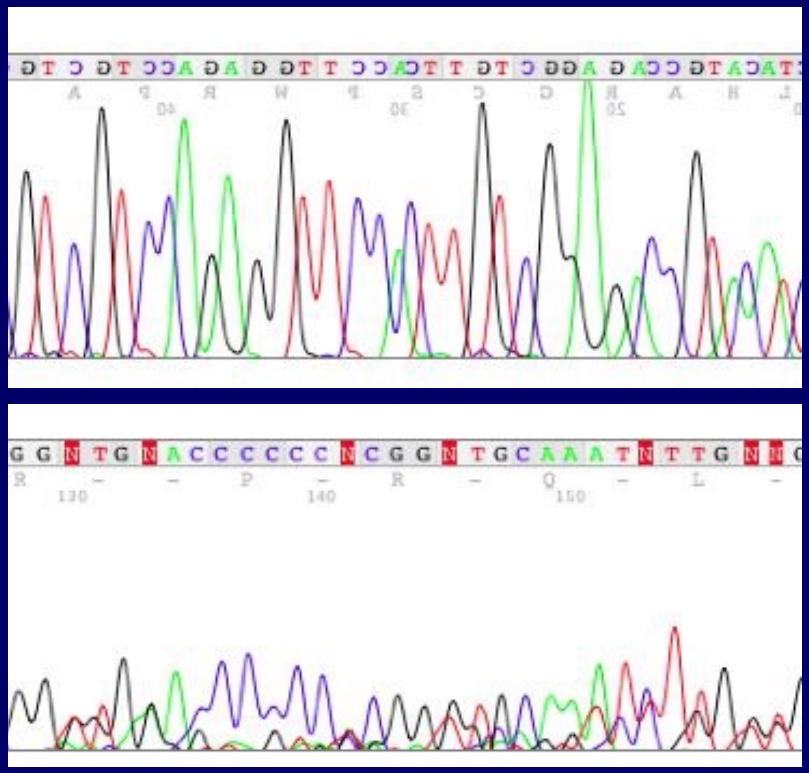
HTS data file

- Sequence and quality information are recorded as multi-FASTQ files.
- For efficient storage and transmission, they are transformed into SRA (Sequence Read Archives) format.

Observe the format of the fastq

Recording sequence reads from the machine – FASTQ



FASTA:

```
>My_sequence
AATTACGCGCGATACGAT
```

FASTQ:

```
@My_sequence
AATTACGCGCGATACGAT
+My_sequence quality
efcfffffcfeeYBBsdf
```

Recording of quality assessment allows filtering based on sequence quality.

Recording sequence and quality information

FASTQ format = FASTA + Quality

```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTT
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQabdefghadfda
```

- Two identification lines (@, +) for each sequence.
- Identification line format depends on specific sequencing platform.
- Quality line using characters representing integer values.

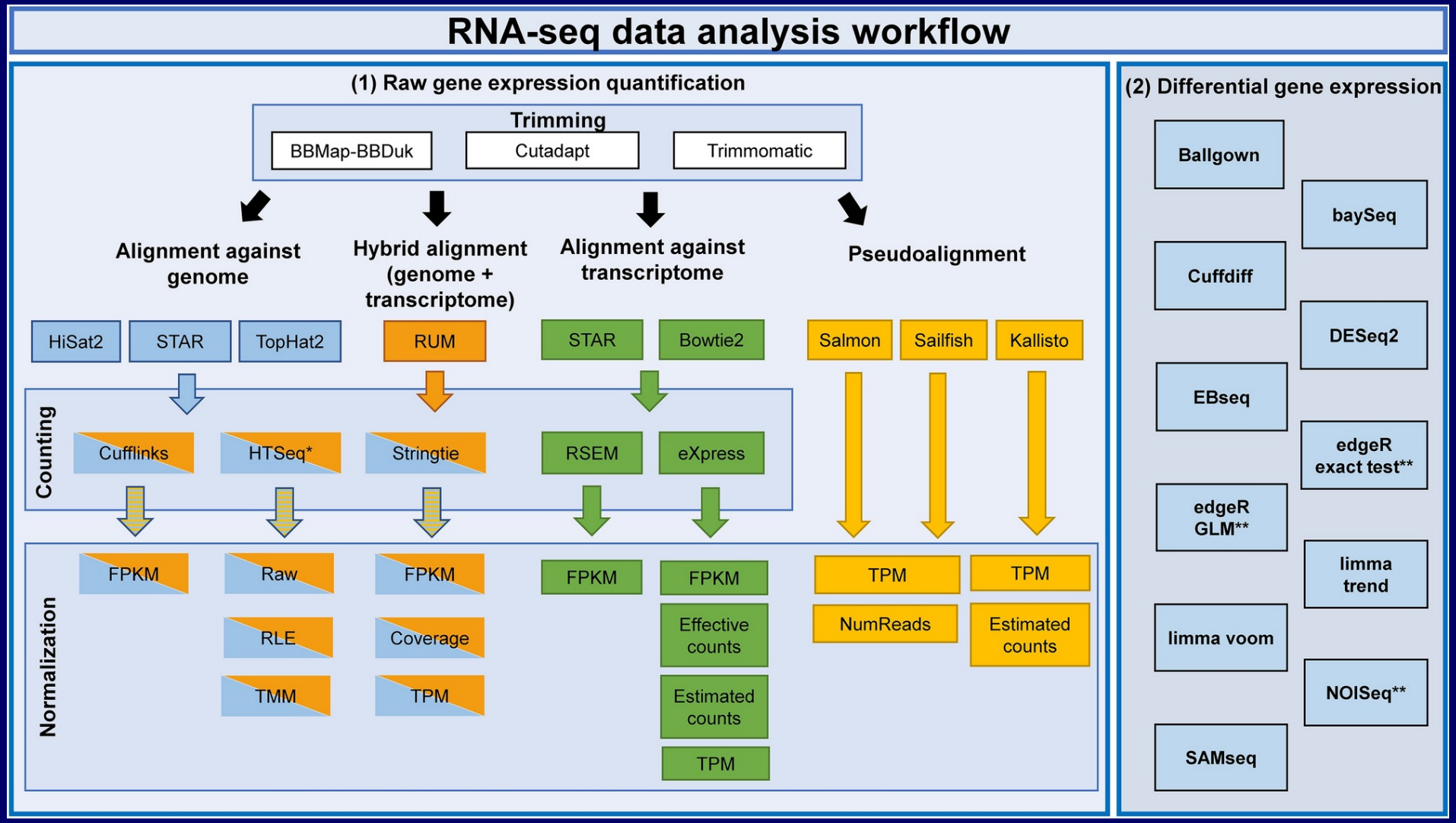
HTS data – map to genome

- Each algorithm applies a different mapping strategy and requires a specific index (multiple files).
- Index can be built with programs using genomic sequence (and annotation) files as input.
- Pay attention to the version of the genome.
  - Consider how you will view and compare your results with other data.

Generate index from .fasta and .gtf

```
#!/bin/sh
#SBATCH --job-name=Dm6_Star_Index
#SBATCH --mail-type=ALL
#SBATCH --mail-user=leizhou@ufl.edu
#SBATCH --mem-per-cpu=6gb
#SBATCH --cpus-per-task=8
#SBATCH --qos=zhou
#SBATCH -t 3:00:00
#SBATCH --output=STAR_Index_%j.log

module load star

STAR --runThreadN 16 \
--runMode genomeGenerate \
--genomeDir Dm6.49.StarIndex \
--genomeFastaFiles ./Dm6.49.fa \
--sjdbGTFfile ./Dm6.49.gtf \
--sjdbOverhang 99
```

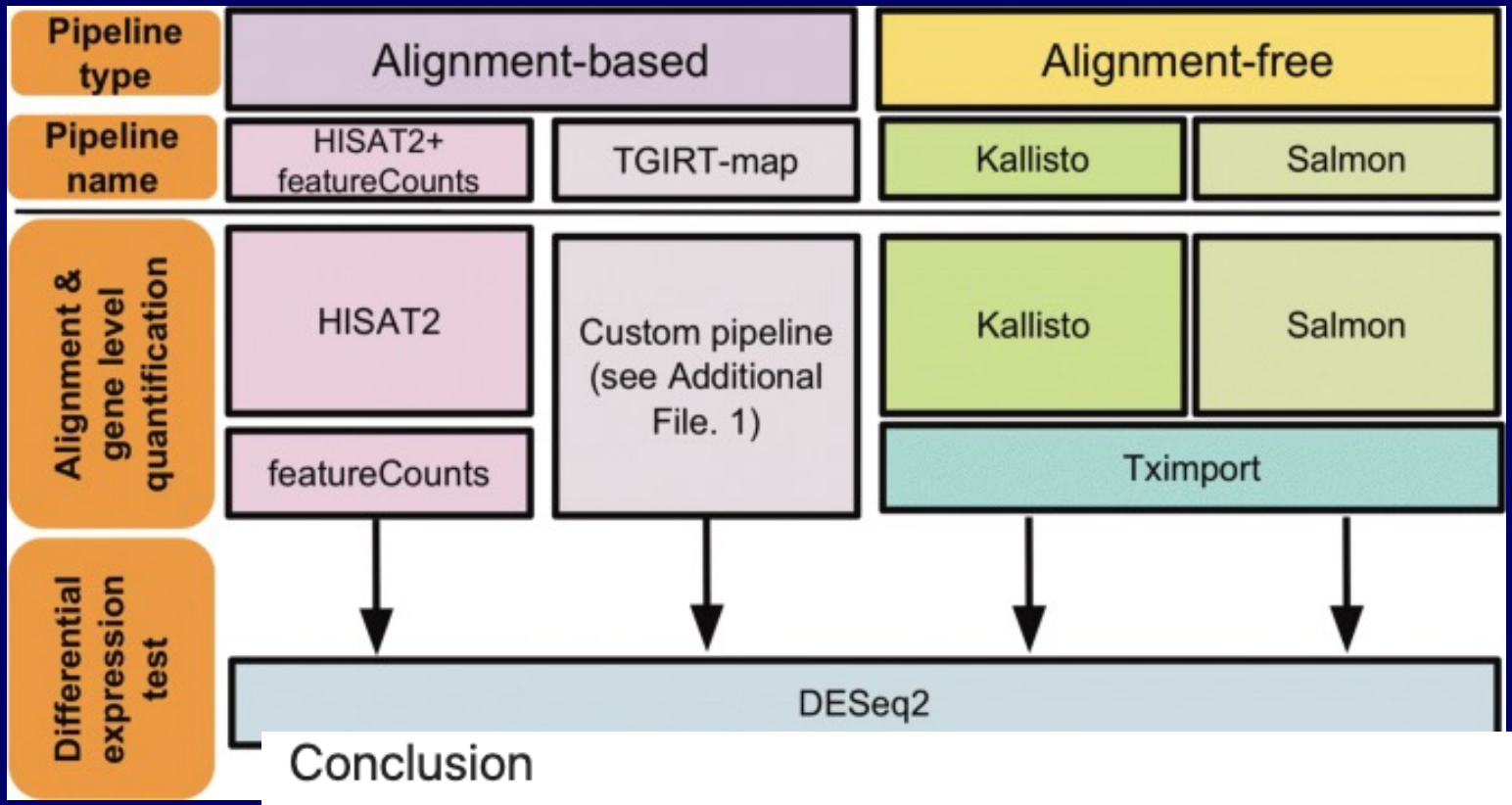
RNA-seq map options



Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Corchete et al (2020) Sci. Rep*

Limitations of alignment-free tools in total RNA-seq quantification

Douglas C. Wu, Jun Yao, Kevin S. Ho, Alan M. Lambowitz & Claus O. Wilke



Conclusion

We have shown that alignment-free and traditional alignment-based quantification methods perform similarly for common gene targets, such as protein-coding genes. However, we have identified a potential pitfall in analyzing and quantifying lowly-expressed genes and small RNAs with alignment-free pipelines, especially when these small RNAs contain biological variations.

RNA-Seq: map to genome

Q: If you are interested in identifying differentially expressed genes, de novo mRNA splicing, and novel ncRNAs in cancer samples. How would you map your RNA-Seq reads.

a.) Map to genome with Star (or Hisat2/Tophat2).

b.) Map to transcriptome with Star (or bowtie).

c.) Get TPM with Salmon.

Map to genome - job file (tophat)

```
#!/bin/bash
#SBATCH --job-name=WG_tophat_20200412
#SBATCH --output=WG_tophat_20200412_%j.log   # Standard output and error log
#SBATCH --mail-type=ALL
#SBATCH --mail-user=leizhou@ufl.edu
#SBATCH --time=24:00:00
#SBATCH --cpus-per-task=4
#SBATCH --mem-per-cpu=3gb
pwd; date

module load tophat/2.1.1

ln -s /ufrc/gms6014/share/genomes/dm6/Annotations/genes.gtf .
ln -s /ufrc/gms6014/share/genomes/dm6/Bowtie2Index/genome.* .

mkdir Bam  ## for Bam files

tophat -G genes.gtf -p 2 -o Bam/WG_young_1.bam genome SRR1618640.fastq.gz
tophat -G genes.gtf -p 2 -o Bam/WG_young_2.bam genome SRR1618641.fastq.gz
tophat -G genes.gtf -p 2 -o Bam/WG_old_1.bam genome SRR1618641.fastq.gz
tophat -G genes.gtf -p 2 -o Bam/WG_old_2.bam genome SRR1618643.fastq.gz
```

Map to genome - job file (Star)

```
STAR --readFilesCommand zcat --genomeDir ./index/Dm6.44.StarIndex/ \
--sjdbGTFfile ./index/Dm6.44.gtf \
      --runThreadN 2 --runMode alignReads --outSAMtype BAM SortedByCoordinate \
--outBAMsortingBinsN 200 --limitBAMsortRAM 16013050982 \
      --readFilesIn SRR1618640.fastq.gz \
       --outFileNamePrefix ./starMap/WG_young_1
```

Map to genome - job file (hisat2)

```
#!/bin/bash
#SBATCH --job-name=WG_Hisat2_20200412
#SBATCH --output=WG_Hisat2_%j.log   # Standard output and error log
#SBATCH --mail-type=ALL
#SBATCH --mail-user=leizhou@ufl.edu
#SBATCH --time=24:00:00
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=4
#SBATCH --mem-per-cpu=3gb
pwd; date

module load hisat2

ln -s /ufrc/gms6014/share/genome/dm6/hisat2/genome.* .

hisat2 -p 4 --dta -x genome -S Sam/WG_young_1.sam   -U SRR1618640.fastq.gz
hisat2 -p 4 --dta -x genome -S Sam/WG_young_2.sam  -U  SRR1618641.fastq.gz
hisat2 -p 4 --dta -x genome -S Sam/WG_old_1.sam  -U  SRR1618642.fastq.gz
hisat2 -p 4 --dta -x genome -S Sam/WG_old_2.sam  -U  SRR1618643.fastq.gz
```

Practice: Map the reads to Dm6.49 genome with STAR.

Count the transcripts - cufflinks

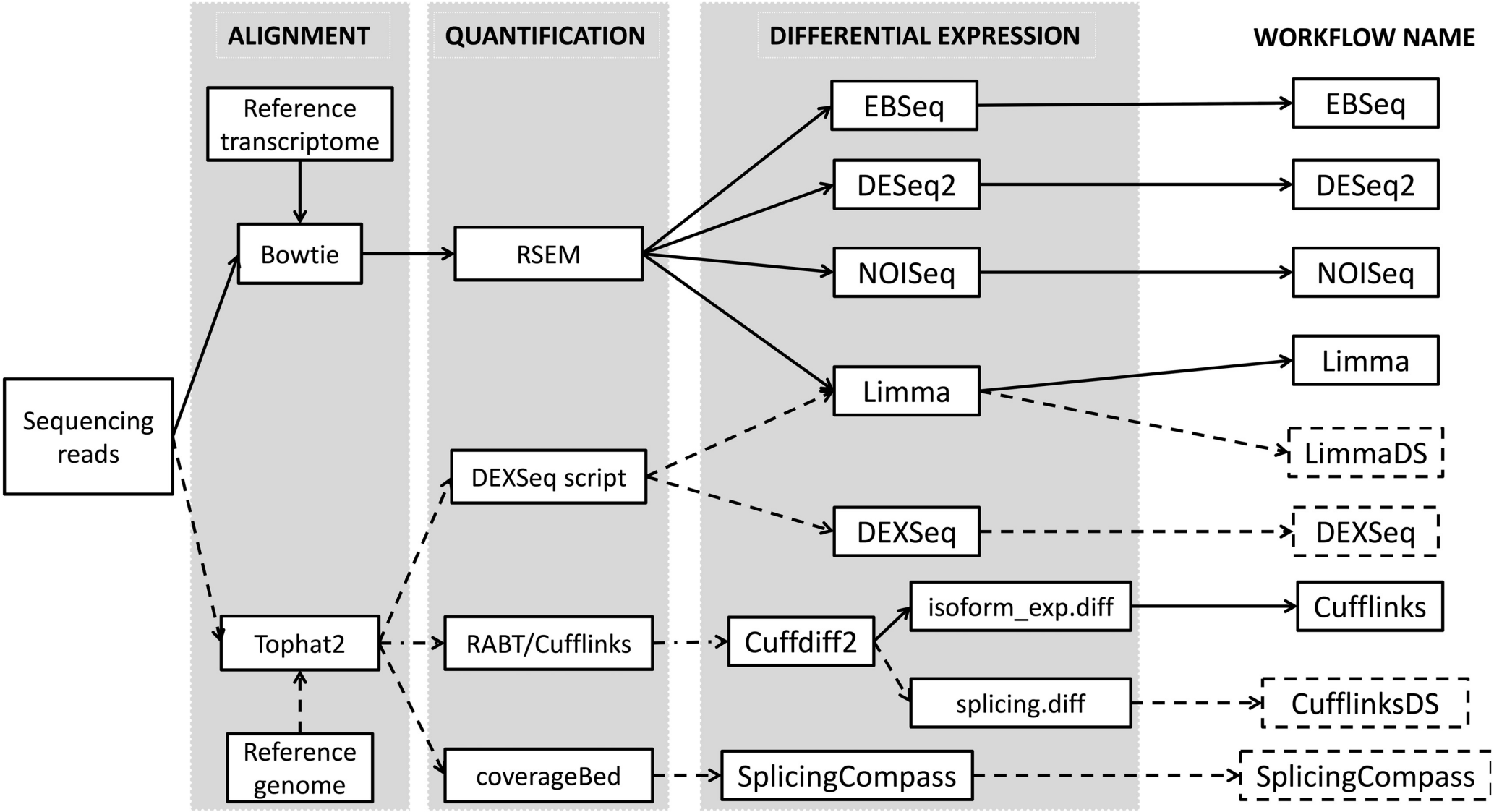
```
#!/bin/bash
#SBATCH --job-name=cufflnks
#SBATCH --mail-type=ALL
#SBATCH --mail-user=leizhou@ufl.edu
#SBATCH --mem-per-cpu=2gb
#SBATCH --cpus-per-task=4
#SBATCH --time=30:00:00
#SBATCH --output=cufflinks_%j.log

pwd; date

module load cufflinks


## count ####

cufflinks -p 4 -o WG_young_1.clout Bam/WG_young_1.bam
cufflinks -p 4 -o WG_young_2.clout Bam/WG_young_2.bam

cufflinks -p 4 -o WG_old_1.clout Bam/WG_old_1.bam
cufflinks -p 4 -o WG_old_2.clout Bam/WG_old_2.bam
```

RNA-Seq Overview

Four major steps, semi-independent of each other.

I. Mapping → produce SAM/BAM or counts data.
II. Quantification → produce RPKM for each gene/transcript.
III. Identifying DEG (Differentially expressed genes) → gene list.
IV. Identifying affected biological processes/pathways.

RNA-Seq overview



Merrino et al. "A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human RNA-seq studies" Brief Bioinform. 2017. doi:10.1093/bib/bbx122

RNA-Seq: Getting counts

- Raw – counts (reads) per gene.
- Normalized
  - FPKM (Fragments Per Kilobase gene length and per Million reads)
  - TPM (Transcripts Per Million)
- Depending on the which program will be used for identifying DEGs.
  - DESeq (DESeq2) requires raw counts
  - CuffLinks generated normalized counts as well as models for CuffDiff.