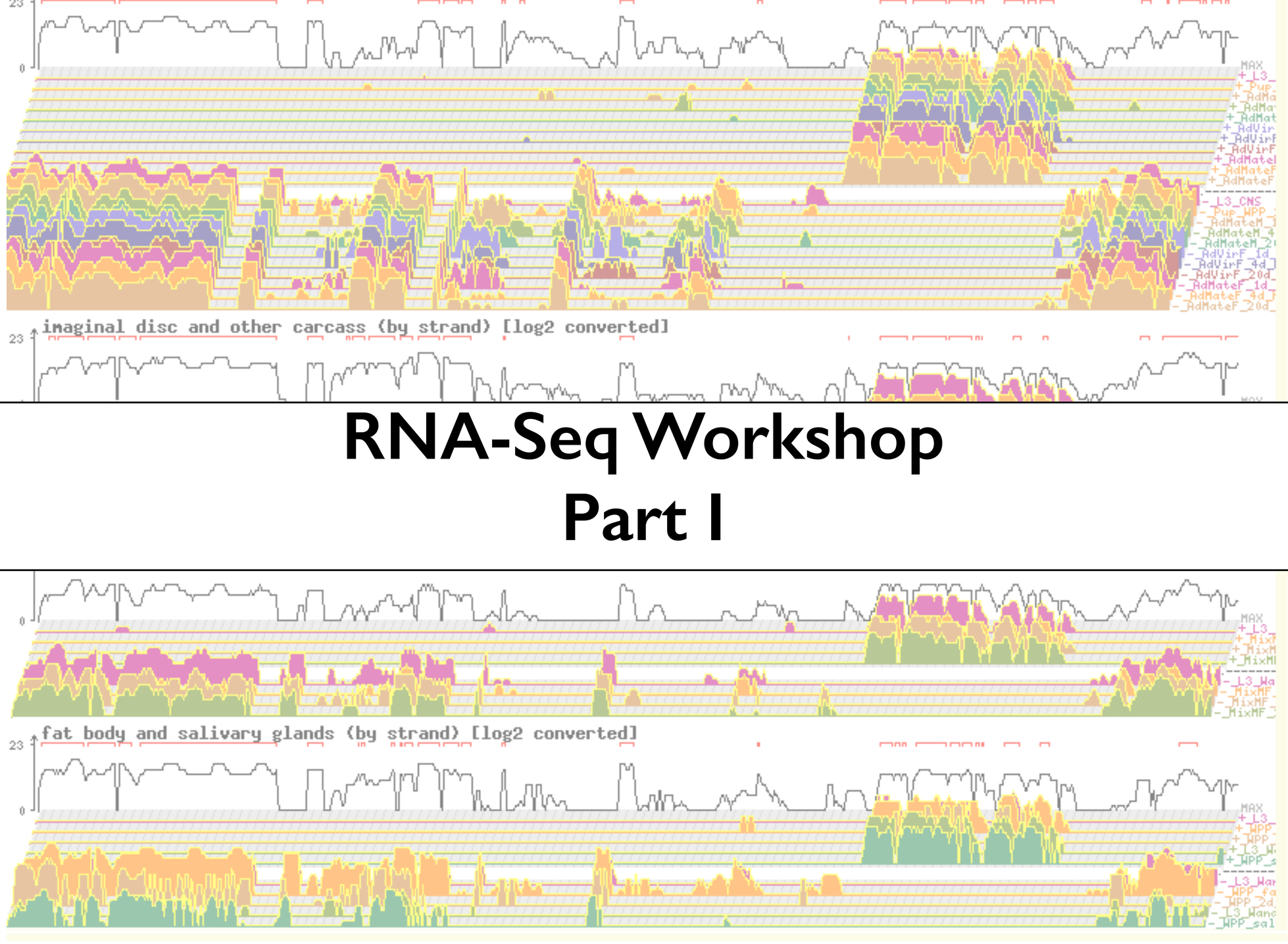


RNA-Seq Workshop
Part I

Exercise 1

Run shell script run_tuxedo.sh (to be explained)

1. Copy ch4_demo_dataset.tar.gz from ~/Data to your Desktop and unzip

```
$cp ~/Data/ch4_demo_dataset.tar.gz ~/Desktop
$cd ~/Desktop
$tar -xvf ch4_demo_dataset.tar.gz
```

2. Copy shell script run_tuxedo.sh to ~/Desktop/ch4_demo_dataset/ and make it executable

```
$chmod 755 run_tuxedo.sh
$cp scripts/shell/run_tuxedo.sh ch4_demo_dataset/
```

4. Change name of SL2.40ch04.fasta to SL2.40ch04.fa

```
$mv ch4_demo_dataset/bwt2_index/SL2.40ch04.fasta ch4_demo_dataset/bwt2_index/SL2.40ch04.fa
```

5. run shell script from **~/Desktop/ch4_demo_dataset/**

```
$cd ch4_demo_dataset/

$./run_tuxedo.sh .
```

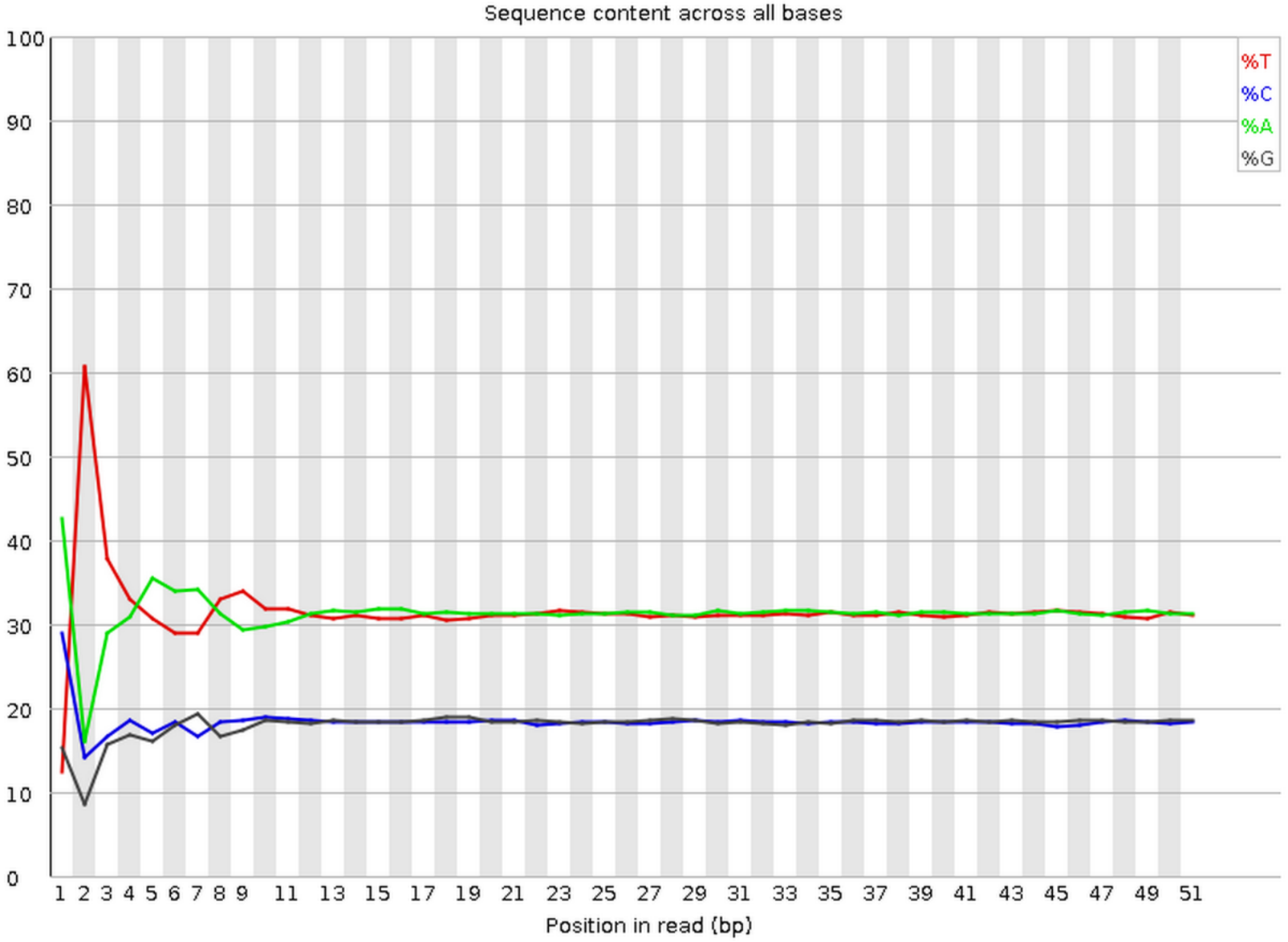
This will run ~15 minutes. We'll discuss what it is doing while it runs.

Transcriptome Assembly

First steps

- FASTQC - give some basic stats about raw reads, can run again after cleaning
  - You can see bias of the random hexamer priming in per base sequence content (next slide)
  - Also GC bias can be present, can effect DE results
- Duplicate removal - not recommended for RNA-seq
- To clean or not to clean? Cleaning only critical for de novo assembly, unless adapter contamination
  - can use fastq-mcf, trimmomatic, prinseq for this, **be sure to run both mate files together if you have paired end data!**

Per base sequence content



Transcriptome Assembly

De novo vs Reference-guided

Reference-guided - map to previously-assembled genome or transcriptome

- pros: computationally easier
- cons: must have good reference assembly to a related species, may miss some transcripts

De novo - cluster reads into transcripts

- pros: no prior assemblies/annotations necessary
- cons: more difficult, more memory needed, LOTS of transcripts

Transcriptome Assembly

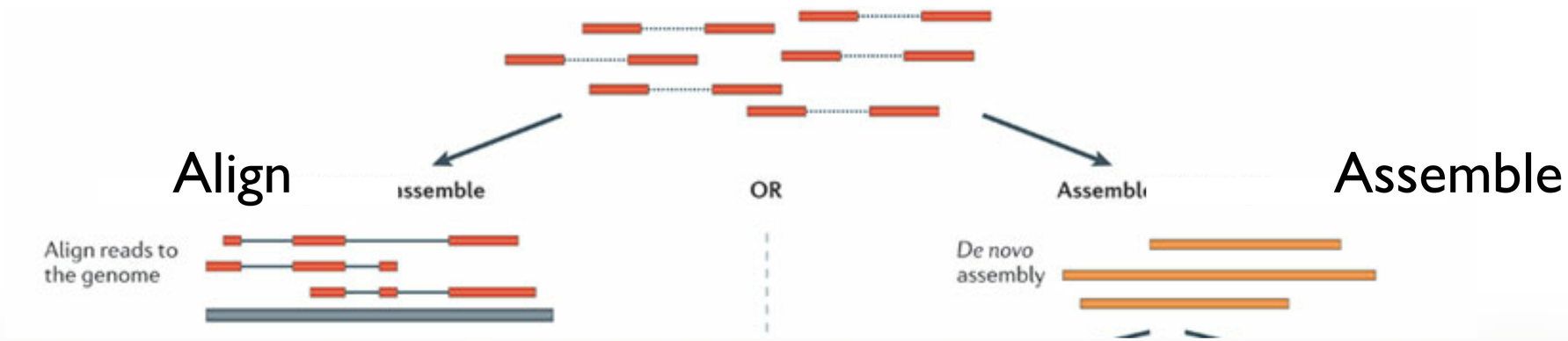
Reference-guided vs De novo

Reference-guided

today

De novo

next time

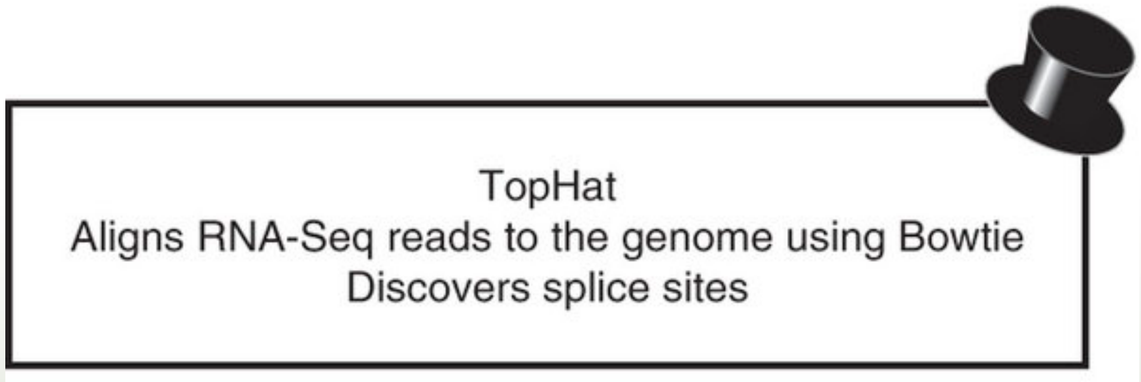


Reference-guided Assembly

- How many reads for DE?
  - 30 x coverage of CDS (Introduction to RNA-Seq)
- What length? SE or PE?
  - short (50 bp), SE ok for differential expression, multi-mapping more problematic
  - longer PE better for annotation or if the reference may not be very good (de novo)

Reference-guided Assembly

- Most use Burrows Wheeler transform (transform and compress into index)
- Tool must be able to deal with splice junctions in eukaryotes
- Does tool accept paired or single-end reads (fastq format)?
- map to reference fasta genome file (needs to be indexed with bowtie2-build for tophat2)



HISAT
Hierarchical Indexing for Spliced Alignment of Transcripts

Reference-guided Assembly

Tophat:

- multi-mapping reads default - tries to find best, or else arbitrarily assigns a location (mapping quality) NH:i:2
- 3 steps: aligns to transcriptome (if provided), aligns to genome, map 25 bp fragments to genome

Reference-guided Assembly

Tophat output:

- 2 bam files: mapped and unmapped reads
  (bam = compressed .sam file)
- bed files: junctions, insertions, deletions
- log files

Reference-guided Assembly

Sam file sample

```
@SQ     SN:SL2.40ch00   LN:21805821
@SQ     SN:SL2.40ch01   LN:90304244
@SQ     SN:SL2.40ch02   LN:49918294
@SQ     SN:SL2.40ch03   LN:64840714
@SQ     SN:SL2.40ch04   LN:64064312
@SQ     SN:SL2.40ch05   LN:65021438
@SQ     SN:SL2.40ch06   LN:46041636
@SQ     SN:SL2.40ch07   LN:65268621
@SQ     SN:SL2.40ch08   LN:63032657
@SQ     SN:SL2.40ch09   LN:67662091
@SQ     SN:SL2.40ch10   LN:64834305
@SQ     SN:SL2.40ch11   LN:53386025
@SQ     SN:SL2.40ch12   LN:65486253
HWI-EAS339_0011:2:92:9381:20526#0       99      SL2.30ch00      78      20      7S77M2S =
      182     187     AAAATGAAAATAAAAATACAAATAGTAATAATAATAATAATAATAATAATAATCATAATAATAATAATAATTACAATAATATT
DDD?DDDDDDD6:32-18:6<BCC>DDDDD>>A?A=CACCAC=CCC6@>@;>C@C;>>6@?CCACDD=D=>7BC8B?BAABACCCA
PG:Z:novoalign  ZO:Z:+- AS:i:249        UQ:i:249        NM:i:7  MD:Z:7T1A1T5A31A17A1T7  PQ:i:459        SM:i:150        AM:i:21
HWI-EAS339_0011:3:88:16221:9366#0       99      SL2.30ch00      90      5       10S76M  =
      127     121     TGACTAAGCAAAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAAT
GGGGGGGGFGGGGGGFGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGFGEGGBGGFGGFFGGGGGFGGGGDGGDGGEGFD
FF  PG:Z:novoalign  ZO:Z:+- AS:i:88 UQ:i:88 NM:i:0  MD:Z:76 PQ:i:88 SM:i:0
  AM:i:0
HWI-EAS339_0011:2:3:16931:15264#0       163     SL2.30ch00      92      11      3S73M1D9M1H     =       139     132
ATTATAATAATAATAATAATAAGAATAATAATAAGAATAATAATAATAATAATAATAATAATAATAATAATAAAATAATAAT
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGFGGGGGGGGGGGGEGGGGGFGGFGGGGGFGGGGGEGGGFGGGDGGD
GF   PG:Z:novoalign  ZO:Z:+- AS:i:143        UQ:i:143        NM:i:3  MD:Z:19T11T41^T9        PQ:i:203        SM:i:0  AM:i:0
```

* reference-guided assembly output

Reference-guided Assembly

Sam format column definitions

Index	Field Name	Description
1	QNAME	Query pair NAME if paired; or Query NAME if unpaired
2	FLAG	Bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition of the clipped sequence
5	MAPQ	MAPping Quality
6	CIGAR	Extended CIGAR string
7	MRNM	Mate Reference sequence NaMe; "=" if the same as RNAME
8	MPOS	1-based leftmost Mate POSition of the clipped sequence
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence
11	QUAL	Query QUALity

Table 1.1: Brief summary of the SAM format

Reference-guided Assembly

Excellent explanation for bitwise flags found here: http://seqanswers.com/forums/showthread.php?t=2301

Tool to translate meaning of bitwise flag: http://picard.sourceforge.net/explain-flags.html

Sam flag field definitions

Field	Hex Code	Description
f_0	0x0001	the read is paired in sequencing, no matter whether it is mapped in a pair
f_1	0x0002	the read is mapped in a proper pair (depends on the protocol, normally inferred during alignment)
f_2	0x0004	the query sequence itself is unmapped
f_3	0x0008	the mate is unmapped
f_4	0x0010	strand of the query (0 for forward; 1 for reverse strand)
f_5	0x0020	strand of the mate
f_6	0x0040	the read is the first read in a pair
f_7	0x0080	the read is the second read in a pair
f_8	0x0100	the alignment is not primary (a read having split hits may have multiple primary alignment records)
f_9	0x0200	the read fails platform/vendor quality checks
f_10	0x0400	the read is either a PCR duplicate or an optical duplicate

Table 1.2: Brief summary of the SAM format

Reference-guided Assembly

Cigar String

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

Return to Exercise 1

Two RNA-seq datasets used in the tomato genome project were downloaded from the SRA in .sra format and extracted using the SRA toolkit (NCBI).

http://www.ncbi.nlm.nih.gov/sra

They were already cleaned using fastq-mcf. All data is from *S. pimpinellifolium*.

Datasets :

- breaker fruit (two files)
- immature fruit (two files)

In this exercise, we will map the reads to tomato chromosome 4 using reference-guided assembly

Exercise 1(cont'd)

Run shell script run_tuxedo.sh (to be explained)

1. Copy ch4_demo_dataset.tar.gz from ~/Data to your Desktop and unzip

```
$cp ~/Data/ch4_demo_dataset.tar.gz ~/Desktop
$cd ~/Desktop
$tar -xvf ch4_demo_dataset.tar.gz
```

2. Copy shell script run_tuxedo.sh to ~/Desktop/ch4_demo_dataset/ and make it executable

```
$chmod 755 run_tuxedo.sh
$cp scripts/shell/run_tuxedo.sh ch4_demo_dataset/
```

4. Change name of SL2.40ch04.fasta to SL2.40ch04.fa

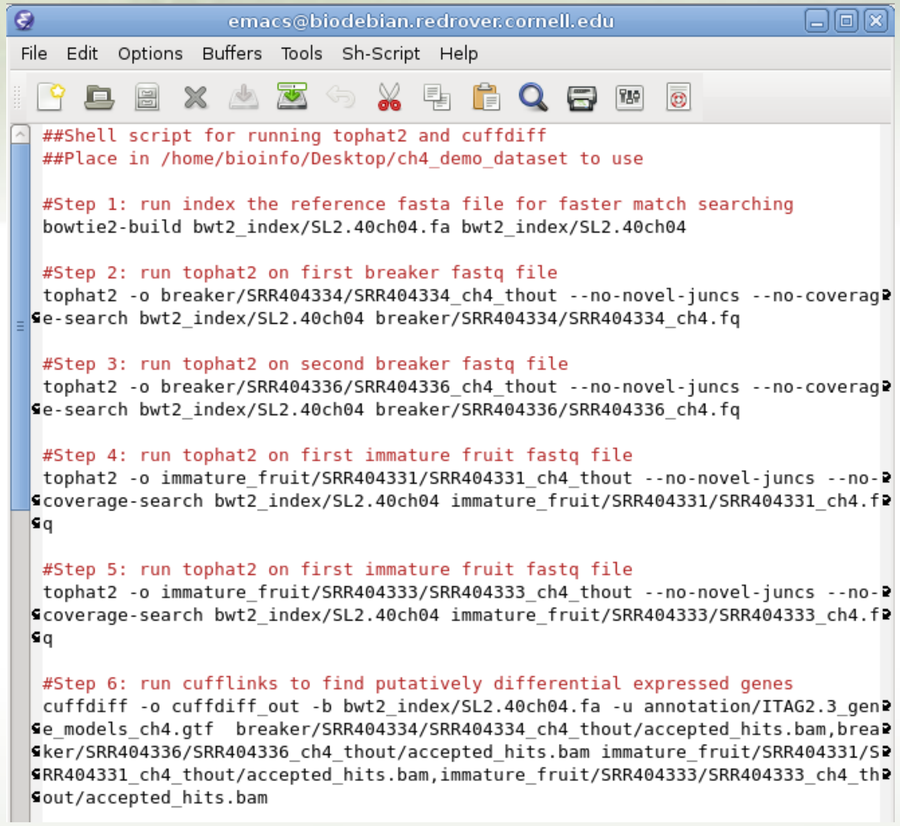
```
$mv ch4_demo_dataset/bwt2_index/SL2.40ch04.fasta ch4_demo_dataset/bwt2_index/SL2.40ch04.fa
```

5. run shell script from **~/Desktop/ch4_demo_dataset/**

```
$cd ch4_demo_dataset/

$./run_tuxedo.sh .
```

Exercise 1 (cont'd)



What does run_tuxedo.sh do?

It is a shell script that runs 6 linux commands (each can be typed in a terminal)

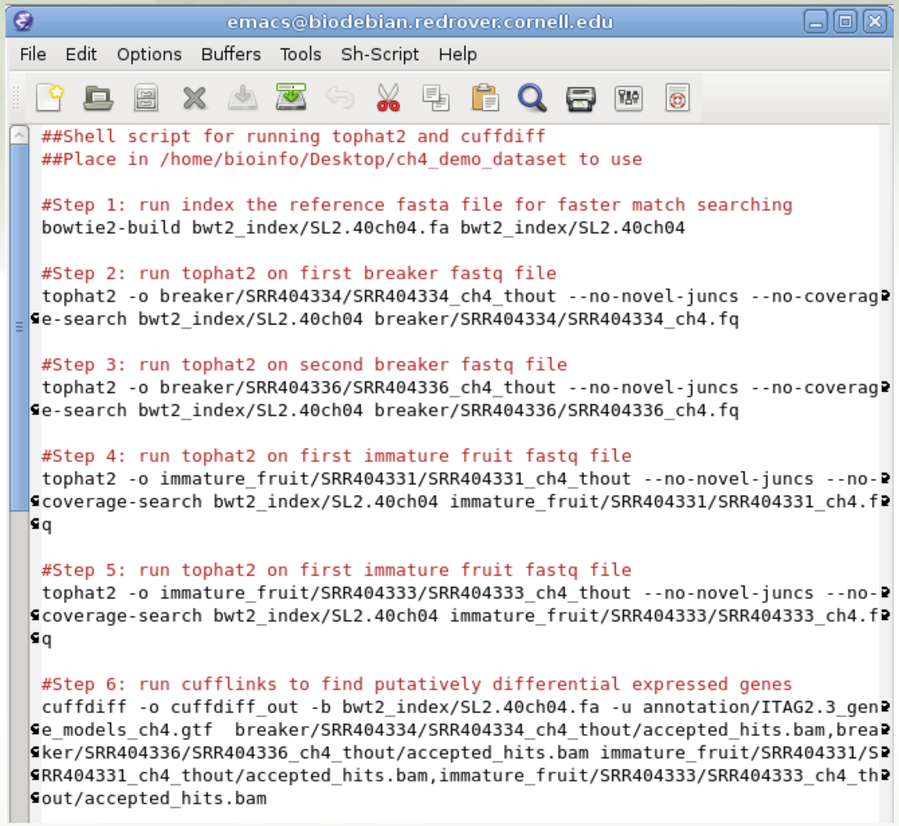
STEP 1: uses bowtie2-build to index the tomato reference file. ~2 minutes

```
$bowtie2-build bwt2_index/SL2.40ch04.fa bwt2_indexch04
```

STEPS 2-5: use tophat2 to map each fastq file in ch4_demo_dataset/ to the reference. ~3 minutes each step

```
$tophat2 -o SRR404334_tophat_out/ --no-novel-juncs --no-coverage-search \
bwt2_index/ SL2.40ch04 breaker/SRR404334/SRR404334_ch4.fq
```

Exercise 1 (cont'd)



STEP 6: runs cufflinks, but we will discuss this part later

Exercise 1 (cont'd)

1. Convert accepted_hits.bam to accepted_hits.sam

```
$samtools view -ho accepted_hits.sam accepted_hits.bam
```

2. How many reads are in the tophat2 output file?

```
$grep -v "^@" accepted_hits.sam |wc
```

2. How many chromosomes are in the reference file?

```
$grep "^@SQ" accepted_hits.sam |wc
```

3. How many reads map to each chromosome?

```
$grep -v "^@" accepted_hits.sam |cut -f3 |sort |uniq -c
```

De novo Assembly

- Must clean (quality trim) reads well before de novo assembly!
- Error correction
- make sure to run assembly on larger memory server
- Many “isoforms”!!! :(
- Also Velvet + Oases

De novo Assembly

“ a method that hierarchically clusters contigs using shared reads and expression, then summarizes read counts to clusters, ready for statistical testing"

2. Assembly

Viewing Output

- Tablet (http://bioinf.scri.ac.uk/tablet/)

Indexing files for Tablet:

1. reference:

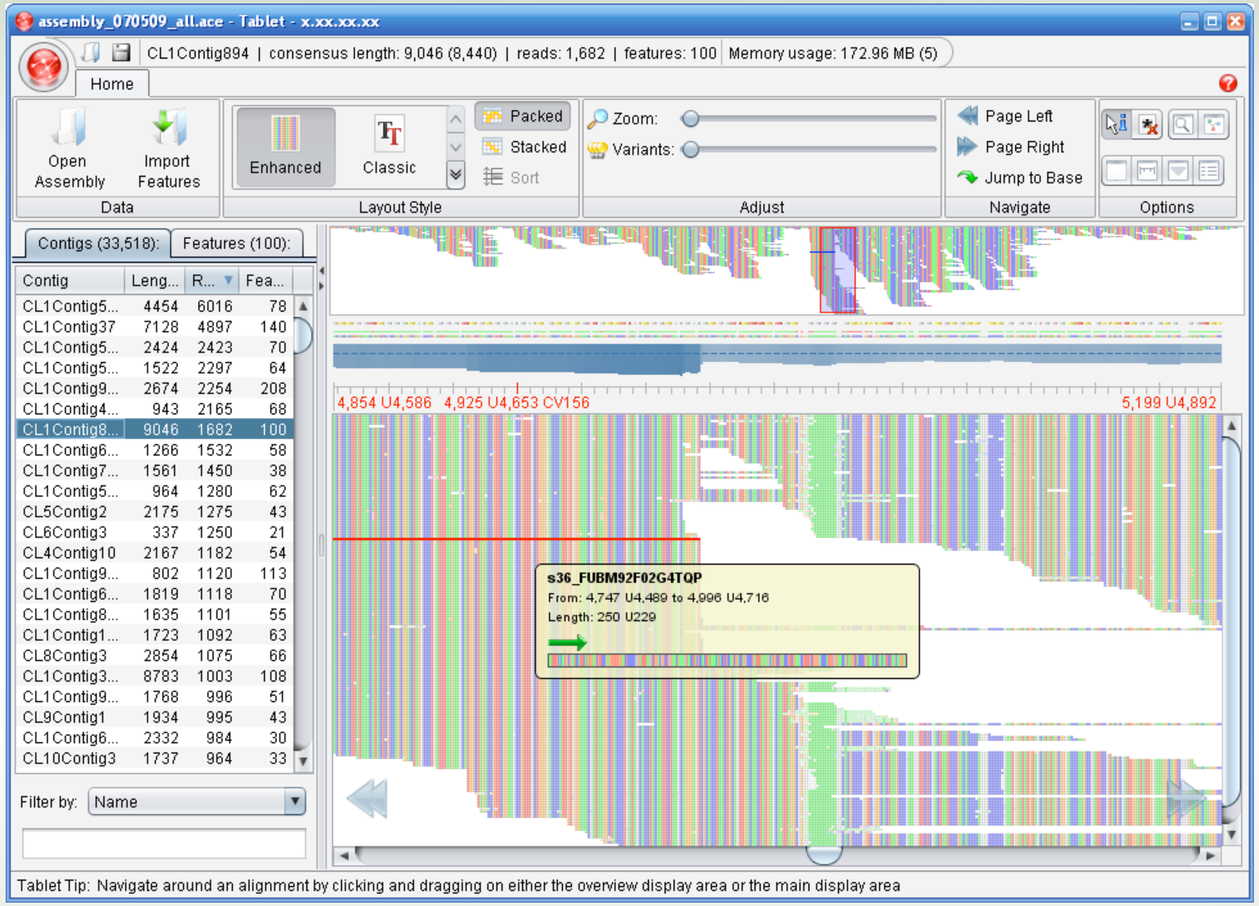
```
samtools faidx SL2.40ch04.fa
```

2. bam:

```
samtools index
```

Load into Tablet:

1. reference .fa file
2. mapped reads .bam file
3. annotations .gff3 file



2. Assembly

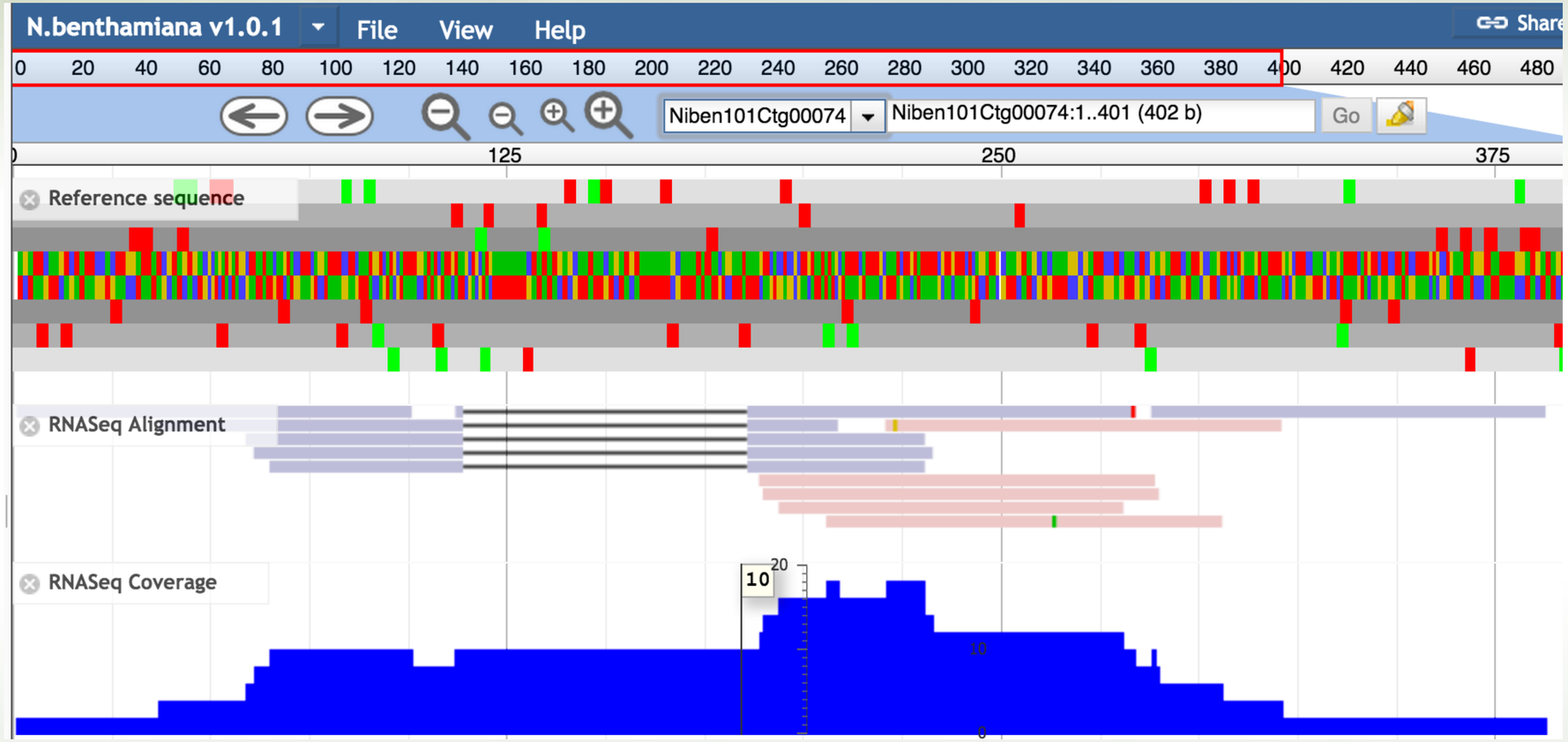
Viewing Output

- Samtools tview

2. Assembly

Viewing Output

JBrowse



also IGV

3. Analysis

How good is the assembly?

- Check for contaminants by using BLAST to search appropriate databases (SeqClean)
- How many contigs are there vs how many genes expected - total sequence length? CEGMA
- Length of contigs - compare to known long transcripts?
- How many reads map (reference-guided)?
- check no DNA contamination
- coverage of features, biases (introns, 3', ribosomal)
- sequencing saturation (new gene discovery)

Method

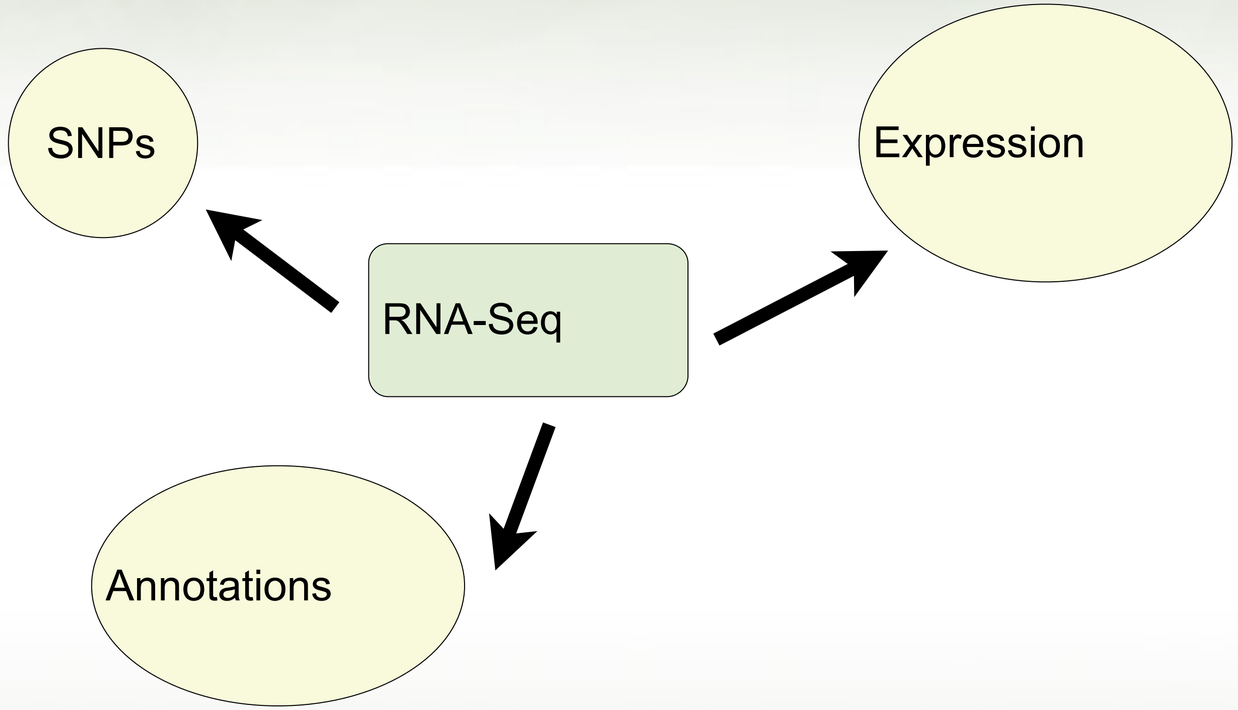
Highly accessed Open Access

Evaluation of *de novo* transcriptome assemblies from RNA-Seq data

Bo Li[1†], **Nathanael Fillmore**[2†], **Yongsheng Bai**[3], **Mike Collins**[4], **James A Thomson**[4,5,6], **Ron Stewart**[4] and **Colin N Dewey**[2,7*]

3. Analysis

What information can be gained?



3. Analysis

Expression

RPKM/FPKM (reads/fragments per 1k of exon per 1M mapped reads.

- cufflinks
- RSEM
- Only used as an expression metric within or between samples, not for de testing.

Counts - number of reads overlapping a feature (use htseq-count or featurecount)

- DESeq
- edgeR
- These don’t take read mapping uncertainty into account

Other:

- cuffdiff - attempts to correct multimap reads (not default), computes separate variance model for isoforms, may produce high number of false positives (Rapaport et al., 2013), no support for factorial designs
- EBSeq - empirical Bayesian methods, de of isoforms

Normalization???????

Total Count, Upper Quartile, Median (Med), DESeq (geometric), Trimmed Mean of M values implemented in the edgeR, Quantile, and the Reads Per Kilobase per Million mapped reads (RPKM)!!!!!

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

Marie-Agnès Dillies*, Andrea Rau*, Julie Aubert*, Christelle Hennequet-Antier*, Marine Jeanmougin*, Nicolas Servant*, Céline Keime*, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom*, Mickaël Guedj* and Florence Jaffrézic* on behalf of The French StatOmique Consortium

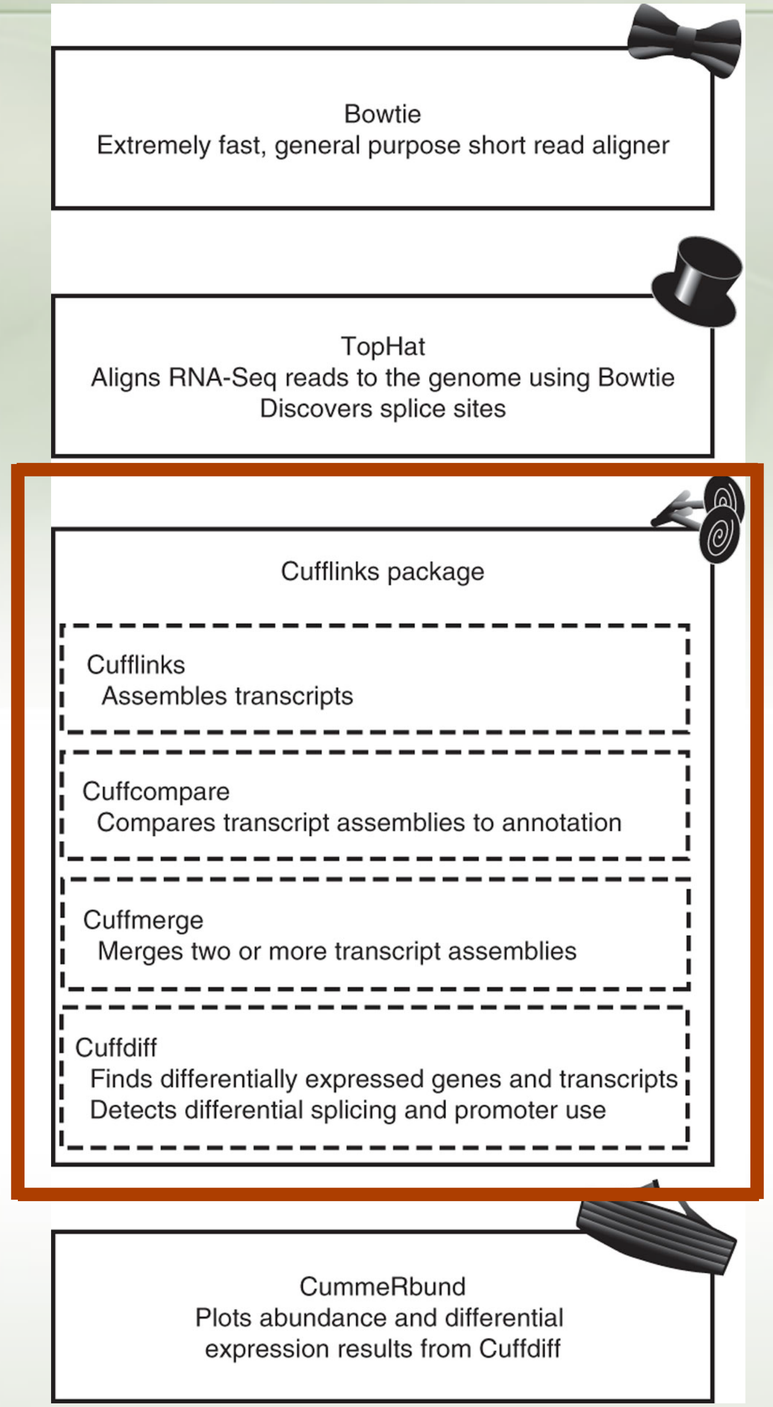
"Based on three real mRNA and one miRNA-seq datasets, we confirm previous observations that RPKM and TC, both of which are still widely in use [40, 41], are ineffective and should be definitively abandoned in the context of differential analysis. "

" The other normalization methods (UQ, Med, DESeq and TMM) perform similarly on the varied datasets considered here, both in terms of the qualitative characteristics of the normalized data and the results of DE analyses."

3. Analysis
Expression

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

- Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel,Steven L Salzberg, John L Rinn & Lior Pachter



3. Analysis

Expression

Cufflinks can also be used for:

✦ strand-specific RNA-seq

✦ novel transcript discovery in annotated genomes

✦ identification of novel splice variants

✦ detecting transcripts in genomes without annotation

For protocols see:

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks (Trapnell et al., 2012).

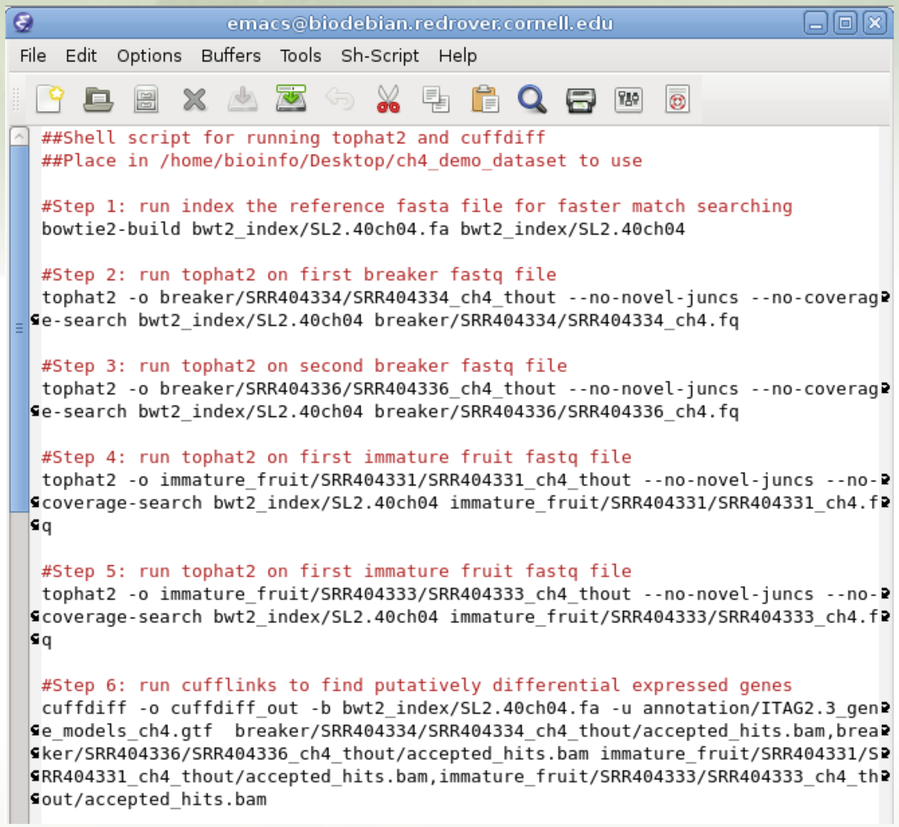
3. Analysis

Expression

Challenges

- paralogs/reads that map to more than one place
- Cufflinks divides these evenly across the matches or tries to assign them a location based on expression of surrounding reads.
- Counts-based methods do nothing (although counting software might). Filter on mapping quality?

Exercise 2



What does run_tuxedo.sh do? (cont'd)

STEP 6: Use cuffdiff to detect differentially expressed genes in immature and breaker fruit using known chromosome 4 tomato gene models

```
$cuffdiff -o cuffdiff_out -b bwt2_index/SL2.40ch04.fa -u annotation/
ITAG2.3_gene_models_ch4.gtf  breaker/SRR404334/SRR404334_ch4_thout/
accepted_hits.bam,breaker/SRR404336/SRR404336_ch4_thout/accepted_hits.bam
immature_fruit/SRR404331/SRR404331_ch4_thout/ accepted_hits.bam,immature_fruit/
SRR404333/SRR404333_ch4_thout/accepted_hits.bam
```

~5 minutes

3. Analysis

Expression

Cufflinks Output:

```
-rw-r--r-- 1 bioinfo bioinfo   12 Apr 16 11:17 cds.count_tracking
-rw-r--r-- 1 bioinfo bioinfo  115 Apr 16 11:17 cds.diff
-rw-r--r-- 1 bioinfo bioinfo  124 Apr 16 11:17 cds_exp.diff
-rw-r--r-- 1 bioinfo bioinfo   91 Apr 16 11:17 cds.fpkm_tracking
-rw-r--r-- 1 bioinfo bioinfo  115 Apr 16 11:17 cds.read_group_tracking
-rw-r--r-- 1 bioinfo bioinfo 343K Apr 16 11:17 gene_exp.diff
-rw-r--r-- 1 bioinfo bioinfo 183K Apr 16 11:17 genes.count_tracking
-rw-r--r-- 1 bioinfo bioinfo 333K Apr 16 11:17 genes.fpkm_tracking
-rw-r--r-- 1 bioinfo bioinfo 564K Apr 16 11:17 genes.read_group_tracking
-rw-r--r-- 1 bioinfo bioinfo 348K Apr 16 11:17 isoform_exp.diff
-rw-r--r-- 1 bioinfo bioinfo 189K Apr 16 11:17 isoforms.count_tracking
-rw-r--r-- 1 bioinfo bioinfo 346K Apr 16 11:17 isoforms.fpkm_tracking
-rw-r--r-- 1 bioinfo bioinfo 586K Apr 16 11:17 isoforms.read_group_tracking
-rw-r--r-- 1 bioinfo bioinfo  115 Apr 16 11:17 promoters.diff
-rw-r--r-- 1 bioinfo bioinfo  466 Apr 16 11:17 read_groups.info
-rw-r--r-- 1 bioinfo bioinfo  451 Apr 16 11:17 run.info
-rw-r--r-- 1 bioinfo bioinfo  115 Apr 16 11:17 splicing.diff
-rw-r--r-- 1 bioinfo bioinfo  124 Apr 16 11:17 tss_group_exp.diff
-rw-r--r-- 1 bioinfo bioinfo   12 Apr 16 11:17 tss_groups.count_tracking
-rw-r--r-- 1 bioinfo bioinfo   91 Apr 16 11:17 tss_groups.fpkm_tracking
-rw-r--r-- 1 bioinfo bioinfo  115 Apr 16 11:17 tss_groups.read_group_tracking
```

extract genes that have a significant value for differential expression

```
$awk -F "\t" 'BEGIN {OFS = "\t"} $14 = "yes" {print $0}' gene_exp.diff > significant_genes.txt
```

We will analyze these results in R tomorrow!

Other useful tools

- Bedtools - used to compare features.
- snpeff - gives predicted effect of SNP (synonymous, nonsynonymous, etc).
- sra toolkit - for converting NGS data files from the SRA.
- Picard - duplicate detection, etc.
- Prinseq - duplicate removal
- Dindel - indel calling from NGS mapping data.
- seqanswers