Introduction to Next Generation Sequencing

Bioinformatics Workshop
31 March 2017

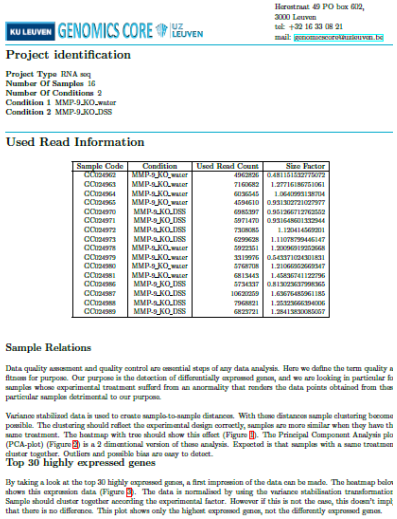
GENOMICS CORELEUVEN

Overview

- This workshop: why?
- Bioinformatics NGS terminology
- Fasta and fastq formats
- Mapping and alignment

Sequencing at the Genomics Core Leuven

A typical RNA-Seq analysis at the Genomics Core delivers:



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
CO24962	MMP-9_KO_water	4962826	0.481151332775072
CO24963	MMP-9_KO_water	7160682	1.37716186751061
CO24964	MMP-9_KO_water	6036545	1.06499913830704
CO24965	MMP-9_KO_water	4594610	0.931302721027977
CO24970	MMP-9_KO_DSS	6981997	0.881266717762552
CO24971	MMP-9_KO_DSS	5971470	0.931648631322844
CO24972	MMP-9_KO_DSS	7309385	1.12041456930
CO24973	MMP-9_KO_DSS	6299628	1.11078799446147
CO24978	MMP-9_KO_water	5922351	1.30096919233668
CO24979	MMP-9_KO_water	3319976	0.543370154301831
CO24980	MMP-9_KO_water	5768708	1.21066052669547
CO24981	MMP-9_KO_water	6813443	1.45836741122796
CO24986	MMP-9_KO_DSS	5734337	0.813023623798365
CO24987	MMP-9_KO_DSS	10626559	1.42657685961185
CO24988	MMP-9_KO_DSS	7968821	1.35329666394606
CO24989	MMP-9_KO_DSS	6825721	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 fi samples whose experimental treatment suffered 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 belo shows this expression data (Figure 3). The data is normalised by using the variance stabilisation transformatio 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,…

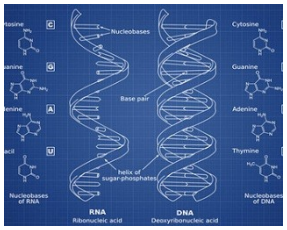
3GB of raw and analyzed data: Fastq, BAM, counts, normalized counts, gene expression files, heatmaps…

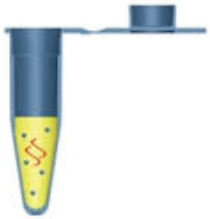
GENOMICS CORELEUVEN

Data interpretation and re-analysis

From Experimental Design to Analyzed Data

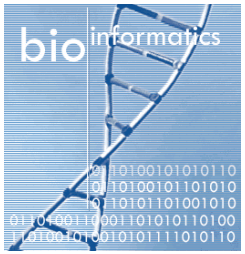
Experimental Design





Library Preparation

Sequencing



Bioinformatics

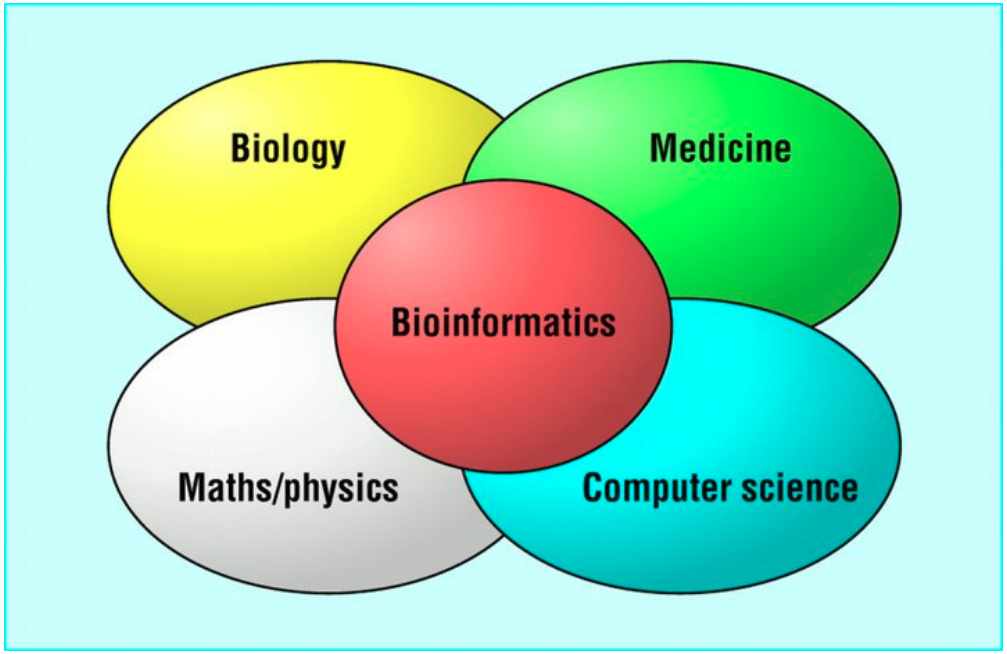
Follow up and support

GENOMICS CORELEUVEN

Bioinformatics

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data.

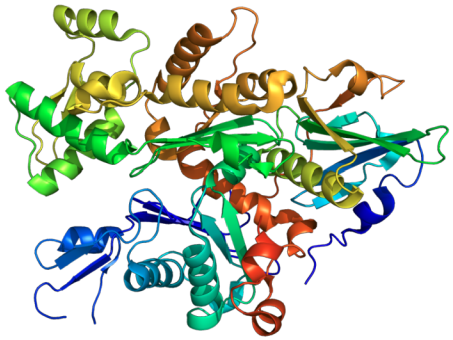
--Wikipedia

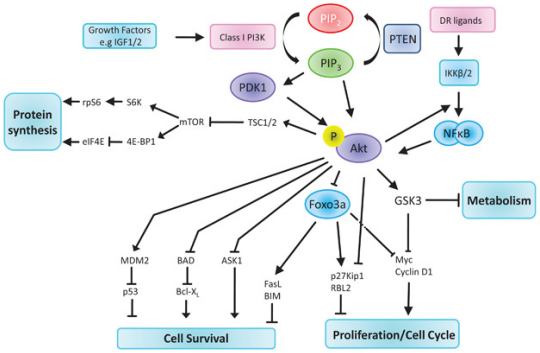


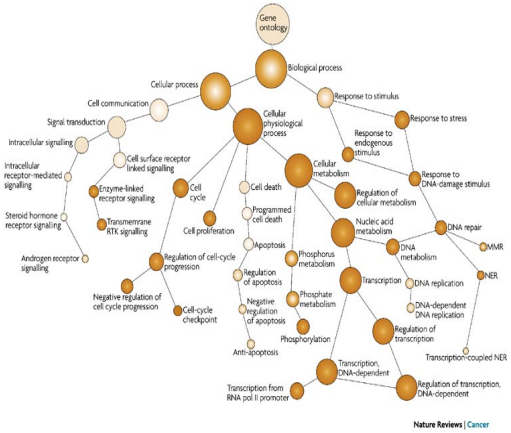
GENOMICS CORELEUVEN

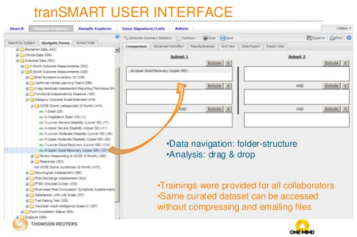
Bioinformatics...

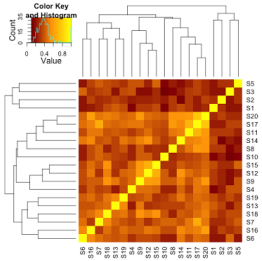
...is a broad field











GENOMICS CORELEUVEN

A day in bioinformatics at GC

Poly-A tail

Ontology

Fragments cluster

Hadoop

Xenograft Murine infiltration

Flow Cell

Demultiplexing

VCF File

Ortholog gene

Negative binomial distribution

HPC Cluster

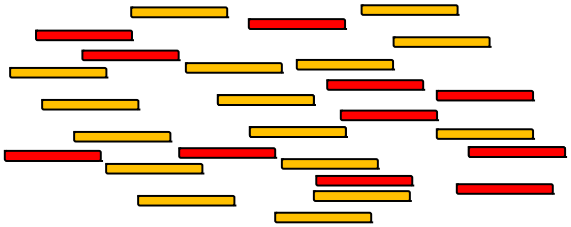
snRNA

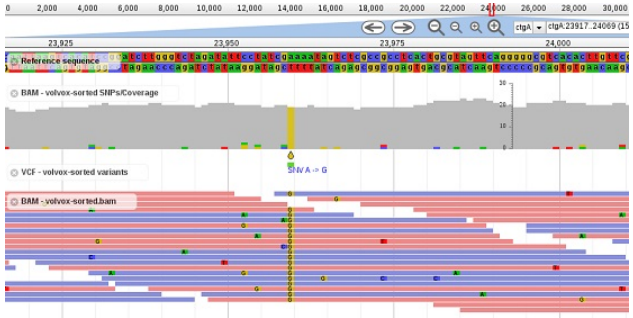
Bonferroni Correction

Barcode

GENOMICS CORELEUVEN

NGS Bioinformatics





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

GENOMICS CORELEUVEN

NGS Bioinformatics

NGS data means big data.

Raw data per sample:

- Whole Human Genome:
  300Gb
- Exome data:
  6GB
- RNA-Seq
  1GB

GENOMICS CORELEUVEN

NGS Bioinformatics

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

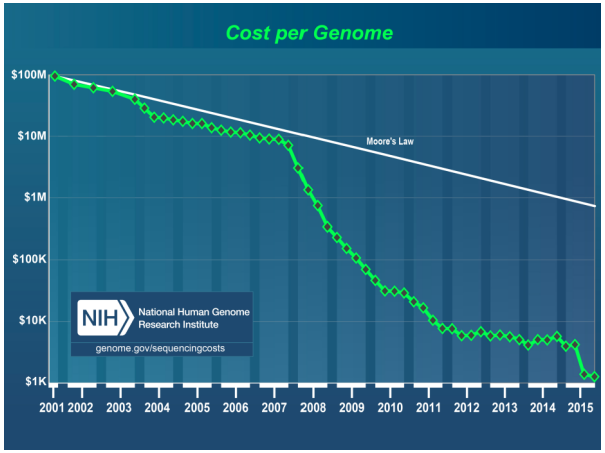
NGS data is usually analyzed on a supercomputer or cluster.

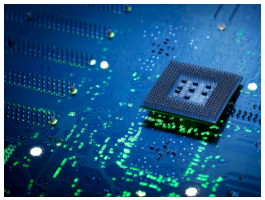
- UZ Leuven: Avalok/Hydra
- KU Leuven: VSC Flemish Super Computer
- Google genomics, etc.

GENOMICS CORELEUVEN

Cost of NGS Analysis

Sequencing technology advances, analysis remains challenging





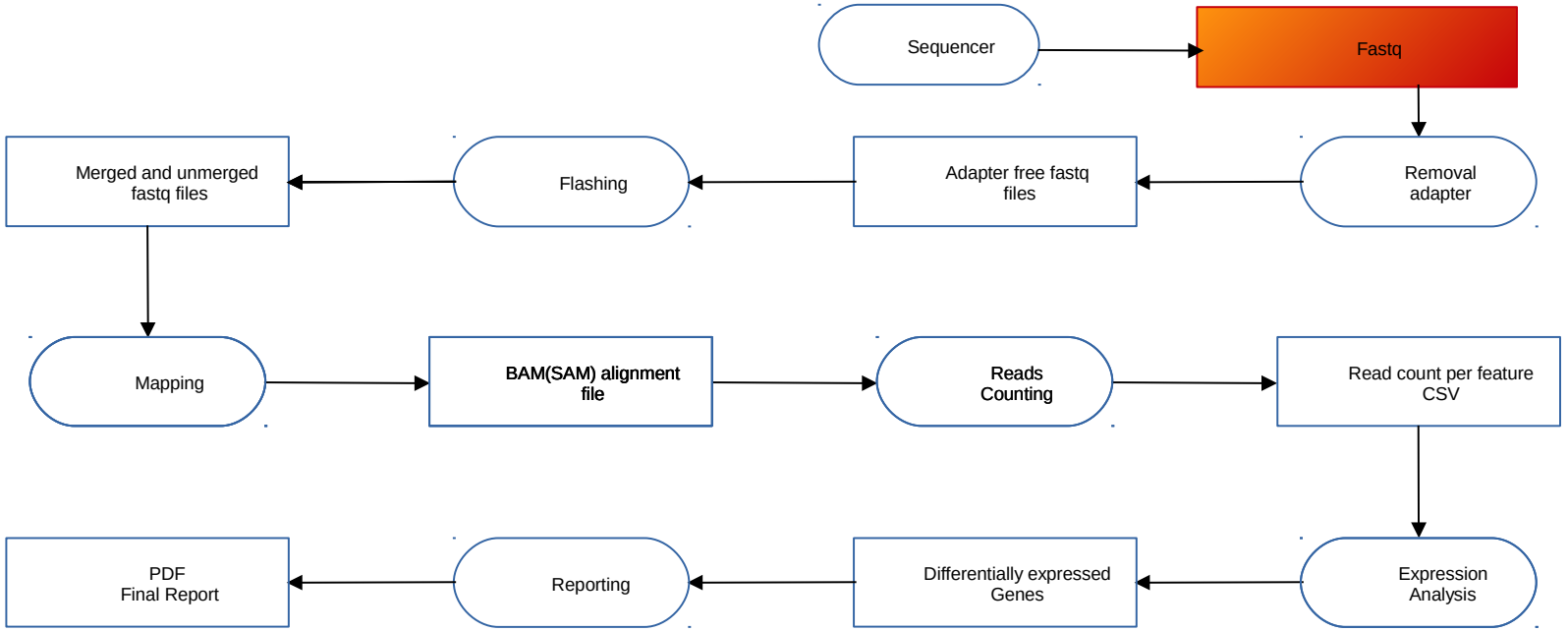
$1.000 sequencing and $10.000 bioinformatics costs?

GENOMICS CORELEUVEN

NGS Terminology

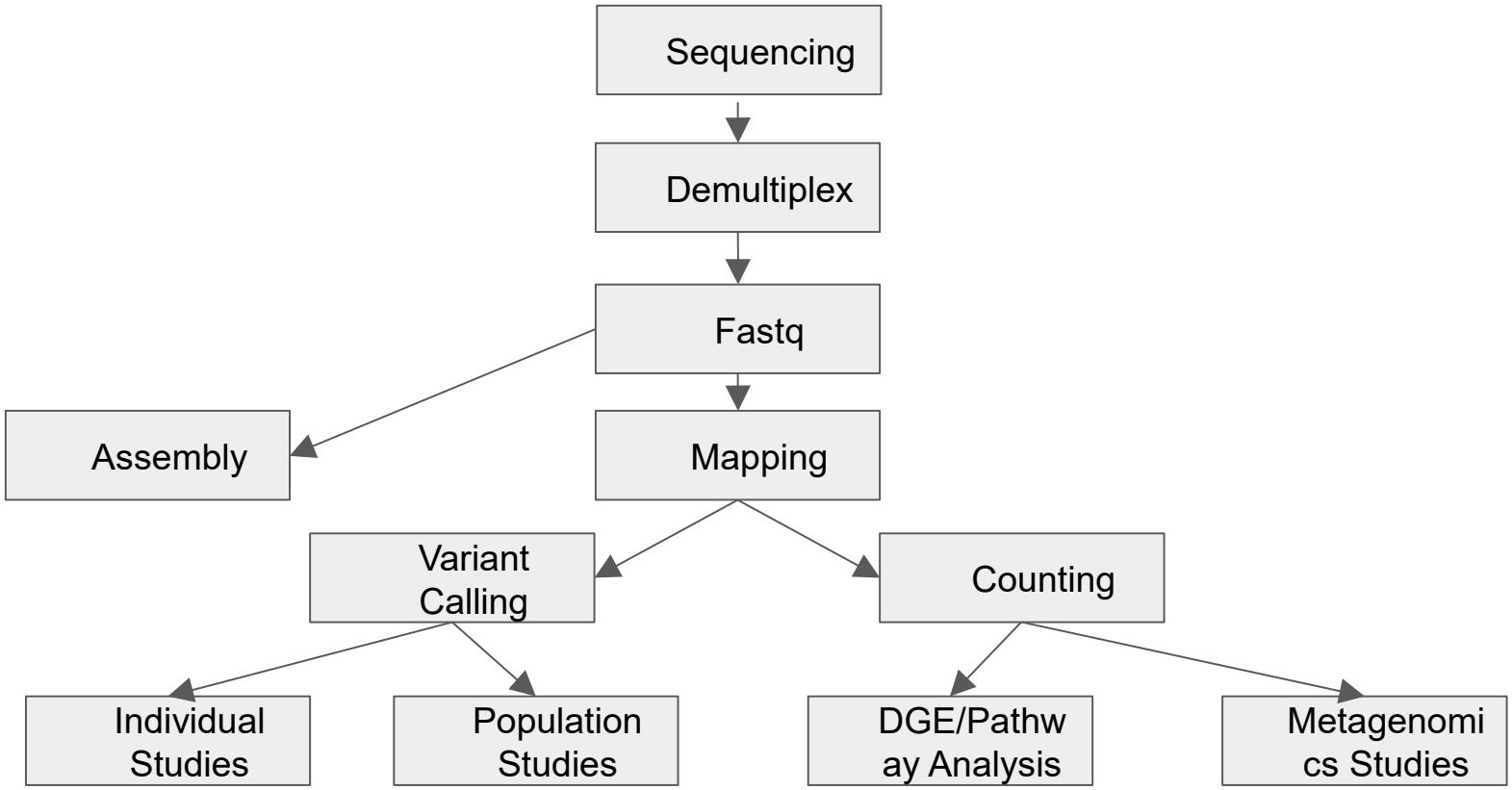
- Bioinformatics Pipeline:

Computational steps to transform input data into processed output information (in the context of biological data)



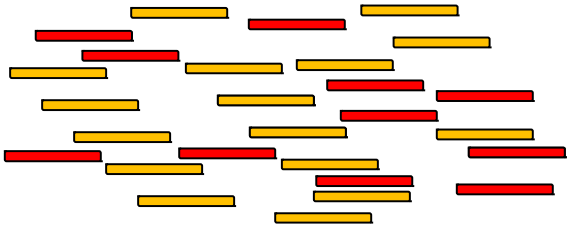
GENOMICS CORELEUVEN

NGS Terminology



GENOMICS CORELEUVEN

NGS Sequencing



DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule. It includes any method or technology that is used to determine the order of the four bases—adenine, guanine, cytosine, and thymine—in a strand of DNA.

- NGS sequencing
- Shotgun sequencing
- Massively parallel sequencing
- Deep Sequencing

NGS platforms perform sequencing of millions of small fragments of DNA in parallel.

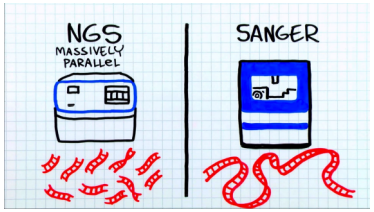
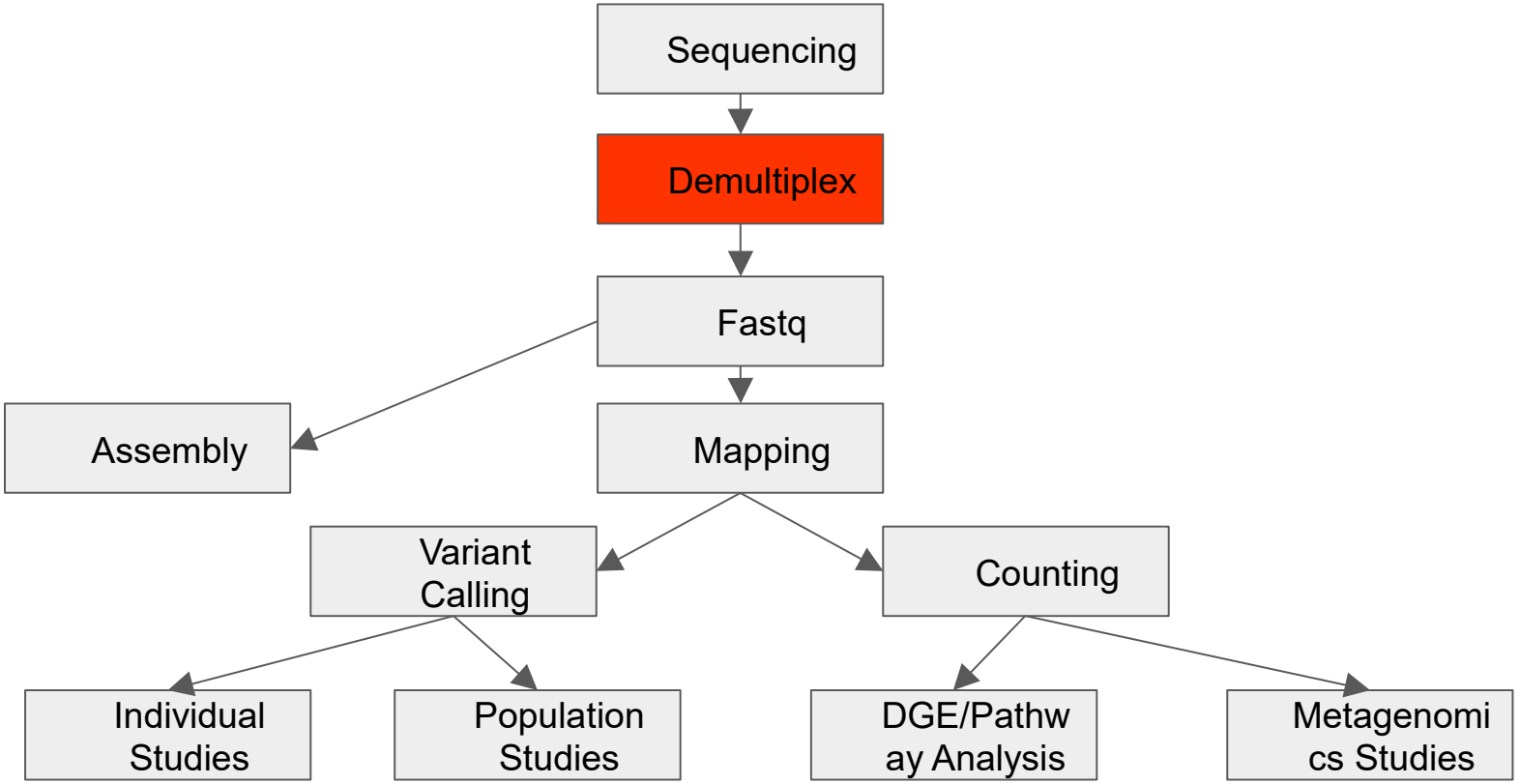


Image:Thermo Fisher Scientific

GENOMICS CORELEUVEN

Demultiplexing



GENOMICS CORELEUVEN

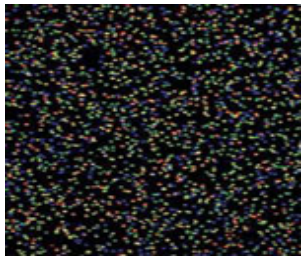
From DNA molecules to digital reads

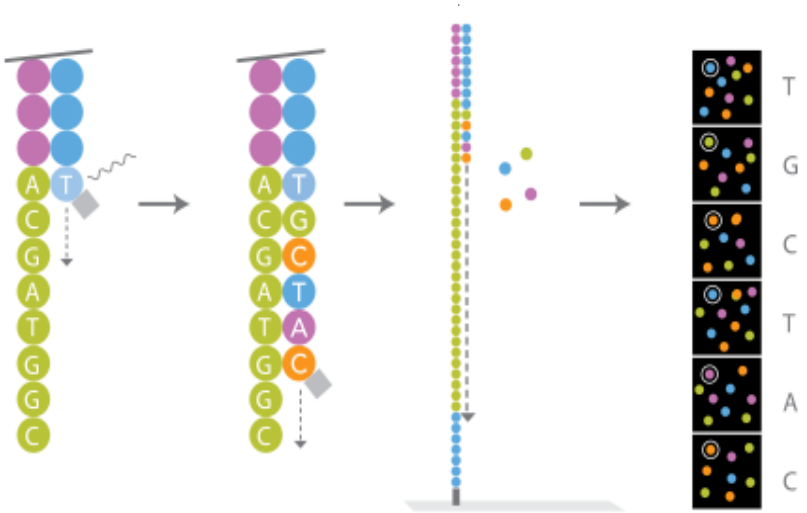
DNA Library

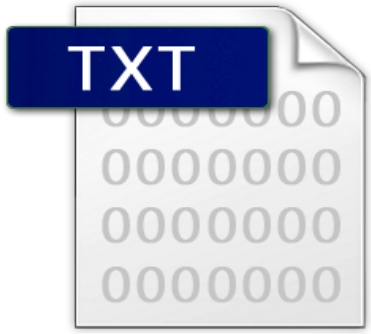
Reads

GENOMICS CORELEUVEN

From DNA molecules to digital reads



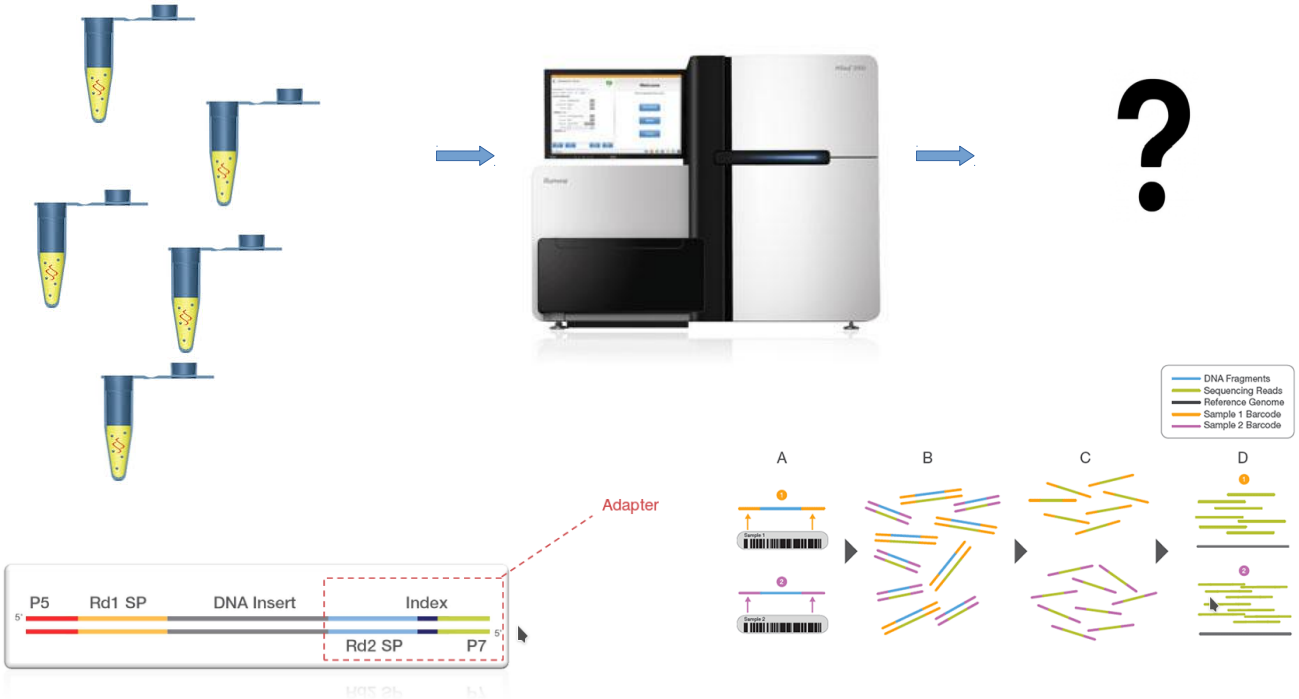




GENOMICS CORELEUVEN

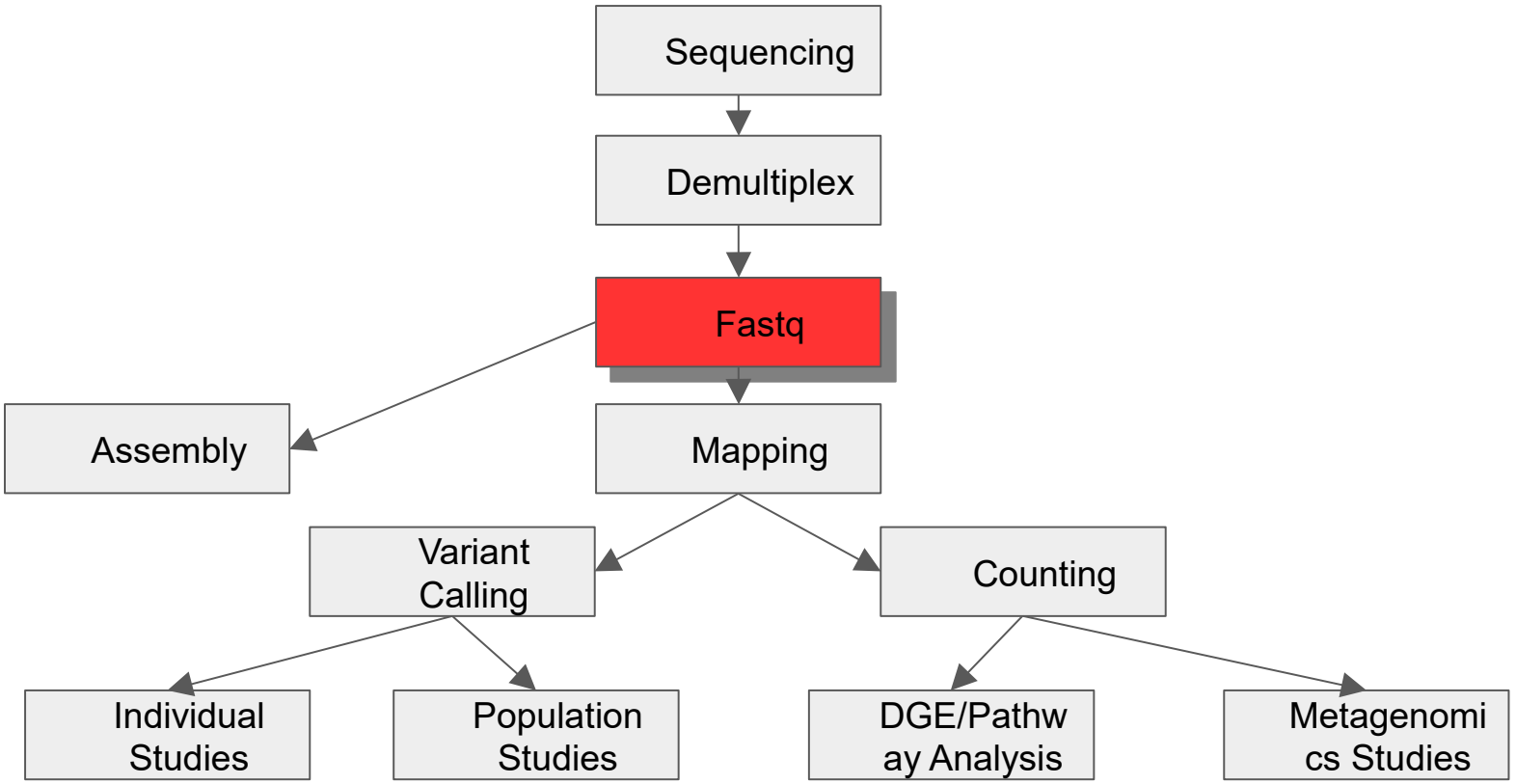
(De)Multiplexing

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



GENOMICS CORELEUVEN

Raw reads: Fastq format



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 every NGS bioinformatics project
- Fastq files are often compressed using zip or gunzip

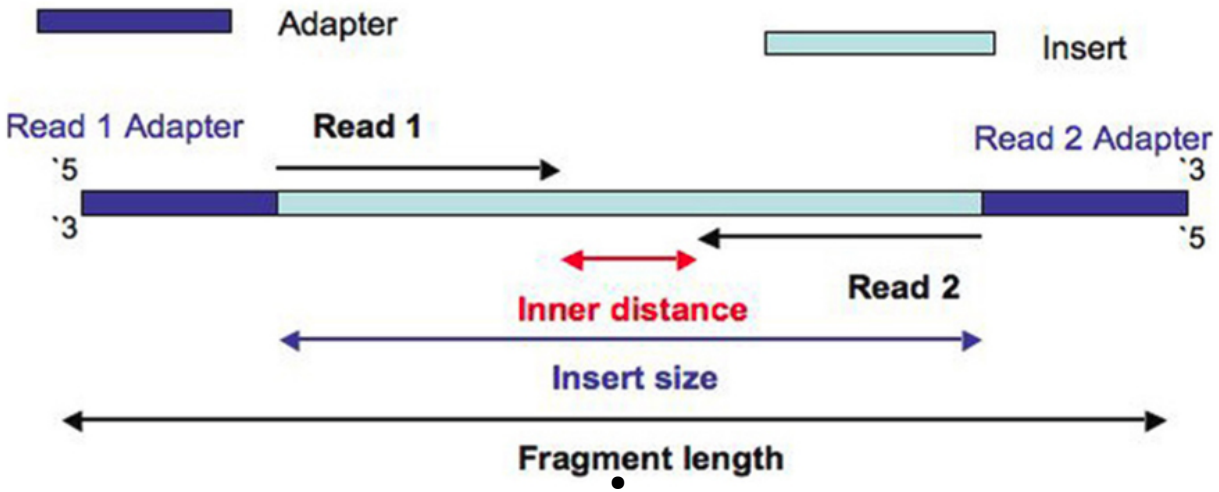
```
-rwxr-xr-x  1 vsc31439 lp_biogenomics  32G 17 janv. 15:05 GC036462.R1.fastq.gz
-rwxr-xr-x  1 vsc31439 lp_biogenomics  39G 17 janv. 15:49 GC036462.R2.fastq.gz
-rwxr-xr-x  1 vsc31439 lp_biogenomics  31G 17 janv. 16:25 GC036463.R1.fastq.gz
-rwxr-xr-x  1 vsc31439 lp_biogenomics  37G 17 janv. 17:07 GC036463.R2.fastq.gz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

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

Reads and Fastq Format

- What is a "read"?
  - A raw sequence (ordered collection) of nucleotides names A,C,G,T, or N.
  - TruSeq Stranded mRNA: 51 characters long.

- 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
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@HISEQ:574:C6VG2ANXX:2:1110:1560:2177 1:N:0:ATCACG
ATGGTCCAGCAAGGGGTATGCTGAGAAGGGGAGCAGTTCAGAACCCATCAG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@HISEQ:574:C6VG2ANXX:2:1110:1583:2223 1:N:0:ATCACG
CTACCTTCACTATCAACATAGCAAACACACCTTTAGCTCCAGCTATTAACA
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@HISEQ:574:C6VG2ANXX:2:1110:1609:2245 1:N:0:ATCACG
AGCTTAAGAGGCACTAGACACAGCCAGCTTCTTCAGGTGATCCATGAACAC
```

- Counting reads in fastq files: zcat my.fastq.gz | echo $((`wc -l`/4)) : 12.748.143

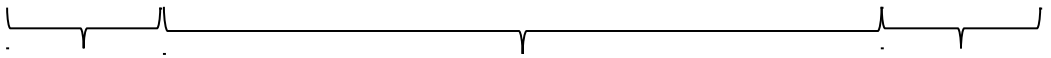
- Sequencing depth: The total number of sequences generated for a sample.
  - Usually expressed in fragments or reads

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:
- 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

Raw Reads: Base Calling Quality

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

- A quality score (Q-score) is a prediction of the probability of an error in base calling.
- It serves as a compact way to communicate very small error probabilities
  - P = 10^(–Q/10)
  - Q = –10 log10(P)

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

Raw Reads: Base Calling Quality

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

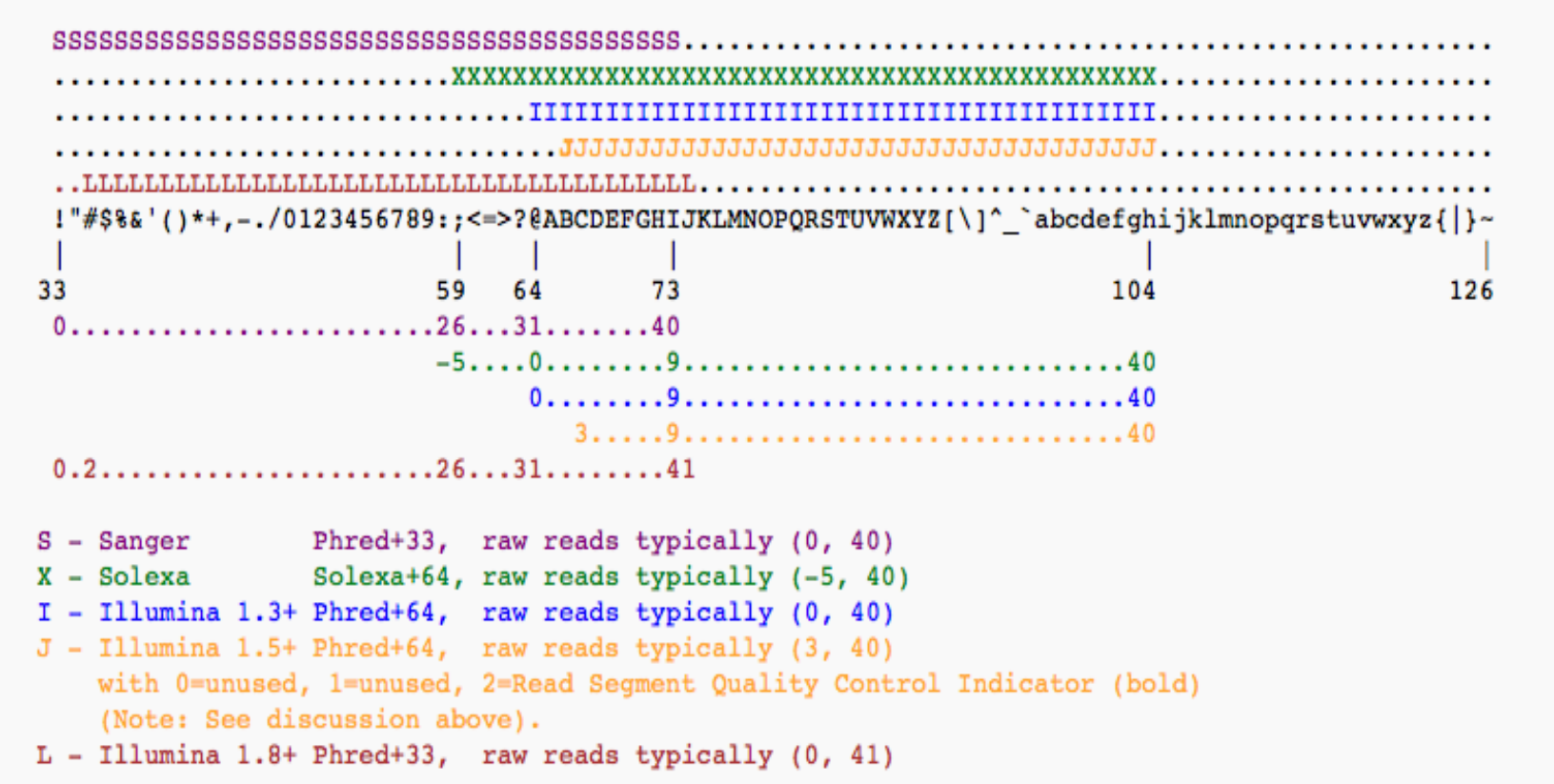
- Base: G
- Quality: C
- ASCII: 67
- Q: 34
- P: 0.00040

- Base: G
- Quality: G
- ASCII: 71
- Q: 38
- P: 0.00016

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

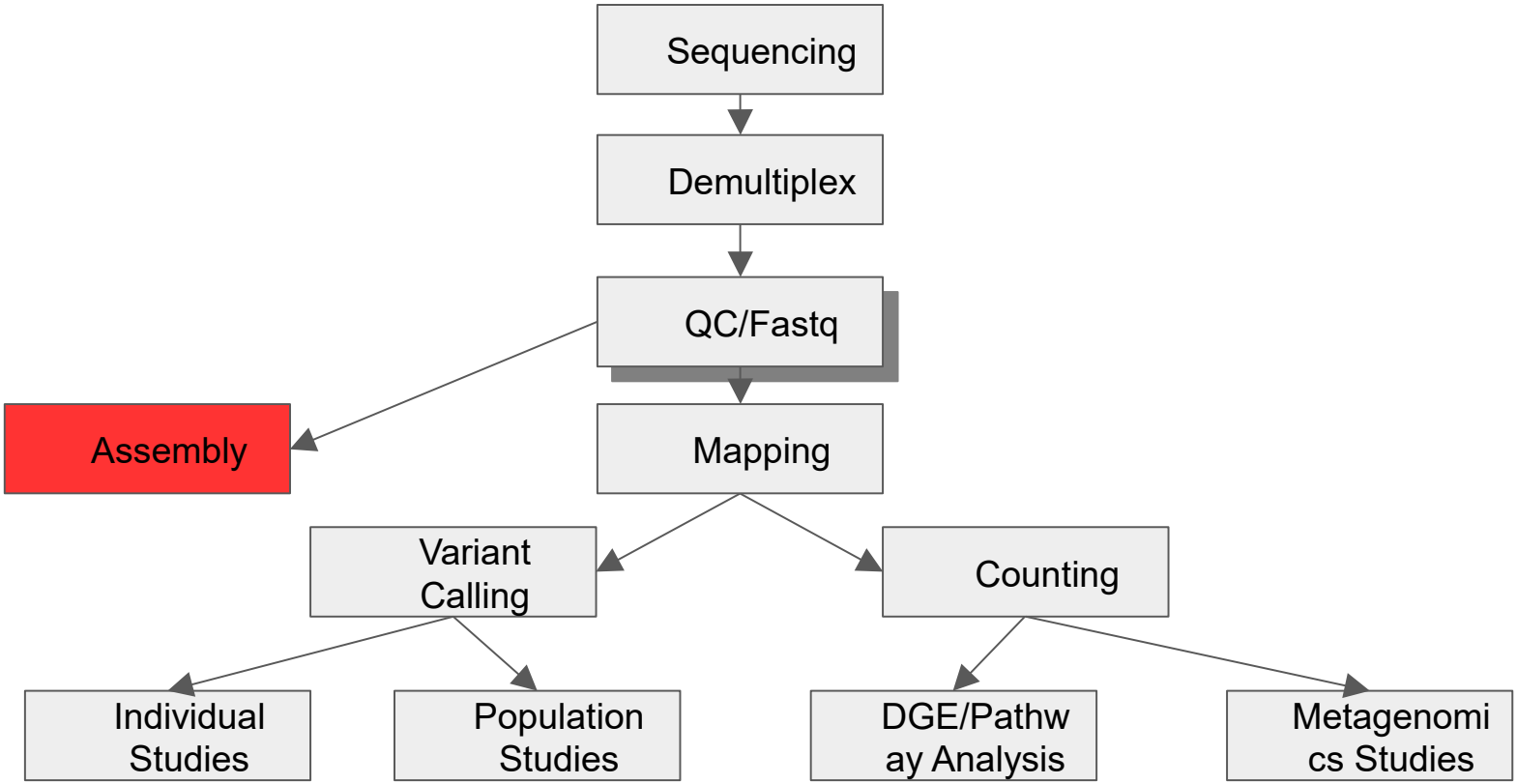
Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

Raw Reads: Base Calling Quality



GENOMICS CORELEUVEN

Assembly



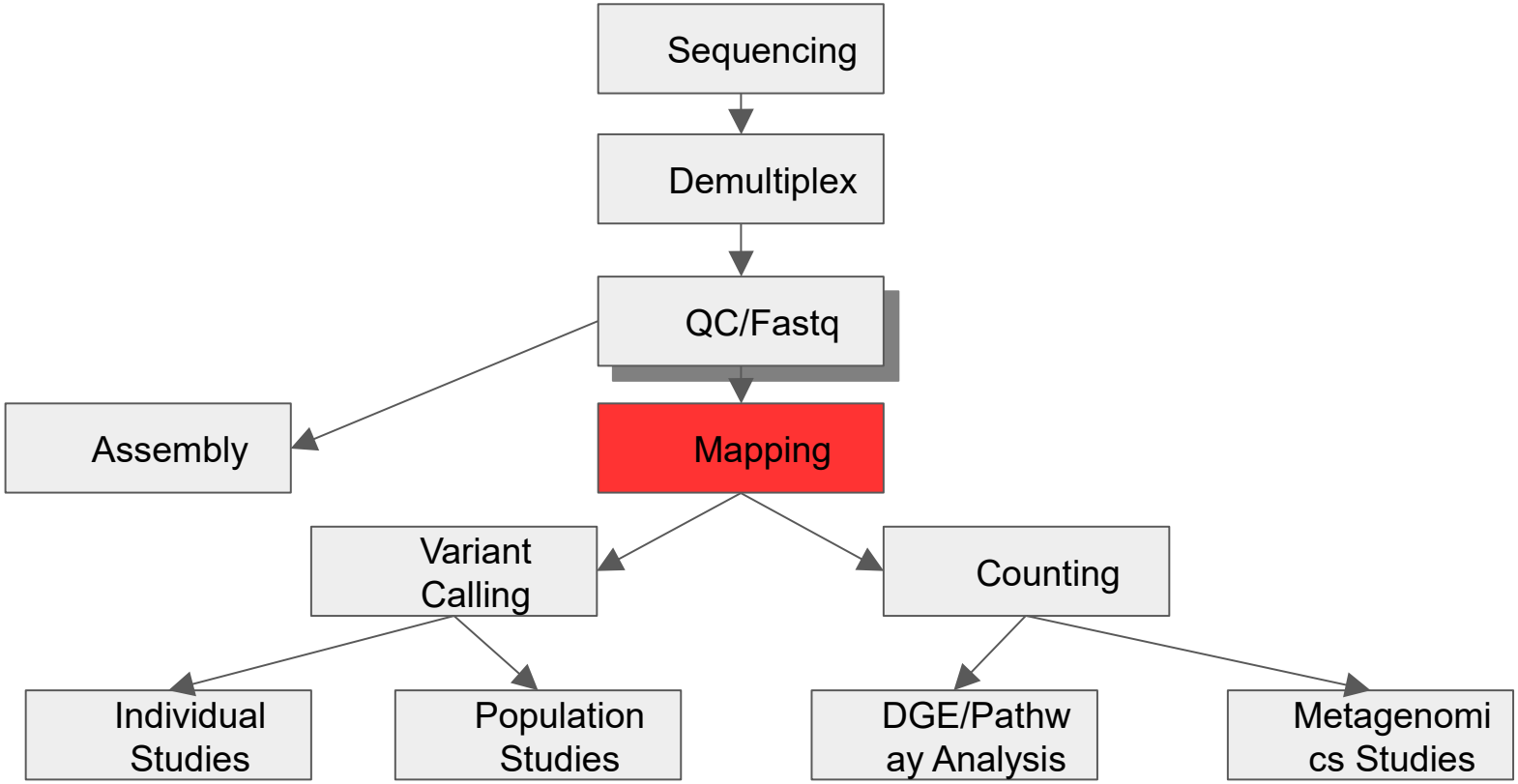
GENOMICS CORELEUVEN

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

Mapping



GENOMICS CORELEUVEN

Reference Sequence in Fasta Format

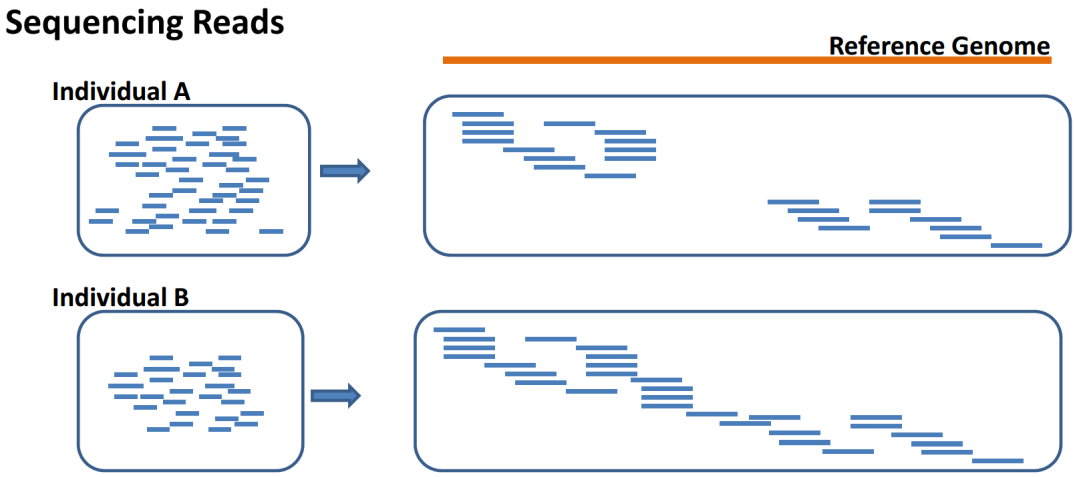
- genome.fa human-readable nucleotide sequence
- Species dependent
- Mouse genome: 2.6GB (Giga Bases!)
- Evolves

```
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, whether the reference is a complete genome, transcriptome, or *de novo* assembly.

--http://sfg.stanford.edu/mapping.html



GENOMICS CORELEUVEN

Mapping & Alignment

- Reference Ch1:

  1234567890123456

  ATGGTTACACCATT
- Read:

  GGTTCA
- Possible alignment:

  ATGGTTACACCATT

  GGTT-CA
- Mapping: Ch1-pos3

"Also of note is that by this time the terms "read alignment" and "read mapping" had become interchangeable. The BWA and Bowtie papers both used both terms, as did many other papers."

https://liorpachter.wordpress.com/2015/11/01/what-is-a-read-mapping/

Thanks!

GENOMICS CORELEUVEN