Pseudogenes within GENCODE Annotation and their Chromatin Features and Evolutionary Constraints

Baikang Pei
Group Meeting
Gerstein Lab
07/13/2011

Outline

- Extraction of pseudogene annotation from GENCODE
- Pseudogenes with transcription evidence
- Chromatin features of pseudogenes
- Evolutionary constraints on pseudogenes
- Summary and Future Directions

Pseudogenes in GENCODE Annotation v7

11216 Pseudogenes are extracted from gencode v7 annotation file with the following criteria:

- Only keep level 1 or level 2 pseudogenes;
- Transcript type must be pseudogene;
- Remove polymorphic pseudogenes.

Level 1 (7184)

Transcript Type	Count
processed_pseudogene	6338
unprocessed_pseudogene	288
transcribed_processed_pseudogene	99
transcribed_unprocessed_pseudogene	34
unitary_pseudogene	11
pseudogene	406
IG_V_pseudogene	7
TR_V_pseudogene	1

Level 2 (4032)

Transcript Type	Count
processed_pseudogene	1769
unprocessed_pseudogene	1571
transcribed_processed_pseudogene	42
transcribed_unprocessed_pseudogene	234
unitary_pseudogene	127
pseudogene	115
IG_V_pseudogene	144
IG_J_pseudogene	3
IG_C_pseudogene	7
TR_V_pseudogene	20

GENCODE v3c vs. v7

Level 1

Transcript Type	v7	v3c
processed_pseudogene	6338	3018
unprocessed_pseudogene	288	65
transcribed_processed_pseudogene	99	12
transcribed_unprocessed_pseudogene	34	2
unitary_pseudogene	11	2
pseudogene	406	24
IG_V_pseudogene	7	48
TR_V_pseudogene	1	-
Total	**7184**	**3171**

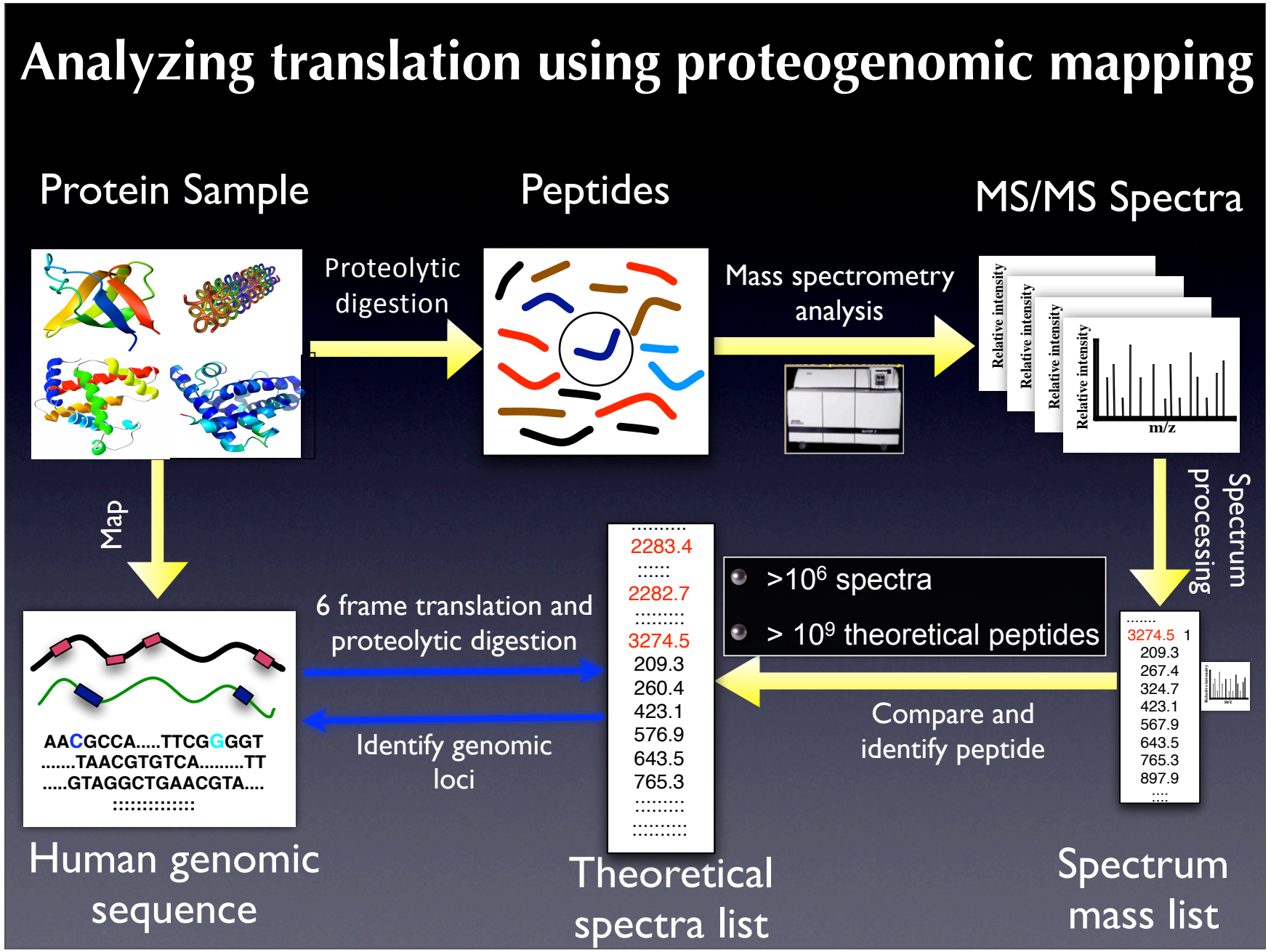
Level 2

Transcript Type	v7	v3c
processed_pseudogene	1769	3211
unprocessed_pseudogene	1571	1201
transcribed_processed_pseudogene	42	50
transcribed_unprocessed_pseudogene	234	146
unitary_pseudogene	127	120
pseudogene	115	609
IG_V_pseudogene	144	112
IG_J_pseudogene	3	
IG_C_pseudogene	7	
TR_V_pseudogene	20	19
Total	**4032**	**5468**

Transcription Evidences

- Transcription annotation from GENCODE file (409)
- BodyMap + PseudoSeq pipeline (381)
- Proteogenomic mapping

Analyzing translation using proteogenomic mapping



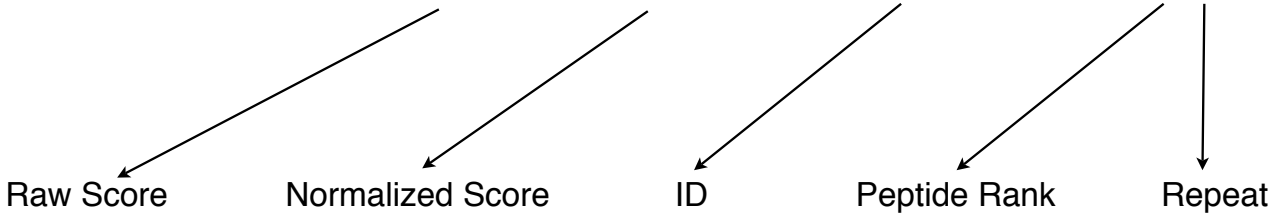
Proteogenomic Data

Proteogenomic analysis was carried out for proteins from nuclear fraction of GM12878.

At 10% FDR, ~ 95K genome loci matches from GM12878.

Data can be downloaded from UCSC genome browser.

chr1 67891915 67891948 RPDQQLQGEGK 182 - 187.443329 628.32638751605 1 3



Mapping to Pgenes

Peptides: peptides identified from nuclear fraction of GM12878

Pgenes: 11,216 HAVANA pseudogenes from GENCODE 7

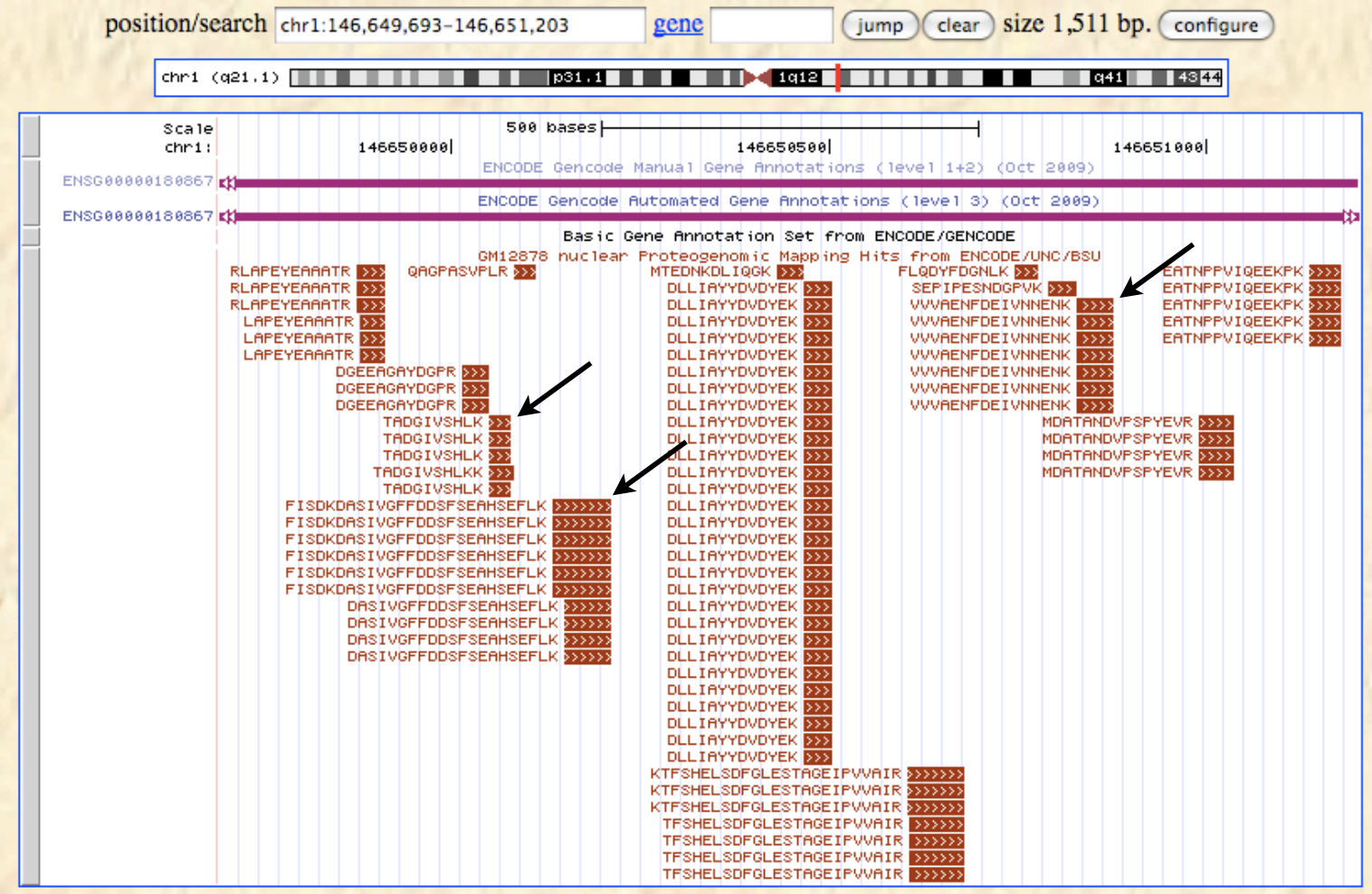
Results Summary

Total peptides searched	95,013
Total peptides mapped to pgene*	1,566
number of pgenes covered	230

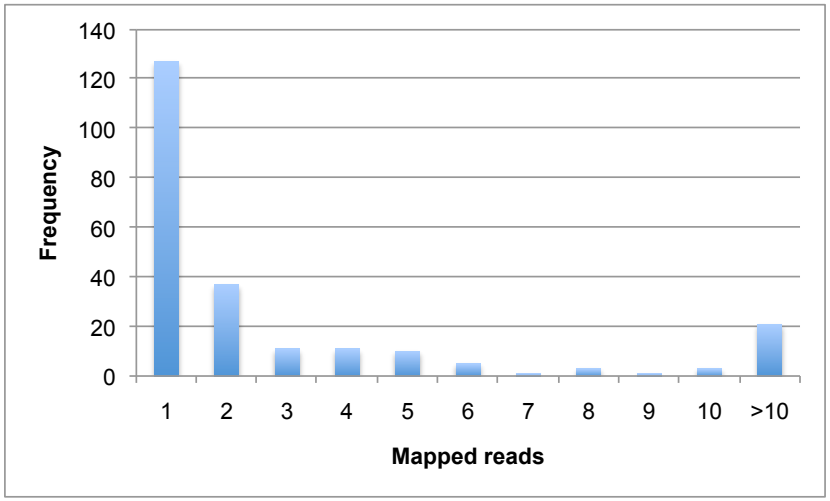
* Only unique mappings are kept

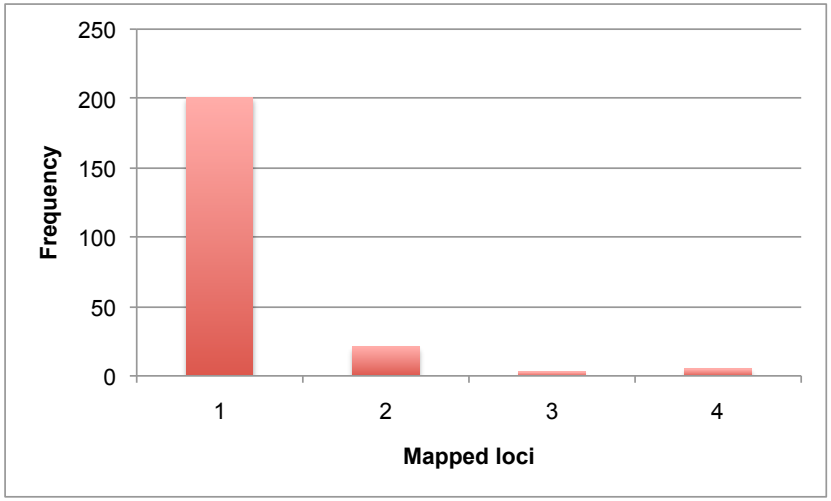
Case Study

gene_id: ENSG00000180867; transcript_id: ENST00000471856; processed_pseudogene



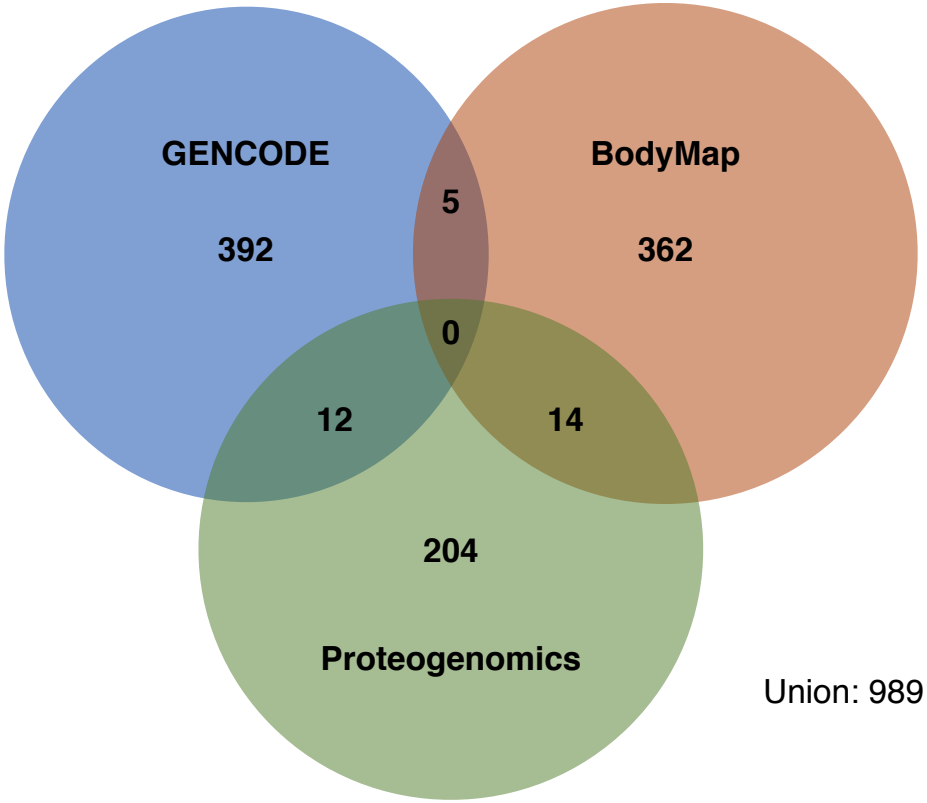
Mapping Distribution





“Expressed” Pseudogenes

1. Transcribed pseudogenes from GENCODE annotation: totally 409 pseudogenes.
2. Transcribed pseudogenes from Pseudoseq using BodyMap data, lifteOver to Hg19.
3. Proteogenomics data from Morgan Gidding’s lab for nuclear fraction of GM12878

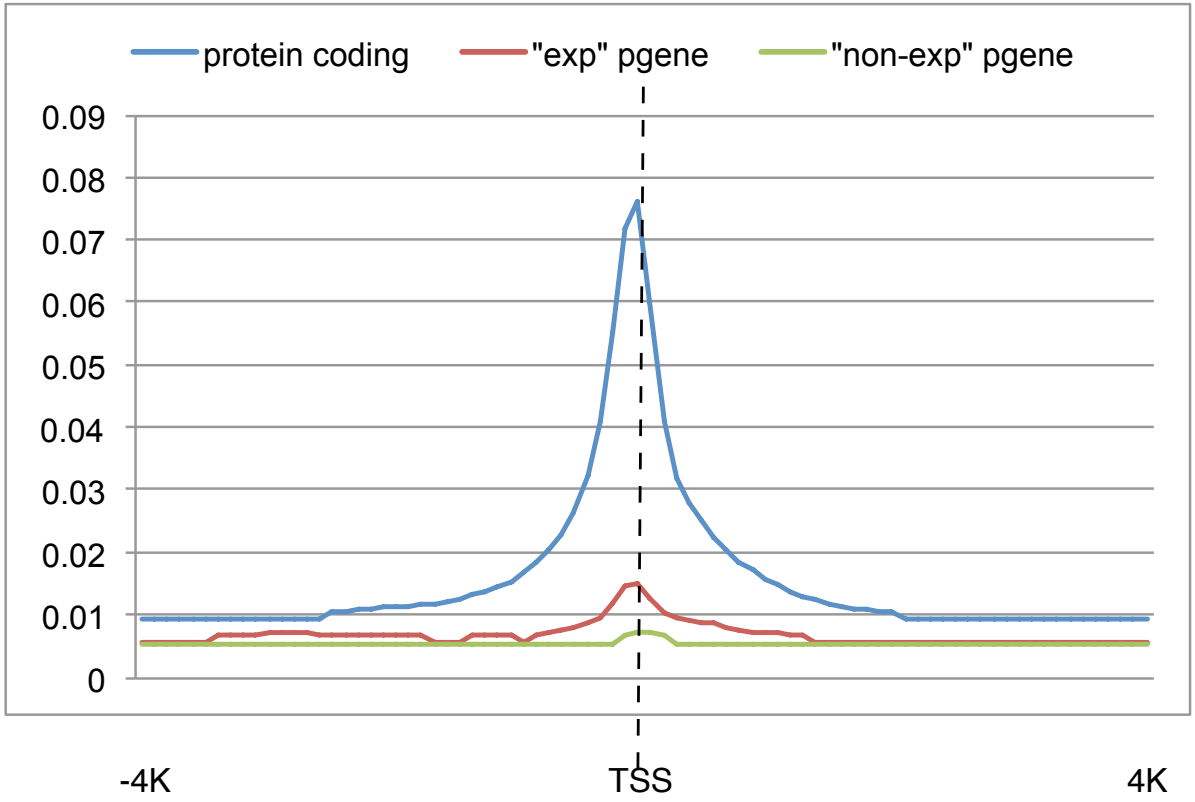


Union: 989 pseudogenes

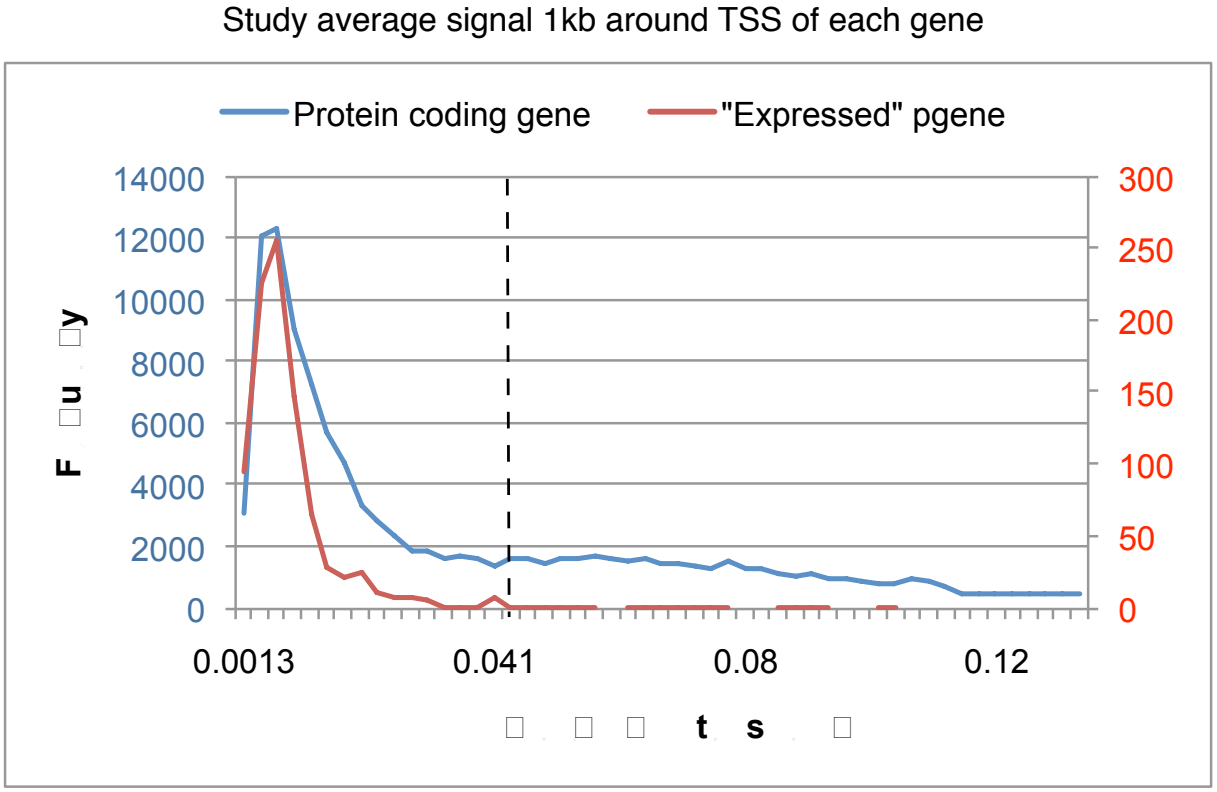
Chromatin Features

Chromatin Accessibility

DnaseI Hyper-sensitivity data on GM12878 by Duke University



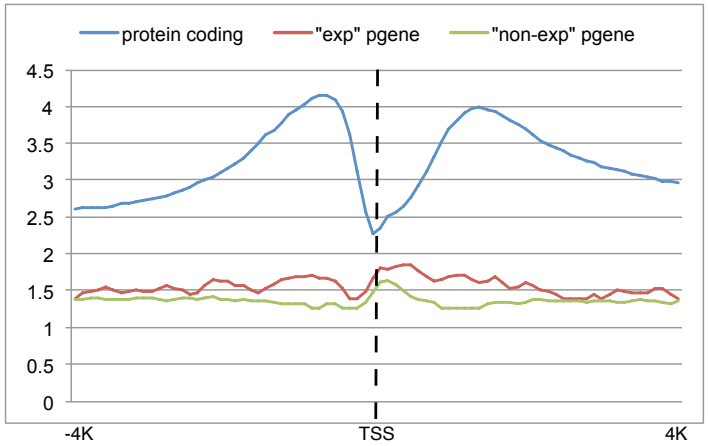
Chromatin Accessibility Distribution



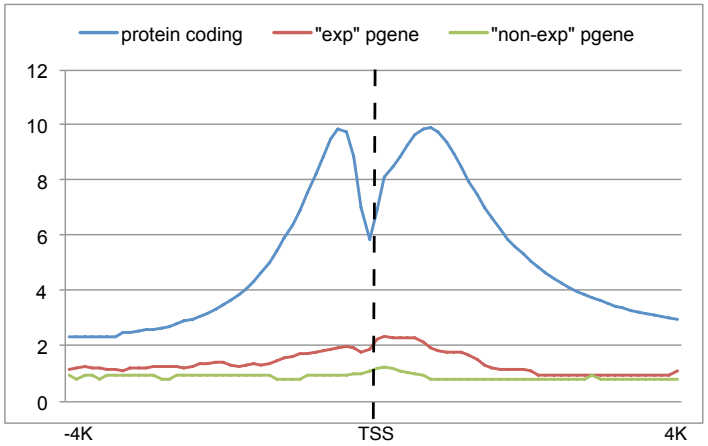
25 pgenes have signal higher than protein coding average.

Histone Modification

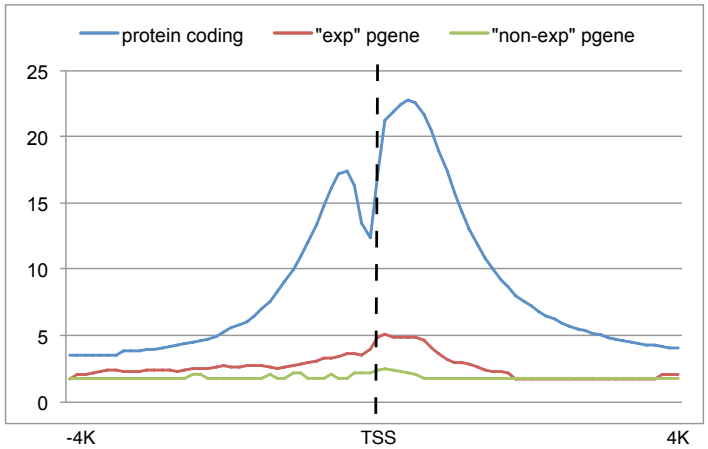
H3K4me1



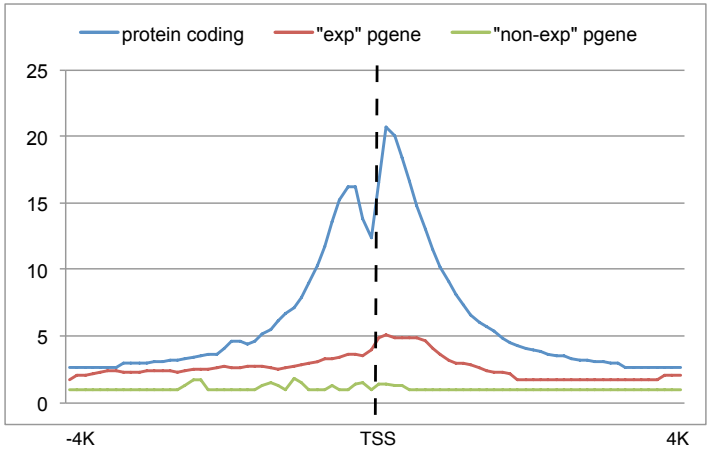
H3K4me2



H3K4me3

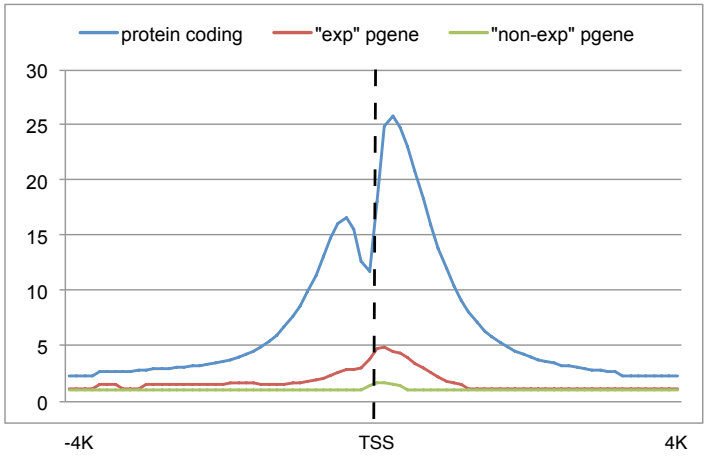


H3K27ac

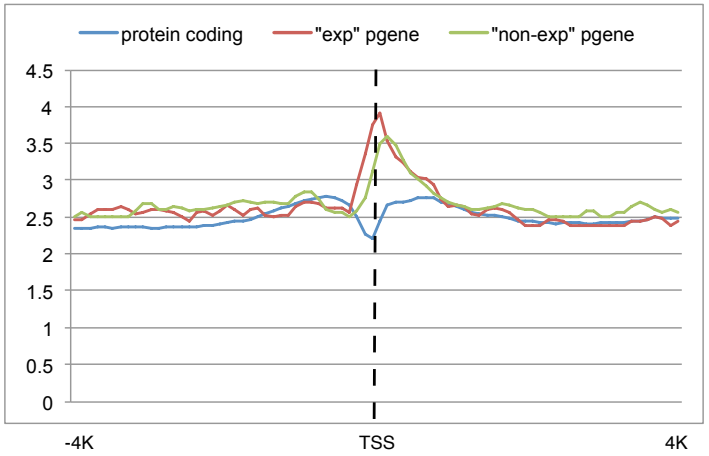


Histone Modification (Cont.)

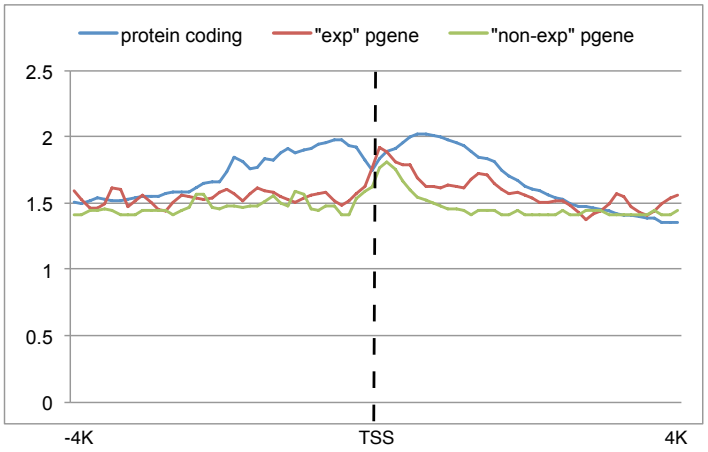
H3K9ac



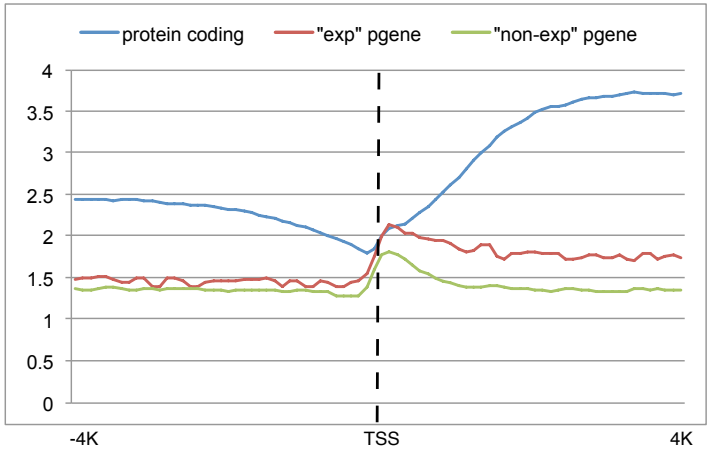
H3K9me3



H3K27me3

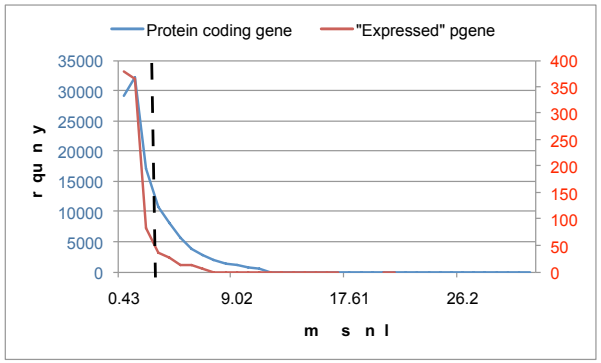


H3K36me3

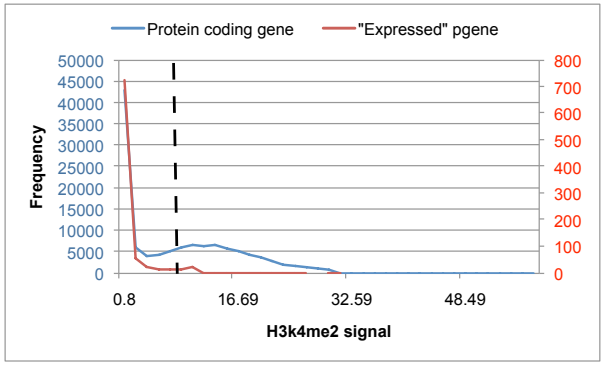


Histone Modification Distribution

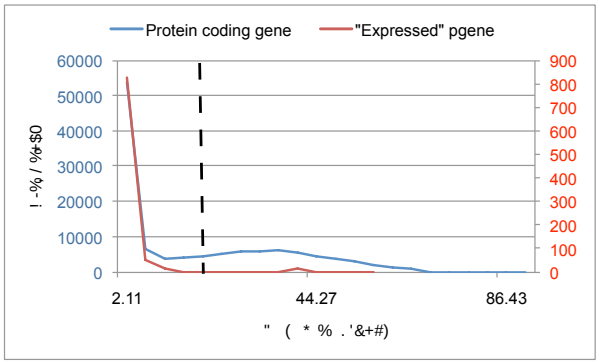
Study average histone mark signal 1kb around TSS of each gene, except for H3k36me3, which the signal is from 1kb after TSS



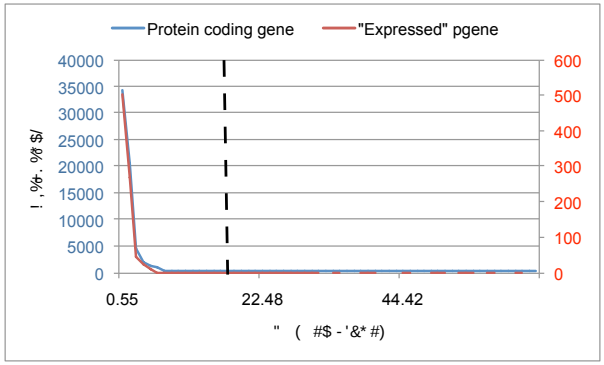
126 pgenes have signal higher than protein coding average.



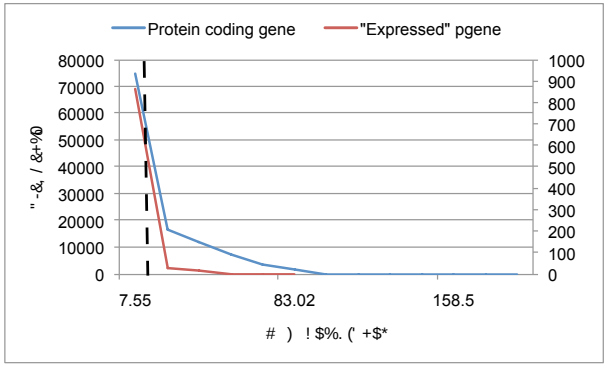
90 pgenes have signal higher than protein coding average.



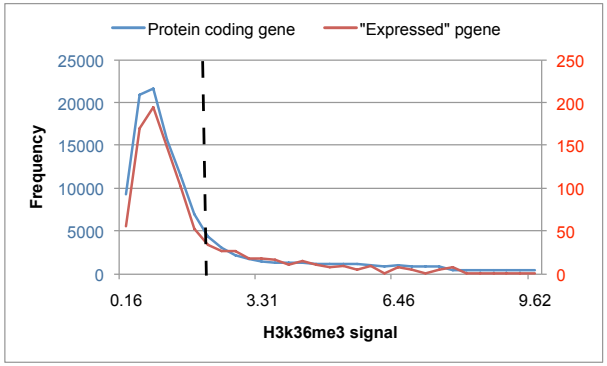
68 pgenes have signal higher than protein coding average.



57 pgenes have signal higher than protein coding average.



47 pgenes have signal higher than protein coding average.

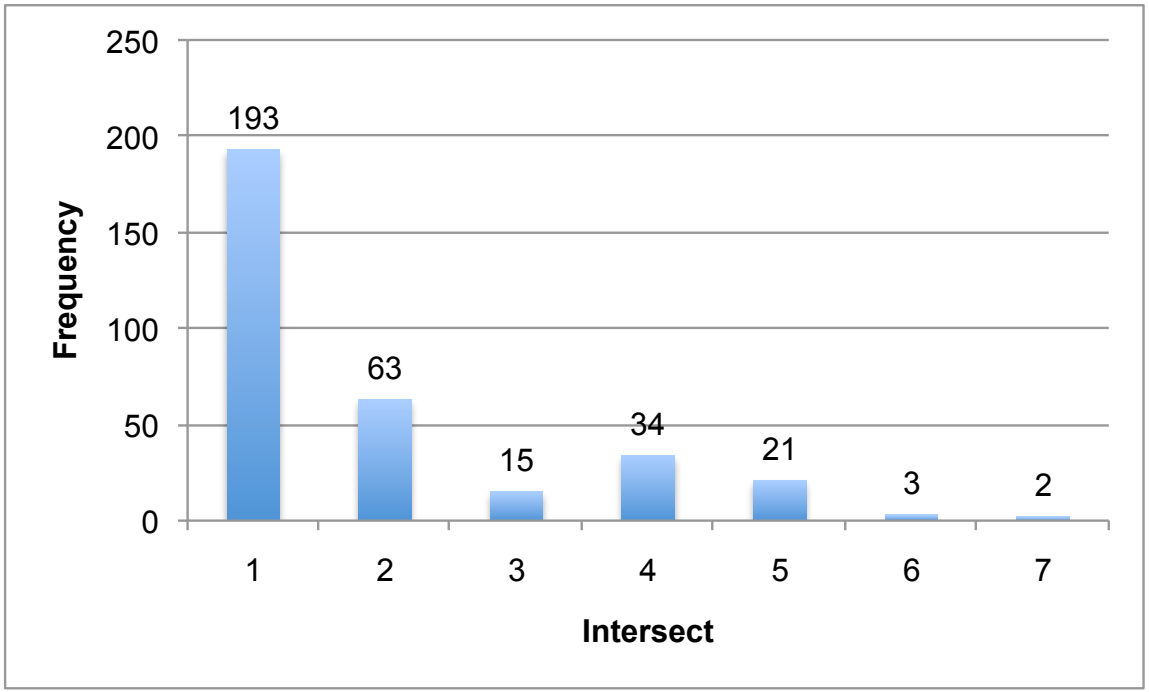


224 pgenes have signal higher than protein coding average.

Intersect of pgenes with active chromatin feature

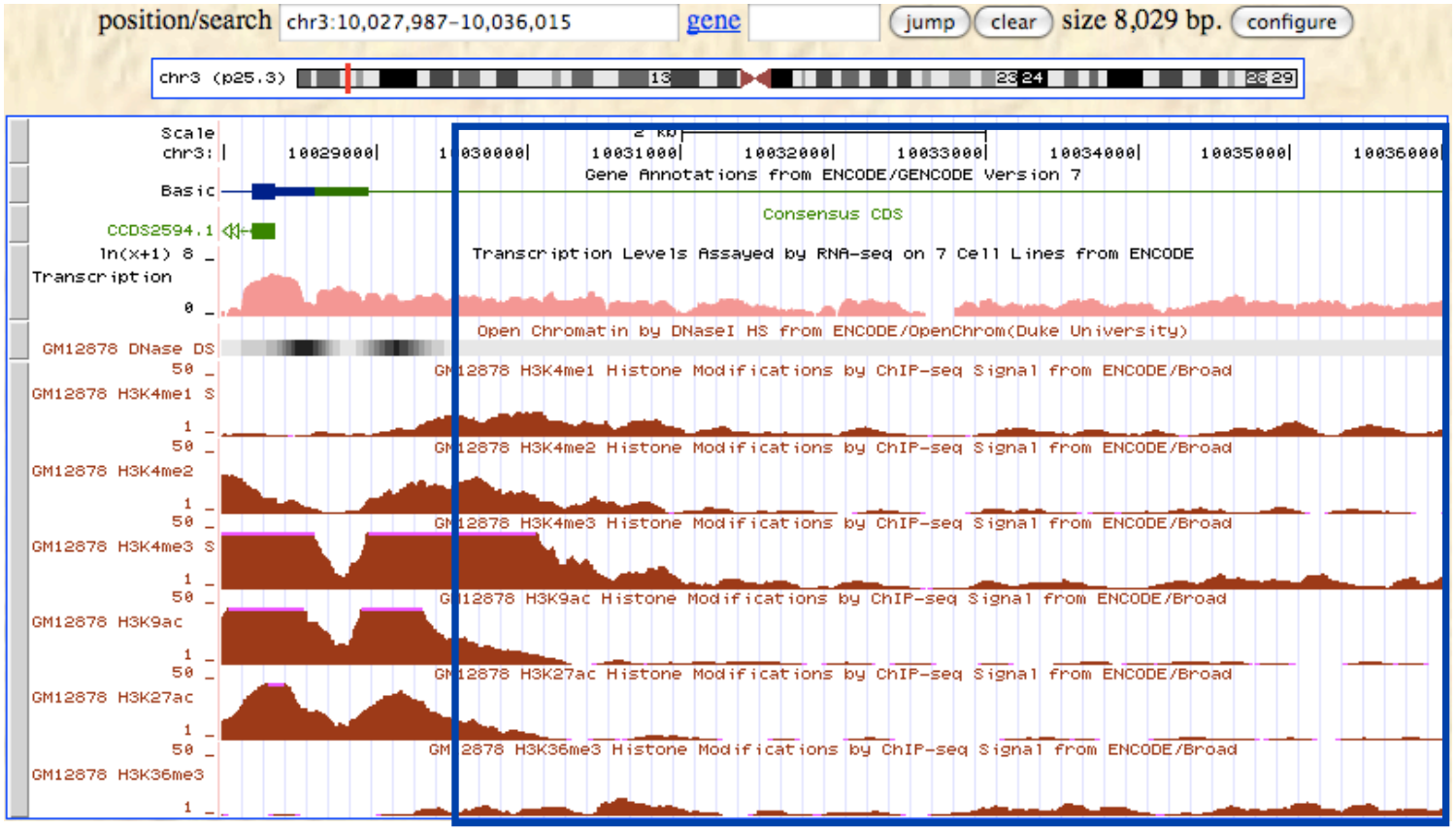
Number of pgenes with active chromatin feature

Open Chromatin	24
H3k4me1	125
H3k4me2	89
H3k4me3	67
H3k9ac	56
H3k27ac	46
H3k36me3	223



Case Study

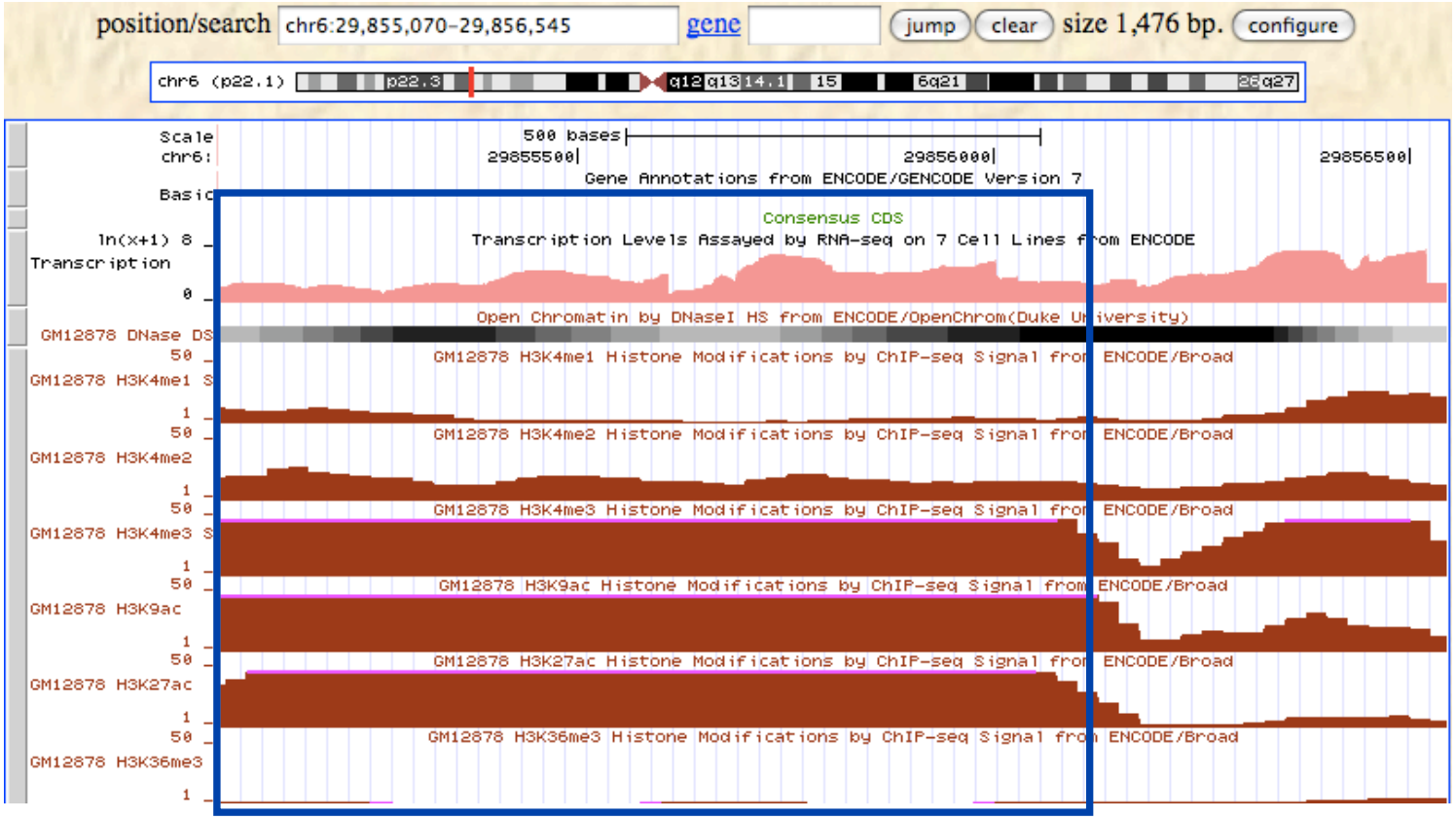
gene_id "ENSG00000180385.4"; transcript_id "ENST00000438698.1". In plus strand



This pseudogene is annotated as transcribed_unprocessed_pseudogene, and also called by the Pseudoseq pipeline.

Case Study

gene_id "ENSG00000233902.1"; transcript_id "ENST00000440087.1". In minus strand



This pseudogene is called by the Pseudoseq pipeline as transcribed pgene

Evolutionary Constraint

Constraint Regions in Pseudogenes

Annotation: from GENCODE v7 annotation file

- Pseudogenes: exons, processed pgenes, non-processed pgenes, etc.
- Protein coding genes: CDS, 5' UTR, 3' UTR and introns

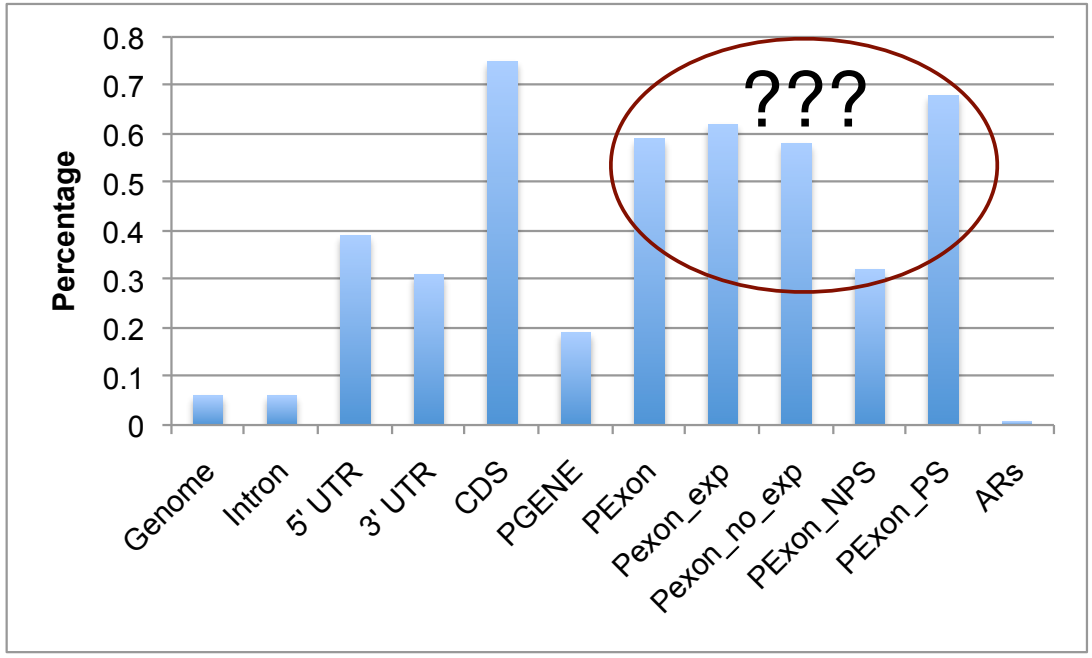
Constrain Annotation: download from ENCODE wiki

- EPO alignment of 33 mammals (of which 22 are 2x mammal)
- Evolutionary tree
- GERP scores and constrained elements

GERP: v2.1 from Sidow lab

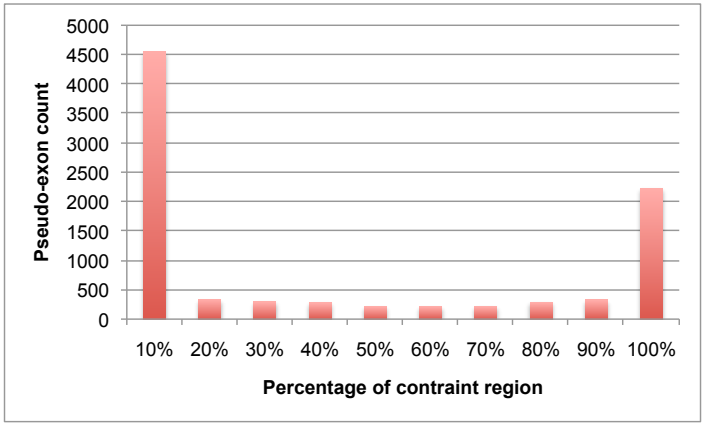
- Get same results when using multiple alignment and evolutionary tree from ENCODE wiki

Percentage of Genomic Region Under Constraint

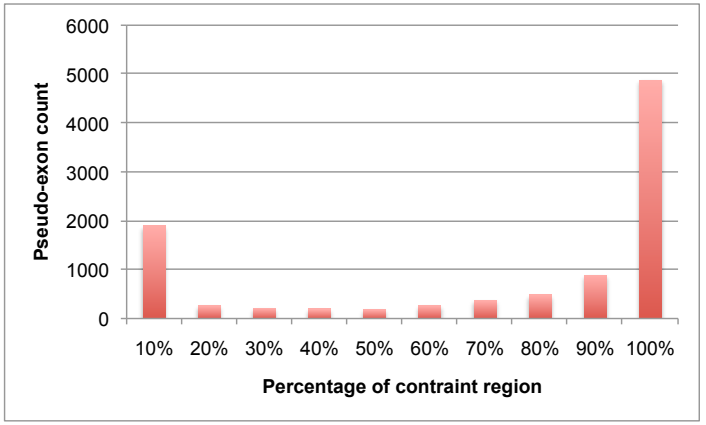


Distribution of Constraint Regions in Pseudo-exons

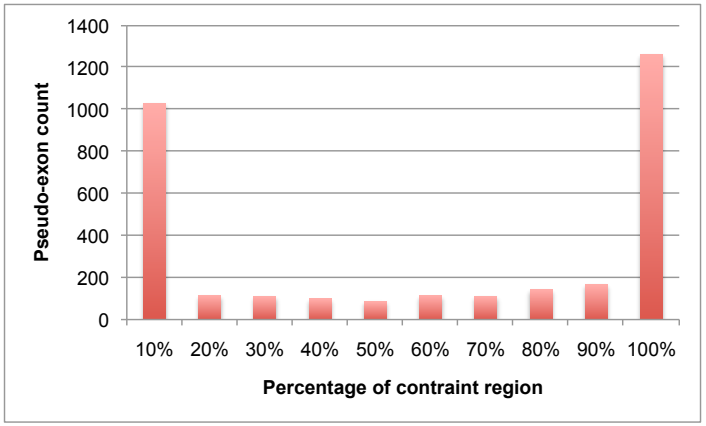
Unprocessed Pseudogens



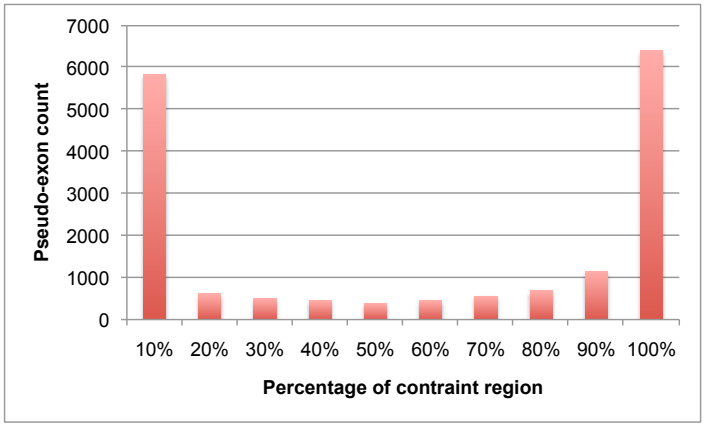
Processed Pseudogens



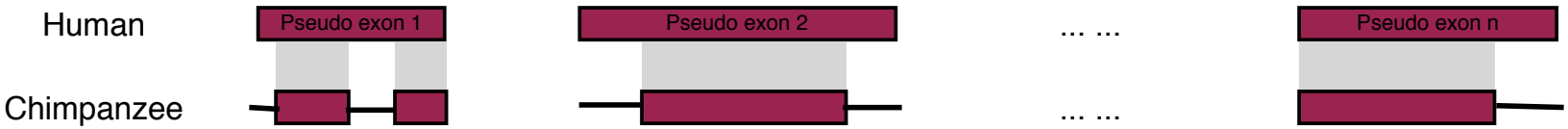
“expressed” Pseudogens



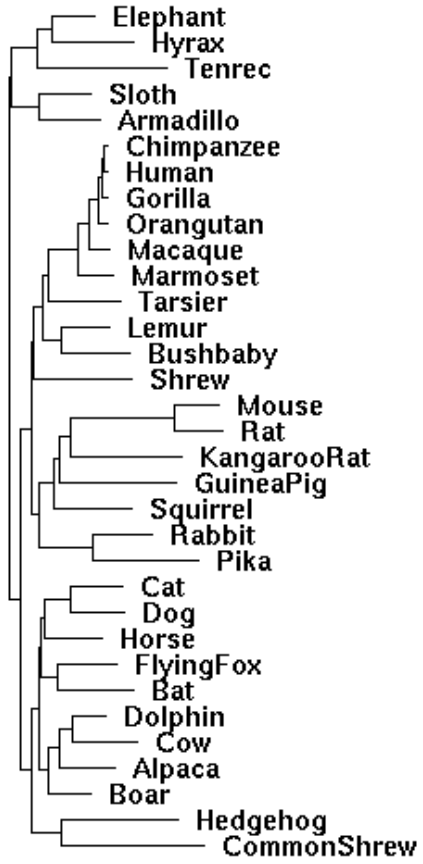
“Unexpressed” Pseudogens

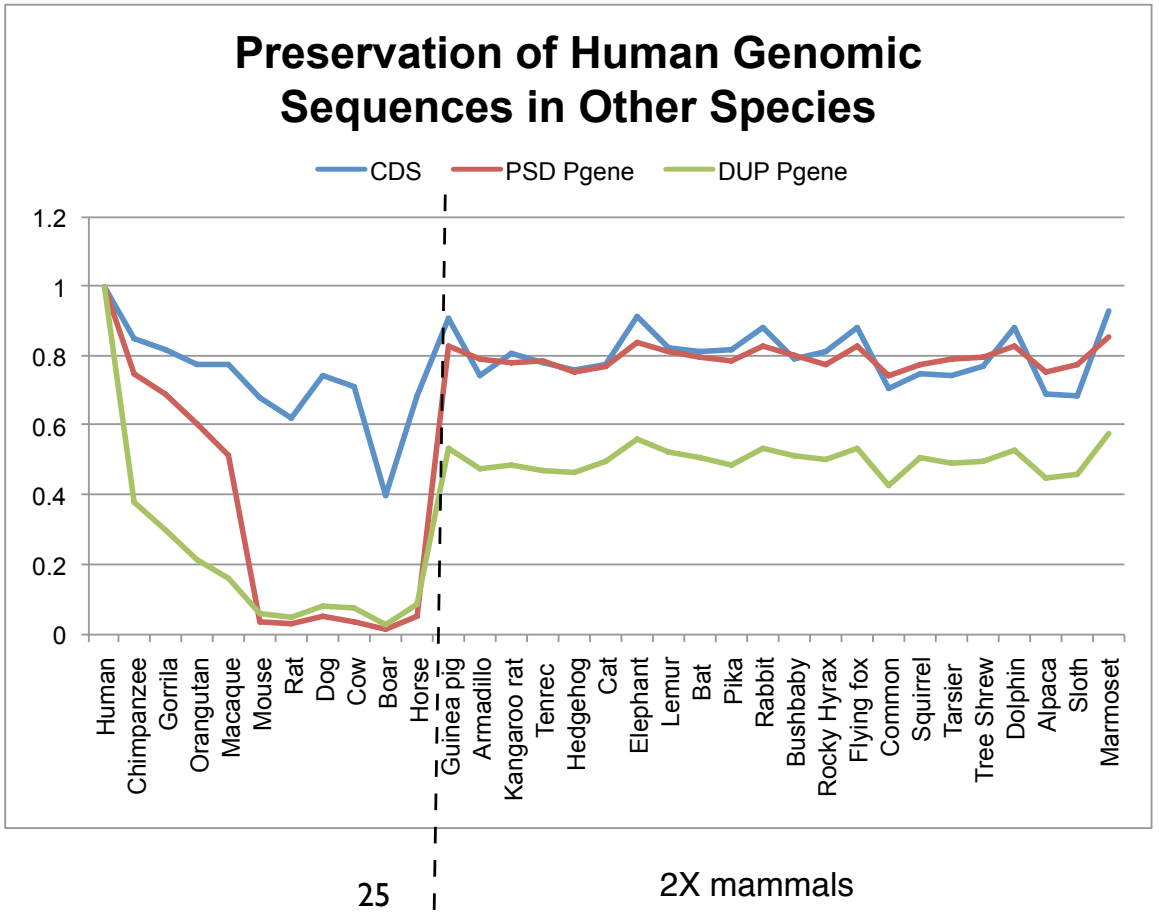


Multiple Sequence Alignments



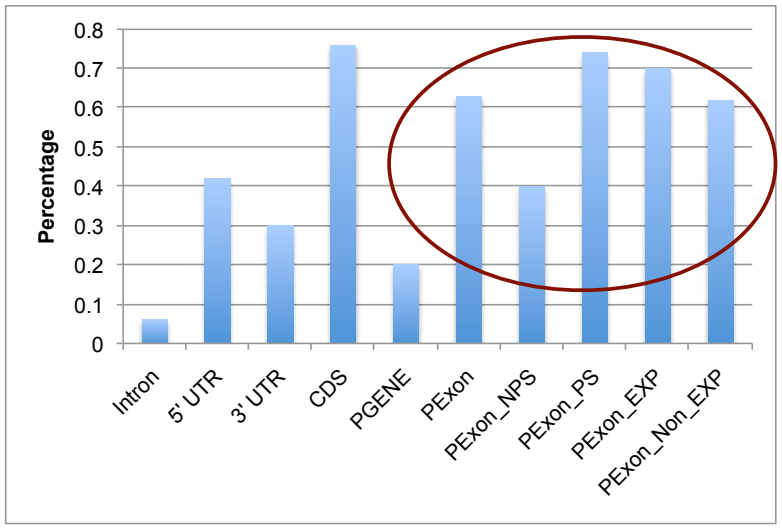
Preservation = overlap / total



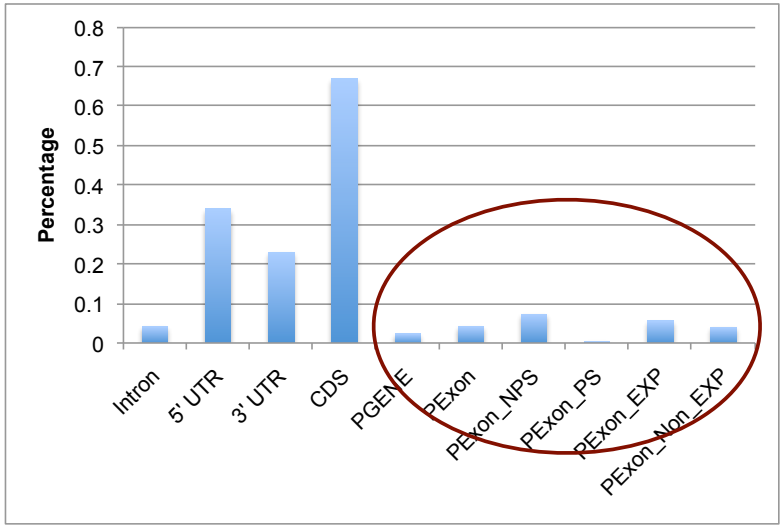


Percentage of Constraint Regions on Chromosome 1

Gerp results from 33 species

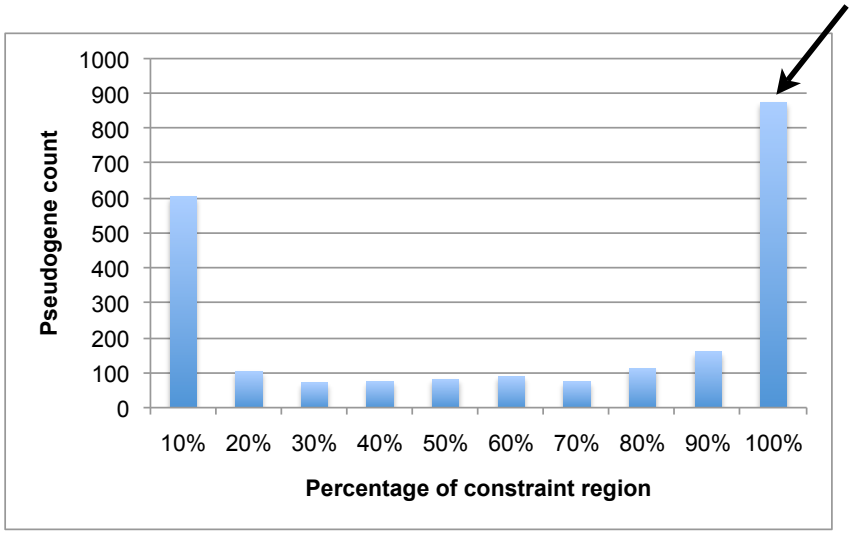


Gerp results from 11 species, without 2x mammals

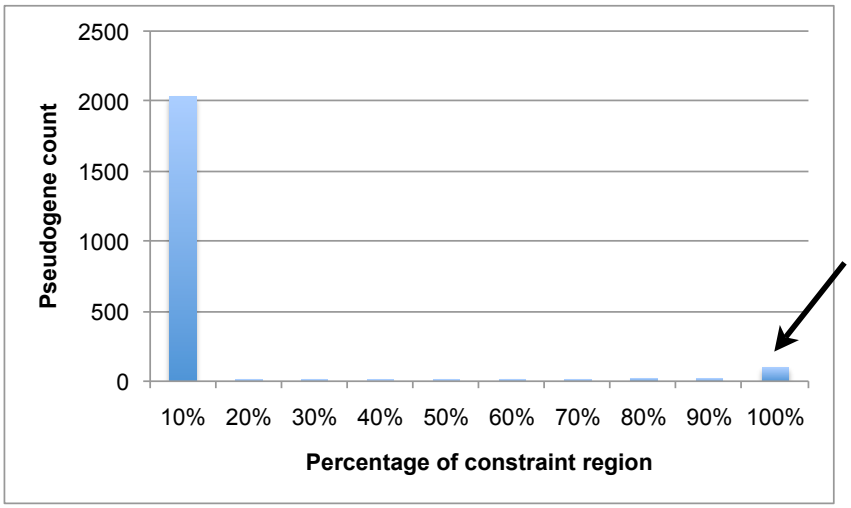


Distribution of Constraint Regions in Pseudo-exons on Chromosome 1

Gerp results from 33 species

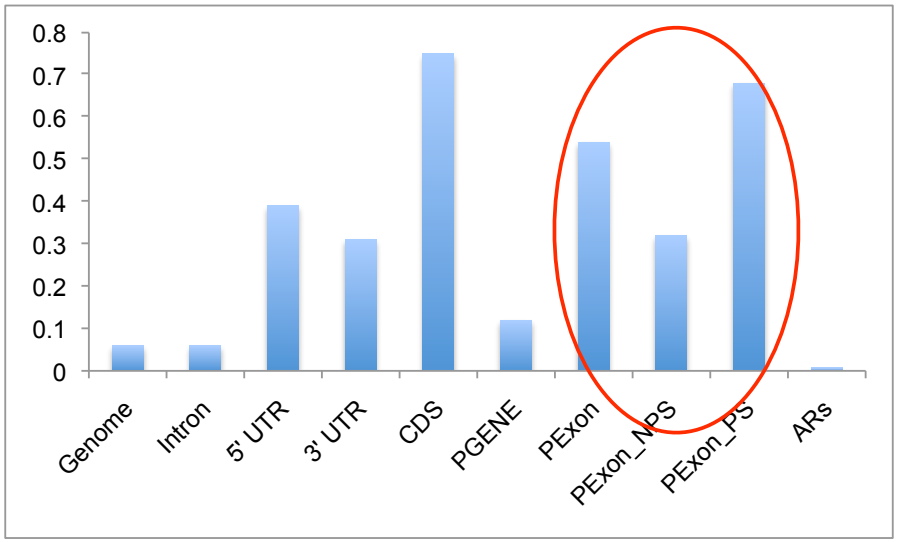


Gerp results from 11 species, without 2x mammals

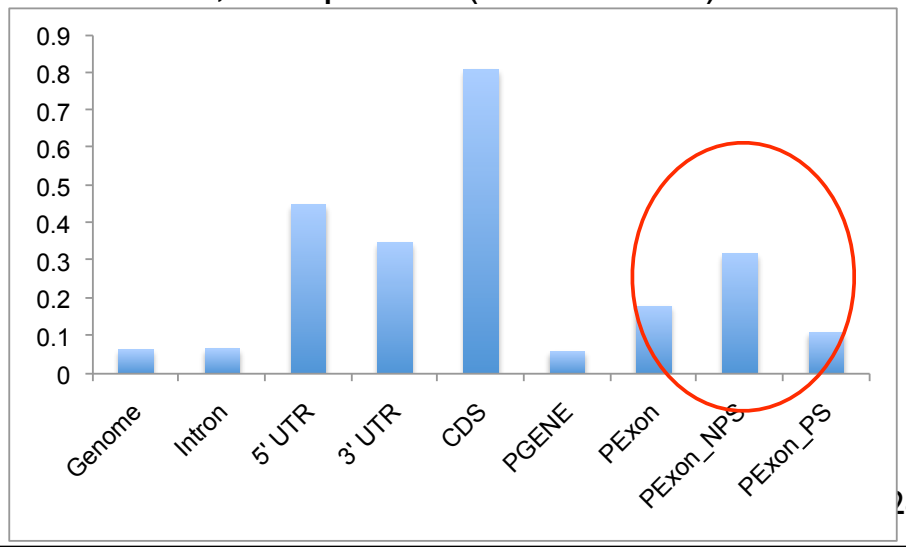


Constraint Regions from Different Alignments

EPO, 33 species (from ENCODE)



TBA, 36 species (from UCSC)



Difference of Species included in Alignment

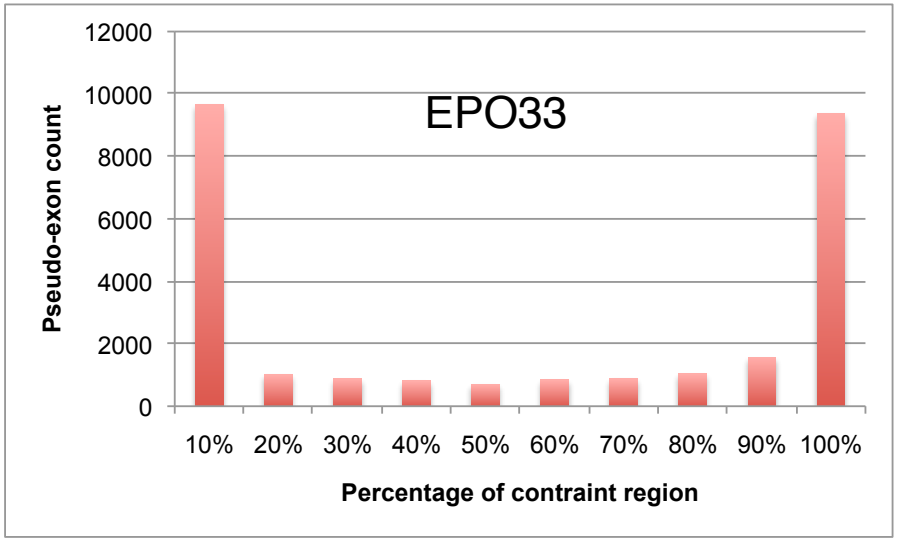
	ENCODE	UCSC
Boar	x	-
Baboon	-	x
Wallaby	-	x
Opossum	-	x
Platypus	-	x

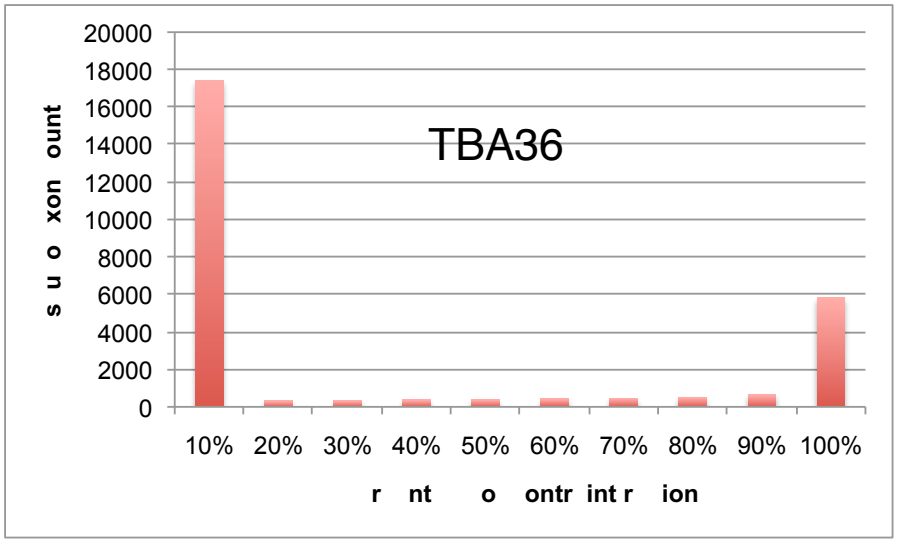
Difference of 2X mammals included in Alignment

	ENCODE	UCSC
Guinea Pig	x	-
Wallaby	-	x
Opossum	-	x
Platypus	-	x
Gorilla	-	x

There is a filter step in the TBA alignment: those for 2X mammalian genomes were filtered to retain only alignments of best quality in both the target and query ("reciprocal best")

Distribution of Constraint Regions in Pseudo-exons





Summary

- Pgenes are extracted from GENCODE 7 annotation file;
- Transcription evidence for pgenes are derived from GENCODE 7, BodyMap data and proteogenomics data;
- Pgenes with transcription evidence show more active chromatin state than those show no sign of transcription;
- A subset of pgenes with active chromatin state are identified
- Constraint regions in pgenes are analyzed
- 2X mammal data introduces noise into pseudogene conservation analysis

Future Directions

- Correlate pseudogene annotation with other ENCODE data, such as polymerase activity, expression data, etc.
- Identify a set of pseudogene with transcription evidence and possible regulatory roles
- Upstream TFBS and conservation analysis of parent genes and duplicated pseudogenes
- How is activity (such as transcription, histone marks, etc.) related to pseudogene duplication

Acknowledgement

- Suganthi
- Lukas H.
- Rachel (UCSC)
- Mark D. (UCSC)
- Annotation/Assembly
- Mark