



IX International Course of **Massive Data Analysis FOR GENOMICS**

Course Presentation



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
 - Marta Bleda: Bioinformatician
 - Ignacio Medina: Researcher and Project Manager at EMBL-EBI Variation
- From Joaquin Dopazo group at CIPF:
 - <http://bioinfo.cipf.es/>
- More than 8 years of experience in microarrays and NGS data analysis and Bioinformatic tool development, and developing methodologies for data analysis
- Many suites and tools developed: GEPAS, Babelomics, Genome Maps, VARIANT, ...
- More than 50 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 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 data
- To learn how to install and use the ecosystem of tools to perform NGS data analysis
- To tune up data analysis pipelines by simulating data
- To perform some basics functional interpretation of variant (DNA re-sequencing) and RNA-seq analysis

Program

First day

- 09:30 Presentation
- 10:00 Introduction to NGS Technologies for Genomic Studies
- 10:30 Introduction to GNU/Linux shell
- 11:00 Coffee Break
- 11:30 Quality Control for NGS Raw Data (FASTQ) and Data Preprocessing
- 12:30 Lunch Break
- 14:00 Mapping NGS Reads for Exome and Transcriptomics Studies I
- 16:00 Tea Break
- 16:15 Mapping NGS Reads for Exome and Transcriptomics Studies II
- 17:15 Finish

Program

Second day

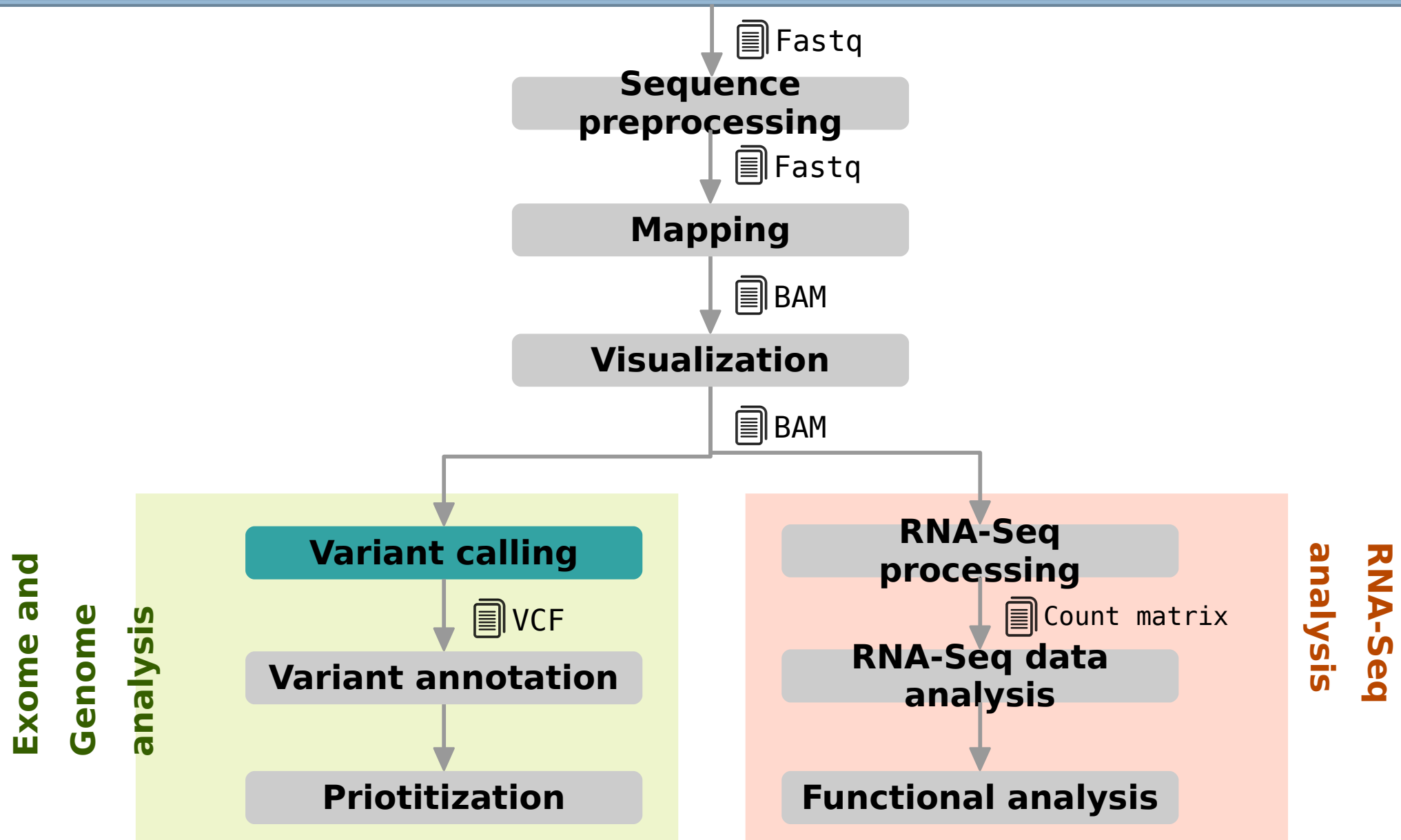
- 09:30 Visualization of NGS data (BAM files)
- 11:00 Coffee Break
- 11:30 Variant Calling (SNPs & INDELs) and Variant Visualization (VCF) I
- 12:30 Lunch Break
- 14:00 Variant Calling (SNPs & INDELs) and Variant Visualization (VCF) II
- 15:15 Variant Annotation
- 16:00 Tea Break
- 16:30 Variant prioritization
- 17:30 Finish

Program

Third day

- 09:30 RNA-seq data preprocessing
- 11:00 Coffee Break
- 11:30 RNA-Seq Quantification and Isoforms Finding
- 12:30 Lunch Break
- 14:00 Functional Analysis
- 15:00 Exercises and questions
- 16:00 Tea Break
- 16:15 Exercises and questions
- 17:00 Finish

Analysis pipeline



Some considerations

- NGS data is 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 (<40-50TB): **clusters**
 - Big datasets (50TB-1PB): distributed 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 to install software

What about you?

Brief presentation

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