Data deluge requires a training tsunami

Eija Korpelainen
CSC – IT Center for Science
Outline

- **Why** is so much training needed now?
- **Who** should be trained?
- **What** should be taught and **how**?
- **Who** provides training?
- **Where** are we now?
My training situation:
National bioinformatics core facility

- CSC serves all universities in Finland
  - lot of users
  - different backgrounds (medicine, agriculture, forestry, biotechnology, fisheries…)
  - scattered around the country
- 64 training courses in the last 4 years
- Training collaboration in GOBLET, EMBnet and ELIXIR
Why is so much training needed now?
Sequencing is popular

- Sequencing is an extremely versatile measurement technology which can be applied to many topics
  - Genomes, metagenomes, transcriptomes, genomic variants
  - Expression and regulation (miRNA, TF, methylation,…)
  - Etc etc

- Sequencing has become affordable
  - More and more researchers use sequencing

But...

- Exploiting sequencing to its full extent requires substantial data analysis skills
  - This is often a bottleneck
What is the problem with NGS data analysis?

- A lot of methods
  - Each “seq” requires its own data analysis methods
- Methods change all the time
  - NGS data analysis field has not matured yet
  - New sequencing platforms require new methods
- Data is voluminous
  - New technical skills (e.g. Hadoop) required

- Life scientists are ill-equipped for the task
  - Typically no background in programming or statistics
Who should be trained?
Who needs to understand data analysis?

- **People who analyze the data**
  - Bioinformaticians
  - Wet lab scientists
  - Technical specialists

- **People who collaborate with those analyzing the data**
  - Wet lab scientists
  - Principal investigators

- **Other people who use the data**
  - Health care professionals etc

→ **Training challenge: heterogeneous audience with different backgrounds and learning objectives**
Should wet lab scientists analyze data?

- “No, because they don’t have the required background”
- “No, because if they do, bioinformaticians will be unemployed”

- Yes, because they know best their research question
- Yes, so that they can plan experiments better
- Yes, so that bioinformaticians can offload routine tasks to them and concentrate on the more demanding ones
What should be taught and how?
What should we teach?

- Depends on the audience: different backgrounds and learning objectives

- Theory of the analysis methods or practical skills?
  - Both, because people need to understand what they are doing and how to do it
  - Finding the right balance is challenging, especially when the training time is limited

- Specific analysis tasks or general skills such as programming and statistics?
  - JIT = just in time training
Does everybody need to learn unix and R?

- Bioinformatics oriented people benefit from investing the time to learn unix, R, etc

- People who analyze data less frequently (or only want to learn the concepts) benefit from using a GUI
  - Minimal time needed to learn the GUI $\rightarrow$ more time for understanding the analysis methods instead
  - Lower threshold, easier to feel encouraged: “I can!”
Course format that works in our setting

- 25-30 students, 2 trainers, 2-3 days
- Homogenous group
  - Describe the goals and prerequisites clearly!
- Keep it informal
- Break the topic into small chunks
  - Short lectures followed by exercises
  - Wrap-up the exercises before moving to the next topic
  - Circulate in the room during the exercises
  - Bonus exercises for faster people
- Make students work in pairs
- Collect feedback
- Follow up after the course
Many good formats available

- **Couple of hours during several weeks**
  - Gives trainees more time to digest and try new things

- **eLearning, MOOCs**

- **Blended learning, inverted classroom**
  - Lectures by video, exercises in classroom

- **Oxford model: training via a project for 3 years**

- **Bioinformatics miracle pants? Ask Pedro…**
Who provides training?
Training is often a side job

- Trainers are typically analysis method developers or core facility bioinformaticians
  - Good substance knowledge
  - May lack time, interest or pedagogical skills
  - Developers might be biased to teach their own methods

- The demand for training is bigger than supply. How to get more trainers?
  - Improve the status of trainers
  - Provide training for trainers
  - Establish trainer networks to exchange ideas
Who organizes training?

- Universities and research institutes
- National and international bioinformatics centers
- International projects like SeqAhead!
- Professional organizations (CPD)
- Companies

And who pays for it?

- All research funding for life sciences should have some money ear-marked for training (and for bioinformatics in general)
Where are we now?
Training situation is improving

- National and international networks
  - GOBLET
  - EMBnet
  - ELIXIR
  - BD2K

- Growth in online training resources

- Bioinformatics training in kindergarten, high school, ...and even in life science degree programs!
GOBLET

- Global Organisation for Bioinformatics Learning, Education & Training
- Provides a global, sustainable support and networking infrastructure for trainers
  - portal for sharing materials, yearly meetings
- Promotes training
- http://mygoblet.org/
Summary

- Making full use of sequencing requires data analysis skills
- Need to train a large number of people who have very different backgrounds and goals
- Wet lab scientists should analyze data
- Need to balance theory and practical skills
  - It’s ok to use a GUI
- More trainers are needed
- Training needs to be funded properly
- Training situation is improving
Acknowledgements

- All my students for training me in training
- GOBLET for peer support
- Erik for all the SeqAhead work