and quantitative analyses required in life science research has been noted for decades, but the arrival of accessible and affordable DNA sequencing technologies has accelerated the demand for skills that are not yet commonly incorporated into training programs. Thus, beyond the roadblock of HPC access and use, life science researchers must also acquire training in genomic analysis. Recent global surveys on the training needs for life science researchers in bioinformatics analysis revealed that most training comes at time-of-need or point-of-need, imposed by the necessity to analyze acquired data, instead of being a core part of the curriculum or formal education<sup>2,3</sup>. The bioinformatics community has reacted to the challenge of meeting this time-of-need training on numerous fronts with great efficacy. Current training options in bioinformatics include massive open online courses (MOOCs)<sup>4</sup>, static online tutorials<sup>5,6</sup> and in-person workshops<sup>3,7,8</sup>. Although online forums provide a variety of resources, current tutorials are not compute-resource specific and require compute environment tailoring for HPC systems which is prohibitive to researchers lacking strong computational skills. Furthermore, surveys regarding efficacy place a high value on the practical analysis components, which are often not delivered in the online capacity due to difficulty of coordination<sup>1</sup>. In-person workshops allow for hands on applications, but require overhead cost and scheduling, which can be a limitation to many researchers. Also, many of these workshops are primarily tailored for analysis on a laptop or desktop, which differ from the environment of HPC platforms<sup>9</sup>. Moreover, some of the most

popular workshops have transitioned away from introductory content, instead focusing on more complex topics and applications. Such workshops often assume background knowledge and experience in Linux and HPC. This transition occurred in part due to an expansion in online curriculum for introductory content, as well as the need to refine hundreds of workshop applicants to a feasible number of local attendees<sup>8,9</sup>. Lastly, most available workshops, both online and in-person, are rigid in structure, not catering to the interdisciplinary and diverse skill levels prevalent in the genomic era. In summary, researchers in the life sciences are faced with the need for acquiring introductory analysis skills in an environment that has the capacity for high throughput analysis, and no current training options are singular and effective at delivering this training in a time-sensitive manner.

In light of the challenges, we sought to create a new educational approach to catalyze the use of academic HPC systems by life scientists. We envisioned a flexible, centrally-organized, easily accessible, interactive, and compute resource specific training for the foundational skills of genomic analysis. To this end, we implemented and taught a modular workshop series titled: *“Introduction to Genomic Analysis”*. The material was taught in both an online (using the Vidyio system) and local (at the BC Children’s Hospital Research Institute) capacity for a total of seven two-hour interactive sessions. We focused the content on practical application, effective use of available compute resources, and data analysis exercises, while avoiding other aspects of genomics such as theory and experimental design. Since these aspects of genomics are fundamental, we encourage students to utilize external open-source materials deposited in curated bioinformatics education repositories<sup>10</sup>. The modularity of our workshop structure allows participants to pick-and-choose the sessions to attend based on prior experience level, thereby catering to both those new to the command-line and those experienced in Linux with an interest in exposure to genomic data analysis. Moreover, a strong emphasis was placed upon student evaluation in our workshop delivery. To benchmark student progress, problem sets were used as module exit and entrance requisites and a culminating exam assessed the ability to effectively analyze next generation DNA sequencing data.

In total, 80 participants attended at least one of the seven modules, and 58 certificates of completion were awarded based on completion of the core modules. Our initial cohort of students was diverse, including a spectrum of prior genomic analysis experience, equal gender representation, and various educational levels which recapitulates the world-wide audience<sup>4</sup>. Successful completion of the workshop showed similar results for in-person and virtual attendees and across levels of prior experience. Post-workshop surveys of the course efficacy provided deeper insight into potential improvements for future implementations of this material, and expansion into more detailed non-introductory topics of analysis.

In summary, we demonstrate the utility of our workshop format and information delivery methodology with the hopes that

other trainers across the globe will improve and adapt its content to fit the needs of the life-science research community. The materials will persist in an open source state, and future implementations of the workshop will be delivered as we continue to fill the gap in genomic analysis training.

In accordance with expectations and guidelines for effective bioinformatics workshops, the materials are all open source, stored online, and video-recorded for future use<sup>11</sup>. All materials are published under Creative Commons ShareAlike 3.0 Unported License (CC BY-SA 3.0) and can be accessed at: <https://phillip-a-richmond.github.io/Introduction-to-Genomic-Analysis/>.

## Implementation

### Content delivery

The materials for this workshop were designed to emphasize accessibility, consistency, modularity, and reusability, designed around an environment with capacity for high throughput analysis. We also place an emphasis on student evaluation by implementing assigned problem sets and a final exam.

**Accessibility.** The primary challenge for delivery was the presence of both local and virtual attendees, which allowed us to expand our attendance beyond typical locally-constrained workshops. We reached both audiences simultaneously by delivering the content through an audio-visual casting of the primary teacher's screen, which showed both lecture slides and an open terminal for executing commands. Local attendees followed the session on a projector and had tables for their laptops, while virtual attendees tuned in via internet broadcast of the screen capture using [Vidyo software](#). Teaching assistants (TAs) monitored both local and virtual participants, the later using [openstack Etherpad](#), and responded to questions and provided assistance during the follow-along lecture (a lecture format in which students repeat commands as they are performed by the instructor). The Etherpad environment provides an online text document updated in real-time that contains links to resources, an attendance section, and a question-and-answer section monitored by the TAs. Lectures were recorded and made available after the session for participants who couldn't attend or desired to re-watch the presentation.

**Consistency.** It was a goal in developing the materials to provide a consistent structure and process for each workshop session. Every session started with 45 minutes of a follow-along lecture that integrated commands for the audience to execute alongside the primary teacher. Commands executed were documented as [GitHub Gists](#), which also contained additional details regarding command usage. At the end of the lecture, students spent 1 hour working through a problem set in small (2-3 person) groups with assistance from a roaming TA (virtually via Etherpad or locally in-person). An important introductory concept for genomic analysis is maintaining a well-organized hierarchical filesystem structure. To enforce this practice, which includes centralization of reference genomes as well as separation of raw from processed data, each session followed an identical file structure (*Extended data:* Supplemental Figure 1). Within this structure, individual student directories

titled with unique identifiers allowed for both tracking of student participation, and simultaneous use of common files throughout the session.

**Modularity.** Modularity was a key component of this workshop as attendees had different prior training and experience. To enforce modularity and allow participants to skip material they had previously mastered, we designed each workshop session with a prerequisite assignment that assessed the comprehension of the preceding material. The problem set at the end of each session would serve as the prerequisite entry to the next. We found this to be necessary due to a mixing of content between basic Linux command-line usage and applied short-read sequencing analysis. A final exam was performed after 4 core modules, which served as a comprehensive evaluation of basic skills necessary for genomic analysis in the HPC environment. Completion of the exam was necessary for attendance of the final three modules, which delved into more advanced topics.

**Reusability.** All course materials are available through open source licensing under Creative Commons ShareAlike 3.0 Unported License (CC BY-SA 3.0), and a single github-based website links together the relevant resources. These resources include the lecture slides, problem sets, Github Gists, course exam, Etherpad links, and recordings of the lecture and problem-set sessions. Additionally, the workshop directory on the HPC environment remains for future individual use within the structured environment.

**HPC environment.** Contrary to numerous laptop-based learning modules and teaching practices, we designed our material around analysis within a HPC environment. The motivation was two-fold: 1) by teaching in this environment we could combine an introduction to computing and Linux with an introduction to genomic analysis; and 2) we prepared researchers to be comfortable with data analysis on the platform upon which they would analyze data in the future. We utilized a national Canadian grid HPC system, Compute Canada, set up with a module system for controlling software dependencies, running Torque-Moab scheduling software for distributing jobs from the head node to the compute nodes. With slight modifications, the material can be adapted for use on most academic HPC systems. We provided temporary guest accounts to participants lacking an account. This allowed us to expose the attendees to the environment and actively engaged them to utilize the resources that are made available to them as academic researchers. Numerous participants followed up with the systems administrators to acquire full accounts during the workshop and after its conclusion.

### Workshop materials

Standards in workshop design and implementation have highlighted the importance of labelling learning objectives and prerequisites explicitly for each session<sup>11</sup>. An overview of the workshop materials is displayed in [Table 1](#). Since the workshop is modular, with benchmarks of understanding and analysis capabilities before and after each session, there is little redundancy in the per-session learning objectives. With this workshop

**Table 1. Workshop materials.** Breakdown of the workshop materials, including the learning objectives, commands learned, and prerequisites for each session.

Session #	Title	Learning Objectives	Commands Learned	Prerequisites
<b>Session 1</b>	Introduction To Linux I: Basic Command Line	Basics of computing with high performance computers	cp, ls, mv, cut, clear, mkdir, rm, wget, grep, more, less, head, tail, cat, gunzip, gzip, chmod	Log-in to the cluster
		Filesystem hierarchy		
		Basic command line operations		
		File handling & permissions		
		Standard out		
		File Formats: GTF, Clinvar Variant Summary		
<b>Session 2</b>	Introduction to Linux II: Interacting with the Queue	Basics of computing with high performance computers	emacs, nano, qstat, qsub, showq, module avail, module list, module load, scp	Session 1 Problem Set
		Editing files with linux file-editors		
		Shell Scripts		
		Interacting with the queue		
		File transfer		
<b>Session 3</b>	Short Read Mapping and Visualization	Next Generation Sequencing Primer	BWA mem, samtools sort, samtools view, samtools index	Session 2 Problem Set
		Map short read DNA sequences to the genome using BWA mem		
		Convert file formats using Samtools		
		Utilize scheduler for pipeline execution		
		Visualize short read data in IGV		
		File Formats: SAM, BAM, indexed BAM, Fastq, Fasta		
<b>Session 4</b>	Variant Calling (Small variants)	Exome sequencing	freebayes, bgzip, tabix	Session 3 Problem Set
		Mapped read post-processing		
		Variant calling for small variants		
		Variant compression and indexing		
		Visualization of variants and short-read data together		
		File Formats: VCF, compressed VCF, indexed VCF		
<b>Session 5</b>	Variant Interpretation with GEMINI	Brief introduction to MySQL queries	gemini load, gemini query, gemini <i>de novo</i>	Course Exam
		Variant annotation with VEP		
		File Formats: PED file, gemini.db		
<b>Session 6</b>	RNA seq I: Analysis with the Tuxedo Suite	RNAseq overview	HISAT2, Stringtie	Course Exam
		RNAseq read mapping and visualization		
		Transcript assembly		
		Transcript quantification		
<b>Session 7</b>	RNA seq II: Differential Expression Analysis in R	R data loading and visualization	R, DESeq2	Course Exam
		Differential expression using DESeq2		

design, we build on commands and topics from previous sessions with increasing complexity. For example, in Session 3, students learn about mapping short read DNA sequencing data to the human genome, and then, in Session 4, they re-run the same mapping commands but with the addition of variant calling. For students who are unable to master the prior material, they are encouraged to follow along and revisit the material from previous sessions. The workshop materials are available online through a coordinated website available at <https://phillip-a-richmond.github.io/Introduction-to-Genomic-Analysis/>.

## Evaluation of efficacy

### Participant breakdown

The group of participants in the workshop series was diverse in prior training, sex, academic standing, and mode of attendance (Figure 1A–C, Figure 2A). The academic standing of our participants was similar in distribution to that reported in global surveys of researchers interested in bioinformatics related training<sup>4</sup>. The primary audience was graduate students actively completing their respective degrees, followed by undergraduates and post docs. Two faculty members attended the workshop, demonstrating the broad distribution in academic standing. Participants were evenly split between local and virtual (Compute Canada Vidyo screencast) attendees. This even distribution allowed us to draw conclusions regarding the efficacy of the teaching model where both local and virtual audiences coexist. Additionally, we had even splits in both prior genomic analysis experience and familiarity with Linux and HPC systems. Lastly, we had an equal representation of both sexes, which is an important factor in bringing equality to the field of computational biology, a historically male-dominated field<sup>12</sup>.

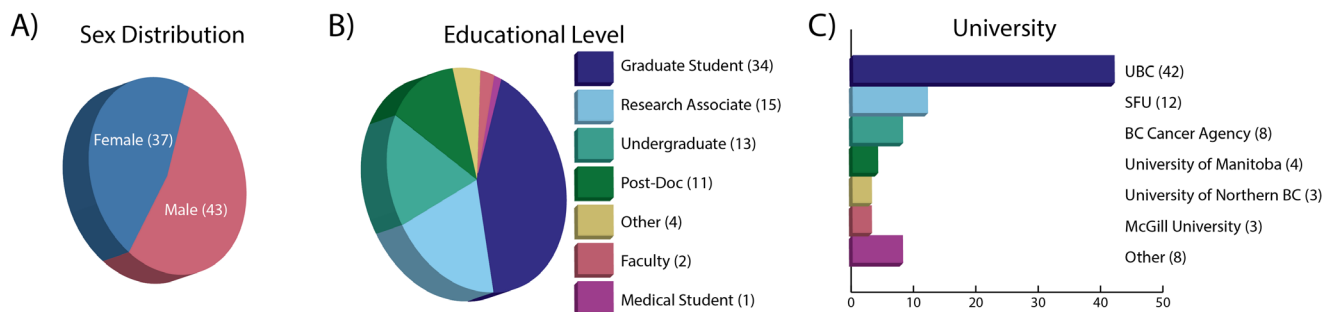
### Workshop efficacy

Workshop efficacy was determined based on three measures: 1) the number of total participants versus the number of participants that completed the exam; 2) the per-session attendance and their ability to complete the problem set; and 3) a set of questions given in a post-workshop anonymous survey. Regarding course completion, 80 attendees participated in the workshop, 59 (73.8%) of which completed the exam. The completion rate was slightly higher for the local audience (80%) than the virtual audience (66%), possibly due to the stimulation of in-person

collaboration between students which was noted as lacking in the post-workshop surveys from some virtual attendees (Figure 2A). When comparing between participants with prior and “zero” experience, the completion rates were surprisingly similar (Figure 2A), with both exceeding 70%. This represents a key success, as a key intention of the novel workshop design was engagement of researchers in the life sciences with diverse levels of experience.

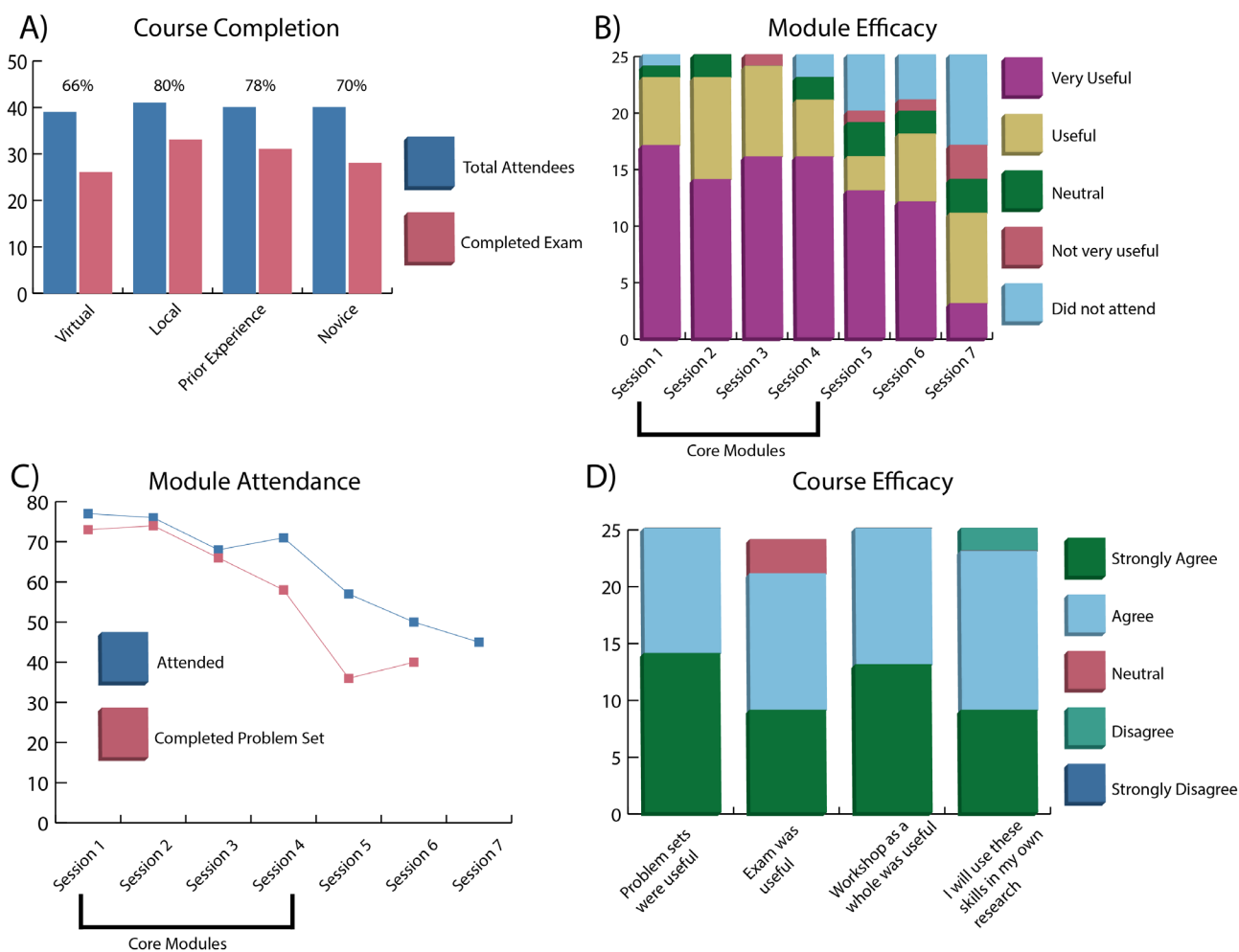
In analyzing the per-session attendance and problem set completion, we observed the efficacy of the module-based teaching methodology. While attendance was a measure of the interest in the subject matter, completion of the problem set identified student ability to effectively reuse the material taught during the first portion of the lesson on a unique data set. Sessions 1 through 3 had similar attendance and completion rates, and session 4 had a slightly lower completion rate since the problem set was also the mid-series exam (Figure 2B and C). The exam covered material from session 1–4, and was more in depth and difficult than preceding problem sets. Attendance for the last three sessions was optional and based on participant interest which was reflected in the lower number of attendees. Problem sets for sessions 5 and 6 were not required for attendance of the following sessions, and therefore had a lower completion rate. There was no problem set for Session 7.

Lastly, we analyzed the survey responses from 25 attendees for both per-session evaluations and aggregated opinions about the utility of the content. Sessions 1 through 4 are the core modules that guided attendees through an introduction to Linux and the command-line, interacting with the scheduler and queue, basics of next generation sequencing (NGS) short-read mapping and visualization, and variant calling. These sessions scored well in both efficacy and attendance, and received favorable reviews (Figure 2B). Session 5 through 7, advertised prior to the workshop as “optional”, went into more depth on subjects beyond the core NGS analysis including human genome variant interpretation and RNA-seq analysis. Despite an increase in the complexity of subject matter, the material was still judged to be accessible by the attendees. Some responses from the open commentary section were critical of the last few sessions, and suggested that those more complex topics need more time than allotted within a single 2-hour session. The progress assessments,



**Figure 1. Participant background.** Description of workshop attendees including **A)** distribution of sexes; **B)** educational level; and **C)** the university from which they participated.





**Figure 2. Workshop results.** A breakdown of the workshop results including **A)** distribution of course completion rates annotated by mode of attendance and prior experience; **B)** efficacy of each module based on survey responses; **C)** per-session attendance and problem set completion; and **D)** course efficacy breakdown including problem set and examination utility. Values were tallied based on attendance sign in sheets, user-submitted assignments, server workshop directories, and survey responses (Underlying data).

including both problem sets and exam, were deemed useful by the majority of those that responded to the survey and received positive commentary throughout the workshop (Figure 2D). Lastly, 24/25 survey completers found the workshop useful and indicated an intent to utilize the materials and skills they learned in their own research projects.

## Future perspectives

### Improvements

Constructive criticism of the workshop from attendees primarily focused on the final three modules, where advanced content was presented. In future iterations of the workshop, the initial 4 core modules will be taught as a set, while advanced topics will be offered separately. Those attendees that requested longer in-person sessions were referred to offerings from consortiums, such as Bioinformatics.ca, GOBLET, and

ELIXIR, each of which provides excellent advanced in-person all-day workshops. Future iterations will test whether the advanced topics are viable to be taught in the practical skills focused format highlighted in this report.

On a more granular level, the problem sets were a key part of many student's learning, but the system for delivering and grading those problem sets, via email and posting to locations on the shared server, was inadequate. Transitioning into a web-based platform (e.g. Moodle) for assignment delivery, completion, and grading, will allow for better feedback and communication regarding problem set questions.

### Following up with participants

*If you don't use it, you lose it.* After being introduced to new concepts, a new language, and new compute environments, it

is critical for continued practice to maintain and refine what you have learned. As a part of our workshop design, we can contact participants in the future to see how they progress in leveraging both the HPC resources and training received to process and analyze their own datasets. Our design of the workshop around the use of centralized compute resources allows us to engage with participants to see how effective the resources are for their research purposes. These long-term metrics will help inform retention and efficacy of materials, and identify gaps in training or resource needs that we can address through partnership with the academic HPC providers.

### Future workshops

We will deliver more workshops on both introductory and more advanced genomic analysis in the same modular format described above. By collecting similar efficacy metrics we can test how this workshop format performs with less introductory topics, and as part of a continued series. The overall goal is to establish training materials that can be delivered at time-of-need, that build strong foundations in genomic analysis utilizing the HPC systems. It is yet to be determined what the total attendee capacity is for this format, but our initial delivery reached 80 attendees, beyond the 30–40 person capacity of local workshops, and with the dual capacity of local and virtual delivery modes we anticipate audiences of over 100–200 participants.

### Conclusion

Training in bioinformatics and genomics will continue to be a critical component of the development of researchers in the biological sciences. Leaders in the research domain have proclaimed that acquiring computational analysis skills should be considered on par with learning the fundamentals of wet lab techniques<sup>13</sup>. Currently, the lack of formal education in genomics and bioinformatics analysis for life-science results in numerous researchers seeking out time-of-need training to answer their research questions<sup>2</sup>.

We have introduced a practical approach to the training of life scientists that focuses on getting researchers actively engaged with an academic HPC environment available to them for continued use beyond the confines of the workshop. The format incorporates both virtual and on-site participation, and the implementation successfully enables students with varying levels of experience to engage with the training at the relevant stages. The materials are available for re-use and are adaptable for use with most academic HPC architectures (see *Data availability*). As more centralized HPC systems become utilized and funded, we anticipate that this format of workshop will be invaluable in providing the foundation of training for researchers in the life sciences. Upon that foundation, researchers can further specialize their training needs and effectively participate in the high throughput technology revolution.

### Ethical approval

For the inclusion of attendee data and survey responses within this manuscript, we received ethics approval from the Department

of Medical Genetics at the University of British Columbia (application id H17-01298 – Student opinions of Genomic Bioinformatics Workshop). Approval was granted by Dr. Marco A. Marra, head of Medical Genetics at the University of British Columbia.

### Data availability

#### Underlying data

Zenodo: Extended data for Manuscript: Introduction to Genomic Analysis Workshop: A catalyst for engaging life-science researchers in high throughput analysis, <http://doi.org/10.5281/zenodo.3341800><sup>14</sup>.

This project contains the following underlying data:

- Attendee data redacted
- Attendee survey responses

#### Extended data

Zenodo: Extended data for Manuscript: Introduction to Genomic Analysis Workshop: A catalyst for engaging life-science researchers in high throughput analysis, <http://doi.org/10.5281/zenodo.3341800><sup>14</sup>.

This project contains the following extended data:

- **Supplementary Figure 1. Workshop Directory Structure.** The workshop directory structure used for each of the sessions.

Data are available under the terms of the [Creative Commons Attribution 4.0 International license \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/).

### Grant information

The work was supported by NSERC Discovery Grant (RGPIN-2017-06824) acquired by WWW, and PAR was supported by a BC Children's Hospital Research Institute Graduate Studentship award during this work.

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

### Acknowledgements

A previous version of this work is available from bioRxiv: <https://doi.org/10.1101/478024>.

We would also like to acknowledge Analise Hofmann for her assistance with ethics application and survey construction, Jana Makar of WestGrid for help organizing the workshop, Compute Canada for supporting the workshop with HPC server access and guest accounts, and the BC Children's Hospital Research Institute Evidence2Innovation Theme for on-site coordination and support.



## References

1. Brazas MD, Blackford S, Attwood TK: **Training: Plug gap in essential bioinformatics skills.** *Nature*. 2017; **544**(7649): 161.  
[PubMed Abstract](#) | [Publisher Full Text](#)
2. Attwood TK, Blackford S, Brazas MD, *et al.*: **A global perspective on evolving bioinformatics and data science training needs.** *Brief Bioinform.* 2019; **20**(2): 398–404.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Brazas MD, Ouellette BF: **Continuing Education Workshops in Bioinformatics Positively Impact Research and Careers.** *PLoS Comput Biol.* 2016; **12**(6): e1004916.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Ding Y, Wang M, He Y, *et al.*: **“Bioinformatics: introduction and methods,” a bilingual Massive Open Online Course (MOOC) as a new example for global bioinformatics education.** *PLoS Comput Biol.* 2014; **10**(12): e1003955.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Searls DB: **A new online computational biology curriculum.** *PLoS Comput Biol.* 2014; **10**(6): e1003662.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Searls DB: **Ten simple rules for online learning.** *PLoS Comput Biol.* 2012; **8**(9): e1002631.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Stefan MI, Gutlerner JL, Born RT, *et al.*: **The quantitative methods boot camp: teaching quantitative thinking and computing skills to graduate students in the life sciences.** *PLoS Comput Biol.* 2015; **11**(4): e1004208.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Carvalho BS, Rustici G: **The challenges of delivering bioinformatics training in the analysis of high-throughput data.** *Brief Bioinform.* 2013; **14**(5): 538–47.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Brazas MD, Ouellette BF: **Navigating the changing learning landscape: perspective from bioinformatics.ca.** *Brief Bioinform.* 2013; **14**(5): 556–62.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. Attwood TK, Bongcam-Rudloff E, Brazas ME, *et al.*: **Correction: GOBLET: The Global Organisation for Bioinformatics Learning, Education and Training.** *PLoS Comput Biol.* 2015; **11**(5): e1004281.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Schiffthaler B, Kostadima M; NGS Trainer Consortium, *et al.*: **Training in High-Throughput Sequencing: Common Guidelines to Enable Material Sharing, Dissemination, and Reusability.** *PLoS Comput Biol.* 2016; **12**(6): e1004937.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Bonham KS, Stefan MI: **Women are underrepresented in computational biology: An analysis of the scholarly literature in biology, computer science and computational biology.** *PLoS Comput Biol.* 2017; **13**(10): e1005134.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
13. Eddy SR: **“Antedisciplinary” science.** *PLoS Comput Biol.* 2005; **1**(1): e6.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Richmond PA, Wasserman W: **Extended data for Manuscript: Introduction to Genomic Analysis Workshop: A catalyst for engaging life-science researchers in high throughput analysis [Data set].** *Zenodo*. 2019.  
<http://www.doi.org/10.5281/zenodo.3341800>

# Open Peer Review

Current Peer Review Status:  

---

## Version 1

Reviewer Report 30 September 2019

<https://doi.org/10.5256/f1000research.21179.r53168>

© 2019 Springer M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



### Michael Springer

Department of Systems Biology, Harvard Medical School, Boston, MA, USA

There is a huge need for teaching resources in computational areas for life scientist. While there are a growing number of resources, it is far from saturated. The authors focus on an important sub-area, high performance computing with an eye towards genomic sequence analysis. This is a smart choice as sequencing data has become an increasingly common tool in the life science and students recognize the need to be able to hand such data.

The authors deliver a course consisting of seven workshops, four core and three optional. All the resources are shared online. The results of the course and feedback is shared making the strength and areas of potential improvements clear.

As someone who delivers similar content to graduate student, and who has been design a similar course, I found the online resources valuable and something that I could use to build on for my own class.

As such, I believe the article is ready for indexing.

**Is the topic of the opinion article discussed accurately in the context of the current literature?**

Yes

**Are all factual statements correct and adequately supported by citations?**

Yes

**Are arguments sufficiently supported by evidence from the published literature?**

Yes

**Are the conclusions drawn balanced and justified on the basis of the presented arguments?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Systems biology

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 24 September 2019

<https://doi.org/10.5256/f1000research.21179.r53165>

© 2019 Stefan M et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Richard Fitzpatrick**

Centre for Discovery Brain Sciences, University of Edinburgh, Edinburgh, UK

**Melanie I. Stefan** 

Centre for Discovery Brain Sciences, University of Edinburgh, Edinburgh, UK

Biomedical research relies increasingly on the use of high-performance computing (HPC), but there is a gap in effective and timely HPC-training for practising biomedical researchers. This article describes the implementation and deployment of a modular HPC training programme. It is useful in two ways: First, the authors show that a thoughtfully designed and delivered course can achieve the goal of providing useful training on HPC that life scientists can apply to their own research. Second, by making all course materials available under a Creative Commons license, the authors have created an invaluable resource for instructors, HPC facility coordinators and self-directed learners worldwide. We also appreciated that the author share the full (de-identified) data from the participant survey.

We do feel that the article category "Opinion Article" does not do this work justice, since it describes a rigorously designed educational intervention and collection of resources, as well as presenting evaluation data. We leave it up to the editors to find a better suited category.

**Minor points:**

- The abstract should include information about the outcomes of the course.
- The very first sentence ("The era of genomics and DNA sequencing is being rapidly incorporated into life science research fields ...") should be re-phrased for clarity.
- Figure 1: A reader's interpretation of a pie chart depends on their ability to interpret areas on the pie chart as representing the underlying numbers. Making the pie chart three-dimensional and tilting it does not add any information. Worse yet, it distorts the information available.
- Consider using "gender" instead of sex, since this is very probably the information that has been collected from participants.
- In describing how the course works, we would have liked a bit more information about the teaching assistants, in particular what exactly their role was, how many there were, and how they were trained and prepared for the task.

- The response rate on the post-course survey was not bad for this type of survey, but there is still a sizable proportion of non-respondents. There should be a short discussion on how this may affect survey outcomes. There is some indication (e.g. Bacon *et al.*, *Marketing Education Review*, 2016<sup>1</sup>) that responders are typically at the more "extreme" ends of the student satisfaction spectrum, i.e. students who really liked or really disliked the course are most likely to respond. As a consequence, a higher response rate would lower the scores for classes with high scores and raise the scores for classes with low scores. But on the other hand, see Nowell *et al.* (*International Review of Economics Education*, 2014<sup>2</sup>) who tried to estimate non-response bias in online teaching evaluations and found it negligibly small. It may be useful for the reader to have at least an acknowledgement of the problem in the article.
- Reference 10 is to an erratum on a paper, not the paper itself. The erratum here is important, because it clarifies the first author's last name, but we feel that the original paper should also be referenced.
- The github.io site for the course is really nice, but we notice that the links to final recordings are sometimes broken and just take the user back to the main site.

### References

1. Bacon D, Johnson C, Stewart K: Nonresponse Bias in Student Evaluations of Teaching. *Marketing Education Review*. 2016; **26** (2): 93-104 [Publisher Full Text](#)
2. Nowell C, Gale L, Kerkvliet J: Non-response bias in student evaluations of teaching. *International Review of Economics Education*. 2014; **17**: 30-38 [Publisher Full Text](#)

**Is the topic of the opinion article discussed accurately in the context of the current literature?**

Yes

**Are all factual statements correct and adequately supported by citations?**

Yes

**Are arguments sufficiently supported by evidence from the published literature?**

Yes

**Are the conclusions drawn balanced and justified on the basis of the presented arguments?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Computational neuroscience; design and evaluation quantitative and computational education for life scientists; learning analytics

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

**F1000Research**