MDA 2013

IX International Course of Massive Data Analysis FOR GENOMICS

Course Presentation

EMBL-EBI

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

UNIVERSITY OF CAMBRIDGE

INB

Ignacio Medina
im411@cam.ac.uk

Presentation

Index

- Introduction
- Agenda
- Analysis pipeline
- Some considerations

Introduction
Who we are

- Teachers:
  - David Montaner: Head of the Biostatistics Unit at CIPF (Valencia, Spain)
  - Marta Bleda: Computational Biologist and Data Analyst at Department of Medicine, Addenbrooke's Hospital (University of Cambridge)
  - F. Javier Lopez: Bioinformatician and Software Engineer at EMBL-EBI Variation team (Cambridge, UK)
  - Ignacio Medina: Head of Computational Biology Lab, HPCS University of Cambridge, UK. Also, Scientific Collaborator at EMBL-EBI Variation team (Cambridge, UK)
  - Joaquin Dopazo: Head of the Computational Genomics Department at CIPF (Valencia, Spain)
- Everything started at Joaquin Dopazo's group at CIPF:
  - http://bioinfo.cipf.es/
- More than 10 years of experience in microarrays & NGS data analysis and developing methodologies and bioinformatics tool for data analysis. Many suites and tools developed: GEPAS, Babelomics, Genome Maps, BierApp, VARIANT, ...
- More than 60 papers in the last 8 years in peer reviewed journals: NAR, Bioinformatics, Nat. Biotech., …
- Many active collaborations with experimental and clinic groups
- Many international courses run during last years: *Massive Data Analysis (MDA)*

Introduction
Goals, ambitious

- To learn the basics to understand and be able to conduct a standard NGS data analysis from scratch in a Linux environment
- To know and understand the different data analysis pipelines and formats (FASTQ, SAM/BAM, VCF)
- To preprocess and perform QC of raw and processed data
- To learn and use the **most widely used** tools to perform NGS data analysis and visualization
- To learn the basics of the functional interpretation of variant (DNA re-sequencing) and RNA-seq analysis
- Optionally, learn how to install NGS software in Linux and how to tune up data analysis pipelines by simulating data

Program
First day

- 09:30 Course presentation
- 09:45 Genomic Data Analysis Overview
- 10:30 Introduction to NGS data analysis & GNU/Linux shell
- 11:00 *Coffee Break*
- 12:00 FastQ Quality Control for NGS Raw Data (theory)
- 12:45 *Lunch Break*
- 14:00 FastQ Quality Control for NGS Raw Data (hands-on)
- 14:45 Mapping NGS Reads for Genomic and Transcriptomics Studies (theory)
- 15:30 *Tea Break*
- 15:45 Mapping NGS Reads for Genomic and Transcriptomics Studies I (hands-on)
- 17:00 Finish
- 17:00 Optional: A more advanced Linux session. NGS software Installation (1h)

Program
Second day

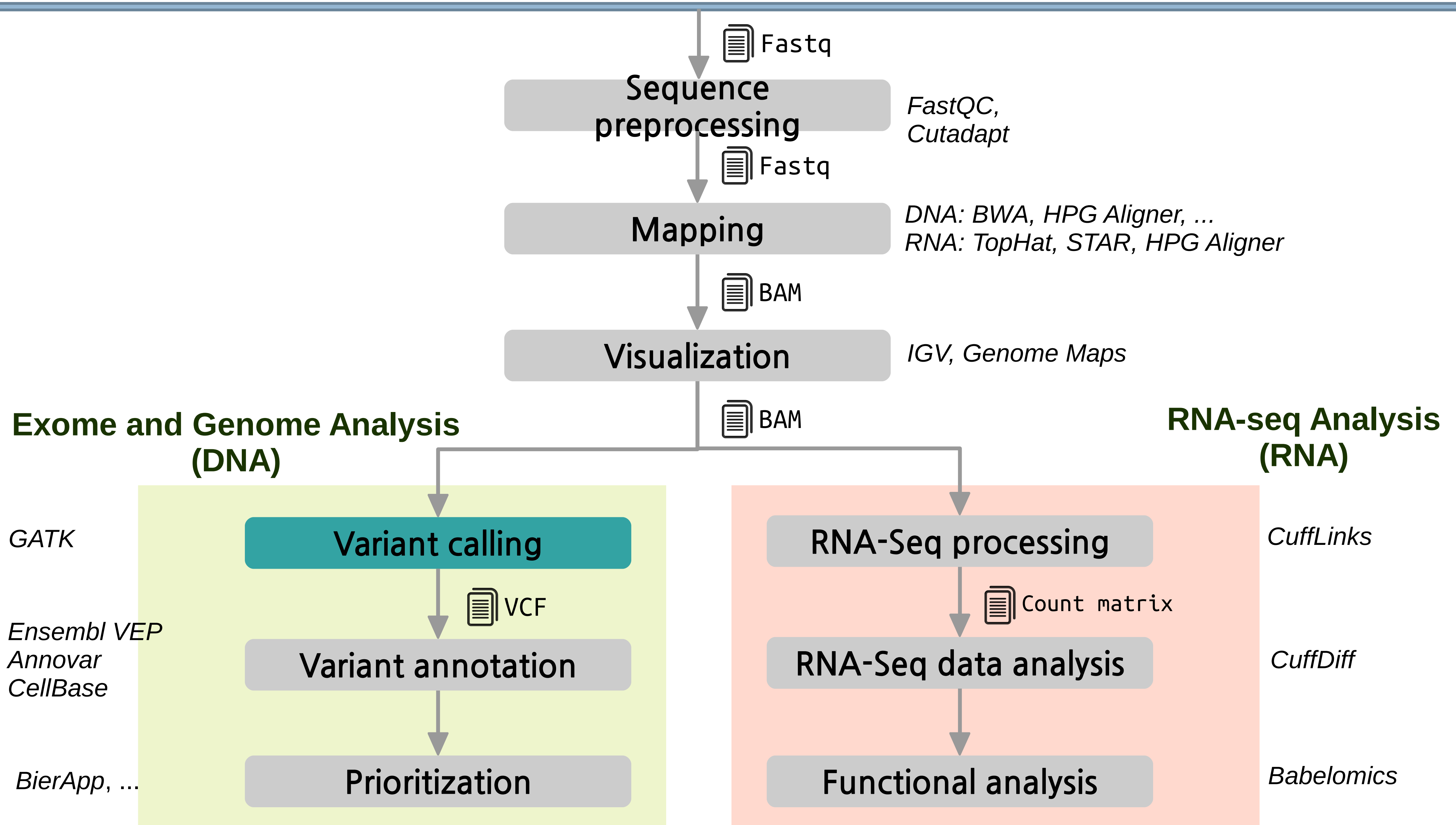
- 09:30 Mapping NGS Reads for Genomic and Transcriptomics Studies II (hands-on)
- 10:30 Visualization of NGS data (theory)
- 11:00 *Coffee Break*
- 11:15 Visualization of NGS data (hand-on)
- 11:45 Variant Calling (SNPs & INDELs) and Variant Visualization (VCF) I
- 12:30 *Lunch Break*
- 14:00 Variant Calling (SNPs & INDELs) and Variant Visualization (VCF) II
- 14:45 Variant Annotation (theory)
- 15:30 *Tea Break*
- 15:45 Variant Annotation (hands-on)
- 16:15 Variant prioritization
- 17:00 Finish

Program

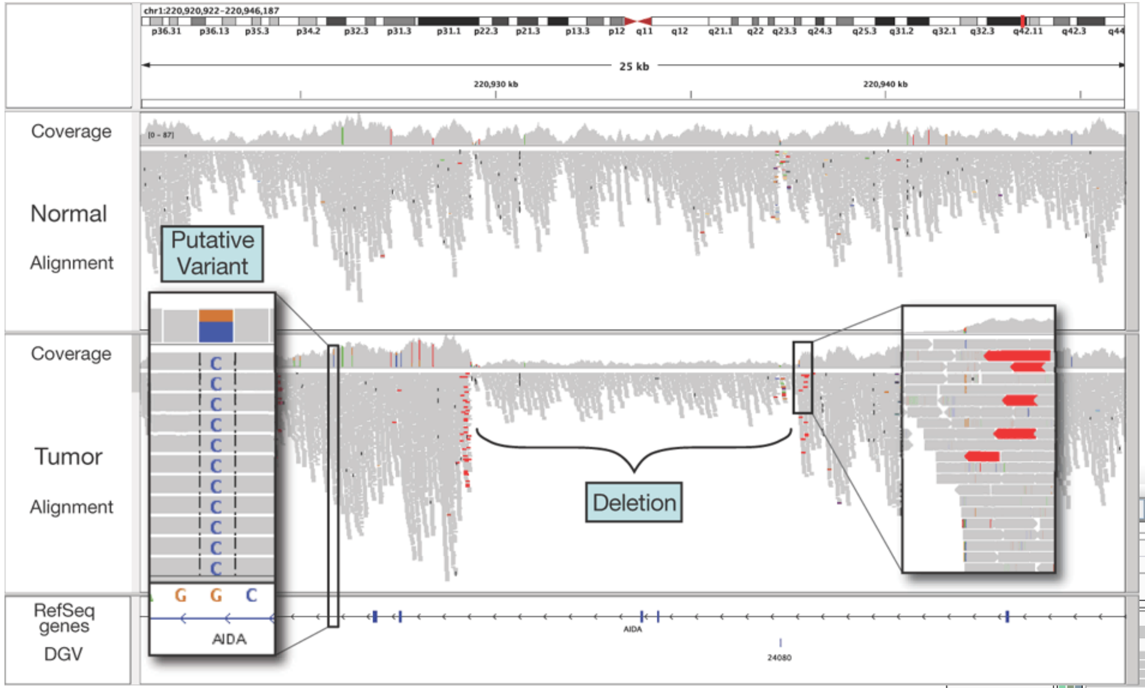
Third day

- 09:30 RNA-seq data preprocessing
- 11:00 *Coffee Break*
- 11:15 RNA-Seq Quantification and Isoforms Finding
- 12:30 *Lunch Break*
- 14:00 Functional Analysis
- 15:00 *Tea Break*
- 15:15 Exercises and questions
- 17:00 Finish

Analysis pipeline

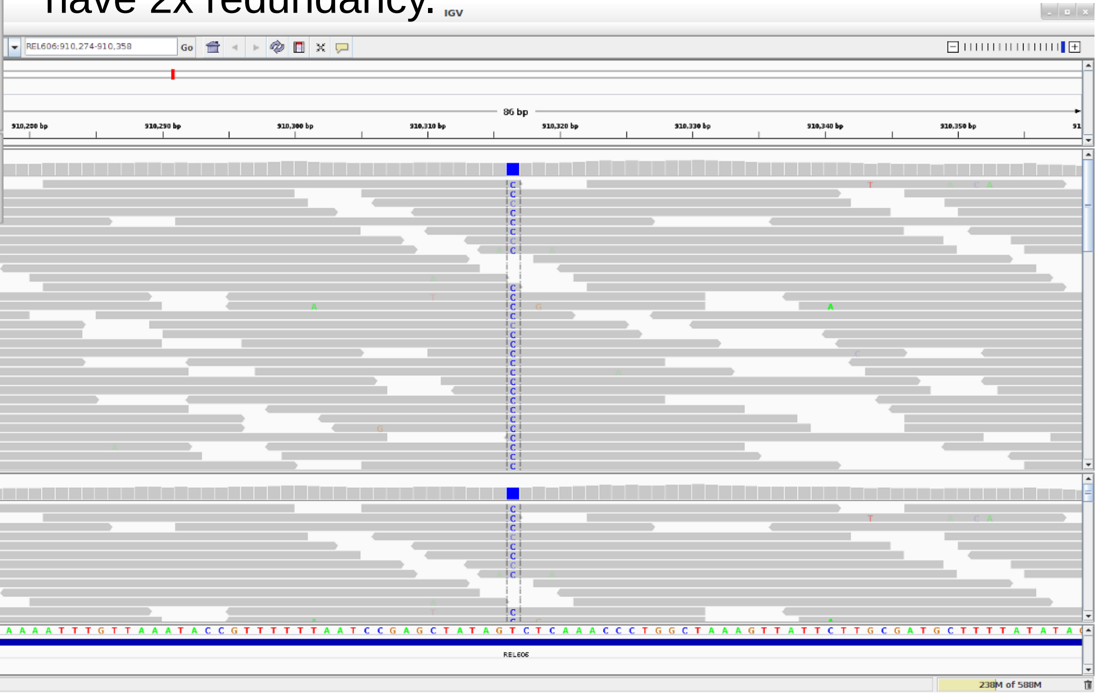


Analysis pipeline
Aligning reads, the coverage



Coverage (read depth or depth) is the ***average*** number of reads representing a given nucleotide in the reconstructed sequence.

It can be calculated from the length of the original genome (G), the number of reads(N), and the average read length(L) as ***NxL/G***. For example, a hypothetical genome with 2,000 base pairs reconstructed from 8 reads with an average length of 500 nucleotides will have 2x redundancy.



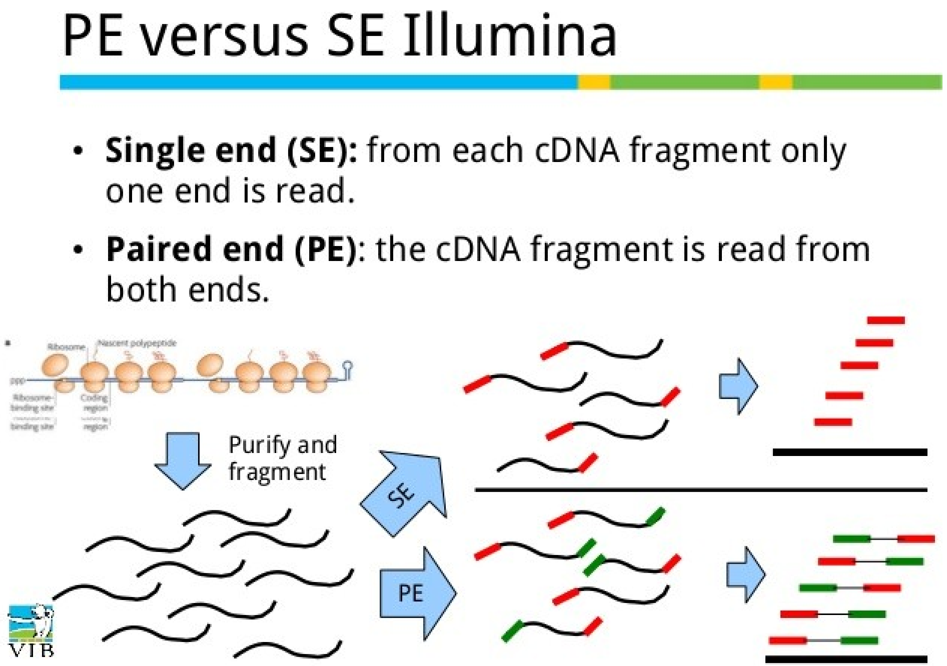
Useful for:

- Error sequencing detection
- Copy number detection
- Genotyping
- ...

Current re-sequencing projects target to 40x depth

Analysis pipeline

paired-end vs single*-end* alignment



PE versus SE Illumina

- **Single end (SE):** from each cDNA fragment only one end is read.
- **Paired end (PE)**: the cDNA fragment is read from both ends.

Paired-end sequencing:

- Improves read alignment and therefore variant calling
- Helps to detect structural variation
- Can detect gene fusions and splice junctions
- Useful for *de novo* assembly
- …

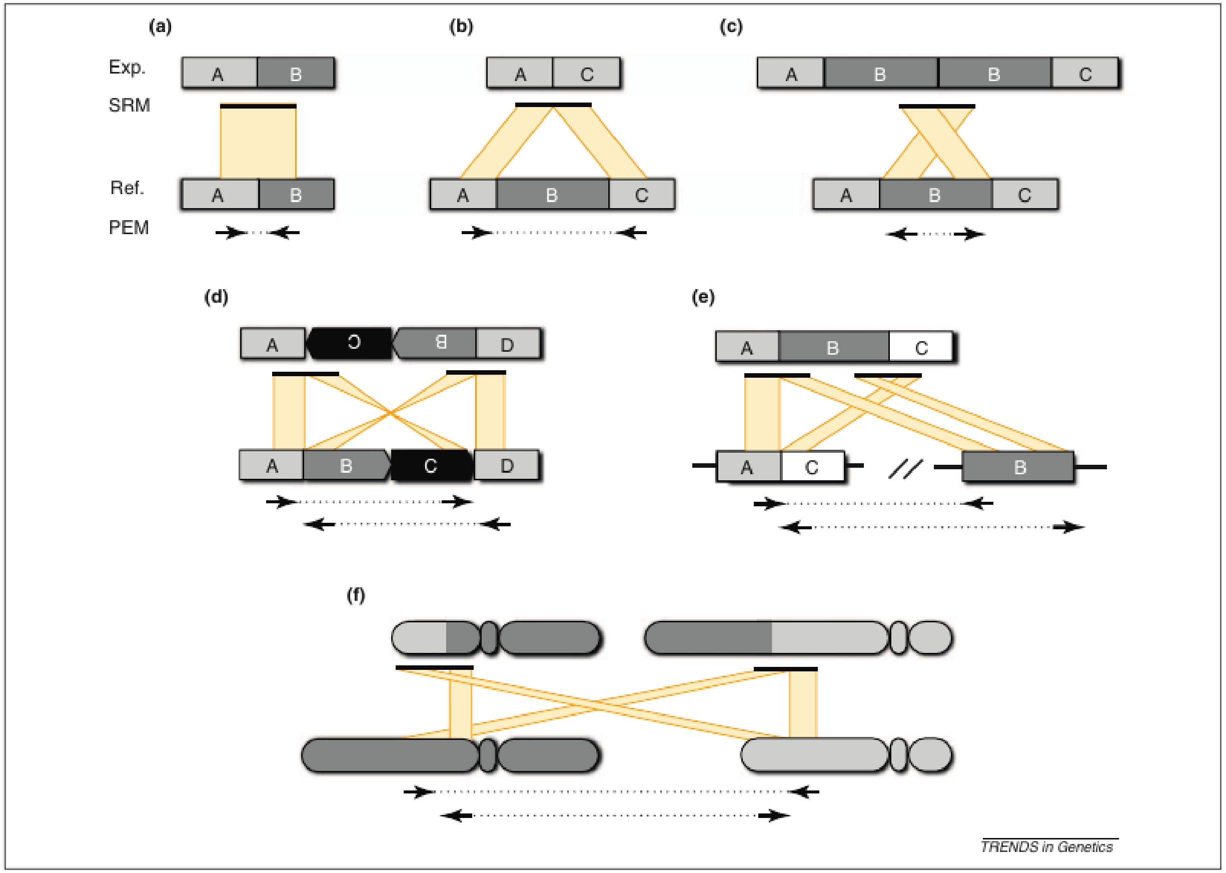


Figure 2. Detecting canonical structural variation (SV) breakpoints through sequencing. When DNA sequences are collected from an experimental (Exp.) genome and aligned to a reference (Ref.) genome, each structural variant class generates a distinct alignment pattern. The patterns observed for paired-end mapping (PEM) and split-read mapping (SRM) are illustrated when both genomes have identical structure **(a)**, and cases where the experimental genome contains a deletion **(b)**, a tandem duplication **(c)**, an inversion **(d)**, a transposon insertion **(e)** or a reciprocal translocation **(f)**. PEM relies upon readpairs whose unsequenced portion (dotted lines) spans a SV breakpoint. When aligned to the reference genome, the alignment distance and orientation of such readpairs indicate the type of rearrangement that has occurred. Reads that map to the plus strand are shown as right-facing arrows, those that map to the negative strand as leftward-facing arrows. All examples depict Illumina paired-end sequence data, where in the absence of SV the normal concordant orientation is plus for the leftmost read and minus for the rightmost read. Note that the expected orientation is different for Illumina mate-pair libraries and for other sequencing platforms, such as SOLiD. In the case of a deletion **(b)**, the readpairs ends will align much farther apart than expected for the DNA library. In contrast to PEM, SRM depends on contiguous sequences that contain an SV breakpoint. Consequently, the sequences before and after the breakpoint will align to disjoint regions of the reference genome. In contrast to PEM, breakpoints are identified at single-base resolution.

Some considerations

- NGS data can be big, very big, huge! Biology is now a Big Data science
  - **There are not many web-based or graphical applications** to perform analysis *yet*, sorry.
  - Most tools developed to work on **Linux**, many command line programs
- How to work in NGS?
  - Small datasets (<1TB): workstations
  - Medium sized datasets (<100TB): **clusters**
  - Big datasets (100TB-20PB): big clusters and/or cloud based solutions
- Exercises during this course the NGS alignment will be done using the human **chromosome 21** as a reference genome. By doing this we can speed up exercises and avoid using too much memory. Under real circumstances, when using the standard reference genome, all the commands are exactly the same
- Software **has been already installed** to save time, so you are not expected to download and install all the software it is going to be used. However, it's usually good to learn the basics of software installation in Linux, there is an optional session at the end of the first day for those that want to learn how to install NGS software in a standard Linux

What about you?

Brief presentation

- Who are you?
- Which is your background?
- Which is your interest?
- What do you expect of this course?