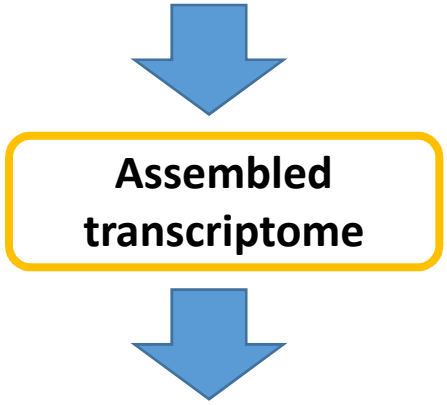
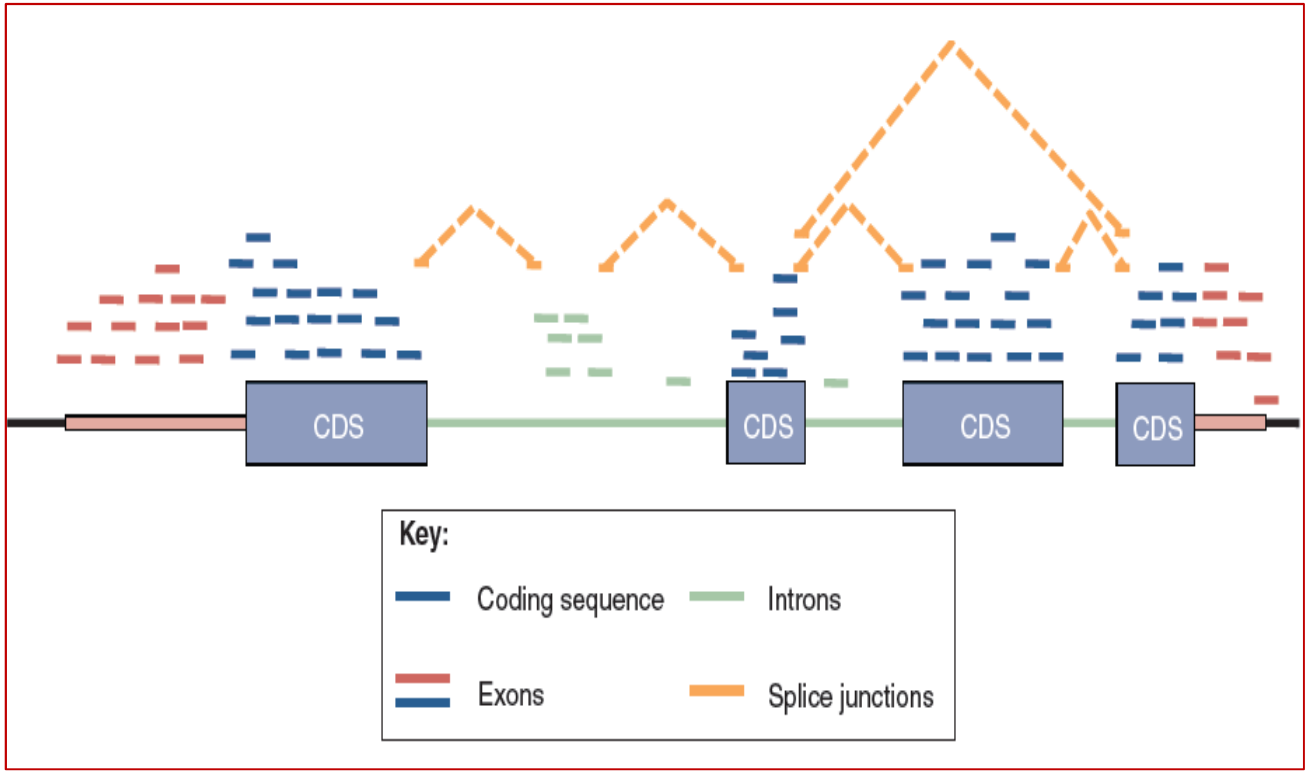


Post-assembly Data Analysis

- Quantification: get expression for each gene in each sample
- Genes differentially expressed between samples
- Clustering/network analysis
- Identifying over-represented functional categories in DE genes
- Evaluation of the quality of the assembly

Part 1. Abundance estimation using RSEM

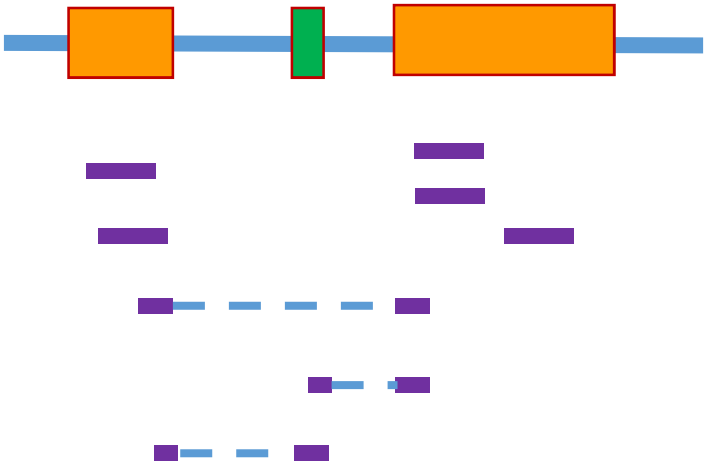


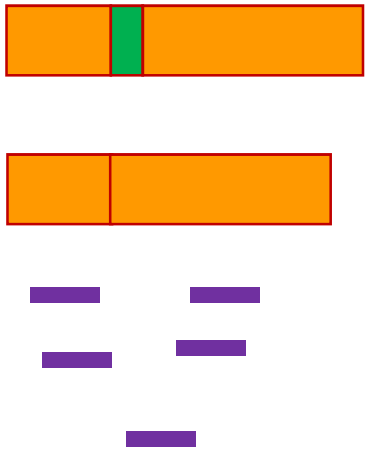
Different summarization strategies will result in the inclusion or exclusion of different sets of reads in the table of counts.

Map reads to genome (TOPHAT)

vs

Map reads to Transcriptome (BOWTIE)





What is easier?

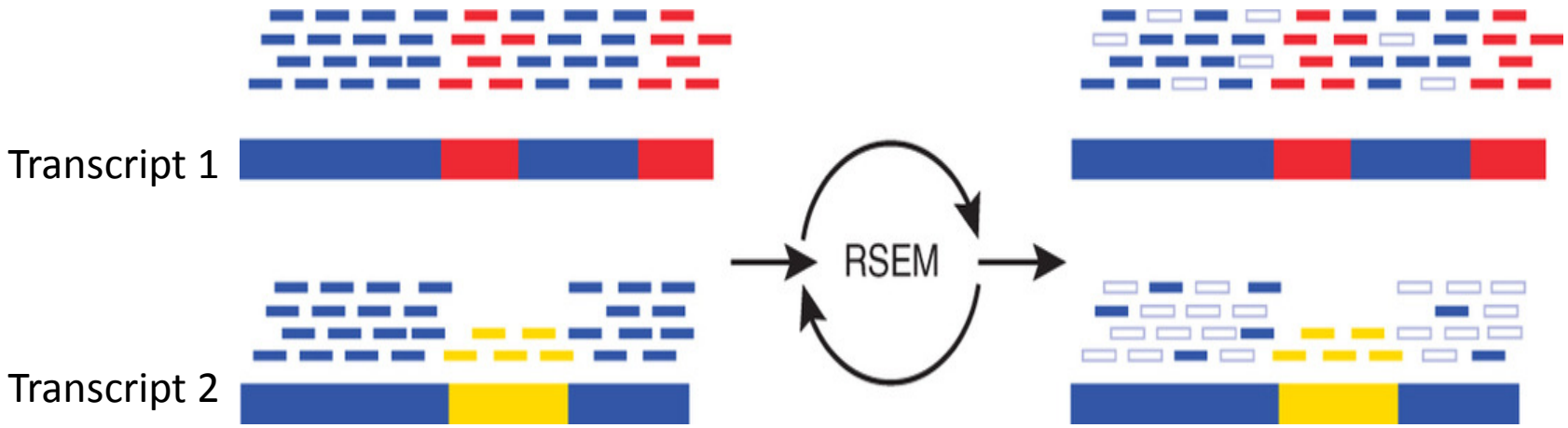
No issues with alignment across splicing junctions.

What is more difficult?

The same reads could be repeatedly aligned to different splicing isoforms, and paralogous genes.

RSEM assign ambiguous reads based on unique reads mapped to the same transcript.

Red & **yellow**: unique regions
Blue: regions shared between two transcripts



Hass et al. Nature Protocols 8:1494

Trinity provides a script for calling BOWTIE and RSEM

align_and_estimate_abundance.pl

```
align_and_estimate_abundance.pl \
        --transcripts Trinity.fasta \
        --seqType fq \
        --left sequence_1.fastq.gz \
        --right sequence_2.fastq.gz \
        --SS_lib_type RF \
        --aln_method bowtie \
        --est_method RSEM \
        --thread_count 4 \
        --trinity_mode \
        --output_prefix tis1rep1 \
```

Parameters for align_and_estimate_abundance.pl

--aln_method: alignment method
Default: “bowtie” .
Alignment file from other aligner might not be supported.

--est_method: abundance estimation method
Default : RSEM, slightly more accurate.
Optional: eXpress, faster and less RAM required.

--thread_count: number of threads

--trinity_mode: the input reference is from Trinity.
Non-trinity reference requires a gene-isoform mapping file (--gene_trans_map).

Output files from RSEM (two files per sample)

*.isoforms.results table

transcript_id	gene_id	length	effective_length	expected_count	TPM	FPKM	IsoPct
gene1_isoform1	gene1	2169	2004.97	22.1	3.63	3.93	92.08
gene1_isoform2	gene1	2170	2005.97	1.9	0.31	0.34	7.92
…							

*.genes.results table

gene_id	transcript_id(s)	length	effective_length	expected_count	TPM	FPKM
gene1	gene1_isoform1,gene1_isoform2	2169.1	2005.04	24	3.94	4.27
…						

Output files from RSEM (two files per sample)

*.isoforms.results table

transcript_id	gene_id	length	effective_length	expected_count	TPM	FPKM	IsoPct
gene1_isoform1	gene1	2169	2004.97	22.1	3.63	3.93	92.08
gene1_isoform2	gene1	2170	2005.97	1.9	0.31	0.34	7.92
…							

Percentage of an isoform in a gene

sum

*.genes.results table

gene_id	transcript_id(s)	length	effective_length	expected_count	TPM	FPKM
gene1	gene1_isoform1,gene1_isoform2	2169.1	2005.04	24	3.94	4.27
…						

Filtering transcriptome reference based on RSEM

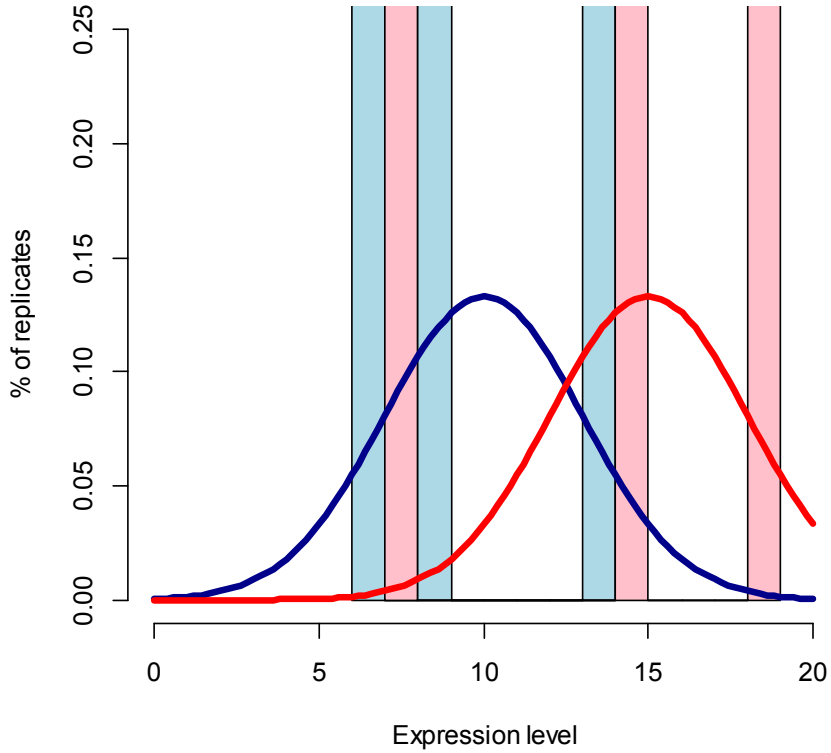
filter_fasta_by_rsem_values.pl

```
filter_fasta_by_rsem_values.pl                                  \
--rsem_output=s1.isoforms.results,s2.isoforms.results           \
--fasta=Trinity.fasta                                           \
--output=Trinity.filtered.fasta                                 \
--isopct_cutoff=5                                               \
--fpkm_cutoff=10                                                \
--tpm_cutoff=10                                                 \
```

- Can be filtered by multiple RSEM files simultaneously. (criteria met in any one file will be filtered)

Part 2. Differentially Expressed Genes

… with 3 biological replicates in each species



Distribution of Expression Level of A Gene

Condition 1

Condition 2

Use EdgeR or DESeq to identify DE genes

Reported for each gene:

- Normalized read count (Log2 count-per-million);
- Fold change between biological conditions (Log2 fold)
- Q(FDR).

Using FDR, fold change and read count values to filter:

E.g. Log2(fold) >2 or <-2
FDR < 0.01
Log2(counts) > 2

Combine multiple-sample RSEM results into a matrix

```
abundance_estimates_to_matrix.pl          \
 --est_method RSEM                        \
        s1.genes.results                  \
        s2.genes.results                  \
        s3.genes.results                  \
        s4.genes.results                  \
--out_prefix mystudy
```

Output file: mystudy.counts.matrix

	s1	s2	s3	s4
TR10295|c0_g1	0	0	0	0
TR17714|c6_g5	51	70	122	169
TR23703|c1_g1	0	0	0	0
TR16168|c0_g2	1	1	1	0
TR16372|c0_g1	10	4	65	91
TR12445|c0_g3	0	0	0	0

……

Use run_DE_analysis.pl script to call edgeR or DESeq

Make a condition-sample mapping file

```
contition1    s1
condition1    s2
condition2    s3
condition2    s4
```

Run script run_DE_analysis.pl

```
run_DE_analysis.pl                          \
--matrix mystudy.counts.matrix              \
--samples_file mysamples                    \
--method edgeR                              \
--min_rowSum_counts 10                      \
--output edgeR_results
```

Skip genes with summed read counts less than 10

Output files from run_DE_analysis.pl

	logFC	logCPM	PValue	FDR
TR8841|c0_g2	13.28489	7.805577	3.64E-20	1.05E-15
TR14584|c0_g2	-12.9853	7.507117	2.82E-19	4.04E-15
TR16945|c0_g1	-13.3028	7.823384	4.21E-18	3.11E-14
TR15899|c3_g2	9.485185	7.708629	5.35E-18	3.11E-14
TR8034|c0_g1	-10.1468	7.836908	5.42E-18	3.11E-14
TR3434|c0_g1	-12.909	7.431009	1.16E-17	5.55E-14

Filtering (done in R or Excel):

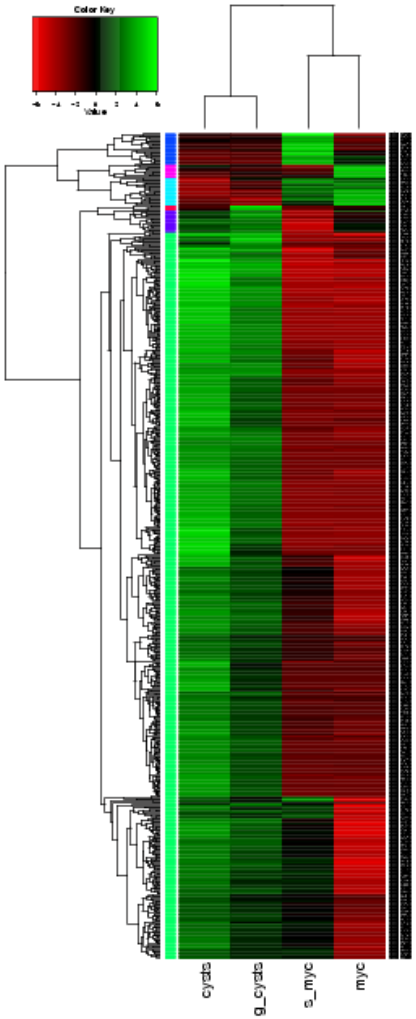
FDR: False-discovery-rate of DE detection (0.05 or below)

logFC: fold change log2(fc)

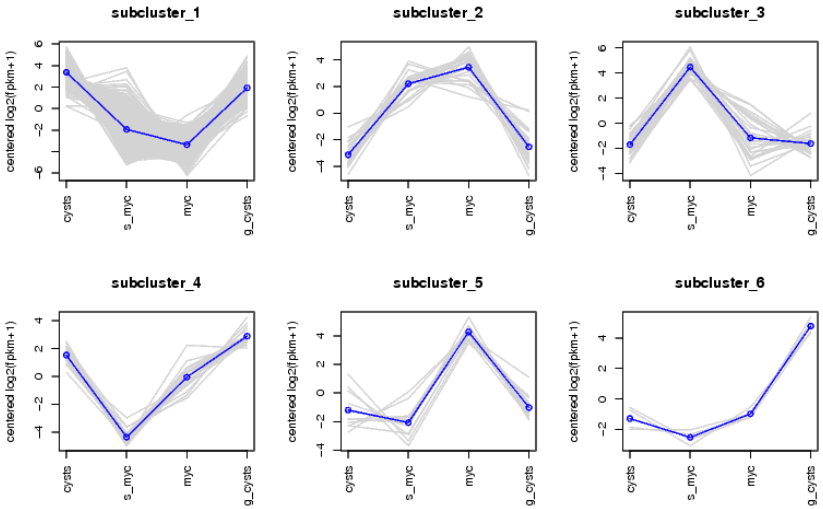
logCPM: average expression level (count-per-million)

Part 3. Clustering/Network Analysis

Heat map and Hierarchical clustering



K-means clustering



Clustering tools in Trinity package

```
# Heap map
analyze_diff_expr.pl                          \
--matrix ../mystudy.TMM.fpkm.matrix           \
--samples ../mysamples                        \
-P 1e-3 -C 2                                  \

# K means clustering
define_clusters_by_cutting_tree.pl            \
-K 6 -R cluster_results.matrix.RData
```

Output:
Images in PDF format;
Gene list for each group from K-means analysis;

Part 4. Functional Category Enrichment Analysis

Gene list

(from DE or clustering)
TR14584|c0_g2
TR16945|c0_g1
TR8034|c0_g1
TR3434|c0_g1
TR13135|c1_g2
TR15586|c2_g4
TR14584|c0_g1
TR20065|c2_g3
TR1780|c0_g1
TR25746|c0_g2

GO enrichment analysis

Enriched GO categories

GO ID	over represented pvalue	FDR	GO term
GO:0044456	3.40E-10	3.78E-06	synapse part
GO:0050804	1.49E-09	8.29E-06	regulation of synaptic transmission
GO:0007268	5.64E-09	2.09E-05	synaptic transmission
GO:0004889	1.00E-07	0.00012394	acetylcholine-activated cation-selective channel activity
GO:0005215	1.23E-07	0.000136471	transporter activity

Function annotation

1. Predict open-reading-frame (DNA -> protein sequence)

```
TransDecoder -t Trinity.fasta --workdir transDecoder -S
```

2. Predict gene function from sequences

<u>Trinotate package</u>

- BLAST Uniprot : homologs in known proteins
- PFAM: protein domain)
- SignalP: signal peptide)
- TMHMM: trans-membrane domain
- RNAMMER: rRNA

Output:

go_annotations.txt (Gene Ontology annotation file)

Step-by-step guidance: https://cbsu.tc.cornell.edu/lab/userguide.aspx?a=software&i=143#c

Enrichment analysis

```
run_GOseq.pl                                  \
  --genes_single_factor DE.filtered.txt       \
  --GO_assignments  go_annotations.txt        \
  --lengths genes.lengths.txt
```

Use the geneIDs in the first column

GO annotation file from Trinotate

3rd column from RSEM *.genes.results file

Part 5. Evaluation of Assembly Quality

Compare the *Drosophila yakuba* assembly used in the exercise to *Drosophila melanogaster* proteins from flybase

#hit_pct_cov_bin	count_in_bin	>bin_below
100	5165	5165
90	1293	6458
80	1324	7782
70	1434	9216
60	1634	10850
50	1385	12235
40	1260	13495
30	1091	14586
20	992	15578
10	507	16085

BLAST against *D. melanogaster* protein database

```
Blastx                                  \
  -query Trinity.fasta                  \
  -db melanogaster.pep.all.fa           \
  -out blastx.outfmt6
  -evalue 1e-20
  -num_threads 6 \
```

Analysis the blast results and generate histogram

```
analyze_blastPlus_topHit_coverage.pl        \
  blastx.outfmt6                            \
  Trinity.fasta                             \
  melanogaster.pep.all.fa                   \
```