Top ten (or so!) tips for setting up a bioinformatics course

Gabriella Rustici, School of Biological Sciences  gr231@cam.ac.uk
Hands on courses on HT data analysis
2007-2014

54 hands on training events, > 1750 people trained

HTS: 76%
Microarray: 24%
Target audience

**Nationality**
- Europe: 80%
- Americas: 12%
- Asia: 6%
- Australia: 1%
- Africa: 1%

**Institution**
- Academia: 93%
- Industry: 6%
- Other: 1%
Target audience

Type of delegate

Bio-informatician
Corporate
Engineer
Masters Student
MD, PhD
PhD student
Postdoc
Senior Academic/PI
Staff Scientist
Undergraduate

[Bar chart showing the distribution of different types of delegates]
What challenges do we face?

Diversified audience

- different backgrounds (biologists, bioinformaticians, clinical researchers, etc.)
- different levels of statistical knowledge
- different levels of familiarity with programming languages

I am starting a project involving HTS applications and I need to learn how to analyse the data that I will generate. I have never done this kind of analysis before, and I have very little familiarity with data analysis tools. Bioinformatics support is lacking in our department, so it is vital for me to acquire these skills if I want my project to be successful.

I am involved in HTS projects. The analysis is done by a bioinformatician, but I would like to learn more about the analysis to be able to have a better interaction with the bioinformatician. I have run some simple analysis tasks using pre-compiled scripts, but I would not know how to modify them to suit my needs.

I am a bioinformatician, supporting various research groups with their analysis needs. I run HTS data analysis using some tools, but I want to learn how to use other tools as well as keep up to date with the latest algorithms that are being developed in this field.

I have been involved in microarray data analysis projects for a long time and now I am switching to HTS data analysis. I feel confident with using tools for microarray analysis, but I want to know what I need to use to analyse HTS data. I am confident in the use of some programming languages.
What challenges do we face?

- Participants’ expectations
- Break down the complexity of the analytical pipelines
- What software to use? command line vs. workflow based solutions, e.g. Bioconductor vs Galaxy
- Computing infrastructure for training purposes often inadequate
What are we aiming for?

• What is the realistic outcome of a course?

• Trainees should learn:
  ✓ how to interpret HTS data
  ✓ what the HTS data analysis entails and how the data processing influences the interpretation?
  ✓ how to critically evaluate the data analysis tools available

• Enable life scientists to establish a partnership with their statistician and/or bioinformatician collaborators, based on mutual understanding
Bioinformatics course’s core elements

**Audience expectations**
- strict selection criteria
- course aim & objectives
- course prerequisites

**Course planning**
- Scientific content
- admin
- IT infrastructure

**Feedback**
- impact measures
- expectation matching
- usefulness and applicability
- short & long

**Course ethos**
- Theory vs practice balance
- openness and reproducibility
- critical evaluation
Course’s aims & objectives

- Provide clear learning objectives and a sufficiently detailed course description

- By reading a course description, participants should immediately understand:
  - what the objectives of the course are
  - who is the course aimed at
  - what they should expect to learn if attending the course
  - what the course content will be
  - what are the course’s prerequisites (if any)
Participants’ selection criteria

• Selection criteria are fundamental
• For NGS courses we require:

  ✓ A basic knowledge of the programming languages that will be used during the course (e.g. R, Unix)

  ✓ Participants to have (or will shortly to acquire) data ready to analyze

  ✓ Having the possibility to immediately apply what they have learned during the course to the analysis of their own data is the most effective approach.
Course’s prerequisites

• What should the participant already be familiar with in order to maximize learning outcomes?

• **Provide tutorials** showing what level of knowledge and familiarity (i.e. with a particular programming language) they should have in order to maximize their learning outcomes.

• Participants should be encouraged to attend **introductory courses** in the programming languages that they will be using during the course, face-to-face or online.
Course’s scientific content

Our hands-on courses consist of a well balance mixture of:

• lectures, which illustrate the fundamental concepts in the analysis of HTS data, and
• hands-on sessions (~50%), which allow the students to practice how to run analysis of real datasets

Focus the training on:

• crucial steps in the analysis of HTS data
• provide the necessary information needed to run each step of the analysis
• clearly connect the theory to hands-on exercises on real data.
Software choice & IT aspects

Software choice

• Focus on the use of open source, stable, actively developed and well-maintained software tools (i.e. Bioconductor, Galaxy,…)

Computing infrastructure

• Local machines, VMs, clusters, cloud computing?
• Positive experience with Docker – students can take home a course’s image and re-run exercises right away
• Piloted in May 2015 – monitoring what the participants do with this through multiple surveys
Course’s planning in numbers

A 5 day course takes:

- Number of trainer
- Days pre- / post- event work [program, materials, practical testing]
- Days of Logistics work [registration; booking; advertising; enquiries]
- Days of image preparation
Short term feedback

Set of questions to establish if:

• The course was well organized
• The program was well balanced and a good fit to the participants’ background
• The presentations/exercises were of good quality and relevant
• The participants had used the resources presented before and whether they are planning to use them in the future
• The participants expectations were met
• Further training is required
Short term feedback results

Course organization:
- Excellent: 20%
- Good: 79%
- Satisfactory: 1%

Course format:
- Satisfactory: 6
- Good: 48
- Excellent: 46

- 81% felt that the duration of the sessions was appropriate.

Resources use:
<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequently</td>
<td>14</td>
</tr>
<tr>
<td>Occasionally</td>
<td>66</td>
</tr>
<tr>
<td>Unaware</td>
<td>6</td>
</tr>
<tr>
<td>Used others</td>
<td>14</td>
</tr>
</tbody>
</table>

95% expectations met

99% recommend the course

Need for further training:
- Yes
- No
Long term feedback

• Were you working in the field of microarray or HTS data analysis BEFORE attending this course? Were you working in the field of microarray or HTS data analysis AFTER attending this course?
• Did you establish collaborations with other participants that attended the same course?
• Did you write a thesis following your participation to this course?
• Did you publish any journal article/book chapter as a result of your participation to this course?
• Did you write a grant proposal as a result of your participation to this course?
• Have you authored any software as a result of your participation to this course?
Long term feedback results

- 25% of people responded
- 70 journal articles
- 16 dissertation/thesis, many more still work in progress
- 10 successful grant proposals
- 4 software packages

Given that participants to our courses have primarily a biological background we did not expect to see the development of software packages as an outcome of the training received
Sustainability

Sustainability is affected by:

• number of trainers available

• the preparation and planning required

• not all have the resources (manpower / training space / computational power) to run these courses

• time consuming effort that needs a good infrastructure

We need to look for methods to ensure sustainability!!!!!!
Sustainability

• “Train the Trainer” courses – training new trainers in the way in which we plan and deliver courses; share training materials to enable them to return to their organisations and teach

• Share training materials – all materials produced made publically accessible via CC licenses

• Establish communities of trainers - GOBLET / ELIXIR to aid in developing these communities

• Ensure reusability of materials – portals for training materials

• What do we need to share materials effectively?
NGS training material hackaton @ UoC

- develop a unified collection of NGS training materials, tailored to different audiences and updated collaboratively
- focus on RNA-seq, ChIP-seq and variation analysis
- integrate materials developed by different trainers
- upload them to a common repository on Git
- agree on common descriptors
- discuss learning assessment methods

See poster **C15**: “An online platform for NGS trainers to share teaching experience and materials” and talk to Bastian!!!!!!!
Step 1: Choose your term(s)

Tip: you can drag, Ctrl + click or Cmd + click to select multiple options.

- Alignment
- Annotation
- Assembly
- BAM
- BED
- Bos-taurus
- CRAM
- ChIP-Seq
- ChIP-Seq-QC
- Culex-quinquefasciatus
- DNA-Seq
- Data-format
- De-novo-transcriptome-assembly
- Differential-binding
- Differential-expression
- Experimental-design
- Exploratory-analysis
- Expression-estimate
- Expression-estimation
- FASTA
- FASTQ
- FastQC
- Feature-summarisation
- GFF3
- Galaxy
- Homo-sapiens
- Linux
- Mus-musculus
- NGS-introduction
- Peak-calling
- Platform
- Populus-tremula
- Pre-processing
- Prerequisite
- QC
- R-programming
- RNA-Seq
- SAM
- Sequencing-technology
- Statistical-model
- Statistics
- Unix
- VCF
- Variant-calling
- Variant-filtering
- Visualisation
- WIG
NGS training material repository - UI

Step 1: Choose your term(s)

Tip: you can drag, Ctrl + click or Cmd + click to select multiple options.

Alignment: 11  Annotation: 5  Assembly  BAM: 8  BED  Bos-taurus  CRAM  ChIP-Seq  ChIP-Seq-QC
Culex-quinquefasciatus  DNA-Seq  Data-format: 1  De-novo-transcriptome-assembly  Differential-binding
Experimental-design  Exploratory-analysis: 8  Expression-estimate  Expression-estimation: 6  FASTA  FASTQ: 7  FastQC
Feature-summarisation: 6  GFF3: 3  Galaxy  Homo-sapiens  Linux  Mus-musculus  NGS-introduction: 1  Peak-calling  Platform
Populus-tremula: 3  Pre-processing: 9  Prerequisite  QC: 10  R-programming: 3  RNA-Seq: 7  SAM  Sequencing-technology
Statistical-model: 9  Statistics  Unix  VCF  Variant-calling  Variant-filtering  Visualisation  WIG

Title: Differential expression analysis on the Robinson, Delhomme et al. dataset.
Author: @bastian, @delhomme
Ontologies: RNA-Seq, Differential-expression, R-programming, Statistical-model
Path: Content/RNA-Seq/Nicolas_Delhomme_EMBO-Oct-2014/06_EMBO-October-2014-Differential-expression.md
NGS training material repository - UI

• Navigate to a module

• Overall learning objective for a module (i.e. RNA-seq) and detailed objectives for each sub-module (i.e. QC, differential expression, …)

• Modules are tagged with keywords for easy retrieval

• Find associated datasets – we need more dataset!!!!!

• Is this a stand alone module or part of a course?

• Core set of descriptors so that sub-modules can be compared
NGS training material repository – module descriptors

* [Title]
* [Author]
* [Description]
* [Aims]
* [Target audience]
* [Prerequisites]
* [Learning objectives]
* [Content: slides, tutorials]. For each file indicate approximate duration, IT requirements for practicals, software needed
* [Datasets: description, origin, how was it modified, what does demonstrate] – datasets themselves not stored in Git
* [Stability of the content]
* [Literature references]
What’s coming up next

- Currently available at: [https://bioinformatics.upsc.se/trainers](https://bioinformatics.upsc.se/trainers)
- Make repository public and write a paper on our experience
- Hackaton to annotate existing collection with ontologies, promoted by Elixir-Estonia, in August 2015
- Hackaton for metagenomics training materials, as collaboration between Elixir and Goblet, in spring 2016 in Denmark.

Get in touch if you would like to join us and collaborate on any of this!

We have a mailing list that you can join!
• We too often deal with the consequences of the low amount, and low quality, of bioinformatics education for undergraduate or master’s students in life science and clinical curricula

• This needs to change

• UoC launching a **Bioinformatics** module for undergraduate in life science, medical and veterinary curricula from 2015-16

• UoC involved in the development of the HEE-funded master in Genomic Medicine – **Bioinformatics** and **Advance Bioinformatics** modules for qualified NHS healthcare professionals
Acknowledgments

Umea Plant Science Centre
Nicolas Delhomme
Bastian Schiffthaler

EMBL-EBI
Sarah Morgan

Trainers consortium and all trainers that contribute to our courses