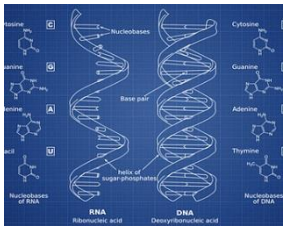
NGS Bioinformatics

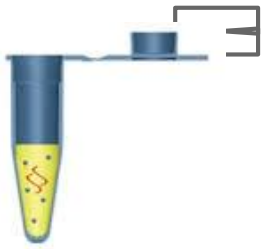
Álvaro Cortés Calabuig
November 2022

GENOMICS CORELEUVEN

NGS Bioinformatics - This Afternoon

Experimental Design





Library Preparation

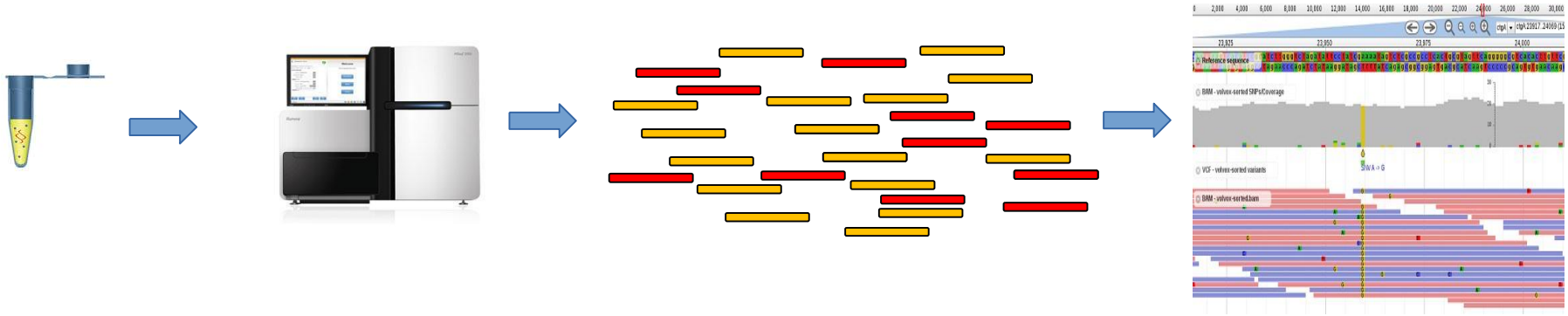
Sequencing

Bioinformatics
- Primary
- Secondary

Follow up and support

GENOMICS CORELEUVEN

NGS Bioinformatics



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

GENOMICS CORELEUVEN

NGS Bioinformatics

What is NGS data?

- Reads produced by a Next Generation Sequencer
- Sequencing of million of short fragments in parallel
- By antonomasia: Illumina Sequencing

What NGS is not?

- Sanger sequencing
- Pacbio, Oxford Nanopore long reads sequencing (LRS)

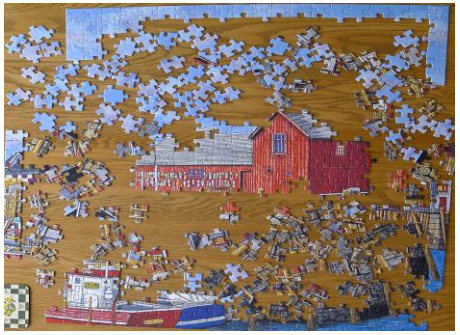
Many of the challenges associated to NGS are also present in LRS analysis

GENOMICS CORELEUVEN

NGS Bioinformatics

NGS Sequencer

Raw reads



NGS Bioinformatics

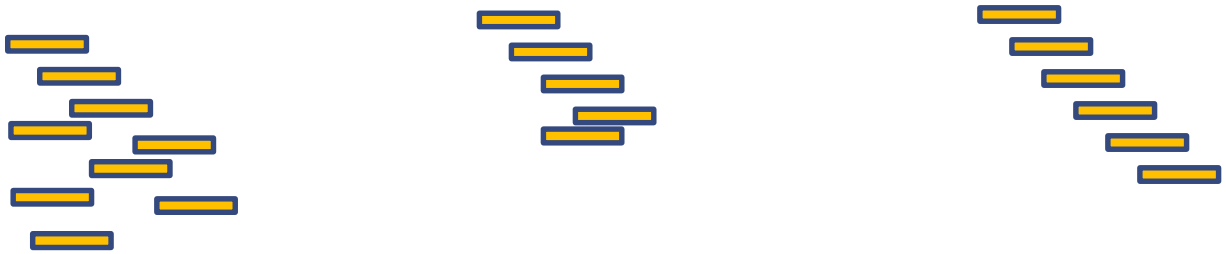
- Some pieces are missing
- Identical pieces
- Pieces fit on multiple locations
- Some pieces do not fit (sequencing errors)
- Pieces from a different puzzle (contamination!)
- Puzzle box or template incomplete or not available

Challenges with NGS data

Reads are too short

Reads contain errors

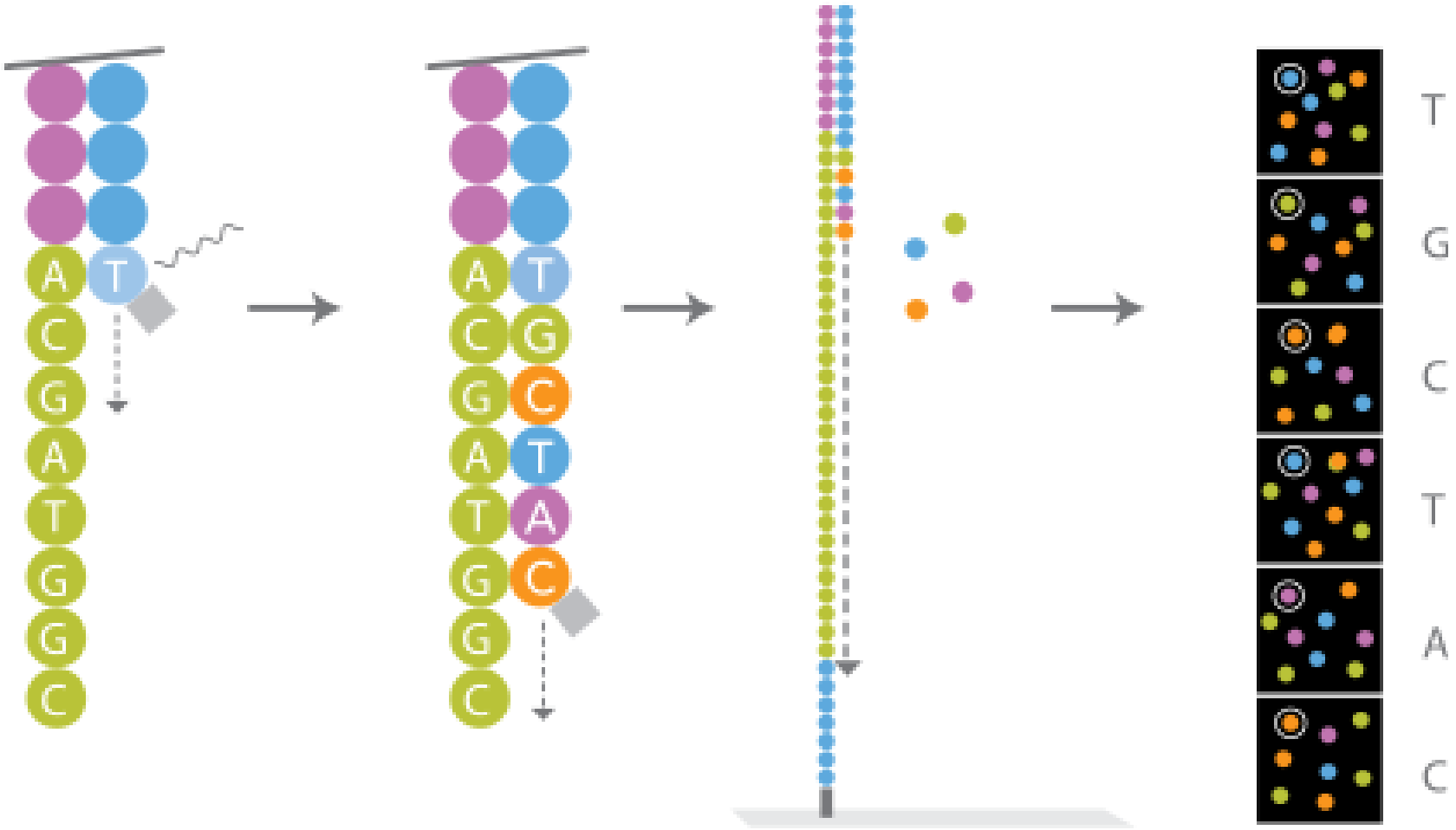
Only fractions of the intended genomic region is sequenced



In this workshop we will cover how these challenges are addressed

GENOMICS COREL EUVEN

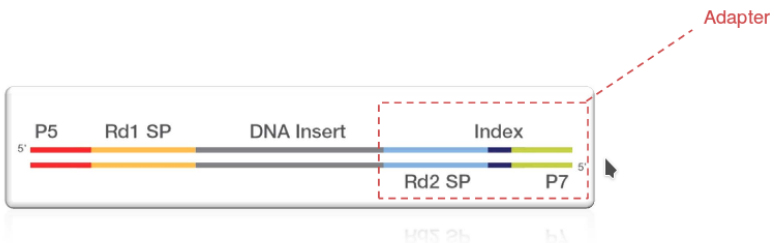
Base Calling

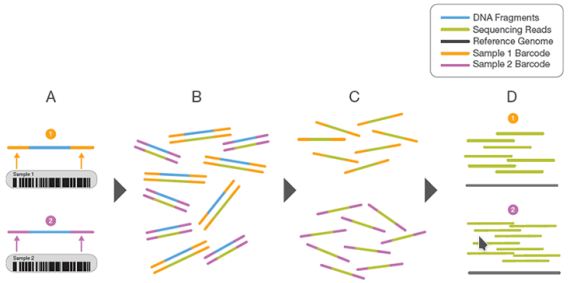


GENOMICS CORELEUVEN

(De)Multiplexing (II)

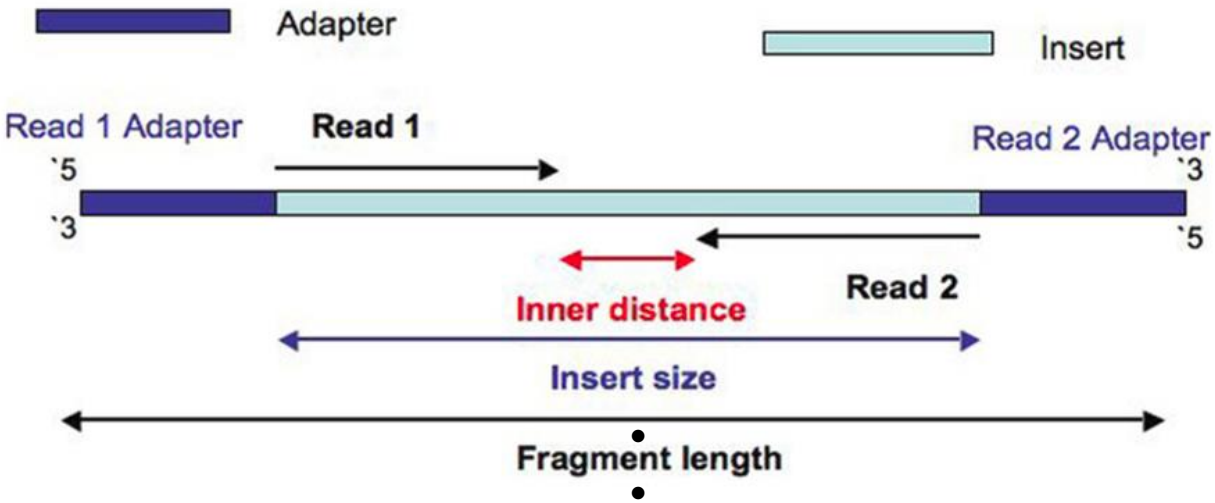
Multiple samples can be *pooled together or multiplexed* into one or more flowcells





GENOMICS CORELEUVEN

Reads and Fragments



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

Reads and Fastq Format

- Fastq format?
  - Plain-text file, where each read and complementary information occupies 4 consecutive lines

```
@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, or
Coverage genomic region

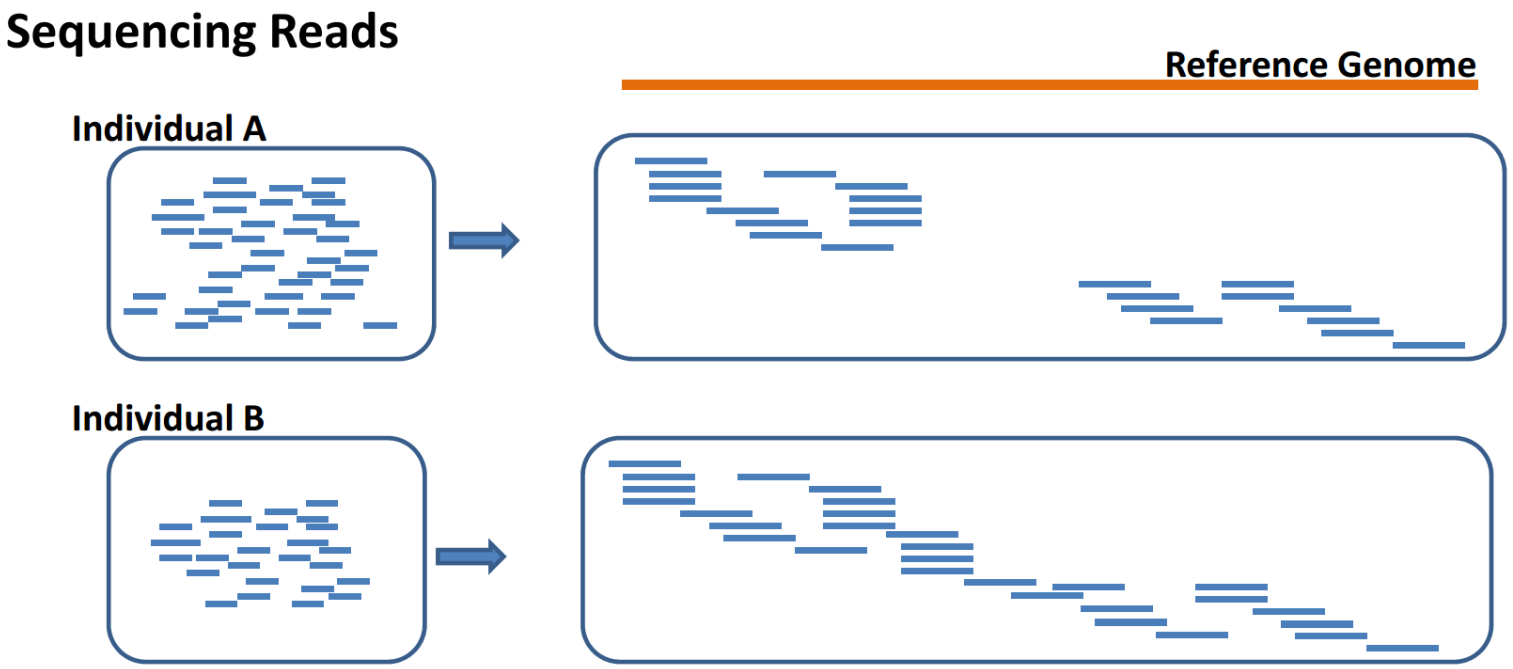
Reference Sequence in Fasta Format

- genome.fa human-readable nucleotide sequence
- Species dependent
- Refinements

```
AATAAGTCAATGGCCTTTCTCTACACAAAGAATAAACAGGCTGAGAAAGAAATTAGGGAA
ACAACACCCTTCTCAATAGTCACAAATAATATAACATATCTCGGCGTGACTCTAACTAAG
GAAGTGAAAGATCTGTATGATAAAAACTTCAAGTCTCTGAAGAAAGAAATTAAAGAAGAT
CTCAGAAGATGGAAAGATCTCCCATGCTCATGGATTGGCAGGATCAATATTGTAAAAATG
GCTATCTTGCCAAAAGCAATCTACAGATTCAATGCAATCCCCATCAAAATTCCAACTCAA
TTCTTCAACGAATTAGAAGGAGCAATTTGCAAATTCATCTGTAATAACAAAAAACCTAGG
ATAGCAAAAAGTCTTCTCAAGGATAAAAGAACCTCTGGTGGAATCACCATGCCTGACCTA
AAGCTTTACTACAGAGCAATTGTGGTAAAAACTGCATGGTACTGGTATAGAGACAGACAA
GTAGACCAATGGAATAGAATTGAAGACCCAGAAATGAACCCACACACCTATGGTCACTTG
ATCTTCGACAAGGGAGCTAAAACCATCCAGTGGAAGAAAGACAGCATTTTCAACAAATGG
TGCTGGCACAACTGGTTGTTATCATGTAGAAGAATGCGAATCGATCCATACTTATCTCCT
TGTACTAAGGTCAAATCTAAATGGATCAAAGAACTTCACATAAAACCAGAGACACTGAAA
CTTATAGAGGAGAAAGTGGGGAAAAGCCTTGAAGATATGGGCACAGGGGAAAAATTCCTG
AACAGAACAGCAATGGCTTGTGCTGTAAGATTGAGAATTGACAAATGGGACCTAATGAAA
CTCCAAAGTTTCTGCAAGGCAAAAGACACCGTCAATAAGAGAAAGAGACCACCAACAGAT
TGGGAAAGGATCTTTACCTATCCTAAATCAGATAGGGGACTAATATCCAACATATATAAA
GAACTCAAGAAGGTGGACTTCAGAAAATCAAACAACCCCATTAAAAAATGGGGCTCAGAA
CTGAACAAAGAATTCTCACCTGAGTTATACCGAATGGCAGAGAAGCACCTGAAAAAATGC
TCAACATCCTTAATCATCAGGGAAATGCAAATCAAAACAACCCTGAGATTCCACCTCACA
CCAGTCAGAATGTCTAAGATCAAAAATTCAGGTGACAGCAGATGCTGGCGAGGATGTGGA
GAAAGAAGAACACTCCTCCATTGTTGGTGGGATTGCAGGCTTGTACAACCACTCTGGAAA
TCCGTCTGGCGGTTCCTCAGAAAATTGGACATAGTACTACCGGAGGATCCAGCAATACCT
CTCCTGGGCATATATCCAGAAGATGCCCCAACTGGTAAGAAGGACACATGCTCCACTATG
TTCATAGCAGCCTTATTTATAATAGCCAGAAGCTGGAAAGAACCCAGATGCCCCTCAACA
GAGGAATGGATACAGAAAATGTGGTACATCTACACAATGGAGTACTACTCAGCTATTAAA
AAGAATGAATTTATGAAATTCCTAGCCAAATGGATGGACCTGGAGGGCATCATCCTGAGT
```

Mapping to Reference Genome

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



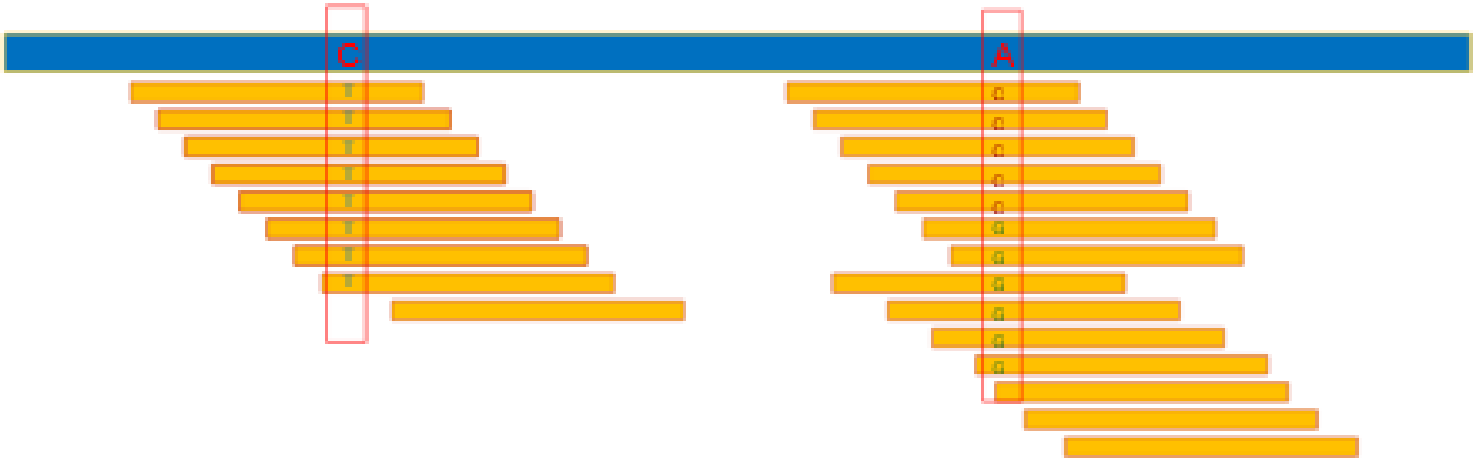
GENOMICS CORELEUVEN

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 CORELEUVEN

Reference-based Variant Calling



- SNP: Single nucleotide polymorphysm
- SNV: Single nucleotide variant
- Pointwise mutation

GENOMICS CORELEUVEN

Computer Cluster

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

- Computer Cluster
  - Computer Node
    - Computer CPU
      - Computer Core

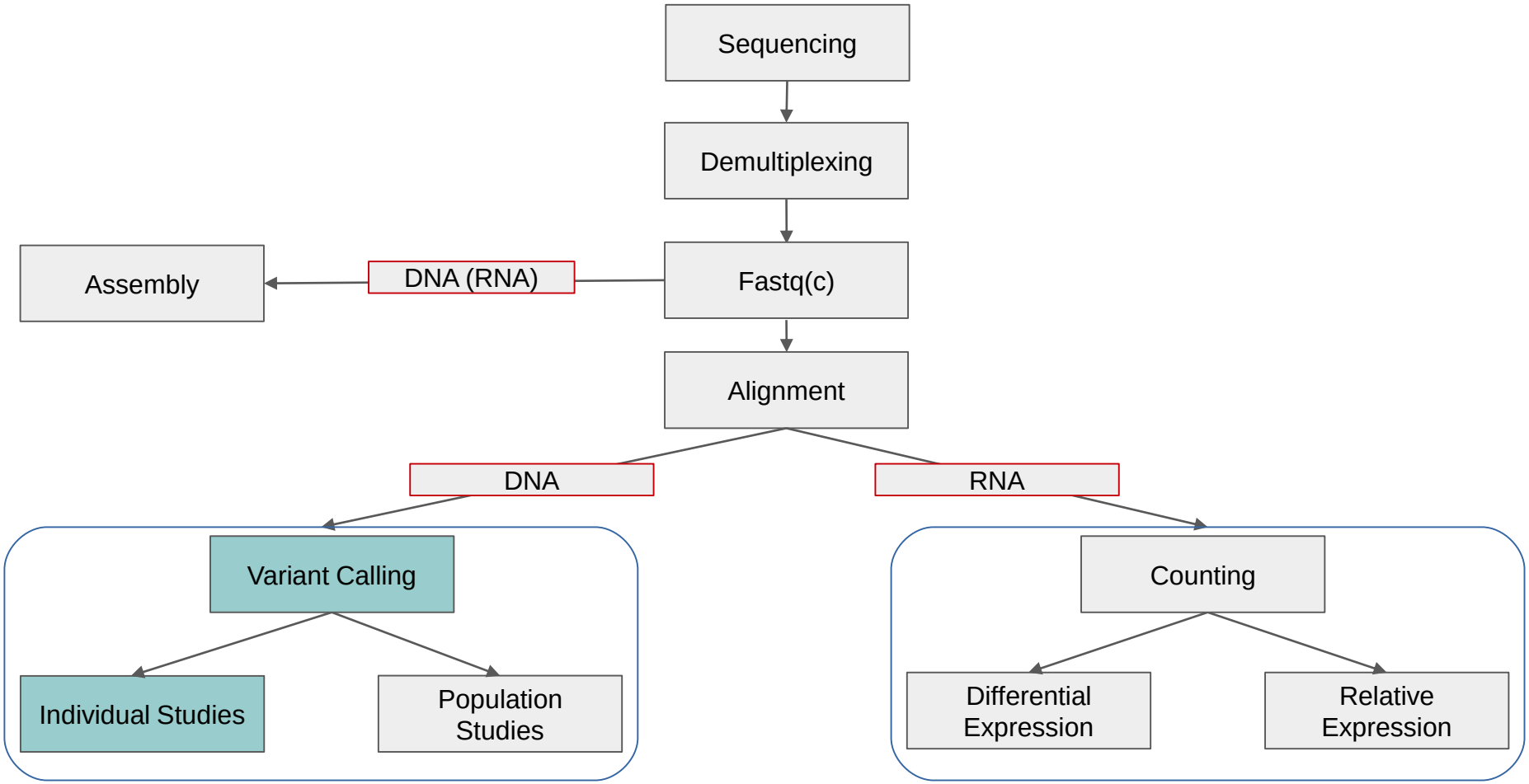
VLAAMS SUPERCOMPUTER CENTRUM

Vlaanderen is supercomputing

Google Health

GENOMICS CORELEUVEN

NGS Common Pipelines



GENOMICS CORELEUVEN

During the following sessions we will use these concepts to solve common bioinformatics problems:

- Mapping and alignment inspection with IGV
- Variant calling
- Calling structural variations
- RNA Seq: differential expression

Thanks!