Workshop UCIBIO
High-Throughput Sequencing Data

Ana Rita Grosso

argrosso@fct.unl.pt

September 12th 2022

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

UCIBIO Applied Molecular Biosciences Unit

Learning Objectives

- Know the essential concepts about High-Throughput Sequencing (HTS) technologies and analysis:
  - HTS technologies and applications;
  - HTS data repositories and analysis

- Perform HTS data analysis *(hands-on):*
  - Get HTS data from public repositories
  - Assess quality of HTS data
  - Alignment to the transcriptome
  - Alignment to the genome

How to measure gene expression (mRNA levels)?

QUANTITATIVE PCR (REAL-TIME)

A lot of work to measure few genes. Very accurate.

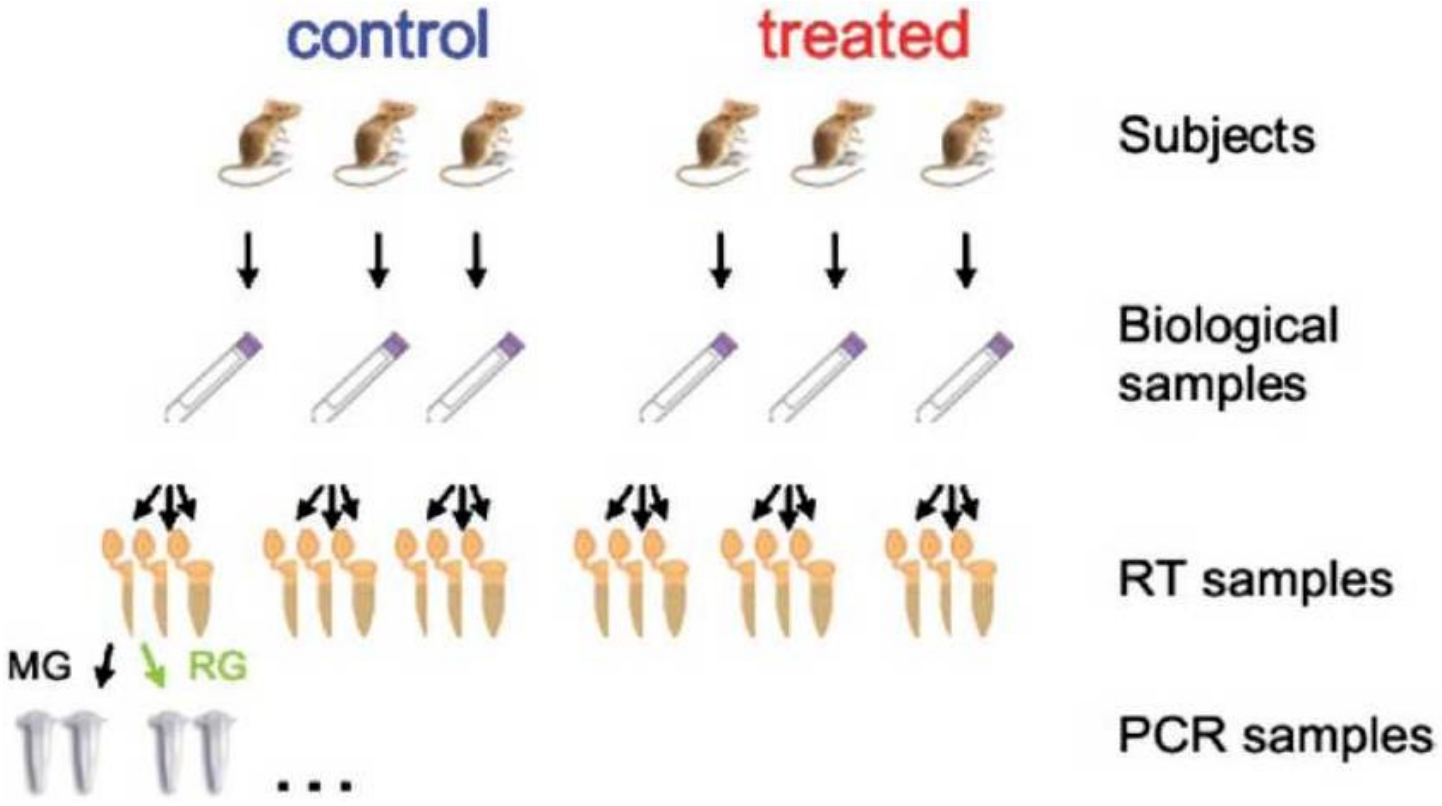


Image from https://www.gene-quantification.de

How to measure gene expression (mRNA levels)?

MICROARRAYS

Easier way to measure *pre-defined genes* in different samples. Robust.

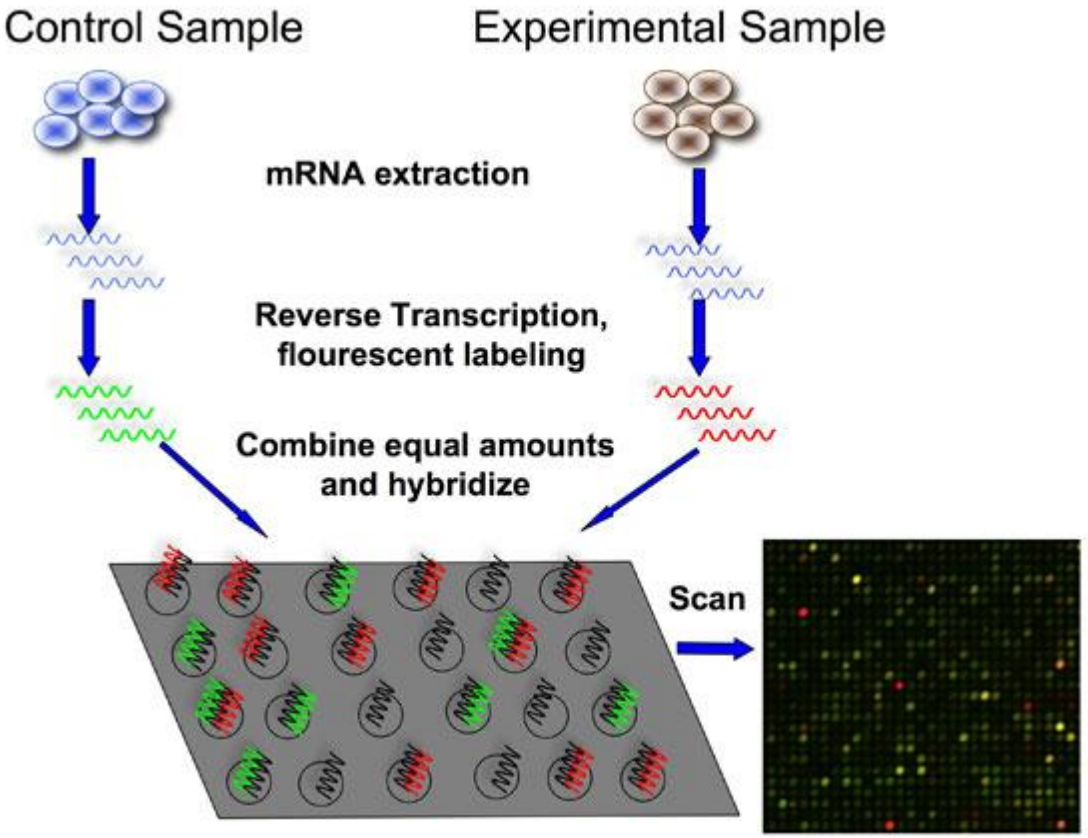
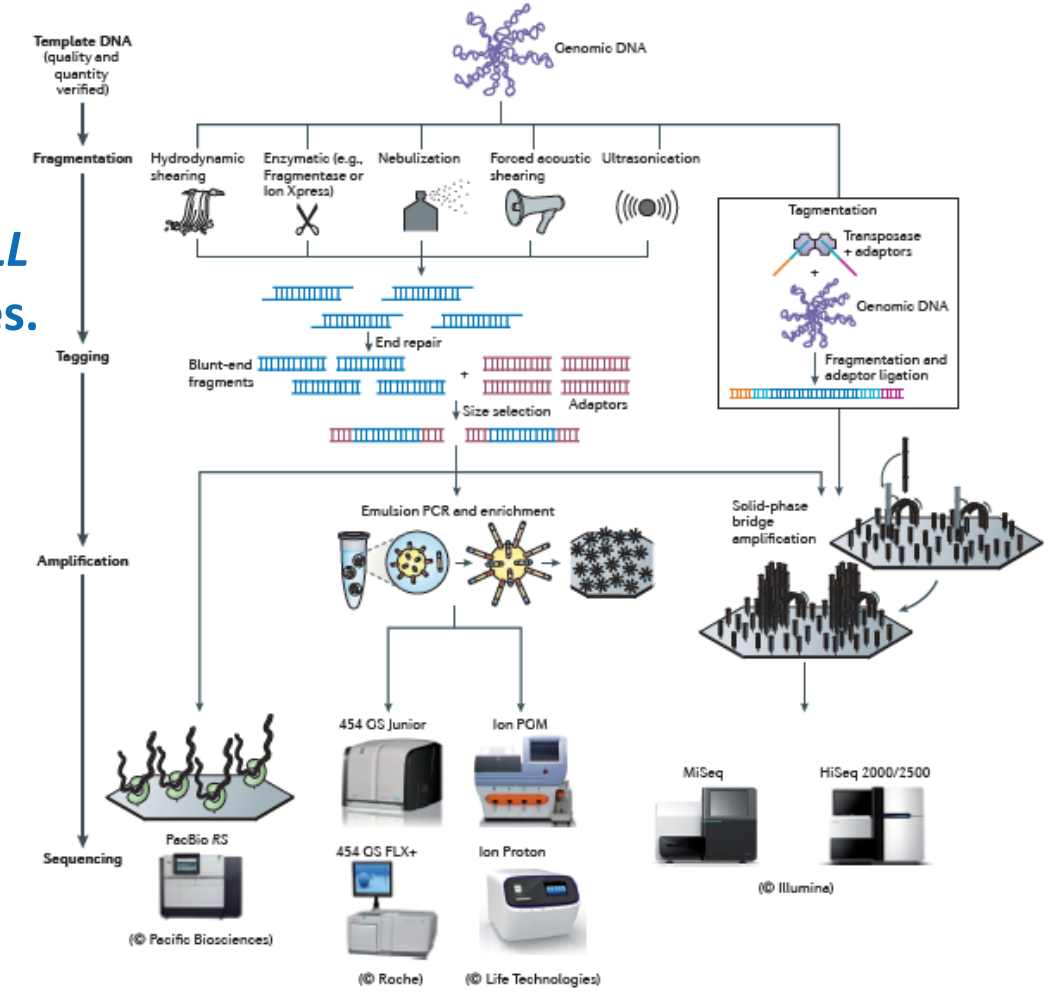


Image from https://www.gene-quantification.de

How to measure gene expression (mRNA levels)?

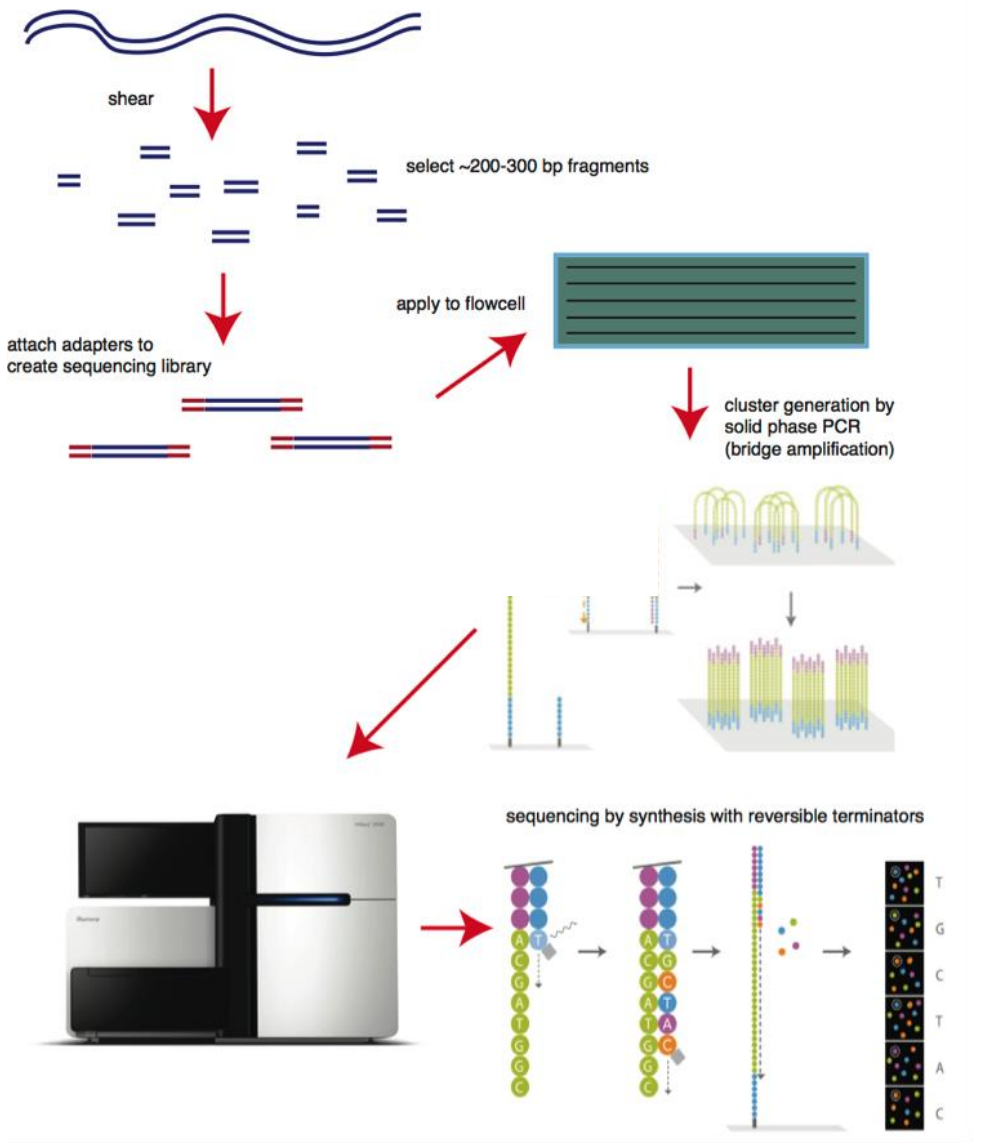
HIGH-THROUGHPUT SEQUENCING

Easier way to measure *ALL genes* in different samples. Detection of new genes.



Loman et al (2012) *Nat Rev Microb*

Illumina Sequencing



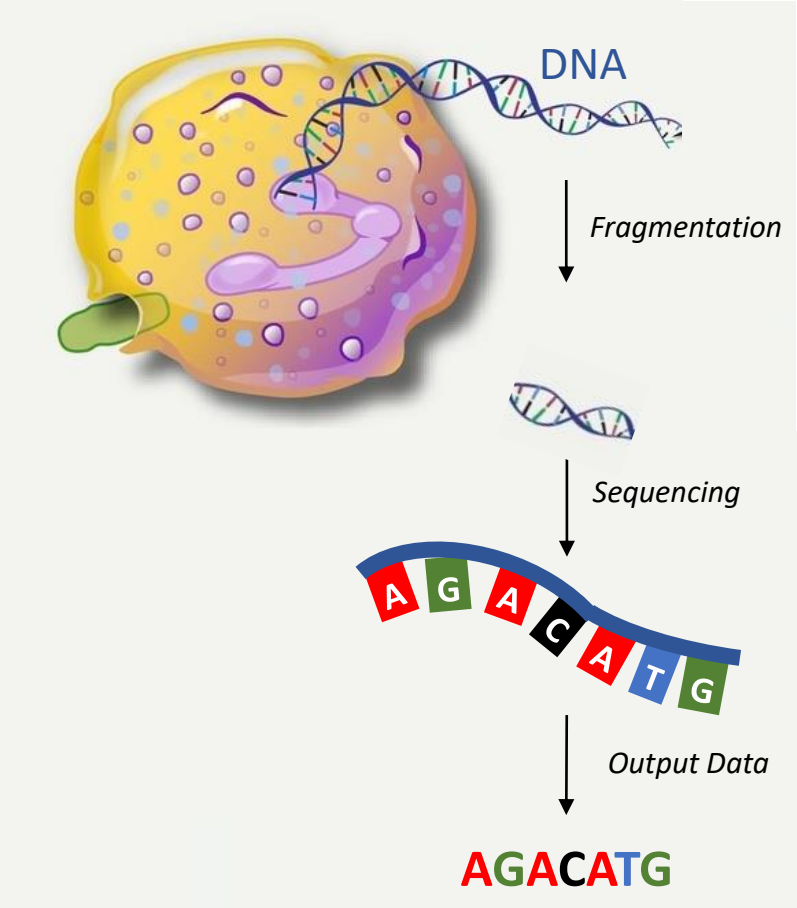
Library Preparation

Cluster Generation

Sequencing by Synthesis

Video: https://www.youtube.com/watch?v=womKfikWlxM

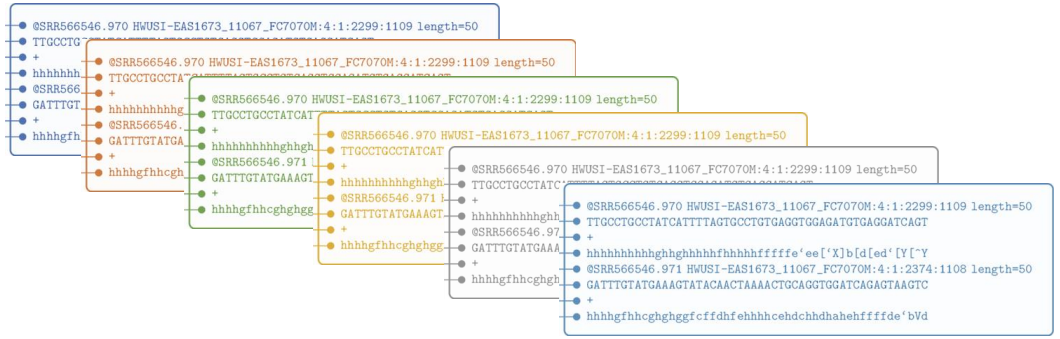
Big Data: Genomics



Output Data

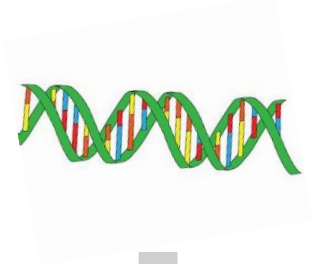
```
Identifier        @SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
Sequence          TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
'+' sign          +
Quality scores    hhhhhhhhhhghhghhhhhfhhhhhfffffe'ee['X]b[d[ed'[Y[^Y
Identifier        @SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
Sequence          GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
'+' sign          +
Quality scores    hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde'bVd
```

File with:
- *~200 Million Sequences*
- *~50 GB*



Projects with:
- *~ 200 samples*
- *~ 12TB*

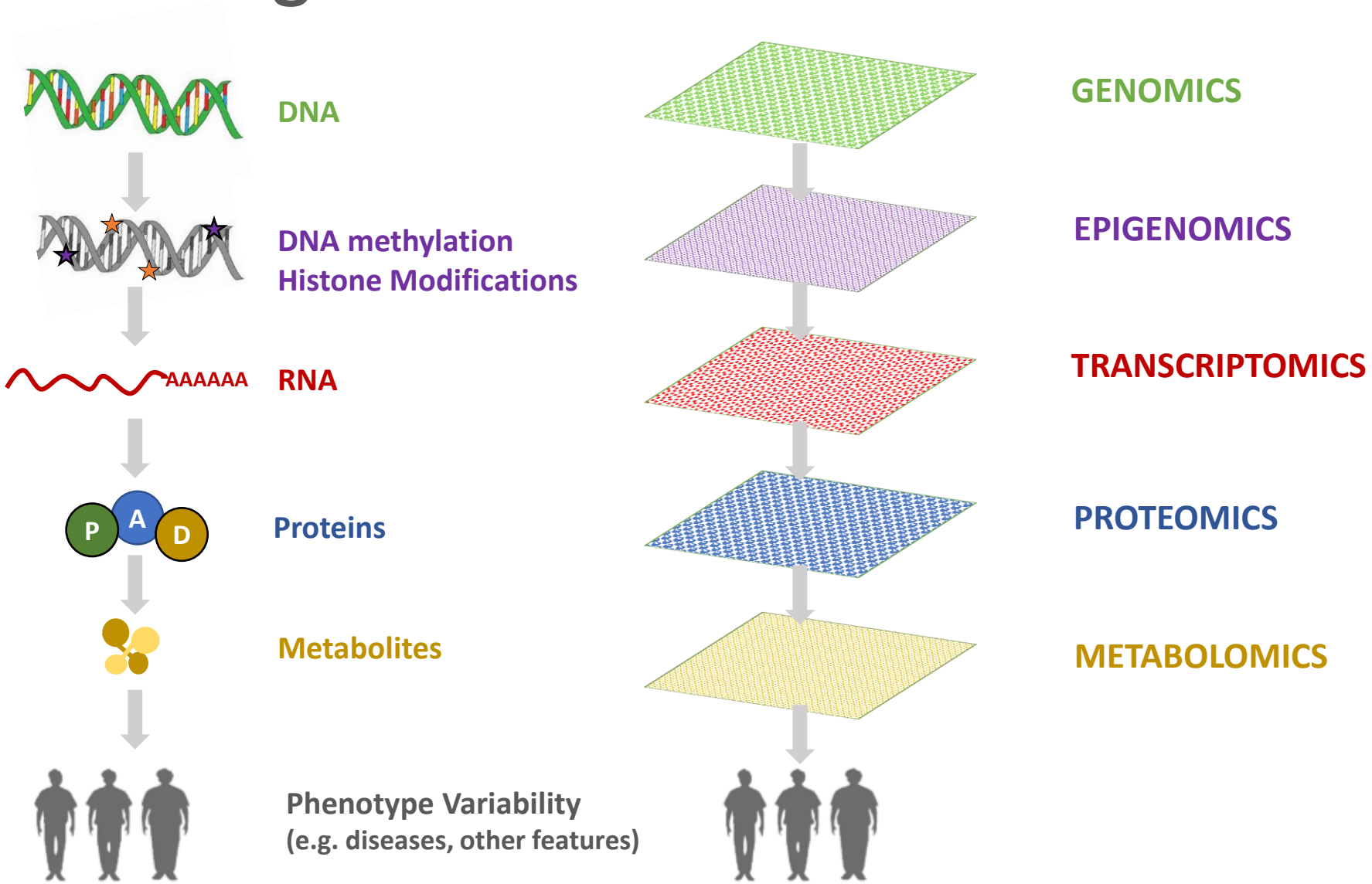
Integration of Multi-Omics Data



DNA

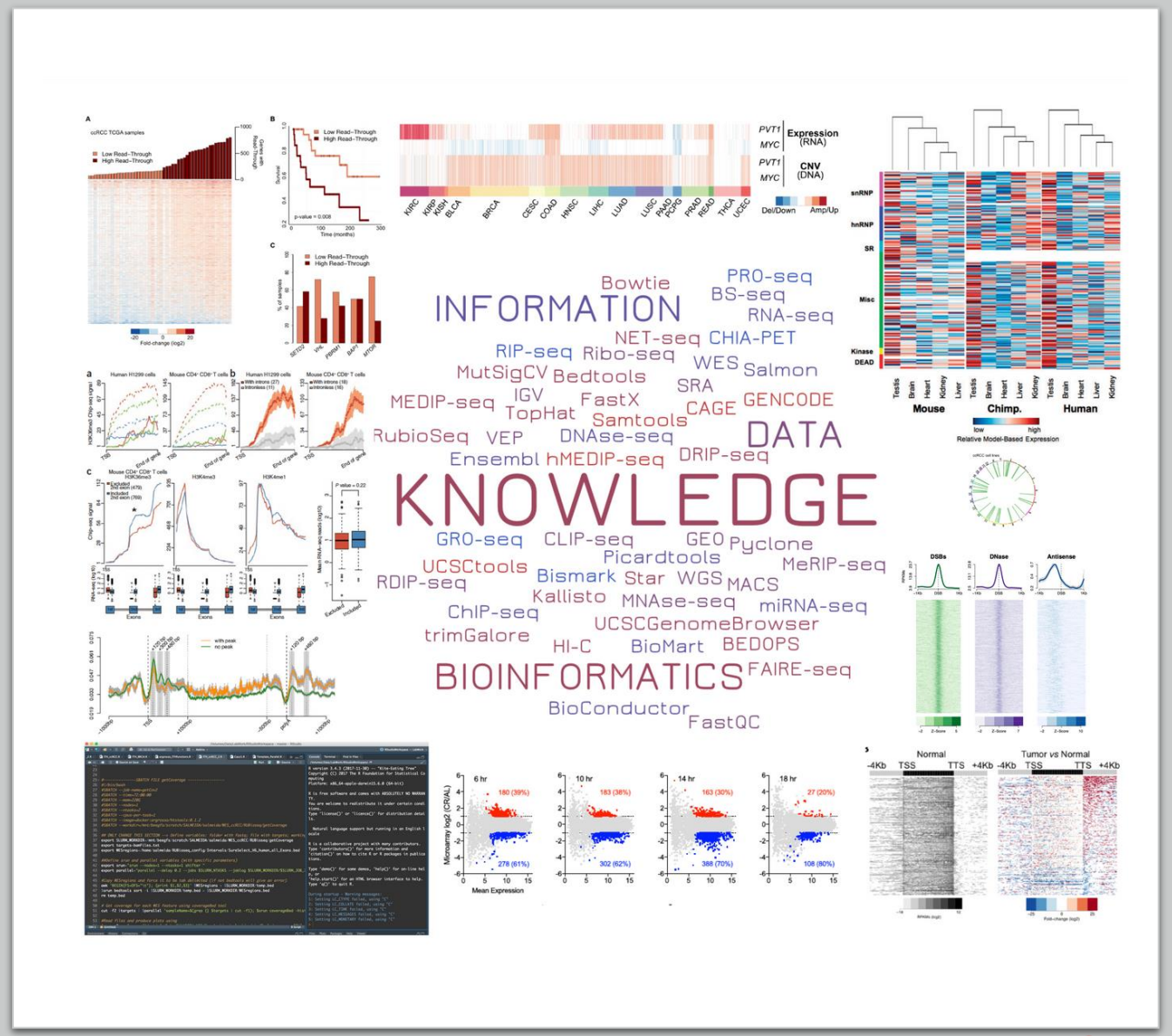
Phenotype Variability
(e.g. diseases, other features)

Integration of Multi-Omics Data



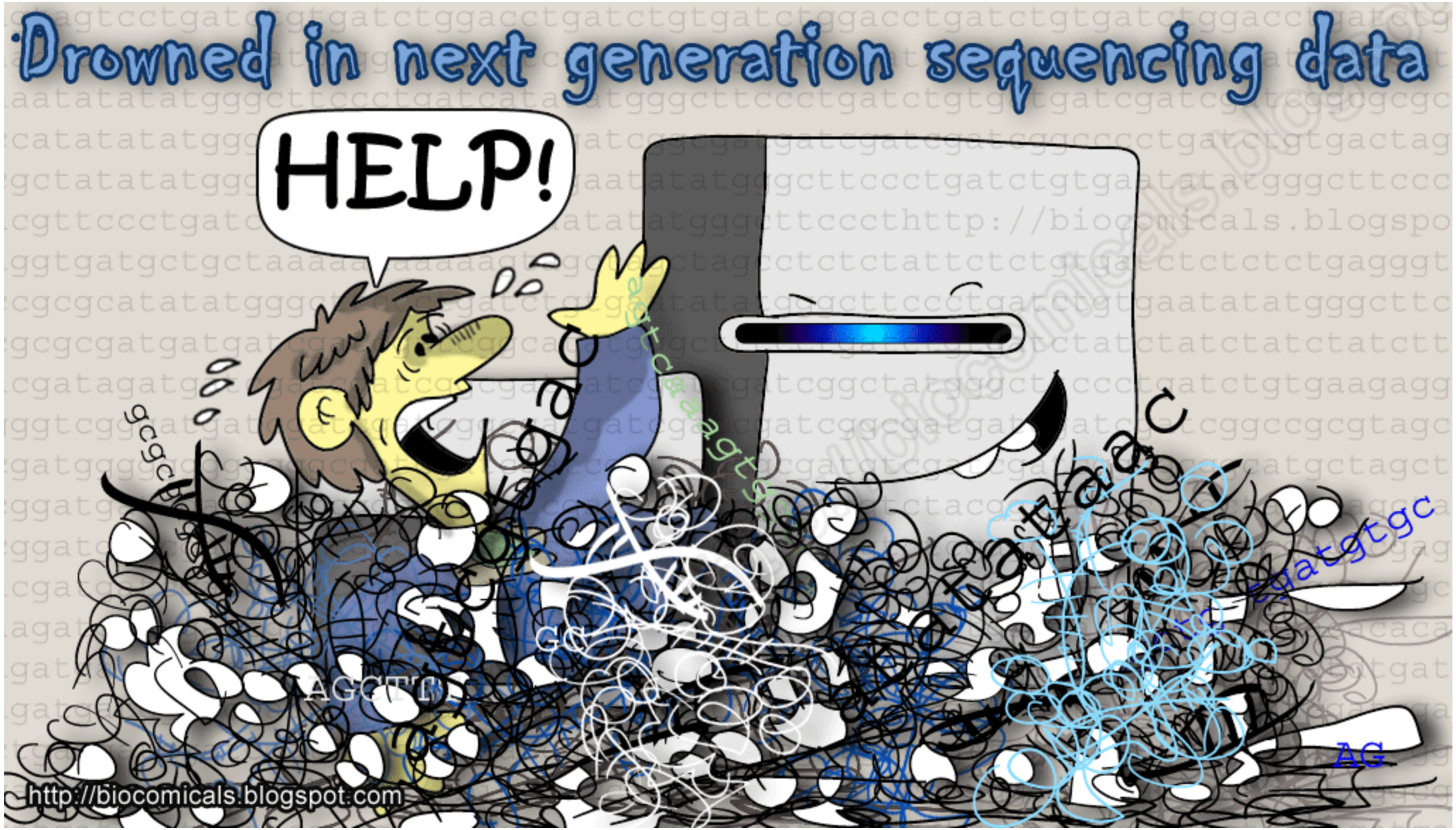
Computational Multi-Omics

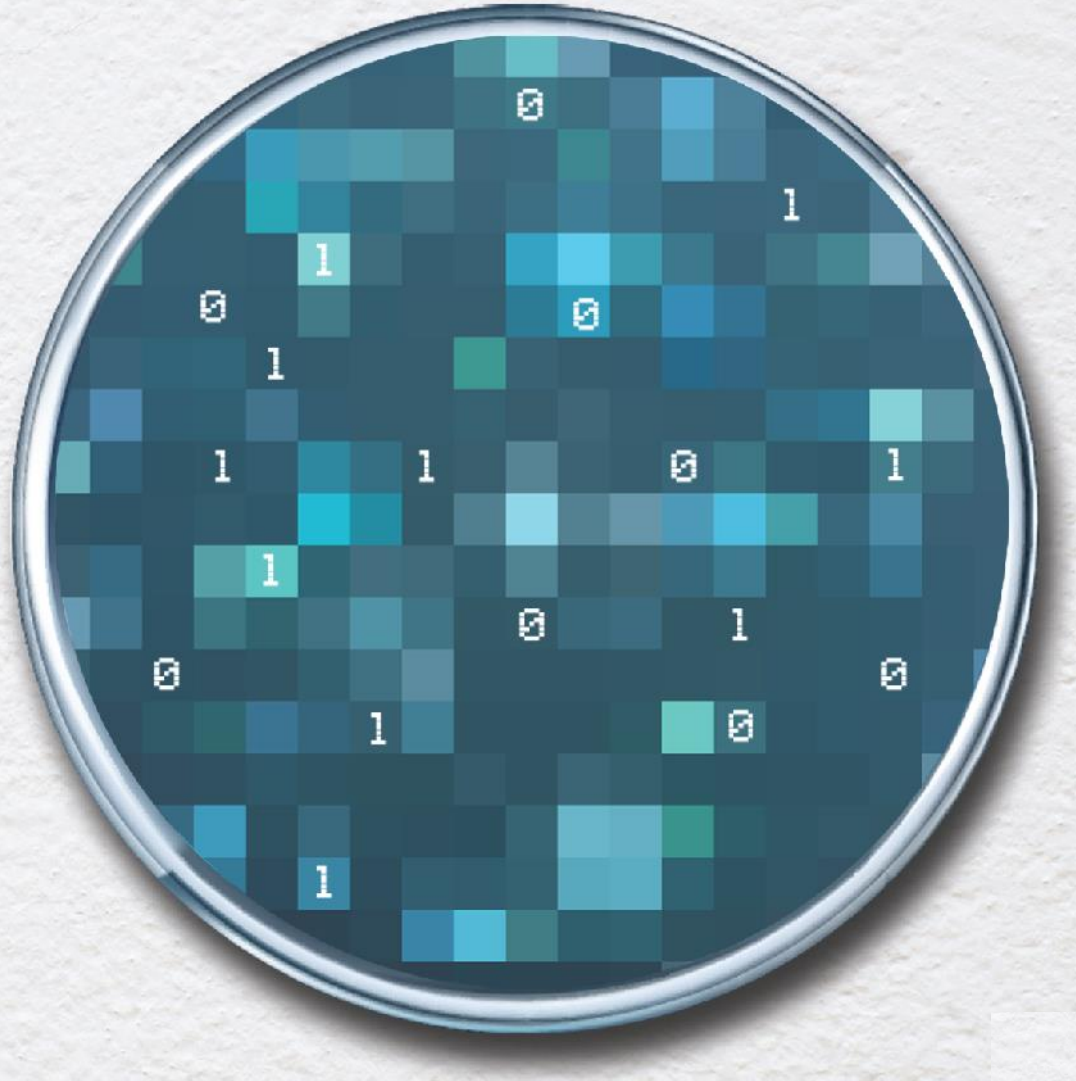
Generate scientific Knowledge by Mining Biological Data and extracting Information



NOVA SCHOOL OF SCIENCE & TECHNOLOGY
DEPARTMENT OF LIFE SCIENCES

UCIBIO
Applied Molecular Biosciences Unit


Drowned in next generation sequencing data
HELP!
http://biocomicals.blogspot.com



MCBBi | NOVA

UNIVERSIDADE NOVA DE LISBOA

MSc. in Computational Biology & Bioinformatics

masters.unl.pt/computationalbiology

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

ITQB UNIVERSIDADE NOVA DE LISBOA

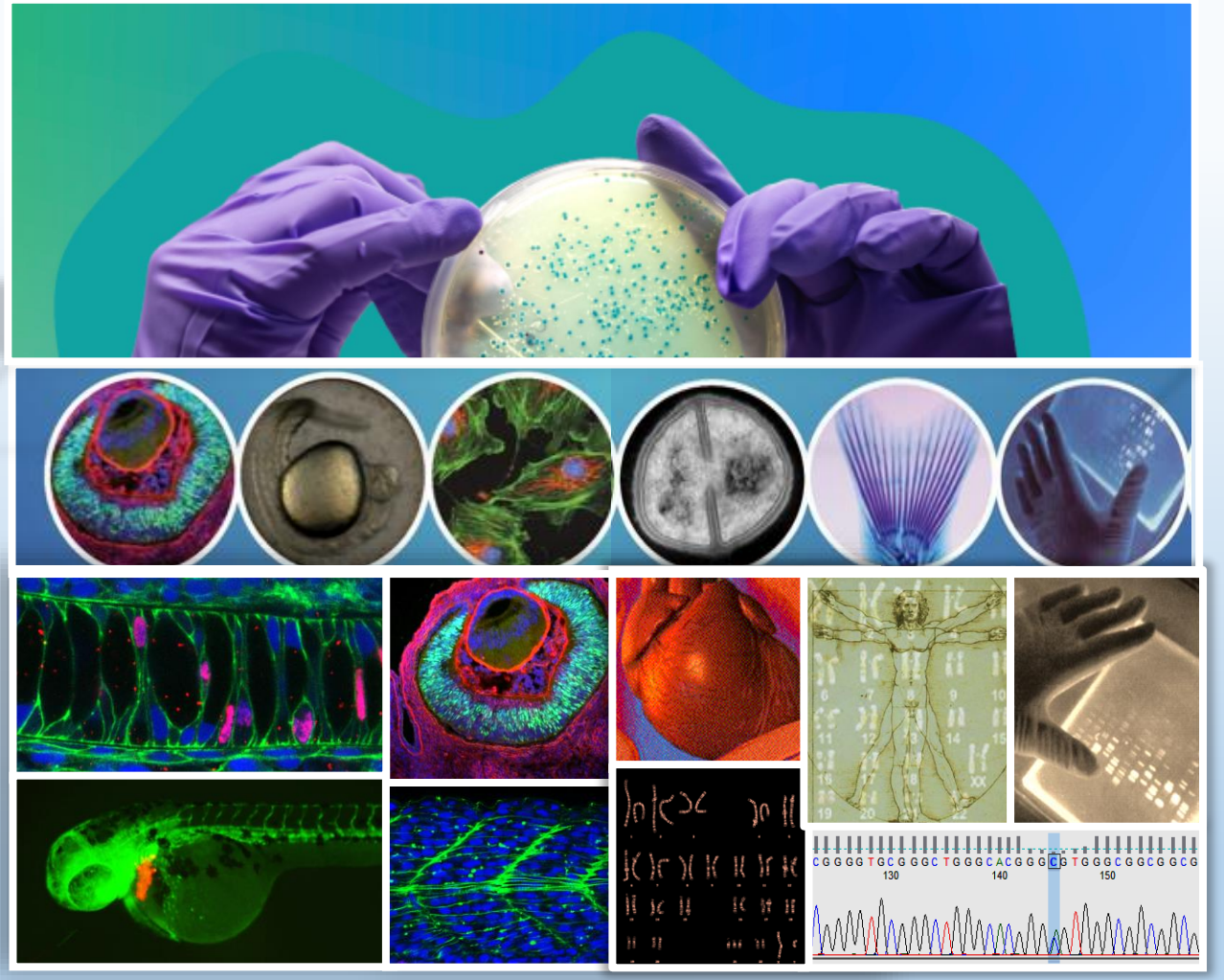
NOVA MEDICAL SCHOOL

NOVA IMS Information Management School

DESDE 1902 INSTITUTO DE HIGIENE E MEDICINA TROPICAL UNIVERSIDADE NOVA DE LISBOA

Science needs Experimental and Computational Biologists!!

Having both skills increases career opportunities!!

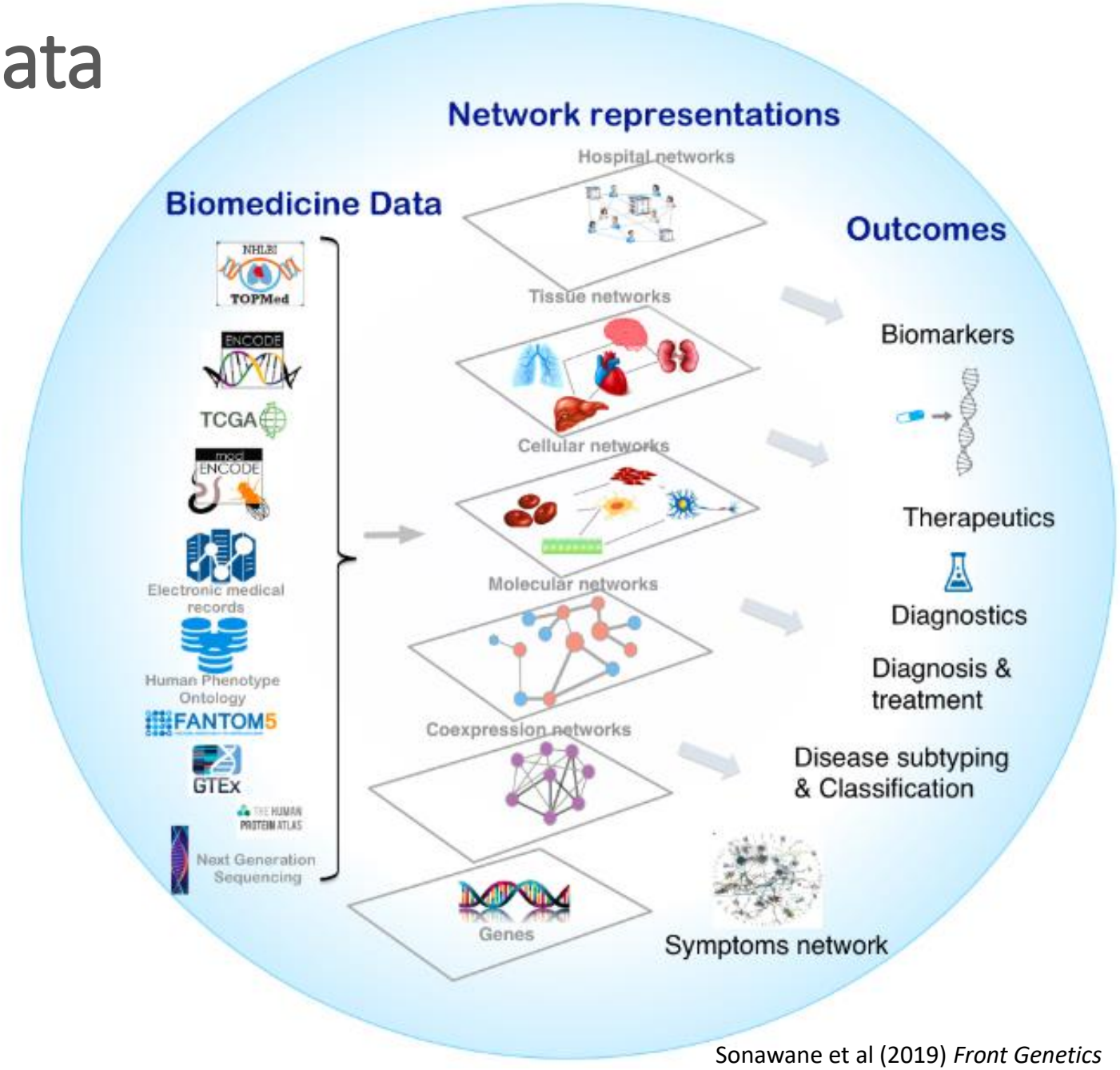


NOVA SCHOOL OF SCIENCE & TECHNOLOGY
DEPARTMENT OF LIFE SCIENCES

MSc in Molecular Genetics and Biomedicine

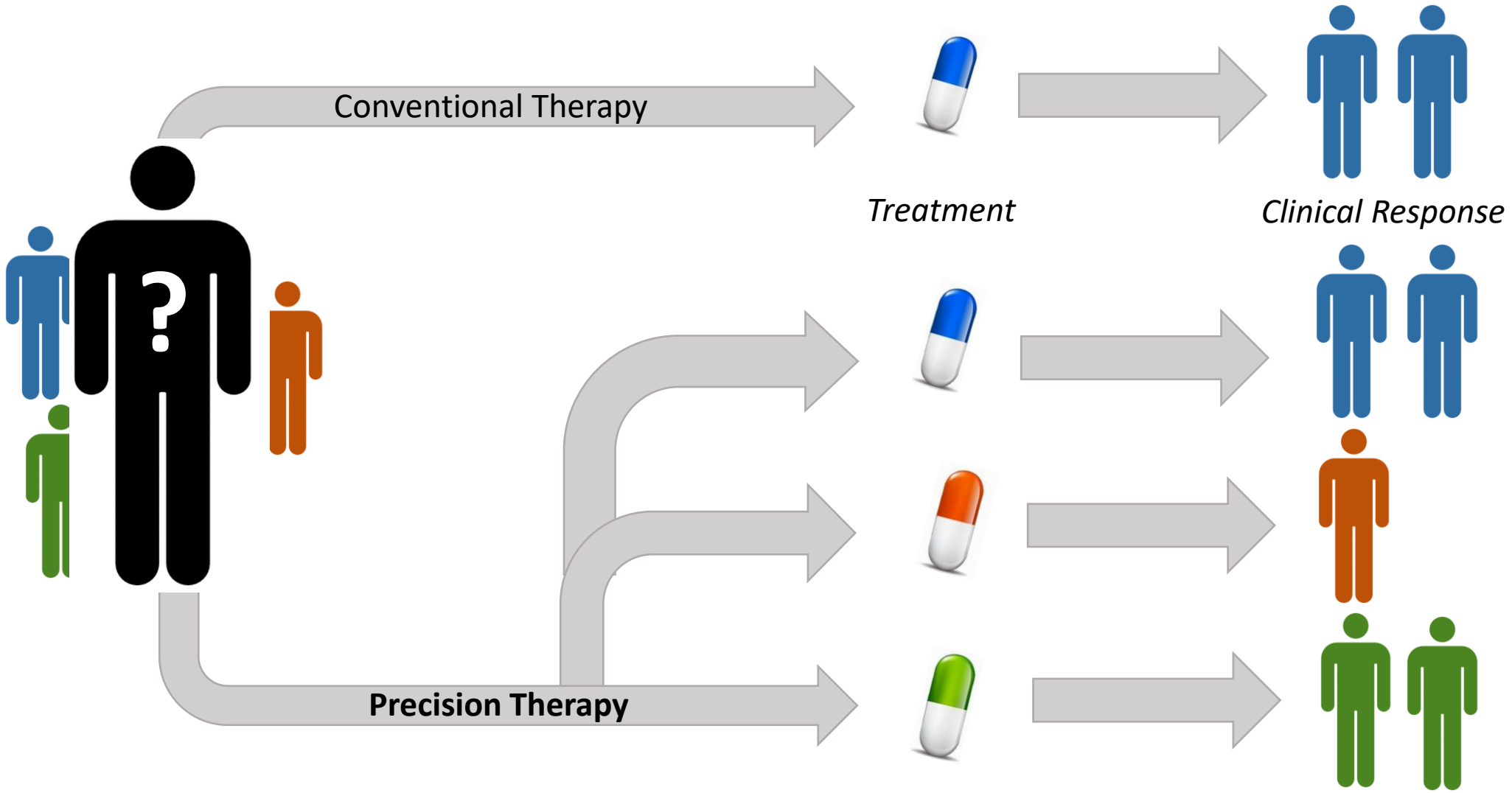
More information in https://www.fct.unl.pt/en/education/course/master-molecular-genetics-and-biomedicine

Biomedical Big-Data

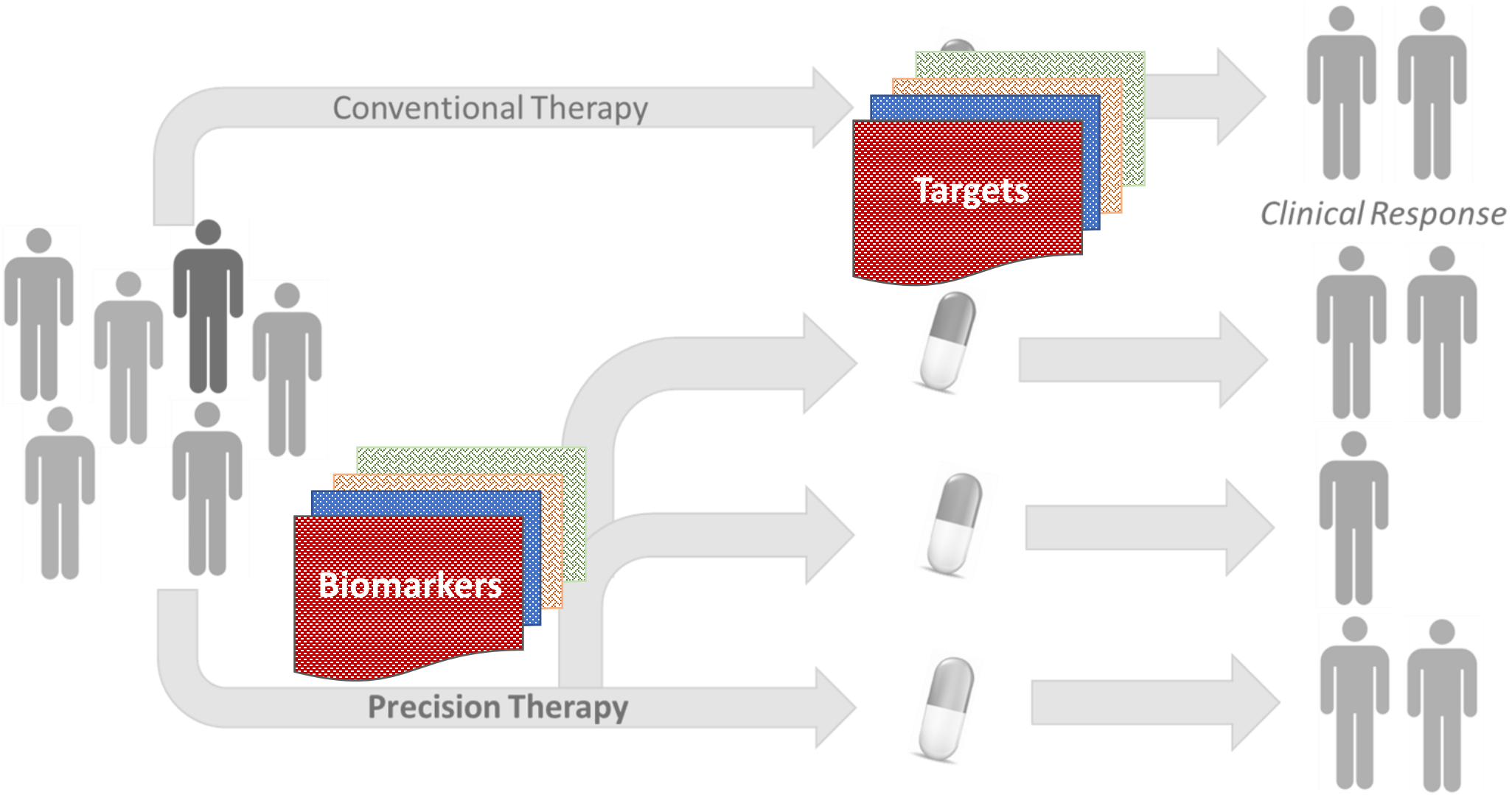


Sonawane et al (2019) *Front Genetics*

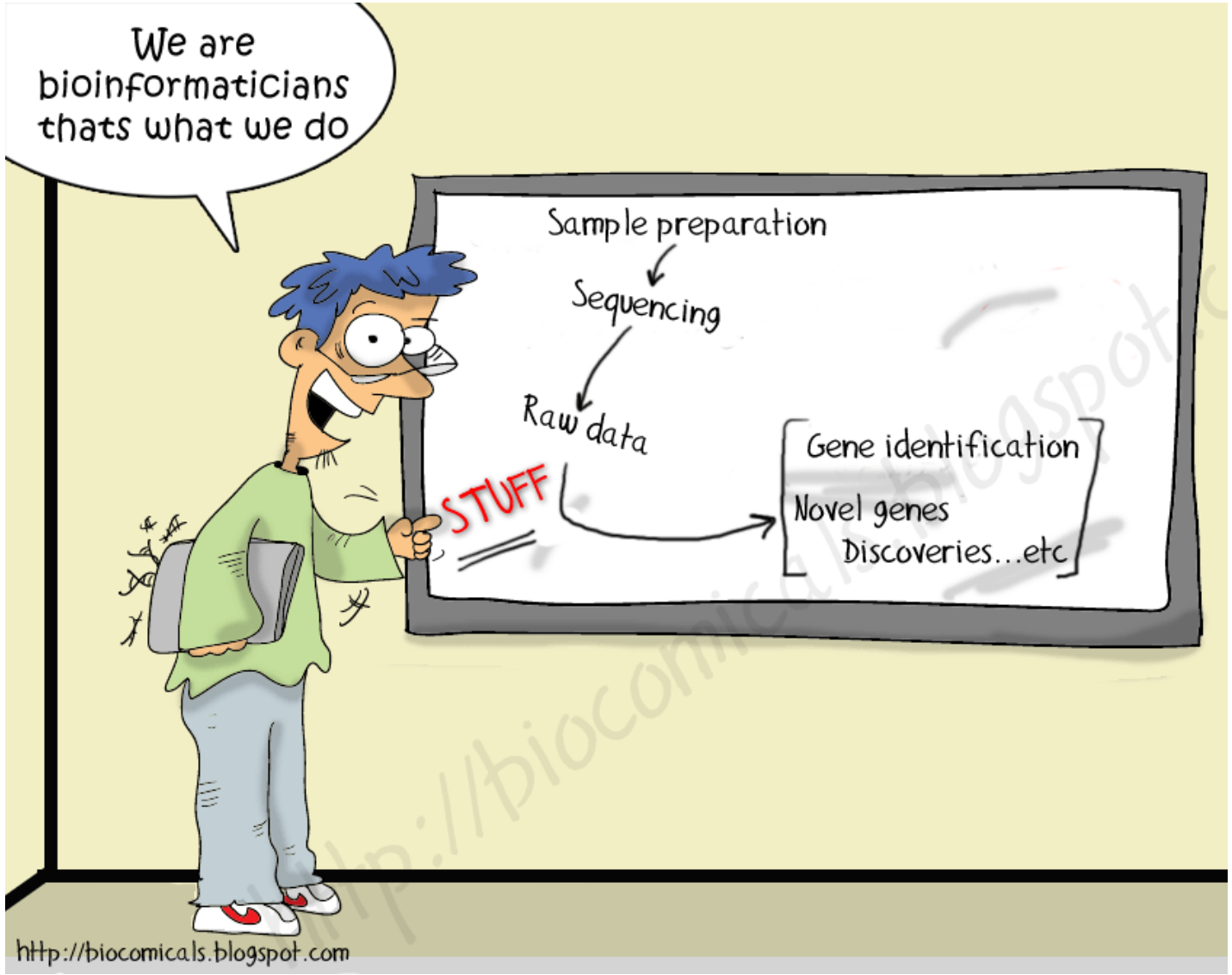
Precision Medicine



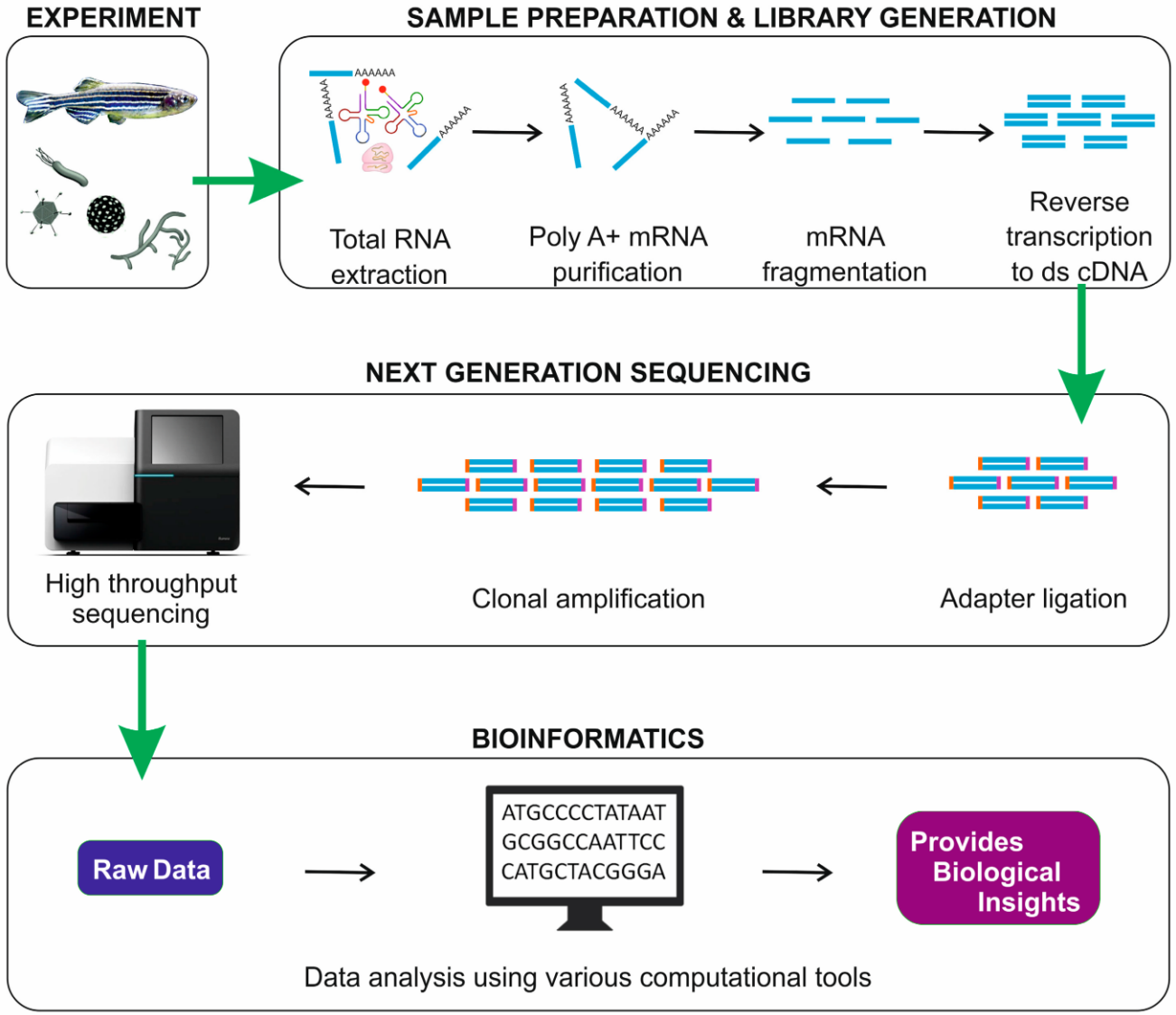
Precision Medicine



High-throughput Sequencing Data Analysis



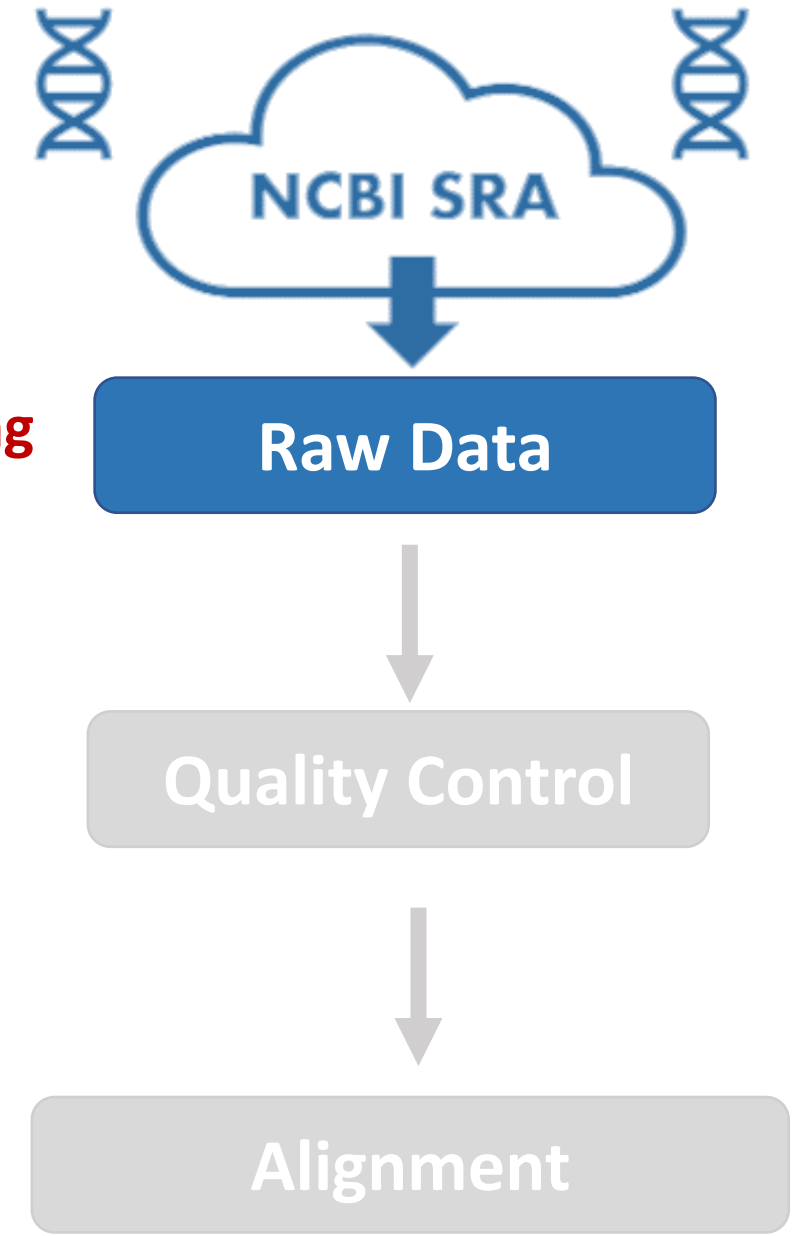
High-throughput RNA sequencing (RNA-seq)



Sudhagar et al (2018) Int J Mol Sci

HTS Data Analysis

1) Get raw data from SRA using <u>SRA tools</u>



HTS Data Repositories

Sequencing Data
(raw or analysed data)

Genome Sequences and Annotations

EMBL-EBI

ENA European Nucleotide Archive

e!Ensembl

GENCODE

NCBI National Center for Biotechnology Information

GEO Gene Expression Omnibus

SRA

NCBI Gene

ENCODE

UCSC Genome Bioinformatics

HTS Data Repositories: SRA

NCBI Resources How To agrosso My NCBI Sign Out

SRA | SRA | glioblastoma | Search

Create alert Advanced Help

Access
Controlled (479)
Public (15,304)

Source
DNA (2,722)
RNA (12,740)

Type
exome (651)
genome (150)

Library Layout
paired (11,281)
single (4,506)

Platform
ABI SOLiD (36)
BGISEQ (84)
Complete Genomics (10)
Helicos (3)
Illumina (15,431)
Ion Torrent (142)
LS454 (40)
Oxford Nanopore (41)

Strategy
EpiGenomics (378)
Exome (1,298)
Genome (151)
RNASeq (58)
other (13,902)

Summary 20 per page Send to:

View results as an expanded interactive table using the RunSelector. Send results to Run selector

Search results

Items: 1 to 20 of 15787

<< First < Prev Page 1 of 790 Next > Last >>

1. Illumina NovaSeq 6000 paired end sequencing; Single cell RNA sequencing of three IDHwt **glioblastoma** patient tumors
   Accession: ERX6426595
2. Illumina NovaSeq 6000 paired end sequencing; Single cell RNA sequencing of three IDHwt **glioblastoma** patient tumors
   Accession: ERX6426594
3. Illumina NovaSeq 6000 paired end sequencing; Single cell RNA sequencing of three IDHwt **glioblastoma** patient tumors
   Accession: ERX6426593
4. GSM5589195: IR rep 3; Mus musculus; RNA-Seq
   1 ILLUMINA (NextSeq 500) run: 23.4M spots, 1.8G bases, 648Mb downloads
   Accession: SRX12261280
5. GSM5589194: IR rep 2; Mus musculus; RNA-Seq
   1 ILLUMINA (NextSeq 500) run: 23.9M spots, 1.8G bases, 663.9Mb downloads
   Accession: SRX12261279

Filter your results:
All (15787)
type_rnaseq (12124)
access: Controlled (479)
access: Public (15304)
aligned data (1393)
source: DNA (2722)
source: metagenomic (0)
source: RNA (12740)
type: exome (651)
type: genome (150)

Manage Filters

Results by taxon
Top Organisms [Tree]
Homo sapiens *(13367)*
Mus musculus *(2399)*
Canis lupus familiaris *(12)*
Rattus norvegicus *(8)*
human metagenome *(1)*
More...

Top Bioprojects
Production ENCODE functional... *(9)*

HTS Data Repositories: SRA

National Library of Medicine
National Center for Biotechnology Information

Log in

SRA | SRA | Search

Advanced | Help

Full | Send to:

Design: Single cell RNA sequencing of three IDHwt glioblastoma patient tumors

Submitted by: Department of Neurological Surgery Washington University in St. Louis (Department of Neurological Surgery Washington Univ)

Study: Single cell RNA sequencing of three IDHwt glioblastoma patient tumors

PRJEB47680 • ERP131977 • All experiments • All runs

show Abstract

Library:

Name: GBM2_total_3p_GEX_CR3_p
Instrument: Illumina NovaSeq 6000
Strategy: OTHER
Source: TRANSCRIPTOMIC SINGLE CELL
Selection: Oligo-dT
Layout: PAIRED
Construction protocol: Cell dissociation and tissue processing: Fresh tumor samples were dissociated using the gentleMACS Octo Dissociator and Brain tumor dissociation kit (Miltenyi). Myelin and red blood cells were removed using the MACS Myelin removal beads II and ACK Red Blood Cell lysis buffer, respectively. Dead cells were removed (MACS Dead cell removal microbeads) and viable cells were counted by Trypan blue exclusion. Fresh tumor samples were dissociated using the gentleMACS Octo Dissociator and Brain tumor dissociation kit (Miltenyi). Myelin and red blood cells were removed using the MACS Myelin removal beads II and ACK Red Blood Cell lysis buffer, respectively. Dead cells were removed (MACS Dead cell removal microbeads) and viable cells were counted by Trypan blue exclusion. Dissociated tumor cells were processed using the 10x Genomics Chromium Controller and the Chromium Single Cell 3'V2 Library & Gel Bead Kit following the manufacturer's protocols (https://tinyurl.com/ybpg2pfz). Dissociated tumor cells were processed using the 10x Genomics Chromium Controller and the Chromium Single Cell 3'V2 Library & Gel Bead Kit following the manufacturer's protocols (https://tinyurl.com/ybpg2pfz).

Related information

BioProject
BioSample
Taxonomy

Recent activity

Turn Off | Clear

glioblastoma ERX6426595 (1) SRA
glioblastoma ERX6426595 AND ("biomol rna"[Properties]) (0) SRA
glioblastoma ERX6426595 AND ("biomol rna"[Properties] AND "librar... (0) SRA
glioblastoma AND ("biomol rna"[Properties] AND "library layout pa... (13583) SRA
glioblastoma AND ("biomol rna"[Properties]) (17745) SRA

See more...

Raw Data

Raw Data File

(file with extension ".fastq", containing ~200M reads/sequences)

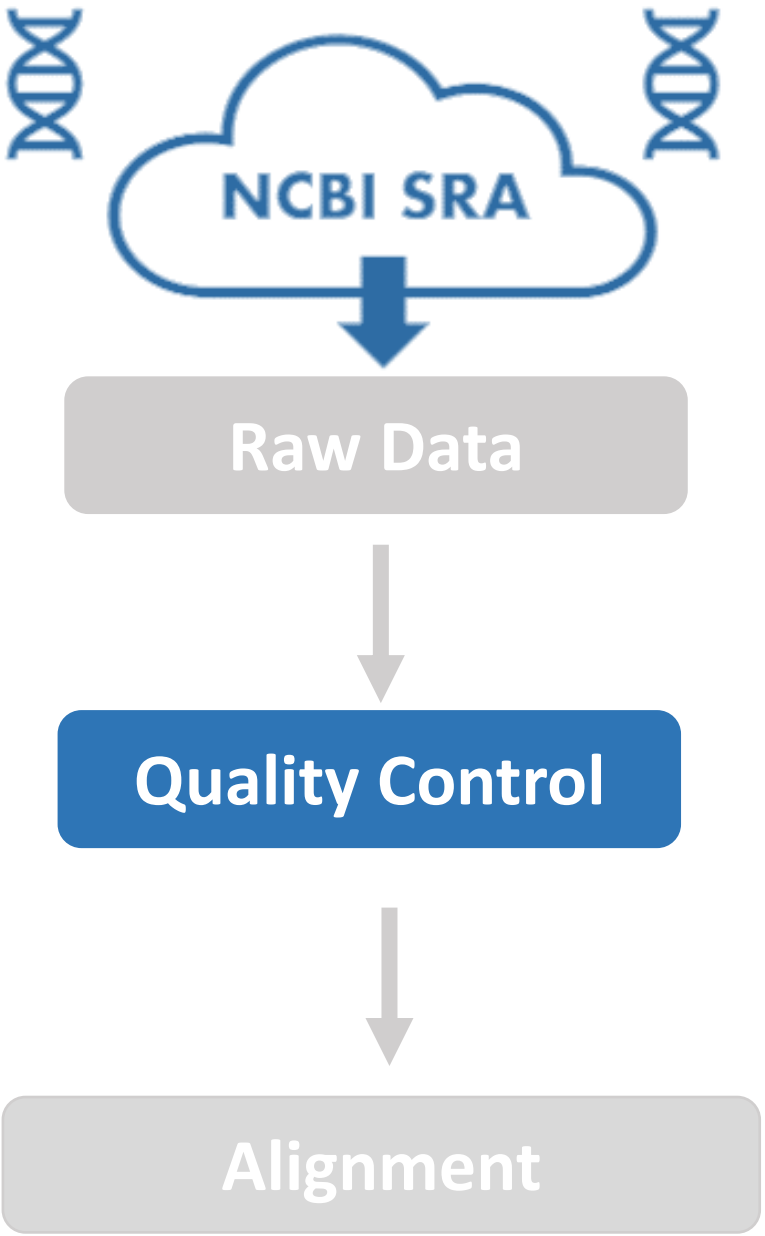
Identifier	@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
Sequence	TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
'+' sign	+
Quality scores	hhhhhhhhhghhghhhhhfhhhhhfffffe'ee['X]b[d[ed'[Y[^Y
Identifier	@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
Sequence	GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
'+' sign	+
Quality scores	hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde'bVd

Single-End Read

Paired-End Read

Hosseini et al (2016) *Information*

HTS Data Analysis



2) Assess data quality using <u>FastQC</u>

HTS Data Analysis – Quality Control

- **HTS artefacts/problems:**
  - Sequence quality
  - Nucleotide composition
  - Technical issues
  - Overrepresented sequences
  - Adaptor sequence presence
  - …

More information in:
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Data Quality Assessment

FastQC Report

FastQC Report

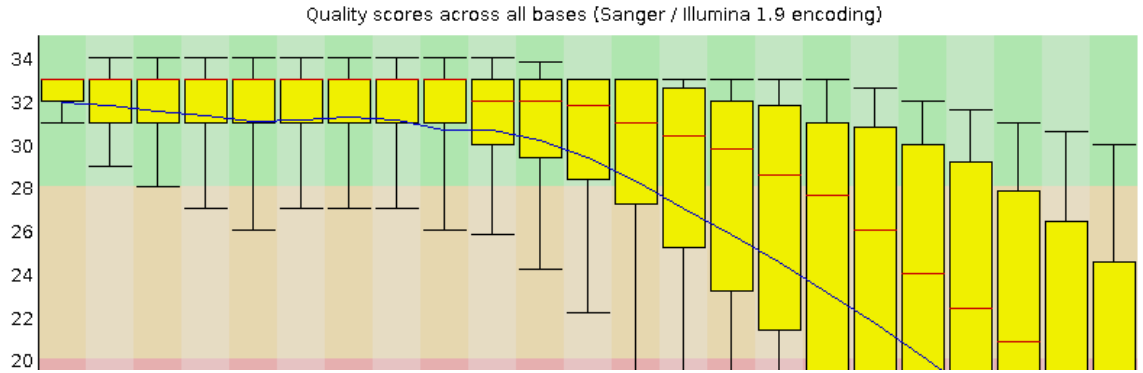
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Basic Statistics

Measure	Value
Filename	SRR040000.sra_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	19756540
Filtered Sequences	0
Sequence length	76
%GC	49

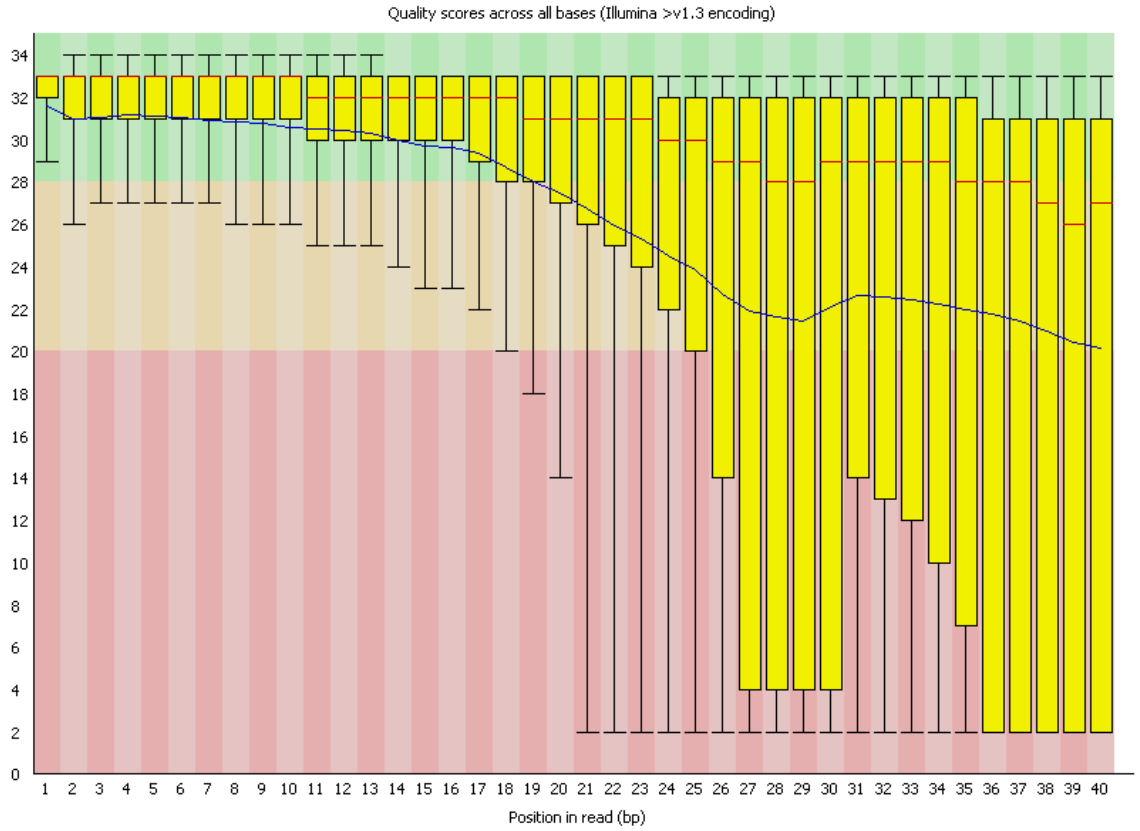
Per base sequence quality



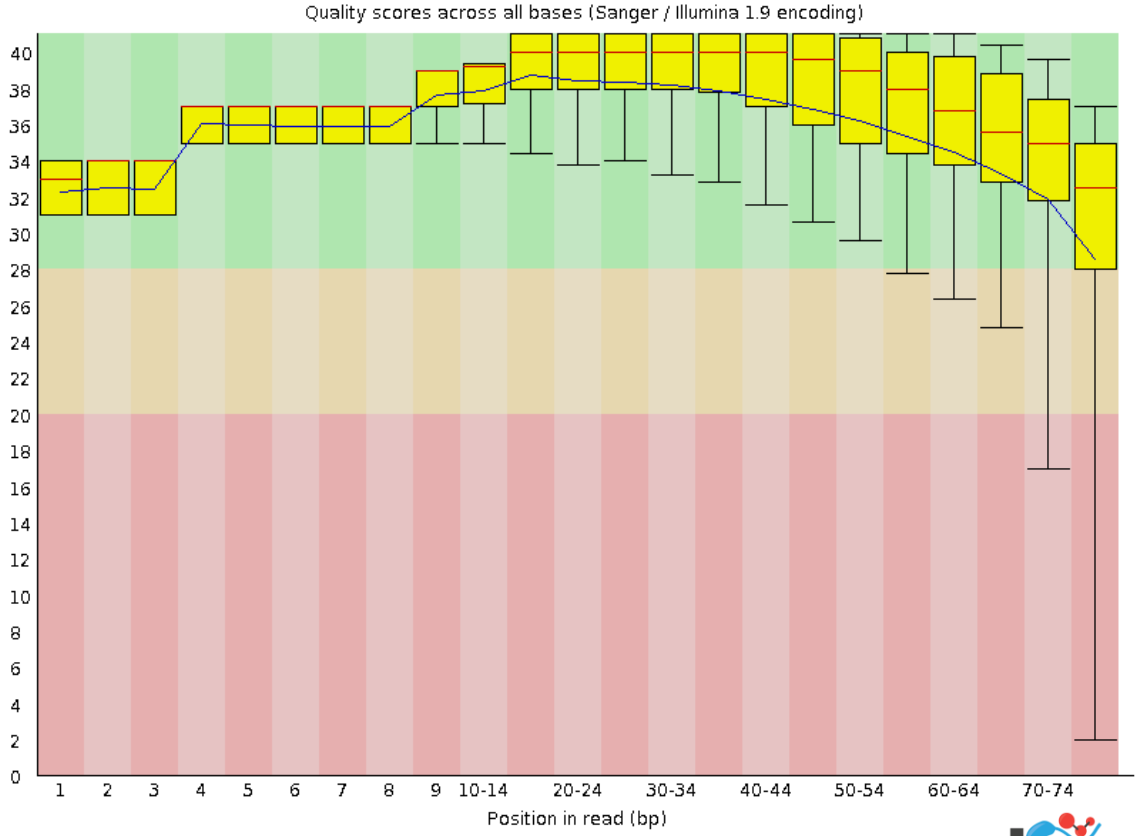
http://www.bioinformatics.babraham.ac.uk

Data Quality Assessment
Per Base Sequence Quality

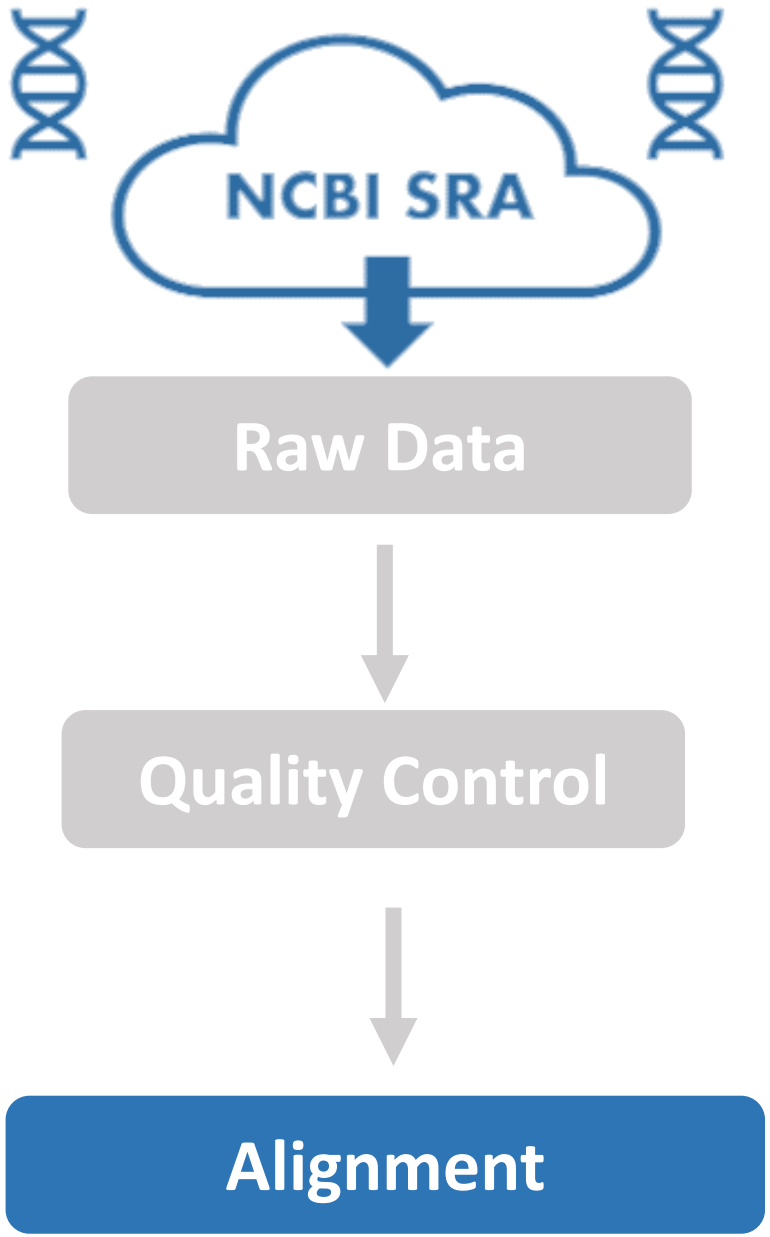
Bad Data



Good Data



HTS Data Analysis



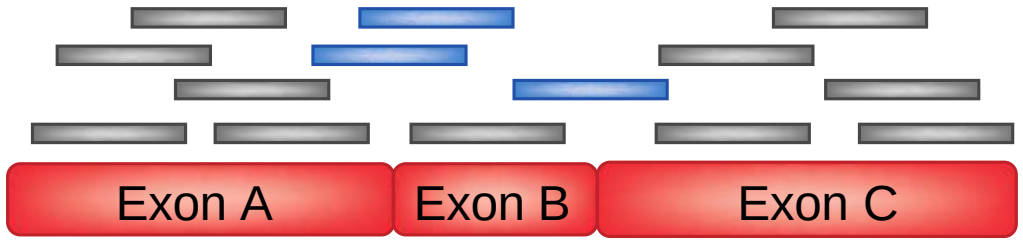
3) Align data to transcriptome and genome (Kallisto and STAR)

RNAseq Data Analysis

Data Alignment: genome *versus* transcriptome

Align reads to Reference:

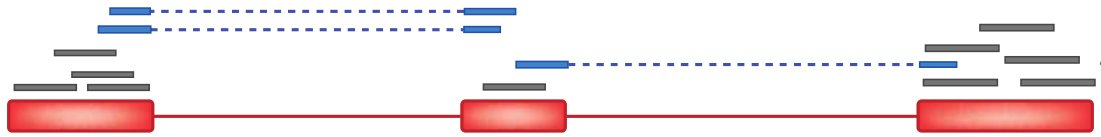
Transcriptome



Processed mRNA

Advantages: Easy, focused on the known transcribed regions
Disadvantages: Reads from novel isoforms may not align or be aligned to wrong isoform

Genome



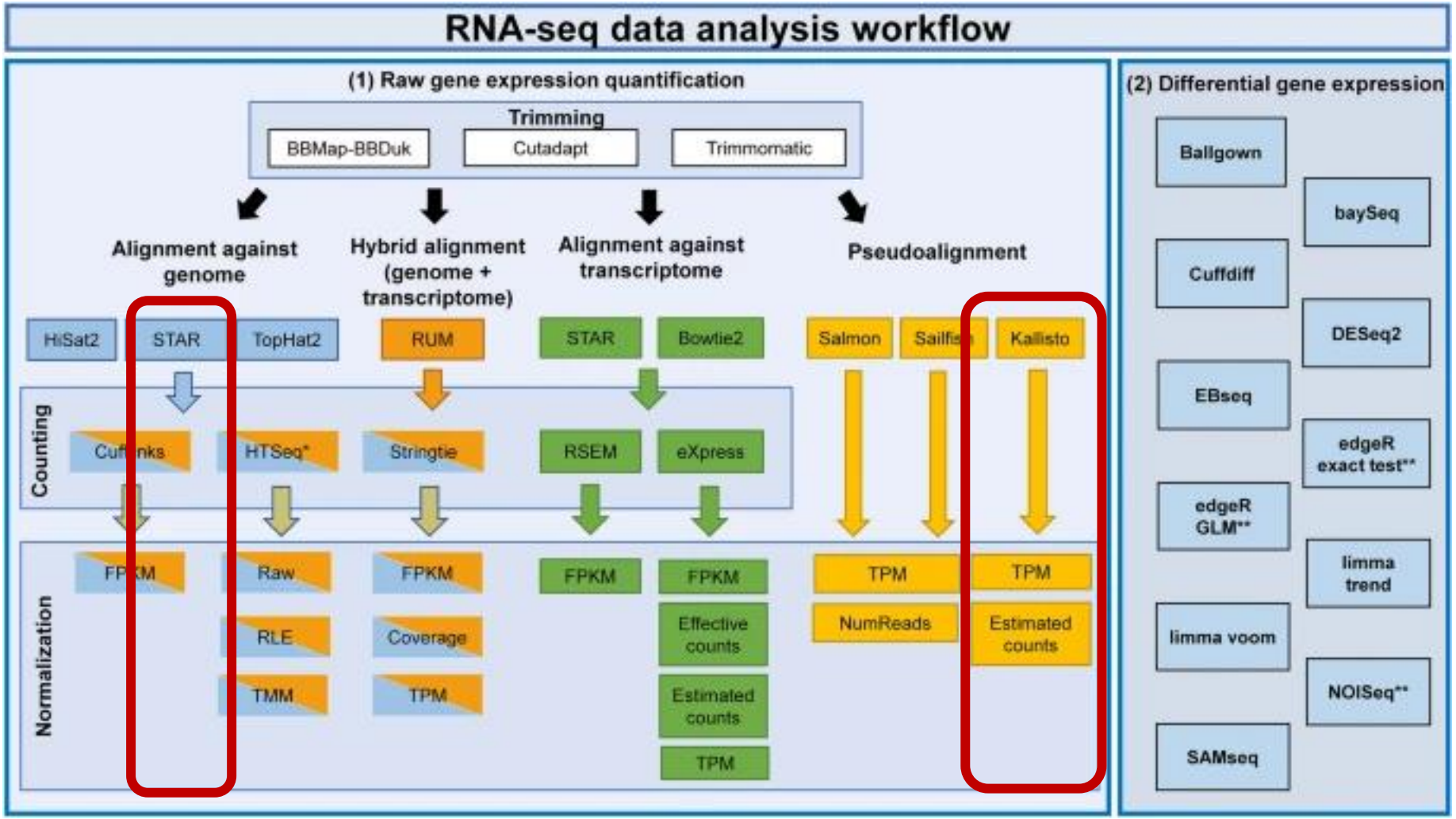
Mapping to genome

Advantages: Can align novel isoforms
Disadvantages: less time-efficient, spliced alignments

Trapnell et al (2009) *Bioinformatics;* Trapnell and Salzberg (2009) *Nat Biotechnol*

RNAseq Data Analysis

Bioinformatics Tools



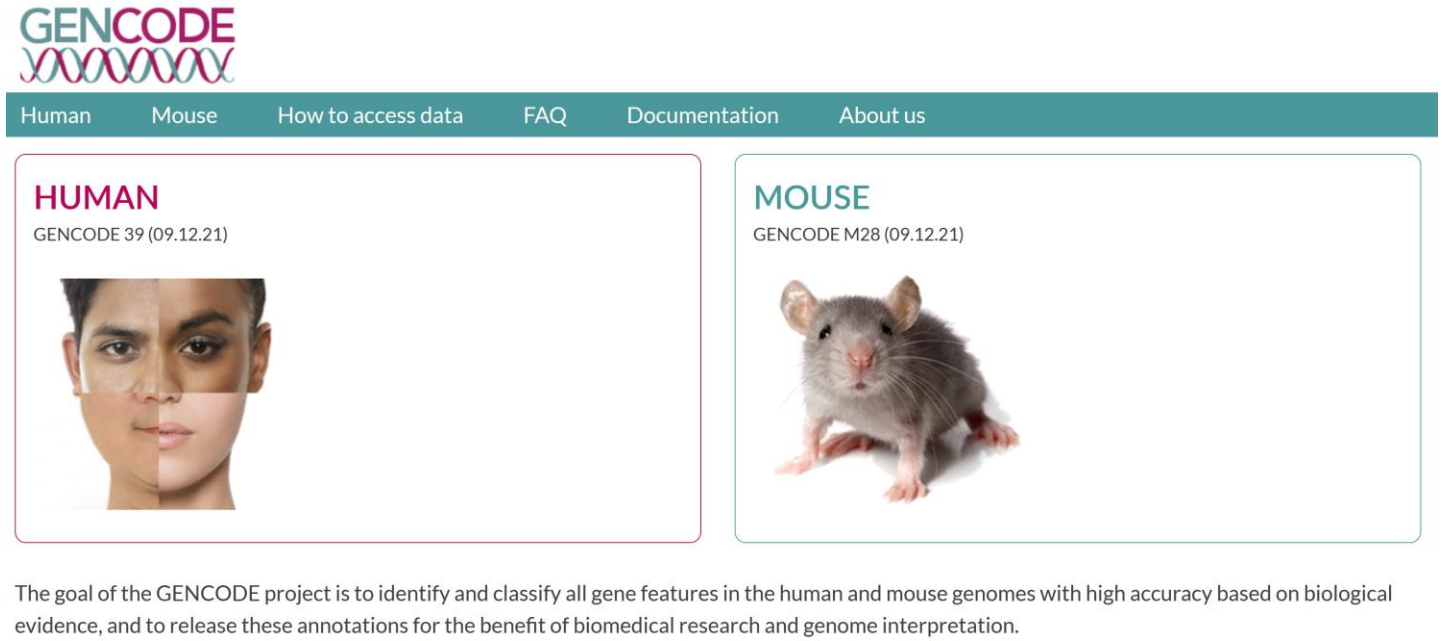
Corchete et al (2020) *Scientific Reports*

RNAseq Data Analysis

Data Alignment: genome/transcriptome versions

Genome/Transcriptome Versions for most Vertebrates

Transcriptome and Gene annotations for Human and Mouse



https://www.gencodegenes.org/human/releases.html

HTS Data Analysis – Transcriptome Alignment

Kallisto output (one file for each sample)

target_id	length	eff_length	est_counts	tpm
ENST00000513300.5	1924	1746.98	102.328	11129.2
ENST00000282507.7	2355	2177.98	1592.02	138884
ENST00000504685.5	1476	1298.98	68.6528	10041.8
ENST00000243108.4	1733	1555.98	343.499	41944.9
ENST00000303450.4	1516	1338.98	664	94221.8
ENST00000243082.4	2039	1861.98	55	5612.36
ENST00000303406.4	1524	1346.98	304.189	42908.2
ENST00000303460.4	1936	1758.98	47	5076.85
ENST00000243056.4	2423	2245.98	42	3553.05
ENST00000312492.2	1805	1627.98	228	26609.9
ENST00000040584.5	1889	1711.98	4295	476675
ENST00000430889.2	1666	1488.98	623.628	79578.2
ENST00000394331.3	2943	2765.98	85.6842	5885.85
ENST00000243103.3	3335	3157.98	962	57879.3

Estimated read counts

- Read counts for each transcript/sample
- Used in Differential expression analysis

Transcripts per Millions (TPMs)

- Normalized read counts
- Used in Exploratory analysis

HTS Data Analysis – Genome Alignment

Sequence Alignment/Map Format (SAM): a file format to represent alignments

SAM file simpli

Flow cycle

@RG ID:3G8KR.IonXpress_006.....**FO:TACGTACGTCTGAGCATCGATCGATGTACAGCTACGTACGTCTGAGCATCGATCGATG**
TACAGCTACGTACGTCTGAGCATCGATCGATGTACAGCTACGTACGTCTGAGCATCGATCGATGTACAGCTACGTACGTCTGAGCATC
GATCGATGTACAGCTACGTACGTCTGAGCATCGATCGATGTACA.....KS:TCAGCTGCAAGTTCGAT

Header line

Key sequence

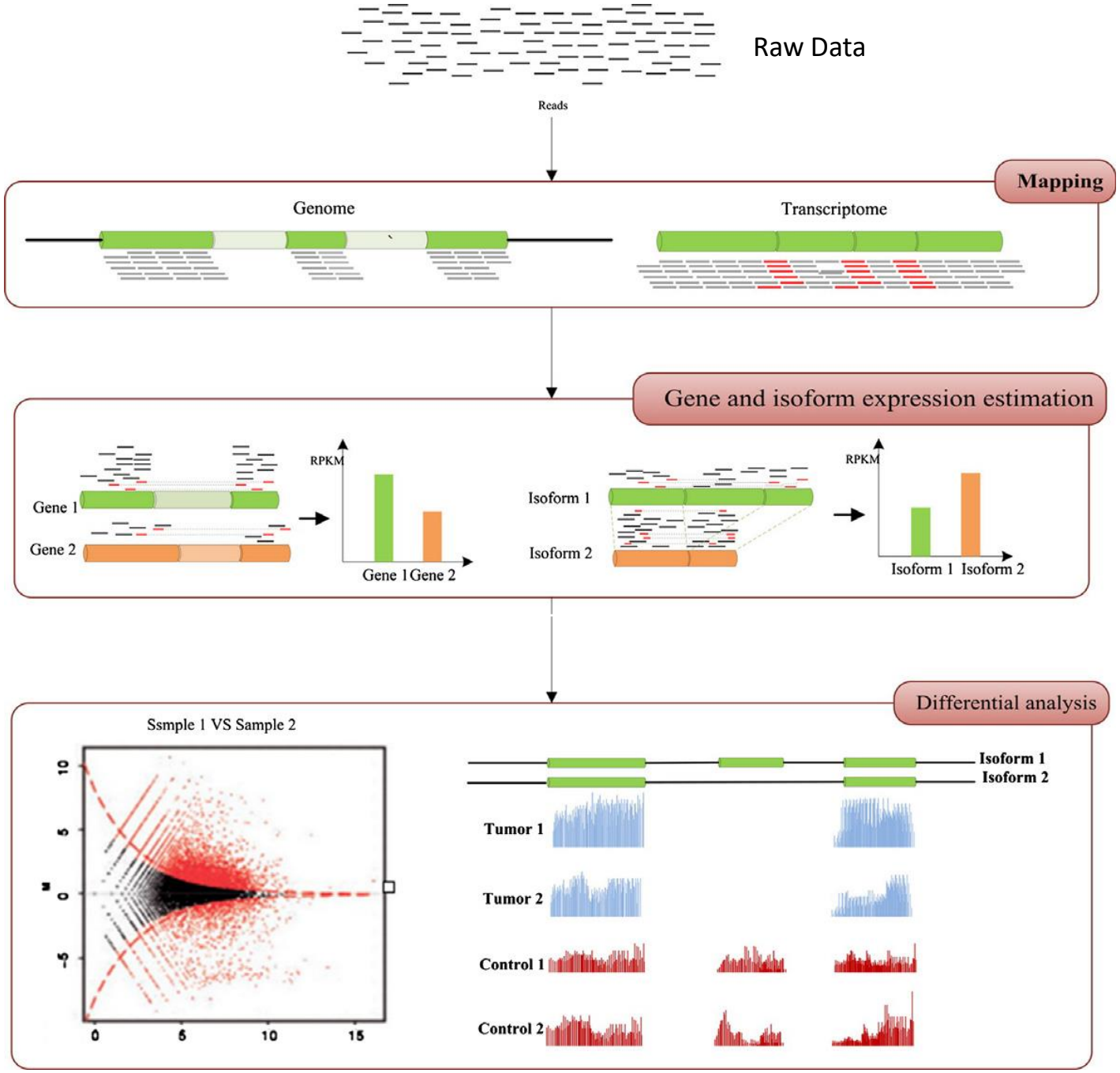
ID		pos		CGAR			Sequence		Quality		Flow Signal
3G8KR:02859	..	chr1	14725	..	192M	.....	TCAGCTGCAAGT	...	@<<?9=;>;9=<	.....	ZM:B:s,266,-48,242,-12,-50,226,48,274,-56
3G8KR:01234	..	chr1	14725	..	192M	.....	TCAGCTGCAAGT	...	<;;;7<:;;8===<;	.....	ZM:B:s,238,-14,260,-32,10,254,16,232,0,21
3G8KR:00083	..	chr1	14725	..	192M	.....	TCAGCTGCAAGT	...	<==?9>===9<9	.....	ZM:B:s,224,-10,244,-24,-10,246,-22,244,-2,19
3G8KR:00315	..	chr1	14725	..	192M	.....	TCAGCTGCAAGT	...	;;<=9?>==9=>>	.....	ZM:B:s,272,-38,214,-22,-6,216,-2,284,-10,244
3G8KR:01099	..	chr1	14725	..	192M	.....	TCAGCTGCAAGT	...	;;==8<===8<;>	.....	ZM:B:s,258,-50,244,-12,-4,234,2,224,-44,208,
3G8KR:07971	..	chr1	14725	..	192M	.....	TCAGCTGCAAGT	...	===>9>><<6<<	.....	ZM:B:s,246, -10,272,-26,-2,248,8,250,-18,194,
3G8KR:01648	..	chr1	14725	..	192M	.....	TCAGCTGCAAGT	...	=<<<9<<<=9=@	.....	ZM:B:s,248,-32,246,0,2,234,-20,246,-12,202,2
3G8KR:02227	..	chr1	14725	..	192M	.....	TCAGCTGCAAGT	...	<<;<6<;;<8;;<;[	.....	ZM:B:s,244,-44,244,0,-8,246,0,250,-10,222,24
3G8KR:02263	..	chr1	14725	..	192M	.....	TCAGCTGCAAGT	...	;<<<8;;;=8===<	.....	ZM:B:s,278,-16,248,-40,4,230,-6,242,0,242,23
3G8KR:06082	..	chr1	14725	..	192M	.....	TCAGCTGCAAGT	...	<<<<6<<<<6;;<	.....	ZM:B:s,310,-21,250,44,14,224,-36,232,4,226,

Read line

Shokrof and Abouelhoda (2020) *BMC Bioinformatics*

HTS Data Analysis

What's next?



Next!!

Feng et al (2012) *Cancer Lett*

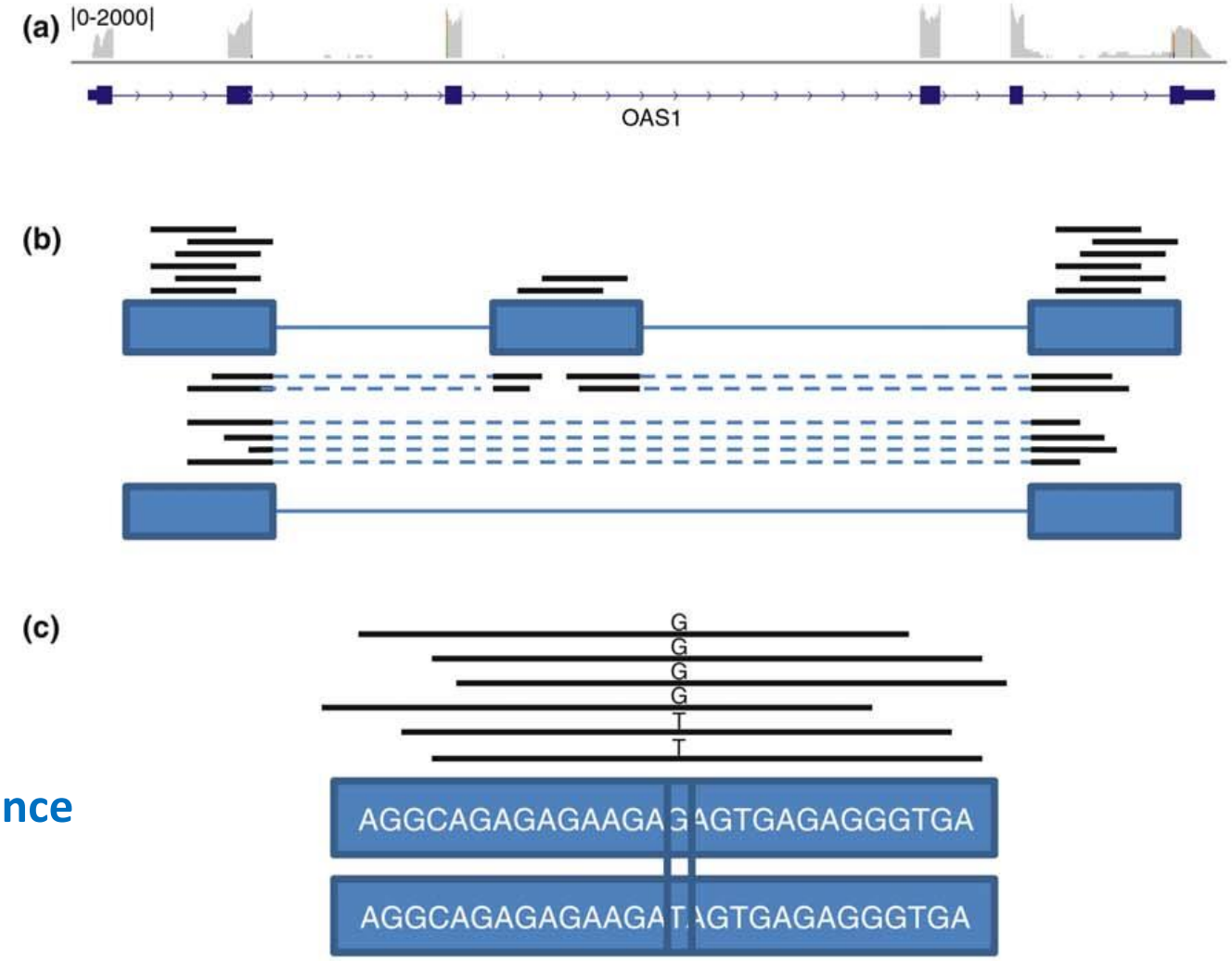
HTS Data Analysis

What's next?

Gene Level

Alternative Splicing

Allelic Imbalance

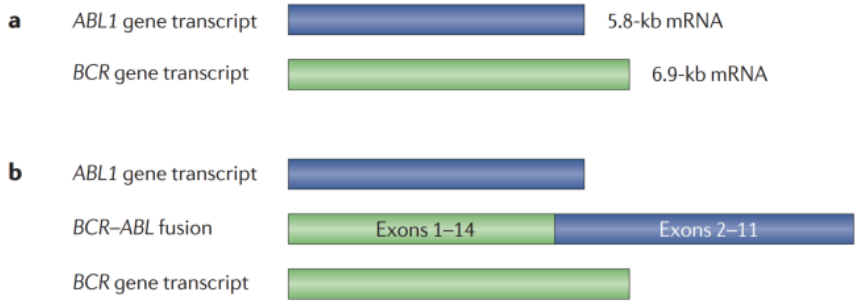


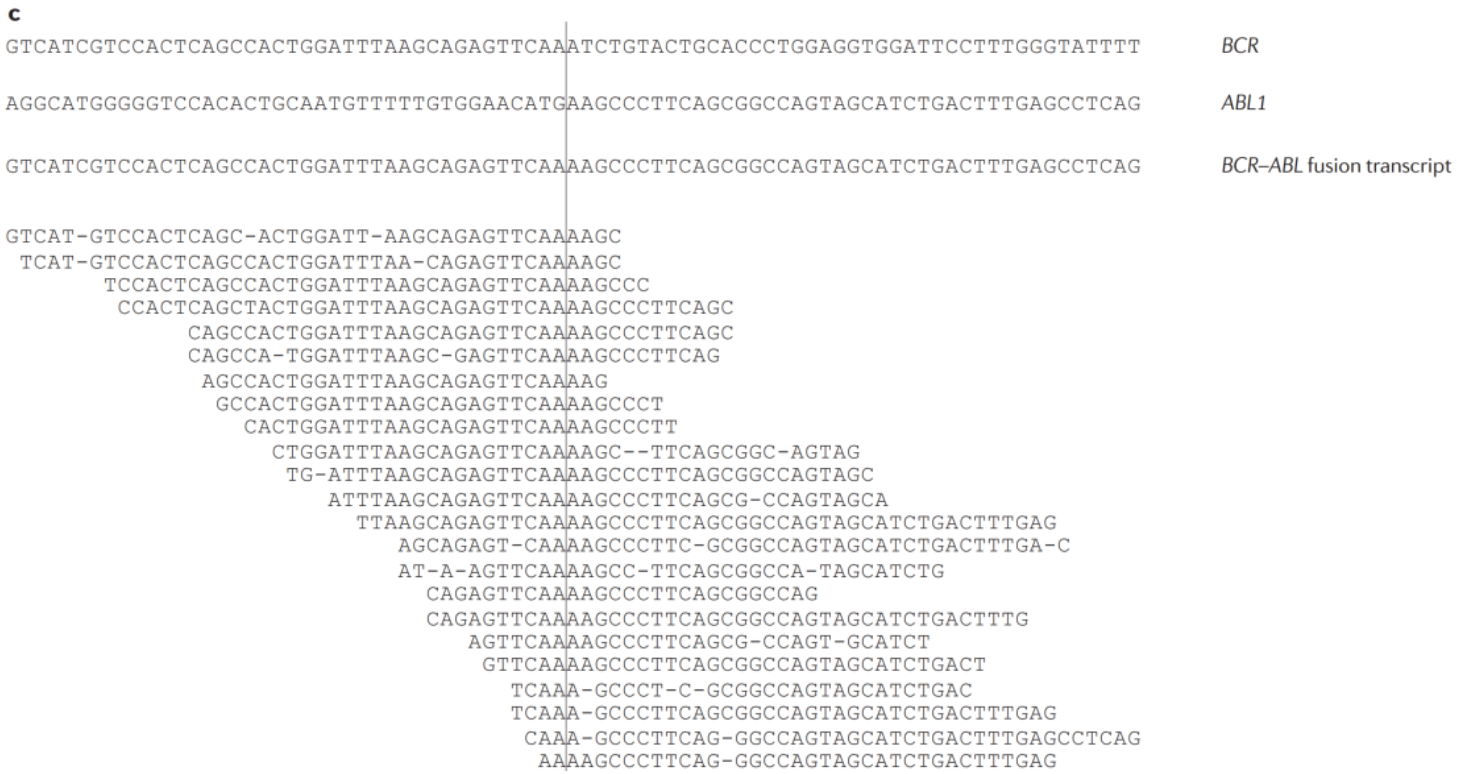
Majewski and Pastinen (2011) *Trend in Genetics*

HTS Data Analysis

What's next?

Gene fusion





Ozsolak and Milos (2011) *Nature Review Genetics*

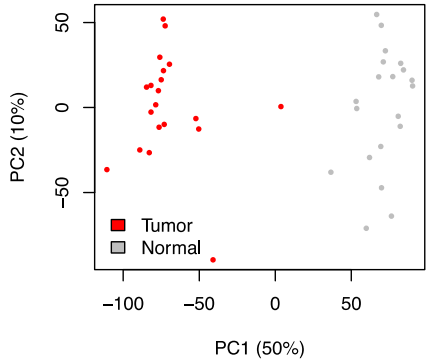
HTS Data Analysis

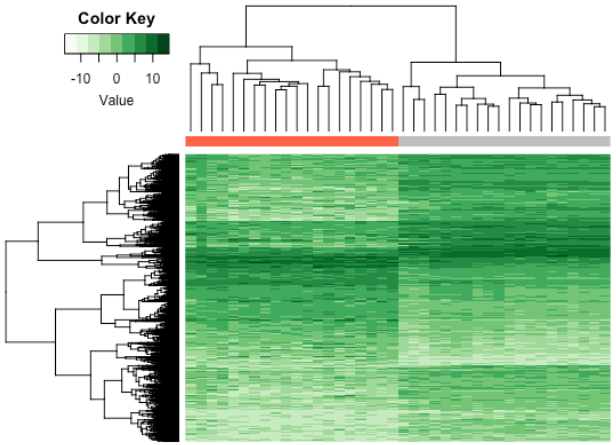
What's next?

Transcriptome Profiles

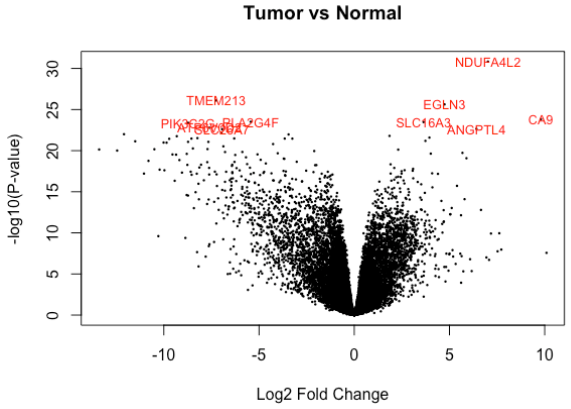
	TCGA.A3.3358.01	TCGA.A3.3387.01	TCGA.B0.4700.01	TCGA.B0.4712.01	TCGA.B0.5402.01	TCGA.B0.5690.01
EEF1A1	13.98220636	13.8186912	12.863860	13.1913250	14.340716	13.0939714
GPX3	11.88457267	11.4090069	9.920329	10.0343242	14.340247	10.3229445
UMOD	1.80871422	-3.4935634	-0.565008	-1.3773438	-1.547451	-0.6272458
ALDOB	6.55599683	2.3503254	4.762704	-1.5762452	9.765960	5.7475390
ADAM6	12.72001691	11.1953384	12.814279	8.9573795	9.923689	8.2191920
CD74	13.09458021	13.0961807	13.943524	12.4496722	12.859715	12.1444018
SLC12A1	0.74457882	0.9017328	-1.463406	-3.2290709	-2.624186	-2.0981981
GAPDH	12.88927117	12.9062785	13.559814	13.9336383	12.702471	12.6477753
ATP1A1	8.64132224	9.7491925	8.891576	9.5413563	8.870964	8.6871353
TGFBI	9.66085759	8.8838095	11.757477	12.3198370	10.664287	6.6940544
B2M	13.12708926	12.4926319	12.683055	11.4133982	12.249986	12.4608324
AQP2	-0.82454090	-4.2464002	-2.140967	-1.3734517	-3.238887	-2.0499339
VIM	12.31619343	12.0776191	12.769947	12.4948214	12.359648	12.5291941
RGS5	9.95871838	10.5353376	10.094421	8.9659204	10.474964	11.7380251
IGFBP3	11.42661141	12.8167620	11.217148	12.2316529	11.015955	12.4014856
FTL	13.75148593	12.9252368	14.057338	13.8082925	12.948375	12.4573550
TPT1	11.86833510	11.3236103	11.574811	11.4911512	12.129083	11.6050301
C3	10.38646612	10.6098185	8.839554	9.5241840	10.841007	7.7632210
SERPINA1	10.09616434	9.6421909	12.108501	12.5736856	10.142289	9.3116322
ACTB	11.75074690	12.1532891	13.071242	12.8642593	11.187768	12.8642851
SPARC	10.24885784	11.4576436	11.842904	11.2142744	11.449070	12.5473839

Discovering New Groups

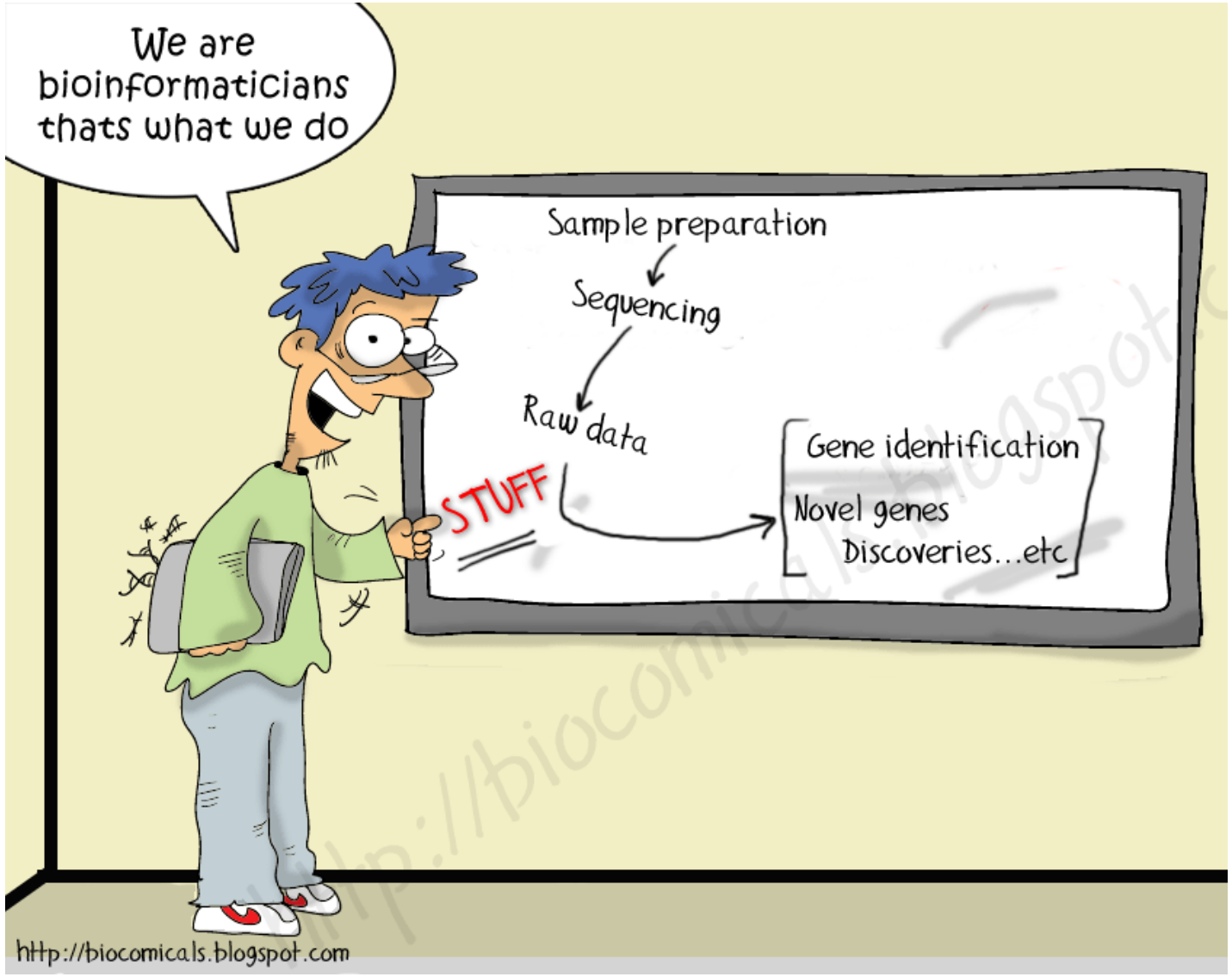




Depicting Significant Transcriptome Alterations



Hands-On…


We are bioinformaticians thats what we do
Sample preparation
Sequencing
Raw data
STUFF
Gene identification
Novel genes
Discoveries…etc
http://biocomicals.blogspot.com

HTS Data Analysis

Hands-on…

1) Get raw data from SRA using <u>SRA tools</u>

Raw Data

2) Assess data quality using <u>FastQC</u>

Quality Control

3) Align data to transcriptome and genome (Kallisto and STAR)

Alignment

NOVA SCHOOL OF SCIENCE & TECHNOLOGY
DEPARTMENT OF LIFE SCIENCES