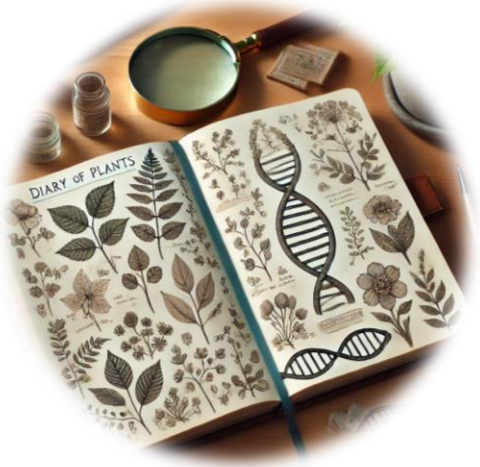
24.09.2025

INSTITUTE OF PLANT SCIENCES

Sant'Anna
School of Advanced Studies – Pisa

Genomics & Transcriptomics:
Peaking into the Diary of Plants



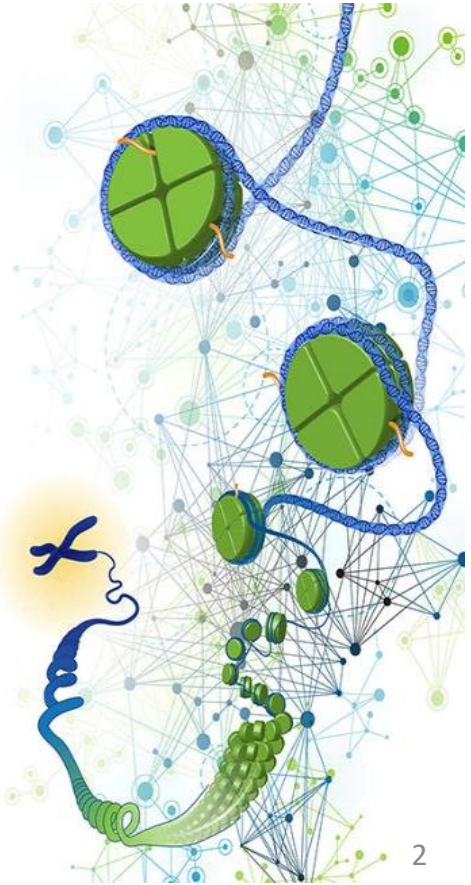
Svenja Mager
svenja.mager@santannapisa.it

About this Lecture

1
- What are Genomics
  - The Genomic Revolution

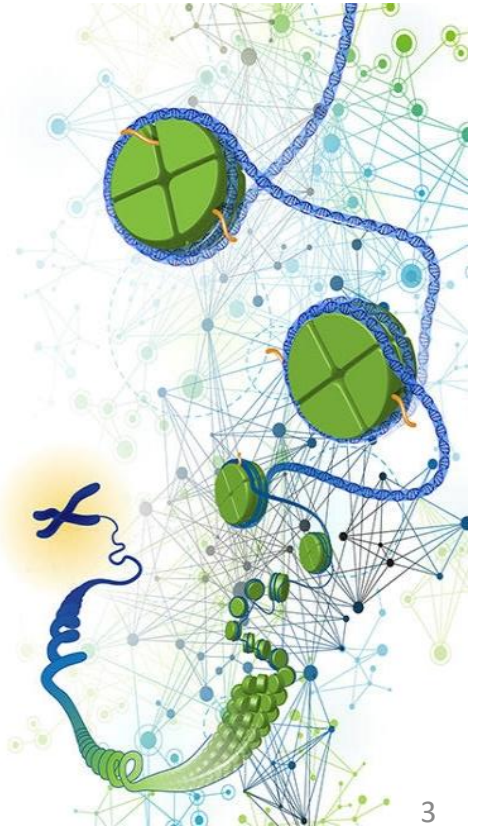
2
- Sequencing
  - Sample Preparation
  - Basic Steps for Librabry Preparation
  - DNA Sequencing
    - Illumina Short Read Sequencing
    - Nanopore Long Read Sequencing
    - DNA-Seq Data Analysis
  - RNA Sequencing (Transcriptomics)
    - RNA-Seq Data Analysis (differences in comparison to DNA-Seq)

3
- Presentation of a real Project as Example to show
  - … how a project using DNA sequencing can look like
  - … what information can be drawn from sequencing data
  - … which downstream analyses can be done with the variant file that we will produce during the hands-on practical part



Part 1

- What are Genomics
  - The Genomic Revolution



What are Genomics

An interdisciplinary field of biology studying the many aspects of genomes

Structural genomics: the study of the physical composition and organization of the genome and 3D-structure of proteins

Functional genomics: Transcriptomics (Gene expression), gene and protein function, gene interactions

Epigenomics: heritable changes in gene expression without DNA sequence changes (DNA methylation, histone modifications)

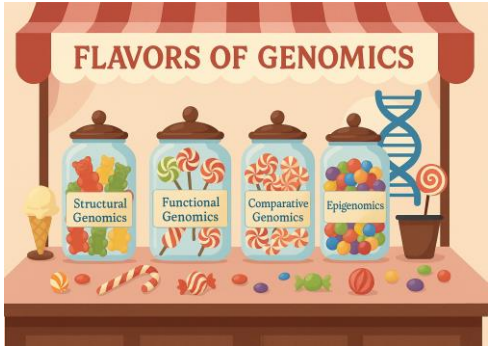
Comparative genomics: the study of the genome structure and function across different species and the evolution of genomes

Metagenomics: the study of environmental samples containing genetic material from several individuals and species

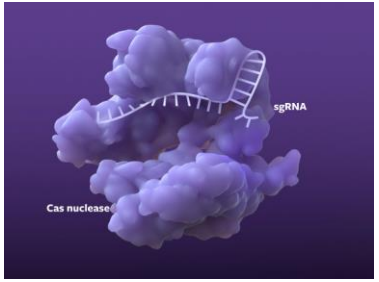
Population genomics: the study of genetic variation within and between populations, understanding evolutionary forces (selection, drift, migration)

Pangenomics/Pantranscriptomics: the determination of the entirety of genes, transcripts, differential splicing and genomes within a certain species

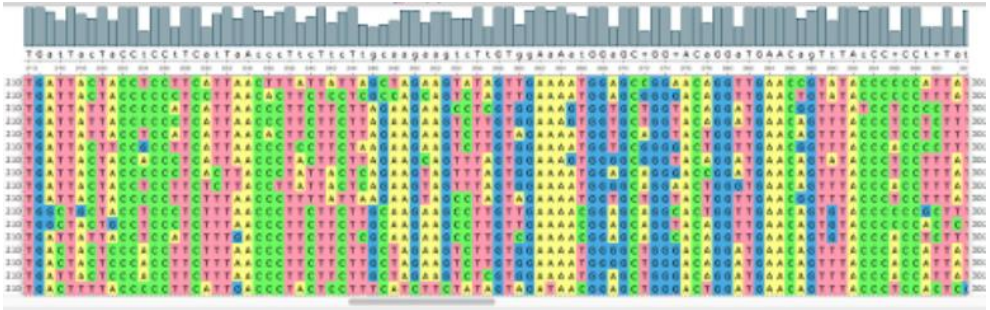
And others…

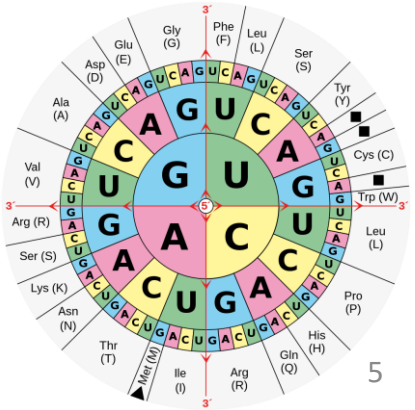


The Genomic Revolution: Definition



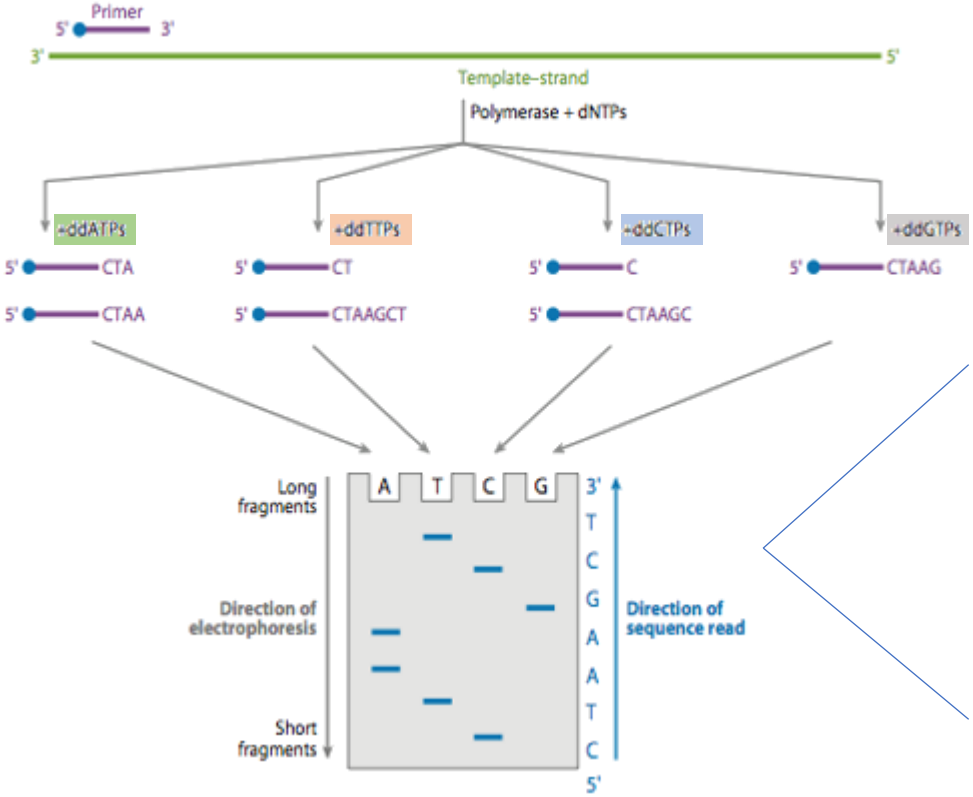
The genomic revolution refers to the **rapid advancement and widespread application of technologies** to quickly and affordably obtain and **analyze entire genomes** → transforming biology, healthcare, agriculture

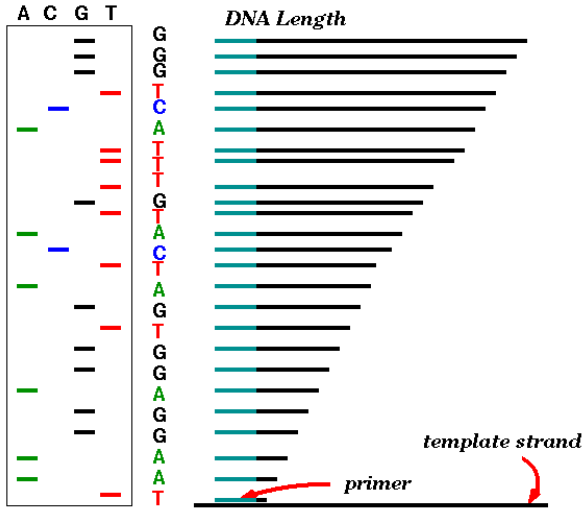


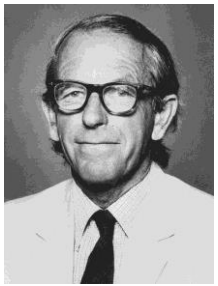


The Genomic Revolution: Beginnings…

First sequencing methods in 1970s and 1980s: Sanger and Maxam-Gilbert Sequencing



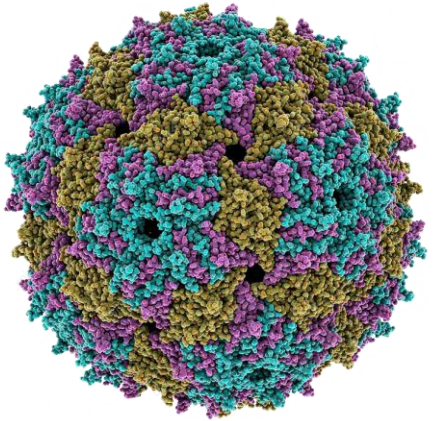




Frederick Sanger
British Biochemist
Nobel prize 1980
for first-ever DNA
sequencing
technique

The Genomic Revolution: Beginnings…

First complete genome sequenced in 1976: Bacteriophage MS2 (*Emesvirus zinderi*)



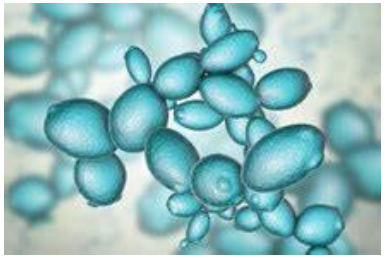
One of the smallest known genomes: **3569 single-stranded RNA nucleotides**!

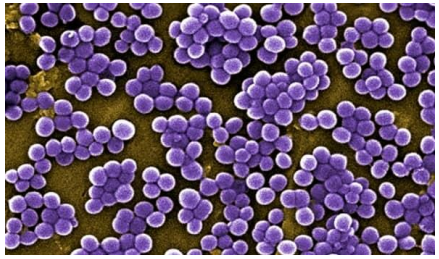
Contains only 4 proteins:

- coat protein (*cp*) → protein shell of the virus, enclosing the genome
  First gene to be completely sequenced (1972)
- maturation protein (*A*-protein) → attaches to bacteria during infection
- replicase (*rep*) protein → replicates the RNA genome
- lysis (*lys*) protein → lysis of infected bacterial cell to release new virions

The Genomic Revolution: Beginnings…

In **1992**, first fully sequenced **chromosome** (yeast chromosome III → 315 Kbp)





In **1995**, first organism's fully sequenced **genome** (*Haemophilus influenzae* → 1.8 Mbp)

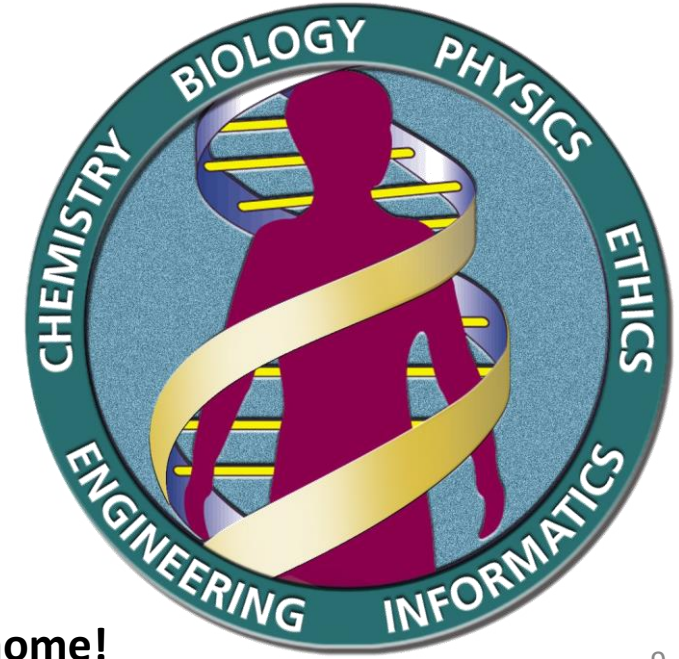
After that, several other, still **relatively small genomes** followed (bacteria and archaea)

The Genomic Revolution: «Real» Start

The “Revolution“ started with the Human Genome Project in 1990

Goals:

- determine the complete genome sequence
- identifying, mapping and sequencing all genes → physical and functional characteristics

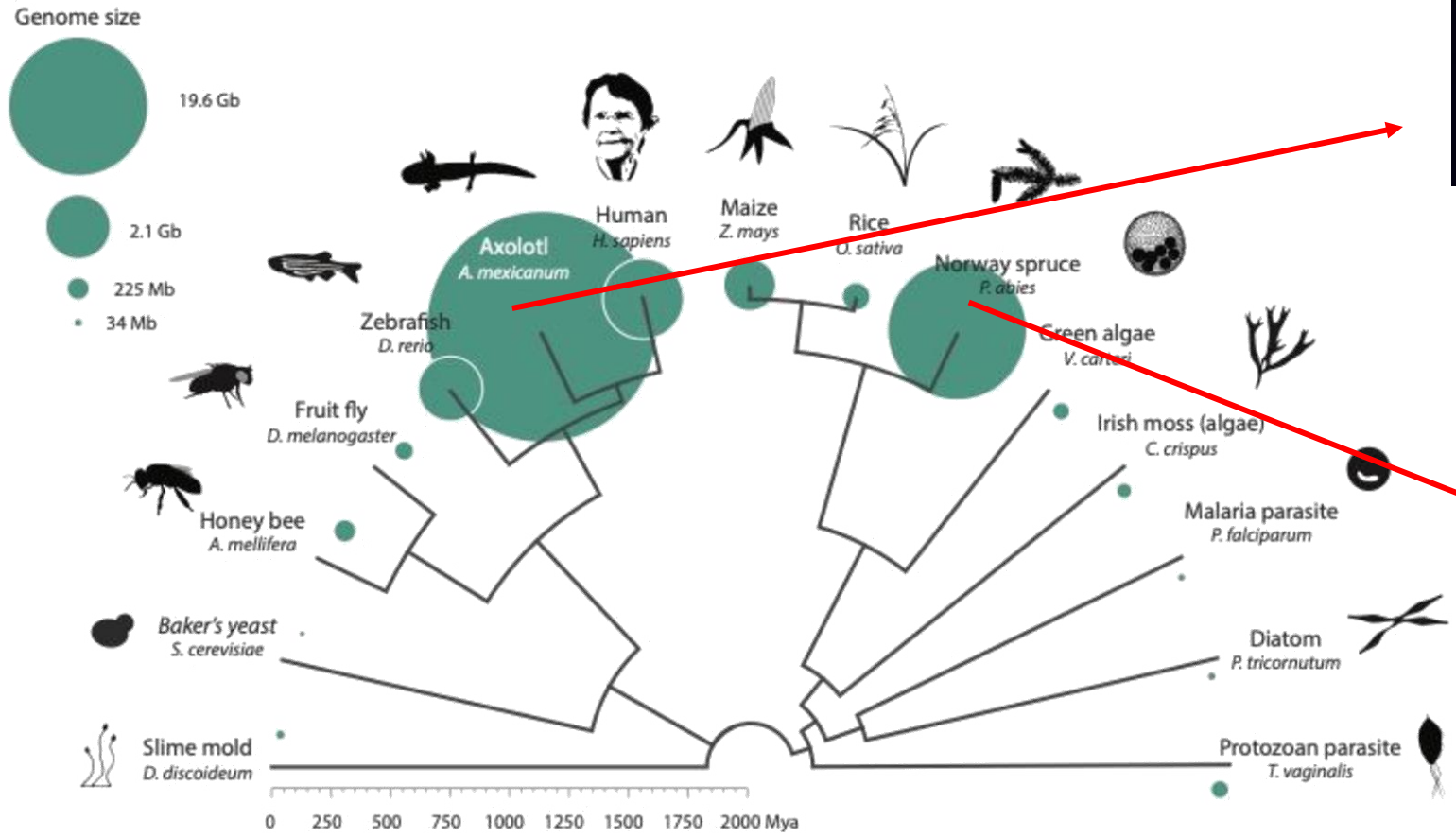


“Completed” in 2003 → ~92% of total genome covered

2021 → 0.3% of bases with potential issues
2022 → gapless assembly

~ 3.1 Gbp genome!

The Genomic Revolution: Genome Sizes



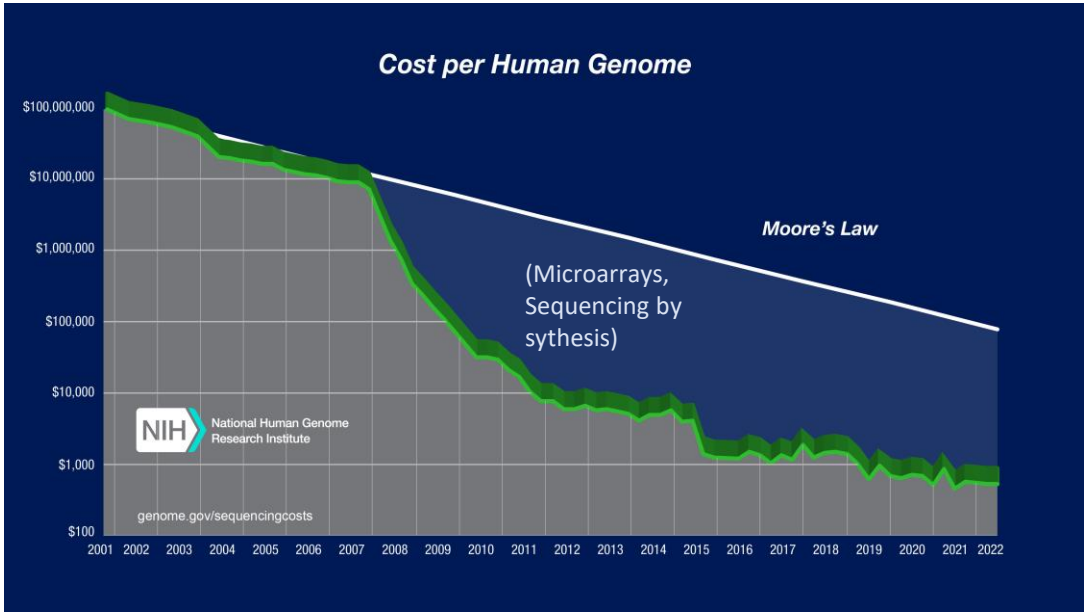
Axolotl – 32 Gbp
Largest ever sequenced!
About the same number of genes as humans

***Picea abies* – 19.6 Gbp**
About the same number of genes as humans

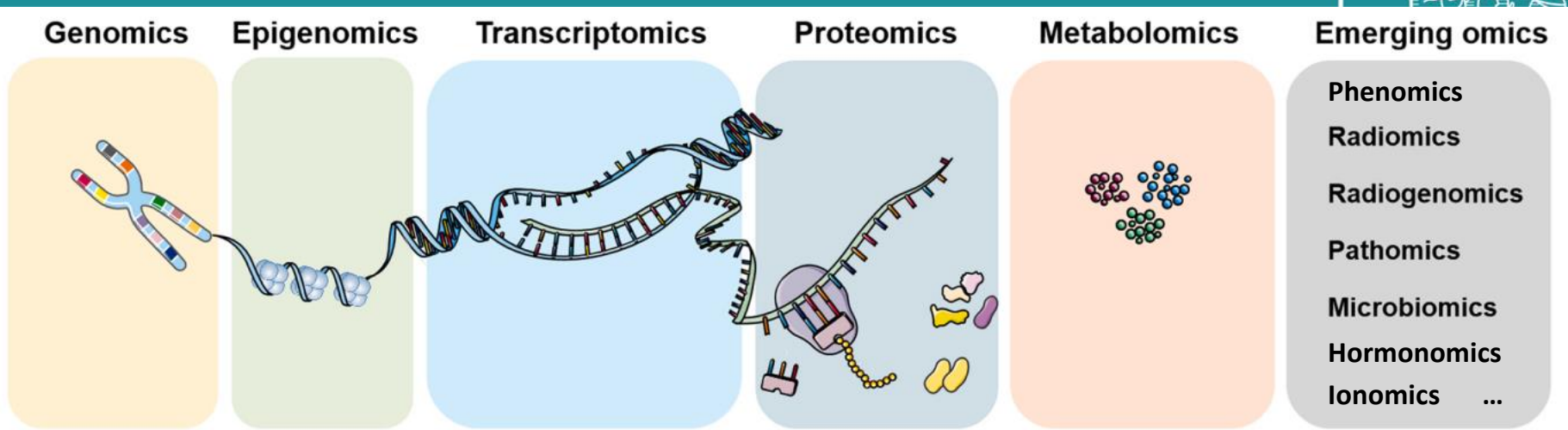
The Genomic Revolution: Effects

Key effects of the Genomic Revolution include:

- ❖ dramatic cost reductions for sequencing
- ❖ advancements in personalized medicine, disease diagnosis and treatment
- ❖ agricultural improvements (e.g. higher yield, stress resistance etc)



The Genomic Revolution: The «-omics» Era



The information provided by omics technologies

- point mutations
- small insertions/deletions
- genomic rearrangements
- viral-genome insertions
- structural variants
- copy-number variants

- DNA modifications
- histone modifications and variants
- nucleosome occupancy
- chromatin interactions
- chromatin domains

- gene expression
- noncoding RNAs
- alternative splicing
- alternative polyadenylation
- gene fusions
- allele-specific expression
- RNA editing
- endogenous retrotransposon transcription

- identification and quantitation of proteins
- protein modifications

- identification and quantitation of metabolites
- drug metabolism and toxicity
- cancer metabolic reprogramming
- immunometabolism

- cell composition, cell morphology, and spatial context
- quantitative features from digital images
- microenvironment information

Representative techniques

- WGS
- WES

- WGBS
- ChIP-seq
- MeRIP-Seq
- ATAC-seq
- 3C and derivatives

- Microarray
- RNA-seq

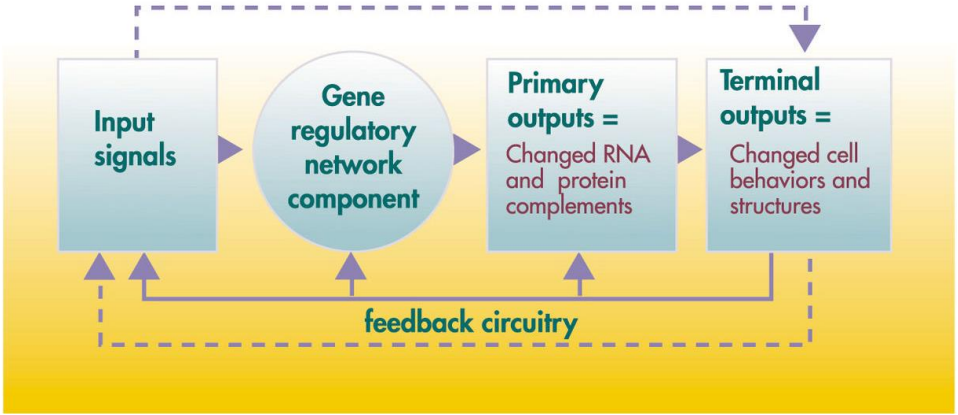
- MS-based proteomics technology
- SMPS

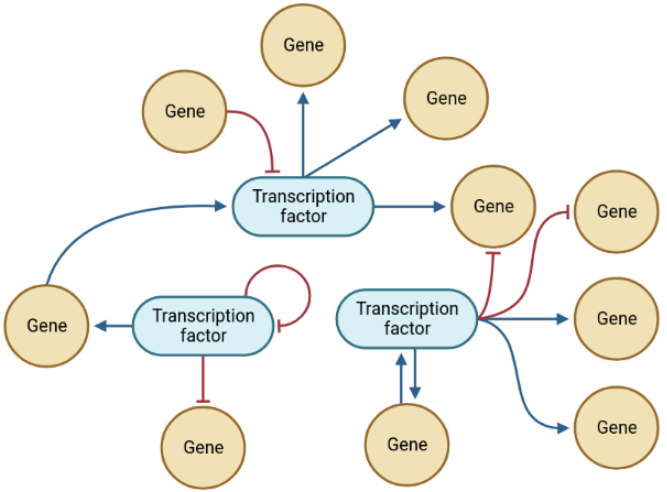
- NMR spectroscopy
- MS-based metabolomics technology

- PET/CT, MRI, Dermoscopic images, Mammograms, H&E
- WMS, 16S rRNA gene sequencing

The Genomic Revolution: Gene regulatory networks

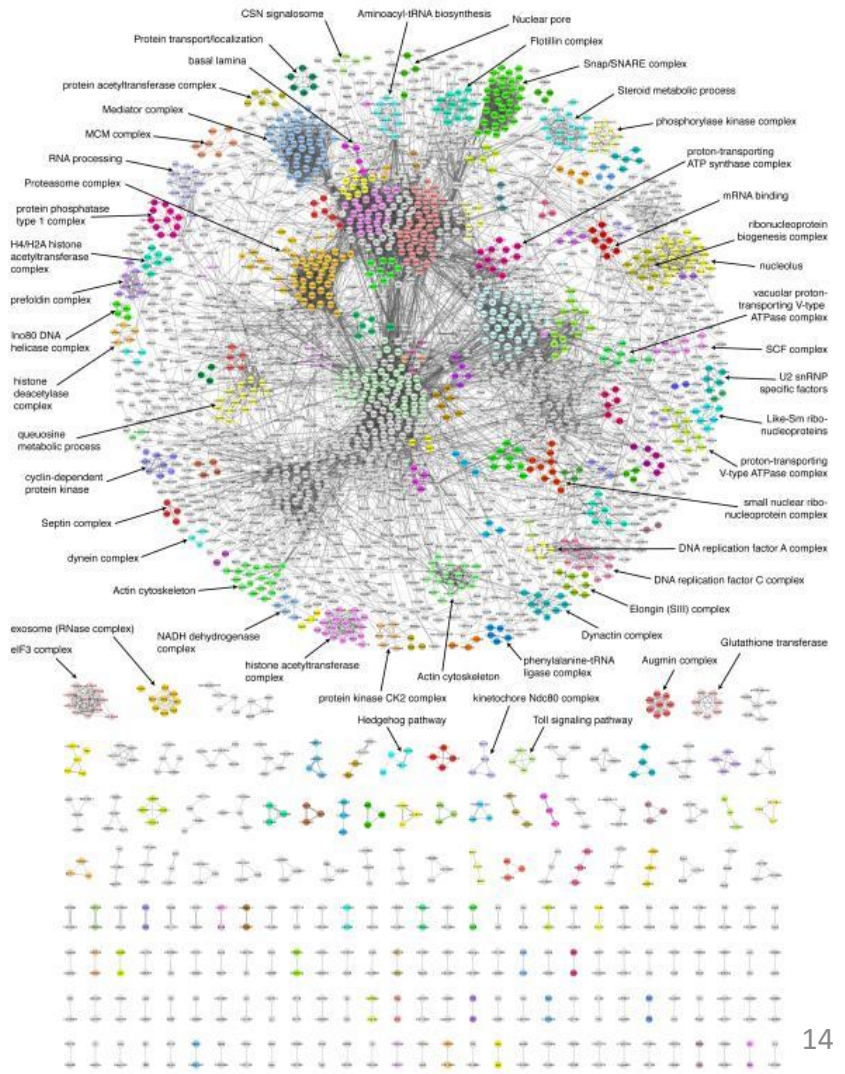
- A predictive model connecting the expression of genes with DNA variation and phenotypic variation
- Investigating all factors that influence gene expression (other genes, epigenetic factors, transcription factor, non-coding RNAs, proteins)





The Genomic Revolution: Interactomics

The whole set of molecular interactions in a cell, e.g. protein – protein interactions or small molecules – proteins interactions

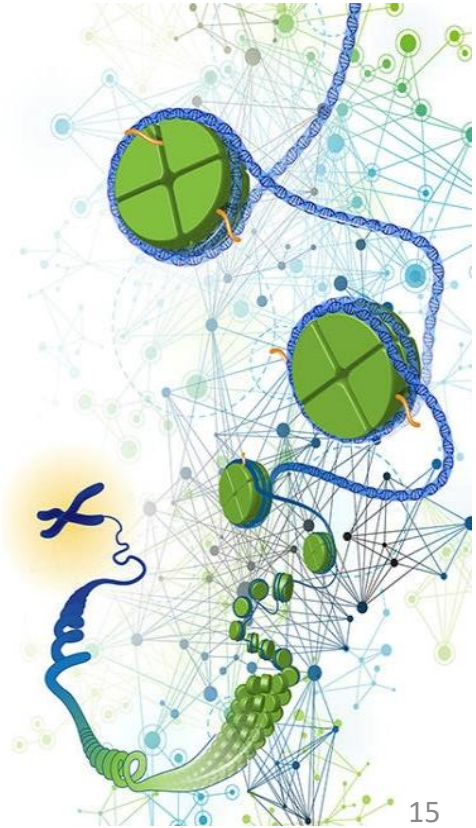


A Protein Complex Network of *Drosophila melanogaster*

K.G. Guruharsha,[1,4] Jean-François Rual,[1,4] Bo Zhai,[1,4] Julian Mintseris,[1,4] Pujita Vaidya,[1] Namita Vaidya,[1] Chapman Beekman,[1] Christina Wong,[1] David Y. Rhee,[1] Odise Cenaj,[1] Emily McKillip,[1] Saumini Shah,[1] Mark Stapleton,[2] Kenneth H. Wan,[2] Charles Yu,[2] Bayan Parsa,[2] Joseph W. Carlson,[2] Xiao Chen,[2] Bhaveen Kapadia,[2] K. VijayRaghavan,[3] Steven P. Gygi,[1] Susan E. Celniker,[2] Robert A. Obar,[1,*] and Spyros Artavanis-Tsakonas[1,*]
[1]Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA
[2]Berkeley *Drosophila* Genome Project, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[3]National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore 560065, India
[4]These authors contributed equally to this work
*Correspondence: robert_obar@hms.harvard.edu (R.A.O.), artavanis@hms.harvard.edu (S.A.-T.)
DOI 10.1016/j.cell.2011.08.047

Part 2

- Sequencing
  - Sample Preparation
  - Basic Steps for Library Preparation
  - DNA Sequencing
    - Illumina Short Read Sequencing
    - Nanopore Long Read Sequencing
    - DNA-Seq Data Analysis
  - RNA Sequencing (Transcriptomics)
    - RNA-Seq Data Analysis (differences in comparison to DNA-Seq)



DNA Sequencing: Sample Preparation

1. Get your plant material (e.g. leaves):

- Shipped (hopefully fresh and cooled)
- Harvested by yourself

2. Store your plant material (if needed):

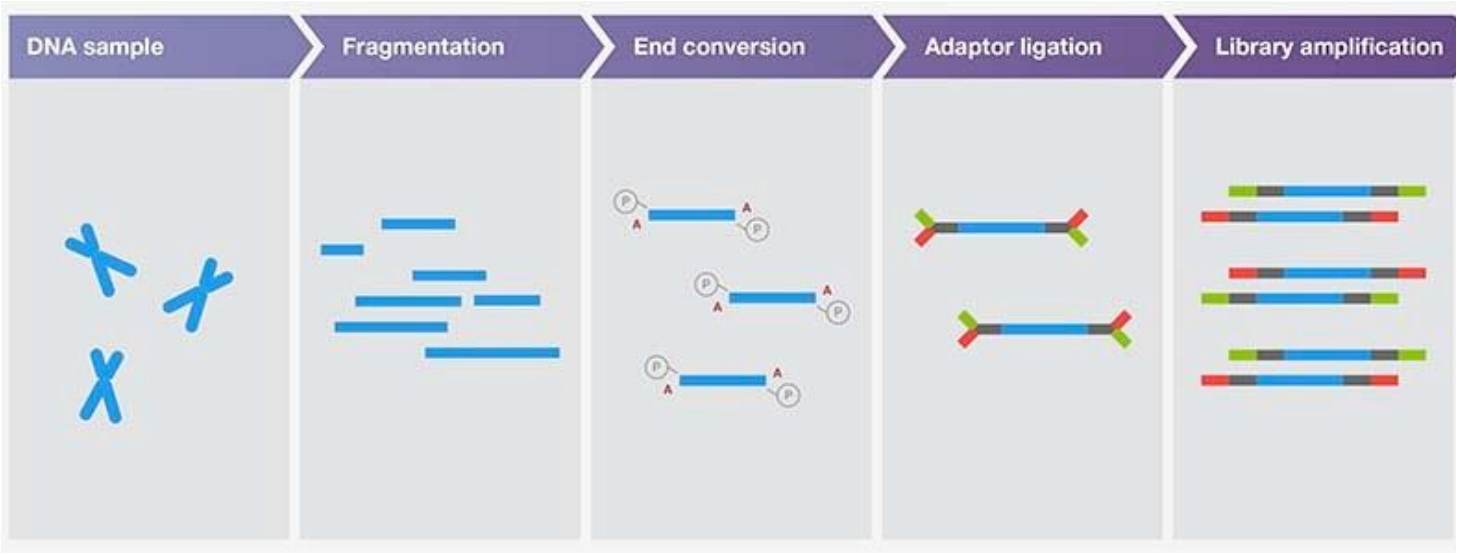
- Shock freeze in liquid nitrogen (optional) and store in -80°C (-20°C for short-term)

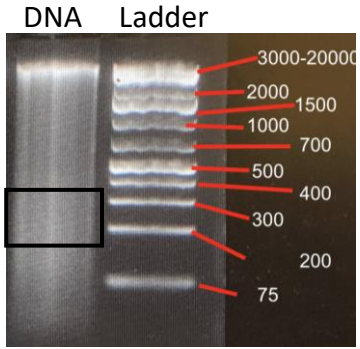
3. Extract your DNA:

- DNA extraction kits (containing ingredients and a step by step manual) or customized scripts
- Before starting extraction, you have to grind your material

DNA Sequencing: Library Preparation

To sequence DNA (or RNA), so-called sequencing libraries have to be prepared (usually done by sequencing company) from the samples with the genetic material





Size Selection

Perform NGS sequencing

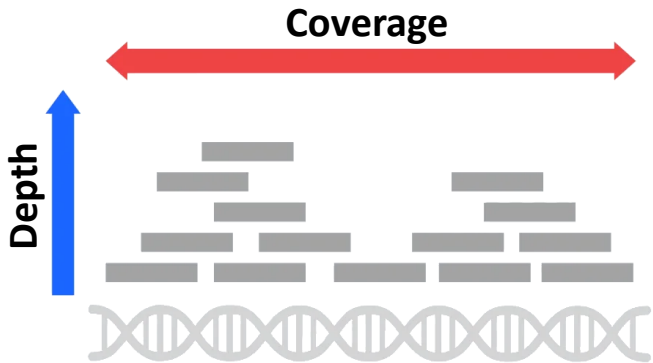
DNA Sequencing: Depth and Coverage

Sequencing Depth

Number of times a particular nucleotide is read → Influences confidence in the accuracy

Sequencing Coverage

Proportion or percentage of a genome being sequenced → Influences general data (information) availability



DNA Sequencing: Costs

Let's consider an example for whole-genome sequencing

- We have a 300Mbp genome
- A depth of ≥ 10 is recommendable.

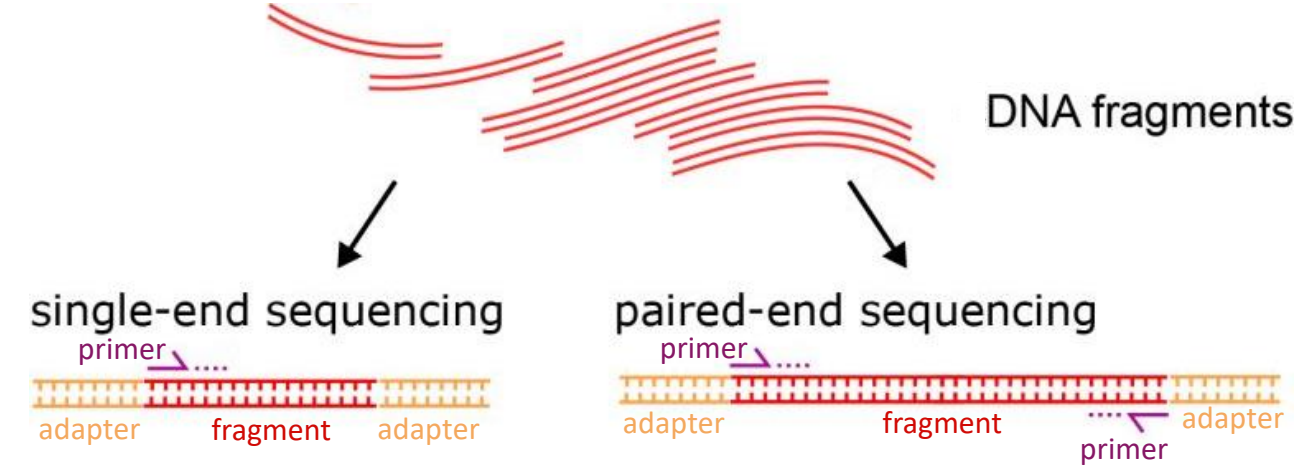
→ When having a 300 000 000 bp (300Mbp) genome and wanting each base being sequenced 10x on average we need per sample:

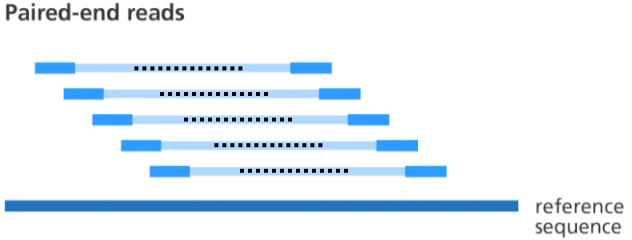
300 000 000 (genome length) x 10 (desired average depth) = 3 000 000 000 (3 Gbp)

Costs: *very* roughly between 150 and 1500$ (depends on many different factors)

BUT: Reduced-representation genome sequencing (e.g. ddRAD) helps reduce costs. And there are tremendous amounts of sequencing data publicly available online (e.g. NCBI, ENA)

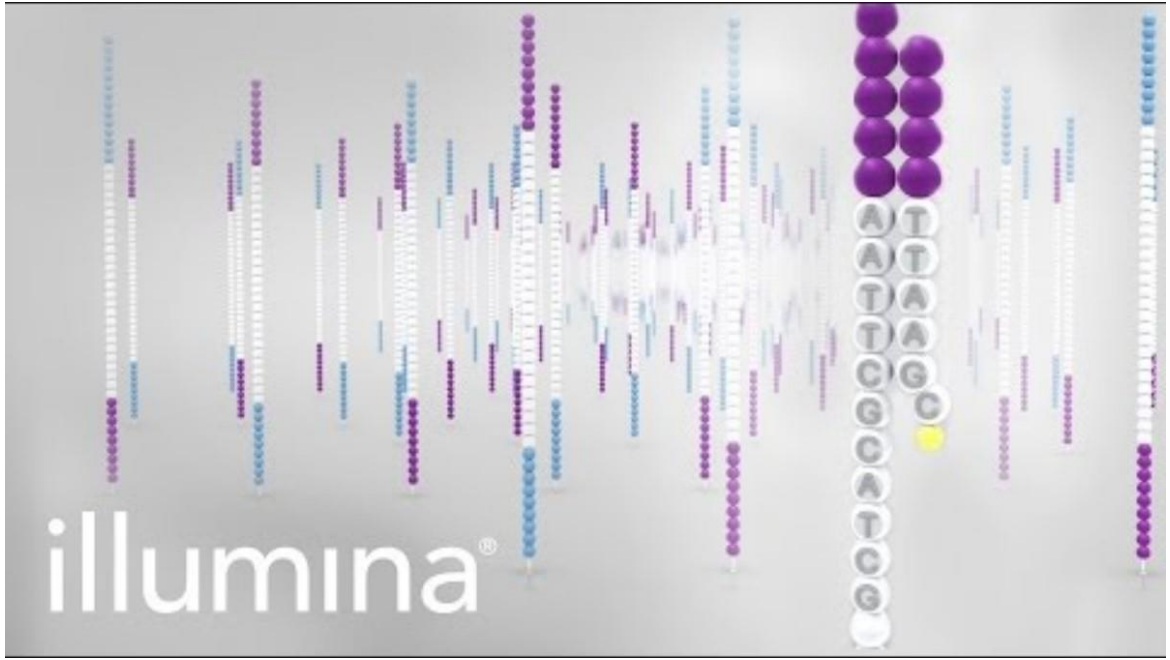
DNA Sequencing: Paired-end versus single-end libraries





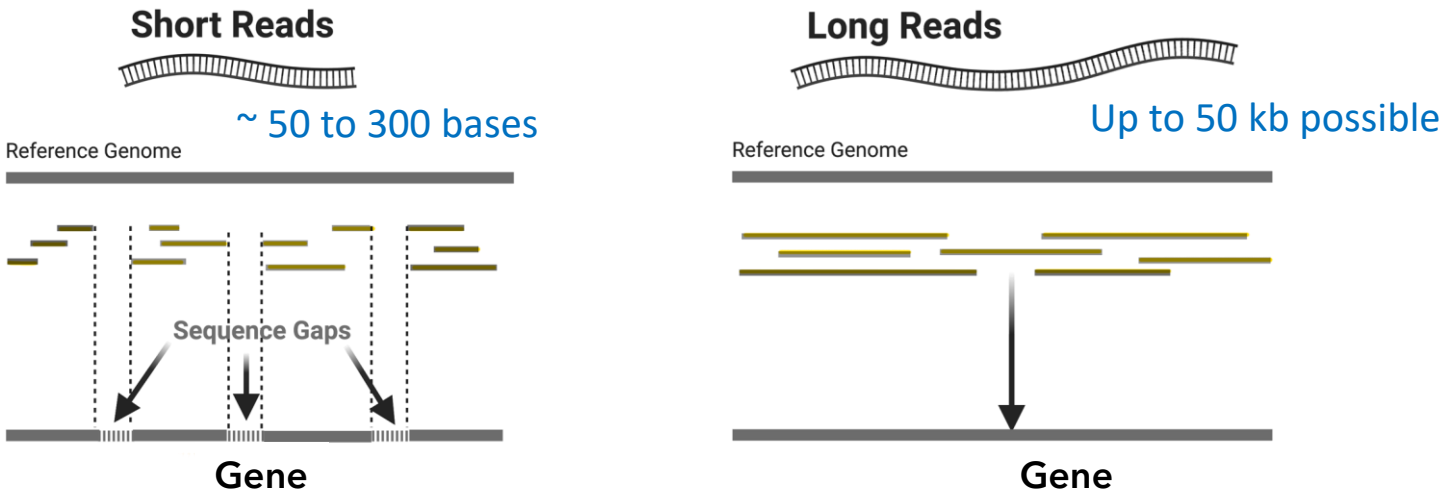
DNA Sequencing: Illumina sequencing by synthesis

For a long time the absolute gold-standard in sequencing



DNA Sequencing: Short- vs Long-reads

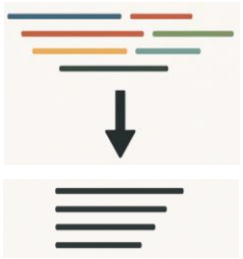
Tremendous advancement for transcript detection **via long-read sequencing**
→ One read = one transcript (full-length cDNA) has become reality



DNA Sequencing: Long-read nanopore

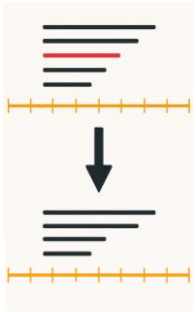
DNA Seq Data Analysis: General Steps

Raw Sequencing Reads (FASTQ)

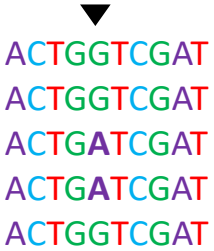


Quality Control & Filtering (Fastqc, Trimmomatic, BBDuk

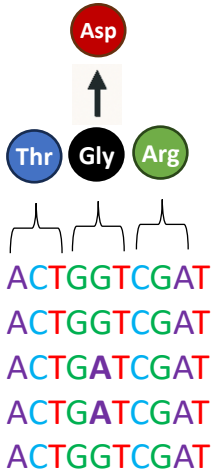
Alignment to reference (BWA, Bowtie → BAM)



Quality Control & Filtering (Qualimap, Picard, samtools)



Variant Calling (GATK, Stacks, Freebayes, bcftools → VCF)



Variant Filtering & Annotation (SNPeff, vcftools)

DNA Seq Data Analysis: Raw sequencing reads file format

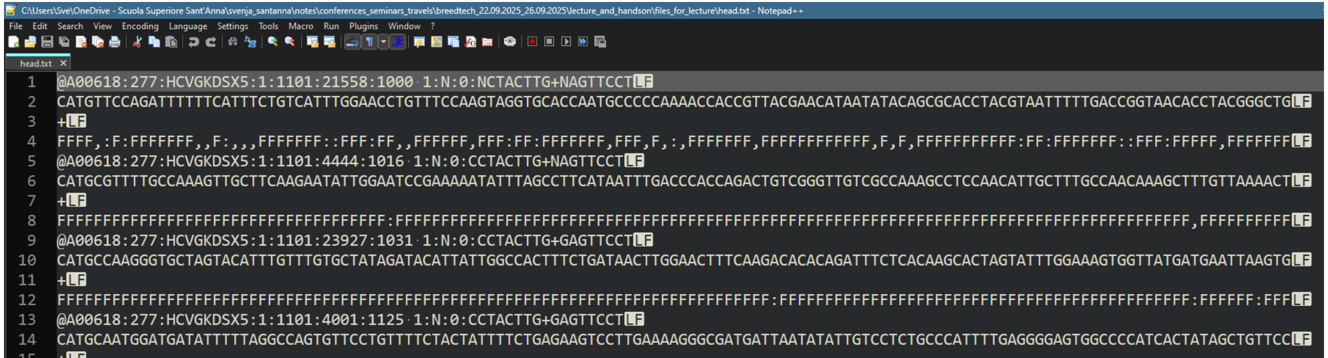
Sequencing services provide data in fastq format:

Header ←
Read sequence ←
Separator/header info ←
Base quality ←

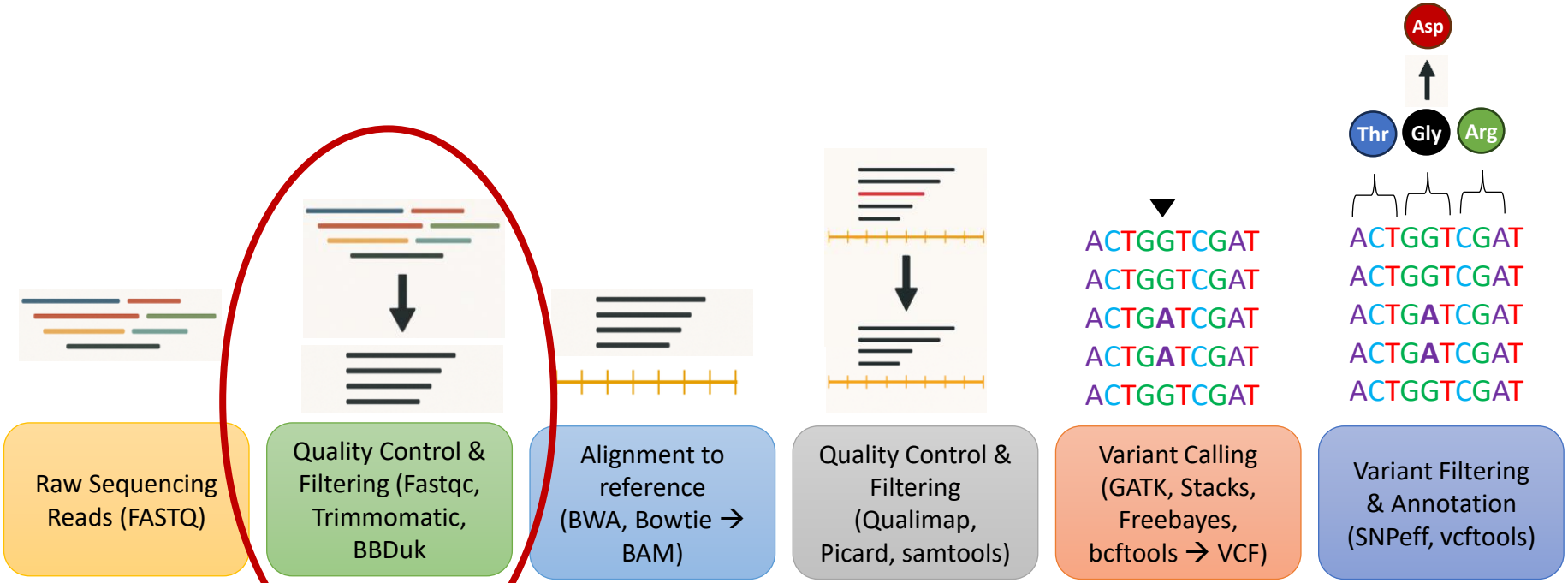
```
@M04743:199:000000000-CGG4F:1:1101:16145:1655 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGAAGGTGGCAAGCGTTGTTCGGATTCACTGGGCGTACAGGGAGCGTA
+
ABCCCFFFCADBGGGGGGGGGGHHGHGGFHGHHHGHGGGAFFHGGGGGHHHHHHHGGGGGHHGGGGGGG
@M04743:199:000000000-CGG4F:1:1101:18938:1729 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGCGCA
+
BBBBBFFFBBBBGGGGGGGGGFHHHHHGGHGGGGGGGGGGHHGGEGFHHHHHHHGGGGHFHGGGGGGG
@M04743:199:000000000-CGG4F:1:1101:13893:1760 1:N:0:233
GGTGCCAGCAGCCGCGGTACTACGTAGGGTGCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGTTCGTA
+
BBBBBFFFB4CCGGGGGGGGCFFHGHHHGGHGGGGGGGGGAFGHGG?EFHFEHHHHHGGGGFHFHFGHGGH
```

Header info:

- Machine used
- Flow cell id
- Lane
- Coordinates
- Read direction (forward/reverse)
- Other optional info



DNA Seq Data Analysis: General Steps



DNA Seq Data Analysis: Raw Reads QC

Fastqc Report Examples

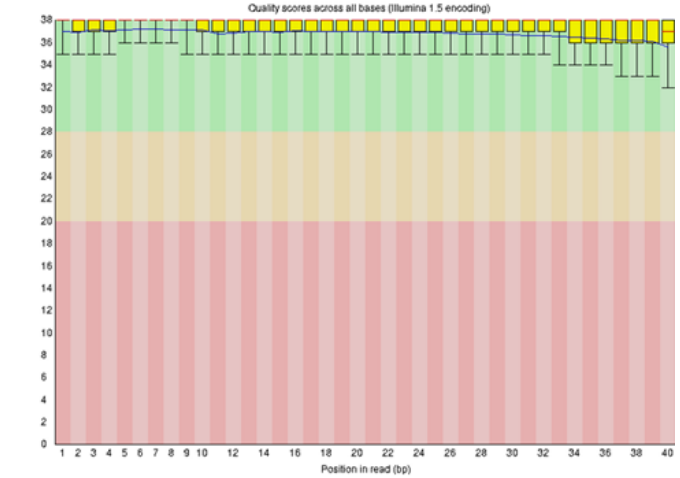
Based on the report, reads can be filtered, e.g. adapter sequences, short reads, low quality reads

Basic Statistics

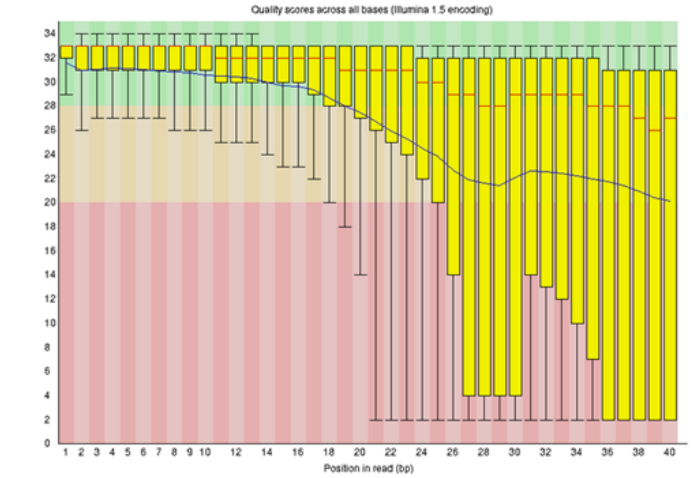
Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Total Bases	10 Mbp
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

Some basic info, like total read number and length of reads

Per base sequence quality



Per base sequence quality



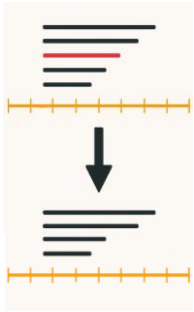
Box plots showing aggregated quality score (Phred score) statistics at each position along all reads in the file

DNA Seq Data Analysis: General Steps

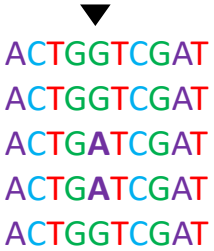
Raw Sequencing Reads (FASTQ)

Quality Control & Filtering (Fastqc, Trimmomatic, BBDuk

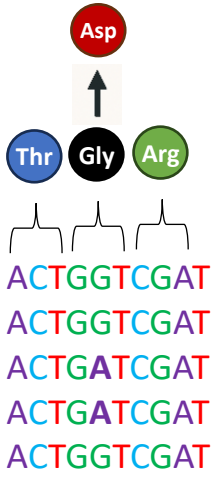
Alignment to reference (BWA, Bowtie → BAM)



Quality Control & Filtering (Qualimap, Picard, samtools)

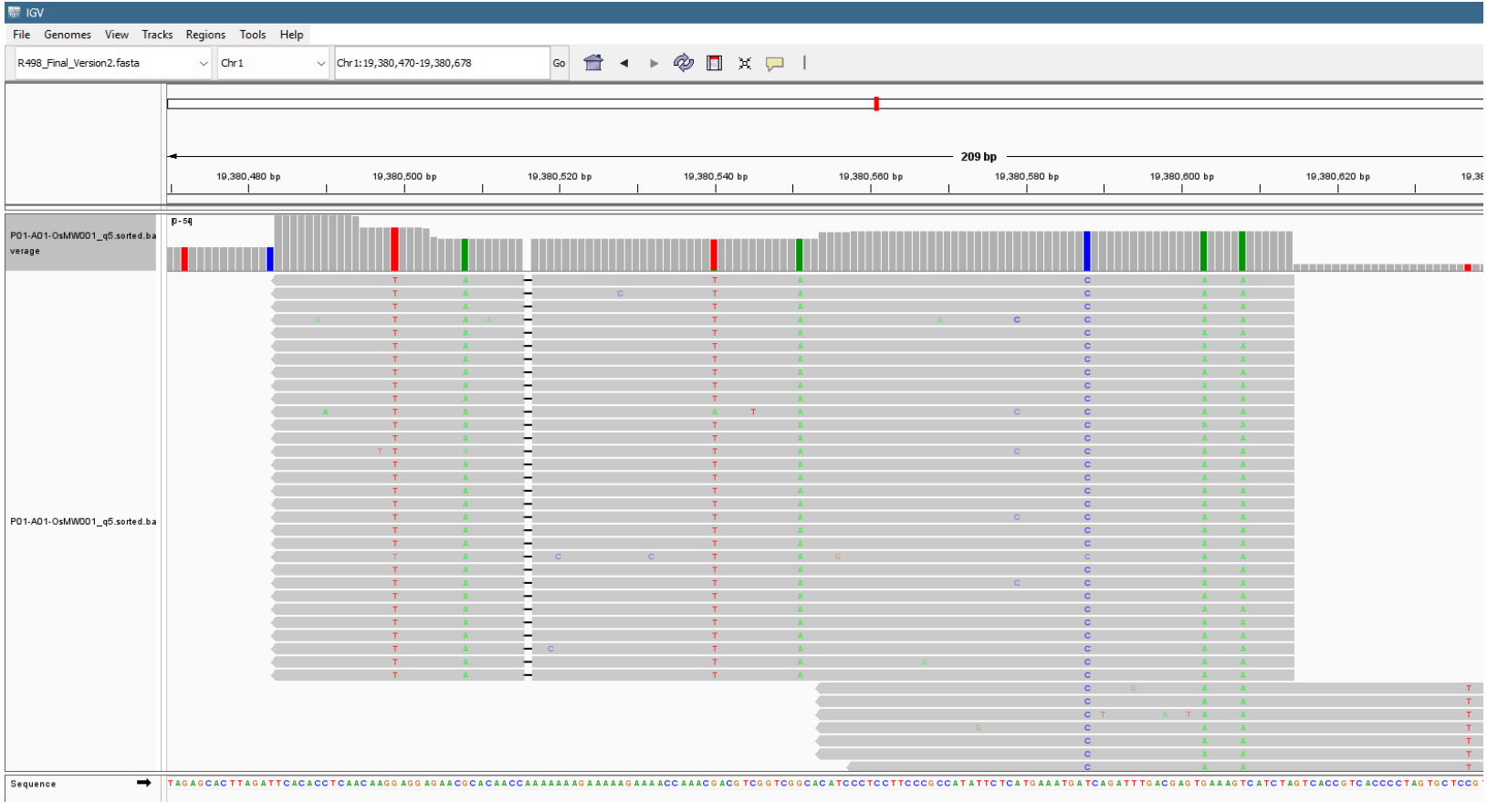


Variant Calling (GATK, Stacks, Freebayes, bcftools → VCF)



Variant Filtering & Annotation (SNPeff, vcftools)

DNA Seq Data Analysis: Aligned Reads Visualization

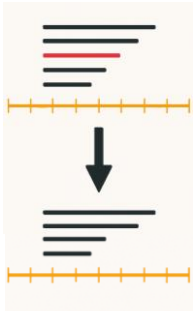


DNA Seq Data Analysis: General Steps

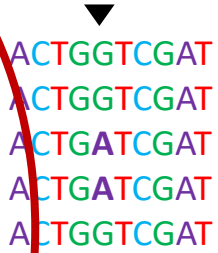
Raw Sequencing Reads (FASTQ)

Quality Control & Filtering (Fastqc, Trimmomatic, BBDuk

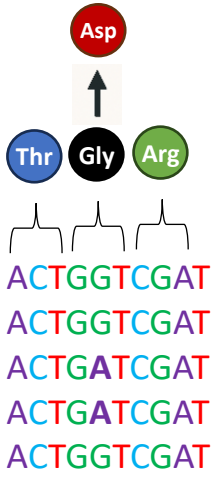
Alignment to reference (BWA, Bowtie → BAM)



Quality Control & Filtering (Qualimap, Picard, samtools)



Variant Calling (GATK, Stacks, Freebayes, bcftools → VCF)



Variant Filtering & Annotation (SNPeff, vcftools)

DNA Seq Data Analysis: Bam file QC

Qualimap Analysis Results

BAM QC analysis
Generated by Qualimap v.2.2.2-dev
2023/10/25 14:18:08

1. Input data & parameters

1.1. QualiMap command line

```
qualimap bamqc -bam P01-A01-OsMW001.sorted.bam -nw 400 -hm 3
```

1.2. Alignment

Command line:	bwa-mem2 mem -t 15 -R @RG\tID:P01-A01-OsMW001\tSM:P01-A01-OsMW001 ./bwa_ref_index_indica/R498_Final_Version2.fasta ./reads/P01-A01-OsMW001_R1.fastq ./reads/P01-A01-OsMW001_R2.fastq

2. Summary

2.1. Globals

Reference size	390,983,850
Number of reads	2,408,138
Mapped reads	2,399,894 / 99.66%
Unmapped reads	8,244 / 0.34%
Mapped paired reads	2,399,894 / 99.66%
Mapped reads, first in pair	1,200,557 / 49.85%
Mapped reads, second in pair	1,199,337 / 49.8%
Mapped reads, both in pair	2,394,864 / 99.45%

2.3. Coverage

Mean	0.8049
Standard Deviation	99.3143

2.4. Mapping Quality

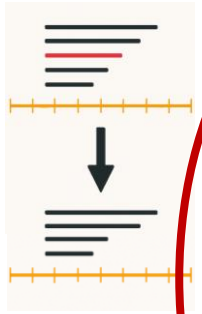
Mean Mapping Quality	49.36

DNA Seq Data Analysis: General Steps

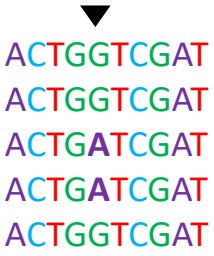
Raw Sequencing Reads (FASTQ)

Quality Control & Filtering (Fastqc, Trimmomatic, BBDuk

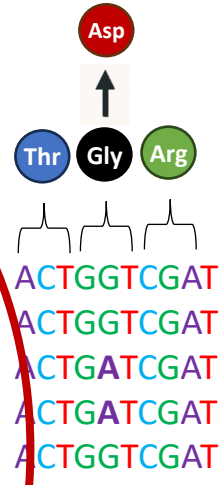
Alignment to reference (BWA, Bowtie → BAM)



Quality Control & Filtering (Qualimap, Picard, samtools)



Variant Calling (GATK, Stacks, Freebayes, bcftools → VCF)

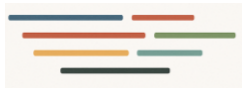


Variant Filtering & Annotation (SNPeff, vcftools)

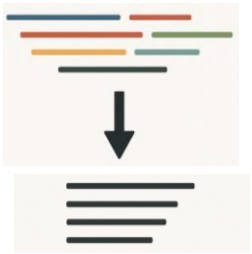
DNA Seq Data Analysis: VCF file for Variants

```
##fileformat=VCFv4.0
##FILTER=<ID=PASS,Description="All filters passed">
##Tassel=<ID=GenotypeTable,Version=5,Description="Reference allele is not known. The major allele was used as reference allele">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the reference and alternate alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth (only filtered reads used for calling)">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=PL,Number=.,Type=Float,Description="Normalized, Phred-scaled likelihoods for AA,AB,BB genotypes where A=ref and B=alt; not applicable
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##contig=<ID=1>
##contig=<ID=2>
##contig=<ID=3>
##contig=<ID=4>
##contig=<ID=5>
##contig=<ID=6>
##contig=<ID=7>
##contig=<ID=8>
##contig=<ID=9>
##contig=<ID=10>
##contig=<ID=11>
##bcftools_viewVersion=1.20+htslib-1.20
##bcftools_viewCommand=view -r 9 -o population.snps.hz.tombul.filtered.leaf_chr9.vcf population.snps.hz.tombul.filtered.leaf.vcf.gz; Date=Wed Sep 2
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  100-HZ102-P2b-D01.1     173-HZ176-P2b-E10.1     244-HZ249-P3b-D07.1     101-HZ103_39-HZ103.1    174-HZ1
9       14977   665713:391:-    G       A       .       PASS    DP=15405        GT:AD:DP:GQ:PL  0/0:108,0:108:100:0,255,255      0/1:8,9:17:100:255,0,237        0/0:10,0:10:99:
9       15013   665713:355:-    C       A       .       PASS    DP=15497        GT:AD:DP:GQ:PL  0/1:48,61:109:100:255,0,255      0/0:17,0:17:99:0,51,255 0/0:10,0:10:99:0,30
9       15047   665713:321:-    G       A       .       PASS    DP=15366        GT:AD:DP:GQ:PL  0/1:48,61:109:100:255,0,255      0/1:8,9:17:100:255,0,237        0/0:10,0:10:99:
9       46854   665743:134:+    A       C       .       PASS    DP=5200 GT:AD:DP:GQ:PL  0/0:48,0:48:99:0,144,255 0/0:1,0:1:66:0,3,36      0/0:8,0:8:99:0,24,255   0/0
9       179083  666050:29:+     G       A       .       PASS    DP=10109        GT:AD:DP:GQ:PL  0/0:133,0:133:100:0,255,255      0/0:4,0:4:94:0,12,144   0/0:5,0:5:96:0,15,180
9       179096  666050:42:+     G       A       .       PASS    DP=10001        GT:AD:DP:GQ:PL  0/0:133,0:133:100:0,255,255      0/0:4,0:4:94:0,12,144   0/0:5,0:5:96:0,15,180
9       179116  666050:62:+     G       A       .       PASS    DP=10148        GT:AD:DP:GQ:PL  0/0:132,0:132:100:0,255,255      0/1:3,1:4:99:24,0,96    ./.:0,0:0:.:.   0/1:5,2
9       179125  666050:71:+     C       T       .       PASS    DP=10106        GT:AD:DP:GQ:PL  0/0:133,0:133:100:0,255,255      0/0:4,0:4:94:0,12,144   0/0:5,0:5:96:0,15,180
```

DNA Seq Data Analysis: General Steps

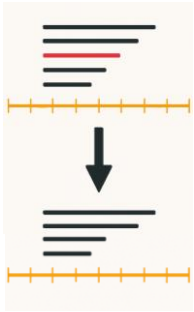


Raw Sequencing Reads (FASTQ)

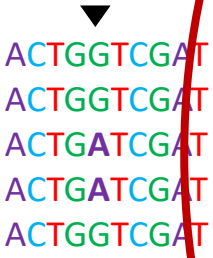


Quality Control & Filtering (Fastqc, Trimmomatic, BBDuk

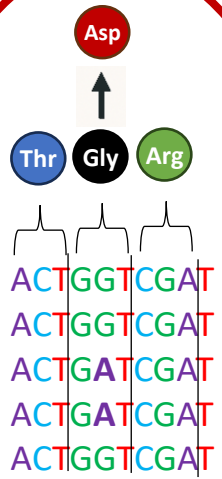
Alignment to reference (BWA, Bowtie → BAM)



Quality Control & Filtering (Qualimap, Picard, samtools)

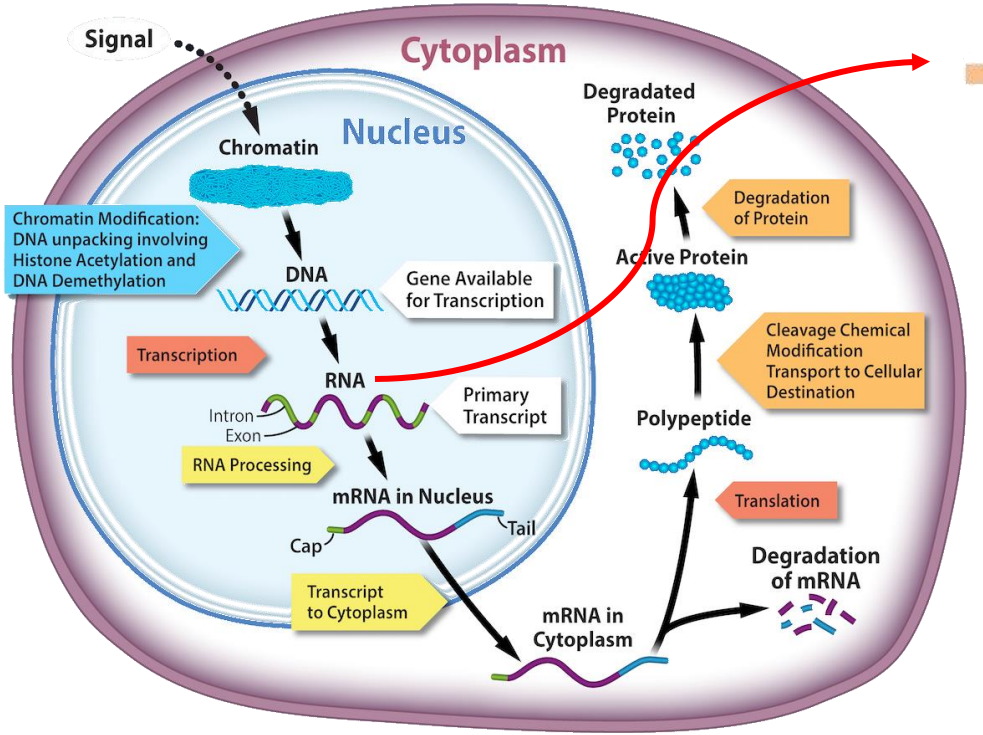


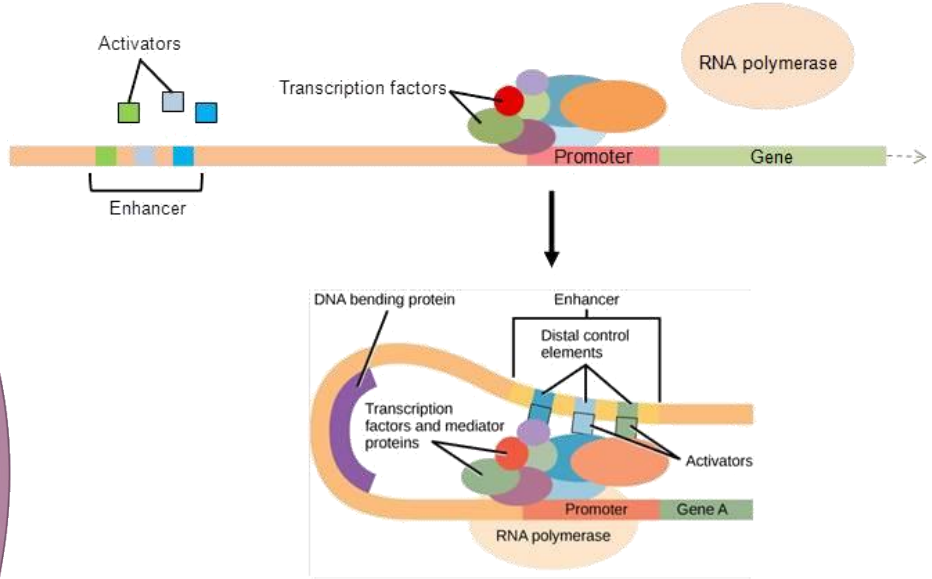
Variant Calling (GATK, Stacks, Freebayes, bcftools → VCF)



Variant Filtering & Annotation (SNPeff, vcftools)

RNA Sequencing: Transcriptomics

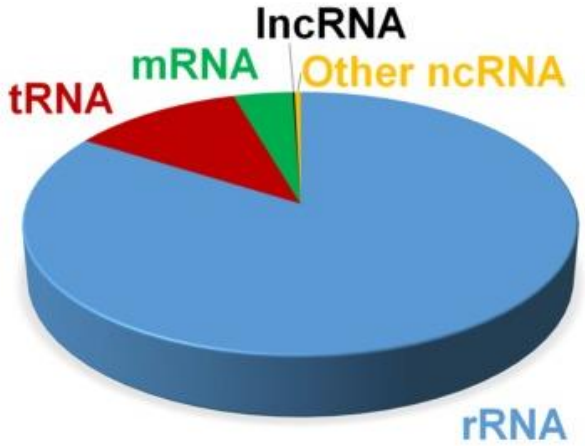




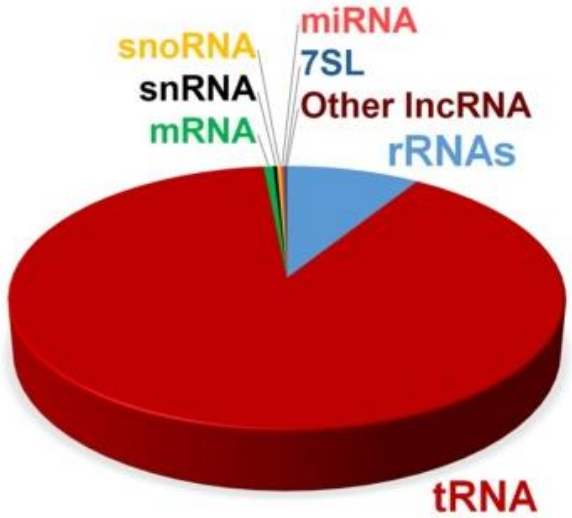
But Transcriptomics do not only look at **coding genes…**

RNA Sequencing: Coding vs Non-coding RNA

Estimate of RNA type proportions in a typical mammalian cell



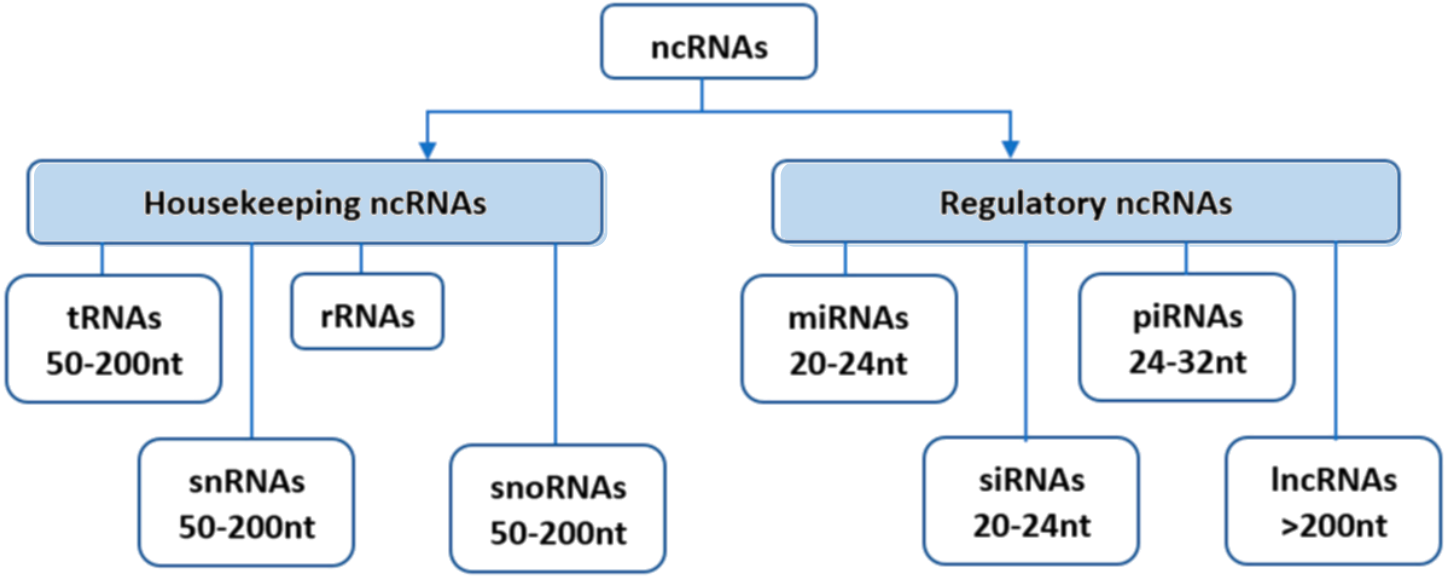
RNA by mass



RNA by number of molecules

→ Typically, only a small part of the transcribed RNA is coding RNA

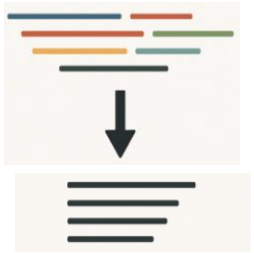
RNA Sequencing: Coding vs Non-coding RNA



RNA Seq Data Analysis: General Steps

Stranded or unstranded

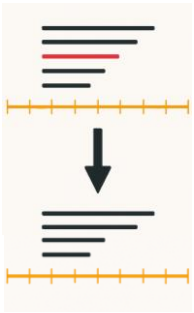
Raw Sequencing Reads (FASTQ)



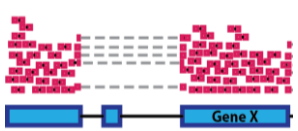
Quality Control & Filtering (Fastqc, Trimmomatic, BBDuk

Reference Genome or Transcriptome

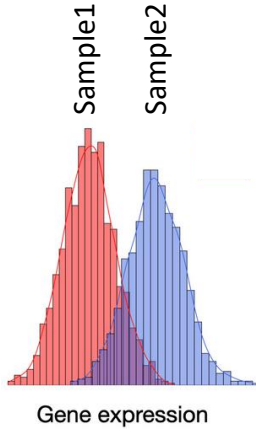
Alignment to reference (BWA, STAR, BBmap → BAM)



Quality Control & Filtering (Qualimap, Picard, samtools)



Quantification (Read Count)

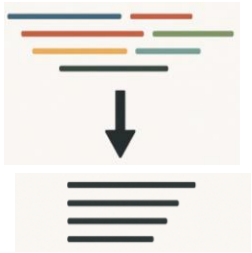


Differential Expression Analysis (EdgeR, Deseq2)

RNA Seq Data Analysis: General Steps

Stranded or unstranded

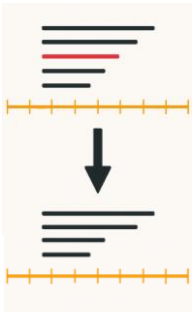
Raw Sequencing Reads (FASTQ)



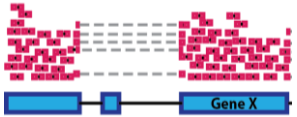
Quality Control & Filtering (Fastqc, Trimmomatic, BBDuk

Reference Genome or Transcriptome

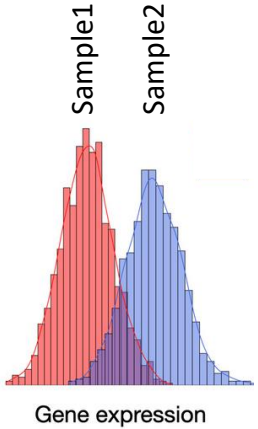
Alignment to reference (BWA, STAR, BBmap → BAM)



Quality Control & Filtering (Qualimap, Picard, samtools)

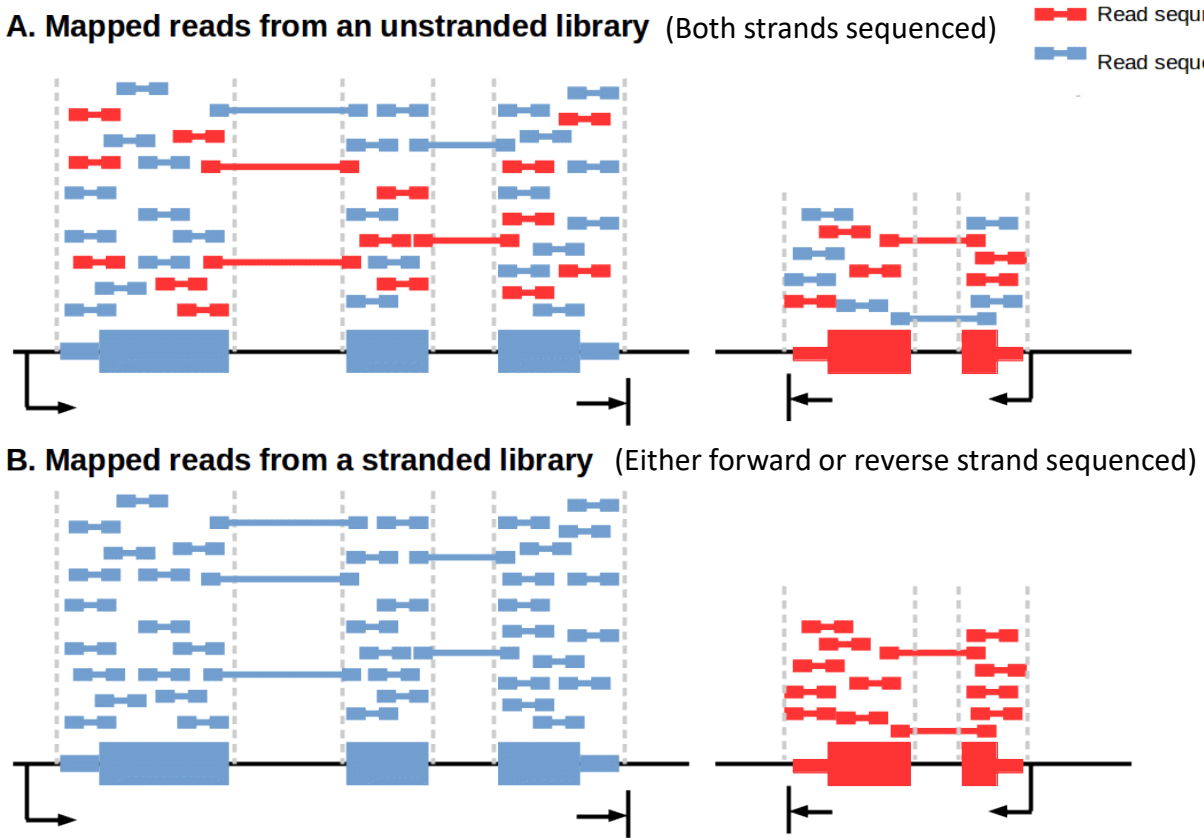


Quantification (Read Count)



Differential Expression Analysis (EdgeR, Deseq2)

RNA Seq Data Analysis: Stranded vs Unstranded



Unstranded:

- Less expensive
- Easier to execute
- Recommended for well annotated references
- Enough for most differential expression analyses

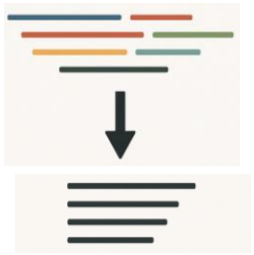
Stranded:

- More accurate
- Identify sense/antisense transcripts
- Advantageous for annotation and novel transcript discovery
- Insights into regulatory mechanisms specific to one strand
- Information about differential expression between genes on different strands

RNA Seq Data Analysis: General Steps

Stranded or unstranded

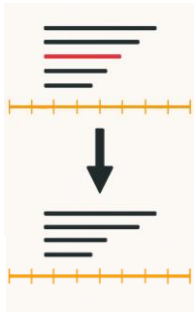
Raw Sequencing Reads (FASTQ)



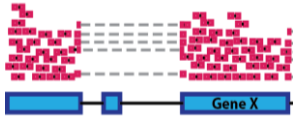
Quality Control & Filtering (Fastqc, Trimmomatic, BBDuk

Reference Genome or Transcriptome

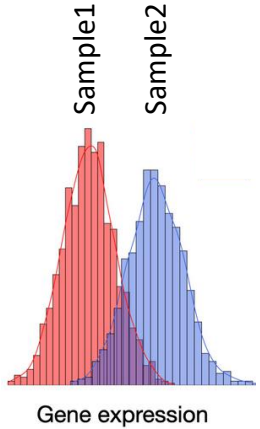
Alignment to reference (BWA, STAR, BBmap → BAM)



Quality Control & Filtering (Qualimap, Picard, samtools)

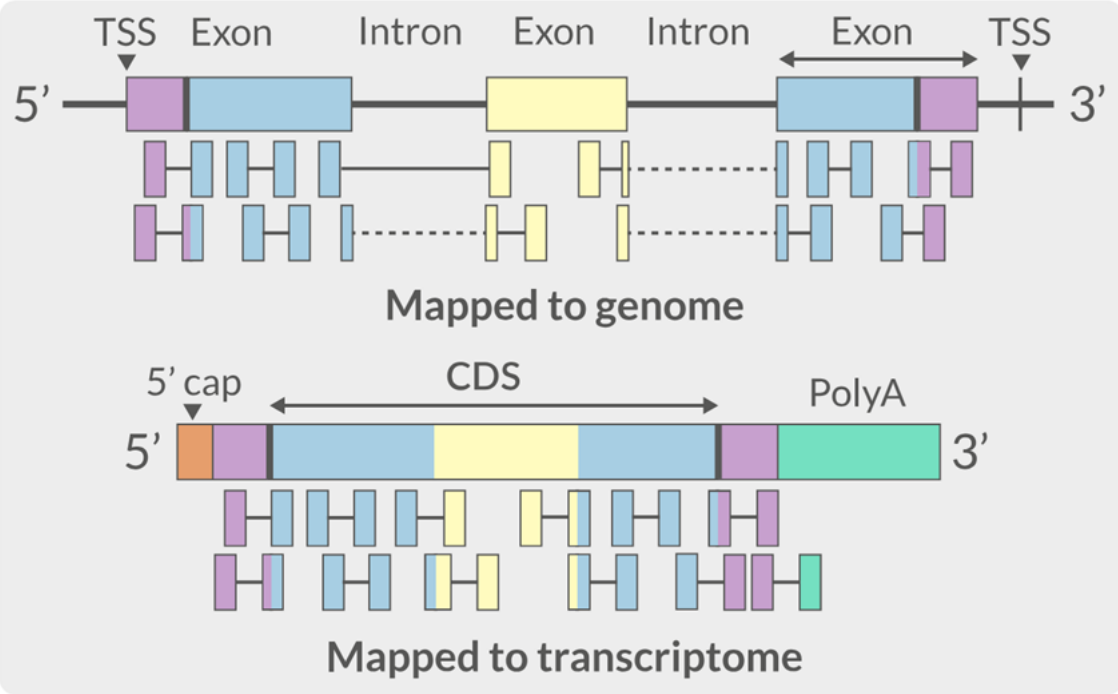


Quantification (Read Count)



Differential Expression Analysis (EdgeR, Deseq2)

RNA Seq Data Analysis: Mapping Reads to Reference

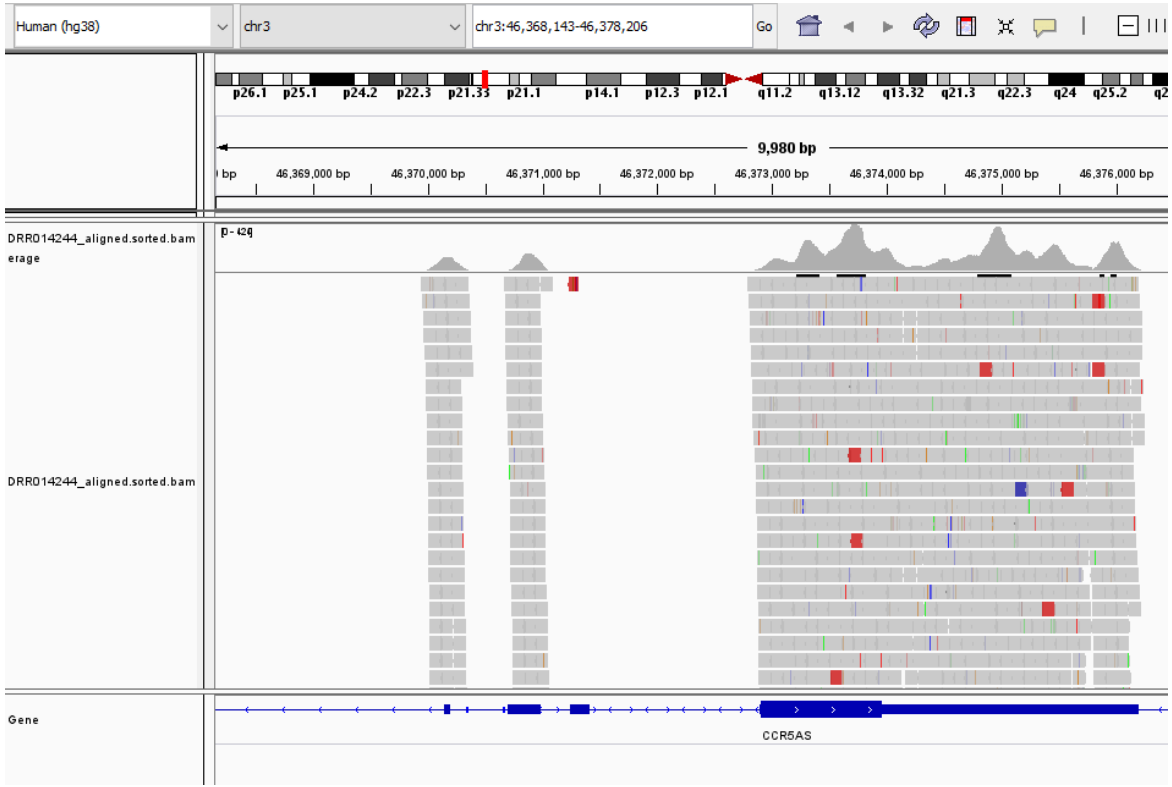


You will end up with a bam file (binary sequence alignment map)

Contains info about mapped reads (and unmapped reads), among others:

- Identifier for each read
- Reference sequence name
- Position respective to reference
- Mapping quality

RNA Seq Data Analysis: Aligned Reads Visualization

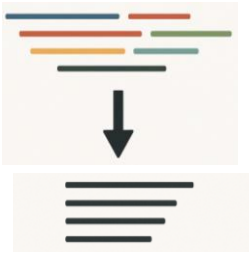


Reads can be visualized in genome browsers, e.g. IGV, showing coverage, single reads, sequencing differences, genes etc.

RNA Seq Data Analysis: General Steps

Stranded or unstranded

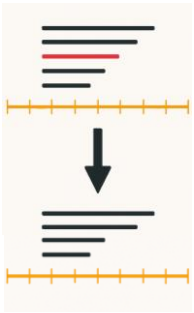
Raw Sequencing Reads (FASTQ)



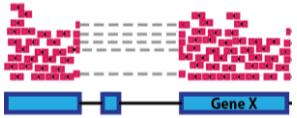
Quality Control & Filtering (Fastqc, Trimmomatic, BBDuk

Reference Genome or Transcriptome

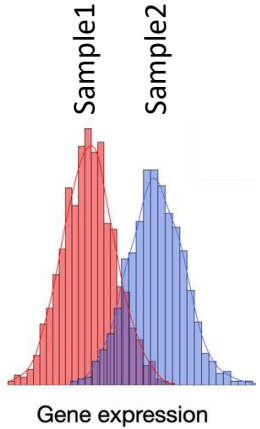
Alignment to reference (BWA, STAR, BBmap → BAM)



Quality Control & Filtering (Qualimap, Picard, samtools)



Quantification (Read Count)



Differential Expression Analysis (EdgeR, Deseq2)

RNA Seq Data Analysis: Read Counting

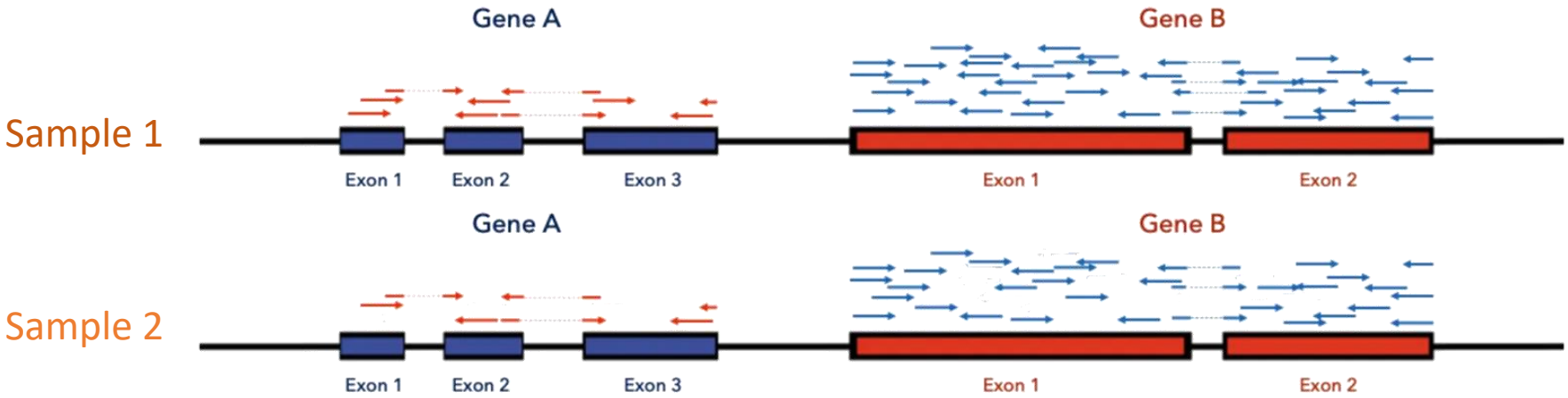
To compare gene expression

a) within one sample (compare two genes)

b) between samples (compare expression between conditions)

the reads aligning to each gene have to be counted.

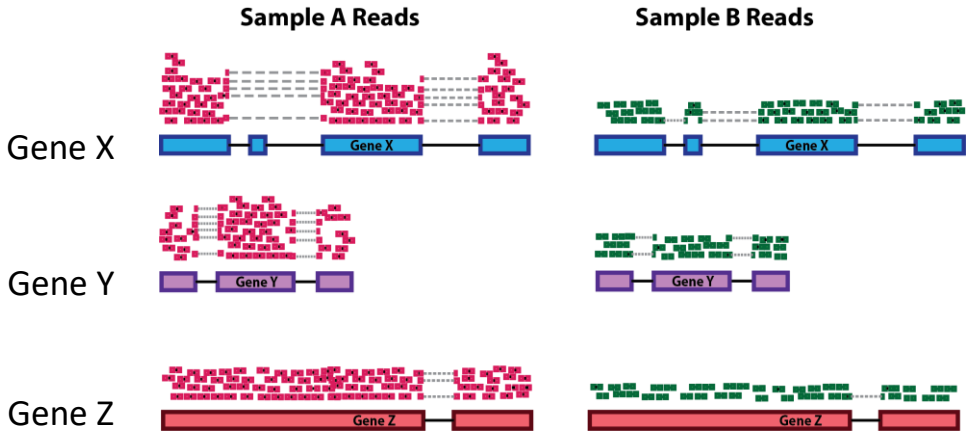
Assumption: the number of mapped reads for each gene is proportional to the expression of RNA



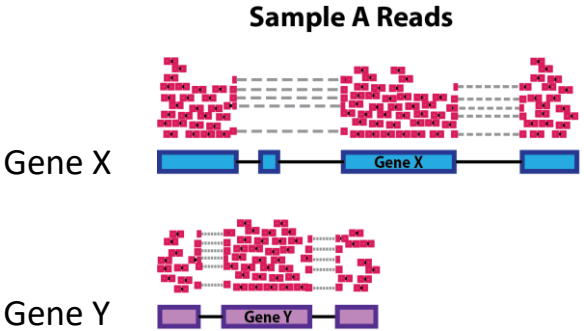
→ NOTE: counts must be **normalized** according to the question to be answered

RNA Seq Data Analysis: Read Count Normalization

If you want to compare expression of a certain gene between two samples, you must normalize for sequencing depth



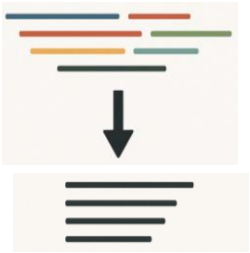
If you want to compare expression of two genes within the same sample you must normalize for gene length



RNA Seq Data Analysis: General Steps

Stranded or unstranded

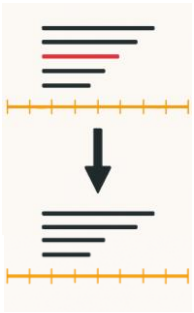
Raw Sequencing Reads (FASTQ)



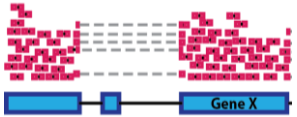
Quality Control & Filtering (Fastqc, Trimmomatic, BBDuk

Reference Genome or Transcriptome

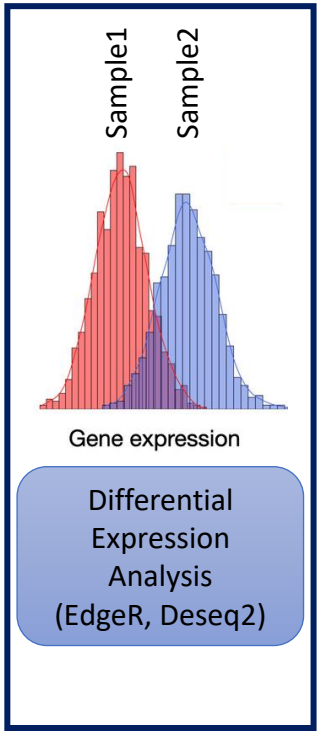
Alignment to reference (BWA, STAR, BBmap → BAM)



Quality Control & Filtering (Qualimap, Picard, samtools)

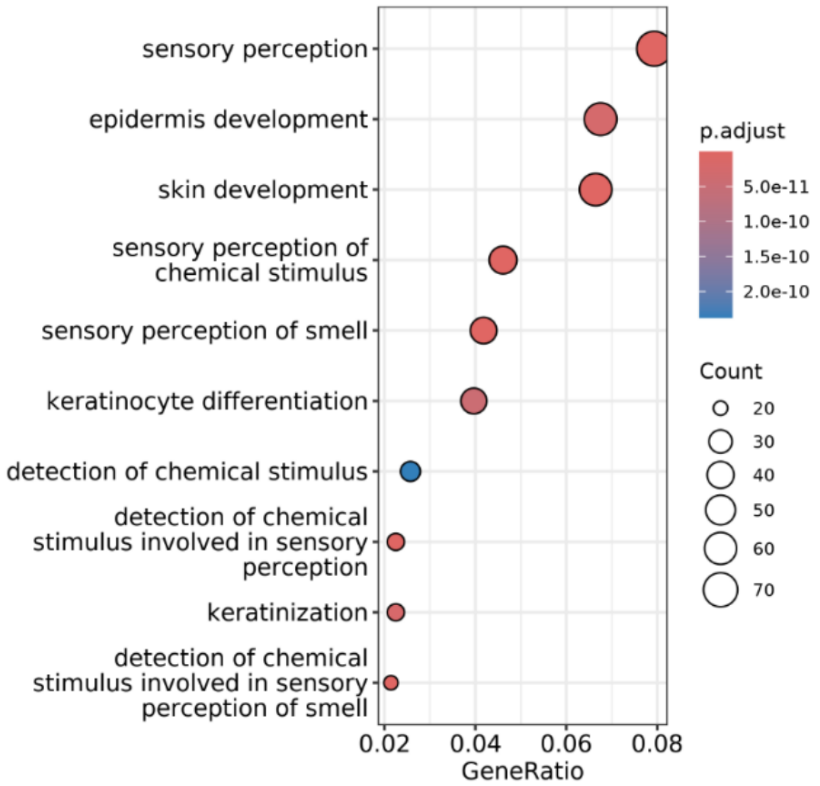


Quantification (Read Count)



Differential Expression Analysis (EdgeR, Deseq2)

RNA Seq Data Analysis: DGE and GO Analysis



Compare genes or samples in a Differential Gene Expression (DGE) Analysis:

- ❑ Compare genes within one sample, e.g. in a gene family, which genes are more expressed?
- ❑ Compare gene expression between two conditions, e.g. plants grown under normal conditions compared to heat stress conditions, or healthy cells vs diseased cells

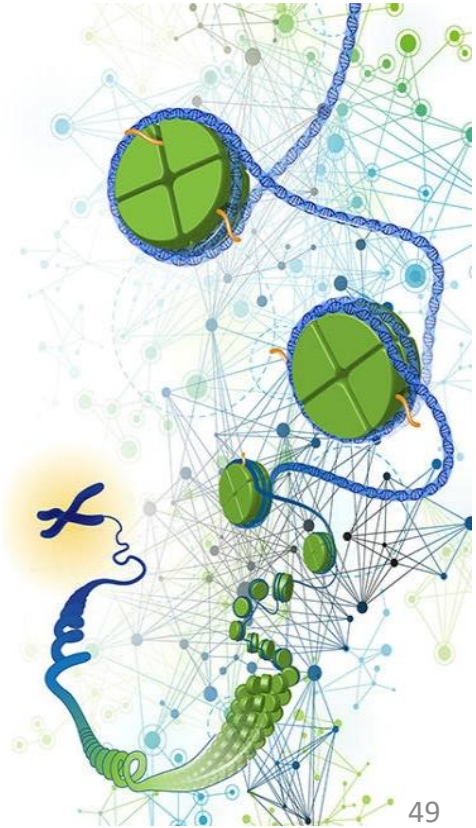
Downstream analysis when having a list of Differentially Expressed Genes (DEGs):

→ For example, make **Gene Ontology (GO) enrichment analysis** to check for overrepresented gene categories

- Can give insights into adaptability of plants or genes causing diseases
- Enables more precise breeding or finding cures for diseases

Part 3

- Presentation of a real Project as Example to show
  - … how a project using DNA sequencing can look like
  - … what information can be drawn from sequencing data
  - … which downstream analyses can be done with the variant file that we will produce during the hands-on practical part



Investigating the Genetic Diversity and the Genetic Factors influencing Nut Quality in Hazelnut (*Corylus avellana* L.)

Sant'Anna
School of Advanced Studies – Pisa

Hazelnut (*Corylus avellana* L.) is one of the most important edible nut species in the world

- Large genetic diversity with **300+ hazelnut cultivars**
- **Nuts are the primary reason for hazelnut cultivation**: food industry uses them to manufacture different **products**
- **Nut quality traits**, such as nut **morphology and flavour/aroma,** are **important** to the food industry.

Challenges:

- **Only 20 cultivars dominate cultivation**
- **Genetic factors** that influence **nut quality traits** are **unknown**
- **Lack of tools** to **help breeders choose** accessions that produce **higher quality nuts**

The reliance on few cultivars is a risk for hazelnut cultivation

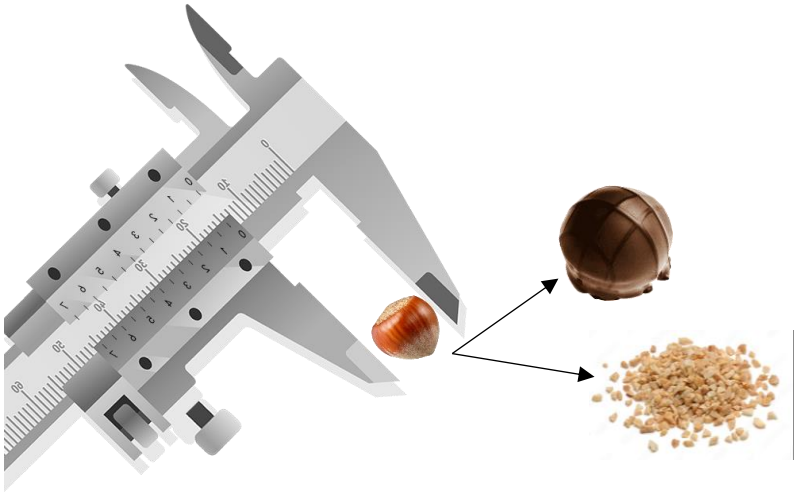
Reliance on 20 widely grown cultivars

Higher vulnerability to stress

Explore the genetic diversity to breed resilient cultivars

Nut morphology matters...

Diversity in nut morphological traits

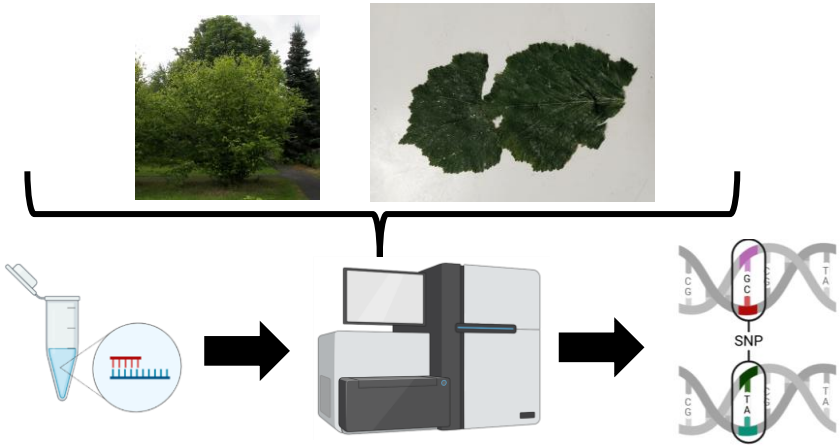


Different suitability to manufacture products

Objective: Describe the existing genetic diversity in a global collection of hazelnuts and the genetic factors influencing nut morphology

Analysis of the genetic diversity

316 leaf samples from varieties from 15 geographical provenances were genotyped

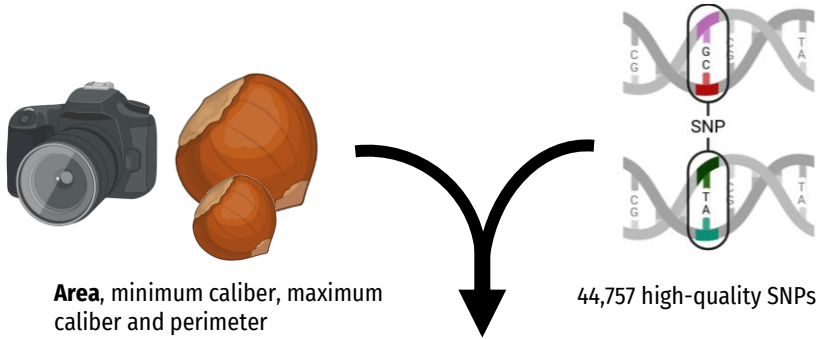


ddRAD-seq (**d**ouble **d**igest **R**estriction **A**ssociated **D**NA **seq**uencing)

141 samples and 16,378 SNPs for the genetic diversity analysis after quality filtering.

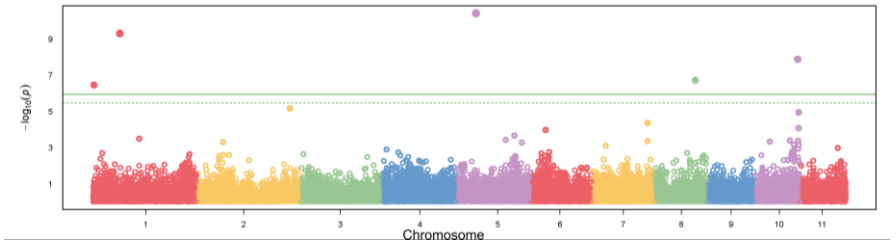
Analysis of genetic factors influencing Nut morphology

Nuts of 151 genotyped samples were phenotyped



Area, minimum caliber, maximum caliber and perimeter

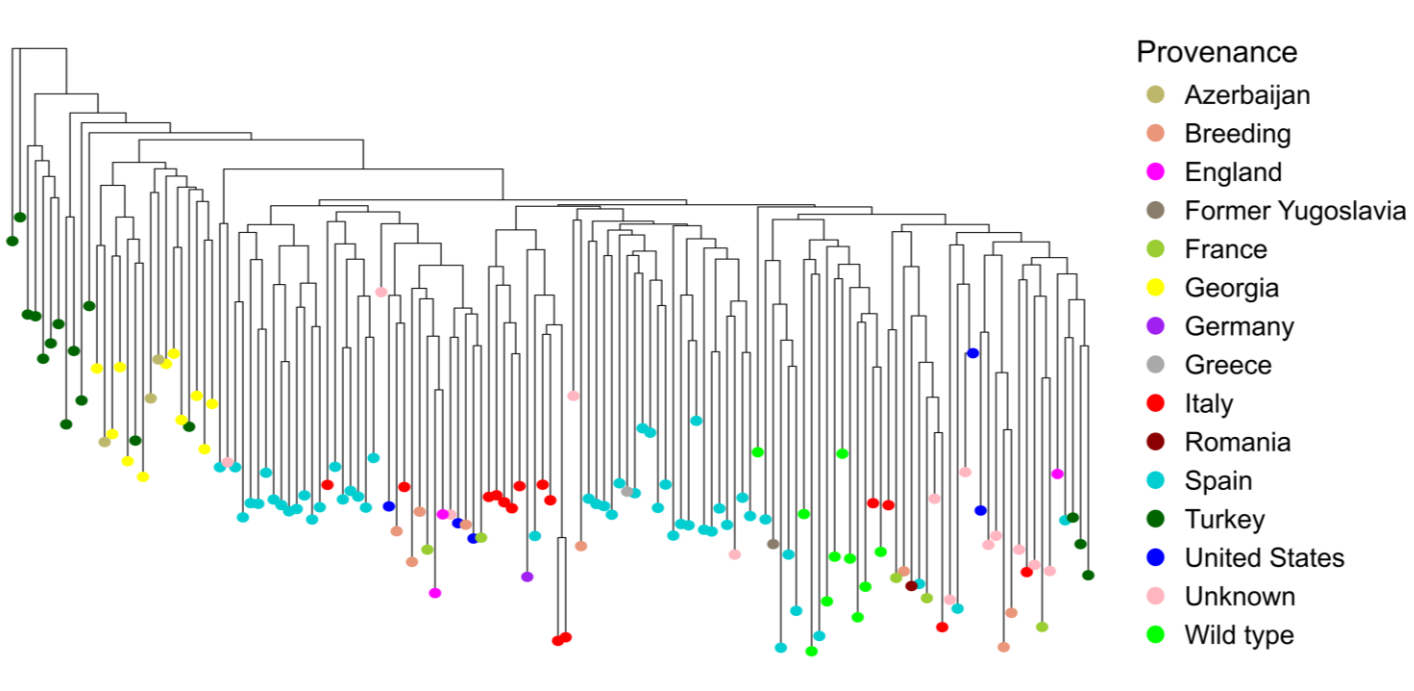
44,757 high-quality SNPs



151 samples and 44757 SNPs for GWAS to associate SNPs to Hazelnut morphology

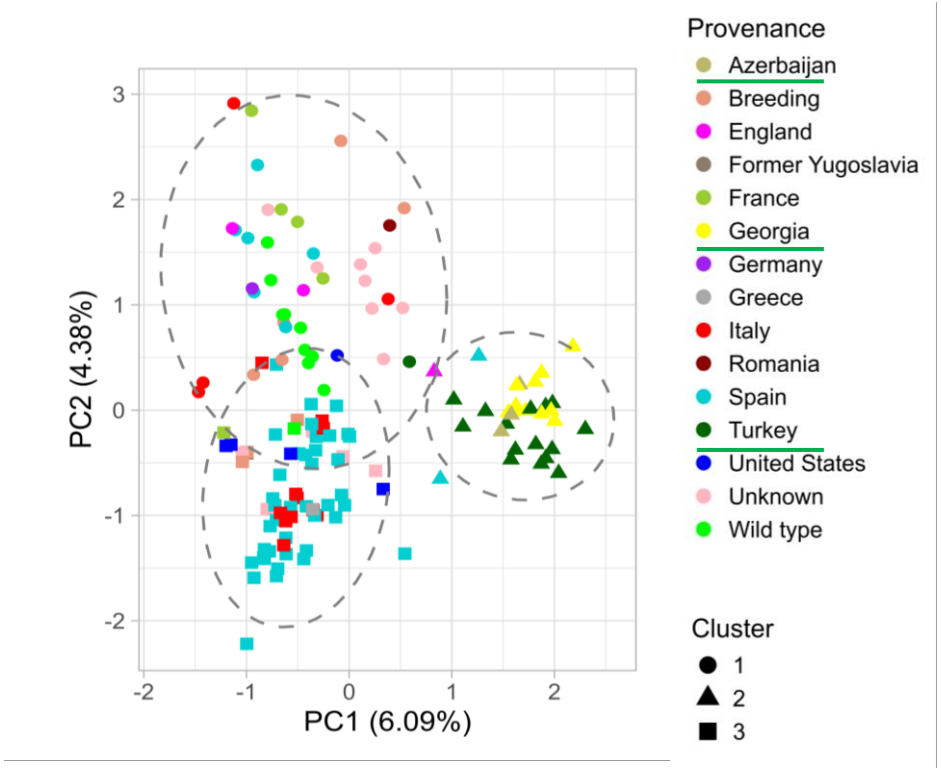
Samples were provided by Ferrero Hazelnut Company

Genetic diversity in the hazelnut collection



→ **Same provenance, higher genetic similarity**

How many genetic clusters currently exist in the collection?



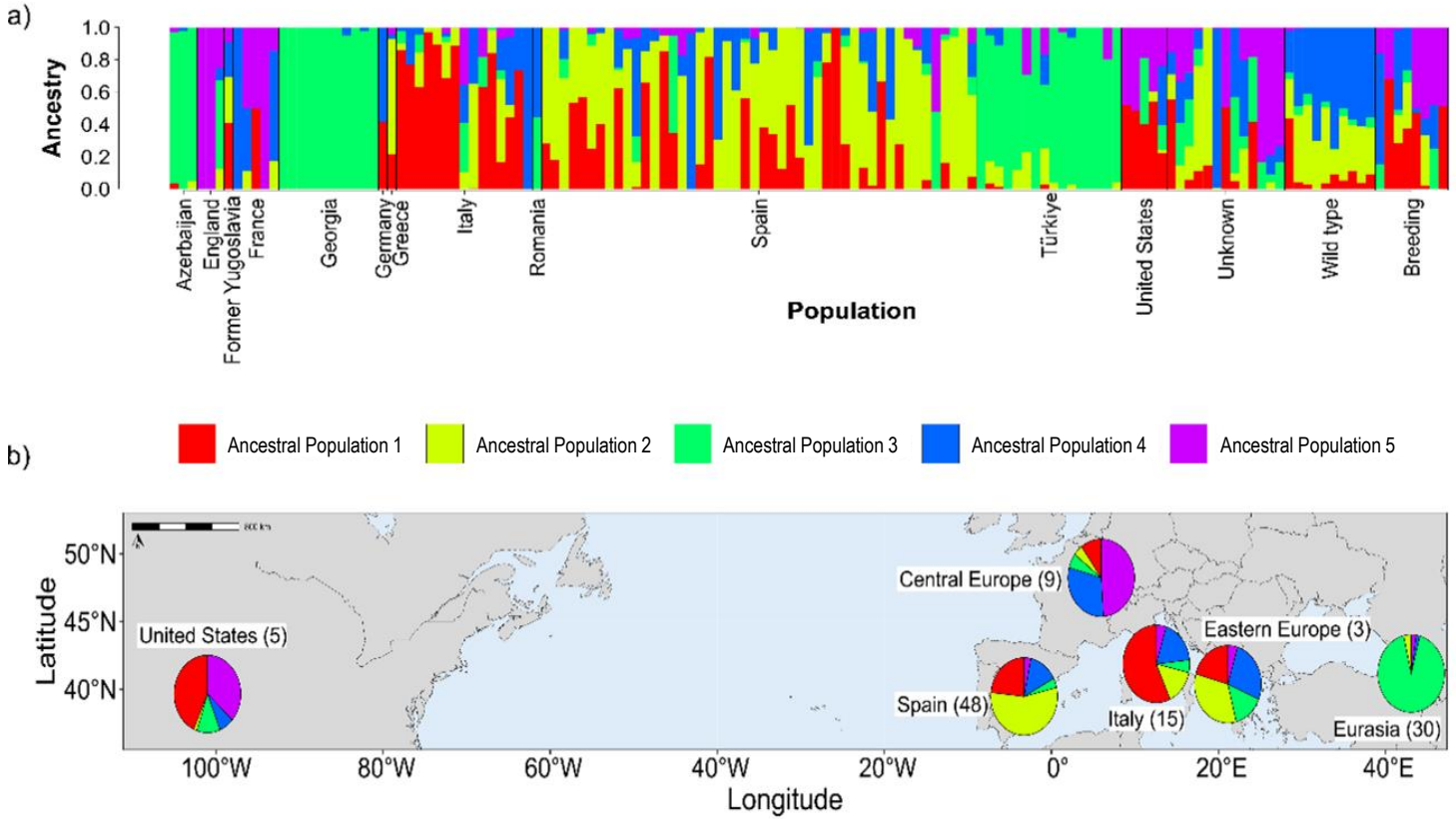
3 genetic clusters

Cluster 1: Italian and Spanish (Italian majority)

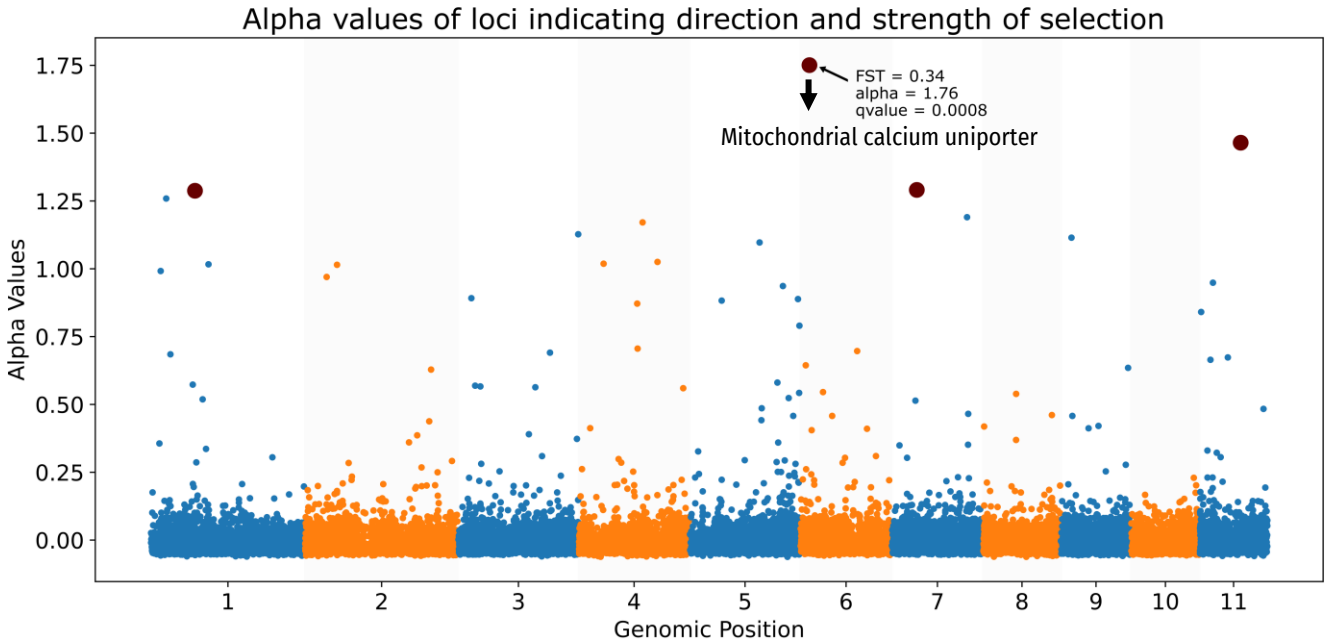
Cluster 2; Eurasian varieties (Azerbaijan, Georgia, and Turkey) form a group with **unique genetic characteristics**

Cluster 3: Spanish and Italian Samples (Spanish majority)

Estimating Ancestry and Admixture of Samples



Are there loci that are differentially selected in Eurasian samples?



Mitochondrial calcium uniporters → **calcium signalling** influences **bud dormancy** and **cold acclimation** in perennial plants → **variation** in this gene might underlie **adaptations to temperate Eurasian climates** → candidate for explaining **divergence in stress resilience** between Eurasian and other hazelnut populations

Are there loci associated to nut morphological traits?

Area Max. Caliber Min. Caliber Perimeter

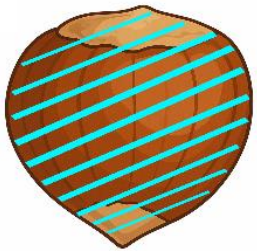
We measured 4 nut traits and searched for associated variants

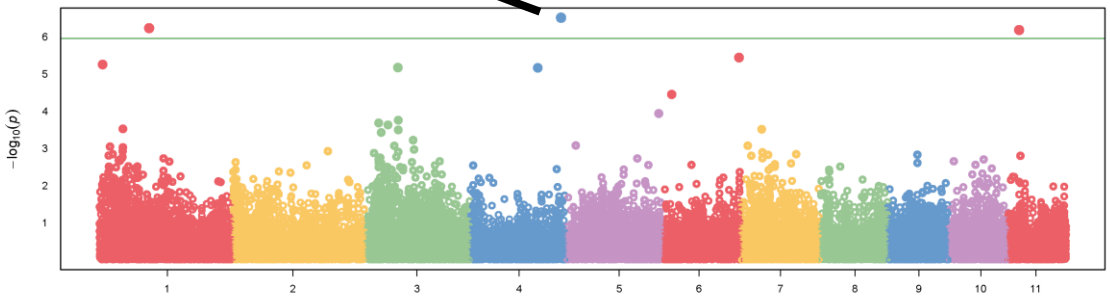
HSP21 is a heat shock protein known to support **chloroplast function** and **seedling development** under **heat stress** in other species.

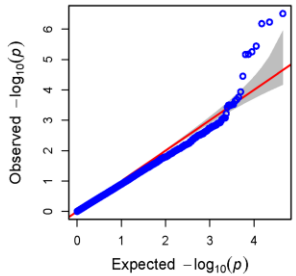
→ ***HSP21*** **may protect developing nut cells** from **environmental stress**.

BUT: Further studies are needed to confirm this hypothesis!

Nut area







Conclusions

- The **genetic uniqueness of Eurasian accessions** can stimulate research **about how the genetic potential can be used in breeding efforts**.

- **GWAS** results provided **suggestions for the genetic factors influencing nut morphology** and **enhanced our understanding of genome-trait interactions.**

Thanks to:

- Monique Salardi-Jost 1
- Gabriele Dini 1
- Mercy Wairimu Macharia 1
- Mercè Rovira 2
- Valerio Cristofori 3
- Mario Enrico Pé 1
- Andrea Cavallero 4
- Claudio Todeschini 5
- Giuseppe Genova 4
- Matteo Dell'Acqua 1

1 INSTITUTE OF PLANT SCIENCES Sant'Anna School of Advanced Studies – Pisa

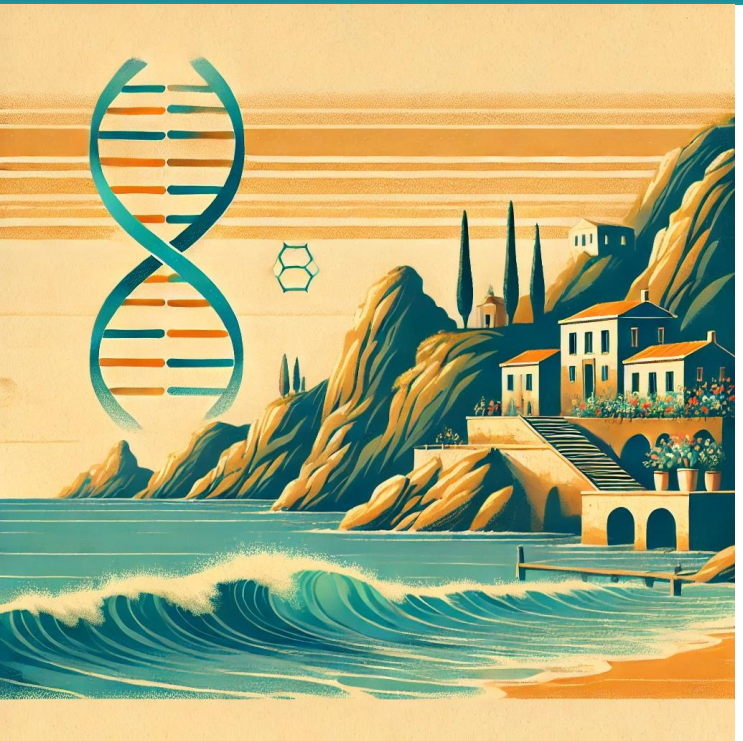
2 IRTA Institute of Agrifood Research and Technology

3 UNIVERSITÀ DEGLI STUDI DELLA TUSCIA DIPARTIMENTO DI SCIENZE AGRARIE E FORESTALI

4 SOREMARTEC Gruppo FERRERO

5 HCo FERRERO Hazelnut Company

THANK YOU!



Questions?