General processing of HTS data

1. Mapping to the targeted genome.
2. Quantification.
3. Comparison.
4. Further analysis (GO enrichment assay, etc.)

RNA-Seq Overview

Four major steps, semi-independent of each other.

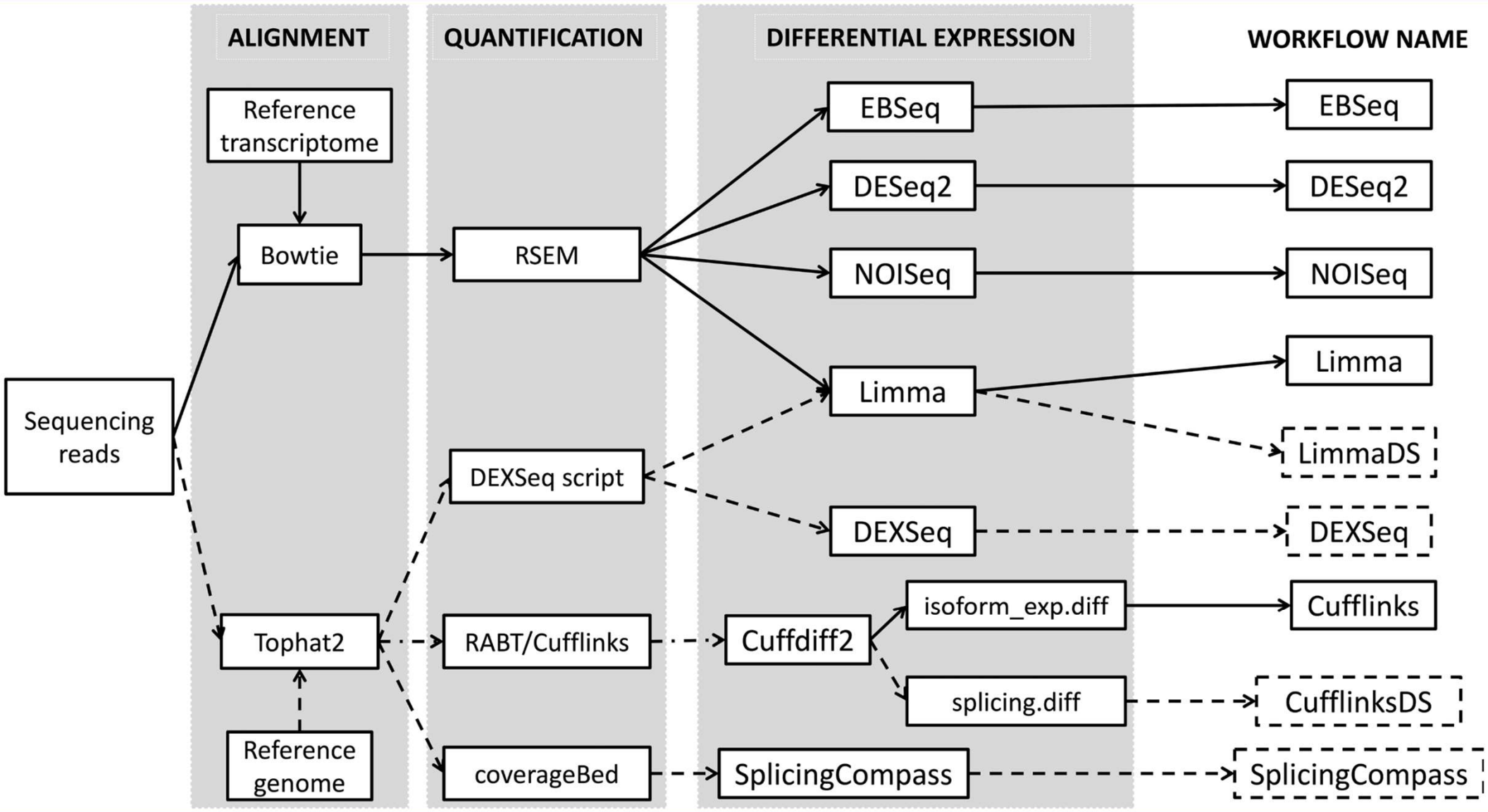
I. Mapping → produce SAM/BAM or counts data.

II. Quantification → produce RPKM for each gene/transcript.

III. Identifying DEG (Differentially expressed genes) → gene list.

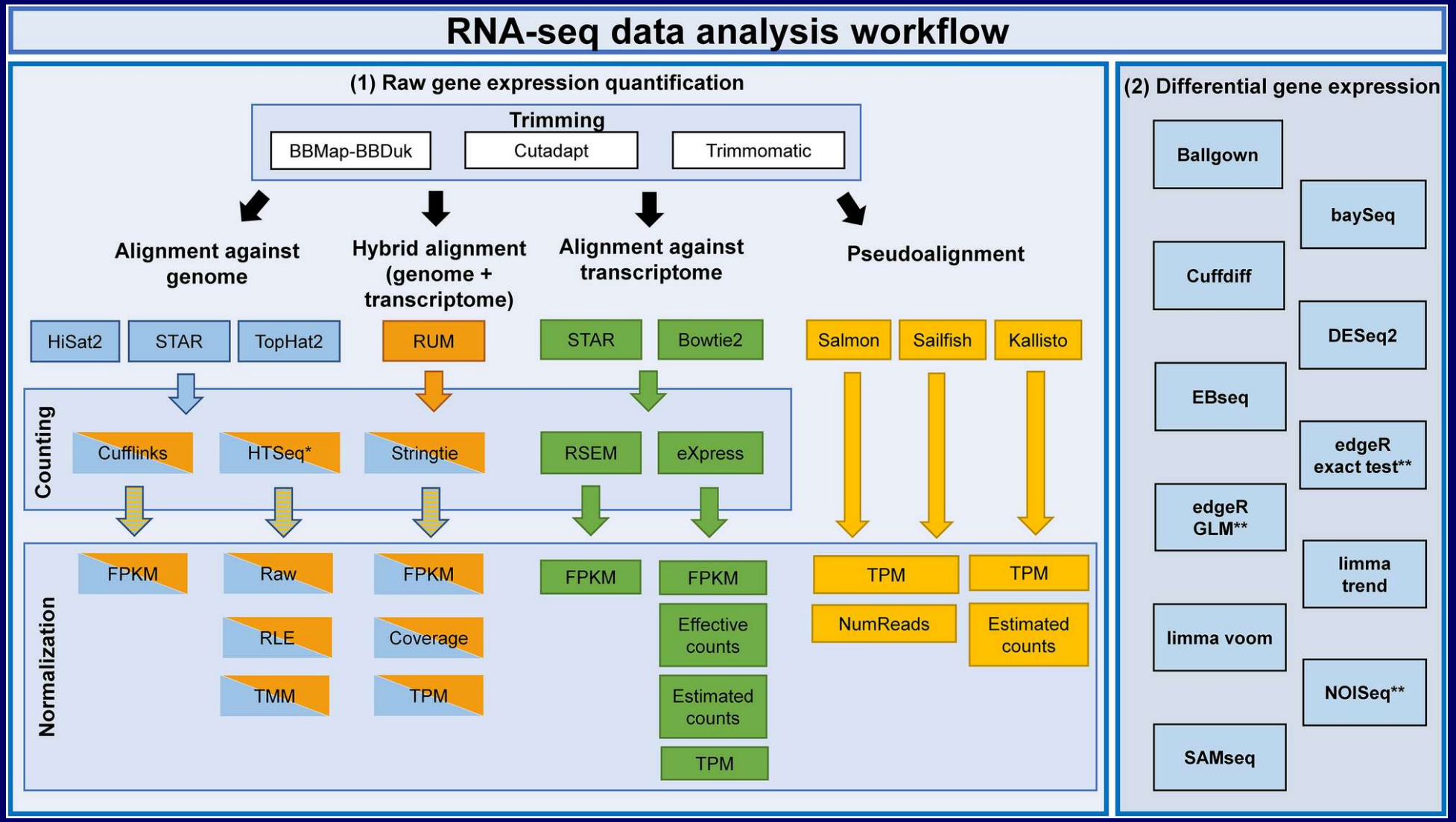
IV. Identifying affected biological processes/pathways.

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 data processing options



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

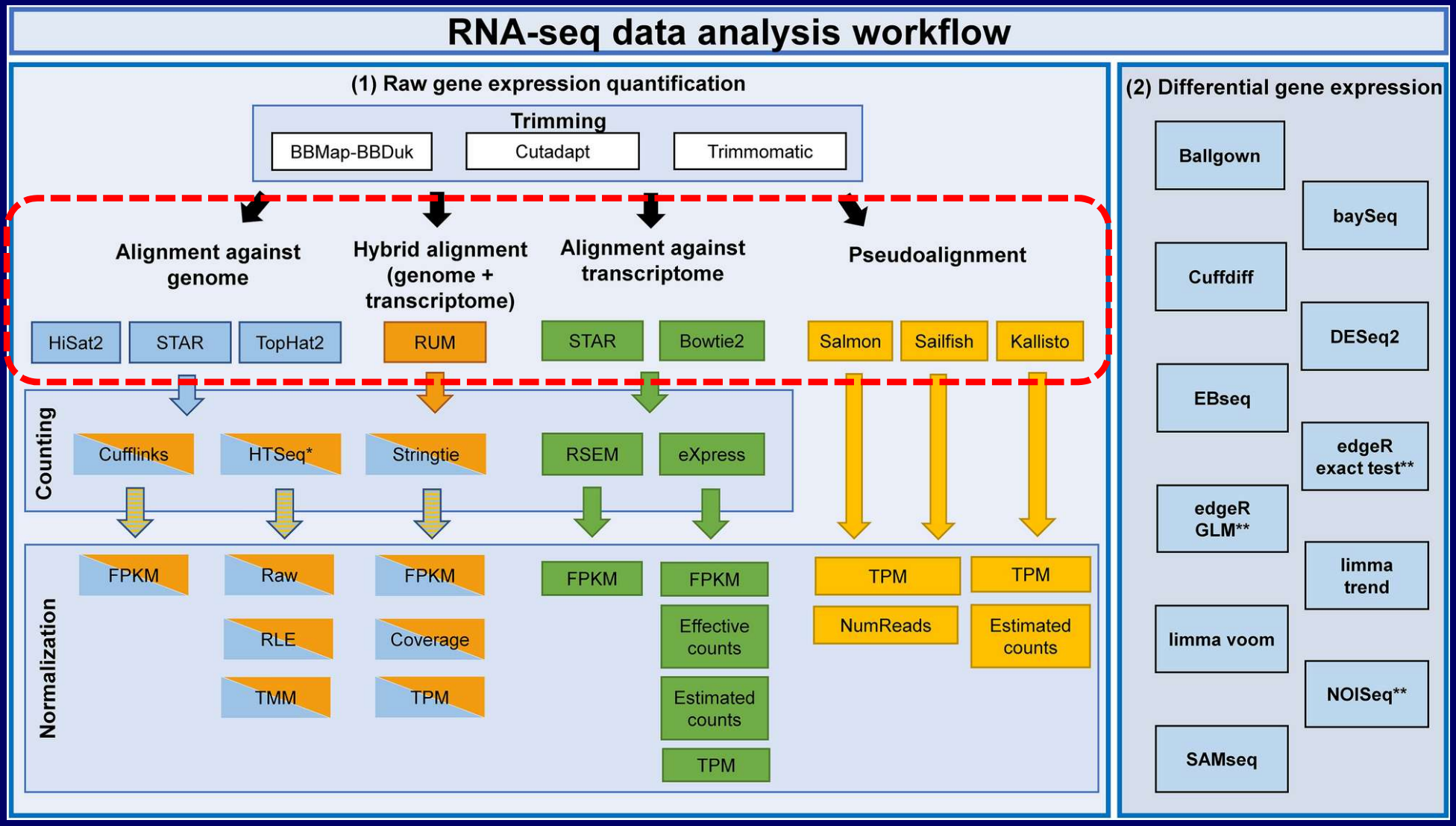
RNA-seq data processing options

How to design your RNA-Seq analysis process:

- **Based on the biological questions:**
  - **Identifying differentially expressed gene**
  - **Using gene expression data for clustering cancer samples.**

- **Based on objective:**
  - **I want to identify different splicing forms and ncRNAs, vs.**
  - **I am only interested in protein-coding genes.**

RNA-seq map options

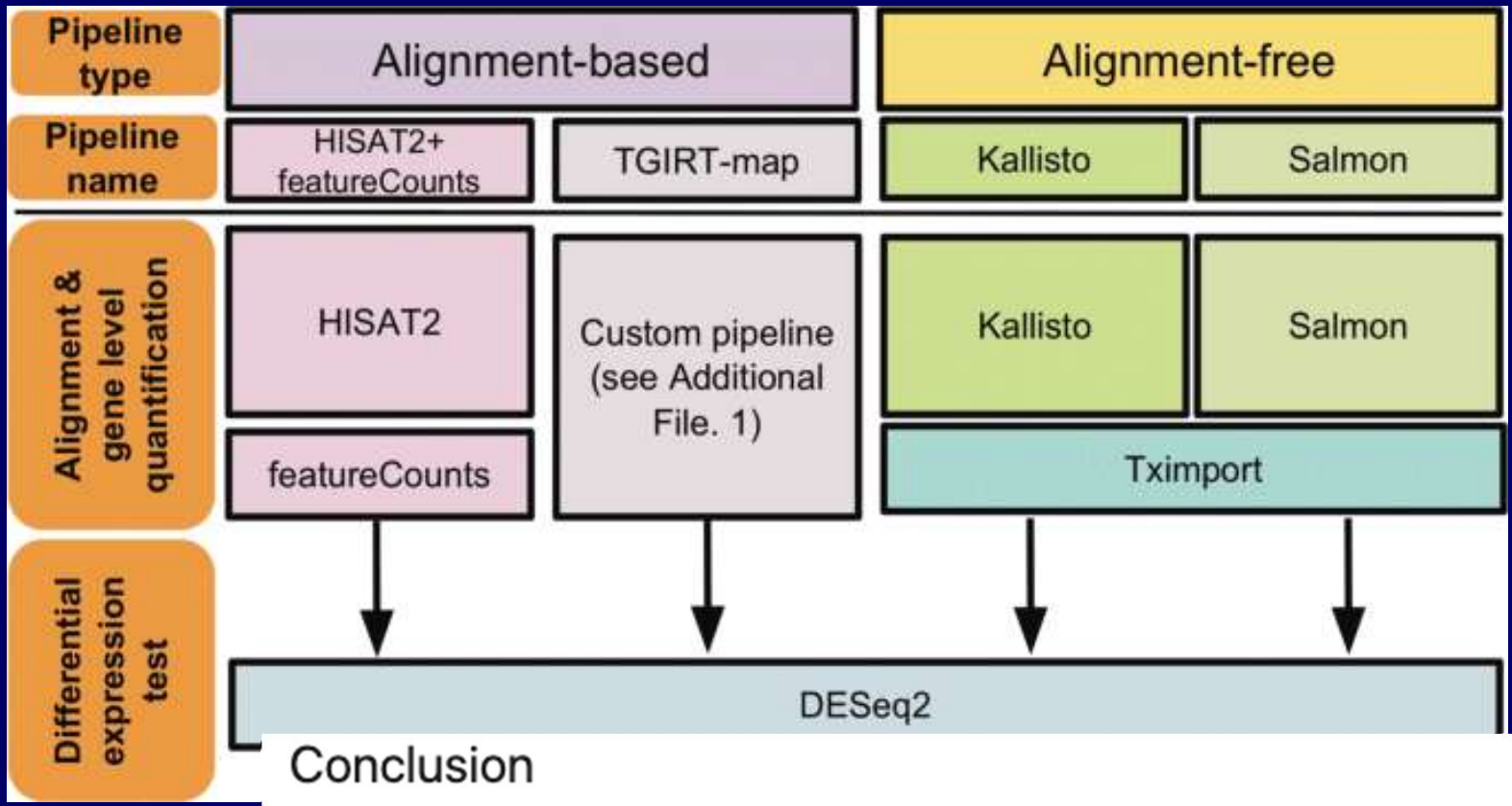


RNA-Seq: map to genome

- Each algorithm applies a different mapping strategy and requires a specific index (multiple files).
- Index can be built with programs using genomic or transcriptomic sequence and GTF files as input, analogous to what you did with “makeblastdb”.
  - GTF- Gene Transfer (annotation) format.
- Given the task and resources the mapping may take minutes, or more likely hours, or even days.
  - – *you might need 12 pots of coffee if you did not use a job file*.

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.

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 Dm6.44.StarIndex \
--genomeFastaFiles ./Dm6.44.fa \
--sjdbGTFfile ./Dm6.44.gtf \
--sjdbOverhang 99
```

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 – Choose resource if you have a primary account

```
#!/bin/bash
#SBATCH --job-name=StarMapping
#SBATCH --output=StarMapping_%j.log
#SBATCH --mail-type=ALL
#SBATCH --mail-user=xxxx@ufl.edu
#SBATCH --time=24:00:00
#SBATCH --qos=gms6014
#SBATCH --cpus-per-task=2
#SBATCH --mem-per-cpu=4gb
```

```
$ sbatch --account=gms6014 --qos=gms6014 jobfile
```

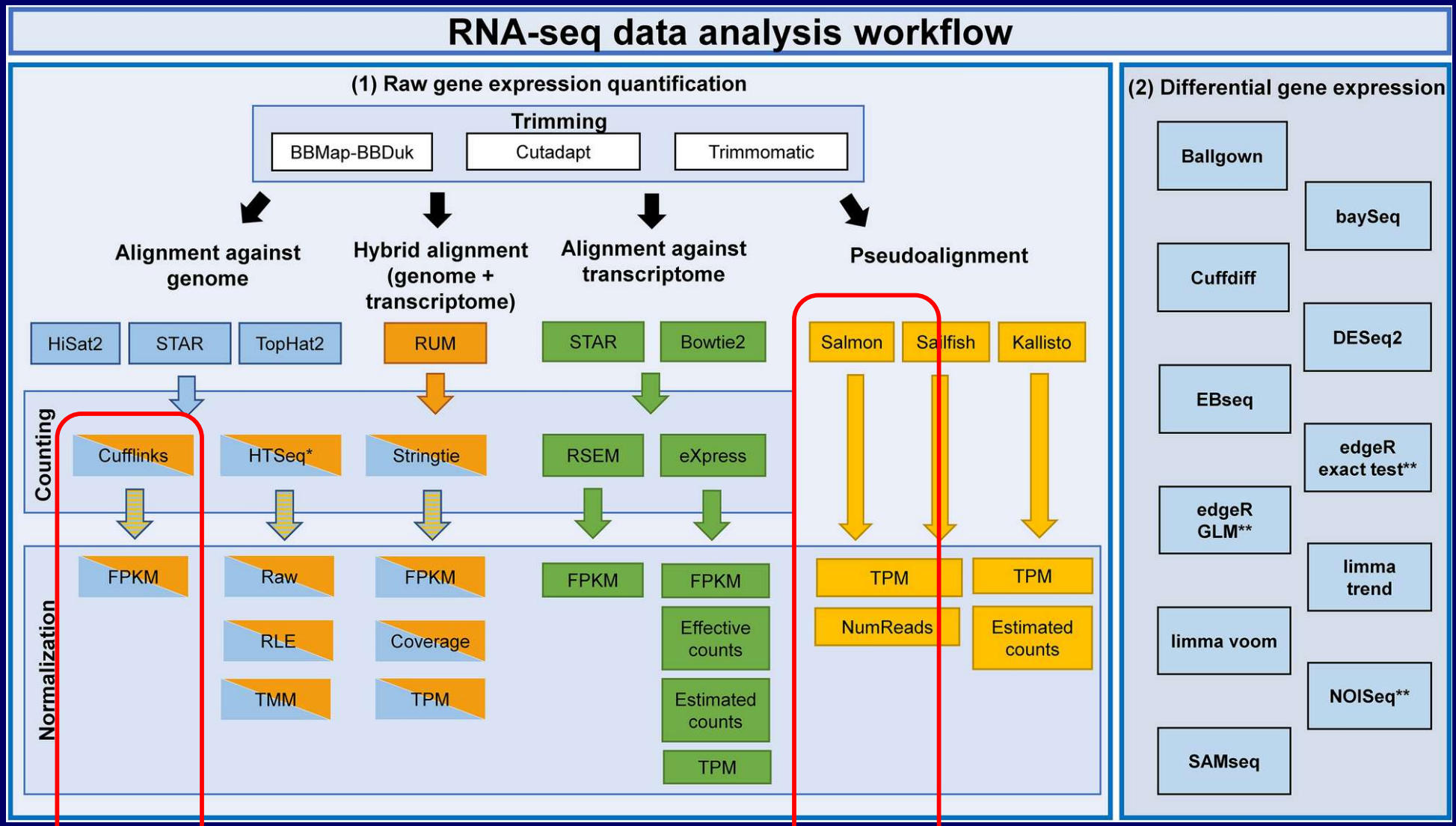
```
#!/bin/bash
#SBATCH --job-name=StarMapping
#SBATCH --output=StarMapping_%j.log
#SBATCH --mail-type=ALL
#SBATCH --mail-user=xxxx@ufl.edu
#SBATCH --time=24:00:00
#SBATCH --qos=YOURGROUP
#SBATCH --cpus-per-task=XXX
#SBATCH --mem-per-cpu=4gb
```

```
$ sbatch jobfile
```

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.

RNA-seq: Getting counts



Getting counts with cufflinks

```
#!/bin/bash
#SBATCH --job-name=cufflink_gms6014
#SBATCH --mail-type=ALL
#SBATCH --mail-user=xxxx@ufl.edu
#SBATCH --mem-per-cpu=4gb
#SBATCH --cpus-per-task=2
#SBATCH --qos=gms6014
#SBATCH --time=8:00:00
#SBATCH --output=cufflinks_%j.log

pwd; date

module load cufflinks

cufflinks -p 2   -o WG_young_1.clout ./starMap/WG_young_1Aligned.sortedByCoord.out.bam
cufflinks -p 2   -o WG_young_2.clout ./starMap/WG_young_2Aligned.sortedByCoord.out.bam
cufflinks -p 2   -o WG_old_1.clout ./starMap/WG_old_1Aligned.sortedByCoord.out.bam
cufflinks -p 2   -o WG_old_2.clout ./starMap/WG_old_2Aligned.sortedByCoord.out.bam
```

Practice: RNA-Seq analysis

Follow instructions on the course web site.