Updates on Pseudogenes in Mouse Strains

Cristina Sisu & Paul Muir
Gerstein Lab

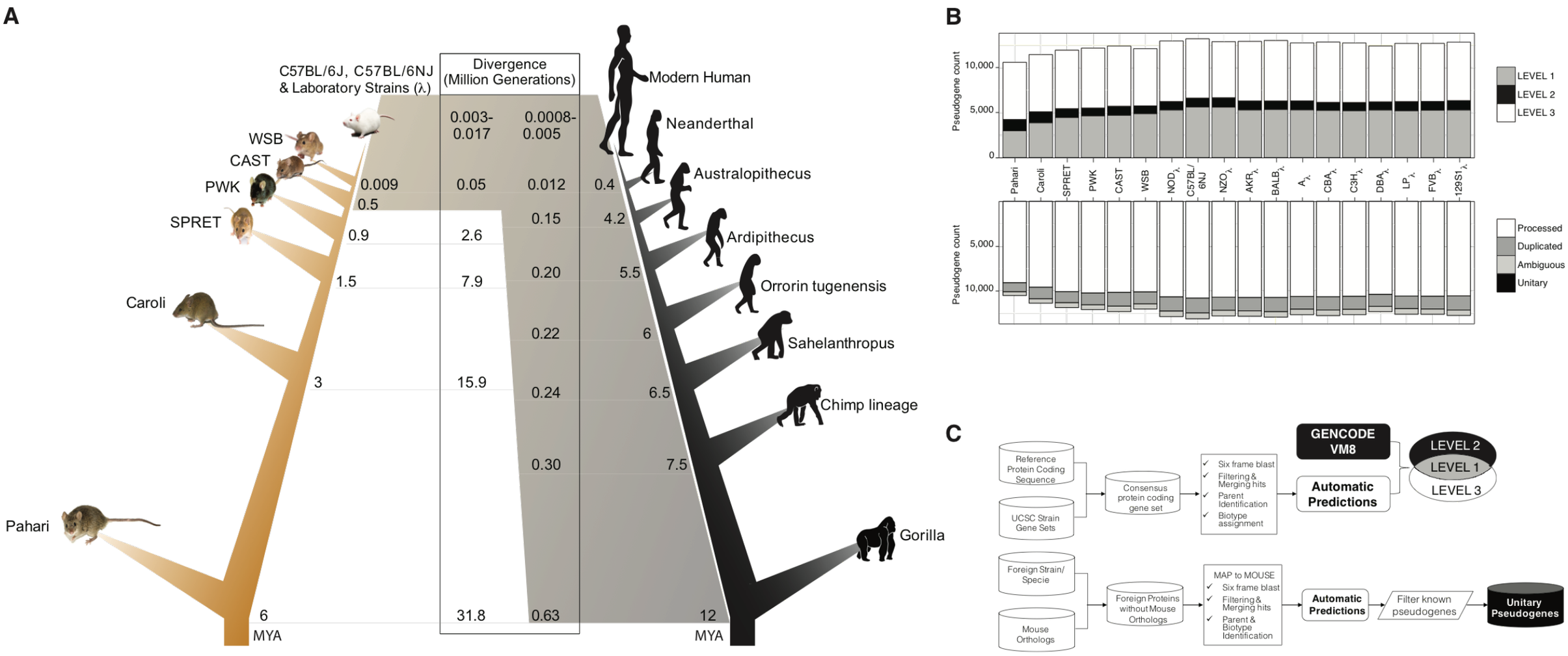
GENCODE meeting 10-11th July 2017

Pseudogenes in the mouse lineage: transcriptional activity and strain-specific history

Cristina Sisu*[1,2,3], Paul Muir*[1], Adam Frankish[4], Ian Fiddes[5], Mark Diekhans[5], David Thybert[4,6] Duncan T. Odom[7,8], and Paul Flicek[4,9], Thomas Keane[4], Mark Gerstein[1,2,10]

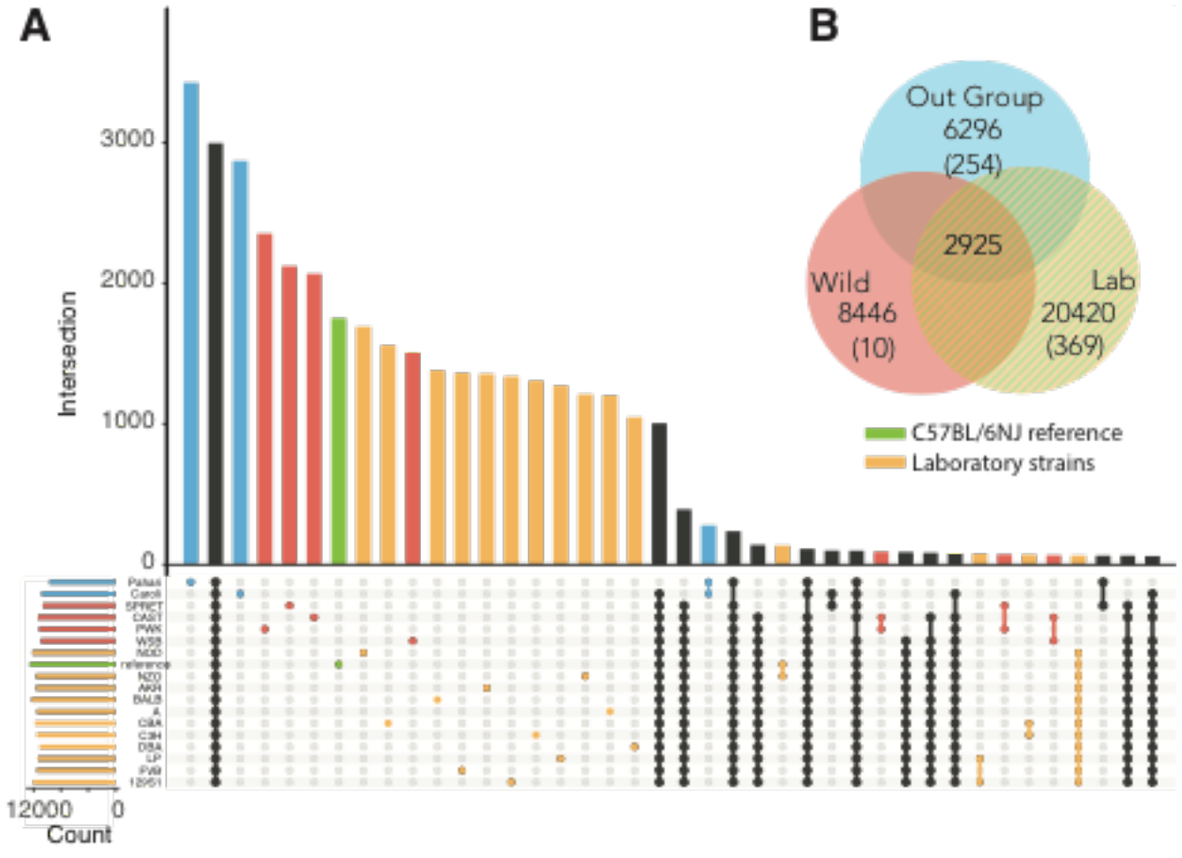
Pseudogenes are ideal markers of genome remodelling. In turn, the mouse is an ideal platform for studying them, particularly with the availability of transcriptional time course data during development (just completed in phase 3 of ENCODE) and the sequencing of 18 strains (completed by the Mouse Genome Project). Here we present a comprehensive genome-wide annotation of the pseudogenes in the mouse reference genome and associated strains. We compiled this by combining manual curation of over 10,000 pseudogenes with results from automatic annotation pipelines. Also, by comparing human and mouse, we annotated 217 new unitary pseudogenes in human and 237 unitary pseudogenes in mouse. (We make our annotation available through a resource website mouse.pseudogene.org.) The overall mouse pseudogene repertoire (in the reference and strains) is similar to human in terms of overall size, biotype distribution (~80% processed, 20% duplicated) and top family composition (with many GAPDH and ribosomal pseudogenes). However, notable differences arise in the age distribution of pseudogenes with multiple retro-transpositional bursts in mouse evolutionary history and only a single one in human. Furthermore, in each strain ~20% of the pseudogenes are unique, reflecting strain-specific functions and evolution – e.g. the pseudogenization of taste receptors is clearly linked to a change in the diet of the NZO strain. Finally, we find ~15% of the pseudogenes are transcribed, a fraction similar to human. Furthermore, we show that processed pseudogenes are commonly associated with highly transcribed genes. While this can be observed through all of mouse development, the relationship is strongest not at the early embryo stages but later on, after depletion of maternal RNA.

Human and mouse share evolutionary similarities

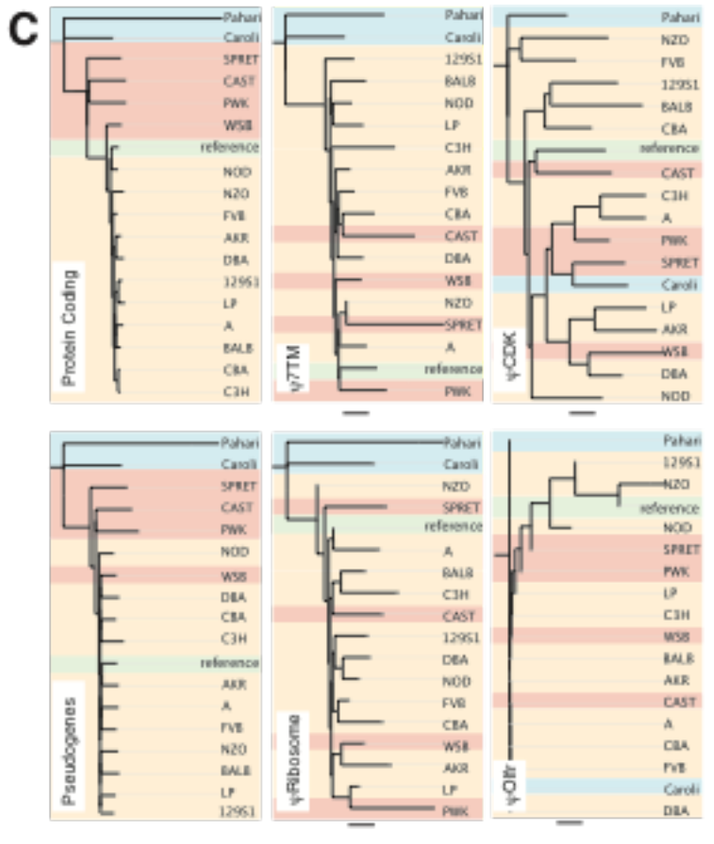


Mouse strains have comparable pseudogene contents in both size and biotype distribution

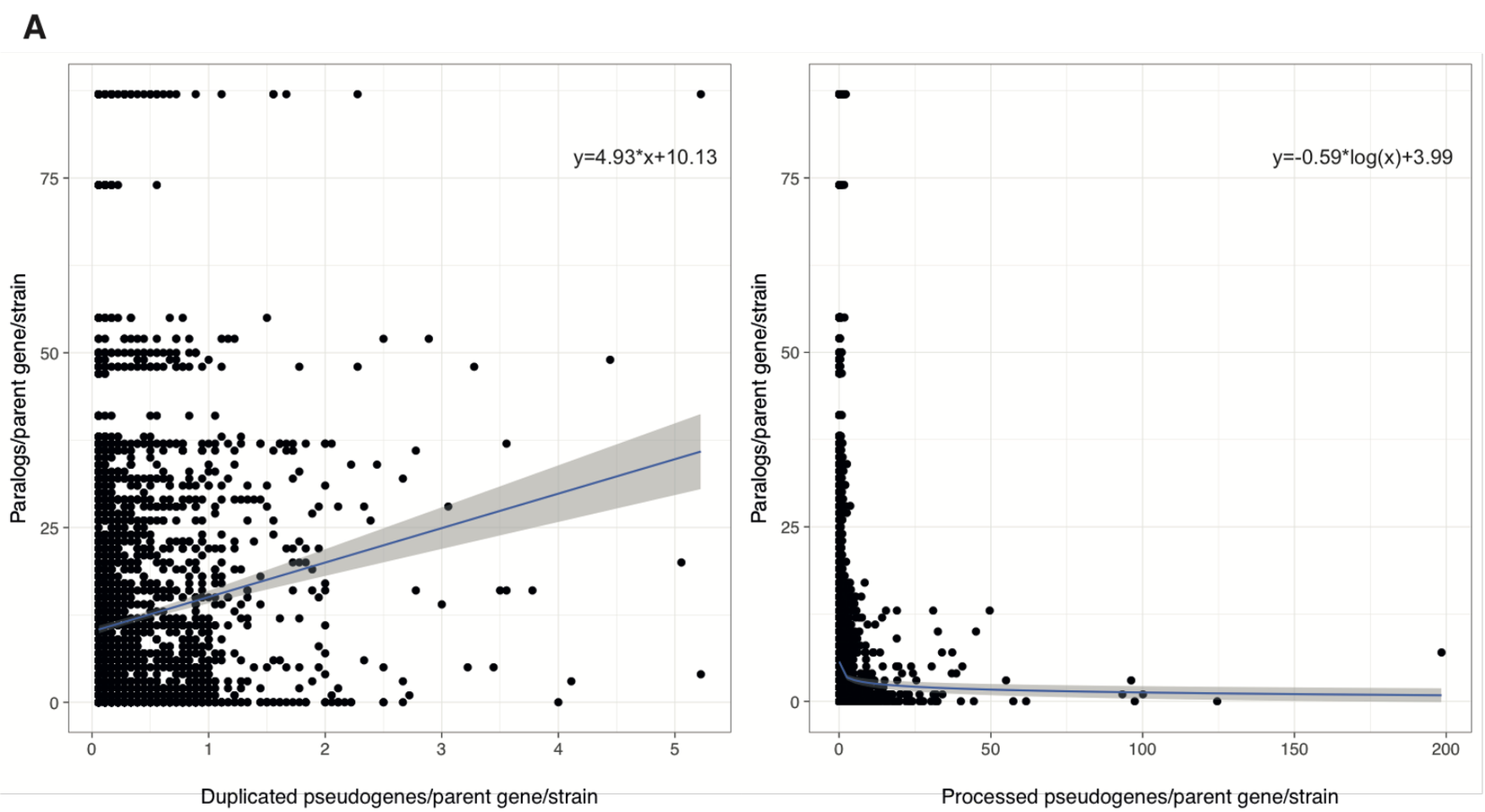
Mouse strain pan-genome dataset

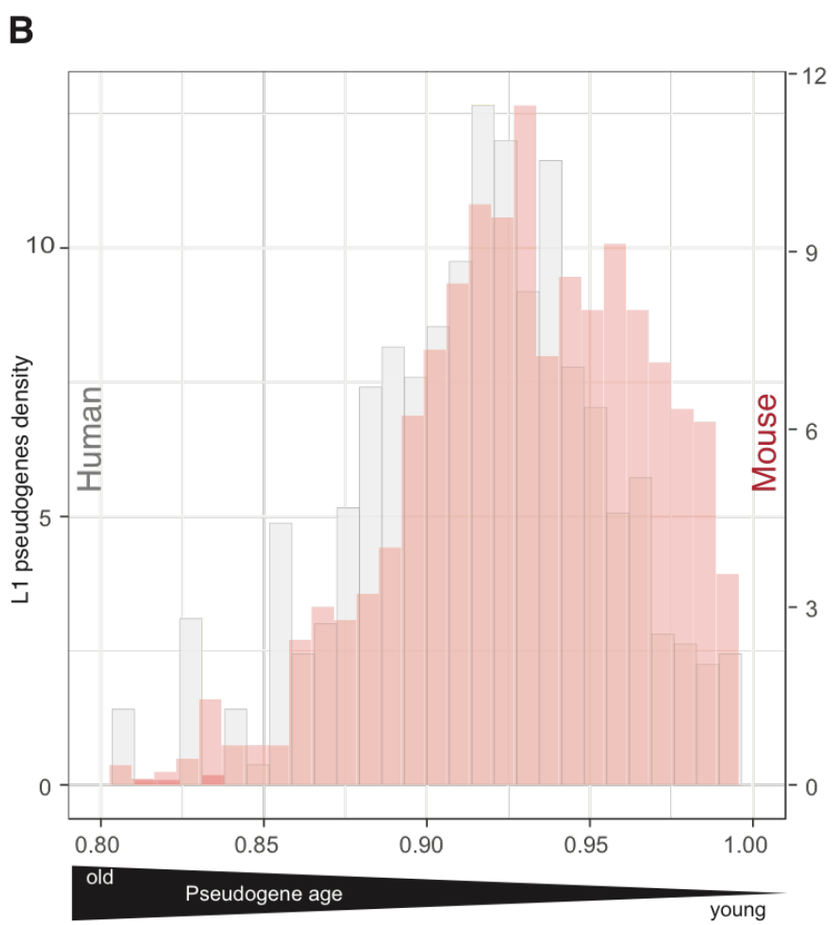


Pseudogenes reflect a strain specific evolution of gene function and phenotype



Genome evolution and plasticity

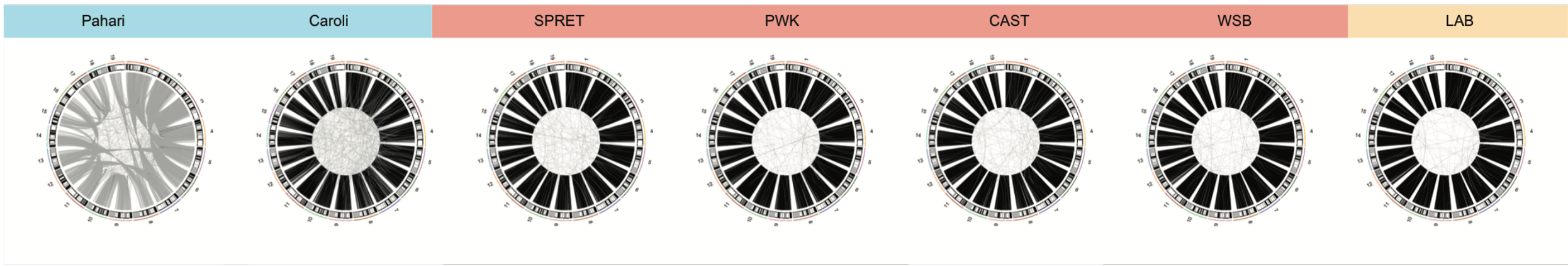




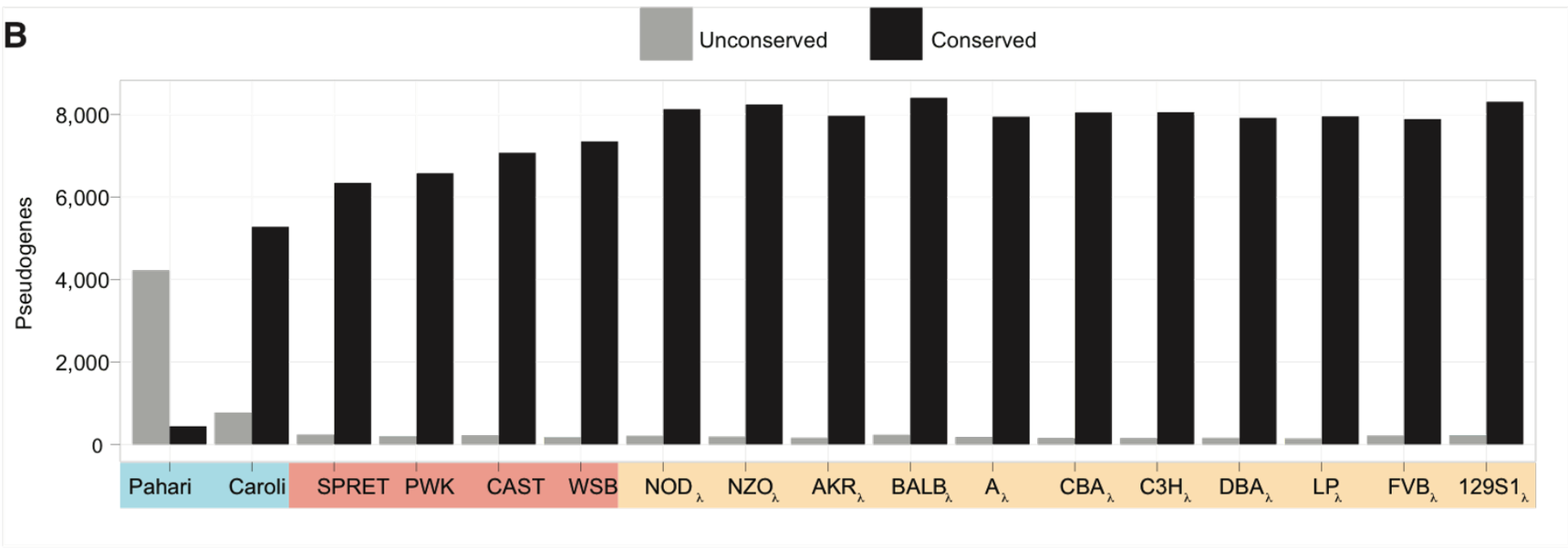
Unprocessed pseudogenes are related to the changes in the selective pressure that drive gene duplication, while processed pseudogenes give an indication of the transposable element activity

Low conservation of pseudogene location in out group species suggests large genomic rearrangements

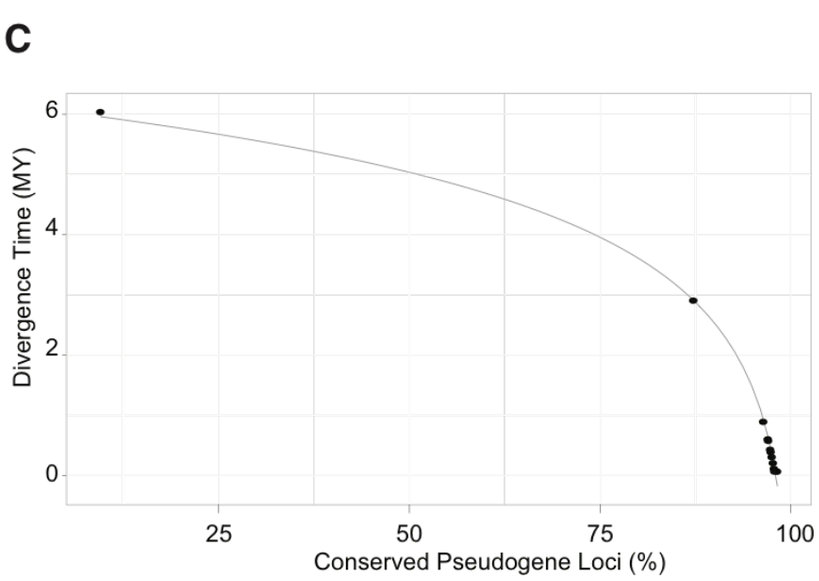
A



B



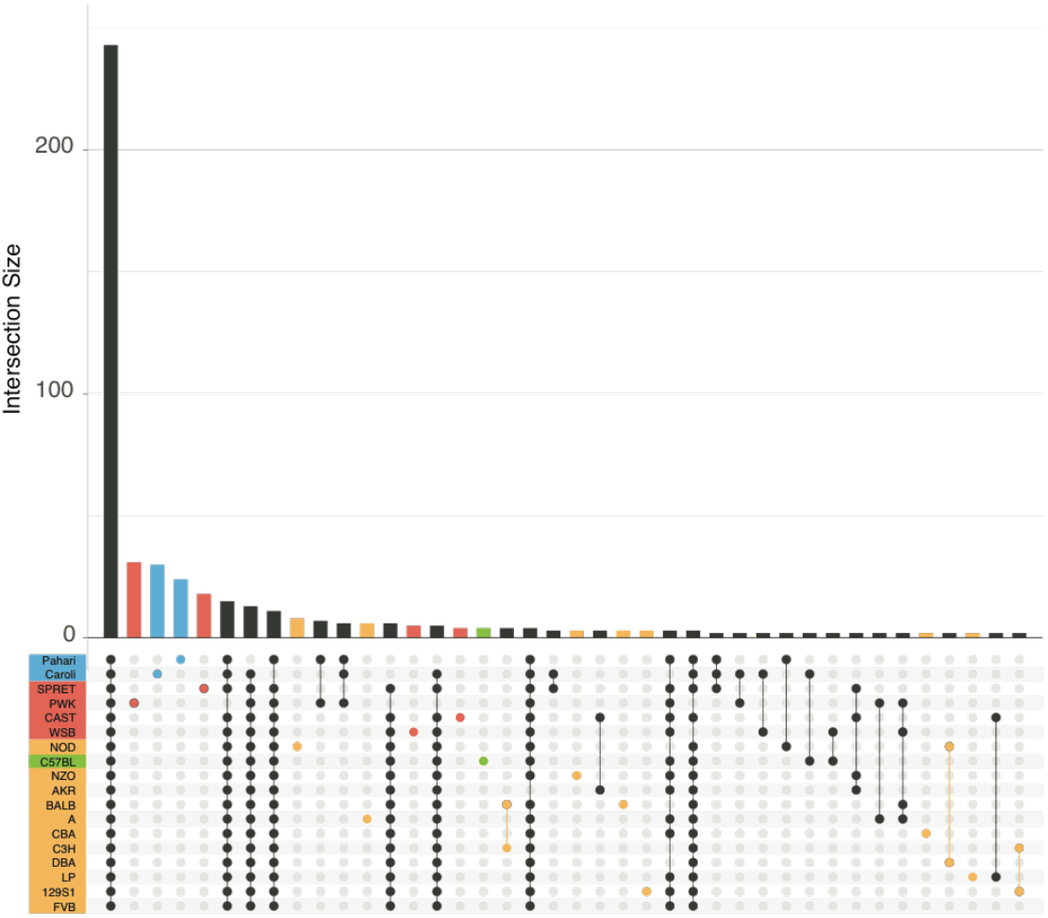
C



% of conserved pseudogene loci follow the diverge times on a logarithmic scale

Gene ontology and pseudogene family analysis reflect a strain specific evolution and phenotype

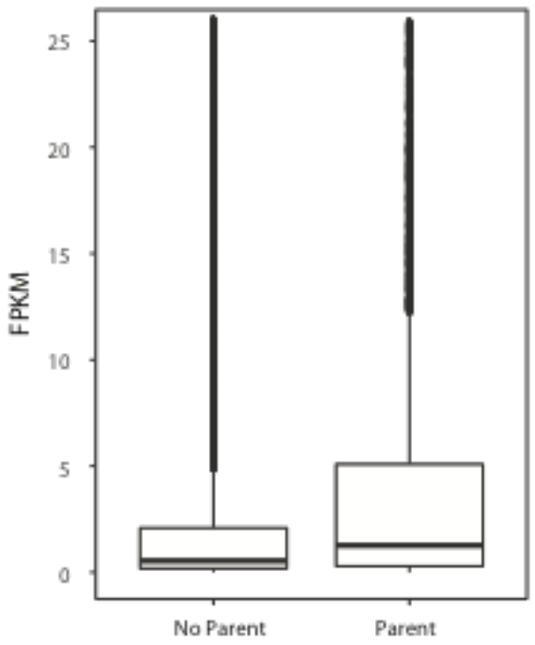
A

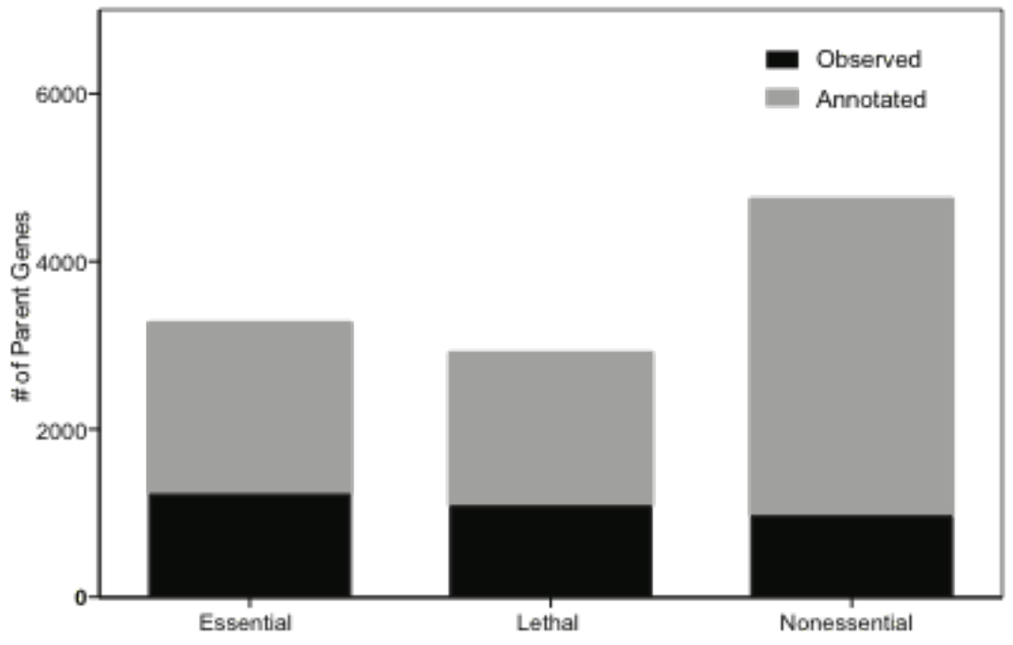


B

Pahari	Caroli	SPRET	PWK	CAST	WSB	NOD	C57BL	NZO	AKR	BALB	A	CBA	C3H	DBA	LP	FVB	129S1
7tm	7tm	7tm	7tm	7tm	GapDH	GapDH	7tm	GapDH	7tm	7tm	7tm	7tm	GapDH	RRM1	7tm	7tm	GapDH
GapDH	GapDH	GapDH	GapDH	GapDH	GapDH	7tm	ZnF	GapDH	GapDH	GapDH	GapDH	GapDH	GapDH	GapDH	GapDH	GapDH	GapDH
GapDH	RRM1	GapDH	GapDH	GapDH	7tm	GapDH	Misc	7tm	GapDH	GapDH	GapDH	GapDH	7tm	7tm	GapDH	GapDH	7tm
Ribo	ZnF	RRM1	300	RRM1	RRM1	RRM1	RRM1	ZnF	RRM1	RRM1	ZnF	ZnF	RRM1	GapDH	RRM1	RRM1	RRM1
RRM1	GapDH	ZnF	RRM1	ZnF	ZnF	ZnF	GapDH	RRM1	ZnF	ZnF	RRM1	RRM1	ZnF	ZnF	ZnF	ZnF	ZnF
ZnF	Ribo	Misc	Misc	Misc	Misc	Misc	GapDH	Misc	Misc	Misc	Misc	Misc	Misc	Misc	Misc	Misc	Misc
Misc	Misc	Ribo	Ribo	Misc	Ribo	Misc	ZnF	Ribo	Ribo	Ribo	Ribo	Ribo	Ribo	ZnF	Ribo	Ribo	Ribo
Ribo	Ribo	Ribo	Kin	Ribo	Ribo	7tm	Ribo	ZnF	Ribo	Ribo	Ribo	Ribo	Ribo	Ribo	Ribo	Ribo	Ribo
Ribo	Misc	Ribo	Misc	Ribo	Ribo	7tm	Ribo	Misc	166	Ribo	Ribo	Misc	Ribo	Ribo	Ribo	Misc	Ribo
Ribo	Ribo	ZnF	7tm	7tm	Misc	Ribo	Ribo	7tm	Ribo	Ribo	Ribo	Ribo	Misc	Ribo	ZnF	7tm	Ribo
Misc	Ribo	Ribo	7tm	7tm	7tm	Ribo	Kin	7tm	Ribo	Ribo	ZnF	ZnF	7tm	Misc	Ribo	7tm	Ribo
Misc	Misc	Misc	Ribo	Ribo	7tm	Ribo	Ribo	Ribo	Ribo	Misc	Misc	7tm	ZnF	Misc	Misc	Misc	ZnF
Misc	Misc	Misc	Ribo	Ribo	Ribo	Misc	Ribo	Ribo	Misc	ZnF	His	7tm	7tm	Misc	Misc	Misc	Misc
Ribo	Ribo	Misc	Misc	His	ZnF	Misc	Misc	Ribo	Ribo	7tm	Misc	Misc	Ribo	Ribo	Ribo	Misc	Misc
Ribo	Ribo	7tm	ZnF	ZnF	Ribo	Ribo	Misc	Misc	7tm	Misc	Ribo	Misc	Misc	7tm	7tm	Ribo	Misc
ZnF	ZnF	7tm	Ribo	Ribo	Kin	Misc	Misc	Misc	7tm	7tm	Misc	Misc	Misc	7tm	7tm	ZnF	Ribo
Ribo	Ribo	Misc	His	Kin	Misc	Ribo	Misc	Misc	Misc	Misc	Ribo	Ribo	Ribo	Ribo	Misc	Ribo	Misc
Ribo	His	Ribo	Misc	Misc	Misc	Kin	His	Misc	Kin	Misc	131	Misc	Misc	Misc	Misc	His	His
HIS	Misc	Ribo	Misc	Ribo	Ribo	His	7tm	Kin	Misc	His	7tm	His	Misc	Kin	Misc	Kin	Kin
Kin	Misc	Kin	Misc	Misc	Misc	Misc	7tm	His	Misc	Misc	7tm	Kin	Misc	Misc	Misc	Ribo	7tm
Misc	Kin	Misc	Ribo	Misc	His	ZnF	Misc	Ribo	Misc	Misc	Misc	Ribo	His	Misc	His	Misc	7tm
Misc	Misc	His	Ribo	Misc	Misc	Misc	Misc	Misc	His	Kin	Misc	Misc	Kin	His	Kin	Misc	Misc
Misc	Ribo	Misc	Ribo	Misc	Misc	Misc	Misc	Misc	ZnF	Misc	Ribo	ZnF	Misc	Ribo	Ribo	Ribo	Misc
Misc	Misc	Misc	Misc	Misc	Ribo	Ribo	Ribo	Misc	Misc	Ribo	Kin	Misc	Ribo	ZnF	91	Ribo	Ribo
Ribo	7tm	Ribo	ZnF	Ribo	Ribo	Misc	Misc	Ribo	Ribo	ZnF	Misc	Ribo	Ribo	Ribo	Ribo	ZnF	Misc
7tm	7tm	Misc	Misc	Ribo	ZnF	Ribo	ZnF	Ribo	Misc	Misc	Ribo	Misc	Ribo	Misc	Ribo	Misc	ZnF
7tm	Misc	Ribo	Ribo	Ribo	Misc	Ribo	Ribo	ZnF	Ribo	Ribo	Misc	Ribo	Misc	Ribo	ZnF	Misc	Ribo
Misc	ZnF	Ribo	Ribo	ZnF	Ribo	ZnF	Ribo	Ribo	Ribo	Ribo	ZnF	Ribo	ZnF	Misc	Ribo	Ribo	Misc
ZnF	Ribo	ZnF	Misc	Misc	Misc	Ribo	Ribo	Ribo	Misc	Ribo	Ribo	Ribo	Ribo	Ribo	Misc	Ribo	Ribo

Essential genes





Unique Pseudogene Parent Gene Transcripts

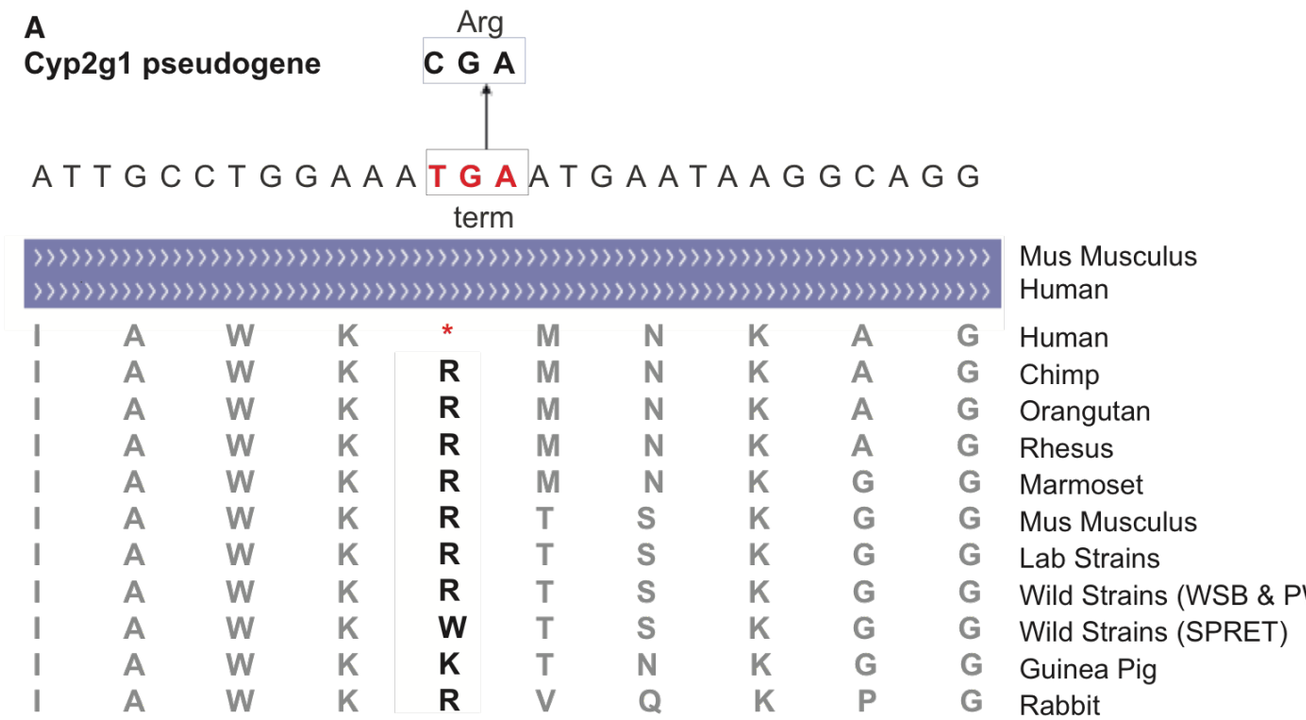
	all	annotated	essential	lethal	nonessential
129S1	6463	4716	0.360711	0.355839	0.194890
A	6514	4762	0.368373	0.363762	0.199113
AKR	6666	4813	0.369599	0.364106	0.200802
BALB	6612	4755	0.366840	0.363762	0.193623
C3H	6511	4672	0.359792	0.352394	0.193412
C57BL	6801	4904	0.375115	0.373062	0.203758
CAROLI	5590	4113	0.310451	0.305890	0.182010
CAST	6434	4694	0.358872	0.353772	0.201647
CBA	6644	4800	0.375115	0.371340	0.197002
DBA2	6504	4669	0.357033	0.351705	0.193201
FVBN	6669	4843	0.374502	0.369962	0.202914
LP	6521	4754	0.372970	0.364451	0.199747
NOD	6882	4952	0.389519	0.384774	0.210304
NZO	6649	4789	0.365002	0.359972	0.196579
PAHARI	5550	4120	0.308612	0.304168	0.189823
PWK	9138	6727	0.512105	0.514640	0.319468
SPRET	6452	4672	0.364695	0.365828	0.203125
WSB	6470	4702	0.356114	0.352739	0.200169

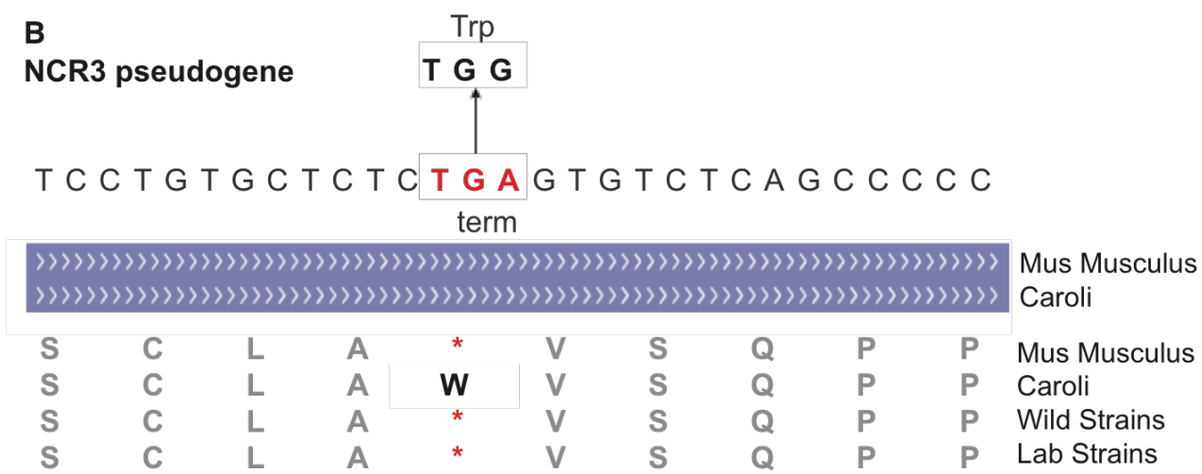
	Total	Essential	Lethal	Nonessential
MGI/IMPC	19,658	3,326	2,940	4,919
Ensembl	15,940	3,263	2,903	4,736

Genes that generate pseudogenes show on average a higher transcription level than the rest of protein coding genes.

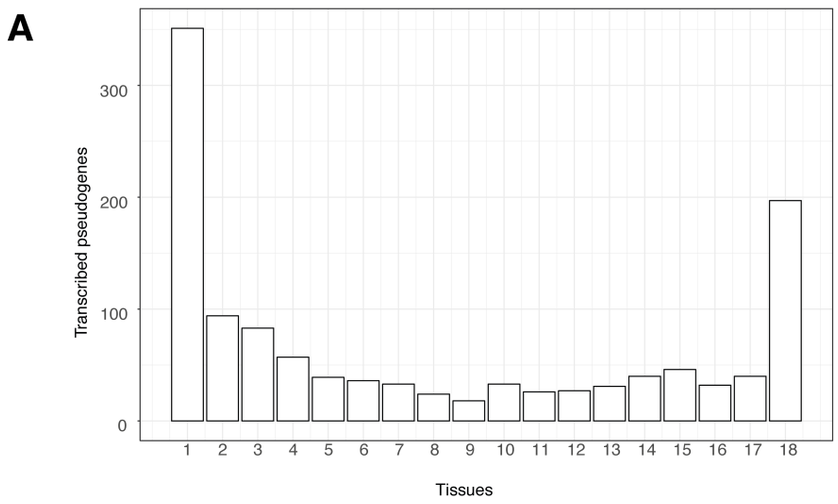
Parent genes are enriched amongst essential genes.

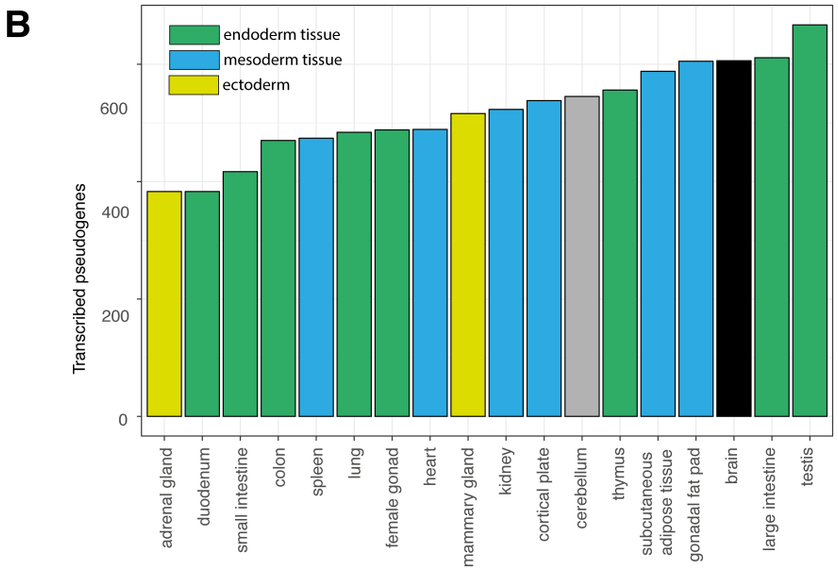
Loss and gain of function in human and mouse lineage



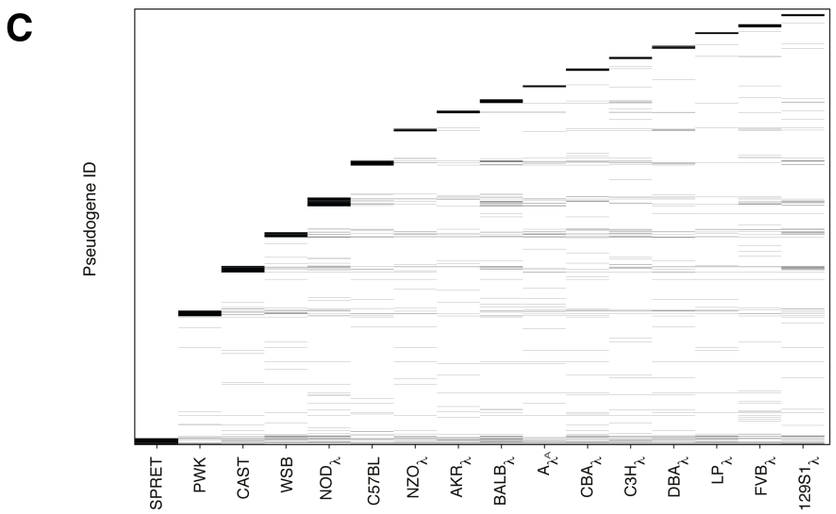


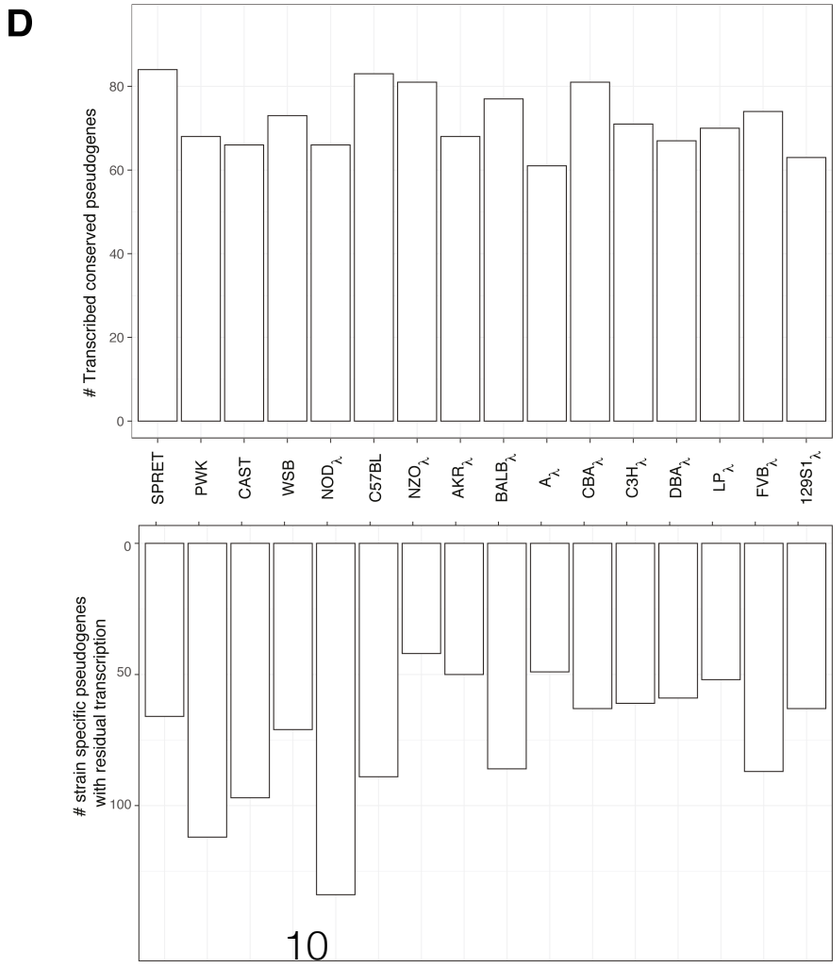
Pseudogene transcriptional activity





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





Strain specific pseudogene show higher level of transcription than pseudogene conserved across all the mouse strains

Summary

- The first draft of pseudogene annotation in 18 mouse strains and the reference genome
- On average 20% of pseudogenes are strain specific and 20% are ancestral pseudogenes, 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 localisation suggests multiple large scale genomic rearrangements between the out group - wild strains and the reference (lab strains) mouse genome.
- A significant proportion of pseudogenes show signs of transcriptional activity.

Future plans

1. Continue refining pseudogene annotations in the human and mouse reference assemblies.

2. Pseudogene annotation and characterization in the set of mouse genomes.

3. Pseudogene annotation in personal genomes.