LEARN BIOINFORMATICS BY DOING BIOINFORMATICS

Myrto Kostadima – Research Associate
Department of Haematology, University of Cambridge, UK
WHO I AM

Computational biologist specialised in the analysis and visualisation of high-throughput sequencing data with a focus on the regulation of gene expression in the human haematopoietic system. Experienced in conceiving, organising, implementing and delivering bioinformatics training courses on high-throughput sequencing data analysis. Member of the Higher Education Academy in UK.
The following ideas, suggestions and points of concern have arisen from my involvement in short (1-5 days long) bioinformatics hands-on training events.

But with the right modifications can be also applicable to other types of training activities.
I KEEP THINKING ABOUT TEACHING.
The UK Professional Standards Framework (UKPSF)

Areas of Activity

A1. Design and plan learning activities and/or programmes of study.
A2. Teach and/or support learning.
A3. Assess and give feedback to learners.
A4. Develop effective learning environments and approaches to student support and guidance.
A5. Engage in continuing professional development in subjects/disciplines and their pedagogy, incorporating research, scholarship and the evaluation of professional practices.

Core Knowledge

K1. The subject material.
K2. Appropriate methods for teaching and learning in the subject area and at the level of academic programme.
K3. How students learn, both generally and within their subject/disciplinary area(s).
K4. The use and value of appropriate learning technologies.
K6. The implications of quality assurance and quality enhancement for academic and professional practice with a particular focus on teaching.

Professional Values

V1. Respect individual learners and diverse learning communities.
V2. Promote participation in higher education and equality of opportunity for learners.
V3. Use evidence-informed approaches and the outcomes from research, scholarship and continuing professional development.
V4. Acknowledge the wider context in which higher education operates recognising the implications for professional practice.
STEP 1: conceive training event

A1. Design and plan learning activities and/or programmes of study
Where do I start and what do I teach?

- Identify your area of expertise
  - K1. The subject material

- Identify the training needs/gaps

- Identify your audience (e.g. pre-workshop surveys)
  - K2. The level of academic programme
How do I teach and with whom?

- Find the “right” tools to enhance training
  - K4. The use and value of appropriate learning technologies

- Find the “right” people to collaborate
STEP 2: prepare material

A1. Design and plan learning activities and/or programmes of study.
Tips for teaching material

- Clear learning goals
  - Achieved with a minimal set of different tools

- A good data set
  - availability/accessibility
  - size
  - good fit

- Less talking, more doing
ENGAGE

CHALLENGE

REWARD
V1. Respect individual learners and diverse learning communities.
V1. Respect individual learners and diverse learning communities.
V1. Respect individual learners and diverse learning communities.
“Limiting information density in presentations promotes effective delivery by restricting cognitive strain.” *
Break the ice.. with a group exercise

- *de novo* genome assembly
- ChIP-seq
- RNA-seq
How does a ChIP-seq analysis pipeline look like?

**INPUT FILES**
- Reference annotation
- Indexed Reference Genome FASTA
- Raw Illumina ChIP-seq reads FASTQ

**DATA ANALYSIS STEPS**
- Motif analysis
- Peak annotation
- Short read alignment
- Peak detection
1. Raw Illumina ChIP-seq reads FASTQ
2. Indexed Reference Genome FASTA
3. Short read alignment
4. Peak detection
5. Peak annotation
6. Motif analysis
General information

The following standard icons are used in the hands-on exercises to help you locating:

- Purple circle: Important Information
- Green box: General information / notes
- Blue question mark: Follow the following steps
- Yellow question mark: Questions to be answered
- Red exclamation point: Warning – PLEASE take care and read carefully
- Green star: Optional Bonus exercise
- Red star: Optional Bonus exercise for a champion

Resources used

Samtools: http://samtools.sourceforge.net/
BEDTools: http://code.google.com/p/bedtools/
UCSC tools: http://hgdownload.cse.ucsc.edu/admin/exe/
IGV genome browser: http://www.broadinstitute.org/igv/
MACS: http://liulab.dfci.harvard.edu/MACS/index.html
PeakAnalyzer: http://www.ebi.ac.uk/bertone/software
MEME: http://meme.sdsc.edu/meme/cgi-bin/meme.cgi
TOMTOM: http://meme.sdsc.edu/meme/cgi-bin/tomtom.cgi
SICER: home.gwu.edu/~wpeng/Software.htm

Additional resources:

Ensembl: http://www.ensembl.org

Original Data from: http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-56138/
Hands-on tutorials should be split in short parts, each one followed by a set of questions.

- Allow enough time for the task completion
- Go through the answers in the classroom.
“Repetition is the mother of learning, the father of action, which makes it the architect of accomplishment.”

3. Why does the quality deteriorate at the end of the read?

4. How can we trim the reads to filter out the low quality data?
Day 1

Sort alignments according to chromosome position and store the result in the file with the prefix Oct4.sorted:

```
samtools sort Oct4.bam Oct4.sorted
```

Index the sorted file.

```
samtools index Oct4.sorte.bam
```

Day 2

We already know that in order to load a BAM file onto IGV we need to have this file sorted by genomic location and indexed. Here’s a reminder of the commands to perform these: Sort the BAM file using samtools:

```
samtools sort [bam file to be sorted] [prefix of sorted bam output file]
```

Index the sorted file.

```
samtools index [sorted bam file]
```

On Thursday, we will cover the basics of Epigenetics.
1. Given that we used the following command to align the 2cells dataset:

```
 tophat --solexa-quals -g 2 -p 8 --library-type fr-unstranded -j
    annotation/Danio_rerio.Zv9.66.spliceSites -o tophat/ZV9_2cells
    genome/ZV9 data/2cells_1.fastq data/2cells_2.fastq
```

What is the command to align the ‘6h’ dataset? Run this command on the terminal. ________________________________

Note: You will have to change the input fastq files and the output folder. If you don’t change the output folder, then these results will overwrite the ones for the 2cells dataset.
Bonus Exercise II

- Filter the `2cells` BAM file to only contain uniquely aligned reads
  - Hint: use `samtools view` to keep only those with a mapping score equal to 255
- Then remove duplicates from the `2cells` BAM files using:
  - `picard-tools MarkDuplicates`
- Extract the splice junctions from this deduplicated BAM file
  - Hint: have a look at this thread: [https://www.biostars.org/p/12626/](https://www.biostars.org/p/12626/)
- Run the guided Cufflinks transcriptome assembly on this new BAM file
- Compare the two `transcripts.gtf` files using `cuffcompare`
STEP 3: teach

A2. Teach and/or support learning
What teaching taught me

- Leave personal preferences/biases out of the classroom
- Inspire learners with your successes, but teach them also through your failures/mistakes
- Make time for project-specific questions
  - Either through poster or Q&A sessions
- Enable continuation in teaching
  - Emails exchange, create post-workshop forum, course material distribution, online help, bioinformatics clinic
For the trainer:
- Good feedback
- Request for more (advanced) training activities
- Personal satisfaction/fulfillment

For the learner:
- Confidence
- Common language in a multi-disciplinary field

For the training community:
- (Hopefully) New bioinformatic trainers
A3. Assess and give feedback to learners

- Difficult and poorly incorporated

- Real-time assessment during the hands-on practicals
  - Software carpentry post-it system on multiple-choice questions
  - Real-time collection of answers and instant plotting
Acknowledgements

Prof. Willem H. Ouwehand
Dr Gabriella Rustici
Dr Remco Loos
Dr Steven Wilder
FAQs

Questions

- Have people walked out of your training workshops before the end?

Answers

- Absolutely!! What is important is to identify the reason. Was the course poorly described? If yes, how can it be improved?