NGS Bioinformatics

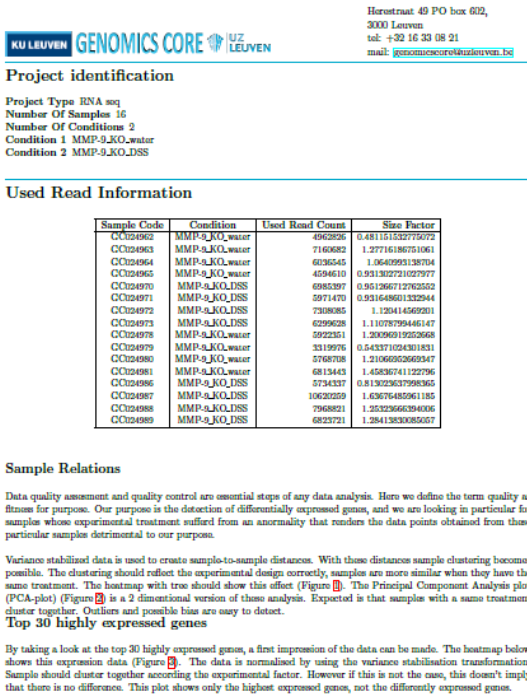
Álvaro Cortés Calabuig
October 2019

GENOMICS CORELEUVEN

Overview

- This workshop: why?
- Bioinformatics NGS terminology
- NGS bioinformatics pipelines

Typical RNA-Seq Experiment Produces:



KU LEUVEN GENOMICS CORE UZ LEUVEN

Herestraat 49 PO box 602,
3000 Leuven
tel: +32 16 33 08 21
mail: genomicscore@uzleuven.be

Project identification

Project Type RNA seq
Number Of Samples 16
Number Of Conditions 2
Condition 1 MMP-9_KO_water
Condition 2 MMP-9_KO_DSS

Used Read Information

Sample Code	Condition	Used Read Count	Size Factor
CC124962	MMP-9_KO_water	4962936	0.481151532775072
CC124963	MMP-9_KO_water	7160682	1.27716186751061
CC124964	MMP-9_KO_water	6036545	1.06409931138704
CC124965	MMP-9_KO_water	4594610	0.931302721027977
CC124970	MMP-9_KO_DSS	6985397	0.951266712762552
CC124971	MMP-9_KO_DSS	5971470	0.931648601332944
CC124972	MMP-9_KO_DSS	7308085	1.12914569201
CC124973	MMP-9_KO_DSS	6299628	1.11078799446147
CC124978	MMP-9_KO_water	5922351	1.29096919252668
CC124979	MMP-9_KO_water	3319976	0.543371624301831
CC124980	MMP-9_KO_water	5768708	1.21066952669347
CC124981	MMP-9_KO_water	6813443	1.4583674122796
CC124986	MMP-9_KO_DSS	5734337	0.813029637998365
CC124987	MMP-9_KO_DSS	10620259	1.63676485961185
CC124988	MMP-9_KO_DSS	7968821	1.25329666394006
CC124989	MMP-9_KO_DSS	6823721	1.28413830085057

Sample Relations

Data quality assessment and quality control are essential steps of any data analysis. Here we define the term quality a fitness for purpose. Our purpose is the detection of differentially expressed genes, and we are looking in particular fo samples whose experimental treatment suffers from an anormality that renders the data points obtained from thes particular samples detrimental to our purpose.

Variance stabilized data is used to create sample-to-sample distances. With these distances sample clustering become possible. The clustering should reflect the experimental design correctly, samples are more similar when they have th same treatment. The heatmap with tree should show this effect (Figure 1). The Principal Component Analysis plo (PCA-plot) (Figure 2) is a 2 dimentional version of these analysis. Expected is that samples with a same treatmen cluster together. Outliers and possible bias are easy to detect.

Top 30 highly expressed genes

By taking a look at the top 30 highly expressed genes, a first impression of the data can be made. The heatmap below shows this expression data (Figure 3). The data is normalised by using the variance stabilisation transformation Sample should cluster together according the experimental factor. However if this is not the case, this doesn't impl that there is no difference. This plot shows only the highest expressed genes, not the differently expressed genes.

10 report files: differentially expressed genes, reads quality report, counting report, …

3GB of raw and analyzed data: Fastq, BAM, counts, normalized counts, gene expression files, heatmaps…

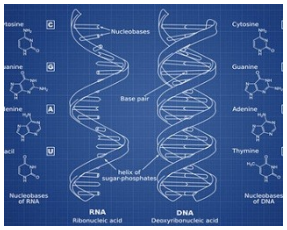
GENOMICS CORELEUVEN

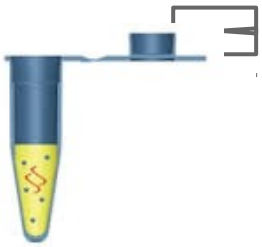
Results Interpretation & Analysis

GENOMICS CORELEUVEN

NGS Bioinformatics - This Workshop

Experimental Design





Library Preparation

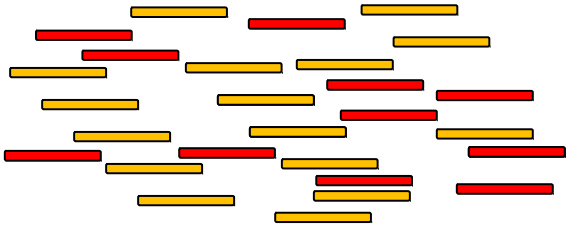
Sequencing

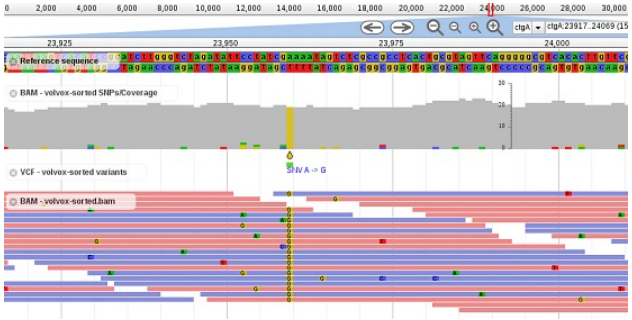
Bioinformatics

Follow up and support

GENOMICS CORELEUVEN

NGS Bioinformatics





NGS bioinformatics: interpretation and analysis of NGS data using informatics tools

GENOMICS CORELEUVEN

NGS Bioinformatics

GENOMICS CORELEUVEN

Overview

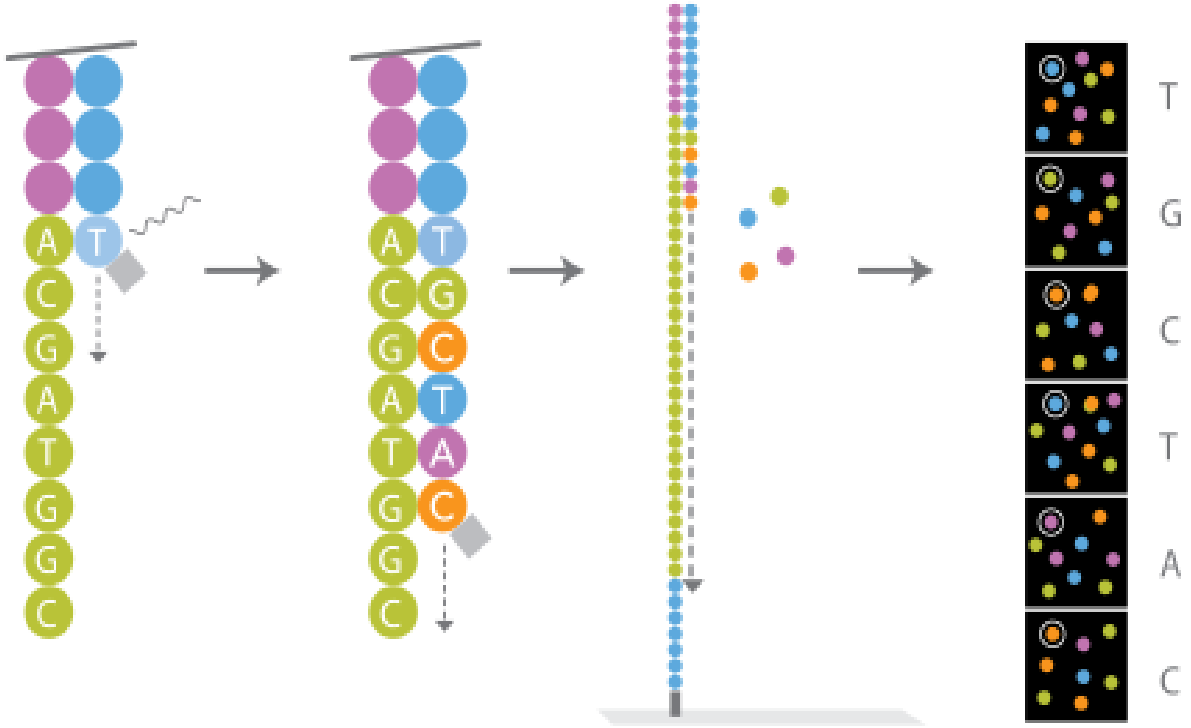
- This workshop: why?
- Bioinformatics NGS terminology
- NGS bioinformatics pipelines

A day in bioinformatics: Terminology

- Poly-A tail
- Ontology
- Index
- Barcode
- Hadoop
- Xenograft Murine infiltration
- Flow Cell
- Demultiplexing
- VCF File
- Ortholog gene
- BAM
- Negative binomial distribution
- Cluster
- snRNA
- Bonferroni Correction
- Barcode

(De)Multiplexing (I)

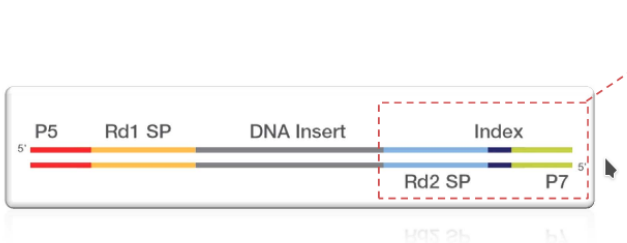
Base Calling

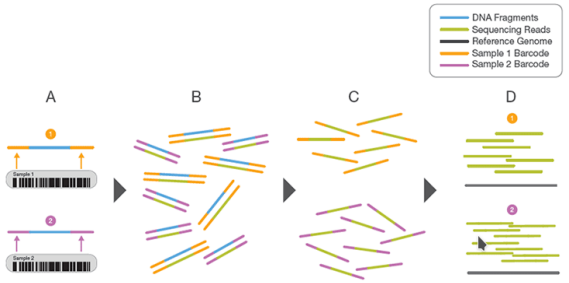


GENOMICS CORELEUVEN

(De)Multiplexing (II)

Multiple samples can be *pooled together or multiplexed* into one or more flowcells





GENOMICS CORELEUVEN

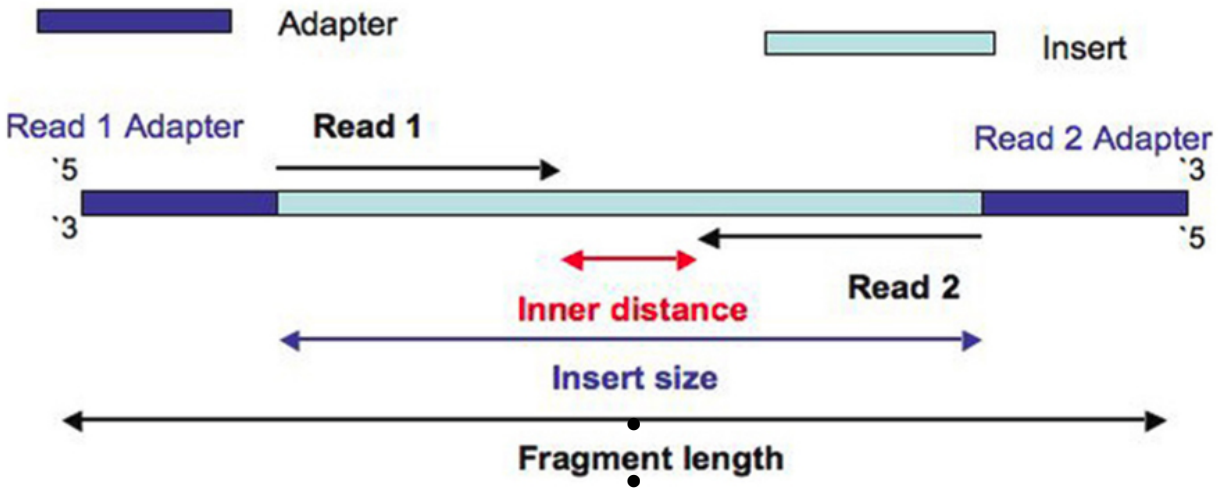
From DNA Molecules to Digital Reads

DNA fragments

Reads

GENOMICS CORELEUVEN

Reads and Fragments



Fragment: biological entity
Read: bioinformatic concept

Fragment: the DNA template + adapters that were loaded on the sequencing machine (is not completely sequenced)

Read: a raw sequence originating from a sequencing machine

Single Read: Sequencing only from one end

Paired-end: Sequencing starting from both ends of the insert

GENOMICS CORELEUVEN

Fastq format & Files

The result of demultiplexing is one (or two for paired-end reads) fastq files containing raw reads

- Fastq files are human readable (not binary) text files.
- Referred as raw data.
- Fastq files are the raw diamonds of any NGS bioinformatics project
- Fastq files are often compressed using zip or gunzip

```
-rwxr-xr-x  1 vsc31439 lp_biogenomics  32G 17 janv. 15:05 GC036462.R1.fastq.gz
-rwxr-xr-x  1 vsc31439 lp_biogenomics  39G 17 janv. 15:49 GC036462.R2.fastq.gz
-rwxr-xr-x  1 vsc31439 lp_biogenomics  31G 17 janv. 16:25 GC036463.R1.fastq.gz
-rwxr-xr-x  1 vsc31439 lp_biogenomics  37G 17 janv. 17:07 GC036463.R2.fastq.gz
-rwxr-xr-x  1 vsc31439 lp_biogenomics  31G 17 janv. 17:41 GC036464.R1.fastq.gz
-rwxr-xr-x  1 vsc31439 vsc31439        37G 17 janv. 18:21 GC036464.R2.fastq.gz
```

4x

```
vsc31420@hpc-p-login-1 /staging/leuven/stg_00019/full_genomes/test_ws 11:48 $ ls -lha  GC036463.R1.fastq
-rwxr-xr-x 1 vsc31420 vsc31420 134G 29 mars  11:24 GC036463.R1.fastq
```

GENOMICS CORELEUVEN

Raw Reads

```
1 @HISEQ:574:C6VG2ANXX:2:1110:1400:2194 1:N:0:ATCACG
2 GGGGGATTCTCACTAGGTCTCAAGGTCTCTCACTCTCGGTAGTGTTCCCAG
3 +
4 CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
```

- Line 1: Read identifier and is followed by a sequence that
  - Unique, platform dependent
  - Begins with a '@' character

@HISEQ:574:C6VG2ANXX:2:1110:1400:2194 1:N:0:ATCACG

- Instrument
- Run, flowcellID, lane, tile, pos, read, filter, control
- Index

- Line 2: Raw sequence of nucleotides: read
- Line 3: begins with a '+' character and is optionally followed by the same sequence identifier.
- Line 4: Quality values for the sequence in Line 2

Reads and Fastq Format

- Fastq format?
  - Plain-text file, where each read and complementary information occupies 4 consecutive lines
  - Typical size 500M compressed, 2000M unzipped

```
@HISEQ:574:C6VG2ANXX:2:1110:1400:2194 1:N:0:ATCACG
GGGGGATTCTCACTAGGTCTCAAGGTCTCTCACTCTCGGTAGTGTTCCCAG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@HISEQ:574:C6VG2ANXX:2:1110:1560:2177 1:N:0:ATCACG
ATGGTCCAGCAAGGGGTATGCTGAGAAGGGGAGCAGTTCAGAACCCATCAG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@HISEQ:574:C6VG2ANXX:2:1110:1583:2223 1:N:0:ATCACG
CTACCTTCACTATCAACATAGCAAACACACCTTTAGCTCCAGCTATTAACA
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@HISEQ:574:C6VG2ANXX:2:1110:1609:2245 1:N:0:ATCACG
AGCTTAAGAGGCACTAGACACAGCCAGCTTCTTCAGGTGATCCATGAACAC
```

- Sequencing depth: The total number of sequences generated for a sample.
  - Usually expressed in fragments or reads
  - How deep to sequence? Experiment!

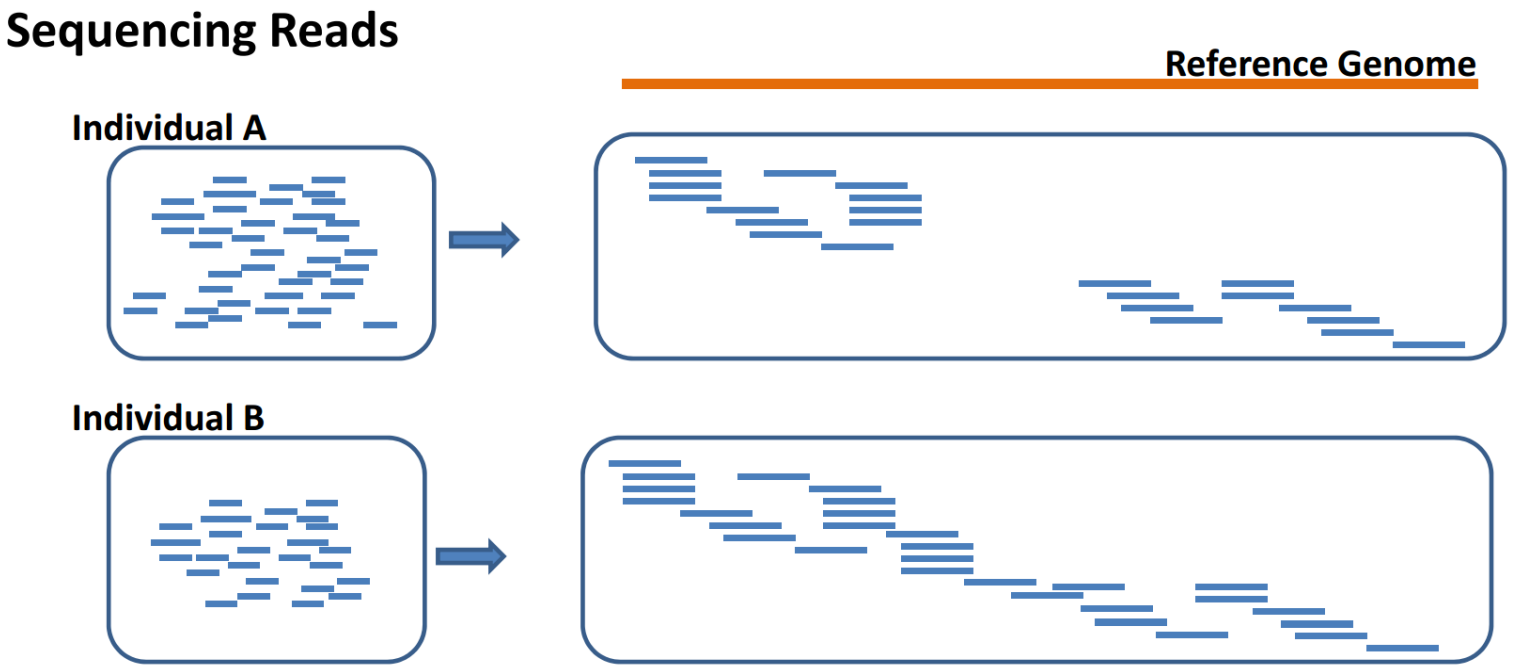
Reference Sequence in Fasta Format

- genome.fa human-readable nucleotide sequence
- Species dependent
- Mouse genome: 2.6GB
- Evolves

GENOMICS CORELEUVEN

Mapping to Reference Genome

Mapping refers to the process of aligning short reads to a reference sequence



GENOMICS CORELEUVEN

Assembly

- The **generation of a reference**, from scratch (*de novo*) or reference assisted.
- Overlapping reads are merged to contigs (smallest unitable unit without unknown bases)
- Contigs that belong together, but where the connecting sequence is unknown, can be connected to scaffolds, inserting N's for the unknown bases

GENOMICS COREL EUVEN

Computer Cluster

NGS data means big data...means big computing power

- Whole Human Genome:
- 300Gb, gzip file

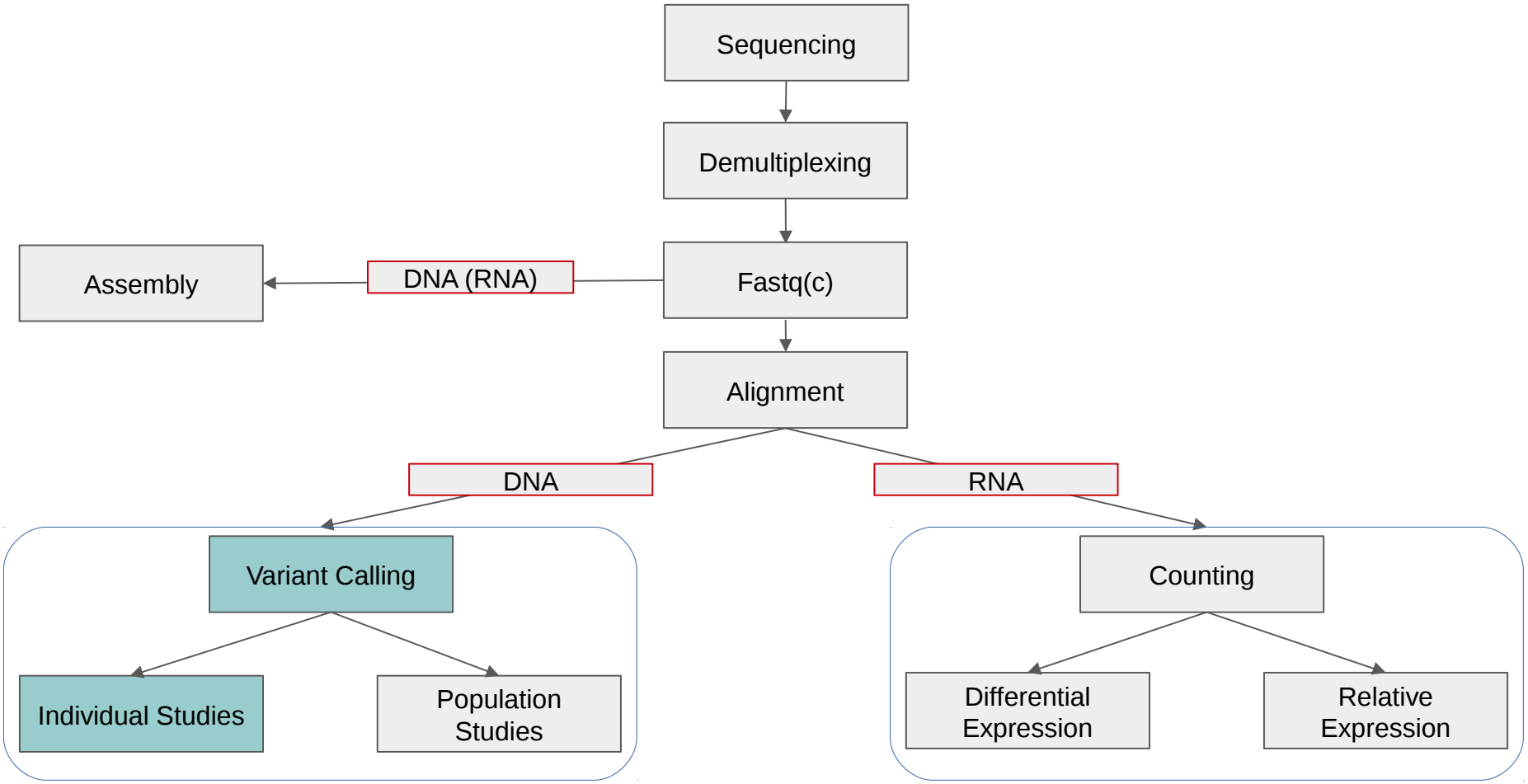
- Exome data:
- 6GB

- RNA-Seq
- 1GB

- NGS data is usually analyzed on a supercomputer or cluster.
  - UZ Leuven: Avalok/Hydra
  - KU Leuven: VSC Flemish Super Computer
  - Google genomics, etc.

GENOMICS CORELEUVEN

NGS Common Pipelines



GENOMICS CORELEUVEN

Digging deeper

- Introduction to RNA-Seq pipelines and bioinformatics
  - Friday the 8th of November
- High performance computing for genomics
  - Hands-on Workshop (Speakers to determine)
  - Friday the 6th of December
  - Organized in conjunction with VSC, Lab. Comput biology, GC
- Introduction to Single-Cell omics
  - January

Thanks!