INF-BIOx121 2017

RNA-seq

differential expression analysis

Arvind Sundaram

Sep 18-20, 2017

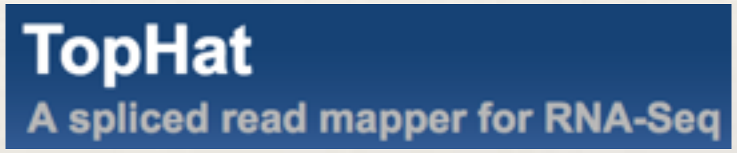
RNA-seq analysis

Case Study

Arvind Sundaram
Sep 18-20, 2017

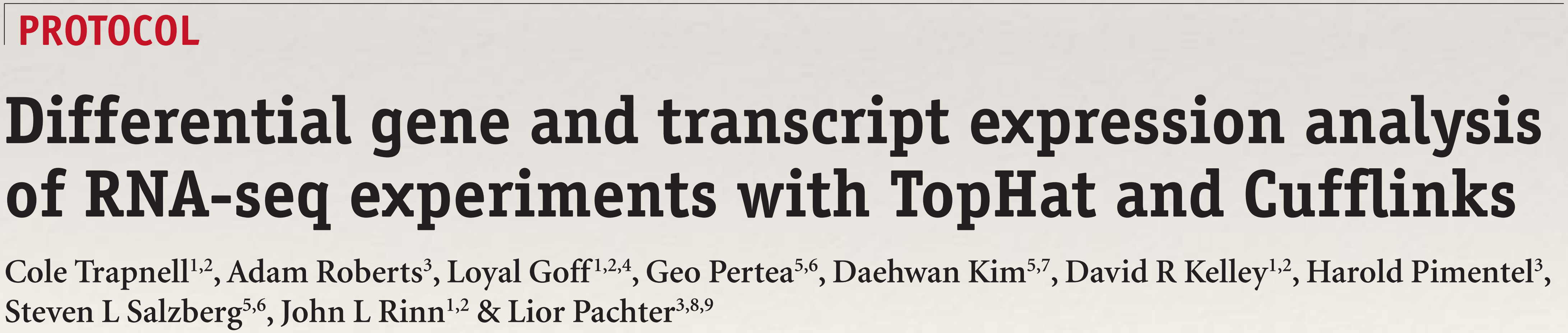
Case study

- Compare two conditions with three replicates
- *in silico* simulated dataset
- NCBI GEO: GSE32038
- DOI: 10.1038/nprot.2012.016



TopHat

A spliced read mapper for RNA-Seq



PROTOCOL

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

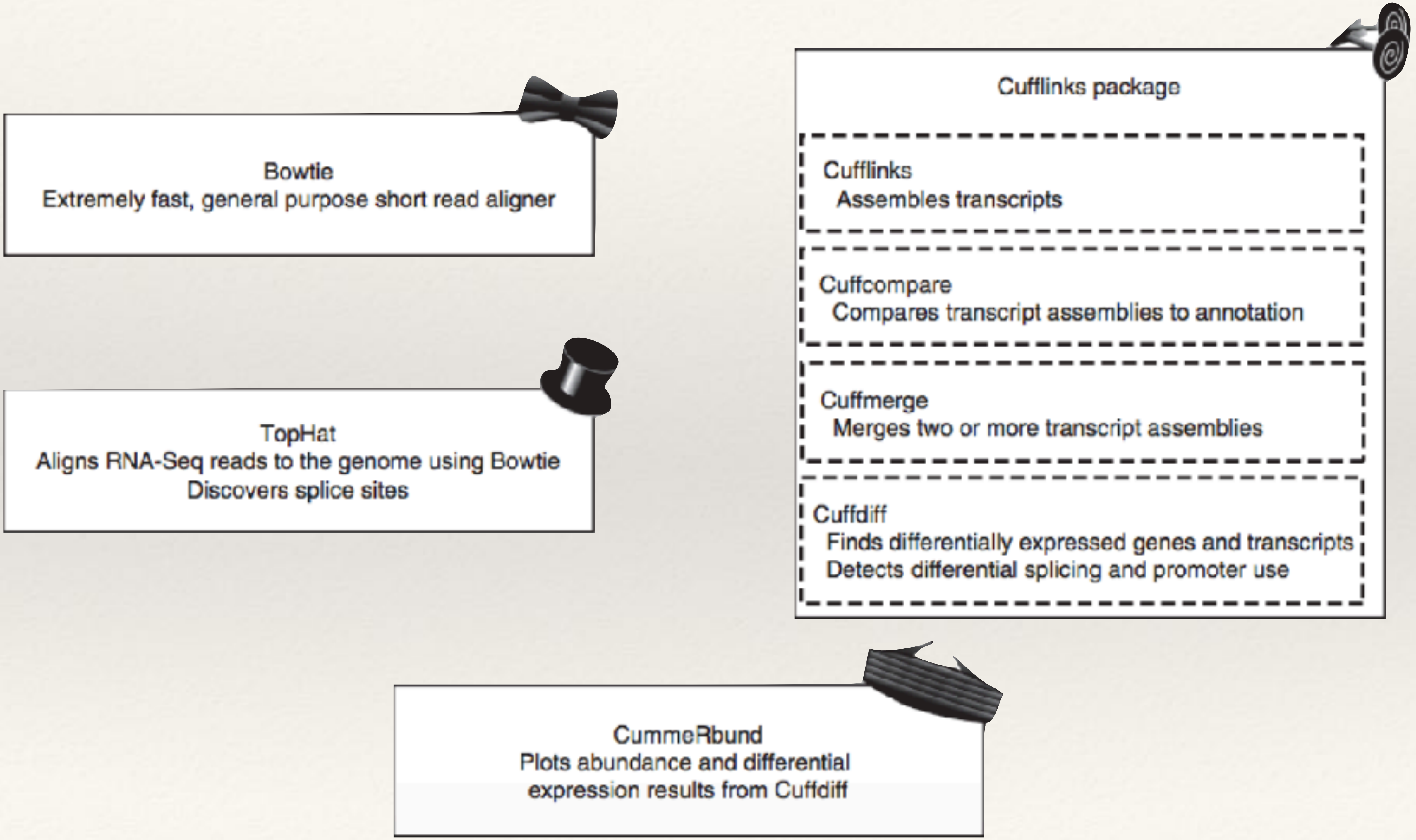
Cole Trapnell[1,2], Adam Roberts[3], Loyal Goff[1,2,4], Geo Pertea[5,6], Daehwan Kim[5,7], David R Kelley[1,2], Harold Pimentel[3], Steven L Salzberg[5,6], John L Rinn[1,2] & Lior Pachter[3,8,9]

HISAT2

graph-based alignment of next generation sequencing reads to a population of genomes

https://doi.org/10.1038/nprot.2012.016

TUXEDO pipeline



Bowtie
Extremely fast, general purpose short read aligner

TopHat
Aligns RNA-Seq reads to the genome using Bowtie
Discovers splice sites

Cufflinks package

Cufflinks
Assembles transcripts

Cuffcompare
Compares transcript assemblies to annotation

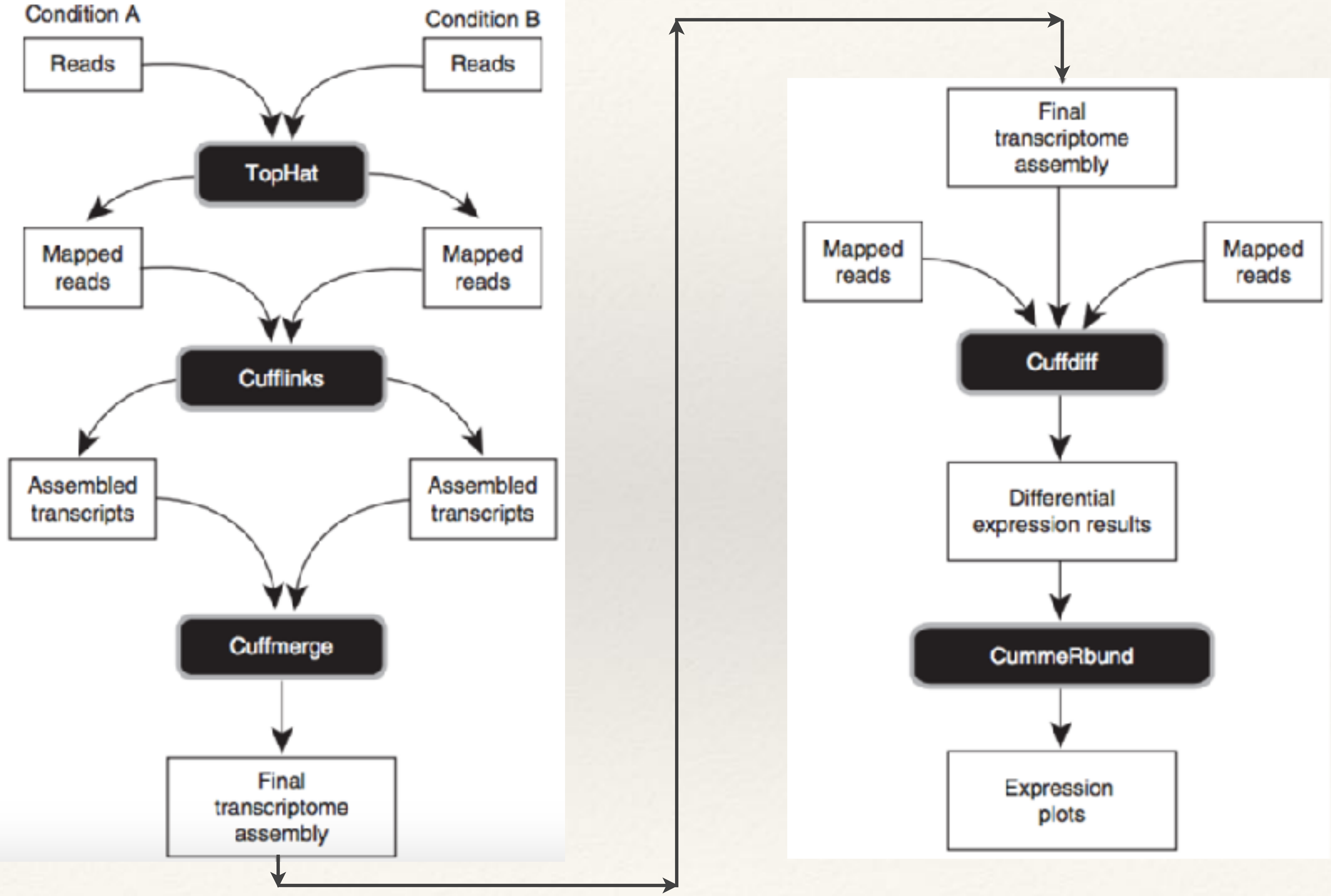
Cuffmerge
Merges two or more transcript assemblies

Cuffdiff
Finds differentially expressed genes and transcripts
Detects differential splicing and promoter use

CummeRbund
Plots abundance and differential
expression results from Cuffdiff

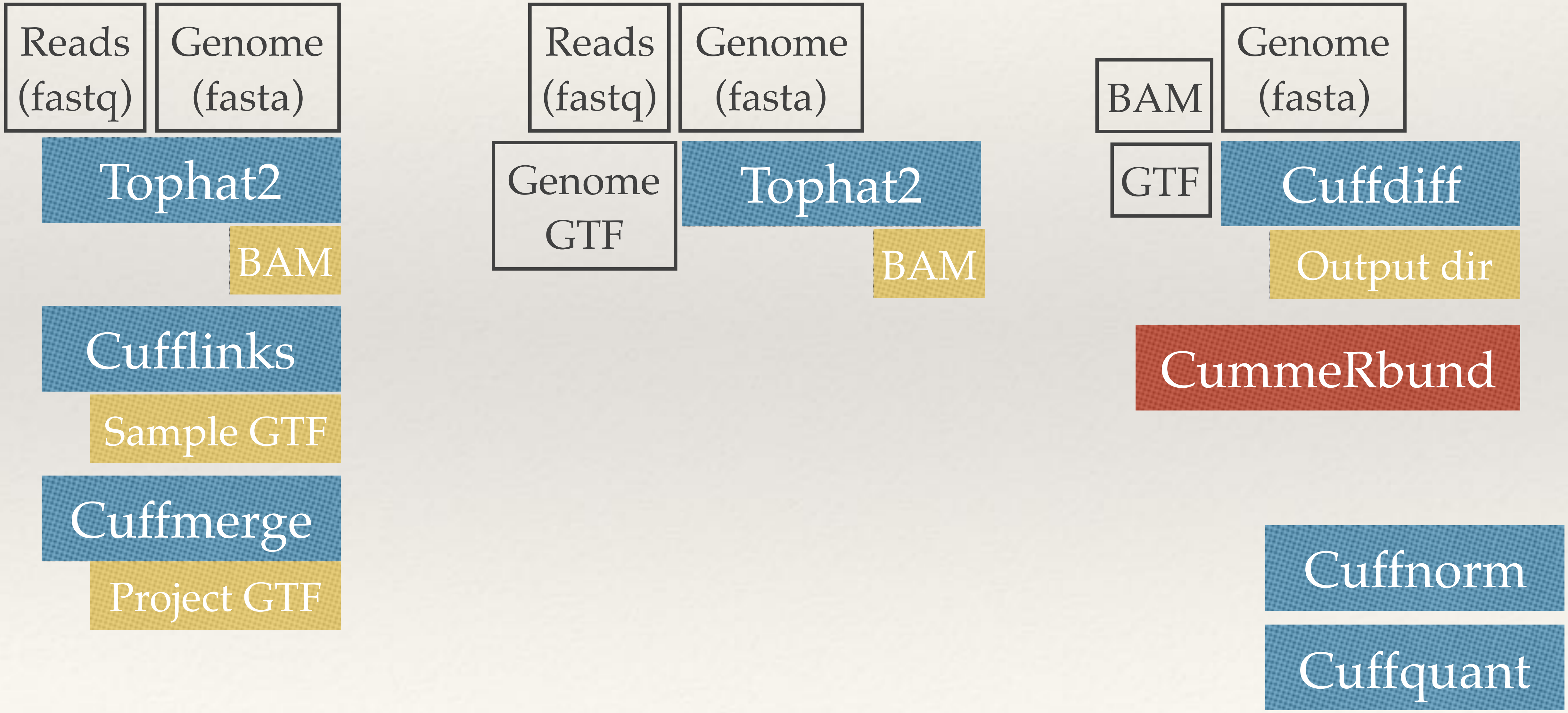
https://doi.org/10.1038/nprot.2012.016

TUXEDO pipeline



https://doi.org/10.1038/nprot.2012.016

TUXEDO pipeline



Genome

Reads (fastq)

Genome (fasta)

Tophat2

BAM

Cufflinks

Sample GTF

Cuffmerge

Project GTF

Genome + Transcriptome

Reads (fastq)

Genome (fasta)

Genome GTF

Tophat2

BAM

Differential Expression

BAM

Genome (fasta)

GTF

Cuffdiff

Output dir

CummeRbund

Cuffnorm

Cuffquant

https://doi.org/10.1038/nprot.2012.016

TUXEDO input

- Sequenced data - Fastq files
  - Single read
  - Paired end reads
  - pre-processed and cleaned*

Not necessary but a good practice

- Reference genome
- Reference annotation (GTF)*
  - Good to provide one if decent annotation exists

https://doi.org/10.1038/nprot.2012.016

Tophat aka Tophat2

- Tophat2 uses bowtie2 aligner engine
  - Bowtie2 is not a splice-aware aligner
  - Tophat2 is a splice-aware *aligner*
- Identifies potential exons and possible splice junctions in the genome and uses aligned data to confirm the same.

Handles **STRANDED** RNA data

https://doi.org/10.1038/nprot.2012.016

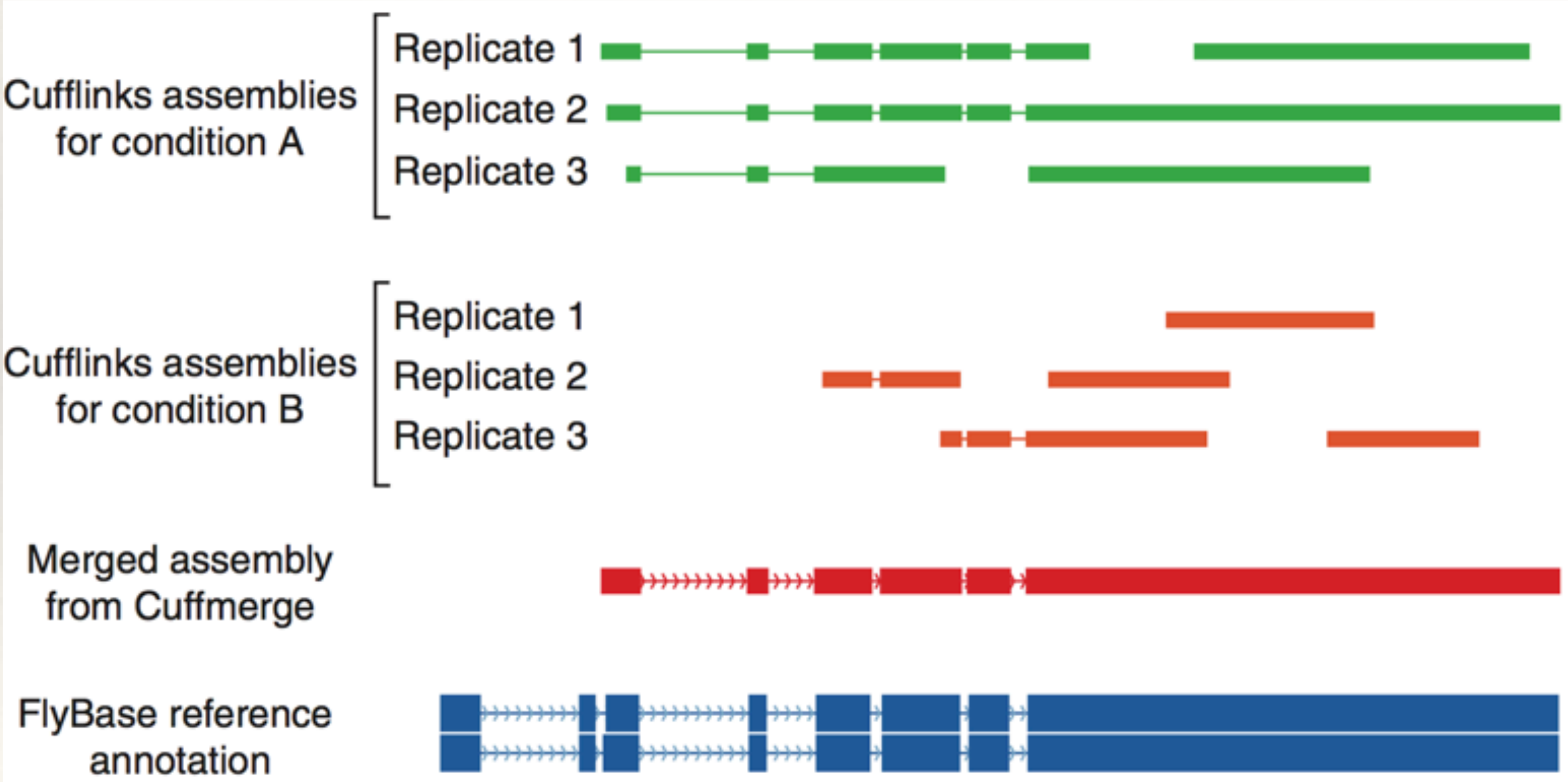
Cufflinks

- Transcript assembly
    - A parsimonious strategy to resolve isoforms
- First level transcript quantification
    - Immature vs mature transcripts

https://doi.org/10.1038/nprot.2012.016

Cuffmerge

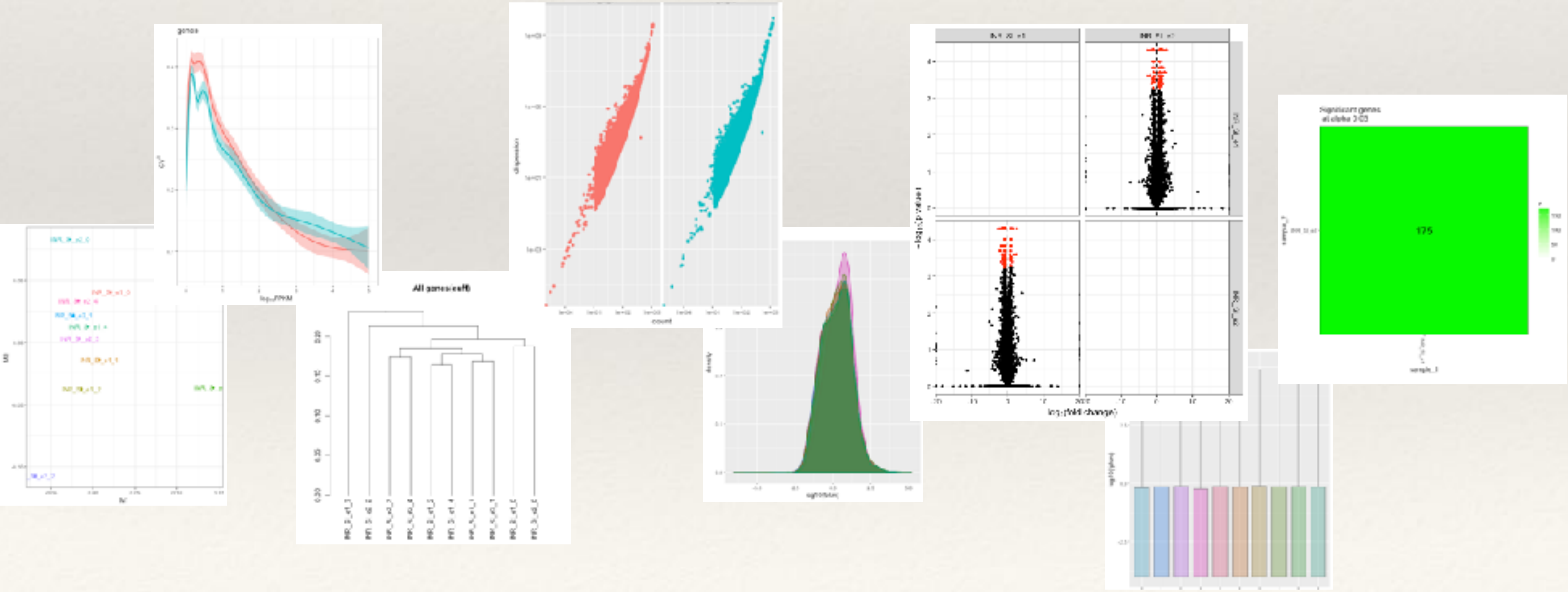
- Pooling of cufflinks data per sample to ensure proper overall experiment ”present transcripts” overview


Cufflinks assemblies for condition A
Replicate 1
Replicate 2
Replicate 3
Cufflinks assemblies for condition B
Replicate 1
Replicate 2
Replicate 3
Merged assembly from Cuffmerge
FlyBase reference annotation

https://doi.org/10.1038/nprot.2012.016

Cuffdiff

- Cuffdiff ″learns the variation for each gene across replicates″ to calculate differential expression
- CummeRbund in R used for visualisation



https://doi.org/10.1038/nprot.2012.016

RNA-seq analysis

Reference

- Prepare reference
  - Index genome using bowtie2-build
  - If you are using the annotation in GTF format, you ‘tophat2’ to create a ‘transcriptome index’

```
$ cd
$ cd Desktop
$ mkdir rna_seq
$ cd rna_seq
$ mkdir reference
$ cd reference
$ ln -s /data/RNA-seq/reference/* .
```

script.sh

```
bowtie2-build genome.fa genome
tophat2 -G genes.gtf --transcriptome-index=known genome
```

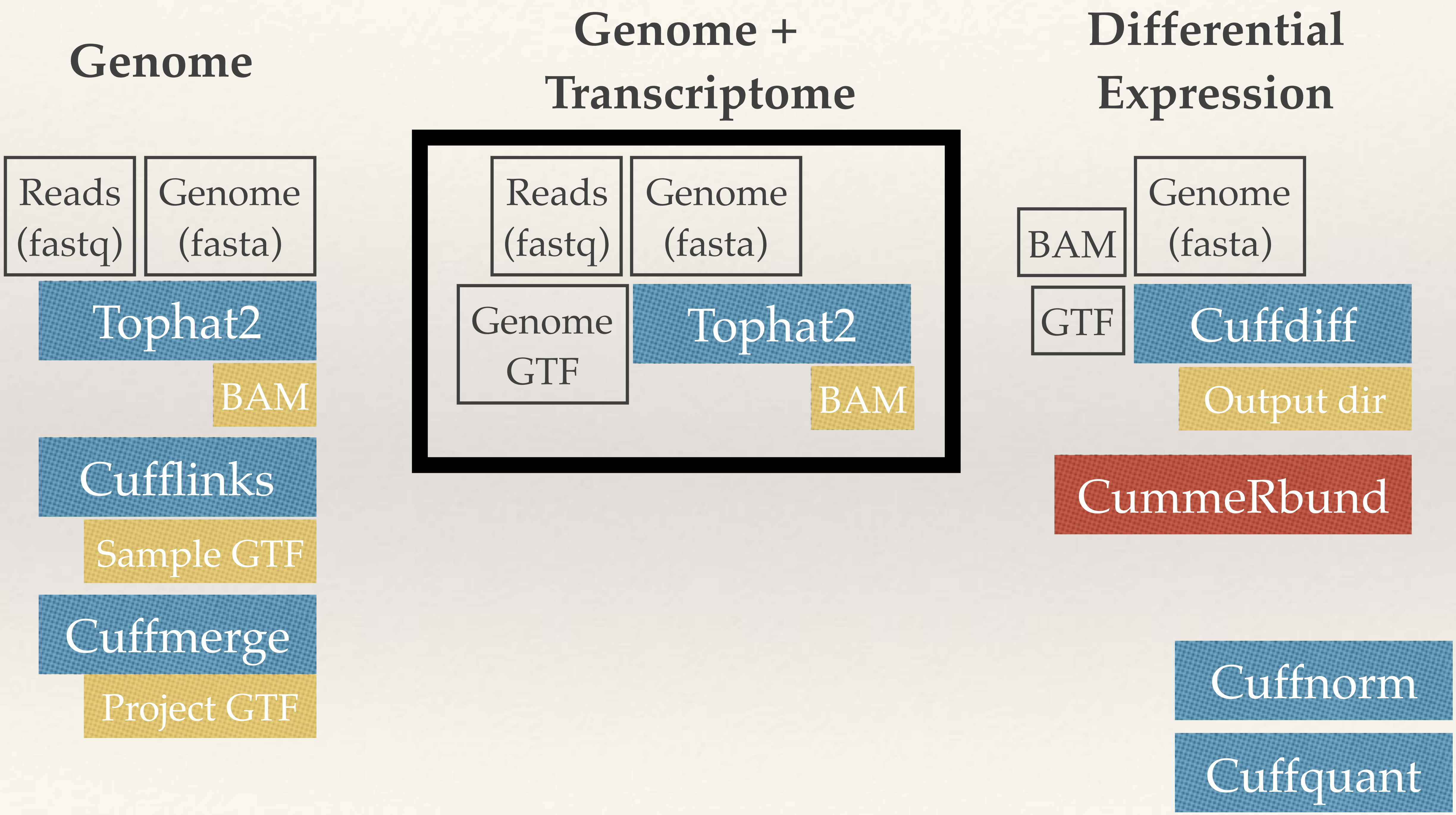
Raw data

- Compare two conditions (C1, C2) with three replicates (R1, R2, R3)
- *in silico* simulated dataset from *Drosophila melanogaster*
- NCBI GEO: GSE32038

```
$ cd
#
$ check if you are in your home page
$ cd Desktop
$ mkdir rna_seq
$ cd rna_seq
$ mkdir 00_raw_data
$ cd 00_raw_data
$ ln -s /data/RNA-seq/00_raw_data/C* .

## Run fastQC to check raw data
```

TUXEDO pipeline



Genome

Reads (fastq)

Genome (fasta)

Tophat2

BAM

Cufflinks

Sample GTF

Cuffmerge

Project GTF

Genome + Transcriptome

Reads (fastq)

Genome (fasta)

Genome GTF

Tophat2

BAM

Differential Expression

BAM

Genome (fasta)

GTF

Cuffdiff

Output dir

CummeRbund

Cuffnorm

Cuffquant

https://doi.org/10.1038/nprot.2012.016

Tophat2

- Raw data is available in folder `raw_data`
- Tophat2 has to be run for individual samples - 6 times for this case study

```
tophat2 <options> -o output_folder genome_bowtie2_idx Read1 Read2
```

```
tophat2
 -p 8
 -G reference/genes.gtf
 --transcriptome-index=reference/known
 -o C1_R1_thout
 reference/genome
 00_raw_data/C1_R1_1.fq.gz
 00_raw_data/C1_R1_2.fq.gz
```

Tophat2

- If your tophat2 has not completed, copy the output as below

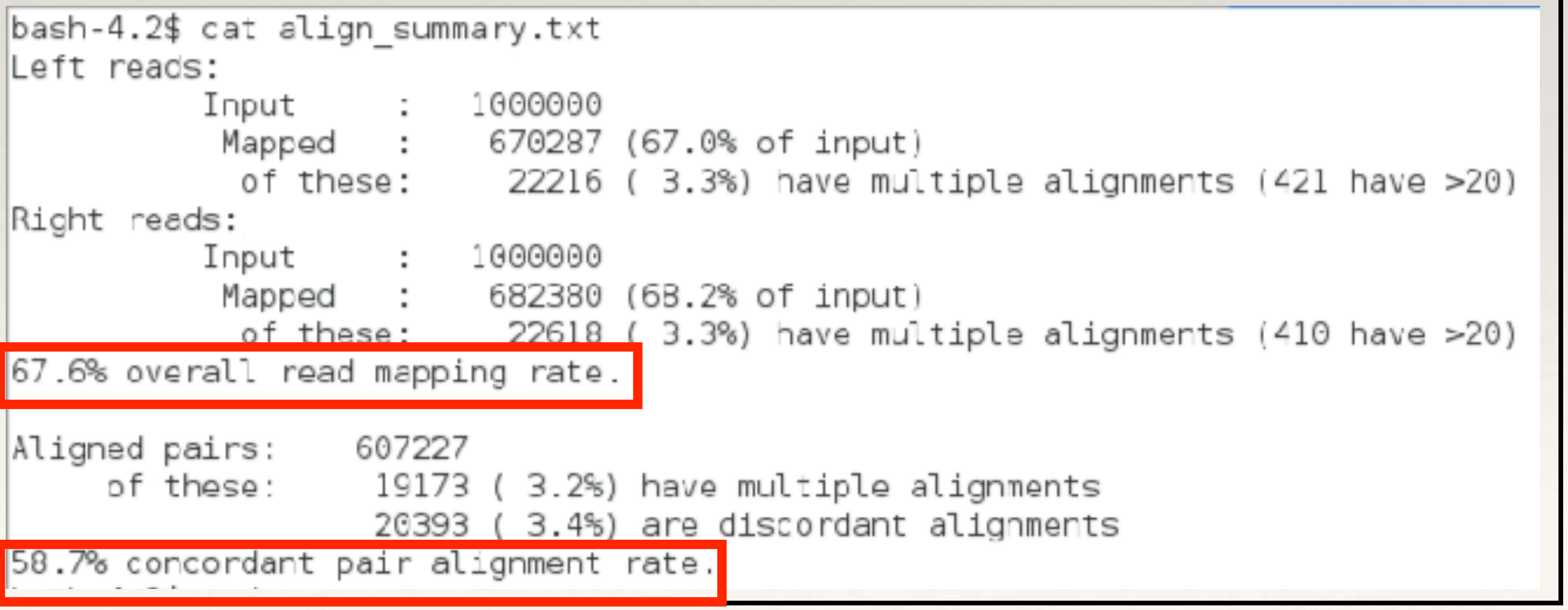
```
$ cd
$ cd Desktop
$ cd rna_seq
$ mkdir 10_tophat
$ cd 10_tophat
$ cp /data/RNA-seq/10_tophat/C1_R1_thout.tar .
$ tar -xvf C1_R1_thout.tar
```

Tophat2

- Tophat2 produces a lot of output files in the directory
  - `accepted_hits.bam` contain the aligned data
    - mapped reads only
  - `align_summary.txt` contains mapping statistics

`tophat2 output dir`

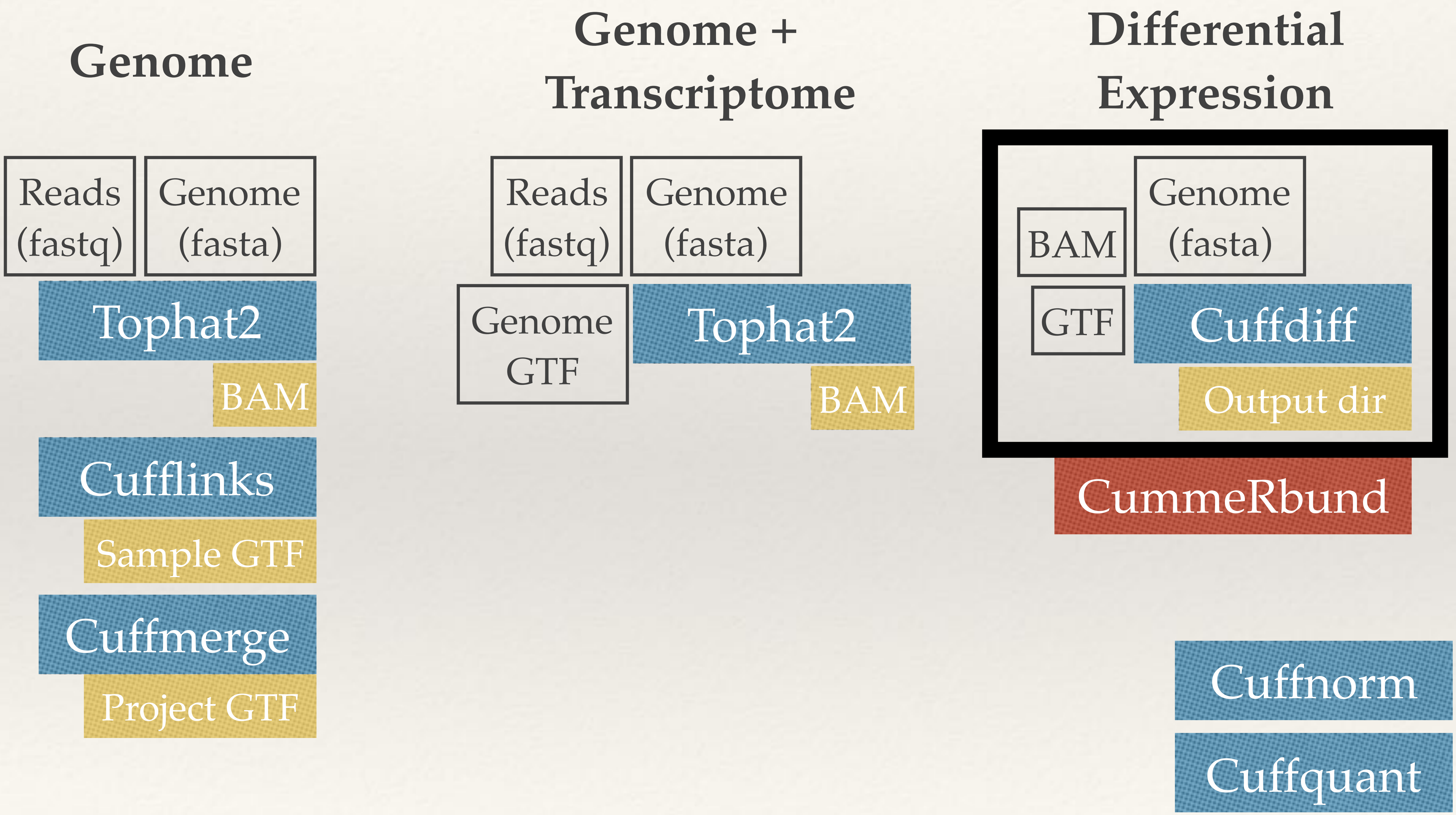
```
C1_R1_thout/
  accepted_hits.bam
  align_summary.txt
  deletions.bed
  insertions.bed
  junctions.bed
  logs/
  prep_read.info
  unmapped.bed
```



```
bash-4.2$ cat align_summary.txt
Left reads:
          Input     :   1000000
           Mapped   :    670287 (67.0% of input)
            of these:     22216 ( 3.3%) have multiple alignments (421 have >20)
Right reads:
          Input     :   1000000
           Mapped   :    682380 (68.2% of input)
            of these:     22618 ( 3.3%) have multiple alignments (410 have >20)
67.6% overall read mapping rate.

Aligned pairs:    607227
     of these:     19173 ( 3.2%) have multiple alignments
                   20393 ( 3.4%) are discordant alignments
58.7% concordant pair alignment rate.
```

TUXEDO pipeline



Genome

Reads (fastq)

Genome (fasta)

Tophat2

BAM

Cufflinks

Sample GTF

Cuffmerge

Project GTF

Genome + Transcriptome

Reads (fastq)

Genome (fasta)

Genome GTF

Tophat2

BAM

Differential Expression

BAM

Genome (fasta)

GTF

Cuffdiff

Output dir

CummeRbund

Cuffnorm

Cuffquant

https://doi.org/10.1038/nprot.2012.016

Cuffdiff

- Cuffdiff calculates differential expression between two conditions
  - takes care of replicates
  - produces statistical information

```
cuffdiff
  -p 8
  -b reference/genome.fa
  -u reference/genes.gtf
  -o diff_out
  -L C1,C2

  C1_R1_thout/accepted_hits.bam,
  C1_R2_thout/accepted_hits.bam,
  C1_R3_thout/accepted_hits.bam

  C2_R1_thout/accepted_hits.bam,
  C2_R2_thout/accepted_hits.bam,
  C2_R3_thout/accepted_hits.bam
```

CummeRbund

```
$ cd
$ cd Desktop/rna_seq
$ cp /data/RNA-seq/20_cuffdiff.tar .
$ tar -xvf 20_cufflinks.tar
$ mkdir 30_cummeRbund
$ cd 30_cummeRbund

## R using Rstudio
$ rstudio

> getwd()
# should point to 30_cummeRbund
> library(“cummeRbund”)
> cuff <- readcuffinks(“../20_cufflinks”)
> cuff
```

CummeRbund

```
> dispersionPlot(genes(cuff))
> csDensity(genes(cuff), replicates=T)
> csBoxplot(genes(cuff), replicates=T)
> csScatterMatrix(genes(cuff))
> csDendro(genes(cuff), replicates=T)
> fpkmSCVPlot(genes(cuff))
> csVolcanoMatrix(genes(cuff))
> MDSplot(genes(cuff), replicates=T) > sigMatrix(cuff)
> sigMatrix(cuff, level=“isoforms”)

> diff.genes <- diffData(genes(cuff))
> annot.genes <- annotation(genes(cuff))[,c(1,4)]
> diff.genes.annot <- merge(diff.genes, annot.genes, by = 'gene_id')
> diff.genes.sig <- subset(diff.genes.annot, significant==“yes”)
> write.table(diff.genes.sig, ‘DE_cuff_genes.txt’, quote=F, sep=“/t”)

> diff.iso <- diffData(isoforms(cuff))
> annot.iso <- annotation(isoforms(cuff))[,c(1,2,4)]
> diff.iso.annot <- merge(diff.iso, annot.iso, by = ‘isoform_id')
> diff.iso.sig <- subset(diff.iso, significant==“yes”)
> write.table(diff.iso.sig, ‘DE_cuff_isoforms.txt’, quote=F, sep=“/t”)
```

featureCounts

```
## Correct

$ featureCounts -p -s 2 -a ../reference/genes.gtf -o
counts_paired_stranded ../35_tophat_for_featureCounts/*bam

## Try the following, run DESeq2 and check the difference from
above

$ featureCounts -p -a ../reference/genes.gtf -o
counts_paired ../35_tophat_for_featureCounts/*bam

$ featureCounts -a ../reference/genes.gtf -o counts ../
35_tophat_for_featureCounts/*bam
```

DESeq2

```
> library(“DESeq2”)
# Check folder
> count <- read.delim(‘counts_paired_stranded’, skip=1, header=T)
> data <- count[,c(7:12)]
> rownames(data) <- count[,1]
> colnames(data) <- c('C1_R1', 'C1_R2', 'C1_R3', 'C2_R1', 'C2_R2',
 'C2_R3')

> colData <- data.frame(condition = c('C1', 'C1', 'C1', 'C2', 'C2',
 'C2'))
> rownames(colData) <- colnames(data)

> dds <- DESeqDataSetFromMatrix(countData = data, colData = colData,
design=~condition)
> dds <- DESeq(dds)
> res <- results(dds)
> summary(res)

> plotDispEsts(dds)
> plotPCA(DESeqTransform(dds)
> plotMA(dds)
```