University of the Philippines
LOS BAÑOS

CMSC244

HeatSeq: A Comparative Benchmark of Alignment Tools for Tissue/Organ-specific Gene Expression Profiling

- Mercado, Mark Cyril R.
  - MS Molecular Biology and Biotechnology minor in Computer Science student, UPLB Graduate School
  - Graduate Research Associate, IPB, CAFS, UPLB

OUTLINE

01 Introduction

02 Considerations

03 Alignment Methods

04 Major Results

INTRODUCTION

Basic Molecular Biology: The Central Dogma of Molecular Biology

- Point of Interest: Transcription or expression of genes
    - Not all genes are expressed at the same time.
    - Different organ or developmental stages have different set of genes

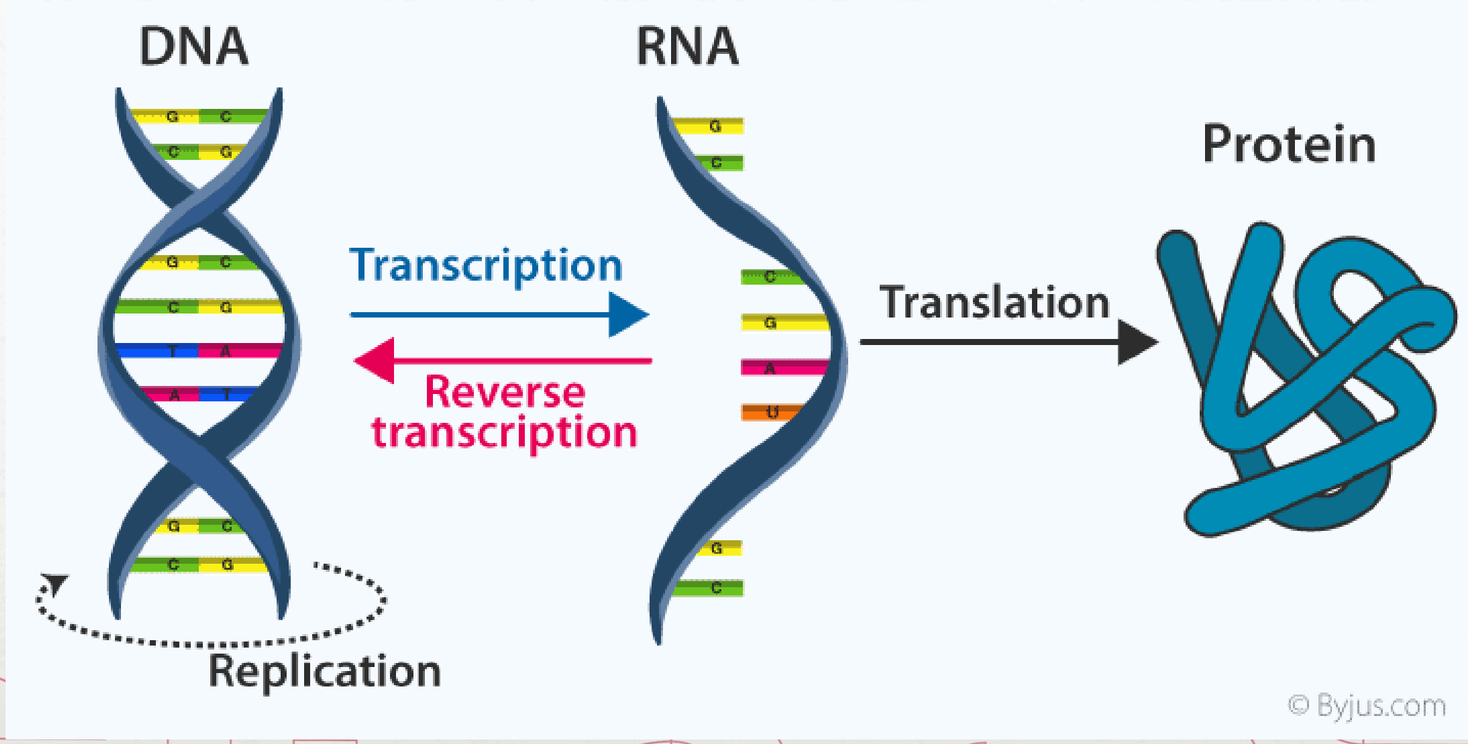


Figure 1. The Central Dogma of Molecular Biology. (Lifted from https://cdn1.byjus.com/wp-content/uploads/2018/11/Central-Dogma-DNA-to-RNA-to-Protein.png

INTRODUCTION

Basic Molecular Biology: Alternative Splicing

- Sometimes genes have alternative splicing
    - Some part of the raw genes are not expressed (introns)
    - while some are expressed (exons)
- e.g. *Dscam* (Down syndrome cell adhesion molecule) gene in *Drosophila melanogaster*
    - 30,000 version of protein
        - from a single-gene

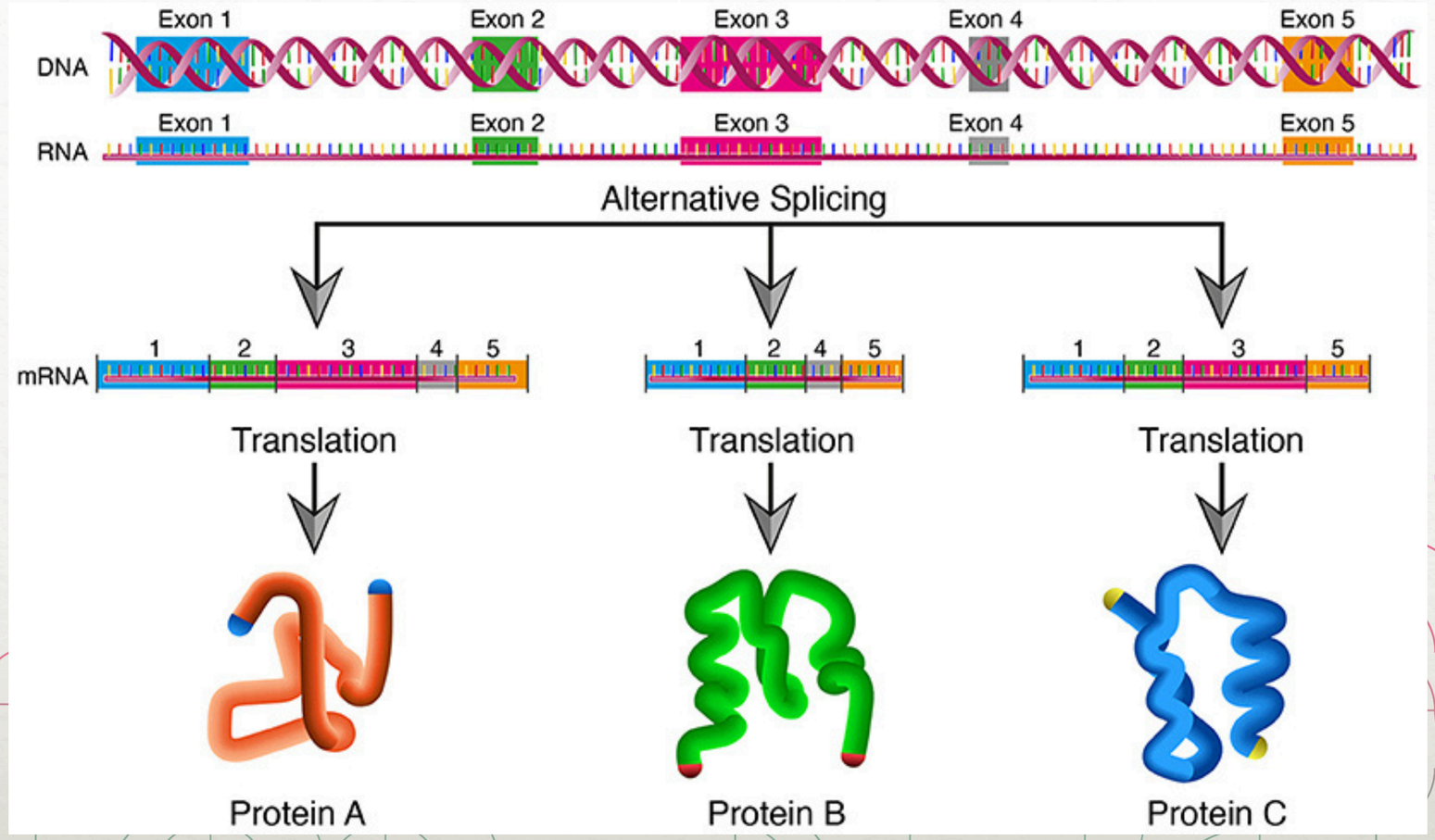


Figure 2. Demo of Alternative Splicing. (Lifted from https://www.frontiersin.org/files/Articles/1063940/frym-11-1063940-HTML-r2/image_m/figure-3.jpg

Research Questions

- Which of the 8 *SmelDMPs* are expressed in the buds and open buds?
- Which of the SmelGRFs and SmelGIFs are expressed in meristematic region?
- And how could RNA-pipeline be optimized and reduced in size given the heaviness of RNA-seq pipeline

Subject ID	Query ID	E-value	Percent Identity
SmelDMP02	Solyc02	0	93.333
	CaDMP6-like	0	92.381
	CsDMP4	2.01E-77	76.548
SmelDMP01.990	CaDMP2	0	93.11
	Solyc01.490	0	92.762
	PtDMP2	3.16E-30	76.494
SmelDMP12	Solyc10.200	0	92.61
	CsDMP2-like	1.26E-59	75.187
SmelDMP04	Solyc10.970	0	90.769
	Solyc02	9.22E-81	76.796
SmelDMP01.730	CaDMP3-like	0	88.613
	Solyc01.580	0	87.888
SmelDMP10.550	CaDMP7-like	7.87E-177	84.762
	Solyc10.200	7.59E-111	88.685
SmelDMP10.560	CaDMP7-like	7.87E-177	84.762
	SlDMP7-like	7.59E-111	88.685
SmelDMP10.200	Solyc05	0	88.253
	NtDMP	0	87.818
	CaDMP9-like	0	87.663
	AtDMP9	9.10E-22	75.728
	AtDMP8	1.17E-20	75.743

Table 1. *SmelDMPs* BLAST Gene Identification.

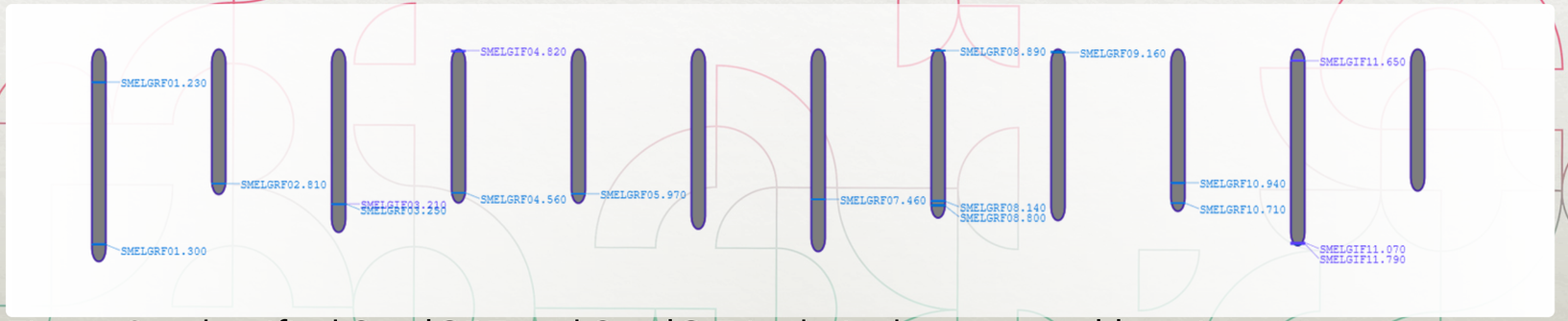


Figure 3. Identified SmelGRF and SmelGIF in their chromosomal location.

RNA-seq data analysis workflow

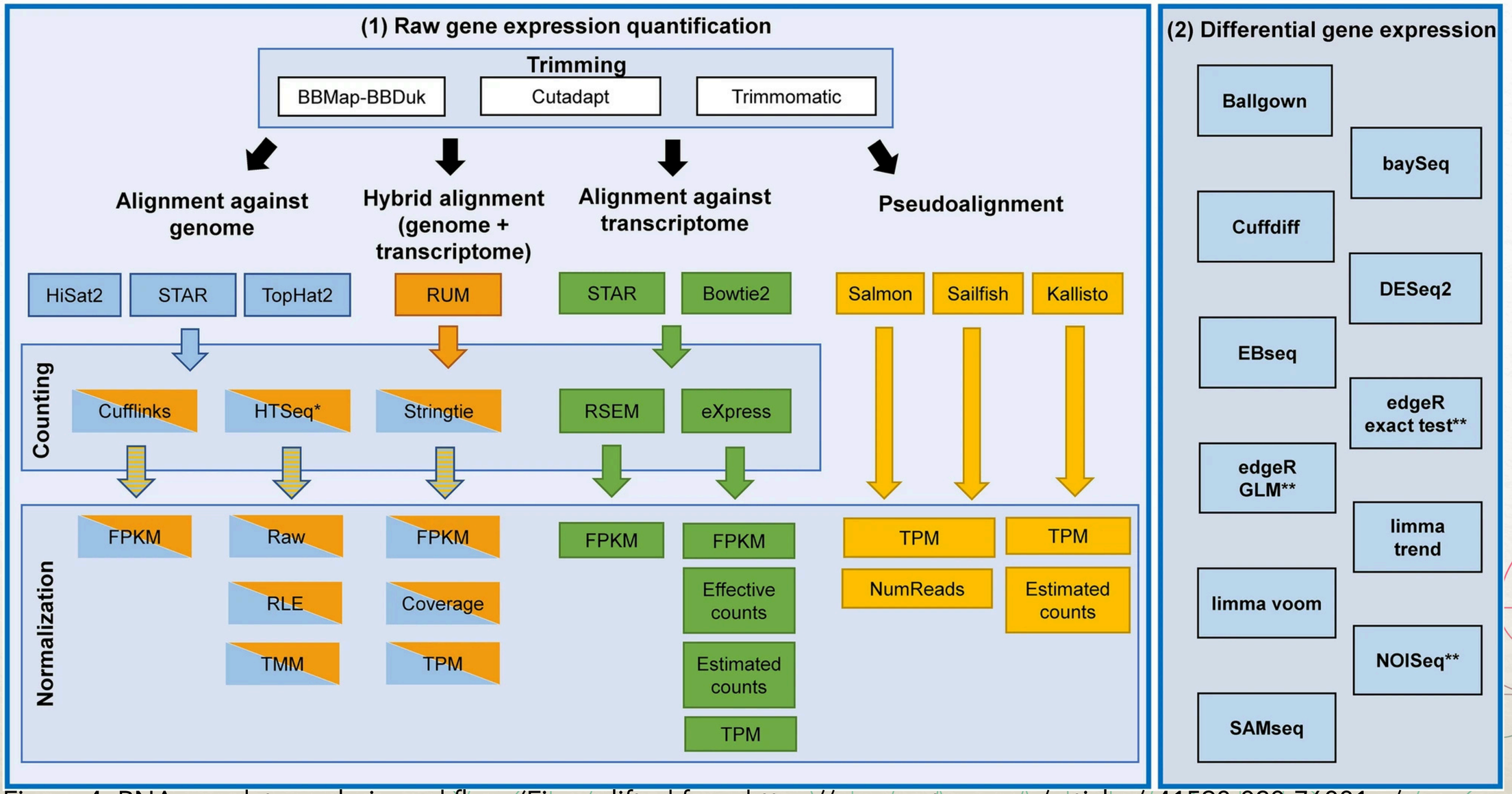
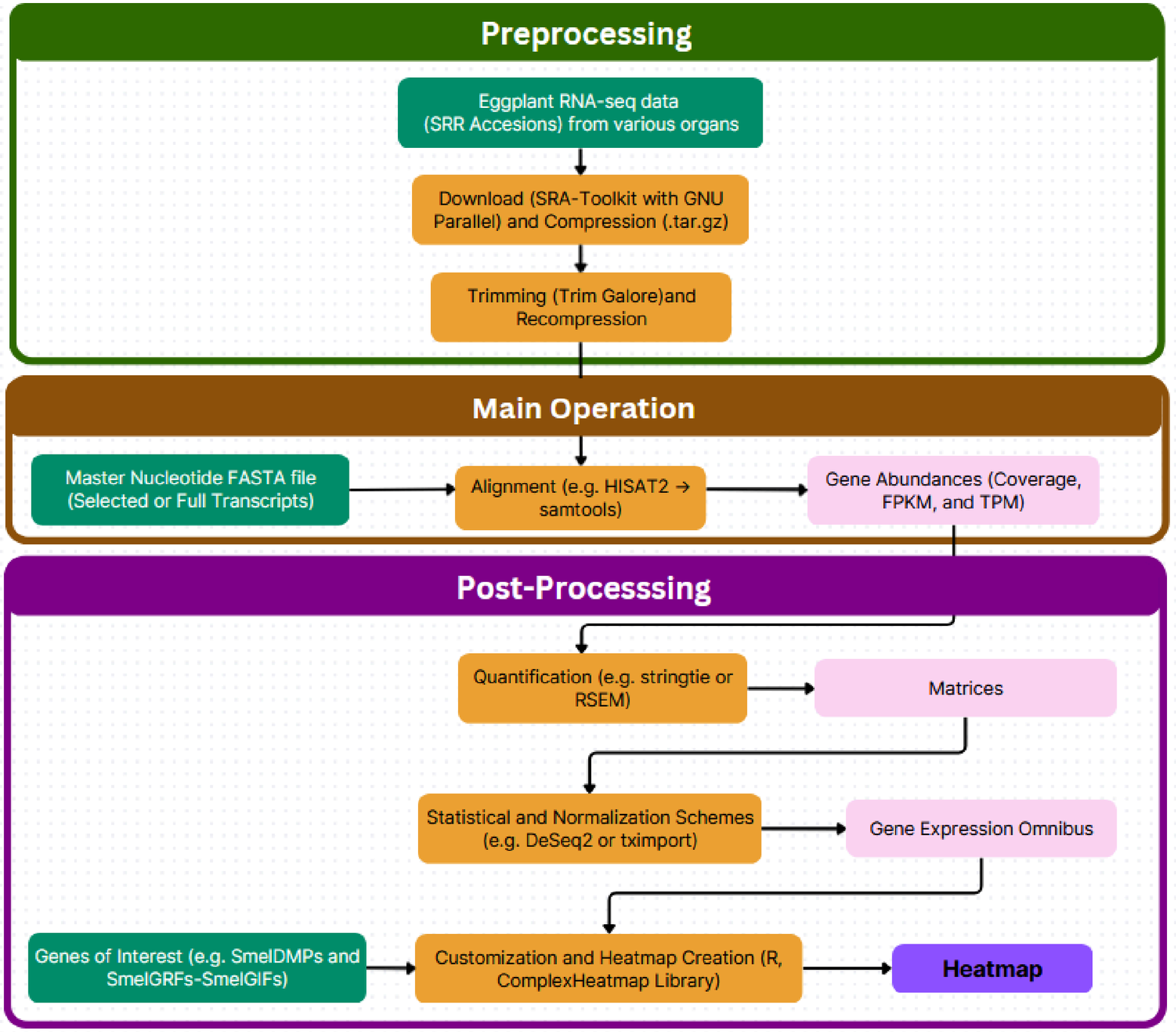


Figure 4. RNA-seq data analysis workflow. (Figure lifted from https://www.nature.com/articles/s41598-020-76881-x/figures/1

Overview of the Pipeline Used

- Preprocessing of the Inputs
- Alignment
- Post-processing of the results



CONSIDERATION

Space Complexity

- Compressed and trimmed files:
  - A single SRR :
    - ~ 2 GB per accession
  - x 2: for Forward and Reverse Read
  - x 12 Organs
  - Minimum of 3 replicate
  - All in all, its around:
    - 145 GB for trimmed inputs

Select

	Runs	Bytes	Bases	Download
Total	190	352.29 Gb	580.16 G	Metadata or Accession List
Selected	0	0	0	Metadata or Accession List or JWT Cart

Found 20 Items | Search within results | Clear

	Run	BioSample	Assay Type	AvgSpotLen	Bases	Bytes	Experiment	Library Name	LibrarySelectie	tissue
1	SRR3884678	SAMN05375150	RNA-Seq	202	5.51 G	3.21 Gb	SRX1940359	Sample_27T	cDNA	Fruits peduncle
2	SRR3884642	SAMN05375145	RNA-Seq	202	5.26 G	3.16 Gb	SRX1940323	Sample_17	cDNA	Fruits Skin Stage 2
3	SRR3884681	SAMN05375153	RNA-Seq	202	5.19 G	3.14 Gb	SRX1940362	Sample_33T	cDNA	Fruits Flesh Stage 3
4	SRR3884608	SAMN05375142	RNA-Seq	202	5.05 G	3.00 Gb	SRX1940289	Sample_13T	cDNA	Fruits ~ 1 cm
5	SRR3884675	SAMN05375148	RNA-Seq	202	4.36 G	2.64 Gb	SRX1940356	Sample_1B	cDNA	Roots
6	SRR3884653	SAMN05375146	RNA-Seq	202	4.40 G	2.63 Gb	SRX1940334	Sample_18T	cDNA	Fruits Flesh Stage 2
7	SRR3884680	SAMN05375152	RNA-Seq	202	4.19 G	2.57 Gb	SRX1940361	Sample_32T	cDNA	Fruits Skin Stage 3
8	SRR3884631	SAMN05375144	RNA-Seq	202	4.24 G	2.55 Gb	SRX1940312	Sample_15T	cDNA	Fruits ~ 6 cm
9	SRR3884677	SAMN05375149	RNA-Seq	202	4.08 G	2.37 Gb	SRX1940358	Sample_22T	cDNA	Cotyledons
10	SRR3884685	SAMN05375156	RNA-Seq	202	3.97 G	2.35 Gb	SRX1940365	Sample_38	cDNA	Radicles
11	SRR3884620	SAMN05375143	RNA-Seq	202	3.86 G	2.33 Gb	SRX1940301	Sample_14	cDNA	Fruits Stage 1
12	SRR3884597	SAMN05375141	RNA-Seq	202	3.72 G	2.23 Gb	SRX1940278	Sample_12	cDNA	Flowers

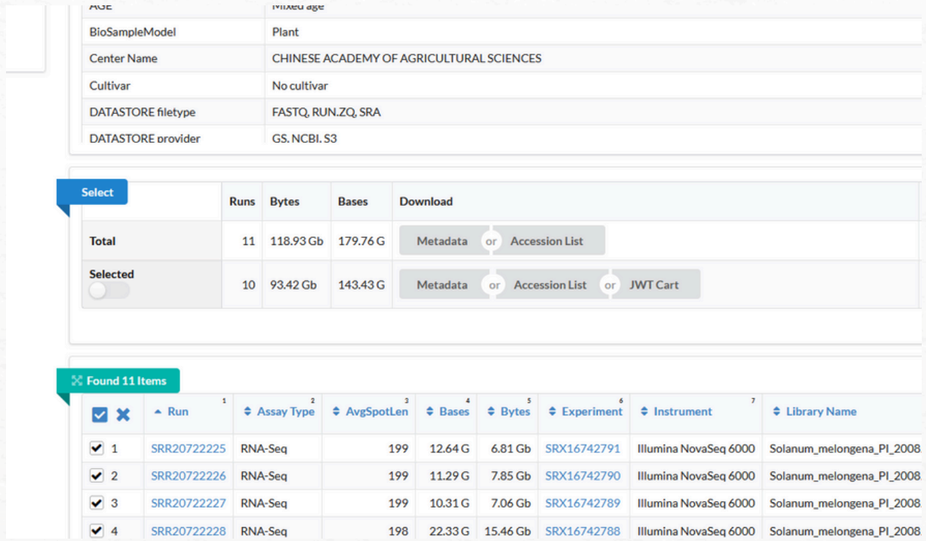
Figure 6. Raw Size of the SRR accessions. (Lifted from NCBI SRA Selector)

CONSIDERATION

Space Complexity

- Second Dataset:
  - Full: 143 GB
    - 6 organs
    - ~ 7 GB per SRR

- Third Dataset:
  - Full: 88 GB
    - 7 organs
    - ~ 3 GB per SRR



BioSampleModel	Plant
Center Name	CHINESE ACADEMY OF AGRICULTURAL SCIENCES
Cultivar	No cultivar
DATASTORE filetype	FASTQ, RUN.ZQ, SRA
DATASTORE provider	GS, NCBI, S3

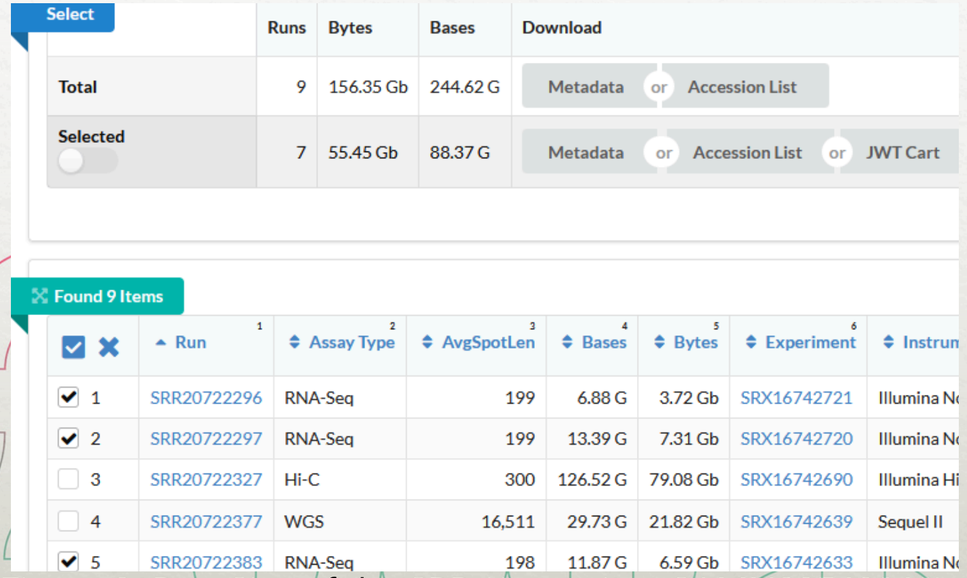
Select

	Runs	Bytes	Bases	Download
Total	11	118.93 Gb	179.76 G	Metadata or Accession List
Selected	10	93.42 Gb	143.43 G	Metadata or Accession List or JWT Cart

Found 11 Items

	Run	Assay Type	AvgSpotLen	Bases	Bytes	Experiment	Instrument	Library Name
1	SRR20722225	RNA-Seq	199	12.64 G	6.81 Gb	SRX16742791	Illumina NovaSeq 6000	Solanum_melongena_PI_2008
2	SRR20722226	RNA-Seq	199	11.29 G	7.85 Gb	SRX16742790	Illumina NovaSeq 6000	Solanum_melongena_PI_2008
3	SRR20722227	RNA-Seq	199	10.31 G	7.06 Gb	SRX16742789	Illumina NovaSeq 6000	Solanum_melongena_PI_2008
4	SRR20722228	RNA-Seq	198	22.33 G	15.46 Gb	SRX16742788	Illumina NovaSeq 6000	Solanum_melongena_PI_2008

Figure 7. Raw Size of the SRR (Accession: SAMN28540077). (Lifted from NCBI SRA Selector)



Select

	Runs	Bytes	Bases	Download
Total	9	156.35 Gb	244.62 G	Metadata or Accession List
Selected	7	55.45 Gb	88.37 G	Metadata or Accession List or JWT Cart

Found 9 Items

	Run	Assay Type	AvgSpotLen	Bases	Bytes	Experiment	Instrun
1	SRR20722296	RNA-Seq	199	6.88 G	3.72 Gb	SRX16742721	Illumina N
2	SRR20722297	RNA-Seq	199	13.39 G	7.31 Gb	SRX16742720	Illumina N
3	SRR20722327	Hi-C	300	126.52 G	79.08 Gb	SRX16742690	Illumina Hi
4	SRR20722377	WGS	16,511	29.73 G	21.82 Gb	SRX16742639	Sequel II
5	SRR20722383	RNA-Seq	198	11.87 G	6.59 Gb	SRX16742633	Illumina N

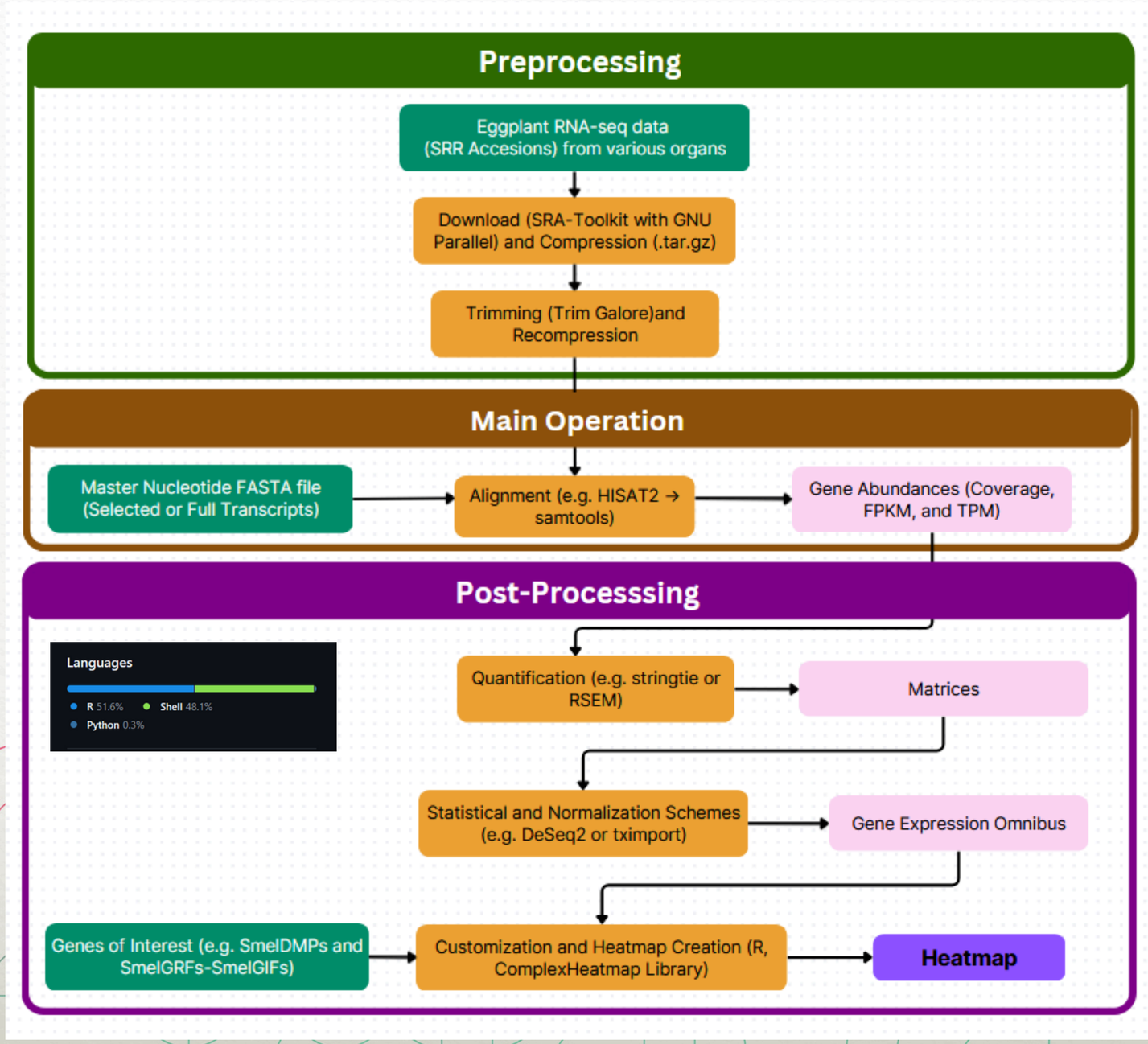
Figure 8. Raw Size of the SRR (Accession: SAMN28540068). (Lifted from NCBI SRA Selector)

CONSIDERATION

Strategies Employed to reduced Space

- Compressing the downloaded and trimmed SRR files.
- Deleting the SRR files once it is trimmed.

- Deleting sam and bam alignment files after processing (one of the big files that contributes to space complexity)
    - If not deleted, the size will increases to x 3.



CONSIDERATION

Time Complexity

- Size of the input data (RNA-seq data (SRR files))

- Alternative Splicing of the Genes (Intron and Exons)
    - e.g. one of the SmelDMP genes that we are interested in have introns or are splice

- And the alignment method.

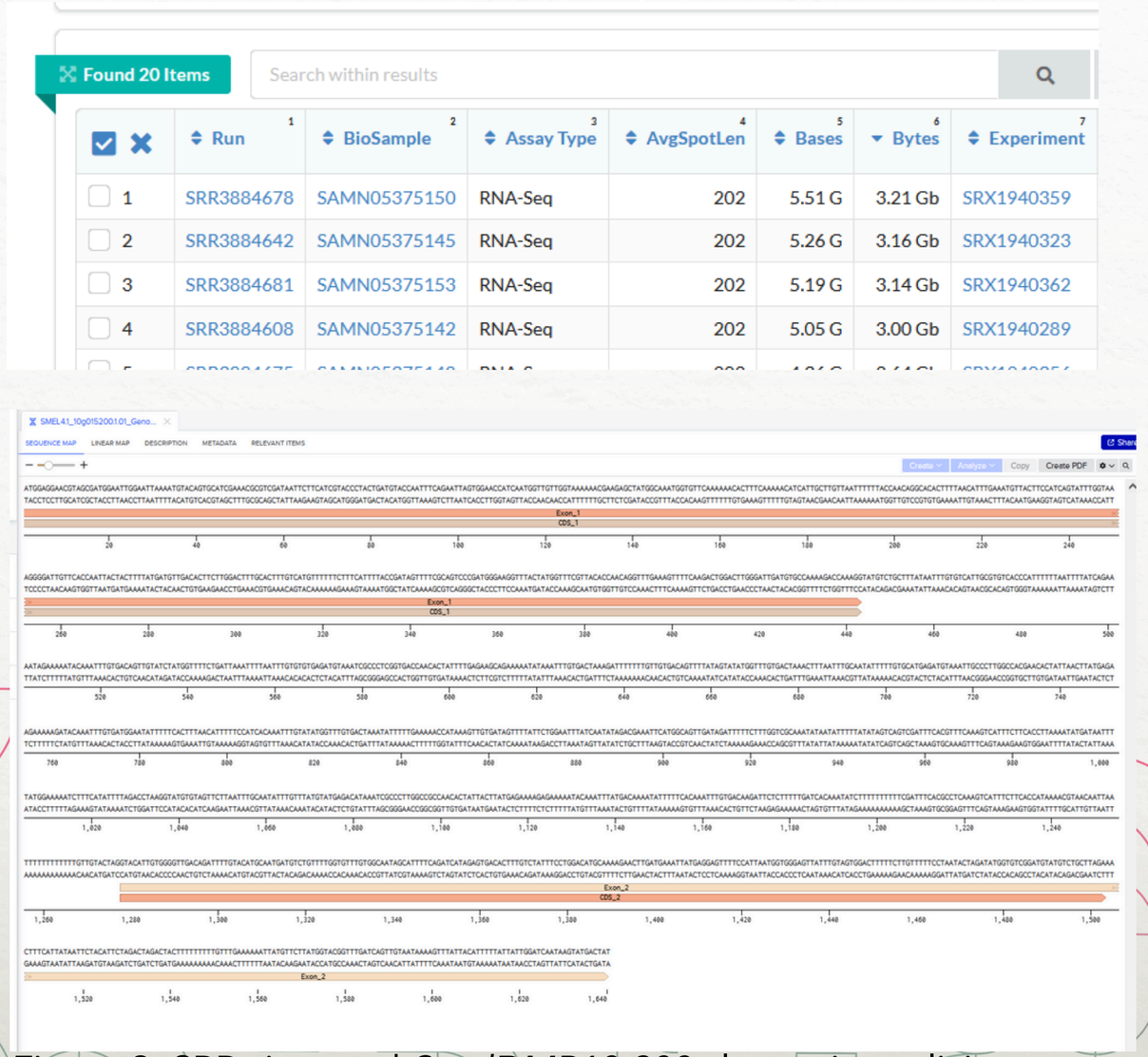


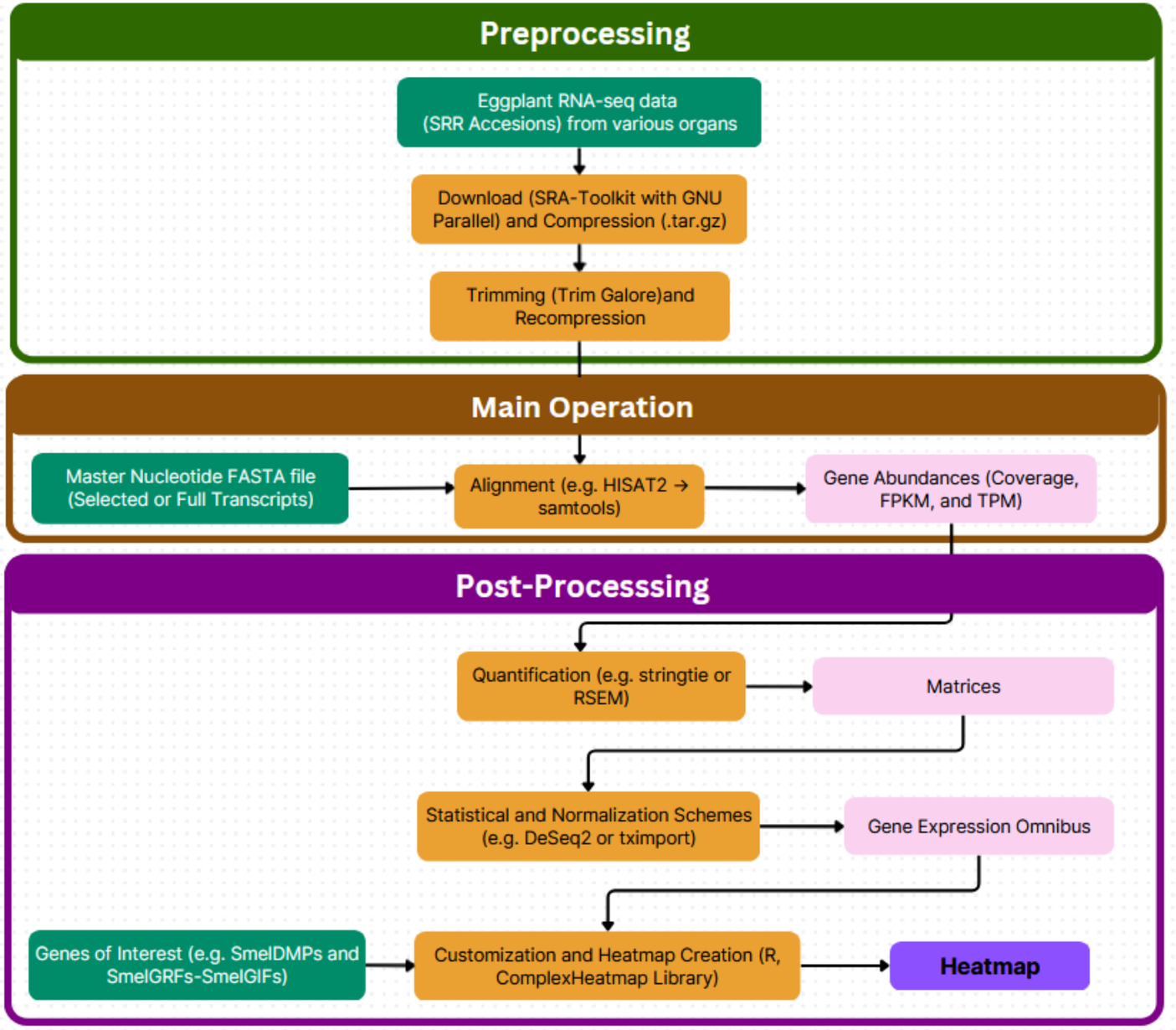
Figure 9. SRR sizes and *SmelDMP10.200* alternative splicing. (Figures screenshot from NCBI Run selector and Benchling, respectively)

CONSIDERATION

Strategies Employed to reduced Runtime

- Using parallelization techniques
    - e.g. In downloading the SRR files.
- Threading.

- Benchmarking of the alignment pipeline.
- Aligning a master ‘FASTA file’ and later querying the matrix file with genes of interest.



CONSIDERATION

Strategies and Factors to ensure Correctness/Accuracy

- Consensus expression thru replicates
- Comparison with other bioinformatics results

- Quality and the methodology of the RNA extraction
- Quality of the RNA sequencing

```
170  SRR_COMBINED_LIST=(
171      "${SRR_LIST_PRJNA328564[@]}"    # Main Dataset for GEA.
172      "${SRR_LIST_SAMN28540077[@]}"   # Chinese Dataset for replicability.
173      "${SRR_LIST_SAMN28540068[@]}"   # Chinese Dataset for replicability.
174      "${SRR_LIST_PRJNA865018[@]}"    # Good Dataset for SmelDMP GEA.
175  )
176
```

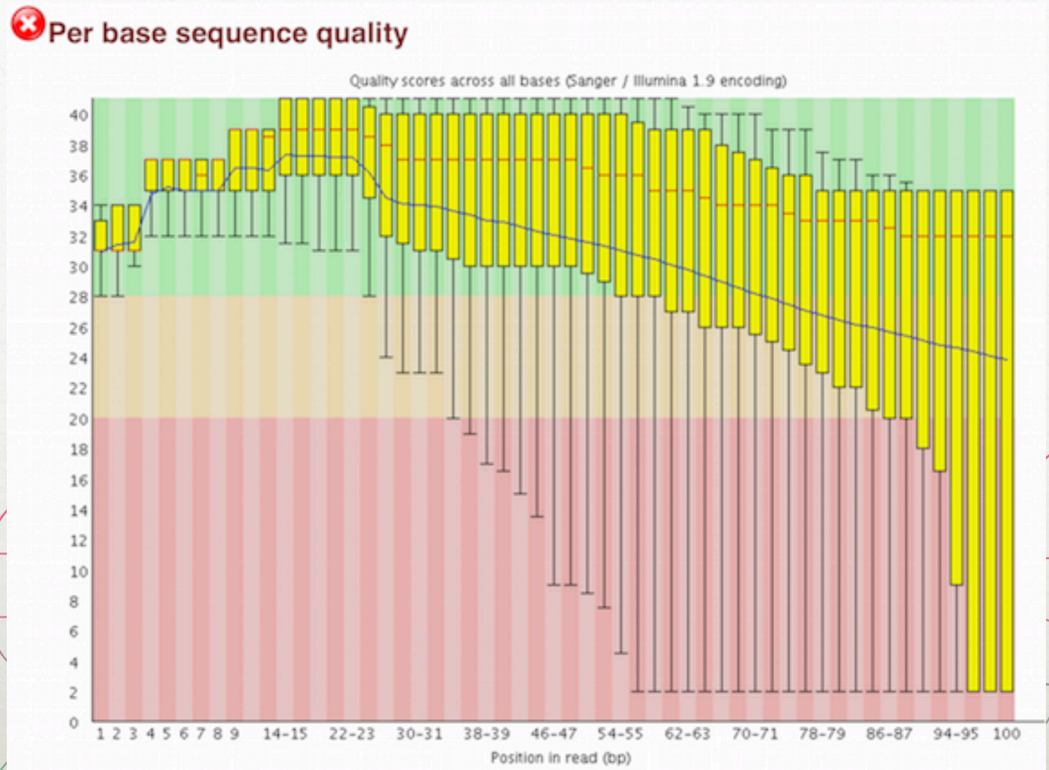


Figure 11. Replicate Datasets and example of Quality Control by FASTQC.

ALIGNMENT METHODS

CONSIDERATIONS

- Is it reference guided or de novo?
- Is it splice aware or not?

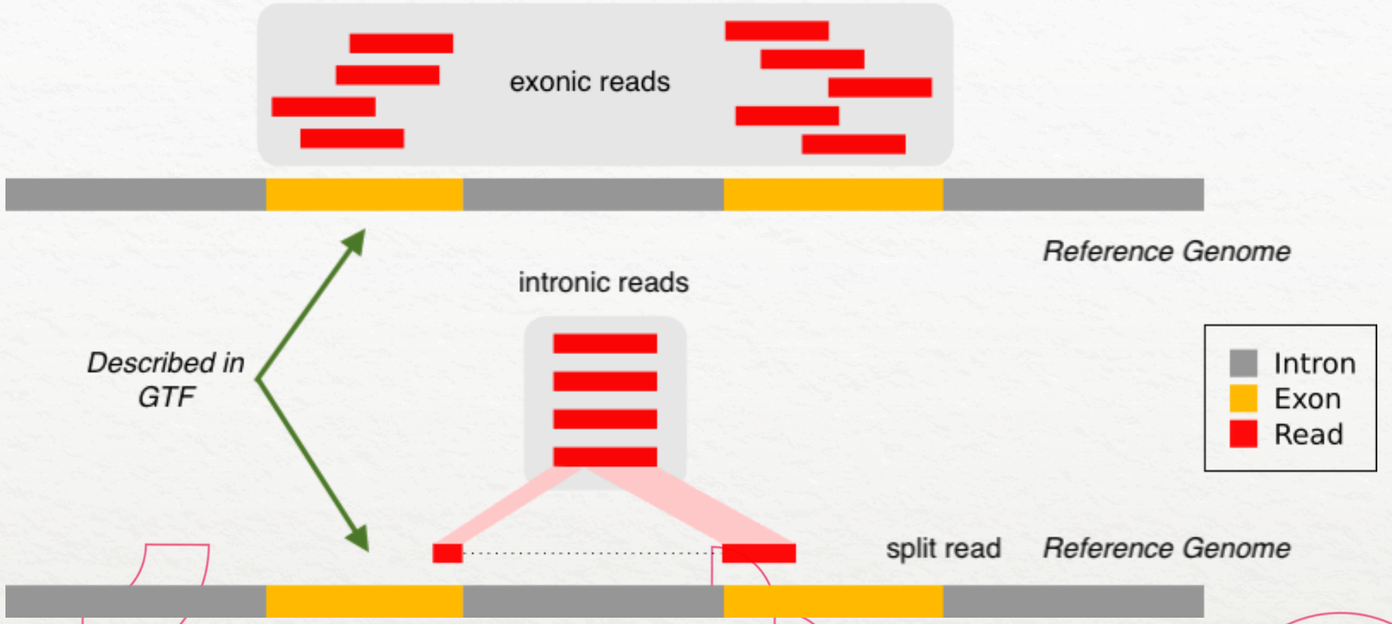


Figure 12. Relationship of the General Feature (GTF) to the introns and exons. Figure lifted from https://artbio.github.io/startbio/bulk_RNAseq-IOC/images/splice_aware_alignment.png

ALIGNMENT METHODS

HISAT2 Ref Guided

- Reference-guided aligner that aligns RNA-seq reads directly to a reference genome.
- Spliced aligner that can detect known and novel splice sites.

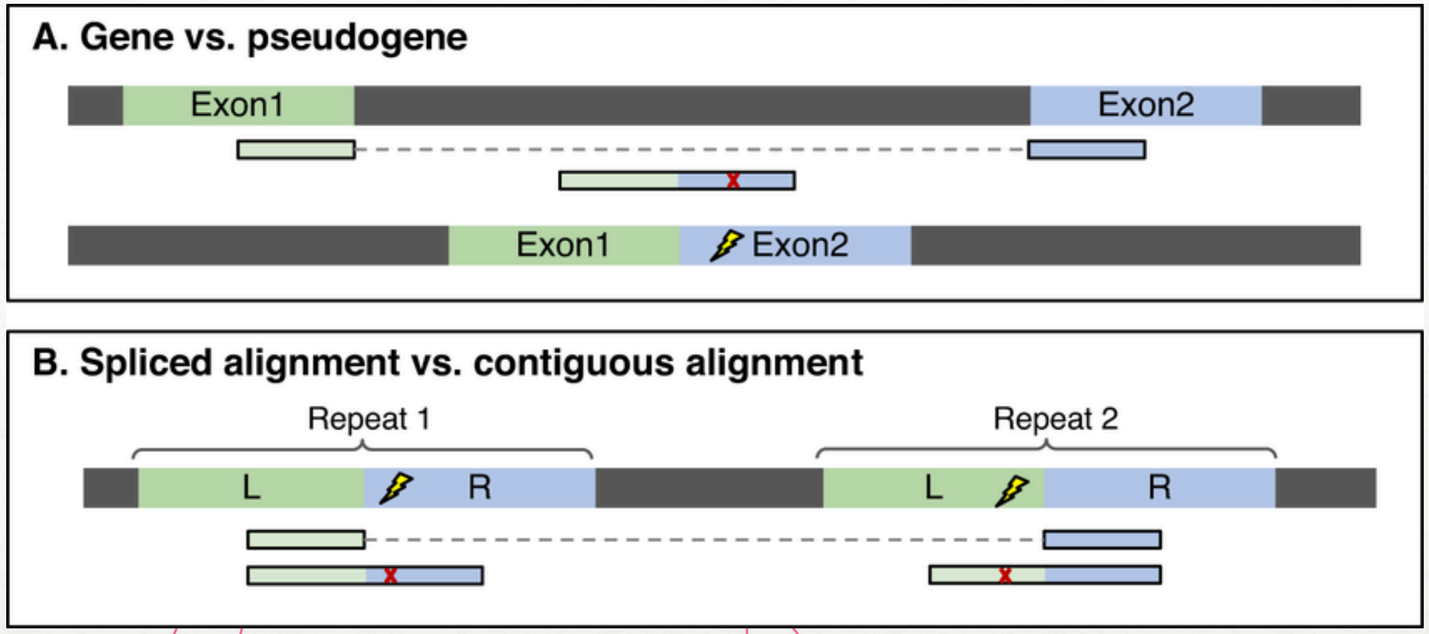


Figure 13. A. Correct spliced alignment. B. Repeats regions are prone to errors in the alignment.

Figure lifted from: "https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.researchgate.net%2Ffigure%2FOverview-of-HISAT2-algorithms-performance-in-two-alignment-scenarios-A-The-correct_fig1_375518866&psig=AOvVaw1MXzOkQVJOSCxqC5nNZyJT&ust=1762971463507000&source=images&cd=vfe&opi=89978449&ved=0CBgQjhxqFwoTCOiG49La6pADFQAAAAAdAAAAABAU"

ALIGNMENT METHODS

Bowtie2-RSEM

- fast and memory-efficient aligner:
  - align sequencing reads to a reference genome or transcriptome
- not-splice aware

- Phase 1: Finding a starting point (the seed) that will match
- Phase 2: Checking the other strand.
- Phase 3: Confirming the whole read.

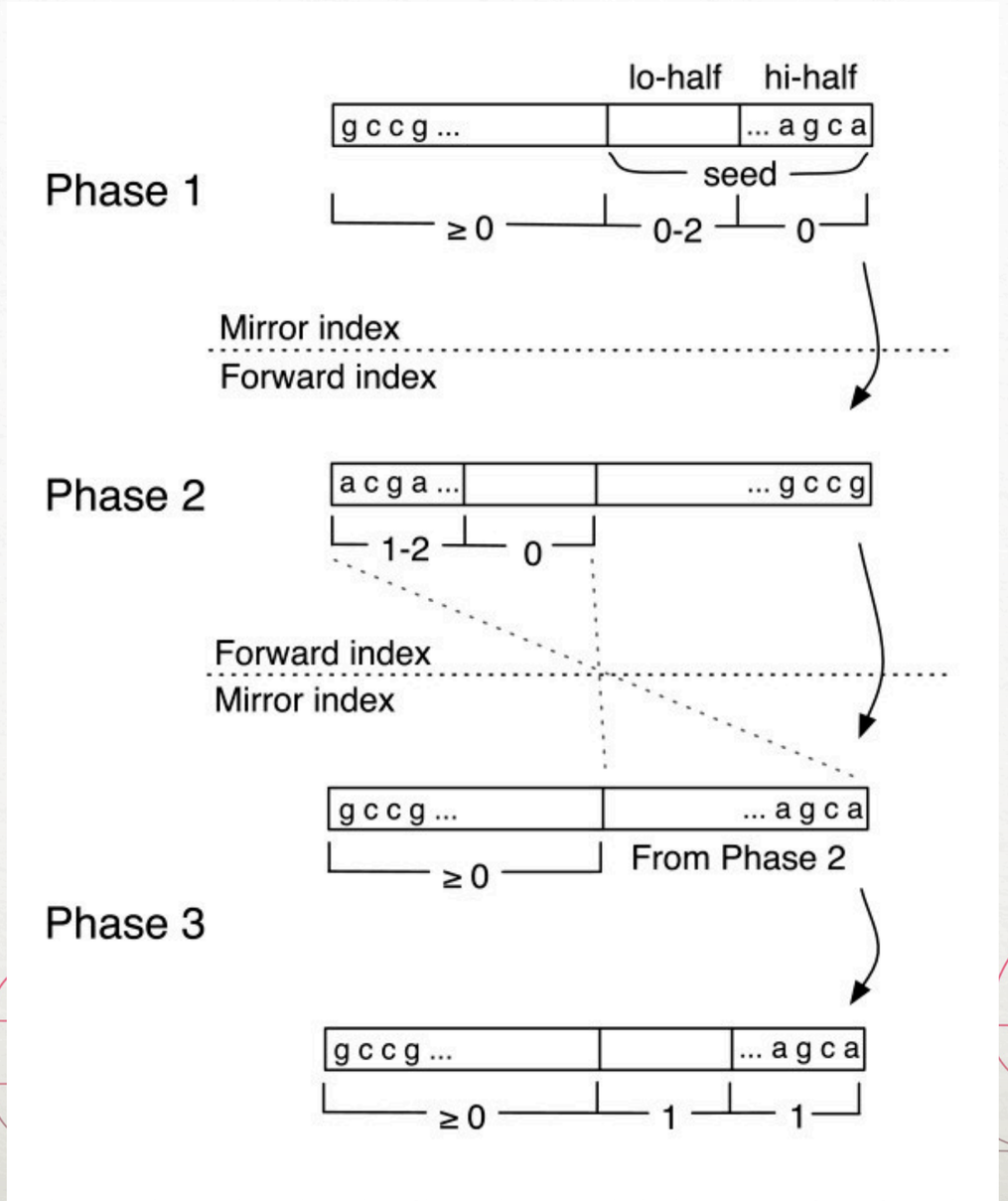


Figure 14. Three phases of Bowtie algorithm. Figure lifted from 'https://www.researchgate.net/figure/The-three-phases-of-the-Bowtie-algorithm-for-the-Maq-like-policy_fig2_24177230"

ALIGNMENT METHODS

Salmon Saf

- quasi-alignment method that quantifies transcript abundances quickly by mapping reads to a reference transcriptome rather than a genome.
    - transcript-level quantification
- faster
- Instead of perfectly aligning every base, it figures out which transcripts each read likely came from.

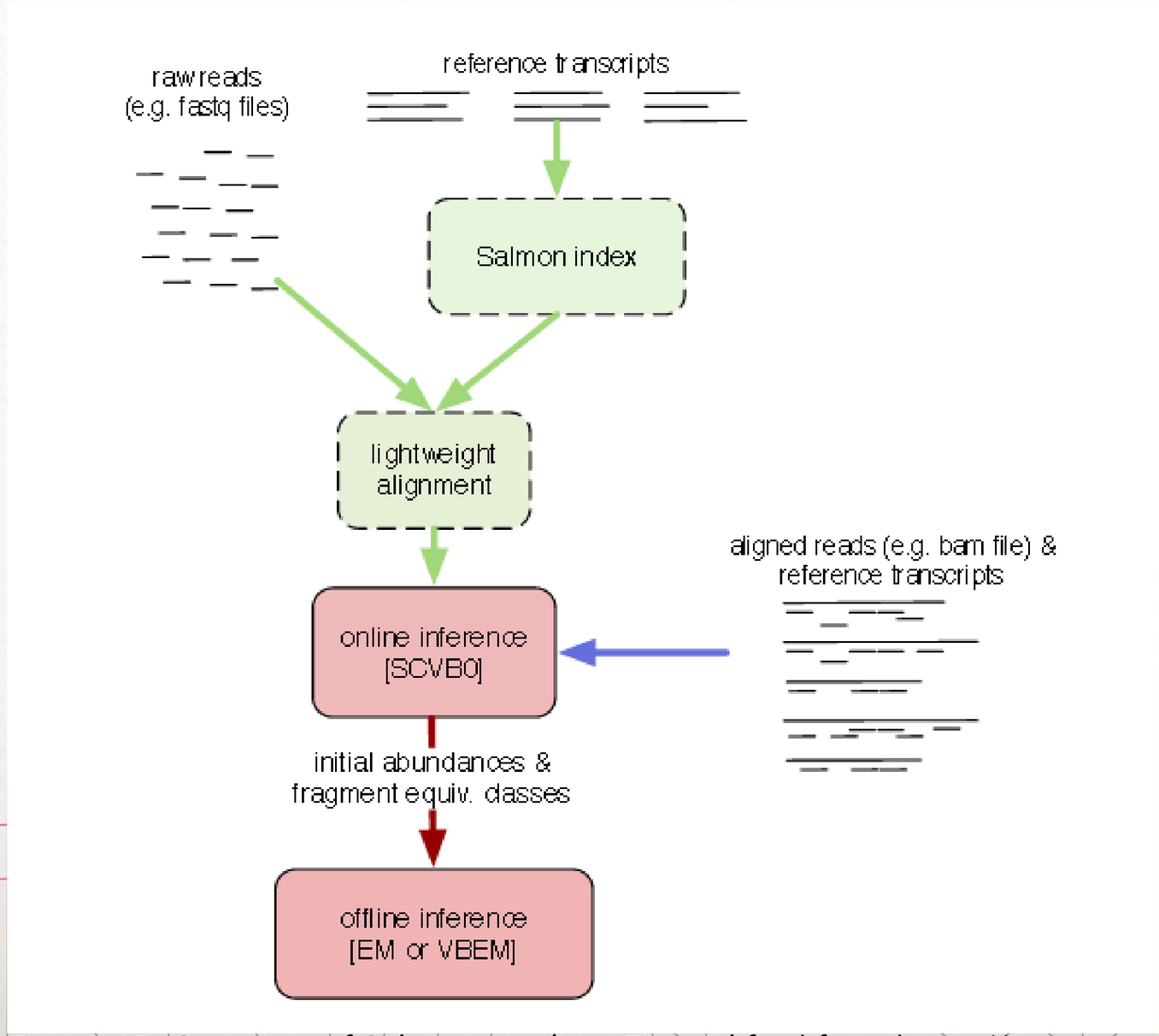


Figure 15. Overview of Salmon pipeline. Figure lifted from https://www.rna-seqblog.com/salmon-accurate-versatile-and-ultrafast-quantification-from-rna-seq-data-using-lightweight-alignment/

BENCHMARK RESULT

Runtime

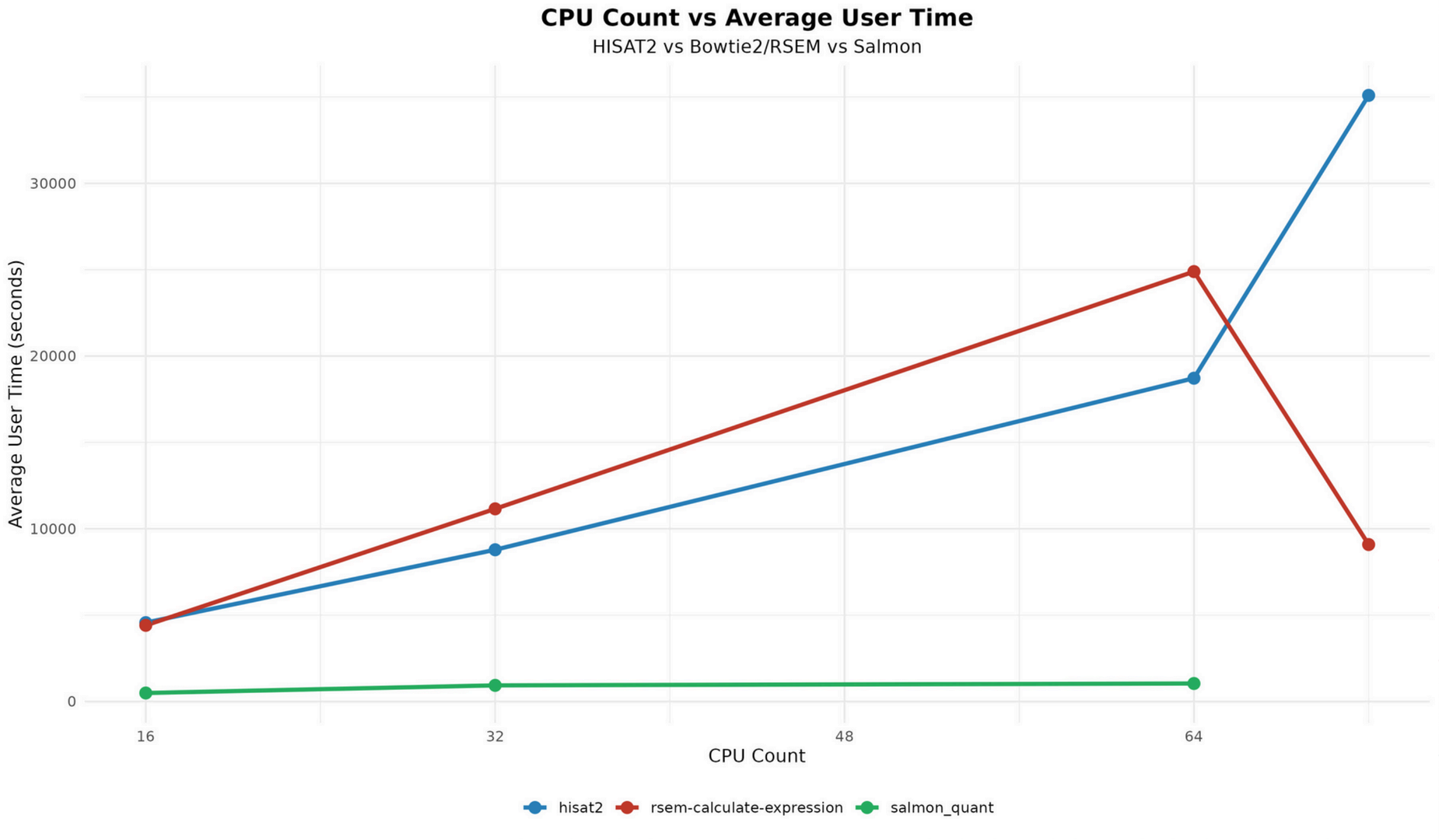
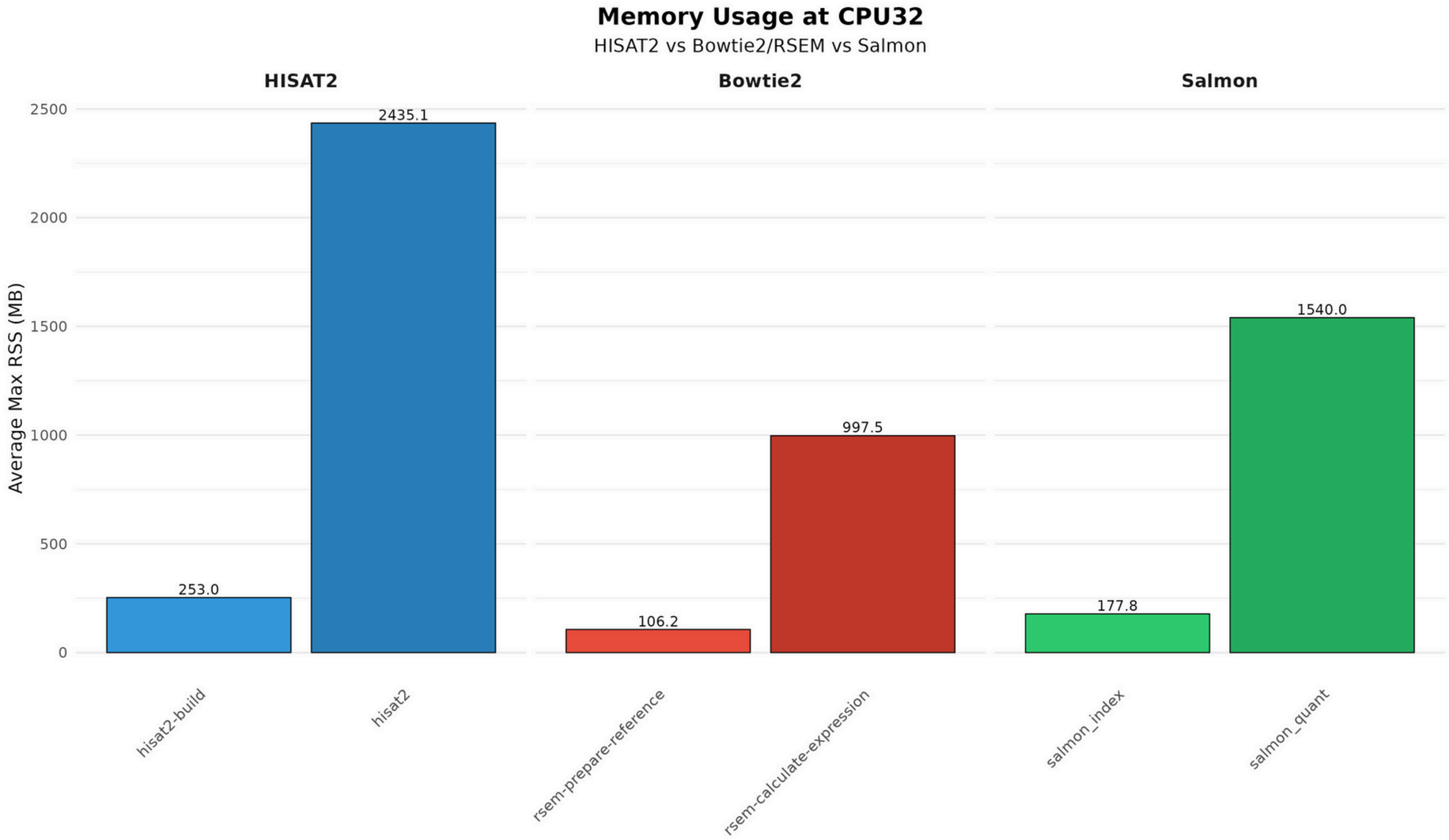


Figure 17. Scaling of CPU for HISAT2 Ref Guided Results.

BENCHMARK RESULT

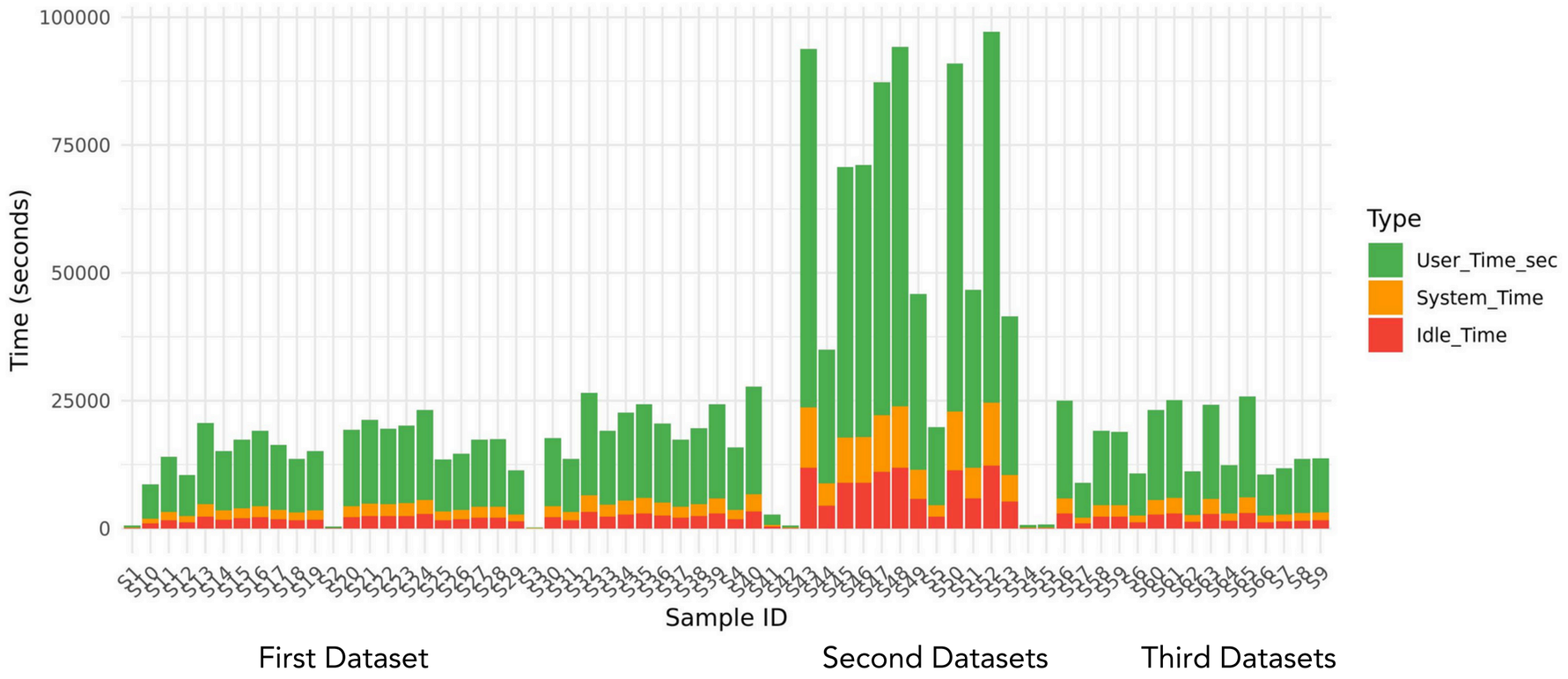
Space/Memory Usage



- Figure 16: RAM Usage of the major alignment tools.

BENCHMARK RESULT

Runtime per Datasets



- Figure 16: Runtime per Datasets

BENCHMARK RESULT

Correctness

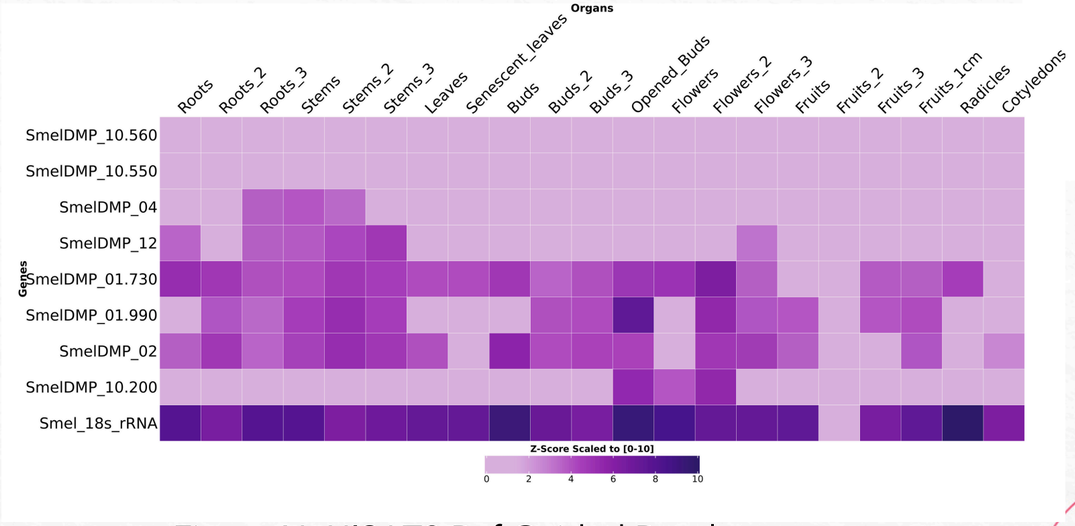


Figure N. HISAT2 Ref Guided Results.

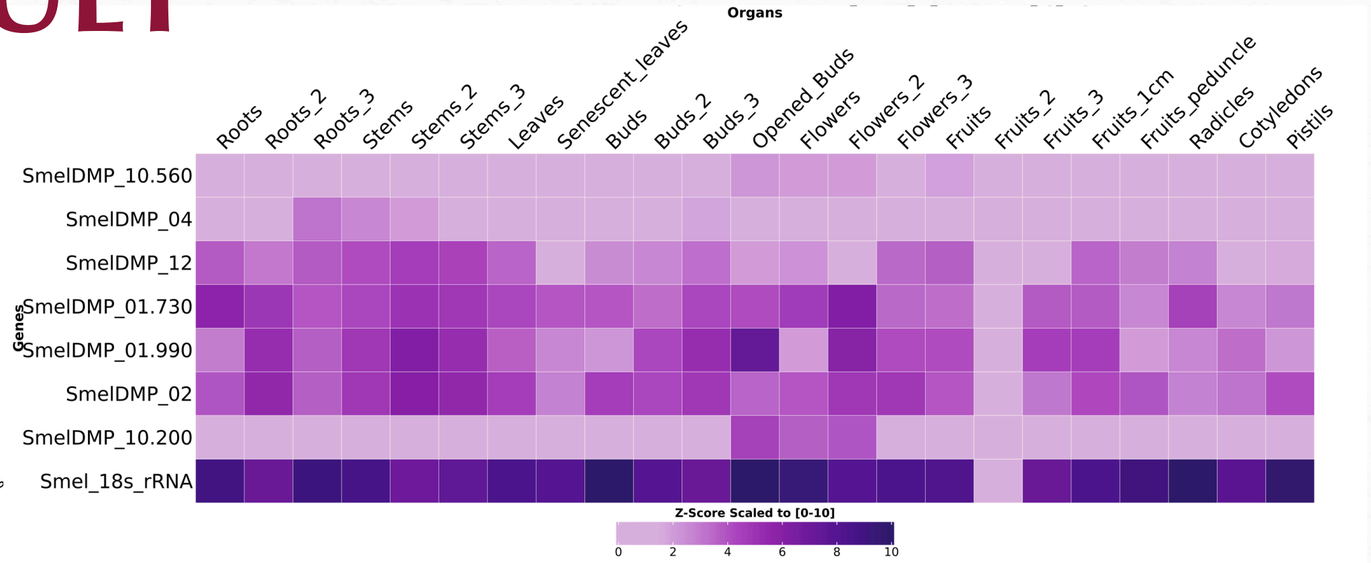


Figure N. Salmon Saf Ref Guided Results.

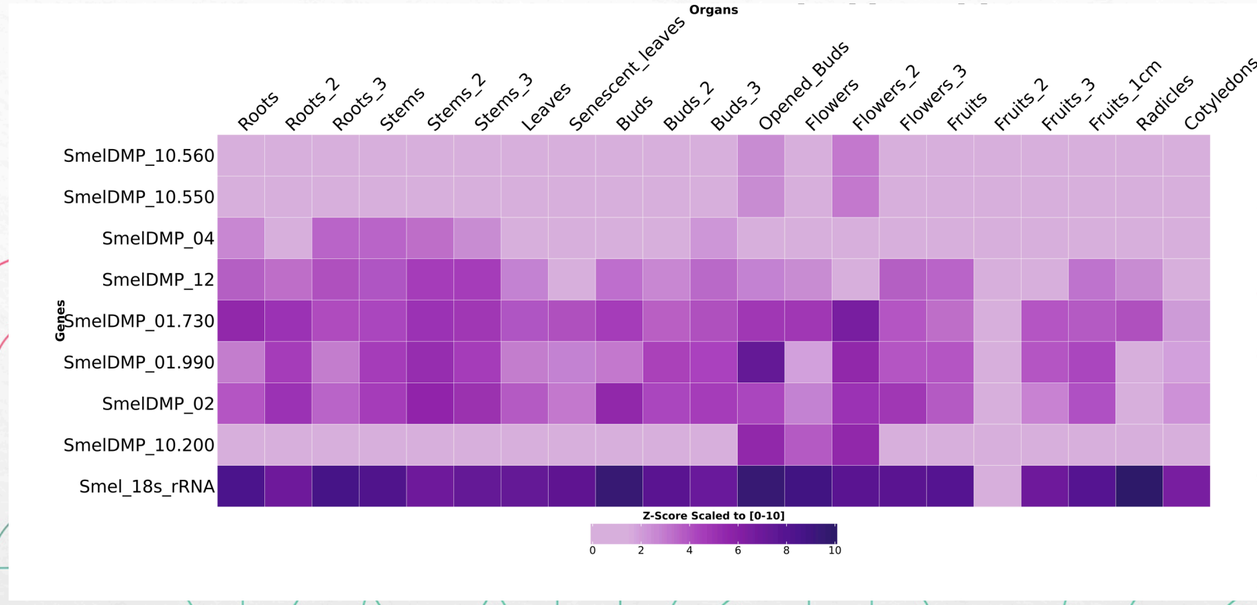


Figure N. Bowtie2 Ref Guided Results.

BENCHMARK RESULT

Correctness

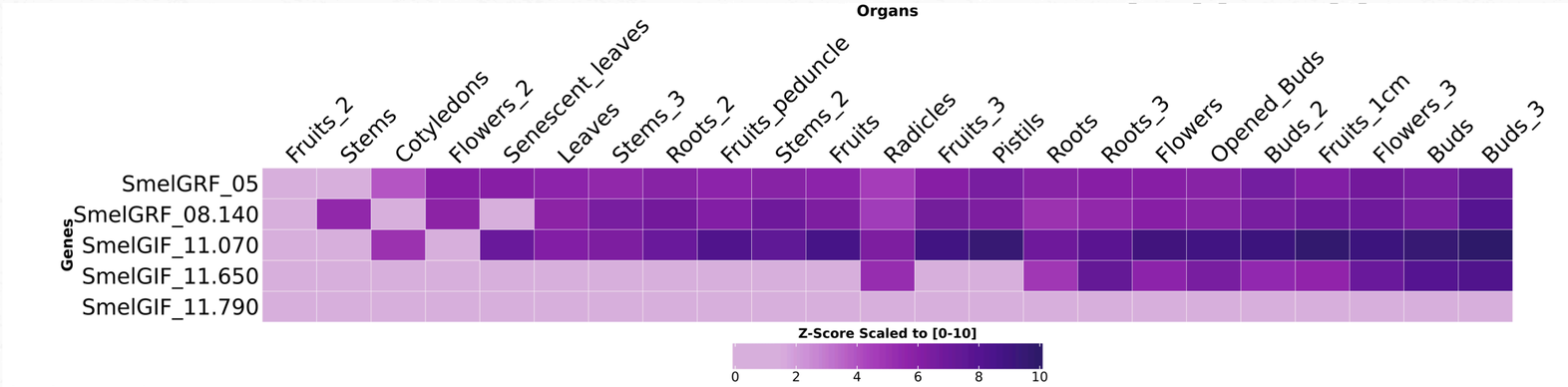


Figure N. HISAT2 Ref Guided Results.

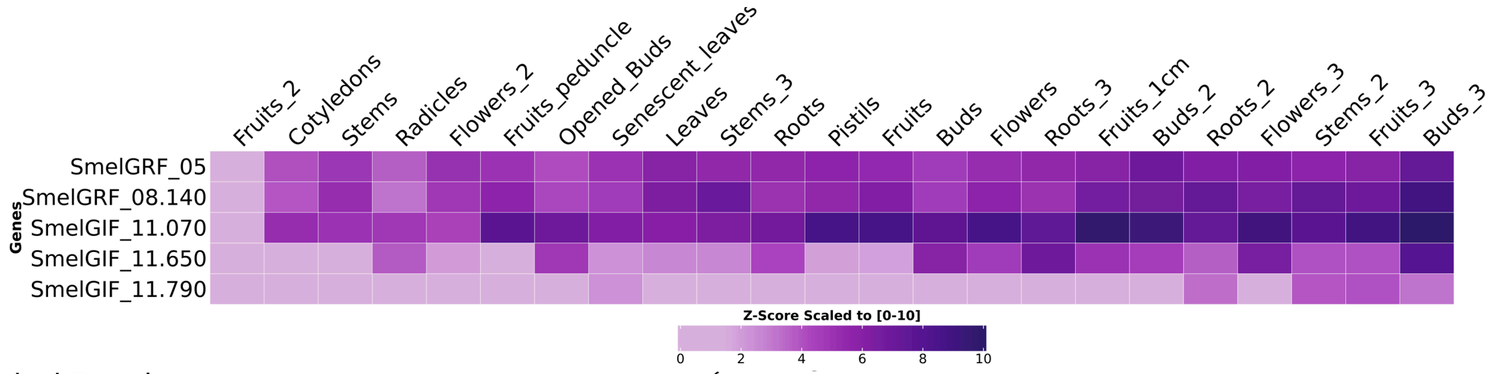


Figure N. Salmon Saf Ref Guided Results.

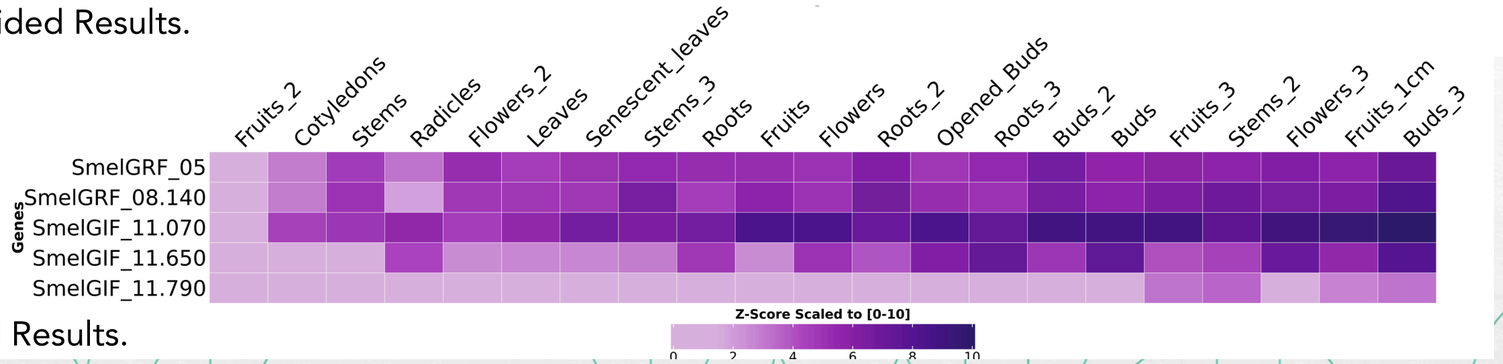


Figure N. Bowtie2 Ref Guided Results.

MAJOR RESULTS

- Number and Size of the Input datasets are the biggest contributor for Memory Usage and Runtime.
- In terms of concordance:
    - All of the alignment methods have consensus results.
        - *SmelDMP10.200* being expressed in the open buds and buds, and
        - Selected GRF-GIF genes are expressed in the meristematic regions.
- Choice between the three:
    - Still depends if genes have alternative splicings.

FUTURE DIRECTION

- Explore additional/various statistical analysis and normalization to increase the confidence in the results.

- Full or all of the transcripts alignment against available RNA-seq data of eggplant
    - Gene expression omnibus that can be lightly queried.
- Coexpression Analysis of Multiple Genes using WGCNA R library.

REFERENCE

- Corchete, L. A., Rojas, E. A., Alonso-López, D., De Las Rivas, J., Gutiérrez, N. C., & Burguillo, F. J. (2020). Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. Scientific Reports, 10(1), 19737. https://doi.org/10.1038/s41598-020-76881-x
- Fonseca, N. A., Marioni, J., & Brazma, A. (2014). RNA-Seq Gene Profiling - A Systematic Empirical Comparison. PLoS ONE, 9(9), e107026. https://doi.org/10.1371/journal.pone.0107026