Developing a new community of practice by engaging clinicians and biomedical scientists with advanced computational methods

Andrea Townsend-Nicholson

Faculty of Life Sciences

Division of Biosciences

Structural & Molecular Biology

Computational Biomedicine

The Vision

The Reality

KEY: starting early to engage biomedical scientists - create fluency, expectation and support innovation and new discoveries in high performance computing

Many barriers to overcome

- Access to the infrastructure
- Training in its use
- Perceived difficulty
- Perceived demographic

CompBioMed Education Aims

To bridge the computational-experimental divide by supporting and developing existing users

To create a new category of users of HPC (“future users”) who will be fluent both computationally and experimentally

To provide ‘tried and tested’ educational templates and expertise for adoptions by other universities and institutions

The Landscape in 2016

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

The Landscape in 2016

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

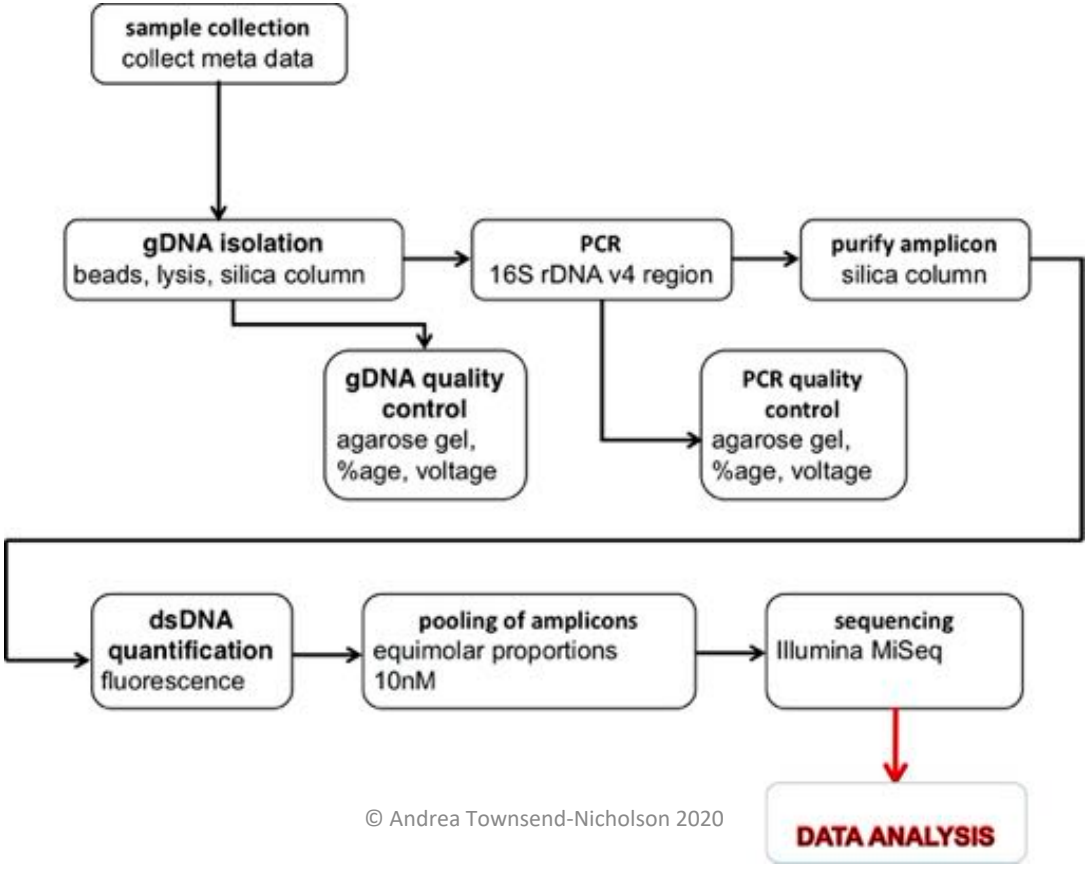
The Teaching

- Life Sciences undergraduates (BSc, iBSc, MSci)
  - BIOC0023 Specialist Research Project in Metagenomics
  - BIOC0031 Metagenomics and Molecular Medicine
  - BIOC0002 The Basics for Molecular Biosciences (Key Skills)
  - BIOC0012 Research Project Foundations (Key Skills)

- Medical students
  - SSC334 From Skin to Metagenomics, Year 1, Year 2
  - Extracurricular projects in Metagenomics (Year 2)

*All module resources are placed on GitHub, on the UCL institutional virtual learning environment and are published on the CompBioMed training portal

Experimental-Computational Workflow



Successful Engagement

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*

The Outcomes

Teaching Outcomes

- 100% success rate for students using HPC as part of their degree (2017-2020)
- Improved diversity:
  >50% female*
  >40% BME
- Employability from embedding computation in the Molecular Biosciences Curriculum

Research Outcomes

- Expansion of the programme:
  EU funding to deliver medical student HPC teaching at UVA, UPF/BSC, UOXF, USFD (CompBioMed2)
- Engagement with UKRI and EC
  To provide resource to support training HPC users, including students (£13m)

CompBioMed Education Aims

To bridge the computational-experimental divide by supporting and developing existing users

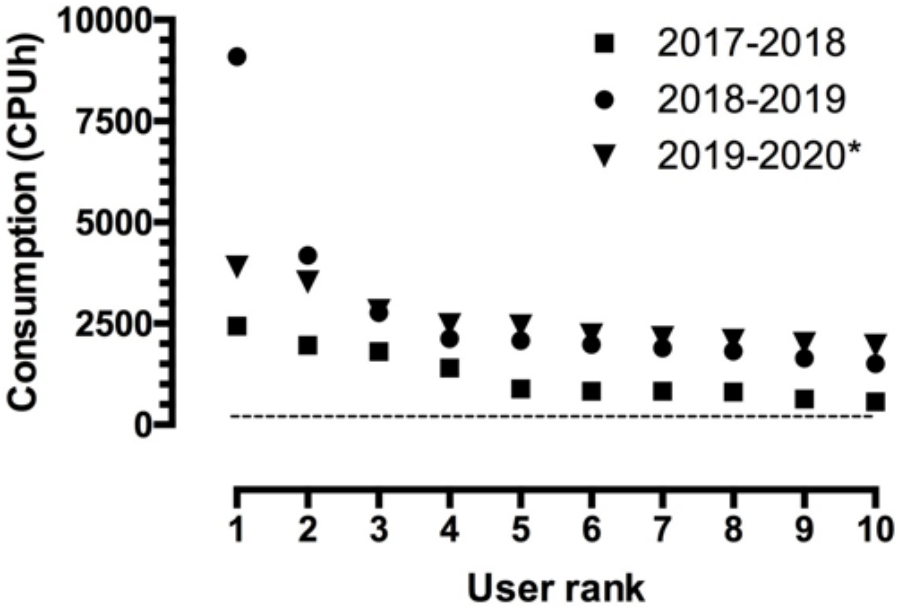
To create a new category of users of HPC (“future users”) who will be fluent both computationally and experimentally

To provide ‘tried and tested’ educational templates and expertise for adoptions by other universities and institutions

The Magnificent Seven

Unanticipated wins that we could not have predicted when we started the CompBioMed university education training programme…..

Unanticipated Win #1:
Significant engagement with HPC for coursework



Unanticipated Win #2:
Medical student-proposed extracurricular project

Hypothesis: The skin microbiome of clinical medical students will be less diverse than that of pre-clinical medical students and may reveal presence of subclinical levels of pathogenic bacterial species (e.g. MRSA, C. difficile, E. coli O157)

Target Journal: The Lancet

Support & Encouragement: Provost, Vice-Provost (Health), Dean (FLS)

Unanticipated Win #3: Continuing science student engagement with HPC

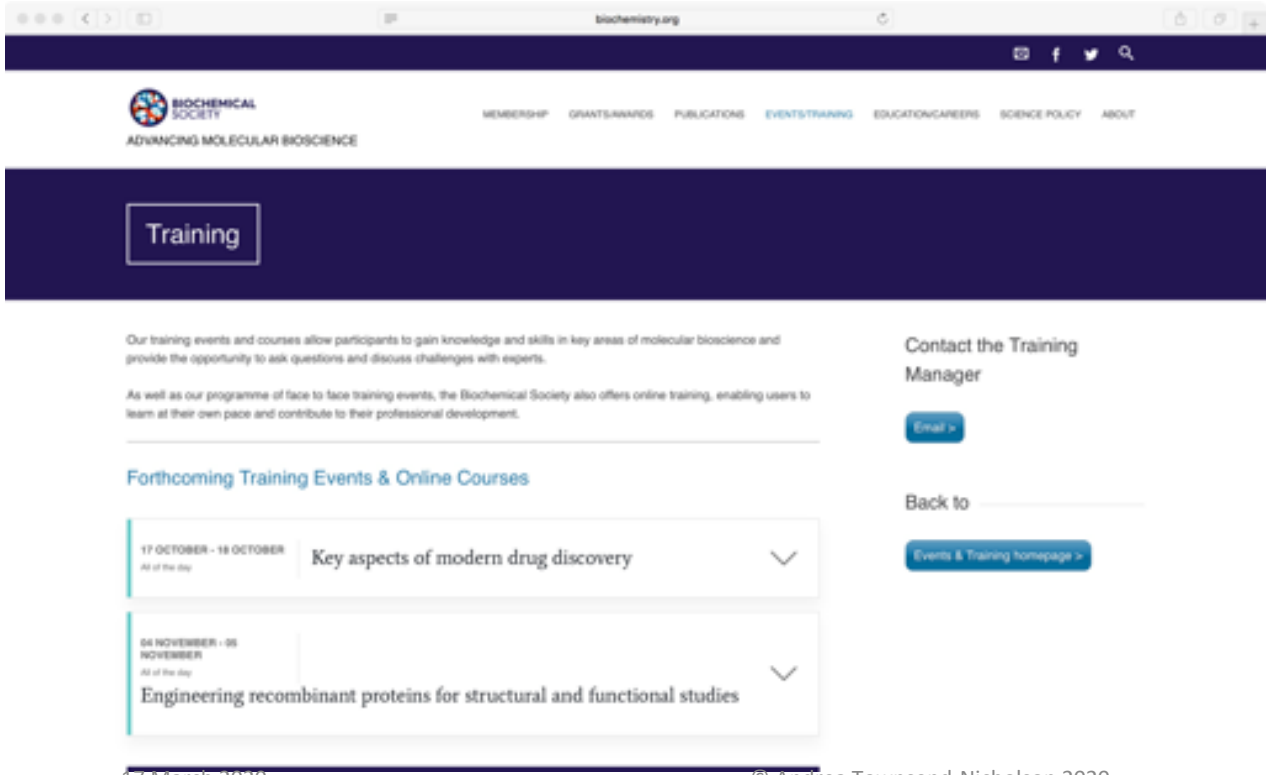
- The Molecular Biosciences Undergraduate Summer Placement Programme

- 9 undergraduate volunteers provided “Tech Support”
  - 3 teams: each with a year 3 BIOC0023 student and two year 1 undergraduates
  - Associate Fellowships of the Higher Education Academy
  - 3 volunteers presented posters at the CompBioMed2019 international conference

Unanticipated Win #4: iSeq@UCL an institutional sequencing-HPC

- UCL's School of Life and Medical Sciences Capital Equipment Fund
  - £40,000 to purchase two Illumina Next Generation Sequencing iSeq-100 machines
  - Establishing the iSeq@UCL facility for use in teaching, for early career researchers and established researchers and for CompBioMed2 HPC-based training
  - Unique in the world
  - Available for CompBioMed2

Unanticipated Win #5:
Adoption of the CompBioMed training by learned societies



- Introduction to Microbiomes and 16S rDNA
- Introduction to NGS (Next Generation Sequencing)
- Workshop *Getting to know your dataset*
- Workshop *Computing and High Performance Computing (HPC)*
- Analysis1 (command line)
- Analysis2 (Jupyter notebooks)
- Analysis3 (in the cloud)
- Comparison of data and summary of findings

Unanticipated Win #6: Public Engagement

The Virtual Human film has been presented to 850 prospective students and relatives at UCL Open Days and to 90,000 members of the public via a newsstand magazine

The Virtual Human went to the UK Parliament to highlight a case study in collaborative data using digital evidence at multiple scales (Evidence Week)

Unanticipated Win #7:
Continuing medical student engagement with HPC

"…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."

(Year 1 MBBS Student - 2017-2018)

2019-2020: 1 student from 2017-1018 enrolled in UCL's iBSc Mathematics, Computers and Medicine *Programme Director: Professor Benny Chain*

Lessons Learned: Successes

- **All students completed the course, for a 100% success rate**
- 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
- At all three Centres, individual accounts were created for students and the Centres ported and profiled the QIIME code
- All students logged in, edited scripts, ran scripts and extracted data without prior HPC experience
- Some of the students embraced the significantly different environment and made substantial use of the resource
- **Students overcame differences in running in different environments at UCL (on multiple clusters, their own laptops, etc.)**
- **Some students were logging onto Cirrus to run jobs from their phones**
- ***Trainers and HPC centres debugged and solved problems on the fly quite successfully***
- Trainers overcame the differences between running at the different HPC centres
- ***The HPC Centres learned about IP white-listing and the Cirrus fail2ban was relaxed for all users of Cirrus, thanks to input from these courses***

Lessons Learned: Challenges

- **The initial very large disk space was still not enough for excited students**
- Fail2ban (Cirrus) was not tested prior to student account creation in the workshop and issues with default draconian fail2ban settings required urgent IP whitelisting to work-around this
- Jobs were hanging due to a batch script bug; students tried increasing wall clock time AND core count without seeking help and burned a lot of the budget
- **Bugs in scripts were discovered, e.g. in TMPDIR, when running at a previously untested scale of usage**
- Scripts were hanging until the wallclock limit was hit due to extraneous line continuation characters because the students were unfamiliar with Linux (they knew how to comment out a line so it wasn't run, but they forgot (or didn't know) that they also had to remove the "/" symbol from the proceeding line)
- **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.**
- **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"**
- Issues arising from the lack of familiarity with a headless system
- 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.
- Students must have some command-line experience (understanding of directories/folders and basic Linux (ls, pwd, etc) plus vi/vim as an editor

Next Steps

- CompBioMed2 Education Training: Expanding from UCL to Oxford, Sheffield, UPF and UvA

- Adding new elements of computational biology to the university curriculum (improved digital skills, machine learning, MD simulations)

- Enhancing the User Experience

A final thought…

"…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."

Acknowledgments

CompBioMed Partners

University College London

Michael Baron and Amanda Cain, Jim Naismith (external examiner), Greg Campbell and Paul Dilworth, Peter Coveney

University of Edinburgh

Gavin Pringle and Terry Sloan

SURFsara

Marco Verdicchio

Barcelona Supercomputing Centre

Mariano Vazquez, Ruth Aris and Alfonso Santiago