Educating and engaging new communities of practice with high performance computing through the integration of teaching and research

Andrea Townsend-Nicholson

Structural & Molecular Biology (Division of Biosciences) and Institute of Structural & Molecular Biology, University College London, Darwin Building, Gower Street, London, UK, WC1E 6BT

ORCID: 0000-0002-7250-2208

Abstract

The identification of strategies by which to increase the representation of women and increase diversity in STEM fields (science, technology, engineering and mathematics), including medicine, has been a pressing matter for global agencies including the European Commission, UNESCO and numerous international Scientific Societies. In my role as UCL training lead for CompBioMed, a European Commission Horizon 2020-funded Centre of Excellence in Computational Biomedicine (compbiomed.eu), and as Head of Teaching for Molecular Biosciences at UCL from 2010 to 2019, I have integrated research and teaching to lead the development of high performance computing (HPC)-based education targeting medical students and undergraduate students studying biosciences in a way that is explicitly integrated into the existing university curriculum as a credit-bearing module. One version of the credit-bearing module has been specifically designed for medical students in their preclinical years of study and one of the unique features of the course is the integration of clinical and computational aspects, with students obtaining and processing clinical samples and then interrogating the results computationally using code that was ported to HPC at CompBioMed's HPC Facility core partners (EPCC (UK), SURFsara (Netherlands) and the Barcelona Supercomputing Centre (Spain)). Another version of the credit-bearing module has, over the course of this project, evolved into a replacement for the third year research project course for undergraduate biochemistry, biotechnology and molecular biology students, providing students with the opportunity to design and complete an entire specialist research project from the formulation of experimental hypotheses to the investigation of these hypotheses in a way that involves the integration of experimental and HPC-based computational methodologies. Since 2017-2018, these UCL modules have been successfully delivered to over 350 students – a cohort with a demographic of >50% female. CompBioMed's experience with these two university modules has enabled us to distil our methodology into an educational template that

can be delivered at other universities in Europe and worldwide. This educational approach to training enables new communities of practice to effectively engage with high performance computing and reveals a means by which to improve the under-representation of women in supercomputing.

Keywords

High performance computing, University education, Higher education, Medical student, Undergraduate, Experimental-Computational Workflow, Microbiome, Next Generation Sequencing, Computational Biology

Introduction

CompBioMed[1] (October 2016 - September 2019) and CompBioMed2 (October 2019 - September 2023) are European Commission Horizon 2020-funded Centres of Excellence focused on the use and development of computational methods for biomedical applications. Although computer-based modelling and simulation is well-established in the physical sciences and engineering, computational methods are only now reaching maturity in the biomedical domain, with predictive models of health and disease increasingly becoming relevant to clinical practice. A key, and futuristic, objective of the consortium of researchers in the Centre of Excellence (CompBioMed/CompBioMed2) has been to integrate computational methods, biomedical data and increasingly powerful supercomputers in order to build a Virtual Human – an *in silico* twin of an individual accurate at every scale from the letters of their DNA code to the architecture of their organs and the way their skeleton moves. Virtual Humans will provide healthcare practitioners with a personalised computer model of each patient, enabling them to predict the outcome of different clinical treatments and pick the most effective one for

that particular individual, and will allow individuals to evaluate the impact of different lifestyle choices on their personal health and wellbeing[2].

The programme of research for both CompBioMed and CompBioMed2 has focused on three healthcare-relevant aspects of the Virtual Human's biology: molecularly-based medicine, cardiovascular medicine and neuromusculoskeletal medicine. These are areas that are reasonably tractable computationally using the current petascale supercomputers and can be readily integrated into the entire Virtual Human when the new exascale machines become available[3]. In addition to its research programme, CompBioMed/CompBioMed2 also includes a training programme developed to enhance the computational biology community. This training programme~~, in the initial CompBioMed award,~~ was originally intended to provide targeted training to the academic, industrial and medical users in the computational biology community by delivering three major training events using CompBioMed's training allocation, which provides access to machines at the three CompBioMed/CompBioMed2 core partner HPC centres: EPCC (UK), SURFsara (The Netherlands), and the Spanish Supercomputing Network (RES, which includes the Barcelona Supercomputing Center (BSC). However, at the CompBioMed Kickoff Meeting in October 2016, a comment from one of the consortium members ("I wish I could teach this to medical students") raised an intriguing prospect: perhaps CompBioMed could amend its training programme to include specific high-performance computing (HPC)-based training for medical students, as part of their programme of taught study.

It was not at all clear whether there would be the support needed to foster this kind of innovation, either within the grant or within the higher education system. However, the benefits would be tremendous for both. An initiative of this kind would provide the opportunity to create

graduates fluent both in clinical practice and in the computational methodologies that are increasingly informing clinical decision-making. It would also allow clinicians to engage with state-of-the-art computational resources at an early stage in their careers – potentially early enough to be able to participate in the co-creation of future developments in computationally-informed healthcare. Although the education of medical students was an appealing prospect, there was no evidence of any national or international HPC facility machines that were routinely being used in medical student education. Was this an innovation that would "meet an unmet need"? How convergent was the CompBioMed wish to educate medical students in the use of HPC with what was happening in the higher education sector at that time?

In 2015, the seventh edition of PA Consulting Group's annual survey of UK vice-chancellors identified seven areas of innovation that could transform higher education, including the use of technology to transform learning experiences. In addition, three trends were identified by 25-30% of vice-chancellors as being "essential for survival", with a further 40-50% of vice-chancellors regarding them as "key to competitiveness". These three trends were: applications of technology to transform learning, closer integration of study with work and use of data analytics[4]. In 2016, the second Global Education Industry Summit (GEIS) brought together ministers of education and industry leaders to start a dialogue on the policies and strategies that could be implemented to foster innovation in education, with the qualitative changes, rather than quantitative expansion, deemed necessary to improve the efficiency and quality of and access to learning opportunities. The resulting report published by the Organisation for Economic Co-Operation and Development (OECD) highlighted that the role of digital skills and the education industries in the process of innovation was of particular relevance since the "digital divide" skills gap is becoming increasingly important as digital technologies advance and the employment and life opportunities for those with good digital skills are significantly

improved over those without these skills[5]. These findings revealed that technology-based innovations in education would be welcomed. Such innovations are transformative and provide the opportunity to educate the next generation in the use of digital methodologies and the application of these to new areas of study.

Introducing HPC to university students studying medicine and bioscience would meet one of the seven areas of innovation, it would address all three trends identified by the surveyed UK vice-chancellors as being of such great importance and it would provide an unusual means of shrinking the digital skills gap identified by the OECD. With respect to the location where such an initiative would be undertaken, it was clear that University College London (UCL) was the perfect place in which to trial the CompBioMed Education Programme. Not only was UCL the lead institution of CompBioMed, but the founding principles of academic excellence and research aimed at addressing real-world problems have been definitional features of degree study at UCL since its establishment in 1826 and research-based education is at the core of UCL's institutional and education strategies[6,7]. This has most recently been encapsulated in the Connected Curriculum, a framework for research-based education that aims to allow students to engage in research and enquiry, be exposed to the state-of-the-art technology and knowledge and work with staff and other students to develop the essential skills needed to become autonomous thinkers[8].

This paper describes the introduction of the CompBioMed Education Programme into the taught curriculum of medical students and of undergraduate bioscience students studying biochemistry, biotechnology and molecular biology (Molecular Biosciences) at UCL with access to state-of-the-art HPC provided through CompBioMed's training allocation on HPC machines in the UK, The Netherlands and Spain. It provides quantitative and qualitative

indicators of the success of the initiative together with lessons learned, for those wishing to adopt this methodology.

Methods

The CompBioMed Education Programme: Course Development

In developing the CompBioMed Education Programme, it was of interest to identify a workflow that involved the computational analysis of data obtained from experiments conducted by the students, themselves, in order to reinforce the integration of experimental and computational methodologies. In addition, it was desirable to find a computational element that would not be computationally expensive, to accommodate student innovation and learning within both the learning objectives of the taught course and the CompBioMed HPC training allocation. Ideally, the content of the course would be of interest to both medical students and bioscience students.

A data analysis and advanced practical research skills module (BIOC3101) had been launched for MSci Biochemistry students in 2015-2016[9]. This module was templated on the Earth Microbiome Project[10] and used Next Generation DNA Sequencing (NGS) to identify the microbial populations present in samples of soil collected by the students. Students performed DNA sequence analysis of the raw DNA sequencing data using the Python-based open source bioinformatics pipeline QIIME1[11] (Quantitative Insights Into Microbial Ecology) on their laptops. A key element of the practical was the computational analysis and computer programming skills that students learned for the DNA sequence analysis. Although performed on standalone machines, there was no reason why the computational elements of BIO3101 could not be conducted using HPC, making this course a suitable base from which to develop the CompBioMed Education Programme.

In 2017-2018, following feedback from staff, students and the external examiner for the Molecular Biosciences degree programmes (MSci Biochemistry, BSc Biochemistry, BSc Biotechnology and BSc Molecular Biology), BIOC3101 was modified substantially, turning it from a 15 credit advanced practical for Year 3 MSci Biochemistry students into a 30 credit specialist research project, BIOC3301, that was taken by all Year 3 Molecular Biosciences students. As part of the modification, the computational elements of the course were expanded significantly and conducted on HPC for the first time, using the CompBioMed HPC resources. Students were provided with the opportunity to develop their own hypotheses, the number of NGS barcodes was expanded three-fold and the students were given monitored but essentially unlimited access to HPC to perform their analyses.

Also in 2017-2018, BIOC3101 was modified to run as a single block (8 weeks at 3 hours per week) Student Selected Component (SSC) for Year 1 and Year 2 pre-clinical medical students as part of UCL's medical school curriculum. Two key modifications were made for this: 1) participants purified and analysed genomic DNA from their skin bacteria (right palm), following the protocol outlined in the Human Microbiome Project[12]; and, 2) the DNA sequence analysis was performed using a simplified QIIME protocol (greatly reduced statistical analyses of the results compared with BIOC3301) run on CompBioMed's HPC resources. From 2018-2019 onwards, following feedback from students, the SSC *From Skin to Metagenomics: You and Your Microbiome* (SSC334) has been run as a double block SSC (16 weeks at 3 hours per week) for Year 1 medical students. In 2019-2020, BIOC3301 (now renamed BIOC0023) deprecated QIIME1 and began using QIIME2[13] for microbiome analyses.

This first introduction to HPC for medical and undergraduate students highlighted the need to provide students with sufficient time to assimilate the computational skills required to conduct the computational elements of the course. This skills gap was rectified by delivering the required computational training to over 300 students in total by the end of the academic year 2017-2018, through the Year 3 module, BIOC3301, and via the Year 1 and Year 2 Molecular Biosciences Key Skills modules that ran in June 2018. The computational training was provided in three hands-on workshops (Introduction to Git and GitHub, Introduction to the Linux Command Line and Python workshops) and ensured that all undergraduate Molecular Bioscience students possessed basic computational skills from their first year of study. With the basic provision covered from Year 1 of their degrees, this training allowed, from 2018-2019, the development and expansion of the computational biology provided in Years 2 and 3 of their degree programmes, an ongoing project that continues to evolve as the computational fluency of the students increases. In 2019-2020, students were asked to process a medium-sized microbiome dataset as a data analysis exercise worth 10% of their final mark on BIOC0023. This provided them with the opportunity to acquire experience with QIIME analyses prior to analysing their own data.

In addition to its delivery as part of the taught curriculum at UCL, the computational element of BIOC0023/SSC334 has been delivered, since 2018, as part of the PRACE (Partnership for Advanced Computing in Europe; prace-ri.eu) training programme, through the annual PATC (PRACE Advanced Training Course) short course on HPC-based computational biomedicine – a Winter School run at the Barcelona Supercomputing Center every February. From 2020, the computational element of SSC334 will be offered through the Biochemical Society (biochem.org) as a training event, “How to use supercomputers for microbiome analyses”, for undergraduate, graduate and postgraduate researchers in academia and industry. These courses

are not particularly expensive to run and fit within the normal teaching hours on a module. Additional demonstrators are required for larger class sizes, with an ideal ratio of approximately one demonstrator per 10 students unless small group, instead of whole class, workshops are provided. These courses do require, however, that the staff and demonstrators (who are experimental bioscientists) have appropriate computational expertise. This was provided through skills development programmes provided for staff to maintain and develop their professional competencies and through independent study.

HPC Machines used in the CompBioMed Education Programme

A machine at each of the three CompBioMed HPC centres was chosen to run QIIME: Cirrus, a Tier 2 national facility machine at EPCC, Cartesius, a Tier 1 machine at SURFsara and MareNostrum, the Tier 0 machine at the BSC.

Cirrus is a state-of-the art SGI ICE XA system with 280 compute nodes, utilising a superfast Infiniband interconnect. There are 36 cores per node (18 cores per Intel Xeon “Broadwell” processor) providing 10,080 cores in total. Hyperthreading is enabled on each node providing a total of 72 threads per node. Each node has 256GB RAM. Three login nodes, with similar hardware and software environment to the compute nodes, are provided for general use. Local Lustre storage is provided by a single Lustre filesystem with 406 TB of disk space and users have access to EPCC's considerable data storage and archiving services. Further description of the machine is available at http://www.cirrus.ac.uk/about/hardware.html.

Cartesius is a bullx system extended with one Bull sequana cell. It is a clustered SMP (Symmetric Multiprocessing) system built by Atos/Bull and consists of a large number of batch nodes and a small number of special purpose nodes, providing 47,776 cores + 132 GPUs and

180 TB memory (CPU + GPGPU + HBM). Cartesius has two interactive front end nodes that provide the ssh login service and are meant for interactive work and five service nodes for data staging. The network interconnect is organised in islands. Every bullx node has a Mellanox ConnectX-3 or Connect-IB (Haswell thin nodes) InfiniBand adapter. Every sequana node has a Mellanox ConnectX-4 InfiniBand adapter. The GPGPU nodes have two ConnectX-3 InfiniBand adapters: one per GPGPU. Cartesius uses a ~ 7.7 PB Lustre (parallel) file system for scratch and project space. Further description of the machine is available at https://userinfo.surfsara.nl/systems/cartesius/description.

The latest version of MareNostrum, MareNostrum 4 (2017), is a supercomputer based on Intel Xeon Platinum processors, Lenovo SD530 Compute Racks, a Linux Operating System and an Intel Omni-Path interconnection. The calculation capacity is distributed in two independent blocks: a general purpose block and an emerging technologies block. The general purpose block has 48 racks with 3,456 nodes - a total of 165,888 Intel Xeon Platinum cores and 384.75 TB of total memory. The compute nodes are interconnected through a high-speed Intel Omni-Path (OPA) interconnection network. The emerging technologies block is divided into three clusters, each with a technology currently being developed to accelerate the next generation of pre-exascale supercomputers, specifically: a cluster consisting of IBM POWER9 processors and NVIDIA Volta GPUs, a cluster made up of AMD Rome processors and AMD Radeon Instinct MI50 and a cluster formed of 64 bit ARMv8 processors in a prototype machine. MareNostrum has a disk storage capacity of 14 Petabytes. Further description of the machine is available at https://www.bsc.es/marenostrum/marenostrum.

Installation of QIIME1 on Cirrus

QIIME1 is an application typically used on desktop machines by biomedical researchers that was initially installed and tested on Cirrus before being ported to Cartesius and MareNostrum. QIIME1 consists of native Python 2 code as well as external applications. QIIME1 is installed and run in a virtual environment created by conda and the initial installation of QIIME1 was performed on Cirrus using miniconda3. This installation broke the Cirrus accounts (stdout and stderr did not appear and remote scp failed). The possibility of a central installation of QIIME1 on Cirrus was investigated but deemed inappropriate, however, miniconda2 and 3 were both centrally installed and it was determined that QIIME1 worked on Cirrus when miniconda2 was used. Students were subsequently instructed to load miniconda2, create their QIIME environment and install QIIME, then activate their QIIME environment and test their QIIME installation. With QIIME successfully installed on Cirrus and the relevant timings obtained, the code was given to SURFsara and BSC to benchmark on Cartesius and MareNostrum, respectively.

Running QIIME on HPC

EPCC and UCL developed, and has presented annually at UCL, a training workshop on Unix and then HPC, aimed at a target audience with zero or negiligible experience of programming and with no prior experience using HPC. This workshop introduced medical students and molecular bioscience students to the Unix command line, and thereafter HPC in general, with overarching concepts, and then how to run QIIME1 using Cirrus in particular, with a basic explanation of job queues, which are necessary to distribute load and harness the compute power appropriately. This workshop also introduced students to vim, a command line editor used to personalise the three job “scripts” supplied by UCL they then employ on Cirrus. Finally, the students were introduced to running QIIME1’s Python routines in parallel.

The computational workflow in QIIME employs three routines to be run one after the other and these were run as separate text files, namely the job scripts, which inform the batch system of how many nodes/cores they wish to employ, the maximum length of time the job should run for, the project the time should be charged to and how to run the job. These three scripts, specifically, involve: 1) demultiplexing and quality filtering of the NGS fastq files, 2) picking operational taxonomic units and 3) conducting core diversity analyses, which calculates several diversity measures as well as taxa tables. The first script is run in serial and employs just one single core; the second and third currently run on Cirrus using the most efficient core count for this amount of data for these particular QIIME1 python routines, namely 16 cores. Bioscientists at UCL went one step further than the Medics, in that they determined what this most efficient core count is. As such, the whole run takes about 78 minutes and can be completed within 20 core hours. This was considered to be sufficiently inexpensive, in terms of compute, that it would allow students the opportunity to achieve the objectives and explore the system whilst staying within the CompBioMed training allocation budget.

Initial benchmarking on Cirrus indicated that students would need a minimum of 6 hours for one complete execution. Subsequently, a condition set was identified that required 78 minutes for one complete execution; this allocation provided a minimum viable baseline for the required computing with minimal provision for more sophisticated analyses. It was agreed to give students more time to allow for multiple executions. An allocation of 50 core hours per medical student and 100 core hours per bioscience student was provided, together with 60GB storage per student for data and scripts. Measures were put in place to prevent the total allocation being used in error: a reserve was set up on the EPCC SAFE, a website designed to support accounting reporting, usage monitoring and resource management on Cirrus, and sub-groups were established. Students were given individual accounts on the SAFE and on Cirrus

and were provided with guidance on how much storage and how many core hours they should be expecting to use to log in, download and build QIIME, load up the NGS dataset, develop and test the QIIME scripts, execute the final QIIME script on the data and download the analysis results. SAFE reports were used to examine how students were using the project budget and activity on Cirrus was monitored closely to ensure that anomalies and errors resulting in jobs running for too long were able to be flagged. Students were asked to email for assistance if they exceeded either the expected disk use or the expected wallclock time. An initial helpdesk was set up at UCL as the first point of call for student queries, with the UCL helpdesk triaging and forwarding issues to the Cirrus helpdesk as appropriate. The Cirrus helpdesk operators were asked to forward any stray QIIME queries to the CompBioMed contact at EPCC. The compute and storage allocations were extended further when it became clear that a number of students were exploring the system and trying out different conditions and parameters for their analyses.

Provisions were made for 120 students in academic session 2017-2018 with approximately 20 medical students and approximately 100 bioscience students. In 2017-2018, all students ran on Cirrus. In 2018-2019, all medical students and half of the bioscience students ran on Cirrus and the rest of the bioscience students ran on Cartesius. In 2019-2020, all students ran on Cartesius. MareNostrum has been used solely to run QIIME for the PRACE PATC course and, from 2020, will be used for the Biochemical Society training event.

Results and Discussion

The successful introduction of medical and Molecular Biosciences students to HPC

The CompBioMed Education Programme successfully achieved its objective to provide medical students and undergraduate bioscience students studying biochemistry, biotechnology

and molecular biology (Molecular Biosciences) at UCL with access to state-of-the-art HPC. As shown in Table 1, in the past three years, 347 undergraduate students have accessed CompBioMed HPC as part of their taught study at UCL. 80 of these were pre-clinical medical students and the remaining 267 were Molecular Biosciences students. There was a 100% success rate for students using HPC as part of their degree and each student was able to use HPC to conduct the computational analyses that were a required component of their taught course. Table 1 also shows that the students' consumption of HPC resource exceeded the initial allocation (50 core hours per medical student and 100 core hours per bioscience student) provided by 1.66 fold in 2017-2018 and by 4.53-fold in 2018-2019. The modules are currently running and the final core hour consumption for 2019-2020 has not yet been determined but is anticipated to exceed 20-fold.

Academic year	Medical students	Molecular Biosciences students	Total number of students	Core hours consumed	Core hours allocated	Fold difference (consumed/ allocated)
2015-2016	0	20	20	0 (local)	0	-
2016-2017	0	29	29	0 (local)	0	-
2017-2018	40	85	125	17,452	10,500	1.66
2018-2019	20	99	119	49,394	10,900	4.53
2019-2020	20	83	103	97,919*	10,830	9.04*

Table 1
QIIME analyses conducted by UCL undergraduate students using CompBioMed HPC. *Consumption and fold difference are based solely on HPC use up to the end Term 1 of the current academic year, which has not yet concluded.

The increase in fold difference between 2017-2018, 2018-2019 and 2019-2020 can be attributed to the changes made to the computational training provided during these three years. The vast majority of Molecular Biosciences students using HPC in 2017-2018 had not received previous instruction in computational methodologies as part of their degree programme and, therefore, were not equipped to make extensive use of such unfamiliar methodologies. By

summer 2018, the students in Years 1, 2 and 3 had all received the same level of instruction. The students using HPC in 2018-2019 were, therefore, more familiar with the computational methodology and, on average, appeared comfortable performing more extensive analyses than in the previous year, including analyses that involved increases in consumption arising from a rerunning of the data, such as different ways of binning the metadata or comparing results obtained with different 16S rRNA databases for taxonomic assignments. The inclusion, in 2019-2020, of a data analysis exercise that involved conducting a full QIIME analysis of a medium-sized dataset has increased consumption significantly and it will be of interest to see how this enhanced familiarity with the computational methodology will be reflected in the types of analyses conducted on the students' own datasets.

It is also interesting to look at the significant users of the HPC resources that were provided. Figure 1 shows the top 10 users in each of the three years that UCL students had access to CompBioMed HPC. In 2017-2018, the top user consumed 2435 core hours, a 24-fold increase in consumption over allocation per student and approximately 23% of the entire allocation for 2017-2018. In 2018-2019, the top user consumed 9090 core hours on Cartesius, a 91-fold increase in consumption over allocation and approximately 83% of the initial 2018-2019 allocation. The top user on Cirrus in 2018-2019 was number 7 overall in the top 10, consuming 1894 core hours (19-fold increase in consumption and 17% of the initial 2018-2019 allocation). Together, these two individuals, the top user on Cartesius and the top user on Cirrus, consumed the entire allocation initially provided for 2018-2019. The monitoring put in place to provide elasticity of the allocations in response to demand worked well and it was clear that these users were applying the additional resource to QIIME analyses, using it to interrogate and analyse different conditions. These include the addition of new data sets or the modification of metadata (for example, amending the way in which time course data are grouped) or any other

change that requires the computational workflow to be run with new parameters. The monitoring of the use of the HPC was sufficiently robust that it flagged up a slow running and oddly named batch script (KFC_fried_chicken) that turned out to be a student who was trying something very sophisticated in terms of analysis. This student was subsequently able to resolve their difficulties with some assistance on parameters. Interestingly, this enabled the identification of an issue with misuse of the continuation character that was leading to excessive consumption of HPC resource by a number of students (see Lessons Learned). It is worth pointing out that all of the Top 10 users have been Year 3 Molecular Biosciences students. At 48 hours in total, the double-block SSC course is much shorter than BIOC0023, which has 300 hours allocated to it, so there is less opportunity to provide training and experience in HPC to medical students than for Molecular Biosciences students. In addition, it is not obvious how to introduce key skills to medical students in the way that was achieved for the Molecular Biosciences students in 2018. A means of providing medical students with a programme to enhance their computational skills is currently being sought, as the outcomes of the CompBioMed "QIIME on HPC" programme suggest that the volume and sophistication of HPC use increases when students have greater familiarity with the methodology and are given more extensive opportunities to access the technology.

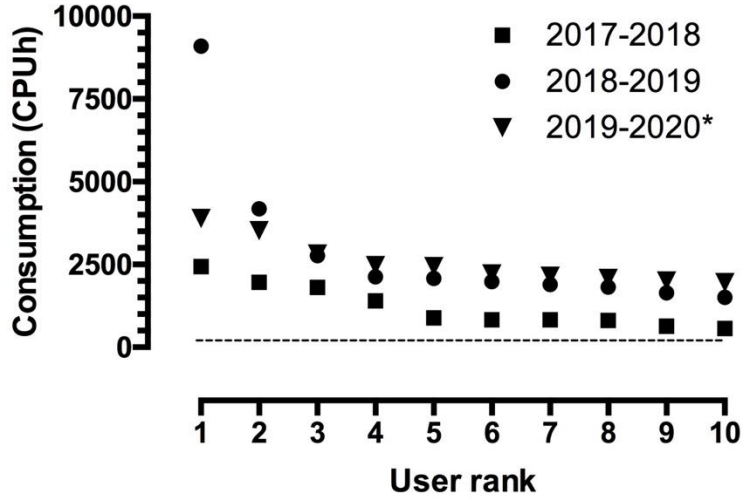

Figure 1
The "Top 10" UCL student users on CompBioMed HPC in each of the three years of the CompBioMed Education Programme. The horizontal dashed line indicates the 20 core hours needed for one complete QIIME execution of the dataset. *Data based solely on HPC use up to the end Term 1 as the 2019-2020 academic year has not yet concluded.

The successful creation of a new community of practice interested in using HPC

A significant number of medical and bioscience undergraduate students at UCL have been introduced to HPC and educated in its use over the past three years. These students lie outside the typical demographic for HPC users and comprise a new community of practice that is starting to engage with computational technology of a scope and at a scale previously unfamiliar to them. In 2016, when this project commenced, the field-specific use of CompBioMed HPC was examined. The results are shown in Table 2. Cirrus came online months prior to the start of the CompBioMed Education Programme, so the 2016-2017 statistics are shown for Archer, the UK's Tier 1 machine at EPCC, instead. Although the relative proportions vary it is clear that, overall, the life and medical sciences were not making as much use of these resources as the other fields listed. Table 3 reveals that the use of Archer does not correlate with the percentage of first degrees or postgraduate Higher Education qualifications obtained by subject in the United Kingdom[14] and that there is a predominance of HPC use by the physical sciences. Medical use of Archer is exceptionally low given the number of first degree and postgraduate degrees awarded in this field and is in the area of medical physics. The biological use of Archer is in the area of biomolecular simulation. It will be interesting to evaluate the Cirrus usage statistics over time to identify whether the usage profile is similar to that of Archer or if the machine is attracting users from other areas of medicine and biology, such as genomics and microbiome analyses.

Field	Archer	Cartesius	RES/BSC
Life and Health	10	12	20
Astronomy, Space and Earth Sciences	10	22	23

Chemistry, Material Science and Technology	60	39	30
Math, Physics and Engineering	19	26	27
Other	1	1	0

Table 2
Field-specific use (as % of total machine use) of CompBioMed HPC at EPCC (Archer), SURFsara (Cartesius) and the Spanish Supercomputing Network (RES), which includes the Barcelona Supercomputing Center (BSC), in 2016-2017. Data were provided by each of the HPC centres.

Field	HE First Degree (%)	HE Postgraduate Degree (%)	Archer Use (%)
Medical	33	24	0.2
Biological	26	22	9.8
Physical	15	25	60.7
Mathematical	5	4	9.5
Engineering	15	20	18.6
Other	6	4	1.2
Total	100	100	100

Table 3
Percentage of Higher Education degrees awarded in 2016-2017 by subject compared with use of Archer in 2016-2017. Medical = "Medicine and dentistry" combined with "Subjects allied to medicine"; Biological = "Biological sciences" combined with "Veterinary science"; Physical = "Physical sciences" and "Computational science"; Mathematical = "Mathematical sciences"; and, Other = "Agriculture and related subjects" and "Architecture, building and planning". All degree data provided by the Higher Education Statistics Agency (HESA)[14]. All Archer data provided by EPCC; Medical use = medical physics, Biological = biomolecular simulation.

Diversity and HPC

The Women in HPC programme (WHPC; https://womeninhpc.org) was initiated by EPCC in 2013 and is the only international organisation working to improve equality, diversity and inclusion in High Performance Computing. According to WHPC, women form less than 17% of the HPC workforce, although the exact numbers are difficult to ascertain. A recent survey of the scholarly literature in biology, computer science and computational biology from 1997 to 2014 concludes that there are fewer women authors on research papers in computational biology as compared to biology in general[15]. HESA data, shown in Figures 2 and 3, reveal that

whilst the percentage of female recipients of first and postgraduate Computer Sciences degrees, respectively, is below the average for all sciences, the percentage of female recipients of first and postgraduate degrees in medicine and in the biological sciences is not. The underrepresentation of women in HPC may arise from subject-specific issues of access to HPC or the absence of HPC-based analyses in particular subjects, as the fields with the greatest representation of women are also those with the least use of HPC. By supporting initiatives that provide non-standard users of HPC with the ability to engage with the technology and continue to have access to it, it should be possible to rapidly and irreversibly improve the diversity of HPC users. The CompBioMed programme is ongoing and the diversity of UCL's medical student and Molecular Biosciences undergraduates using CompBioMed HPC reflects the demographics of UCL's intake in these fields, which is >50% female. Whether through education formally embedded in the curriculum for medical and bioscience undergraduates or through workshops designed to redress the lack of formal education in HPC-based computational analyses for the life sciences[16], targeted efforts to improve access to HPC for all practitioners in these subject areas results in a quantifiable improvement in the percentage of women using HPC.

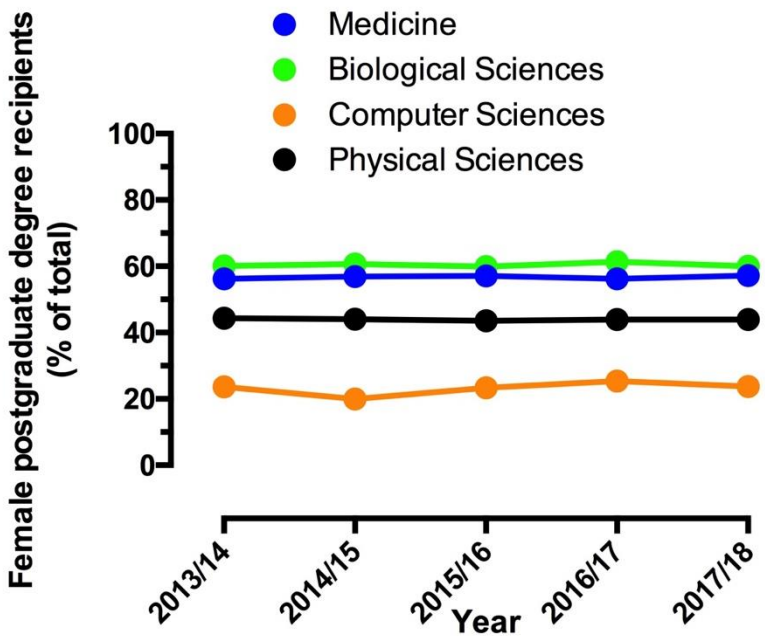

Figure 2
The percentage of female recipients of first (bachelor-level) degrees by subject.

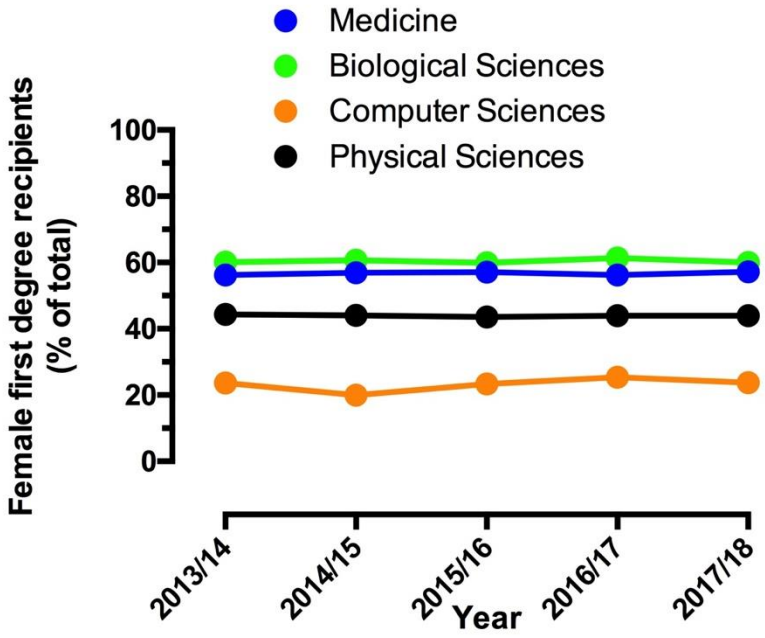

Figure 3
The percentage of female recipients of postgraduate (masters and doctoral level) degrees by subject.

Lessons Learned

Successes

1. All students completed the course, for a 100% success rate

2. A 100% questionnaire response rate was obtained for the medical students by asking them to include in their assessed report a personal reflection on the three things they most enjoyed and the three things they found most difficult and how could the latter have been made easier for them
3. At all three Centres, individual accounts were created for students and the Centres ported and profiled the QIIME code before the students arrived.
4. All students logged in, edited scripts, ran jobs and retrieved data without prior HPC experience
5. Some of the students embraced the significantly different environment and made substantial use of the resource
6. Students overcame differences in running in different environments at UCL (on multiple clusters, their own laptops, etc.)
7. Some students were logging onto Cirrus to run jobs from their mobile phones
8. Trainers and HPC centres debugged and solved problems on the fly quite successfully
9. Trainers overcame the differences between running at the different HPC centres
10. The Cirrus fail2ban settings were relaxed for all users of Cirrus, thanks to input from these courses

Challenges

1. More assistants were needed at the computational workshops: 1 assistant per 4-5 students would be ideal
2. The initial very large disk space allocation was still not enough for excited students
3. Student's Cirrus accounts took longer than expected to create. This was resolved by using Cirrus for the Linux course, run before the HPC course. The default fail2ban settings were found to be too severe in classrooms where multiple terminals share the

same IP address: urgent IP Whitelisting was employed as a work-around. Jobs were hanging due to their batch script bug: rather than seeking help, the students tried increasing wall clock time AND core count and burned a lot of the budget. The HPC course was updated to illuminate common script errors.

4. Bugs in scripts were discovered, e.g. in TMPDIR, when running at a previously untested scale of usage
5. Other issues arising from a lack of familiarity with Linux, including filepath errors, being in the wrong directory, pulling files from Cirrus to a local machine from Cirrus rather than the local machine etc. This was addressed by running a Linux course before the HPC course.
6. A general bewilderment at the need to work with the command line instead of something more visual, with students wondering "why things hadn't moved on", despite the fact that this topic was addressed in the HPC course.
7. Issues arising from the lack of familiarity with batch job system
8. A number of students provided an email address that they didn't read so they didn't see emails describing outages and maintenance periods. The HPC course was updated to make this requirement clear.
9. The use of multiple HPC facilities presented a challenge for trainers as each had different access mechanisms, different OS, homogeneous vs heterogeneous login nodes inside single systems, different batch systems, different ways of moving data around difficult, different usage policies, different facility schedules for upgrades/downtime, testing was performed on different OSs, etc.
10. Students must have some command-line experience (understanding of directories/folders and basic Linux (ls, pwd, etc) plus vi/vim as an editor, which is why we ran a Linux course before the HPC course.

Unanticipated wins

1. Student engagement with HPC went well beyond what was expected with a number of students making exceptional and significant use of the resource
2. The SSC course has engaged medical students in microbiome analyses and HPC and a group of Year 1 medical students from 2017-2018 proposed and are currently conducting an extracurricular microbiome project. Year 1 medical students from 2018-2019 have also proposed an extracurricular microbiome project that they wish to conduct
3. Funding was obtained from UCL's School of Life and Medical Sciences Capital Equipment fund to purchase two Illumina iSeq-100 NGS DNA sequencers for use in teaching, providing students with the opportunity to have hands on experience generating the data they will be analysing on HPC
4. Continuing medical student engagement with HPC, as one of the 2017-2018 Year 1 medical students has enrolled on the Mathematics, Computers and Medicine iBSc programme.
5. A BIOC0023 graduate contacted EPCC to ask if they could continue to have access to Cirrus for molecular dynamics simulations they wanted to conduct as part of their PhD studies.

Student Feedback

The increase in fold difference of compute use between 2017-2018, 2018-2019 and 2019-2020 can be directly compared with student feedback on the modules. Student evaluation questionnaire responses with quantitative data and qualitative data are available for 2017-2018 (n=85 students; 48.2% response rate) and for 2018-2019 (n=100 students; 16% response rate).

Student ratings for "Overall, I am satisfied with the quality of the module" rose from 4.9% definitely satisfied and 12.2% mostly satisfied in 2017-2018 to 28.6% and 71.4%, respectively, in 2018-2019. Similarly, student ratings for "The module is well organised and is running smoothly" rose from 2.2% of students definitely satisfied and 11.1% of students mostly satisfied in 2017-2018 to 35.7% and 64.3%, respectively, in 2018-2019. The response rates are consistent with previous observations that students are more likely to complete module questionnaires when dissatisfied. Many of the students in 2017-2018 were acquiring computational skills for the first time, whereas the students in 2018-2019 were able to build on the computational expertise they had obtained in the previous year's key skills module. Representative qualitative feedback is shown below.

Examples of student feedback

"…I have never had formal education or training in computer science or coding, and so I feel very proud that I now have: a basic understanding of Linux, the tools I need to improve my understanding of it and my ability at using it, and (most importantly) the desire to further develop this skill."

"…perhaps most enjoyable thing was the very exposure to high-performance computing. The field of computing is so fundamental to our modern, technological world in every facet of society. I believe it is a field to which we are not exposed enough in both secondary and higher education, and so I took great interest in this aspect of the study…..On the flipside, high-performance computing…was extremely tedious. Though we were given step-by-step instructions, it was all too easy to make a typing error which would impede the process of graphing our results. Unfortunately, as I have come to learn, this is the nature of writing code, and nothing can be done to change this."

Future Work

CompBioMed researchers have now participated in the education of ~400 university students over the past three years. These students, who are now HPC-enabled and keen to continue to take advantage of this technology, are a new community of practice in HPC that has been established by bringing the technology closer to younger practitioners. Continuing to introduce university students to HPC is necessary to grow and support this new community of HPC users. The CompBioMed Education Programme has recently been enhanced with the renewal of the award and from 2021, CompBioMed2 will be expanding beyond UCL to provide the medical student education at the University of Oxford, the University of Sheffield, the University of Amsterdam and the University Pompeu Fabra in Barcelona. CompBioMed2 researchers also keen to provide assistance to any university wishing to adopt this programme at their own institution.

Conclusions

The key to successful engagement with HPC for non-traditional users is to start early to create fluency, expectation and to provide the opportunity to support innovation and new discoveries. Embedding HPC in the university curriculum in different programmes and over different years of study allows the building of a relationship with learners on the timeline needed to foster mentoring and coaching for those wishing to explore the area further and facilitates diversity. The integration of teaching with research for medical students creates a motivating environment for the clinicians of the future to engage with advanced technologies such as HPC and provides the opportunity to nurture new ideas and new technologies in personalised medicine and healthcare. The integration of experimental and computational workflows allows the creation of biomedical researchers with dual fluency and expertise. The inclusion of HPC

training in biomedical degrees improves diversity, increasing the representation of women using HPC resources.

After three years of running this innovative approach to educating students in the use of HPC and computational methodologies, a new generation of biomedical scientists and medical students who have used supercomputing as part of their taught degree are now graduating from UCL and will be wanting to incorporate this technology into their future endeavours. It is possible to achieve this outcome at universities worldwide.

Acknowledgments

I am exceptionally privileged to have been in a position to work with a great team at UCL and a great team at CompBioMed and to have had such fabulous students, all of whom graciously received these courses in the spirit in which they were intended and actively helped to improve them. The integration of HPC into the higher education curriculum could not have been achieved without any of these people!

At University College London, I would like to thank the following: Michael Baron, the module organizer for BIOC3101/BIOC3301/BIOC0023 who provided stellar technical support for both the computational and experimental parts of the course and wrote the scripts and documentation for running QIIME on HPC; Amanda Cain, the deputy head of teaching for Molecular Biosciences and Jim Naismith, our External Examiner for BIOC0023 and the MSci Biochemistry degree programme, who provided invaluable support and advice on how to improve the content and delivery of BIOC0023 as part of the development of the Molecular Biosciences curriculum; Andrew Martin who wrote the self-paced computational tutorials on the Linux command line, Git and GitHub; Greg Campbell and Paul Dilworth who approved

the establishment of the medical student SSC; and, Peter Coveney, the coordinator of CompBioMed and CompBioMed2 for his strong and continuing support for this work.

In CompBioMed, I would like to thank all of my colleagues for their advice and support. At the University of Edinburgh (EPCC), Gavin Pringle installed, tested and benchmarked QIIME on Cirrus and gave the code to SURFsara and BSC to install and benchmark, and also wrote and delivered the EPCC "Introduction to HPC for Medics and Bioscience students" workshop at UCL, approved student accounts on Cirrus and was the primary point of contact between the EPCC and UCL; Terry Sloan helped to coordinate CompBioMed-UCL communications and reviewed CompBioMed documentation for this project. Marco Verdicchio at SURFsara installed QIIME1 and QIIME2 on Cartesius, helped get QIIME2 running on Cartesius from Jupyter notebooks after QIIME1 was deprecated and provided help and advice for issues students encountered. Mariano Vazquez, Ruth Aris and Alfonso Santiago at the Barcelona Supercomputing Center helped with the installation of QIIME1 on Nord3 and MareNostrum 4 and provided support during the PATC training course.

All of the training materials used on BIOC0023, SSC334 and the PATC training course on HPC-based computational biomedicine are available on the CompBioMed training portal at https://www.compbiomed.eu/training-3/.

Funding

I am grateful to the European Commission for Horizon 2020 funding support for the CompBioMed (675451) and CompBioMed2 Centres of Excellence (823712). This work used the Cirrus UK National Tier-2 HPC Service at EPCC (http://www.cirrus.ac.uk) funded by the University of Edinburgh and EPSRC (EP/P020267/1)

References

1. https://www.compbiomed.eu
2. https://www.youtube.com/watch?v=1FvRSJ9W734
3. https://www.datacenterdynamics.com/analysis/superpowers-supercomputers-and-race-exascale/
4. Morgan J. V-cs wanting to lead from the front are frustrated by a lack of followers. Times Higher Education; London Iss. 2208, (Jun 18, 2015), p11.
5. OECD. Innovating Education and Educating for Innovation: The Power of Digital Technologies and Skills, Paris: OECD Publishing; 2016. Available from: http://dx.doi.org/10.1787/9789264265097-en
6. https://www.ucl.ac.uk/2034/
7. https://www.ucl.ac.uk/teaching-learning/ucl-education-strategy-2016-21
8. https://www.ucl.ac.uk/teaching-learning/connected-curriculum-framework-research-based-education
9. Townsend-Nicholson, A. Vignette E: Designing a throughline and a research-culture in Biochemistry. In: Carnell, B. and Fung, D, editors. Developing the Higher Education Curriculum Research-based Education in Practice. London: UCL Press; 2017. p.238.
10. http://www.earthmicrobiome.org
11. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010 May; 7(5):335-6. Available from: https://doi.org/10.1038/nmeth.f.303
12. https://hmpdacc.org
13. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using

QIIME 2. Nat Biotechnol. 2019 Aug; 37(8): 852–857. Available from: https://doi.org/10.1038/s41587-019-0209-9

14. https://www.hesa.ac.uk/data-and-analysis/sb252/figure-18
15. 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 Oct; 13(10) :e1005134. doi: 10.1371/journal.pcbi.1005134.
16. Richmond PA, Wasserman WW. Introduction to Genomic Analysis Workshop: A catalyst for engaging life-science researchers in high throughput analysis. F1000Res. 2019 Jul;8:1221. doi: 10.12688/f1000research.19320.1. eCollection 2019.