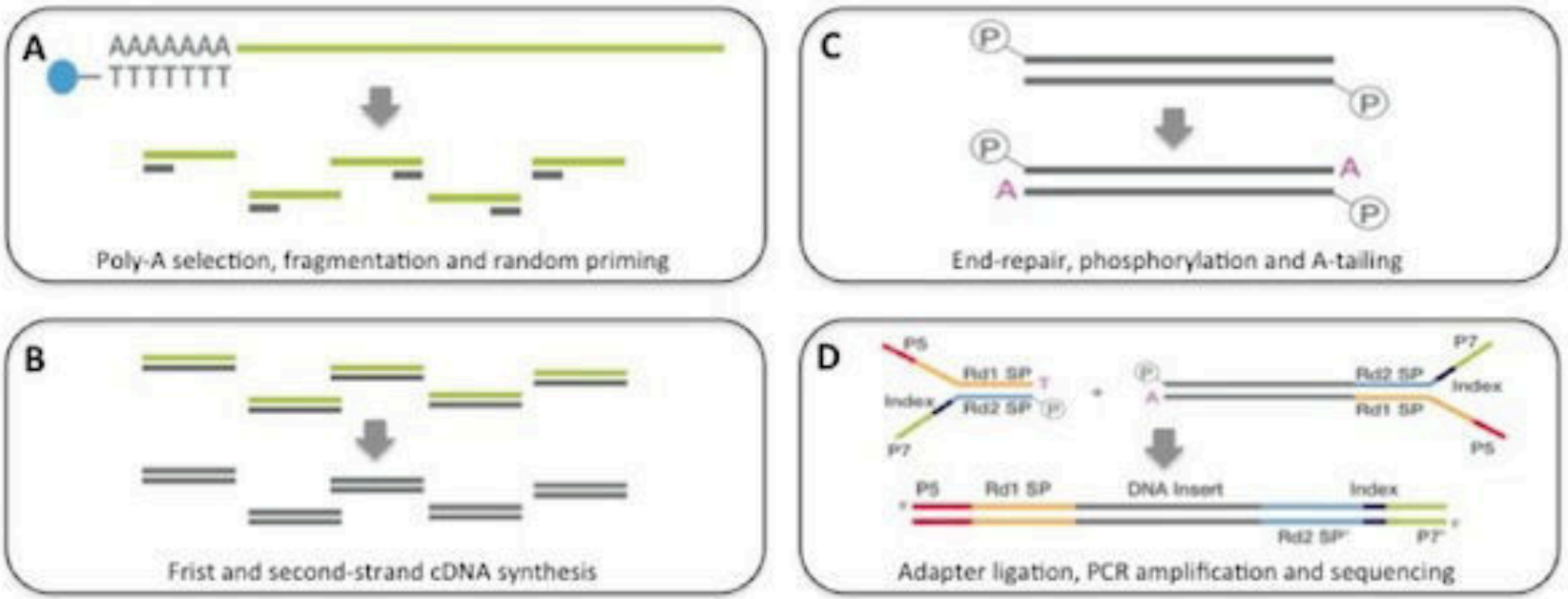
Galaxy RNAseq tutorial - GCC2014

Saskia Hiltemann (ErasmusMC)
Youri Hoogstrate (ErasmusMC)
Leon Mei (LUMC)

June 29, 2014

RNA-seq protocol

Illumina Tru-Seq RNA-seq protocol



Library prep begins from 100ng-1ug of Total RNA which is poly-A selected (A) with magnetic beads. Double-stranded cDNA (B) is phosphorylated and A-tailed (C) ready for adapter ligation. The library is PCR amplified (D) ready for clustering and sequencing.

http://bitesizebio.com/13542

Sequencers: HiSeq

Figure : HiSeq 2000.

Characteristics:

- High throughput.
- Paired end.
- High accuracy.
- Read length 2×150bp.
- Relatively long run time.
- Relatively expensive.

Sequencers: Ion Proton

Figure : Ion torrent.

Characteristics:

- Moderate throughput.
- High accuracy.
- Read length 200bp.
- Short run time
- Cheap runs.
- Homopolymer issue.

General layout of an RNA-seq pipeline.

1. Pre-alignment.
   - QC.
   - Data cleaning.

General layout of an RNA-seq pipeline.

1. Pre-alignment.
   - QC.
   - Data cleaning.
2. Alignment.
   - Use a specialised (RNA) aligner.

General layout of an RNA-seq pipeline.

1. Pre-alignment.
   - QC.
   - Data cleaning.
2. Alignment.
   - Use a specialised (RNA) aligner.
3. Expression (gene, transcripts) analysis.
   - Known transcripts.

General layout of an RNA-seq pipeline.

1. Pre-alignment.
   - QC.
   - Data cleaning.
2. Alignment.
   - Use a specialised (RNA) aligner.
3. Expression (gene, transcripts) analysis.
   - Known transcripts.
4. Transcript assembly.
   - New transcripts, alternative splicing, etc.

Data cleaning

We use Sickle for data cleaning.

Data cleaning

We use Sickle for data cleaning.

For adapter clipping, we can use Cutadapt, Trimmomatic or the FastX toolkit (not in this practical).

- Remove linker/adapter sequences.
- Trim low quality reads at the end of the read.
- Evaluate the part of the read that is left.

Data cleaning

We use Sickle for data cleaning.

For adapter clipping, we can use Cutadapt, Trimmomatic or the FastX toolkit (not in this practical).

- Remove linker/adapter sequences.
- Trim low quality reads at the end of the read.
- Evaluate the part of the read that is left.

The FastQC tool kit is used for quality control (both before and after the data cleaning step).

- GC content.
- GC distribution.
- Quality scores distribution.

FastQC report

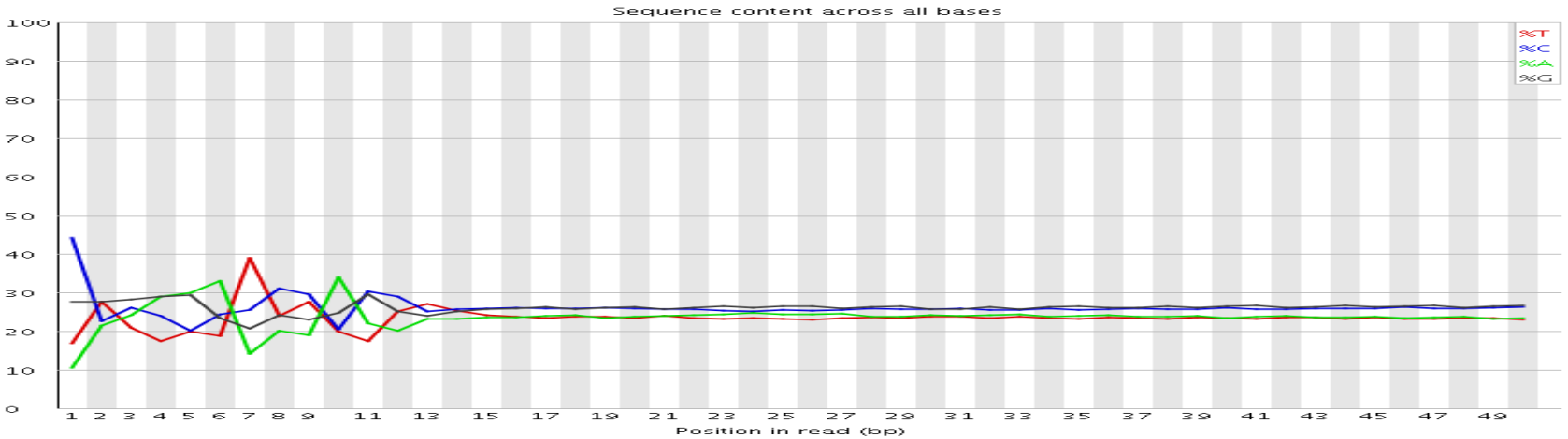


Figure : Per base sequence content. (ref. PMC2896536)

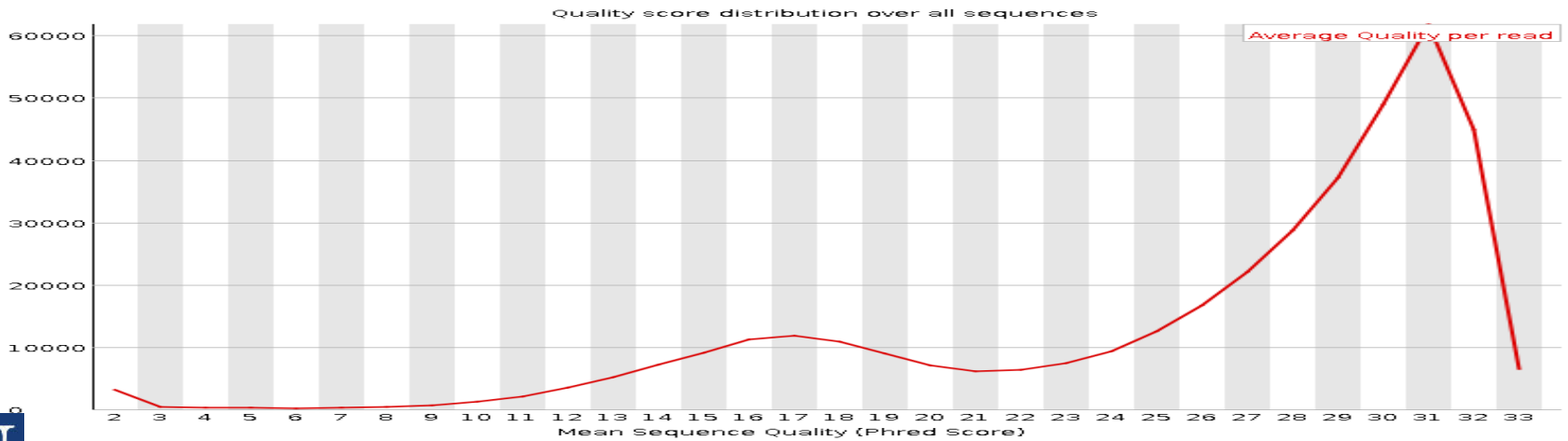
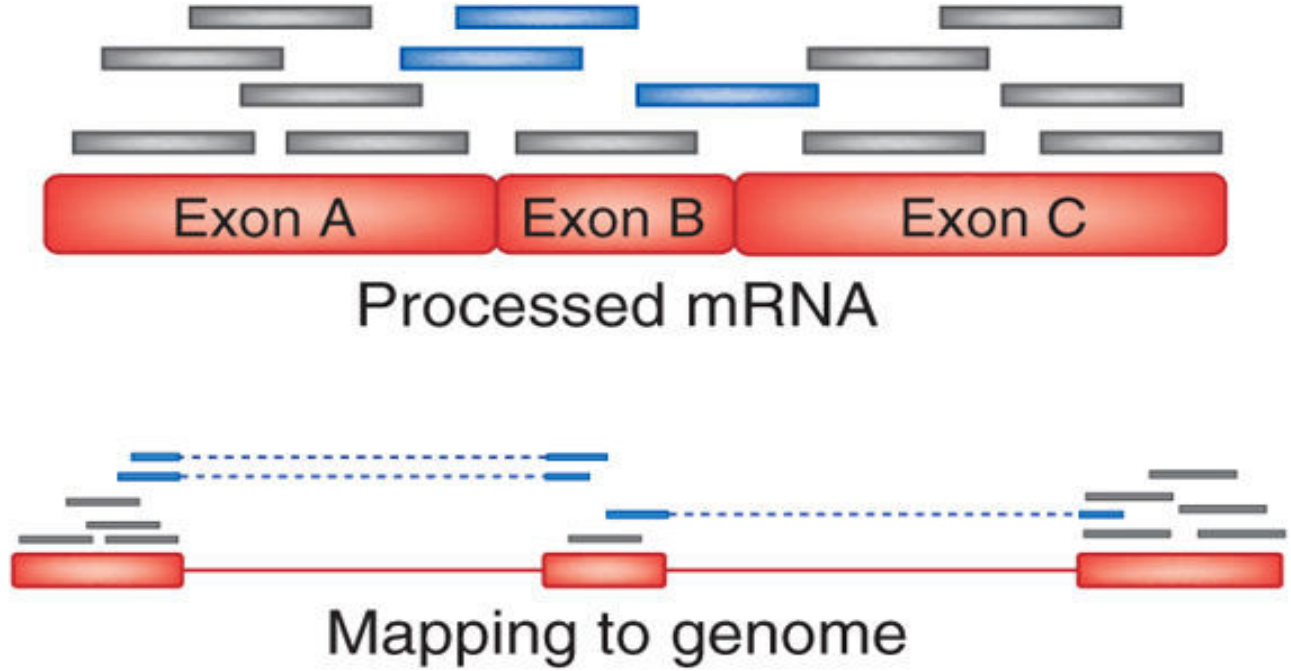


Figure : Per sequence quality.

RNAseq alignment



http://www.nature.com/nbt/journal/v27/n5/full/nbt0509-455.html

RNAseq alignment

Aligning RNAseq reads to reference genome

- Alternative splicing.

RNAseq alignment

Aligning RNAseq reads to reference genome

- Alternative splicing.

This affects:

- Insert sizes.
- Mapping of reads that cover an exon-exon boundary.

RNAseq alignment

Aligning RNAseq reads to reference genome

- Alternative splicing.

This affects:

- Insert sizes.
- Mapping of reads that cover an exon-exon boundary.

Available RNAseq aligners

- Tophat.
- Gmap / Gsnap.
- STAR.
- GEM.
- HMMSplicer.

http://www.nature.com/nmeth/journal/v10/n12/full/nmeth.2722.html

Tophat

http://ccb.jhu.edu/software/tophat/

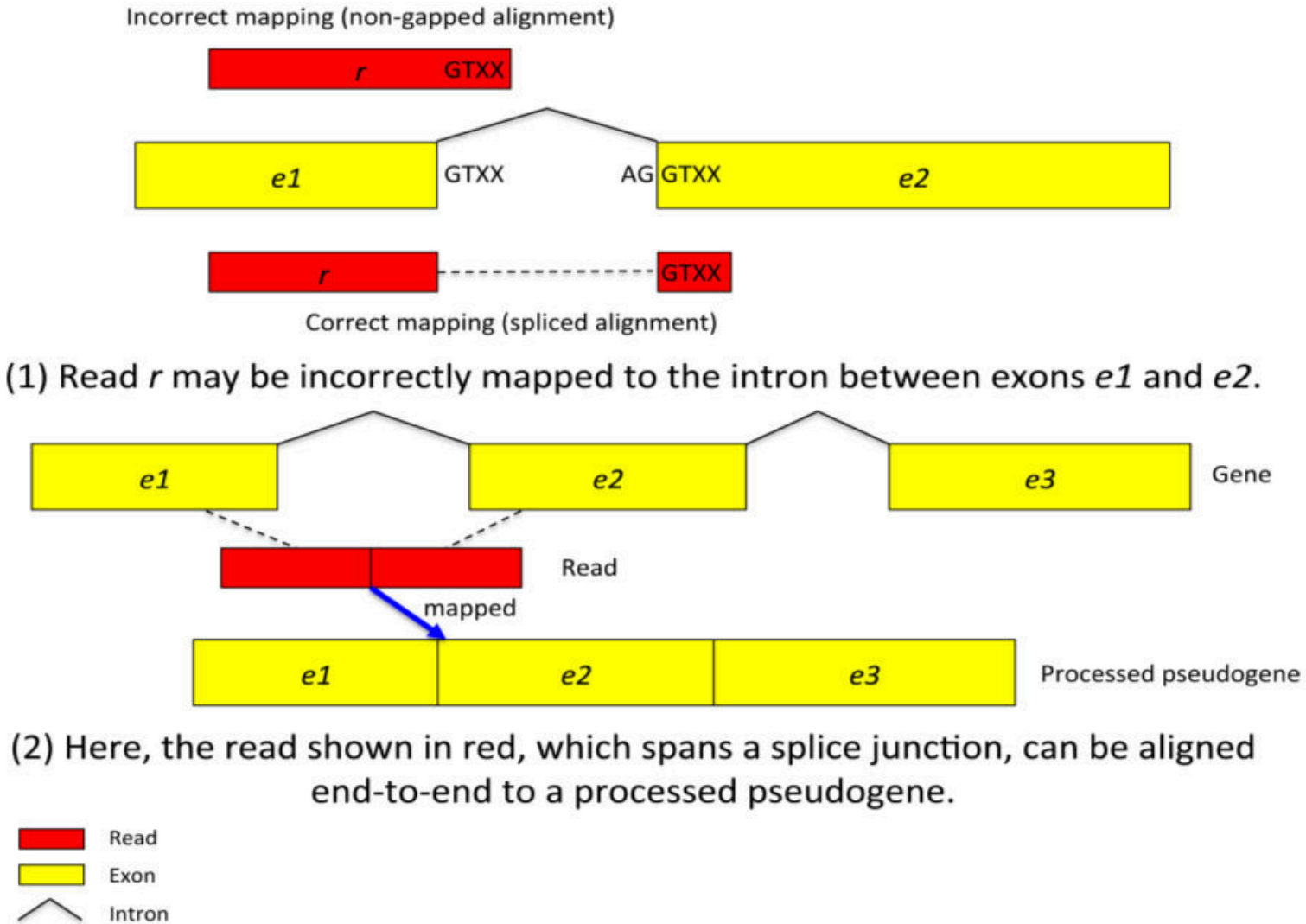
Two-step approach:

- (optional) Align to transcriptome first.
- Use bowtie to align whole reads, identify potential exon.
- Split left-over reads into small segments, align independently.
  - Make a database of splice junctions.
  - Map the reads to confirm the splice junctions.

Some considerations:

- Does not support soft-clipping

Tophat alignment



http://genomebiology.com/2013/14/4/R36

STAR

https://code.google.com/p/rna-star/

- Selected by many large sequencing consortium, e.g., Geuvadis.
- High alignment yield with good alignment quaility (close to tophat).
- Very very fast.

Some considerations:

- Require large amount of memory (more than 30GB).

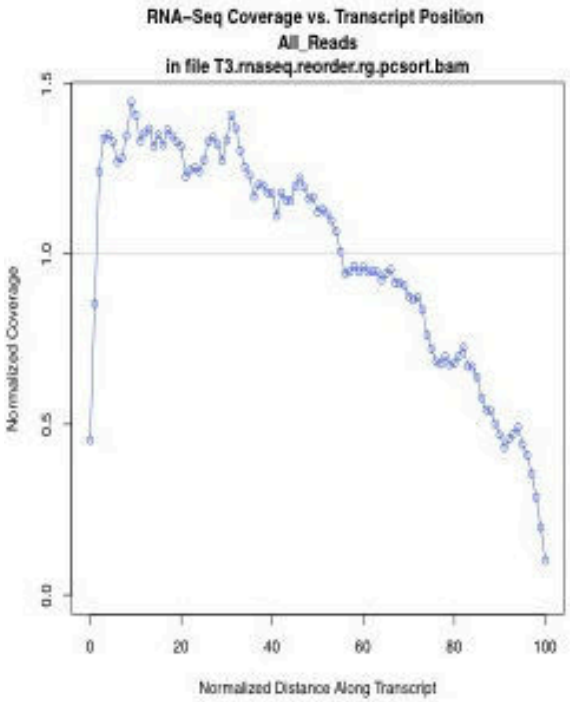
Gmap/Gsnap

http://research-pub.gene.com/gmap/

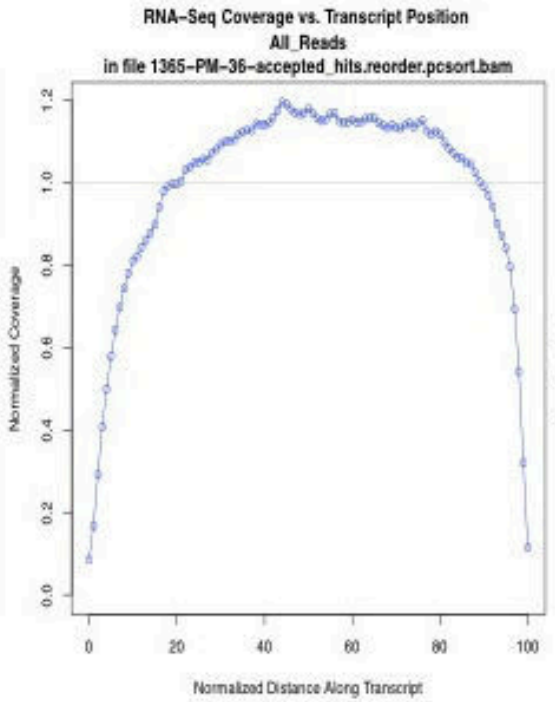
- Before tophat2 and STAR
  - Split both ends.
  - Split a read into many pieces.
- Fast.
- Memory efficient.

Post alignment QC (5'-3' coverage bias)

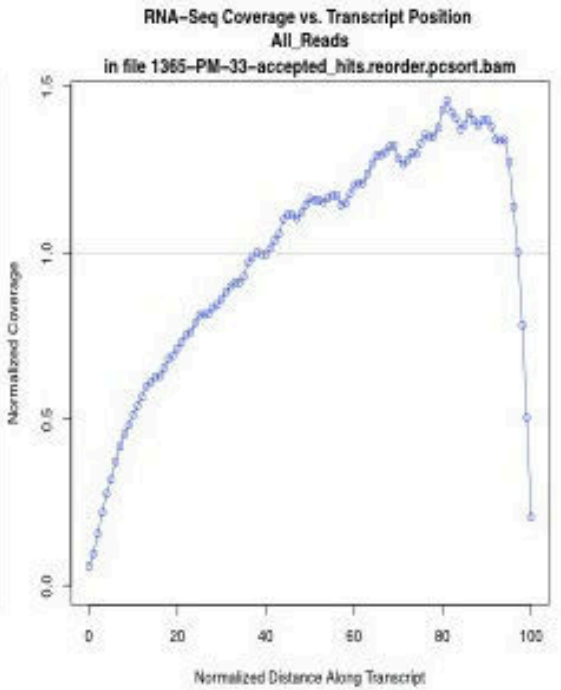
5' bias
(strand oriented protocol)



no bias
(low coverage at ends of transcript)



3' bias
(poly-A selection)



Zuojian Tang NYU Langone MEDICAL CENTER

https://www.biostars.org/p/102812/

Cufflinks

Input:

- Aligned reads.
  - Gmap / Gsnap.
  - Tophat.

Cufflinks

Input:

- Aligned reads.
  - Gmap / Gsnap.
  - Tophat.

What it can do:

- Assemble transcripts.
- Estimate transcript abundance.

Cufflinks

Modes of operation:

- Use predefined transcripts.
- Assemble transcripts assisted by known transcripts.
- Assemble transcripts with no prior knowledge.

Cufflinks

Modes of operation:

- Use predefined transcripts.
- Assemble transcripts assisted by known transcripts.
- Assemble transcripts with no prior knowledge.

When to use:

- Only interested in expression.
- Alternative splicing.

Cufflinks

Modes of operation:

- Use predefined transcripts.
- Assemble transcripts assisted by known transcripts.
- Assemble transcripts with no prior knowledge.

When to use:

- Only interested in expression.
- Alternative splicing.

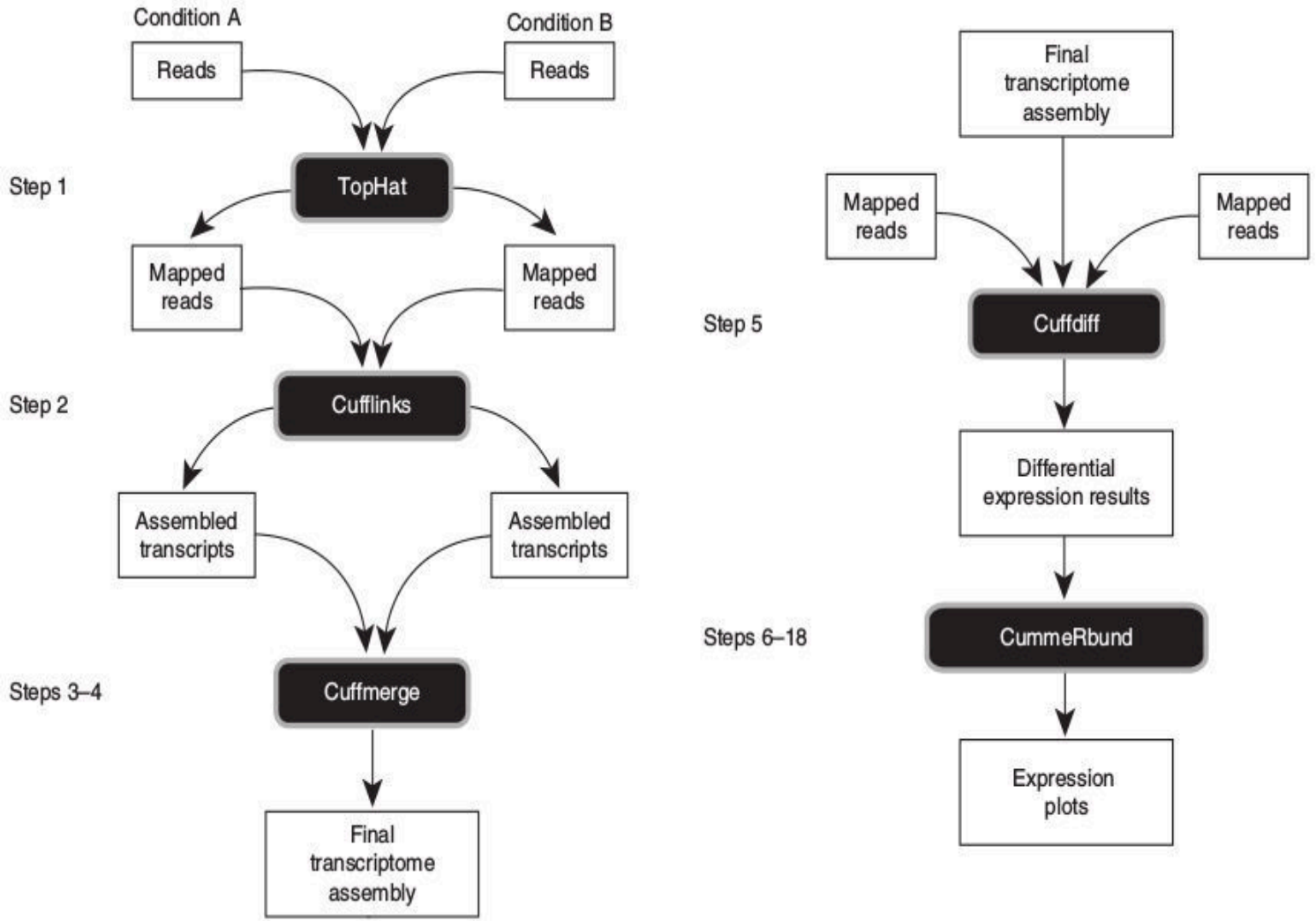
Discover new transcripts (Cuffcompare).

Cuffdiff

Find significant changes in transcript expression, splicing, and promoter use.

- Support multiple samples and replicates
- Models to estimate the distribution
  - pooled, per-condition, blind
- Output a number of statistic tests
  - fold change, p values, q values

Tuxedo pipeline



http://www.ncbi.nlm.nih.gov/pubmed/22383036

Combining tools in a pipeline

Listing 1 : Shell script

```
bwa mem -t 8 $reference $i > $i.sam
samtools view -bt $reference -o $i.bam
    $i.sam
```

Listing 2 : Makefile

```
%.sam: %.fq
  $(BWA) mem -t $(THREADS) $(call MKREF,
      $@) $< > $@

%.bam: %.sam
  $(SAMTOOLS) view -bt $(call MKREF, $@)
      -o $@ $<
```

Snakemake, GATK Queue

Snakemake

- Inspired by good features of makefiles
- Python like syntax

Snakemake, GATK Queue

Snakemake

- Inspired by good features of makefiles
- Python like syntax

GATK Queue

- Scala based
- full DRMAA support

Acknowledgements:

Jeroen Laros (LUMC)
Wibowo Arindrarto (LUMC)
Irina Pulyakhina (LUMC)
Peter-Bram 't Hoen (LUMC)
Johan den Dunnen (LUMC)
Andrew Stubbs (ErasmusMC)

Outline of the practical

1. Do a typical RNA-seq analysis.
   - Expression.
2. Workflows.
   - Rerun the analysis with no effort.
3. Differential expression analysis.