RNA Seq Data Analysis

Yedomon Ange Bovys Zoclanclounon, PhD

@angeomics angez9914@gmail.com

November, 2023

Content

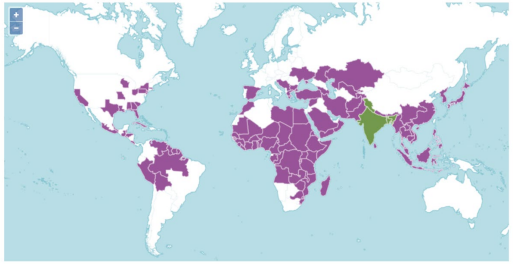
1 | A little bit of context

Sesame world trade

- **2014 USD 2500 per tonnes**
- **2017 USD 2300 million Import**
- **2017 USD 2100 million Export**
- **2020 USD 373.3 million (Lignans)**

Step 1: Know your genetic resources

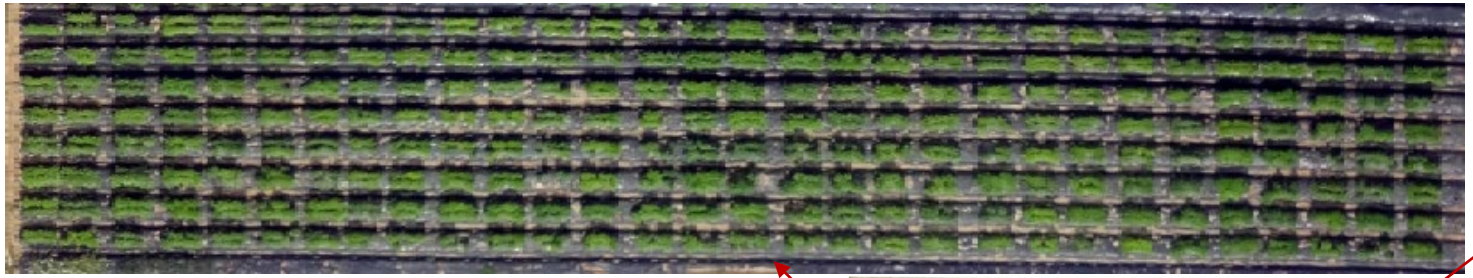
Where ? Origin | Native? | Introduced?

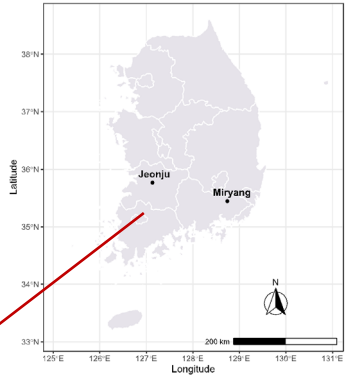


What ?

Wild	Pedigree
Landrace	Characteristic Trait
Cultivar	Core collection
Variety	Mapping population (RILS, MAGIC, ...)

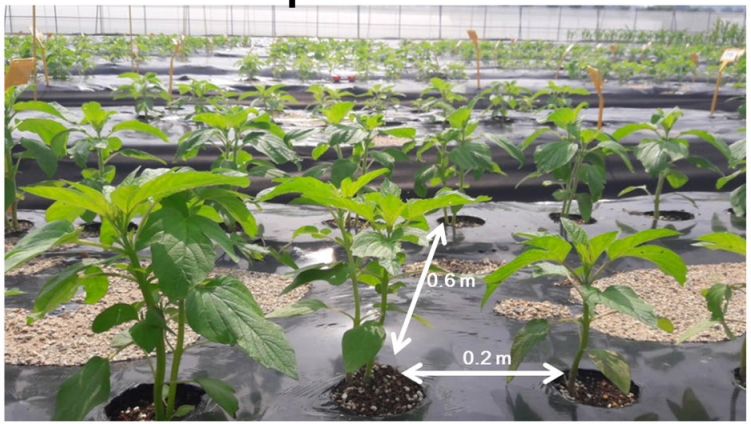
Step 1: Know your genetic resources

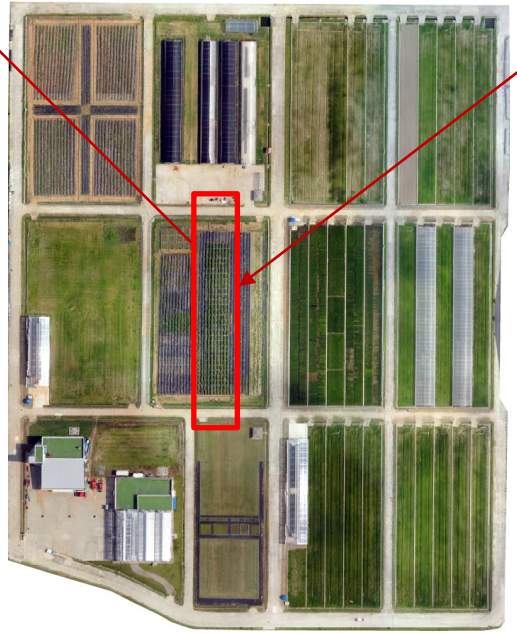




- **Federer Augmented Block Design**

  - Checks replicated 8 times



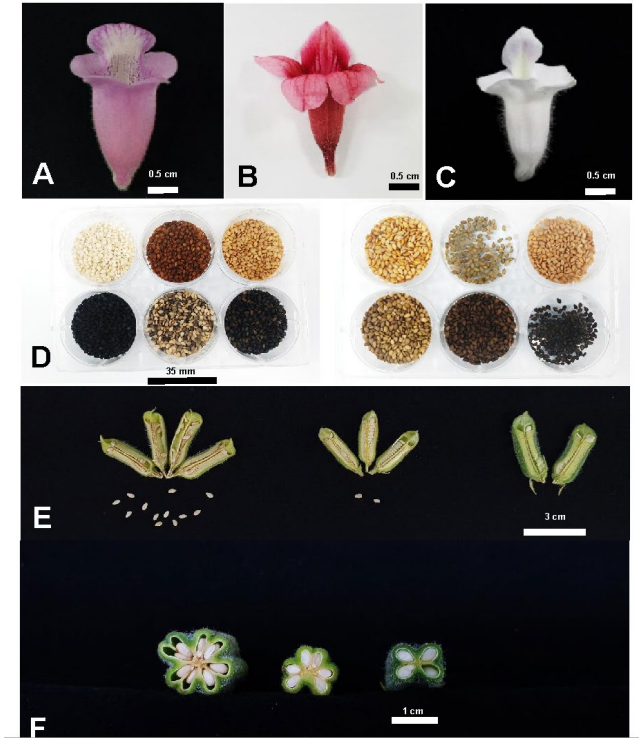


Total: 506 ccessions

Total: 24 traits

- **Agronomic (18)**
- **Seed quality (06)**
  - **oil**
  - **fatty acids**
  - **sesamin**
  - **sesamolin**

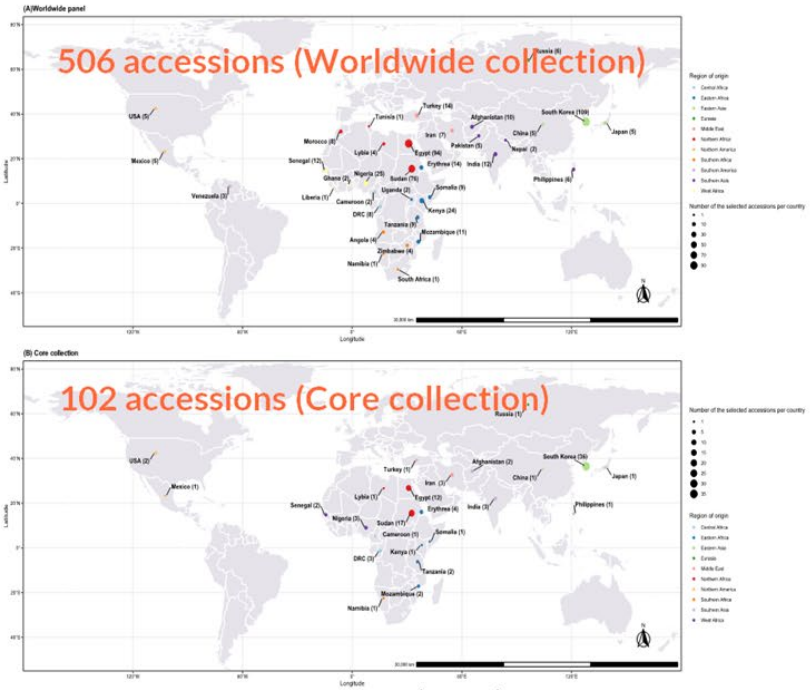
Step 1: Know your genetic resources



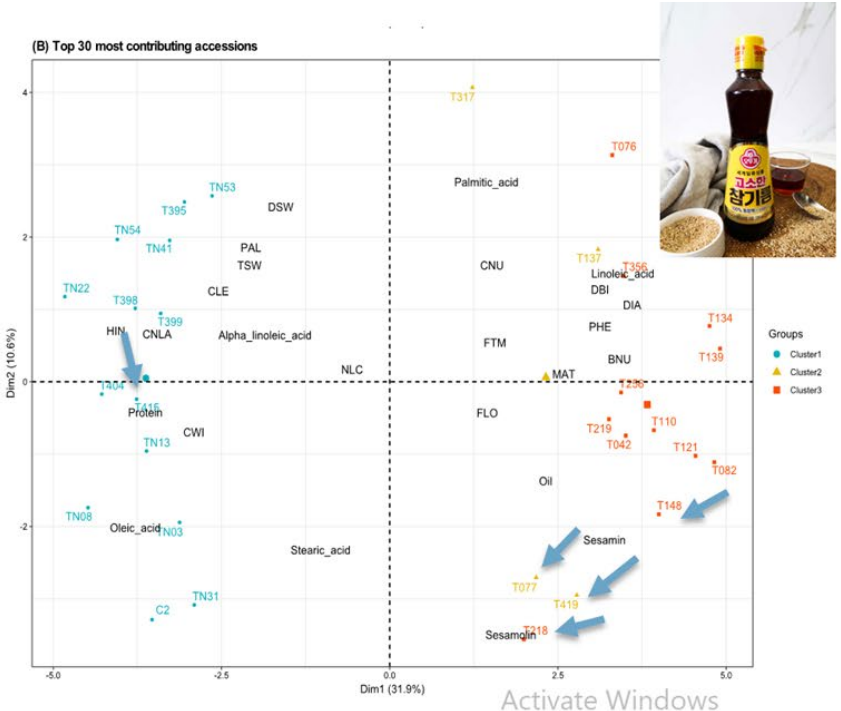
© Zoclanclounon (2022)

Eastern and Northern Africa also contribute to the high yield accessions in the cluster 3

Step 1: Know your genetic resources



©Zoclanclounon (2022)



©Zoclanclounon (2022)

Crude Oil - Fatty acids [Palmitic, Stearic, Oleic, Linoleic, alpha linoleic] – TN03 & T415 Lignan [sesamin, sesamolin] - T218, T077, T419, & T148

Step 1: Know your genetic resources

Which omics concept have we covered so far?

Their utility

Genotyping

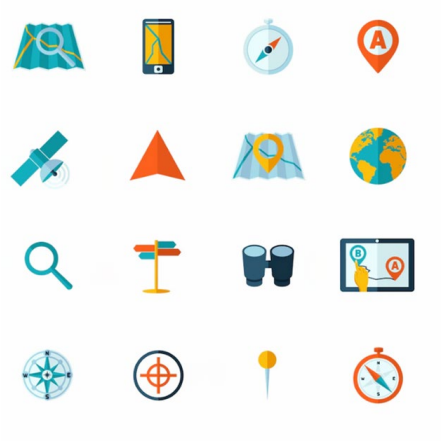
- **Array-based:** Affymetrix axiom – Affimetrix GeneChip – Illumina Infinium Beadchip
- **NGS-based:** GBS – DArT-seq – RAD-seq – ddRAD – REST-Seq

Whole genome sequencing – Pangenomes – Structural Variations

Trait mapping: GWAS – QTL detection

Step 1: Know your genetic resources

Find a
gene?
Where?



Step 2: Generate genomic resources

Genome assembly

Genes

SSRs

QTL

Database

SNPs

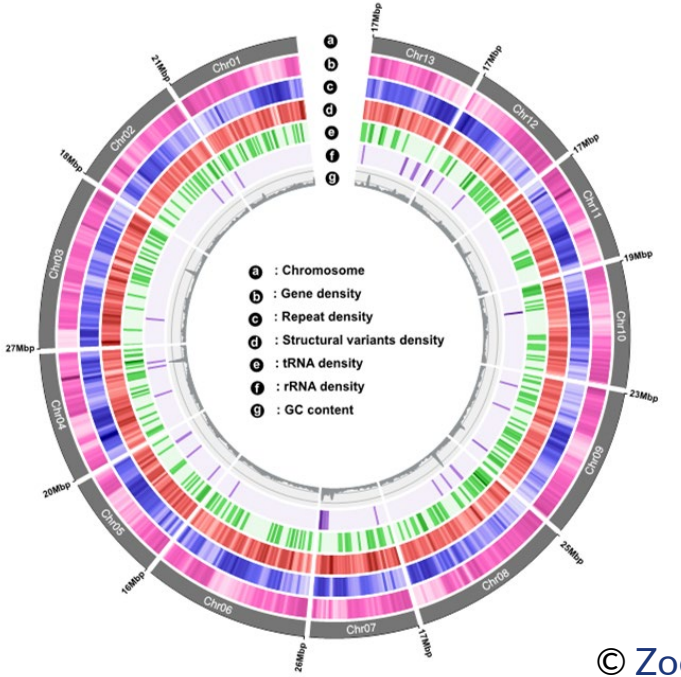
Annoted genome

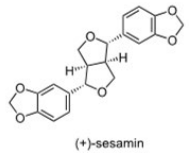
Molecular markers

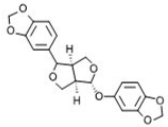
Step 2: Generate genomic resources

1.16 ton per hectare | high oil content of 50.2% | sesamin : 3.96 mg/g sesamolin 2.57mg/g | Linoleic acid: 44.5%

Sesamum indicum cv Goenbaek





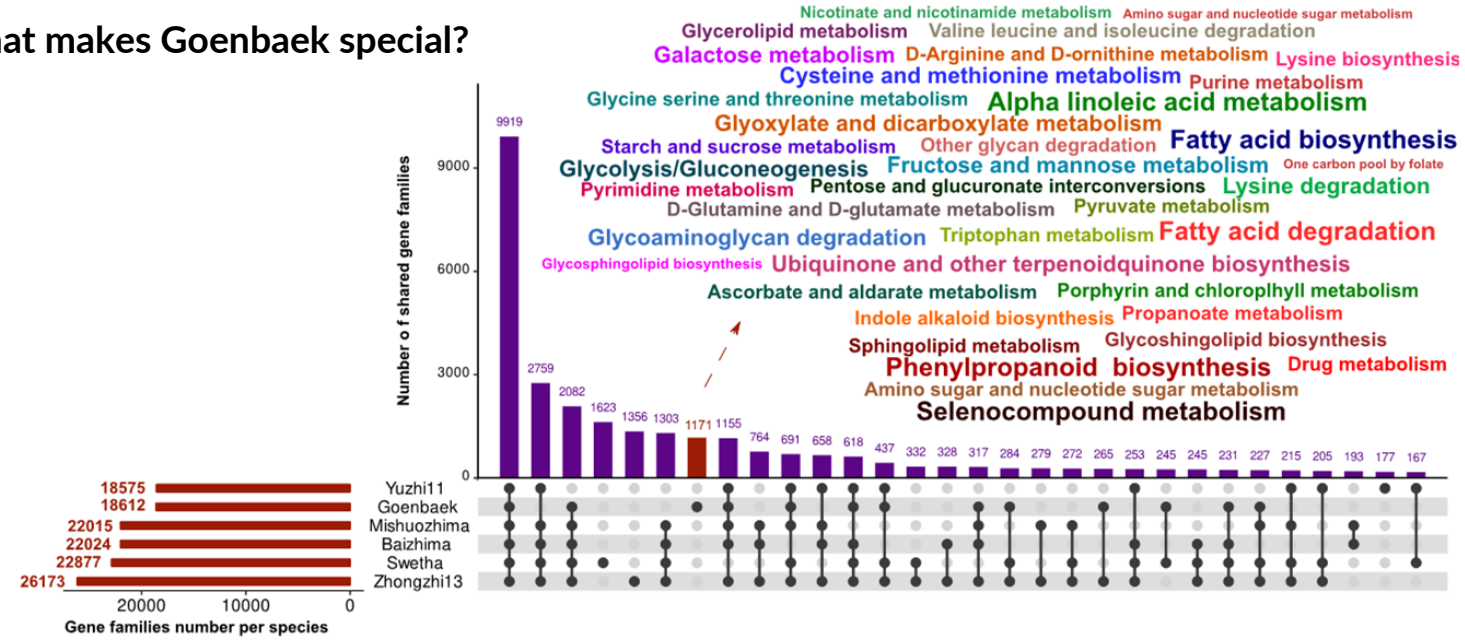


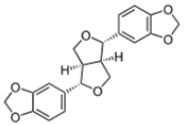
Circos plot of Goenbaek genome

Data: https://www.ncbi.nlm.nih.gov/bioproject/810203

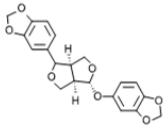
Step 2: Generate genomic resources

What makes Goenbaek special?





(+)-sesamin

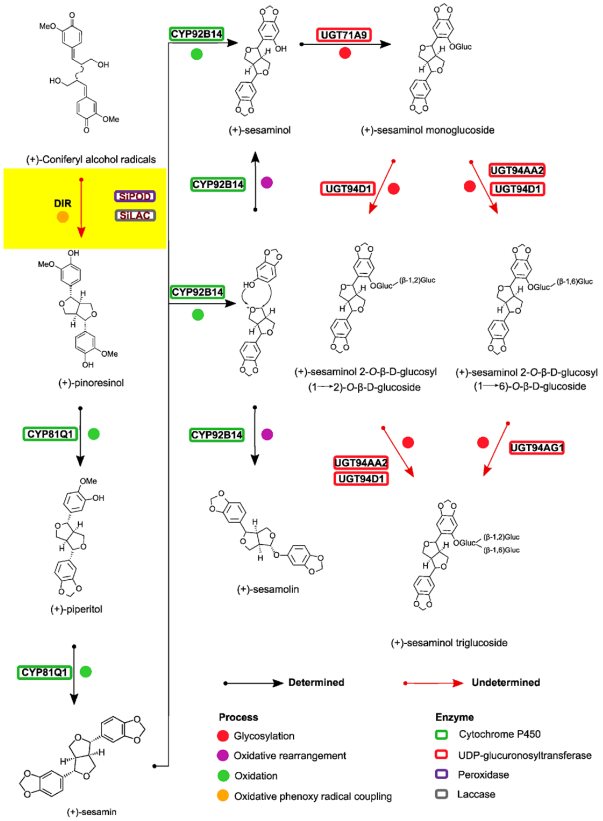


(+)-sesamolin

© Zoclanclounon et al. (2024)

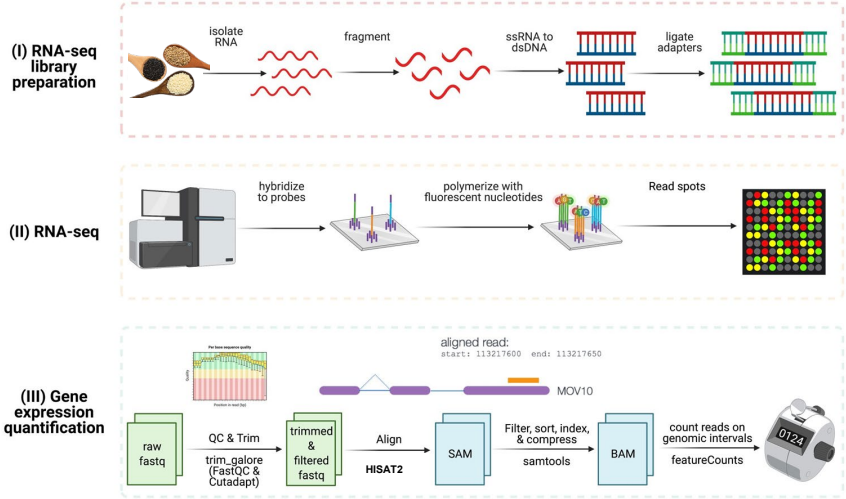
Step 3: Investigate key genes of interest

Step 3: Investigate key genes of interest



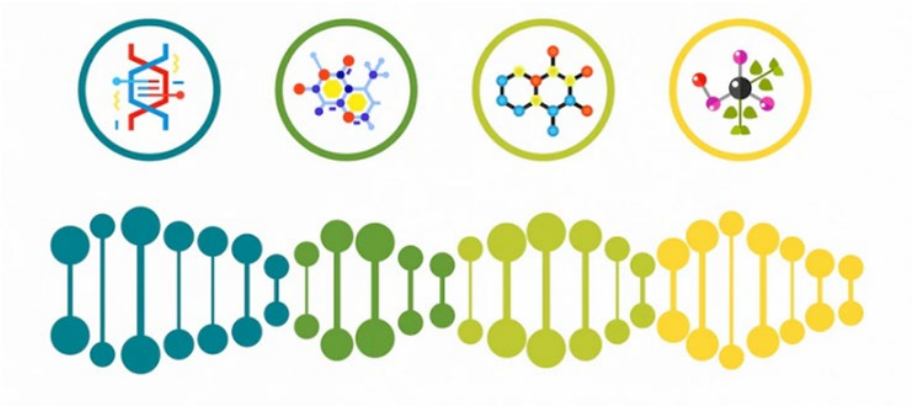
© Zoclancloun*on et al.* (2022)

Transcriptomics in action



White vs Black | Rich oil vs Low oil

2 | Data Acquisition



Data Acquisition | Metadata matters !!!

Variety name	Key characteristic	NCBI Project	SRA ID
ZZM4728	High oil content (59g/100g seed)	PRJNA400575	SRR6010085,SRR6010086,SRR6010087
ZZM2161	Low oil content (48g/100g seed)	PRJNA400575	SRR6010088,SRR6010089,SRR6010090

https://www.ebi.ac.uk/ena/browser/view/SRR6010085

https://www.ncbi.nlm.nih.gov/sra/?term=SRR6010085

ENA European Nucleotide Archive

NIH National Library of Medicine National Center for Biotechnology Information

Data Acquisition | srahunter

srahunter

```
srahunter download -i accession_list.txt
```

Data Acquisition | srahunter

```
(srahunter_env) angeomics@DESKTOP-UB1829G:/home/angeomics/data$ vi accession_list.txt
(srahunter_env) angeomics@DESKTOP-UB1829G:/home/angeomics/data$ srahunter download -i accession_list.txt
                                                  _
 ___ _ __ __ __ _| |__   _   _ _ __ | |_ ___ _ __
/ __| '__/ _` | '_ \| | | | '_ \| __/ _ \ '__|
\__ \ | | (_| | | | | |_| | | | | ||  __/ |
|___/_|  \__,_|_| |_|\__,_|_| |_|\__\___|_|


Downloading with list: accession_list.txt
Number of t: 6
Download path: /home/angeomics/data/tmp_srahunter
Max size: 50G
Output directory: /home/angeomics/data
Currently downloading: SRR6010085
The command used was: prefetch -p -X 50G SRR6010085 --output-file /home/angeomics/data/tmp_srahunter/SRR6010085.sra
2024-10-30T18:30:23 prefetch.3.1.1: 1) Resolving 'SRR6010085'...
2024-10-30T18:30:25 prefetch.3.1.1: Current preference is set to retrieve SRA Normalized Format files with full base quality scores
2024-10-30T18:30:25 prefetch.3.1.1: 1) Downloading 'SRR6010085'...
2024-10-30T18:30:25 prefetch.3.1.1:  SRA Normalized Format file is being retrieved
2024-10-30T18:30:25 prefetch.3.1.1:  Downloading via HTTPS...
|---------------------------------------------------- 100%
2024-10-30T18:42:10 prefetch.3.1.1:  HTTPS download succeed
2024-10-30T18:42:18 prefetch.3.1.1:  'SRR6010085' is valid: 1750385155 bytes were streamed from 1750378499
2024-10-30T18:42:18 prefetch.3.1.1: 1) 'SRR6010085' was downloaded successfully
Generating fastq for: SRR6010085
The command used was: fasterq-dump --skip-technical -p -e 6 /home/angeomics/data/tmp_srahunter/SRR6010085.sra --outdir /home/angeomics/data
join   :|---------------------------------------------------- 100%
concat :|---------------------------------------------------- 100%
spots read      : 13,305,758
reads read      : 26,611,516
reads written   : 26,611,516
Processing SRR6010085 completed successfully.
(srahunter_env) angeomics@DESKTOP-UB1829G:/home/angeomics/data$ ls -rltha
total 7.6G
drwxr-x--- 7 angeomics angeomics 4.0K Oct 30 18:29 ..
-rw-r--r-- 1 angeomics angeomics   11 Oct 30 18:29 accession_list.txt
drwxr-xr-x 2 angeomics angeomics 4.0K Oct 30 18:42 tmp_srahunter
-rw-r--r-- 1 angeomics angeomics 3.8G Oct 30 18:45 SRR6010085_2.fastq
-rw-r--r-- 1 angeomics angeomics 3.8G Oct 30 18:46 SRR6010085_1.fastq
drwxr-xr-x 3 angeomics angeomics 4.0K Oct 30 18:46 .
(srahunter_env) angeomics@DESKTOP-UB1829G:/home/angeomics/data$
```

Data Acquisition | Fastq file explained!

```
angeomics@DESKTOP-UB182!

(base) angeomics@DESKTOP-UB1829G:/home/angeomics/data$ head -n 5 SRR6010085_1.fastq
@SRR6010085.1 FCC0CF3ACXX:8:1101:1600:2212 length=90
GTTTGATGGAAATTTCCTTGTAAATTTCATGGAAGGCTTAGGAATAAAAGTGACAGATGGCATTGACAAAACAAAAGTTGGGATCGAATT
+SRR6010085.1 FCC0CF3ACXX:8:1101:1600:2212 length=90
CCCFFFFFHHHHHJJJJJJJHIJJIJJJJJJJJJJJJJJIJJIIJJJJJIEHHJJJJJIJJJJJIIJIHIIJIJJIHHHHHFFFDDDECD
@SRR6010085.2 FCC0CF3ACXX:8:1101:1849:2096 length=90
(base) angeomics@DESKTOP-UB1829G:/home/angeomics/data$
```

Data Acquisition | Fastq file explained!

```
angeomics@DESKTOP-UB182!

(base) angeomics@DESKTOP-UB1829G:/home/angeomics/data$ head -n 5 SRR6010085_1.fastq
@SRR6010085.1 FCC0CF3ACXX:8:1101:1600:2212 length=90
GTTTGATGGAAATTTCCTTGTAAATTTCATGGAAGGCTTAGGAATAAAAGTGACAGATGGCATTGACAAAACAAAAGTTGGGATCGAATT
+SRR6010085.1 FCC0CF3ACXX:8:1101:1600:2212 length=90
CCCFFFFFHHHHHJJJJJJJHIJJIJJJJJJJJJJJJJJIJJIIJJJJJIEHHJJJJJIJJJJJIIJIHIIJIJJIHHHHHFFFDDDECD
@SRR6010085.2 FCC0CF3ACXX:8:1101:1849:2096 length=90
(base) angeomics@DESKTOP-UB1829G:/home/angeomics/data$
```

Line 1 (@SRR6010085.1 FCC0CF3ACXX:8:1101:1600:2212 length=90):

- Starts with @, marking it as the header line.
- **SRR6010085.1**: The read identifier, which includes the SRA accession number (SRR6010085) and a unique read number (.1).
- **FCC0CF3ACXX:8:1101:1600:2212**: Details about the sequencing run and position of the read in the Illumina flow cell:
    - **FCC0CF3ACXX**: Flow cell ID.
    - **8**: Lane number.
    - **1101:1600:2212**: X and Y coordinates of the cluster within the flow cell.
- **length=90**: Specifies the read length in base pairs.

Data Acquisition | Fastq file explained!

```
angeomics@DESKTOP-UB182!

(base) angeomics@DESKTOP-UB1829G:/home/angeomics/data$ head -n 5 SRR6010085_1.fastq
@SRR6010085.1 FCC0CF3ACXX:8:1101:1600:2212 length=90
GTTTGATGGAAATTTCCTTGTAAATTTCATGGAAGGCTTAGGAATAAAAGTGACAGATGGCATTGACAAAACAAAAGTTGGGATCGAATT
+SRR6010085.1 FCC0CF3ACXX:8:1101:1600:2212 length=90
CCCFFFFFHHHHHJJJJJJJHIJJIJJJJJJJJJJJJJJIJJIIJJJJJIEHHJJJJJIJJJJJIIJIHIIJIJJIHHHHHFFFDDDECD
@SRR6010085.2 FCC0CF3ACXX:8:1101:1849:2096 length=90
(base) angeomics@DESKTOP-UB1829G:/home/angeomics/data$
```

Line 2 (GTTTGAT...TGAATT):

- Contains the DNA sequence for this read.

Data Acquisition | Fastq file explained!

```
angeomics@DESKTOP-UB182!

(base) angeomics@DESKTOP-UB1829G:/home/angeomics/data$ head -n 5 SRR6010085_1.fastq
@SRR6010085.1 FCC0CF3ACXX:8:1101:1600:2212 length=90
GTTTGATGGAAATTTCCTTGTAAATTTCATGGAAGGCTTAGGAATAAAAGTGACAGATGGCATTGACAAAACAAAAGTTGGGATCGAATT
+SRR6010085.1 FCC0CF3ACXX:8:1101:1600:2212 length=90
CCCFFFFFHHHHHJJJJJJJHIJJIJJJJJJJJJJJJJJIJJIIJJJJJIEHHJJJJJIJJJJJIIJIHIIJIJJIHHHHHFFFDDDECD
@SRR6010085.2 FCC0CF3ACXX:8:1101:1849:2096 length=90
(base) angeomics@DESKTOP-UB1829G:/home/angeomics/data$
```

- **Line 3** (+SRR6010085.1 FCC0CF3ACXX:8:1101:1600:2212 length=90):
- Begins with a + symbol, marking it as the separator line.
- Optionally, this line repeats the read identifier, but it can be left blank.

Data Acquisition | Fastq file explained!

```
angeomics@DESKTOP-UB182!

(base) angeomics@DESKTOP-UB1829G:/home/angeomics/data$ head -n 5 SRR6010085_1.fastq
@SRR6010085.1 FCC0CF3ACXX:8:1101:1600:2212 length=90
GTTTGATGGAAATTTCCTTGTAAATTTCATGGAAGGCTTAGGAATAAAAGTGACAGATGGCATTGACAAAACAAAAGTTGGGATCGAATT
+SRR6010085.1 FCC0CF3ACXX:8:1101:1600:2212 length=90
CCCFFFFFHHHHHJJJJJJJHIJJIJJJJJJJJJJJJJJIJJIIJJJJJIEHHJJJJJIJJJJJIIJIHIIJIJJIHHHHHFFFDDDECD
@SRR6010085.2 FCC0CF3ACXX:8:1101:1849:2096 length=90
(base) angeomics@DESKTOP-UB1829G:/home/angeomics/data$
```

Line 4 (CCCFFFF...DECD):

•Shows the quality scores for each base in the sequence. Each character corresponds to a base in Line 2 and represents its quality score (often in ASCII, with F and higher generally indicating high quality).

Click here (https://www.drive5.com/usearch/manual/quality_score.html)to learn about quality scores

Data Acquisition | Fastq file explained!

```
angeomics@DESKTOP-UB182!

(base) angeomics@DESKTOP-UB1829G:/home/angeomics/data$ head -n 5 SRR6010085_1.fastq
@SRR6010085.1 FCC0CF3ACXX:8:1101:1600:2212 length=90
GTTTGATGGAAATTTCCTTGTAAATTTCATGGAAGGCTTAGGAATAAAAGTGACAGATGGCATTGACAAAACAAAAGTTGGGATCGAATT
+SRR6010085.1 FCC0CF3ACXX:8:1101:1600:2212 length=90
CCCFFFFFHHHHHJJJJJJJHIJJIJJJJJJJJJJJJJJIJJIIJJJJJIEHHJJJJJIJJJJJIIJIHIIJIJJIHHHHHFFFDDDECD
@SRR6010085.2 FCC0CF3ACXX:8:1101:1849:2096 length=90
(base) angeomics@DESKTOP-UB1829G:/home/angeomics/data$
```

Task: How many reads we have in each fastq file ?

Data Acquisition | Fastq file explained!

```
angeomics@DESKTOP-UB182!
(base) angeomics@DESKTOP-UB1829G:/home/angeomics/data$ head -n 5 SRR6010085_1.fastq
@SRR6010085.1 FCC0CF3ACXX:8:1101:1600:2212 length=90
GTTTGATGGAAATTTCCTTGTAAATTTCATGGAAGGCTTAGGAATAAAAGTGACAGATGGCATTGACAAAACAAAAGTTGGGATCGAATT
+SRR6010085.1 FCC0CF3ACXX:8:1101:1600:2212 length=90
CCCFFFFFHHHHHJJJJJJJHIJJIJJJJJJJJJJJJJJIJJIIJJJJJIEHHJJJJJIJJJJJIIJIHIIJIJJIHHHHHFFFDDDECD
@SRR6010085.2 FCC0CF3ACXX:8:1101:1849:2096 length=90
(base) angeomics@DESKTOP-UB1829G:/home/angeomics/data$
```

```
angeomics@DESKTOP-UB182!
(base) angeomics@DESKTOP-UB1829G:/home/angeomics/data$ grep "@SRR" SRR6010085_1.fastq | wc -l
13305758
(base) angeomics@DESKTOP-UB1829G:/home/angeomics/data$ grep "@SRR" SRR6010085_2.fastq | wc -l
13305758
(base) angeomics@DESKTOP-UB1829G:/home/angeomics/data$
```

https://www.ncbi.nlm.nih.gov/sra/?term=SRR6010085

Data Acquisition | fastqc

FastQC

```
fastqc -t 8 -o fastqc_dir/ SRR6010085_*.fastq
```

Data Acquisition | fastqc

```
(fastqc_env) angeomics@DESKTOP-UB1829G:/home/angeomics/data$ fastqc -t 8 -o fastqc_dir/ SRR6010085_*.fastq
null
null
Started analysis of SRR6010085_1.fastq
Started analysis of SRR6010085_2.fastq
Approx 5% complete for SRR6010085_1.fastq
Approx 5% complete for SRR6010085_2.fastq
Approx 10% complete for SRR6010085_1.fastq
Approx 10% complete for SRR6010085_2.fastq
Approx 15% complete for SRR6010085_1.fastq
Approx 15% complete for SRR6010085_2.fastq
Approx 20% complete for SRR6010085_1.fastq
Approx 20% complete for SRR6010085_2.fastq
Approx 25% complete for SRR6010085_2.fastq
Approx 25% complete for SRR6010085_1.fastq
Approx 30% complete for SRR6010085_2.fastq
Approx 30% complete for SRR6010085_1.fastq
Approx 35% complete for SRR6010085_2.fastq
Approx 35% complete for SRR6010085_1.fastq
Approx 40% complete for SRR6010085_2.fastq
Approx 40% complete for SRR6010085_1.fastq
Approx 45% complete for SRR6010085_2.fastq
Approx 45% complete for SRR6010085_1.fastq
Approx 50% complete for SRR6010085_2.fastq
Approx 50% complete for SRR6010085_1.fastq
Approx 55% complete for SRR6010085_2.fastq
Approx 55% complete for SRR6010085_1.fastq
Approx 60% complete for SRR6010085_2.fastq
Approx 60% complete for SRR6010085_1.fastq
Approx 65% complete for SRR6010085_2.fastq
Approx 65% complete for SRR6010085_1.fastq
Approx 70% complete for SRR6010085_2.fastq
Approx 70% complete for SRR6010085_1.fastq
Approx 75% complete for SRR6010085_2.fastq
Approx 75% complete for SRR6010085_1.fastq
Approx 80% complete for SRR6010085_2.fastq
Approx 80% complete for SRR6010085_1.fastq
Approx 85% complete for SRR6010085_2.fastq
Approx 85% complete for SRR6010085_1.fastq
Approx 90% complete for SRR6010085_2.fastq
Approx 90% complete for SRR6010085_1.fastq
Approx 95% complete for SRR6010085_2.fastq
Approx 95% complete for SRR6010085_1.fastq
Analysis complete for SRR6010085_2.fastq
Analysis complete for SRR6010085_1.fastq
(fastqc_env) angeomics@DESKTOP-UB1829G:/home/angeomics/data$
```

Please check the report

Data Cleaning | fastp

```
fastp --detect_adapter_for_pe \
        --overrepresentation_analysis \
        --correction --cut_right --thread 2 \
        --html trimmed_dir/SRR6010085.fastp.html
        --json trimmed_dir/SRR6010085.fastp.json \
        -i SRR6010085_1.fastq -I SRR6010085_2.fastq \
        -o trimmed_dir/SRR6010085_1.fastq -O trimmed_dir/SRR6010085_2.fastq
```

Data Cleaning | fastp

```
angeomics@DESKTOP-UB1829

(fastp_env) angeomics@DESKTOP-UB1829G:/home/angeomics/data$ fastp --detect_adapter_for_pe \
        --overrepresentation_analysis \
        --correction --cut_right --thread 2 \
        --html trimmed_dir/SRR6010085.fastp.html --json trimmed_dir/SRR6010085.fastp.json \
        -i SRR6010085_1.fastq -I SRR6010085_2.fastq \
        -o trimmed_dir/SRR6010085_1.fastq -O trimmed_dir/SRR6010085_2.fastq
Detecting adapter sequence for read1...
AAAGGCTTACGGTGGATACCTAGGCACCCAGAGACGAGGAAGGGCGTAGTAATCGACGAA

Detecting adapter sequence for read2...
No adapter detected for read2

Read1 before filtering:
total reads: 13305758
total bases: 1197518220
Q20 bases: 1157527932(96.6606%)
Q30 bases: 1073681705(89.6589%)

Read2 before filtering:
total reads: 13305758
total bases: 1197518220
Q20 bases: 1102927865(92.1011%)
Q30 bases: 996543643(83.2174%)

Read1 after filtering:
total reads: 12402218
total bases: 1059211267
Q20 bases: 1050942053(99.2193%)
Q30 bases: 997961572(94.2174%)

Read2 after filtering:
total reads: 12402218
total bases: 1008532481
Q20 bases: 995490379(98.7068%)
Q30 bases: 928508411(92.0653%)

Filtering result:
reads passed filter: 24804436
reads failed due to low quality: 1742
reads failed due to too many N: 0
reads failed due to too short: 1805338
reads with adapter trimmed: 192723
bases trimmed due to adapters: 6579734
reads corrected by overlap analysis: 4204
bases corrected by overlap analysis: 4314

Duplication rate: 28.224%

Insert size peak (evaluated by paired-end reads): 90

JSON report: trimmed_dir/SRR6010085.fastp.json
HTML report: trimmed_dir/SRR6010085.fastp.html

fastp --detect_adapter_for_pe --overrepresentation_analysis --correction --cut_right --thread 2 --html trimmed_dir/SRR6010085.fastp.html --json trimmed_dir/SRR6010085.fastp.json -i SRR6010085_1.fastq -I SRR6010085_2.fastq -o trimmed_d
/SRR6010085_1.fastq -O trimmed_dir/SRR6010085_2.fastq
fastp v0.23.4, time used: 278 seconds
(fastp_env) angeomics@DESKTOP-UB1829G:/home/angeomics/data$
```

Data Cleaning | FastQC on trimmed data

FastQC on trimmed data

```
fastqc -t 8 -o fastqc_trimmed_dir/ trimmed_dir/SRR6010085_*.fastq
```

Data Cleaning | FastQC on trimmed data

```
angeomics@DESKTOP-UB1829

(fastqc_env) angeomics@DESKTOP-UB1829G:/home/angeomics/data$ fastqc -t 8 -o fastqc_trimmed_dir/ trimmed_dir/SRR6010085_*.fastq
null
null
Started analysis of SRR6010085_1.fastq
Started analysis of SRR6010085_2.fastq
Approx 5% complete for SRR6010085_1.fastq
Approx 5% complete for SRR6010085_2.fastq
Approx 10% complete for SRR6010085_2.fastq
Approx 10% complete for SRR6010085_1.fastq
Approx 15% complete for SRR6010085_2.fastq
Approx 15% complete for SRR6010085_1.fastq
Approx 20% complete for SRR6010085_2.fastq
Approx 20% complete for SRR6010085_1.fastq
Approx 25% complete for SRR6010085_2.fastq
Approx 25% complete for SRR6010085_1.fastq
Approx 30% complete for SRR6010085_2.fastq
Approx 30% complete for SRR6010085_1.fastq
Approx 35% complete for SRR6010085_2.fastq
Approx 35% complete for SRR6010085_1.fastq
Approx 40% complete for SRR6010085_2.fastq
Approx 40% complete for SRR6010085_1.fastq
Approx 45% complete for SRR6010085_2.fastq
Approx 45% complete for SRR6010085_1.fastq
Approx 50% complete for SRR6010085_2.fastq
Approx 50% complete for SRR6010085_1.fastq
Approx 55% complete for SRR6010085_2.fastq
Approx 55% complete for SRR6010085_1.fastq
Approx 60% complete for SRR6010085_2.fastq
Approx 60% complete for SRR6010085_1.fastq
Approx 65% complete for SRR6010085_2.fastq
Approx 65% complete for SRR6010085_1.fastq
Approx 70% complete for SRR6010085_2.fastq
Approx 70% complete for SRR6010085_1.fastq
Approx 75% complete for SRR6010085_2.fastq
Approx 75% complete for SRR6010085_1.fastq
Approx 80% complete for SRR6010085_2.fastq
Approx 80% complete for SRR6010085_1.fastq
Approx 85% complete for SRR6010085_2.fastq
Approx 85% complete for SRR6010085_1.fastq
Approx 90% complete for SRR6010085_2.fastq
Approx 90% complete for SRR6010085_1.fastq
Approx 95% complete for SRR6010085_2.fastq
Approx 95% complete for SRR6010085_1.fastq
Analysis complete for SRR6010085_2.fastq
Analysis complete for SRR6010085_1.fastq
(fastqc_env) angeomics@DESKTOP-UB1829G:/home/angeomics/data$
```

Data Cleaning | MultiQC

MultiQC

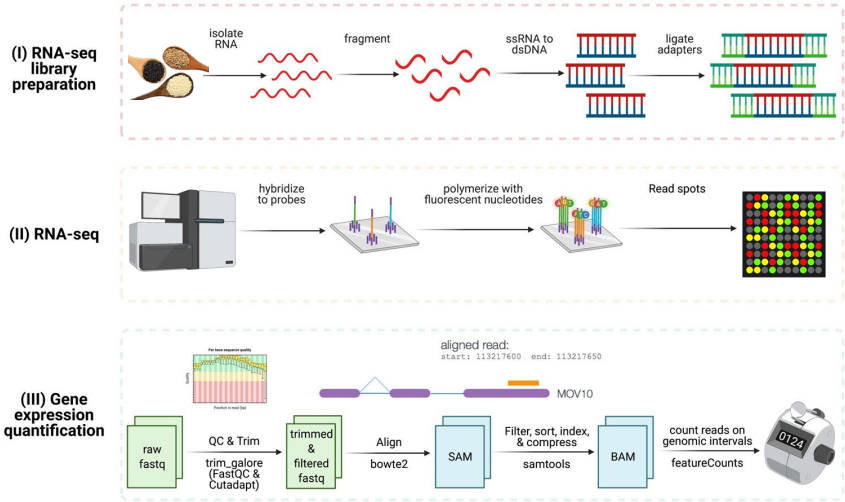
```
multiqc fastqc_trimmed_dir/ trimmed_dir/ --outdir multiqc_dir
```

Data Cleaning | MultiQC

```
angeomics@DESKTOP-UB182!

(multiqc_env) angeomics@DESKTOP-UB1829G:/home/angeomics/data$ multiqc fastqc_trimmed_dir/ trimmed_dir/ --outdir multiqc_dir

/// MultiQC 🔍 v1.25.1

       file_search | Search path: /home/angeomics/data/fastqc_trimmed_dir
       file_search | Search path: /home/angeomics/data/trimmed_dir
         searching | ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 100% 8/8
             fastp | Found 1 reports
            fastqc | Found 2 reports
     write_results | Data        : multiqc_dir/multiqc_data
     write_results | Report      : multiqc_dir/multiqc_report.html
           multiqc | MultiQC complete
(multiqc_env) angeomics@DESKTOP-UB1829G:/home/angeomics/data$
```

Please check the report

3 | Alignment


(I) RNA-seq library preparation
isolate RNA
fragment
ssRNA to dsDNA
ligate adapters
(II) RNA-seq
hybridize to probes
polymerize with fluorescent nucleotides
Read spots
(III) Gene expression quantification
aligned read:
start: 113217600 end: 113217650
MOV10
raw fastq
QC & Trim
trim_galore (FastQC & Cutadapt)
trimmed & filtered fastq
Align
bowte2
SAM
Filter, sort, index, & compress
samtools
BAM
count reads on genomic intervals
featureCounts
0124

Alignment with HISAT2 | Genome Data

Assembly and annotation data : https://zenodo.org/records/6350881

Source : https://www.researchsquare.com/article/rs-4887813/v1

- Assembly: Sesamum_indicum_goenbaek.fasta
- Annotation: Sesamum_indicum_goenbaek.gff3

Alignment with HISAT2 | Build index

```
hisat2-build Sesamum_indicum_goenbaek.fasta Sesamum_indicum_goenbaek
```

```
angeomics@DESKTOP-UB182!

(hisat2_env) angeomics@DESKTOP-UB1829G:/home/angeomics/analyses$ ls -rlth
total 669M
-rw-r--r-- 1 angeomics angeomics 255M Feb 22  2022 Sesamum_indicum_goenbaek.fasta
-rw-r--r-- 1 angeomics angeomics  28M Mar 14  2022 Sesamum_indicum_goenbaek.gff3
-rw-r--r-- 1 angeomics angeomics  63M Oct 30 21:44 Sesamum_indicum_goenbaek.4.ht2
-rw-r--r-- 1 angeomics angeomics 1.8K Oct 30 21:44 Sesamum_indicum_goenbaek.3.ht2
-rw-r--r-- 1 angeomics angeomics    8 Oct 30 21:44 Sesamum_indicum_goenbaek.8.ht2
-rw-r--r-- 1 angeomics angeomics   12 Oct 30 21:44 Sesamum_indicum_goenbaek.7.ht2
-rw-r--r-- 1 angeomics angeomics  63M Oct 30 21:47 Sesamum_indicum_goenbaek.2.ht2
-rw-r--r-- 1 angeomics angeomics  88M Oct 30 21:47 Sesamum_indicum_goenbaek.1.ht2
-rw-r--r-- 1 angeomics angeomics  64M Oct 30 21:47 Sesamum_indicum_goenbaek.6.ht2
-rw-r--r-- 1 angeomics angeomics 110M Oct 30 21:47 Sesamum_indicum_goenbaek.5.ht2
(hisat2_env) angeomics@DESKTOP-UB1829G:/home/angeomics/analyses$
```

Alignment with HISAT2 | Mapping

```
hisat2 \
-p 8 \
-x Sesamum_indicum_goenbaek \
-1 /home/angeomics/data/01.SRA/trimmed_dir/SRR6010085_1.fastq \
-2 /home/angeomics/data/01.SRA/trimmed_dir/SRR6010085_2.fastq \
-S SRR6010085.bam
```

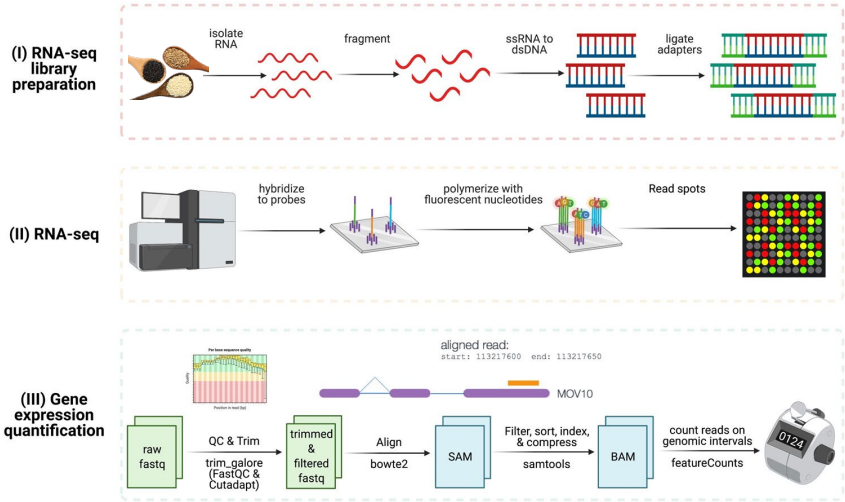
```
angeomics@DESKTOP-UB182

(hisat2_env) angeomics@DESKTOP-UB1829G:/home/angeomics/analyses$ hisat2 -p 8 -x Sesamum_indicum_goenbaek -1 /home/angeomics/data/01.SRA/trimmed_dir/SRR6010085_1.fastq -2 /home/angeomics/data/01.SRA/trimmed_dir/SRR6010085_2.fastq -S SRR6010085.bam
12402218 reads; of these:
  12402218 (100.00%) were paired; of these:
    2715669 (21.90%) aligned concordantly 0 times
    9387390 (75.69%) aligned concordantly exactly 1 time
    299159 (2.41%) aligned concordantly >1 times
    ----
    2715669 pairs aligned concordantly 0 times; of these:
      25232 (0.93%) aligned discordantly 1 time
    ----
    2690437 pairs aligned 0 times concordantly or discordantly; of these:
      5380874 mates make up the pairs; of these:
        5167736 (96.04%) aligned 0 times
        201644 (3.75%) aligned exactly 1 time
        11494 (0.21%) aligned >1 times
79.17% overall alignment rate
(hisat2_env) angeomics@DESKTOP-UB1829G:/home/angeomics/analyses$
```

Alignment with HISAT2 | Mapping

```
samtools view --threads 8 -bS -o SRR6010085.bam SRR6010085.sam
```

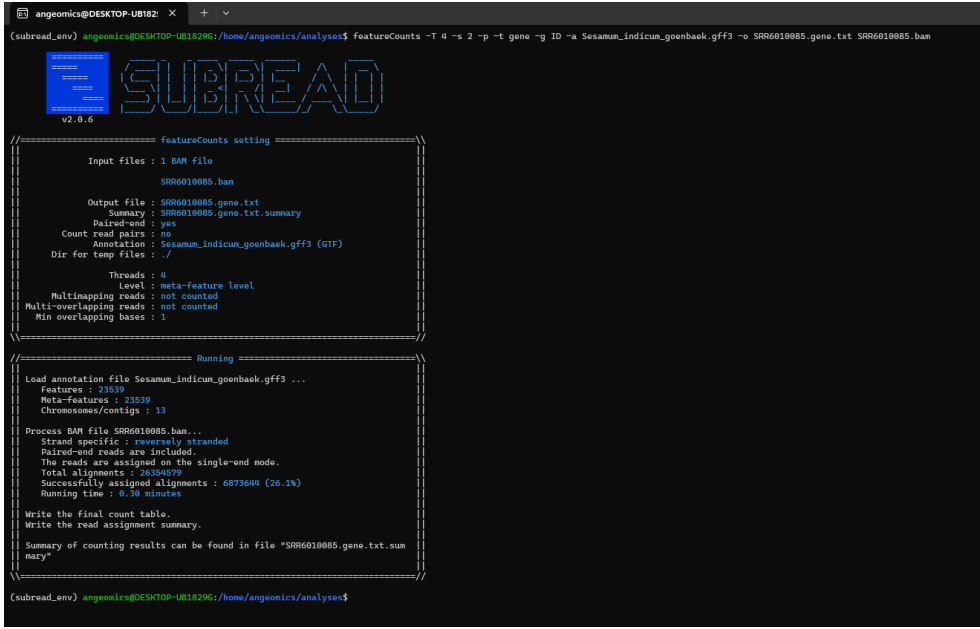
```
angeomics@DESKTOP-UB182!

(samtools_env) angeomics@DESKTOP-UB1829G:/home/angeomics/analyses$ samtools view --threads 8 -bS -o SRR6010085.bam SRR6010085.sam
(samtools_env) angeomics@DESKTOP-UB1829G:/home/angeomics/analyses$ ls -rlth
total 9.9G
-rw-r--r-- 1 angeomics angeomics 255M Feb 22  2022 Sesamum_indicum_goenbaek.fasta
-rw-r--r-- 1 angeomics angeomics  28M Mar 14  2022 Sesamum_indicum_goenbaek.gff3
-rw-r--r-- 1 angeomics angeomics  63M Oct 30 21:44 Sesamum_indicum_goenbaek.4.ht2
-rw-r--r-- 1 angeomics angeomics 1.8K Oct 30 21:44 Sesamum_indicum_goenbaek.3.ht2
-rw-r--r-- 1 angeomics angeomics    8 Oct 30 21:44 Sesamum_indicum_goenbaek.8.ht2
-rw-r--r-- 1 angeomics angeomics   12 Oct 30 21:44 Sesamum_indicum_goenbaek.7.ht2
-rw-r--r-- 1 angeomics angeomics  63M Oct 30 21:47 Sesamum_indicum_goenbaek.2.ht2
-rw-r--r-- 1 angeomics angeomics  88M Oct 30 21:47 Sesamum_indicum_goenbaek.1.ht2
-rw-r--r-- 1 angeomics angeomics  64M Oct 30 21:47 Sesamum_indicum_goenbaek.6.ht2
-rw-r--r-- 1 angeomics angeomics 110M Oct 30 21:47 Sesamum_indicum_goenbaek.5.ht2
-rw-r--r-- 1 angeomics angeomics    0 Oct 30 22:10 Sesamum_indicum_goenbaek
-rw-r--r-- 1 angeomics angeomics 7.1G Oct 30 22:23 SRR6010085.sam
-rw-r--r-- 1 angeomics angeomics 2.2G Oct 31 10:57 SRR6010085.bam
(samtools_env) angeomics@DESKTOP-UB1829G:/home/angeomics/analyses$
```

4: Abundance estimation



Abundance count | featuresCounts

```
featureCounts \
-T 4 -s 2 -p -t gene \
-g ID -a Sesamum_indicum_goenbaek.gff3 \
-o SRR6010085.gene.txt SRR6010085.bam
```



Task: Create a loop for the mapping and the abundance estimation stages

Task: Create a loop for the mapping step

```
#!/bin/bash

# Step 1: Build index for HISAT2 (only needs to be done once)
REFERENCE="Sesamum_indicum_goenbaek.fasta"
INDEX_NAME="Sesamum_indicum_goenbaek"

source /home/angeomics/app/miniconda3/bin/activate
source activate hisat2_env

# Check if index files already exist, to avoid rebuilding
if [ ! -f "${INDEX_NAME}.1.ht2" ]; then
  echo "Building HISAT2 index..."
  hisat2-build "$REFERENCE" "$INDEX_NAME"
  echo "Index building completed."
else
  echo "Index files found, skipping index building."
fi

# Step 2: Mapping and conversion loop
THREADS=8
FASTQ_DIR="/home/angeomics/data/01.SRA/"

for FILE in ${FASTQ_DIR}/*_1.fastq; do
  # Get the base name (e.g., SRR6010085) from the FASTQ file name
  SAMPLE_NAME=$(basename "$FILE" _1.fastq)

  # Define paths for paired-end FASTQ files
  FASTQ1="${FASTQ_DIR}/${SAMPLE_NAME}_1.fastq"
  FASTQ2="${FASTQ_DIR}/${SAMPLE_NAME}_2.fastq"

  # Define output SAM and BAM file names
  SAM_FILE="${SAMPLE_NAME}.sam"
  BAM_FILE="${SAMPLE_NAME}.bam"

  # Run HISAT2 for mapping
  source /home/angeomics/app/miniconda3/bin/activate
  source activate hisat2_env
  echo "Mapping reads for $SAMPLE_NAME..."
  hisat2 -p $THREADS -x "$INDEX_NAME" -1 "$FASTQ1" -2 "$FASTQ2" -S "$SAM_FILE"
  echo "Mapping completed for $SAMPLE_NAME."

  conda deactivate

  # Convert SAM to BAM using samtools

  source activate samtools_env

  echo "Converting $SAM_FILE to BAM format..."
  samtools view --threads $THREADS -bS -o "$BAM_FILE" "$SAM_FILE"

  conda deactivate

  # Optionally, remove the SAM file to save space
  # rm "$SAM_FILE"

  echo "BAM conversion completed for $SAMPLE_NAME."
done
```

Task: Create a loop for the abundance estimation stage

```
#!/bin/bash

# Define input GFF3 annotation file and set options for featureCounts
GFF3="Sesamum_indicum_goenbaek.gff3"
THREADS=4

# Activate subread_env before running

# source /home/angeomics/app/miniconda3/bin/activate
# conda activate subread_env

# Loop through each BAM file in the current directory
for BAM_FILE in *.bam; do
  # Extract the base name of the BAM file (e.g., SRR6010085 from SRR6010085.bam)
  SAMPLE_NAME=$(basename "$BAM_FILE" .bam)

  # Run featureCounts for each BAM file
  featureCounts \
    -T $THREADS \
    -s 2 \
    -p \
    -t gene \
    -g ID \
    -a "$GFF3" \
    -o "${SAMPLE_NAME}.gene.txt" \
    "$BAM_FILE"

  echo "Abundance estimation completed for $BAM_FILE"
done
```

Make a table of gene count

```
paste <(awk 'BEGIN {OFS="\t"} {print $1,$7}' SRR6010085.gene.txt)
\
  <(awk 'BEGIN {OFS="\t"} {print $7}' SRR6010086.gene.txt) \
  <(awk 'BEGIN {OFS="\t"} {print $7}' SRR6010087.gene.txt) \
  <(awk 'BEGIN {OFS="\t"} {print $7}' SRR6010088.gene.txt) \
  <(awk 'BEGIN {OFS="\t"} {print $7}' SRR6010089.gene.txt) \
  <(awk 'BEGIN {OFS="\t"} {print $7}' SRR6010090.gene.txt) | \
  grep -v '^\#' > sesame_count.txt


# Convert into csv format with comma separation

awk -v OFS=',' '{$1=$1}1' sesame_count.txt > sesame_count.csv
```

5: DEG analysis with DESeq2 Package

Task: Install DESeq2 and pheatmap packages

```
# Install DESeq2

if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("DESeq2")
```

Task: Install DESeq2 and pheatmap packages

```
# Install pheatmap

install.packages(pheatmap, dependencies =TRUE)
```

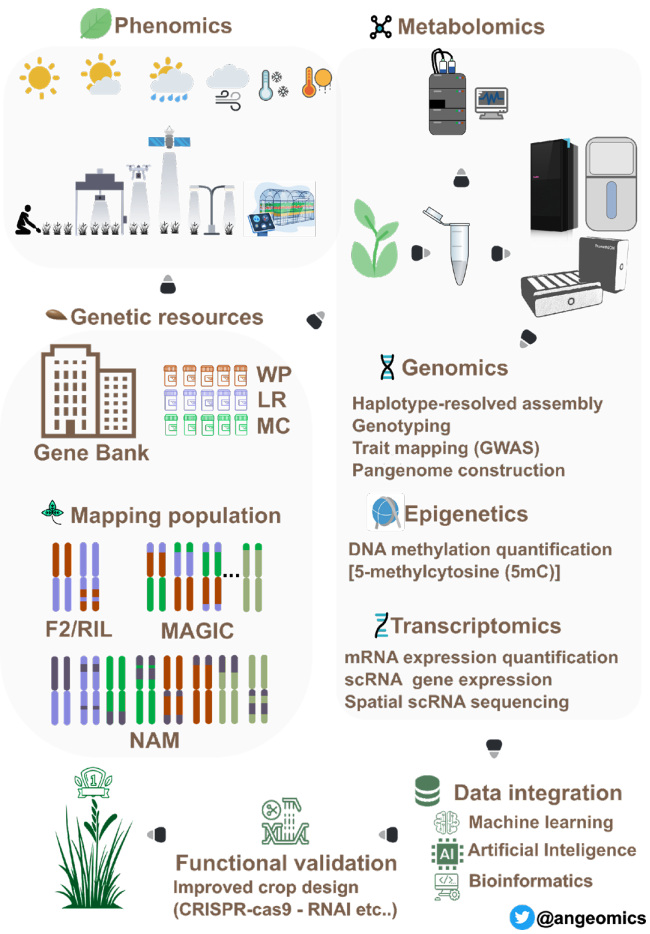
Task: Go through the script and replicate it

https://github.com/Yedomon/GBioS_Training_Genomics_Plant_Breeding_2024/tree/main/Section03/deg_work

But before ... Bonus >>

Bonus

- ❑Omics enables fast-forward breeding for a food-secure world
- ❑Genetic diversity is a paramount
- ❑Big data – Bioinformatics – Machine learning
- ❑Genetic engineering – Gene editing



Thank you