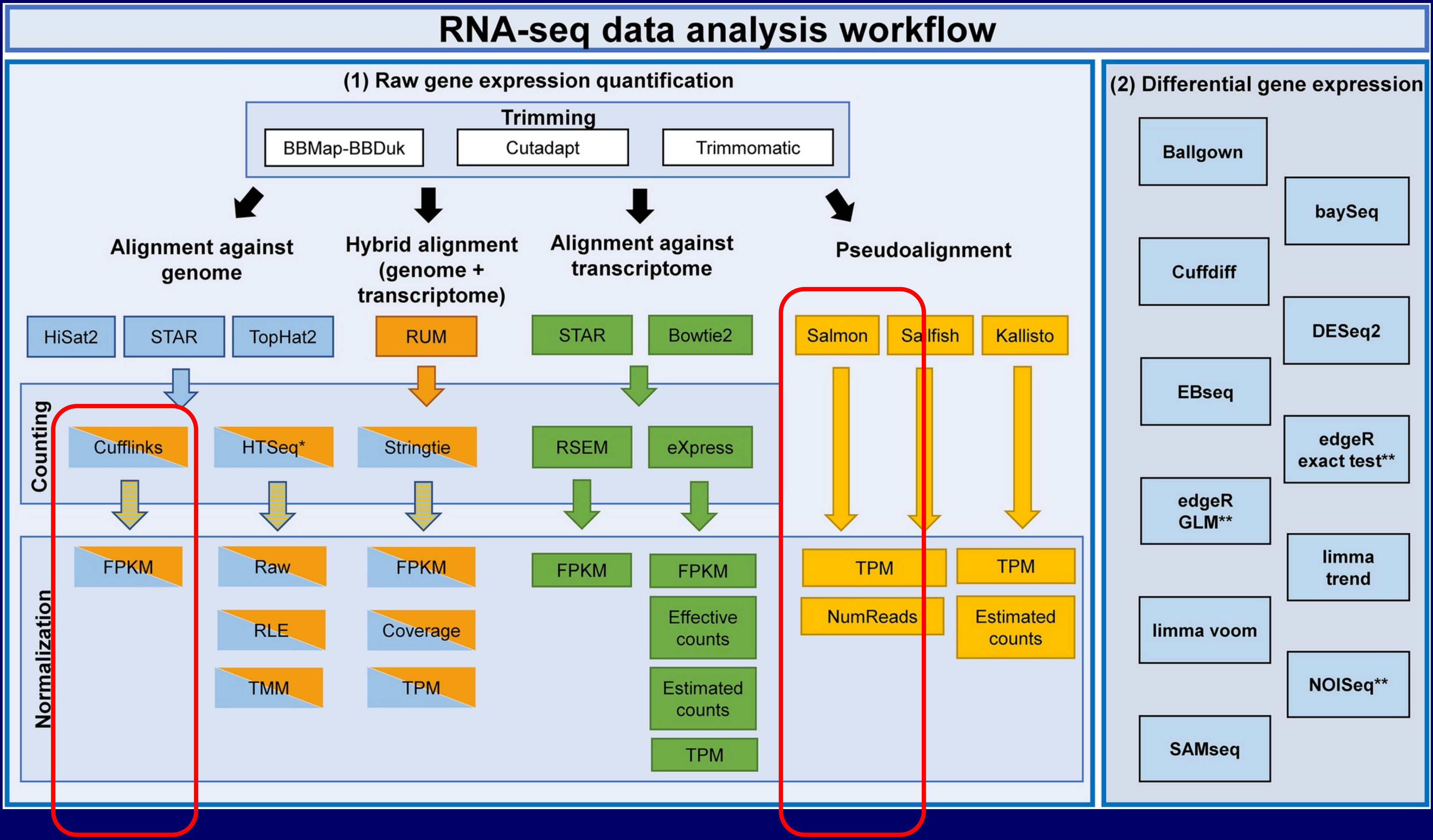
RNA-seq: Getting counts



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 Overview

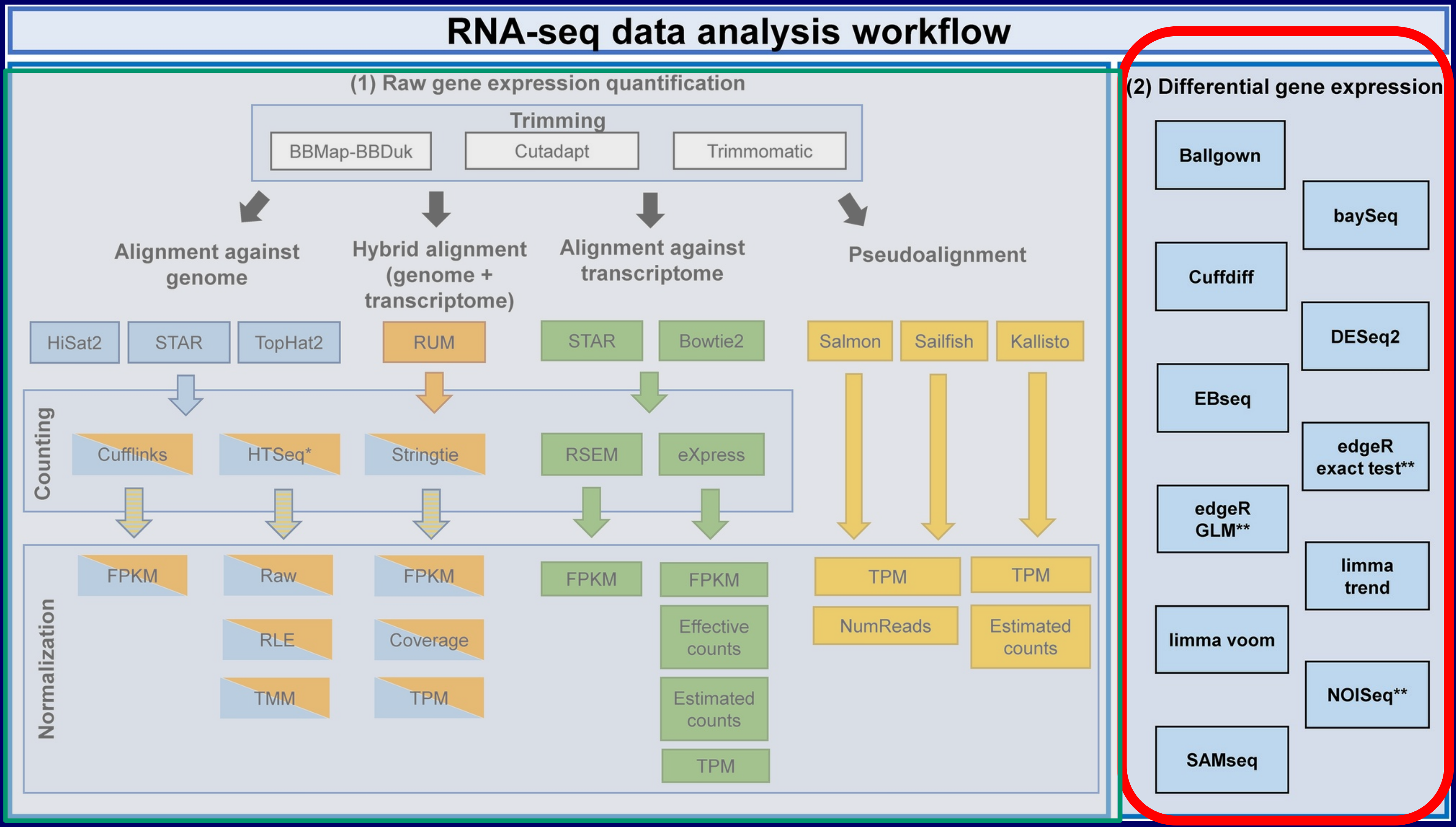
Four major steps, semi-independent of each other.

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

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

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

RNA-seq: Identify DEGs



RNA-seq data analysis workflow

(1) Raw gene expression quantification

Trimming

BBMap-BBDuk

Cutadapt

Trimmomatic

Alignment against genome

Hybrid alignment (genome + transcriptome)

Alignment against transcriptome

Pseudoalignment

HiSat2

STAR

TopHat2

RUM

STAR

Bowtie2

Salmon

Sailfish

Kallisto

Counting

Cufflinks

HTSeq*

Stringtie

RSEM

eXpress

Normalization

FPKM

Raw

FPKM

FPKM

FPKM

TPM

TPM

RLE

Coverage

Effective counts

NumReads

Estimated counts

TMM

TPM

Estimated counts

TPM

(2) Differential gene expression

Ballgown

baySeq

Cuffdiff

DESeq2

EBseq

edgeR exact test**

edgeR GLM**

limma trend

limma voom

NOISeq**

SAMseq

Many options at this stage. Personal favorites – Cuffdiff and DESeq2

Identification of Differentially Expressed Genes (DEGs)

```
module load cufflinks

## Frist merge the gtf files for samples to be compared.
ln /ufrc/gms6014/share/genome/dm6/annotation/genes.gtf dm6.gtf
ln /ufrc/gms6014/share/genome/dm6/sequence/genome.fa dm6.fa

cuffmerge -g dm6.gtf -s dm6.fa -p 2 WG_assemblies.txt
```

```
./WG_young_1.clout/transcripts.gtf
./WG_young_2.clout/transcripts.gtf
./WG_old_1.clout/transcripts.gtf
./WG_old_2.clout/transcripts.gtf
```

Representation of (HTS) data – BED (Browser Extensible Data) file

The picture can't be displayed.

Chrom.	Start	End	name	Scor	Strand
chr2	10000192	10000217	U0	0	+
chr2	10000227	10000252	U1	0	-
chr2	10000310	10000335	U2	0	+
chr3	10000496	10000521	U1	0	-
chr2	10000556	10000581	U2	0	+

Wit o
nee e
same as the reference genome).

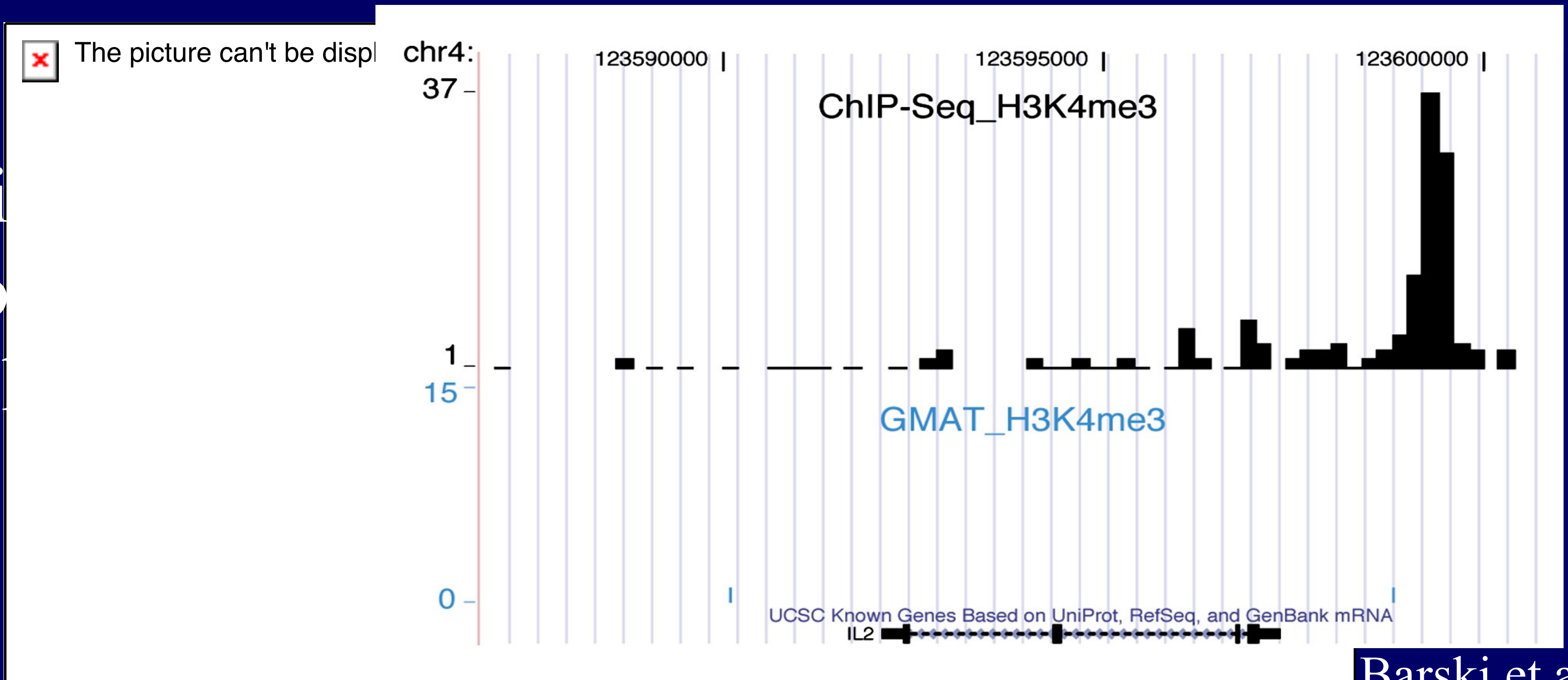
Detailed description of genomic data formats:
http://genome.ucsc.edu/FAQ/FAQformat.html

Representation of (HTS) data – Wig file

```
track type=wiggle_0 name="S_P53_XR60_A_treat_all"
description=”P53_XR"variableStep chrom=chr2L span=10
11      2
21      3
31      4
41      4
51      5
61      4
71      4
81      3
91      3
101     2
111     1
```

Visualization of HTS data.

Simple visualizati distributio (or normal values).


The picture can't be displ
chr4:
123590000
123595000
123600000
37
ChIP-Seq_H3K4me3
1
15
GMAT_H3K4me3
0
UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA
IL2

Barski et al. (2007) Cell

Chr.	ChrStart	ChrEnd	Value
chr4	0	200	0
chr4	200	400	2
chr4	400	600	13
chr4	600	800	35
chr4	800	1000	27

BedGraph file (Wig)

Visualizing Deep Seq data with UCSC genome browser

Practice & Observe I:

1. Load the track file as custom track to the browser by copy/past the URL link or upload the file.
2. View ‘dense’ and then ‘full’ presentation of the track.

Identification of differentially expressed genes (DEGs)

```
module load cufflinks

cuffdiff -o Old_v_Young -b ./index/Dm6.44.fa  -u Merged/merged.gtf  -p 2 -L youngWG,oldWG \
 ./starMap/WG_young_1Aligned.sortedByCoord.out.bam,./starMap/WG_young_2Aligned.sortedByCoord.out.bam \
 ./starMap/WG_old_1Aligned.sortedByCoord.out.bam,./starMap/WG_old_2Aligned.sortedByCoord.out.bam
```