Supplementary Information

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

Cristina Sisu*[1,2,3], Paul Muir*[4,5], Adam Frankish[6], Ian Fiddes[7], Mark Diekhans[7], David Thybert[6,8] Duncan T. Odom[9,10], Paul Flicek[6,10], Thomas Keane[6], Tim Hubbard[11], Jennifer Harrow[12], Mark Gerstein[1,2,13]

FIGURES

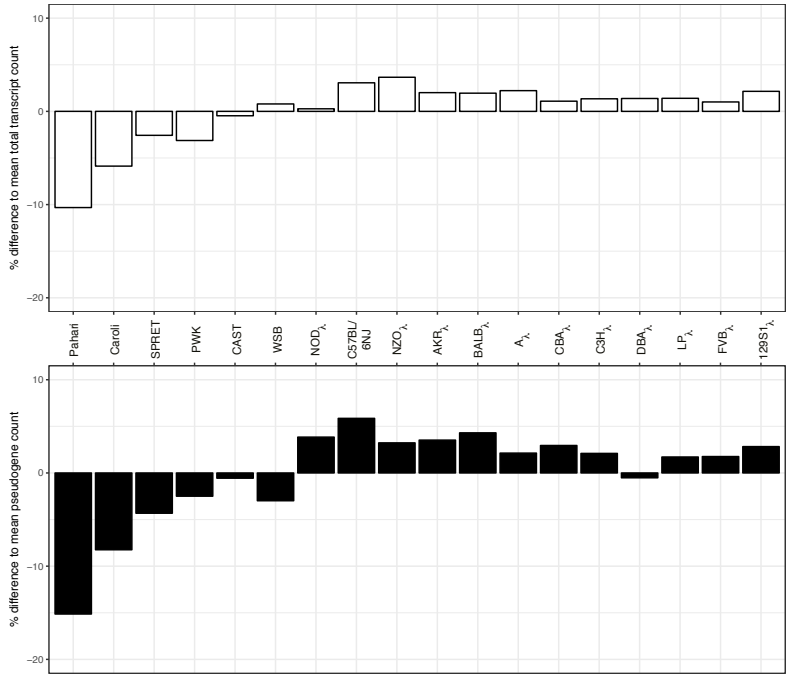

Fig SF1. A – The percentage difference between the number of pseudogene/conserved protein coding transcripts per strain and the average across all strains.

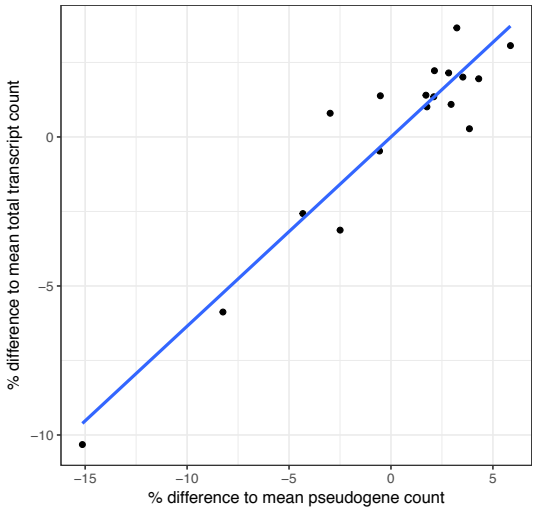

Fig SF1. B – Scatterplot of the percentage difference between the number of pseudogene/conserved protein coding transcripts per strain and the average across all strains. Pearson correlation coefficient = 0.94.

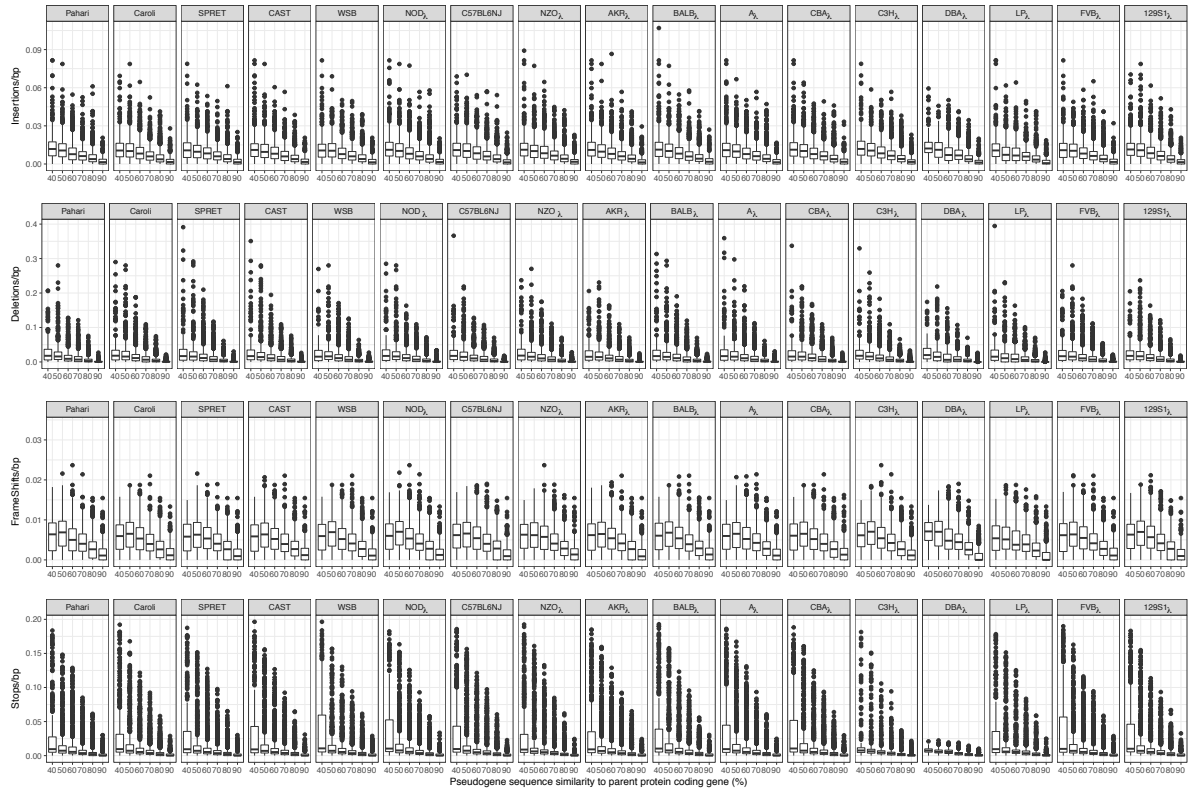

Fig SF2. A – Box plot distribution of pseudogene disablements per bp in 18 mouse strains.

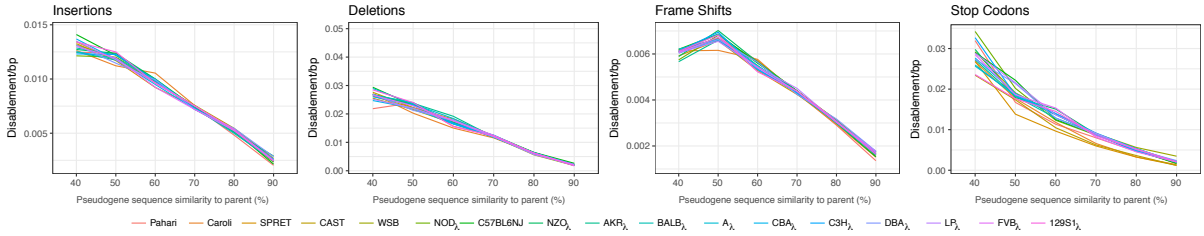

Fig SF2. B – Trends of disablement density per bp as function pseudogene sequence similarity to the parent in 18 mouse strains.

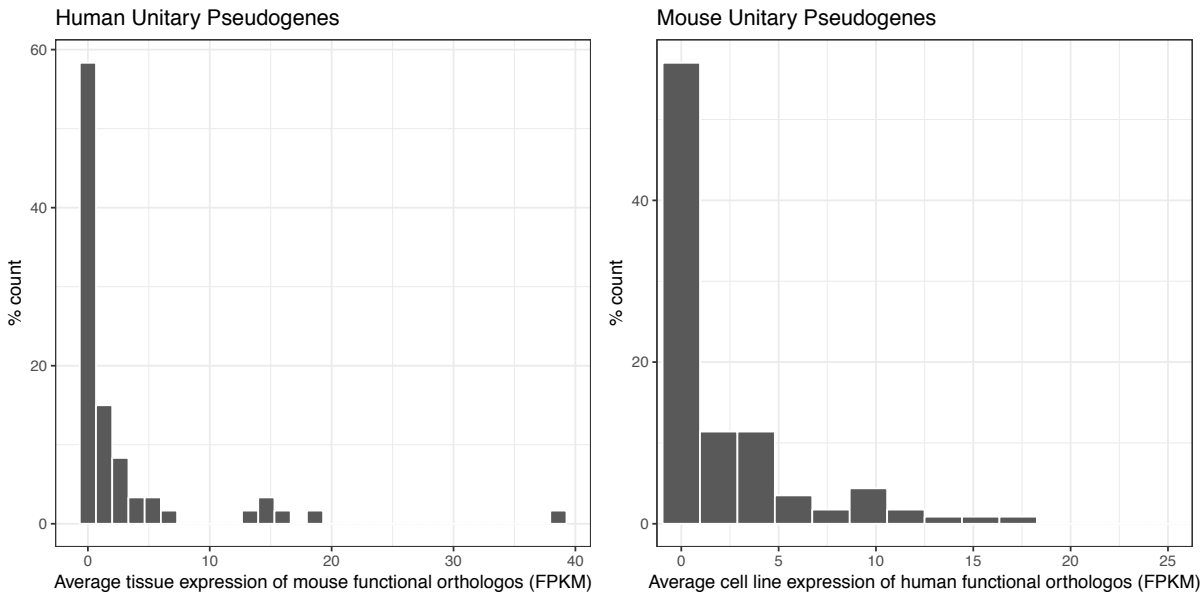

Fig SF3. A – Distribution of expression levels for the functional paralogs of unitary pseudogenes. The left hand graph gives the average tissue expression level for the mouse functional paralogs that are pseudogenised in human, while the right hand graph show the average ENCODE cell line expression level for the human functional paralogs that are unitary pseudogenes in mouse.

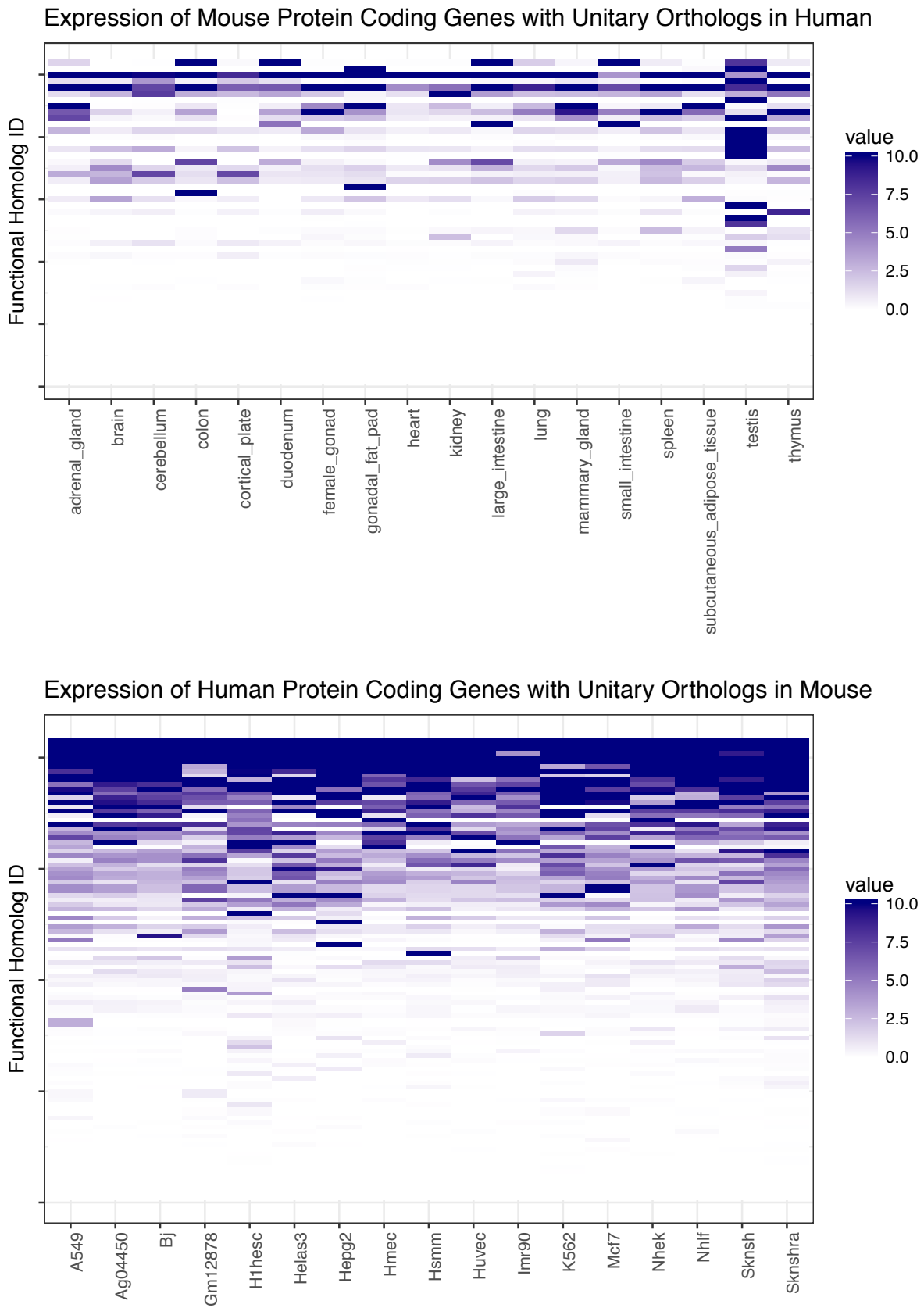

Fig SF3. B – Distribution of expression levels for the functional paralogs of unitary pseudogenes per tissue in mouse (top) and per ENCODE cell line in human(bottom). The colour scale top value corresponds to an expression score of greater or equal to 10FPKM.

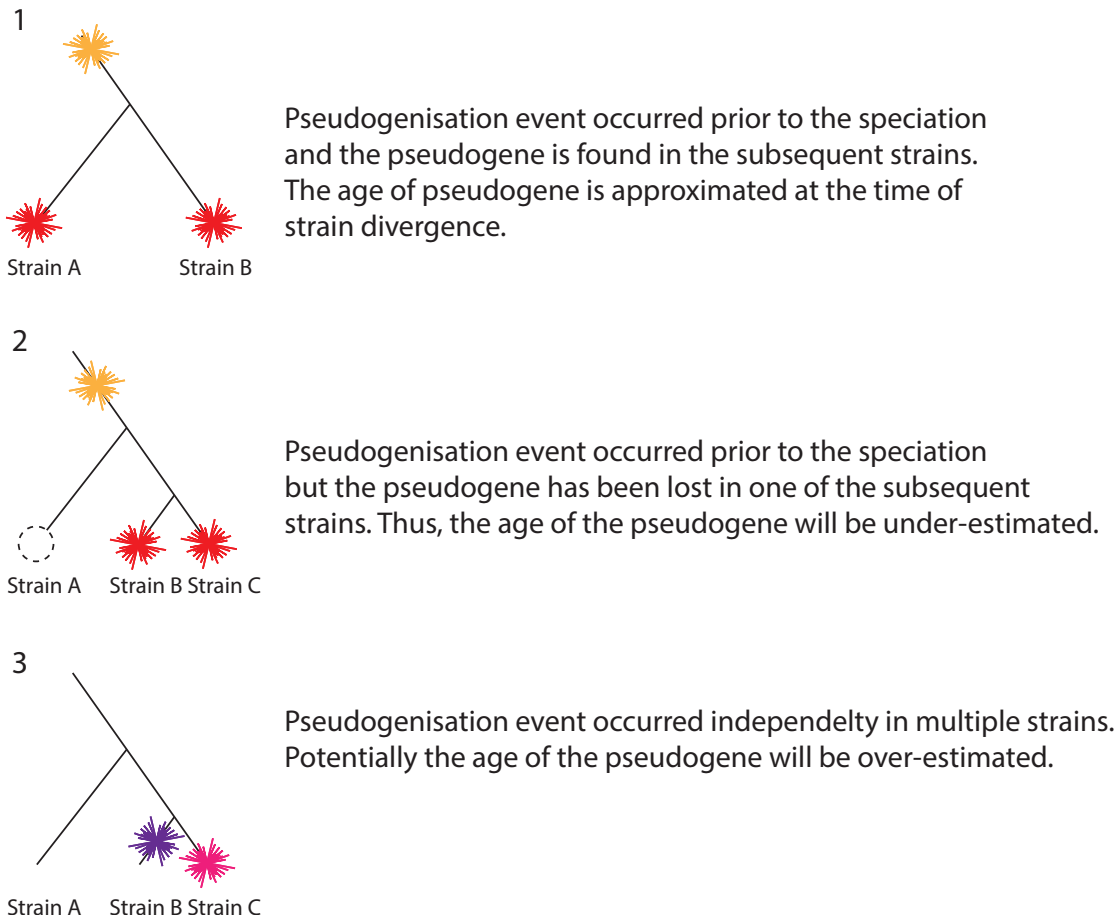

Fig SF3. C1 – Bias inducing events in estimating the age of pseudogene based on its presence or absence in various strains.

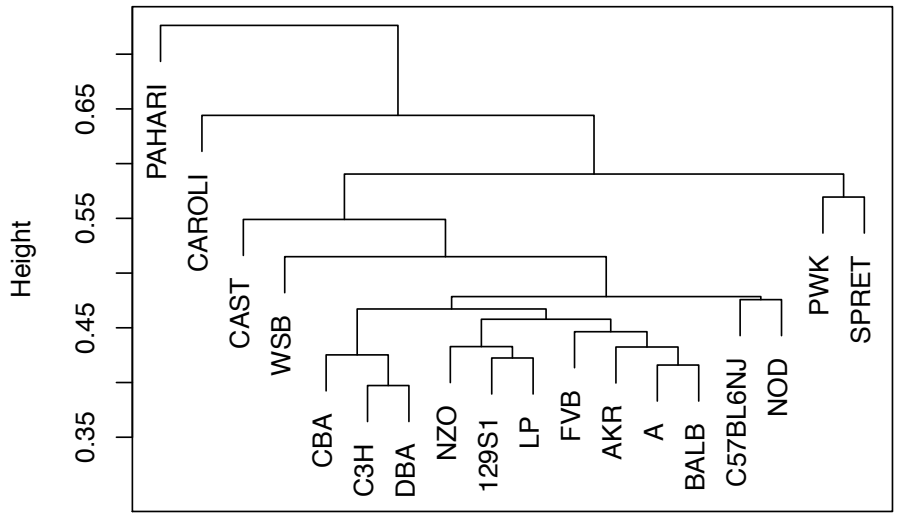

Fig SF3. C2 – Mouse lineage evolutionary tree based on the presence and absence of orthologous and strain specific pseudogenes across the strains using as input a binary matrix (1-pseudogene is present and 0 –the pseudogene is absent from the strain).

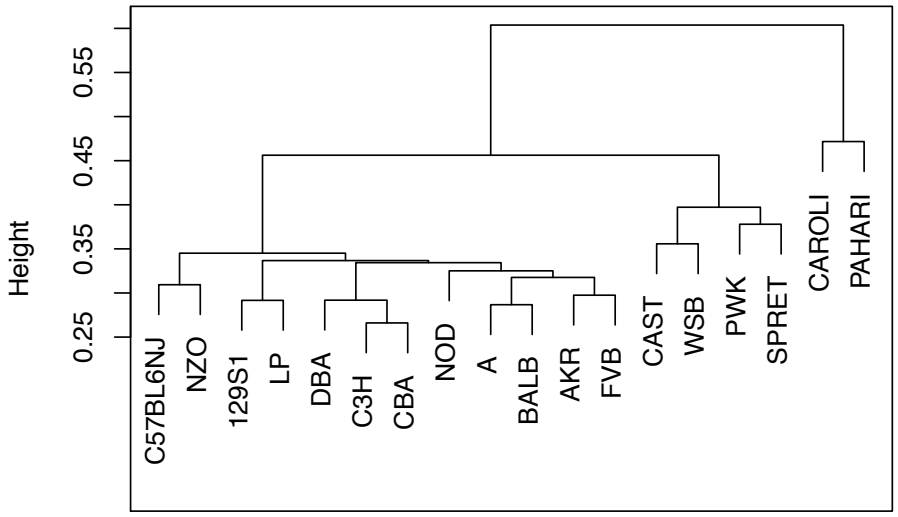

Fig SF3. C3 – Mouse lineage evolutionary tree based solely on the presence and absence of orthologous pseudogenes across the strains using as input a binary matrix (1-pseudogene is present and 0 –the pseudogene is absent from the strain).

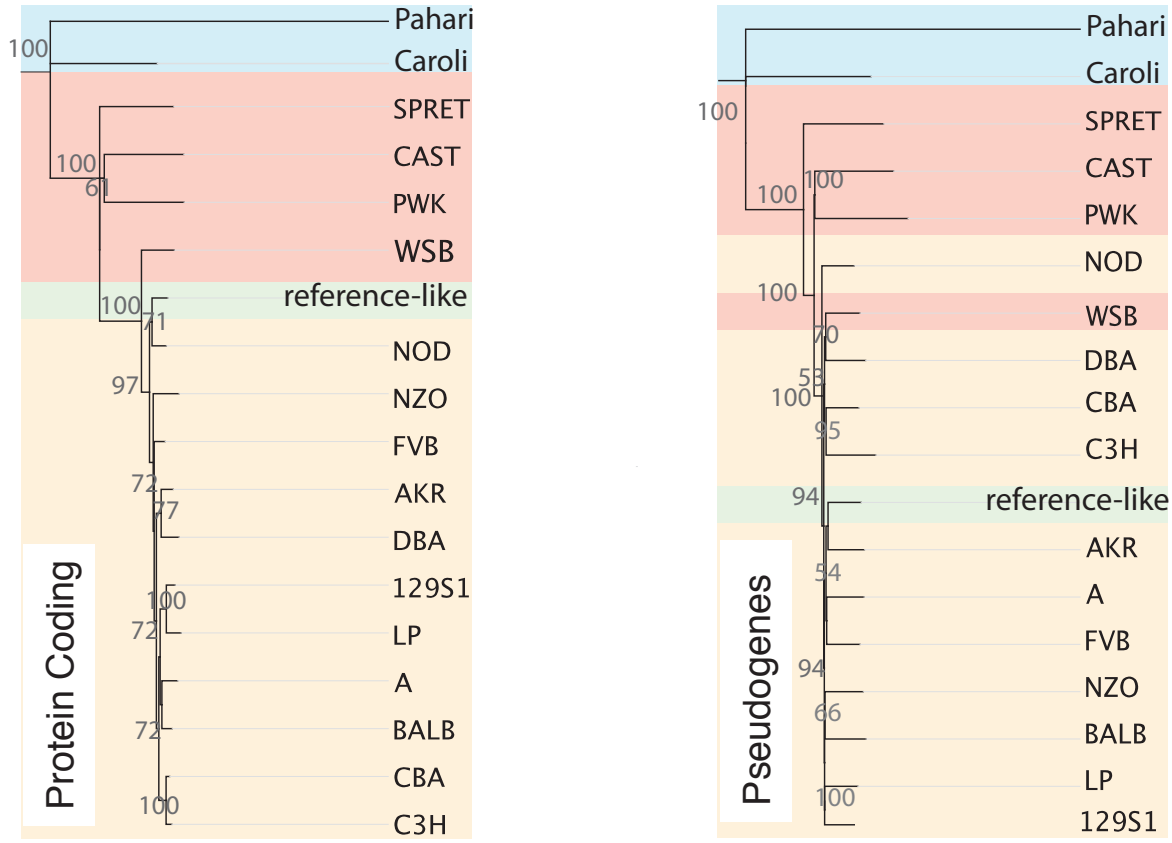

Fig SF3. D – Mirror of Figure 3C highlighting the phylogenetic trees of evolutionary conserved pseudogenes and pseudogenes parents with the associated bootstrap values on the branches.

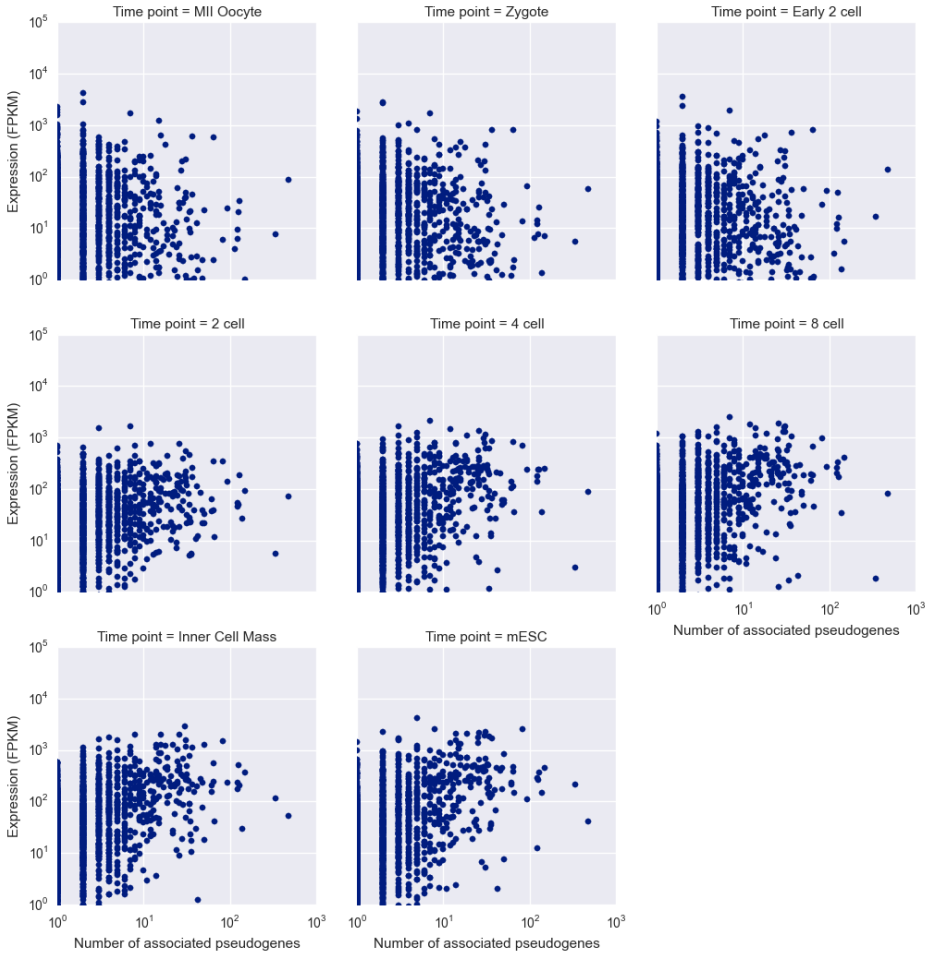

Fig SF4. A. Transcriptional activity of a gene vs the number of its associated pseudogenes at different early embryonic developmental time points.

All Genes (7,797)			
Embryonic Stage	**Slope**	**R^2**	**P-Value**
MII_oocyte	0.000680	.0004	0.0803
zygote	0.003195	0.0027	4.72e-06
early_2cell	0.003324	.0029	2.30e-06
2cell	0.016201	.0185	1.42e-33
4cell	0.013029	.0267	7.37e-48
8cell	0.011471	.0292	3.18e-52
ICM	0.012790	.0431	1.26e-76
mESC	0.012985	.0477	7.24e-85

Parent Genes (1,015)			
Embryonic Stage	**Slope**	**R^2**	**P-Value**
MII_oocyte	0.000195	0.000	0.930
zygote	0.003353	0.001	0.281
early_2cell	0.002932	0.001	0.298
2cell	0.011617	0.007	0.00634
4cell	0.011475	0.015	8.63e-05
8cell	0.010365	0.018	2.30e-05
ICM	0.016475	0.041	6.83e-11
mESC	0.015057	0.044	1.22e-11

Fig SF4. B. Regression statistics defining the transcriptional activity of a gene vs the number of its associated pseudogenes at different early embryonic developmental time points.

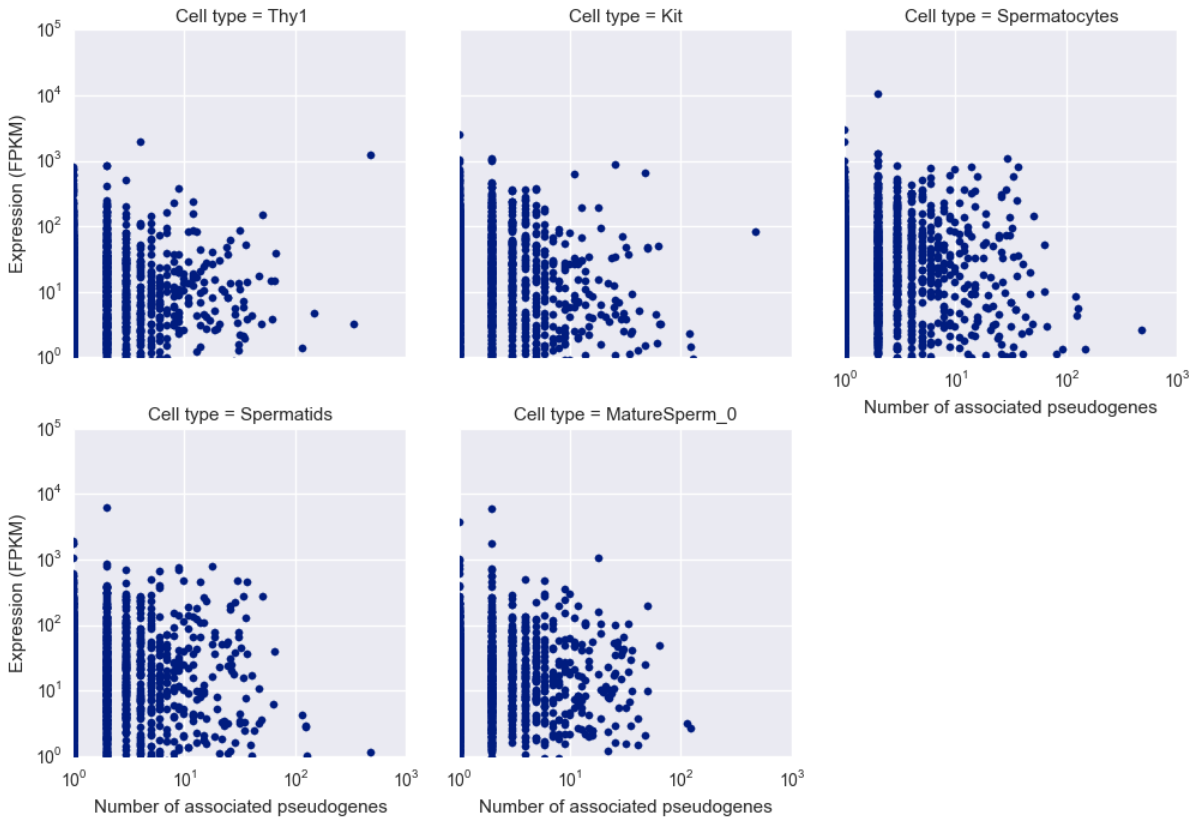

Fig SF4. C. Transcriptional activity of a gene vs the number of its associated pseudogenes during spermatogenesis.

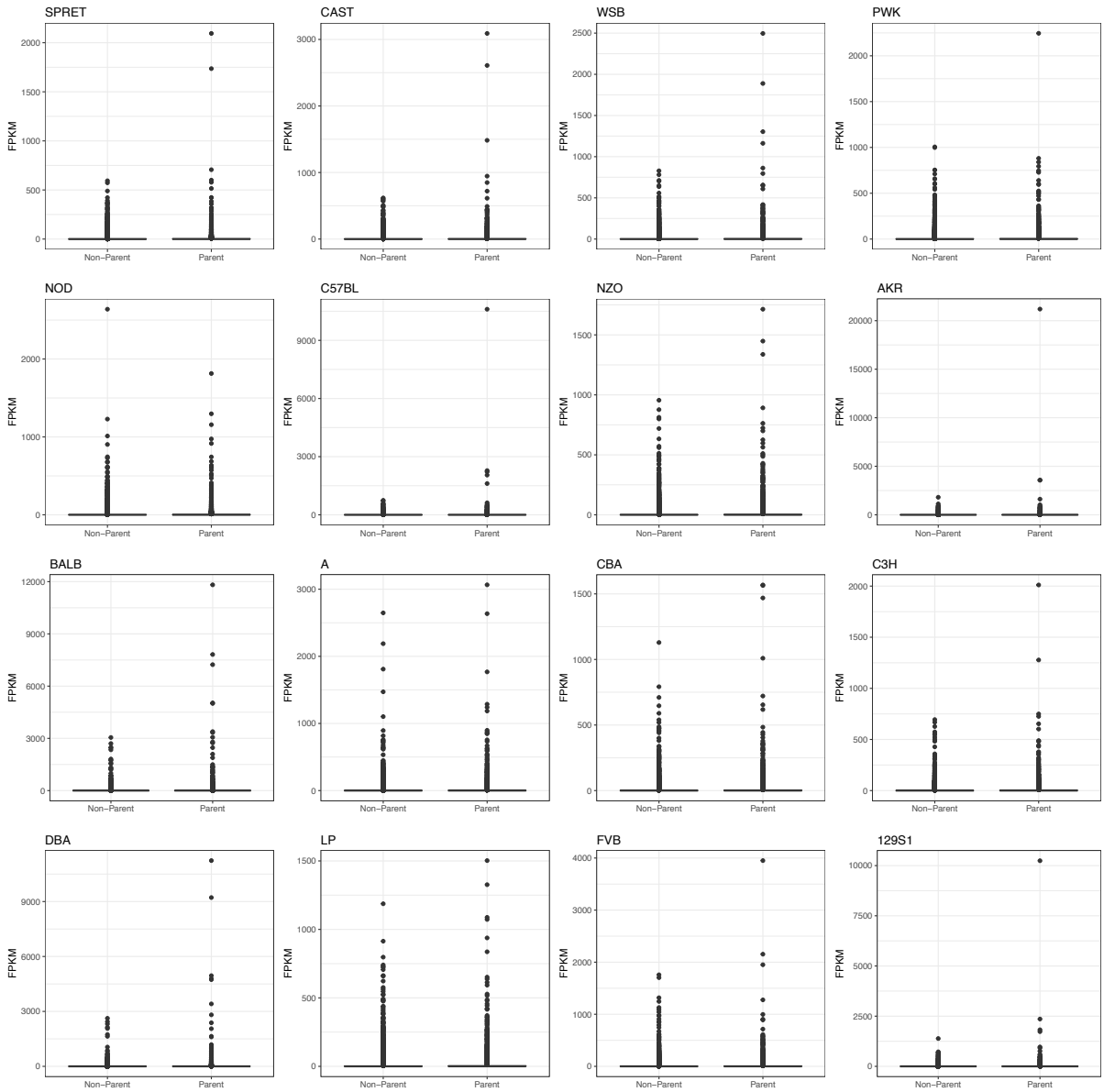

Fig SF4. D – Average expression levels in adult mouse brain for pseudogene parent and non-parent protein coding genes

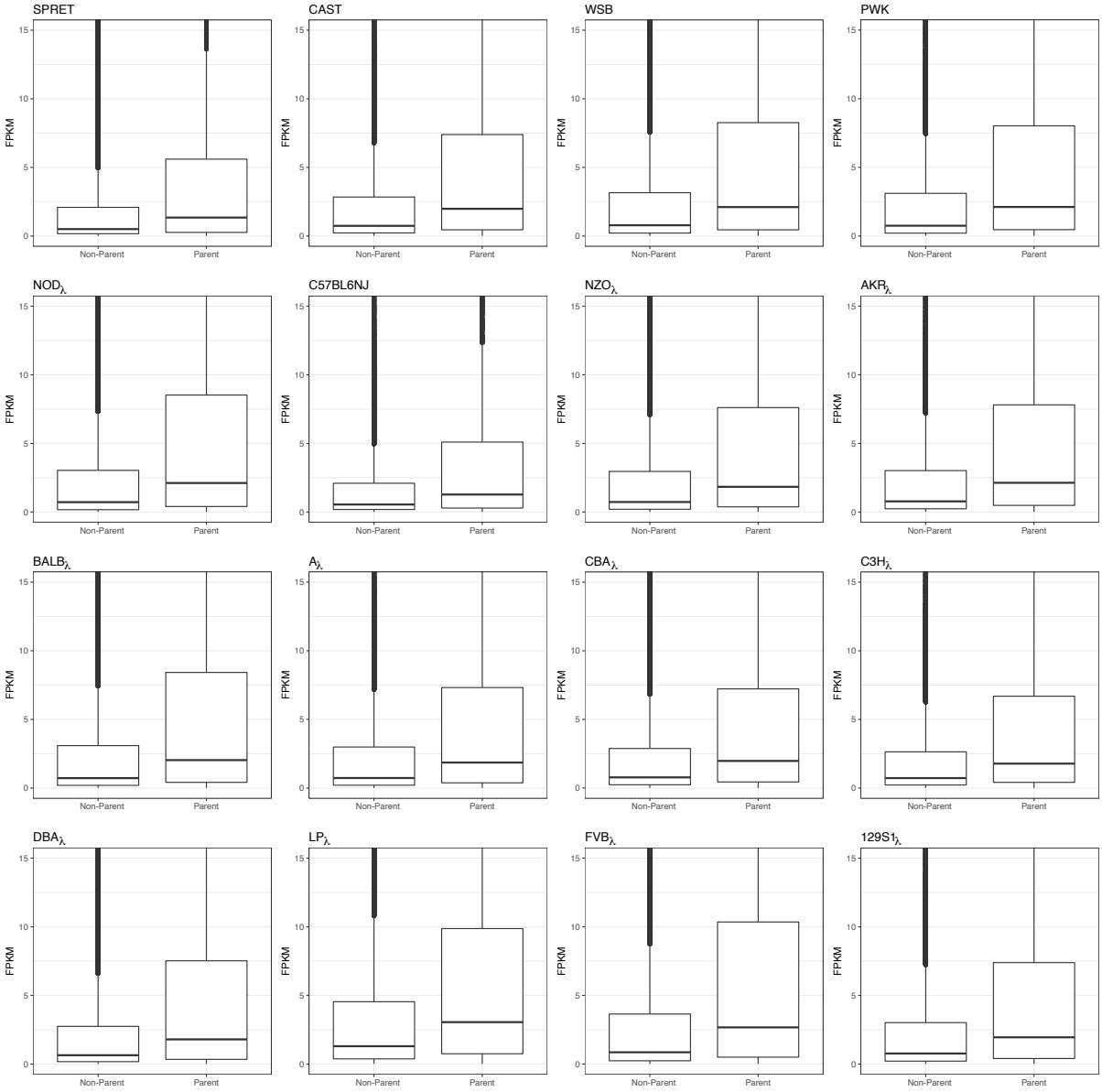

Fig SF4. E – zoom in: Average expression levels in adult mouse brain for pseudogene parent and non-parent protein coding genes.

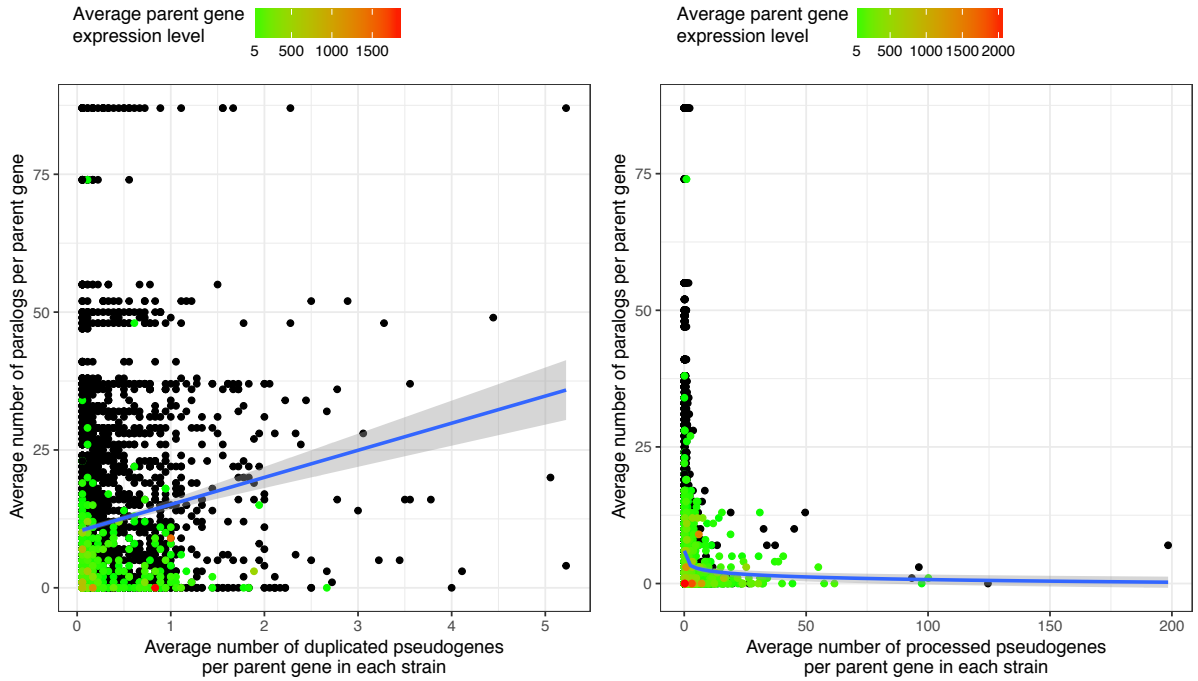

Fig SF5. A – Relationship between the number of pseudogenes and functional paralogs for a given parent gene (left – duplicated pseudogenes, right – processed pseudogenes). Fitting lines show a vague correlation between the number of functional vs disabled copies of a gene, with a linear fit for duplicated pseudogenes and a negative logarithmic fit for processed pseudogene. The gray area is the standard deviation. The dots are coloured by the average expression level of the parent gene in brain adult tissue in the range described in the heat scale above each figure. The black dots correspond to protein coding gene with an average expression level across the strains lower than 5 FPKM.

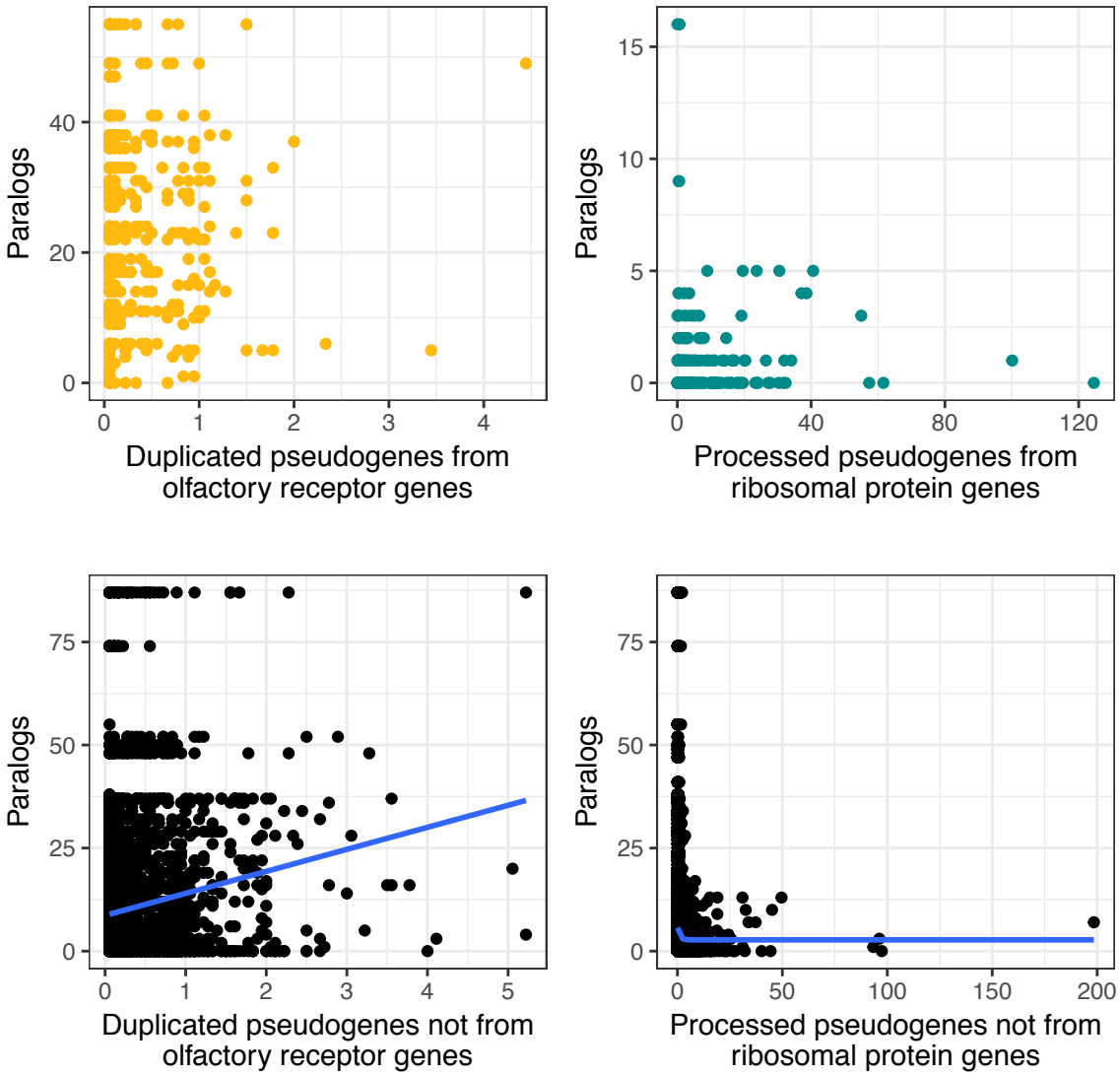

Fig SF5. B – Relationship between the number of pseudogenes and functional paralogs for a given parent gene (left – duplicated pseudogenes, right – processed pseudogenes) for olfactory receptors (OR) and ribosomal protein (RP) derived pseudogenes. The top left plot shows the distribution of OR pseudogenes vs paralogs of olfactory receptors per strain. Correspondingly, the top right plot shows the distribution of RP pseudogenes vs paralogs of ribosomal proteins per strain. The bottom plots show the distribution of the pseudogenes and paralogs that are not generated from olfactory receptor or ribosomal proteins. Correlation lines are drawn in blue.

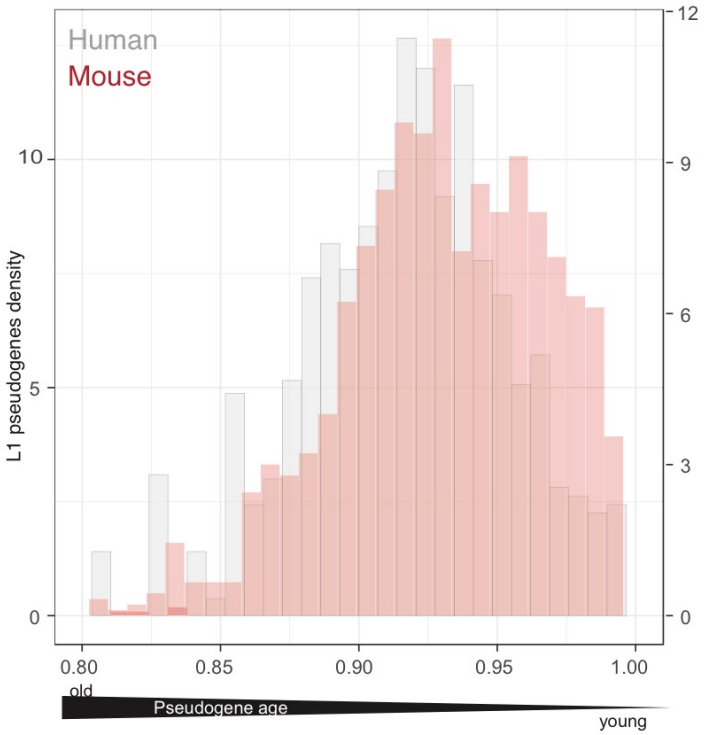

Fig SF5. C – Distribution of L1-flanked pseudogenes (y-axis) as function of age (x-axis). The pseudogene age is approximated as DNA sequence similarity to the parent gene.

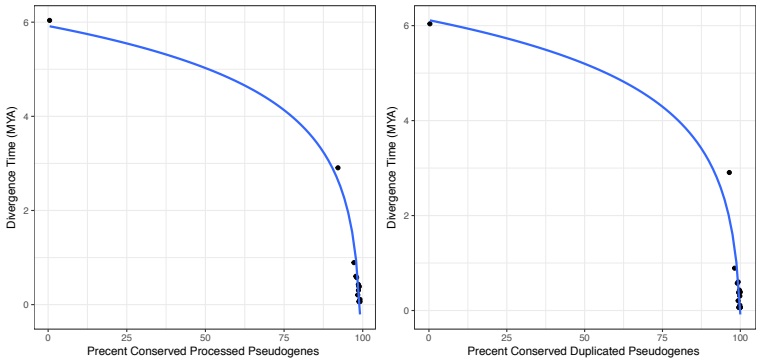

Fig SF6. Distribution of conserved pseudogenes as function of biotype and strain divergence. The "Misc" biotype includes unitary pseudogenes as well as pseudogene for which the biotype could not be accurately determined. All three pseudogene classes follow a logarithmic curve with respect to the strain divergence times, with the best fit being observed for processed pseudogenes.

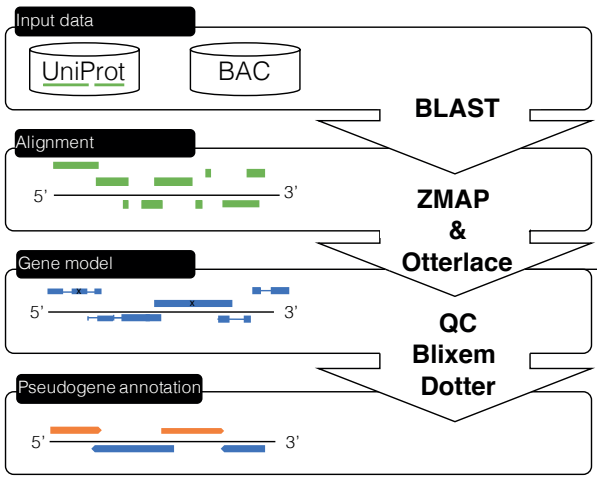

Fig SF7. Manual annotation curation workflow as previously described in [33, 34].

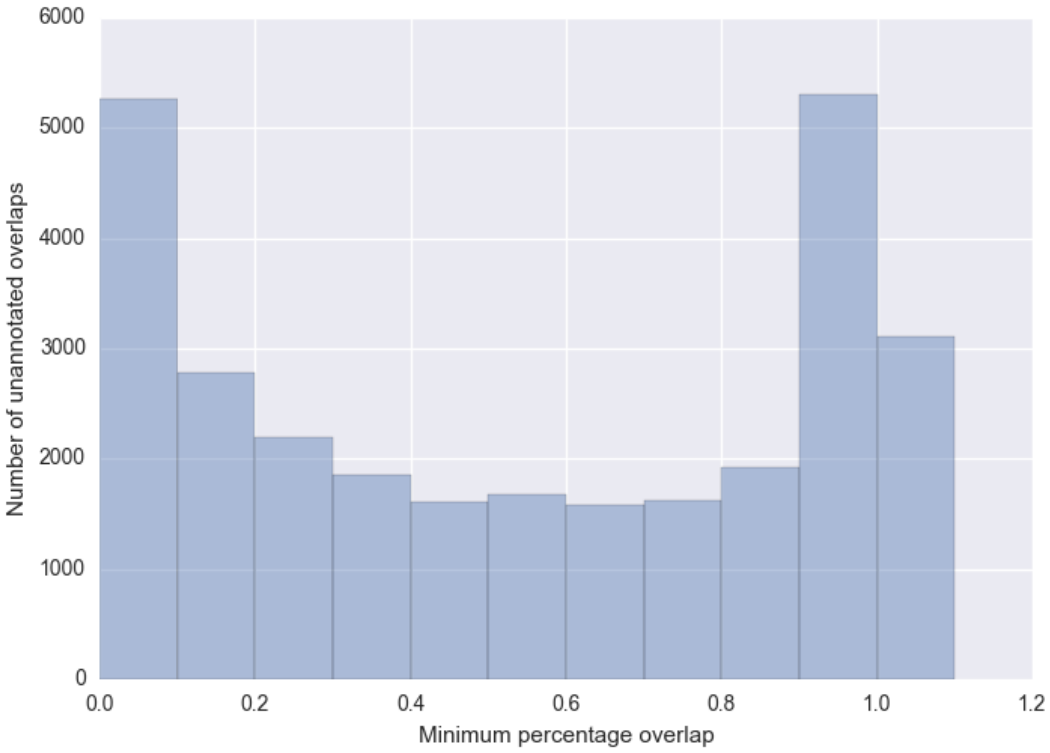

Fig SF8. Histogram of percentage overlap for lower of the reciprocal overlap cut-offs.

Tables

Table S1. A. Reference genome pseudogene annotation in mouse and human.

Organism	Manual curation (M)	PseudoPipe* (PP)	RetroFinder* (RF)	Union PP&RF	Intersection M&PP (%)
Mouse	10,524	18,659	18,467	26,103	8,786 (83.5)
Human	14,650	15,978	15,474	22,396	13,177 (89.9)

*Chromosomal assembled DNA only

Table S1. B. Reference genome automatic pseudogene annotation in mouse and human.

	PseudoPipe (PP)			RetroFinder (RF)	PP-RF overlap
	Autosomes	Sex Chr.	Others*		
Mouse	14,094	4,565	4,162	18,467	10,522
Human	14,638	1,341	2,054	15,474	9,057

*Includes patches, scaffolds, and unassembled DNA.

Table S1. C. Human and mouse pseudogene annotation summary.

	Human (v25)	Mouse (M12)
Total GENCODE	**14,650**	**10,524**
processed pseudogenes	10,725	7,486
unprocessed pseudogenes	3,400	2,625
unitary pseudogenes	214	34
polymorphic pseudogenes	51	77
ambiguous pseudogenes	21	99
Total PseudoPipe	**15,978 (+2,054*)**	**18,659 (+4,162*)**
processed pseudogenes	8,081 (+ 683*)	9,979 (+ 559*)
unprocessed pseudogenes	2,534 (+ 550*)	1,929 (+ 274*)
ambiguous pseudogenes	5,363 (+ 821*)	6,751 (+3,329*)

*Includes patches, scaffolds, and unassembled DNA.

Table S2. Mouse strains description and nomenclature.

Strain ID	Description	Class
Pahari	PAHARI/EiJ – Mus Pahari	Wild-derived outgroup
Caroli	CAROLI/EiJ – Mus Caroli	
SPRET	SPRET/EiJ – Mus Spretus	Wild-derived inbred strains
PWK	PWK/PhJ – Mus Musculus Musculus	
CAST	CAST/EiJ – Mus Musculus Castaneus	
WSB	WSB/EiJ – Mus Musculus Domesticus	
NOD_{λ}	NOD/ShiLtJ – Mus Musculus Non-obese Diabetic	Laboratory inbred strains
C57BL	C57BL/6NJ – Mus Musculus Black 6N	
NZO_{λ}	NZO/HlLtJ – Mus Musculus New Zealand Obese	
AKR_{λ}	AKR/J – Mus Musculus	
$BALB_{\lambda}$	BALB/cJ – Mus Musculus	
A_{λ}	A/J – Mus Musculus	
CBA_{λ}	CBA/J – Mus Musculus	
$C3H_{\lambda}$	C3H/HeJ – Mus Musculus	
DBA_{λ}	DBA/2J – Mus Musculus	
LP_{λ}	LP/J – Mus Musculus	
FVB_{λ}	FVB/NJ – Mus Musculus	
$129S1_{\lambda}$	129S1/SvImJ – Mus Musculus	

Table S3A: Estimation of the total number of pseudogenes according to PseudoPipe per strain.

Strain	PseudoPipe predictions	Input protein coding transcripts conserved between mouse reference and strains	% Protein coding transcripts conserved	% Pseudogenes annotated with respect to the total number of pseudogenes in reference genome	Estimate of total number of pseudo-pipe pseudo-genes
Mouse	18659	56999	100.00	100.00	**18659**
C57BL/6NJ	14722	47145	82.71	79.27	**18659**
PAHARI	12414	41022	71.97	68.97	**18082**
CAROLI	13399	43056	75.54	72.39	**18595**
SPRET	14170	44567	78.19	74.93	**18998**
PWK	14485	44313	77.74	74.50	**19532**
CAST	14427	45527	79.87	76.55	**18935**
WSB	14202	46107	80.89	77.52	**18405**
NOD_{λ}	14965	45869	80.47	77.12	**19495**
NZO_{λ}	13909	47417	83.19	79.72	**17527**
AKR_{λ}	14380	46662	81.86	78.45	**18414**
$BALB_{\lambda}$	14393	46636	81.82	78.41	**18441**
A_{λ}	13823	46760	82.04	78.62	**17664**
CBA_{λ}	14479	46243	81.13	77.75	**18709**
$C3H_{\lambda}$	14400	46360	81.33	77.95	**18560**
DBA_{λ}	13872	46375	81.36	77.97	**17874**
LP_{λ}	13923	46384	81.38	77.99	**17936**
FVB_{λ}	14202	46205	81.06	77.69	**18366**
$129S1_{\lambda}$	13820	46726	81.98	78.56	**17673**

Table S3B: Distribution of curated pseudogenes in each strain based on their confidence level.

Strain	Level 1	Level 2	Level 3
C57BL/6NJ	5615	993	6597
PAHARI	2971	1254	6361
CAROLI	3860	1224	6362
SPRET	4444	980	6511
PWK	4630	865	6668
CAST	4694	1003	6707
WSB	4869	873	6360
NOD_{λ}	5285	937	6732
NZO_{λ}	5592	1048	6237
AKR_{λ}	5289	996	6629
$BALB_{\lambda}$	5344	939	6728
A_{λ}	5295	997	6448
CBA_{λ}	5231	898	6713
$C3H_{\lambda}$	5201	917	6618
DBA_{λ}	5282	908	6219
LP_{λ}	5199	1015	6474
FVB_{λ}	5257	977	6460
$129S1_{\lambda}$	5284	1042	6501

Table S4. Unitary pseudogenes in human and mouse. (see SupTable_S4_Unitary.xlsx available at http://mouse.pseudogene.org/Supplement/)

Table S5. Pseudogene family and clan characterization. (see SupTable_S5_Family.xlsx available at http://mouse.pseudogene.org/Supplement/)

Table S6. Unitary pseudogenes in mouse strains. (see SupTable_S6_StarinsUnitary.xlsx available at http://mouse.pseudogene.org/Supplement/)

Table S7. A. Enrichment of pseudogene parent gene class in essential genes.

Pseudogenes	Genes	Essential	Nonessential	Odds Ratio	p-Value
Total	Parent	1162	1061	1.93	$7.7*10^{-39}$
	Non-Parent	2050	3620		
Processed	Parent	1034	869	2.08	$2.3*10^{-43}$
	Non-Parent	2178	3812		
Duplicated	Parent	334	349	1.44	$6.0*10^{-6}$
	Non-Parent	2878	4332		

Table S7. B. Correlations between gene essentiality and parent gene status controlling for transcription level.

	Linear Prob. Model	Probit	Probit Marginal Effect
Parent gene (Y/N)	0.2035 (0.0168)	0.5108 (0.0441)	0.1943 (0.016)
Transcription	0.0003 (0.0001)	0.0010 (0.0002)	0.0004 (8.11e-05)

Marginal effect for probit (column 3) calculated at mean values for each independent variable. Number of observation: 7797. Standard errors are given in parentheses. Parent gene (Y/N) is a binary categorical variable that is equal to 1 if a gene has any associated pseudogenes and 0 if not.

Table S8. Encode transcription data. (see SupTable_S8_EncodeTranscription.xlsx available at http://mouse.pseudogene.org/Supplement/)