Mouse Pseudogenes
~ Update ~

Cristina Sisu

Gerstein Lab
Yale

GENCODE Call 5th April 2016

Mus castaneus

Mus spretus

129S1/SvImJ
129S5/SvEv
129P2/OlaHsd
LP/J
DBA/2J
C3H/HeJ
CBA/J
A/J
BALB/cJ
AKR/J
NZO/lLtJ
C57BL/6NJ
NOD/ShiLtJ
0.009MYA
WSB/Eij
CAST/Eij
0.5MYA
PWK/PhJ
1.5MYA
SPRET/EiJ
3MYA
CAROLI/Eij
6MYA
PAHARI/EiJ
12MYA
Rat

Classical laboratory strains

Wild strains

Outgroup *Mus* species

Rat

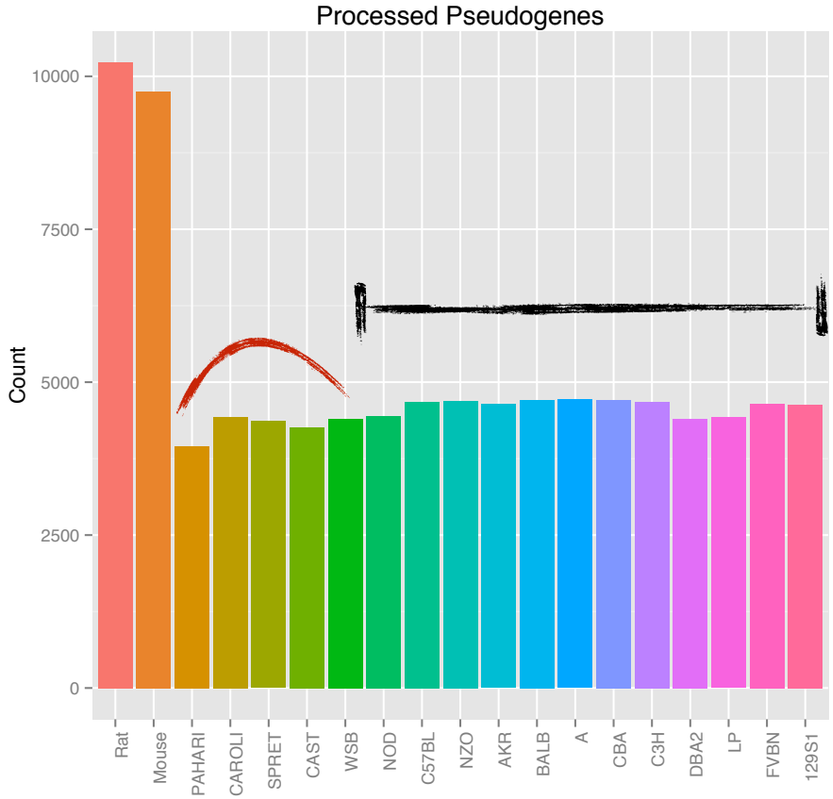
Mus pahari

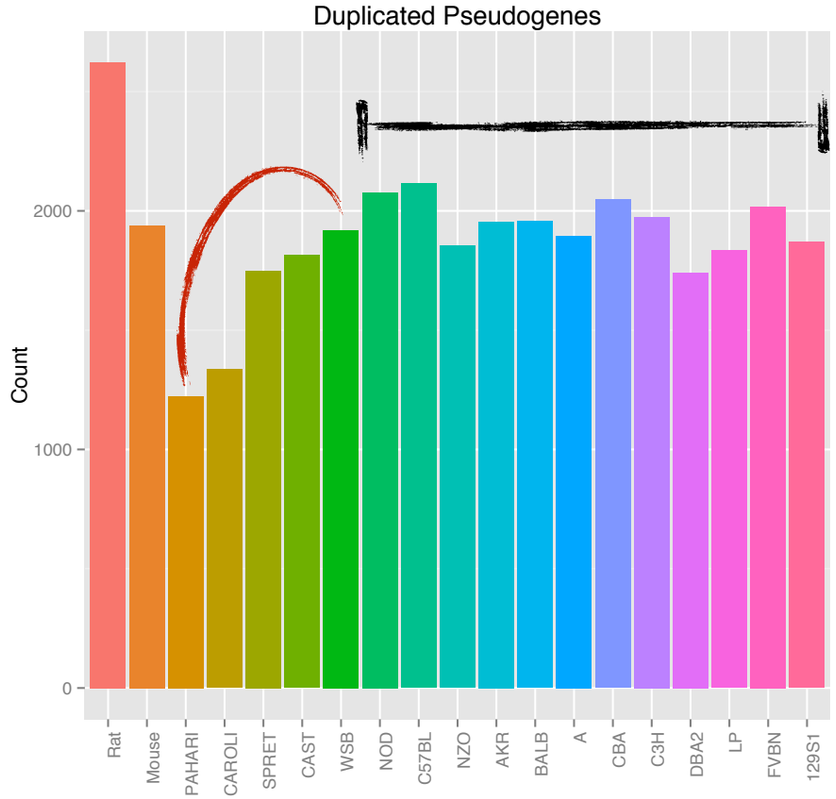
Annotation

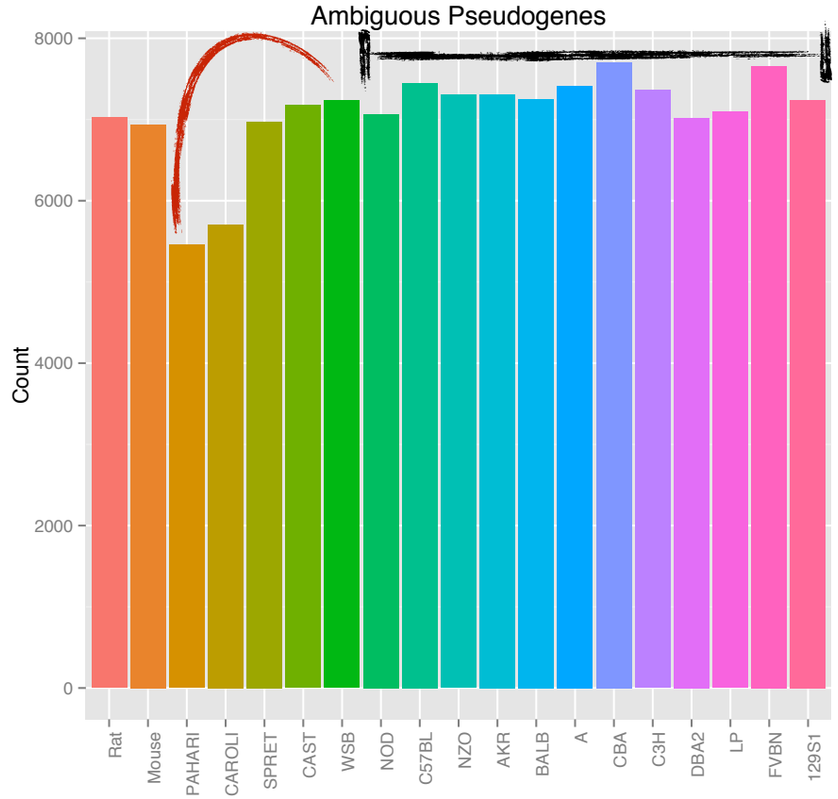
- REL 1509 assembly
- 1509 consensus gene sequences from Ian Fides (http://hgwdev.cse.ucsc.edu/~ifiddes/1509_consensus_gene_set)
- Mouse reference genome peptide annotation file from Ensembl 83 Mus_musculus.GRCm38.pep.all.fa
    - use only peptides from genes present in the consensus gene set for each strain

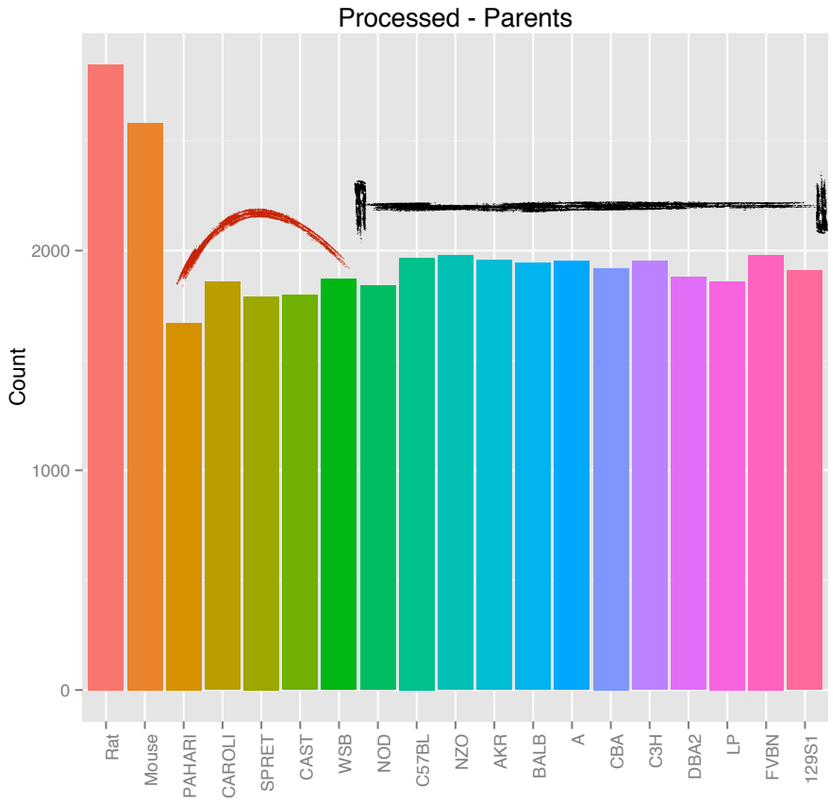
Strain	Rat (ENS83)	PAHARI	CAROLI	SPRET	CAST	WSB	NOD	C57BL	Mouse (MM8)
Processed	10226	3954	4429	4362	4261	4397	4442	4679	**9748**
Processed Parents	2847	1671	1857	1791	1797	1872	1840	1967	**2581**
Duplicated	2622	1223	1336	1748	1815	1918	2077	2114	**1940**
Duplicated Parents	1554	950	983	1114	1042	1199	1248	1280	**1146**
Ambiguous	7023	5459	5707	6973	7177	7232	7069	7448	**6936**
Ambiguous Parents	3674	2674	2785	3249	3098	3221	3251	3285	**2884**
Total Pseudognes	19872	10637	11473	13084	13254	13548	13589	14242	**18625**
Total Parents	6236	4285	4531	4836	4615	4850	4886	5024	**5267**
Consensus transcritps	**28753**	41022	43056	44567	45527	46107	45869	47145	**56999**

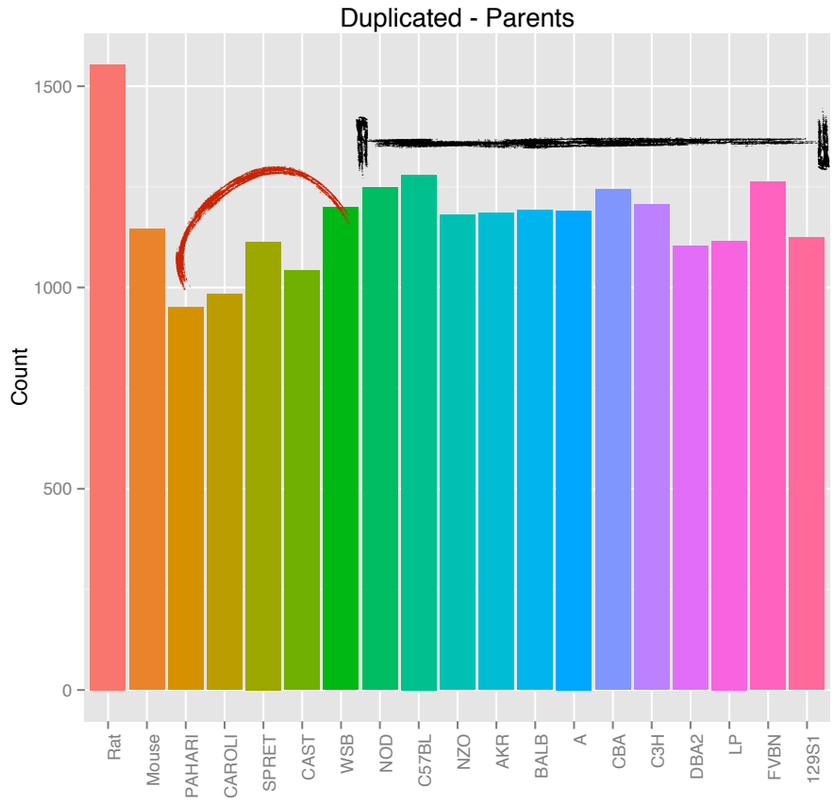
Strain	AKR	BALB	A	CBA	C3H	DBA2	LP	FVBN	Mouse (MM8)
Processed	4641	4706	4715	4713	4675	4398	4425	4651	**9748**
Processed Parents	1957	1944	1951	1920	1953	1878	1860	1980	**2581**
Duplicated	1954	1957	1896	2047	1975	1741	1834	2019	**1940**
Duplicated Parents	1185	1192	1191	1244	1206	1103	1116	1262	**1146**
Ambiguous	7313	7251	7414	7702	7363	7021	7094	7656	**6936**
Ambiguous Parents	3323	3190	3320	3318	3241	3096	3182	3392	**2884**
Total Pseudognes	13909	13915	14026	14463	14014	13161	13354	14327	**18625**
Total Parents	5052	4922	5027	4985	4954	4745	4795	5156	**5267**
Consensus transcritps	46662	46636	46760	46243	46360	46375	46384	46205	**56999**

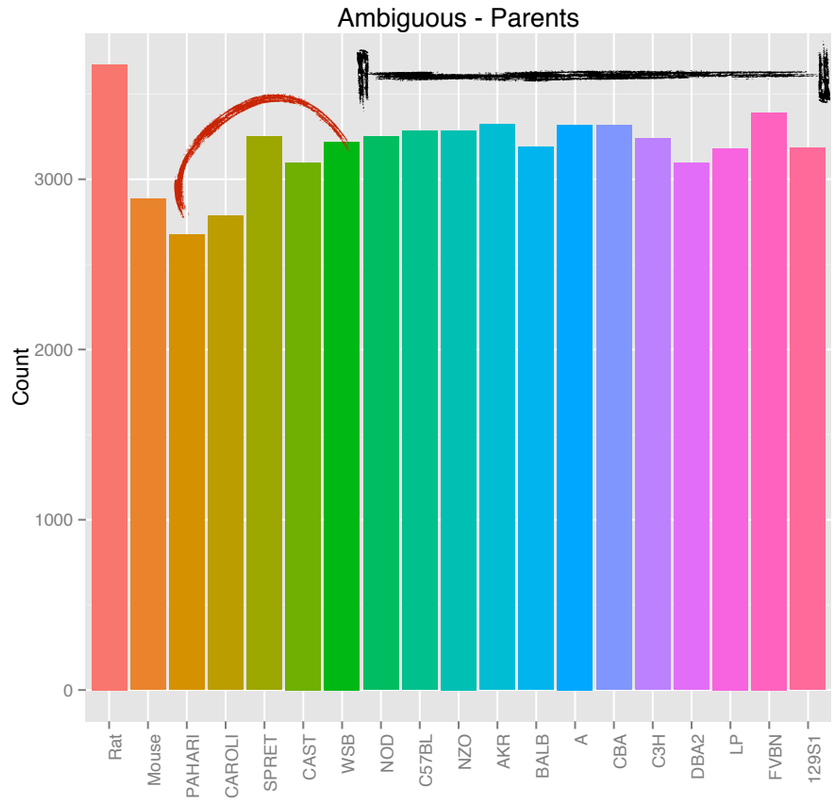

Processed Pseudogenes
Count
10000
7500
5000
2500
0
Rat
Mouse
PAHARI
CAROLI
SPRET
CAST
WSB
NOD
C57BL
NZO
AKR
BALB
A
CBA
C3H
DBA2
LP
FVBN
129S1

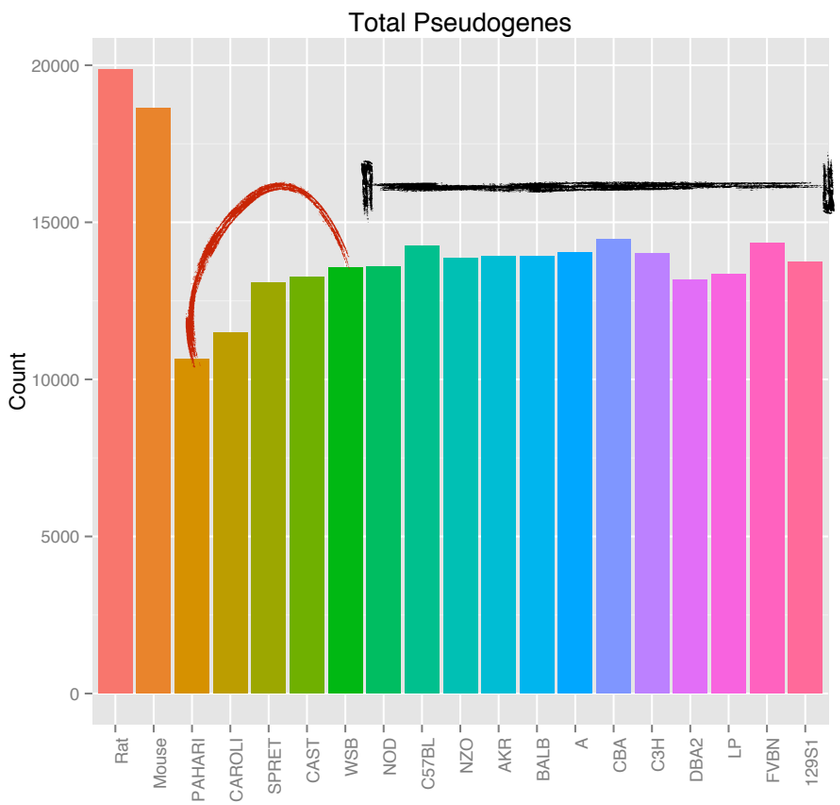

Duplicated Pseudogenes
Count
2000
1000
0
Rat
Mouse
PAHARI
CAROLI
SPRET
CAST
WSB
NOD
C57BL
NZO
AKR
BALB
A
CBA
C3H
DBA2
LP
FVBN
129S1

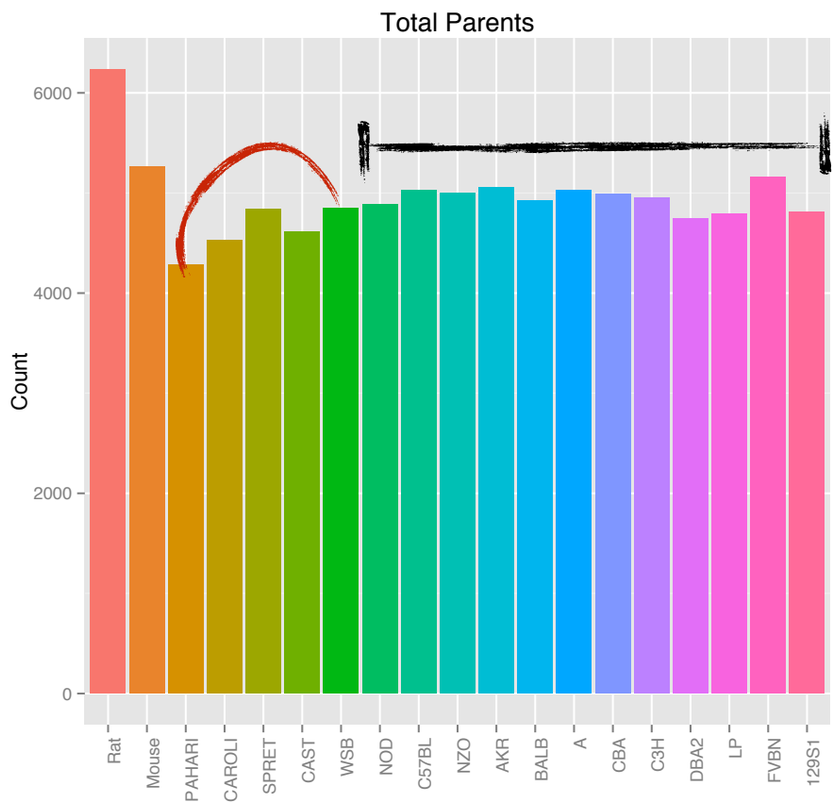

Ambiguous Pseudogenes
Count
8000
6000
4000
2000
0
Rat
Mouse
PAHARI
CAROLI
SPRET
CAST
WSB
NOD
C57BL
NZO
AKR
BALB
A
CBA
C3H
DBA2
LP
FVBN
129S1

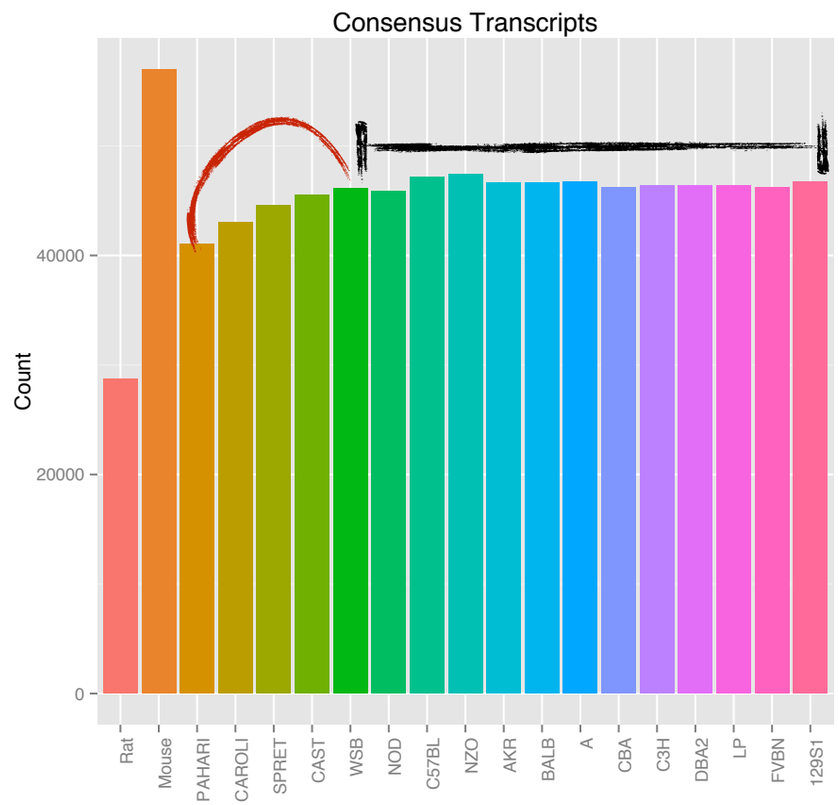

Processed - Parents
Count
2000
1000
0
Rat
Mouse
PAHARI
CAROLI
SPRET
CAST
WSB
NOD
C57BL
NZO
AKR
BALB
A
CBA
C3H
DBA2
LP
FVBN
129S1

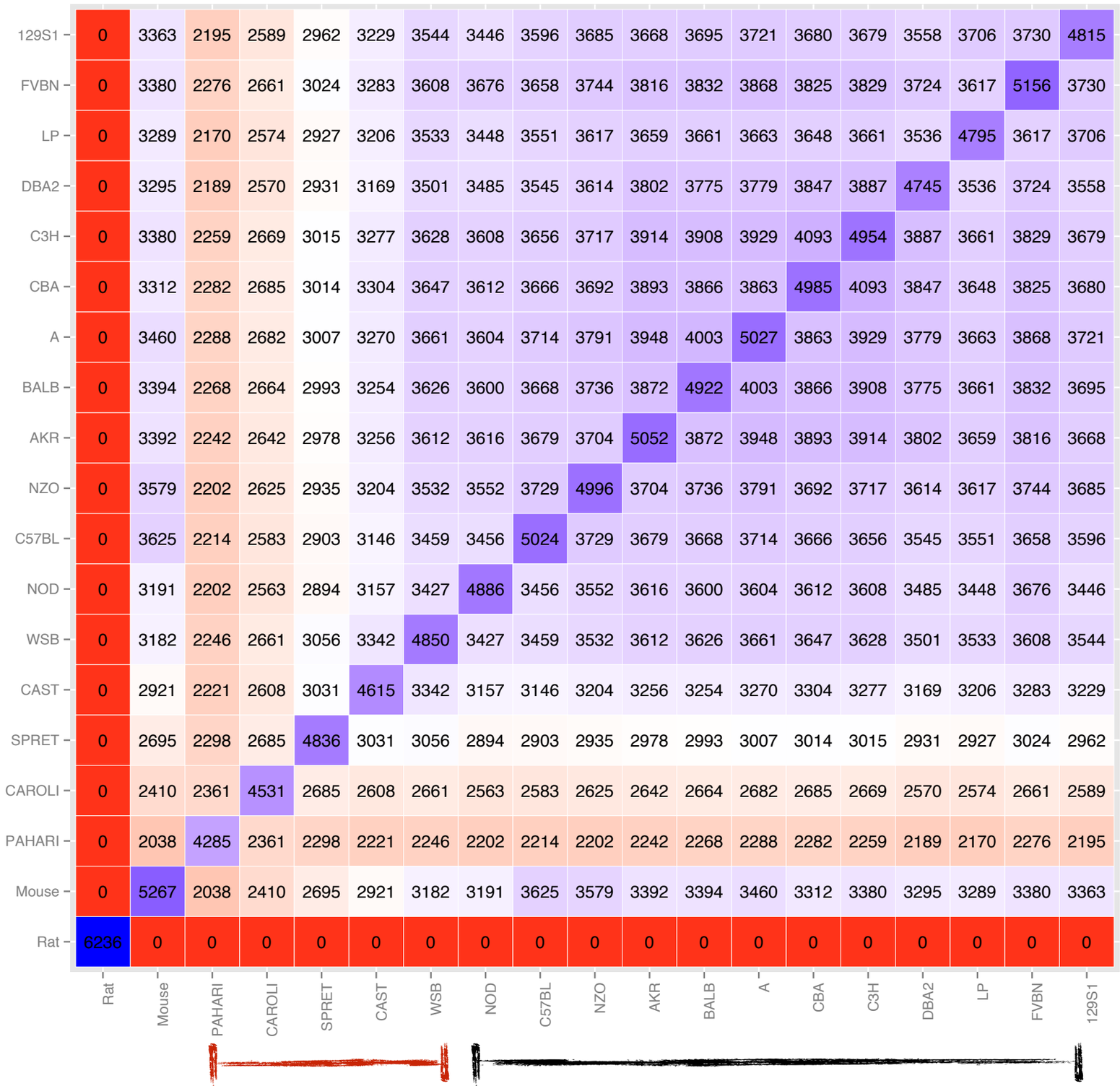

Duplicated - Parents
Count
1500
1000
500
0
Rat
Mouse
PAHARI
CAROLI
SPRET
CAST
WSB
NOD
C57BL
NZO
AKR
BALB
A
CBA
C3H
DBA2
LP
FVBN
129S1


Ambiguous - Parents
Count
3000
2000
1000
0
Rat
Mouse
PAHARI
CAROLI
SPRET
CAST
WSB
NOD
C57BL
NZO
AKR
BALB
A
CBA
C3H
DBA2
LP
FVBN
129S1


Total Pseudogenes
Count
20000
15000
10000
5000
0
Rat
Mouse
PAHARI
CAROLI
SPRET
CAST
WSB
NOD
C57BL
NZO
AKR
BALB
A
CBA
C3H
DBA2
LP
FVBN
129S1

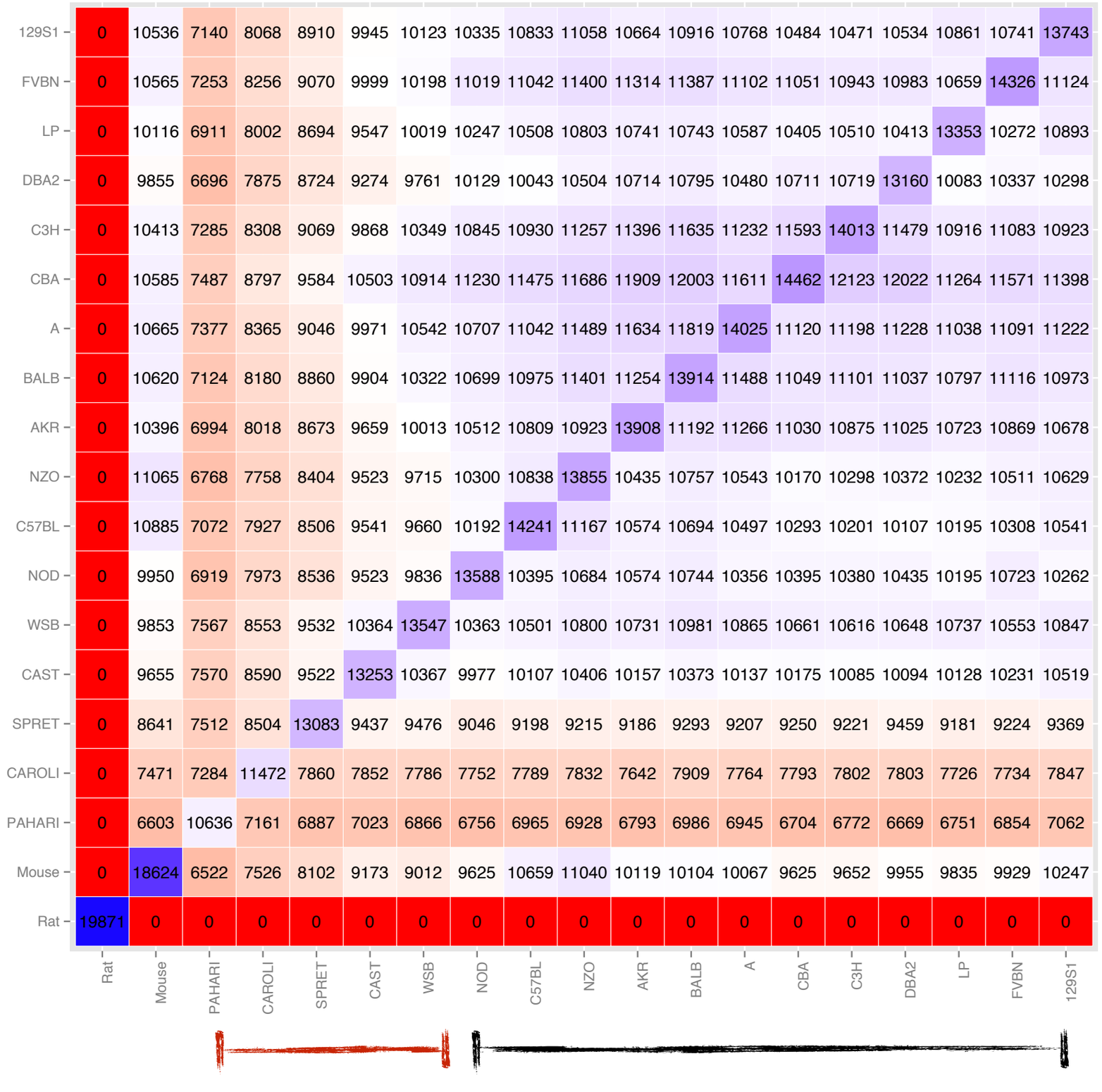

Total Parents
Count
6000
4000
2000
0
Rat
Mouse
PAHARI
CAROLI
SPRET
CAST
WSB
NOD
C57BL
NZO
AKR
BALB
A
CBA
C3H
DBA2
LP
FVBN
129S1


Consensus Transcripts
Count
40000
20000
0
Rat
Mouse
PAHARI
CAROLI
SPRET
CAST
WSB
NOD
C57BL
NZO
AKR
BALB
A
CBA
C3H
DBA2
LP
FVBN
129S1



60% of the parent genes are shared between any two strains.

The number of shared genes raises considerably for the lab strains



The majority of pseudogenes in the lab strains are results of conserved parent genes across all the strains

Top families

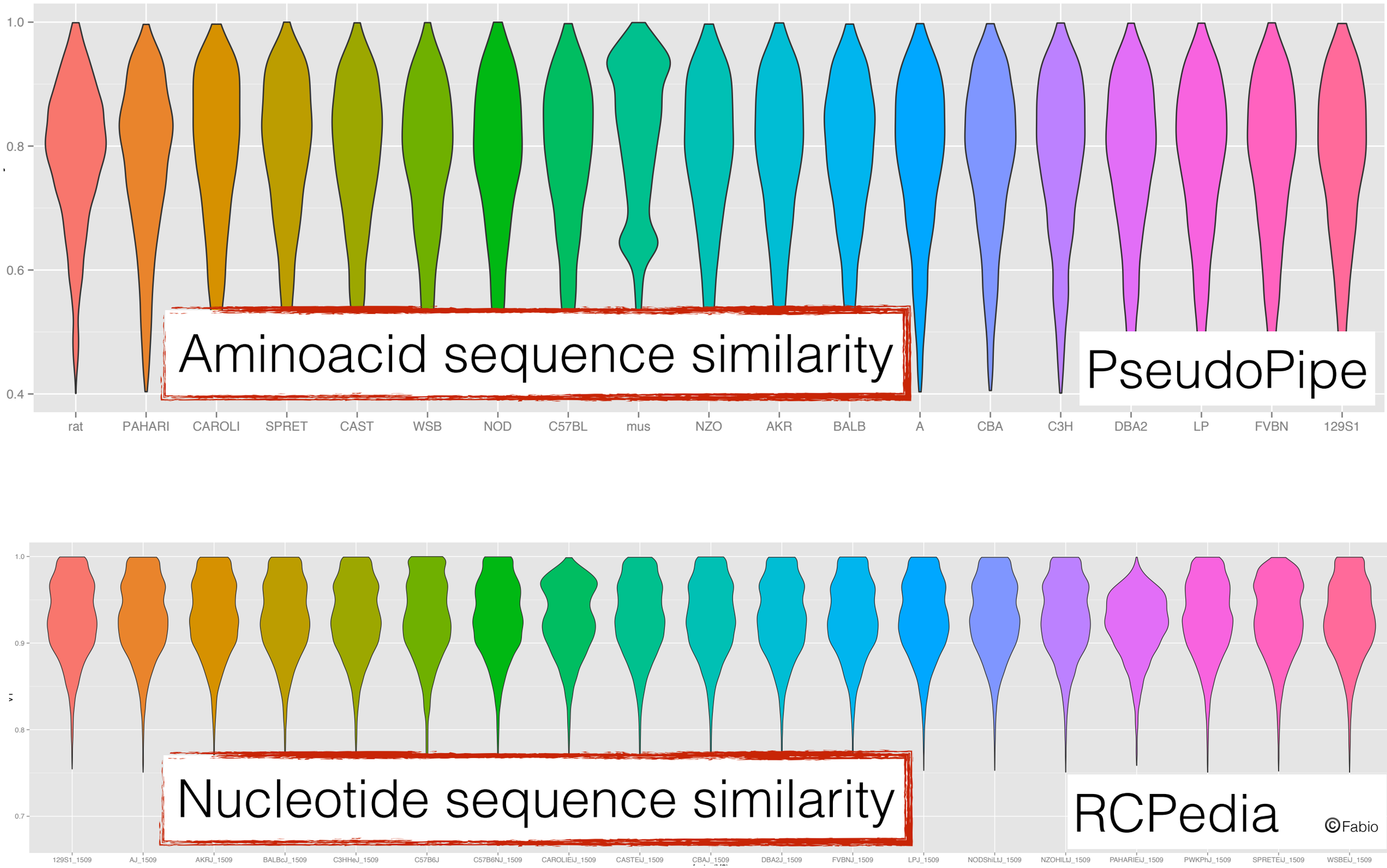
Mouse		PAHARI		CAROLI		SPRET		CAST		WSB	
660	ZnF	437	GPCR	461	GPCR	523	GPCR	463	GPCR	524	GPCR
486	Struct	252	ZnF	241	RRM	337	ZnF	271	ZnF	312	ZnF
480	Trm112p	228	RRM	237	ZnF	290	RRM	270	RRM	259	RRM
467	TF	176	TF	157	Ribo	247	TF	224	TF	222	ANF
434	GpDhC	153	Ribo	152	EGF	222	Kinase	224	ANF	220	GPCR
429	GpDhN	139	Struct	152	Struct	201	ANF	221	GPCR	219	GPCR
425	SRSY	137	EGF	150	TF	198	GPCR	221	GPCR	217	TF
392	TLVcoat	137	Ribo	142	Ribo	198	GPCR	217	Kinase	215	Kinase
333	ANF	135	EGF	139	Kinase	147	Ribo	145	Struct	145	Ribo
329	GPCR	129	Kinase	137	EGF	143	EGF	145	Ribo	140	Struct

NOD		C57BL		NZO		ZKR		BALB		A	
524	GPCR	422	GPCR	467	GPCR	563	GPCR	523	GPCR	504	GPCR
312	ZnF	409	ZnF	380	ZnF	343	ZnF	340	ZnF	355	ZnF
278	RRM	314	TF	278	TF	297	RRM	286	TF	289	RRM
257	ANF	284	RRM	264	RRM	269	TF	280	RRM	272	TF
253	GPCR	218	Kinase	220	Kinase	239	Kinase	224	ANF	234	ANF
253	GPCR	213	ANF	215	ANF	226	ANF	204	Kinase	206	GPCR
233	TF	203	GPCR	199	GPCR	188	GPCR	203	GPCR	206	GPCR
215	Kinase	203	GPCR	199	GPCR	188	GPCR	202	GPCR	202	Kinase
150	Struct	150	EGF	147	ZnF	145	ZnF	150	Ribo	153	EGF
148	EGF	150	Ribo	142	EGF	144	Ribo	143	Struct	147	Struct

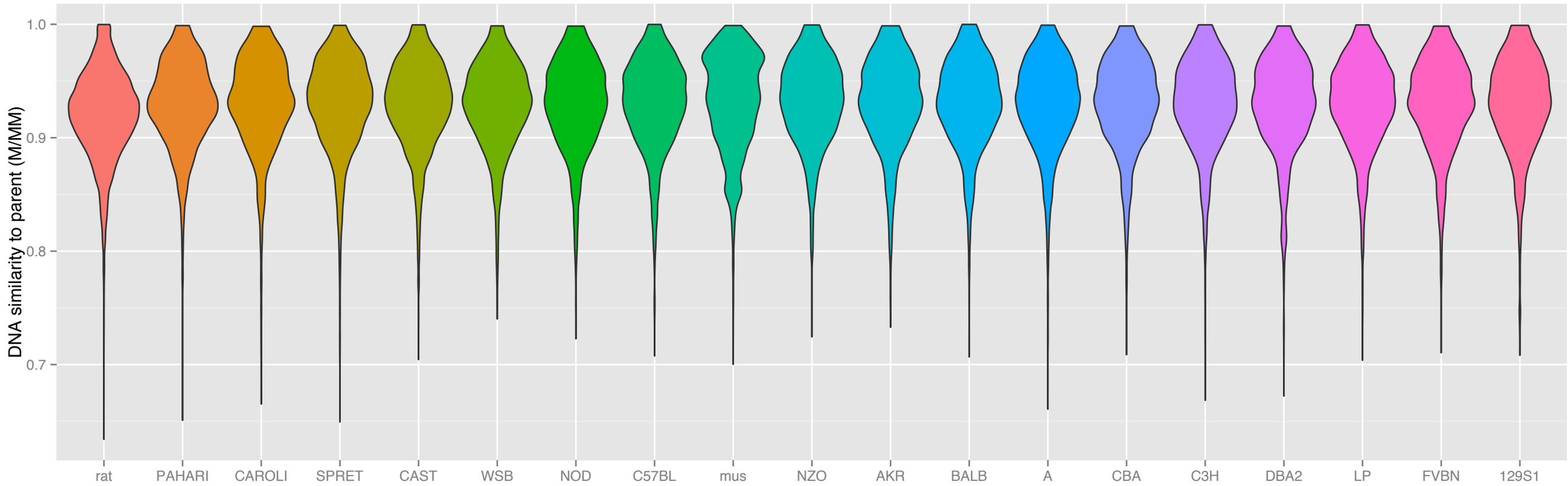
CBA		C3H		DBA		LP		FVB		129S1	
614	GPCR	541	GPCR	535	GPCR	476	GPCR	543	GPCR	499	GPCR
352	ZnF	323	ZnF	371	ZnF	297	ZnF	356	ZnF	318	ZnF
286	RRM	313	RRM	285	TF	263	RRM	285	RRM	264	RRM
266	TF	259	TF	273	RRM	239	TF	279	TF	232	TF
238	ANF	214	Kinase	214	ANF	212	ANF	256	ANF	194	Kinase
227	Kinase	203	ANF	212	GPCR	209	GPCR	253	GPCR	179	ANF
215	GPCR	195	GPCR	212	GPCR	209	GPCR	251	GPCR	176	GPCR
215	GPCR	194	GPCR	205	Kinase	204	Kinase	235	Kinase	176	GPCR
151	GPCR	155	Ribo	145	Struct	150	Ribo	166	EGF	142	Struct
151	Ribo	151	EGF	141	GPCR	143	Ribo	164	Struct	141	Ribo

PSEUDOGENES IN MOUSE STRAINS

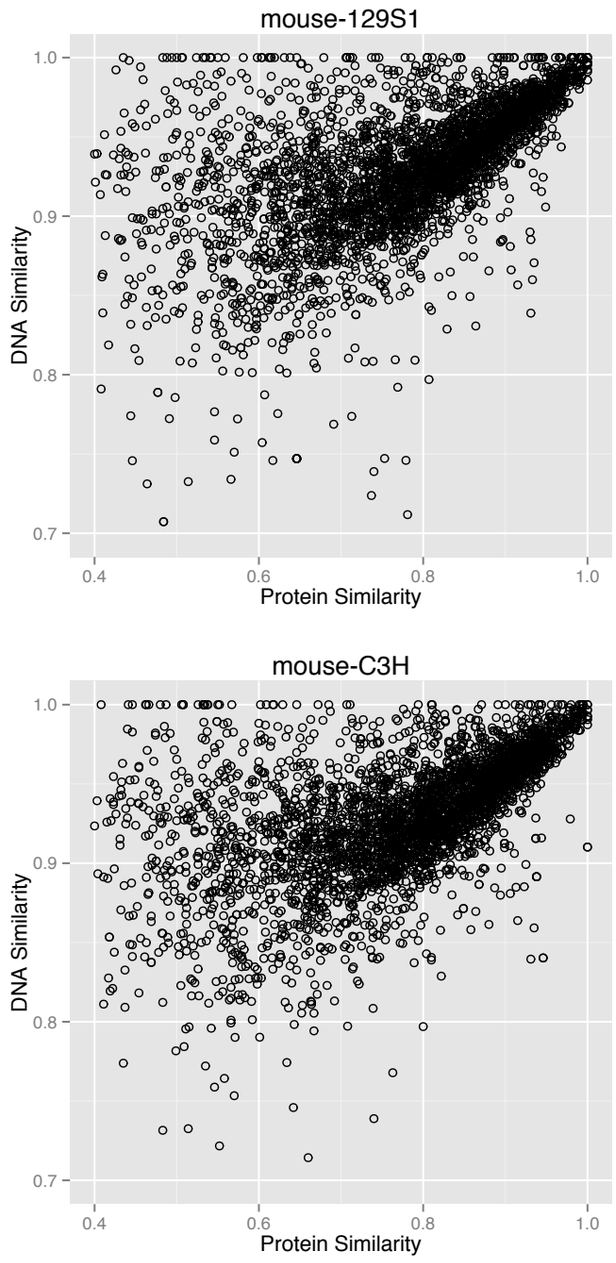
Pseudogene similarity to parents

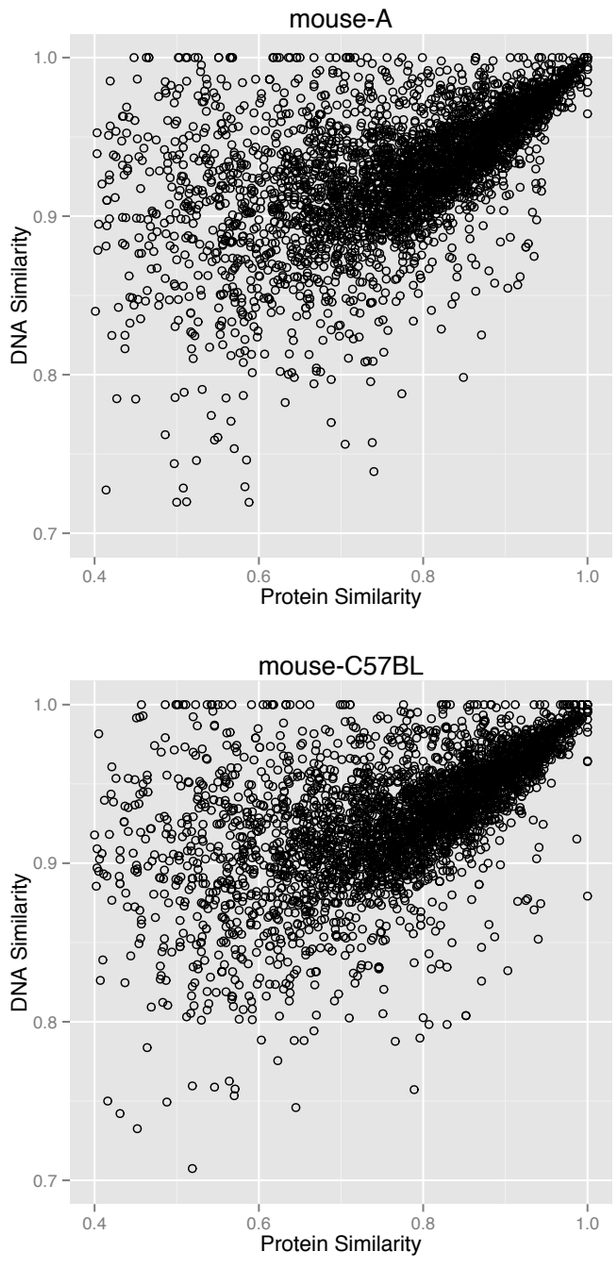


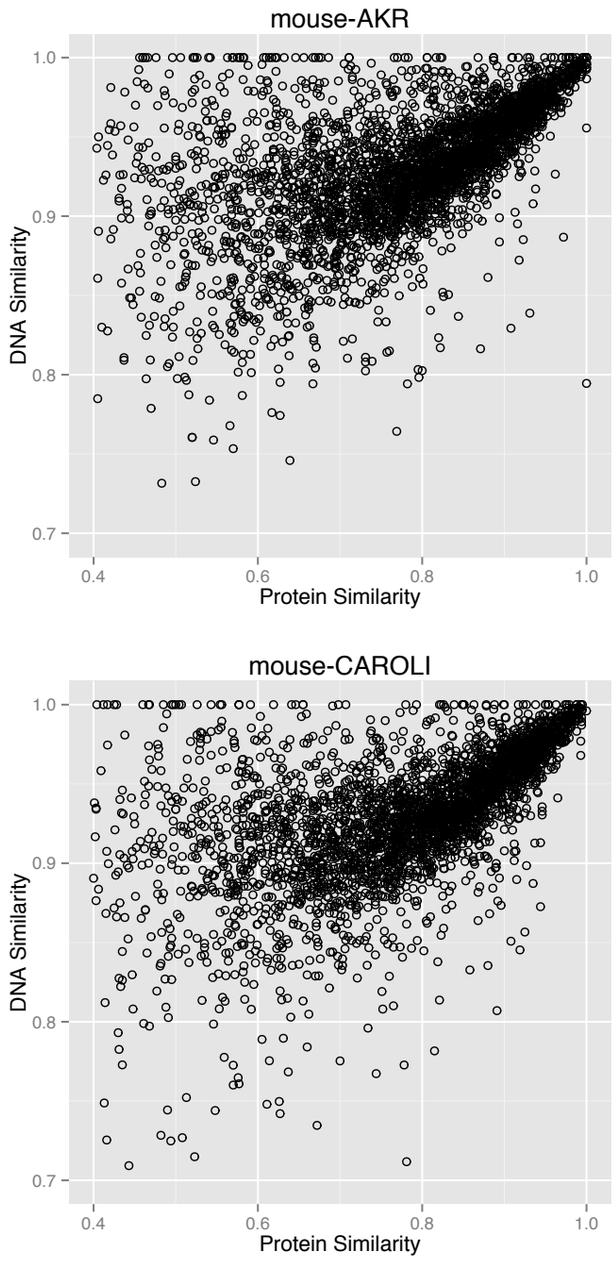
DNA sequence similarity to parent genes

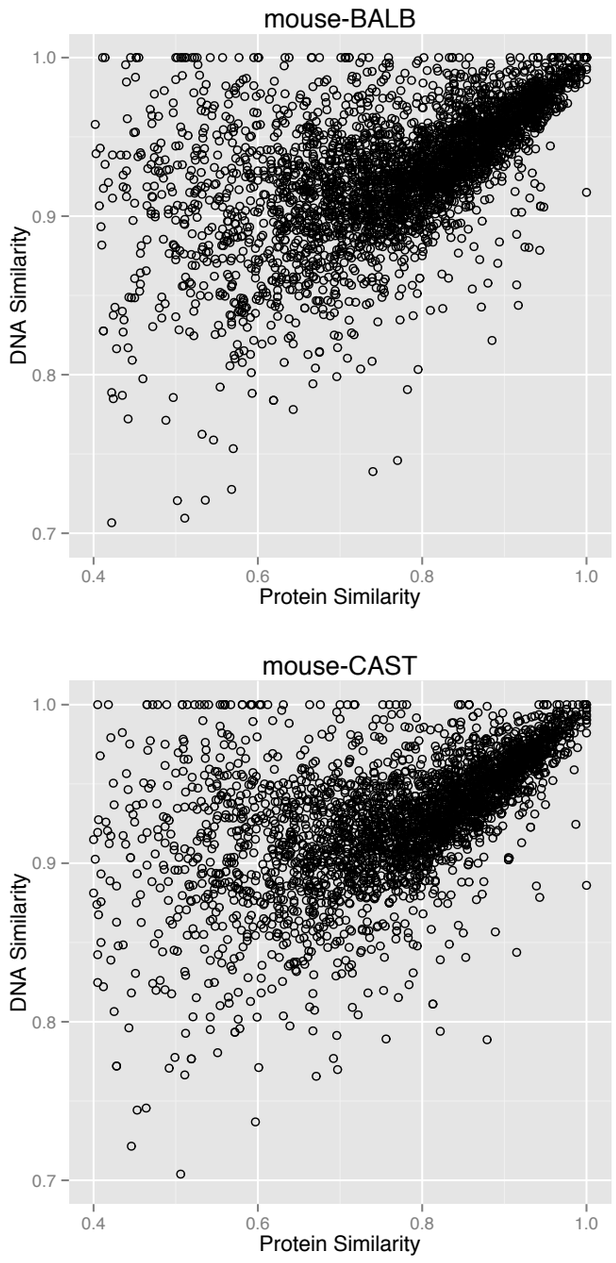


Comparing the protein sequence similarity vs DNA sequence similarity

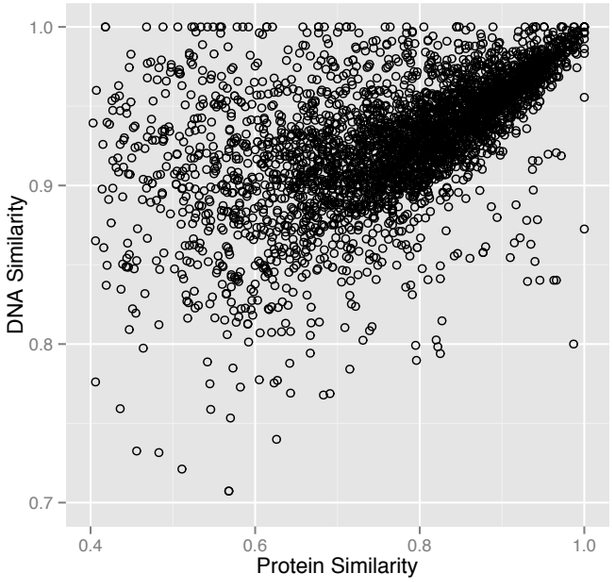




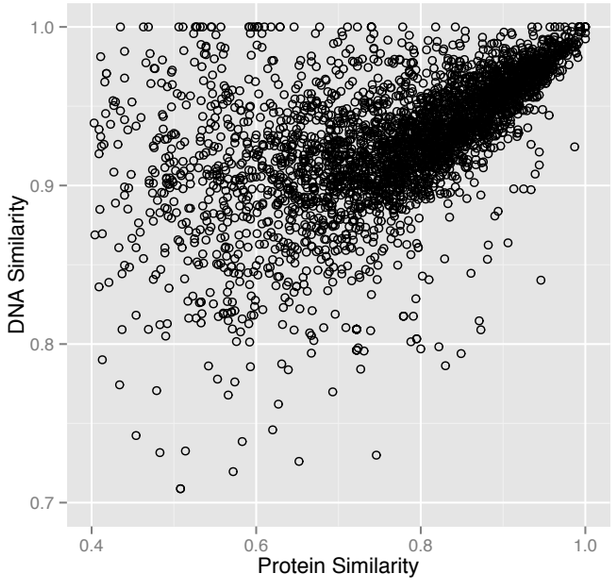




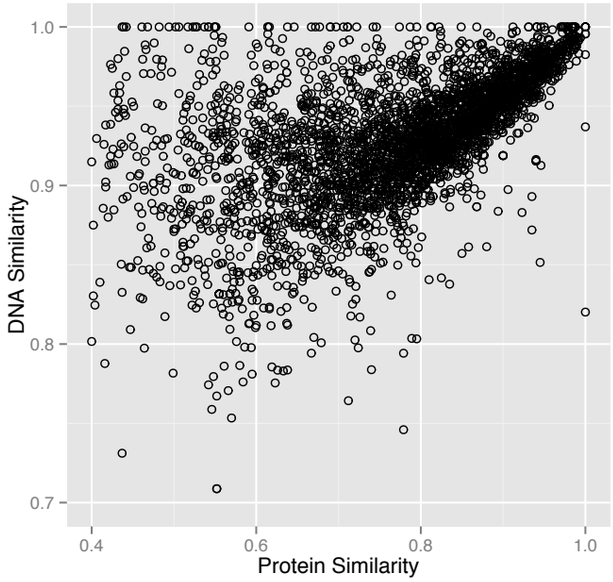
mouse-CBA



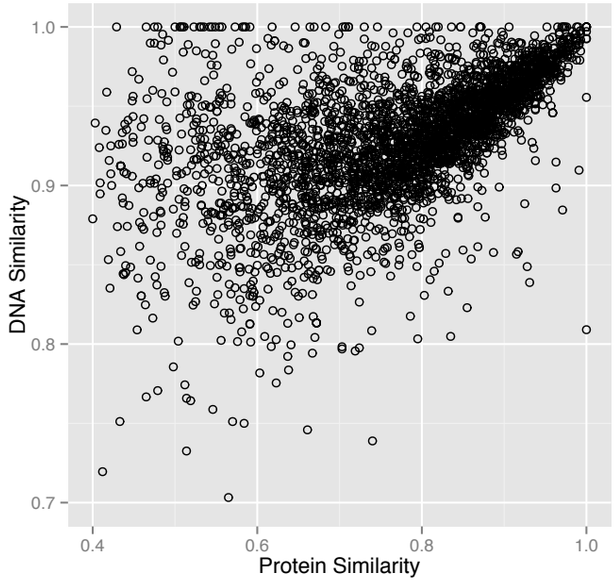
mouse-DBA2



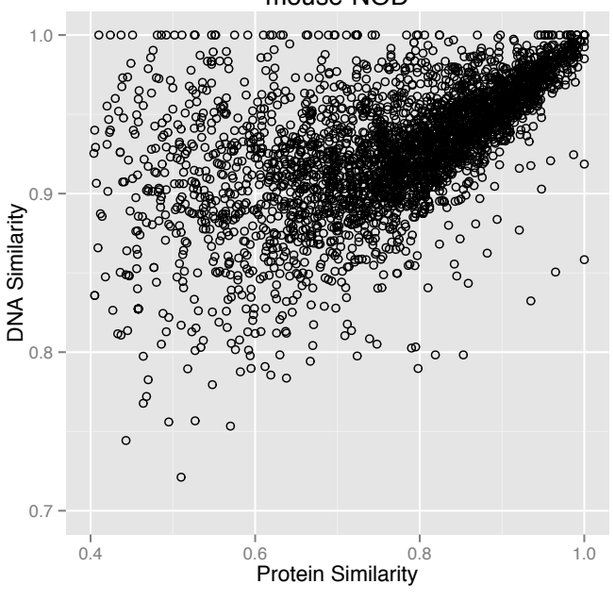
mouse-FVBN



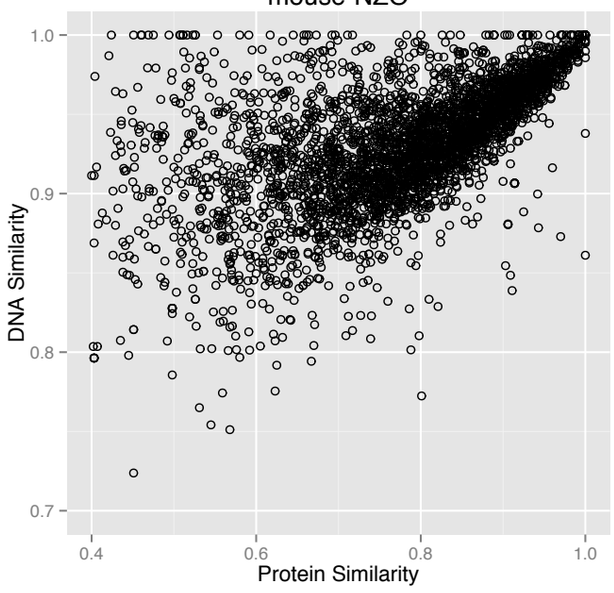
mouse-LP



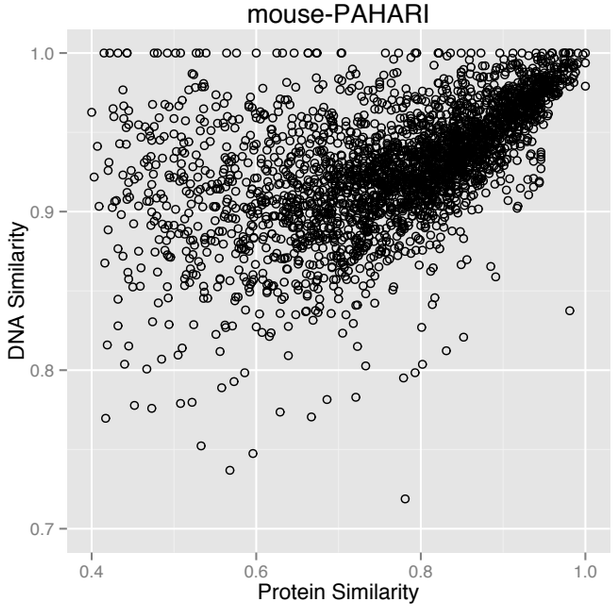
mouse-NOD



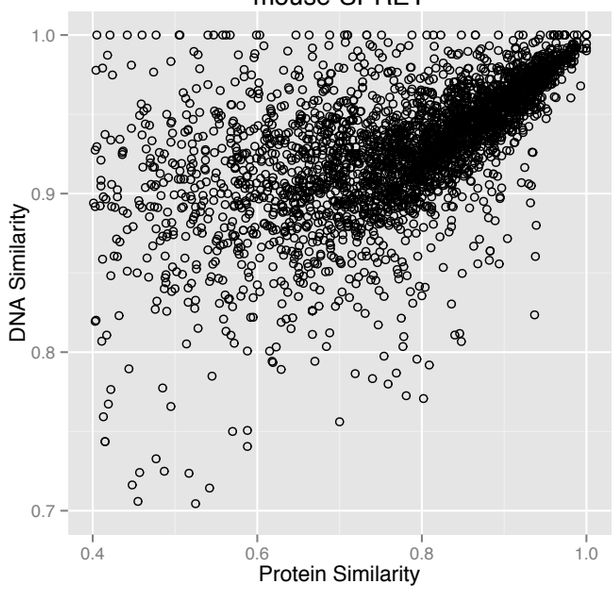
mouse-NZO



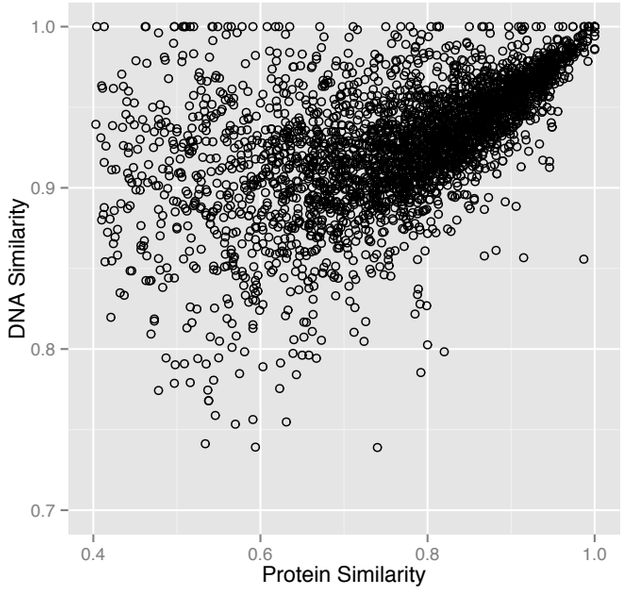
mouse-PAHARI



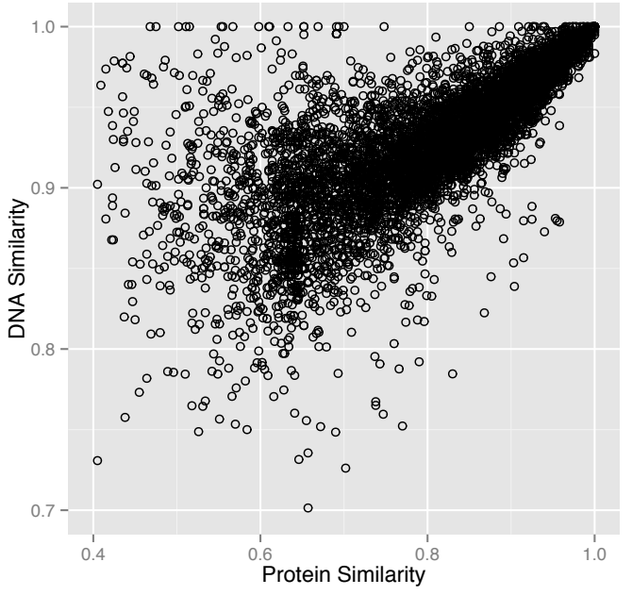
mouse-SPRET



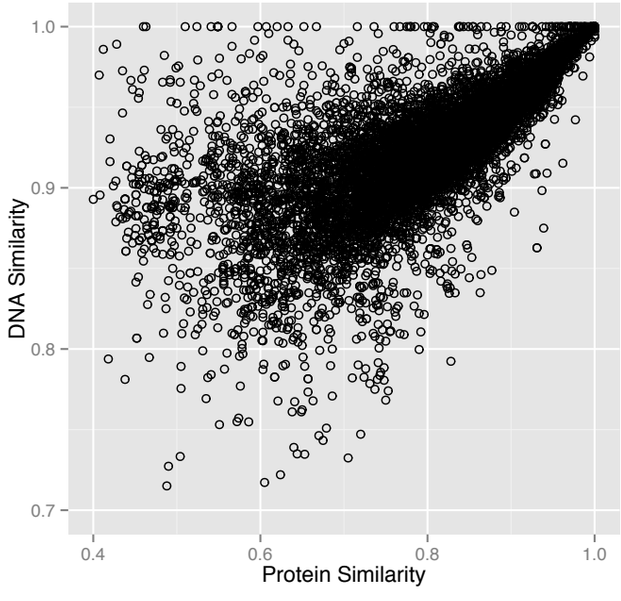
mouse-WSB

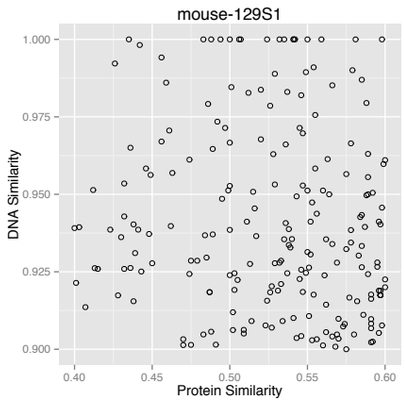


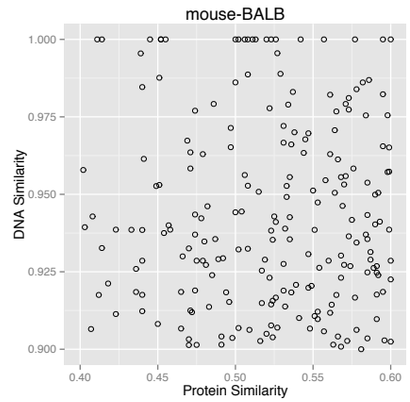
mus-ENS83

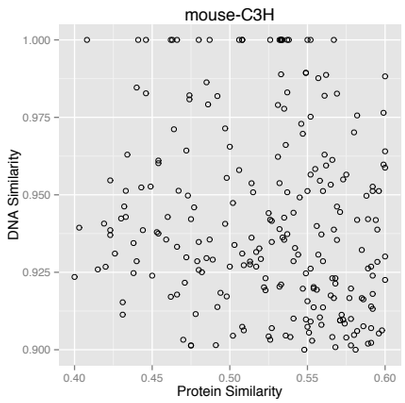


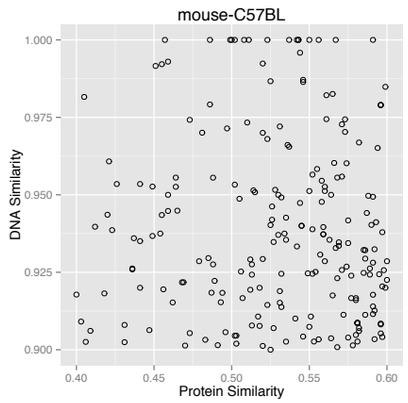
rat-ENS83

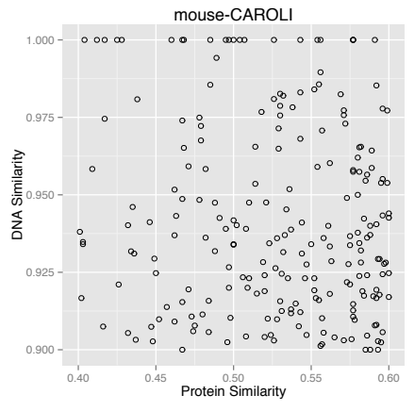


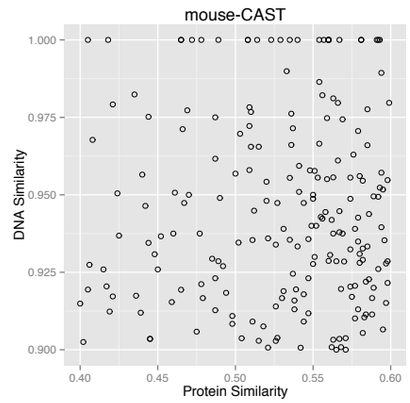


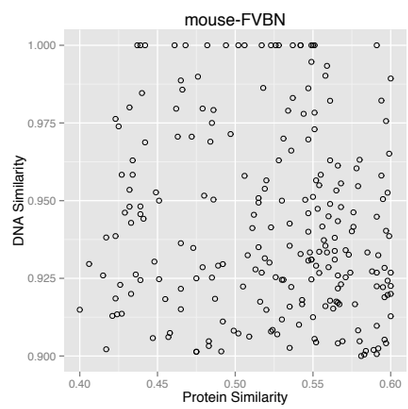


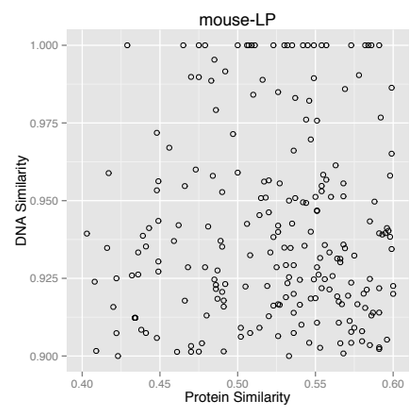


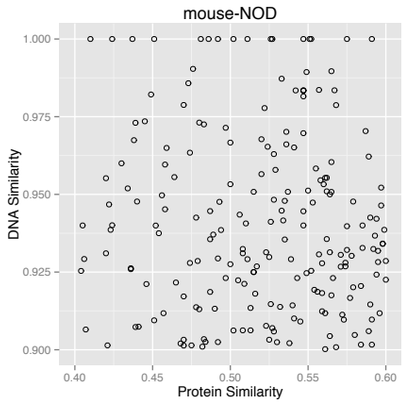


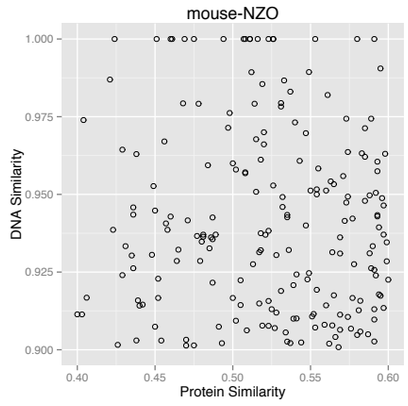


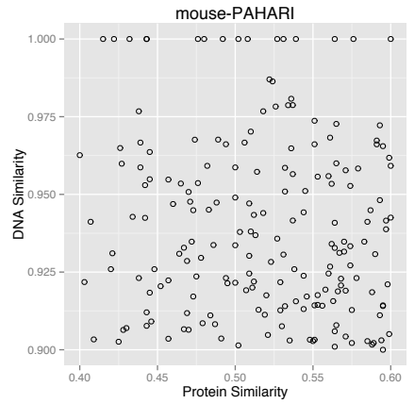




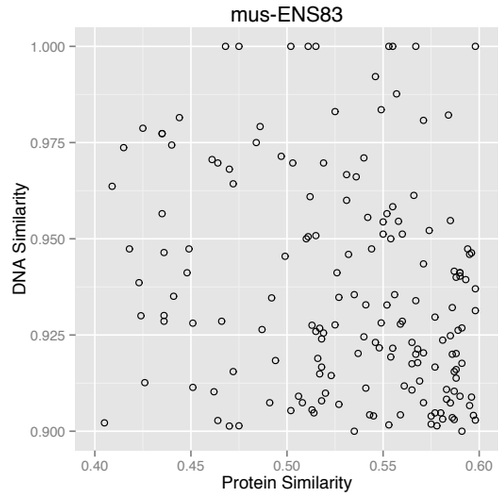


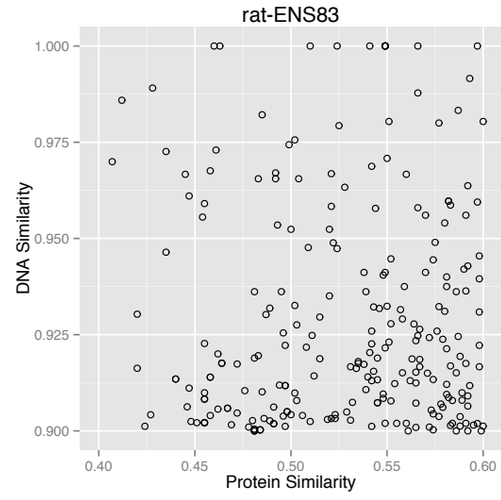






Zoom in at DNA sim >0.9 and protein sim <0.6





Discrepancies in sequence similarity

- On average 2% of the annotated pseudogenes have low amino acid sequence similarity to the parent while having a highly conserved DNA sequence

- However there are a number of pseudogenes that have >0.6 amino acid sequence sequence with the parent gene but *no* DNA sequence similarity

129S1

chr2:113744999-113745796
ENSMUSP00000097213.1

0 0 0 0 0.541

Olfactory pseudogene - 100% DNA sequence similarity

>parent
EDSNRTVSSEFIFQGLCSSRQLEIFLLLPFSILYLMAVVGNLFVVILIIIDHHLHSPMYFLLANLSFIDFCLSSVTTPKLTIDL
LKENKTISFVGYMRQIVCVHFFAGGEMVLLVTMAYDRYVAICRPLHYSSIMDRQKCIWLVVISWIVGFVHAISQMLLILDLPFC
GPRVIDSFFCDIPLVMKLACMNTDTLEILINADSGILATTCFILLLISYTYILLTVQHRSKDGSSKALSTCTSHIIVVLLFFGP
IIFIYLWPVSITWV

>pseudogene
EDSNQTVVSAFIFQGLCTSRQLEIFLLLPFSVLYLMTLVDNLFVVILIITDHHFNSPMYFLSANLSFVXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
GPRVIDSFFCNIPLVMKLACMNTDTLGILINADSGILATYSFTLLLISYTYILLTVQHHSKDGSLKALSTCTSHIIVVLLFFGP
IIFIYLWPINITWV

EDSNQTVVSAFIFQGLCTSRQLEIFLLLPFSVLYLMTLVDNLFVVILIITDHHFNSPMYFLSANLSFVDFYLSSVNTPKLTIDL
LKENKTISFGGCMSQILCVHFFGGSEMVLLVTMAYDQYVAICRPLHCSGSMDRQKCIWLVVISWIVGFVHAISQLLLILDLPFC
GPRVIDSFFCNIPLVMKLACMNTDTLGILINADSGILATYSFTLLLISYTYILLTVQHHSKDGSLKALSTCTSHIIVVLLFFGP
IIFIYLWPINITWV

C57BL

chr10:117497656-117498497: ENSMUSP00000040488.9

22 **16** **12** **4** 0.537

ENSMUSP00000040488.9 - Cyclin D3 — 100 % sequence to parent

>Parent
MELLCCEGTRHAPRAGPDP------RLLGDQRVLQ-SLLRLEER-YVPRASYFQ-CVQKEIKPHM---RKMLAY
-WMLEVCEEQRCEEDVFP-LAMNYLDRYLSCVPT-RKAQLQLLGTVCLLLASKLRETT-PLTIEKLCIYTDQAV
APWQLREWEVLVLGK--LKWDLAAVIAHDFLALIL-HRLS-LPSDRQALVKKHAQTF-LALCATDYTFAMYPPS
MIATGSIGA-AVLGLGACSMSADELTELLAGITGTEVDCLRACQEQIEAALRESLREAAQTAPSPVPKAPRGSS
SQGPSQTSTPTD

>Pseudogene
MELLCC*GTLHGPPGWLDSWLFTGEHLLRDQYIL-\SLLCLEEH\FMPCTSYSG/CEHPETKMHAWC/RKVLAL
/WILEEYEEQCCKEEAFPCL*TIWIALRLPCIPT\KKGQVQLIGRIRWLVSSKPHKTT\P----------HQAE
SPCQLWKWELRGLGKA\LK*ALAAAVASNFLDLSL/HRL-/LPSDQQTMVRKHAQTF\LALCATNYTFAMYWPS
MIVVG----\AVQGLDACCTSGNKFIELSAGIRDSEVDCLWTCQEQIEAAIRESLRNAAQIIPSPVLKAPHGSR
S*GFSLSSIPTH

chr1: 7847498-7848944

ENSMUSP00000079306.5

28 9 3 18 0.690

ENSMUSP00000079306.5 - Heat shock protein 2

>Parent

LNVLRIINEPTAAAIAYGLDKKGCAGGEKNVLIFDLGGGTFDVSILTIEDGIFEV-
KSTAGDTHLGGEDFDNRMVSHLAEEFKRKHKKDIGPNKRAVRRLRTACERAKRTLS
SSTQASIEIDSLY----------EGVDFYTSITRARFEELNADLFRGTLEPVEKAL
RDAKLDKGQIQEIVLVGGSTRIPKIQKLLQDFFNGKELNKSINPDEAVAYGAAVQA
AILIGDKSENVQDLLLLDVTPLSLGIETAGGVMTPLIKRNTTIPTKQTQTFTTYS-
DNQSSVLVQVYEGERAMTKDNNLLGKFDLTGIPPAPRGVPQIEVTFDIDANGILNV
TAADKSTGKENKITITNDKGRLSKDDIDRMVQEAERYKSEDEANRDRVAAKNAVES
YTYNIKQTVEDEKLRGKISEQDKNKILDKCQEVINWLDRNQMAEKDEYEHKQKELE
RVCNPIISKLYQ-------GGPGG---GG---SSG---GPTIEEVD

>Pseudogene

LNVFGIINEPTAVAIVYGLDKK—VGAERNVLIFDLGGGTFEVLILTIEDQIF—\KS
TAGDTHLCREDFDNQMVNHFIAEFK*KHSKDISEN*RAVWNLHTACEWAKYTLSFS
TQASIEXXXXXXXXXXXXX\EGIVFYNSITQA*FKELNVDLFHGALDPVEKVL*D
AKLDKSQIHDTVLVGGSTRIPKIQKLLQYFFHGKELNKSINPDEAVAYNAAVQAAV
LSGDKSTNIQDLLLLDVSPLSLGIETASGVMTVLIKSNTTIPTKQTQTF----/DH
QPCVLIQVYEGERAMTKDYNLLGKFELTGIPPAPRGVTQIEVTFDIHTNGILNVSA
VDKSTGKKNKITITNDKGHLSKEYIEGIVQEA-KYKVENEKQRDKFSSKISLESCA
FNMKATVEDEKLQGKINDEVKQKILDKCNEIISWLDKNQTADKGEFEHQQKEMDKV
YNPIFTKLYQSSSGMLGGMPGGFPVGGAHPSSGAASGHTIEEVD

Lacks any DNA sequence similarity to the *parent* gene.

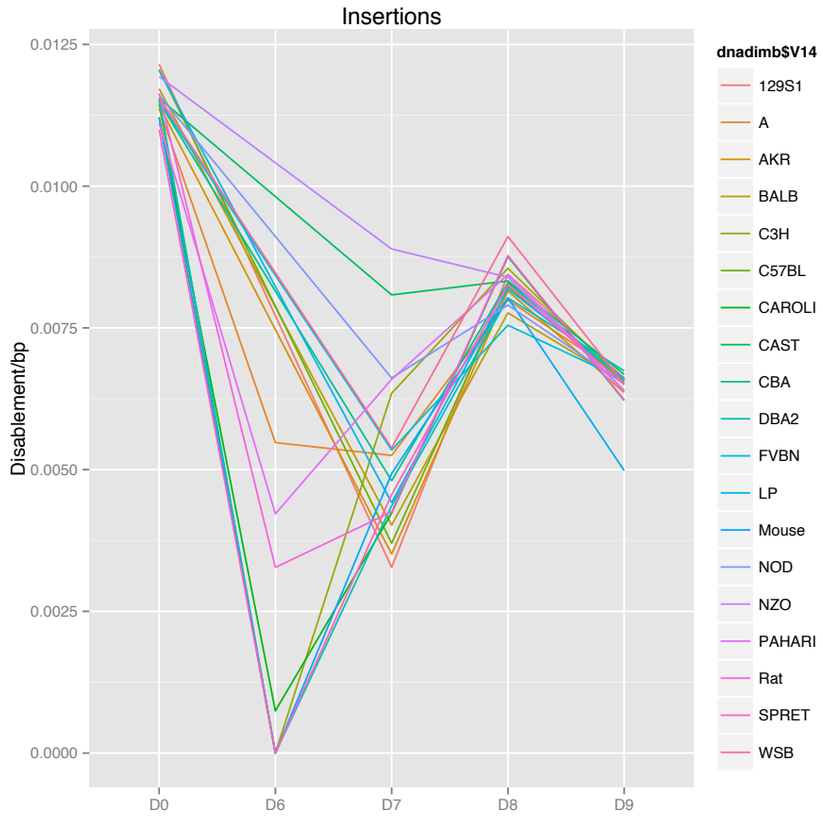
Options:

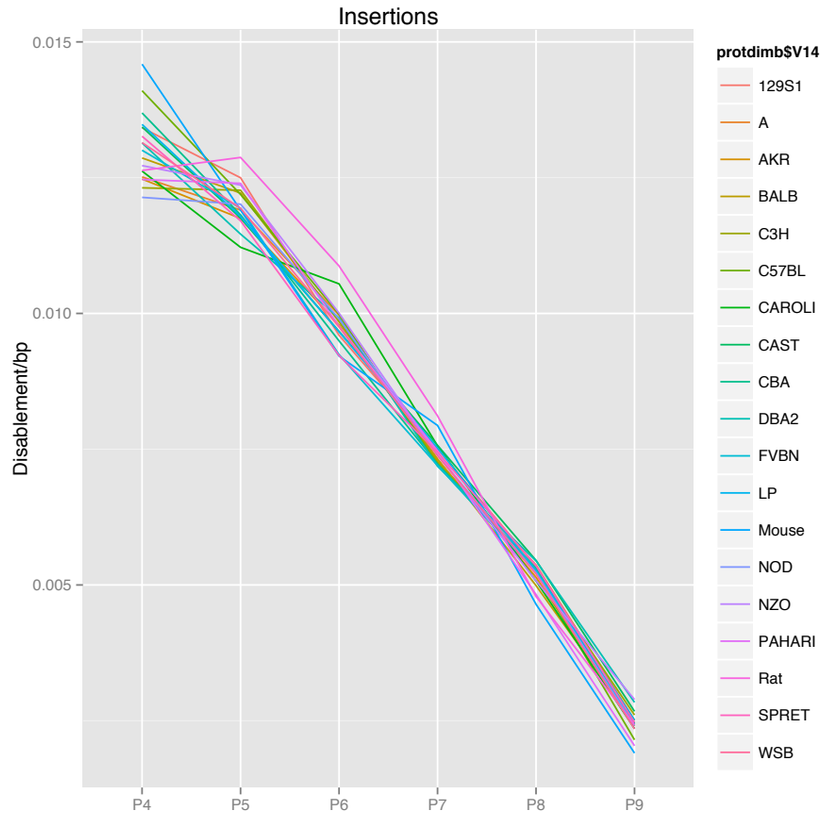
1 - artefact

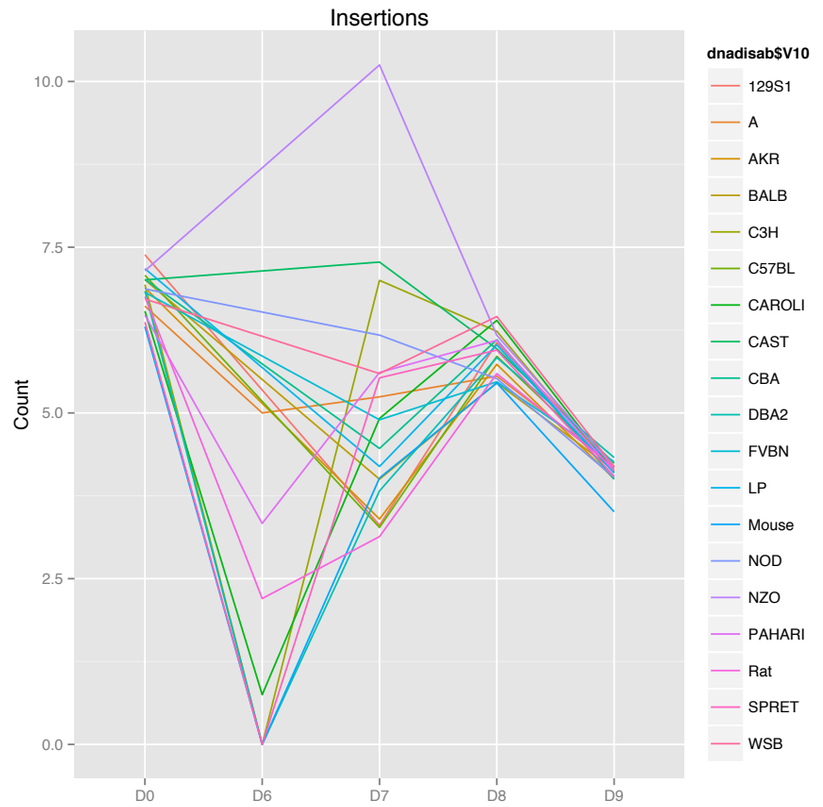
2 - the parent gene is another protein from the same family — > incorrect parent assignment

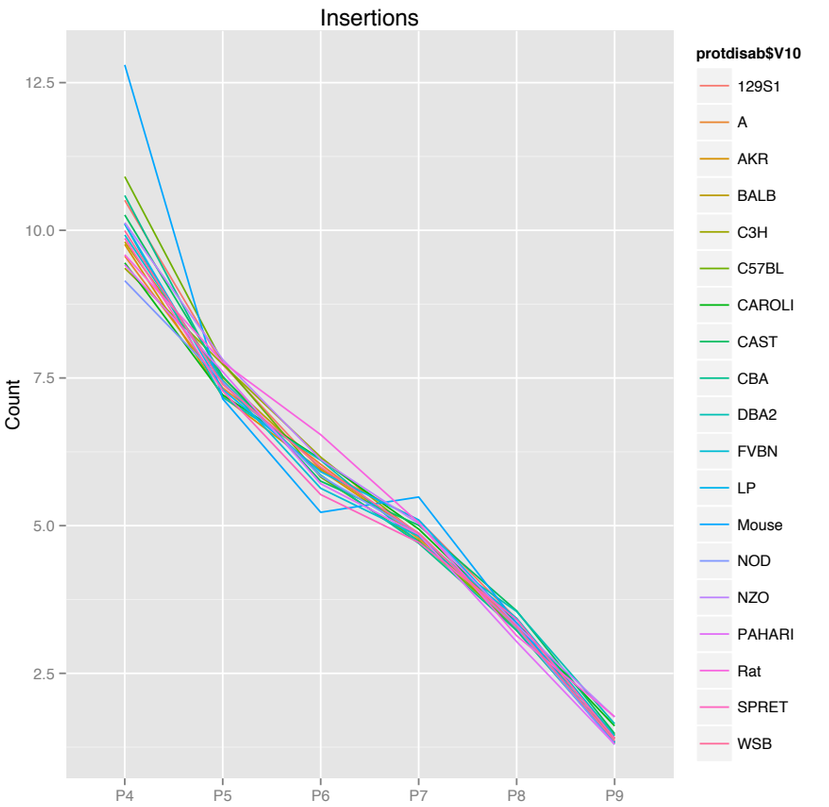
3 - it’s a unitary pseudogene / LOF event

Average disablements distribution as function of sequence similarity to parent genes



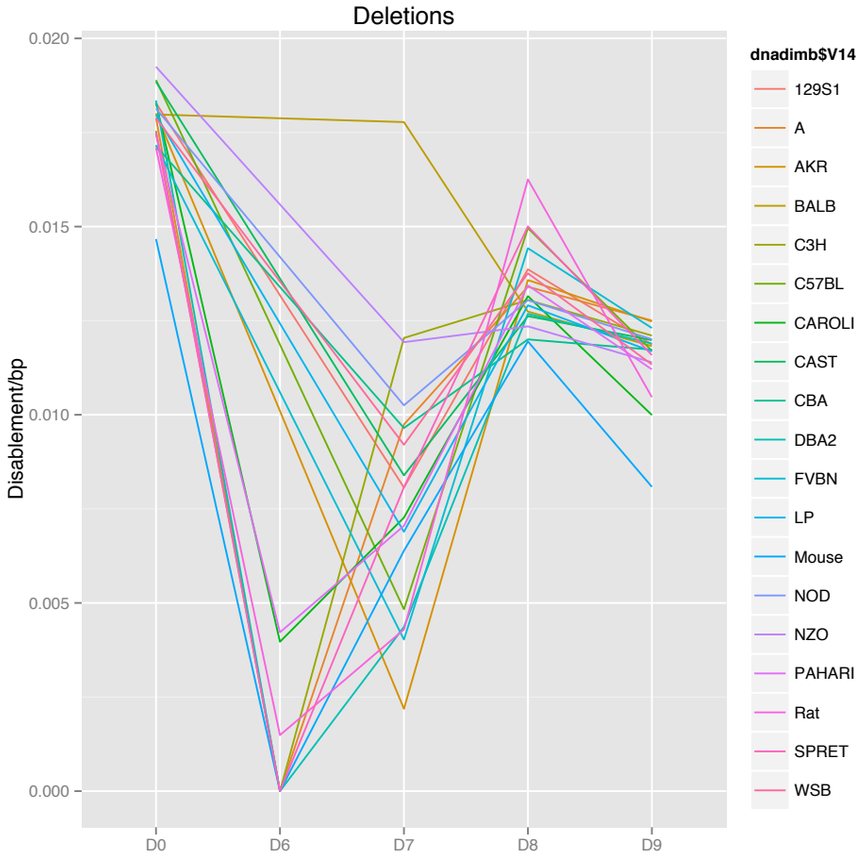


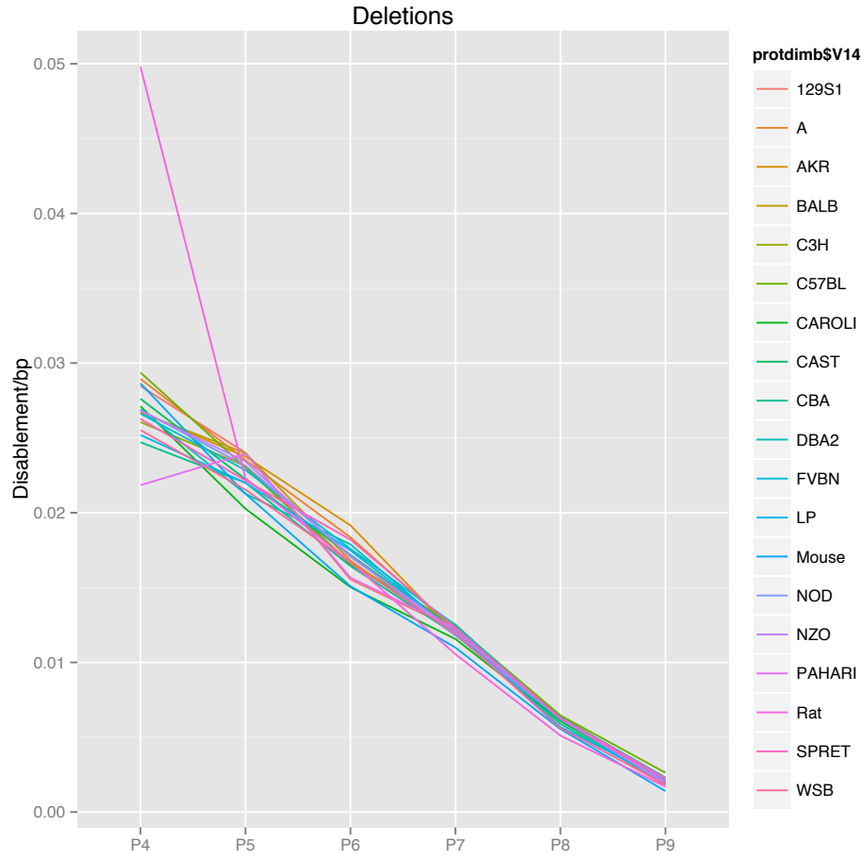


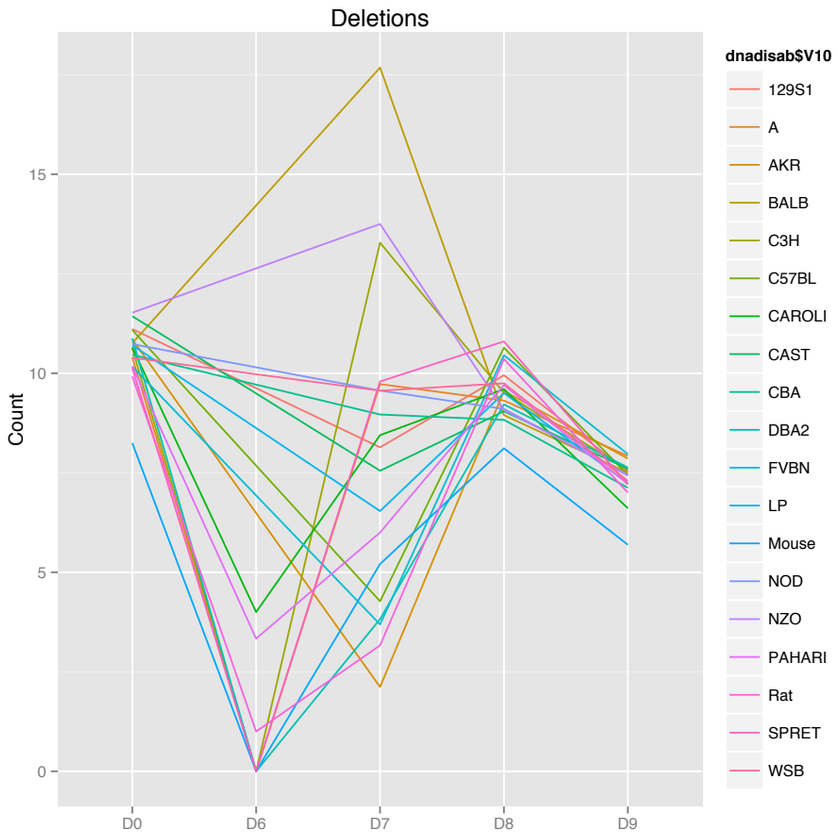


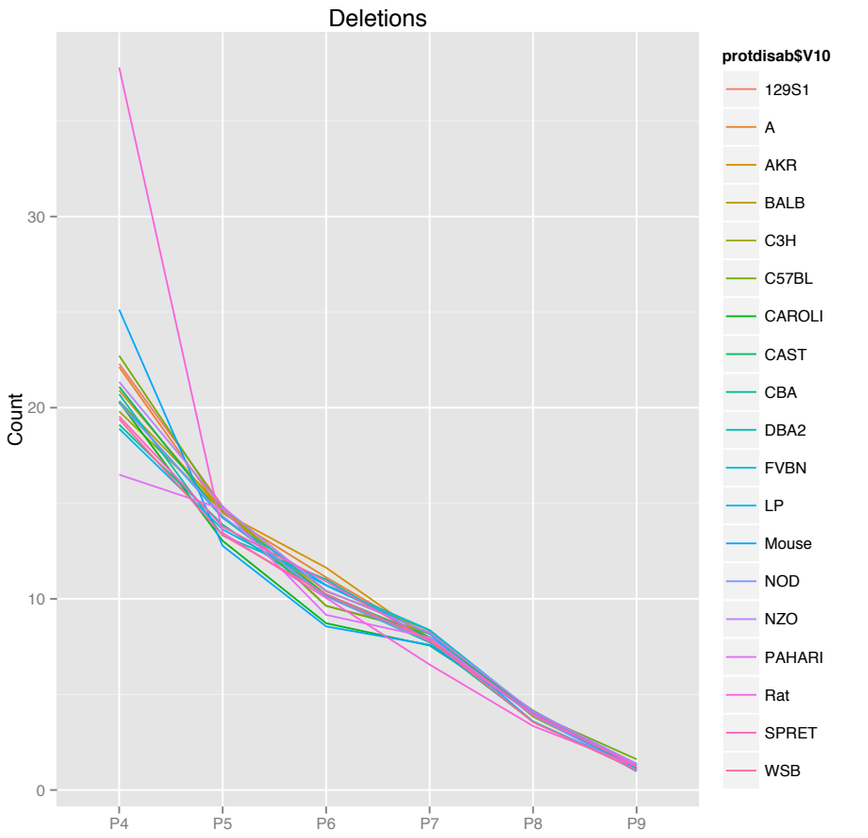
DNA similarity

AA similarity



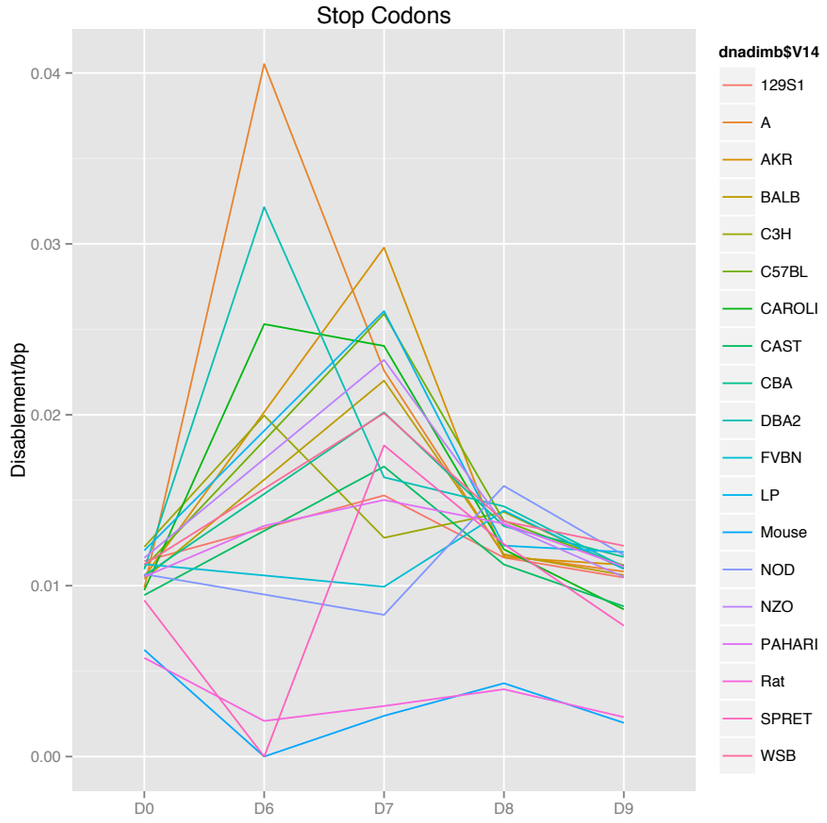


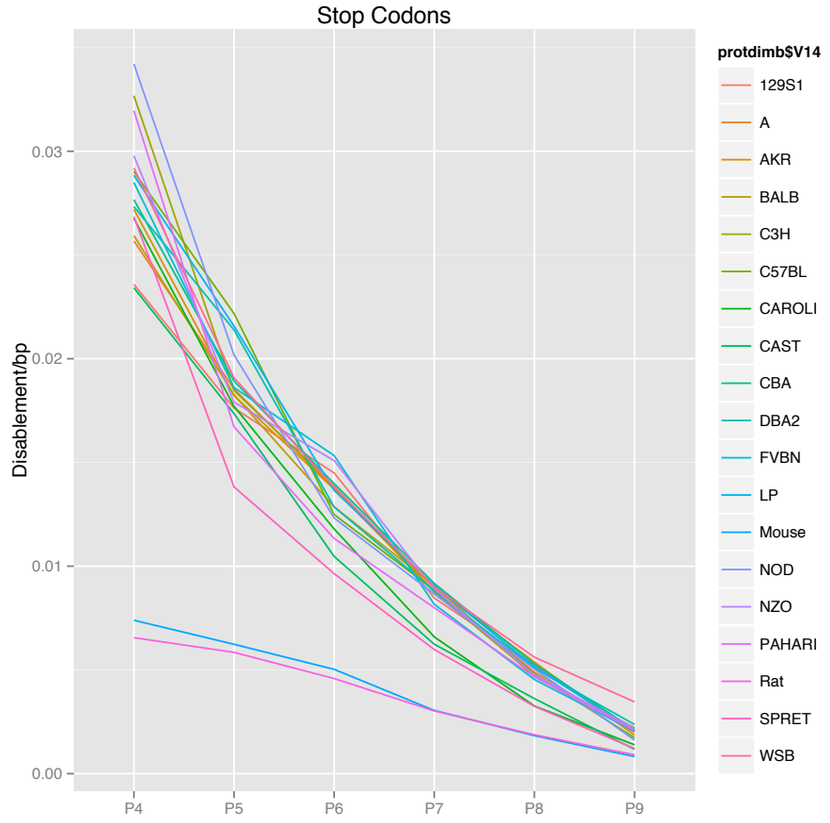


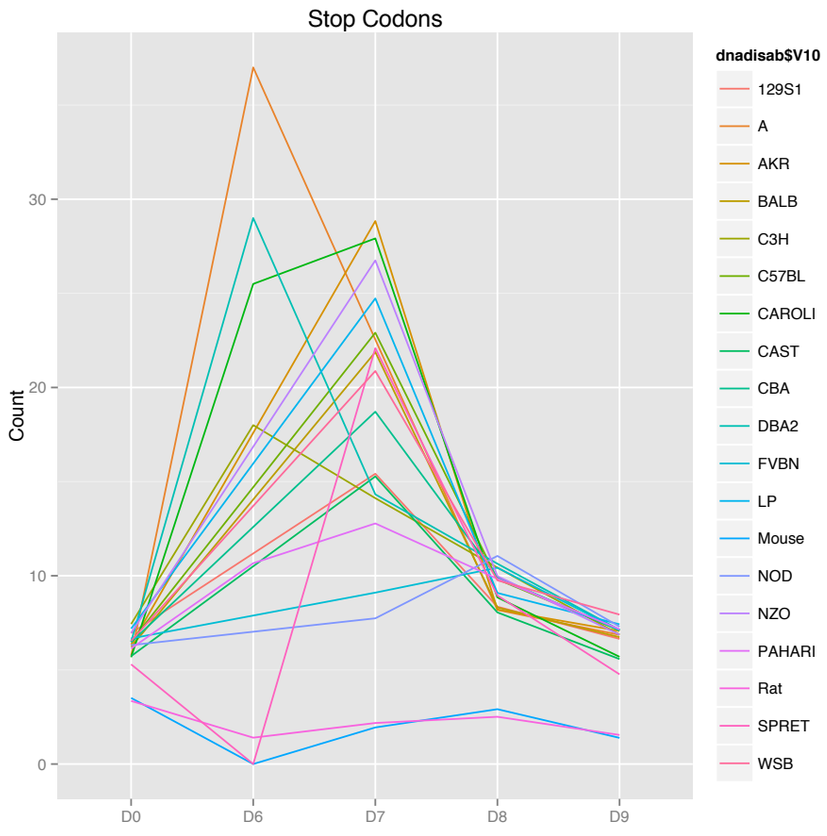


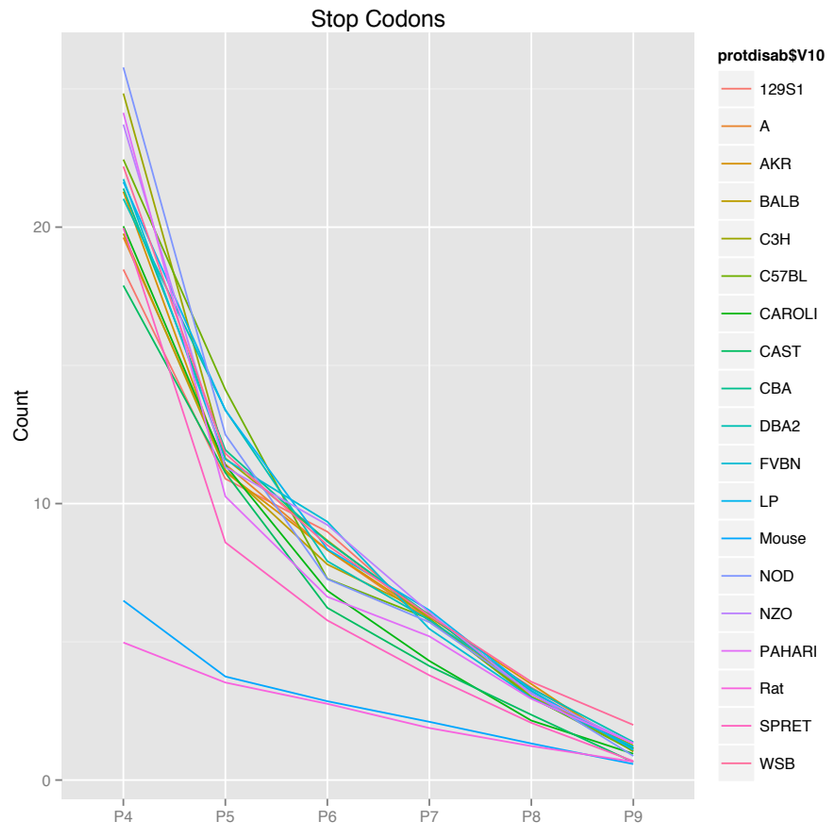
DNA similarity

AA similarity



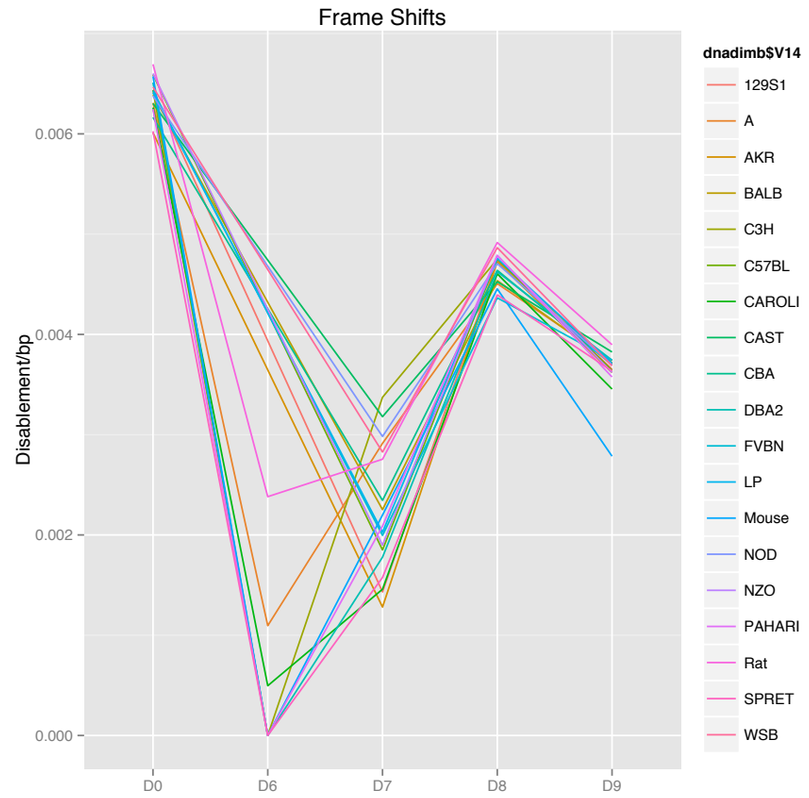


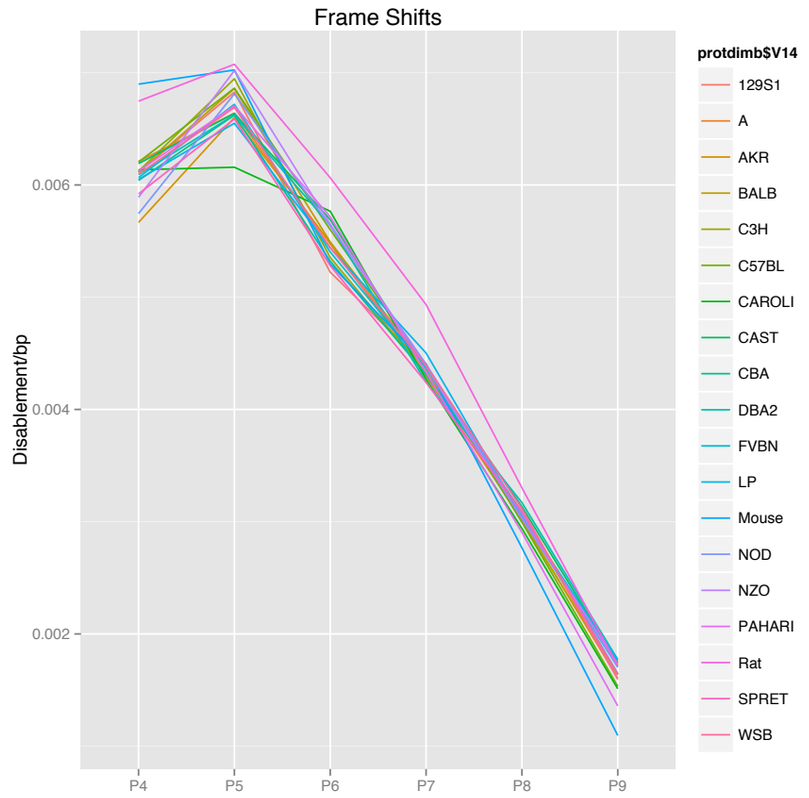


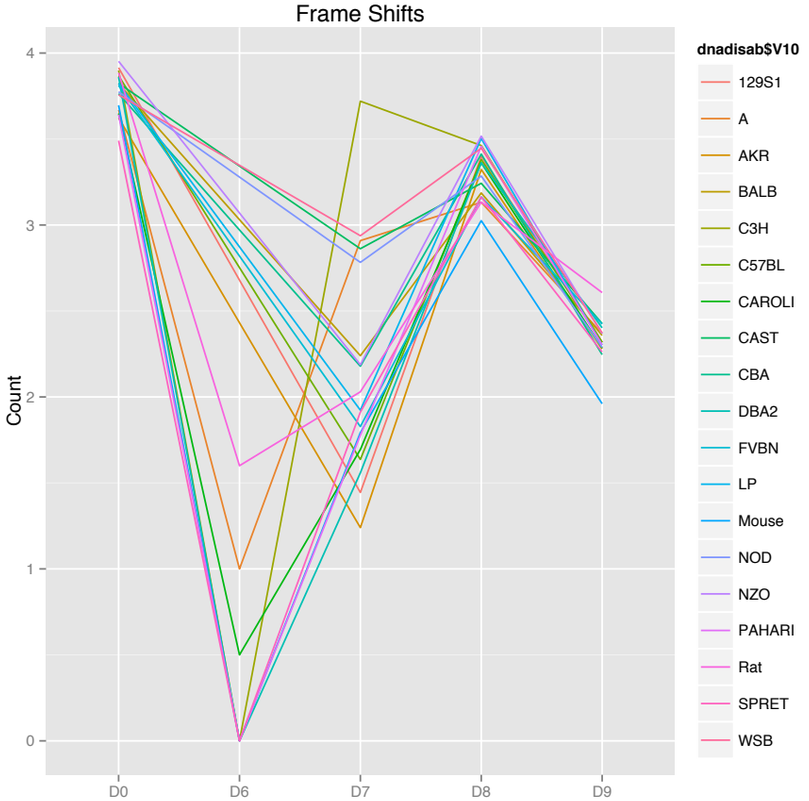


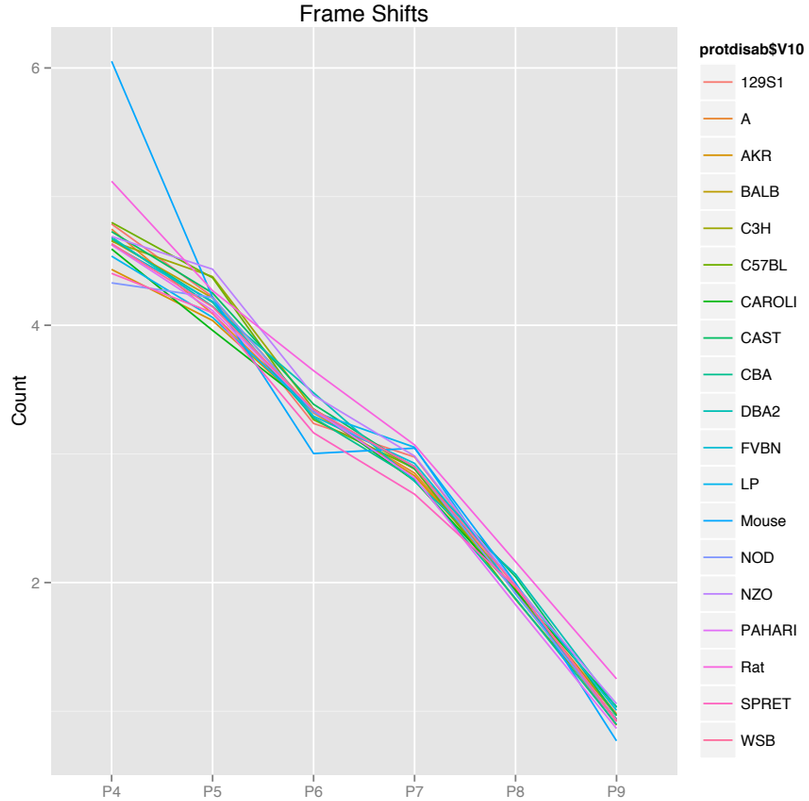
DNA similarity

AA similarity









DNA similarity

AA similarity