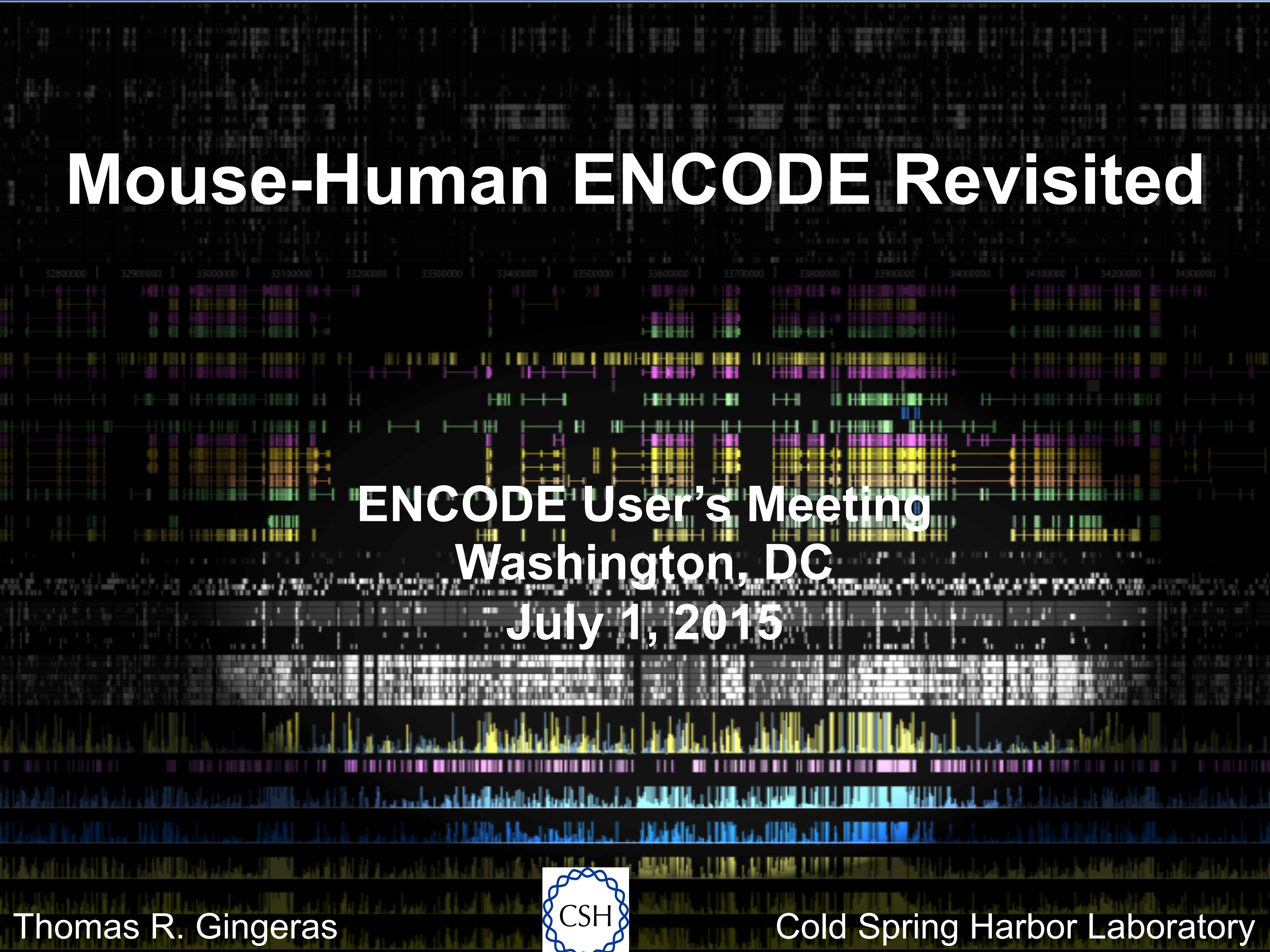
Mouse-Human ENCODE Revisited

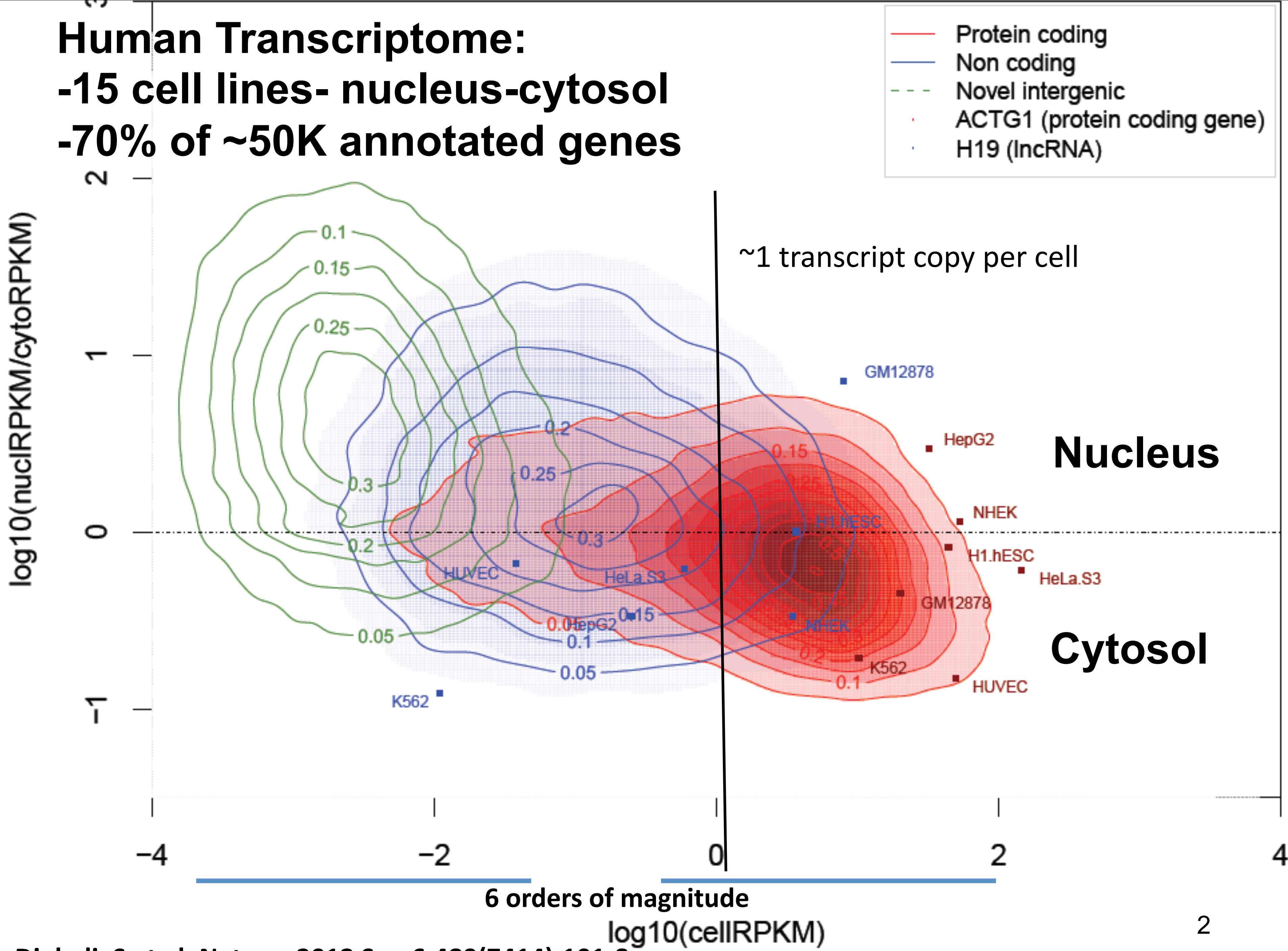


ENCODE User's Meeting
Washington, DC
July 1, 2015

Thomas R. Gingeras

Cold Spring Harbor Laboratory

Human Transcriptome:
-15 cell lines- nucleus-cytosol
-70% of ~50K annotated genes



~1 transcript copy per cell

Nucleus

Cytosol

6 orders of magnitude

Djebali, S et al. Nature. 2012 Sep 6;489(7414):101-8

Mouse vs. Human

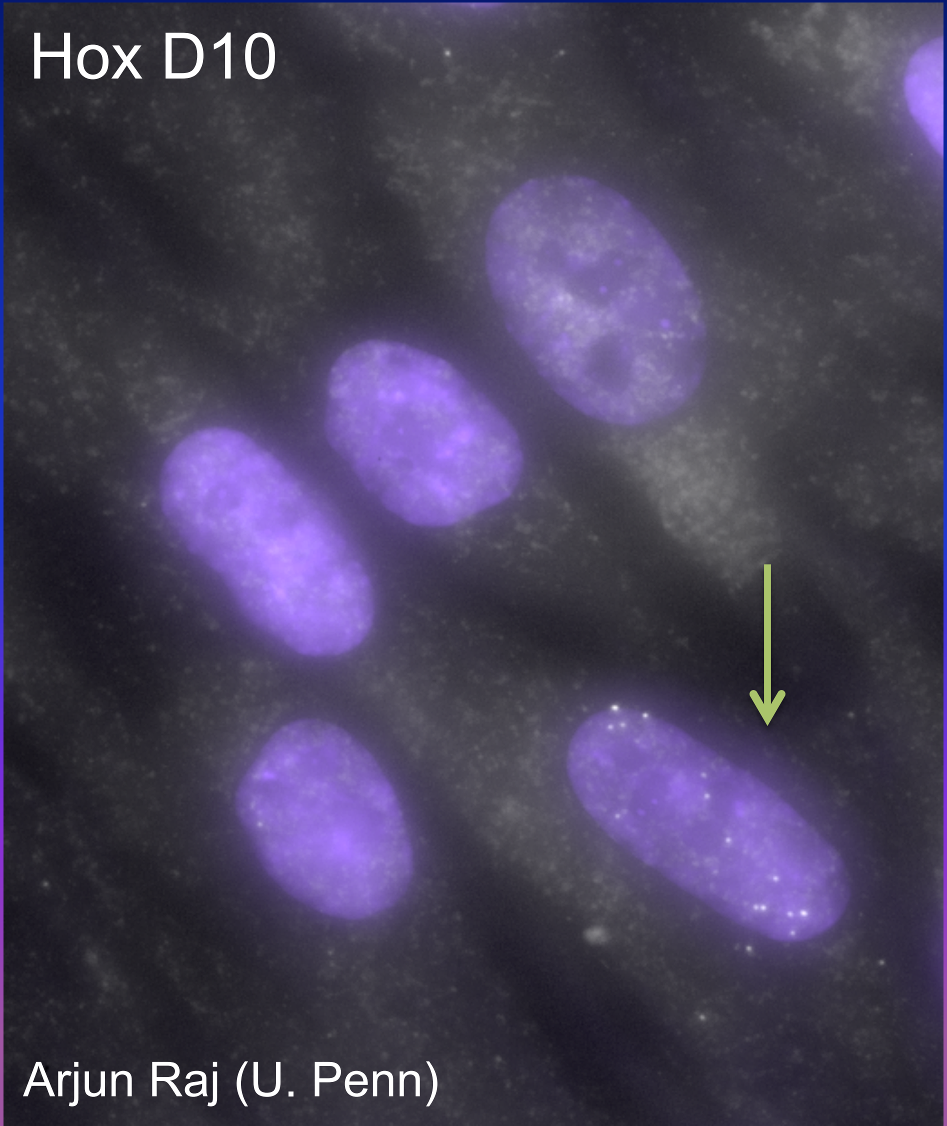
Study Design

- 18 human cell lines (ENCODE) vs. 25 mouse tissue samples in 5 developmental stages
- Two bio-replicates per sample
- Only data passing IDR at 90%> reproducibility (5 read min)
- Poly A+ from total RNA extracted from each sample used to make Illumina libraries consisting of PE 100mers (400 million reads/replica)
- **"Conservation"** is not used in this study in an evolutionary sense (i.e., it does not mean that the similarity of any feature shared by the compared genes found in the two species has been maintained by purifying selection)

Key Points to Remember

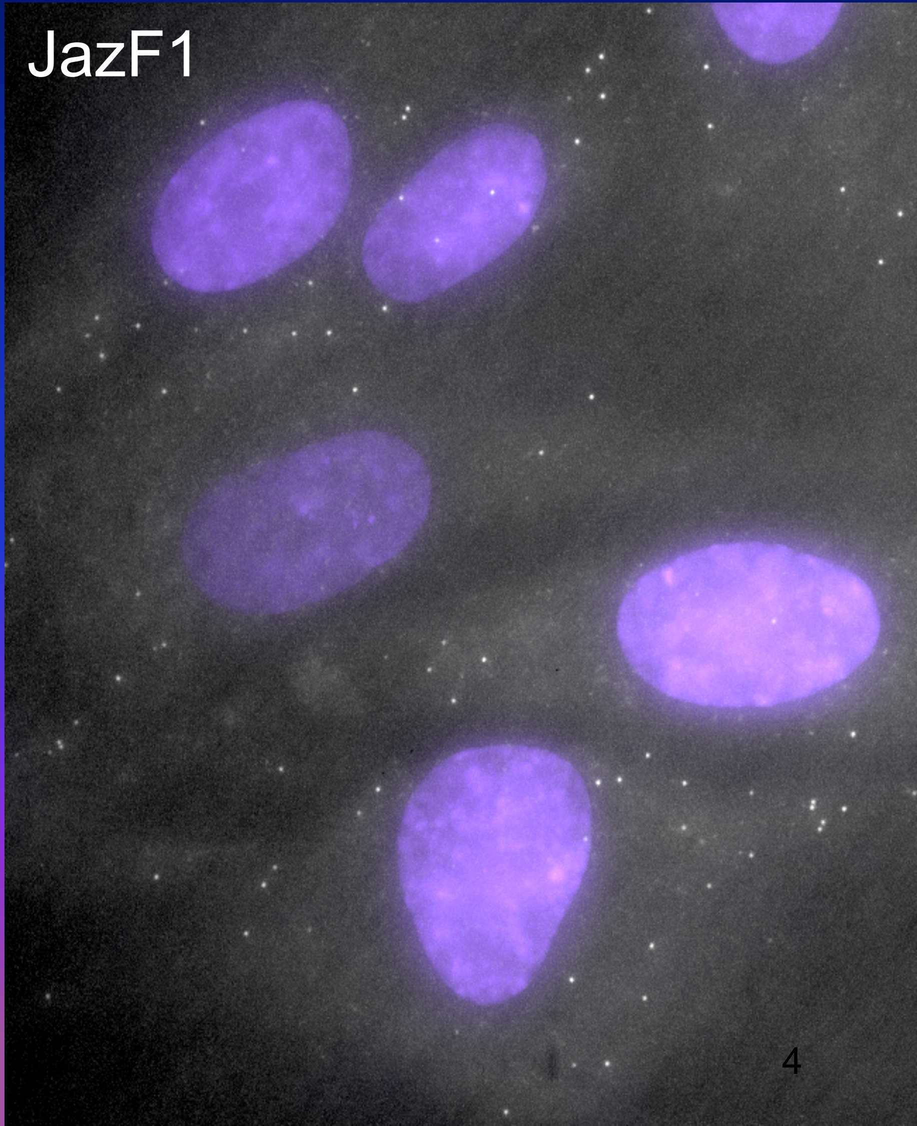
1. The difference in sample types and species underscores the significance of any similarities
2. Conserved features highlighted are not dependent upon common **sequences**

Distribution of RNAs Within Individual Human Foreskin Fibroblasts



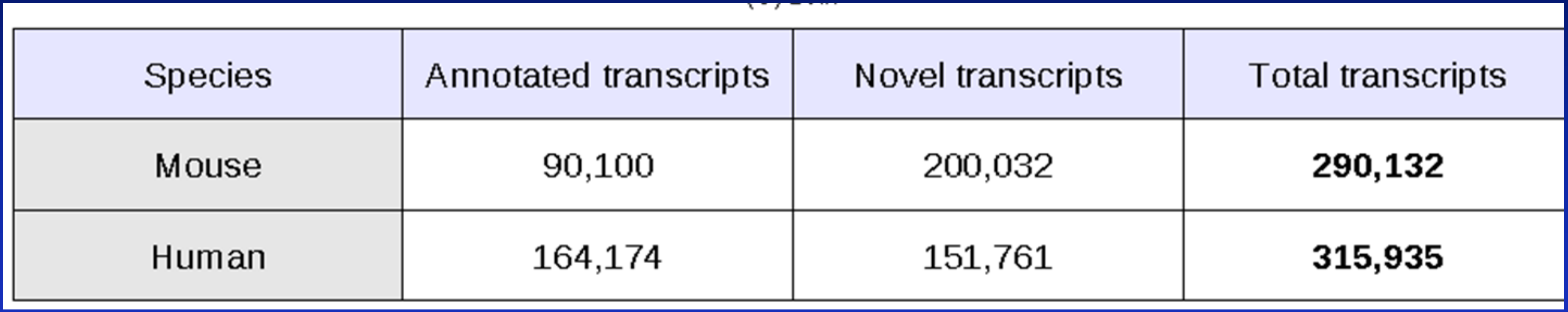
Hox D10

Arjun Raj (U. Penn)

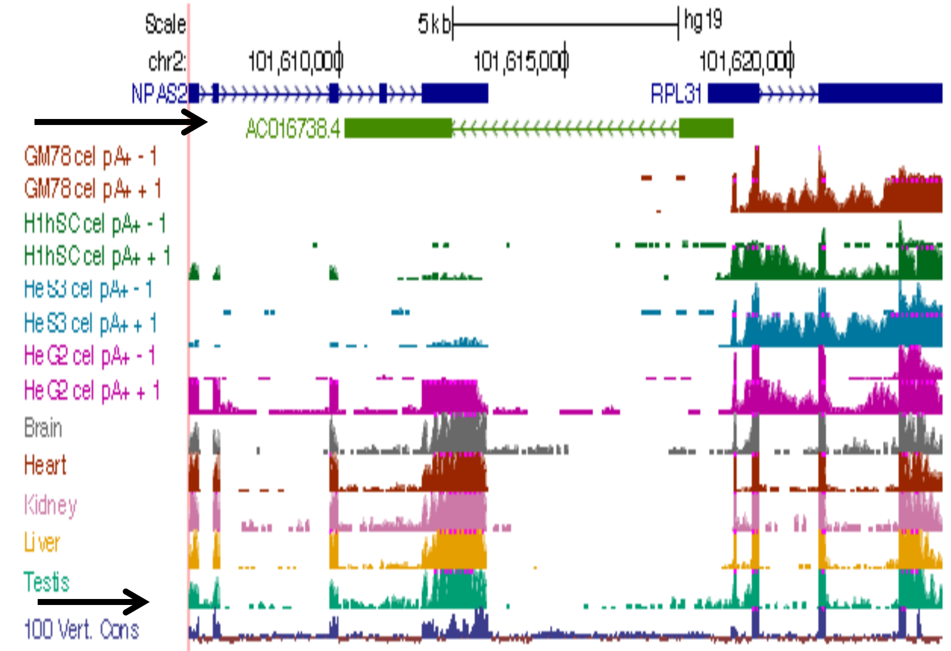


JazF1

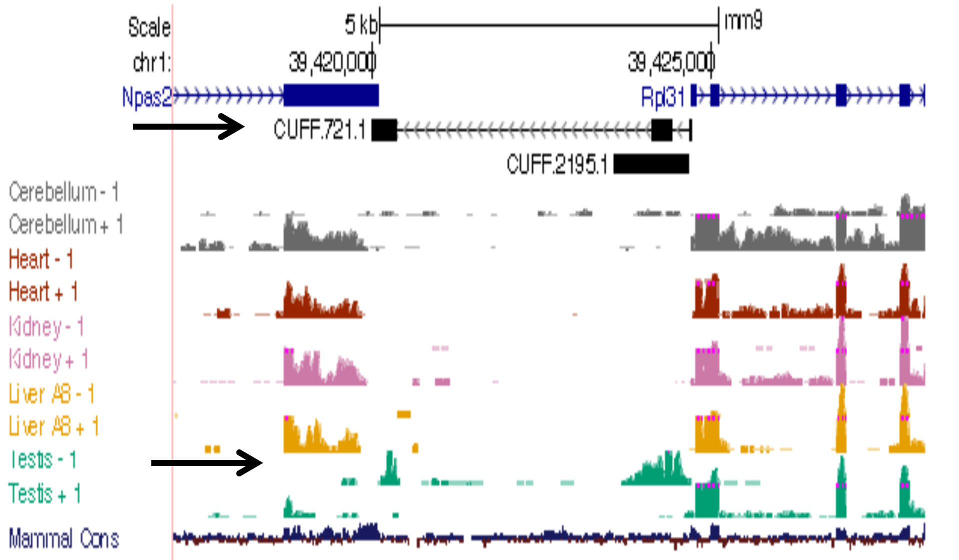
Completing the Mouse Genome Annotation



Species	Annotated transcripts	Novel transcripts	Total transcripts
Mouse	90,100	200,032	**290,132**
Human	164,174	151,761	**315,935**



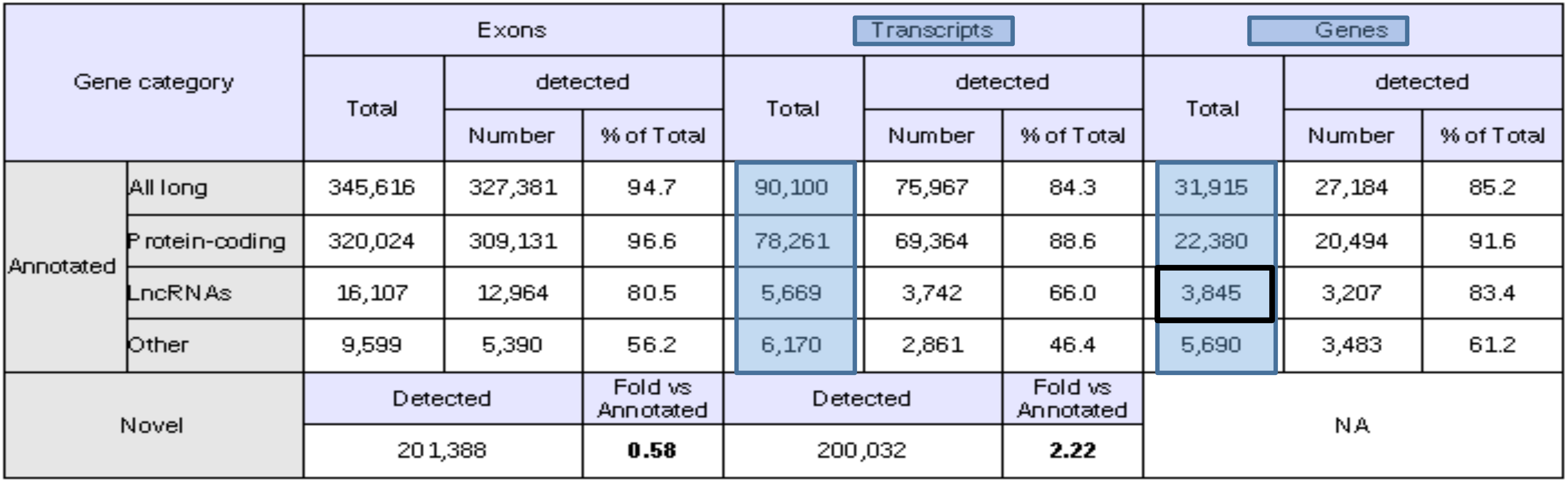
HUMAN



MOUSE

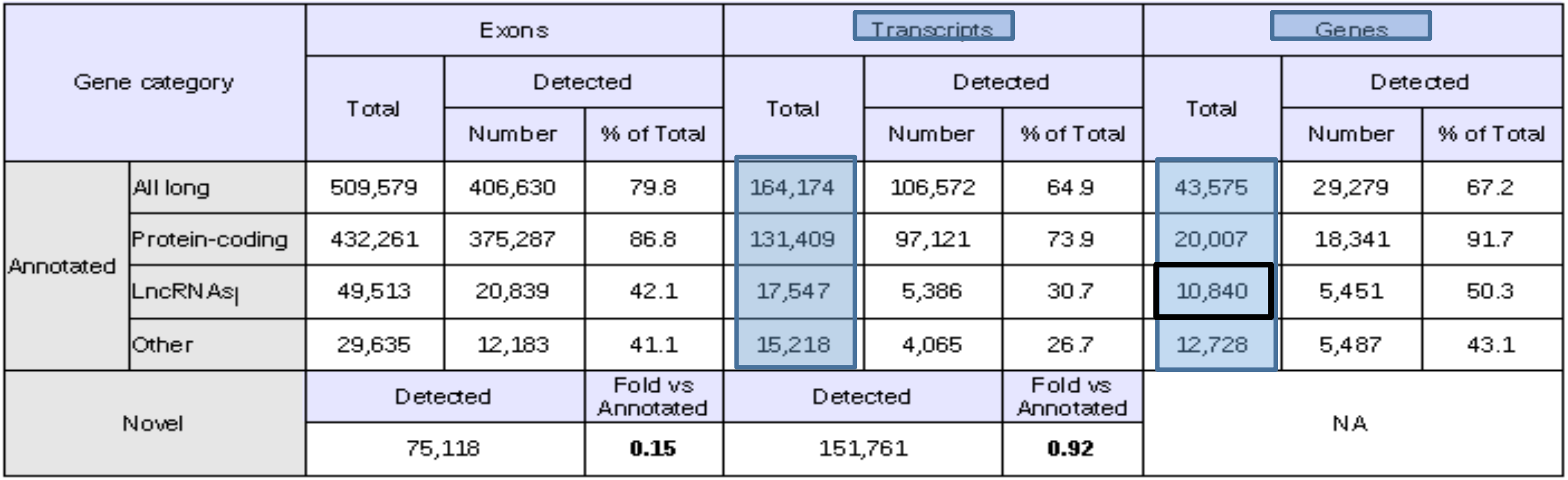
Supplementing Mouse Genome Annotation

(A) Mouse



Gene category		Exons			Transcripts			Genes		
		Total	detected		Total	detected		Total	detected	
			Number	% of Total		Number	% of Total		Number	% of Total
Annotated	All long	345,616	327,381	94.7	90,100	75,967	84.3	31,915	27,184	85.2
	Protein-coding	320,024	309,131	96.6	78,261	69,364	88.6	22,380	20,494	91.6
	LncRNAs	16,107	12,964	80.5	5,669	3,742	66.0	3,845	3,207	83.4
	Other	9,599	5,390	56.2	6,170	2,861	46.4	5,690	3,483	61.2
Novel		Detected		Fold vs Annotated	Detected		Fold vs Annotated	NA		
		201,388		**0.58**	200,032		**2.22**			

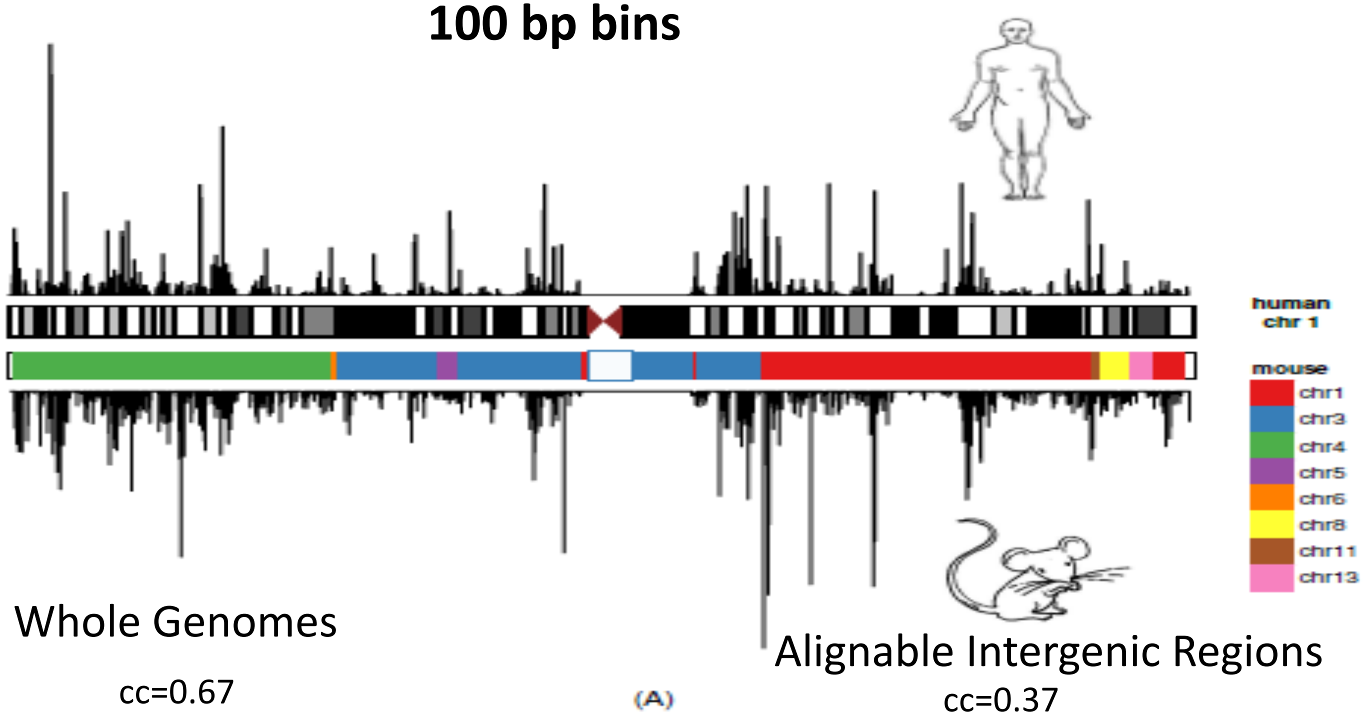
(B) Human



Gene category		Exons			Transcripts			Genes		
		Total	Detected		Total	Detected		Total	Detected	
			Number	% of Total		Number	% of Total		Number	% of Total
Annotated	All long	509,579	406,630	79.8	164,174	106,572	64.9	43,575	29,279	67.2
	Protein-coding	432,261	375,287	86.8	131,409	97,121	73.9	20,007	18,341	91.7
	LncRNAs	49,513	20,839	42.1	17,547	5,386	30.7	10,840	5,451	50.3
	Other	29,635	12,183	41.1	15,218	4,065	26.7	12,728	5,487	43.1
Novel		Detected		Fold vs Annotated	Detected		Fold vs Annotated	NA		
		75,118		**0.15**	151,761		**0.92**			

Correlation of Expression across the Mouse and Human Genomes

100 bp bins

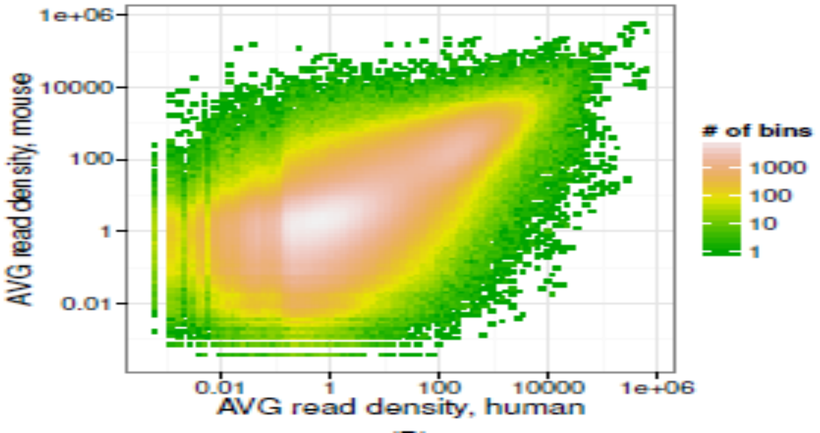


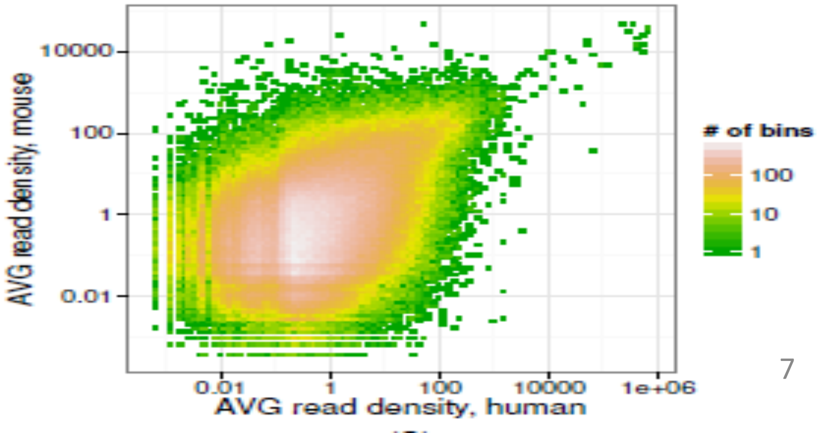
Whole Genomes

cc=0.67

Alignable Intergenic Regions

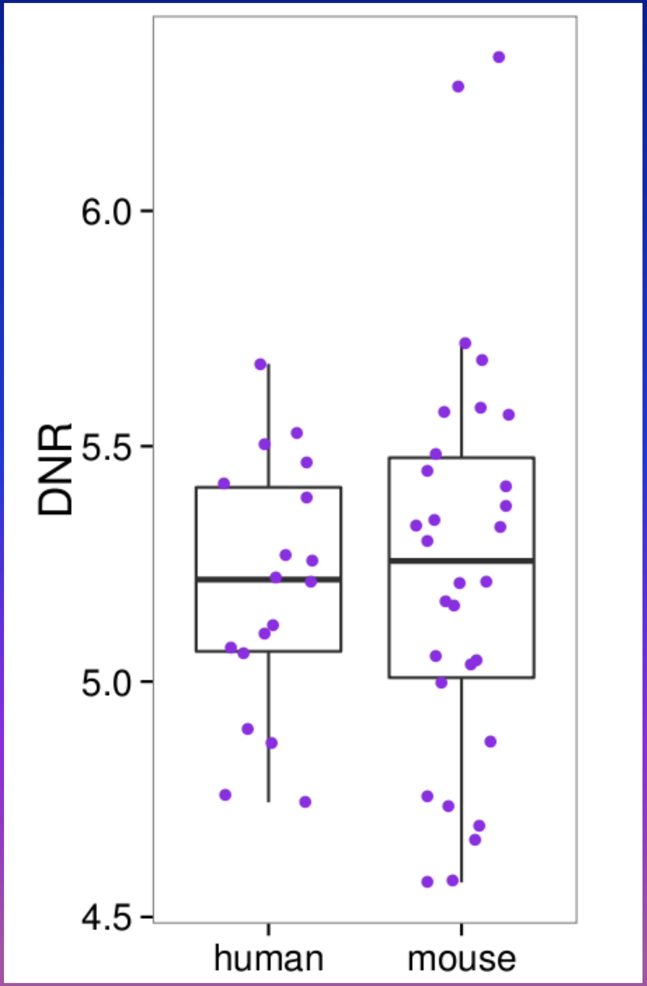
cc=0.37





Comparison of Dynamic Range of Expression Levels of Mouse and Human Orthologous Genes

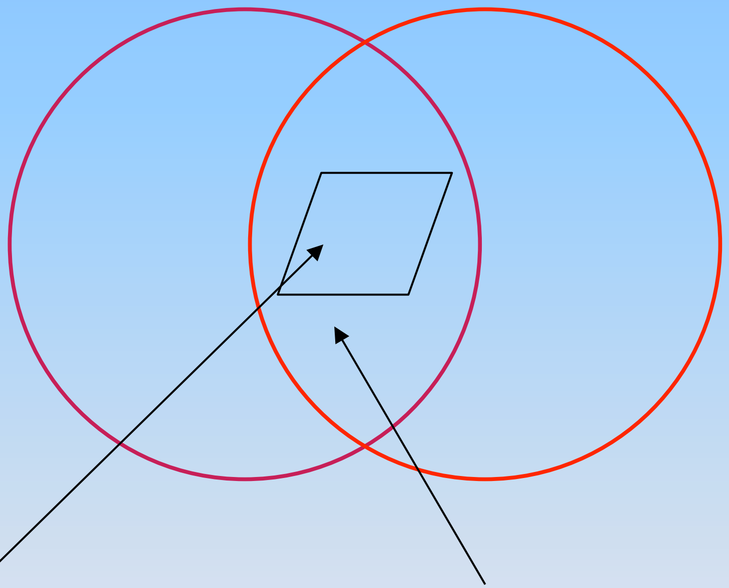
The dynamic range (DNR) of gene expression in a cell line or tissue sample can be up to 6 orders of magnitudes



Each dot is the DNR using all expressed orthologs found in each of the mouse and human samples

Number of Expressed Orthologous Protein Coding (PC) Genes in Multiple Species

Mouse
expressed
PC genes:
20,494



Human
expressed
PC genes:
18,341

1:1 matches

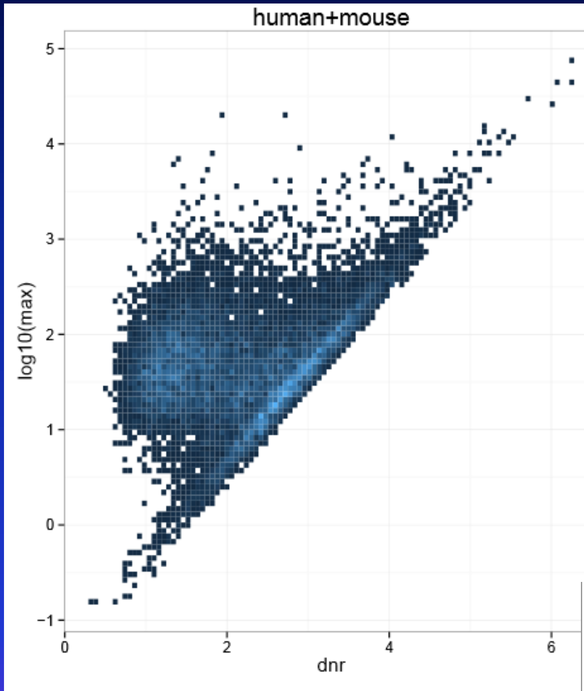
Present in all
6 species

Orthologs
In 6 species:
5,971

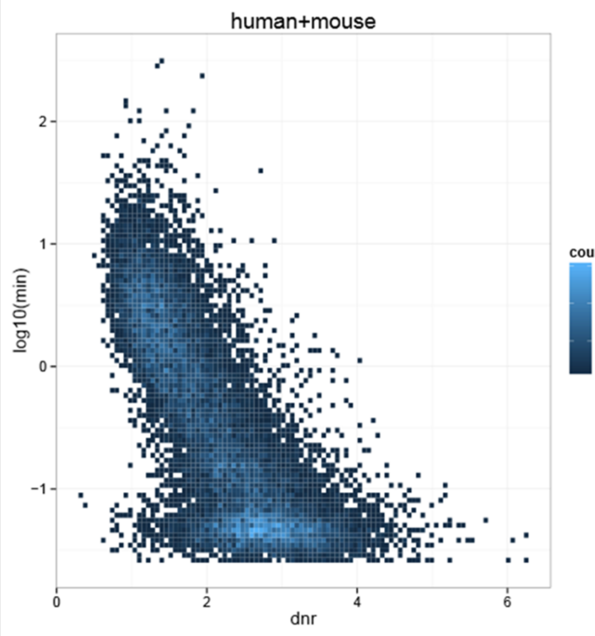
Expressed Orthologous Genes Expressed
In Human and Mouse: **14,984**

Species: human, mouse, macaque, rat , chicken, cow

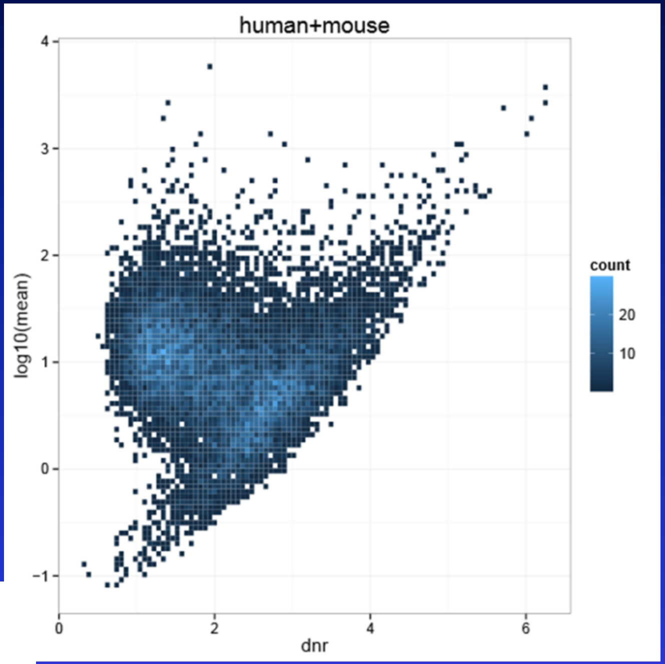
Correlation of Log_{10} Mean, Max and Min RPKM vs. Dynamic Range of Expression



Max

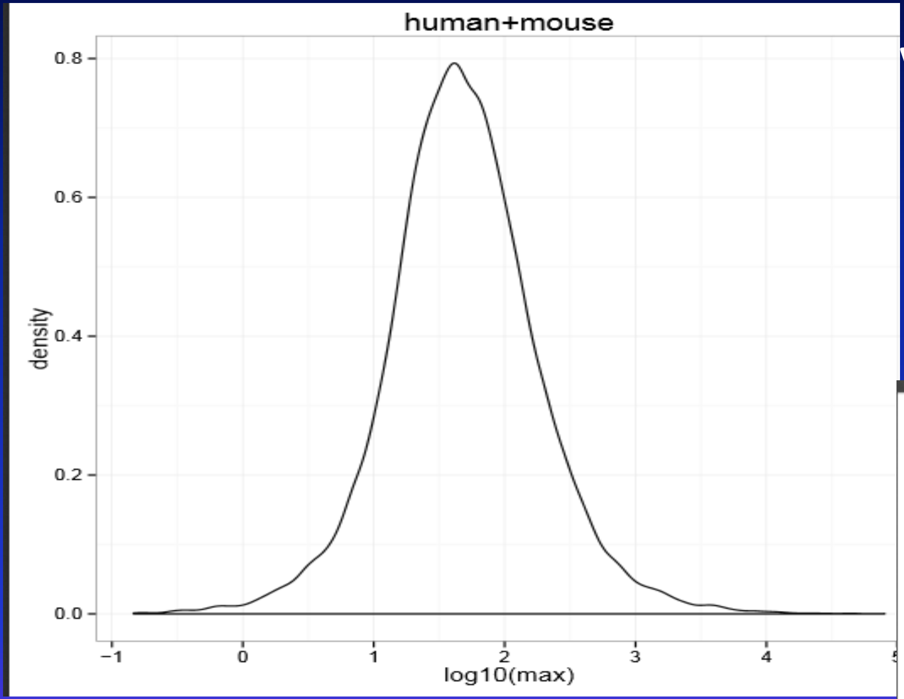


Min

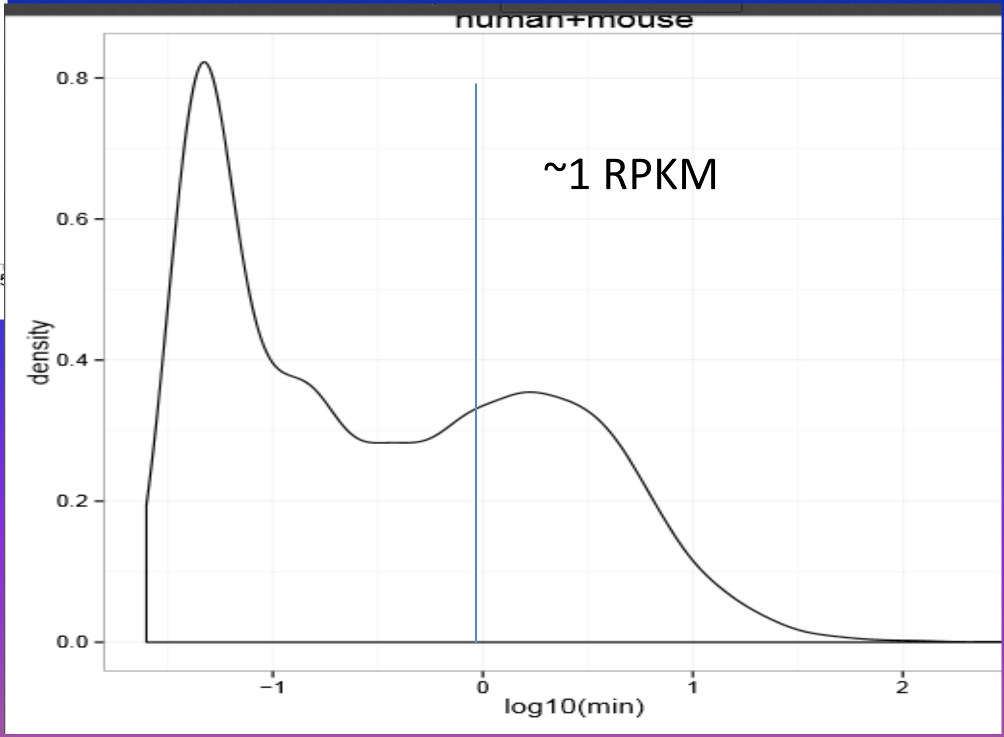


Mean

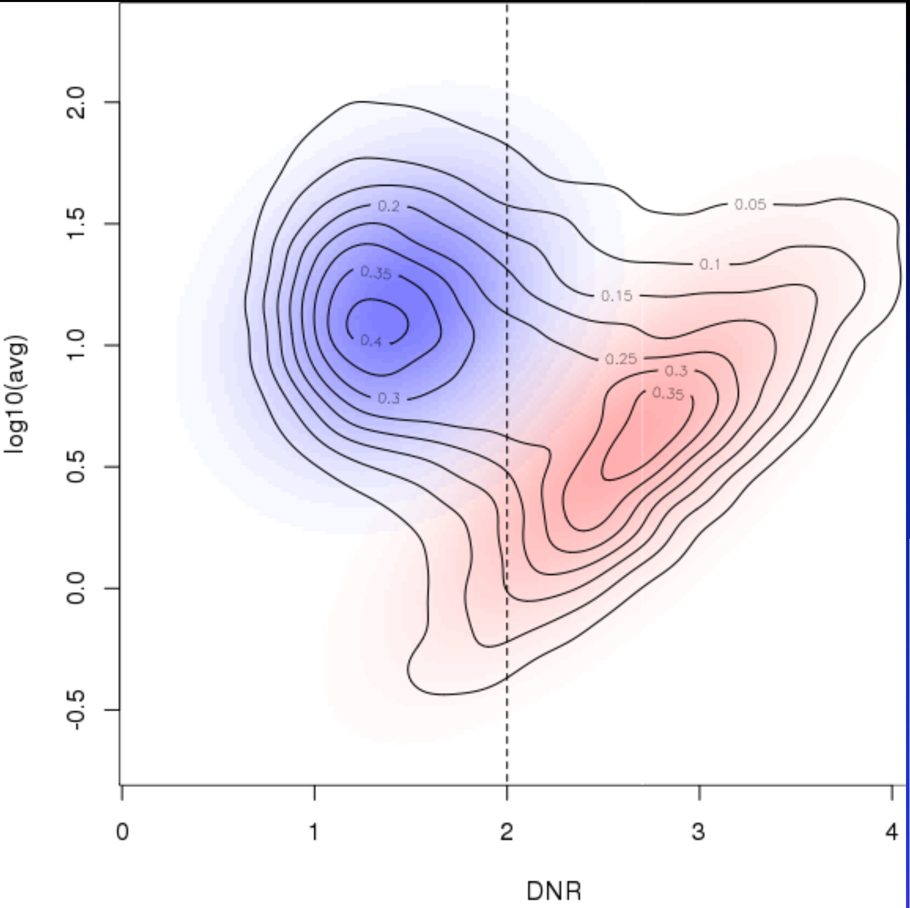
Distribution of # Genes and Log_{10} Max and Min RKPM Values



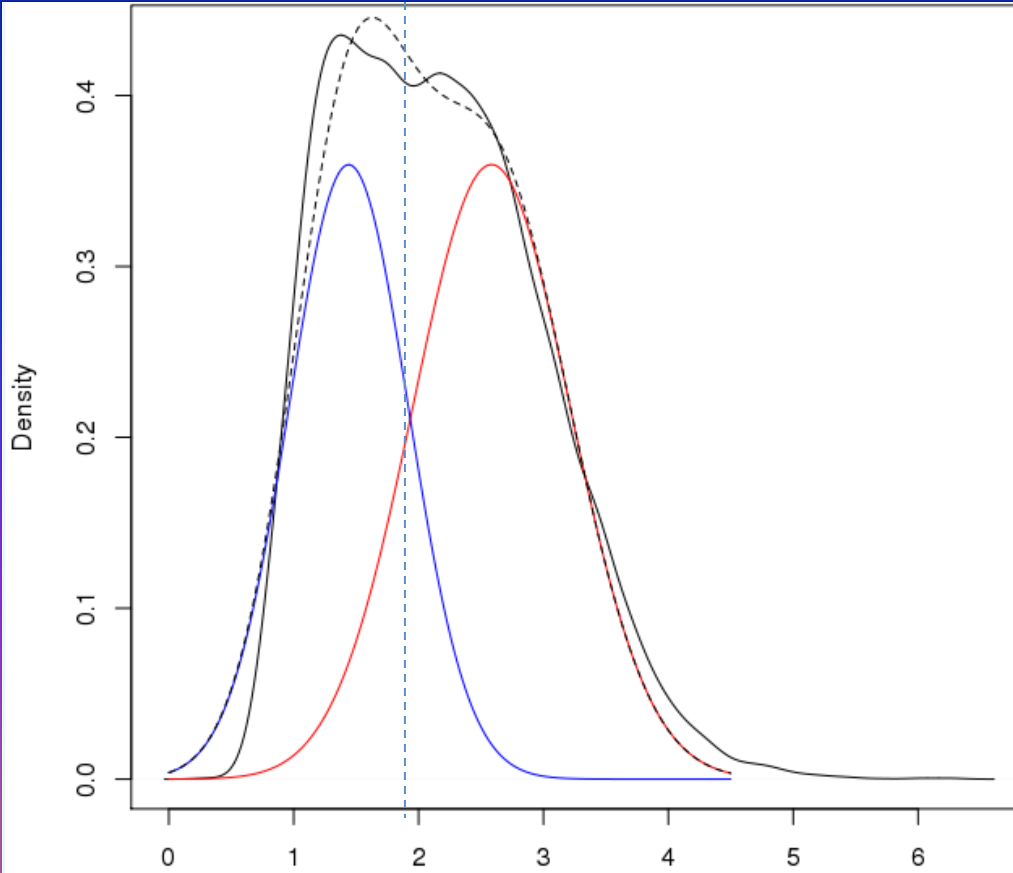
Log_{10} Max



Log_{10} Min



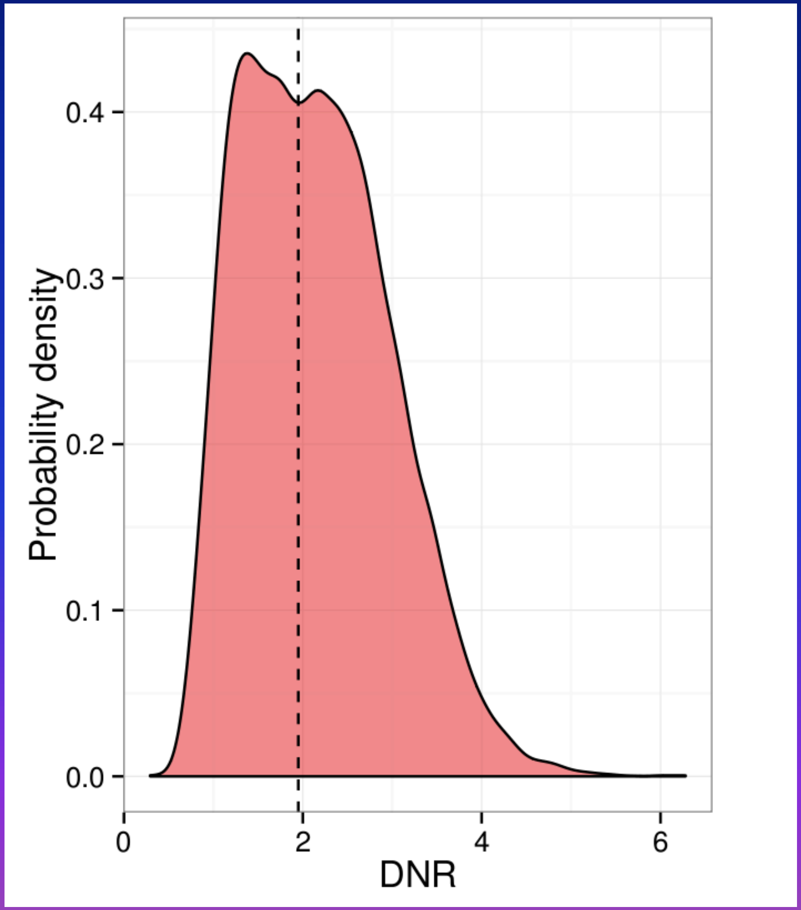
2 dimentional plot of log mean of expression vs DNR



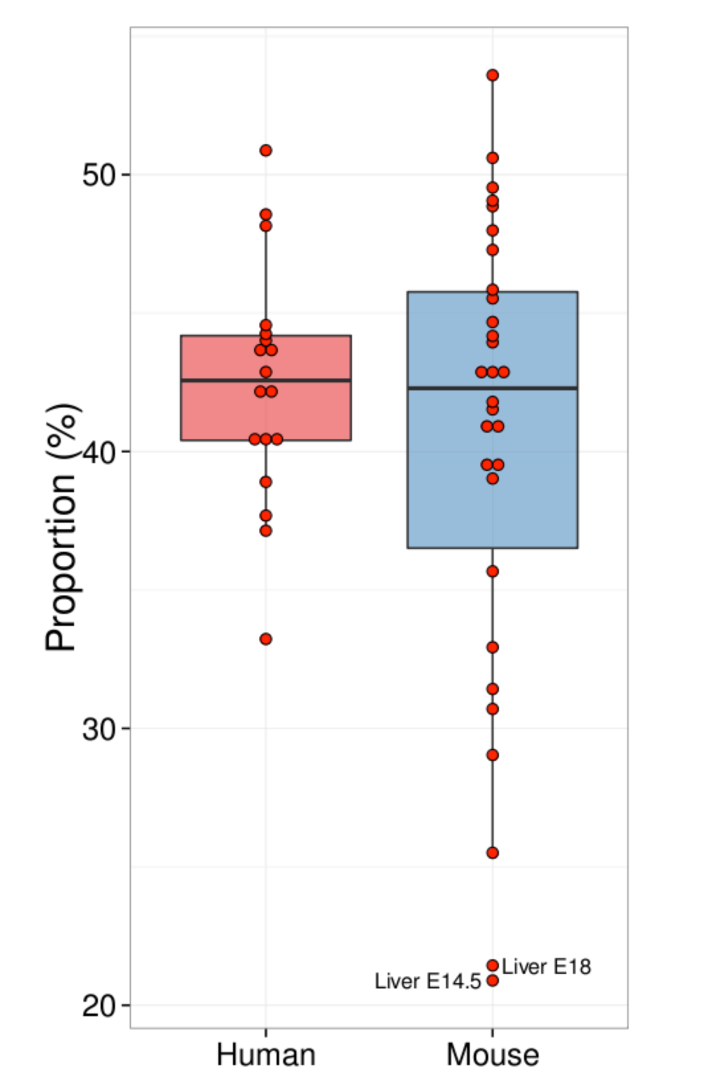
Deconvolved plots of gene number vs,. DNR

Two Gene Populations with Conserved Unconstrained and Constrained Variation in Levels of Expression

The dynamic range (DNR) of a gene expression levels across multiple sample types (cell lines and tissues) in human and mouse has a bimodal distribution, identifying two populations of genes with constrained (DNR=<2) and unconstrained (DNR=>2 levels of expression



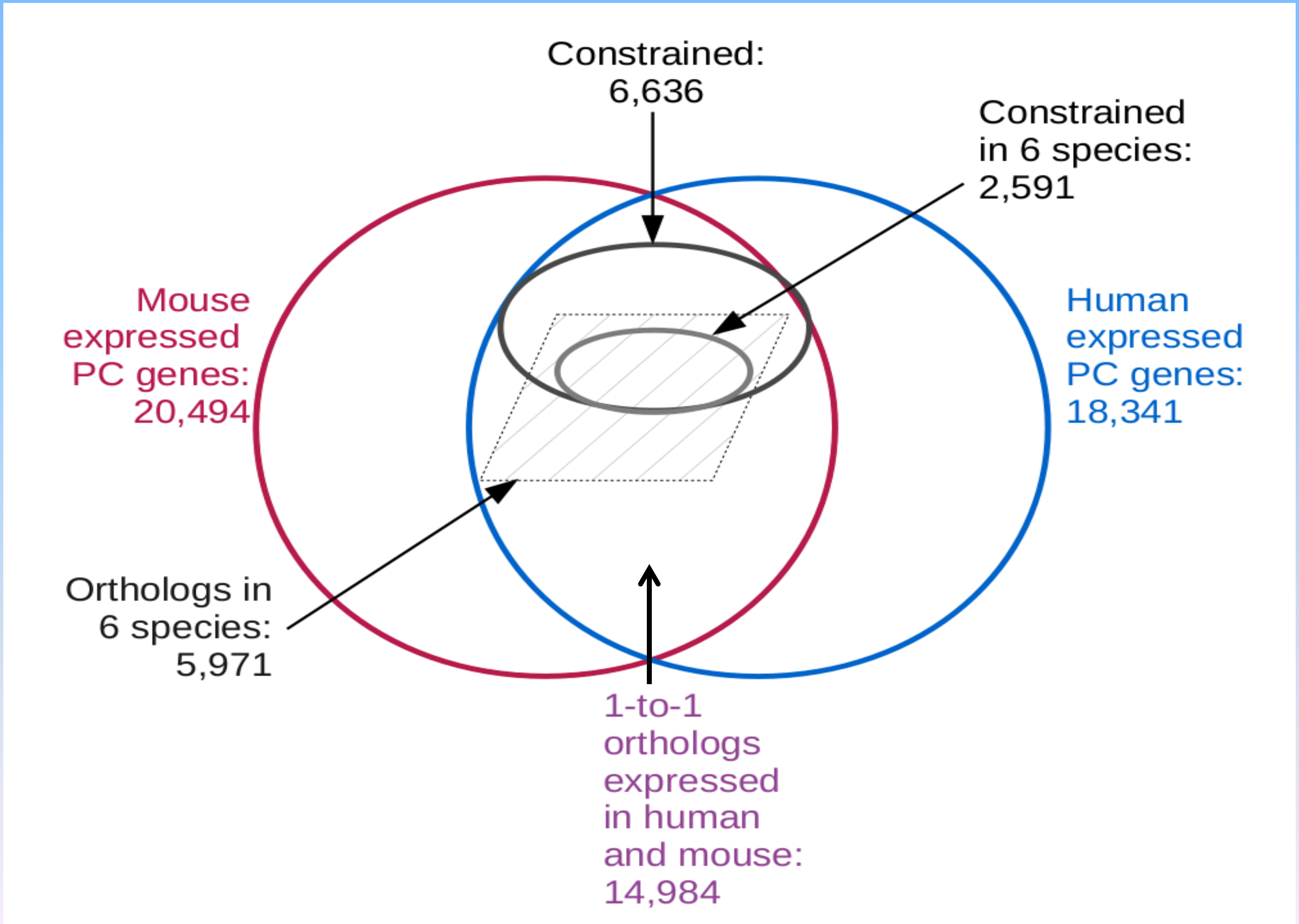
Constrained Genes Provide Substantial Fraction of Cell's/ Tissue's Total RNA Output



- Approximately 40% of RNA mass is attributed to the 17% of all annotated genes

- This RNA output is smaller for less differentiated cells (embryonic liver cells)

Number of Constrained Orthologous Protein Coding (PC) Genes in Six Species



Conclusions

- 73% and 81% of human and mouse 1:1 orthologue genes are expressed comparing cell lines vs tissues.

- 40% of orthologue genes expressed in mouse and human are expressed in 4 other species (macaque, rat , **chicken, cow)**

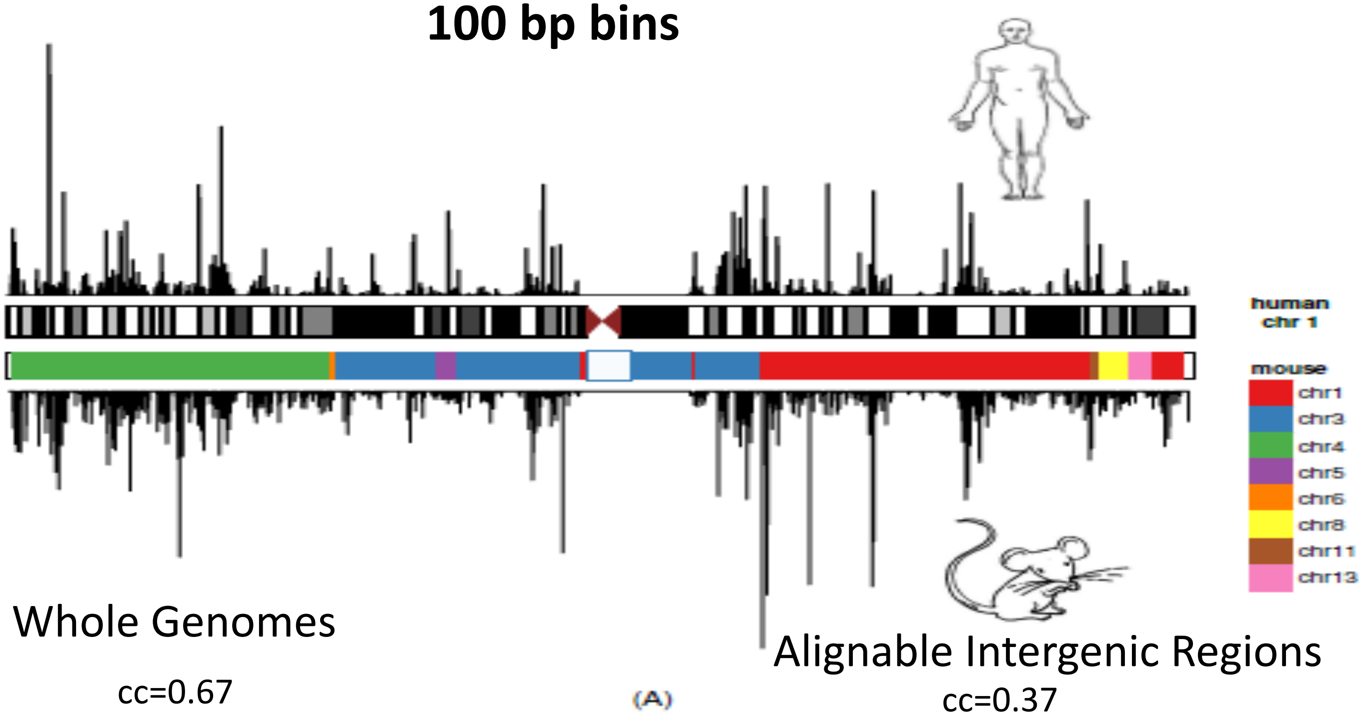
- 44% of expressed mouse and human orthologues have constrained expression (<2 log variation in expression)

- 17% of ortholgue genes expressed in mouse and human are constrained in their expression

- 39% of expressed mouse and human othorologue genes constrained in their expression are constrained in 4 other species

Correlation of Expression across the Mouse and Human Genomes

100 bp bins

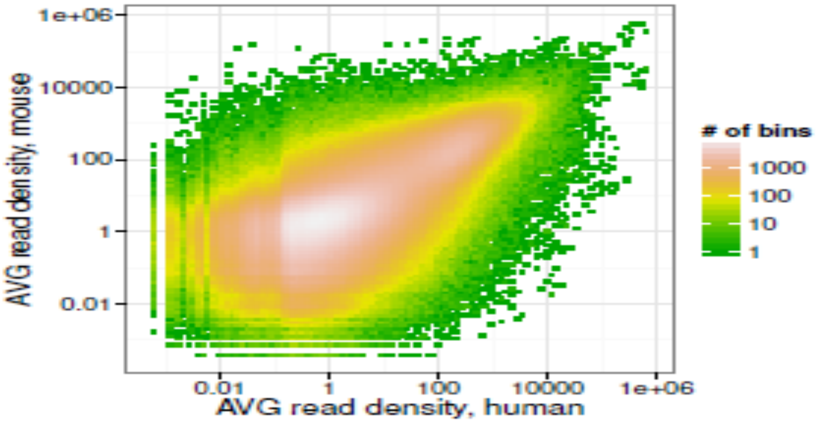


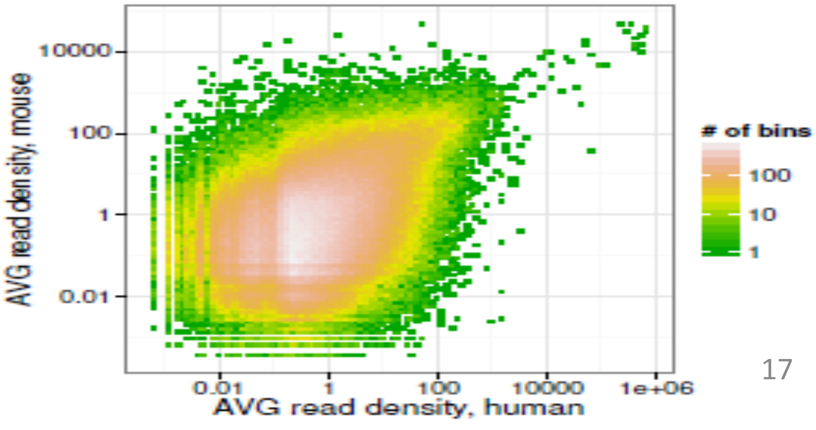
Whole Genomes

cc=0.67

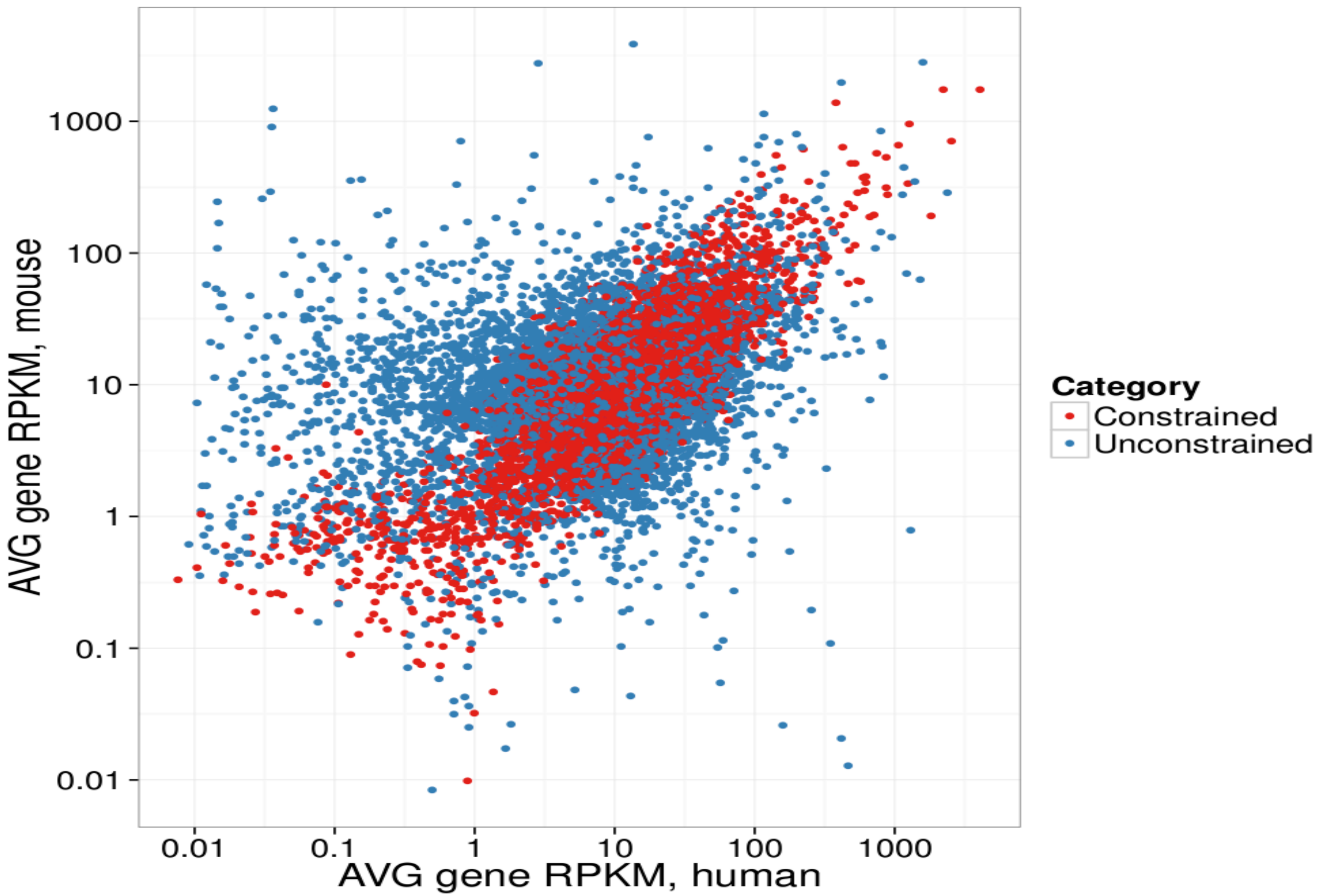
Alignable Intergenic Regions

cc=0.37

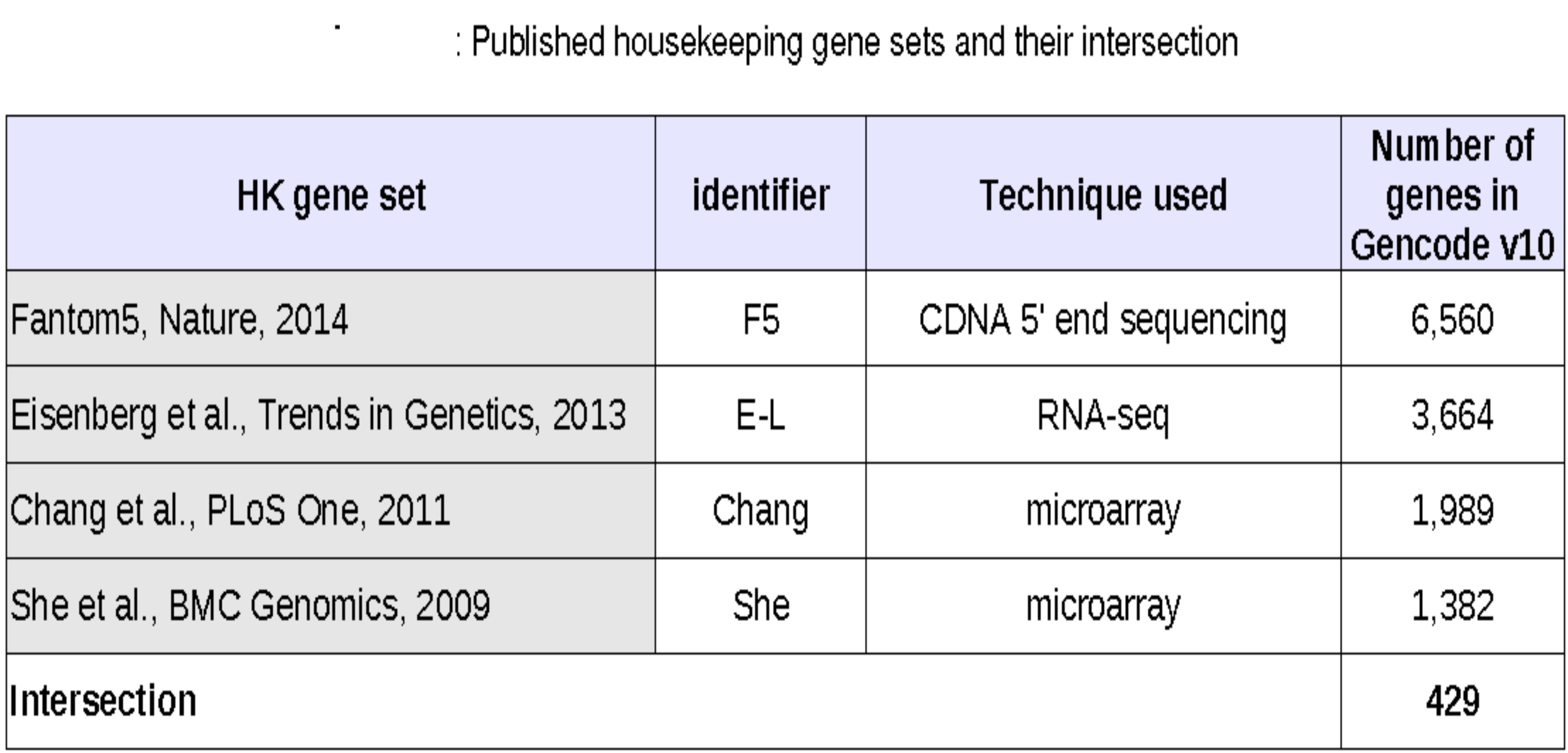




Constrained Genes are Drivers of the Correlation in Levels of Gene Expression seen for All Orthologous Genes



Is There a Consensus in Gene Membership for HKG



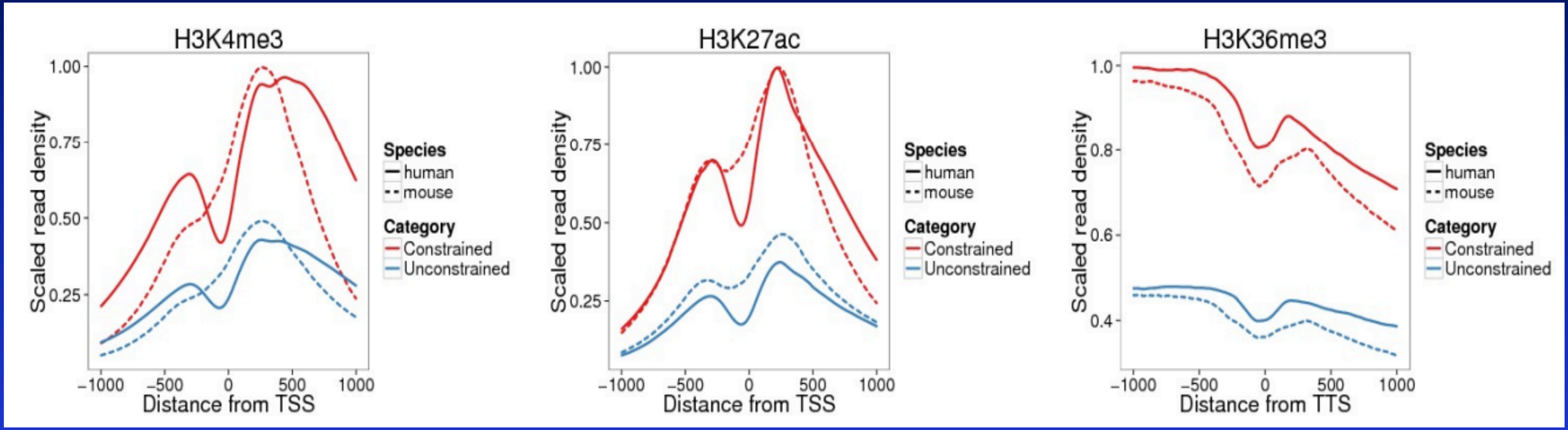
: Published housekeeping gene sets and their intersection

HK gene set	identifier	Technique used	Num ber of genes in Gencode v10
Fantom5, Nature, 2014	F5	CDNA 5' end sequencing	6,560
Eisenberg et al., Trends in Genetics, 2013	E-L	RNA-seq	3,664
Chang et al., PLoS One, 2011	Chang	microarray	1,989
She et al., BMC Genomics, 2009	She	microarray	1,382
Intersection			**429**

Proposal:
Principled Definition of Housekeeping Genes

Genes that have the variation in expression levels constrained irrespective of the tissue or species in which they are active.

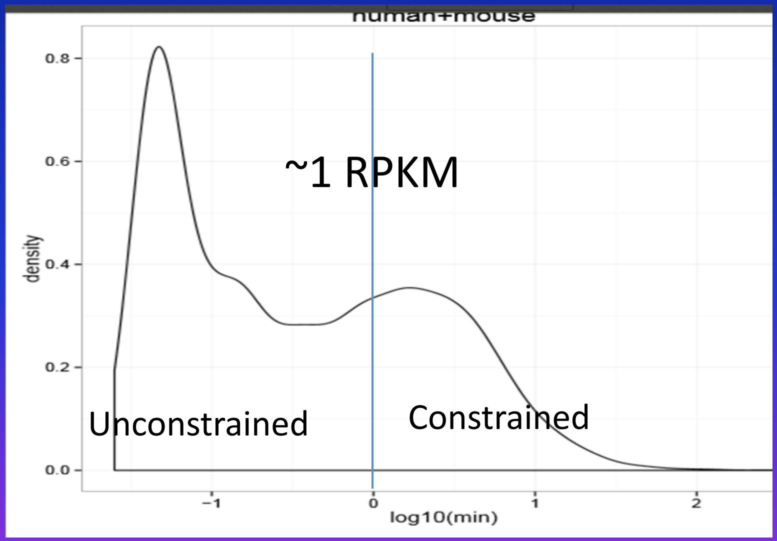
Possible Controls of the Conserved Constrained Gene Expression



- Constraint in gene expression is not reflected by sequence conservation
- Constrained set of genes have patterns of histone modification different from unconstrained genes
- Using human and mouse ENCODE epigenetic data for all cell types studied, stronger histone modification signals (H3K4me3, H3K27ac and H3K36me3) for constrained vs. unconstrained genes (controlling for levels of gene expression sample by sample)
- Suggesting constrained vs. unconstrained gene are under different epigenetic regulatory programs

Other Questions

- Mechanism(s) responsible for establishing, maintaining and inheriting the restricted variation in expression



- What genes are constrained at 1 RPKM in what cell/tissue types
- Are there uncontrained genes that determine cell type and to what levels of expression are they in different cell types
- Do these properties extend to lnc-RNA genes
- What about non-orthologous genes?

Acknowledgements

Cold Spring Harbor

Functional Genomics Group

C. Davis
A. Dobin
J. Drenkow
A. Scavelli
L. H. See
C. Zaleski

CRG, Barcelona

Computational Genomics Group

R. Guigo
A. Breschi
S. Djebali
J. Lagarde
D. Pervouchine