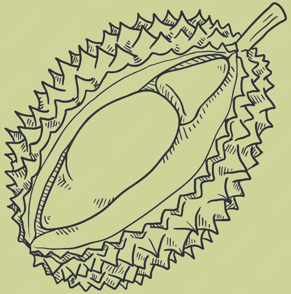
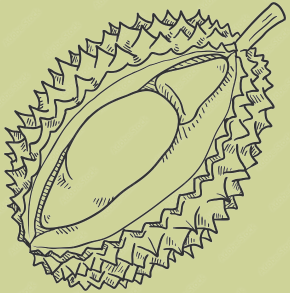
The draft genome of tropical fruit durian (*Durio zibethinus*)



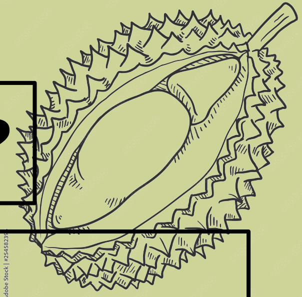
Genome Analysis 2023

Paper IV

Andreas Bergfeldt



What is a Durian and why do this assembly?

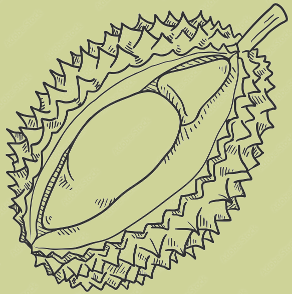


- **What?**
  - Fruit popular in southeast asia
  - Pungent odor, not allowed in some public spaces

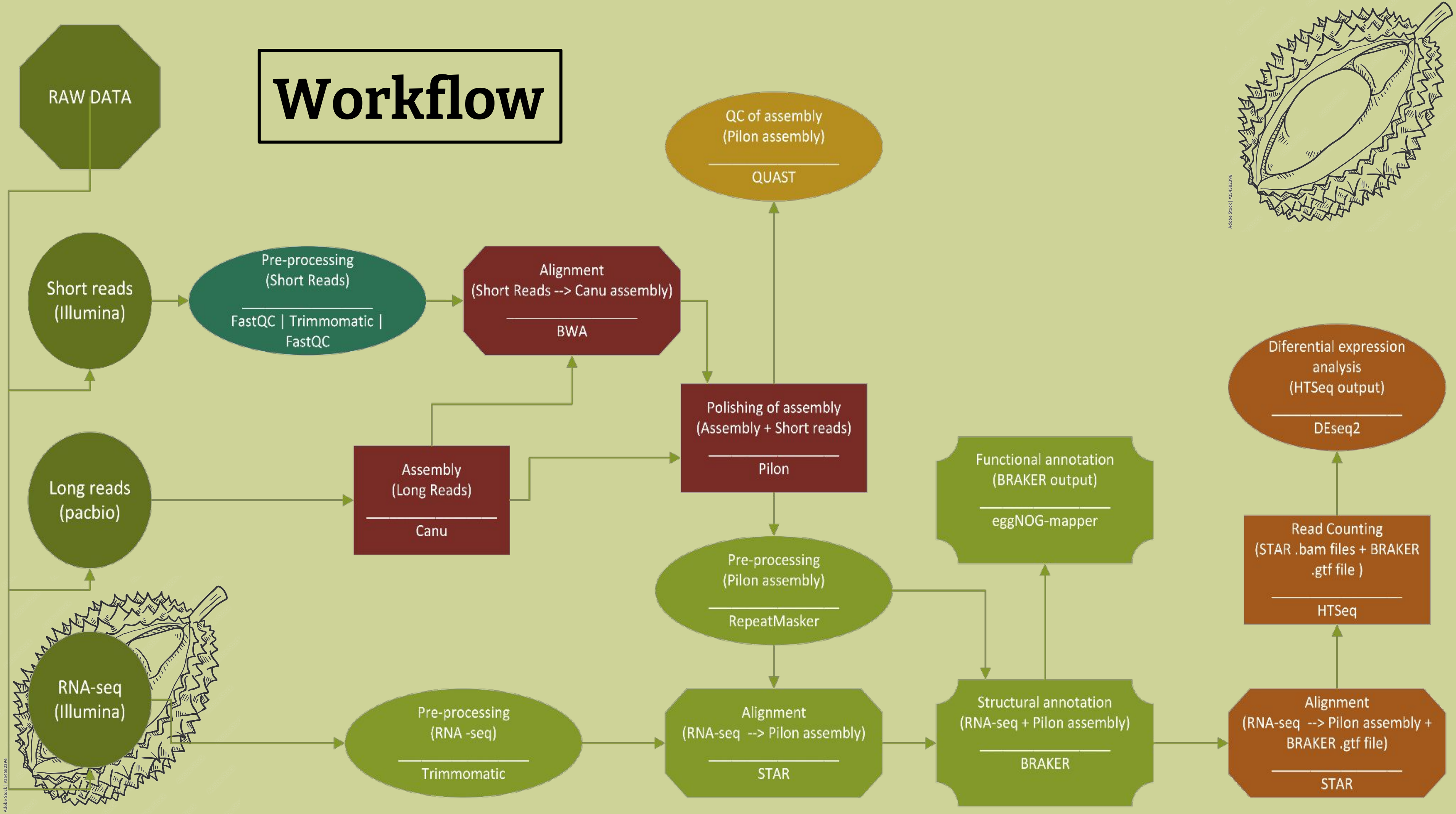
- **Why?**
  - 2016 China imported 600 mUSD
  - Almost non existent genetic research

My assembly

- One scaffold
  - Scaffold 10
- Too much to do whole genome
  - Computing power
  - Time



Workflow



RAW DATA

Short reads (Illumina)

Pre-processing (Short Reads) — FastQC | Trimmomatic | FastQC

Alignment (Short Reads --> Canu assembly) — BWA

Long reads (pacbio)

Assembly (Long Reads) — Canu

Polishing of assembly (Assembly + Short reads) — Pilon

QC of assembly (Pilon assembly) — QUAST

Pre-processing (Pilon assembly) — RepeatMasker

RNA-seq (Illumina)

Pre-processing (RNA -seq) — Trimmomatic

Alignment (RNA-seq --> Pilon assembly) — STAR

Structural annotation (RNA-seq + Pilon assembly) — BRAKER

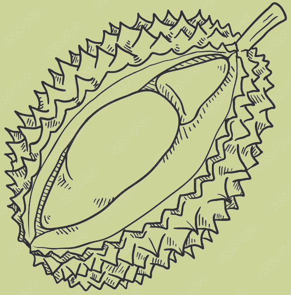
Functional annotation (BRAKER output) — eggNOG-mapper

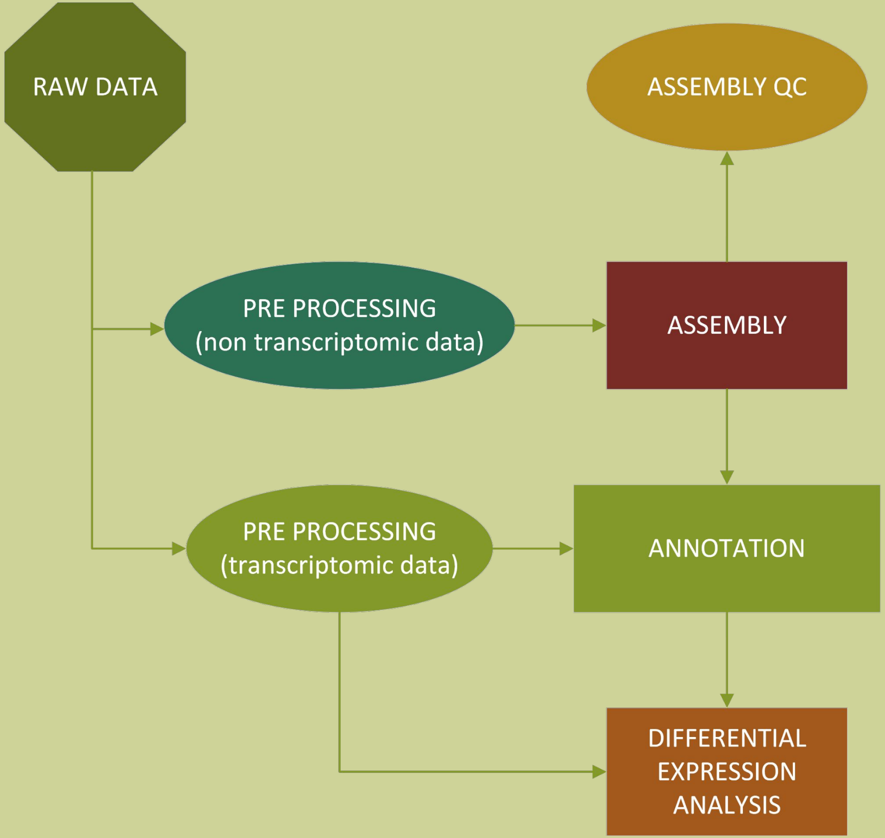
Alignment (RNA-seq --> Pilon assembly + BRAKER .gtf file) — STAR

Read Counting (STAR .bam files + BRAKER .gtf file) — HTSeq

Diferential expression analysis (HTSeq output) — DEseq2

Workflow - simplified





RAW DATA

PRE PROCESSING
(non transcriptomic data)

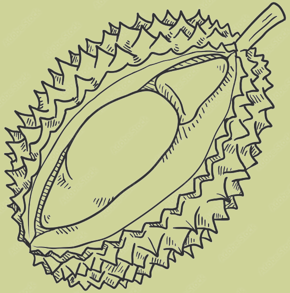
PRE PROCESSING
(transcriptomic data)

ASSEMBLY

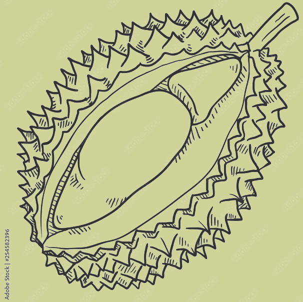
ASSEMBLY QC

ANNOTATION

DIFFERENTIAL
EXPRESSION
ANALYSIS



Pre-processing - Trimmomatic

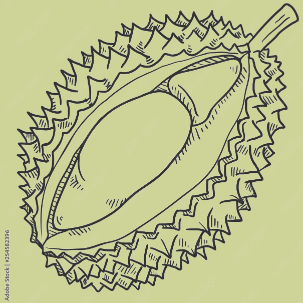


- Pre-trimmed data
    - Trimmed again to make sure it is good
- Trimming to remove adapters and low quality bases

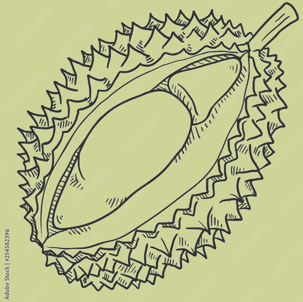
- Quality control was done before and after

Parameters - Trimmomatic

- ILLIUMINACLIP = TruSeq3-SE:2:30:10
- LEADING = 3
- TRAILING = 3
- SLIDINGWINDOW= 4:15
- MINLEN = 36



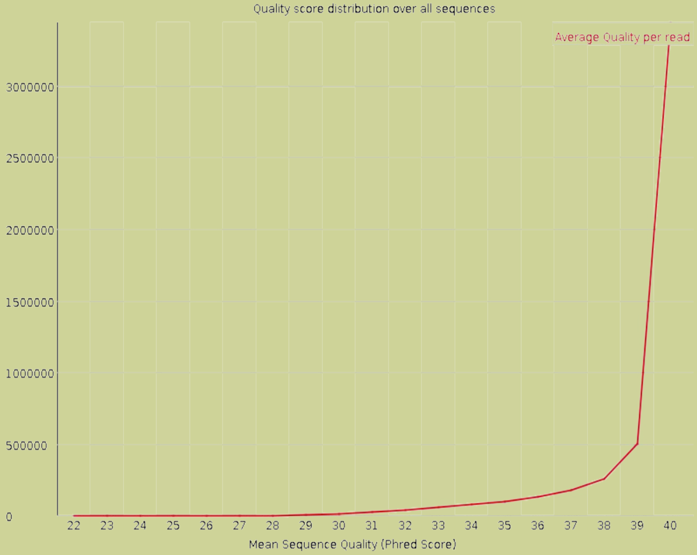
Pre-processing - FastQC

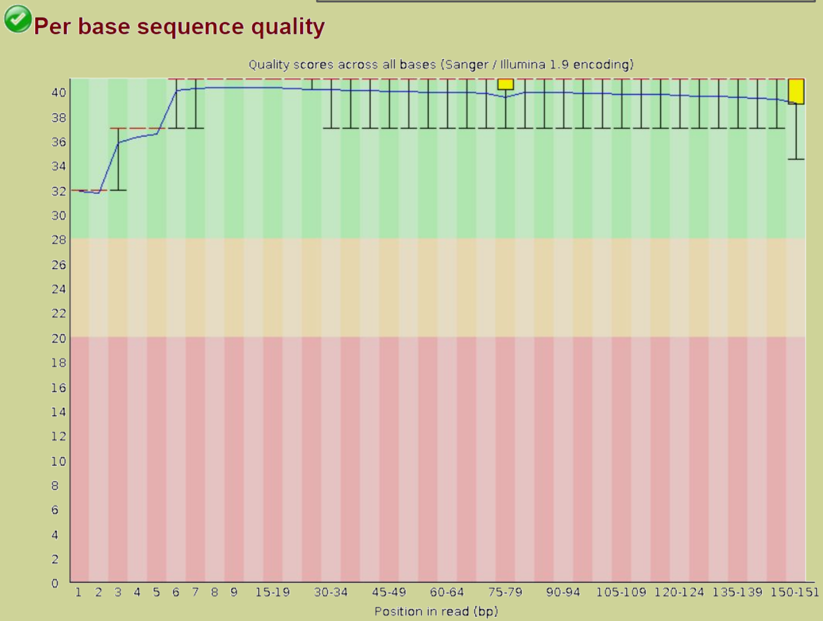


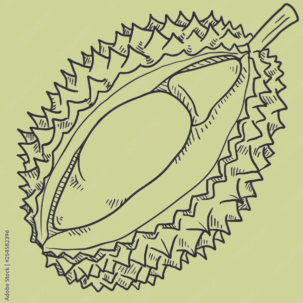
- Important to check data before proceeding!

- Quality deemed good enough

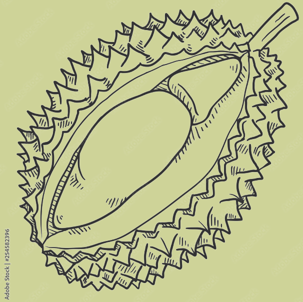
Pictures are from after trimming







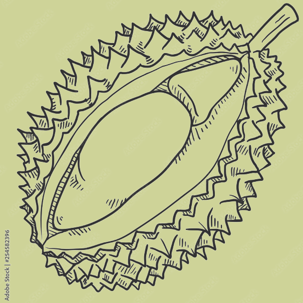
Assembly - Canu



- Assembly of the long reads
- Plant genome with many repeats:
  - corrMaxEvidanceErate parameter
- useGrid for running in UPPMAX

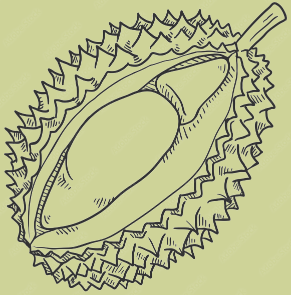
Parameters - Canu

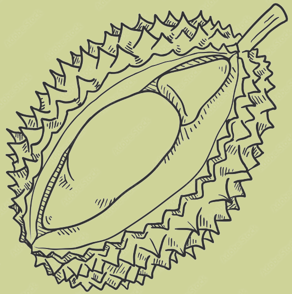
- useGrid = false
- genomeSize = 30m
- corMaxEvidenceErate = 0.15



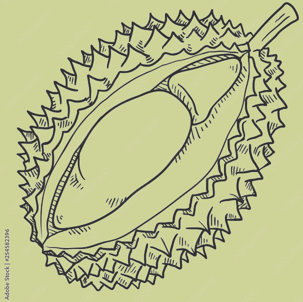
Assembly - BWA (+ samtools)

- **BWA**
- Mapping short reads to long reads
  - Necessary for polishing later
- Manipulating BWA output with samtools to get .bam file
  - .fasta → .sai → .sam → .bam

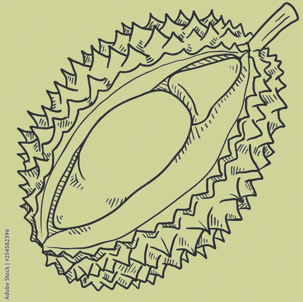




Assembly - Pilon



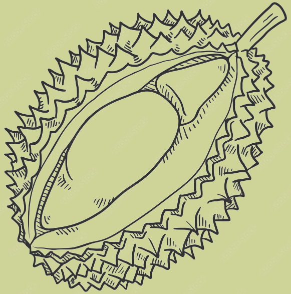
- **Pilon**
- Polishing assembly with short read data
  - Higher accuracy to long scaffolds

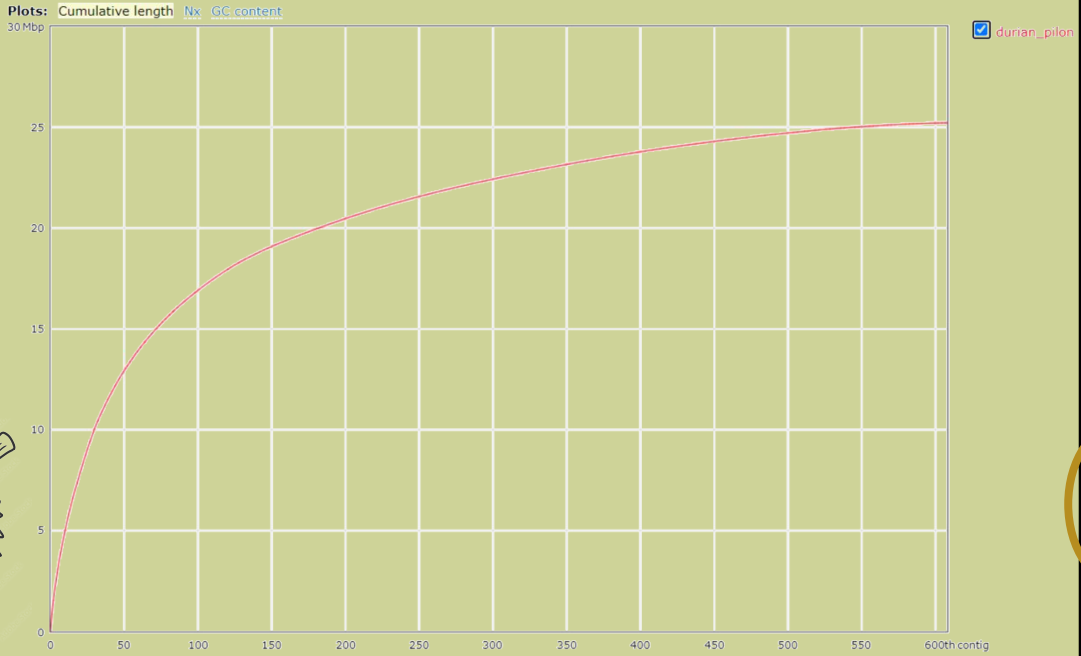


- **Results:** *(No good summary logfile so taken random reads)*
  - Total Reads: 18201
  - Confirmed 177014 of 192342 bases (92.03%)
  - Corrected 11 snps; 0 ambiguous bases
    corrected 96 small insertions totaling 102 bases
    41 small deletions totaling 42 bases

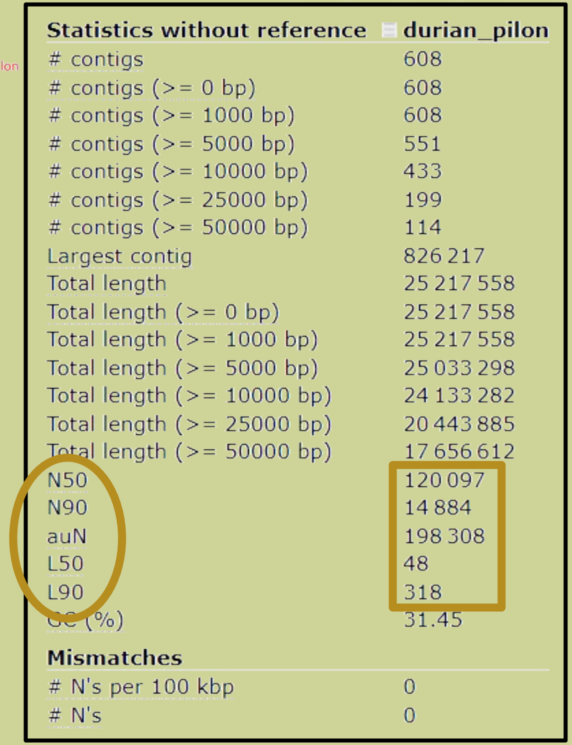
Assembly QC - QUAST

- Was the assembly any good?

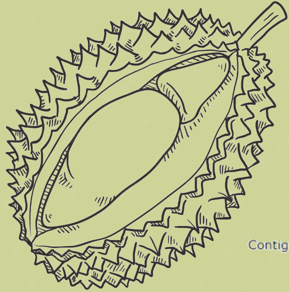




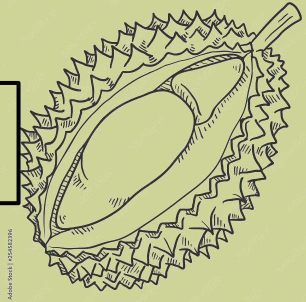
Contigs are ordered from largest (contig #1) to smallest.



Statistics without reference	durian_pilon
# contigs	608
# contigs (>= 0 bp)	608
# contigs (>= 1000 bp)	608
# contigs (>= 5000 bp)	551
# contigs (>= 10000 bp)	433
# contigs (>= 25000 bp)	199
# contigs (>= 50000 bp)	114
Largest contig	826 217
Total length	25 217 558
Total length (>= 0 bp)	25 217 558
Total length (>= 1000 bp)	25 217 558
Total length (>= 5000 bp)	25 033 298
Total length (>= 10000 bp)	24 133 282
Total length (>= 25000 bp)	20 443 885
Total length (>= 50000 bp)	17 656 612
N50	120 097
N90	14 884
auN	198 308
L50	48
L90	318
GC (%)	31.45
Mismatches	
# N's per 100 kbp	0
# N's	0



Annotation - Trimmomatic + repeatMasker

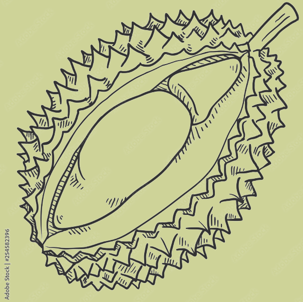


- Pre-processing before annotation
- Trimmomatic → same as before

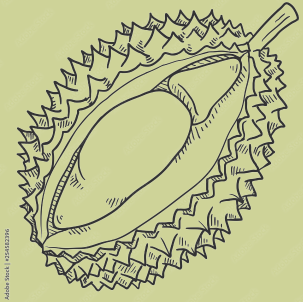
- RepeatMasker
  - softmasking repeats for better annotation
  - Important to specify softmasking

Softmasking

- Identifies and "masks" repeats
- Changes bases in the fasta file from uppercase to lowercase
- Hardmasking removes info, softmasking does not



Structural annotation - BRAKER



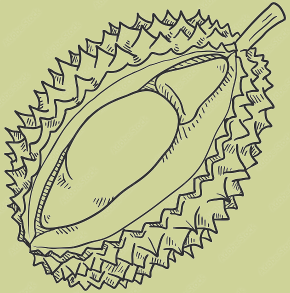
- Pipeline consisting of Augustus and GeneMark
- Annotates based on reference genome (masked assembly), and transcriptomic data

- Gives way to many outputs, still don't know exactly what is what

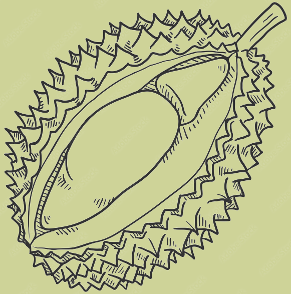
BRAKER results

- Hardmasked scaffold → 96 genes identified
- Softmasked scaffold → 110 genes identified

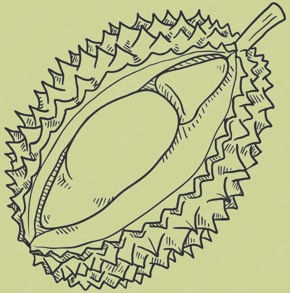
- What are the genes?
  - who knows...
    - will come later



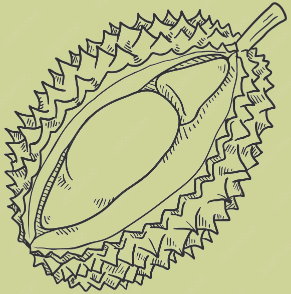
Functional annotation - eggNOG-mapper



- Web based UI
- Loads of information about found hits
- Does not say a lot at this stage
    - Vital for differential expression analysis
        - Yes more patience is needed..



Counting reads - STAR and HTSeq

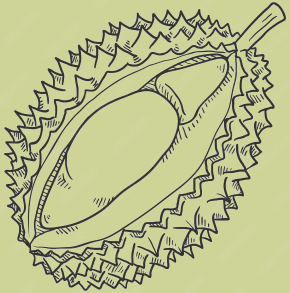


- For transcriptomic reads to be counted they first need to be aligned
    - Done with STAR

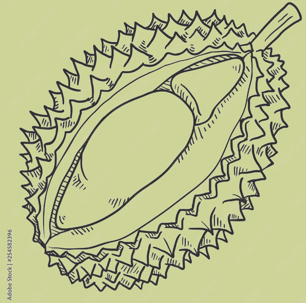
- HTSeq counts the reads that are aligned to each predicted gene (from BRAKER), using a .gtf file

HTSeq results

- 8 files with reads of varying length
- Total 29 269 185 record pairs
- 835 record pairs with missing mate record
    - 0.003% of total record pairs



Differential expression analysis



- DESeq2
  - R-module for expression analysis

- Finally some tangible results!

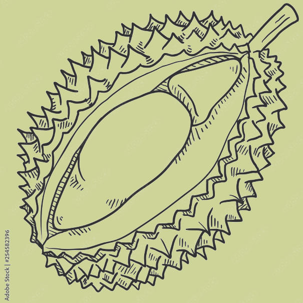
The analysis in short

(In depth on future slides)

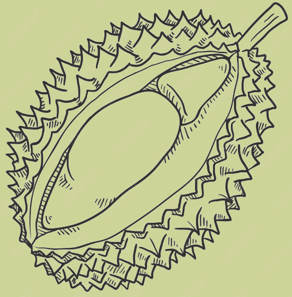
- 2 different species
  - Musang King and Mothong

- 2 different type of analysis
  - Between species *(Musang king, Mothong)*
  - Within species *(Musang King)*
    - Different plant organs *(aril, stem, leaf, root)*

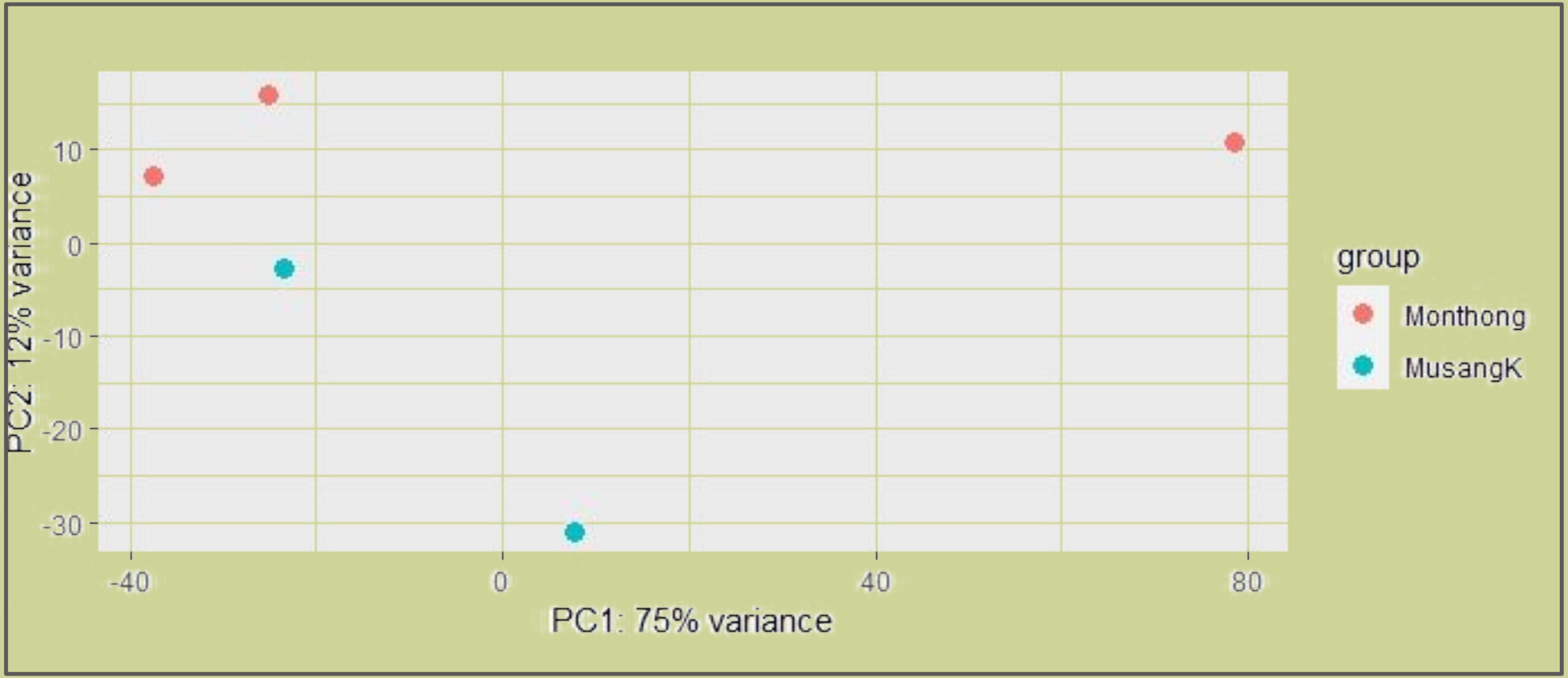
- Types of visualization
  - PCA
  - Heatmap

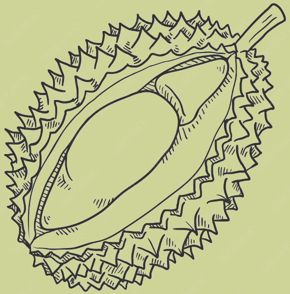


DE - Between species - PCA

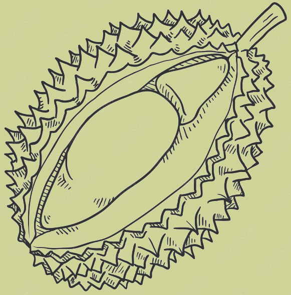


Adobe Stock | #254582396





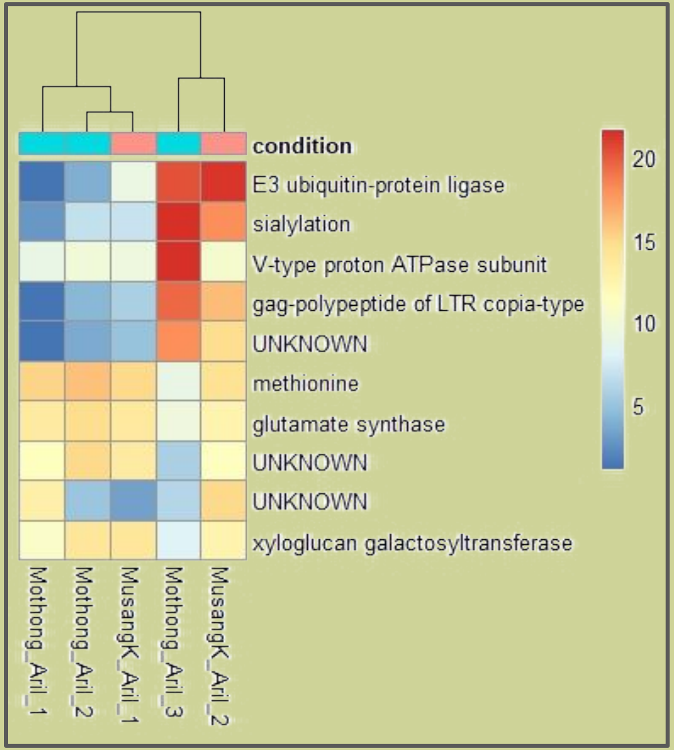
Adobe Stock | #254582396

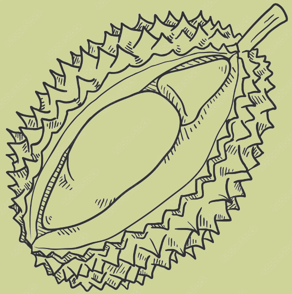


DE - Between species - Heatmap

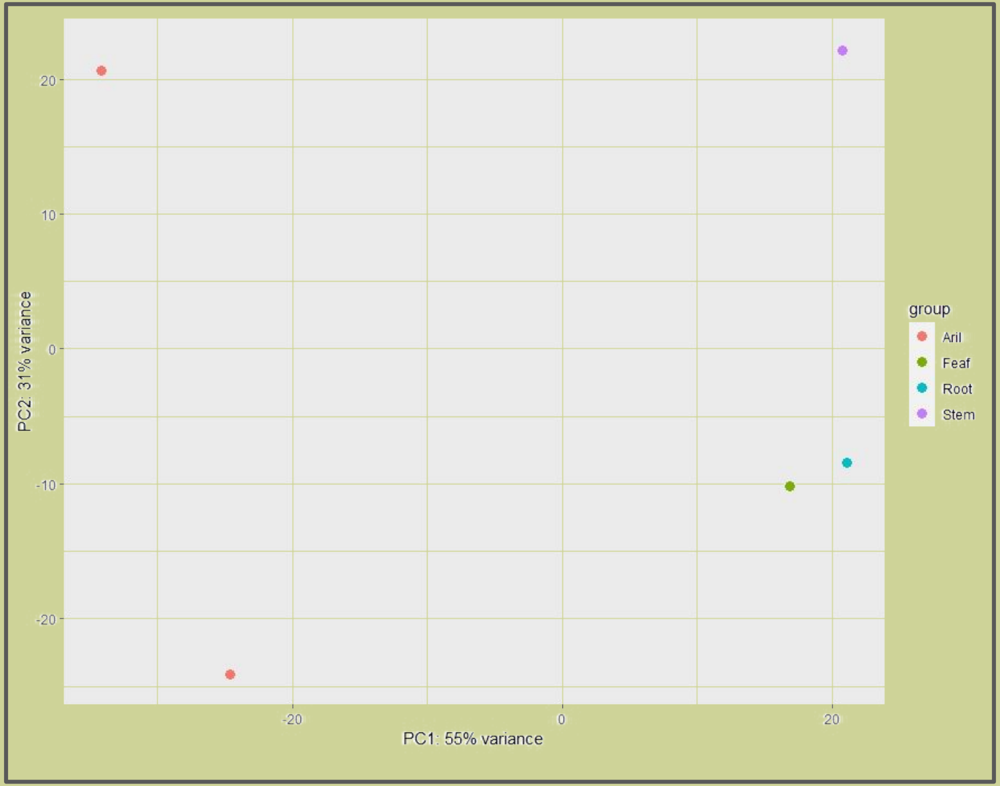
- Overall hard to separate the two species
- On PCA the grouping is not good
  - Groups Musang with Mothong
  - Does not group all of same species

- The heatmap does not show any patterns in which the species can be separated





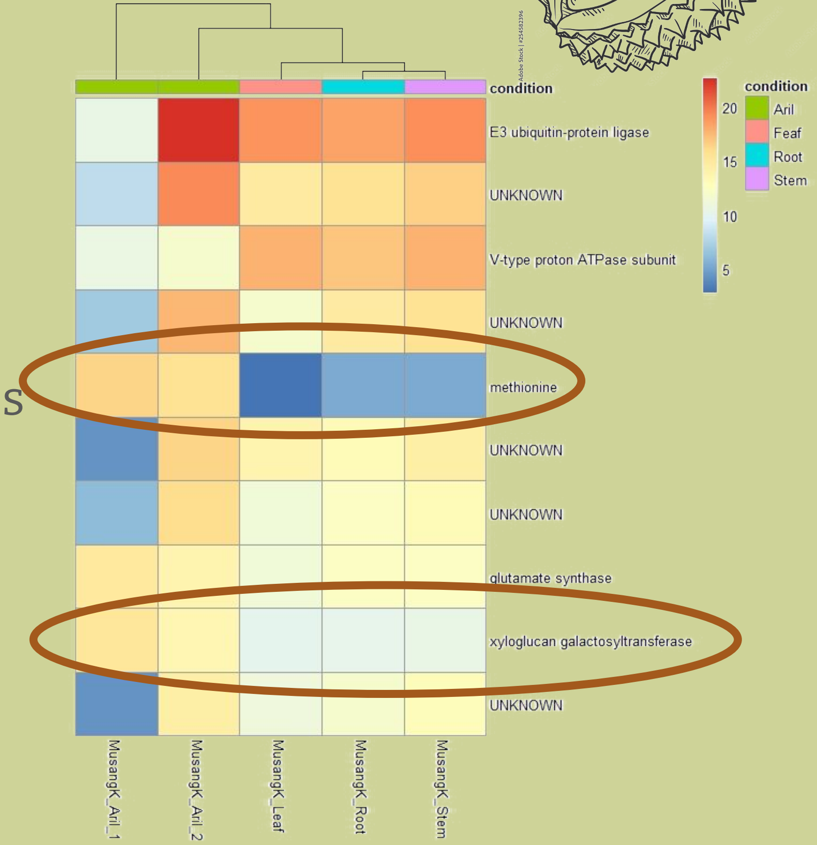
DE - Within species - PCA

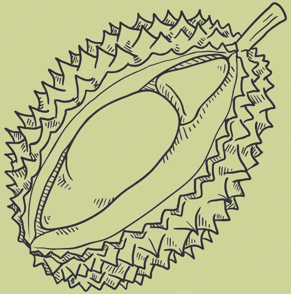


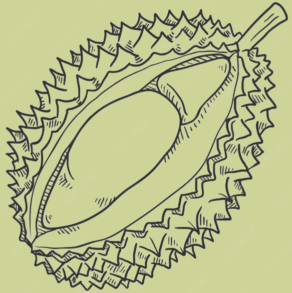
DE - Within species - Heatmap

- Better separation between plant organs compared to plant species
  - PCA did okay
  - Possible separations at least

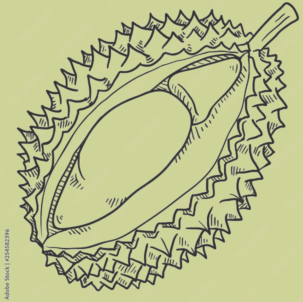
- Some patterns can be seen that separates arils from the rest of the plant organs
  - Eg. methionine and xyloglucan





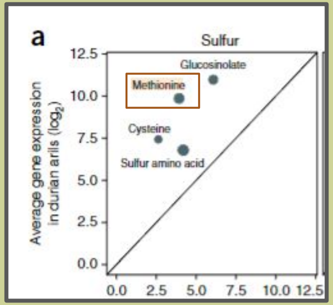


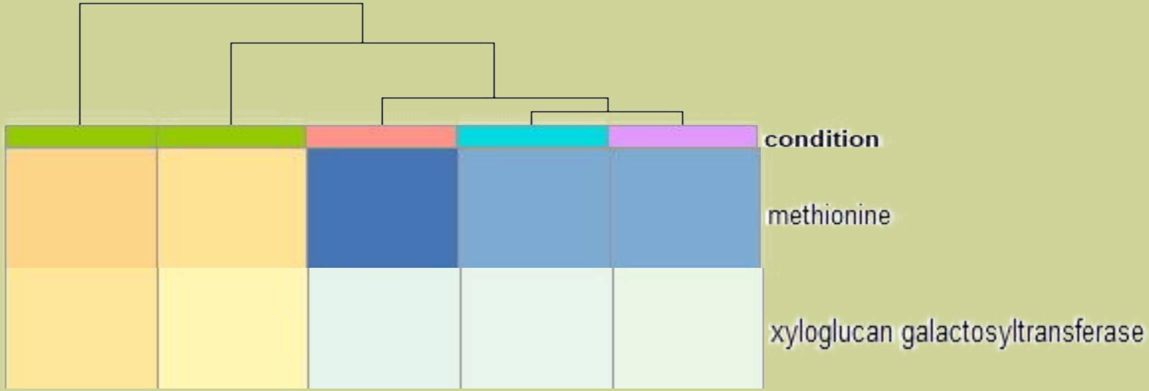
DE - Biological conclusions

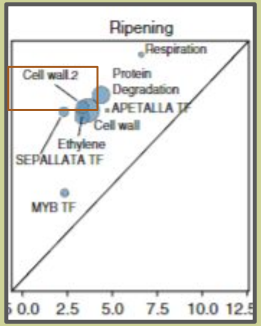


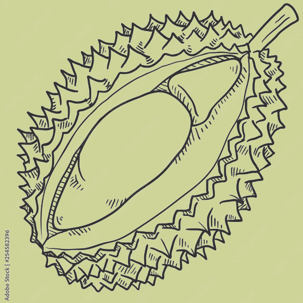
- Methionine converted to methanethiol with the enzyme methionine γ-lyase
  - Methanethiol has key role in odor of durian
- Xyloglucan building block of cell wall 2
  - *“interlace cellulose microfibrils in most flowering plants.”*
  - Genes with association to cell wall 2 upregulated in arils

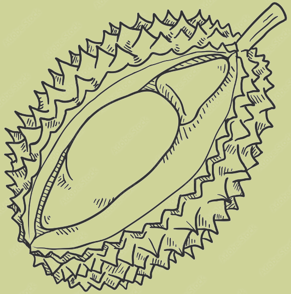
Figures from paper IV











Thank you for your attention!

For more in depth information about programs etc. please look at my github wiki for this project: https://github.com/A-Bergfeldt