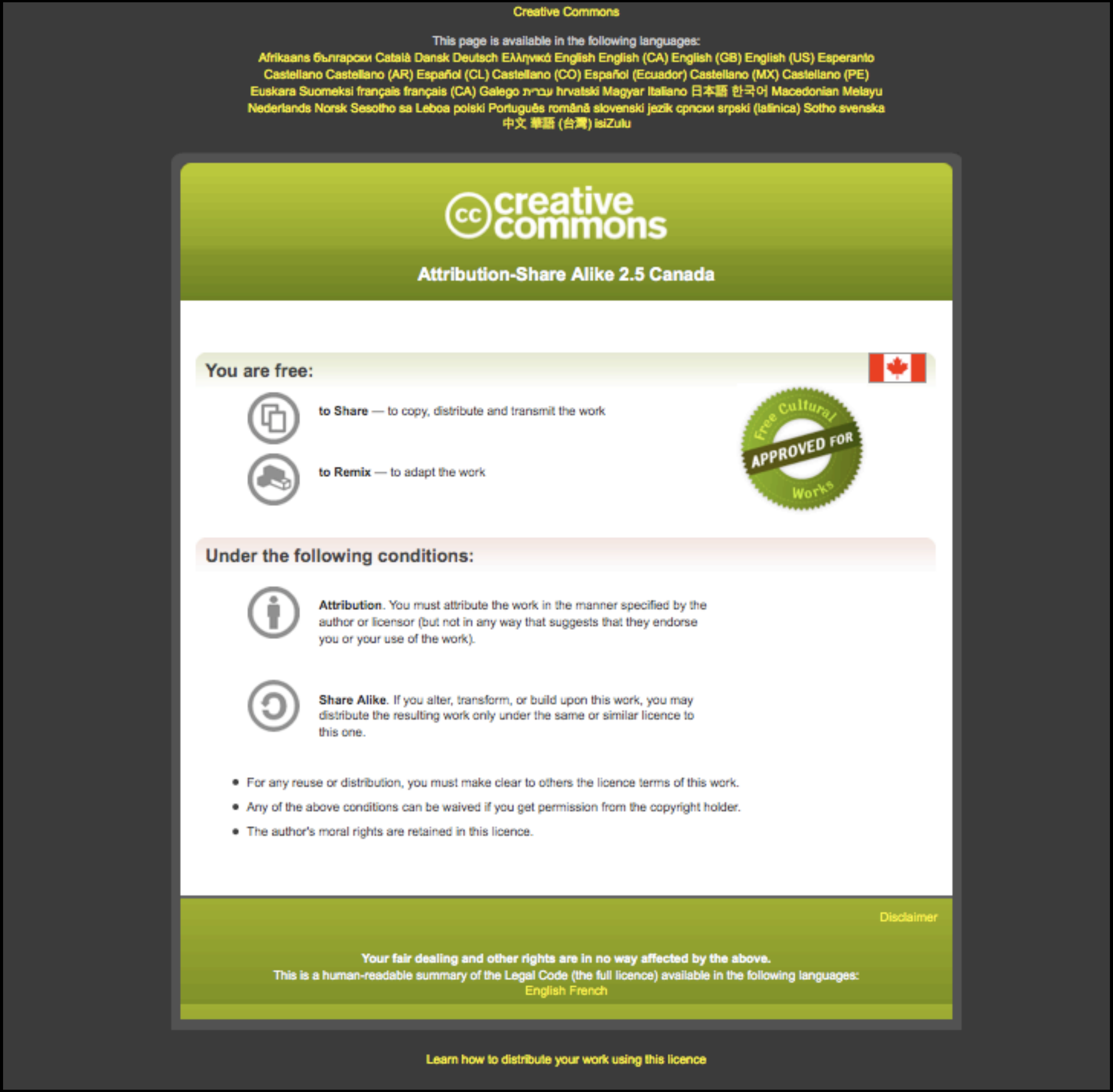
bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca



Creative Commons

This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE) Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska 中文 華語 (台灣) isiZulu

creative commons

Attribution-Share Alike 2.5 Canada

You are free:

to Share — to copy, distribute and transmit the work

to Remix — to adapt the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

Disclaimer

Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French

Learn how to distribute your work using this licence

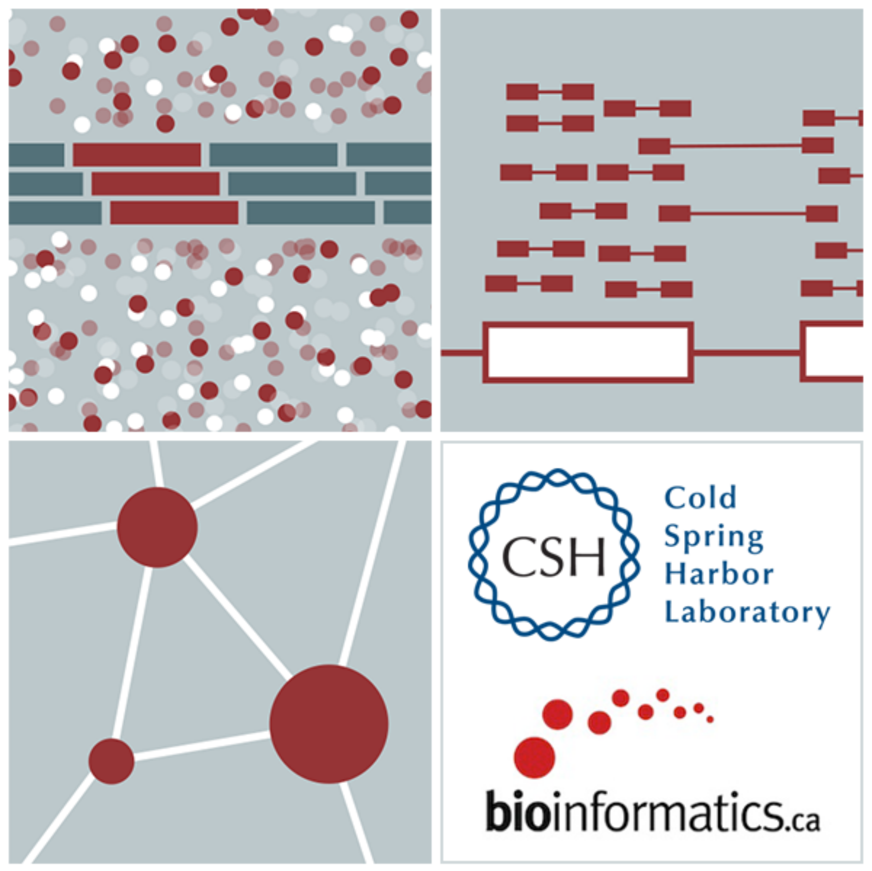
RNA-Seq Module 3
Expression and Differential Expression (lecture)

Malachi Griffith, Obi Griffith, Fouad Yousif

Informatics for RNA-seq Analysis

July 10-12, 2017

bioinformatics.ca



THE GENOME INSTITUTE at Washington University

Learning objectives of the course

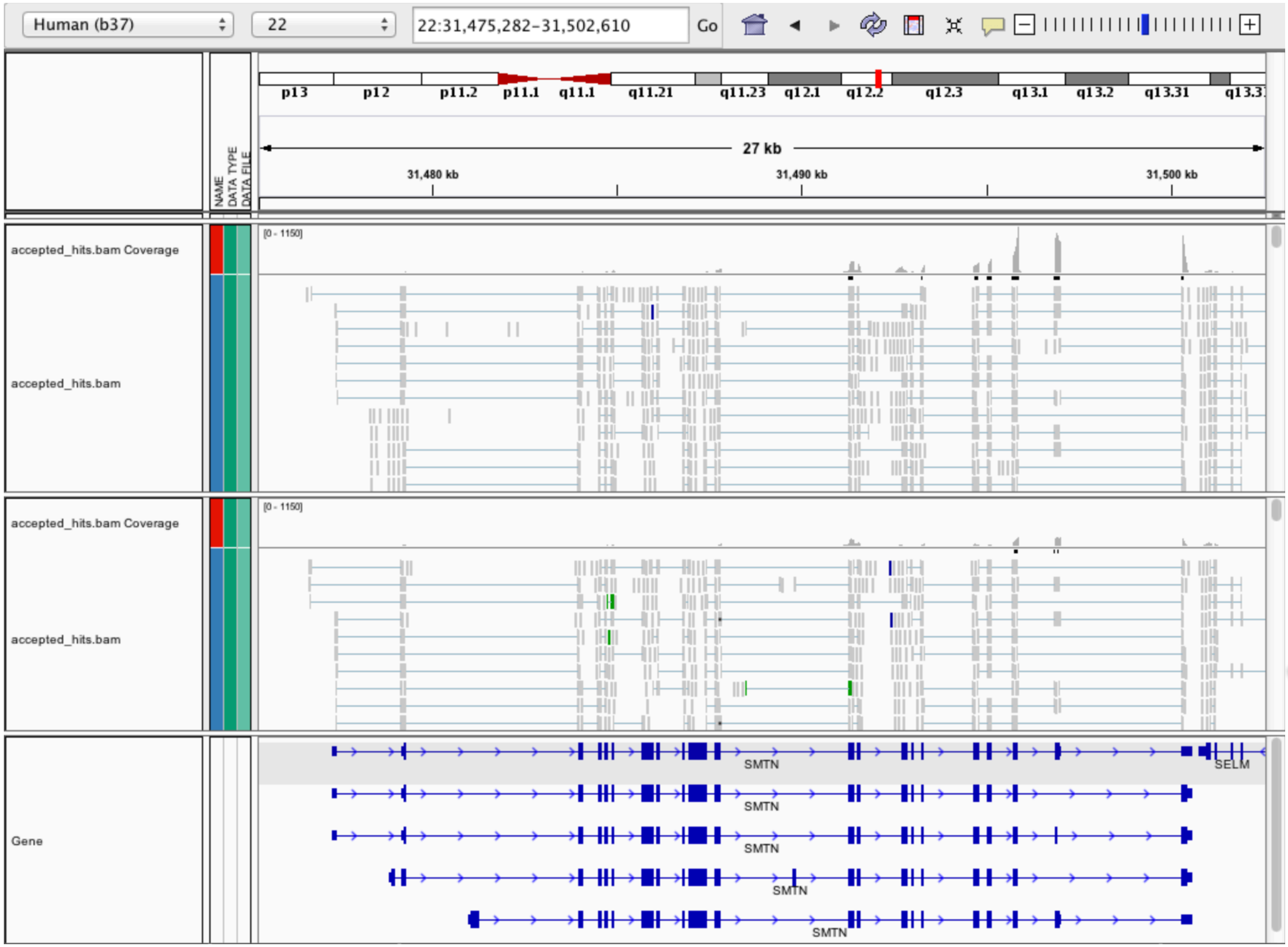
- Module 1: Introduction to RNA Sequencing
- Module 2: Alignment and Visualization
- **Module 3: Expression and Differential Expression**
- Module 4: Isoform Discovery and Alternative Expression

- Tutorials
  - Provide a working example of an RNA-seq analysis pipeline
  - Run in a 'reasonable' amount of time with modest computer resources
  - Self contained, self explanatory, portable

Learning Objectives of Module

- Expression estimation for known genes and transcripts
- ‘FPKM’ expression estimates vs. ‘raw’ counts
- Differential expression methods
- Downstream interpretation of expression and differential estimates
  - multiple testing, clustering, heatmaps, classification, pathway analysis, etc.

Expression estimation for known genes and transcripts



Human (b37) | 22 | 22:31,475,282-31,502,610 | Go

3' bias

Down-regulated

What is FPKM (RPKM)

- RPKM: Reads Per Kilobase of transcript per Million mapped reads.
- FPKM: Fragments Per Kilobase of transcript per Million mapped reads.
- In RNA-Seq, the relative expression of a transcript is proportional to the number of cDNA fragments that originate from it. However:
  - The number of fragments is also biased towards larger genes
  - The total number of fragments is related to total library depth
- FPKM (RPKM) attempt to normalize for gene size and library depth

- FPKM (RPKM) = (10^9 * C) / (N * L)
  - C = number of mappable reads/fragments for a gene/transcript/exon/etc
  - N = total number of mappable reads/fragments in the library
  - L = number of base pairs in the gene/transcript/exon/etc

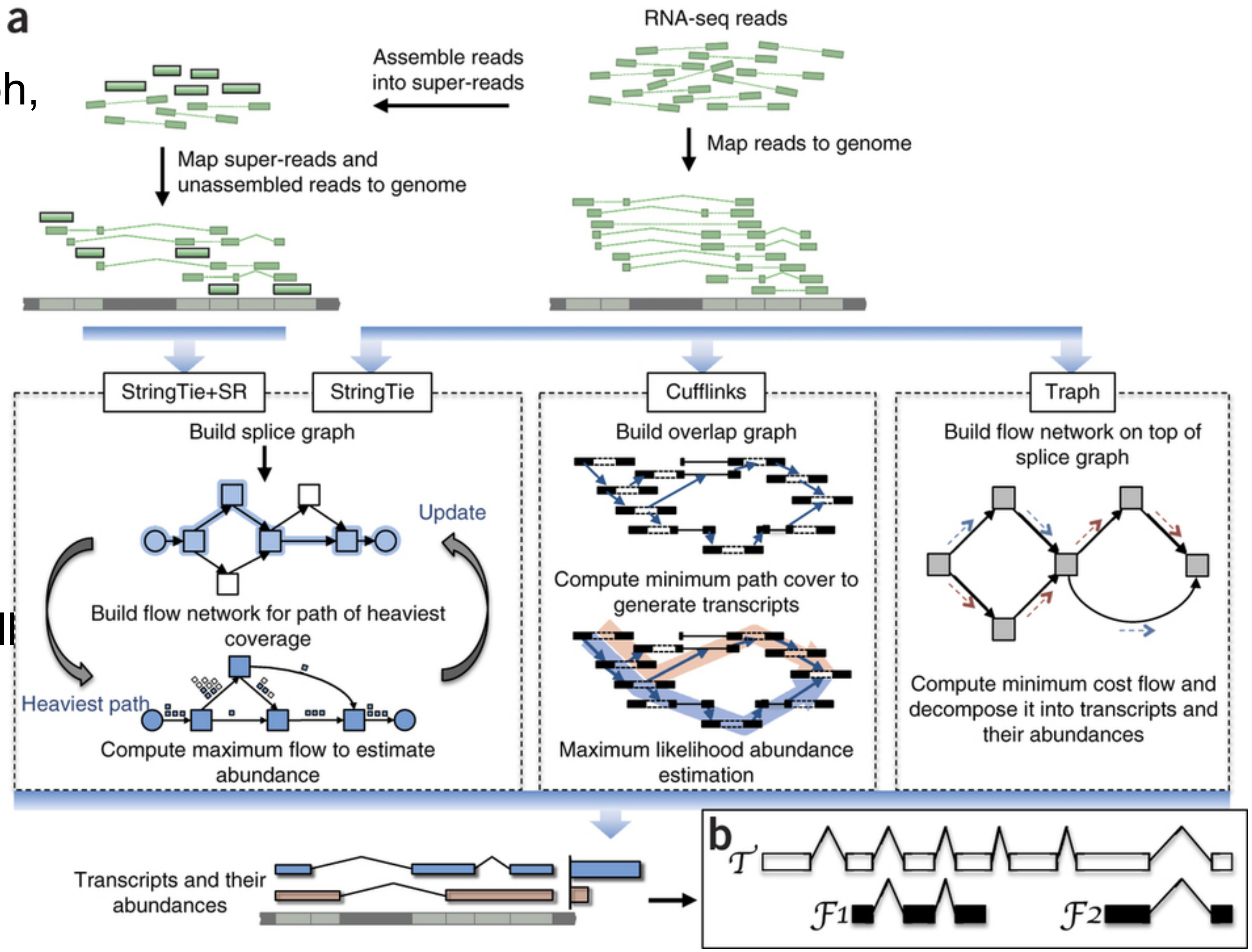
- http://www.biostars.org/p/11378/
- http://www.biostars.org/p/68126/

How do FPKM and TPM differ?

- TPM: Transcript per Kilobase Million
- The difference is in the order of operations:
  - FPKM
    - 1) Sum sample/library fragments per million
    - 2) Divide gene/transcript fragment count by #1
      - fragments per million, FPM
    - 3) Divide FPM by length of gene in kilobases (FPKM)
  - TPM
    - 1) Divide fragment count by length of transcript
      - fragments per kilobase, FPK
    - 2) Sum all FPK for sample/library per million
    - 3) Divide #1 by #3 (TPM)
- http://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

How does StringTie work?

- StringTie iteratively extracts the heaviest path from a splice graph, constructs a flow network, computes maximum flow to estimate abundance, and then updates the splice graph by removing reads that were assigned by the flow algorithm. This process repeats until all reads have been assigned.
- Annotated transcript T for which read data covers only the fragments F1 and F2.


a
RNA-seq reads
Assemble reads into super-reads
Map super-reads and unassembled reads to genome
Map reads to genome
StringTie+SR
StringTie
Cufflinks
Traph
Build splice graph
Update
Build flow network for path of heaviest coverage
Heaviest path
Compute maximum flow to estimate abundance
Build overlap graph
Compute minimum path cover to generate transcripts
Maximum likelihood abundance estimation
Build flow network on top of splice graph
Compute minimum cost flow and decompose it into transcripts and their abundances
Transcripts and their abundances
b
T
F1
F2

Pertea et al. Nature Biotechnology, 2015

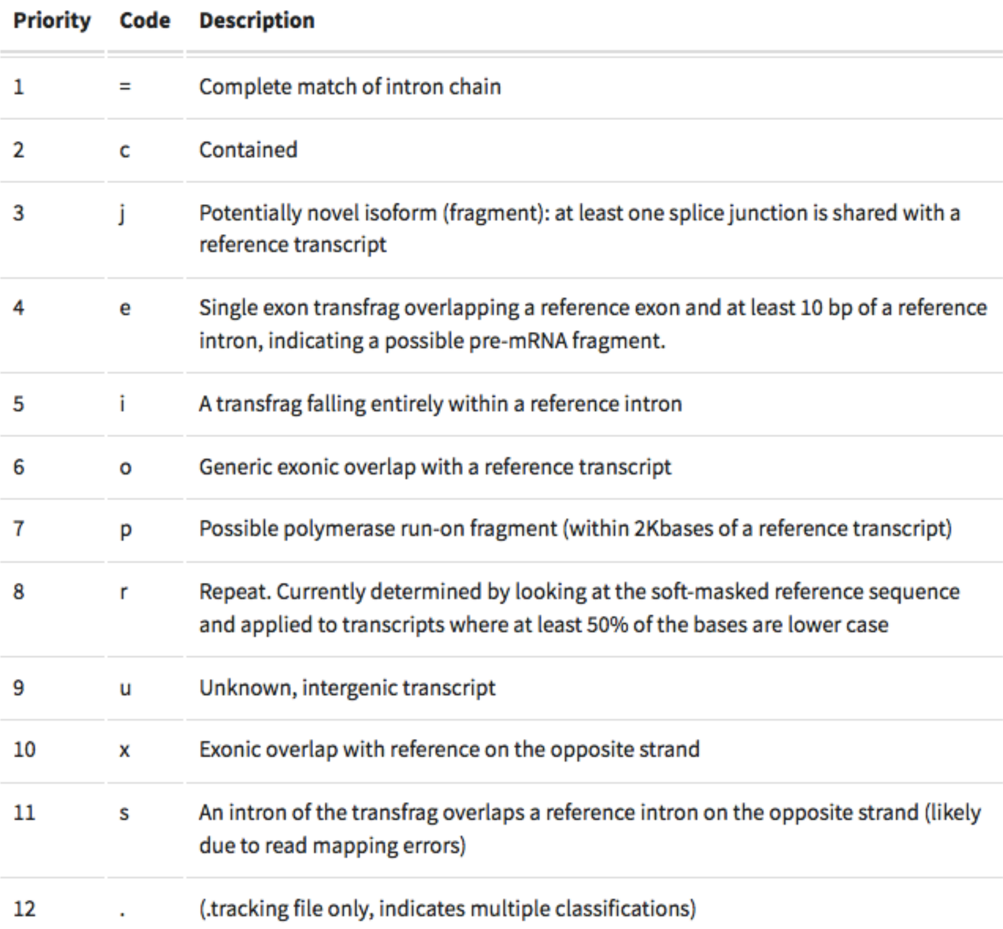
StringTie -merge

- Merge together all gene structures from all samples
  - Some samples may only partially represent a gene structure
- Allows for the incorporation of known transcripts with assembled, potentially novel transcripts
- For de novo or reference guided mode, we will rerun StringTie with the merged transcript assembly.

Pertea et al. Nature Protocols, 2016

gffcompare

- gffcompare will compare a merged transcript GTF with known annotation, also in GTF/GFF3 format
- http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/index.html#cuffcompare-output-files



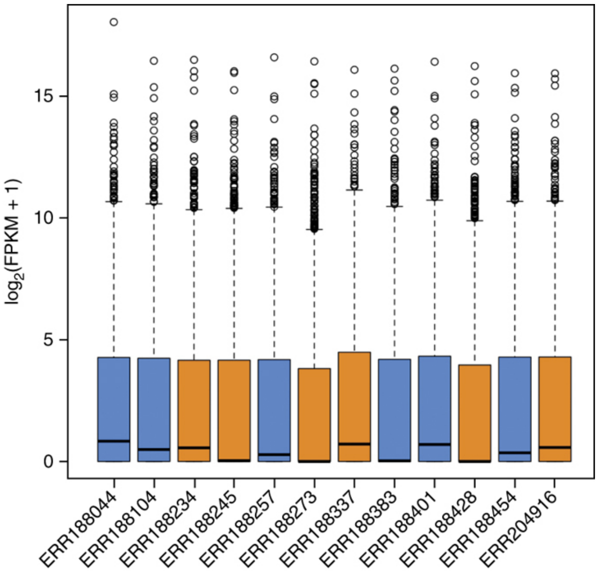
Priority	Code	Description
1	=	Complete match of intron chain
2	c	Contained
3	j	Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript
4	e	Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment.
5	i	A transfrag falling entirely within a reference intron
6	o	Generic exonic overlap with a reference transcript
7	p	Possible polymerase run-on fragment (within 2Kbases of a reference transcript)
8	r	Repeat. Currently determined by looking at the soft-masked reference sequence and applied to transcripts where at least 50% of the bases are lower case
9	u	Unknown, intergenic transcript
10	x	Exonic overlap with reference on the opposite strand
11	s	An intron of the transfrag overlaps a reference intron on the opposite strand (likely due to read mapping errors)
12	.	(.tracking file only, indicates multiple classifications)

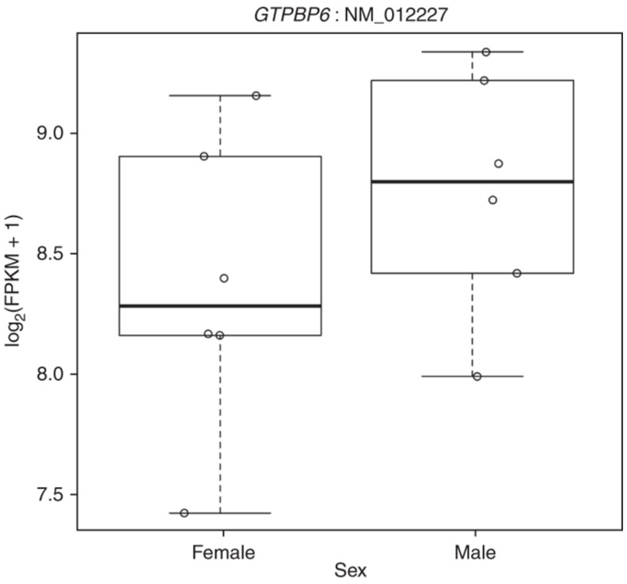
Ballgown for Differential Expression

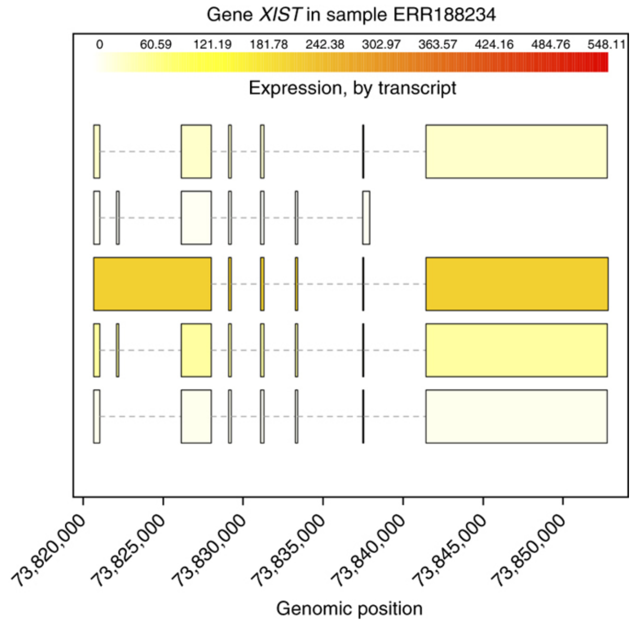
- Parametric F-test comparing nested linear models
- Two models are fit to each feature, using expression as the outcome
  - one including the covariate of interest (e.g., case/control status or time) and one not including that covariate.
- An F statistic and p-value are calculated using the fits of the two models.
  - A significant p-value means the model including the covariate of interest fits significantly better than the model without that covariate, indicating differential expression.
- We adjust for multiple testing by reporting q-values:
  - q < 0.05 the false discovery rate should be controlled at ~5%.

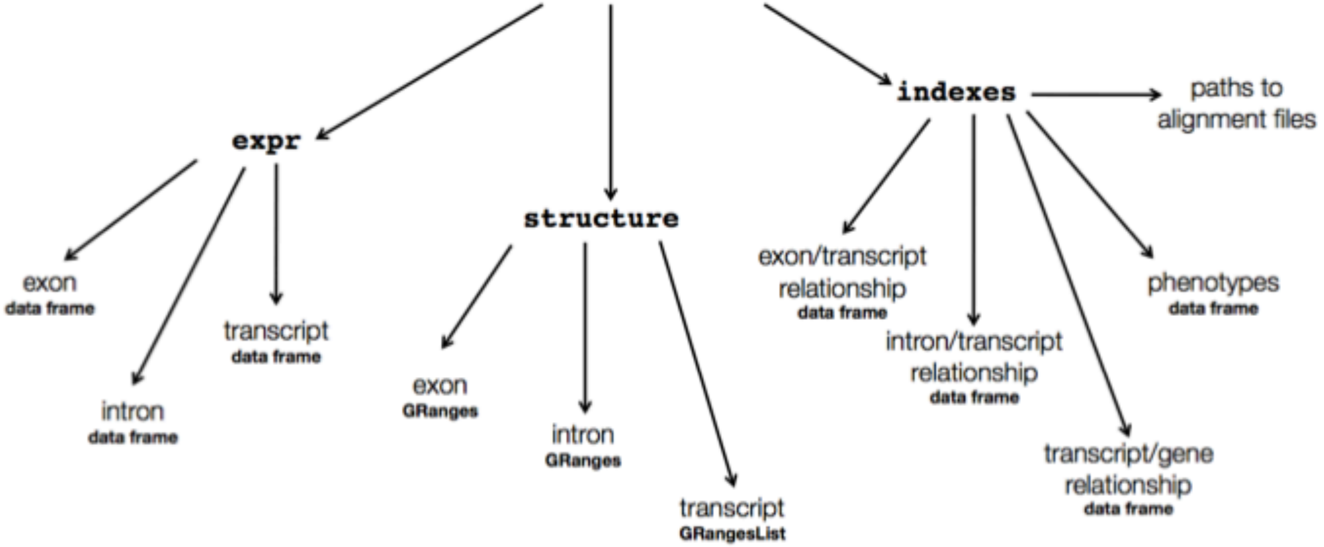
Frazee et al. (2014)

Ballgown for Visualization with R









ballgown object

expr

exon
data frame

intron
data frame

transcript
data frame

structure

exon
GRanges

intron
GRanges

transcript
GRangesList

indexes

paths to alignment files

exon/transcript relationship
data frame

intron/transcript relationship
data frame

transcript/gene relationship
data frame

phenotypes
data frame

Alternatives to FPKM

- Raw read counts as an alternate for differential expression analysis
  - Instead of calculating FPKM, simply assign reads/fragments to a defined set of genes/transcripts and determine “raw counts”
    - Transcript structures could still be defined by something like cufflinks
- HTSeq (htseq-count)
  - http://www-huber.embl.de/users/anders/HTSeq/doc/count.html

  -
    ```
    htseq-count --mode intersection-strict --stranded no --minaqual
    1 --type exon --idattr transcript_id accepted_hits.sam chr22.gff
    > transcript_read_counts_table.tsv
    ```
  - Important caveat of ‘transcript’ analysis by htseq-count:
    - http://seqanswers.com/forums/showthread.php?t=18068

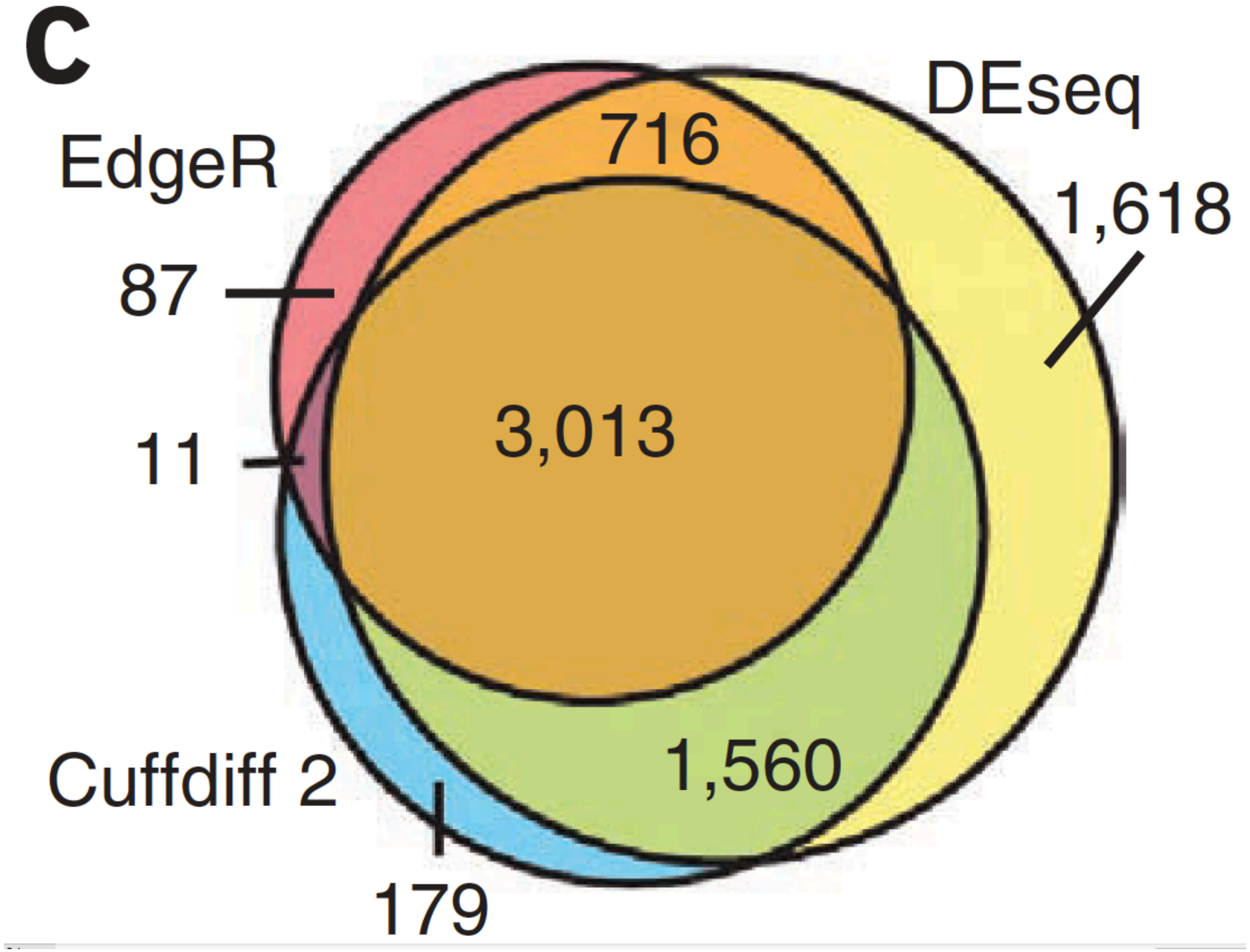
‘FPKM’ expression estimates vs. ‘raw’ counts

- Which should I use?
- FPKM
  - When you want to leverage benefits of tuxedo suite
  - Good for visualization (e.g., heatmaps)
  - Calculating fold changes, etc.
- Counts
  - More robust statistical methods for differential expression
  - Accommodates more sophisticated experimental designs with appropriate statistical tests

Alternative differential expression methods

- Raw count approaches
  - DESeq - http://www-huber.embl.de/users/anders/DESeq/
  - edgeR - http://www.bioconductor.org/packages/release/bioc/html/edgeR.html
  - Others…

Multiple approaches advisable



Lessons learned from microarray days

- Hansen et al. “Sequencing Technology Does Not Eliminate Biological Variability.” Nature Biotechnology 29, no. 7 (2011): 572–573.
- Power analysis for RNA-seq experiments
  - http://euler.bc.edu/marthlab/scotty/scotty.php
- RNA-seq need for biological replicates
  - http://www.biostars.org/p/1161/
- RNA-seq study design
  - http://www.biostars.org/p/68885/

Multiple testing correction

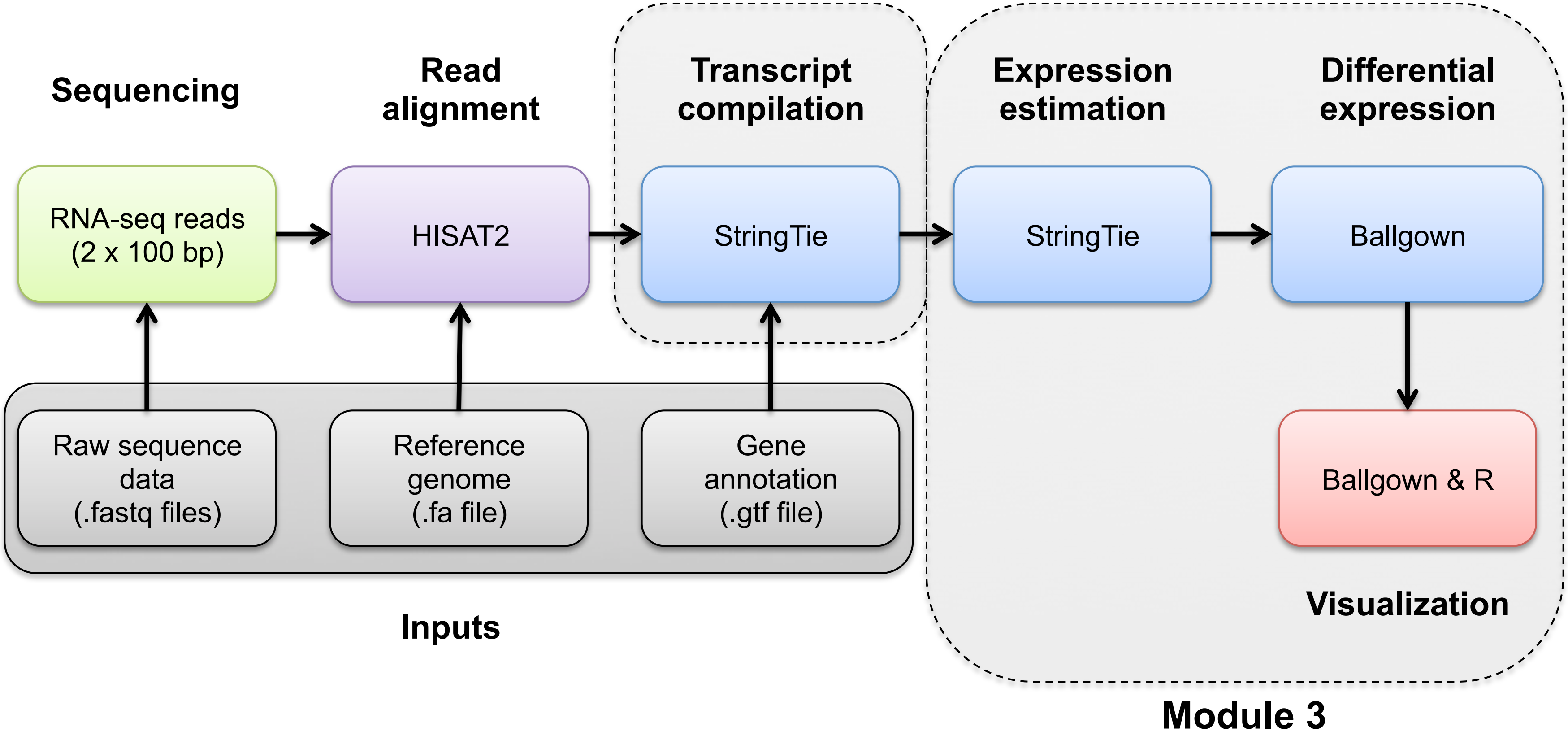
- As more attributes are compared, it becomes more likely that the treatment and control groups will appear to differ on at least one attribute by random chance alone.
- Well known from array studies
  - 10,000s genes/transcripts
  - 100,000s exons
- With RNA-seq, more of a problem than ever
  - All the complexity of the transcriptome
  - Almost infinite number of potential features
    - Genes, transcripts, exons, junctions, retained introns, microRNAs, lncRNAs, etc
- Bioconductor multtest
  - http://www.bioconductor.org/packages/release/bioc/html/multtest.html

Downstream interpretation of expression analysis

- Topic for an entire course
- Expression estimates and differential expression lists from StringTie, Ballgown or other alternatives can be fed into many analysis pipelines
- See supplemental R tutorial for how to format expression data and start manipulating in R
- Clustering/Heatmaps
  - Provided by cummeRbund
  - For more customized analysis various R packages exist:
    - hclust, heatmap.2, plotrix, ggplot2, etc.
- Classification
  - For RNA-seq data we still rarely have sufficient sample size and clinical details but this is changing
    - Weka is a good learning tool
    - RandomForests R package (biostar tutorial being developed)
- Pathway analysis
  - IPA
  - Cytoscape
  - Many R/BioConductor packages: http://www.bioconductor.org/help/search/index.html?q=pathway

Introduction to tutorial (Module 3)

HISAT2/StringTie/Ballgown RNA-seq Pipeline



We are on a Coffee Break & Networking Session