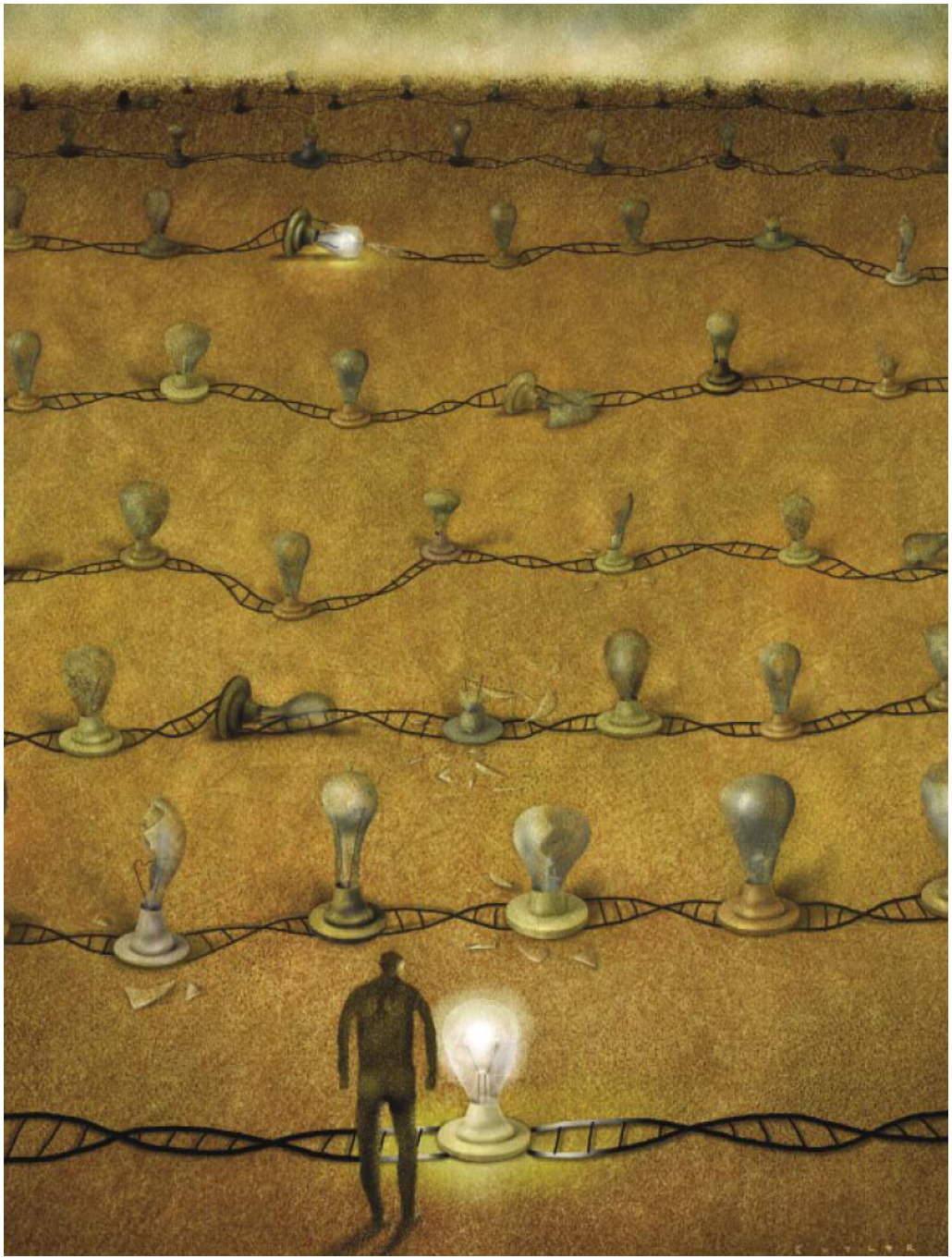


Yale Pseudogene Analysis as part of GENCODE Project

Sanger Center
2009.01.20, 11:20-11:40

Mark B Gerstein
Yale

Illustration from Gerstein & Zheng (2006). Sci Am.

Overview

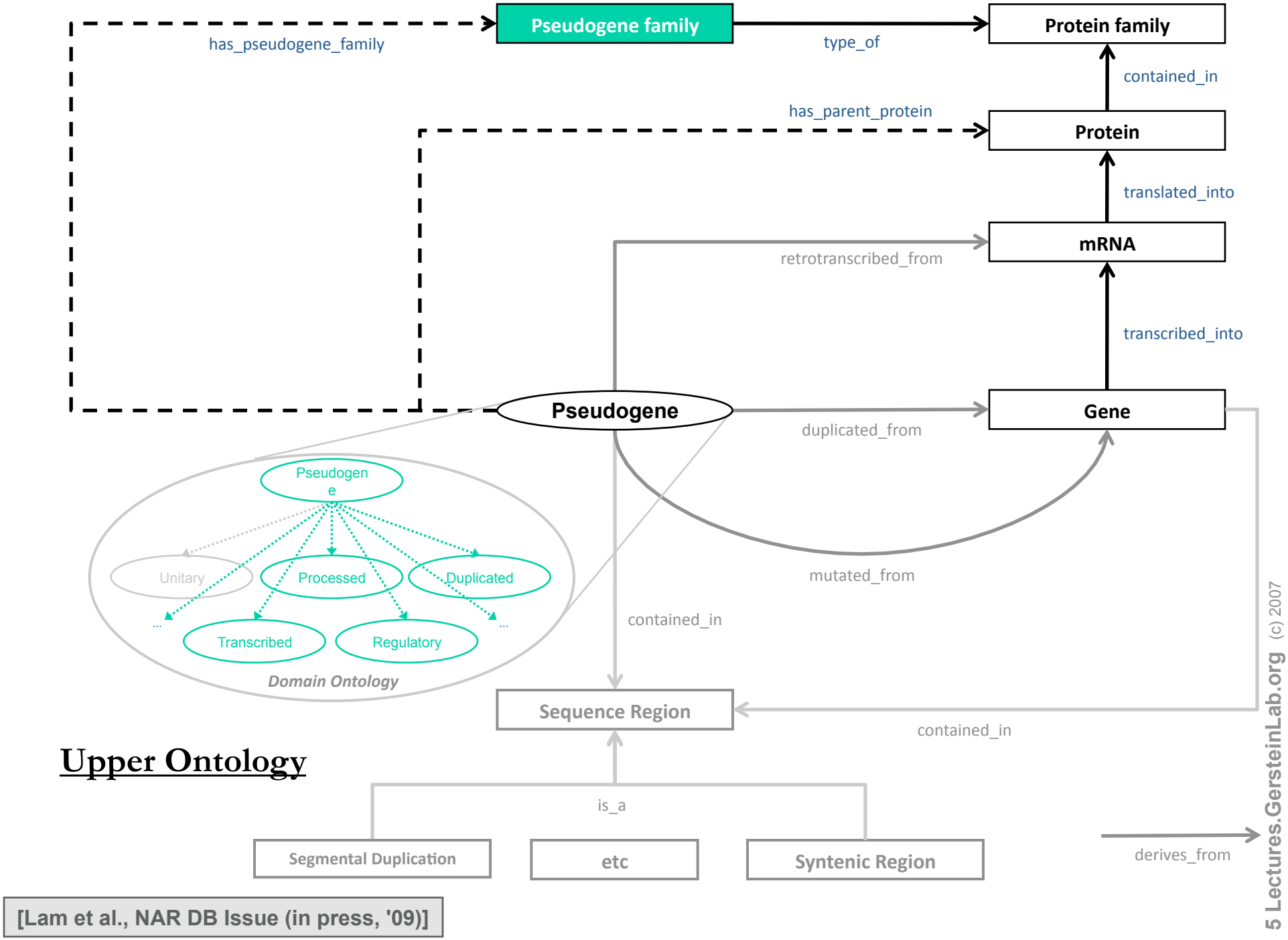
- Outline
  - ◊ Flow
  - ◊ Additional Pseudogene Annotation
  - ◊ Pseudogene Sets
- Regular conference calls
  - ◊ Sanger / Havana: Jenn, Adam
  - ◊ UCSC: Rachel, Mark
  - ◊ Discuss difficult pseudogene cases, pipeline improvements and additional pseudogene annotation
- Questions for you:
  - ◊ consis. Labels ?
  - ◊ Pfam/ensembl usage ?
  - ◊ trans. set ?
  - ◊ consensus set ?

Overall Flow: Pipeline Runs, Coherent Sets, Annotation, Transfer to Sanger

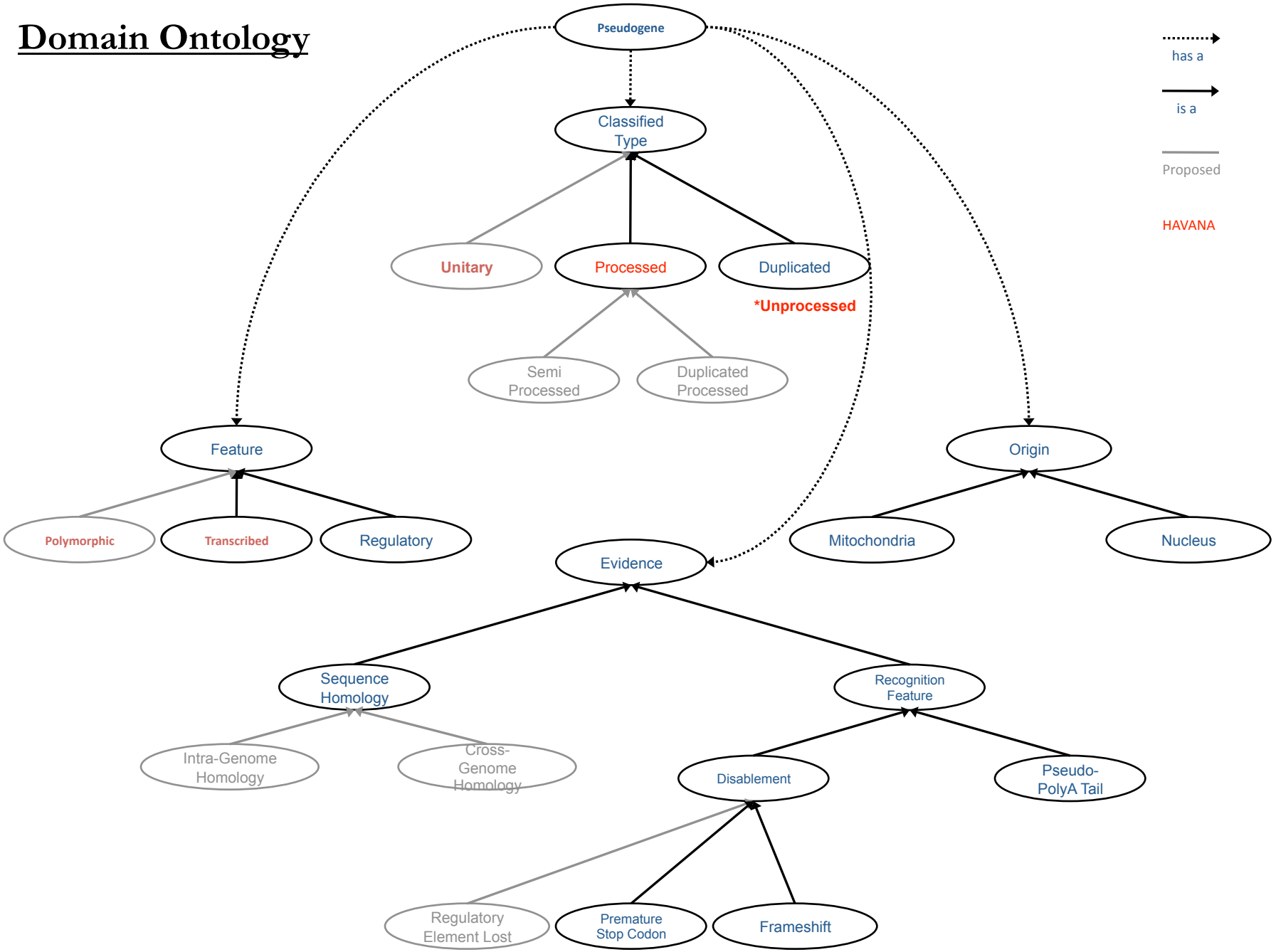
- Overall Approach
  1. Overall Pipeline runs at Yale and UCSC, yielding raw pseudogenes
  2. Extraction of coherent subsets for further analysis and annotation
  3. Passing to Sanger for detailed manual analysis and curation
  4. Incorporation into final GENCODE annotation
  5. Pipeline modification

- Chronology of Sets
  1. Unitary pseudogenes
  2. **Ribosomal Protein pseudogenes**
     (Hard then easy)
  3. Transcribed pseudogenes annotated earlier
  4. Strong overlap consensus pseudogenes
  5. Pseudogenes associated with SDs
  6. GAPDH pseudogenes

Additional Annotation on Pseudogenes: Consistent Labeling (Ontology)

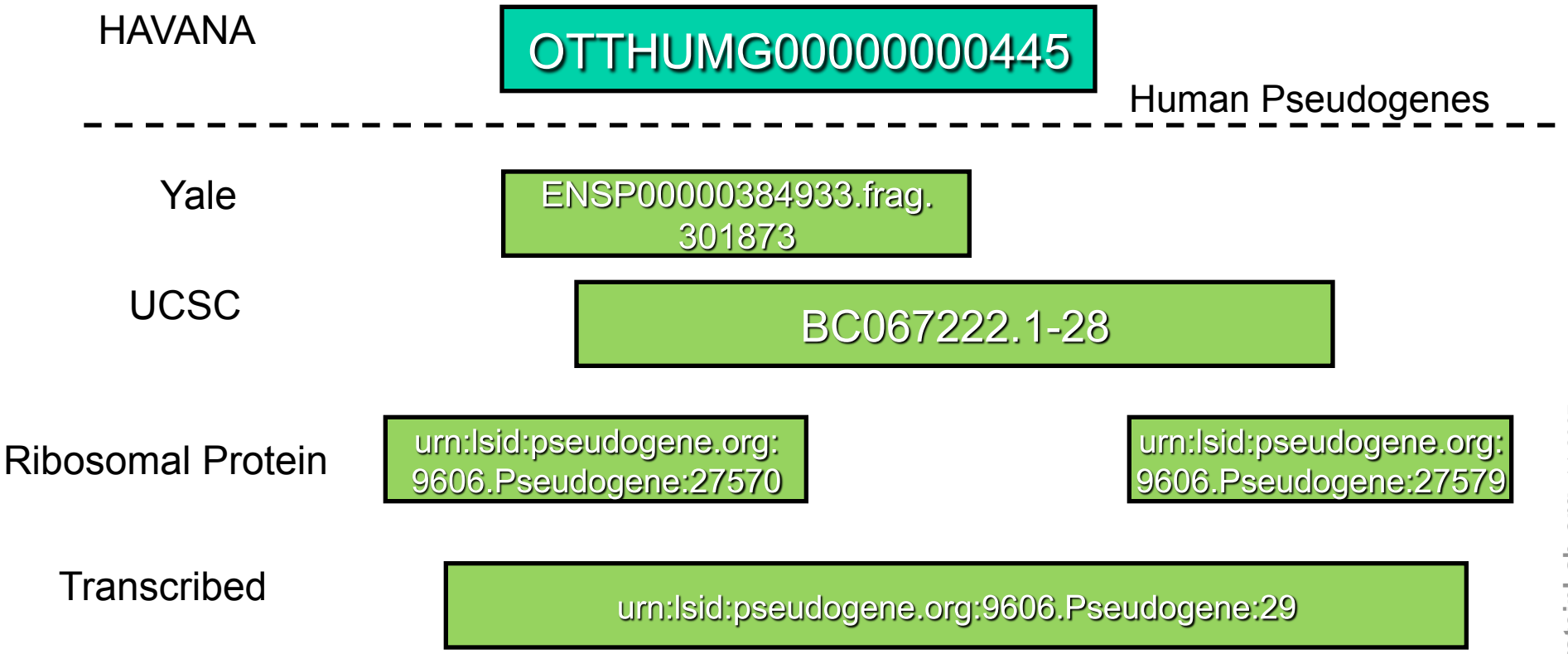

Pseudogene family
Protein family
has_pseudogene_family
type_of
contained_in
has_parent_protein
Protein
translated_into
mRNA
retrotranscribed_from
transcribed_into
Pseudogene
Gene
duplicated_from
Pseudogene
Unitary
Processed
Duplicated
Transcribed
Regulatory
Domain Ontology
mutated_from
contained_in
Sequence Region
contained_in
Upper Ontology
is_a
Segmental Duplication
etc
Syntenic Region
derives_from
[Lam et al., NAR DB Issue (in press, '09)]
5 Lectures.GersteinLab.org (c) 2007

Domain Ontology



[Lam et al., NAR DB Issue (in press, '09)]

Collections to provide further annotation



Additional Annotation on Pseudogenes: Families & Histories

SVN

Flat Files ⟷ DAS

Table Browser

single pgene query

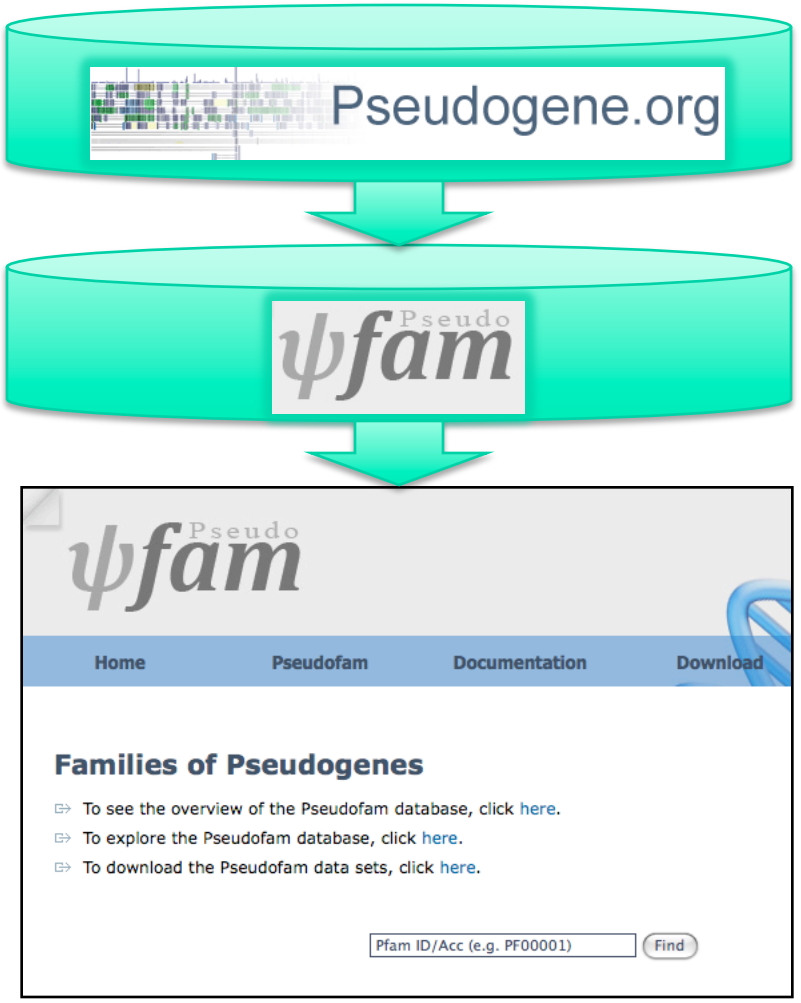
tables.pseudogene.org ⟷ UCSC Genome Browser

- 12 eukaryotic species
  - Human, mouse, rat, chimp…
  - 100,052 pseudogenes
- 64 prokaryotic species
  - 6,412 pseudogenes
- 28,237 human pseudogenes
  - 13+ unique human sets

[Karro et al., *NAR* ('07)]

Pseudofam Database

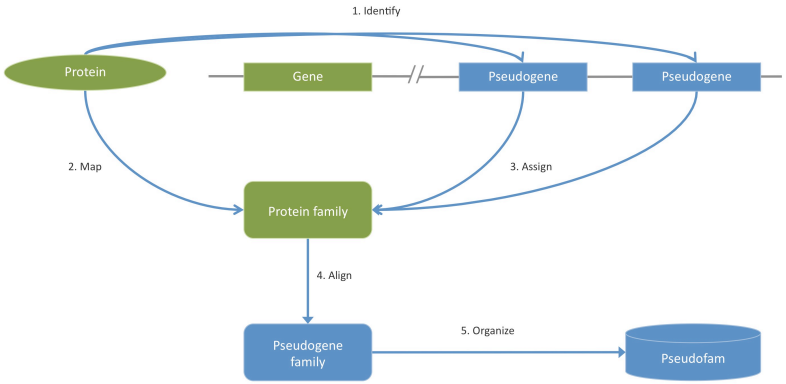
- **Data Sources**
  - ◊ Ensembl, Pfam, BioMart
  - ◊ Pseudopipe
- **Highlighted Features**
  - ◊ Browse families
    - Statistics
      - Enrichment
      - P-value, etc
  - ◊ Search families
    - Pfam ID/Acc
    - Ensembl ID
    - Pseudogene ID
  - ◊ Correlate families
    - Genes
    - Pseudogenes
    - Parents

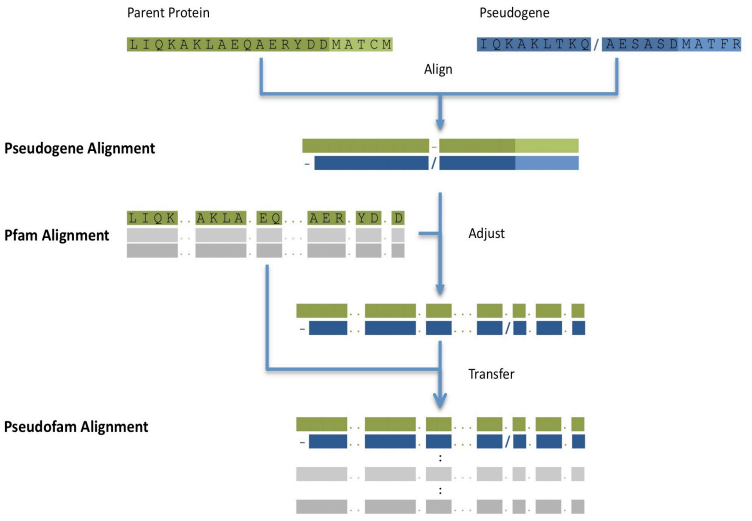


[Lam et al., NAR DB Issue (in press, '09)]

Pseudofam Construction

- **Data Generation**
  - ◊ Identify pseudogenes by proteins
  - ◊ Map parent proteins to protein families
  - ◊ Assign pseudogenes to their parent families
  - ◊ Align the pseudogenes in pseudogene families
  - ◊ Calculate the key statistics and organize the data into database
- **Alignment**
  - ◊ Align pseudogene to parent
  - ◊ Transfer alignment from Pfam
  - ◊ Combine and adjust the alignments to build the pseudofam alignment



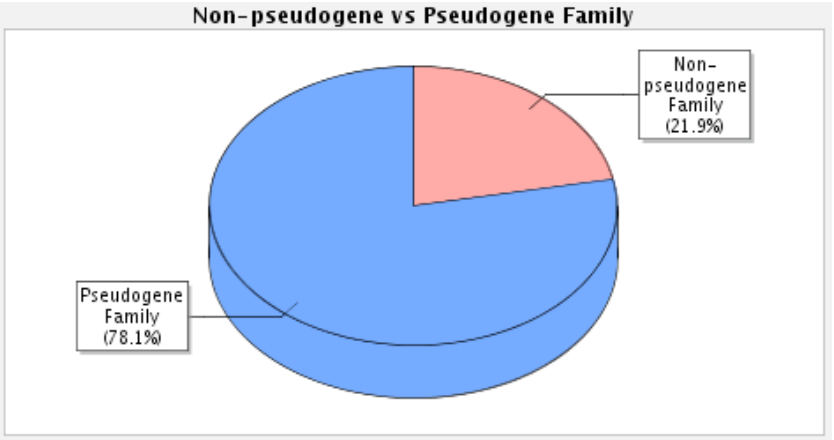


[Lam et al., NAR DB Issue (in press, '09)]

Pseudofam Statistics: Enrichment of pseudogenes within a family ("Living vs Dead")

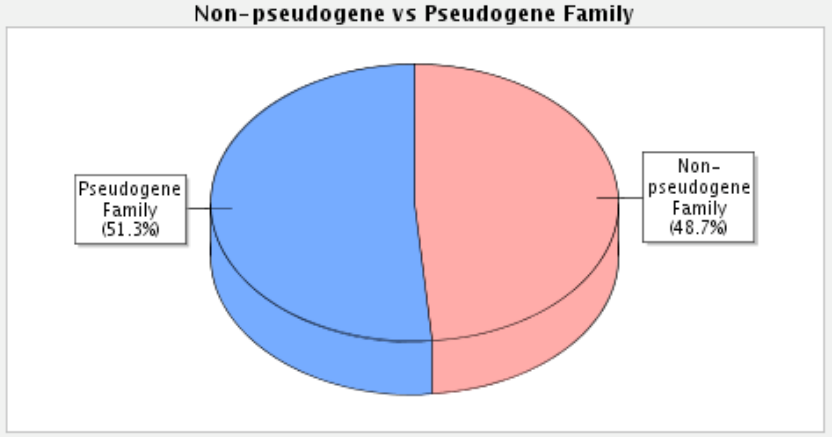
Total (10 Eukaryotes)

Protein Families:	3,820
Pseudogene Families:	2,985
Total Genes:	219,662
Total Parents:	26,679
Total Pseudogenes:	102,679
Pseudogene-to-gene Ratio:	0.47
Pseudogene-to-parent Ratio:	3.85
Parent-to-gene Ratio:	0.12



Human

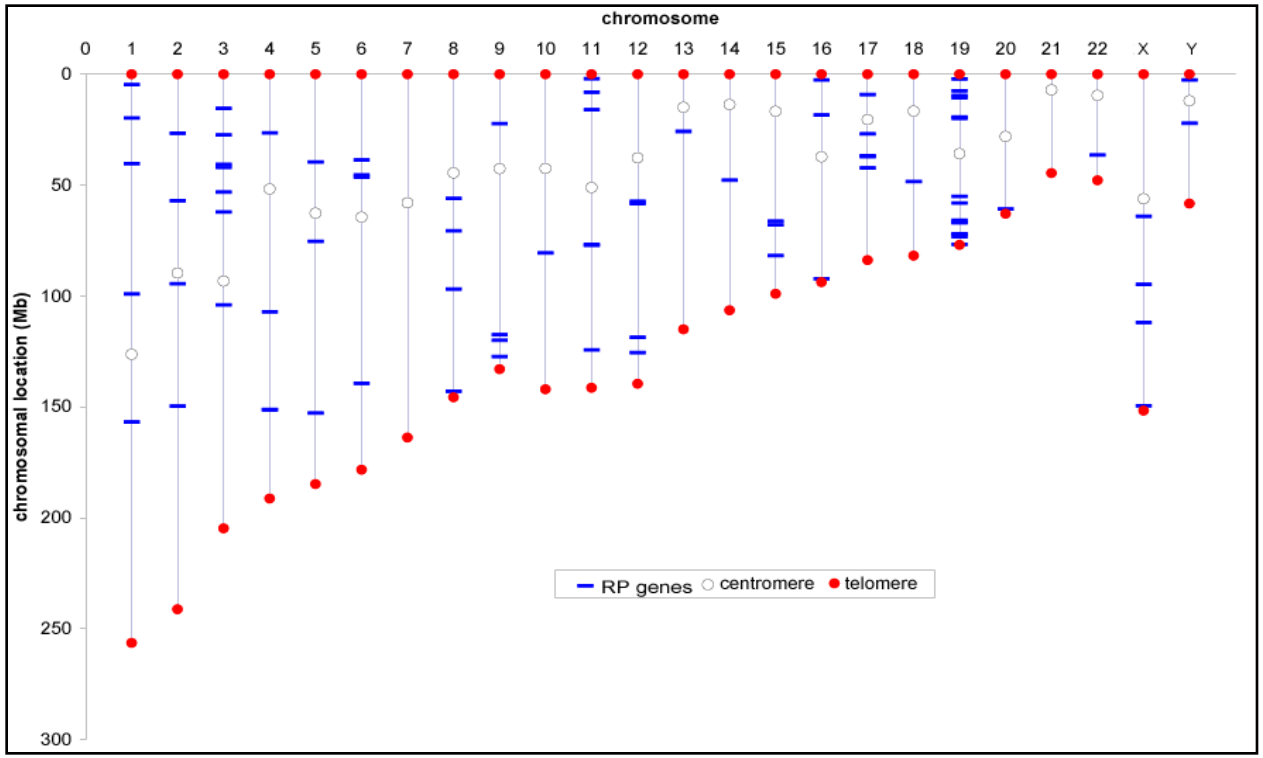
Protein Families:	3,486
Pseudogene Families:	1,790
Total Genes:	34,686
Total Parents:	4,218
Total Pseudogenes:	12,534
Pseudogene-to-gene Ratio:	0.36
Pseudogene-to-parent Ratio:	2.97
Parent-to-gene Ratio:	0.12



[Lam et al., NAR DB Issue (in press, '09)]

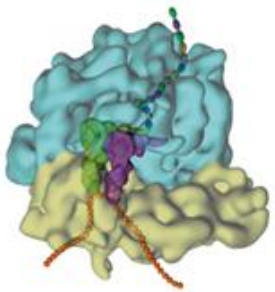
Pseudogene Set #2: RP pseudogenes

Human Ribosomal Proteins (RP)



79 Functional RP genes

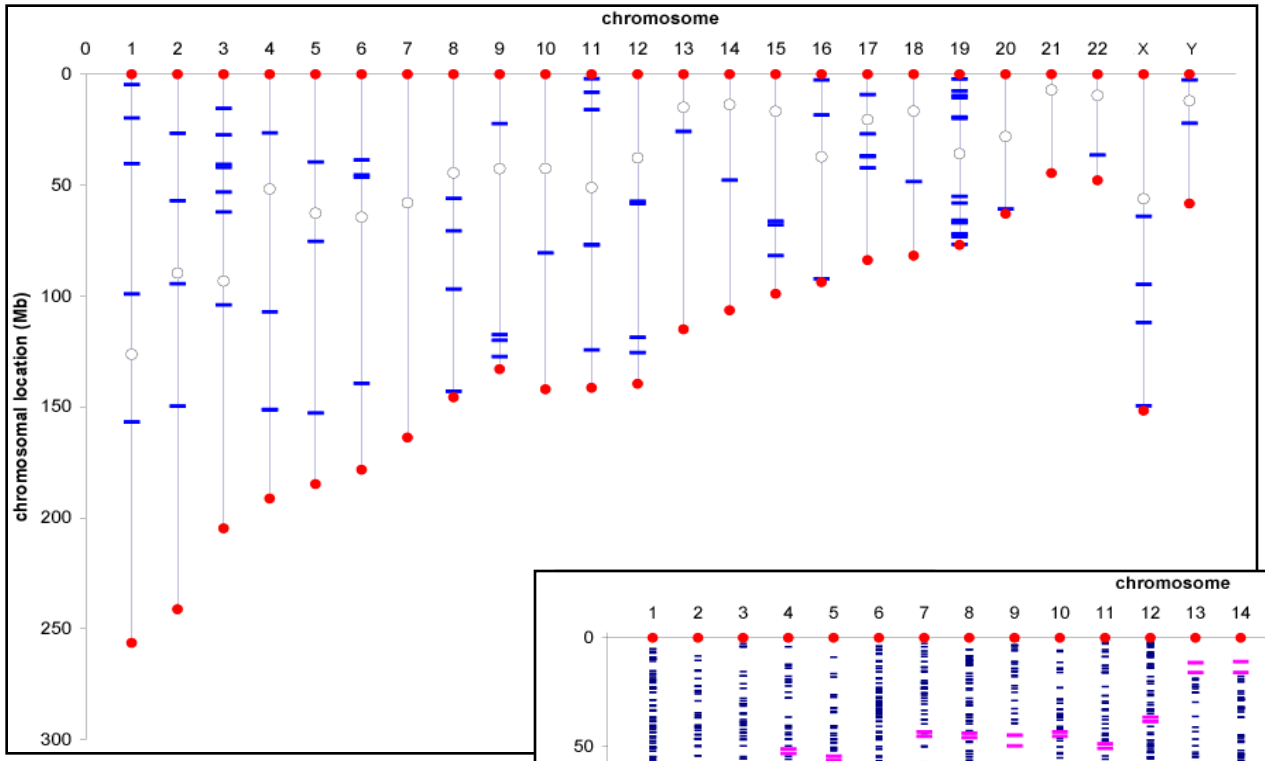
Nakao A, Yoshihama M, Kenmochi N: RPG: the Ribosomal Protein Gene database. *Nucleic Acids Res 2004, 32:D168-170.*

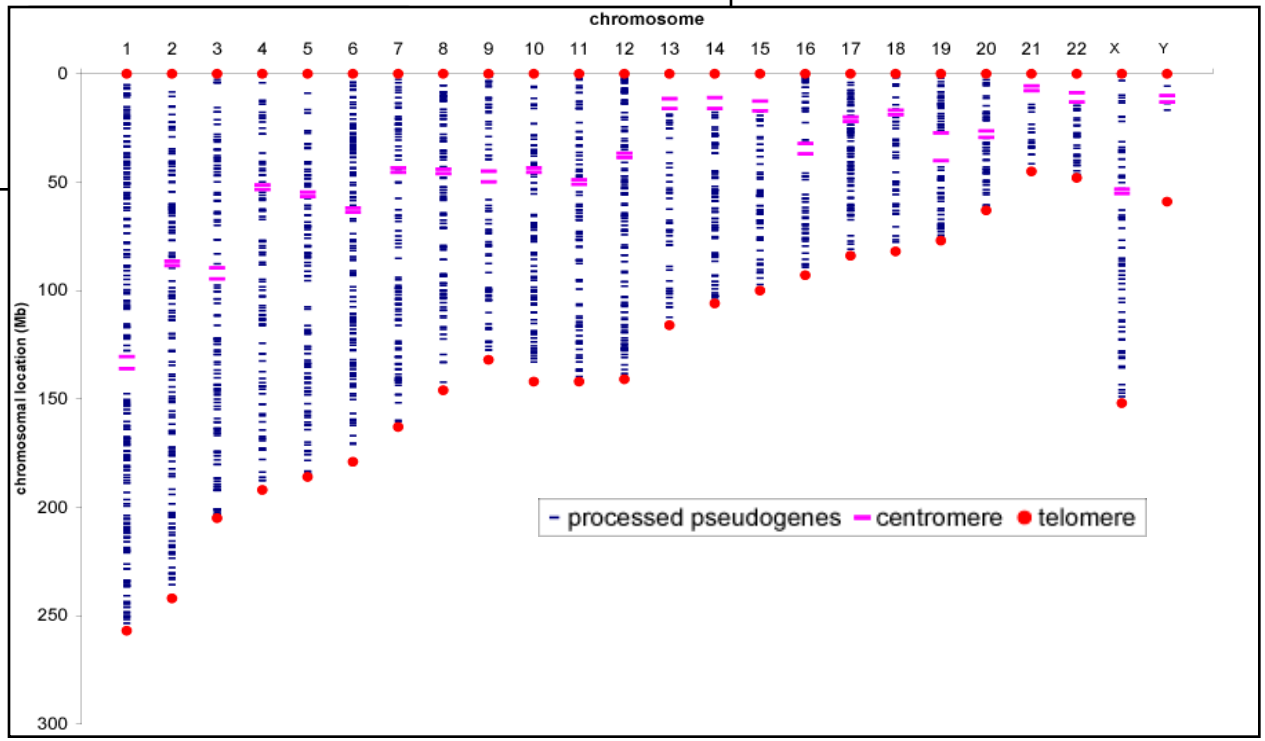


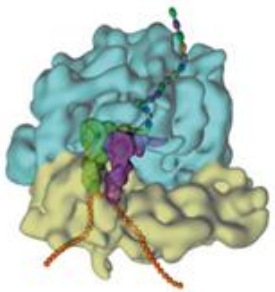
[Zhang *et al*. *Genome Research*, 2002]

Human Ribosomal Proteins (RP)

79 Functional RP genes







~ 2000 RP Ψgenes

[Zhang *et al*. *Genome Research*, 2002]

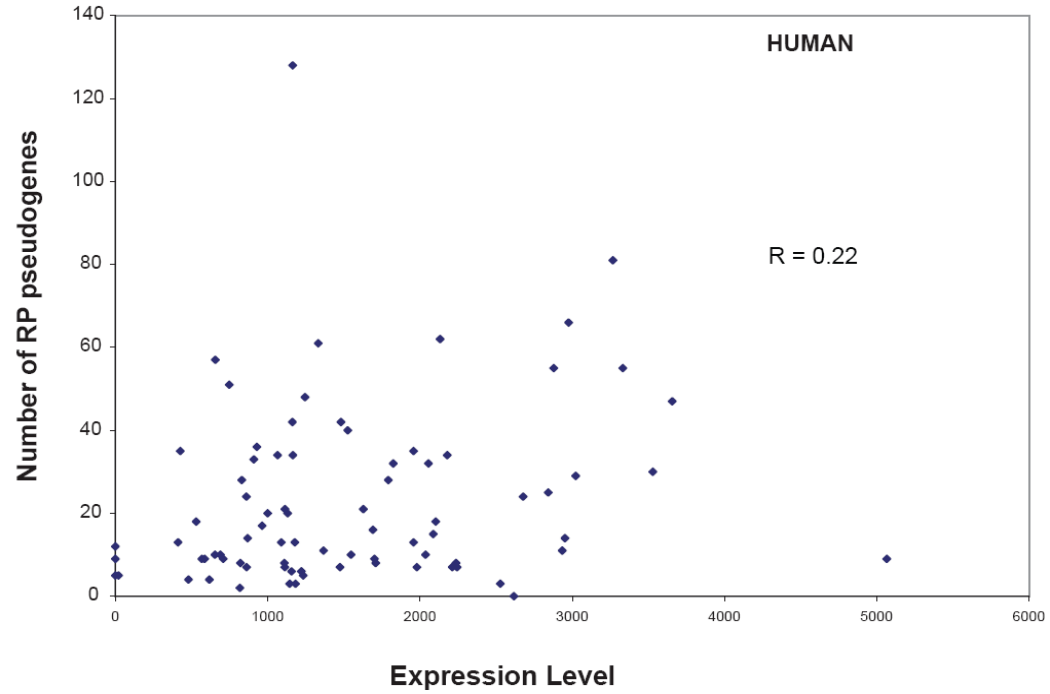
Number of RP pseudogenes

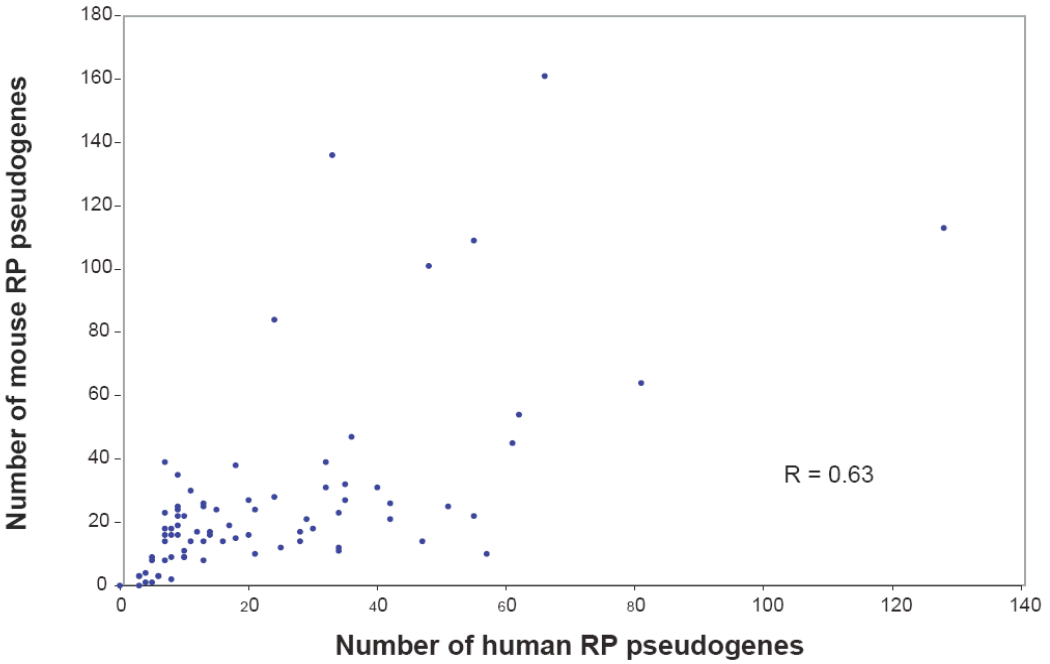
(identified by pipeline)

Organism	Processed	Fragments	Low confidence
Human	1822	218	212
Chimp	1462	219	160
Mouse	2092	326	413
Rat	2848	343	450

RP pseudogenes constitute the largest family of pseudogenes. Almost all are processed: There are ~90 clearly duplicated ones in the human genome

[Balasubramanian et al., *Genome Biol.* ('09)]

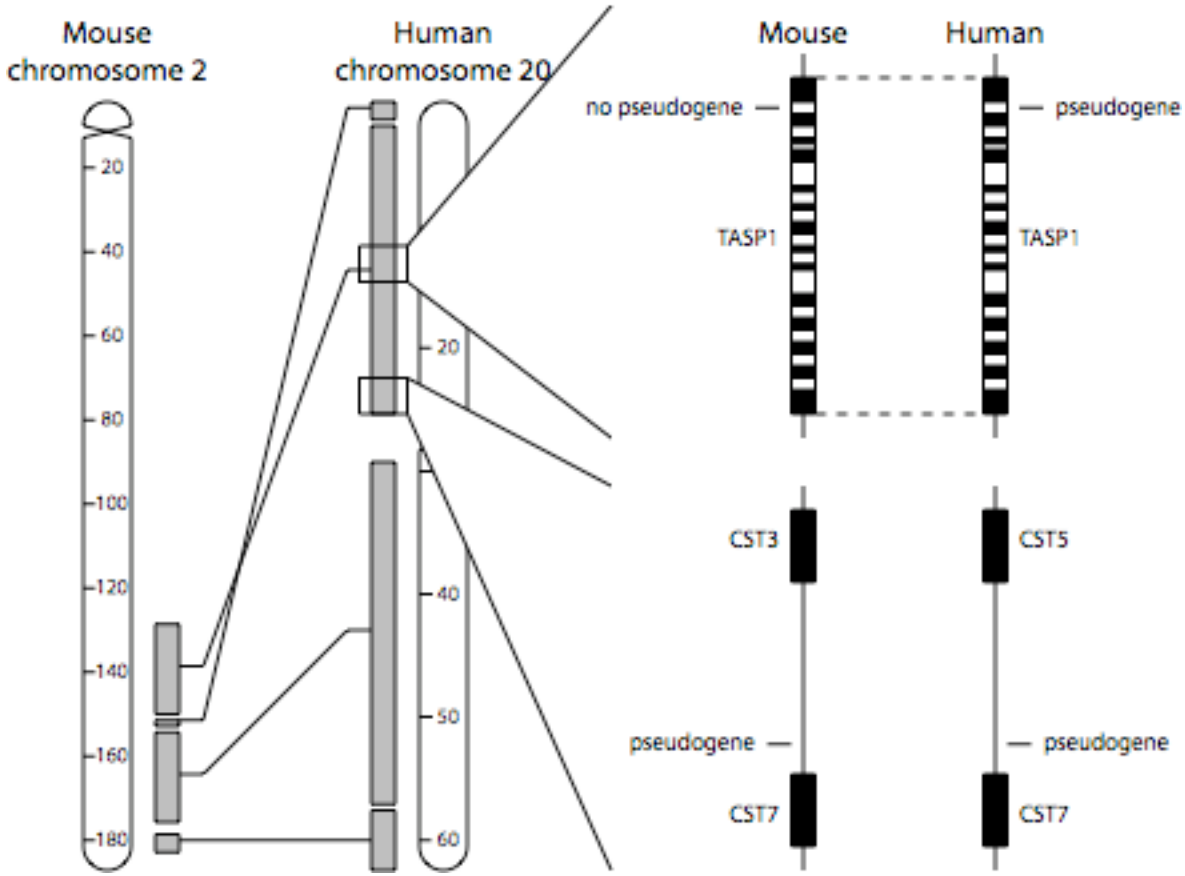




Number of each type of human ribosomal protein processed pseudogenes appears unrelated to expression level or to number in mouse

[Balasubramanian et al., *Genome Biol.* ('09)]

Using Synteny to Improve Annotation of RP pgenes



Synteny derived based on local gene orthology

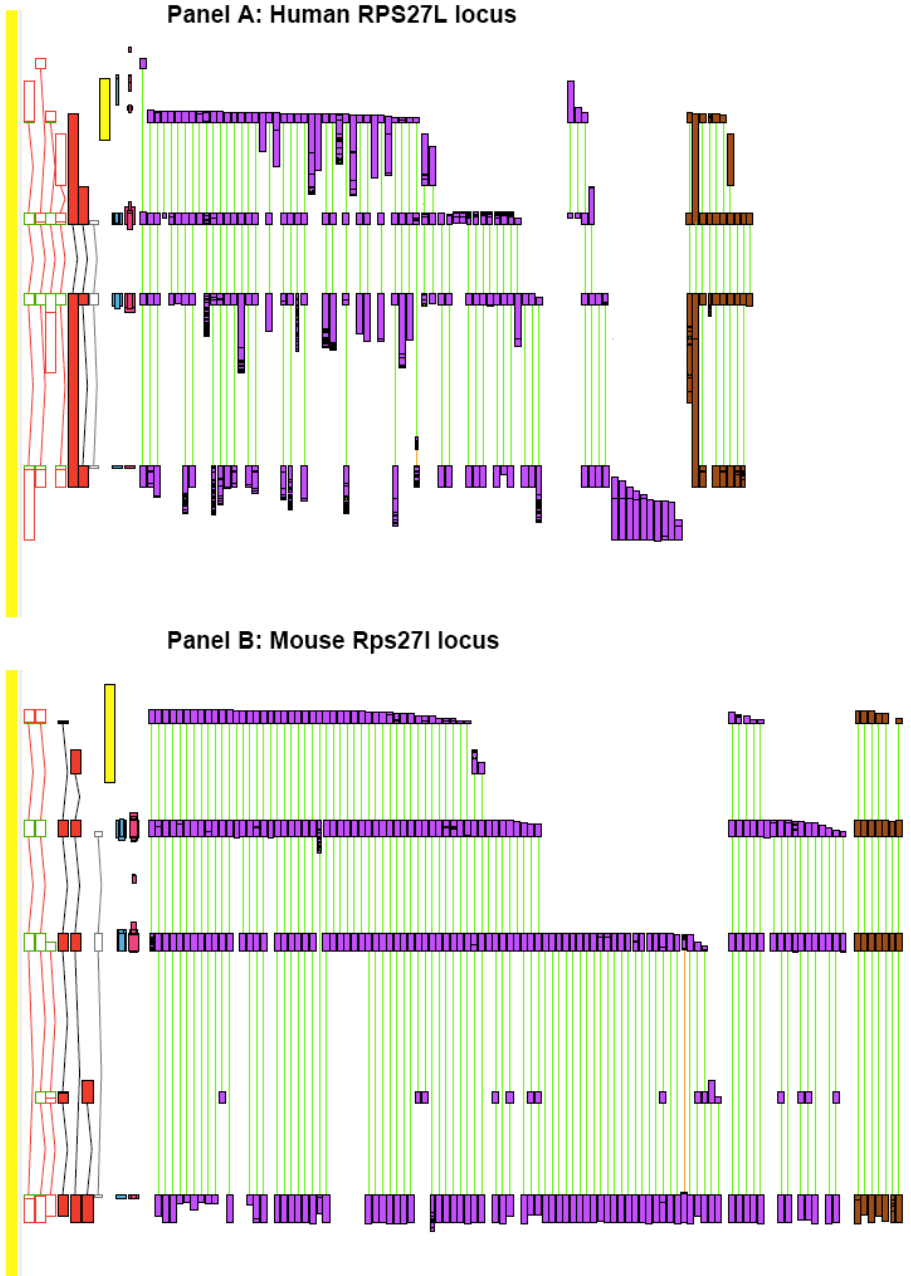
[Balasubramanian et al., *Genome Biol.* ('09)]

Syntenic proc pseudogenes

Species1- Species2	Number of syntenic pgenes
Human-chimp	1282
Human-mouse	6
Human-rat	11
Rat-mouse	394

[Balasubramanian et al., *Genome Biol.* ('09)]

A human-mouse syntenic pgene, likely to be coding



[Balasubramanian et al., *Genome Biol.* ('09)]

Nakao A, Yoshihama M, Kenmochi N: RPG: the Ribosomal Protein Gene database. *Nucleic Acids Res 2004, 32:D168-170.*

Set #3: Transcribed Pseudogenes Annotated Earlier

Harrison '05 set of Transcribed Pseudogenes

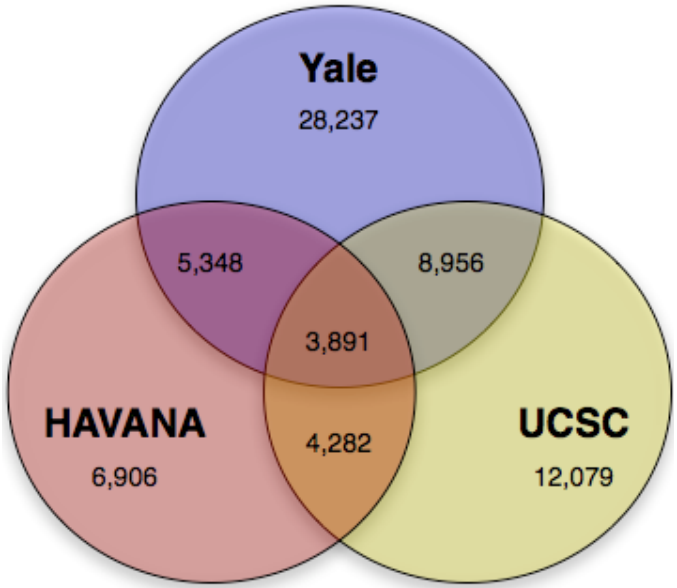
- **233** Transcribed from ~8000 Processed Pseudogenes
- Evidence for Transcription
  - ◊ 8% Refseq mRNAs
  - ◊ 32% Unigene consensus sequences
  - ◊ 72% dbEST expressed sequence tags
  - ◊ 32% Oligonucleotide microarray data (extra support)
- Highly decayed
  - ◊ Fraction with Ka/Ks ≥ 0.5 is 54%

Harrison et al. (2005) NAR

Set #4: Strong Consensus Pseudogenes

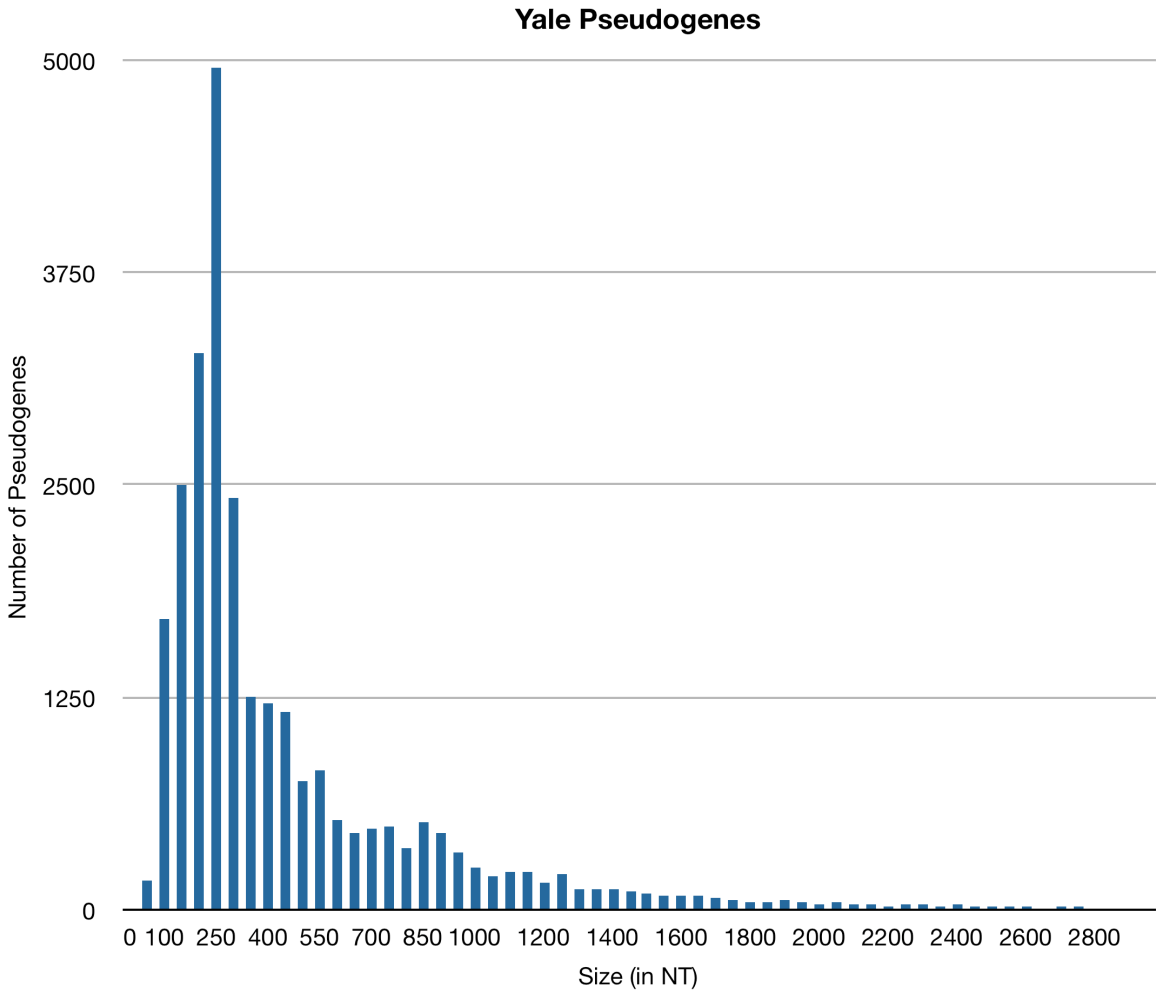
Consensus Pseudogenes

- Yale & UCSC agreement
- Yale, UCSC & HAVANA agreement
- HAVANA disagreement

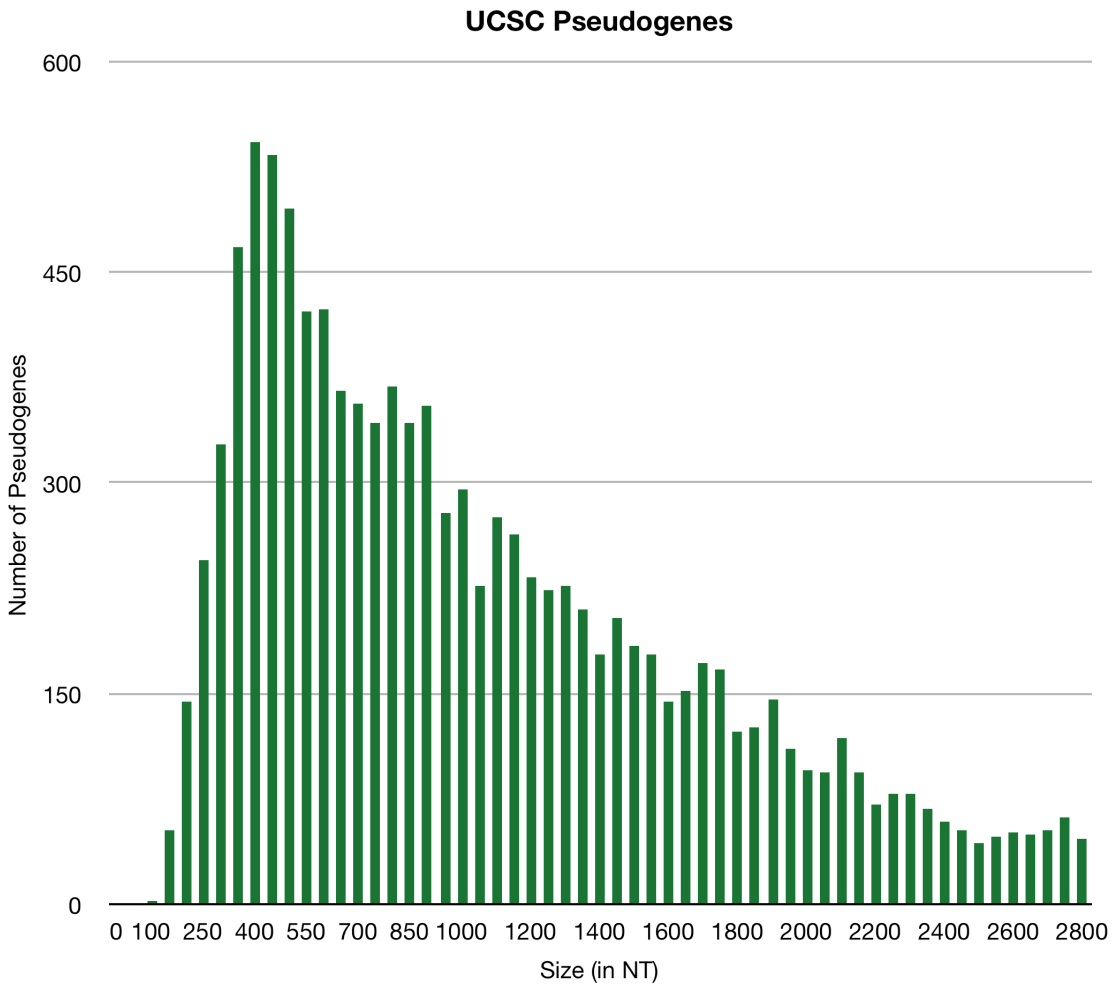


Overlap

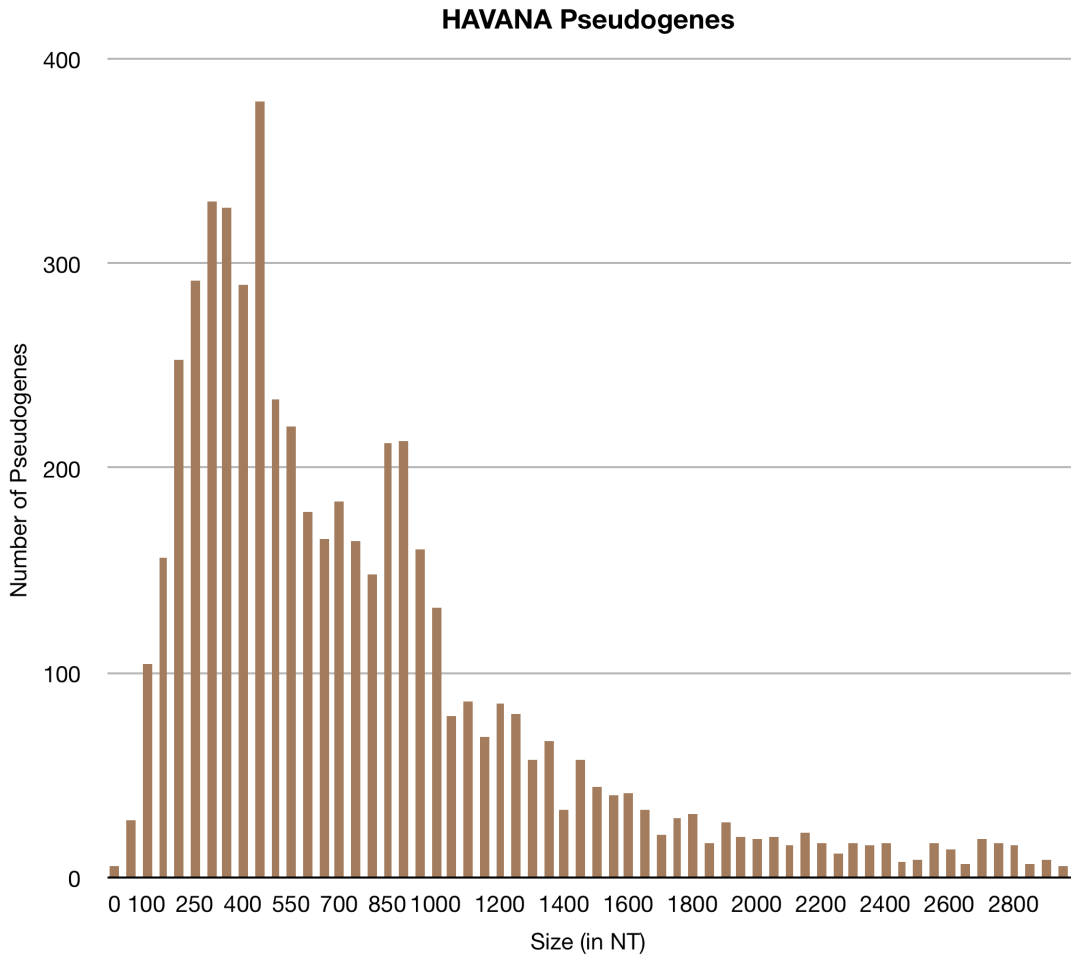
- Initially, 50 nucleotide overlap
- Why so low?
- What about percentage overlap?



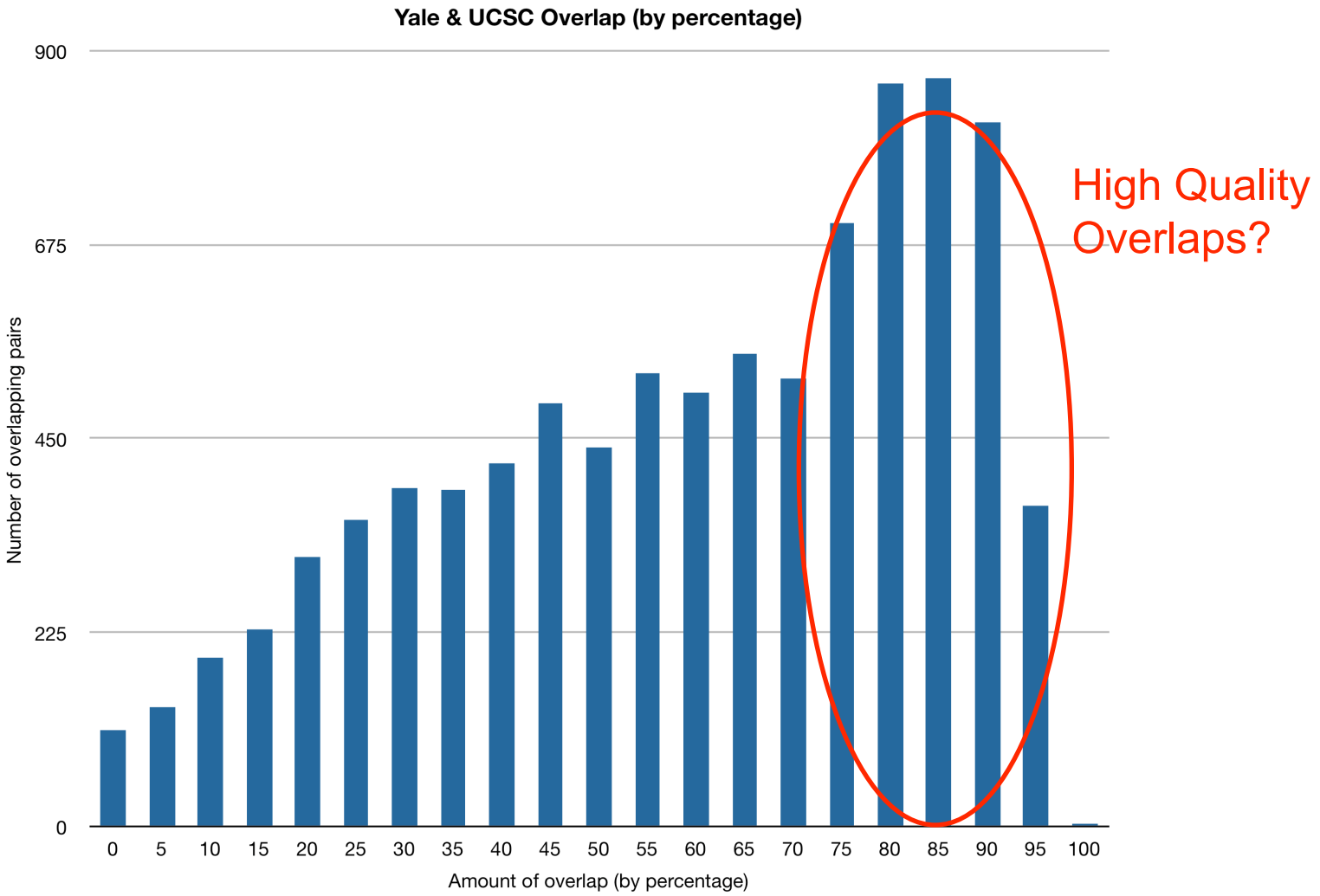
AVG	
	1261
MAX	
	181881
MIN	
	35
MEDIAN	
	322

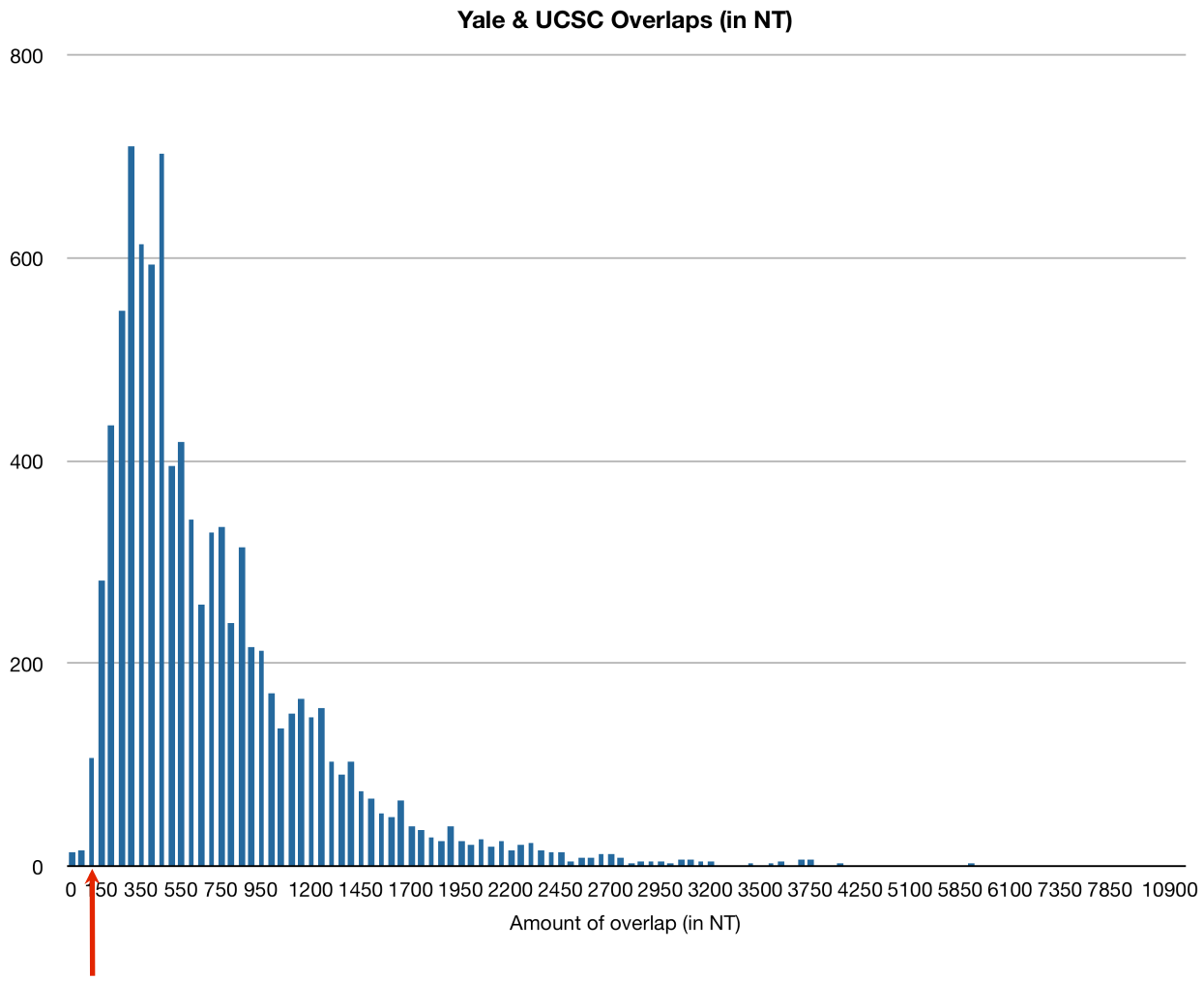


AVG
1349
MAX
35931
MIN
120
MEDIAN
991



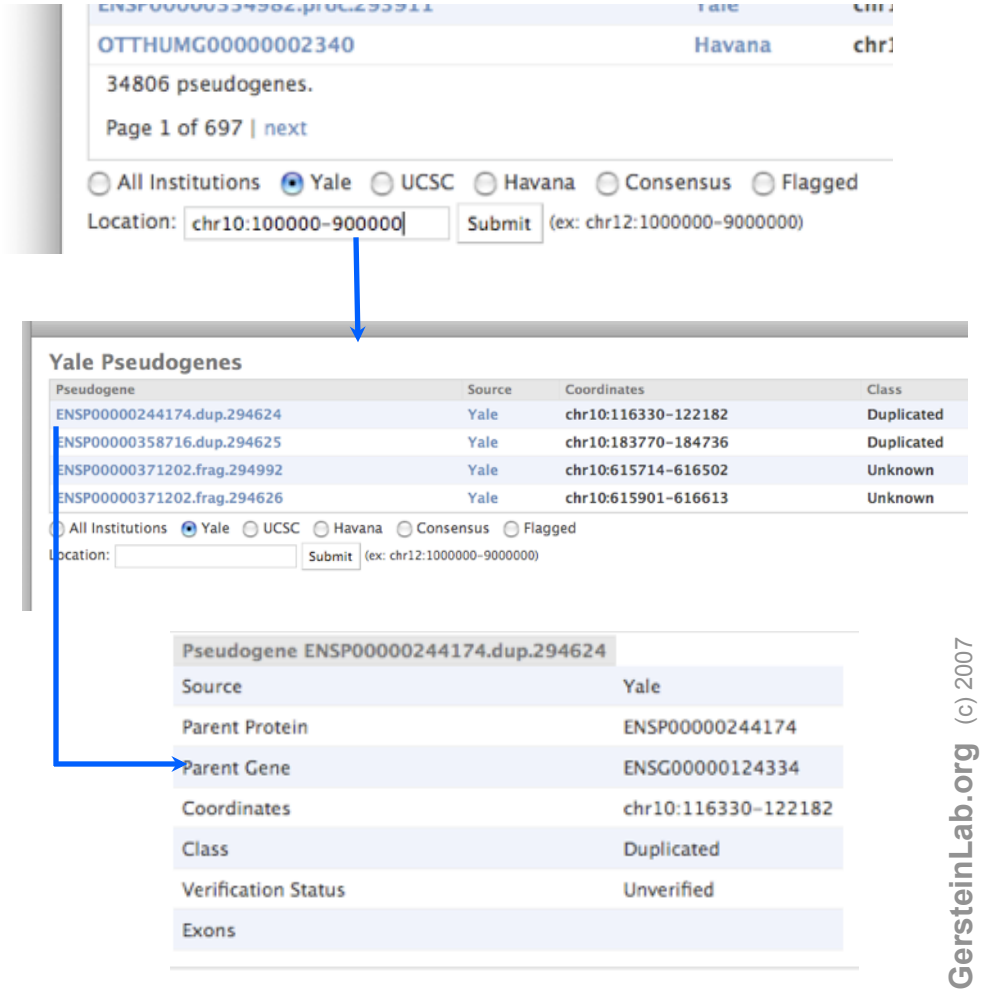
AVG	2191
MAX	189503
MIN	27
MEDIAN	706





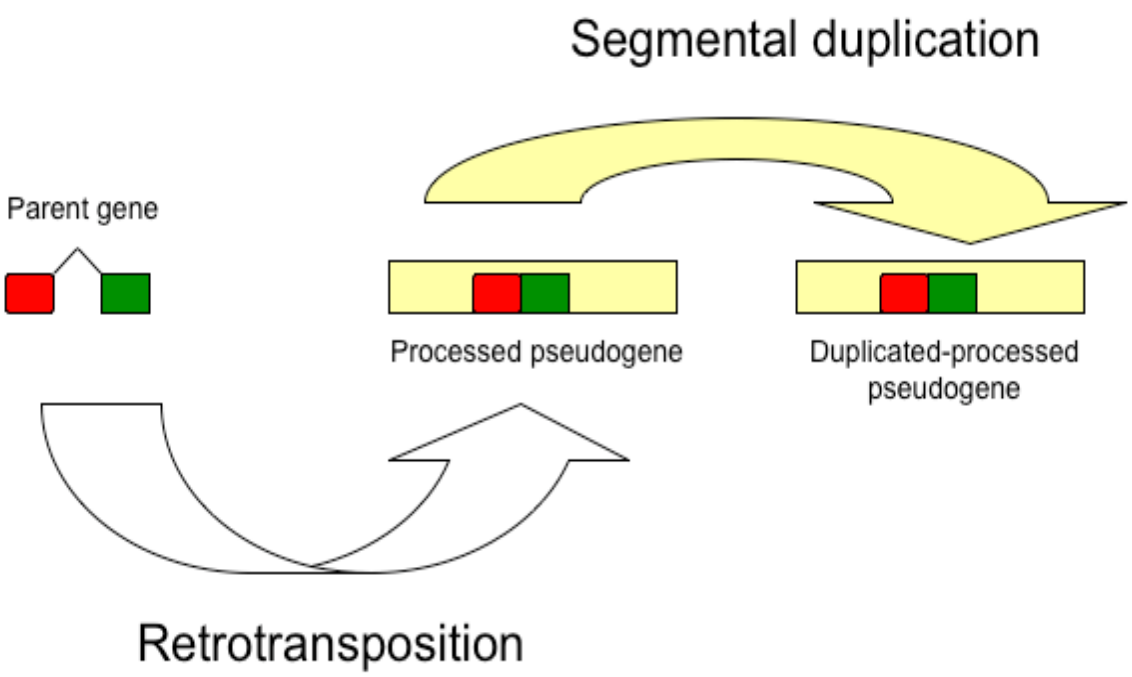
gencode.gersteinlab.org/consensus

- Django-based
- Filter by genomic coordinates, source
- Yale & UCSC Consensus track
- Flagged pseudogenes
- Verification status



Set #5: SD pseudogenes

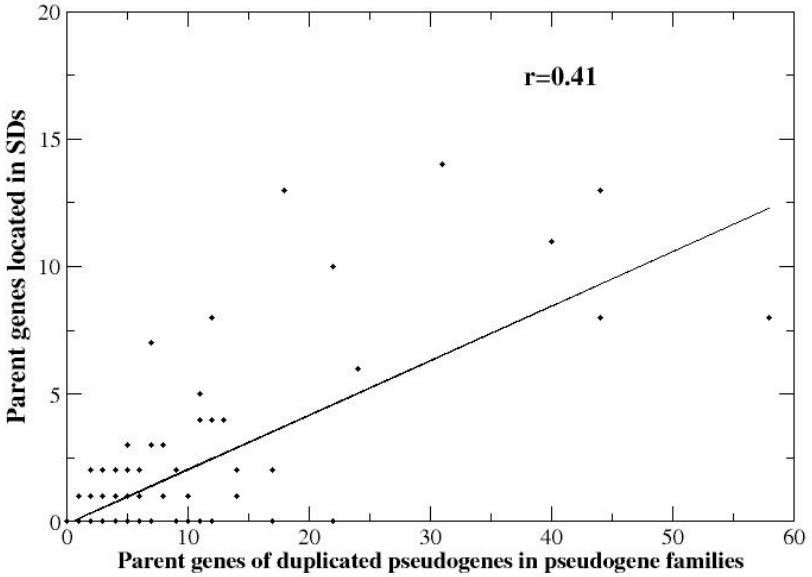
Segmental Duplications (SDs)

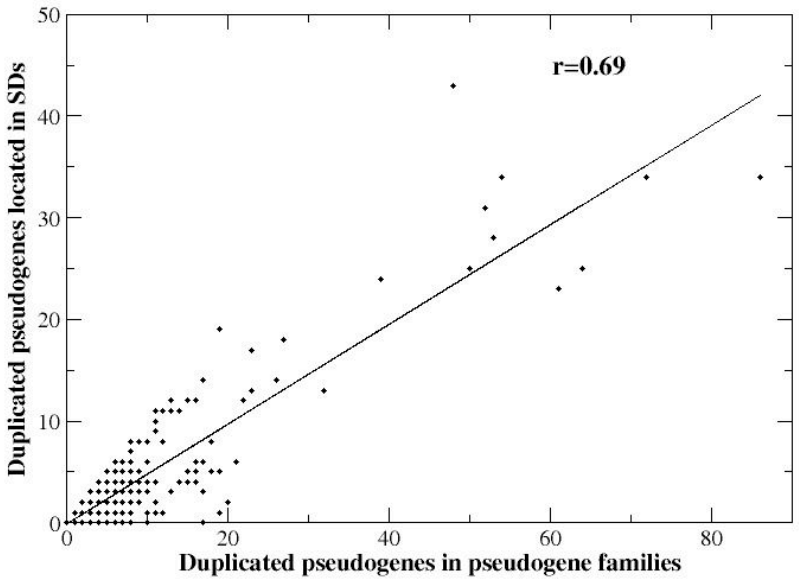


SDs are regions of the genome with ≥ 90% sequence identity and ≥ 1kb in length

Clues about pseudogene formation from pseudogenes located in SDs, for example, annotation of "duplicated-processed" pseudogenes

Pseudogene families and Segmental Duplications (SDs)





- SDs comprise ~5-6% of the human genome but contain ~17.8% genes, 45.7% duplicated pgenes and 21.6% processed pgenes
- Relative values of correlation coefficients in the plots above consistent with the observation that SDs contain more pgenes than parent genes

[Lam et al., NAR DB Issue (in press, '09)]

Summary

- Sets
  - ◊ RP, transcribed ‘05, consensus, SD....
- Additional Annotation
  - ◊ families & labels

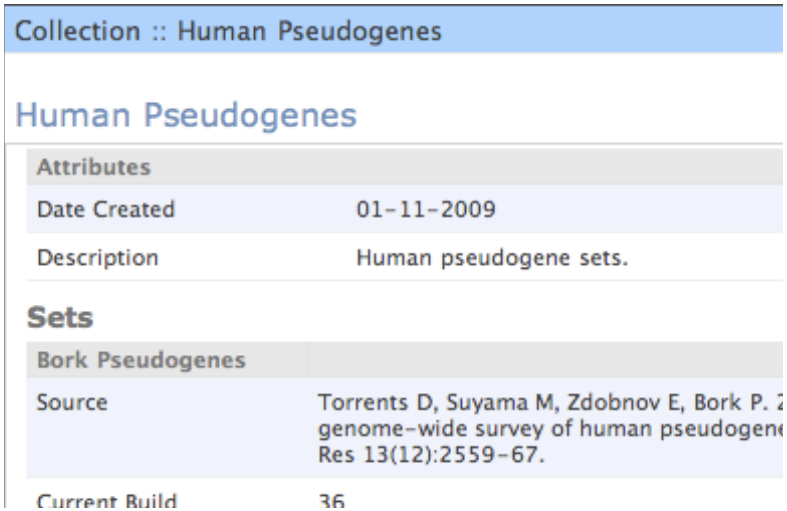
Credits

- Yale: **S Balasubramanian, P Cayting, H Lam,** Y Liu, G Fang, N Carriero, R Robilotto, E Khurana
- Alums: D Zheng, P Harrison, Z Zhang
- Sanger: A Frankish, J Harrow
- UCSC: R Harte, M Diekhans

Extra

Collections

- Combination of multiple sets
  - ◊ Human50, UCSC, GAPDH, Transcribed, Unitary, etc.
- Also can have attributes
- View by set or as a whole



Pseudofam & Pseudogene.org Integration

PseudoPipe

```
ENSP00000217182.dup.270258      11      92287417        92289417        +
        ENSP00000217182 317     463     ENSG00000101210 0.68    4       5
        3       0       0.72    1       Yes     [(92287417, 92287559),
(92287728, 92287905), (92288378, 92288668), (92288749, 92288907),
(92289239, 92289417)]   [(92287560, 92287727), (92287906, 92288377),
(92288669, 92288748), (92288908, 92289238)]     Duplicated
```

svn

annotations
via shared ID

gencode.gersteinlab.org/consensus

Select pseudogene

Pseudogene	Source	Coordinates	Class	Verifi
OTTHUMG00000000961	Havana	chr1:1873+3533	Duplicated	Unver
ENSP00000339557.dup.293897	Yale	chr1:2828+3241	Duplicated	Unver
OTTHUMG00000000958	Havana	chr1:4267-19433	Duplicated	Unver
ENSP00000368792.proc.293898	Yale	chr1:42381+43196	Processed	Unver
OTTHUMG00000001095	Havana	chr1:52811+53750	Duplicated	Unver
ENSP00000279067.dup.293236	Yale	chr1:118892-123443	Duplicated	Unver
NM_015125.3-1	UCSC	chr1:120932+124699	Processed	Unver
OTTHUMG00000001257	Havana	chr1:120967+123786	Processed	Unver
ENSP00000160740.proc.293899	Yale	chr1:120988+125486	Processed	Unver
BC054485.1-1	UCSC	chr1:124808-126924	Processed	Unver
OTTHUMG00000002478	Havana	chr1:125110-127902	Processed	Unver
ENSP00000372234.dup.293237	Yale	chr1:125578-127574	Duplicated	Unver
ENSP00000302684.proc.293238	Yale	chr1:128719-129038	Processed	Unver
AL137655.1-7	UCSC	chr1:129958-130616	Processed	Unver
BC038571.1-2	UCSC	chr1:148114-148539	Processed	Unver
ENSP00000322542.frag.293239	Yale	chr1:154551-154655	Unknown	Unver
BC014459.1-1	UCSC	chr1:218129-218641	Processed	Unver
OTTHUMG00000002552	Havana	chr1:218182-218638	Processed	Unver
ENSP00000259951.frag.293900	Yale	chr1:218384+218624	Unknown	Unver
ENSP00000307202.frag.293240	Yale	chr1:224218-224322	Unknown	Unver
ENSP00000306241.frag.293241	Yale	chr1:257004-257114	Unknown	Unver
BC018677.1-2	UCSC	chr1:311301-311887	Processed	Unver
ENSP00000341218.proc.293901	Yale	chr1:315061+315508	Processed	Unver