Train-the-trainer: from bioinformatics student to bioinformatics trainer

Annette McGrath | CSIRO Bioinformatics Core Leader
12 July 2015
A journey from student to trainer

• Who we are
• Why bioinformatics training?
• Fellow travellers
• Bringing it all back home
• Our next roadtrip(s)
Who we are
• Commonwealth Scientific and Industrial Research Organisation
• Australia’s national science agency
• http://www.csiro.au
• Diverse range of basic and applied research interests:
  • Agriculture
  • Biosecurity
  • Digital Productivity
  • Energy
  • Food & Nutrition
  • Land & Water
  • Manufacturing
  • Mineral Resources
  • Oceans & Atmosphere
National Collections and Facilities

• Australian National Insect Collection
• Australian National Fish Collection
• Australian National Algae Culture Collection
• Australian National Wildlife Collection
• Australian National Herbarium
• Australian Tree Seed Centre
• The Atlas of Living Australia

• Australian Animal Health Laboratory
• Canberra Deep Space Communication Complex (on behalf of NASA)
• Australia Telescope National Facility
• Marine National Facility
Who might use bioinformatics in their research?

- Agriculture
- Biosecurity
- Digital Productivity
- Energy
- Food & Nutrition
- Land & Water
- Manufacturing
- Mineral Resources
- Oceans & Atmosphere

- Australian National Insect Collection
- Australian National Fish Collection
- Australian National Algae Culture Collection
- Australian National Wildlife Collection
- Australian National Herbarium
- Australian Tree Seed Centre
- The Atlas of Living Australia

- Australian Animal Health Laboratory
- Canberra Deep Space Communication Complex (on behalf of NASA)
- Australia Telescope National Facility
- Marine National Facility
Who might use bioinformatics in their research?

- Agriculture
- Biosecurity
- Digital Productivity
- Energy
- Food & Nutrition
- Land & Water
- Manufacturing
- Mineral Resources
- Oceans & Atmosphere

- Australian National Insect Collection
- Australian National Fish Collection
- Australian National Algae Culture Collection
- Australian National Wildlife Collection
- Australian National Herbarium
- Australian Tree Seed Centre
- The Atlas of Living Australia

- Australian Animal Health Laboratory
- Canberra Deep Space Communication Complex (on behalf of NASA)
- Australia Telescope National Facility
- Marine National Facility
Why bioinformatics training?
CSIRO’s bioinformatics capability (2010)

Bioinformaticians by location:

- **Perth**: PI: 1
- **Geelong**: LI: 2, CFNS: 1
- **Canberra**: CES: 1, CMIS: 1, IM&T: 1, PI: 6
- **Brisbane**: CMIS: 2, LI: 2
- **Sydney**: CMIS: 2, CFNS: 3
- **Hobart**: CMAR: 3
### CSIRO’s bioinformaticians by location, Division and position (2010)

<table>
<thead>
<tr>
<th>Main discipline</th>
<th>Bioinformatics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staff</td>
<td></td>
</tr>
<tr>
<td>City</td>
<td>City</td>
</tr>
<tr>
<td>Adelaide</td>
<td>Adelaide</td>
</tr>
<tr>
<td>Brisbane</td>
<td>Brisbane</td>
</tr>
<tr>
<td>Canberra</td>
<td>Canberra</td>
</tr>
<tr>
<td>Geelong</td>
<td>Geelong</td>
</tr>
<tr>
<td>Hobart</td>
<td>Hobart</td>
</tr>
<tr>
<td>Perth</td>
<td>Perth</td>
</tr>
<tr>
<td>Sydney</td>
<td>Sydney</td>
</tr>
<tr>
<td>Werribee</td>
<td>Werribee</td>
</tr>
<tr>
<td>Grand Total</td>
<td>Grand Total</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>City</th>
<th>Site</th>
<th>Division</th>
<th>Projects</th>
<th>Postdoc</th>
<th>Scientist</th>
<th>Leader</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adelaide</td>
<td>Waite Campus</td>
<td>PI</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Brisbane</td>
<td>Herston - RBWH</td>
<td>CMIS</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>St. Lucia</td>
<td>CMIS</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LI</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Canberra</td>
<td>Acton</td>
<td>CMIS</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Black Mountain</td>
<td>PI</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CES</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IM&amp;T</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Geelong</td>
<td>Geelong AAHL</td>
<td>LI</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Hobart</td>
<td>Hobart</td>
<td>CMAR</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Perth</td>
<td>Floreat</td>
<td>PI</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Sydney</td>
<td>Macquarie Uni.</td>
<td>CMIS</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>North Ryde</td>
<td>CFNS</td>
<td>2</td>
<td></td>
<td>1</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Werribee</td>
<td>Werribee SnydRd</td>
<td>CFNS</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Grand Total</td>
<td></td>
<td></td>
<td>5</td>
<td>12</td>
<td>8</td>
<td>2</td>
<td>27</td>
</tr>
</tbody>
</table>
The CSIRO Bioinformatics Core

• A cross-organisational effort
• Bioinformatics needs and activities are similar across many life science organisational units
• Enterprise level infrastructure and activities beyond remit for those bioinformaticians
• Aim to provide **capability** of broad benefit to CSIRO’s bioinformatics efforts
• A strong need to connect bioinformaticians – no more ‘lonely bioinformaticians’
• Collaborative research projects which enhance CSIRO’s bioinformatics capabilities in a particular area of research with divisional bioinformaticians and IM&T
What is “capability”?

- The term “research capability” has three components
  - $S$: Skills, experience, know-how of our people
  - $I$: Infrastructure that supports research
  - $R$: Relationships and collaborations
- Integration is important. A capability
  - is more than the sum of its parts
  - typically draws on *multiple* disciplines to tackle problems
  - is not another term for a discipline.
Barriers to effective uptake of bioinformatics in CSIRO

• Survey in 2011 revealed the following barriers
  – Tools
    – Access to enterprise-wide tools
  – Expertise and access to it
  – Training
    – For bioinformaticians
    – For end-users of bioinformatics tools
  – Awareness
    – Tools and resources

• A threat to the practice of good science
Training needs survey
Increasing bioinformatics literacy
Fellow travellers
A happy coincidence

• Manages the federal government’s investment in
  • Genomics
  • Metabolomics
  • Proteomics
  • Bioinformatics

http://www.bioplatforms.com
<table>
<thead>
<tr>
<th>Capability</th>
<th>APAF NSW</th>
<th>TGR Biosciences SA</th>
<th>Monash Antibody Technology Facility (MATF) and Monash Univ Vic</th>
<th>Queensland Institute of Medical Research Qld</th>
<th>Proteomics International (WAIMR) WA</th>
<th>Adelaide Proteomics Centre SA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass Spectrometry</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Electrophoresis</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Liquid Chromatography</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Post Translational Modifications</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MS Imaging</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Bioactive Discovery</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>High Throughput Cell Based Screening</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Capability</td>
<td>Centre for Comparative Genomics, Murdoch Uni WA</td>
<td>EBI Mirror, UQ Qld</td>
<td>Systems Biology Initiative (SBI), UNSW NSW</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------------------------------------------------</td>
<td>-------------------</td>
<td>--------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computational biology</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparative genomics</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Next-gen data analysis</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proteomic data analysis</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Integration</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Management</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software development</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High performance computing</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Network biology</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Repository</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education &amp; Training</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Environment</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human Health</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Metabolomics Capability Matrix

<table>
<thead>
<tr>
<th>Capability</th>
<th>University of Melbourne, School Botany and Bio21 Vic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolite Analysis</td>
<td>✓</td>
</tr>
<tr>
<td>Mass Spectrometry</td>
<td>✓</td>
</tr>
<tr>
<td>Lipidomics</td>
<td>✓</td>
</tr>
<tr>
<td>Volatile Analysis</td>
<td>✓</td>
</tr>
<tr>
<td>Fluxomics</td>
<td></td>
</tr>
<tr>
<td>Methods Development</td>
<td>✓</td>
</tr>
<tr>
<td>Systems Biology</td>
<td>✓</td>
</tr>
<tr>
<td>Environment</td>
<td>✓</td>
</tr>
<tr>
<td>Human Health</td>
<td>✓</td>
</tr>
<tr>
<td>Agriculture</td>
<td>✓</td>
</tr>
</tbody>
</table>

### Genomics Capability Matrix

<table>
<thead>
<tr>
<th>Capability</th>
<th>AGRF Vic, NSW, SA, WA, Qld</th>
<th>Ramaciotti Centre, UNSW NSW</th>
<th>John Curtin School of Medical Research, ANU ACT</th>
<th>Garvan Institute of Medical Research NSW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Next-gen sequencing</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Genotyping/SNP analysis</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>Gene expression</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>Microarrays</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>DNA/RNA extraction</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Transcriptomics</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Epigenomics</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>Metagenomics</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>Methods Development</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>Environment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
</tr>
</tbody>
</table>
A training collaboration

• CSIRO’s challenge
  • Increasing bioinformatics literacy, capability and competencies in a
    – diverse,
    – relatively small and
    – geographically dispersed
    Group of bioinformaticians and many more biology end-users

• Bioplatforms Australia (BPA)
  • Provide training for a
    – diverse,
    – relatively small and
    – geographically dispersed
    Group of bioinformaticians and many more biology end-users

• Let’s collaborate!
Our goal

• Create a critical mass of trainers
• Engender a cohesive training community
• ‘Trainer’ – volunteers
• To deliver hands-on training workshops in the major cities in Australia
Training the trainers

Turning bioinformaticians into bioinformatics trainers
Where do we start?

• A great deal of bioinformatics training experience
• Many successful workshops in a wide range of bioinformatics topics
• Dedicated resources
• Very willing to help us on our journey!
Introduction to NGS 2012

- Introduction to Next Generation Sequencing March 12th-14th 2012 2-day workshop
- 8 trainers from CSIRO & BPA facilities
- Participants and observers
- 2 days of ‘Train-the-trainer’ workshop
Training the trainers

• Short biography
• Summary of training interests and experience
• TtT workshop:
  – Learning styles
  – Motivation for adult learners
  – Planning our course material
    • Learning objectives
    • Delivery tips and tricks
    • Targeting the right audience
    • Measuring impact
Metagenomics: Managing, Analysing and Visualising Data Course

• 9-11th September 2013
• Expanding the trainer network
• Developed an outline for a 2 day “Introduction to Metagenomics” workshop
DAY ONE 3

1. WELCOME

2. INTRO
   (Overview of NEMO, platforms, data formats, pipelines/workflows)

   QC (Good/BAD)
   Intro + HAND-ON
   Trimmed primers/adaptors
   QC (Q3/ME)

COMMAND LINE WARM UP
DAY ONE (cont'd)

4.3 INTRO OF 16S

Overview

Pros/Cons

PLATFORMS USING 16S

S. HAND-ON!!!

TAXONOMY CLASSIFICATION

- Mapping File
- Library Splitting
- Picking OTU's
- OTU's Visualisation
- α diversity
- β diversity

6. Q&A

- OTU's stats
- Vein OTU's
- Vein OTU's heatmap
- Summary of tax
DAY TWO

1. Intro into WGS
   - Workflows (1 hr)
   - Types of analysis & questions

Diff tools
Diff answers

Data submission
Data download (SRA, ENA webin)

Meta-data

Q.C. → Assemble

Moved to a single slot for 30 min
Q.C. →
   ↓
   PROCESSED READS
   ↓
   CLUSTERING
   ↓
   RNA SELECTOR
   ↓
   FINDING GENE
   ↓
   ANNOTATION
Bringing it all back home
Post-workshop

• Continue to develop our workshop content
  • EBI content
  • New content
• Datasets, tools and testing
• Documenting and developing trainee workbook
• Testing, revising, testing, revising ......
• EBI trainers come to Australia to deliver 2 workshops, back-to-back
• Our trainers are ‘helpers’
Target audience

• Bench biologists
• Aiming to increase research productivity
  – Scientists can independently look at their data
  – Engage in more productive conversations with bioinformaticians
• Used an application process
• Selection questions
  – to ensure that the audience expectations could be met
  – Attendees were of similar skill levels
Introduction to NGS outline

- Intro to Linux (lecture + hands-on)
- Intro to NGS (lecture)
- QC (lecture + hands-on)
- Alignment (lecture + hands-on)
- ChIPSeq (lecture + hands-on)
- RNASeq (lecture + hands-on)
- *de novo* genome assembly (lecture + hands-on)
Delivering the workshop

• We do not have dedicated training facilities

• We want to deliver the workshop in a roadshow format

• Minimise maintenance of the training environment
  • No monolithic installs

• Minimise cognitive burden on trainees
  • The training environment should go unseen

• Make everything publically accessible and as reusable as possible

• Engage local bioinformaticians
NeCTAR

- Australian Research Cloud
- A national infrastructure initiative
- A federated cloud
- Cloud resources, tool sets, virtual labs
THE REUSABLE RESOURCES: VM SETUP

Office

• Vanilla Ubuntu VM
  • Gnome
  • FreeNX
  • Generic Tools

• NGS tools
• NGS data
• NGS handout

NGS Workshop VM
Architecture

- GitHub
  - Orchestration
  - Catalog
  - Training Manual
  - BPA and CSIRO Workshops

- Modules
  - Datasets
  - Object Storage

- Virtual Machine Images

- Analysis Tools

- Platforms
  - VMware
  - Nectar
  - Amazon Web Services
Bioinformatics Training Platform

- OpenNX
- Remote Desktop Access
Resources

- NeCTAR – NGSTrainingV1.4
- Amazon Machine Image (AMI):
  - BTP-1434757863 (ami-67b3c95d)
  - Currently available on Asia Pacific (Sydney) Region
Cloud Pros and Cons

Pro
- Consistent training environment
- No “alien” OS on host network
- Minimal host network configuration and traffic
  • Firewall (port 22)
- Minimal local computer specification and configuration
  • NX Client plus session files
- Scalable resources
- Encourages reproducible work

Con
- Remote vs local confusion
  • Hide this using NX
- How to analyse own data?
- Requires a computer suite
- Sysadmin skills required
Workshops

BPA/CSIRO Competitive Courses:
520 places
8 cities

EMBL Australia PhD Program:
180 places

Other workshops:
380 places

Total: ~1100 places in 3yrs
Feedback

“I found this workshop very helpful. Thank you for organising”
“Provided introduction to NGS for all levels of learners”
“Great instructors and well organised.”
“(The best part was) Finding out which tools are commonly used and issues to be conscious of.”
“The best part was having the material presented beforehand, the trainee handbook to work through and demonstrators to walk around to help if required.”
“(The best part was) The preparation of the environment to work in, no wasted time downloading sequences/files/etc.”
“(The best part was) all the trainers.”
“The fact that there was a hands on practical component.”
“Best part of the course was getting the chance to gain some hands-on experience and being able to ask questions, get some insight. The course help me to see different aspects for example on RNA-seq playing around DAVID was interesting.”
Sharing and extending our material

• All of our materials are available on GitHub
• Our trainers can all contribute
  – edgeR tutorial has been contributed
  – With many courses being run, how do we manage contributions
• You can contribute 😊

https://github.com/BPA-CSIRO-Workshops
Our next roadtrip(s)
Continuing to expand

• Expanding both
  – our network of trainers and
  – the training programs that we will run

• To ensure sustainability

• To respond to needs identified by our trainees and our network

• Upcoming Train-the-trainer events
  – Cancer Genomics at EBI (week of July 20\textsuperscript{th})
  – Non-model organisms workshop collaboration with TGAC (week of July 27\textsuperscript{th})
ACKNOWLEDGEMENTS

Annette McGrath
Konsta Duesing (F&N Flagship)
Sean Li (DP Flagship)
Sean McWilliam (AP Flagship)
Paul Greenfield (O&A Flagship)
David Lovell (DP Flagship)
Phillippe Moncuquet (DP Flagship)
Paul Berkman (AP Flagship)

Catherine Shang
Nathan Watson-Haigh (ACPFG)
Nandan Deshpande (SBI, UNSW)
Paula Moolhuijzen (Murdoch University)
Sonika Tyagi (AGRF)
Matthew Field (ANU)
Susan Corley (SBI, UNSW)
Zhiliang Chen (SBI, UNSW)

Cath Brooksbank (EBI)
Vicky Schneider (TGAC)

Sarah Morgan (EBI)
Matthias Haimel (University of Cambridge)
Myrto Kostadima (University of Cambridge)
Remco Loos (EBI)
Alex Mitchell (EBI)
Hubert Denise (EBI)

Mark Crowe (QFAB)
Peter Sterk (Oxford e-Research Centre)

Jerico Revote (Monash e-Research Centre)
Simon Michnowicz (Monash e-Research Centre)
Steve Quenette (Monash University)
Thank you

Digital Productivity Flagship
Annette McGrath
CSIRO Bioinformatics Core Leader

t  +61 2 6216 7251
e  annette.mcgrath@csiro.au