



The Elizabeth H.
and James S. McDonnell III

**McDONNELL
GENOME INSTITUTE**
at Washington University

GenViz Module 4: Expression profiling, visualization, and interpretation

Malachi Griffith, Obi Griffith, Zachary Skidmore
Genomic Data Visualization and Interpretation

April 8-12, 2019
Freie Universität Berlin



Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

This is a human-readable summary of (and not a substitute for) the [license](#). [Disclaimer](#).

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material for any purpose, even commercially.



The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:



Attribution — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.



ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](#) as the original.

No additional restrictions — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

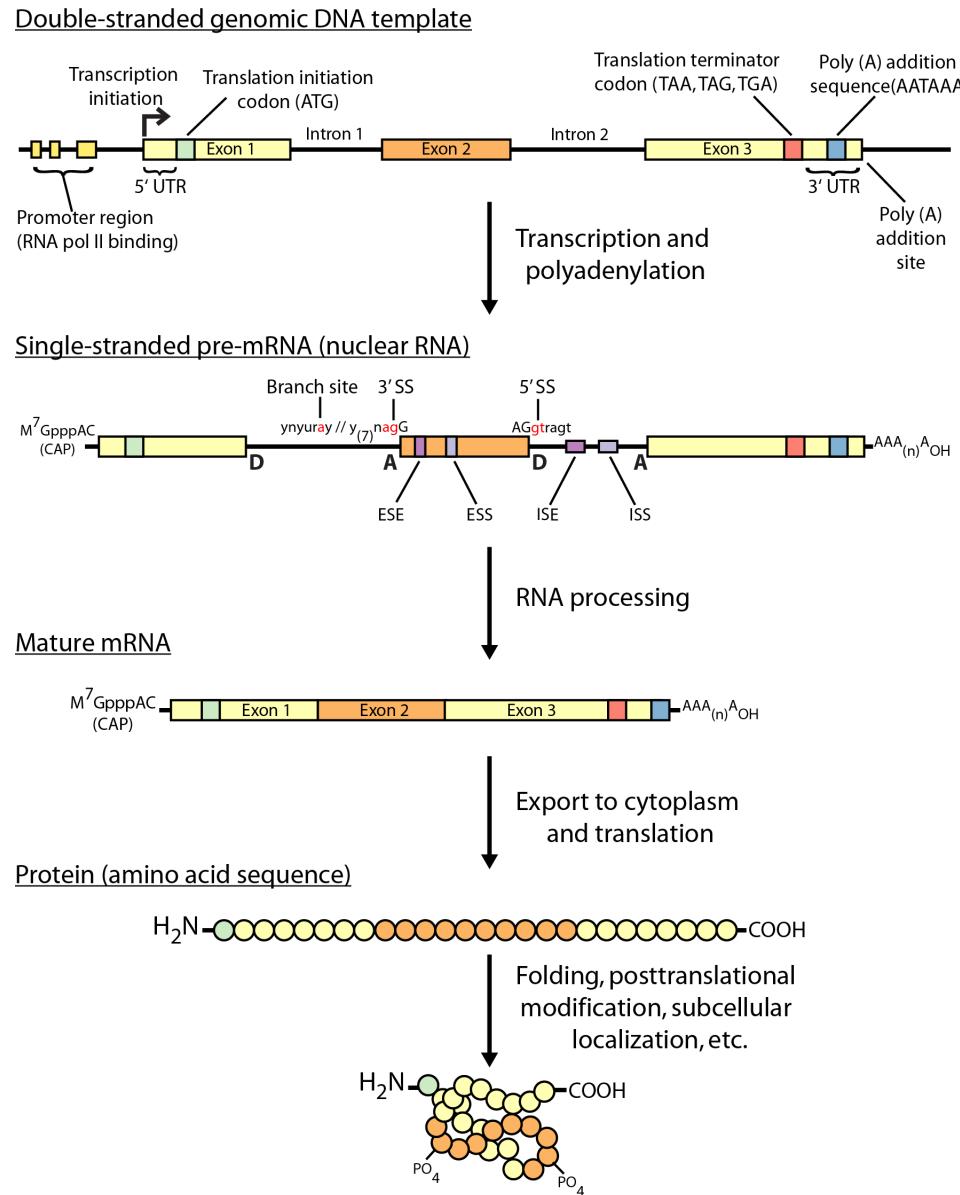
Learning objectives of the course

- Module 1: Introduction to genomic data visualization and interpretation
- Module 2: Using R for genomic data visualization and interpretation
- Module 3: Introduction to GenVisR
- **Module 4: Expression profiling, visualization, and interpretation**
- Module 5: Variant annotation and interpretation
- Module 6: Q & A, discussion, integrated assignments, and working with your own data
- Tutorials
 - Provide working examples of data visualization and interpretation
 - Self contained, self explanatory, portable

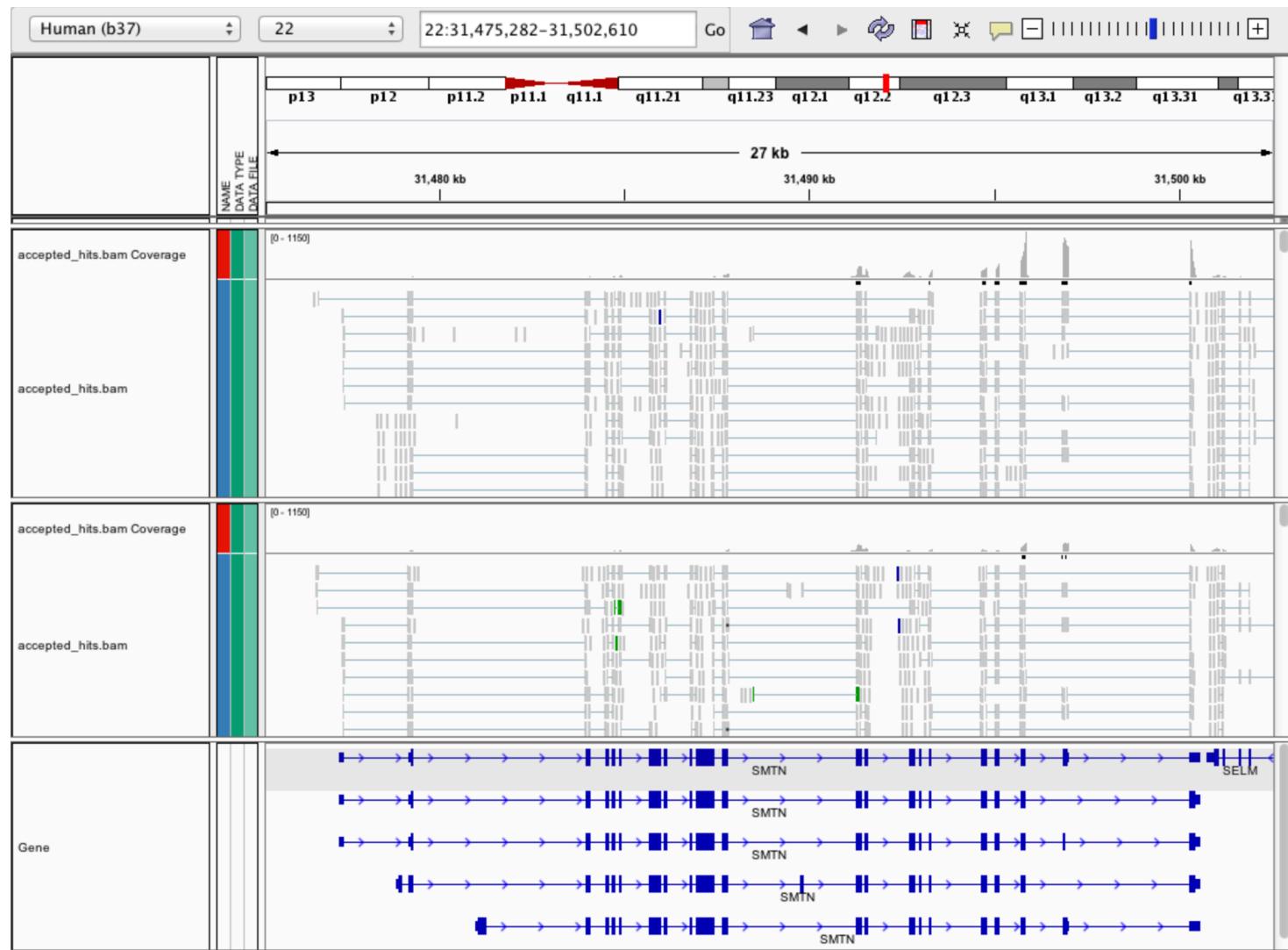
Learning objectives of module 4

- Expression profiling, visualization, and interpretation
 - Expression estimation for known genes (concepts)
 - FPKM' expression estimates vs. 'raw' counts
 - Differential expression methods (DESeq2)
 - Downstream interpretation of expression and differential estimates

Gene expression



Expression estimation for known genes and transcripts



Abundance/expression estimation methods

- Raw Counts
 - [HTSeq-Count](#)
 - [FeatureCounts](#)
 - [StringTie](#)
- RPKM/FPKM values
 - [StringTie](#)
- TPM values
 - [Kallisto](#)
 - [Salmon](#)
 - [StringTie](#)

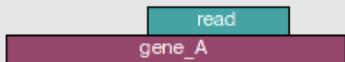
What is FPKM (RPKM)

- 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
- $\text{FPKM (RPKM)} = (10^9 * C) / (N * L)$
 - 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/>

What are raw counts?

- 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>

HTSeq-count basically counts reads supporting a feature (exon, gene) by assessing overlapping coordinates

	union	intersection _strict	intersection _nonempty
 A single read (teal bar) overlaps a single gene (purple bar). The read starts within the gene and ends outside.	gene_A	gene_A	gene_A
 A single read (teal bar) overlaps a single gene (purple bar). The read starts outside the gene and ends within it.	gene_A	no_feature	gene_A
 A single read (teal bar) overlaps two genes (purple bars). The read starts within the first gene and ends within the second gene.	gene_A	no_feature	gene_A
 Two separate reads (teal bars) overlap two separate genes (purple bars). Each read overlaps one gene.	gene_A	gene_A	gene_A
 A single read (teal bar) overlaps two genes (purple and blue bars). The read starts within gene_A and ends within gene_B.	gene_A	gene_A	gene_A
 A single read (teal bar) overlaps two genes (purple and blue bars). The read starts within gene_A and ends within gene_B.	ambiguous	gene_A	gene_A
 A single read (teal bar) overlaps two genes (purple and blue bars). The read starts within gene_A and ends within gene_B.	ambiguous	ambiguous	ambiguous

Whether a read is counted depends on the nature of overlap and “mode” selected

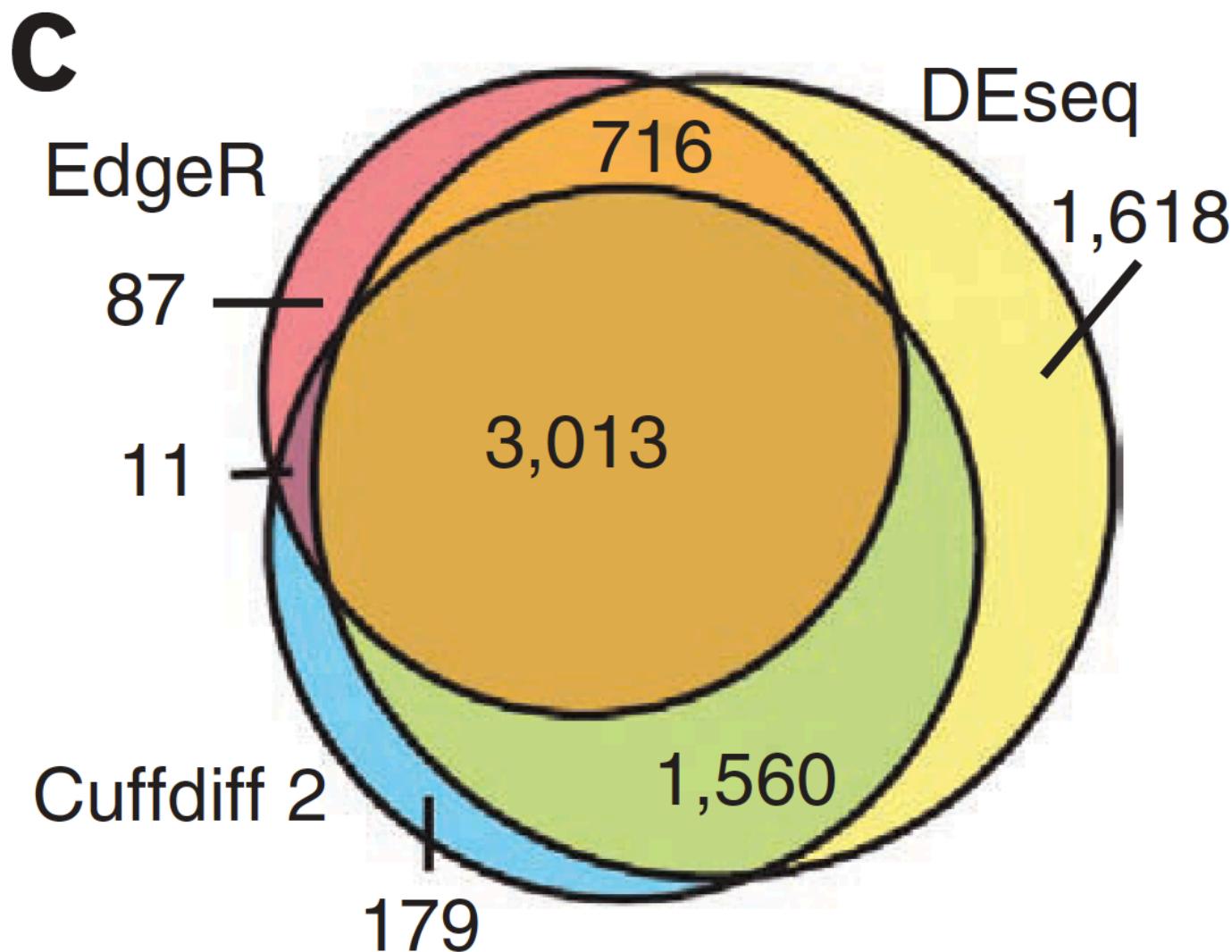
Differential expression methods

- Raw count approaches (gene level)
 - [DESeq2](#)
 - [edgeR](#)
 - Many others...
- FPKM approaches (for transcript level)
 - [Ballgown](#)
 - Helpful explanation ([PMID: 25748911](#))
 - Many others (EBSeq, etc.)
- TPM approaches
 - [Kallisto/Sleuth](#)

'FPKM' expression estimates vs. 'raw' counts

- Which should I use?
 - Long running debate with countless blogs and analyses arguing the advantages of each. The general consensus:
- FPKM
 - Isoform deconvolution
 - Good for straight 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

Multiple approaches advisable



Refer to www.rnaseq.wiki for many, many more details,
resources and exercises