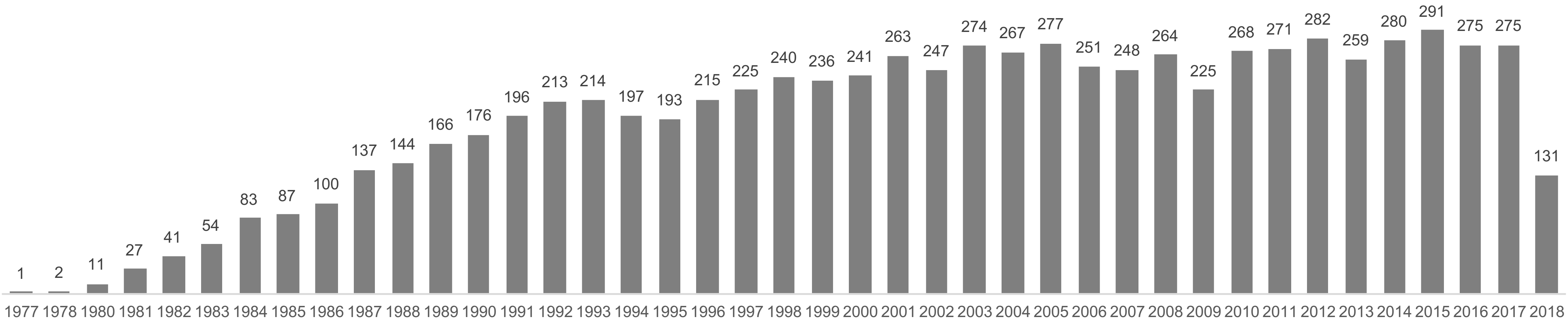
Updates on Pseudogene Analysis in Human and Mouse

Mark Gerstein & Paul Muir

GENCODE meeting

21st June 2018



1977

A Pseudogene Structure in 5S DNA of Xenopus laevis
C. Jacq, J. R. Millier and G.G. Browniae

Pseudogene "has homologous structure, [is] nearly as long as, and almost an exact repeat of, the gene itself"

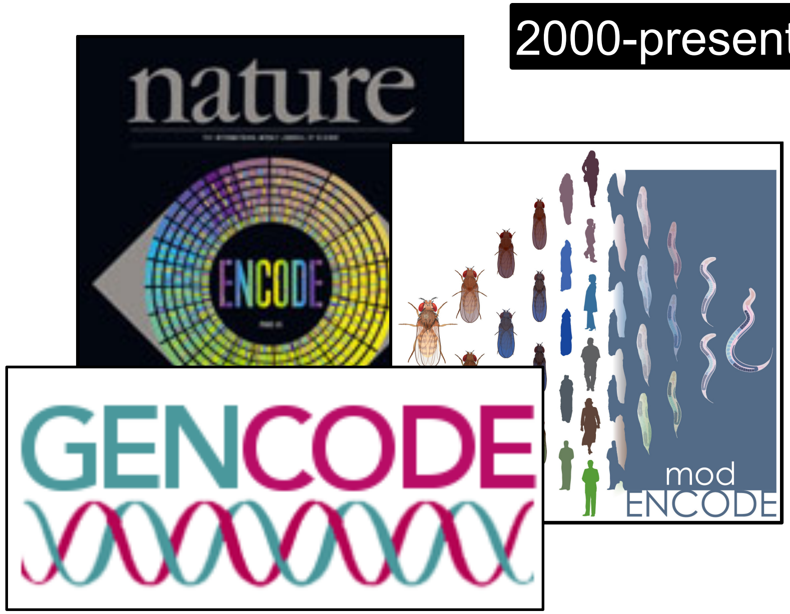
1980s

Pseudogenes' **structure** & **formation mechanisms**

1990s

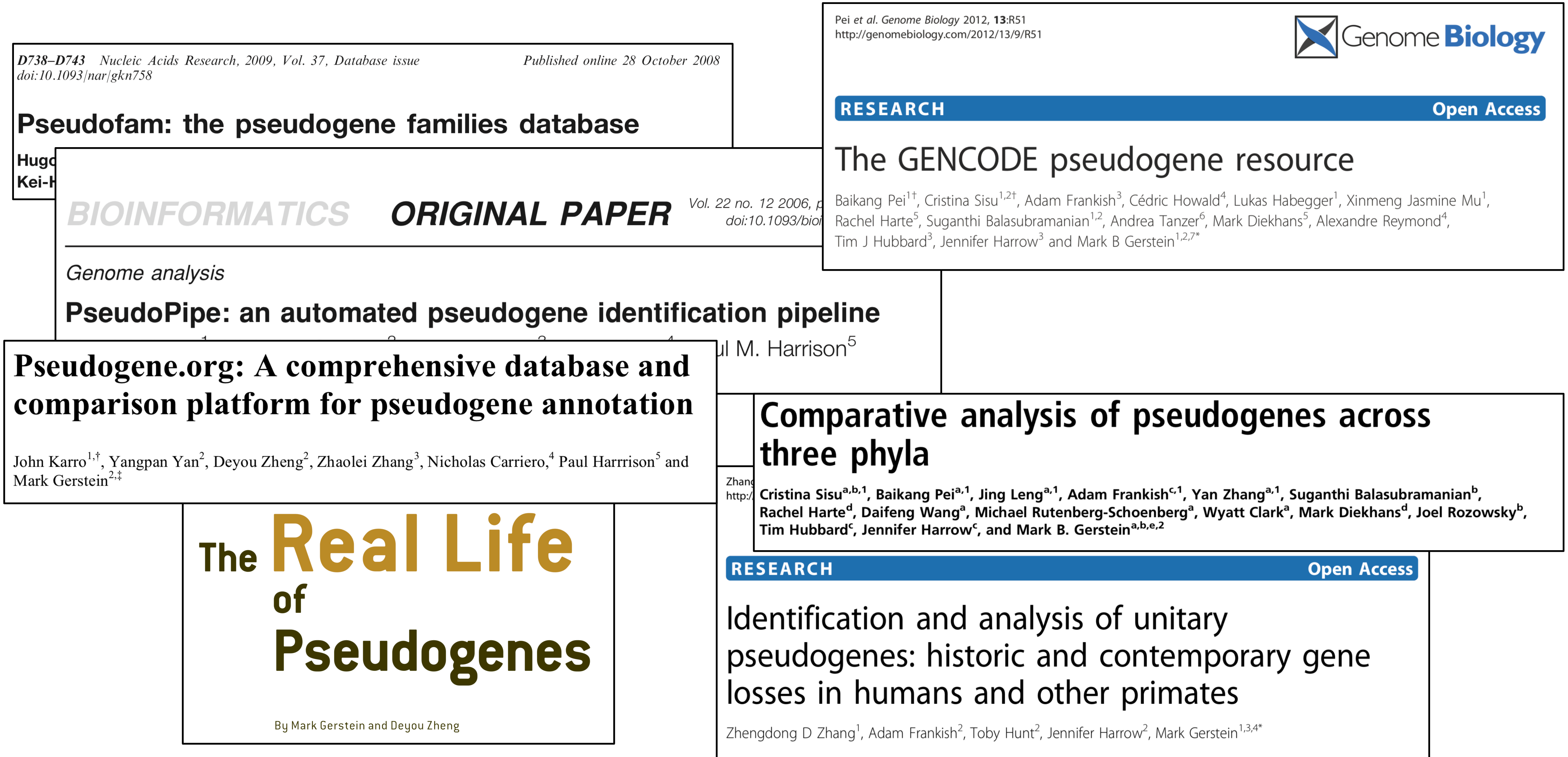
Pseudogenes are **non functional, evolutionary fossils**

2000-present



Systematic annotation and analysis of **pseudogene** complements in genomes of **human** and **model organisms**

The Gerstein lab has a long history in pseudogene annotation and analysis



D738–D743 Nucleic Acids Research, 2009, Vol. 37, Database issue Published online 28 October 2008
doi:10.1093/nar/gkn758

Pseudofam: the pseudogene families database

Hugo
Kei-H

BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 12 2006, p
doi:10.1093/bioi

Genome analysis

PseudoPipe: an automated pseudogene identification pipeline

ul M. Harrison[5]

Pseudogene.org: A comprehensive database and comparison platform for pseudogene annotation

John Karro[1,†], Yangpan Yan[2], Deyou Zheng[2], Zhaolei Zhang[3], Nicholas Carriero,[4] Paul Harrrison[5] and Mark Gerstein[2,‡]

The Real Life of Pseudogenes

By Mark Gerstein and Deyou Zheng

Pei *et al. Genome Biology* 2012, **13**:R51
http://genomebiology.com/2012/13/9/R51

Genome **Biology**

RESEARCH Open Access

The GENCODE pseudogene resource

Baikang Pei[1†], Cristina Sisu[1,2†], Adam Frankish[3], Cédric Howald[4], Lukas Habegger[1], Xinmeng Jasmine Mu[1], Rachel Harte[5], Suganthi Balasubramanian[1,2], Andrea Tanzer[6], Mark Diekhans[5], Alexandre Reymond[4], Tim J Hubbard[3], Jennifer Harrow[3] and Mark B Gerstein[1,2,7*]

Comparative analysis of pseudogenes across three phyla

Cristina Sisu[a,b,1], Baikang Pei[a,1], Jing Leng[a,1], Adam Frankish[c,1], Yan Zhang[a,1], Suganthi Balasubramanian[b], Rachel Harte[d], Daifeng Wang[a], Michael Rutenberg-Schoenberg[a], Wyatt Clark[a], Mark Diekhans[d], Joel Rozowsky[b], Tim Hubbard[c], Jennifer Harrow[c], and Mark B. Gerstein[a,b,e,2]

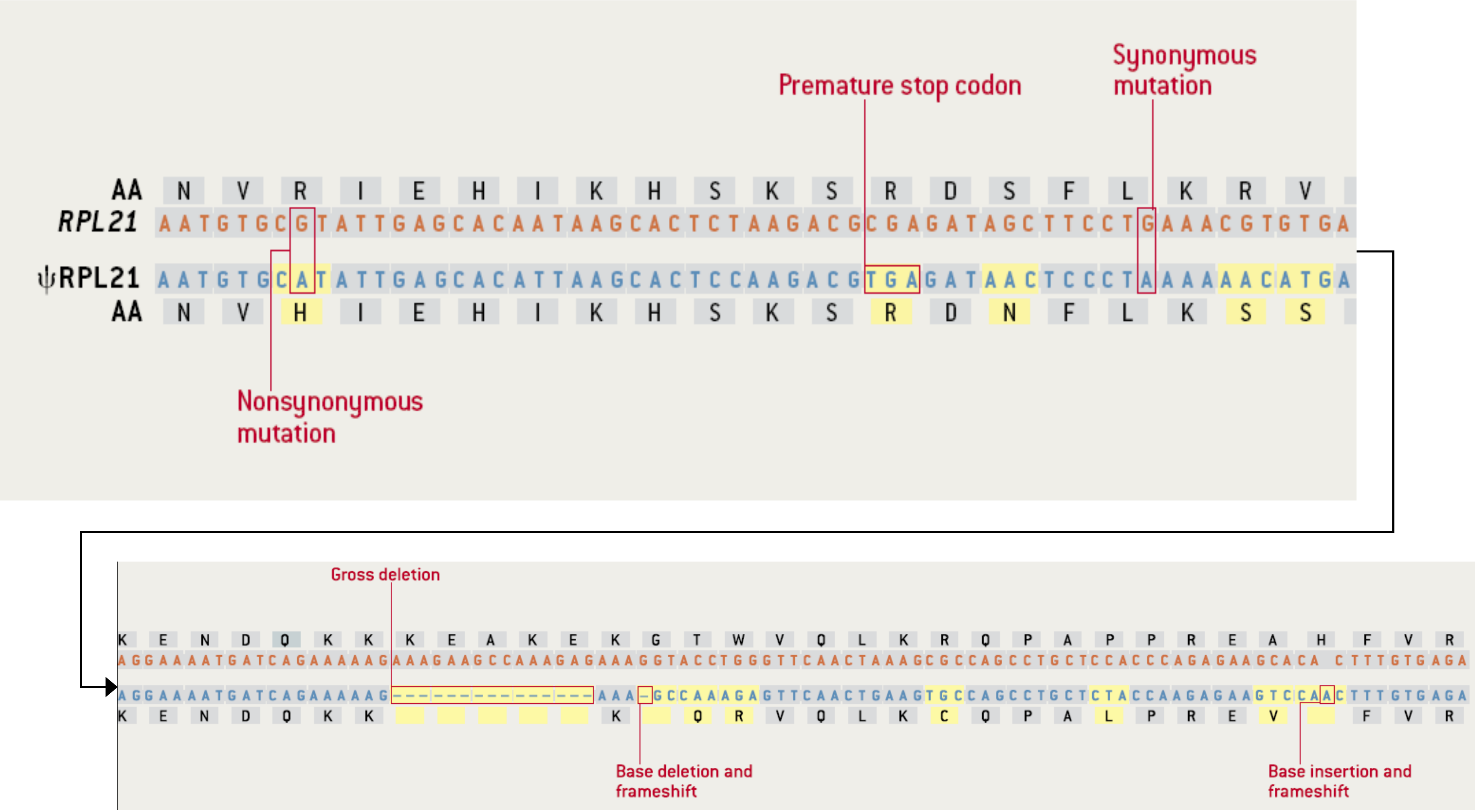
Zhang
http:/

RESEARCH Open Access

Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates

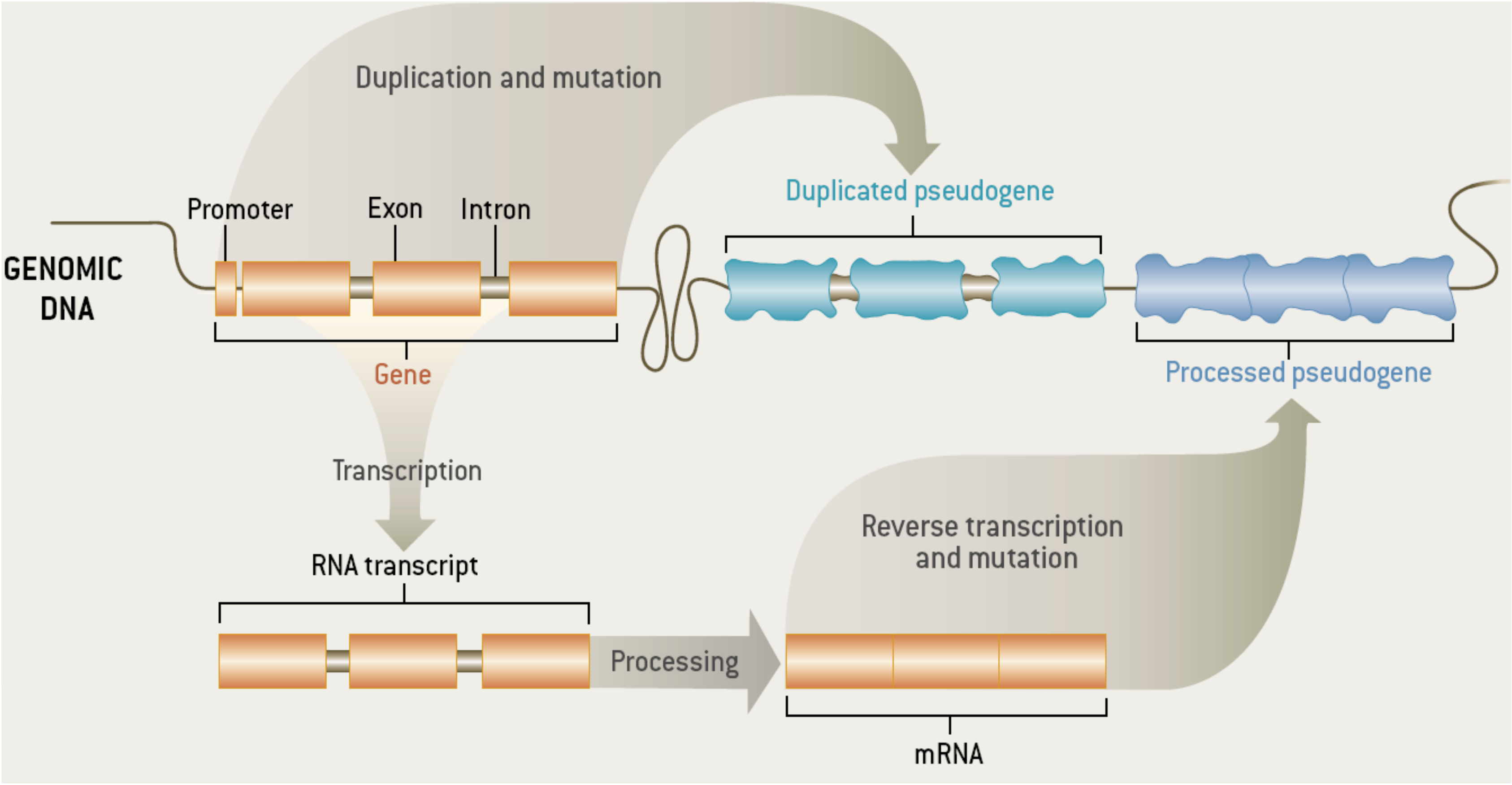
Zhengdong D Zhang[1], Adam Frankish[2], Toby Hunt[2], Jennifer Harrow[2], Mark Gerstein[1,3,4*]

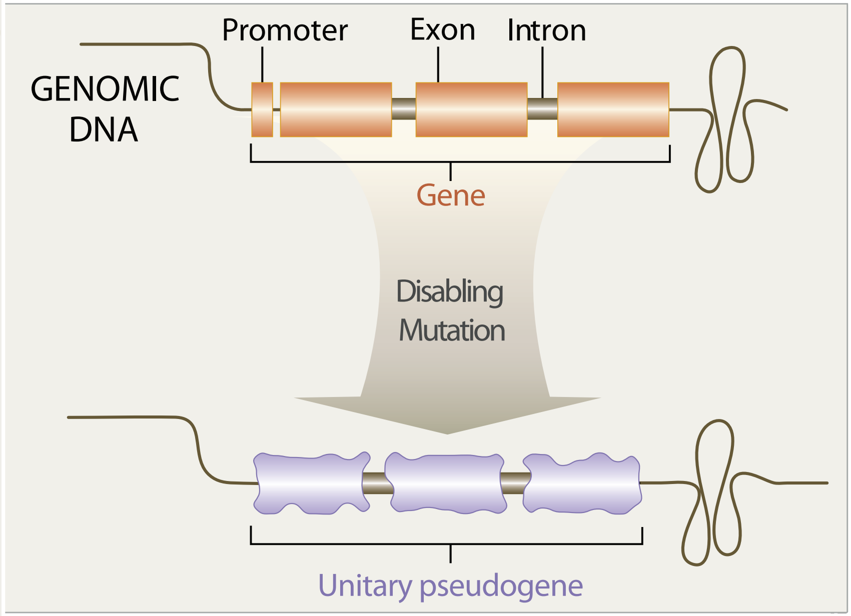
A canonical pseudogene


Premature stop codon
Synonymous mutation
AA N V R I E H I K H S K S R D S F L K R V
RPL21 AATGTGCGTATTGAGCACAATAAGCACTCTAAGACGCGAGATAGCTTCCTGAAACGTGTGA
ψRPL21 AATGTGCATATTGAGCACATTAAGCACTCCAAGACGTGAGATAACTCCCTAAAAAACATGA
AA N V H I E H I K H S K S R D N F L K S S
Nonsynonymous mutation
Gross deletion
K E N D Q K K K E A K E K G T W V Q L K R Q P A P P R E A H F V R
AGGAAAATGATCAGAAAAAGAAAGAAGCCAAAGAGAAAGGTACCTGGGTTCAACTAAAGCGCCAGCCTGCTCCACCCAGAGAAGCACA CTTTGTGAGA
AGGAAAATGATCAGAAAAAG---------------AAA-GCCAAAGAGTTCAACTGAAGTGCCAGCCTGCTCTACCAAGAGAAGTCCAACTTTGTGAGA
K E N D Q K K K Q R V Q L K C Q P A L P R E V F V R
Base deletion and frameshift
Base insertion and frameshift

Gerstein and Zheng SciAm 2006

Review of pseudogene biogenesis





Gerstein and Zheng SciAm 2006

Future work

- Finalize the annotation for mouse

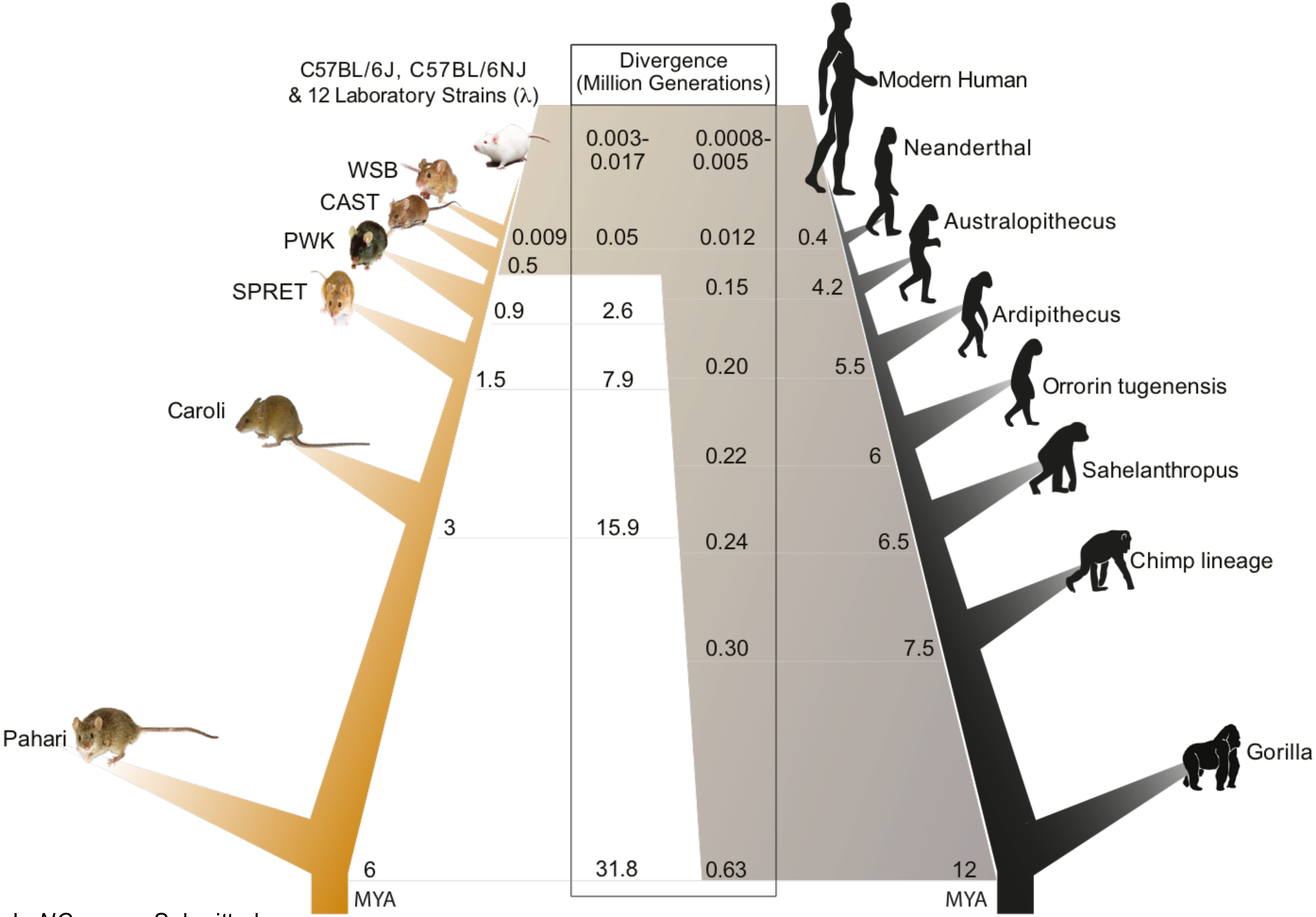
- Improve the annotation in mouse strains

- Pseudogenes as personalized annotations

- Pseudogene annotation customized for human disease studies

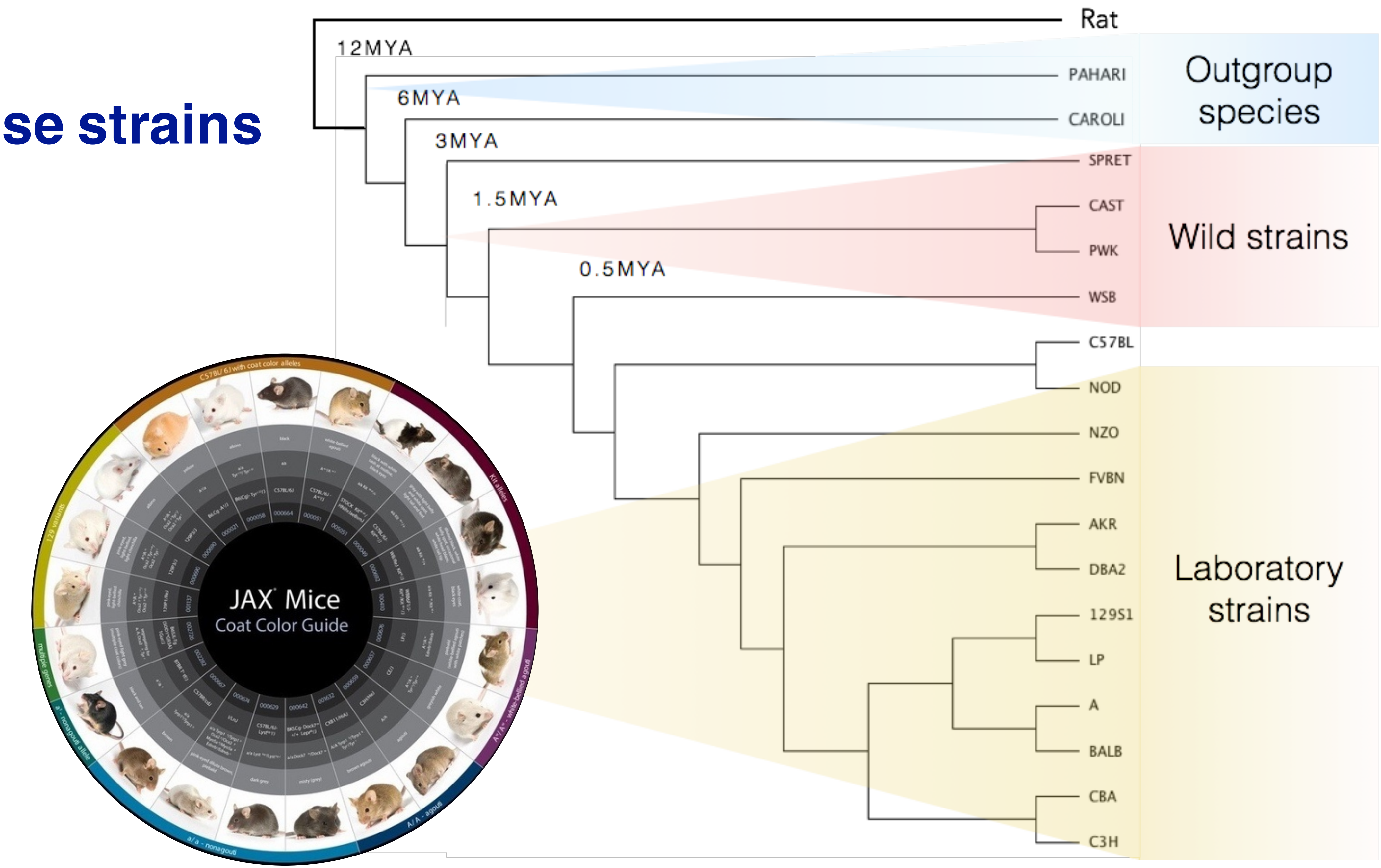
Pseudogenes in the mouse lineage

Comparisons across the mouse and primate lineages



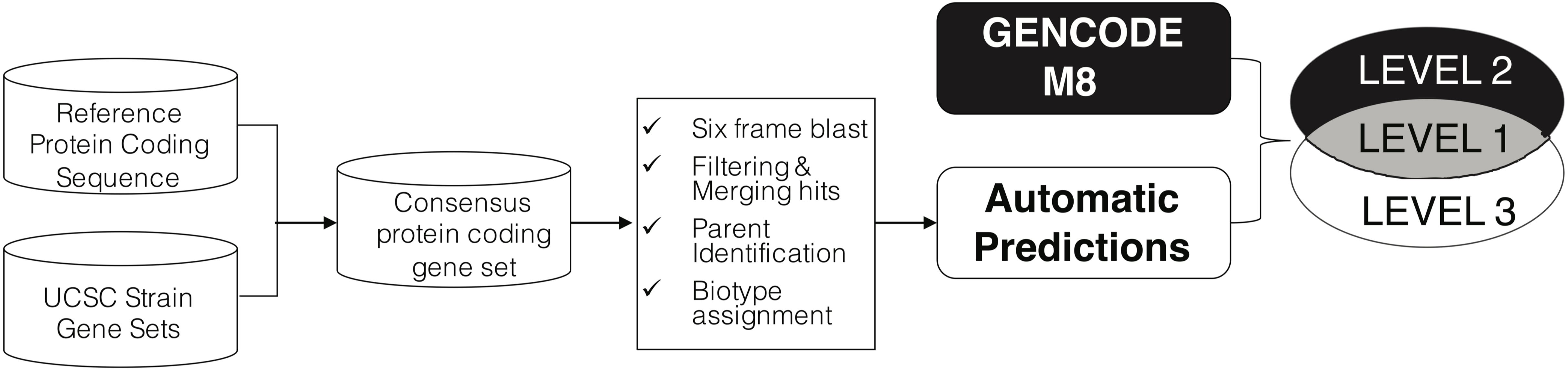
Sisu C.*, Muir P.* et al., *NComms.*, Submitted

Mouse strains



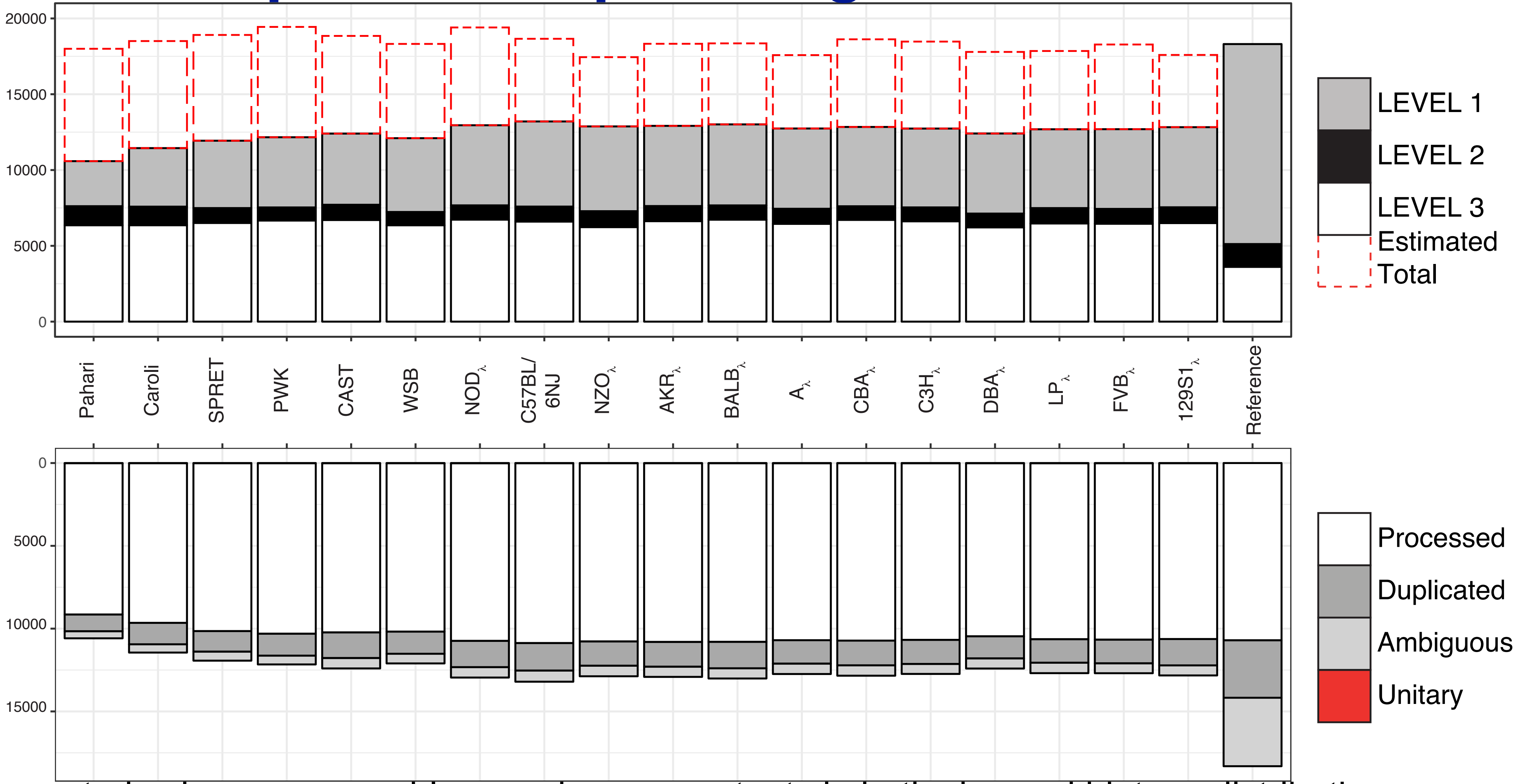
Sisu C.*, Muir P.* et al., *NComms.*, Submitted

Pseudogene annotation pipeline



Sisu C.*, Muir P.* et al., *NComms.*, Submitted

Comparison of pseudogene annotations

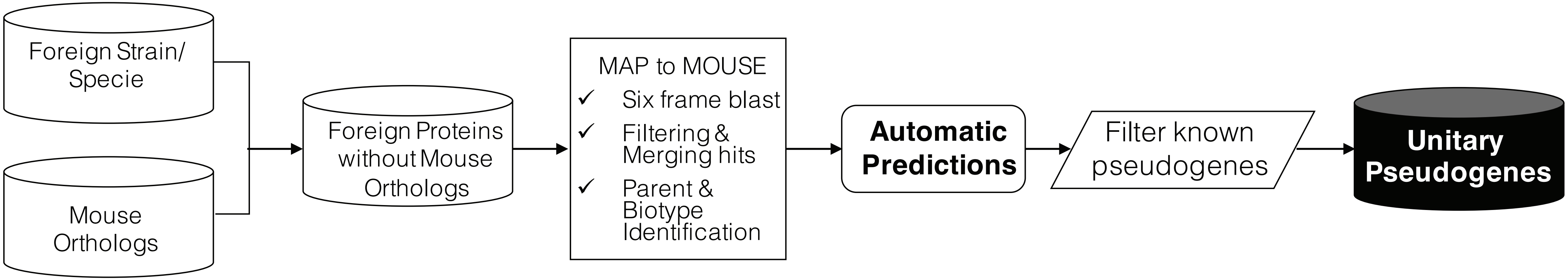


Mouse strains have comparable pseudogene contents in both size and biotype distribution.

Fewer annotations in more divergent species due to use of reference mouse coding set.

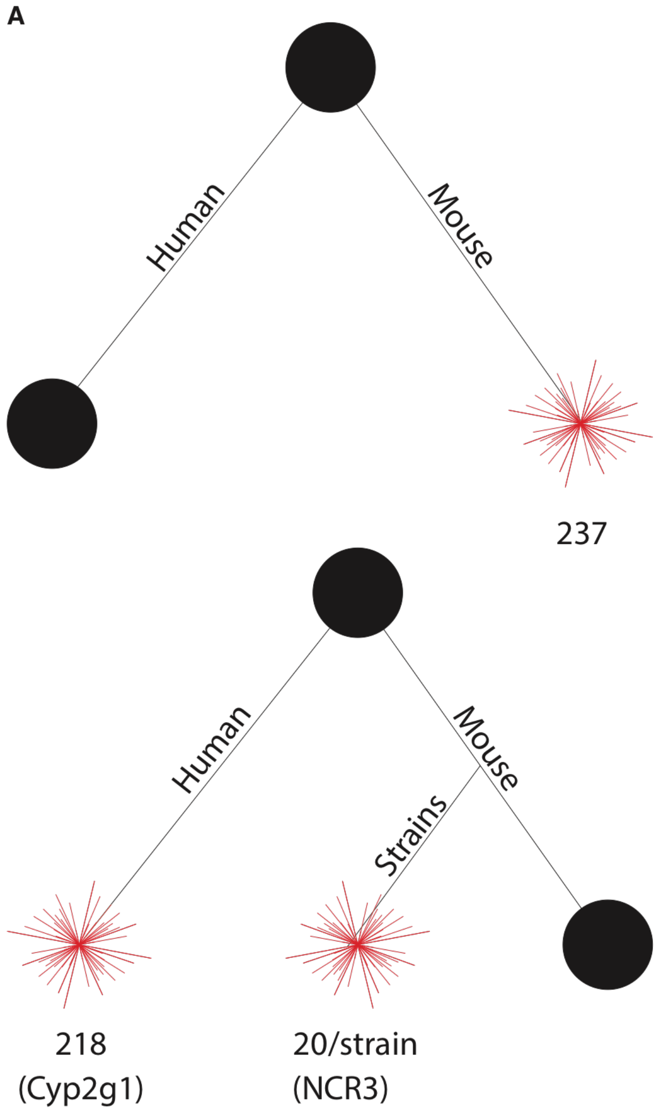
Sisu C.*, Muir P.* et al., *NComms.*, Submitted

Unitary pseudogenes in human and mouse lineages



Sisu C.*, Muir P.* et al., *NComms.*, Submitted

Loss and gain of function in human and mouse lineages



B

Cyp2g1 unitary pseudogene in human with respect to mouse

Arg

CGA

ATTGCCTGGAAA**TGA**ATGAATAAGGCAGG

term

I	A	W	K	*	M	N	K	A	G	Human
I	A	W	K	**R**	M	N	K	A	G	Chimp
I	A	W	K	**R**	M	N	K	A	G	Orangutan
I	A	W	K	**R**	M	N	K	A	G	Rhesus
I	A	W	K	**R**	M	N	K	G	G	Marmoset
I	A	W	K	**R**	T	S	K	G	G	Mus Musculus
I	A	W	K	**R**	T	S	K	G	G	Lab Strains
I	A	W	K	**R**	T	S	K	G	G	Wild Strains (WSB & PWK
I	A	W	K	**W**	T	S	K	G	G	Wild Strains (SPRET)
I	A	W	K	**K**	T	N	K	G	G	Guinea Pig
I	A	W	K	**R**	V	Q	K	P	G	Rabbit

C

NCR3 pseudogene unitary pseudogene with respect to Mus Caroli

Trp

TGG

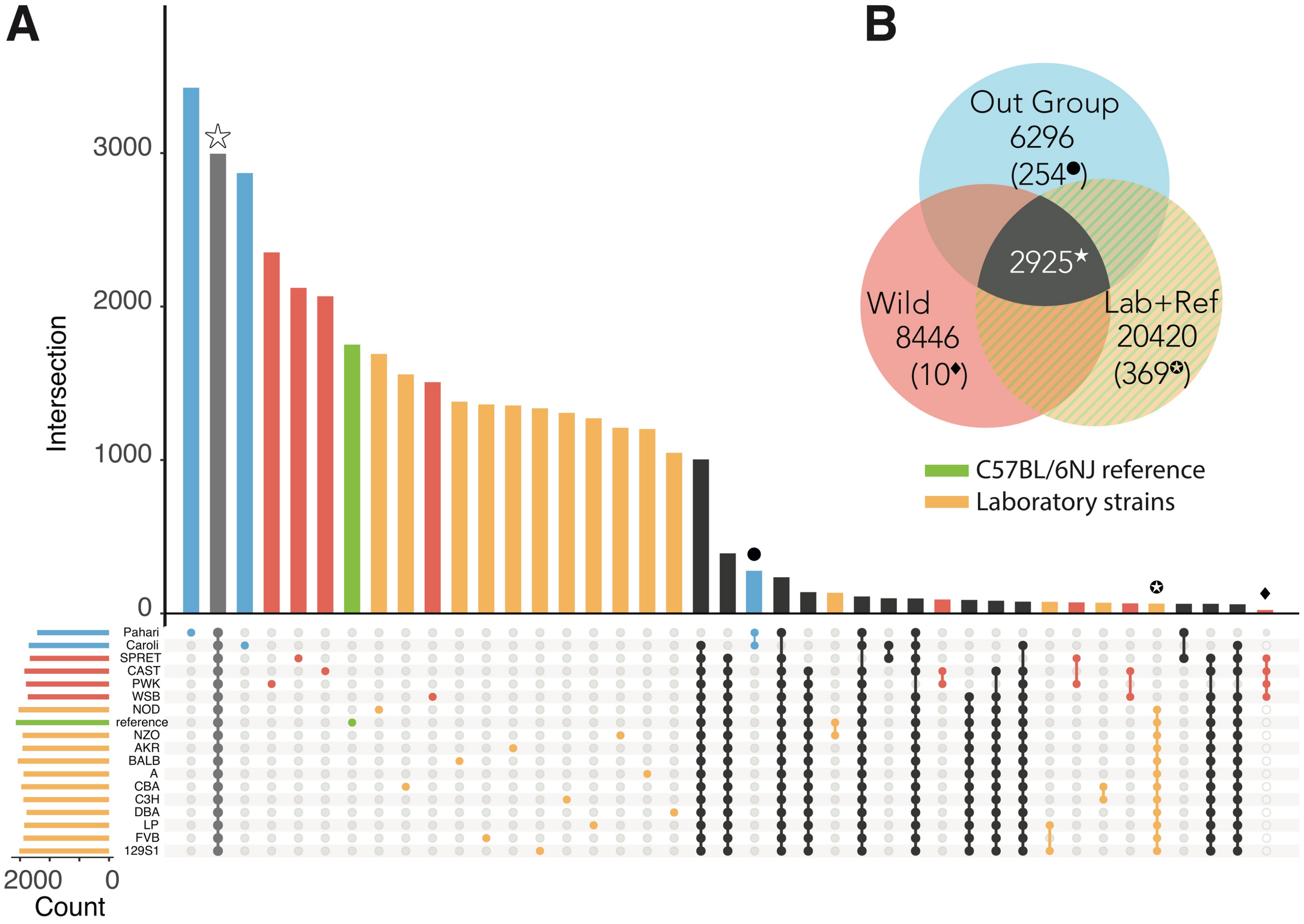
TCCTGTGCTCTC**TGA**GTGTCTCAGCCCCC

term

S	C	L	A	*	V	S	Q	P	P	Mus Musculus
S	C	L	A	**W**	V	S	Q	P	P	Caroli
S	C	L	A	*	V	S	Q	P	P	Wild Strains
S	C	L	A	*	V	S	Q	P	P	Lab Strains

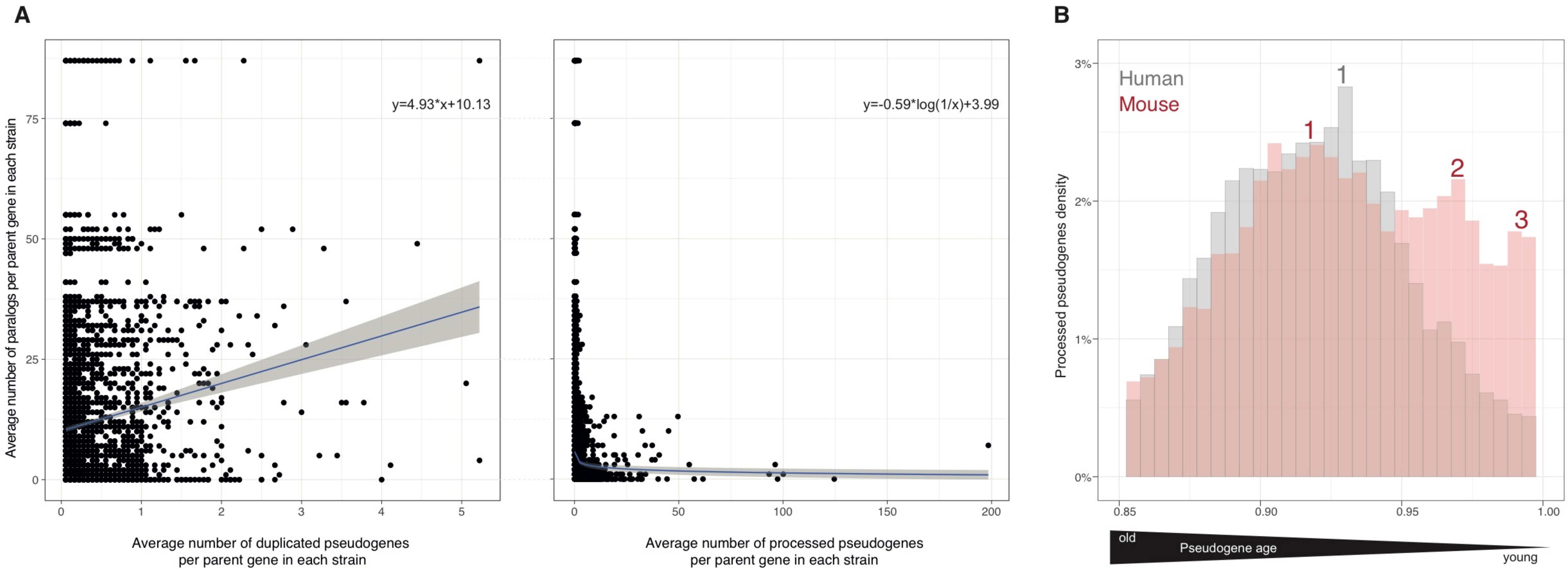
Sisu C.*, Muir P.* et al., *NComms.*, Submitted

Pan-genome pseudogene annotation distribution



Sisu C.*, Muir P.* et al., *NComms.*, Submitted

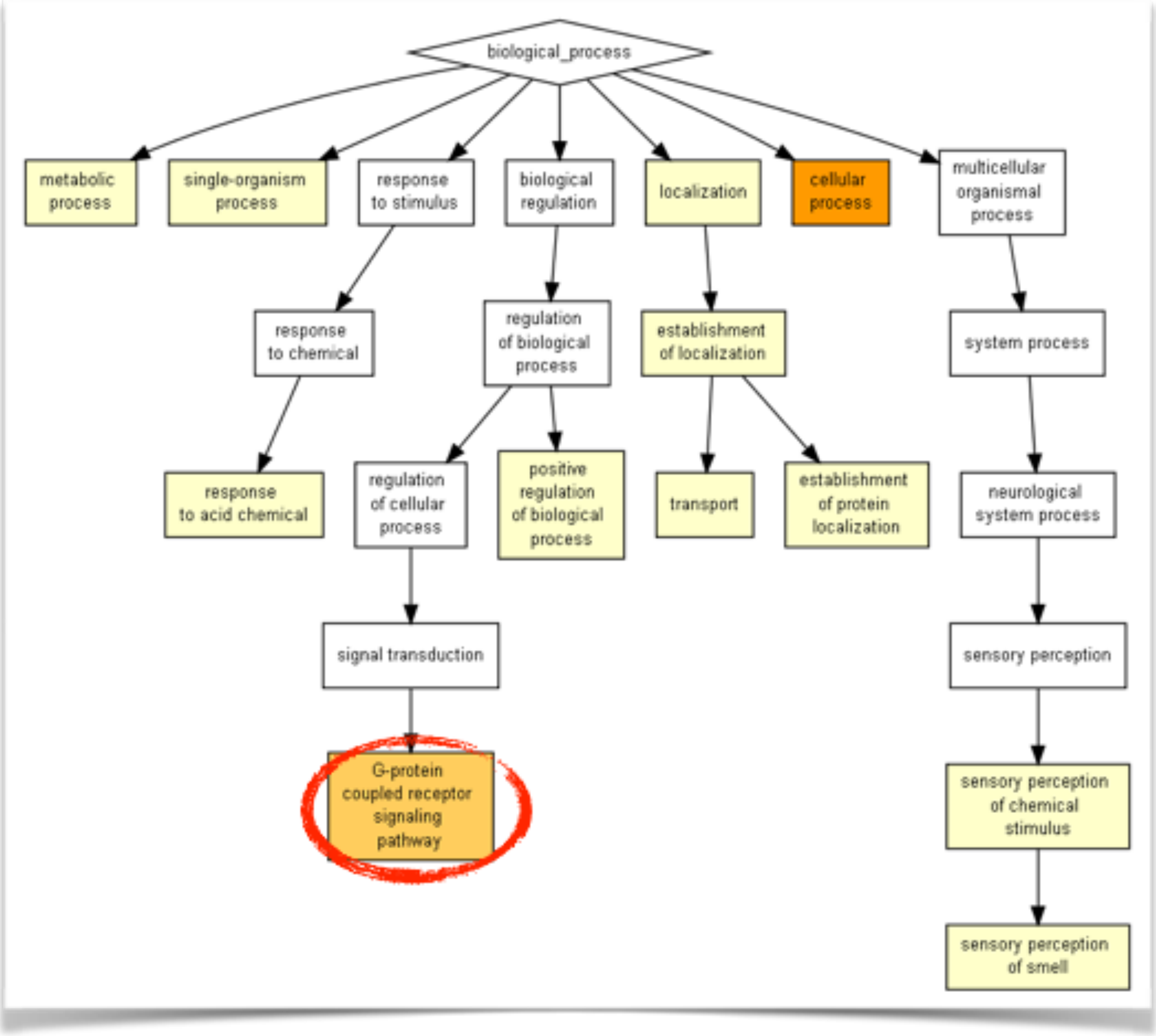
Historical patterns of transposon-mediated pseudogene genesis



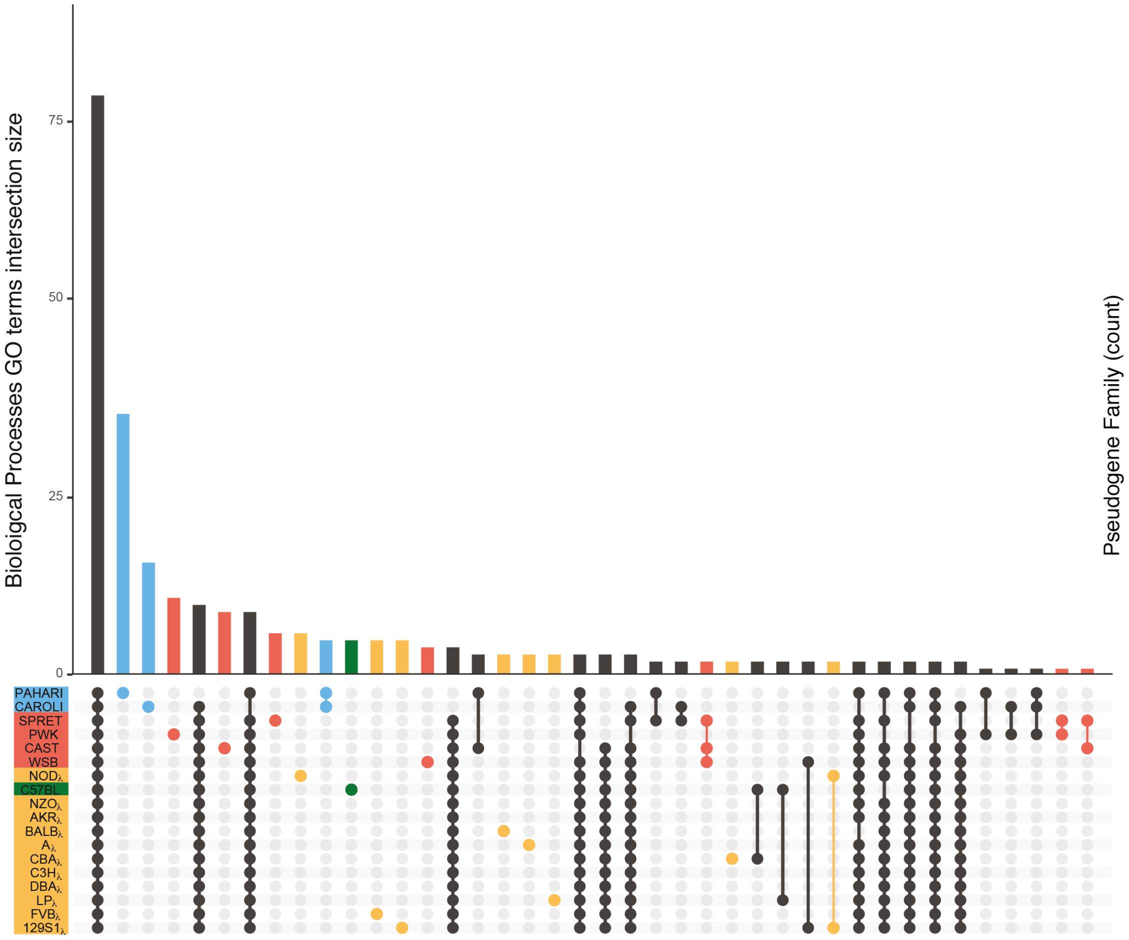
Sisu C.*, Muir P.* et al., *NComms.*, Submitted

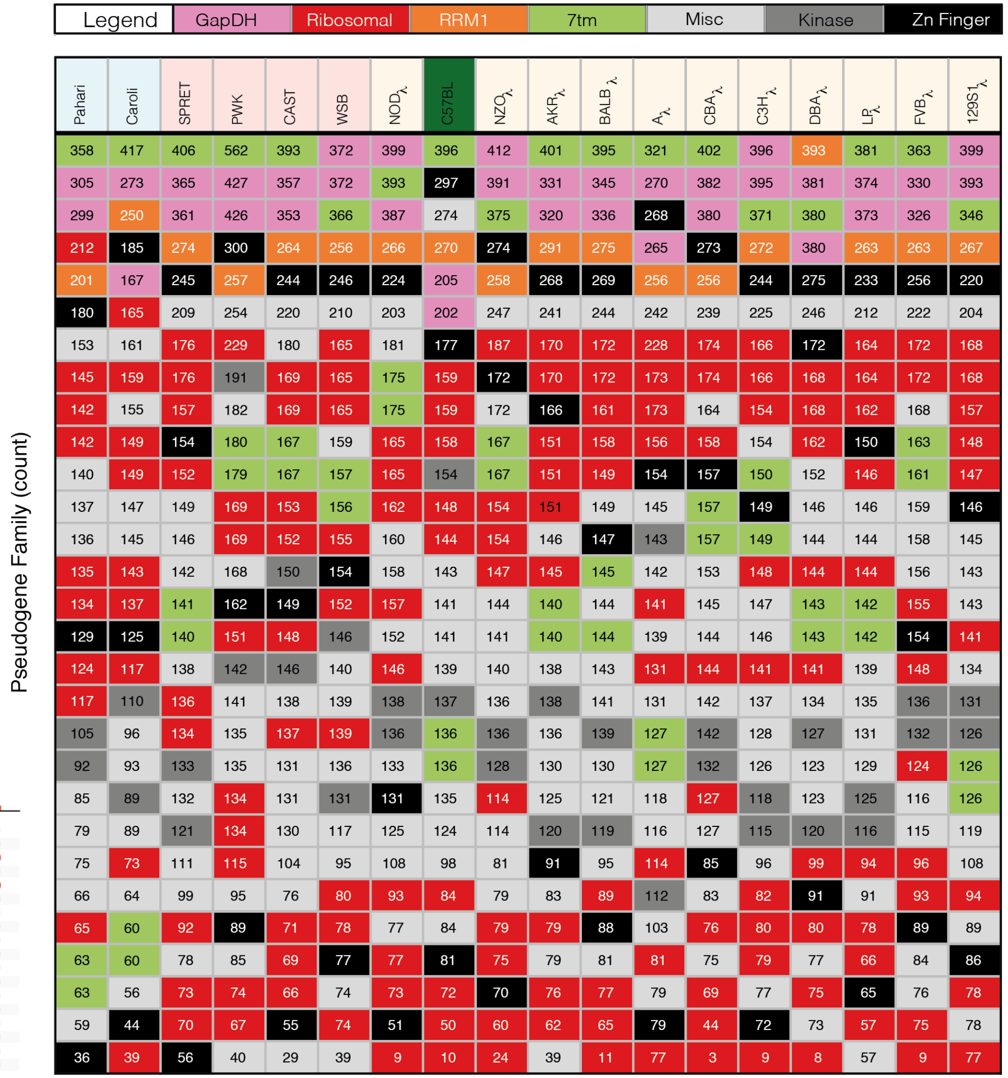
Gene Ontology enrichment analysis of parent genes

Highly abundant protein families show up in GO analysis of pseudogenes.


biological_process
metabolic process
single-organism process
response to stimulus
biological regulation
localization
cellular process
multicellular organismal process
response to chemical
regulation of biological process
establishment of localization
system process
response to acid chemical
regulation of cellular process
positive regulation of biological process
transport
establishment of protein localization
neurological system process
signal transduction
sensory perception
G-protein coupled receptor signaling pathway
sensory perception of chemical stimulus
sensory perception of smell

Cross strain gene ontology and Pfam family analysis of pseudogenes




Pseudogene Family (count)

Legend	GapDH	Ribosomal	RRM1	7tm	Misc	Kinase	Zn Finger

Pahari	Caroli	SPRET	PWK	CAST	WSB	NOD$_\lambda$	C57BL	NZO$_\lambda$	AKR$_\lambda$	BALB $_\lambda$	A$_\lambda$	CBA$_\lambda$	C3H $_\lambda$	DBA$_\lambda$	LP$_\lambda$	FVB$_\lambda$	129S1$_\lambda$
358	417	406	562	393	372	399	396	412	401	395	321	402	396	393	381	363	399
305	273	365	427	357	372	393	297	391	331	345	270	382	395	381	374	330	393
299	250	361	426	353	366	387	274	375	320	336	268	380	371	380	373	326	346
212	185	274	300	264	256	266	270	274	291	275	265	273	272	380	263	263	267
201	167	245	257	244	246	224	205	258	268	269	256	256	244	275	233	256	220
180	165	209	254	220	210	203	202	247	241	244	242	239	225	246	212	222	204
153	161	176	229	180	165	181	177	187	170	172	228	174	166	172	164	172	168
145	159	176	191	169	165	175	159	172	170	172	173	174	166	168	164	172	168
142	155	157	182	169	165	175	159	172	166	161	173	164	154	168	162	168	157
142	149	154	180	167	159	165	158	167	151	158	156	158	154	162	150	163	148
140	149	152	179	167	157	165	154	167	151	149	154	157	150	152	146	161	147
137	147	149	169	153	156	162	148	154	151	149	145	157	149	146	146	159	146
136	145	146	169	152	155	160	144	154	146	147	143	157	149	144	144	158	145
135	143	142	168	150	154	158	143	147	145	145	142	153	148	144	144	156	143
134	137	141	162	149	152	157	141	144	140	144	141	145	147	143	142	155	143
129	125	140	151	148	146	152	141	141	140	144	139	144	146	143	142	154	141
124	117	138	142	146	140	146	139	140	138	143	131	144	141	141	139	148	134
117	110	136	141	138	139	138	137	136	138	141	131	142	137	134	135	136	131
105	96	134	135	137	139	136	136	136	136	139	127	142	128	127	131	132	126
92	93	133	135	131	136	133	136	128	130	130	127	132	126	123	129	124	126
85	89	132	134	131	131	131	135	114	125	121	118	127	118	123	125	116	126
79	89	121	134	130	117	125	124	114	120	119	116	127	115	120	116	115	119
75	73	111	115	104	95	108	98	81	91	95	114	85	96	99	94	96	108
66	64	99	95	76	80	93	84	79	83	89	112	83	82	91	91	93	94
65	60	92	89	71	78	77	84	79	79	88	103	76	80	80	78	89	89
63	60	78	85	69	77	77	81	75	79	81	81	75	79	77	66	84	86
63	56	73	74	66	74	73	72	70	76	77	79	69	77	75	65	76	78
59	44	70	67	55	74	51	50	60	62	65	79	44	72	73	57	75	78
36	39	56	40	29	39	9	10	24	39	11	77	3	9	8	57	9	77

Sisu C.*, Muir P.* et al., *NComms.*, Submitted

Gene Ontology term enrichment amongst pseudogenes (biological processes)

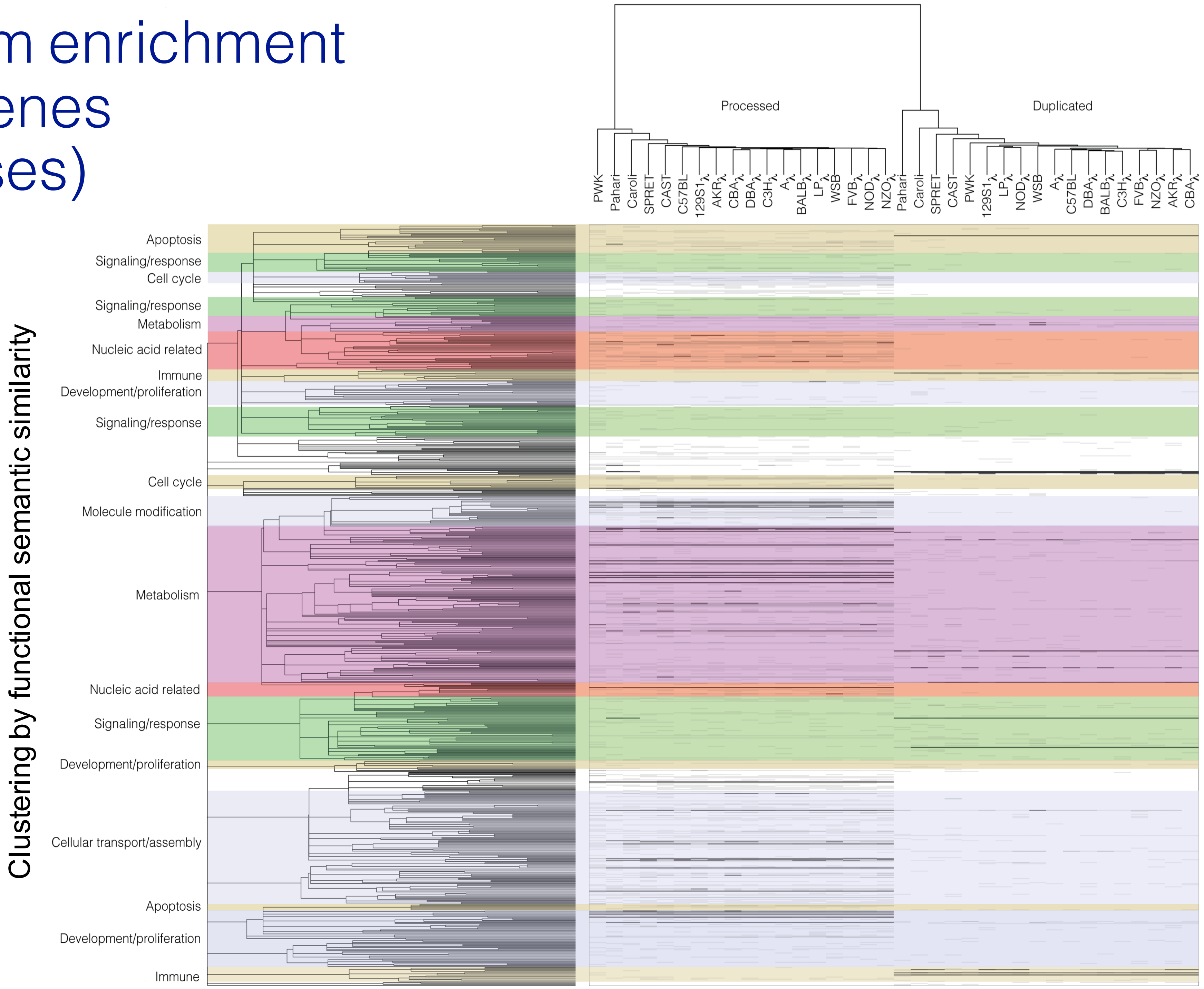
Processed and duplicated pseudogenes enriched for different functions.

Processed pseudogenes enriched for:

- ribosomal functions
- cell cycle
- translation and RNA processing
- ubiquitination.

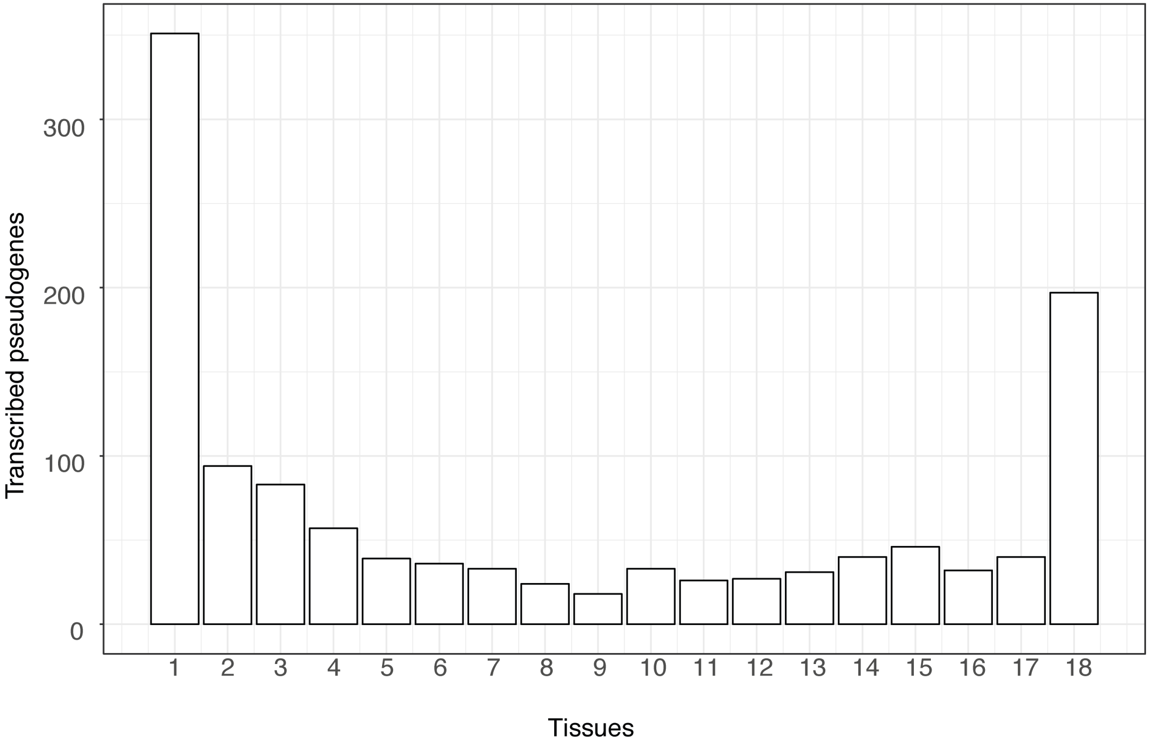
Duplicated pseudogenes enriched for:

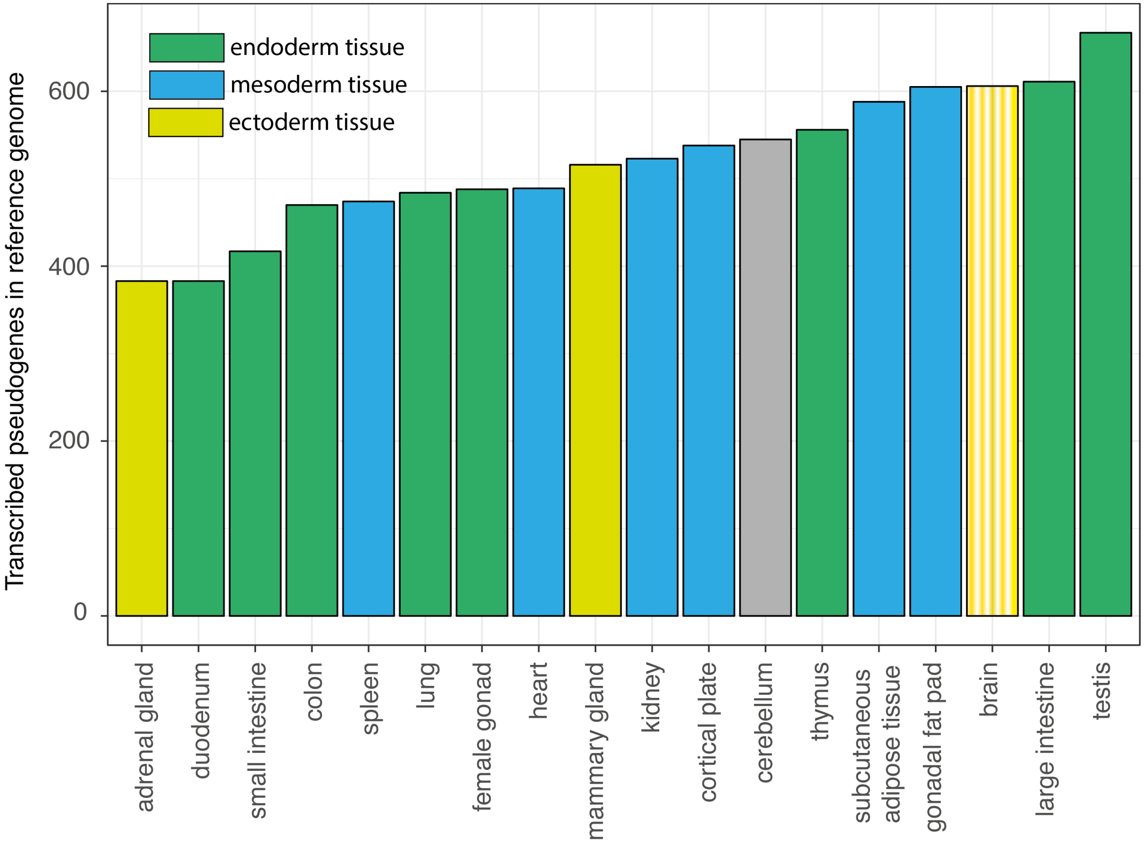
- apoptosis
- sensory and olfactory processes
- immune functions.



Sisu C.*, Muir P.* et al., *NComms.*, Submitted

Transcriptional activity in reference genome





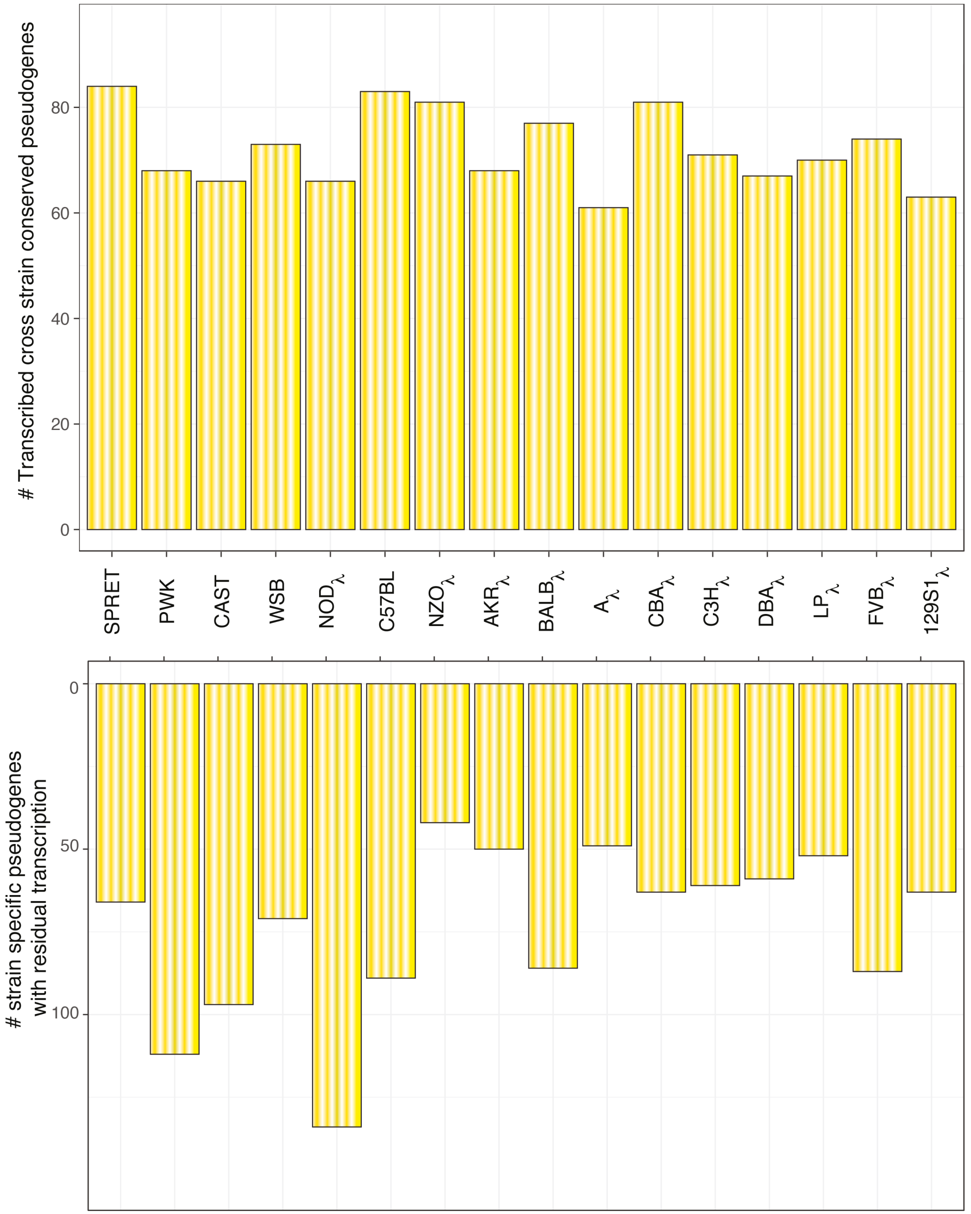
15% of mouse pseudogenes show evidence of residual transcription across multiple tissues

Sisu C.*, Muir P.* et al., *NComms.*, Submitted

Transcriptional activity in mouse strains

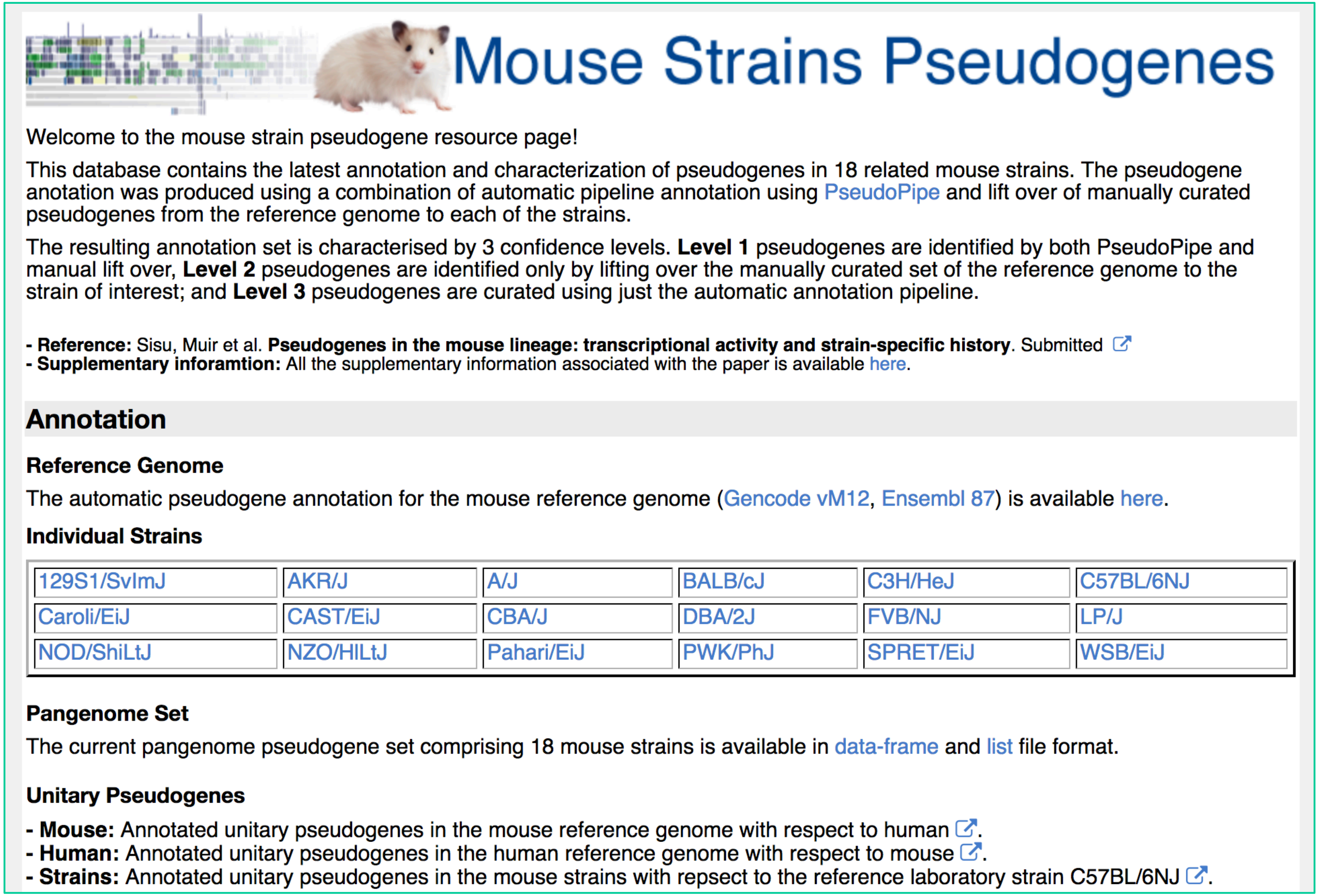
Conserved pseudogenes with transcriptional activity – this set of pseudogenes may need further review to ensure they are not misclassified functional elements.

Strain-specific pseudogenes with transcriptional activity – largely residual activity from pseudogenes with regulatory regions which have not decayed.



Sisu C.*, Muir P.* et al., *NComms.*, Submitted

mouse.pseudogene.org



Mouse Strains Pseudogenes

Welcome to the mouse strain pseudogene resource page!

This database contains the latest annotation and characterization of pseudogenes in 18 related mouse strains. The pseudogene anotation was produced using a combination of automatic pipeline annotation using PseudoPipe and lift over of manually curated pseudogenes from the reference genome to each of the strains.

The resulting annotation set is characterised by 3 confidence levels. **Level 1** pseudogenes are identified by both PseudoPipe and manual lift over, **Level 2** pseudogenes are identified only by lifting over the manually curated set of the reference genome to the strain of interest; and **Level 3** pseudogenes are curated using just the automatic annotation pipeline.

- **Reference:** Sisu, Muir et al. **Pseudogenes in the mouse lineage: transcriptional activity and strain-specific history**. Submitted
- **Supplementary inforamtion:** All the supplementary information associated with the paper is available here.

Annotation

Reference Genome

The automatic pseudogene annotation for the mouse reference genome (Gencode vM12, Ensembl 87) is available here.

Individual Strains

129S1/SvImJ	AKR/J	A/J	BALB/cJ	C3H/HeJ	C57BL/6NJ
Caroli/EiJ	CAST/EiJ	CBA/J	DBA/2J	FVB/NJ	LP/J
NOD/ShiLtJ	NZO/HlLtJ	Pahari/EiJ	PWK/PhJ	SPRET/EiJ	WSB/EiJ

Pangenome Set

The current pangenome pseudogene set comprising 18 mouse strains is available in data-frame and list file format.

Unitary Pseudogenes

- **Mouse:** Annotated unitary pseudogenes in the mouse reference genome with respect to human.
- **Human:** Annotated unitary pseudogenes in the human reference genome with respect to mouse.
- **Strains:** Annotated unitary pseudogenes in the mouse strains with repsect to the reference laboratory strain C57BL/6NJ.

Sisu C.*, Muir P.* et al., *NComms.*, Submitted

Summary

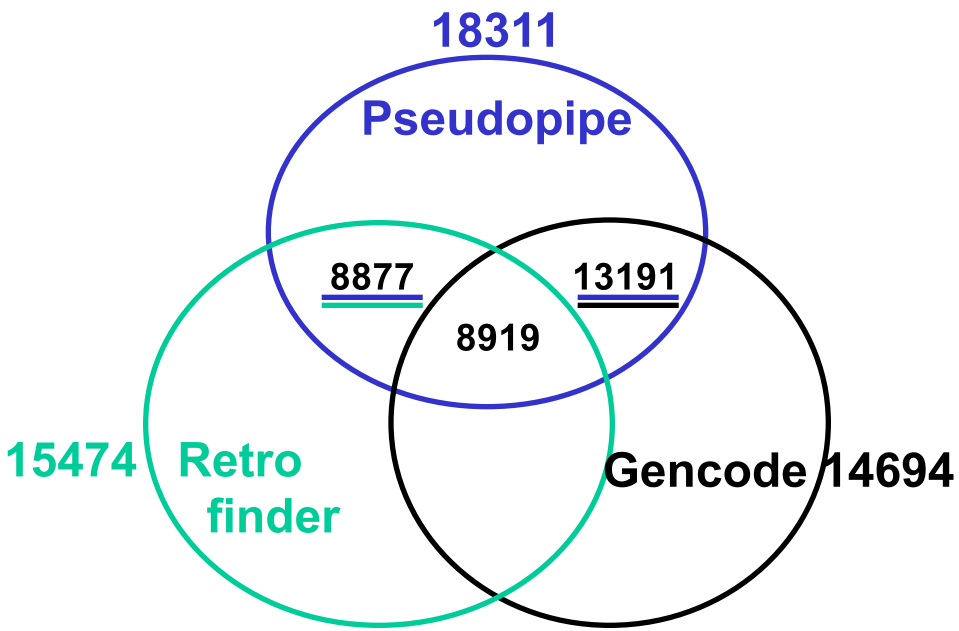
- The first draft of pseudogene annotation in 18 mouse strains and the reference genome
- On average 15-20% of are strain specific and ~ 25% are ancestral, being conserved in all the strains.
- Top pseudogene families are matching closely the human counterparts.
- While human TE activity became silent after the retrotransposition burst, TE are still active in mouse strains.
- Similar to human, pseudogene prolific genes are not enriched in paralogs and vice versa.
- Pseudogene localization suggests multiple large scale genomic rearrangements between the out group - wild strains and the reference (lab strains) mouse genome.
- A significant proportion of show signs of transcriptional activity.

Acknowledgements

Cristina Sisu, Paul Muir, Adam Frankish, Ian Fiddes, Mark Diekhans, David Thybert, Duncan T. Odom, Paul Flicek, Thomas Keane, Tim Hubbard, Jennifer Harrow, Mark Gerstein

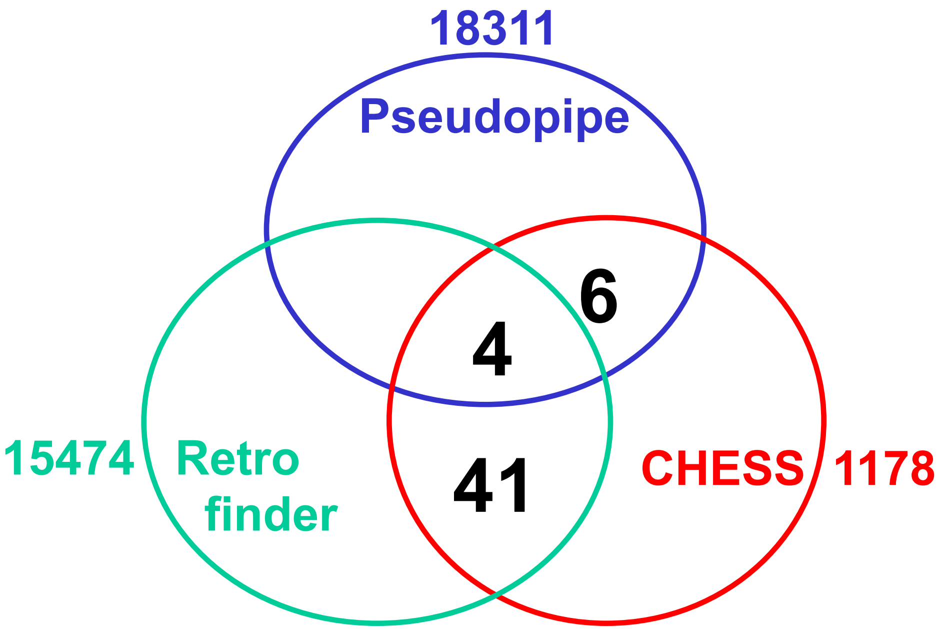
GENCODE

CHESS – pseudogenes or coding genes?



- CHESS genes exclude GENCODE pseudogenes
- Test the overlap between CHESS genes and PseudoPipe & Retrofinder genes

43 unique genes intersect pseudoexons with a 1bp minimum overlap



- 5/6 PseudoPipe pseudoexons have 100% sequence overlap
- 28/41 Retrofinder exons have 100% sequence overlap
- Others: 8-83% sequence overlap

Is CHS.7402 a pseudogene?

- Similar to a protein from the crab eating macaque

```
>XP_005566708.1 PREDICTED: carbohydrate-binding protein AQN-1-
like isoform X2 [Macaca fascicularis]

MRLSRAFAWSLLCSIATIVTAPFATAPSDCGGHYTDEYGRIFNYVGPKTECVWIIELNPGDIVVV
AIPELKGFVCGKEYVEVLDGPPGSESLGRICEAFSTFYHSSSNIITIKYSREPSHPPTFFEIYYF
VDAWSTH
```

```
APSDCGGHYTDEYGRIFNYVGPKTECVWIIELNPGDIVVVAIPELK   KGFVCGKEYVEVLDGP
PGSESLGRICEAFSTFYHSSSNIITIKYSREPSHPPTFFEIYYFVDAWSTH (macaque)
APSDCGGHYTDEYGRIFNYAGPKTECVWIIELNPGEIVTVAIPDLK   RGFACGKEYVEVLDGP
PGSESLDRICKAFSTFYYSSSNIITIKYSREPSHPPTFFEIYYFVDAWSTH  (human)
```

Misses the first 20 amino acids

Does not contain any indels or stop condon disablements

Potentially a duplicated pseudogene?