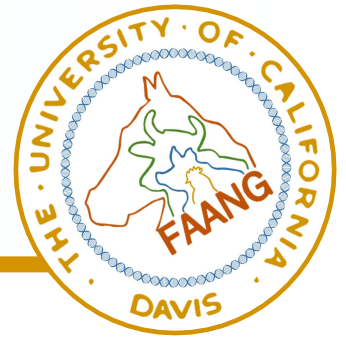
Livestock RAMPAGE for High Definition Transcription Start Site Annotation

Pablo Ross

University of California, Davis

Transcription start site (TSS) annotation is lacking in livestock genomes





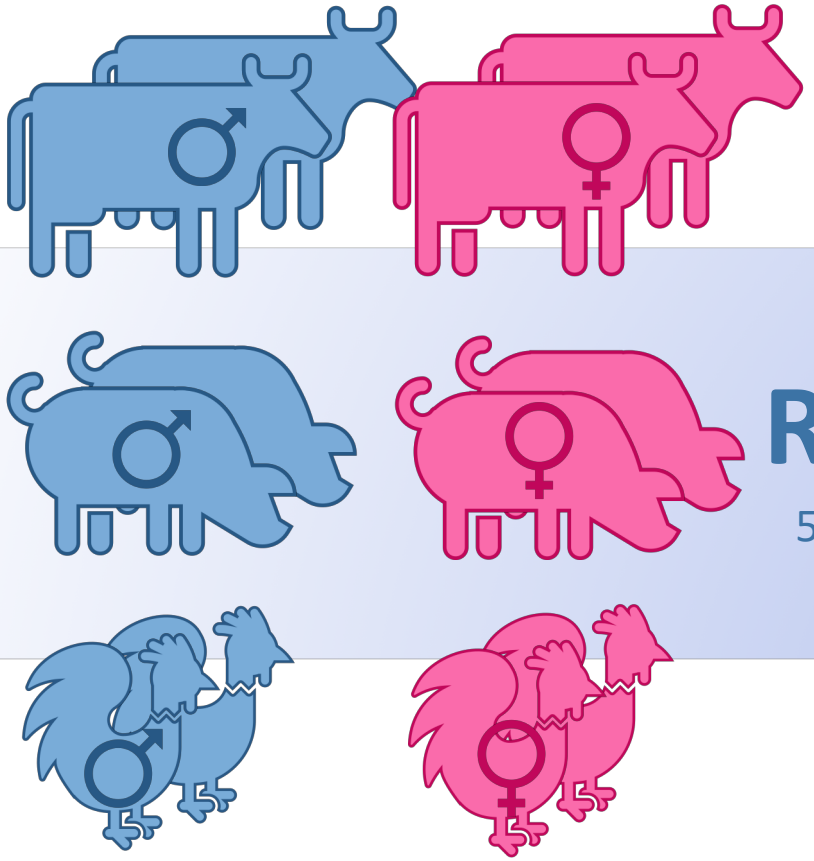

	Cattle	Pig	Chicken	Human
Protein-coding genes	19,994	21,607	15,508	19,814
Protein-coding transcripts	22,118	25,511	16,354	79,712
Transcripts per gene	1.1	1.2	1.1	4.0
Data source	Bos_Taurus.UMD3.1.85.gtf	Sus_scrofa.Sscrofa10.2.85.gtf	Gallus_gallus.Galgal4.85.gtf	Homo_sapiens.GRCh38.85.gtf

Genome Research 23:169–180 © 2013,

Method

High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression

Philippe Batut,[1,3] Alexander Dobin,[1] Charles Plessy,[2] Piero Carninci,[2] and Thomas R. Gingeras[1]



RAMPAGE

5’ complete sequencing

Objectives

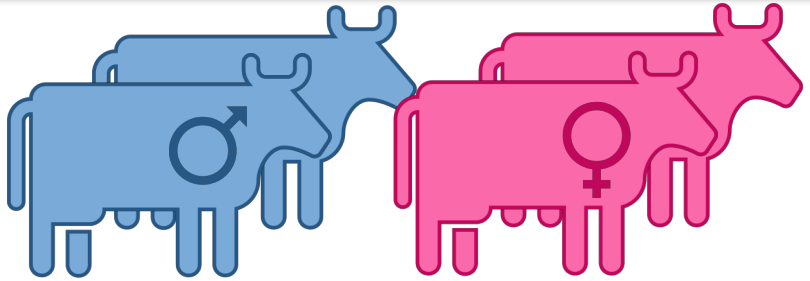
- Identify transcription start sites (TSS) genome-wide
  - Tissue-specific TSS, ‘Novel’ TSS, Alternative TSS
- Reconstruct partial transcript models for TSS of interest
  - Genes associated with multiple TSS
- Quantify transcript expression using 5’ signal at TSS
  - Compare with RNA-seq expression data
  - Obtain TSS-specific expression values

Genome Research 23:169–180 © 2013,

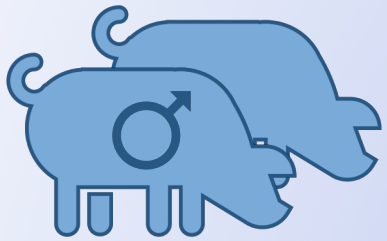
Method

High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression

Philippe Batut,[1,3] Alexander Dobin,[1] Charles Plessy,[2] Piero Carninci,[2] and Thomas R. Gingeras[1]



32 Tissues



16 Tissues

RAMPAGE

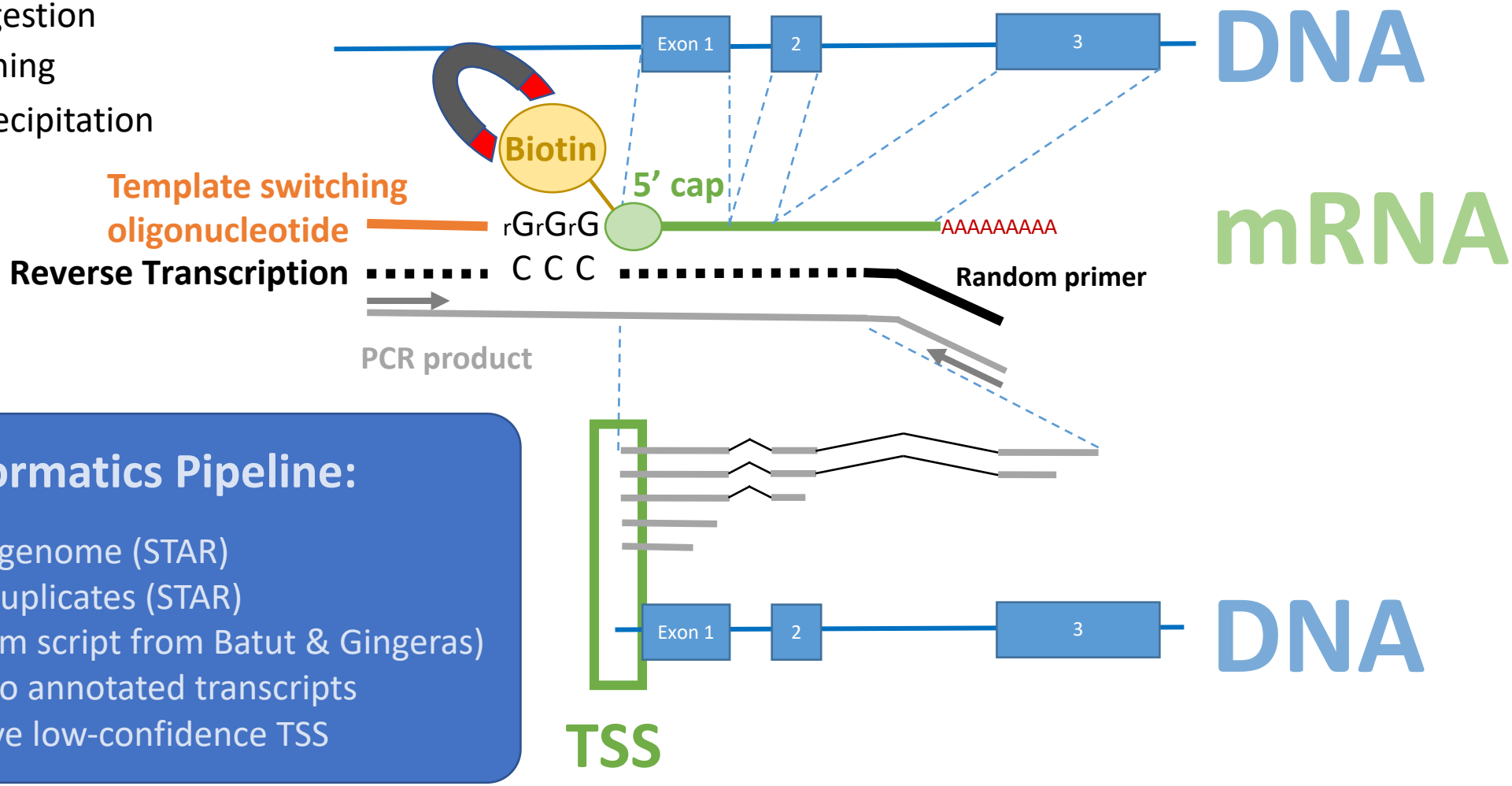
5’ complete sequencing

Objectives

- Identify transcription start sites (TSS) genome-wide
  - Tissue-specific TSS, ‘Novel’ TSS, Alternative TSS
- Reconstruct partial transcript models for TSS of interest
  - Genes associated with multiple TSS
- Quantify transcript expression using 5’ signal at TSS
  - Compare with RNA-seq expression data
  - Obtain TSS-specific expression values

RNA Annotation and Mapping of Promoters for Analysis of Gene Expression

- Exonuclease digestion
- Template switching
- CAP Immunoprecipitation



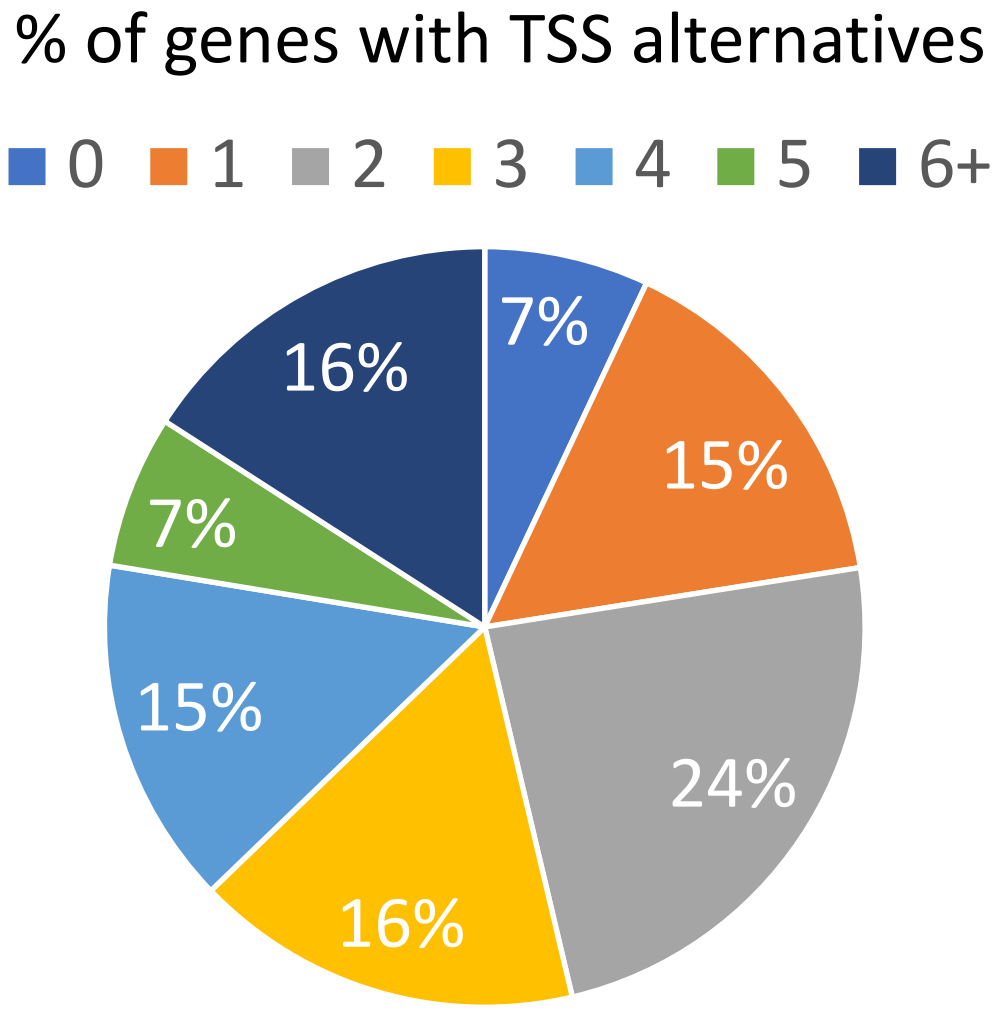
Bioinformatics Pipeline:

1. Align reads to genome (STAR)
2. Remove PCR duplicates (STAR)
3. Call TSS (custom script from Batut & Gingeras)
4. Attribute TSS to annotated transcripts
5. Filter to remove low-confidence TSS

Sequencing data

	Raw reads	Uniquely Mapped Reads (%)	Deduplicated reads for TSS calling
32 Cattle tissues (Average ± SEM)	1,823,954,572 (13,923,317 ± 1,118,716)	1,026,704,029 (56.29 ± 1.4%)	336,439,806 (3,266,406 ± 242,341)
16 Pig tissues (Average ± SEM)	60,505,377 (2,881,208 ± 367,019)	42,196,450 (69.74 ± 3.4%)	17,261,511 (756,894 ± 107,805)

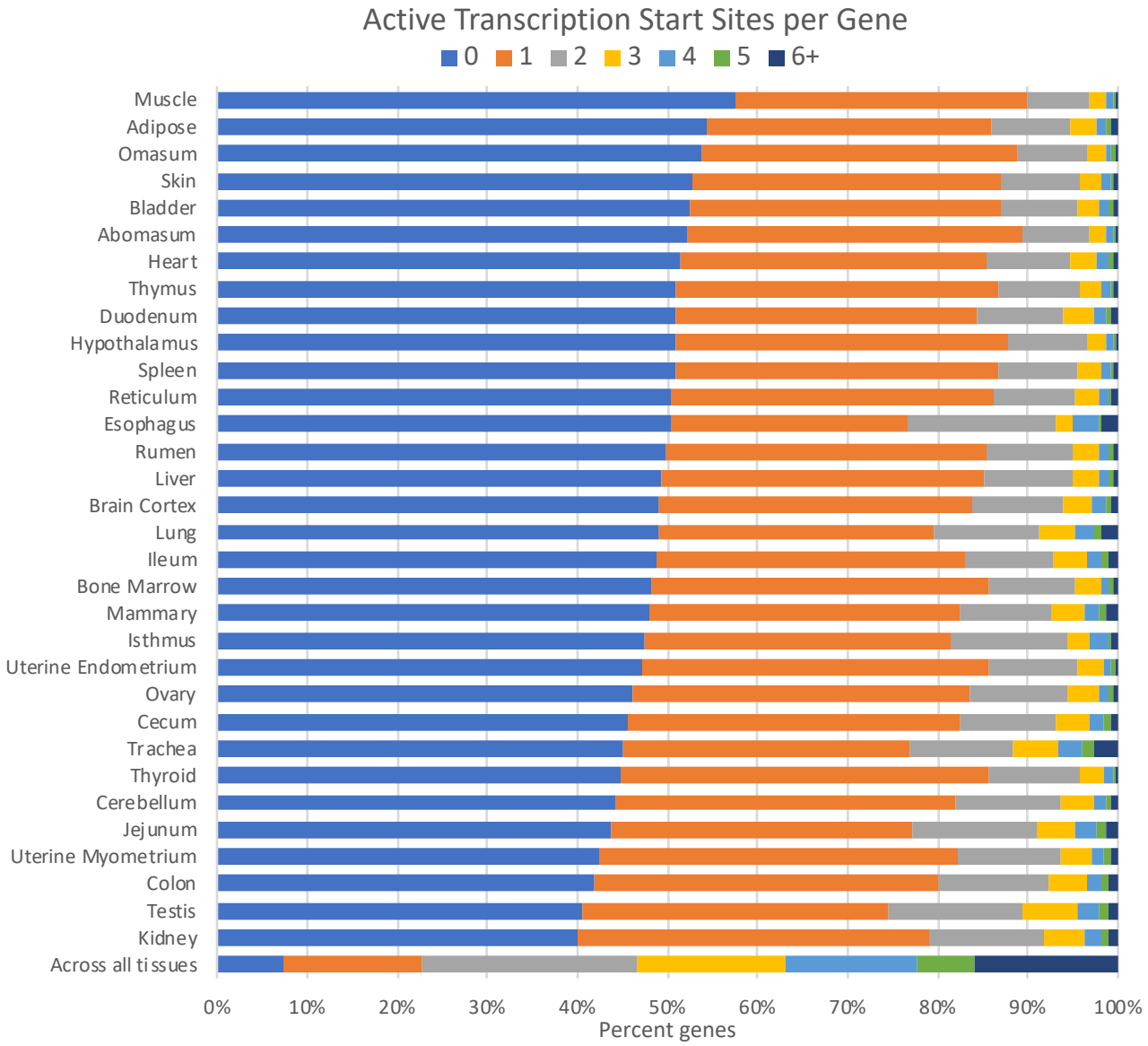
Transcription start site detected



Overall TSS detected		
Conserved between biological replicates	66,411	
Conserved TSS that are ‘novel’ (>50 bp from annotated 5’ end)	55,023	(83%)
Overall genes associated with TSS		
Detected in both biological replicates	15,960	(80%)
All	18,504	(93%)

Total genes detected	**TSS**	**Ratio**
18,504	68,039	**3.68**

Transcription start site usage across tissues



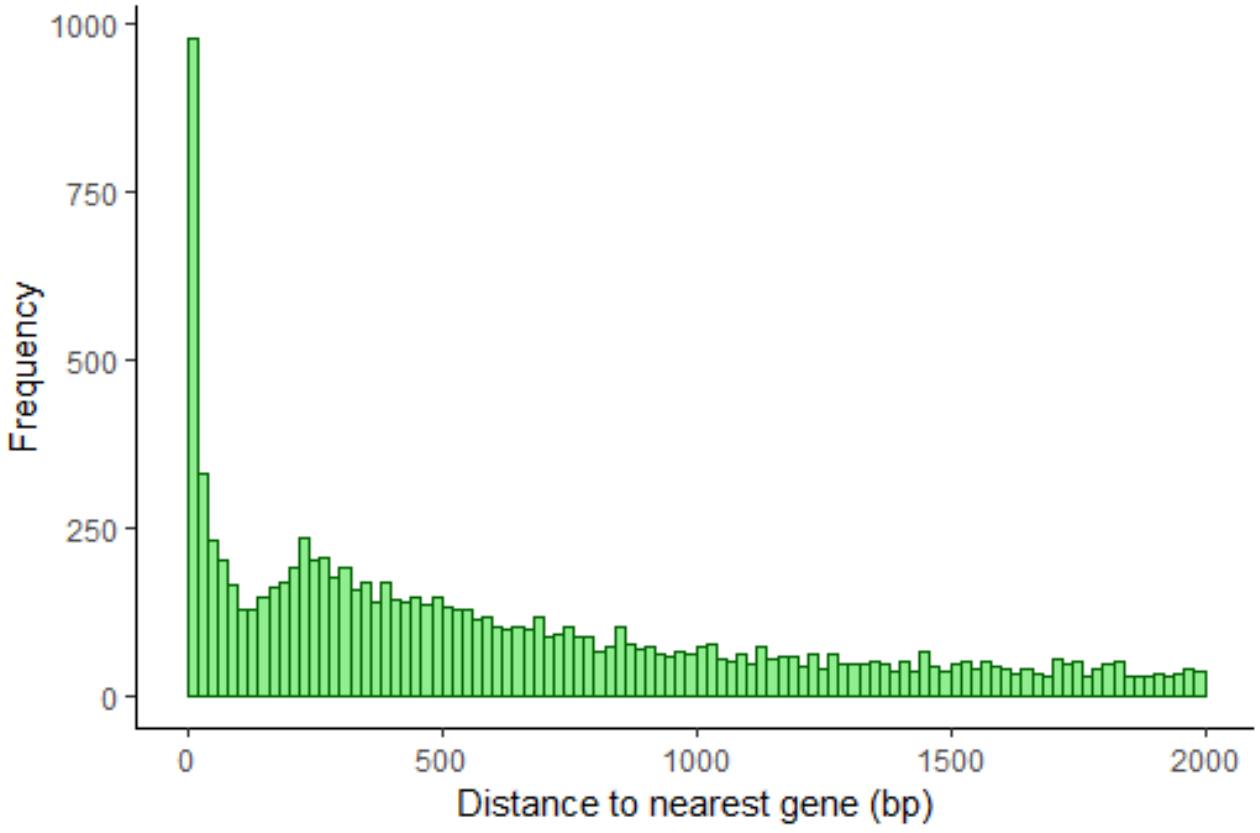
Overall TSS detected		
Conserved between biological replicates	66,411	
Conserved TSS that are ‘novel’ (>50 bp from annotated 5’ end)	55,023	(83%)

Overall genes associated with TSS		
Detected in both biological replicates	15,960	(80%)
All	18,504	(93%)

Total genes detected	**TSS**	**Ratio**
18,504	68,039	**3.68**

Transcription start site usage across tissues

TSS localization relative to currently annotated 5’ transcript ends

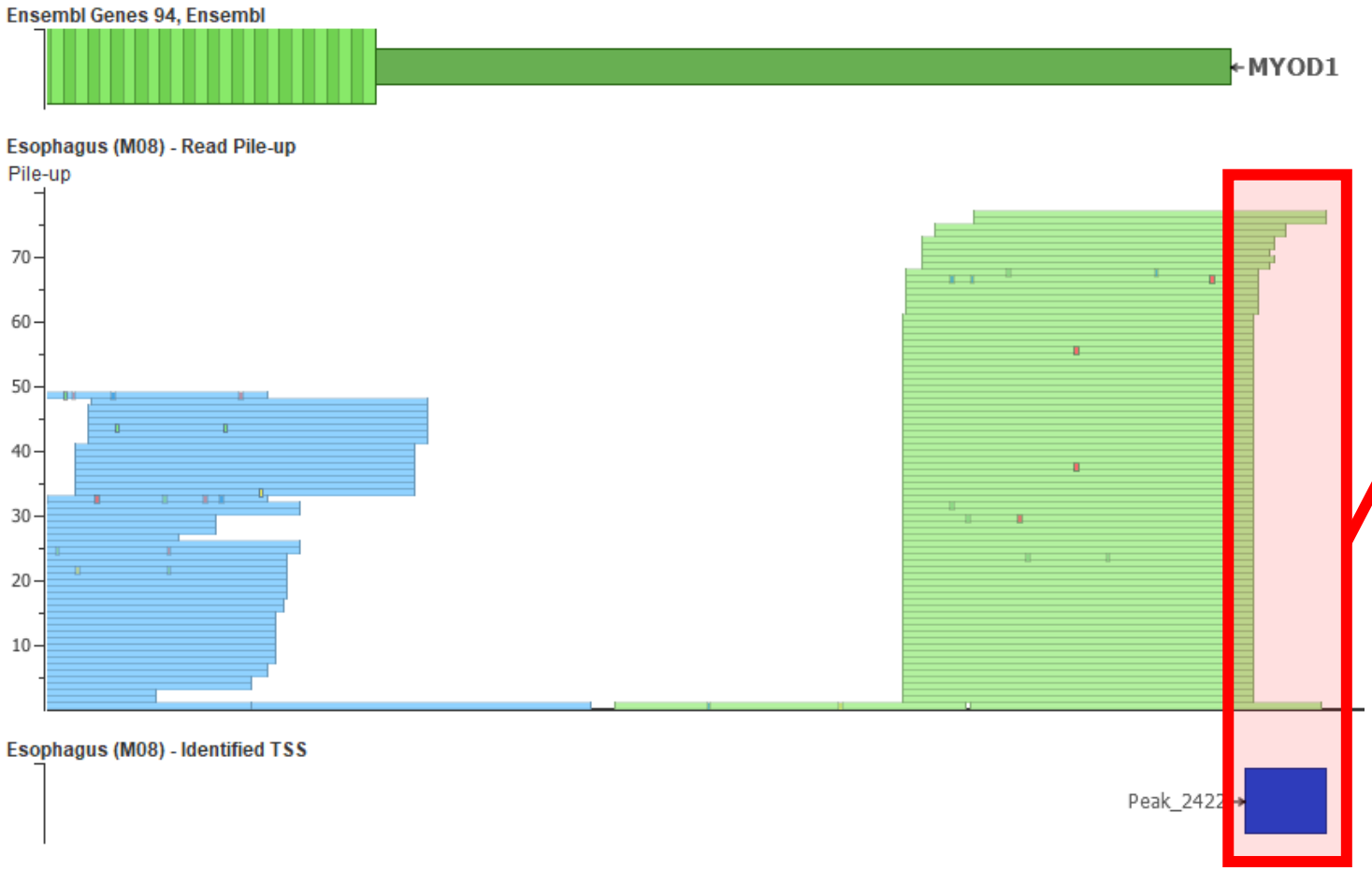


Overall TSS detected		
Conserved between biological replicates	66,411	
Conserved TSS that are ‘novel’ (>50 bp from annotated 5’ end)	55,023	(83%)

Overall genes associated with TSS		
Detected in both biological replicates	15,960	(80%)
All	18,504	(93%)

Total genes detected	**TSS**	**Ratio**
18,504	68,039	**3.68**

Calculating expression from 5’ signal at TSS



Count reads whose 5’ ends fall within a TSS of a given gene

↓

Raw expression values per gene

↓

Normalize using DESeq2

↓

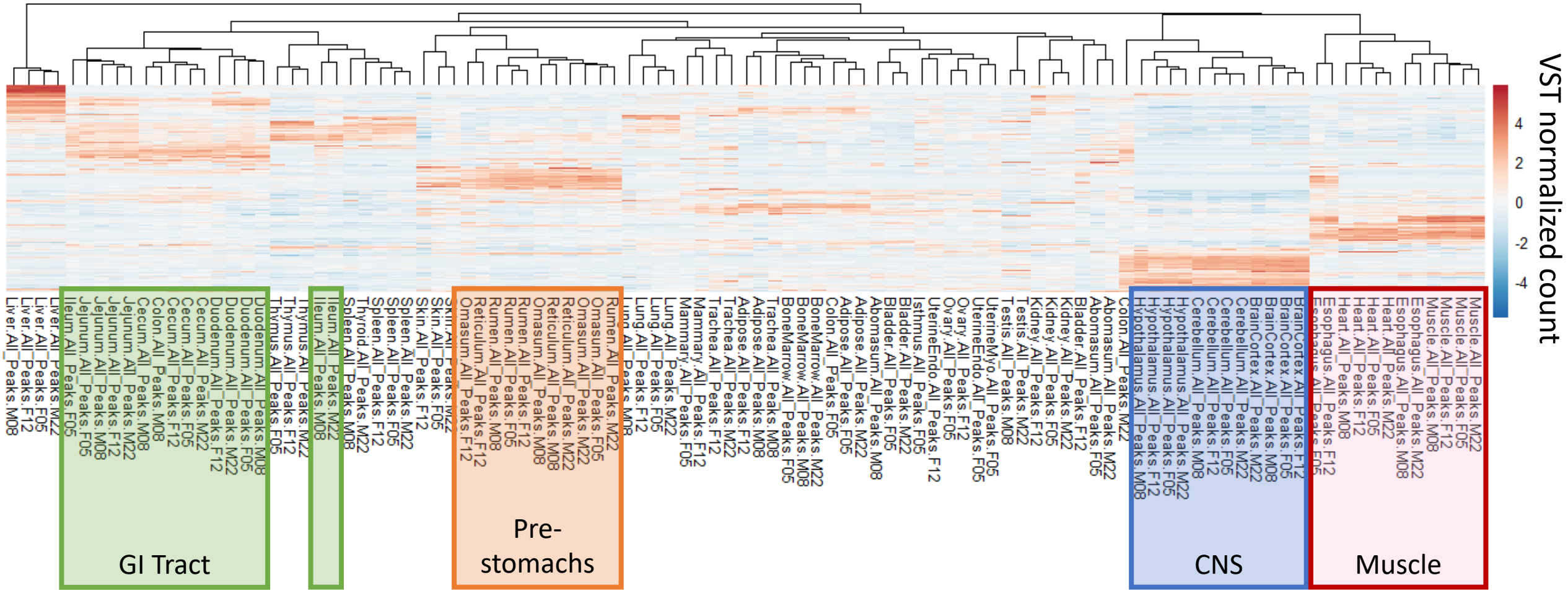
Variance-stabilizing transformation

↓

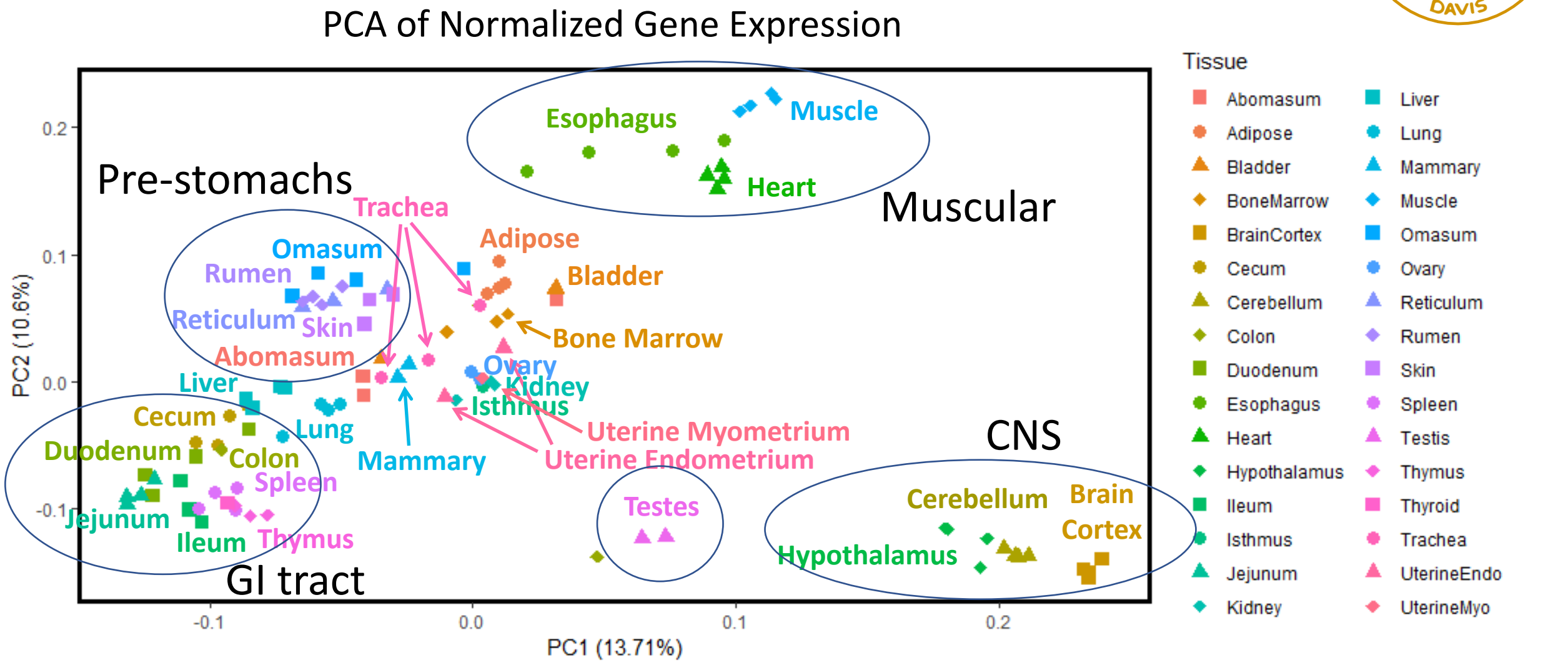
Normalized expression counts per gene

Normalized expression from 5’ signal at TSS

Top 1000 genes by variance

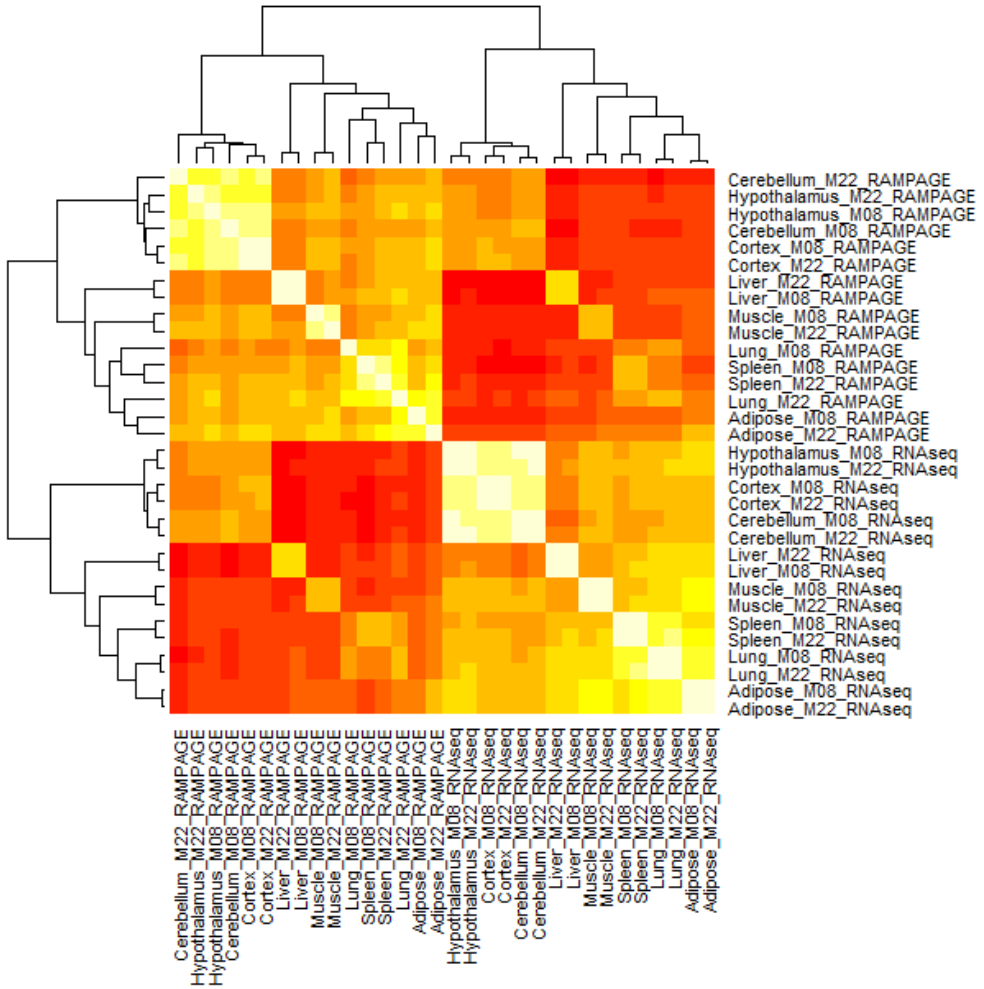


PCA of normalized gene counts from 5’ signal



Gene expression is correlated across assays

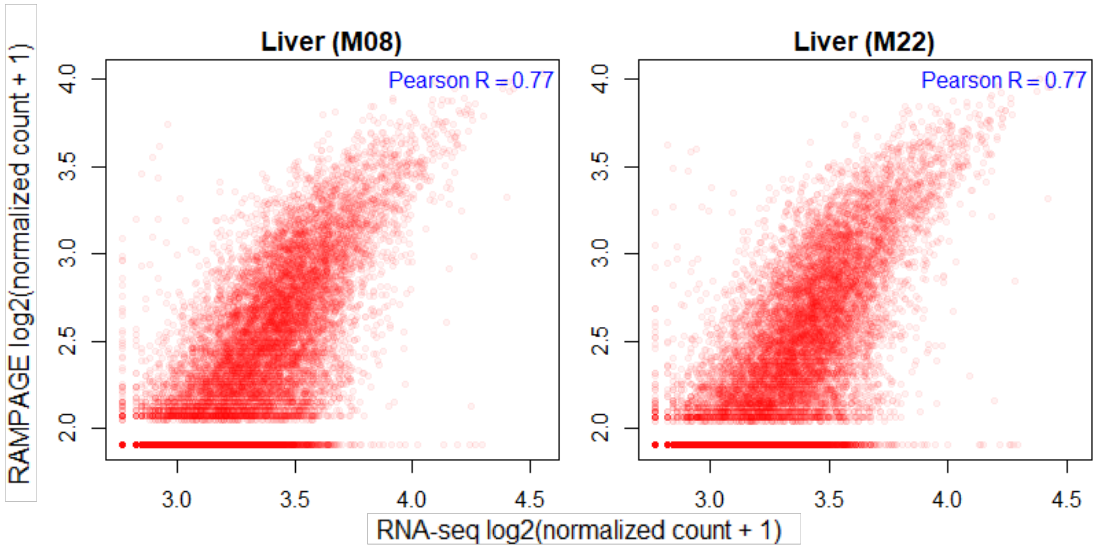
Normalized gene expression:
RAMPAGE versus RNA-seq (Pearson R)

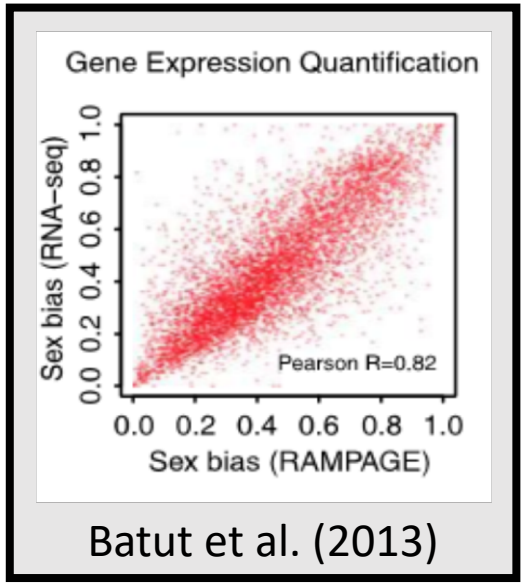


Pearson Correlation Coefficient

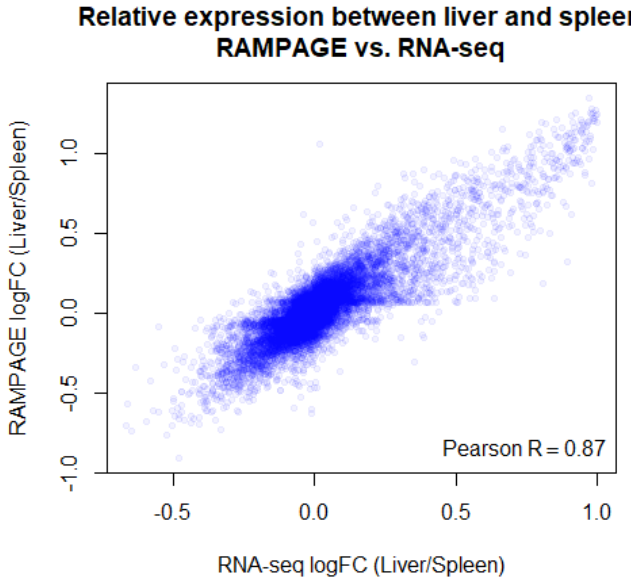
	RAMPAGE v RNA-seq		Between biological replicates	
	M08	*M22*	*RAMPAGE*	*RNA-seq*
Adipose	0.747	0.625	0.806	0.989
Cerebellum	0.713	0.701	0.933	0.989
Cortex	0.684	0.685	0.951	0.973
Hypothalamus	0.627	0.673	0.882	0.975
Liver	0.756	0.766	0.954	0.986
Lung	0.738	0.636	0.823	0.987
Spleen	0.707	0.685	0.896	0.979
Muscle	0.721	0.728	0.899	0.983
Average	0.700		0.893	0.983

Accurate quantification of relative expression





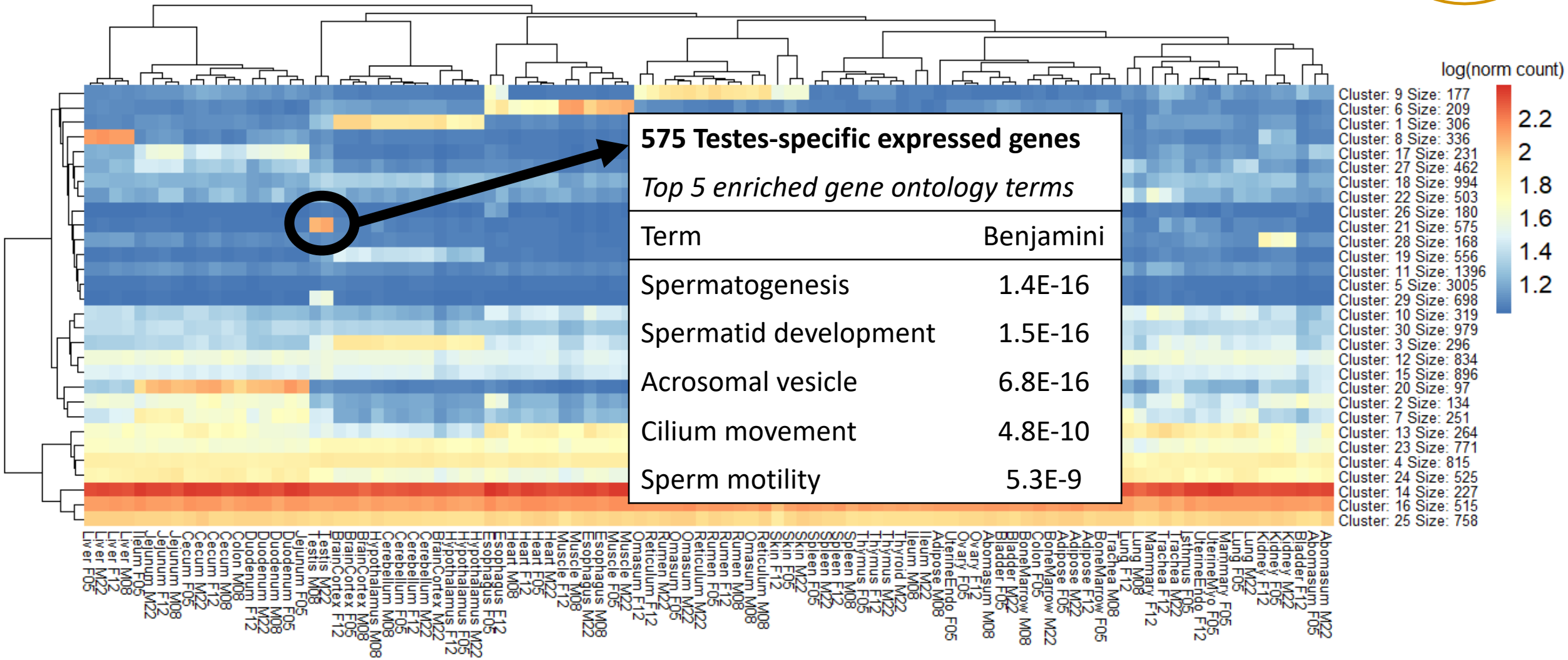
Batut et al. (2013)



Pearson Correlation Coefficient

	RAMPAGE v RNA-seq		Between biological replicates	
	M08	*M22*	*RAMPAGE*	*RNA-seq*
Adipose	0.747	0.625	0.806	0.989
Cerebellum	0.713	0.701	0.933	0.989
Cortex	0.684	0.685	0.951	0.973
Hypothalamus	0.627	0.673	0.882	0.975
Liver	0.756	0.766	0.954	0.986
Lung	0.738	0.636	0.823	0.987
Spleen	0.707	0.685	0.896	0.979
Muscle	0.721	0.728	0.899	0.983
Average	0.700		0.893	0.983

K-means clustering by 5’ signal at TSS



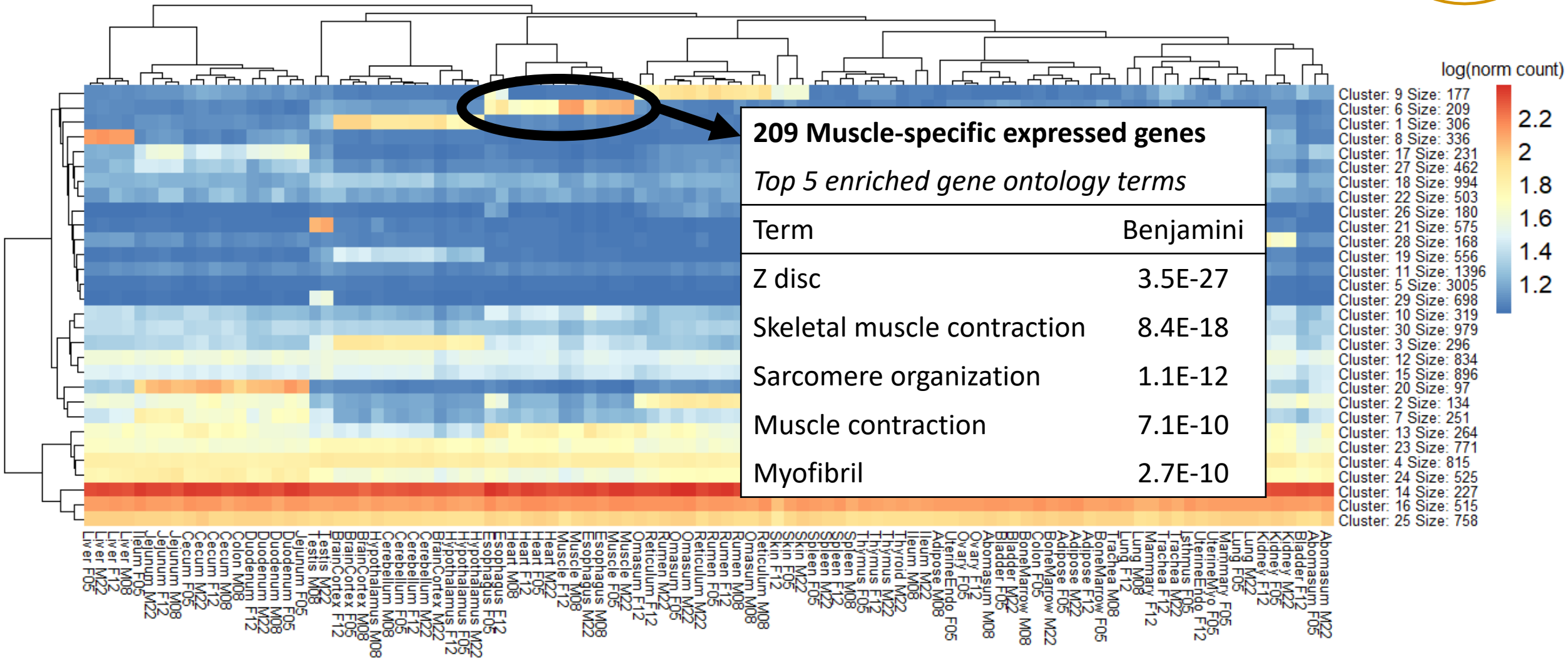
575 Testes-specific expressed genes

Top 5 enriched gene ontology terms

Term	Benjamini
Spermatogenesis	1.4E-16
Spermatid development	1.5E-16
Acrosomal vesicle	6.8E-16
Cilium movement	4.8E-10
Sperm motility	5.3E-9

MYCBP Associated Protein
(Testis Secretory Sperm-Binding Protein Li 214e)

K-means clustering by 5’ signal at TSS



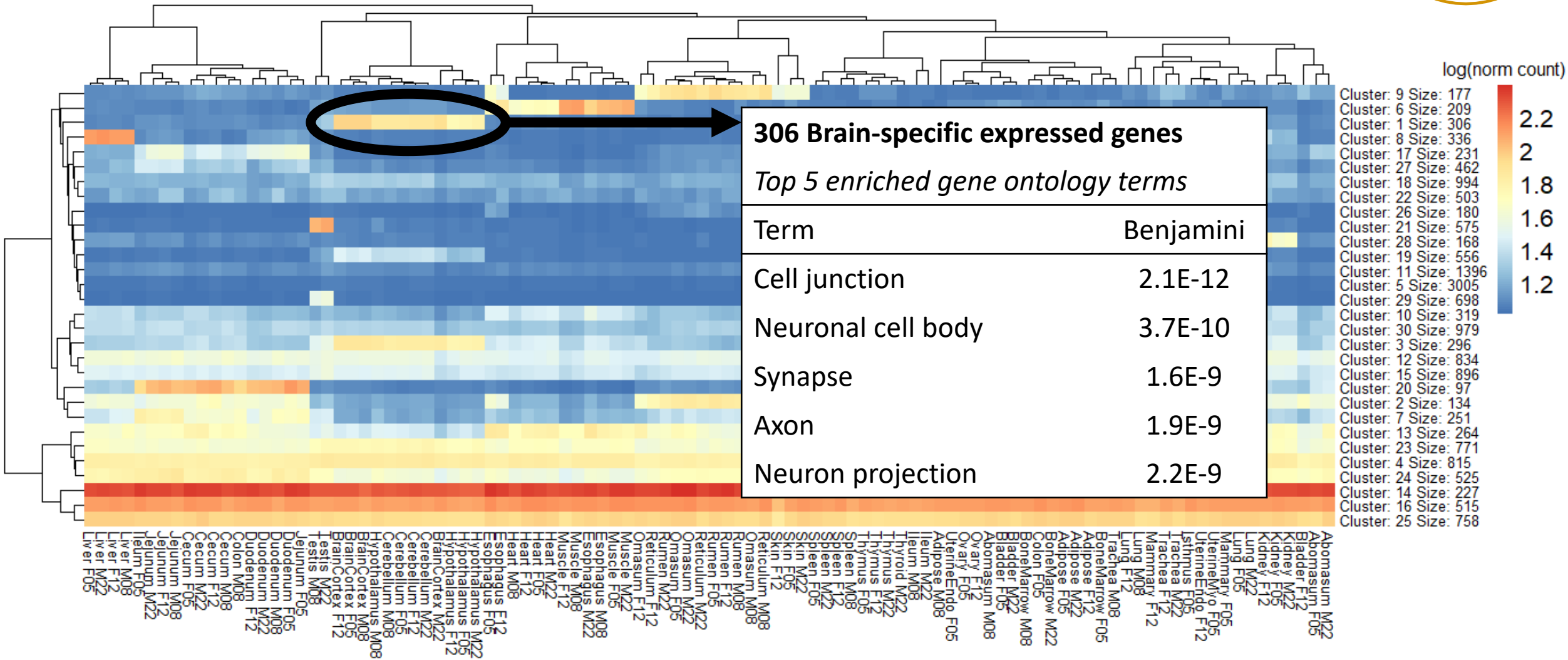
209 Muscle-specific expressed genes

Top 5 enriched gene ontology terms

Term	Benjamini
Z disc	3.5E-27
Skeletal muscle contraction	8.4E-18
Sarcomere organization	1.1E-12
Muscle contraction	7.1E-10
Myofibril	2.7E-10

Myopalladin

K-means clustering by 5’ signal at TSS

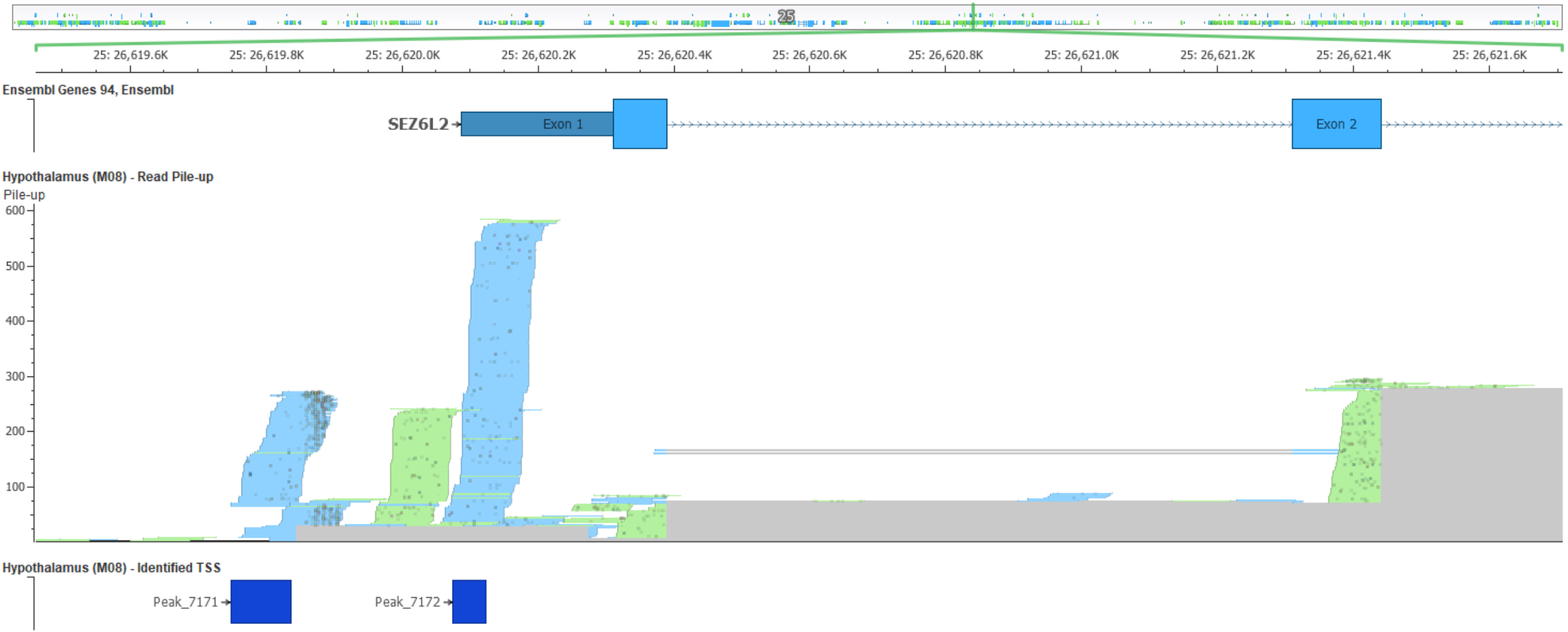


306 Brain-specific expressed genes

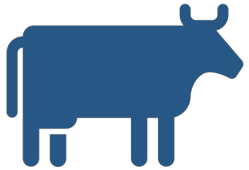
Top 5 enriched gene ontology terms

Term	Benjamini
Cell junction	2.1E-12
Neuronal cell body	3.7E-10
Synapse	1.6E-9
Axon	1.9E-9
Neuron projection	2.2E-9

Seizure Related 6 Homolog Like 2



Improving the Bovine Annotation



	Cattle (old)	Cattle (new)	Human
Protein-coding genes	19,994	18,504	19,814
Protein-coding transcripts	22,118	68,039	79,712
Transcripts per gene	1.1	3.7	4.0
Data source	Bos_Taurus. UMD3.1.85.gtf	UC Davis RAMPAGE	Homo_sapiens. GRCh38.85.gtf

- Using 32 tissues, TSS were identified for 93% of bovine genes
- Over 80% of these TSS are novel annotations

Next:

- Currently analyzing RAMPAGE sequencing data for 16 pig tissues
- In the process of generating chicken tissue RAMPAGE data

- Integrate RAMPAGE TSS Data with Iso-seq sequencing
- Use TSS information to inform ChIP-seq data for defining promoter states in different tissues

Acknowledgements

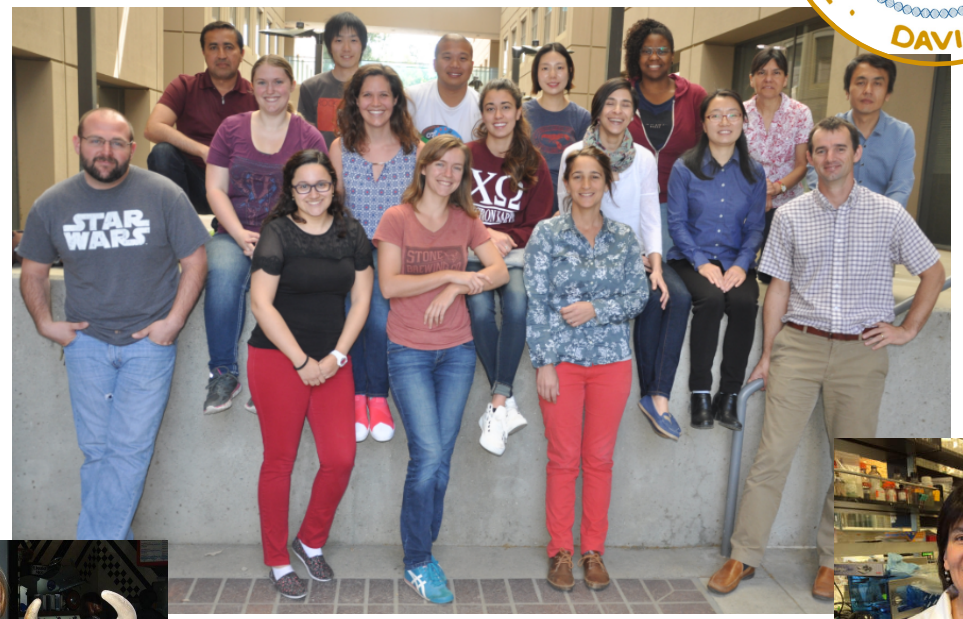
Ross Laboratory
- Michelle Halstead
- Daniel Goszczynski
- Alma Islas-Trejo
- Shally Xu
- Susan Waters
- Marcela Vilarino
- Delia Soto
- Insung Park
- Ahmed Mahdi

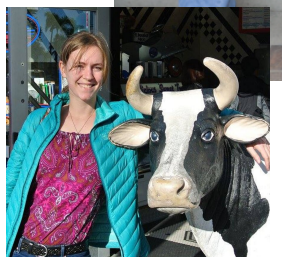
Zhou Laboratory
- Dr. Huaijun Zhou
- Dr. Colin Kern
- Perot Saelao
- Ying Wang
- Kelly Chanthavixay

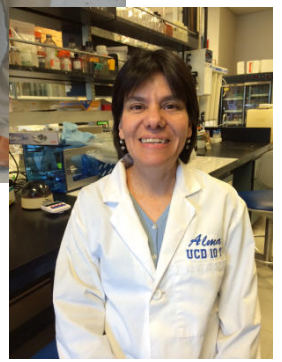
UCDAVIS ANIMAL SCIENCE

Collaborators
- Ian Korf, UCD
- Juan Medrano, UCD
- Chris Tuggle , ISU
- Cathy Ernst, MSU
- Stephany McKay, UV
- Monique Rijnkels, TAMU
- Clare Gill, TAMU
- Brenda Murdoch, UI
- Tim Smith, USDA-ARS
- Zhihua Jiang, WSU
- James Reecy, ISU
- Wansheng Liu, PSU
- Honglin Jiang, VT
- Milton Thomas, CSU
- Angela Canovas, UG
- Graham Plastow, UA

FAANG Functional Annotation of Animal Genomes







USDA NIFA

Funding
- USDA-NIFA-AFRI 2015-67015-22940
- USDA-NIFA-AFRI 2017-67015-26297
- USDA-NIFA-AFRI 2018-67015-27500