RNA-Seq: Methods and Applications

Prat Thiru

Outline

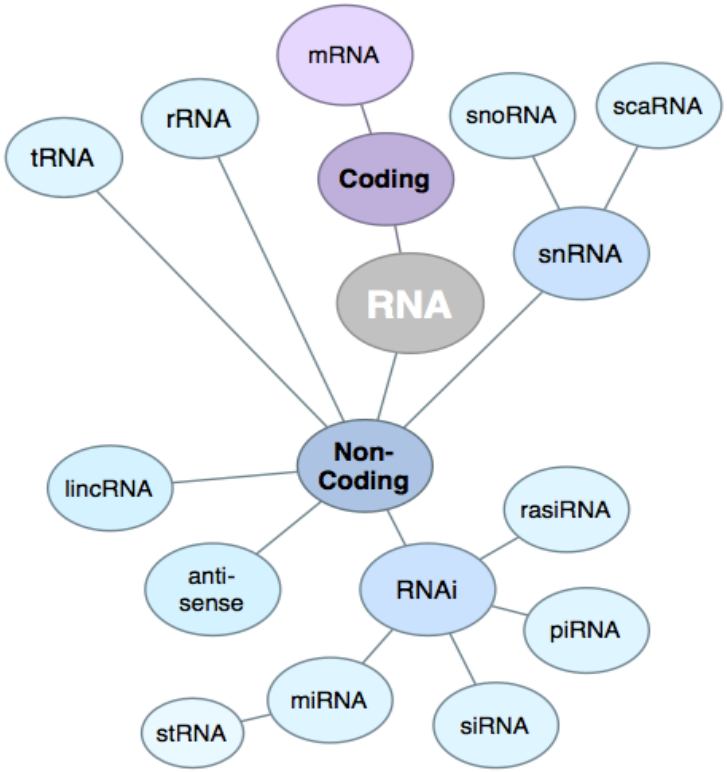
- Intro to RNA-Seq
  - Biological Questions
  - Comparison with Other Methods
  - RNA-Seq Protocol
- RNA-Seq Applications
  - Annotation
  - Quantification
  - Other Applications
- Expression Profiling Steps and Software
- Running TopHat and Cufflinks (Commands)

Goals of Sequencing the Transcriptome

- Annotation
  - Identify genes, exons, splicing events, ncRNAs, etc.
  - Novel genes or transcripts

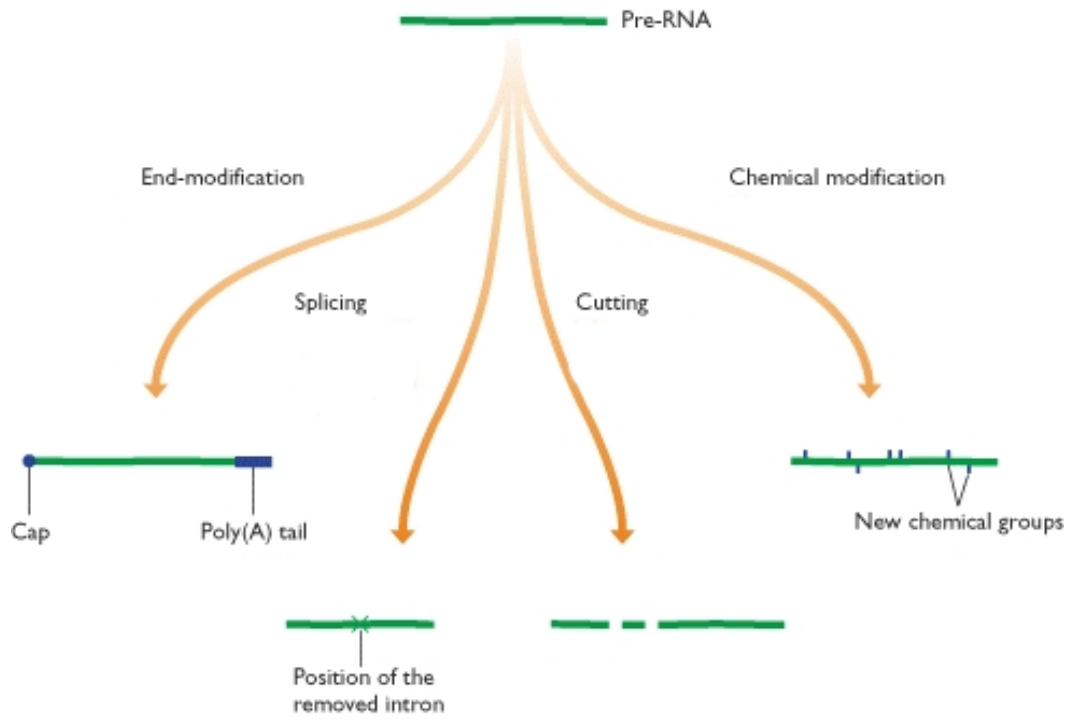
- Quantification
  - Abundance of transcripts between different conditions

Transcriptome: RNA World



http://finchtalk.geospiza.com/2009/05/small-rnas-get-smaller.html

Transcriptome: Complexity



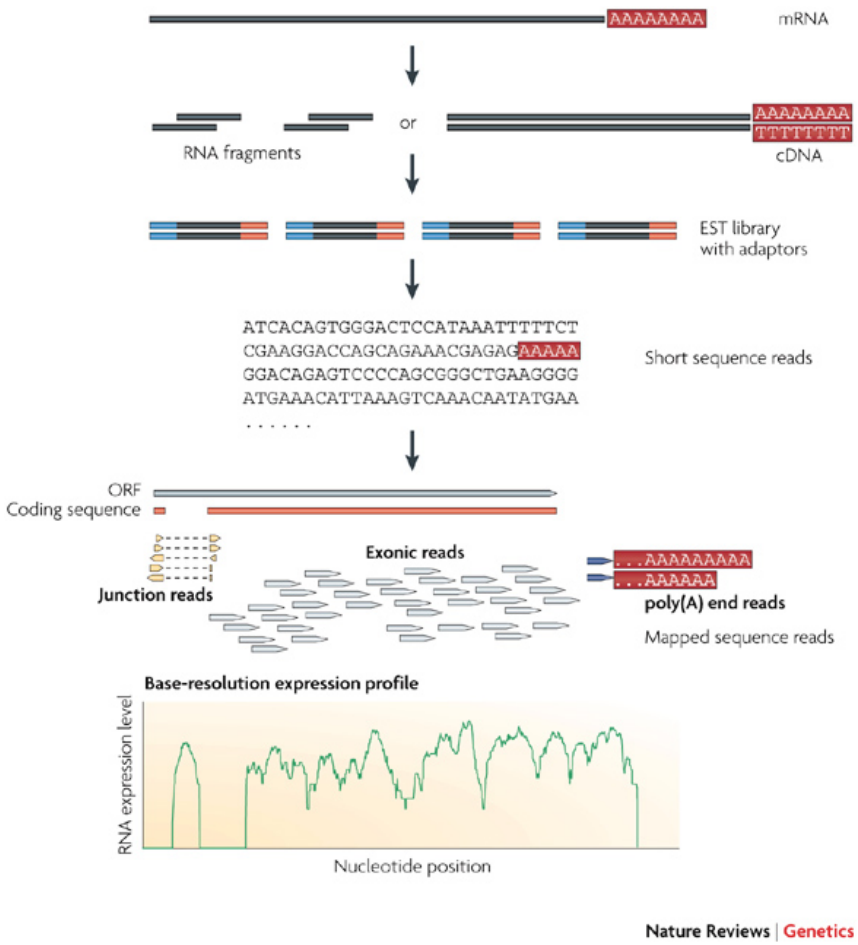
http://www.ncbi.nlm.nih.gov/books/NBK21128/

Comparison of Methods for Studying the Transcriptome

Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

Wang, Z. et al. *RNA-Seq: a revolutionary tool for transcriptomics* Nature Reviews Genetics (2009)

RNA-Seq Experiment

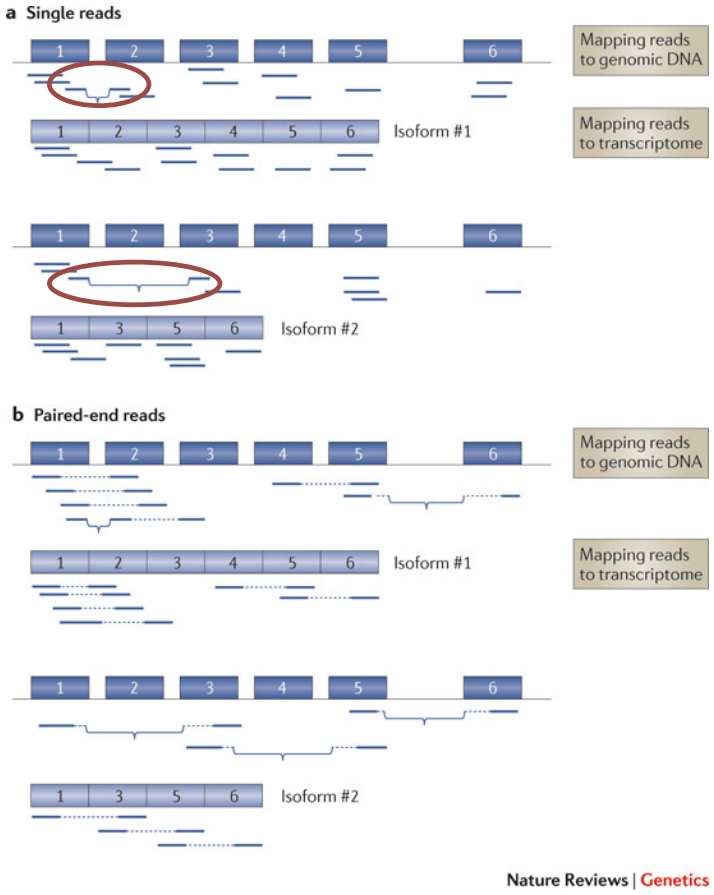


Wang, Z. et al. *RNA-Seq: a revolutionary tool for transcriptomics* Nature Reviews Genetics (2009)

Outline

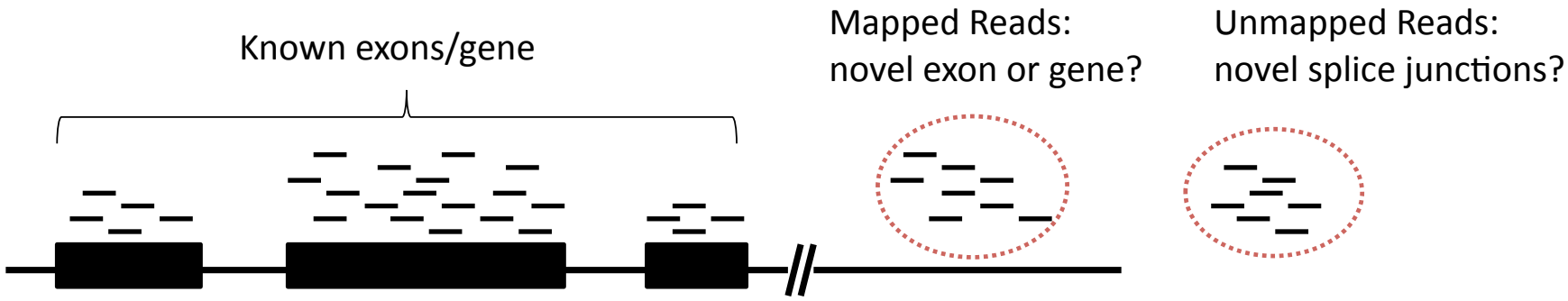
- Intro to RNA-Seq
  - Biological Questions
  - Comparison with Other Methods
  - RNA-Seq Protocol
- RNA-Seq Applications
  - Annotation
  - Quantification
  - Other Applications
- Expression Profiling Steps and Software
- Running TopHat and Cufflinks (Commands)

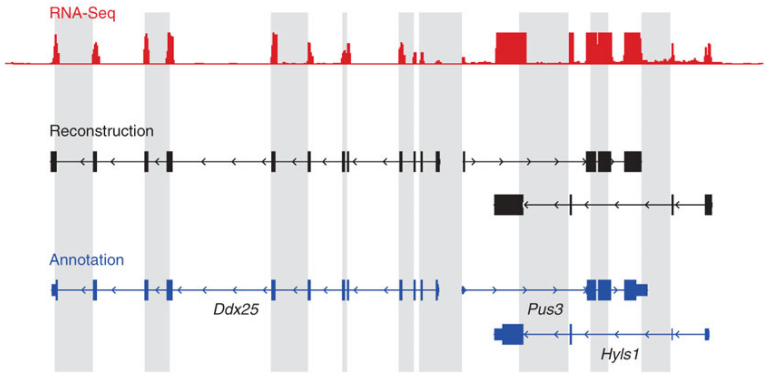
RNA-Seq Applications – Annotation: Alternative Splicing Events



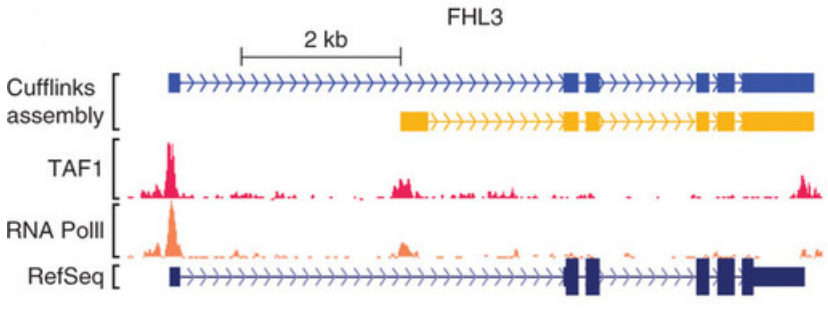
Ozsolak, F. and Milos, P. *RNA sequencing: advances, challenges and opportunities* Nature Reviews Genetics (2011)

RNA-Seq Applications – Annotation: Identify Known and Novel Transcripts





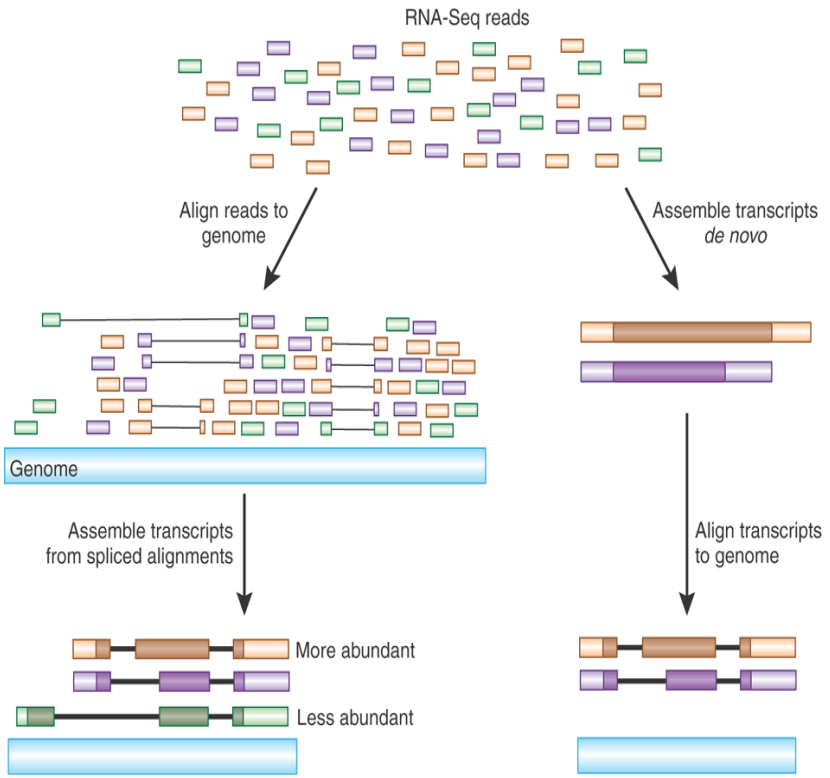
Guttman, M. et al *Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs* Nature Biotechnology (2010)



Trapnell, C. et al *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation* Nature Biotechnology (2010)

Assembly and Mapping RNA-Seq

- Options:
  - Align and then assemble
  - Assemble and then align
- Align to
  - genome
  - transcriptome



Haas, B.J., and Zody, M.C. *Advancing RNA-Seq analysis* Nature Biotechnology (2010)

RNA-Seq Applications - Quantification: Expression Profiling

Mortazavi A., et al. *Mapping and quantifying mammalian transcriptomes by RNA-Seq* Nature Methods (2008)

Need for Normalization

- More reads mapped to a transcript if it is

  i) long

  ii) at higher depth of coverage

- Normalize such that

  i) *features* of different lengths

  ii) total sequence from different conditions

  can be compared

Quantifying Expression: RPKM

- RPKM: Reads Per Kilobase per Million mapped reads

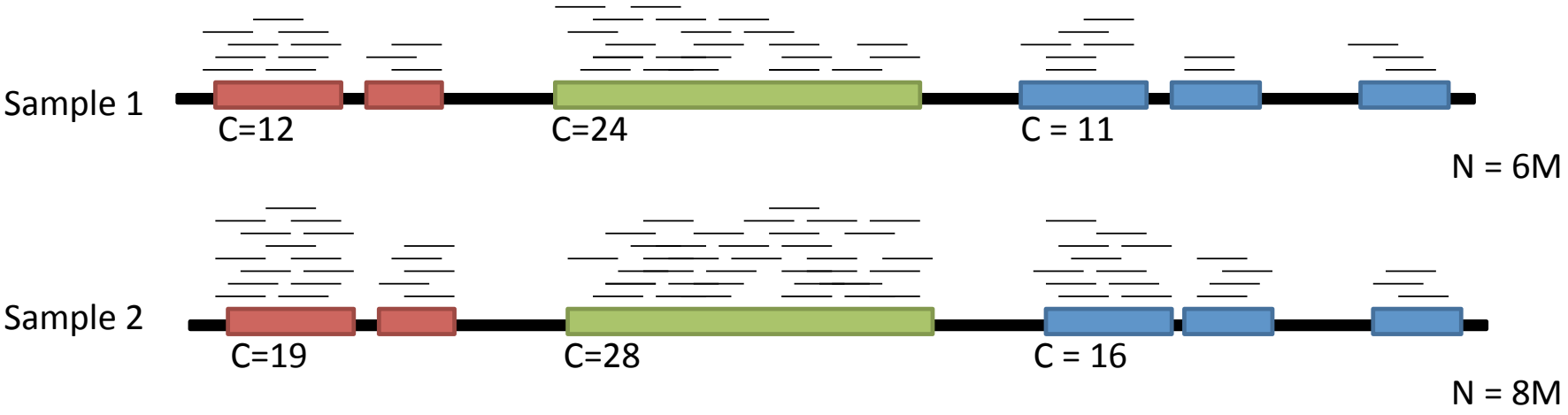
- RPKM = $\frac{C}{LN}$
  - C : Number of mappable reads on a feature (eg. transcript, exon, etc.)
  - L: Length of feature (in kb)
  - N: Total number of mappable reads (in millions)

Mortazavi A., et al. *Mapping and quantifying mammalian transcriptomes by RNA-Seq* Nature Methods (2008)

RPKM Example

Gene A 600 bases Gene B 1100 bases Gene C 1400 bases

RPKM = 12/(0.6*6) = 3.33 RPKM = 24/(1.1*6) =3.64 RPKM = 11/(1.4*6) = 1.31



RPKM = 19/(0.6*8) = 3.96 RPKM = 28/(1.1*8) =1.94 RPKM = 16/(1.4*8) = 1.43

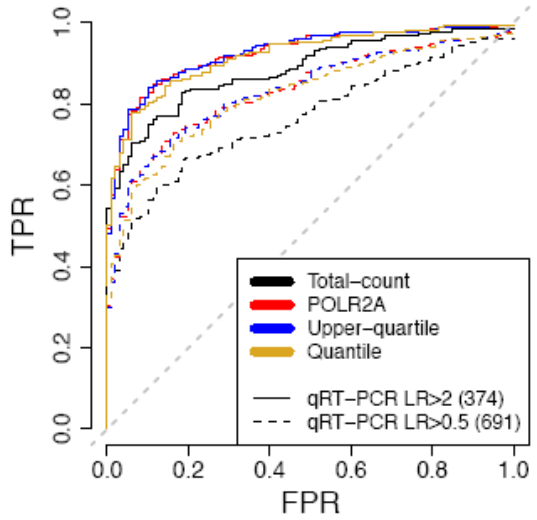
Quantifying Expression: FPKM

- FPKM: Fragments Per Kilobase of transcript per Million fragments mapped
  - Analogous to RPKM but does not use read counts.
  - the relative abundances of transcripts are described in terms of the expected biological objects (*fragments*) observed from an RNA-Seq experiment, which in the future may not be represented by single read

Trapnell, C. et al *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation* Nature Biotechnology (2010)

Quantifying Expression: Normalization Methods

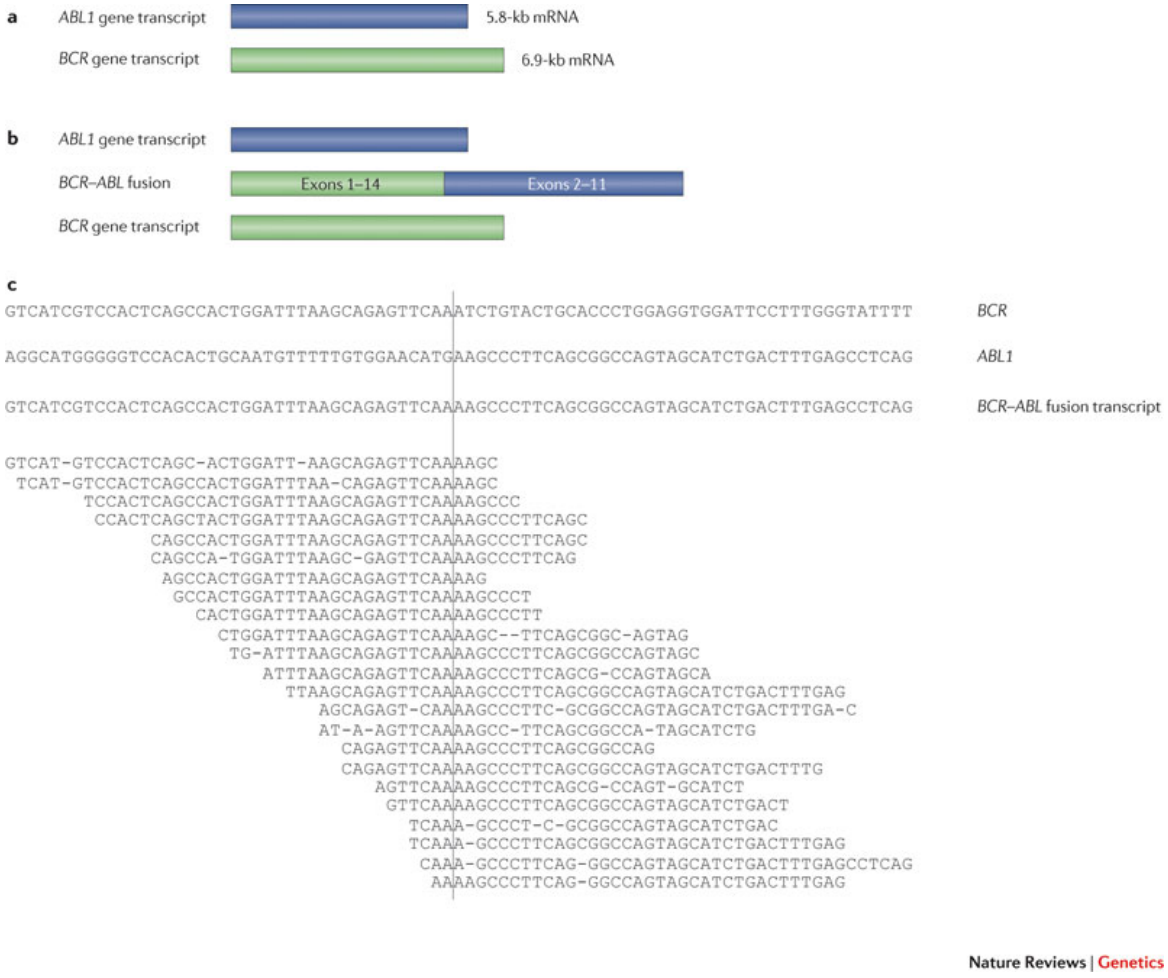
- Total-count (eg. RPKM)
- Upper Quartile (eg. 75th percentile):Similar to Total-count but per-lane upper-quartile of counts for genes with reads in at least one lane.
- Quantile: For each lane the distribution of read counts is matched to a reference distribution defined in terms of median counts



(a) mRNA-Seq

Bullard, J., et al. *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments*
BMC Bioinformatics (2010)

RNA-Seq Applications: Gene Fusion

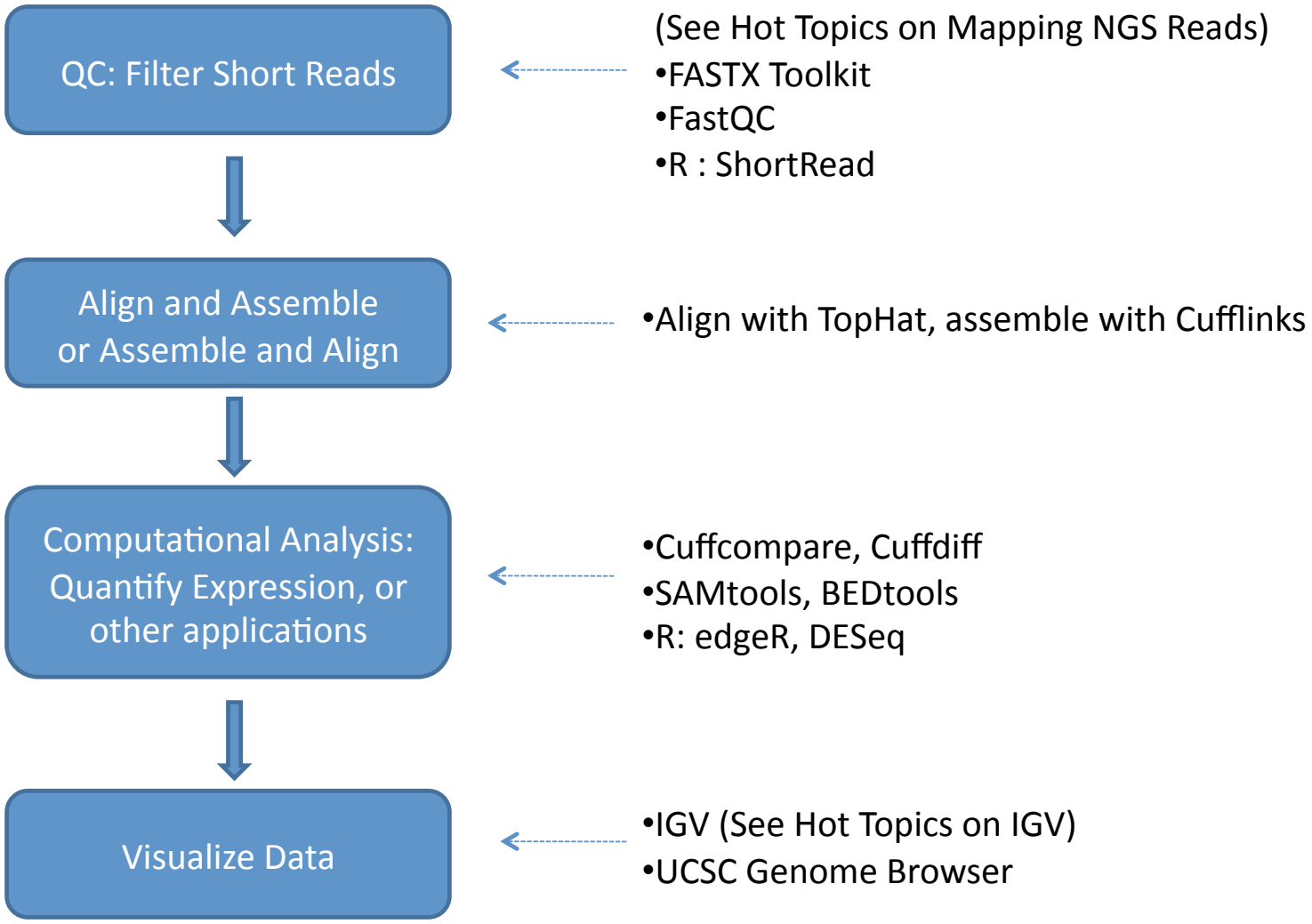


Ozsolak, F. and Milos, P. *RNA sequencing: advances, challenges and opportunities* Nature Reviews Genetics (2011)

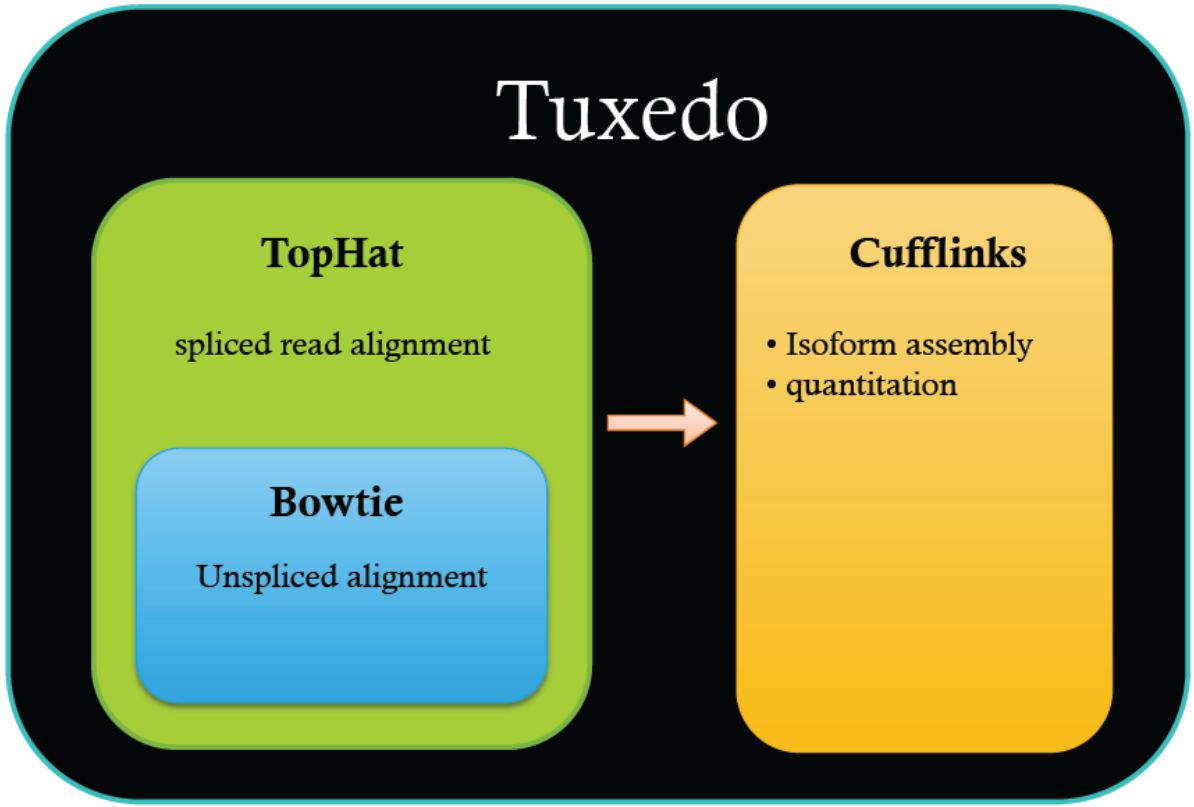
Outline

- Intro to RNA-Seq
  - Biological Questions
  - Comparison with Other Methods
  - RNA-Seq Protocol
- RNA-Seq Applications
  - Identifying Transcripts
  - Quantification
  - Other Applications
- Expression Profiling Steps and Software
- Running TopHat and Cufflinks (Commands)

Expression Profiling Workflow

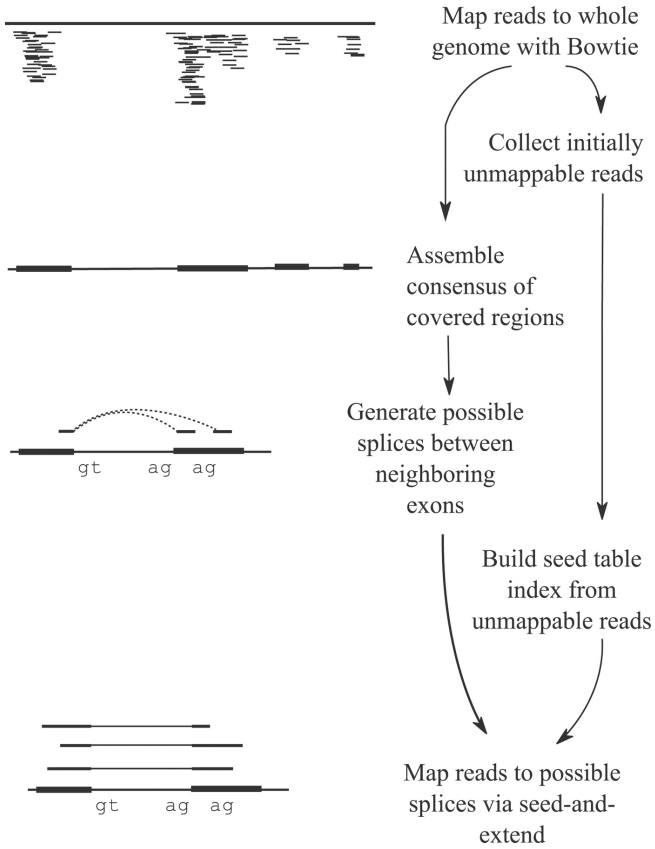


The Tuxedo Tools



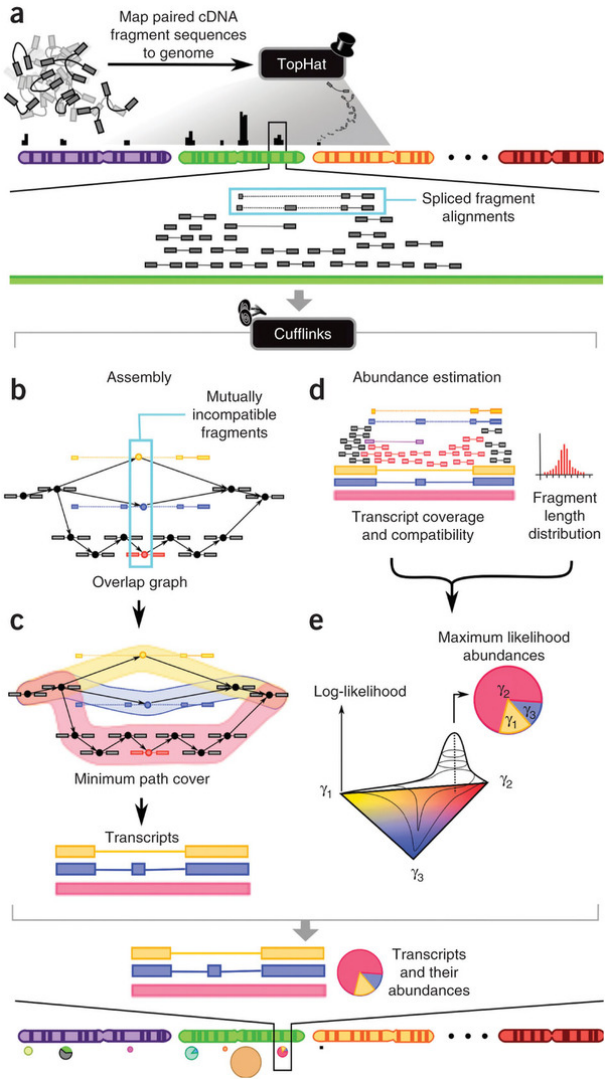
http://mged12-deep-sequencing-analysis.wikispaces.com/file/view/Cole_MGED_tutorial_slides.pdf

TopHat Algorithm



Trapnell, C., et al *TopHat: discovering splice junctions with RNA-Seq* Bioinformatics (2009)

Cufflinks Algorithm



Trapnell, C., et al *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation* Nature Biotechnology (2010)

Outline

- Intro to RNA-Seq
  - Biological Questions
  - Comparison with Other Methods
  - RNA-Seq Protocol
- RNA-Seq Applications
  - Identifying Transcripts
  - Quantification
  - Other Applications
- Expression Profiling Steps and Software
- Running TopHat and Cufflinks (Commands)

Running TopHat: Align Reads

- TopHat Manual: http://tophat.cbcb.umd.edu/manual.html
- Running TopHat on Tak

Usage:

tophat [options] <bowtie_index> <reads1[,reads2,...,readsN]> [reads1[,reads2,...,readsN]]

eg.

bsub "tophat -p 2 --solexa1.3-quals --max-multihits 5 -o s_1_TopHat_Out /nfs/genomes/mouse_gp_jul_07_no_random/bowtie/mm9 s_1_sequence.txt"

Options (See Manual for all available options):

-o/--output-dir	Sets the name of the directory in which TopHat will write all of its output.
--solexa-quals	Use the Solexa scale for quality values in FASTQ files.
--solexa1.3-quals	As of the Illumina GA pipeline version 1.3, quality scores are encoded in Phred-scaled base-64. Use this option for FASTQ files from pipeline 1.3 or later.
-p/--num-threads	Use this many threads to align reads. The default is 1.
-g/--max-multihits	Instructs TopHat to allow up to this many alignments to the reference for a given read, and suppresses all alignments for reads with more than this many alignments. The default is 40.

TopHat Output

- Output of TopHat is a bam file. Binary version of Sequence Alignment/Map (SAM) file
- Use Integrative Genomics Viewer (IGV) to view bam file or use SAMtools to analyze bam file

eg. SAM File

```
WICMT-SOLEXA:1:20:670:1533#  137  chr1  3240920 3  30M  *  0  0  CTGGATCTGGACCTGGACCTGGATCTATAT  ::::::::::::::::-::::::::::::  NM:i:1  NH:i:2  CC:Z:chr6  CP:i:83893005
WICMT-SOLEXA:1:69:135:1285#  89  chr1  3269437 1  30M  *  0  0  TGCCTAAACTTATTAAGGCAGGCCATGGGC  :((/+:::(+:+':/:+++&+//':++:::  NM:i:2  NH:i:4  CC:Z:chr7  CP:i:20934843
WICMT-SOLEXA:1:84:584:747#  153  chr1  3270083 0  30M  *  0  0  AGCAAGTTTTTTNTTAGCCCTAGATTCCAG  ::::::::::::%::::::::::::::::  NM:i:1  NH:i:5  CC:Z:=  CP:i:136301734
WICMT-SOLEXA:1:75:1357:1675#  163  chr1  3522128 255  30M  =  3522287 0  GTGGCTTTGTGGTCTTCACCAACCTTTCTC  ::::::::::::::::::::::::::::::  NM:i:1  NH:i:1
WICMT-SOLEXA:1:75:1357:1675#  83  chr1  3522287 255  30M  =  3522128 0  CTGTAGGTGTAATCCTAAATTCTTATTACG  ::::::::::::::::::::::::::::::  NM:i:0  NH:i:1
WICMT-SOLEXA:1:8:59:283#  153  chr1  3522536 3  30M  *  0  0  TTTCTGCTTTGATTATGGTACTGATGTCTG  ::::::::::4::::::::::::::::  NM:i:2  NH:i:2  CC:Z:chr5  CP:i:134317691
WICMT-SOLEXA:1:12:1161:945#  89  chr1  3523371 1  30M  *  0  0  TCTACATAGCCCAAACTGGCTTTGGACTCT  ::::::::::::::::::::::::::::::  NM:i:0  NH:i:3  CC:Z:chr10  CP:i:117172515
WICMT-SOLEXA:1:45:1469:1826#  73  chr1  3620888 3  30M  *  0  0  CAAGTATTTAATGTTTTCATTAAATTGTTT  :::::::::::::::::::::::::::4:::  NM:i:0  NH:i:2  CC:Z:chr11  CP:i:22903295
WICMT-SOLEXA:1:14:536:150#  73  chr1  3620943 3  30M  *  0  0  CTGGAAGACAATGTCCAAAAACTCTGAATC  :::::::::::::::::::::::::::%::&:  NM:i:1  NH:i:2  CC:Z:chr11  CP:i:22903240
WICMT-SOLEXA:1:66:646:1188#  137  chr1  3662923 0  30M  *  0  0  AAAAAAAAAACACCACCCCCAACAAAAAAA  +00++0+0+''0++++:00::.&:::,:,:  NM:i:2  NH:i:5  CC:Z:chr10  CP:i:94881279
```

Cufflinks: Assemble and Quantify Reads

- Cufflinks Manual:
  http://cufflinks.cbcb.umd.edu/manual.html
- Running Cufflinks on Tak
- Optional: Supply annotation in GTF format with "-G" option

Usage:

cufflinks [options] <hits.bam>

eg.

bsub "cufflinks -p 2 -o s_1_Cufflinks_Out s_1_TopHat_Out/accepted_hits.bam"

eg. cufflinks will assemble and quantify using known transcripts using gtf file supplied

bsub "cufflinks -p 2 -G transcripts.gtf accepted_hits.bam"

Cufflinks Output

- Output of Cufflinks is a GTF file with assembled isoforms

eg.

chr1 Cufflinks transcript 36321447 36330270 1000 - . gene_id "Neurl3"; transcript_id "NM_153408"; FPKM "3.7155221121"; frac "1.000000"; conf_lo "0.000000"; conf_hi "7.570660"; cov "0.649922";
chr1 Cufflinks exon 36321447 36323398 1000 - . gene_id "Neurl3"; transcript_id "NM_153408"; exon_number "1"; FPKM "3.7155221121"; frac "1.000000"; conf_lo "0.000000"; conf_hi "7.570660"; cov "0.649922";
chr1 Cufflinks exon 36325501 36325554 1000 - . gene_id "Neurl3"; transcript_id "NM_153408"; exon_number "2"; FPKM "3.7155221121"; frac "1.000000"; conf_lo "0.000000"; conf_hi "7.570660"; cov "0.649922";
chr1 Cufflinks exon 36326058 36326546 1000 - . gene_id "Neurl3"; transcript_id "NM_153408"; exon_number "3"; FPKM "3.7155221121"; frac "1.000000"; conf_lo "0.000000"; conf_hi "7.570660"; cov "0.649922";
chr1 Cufflinks exon 36330183 36330270 1000 - . gene_id "Neurl3"; transcript_id "NM_153408"; exon_number "4"; FPKM "3.7155221121"; frac "1.000000"; conf_lo "0.000000"; conf_hi "7.570660"; cov "0.649922";
chr1 Cufflinks transcript 36364578 36380874 4 + . gene_id "Arid5a"; transcript_id "NM_145996"; FPKM "0.0015751054"; frac "0.002360"; conf_lo "0.000000"; conf_hi "0.081996"; cov "0.000263";
chr1 Cufflinks exon 36364578 36364681 4 + . gene_id "Arid5a"; transcript_id "NM_145996"; exon_number "1"; FPKM "0.0015751054"; frac "0.002360"; conf_lo "0.000000"; conf_hi "0.081996"; cov "0.000263";
chr1 Cufflinks exon 36373054 36373172 4 + . gene_id "Arid5a"; transcript_id "NM_145996"; exon_number "2"; FPKM "0.0015751054"; frac "0.002360"; conf_lo "0.000000"; conf_hi "0.081996"; cov "0.000263";
chr1 Cufflinks exon 36374929 36375026 4 + . gene_id "Arid5a"; transcript_id "NM_145996"; exon_number "3"; FPKM "0.0015751054"; frac "0.002360"; conf_lo "0.000000"; conf_hi "0.081996"; cov "0.000263";
chr1 Cufflinks exon 36375333 36375498 4 + . gene_id "Arid5a"; transcript_id "NM_145996"; exon_number "4"; FPKM "0.0015751054"; frac "0.002360"; conf_lo "0.000000"; conf_hi "0.081996"; cov "0.000263";
chr1 Cufflinks exon 36375837 36380874 4 + . gene_id "Arid5a"; transcript_id "NM_145996"; exon_number "5"; FPKM "0.0015751054"; frac "0.002360"; conf_lo "0.000000"; conf_hi "0.081996"; cov "0.000263";

Local Resources

- Description of available files, see

`/nfs/genomes/BaRC_Genomes_README.txt`

➢ Bowtie index

`/nfs/genomes/<species>/bowtie`

eg.

`/nfs/genomes/mouse_gp_jul_07_no_random/bowtie`

➢ GTF files

`/nfs/genomes/<species>/gtf`

eg.

`/nfs/genomes/mouse_gp_jul_07/gtf`

Further Reading

- RNA-Seq

Mortazavi, A., et al. *Mapping and quantifying mammalian transcriptomes by RNA-Seq* Nature Methods 5(7): 621-628 (2008)

Wang, Z., at al. *RNA-Seq: a revolutionary tool for transcriptomics* Nature Reviews Genetics 10:57-63 (2009)

Ozsolak, F. and Milos P.M. *RNA sequencing: advances, challenges, and opportunities* Nature Reviews Genetics 12:87-98 (2011)

- TopHat

Trapnell, C., et al. *TopHat: discovering splice junctions with RNA-Seq* Bioinformatics 25(9) 1105-1111 (2009)

- Cufflinks

Trapnell, C., et al. *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation* Nature Biotechnology 28(5) 511-515 (2010)

Online Community
Forum and Discussion

- http://seqanswers.com/