IX International Course of Massive Data Analysis FOR GENOMICS

Course Presentation

EMBL-EBI

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

UNIVERSITY OF CAMBRIDGE

INB

Ignacio Medina
imedina@ebi.ac.uk

Presentation

Index

- Introduction
- Program
- Analysis pipeline
- Some considerations

Introduction
Who we are

- Teachers:
  - David Montaner: Head of the Biostatistics Unit at CIPF (Valencia, Spain)
  - Marta Bleda: Bioinformatician and Data Analyst at Addenbrooke's Hospital (Univ. Cambridge)
  - Ignacio Medina: Bioinformatician and Project Manager at EMBL-EBI Variation (Cambridge, UK)
- Everything started at Joaquin Dopazo group at CIPF:
  - http://bioinfo.cipf.es/
- More than 8 years of experience in microarrays and NGS data analysis, and also developing metholodogies 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 collaborations with experimental and clinic groups
- Many international courses run 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 analysis pipelines and data 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

Ignacio Medina
imedina@ebi.ac.uk

Program
First day

- 09:30 Presentation
- 10:00 Introduction to NGS Technologies
- 10:30 Introduction to GNU/Linux shell
- 11:00 Coffee Break
- 11:15 Quality Control for NGS Raw Data (FASTQ) and Data Preprocessing
- 12:30 Lunch Break
- 14:00 Mapping NGS Reads for Genomic and Transcriptomics Studies I
- 15:30 Tea Break
- 15:45 Mapping NGS Reads for Genomic and Transcriptomics Studies II
- 17:15 Finish

Program
Second day

- 09:30 Visualization of NGS data (BAM files)
- 11:00 Coffee Break
- 11:15 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
- 15:30 Tea Break
- 15:45 Variant prioritization
- 16:30 Big Data analysis and visualization
- 17:15 Finish

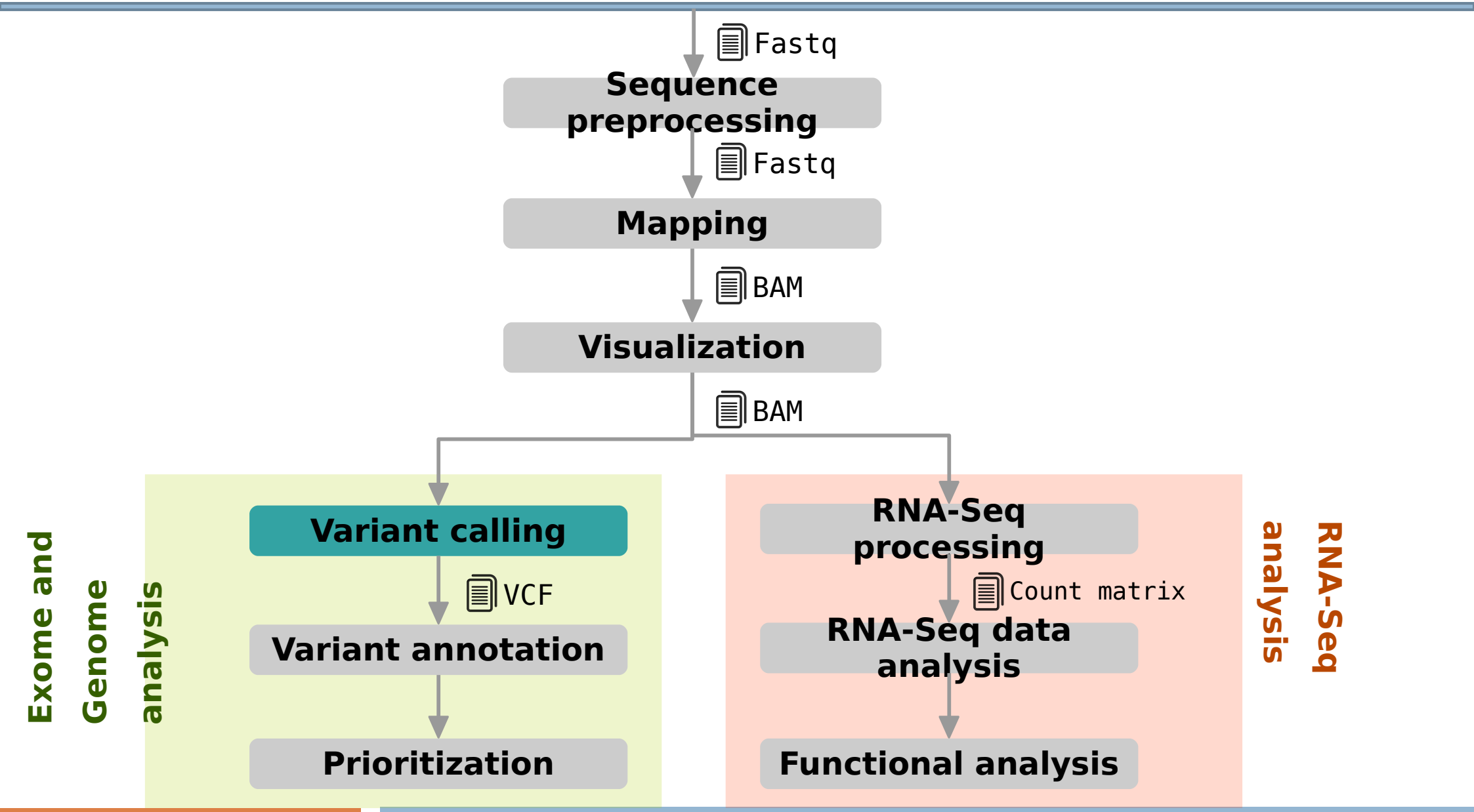
Ignacio Medina
imedina@ebi.ac.uk

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

Ignacio Medina
imedina@ebi.ac.uk

Analysis pipeline



Ignacio Medina
imedina@ebi.ac.uk

Presentation

Some considerations

- NGS data can be big, very big, huge! Biology is now a Big Data science
  - **No web 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 (<100-200TB): **clusters**
  - Big datasets (200TB-5PB): big clusters and cloud based solutions
- Exercises during this course will be done using with human **chromosome 21** to speed up analysis and not use too much memory. Under real circumstances using the whole genome the commands are exactly the same
- Software **has been already installed** to save time so you are not expected to download and install the software we are going to use. However it's usually needed 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?