Mouse Pseudogenes
~ Update ~

Cristina Sisu

Gerstein Lab
Yale

March 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

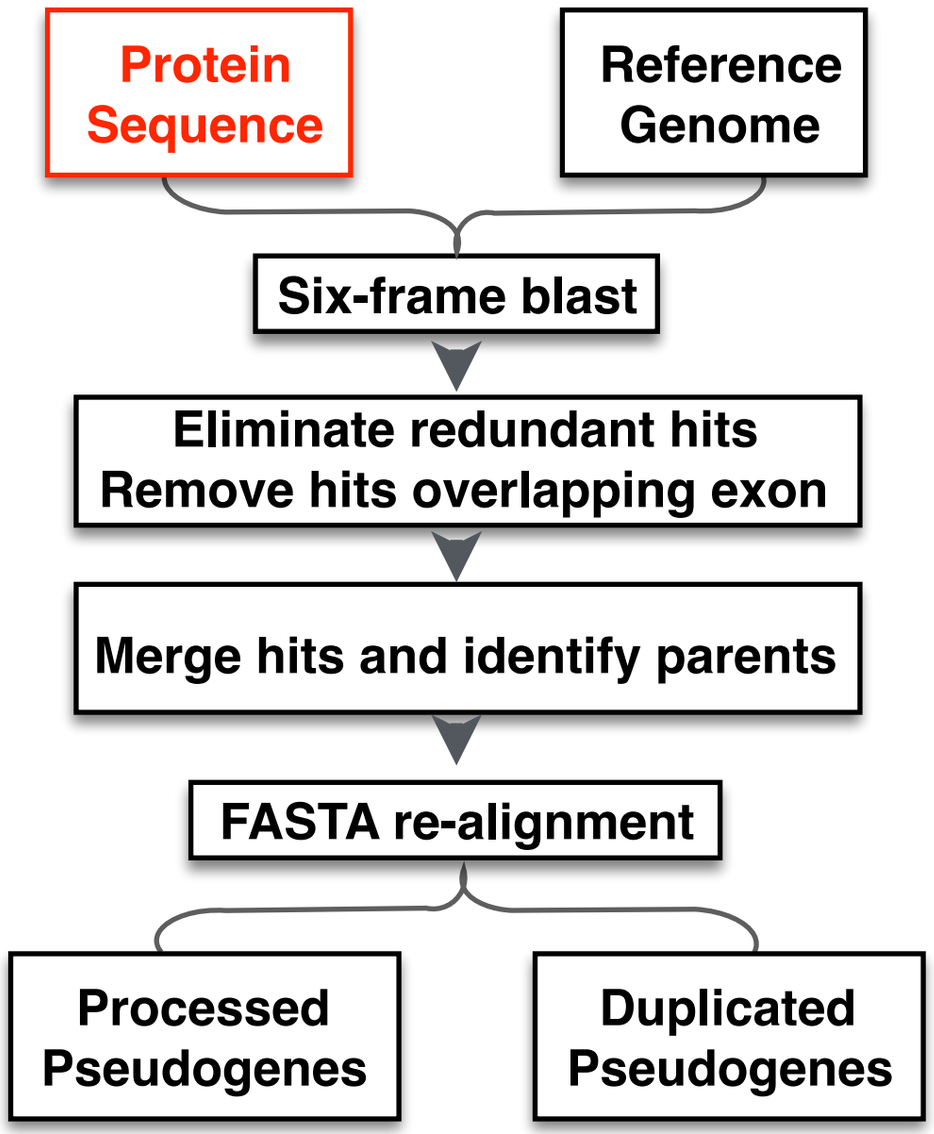
Adapted from Ian Fiddes

Pseudogene annotation of mouse strains

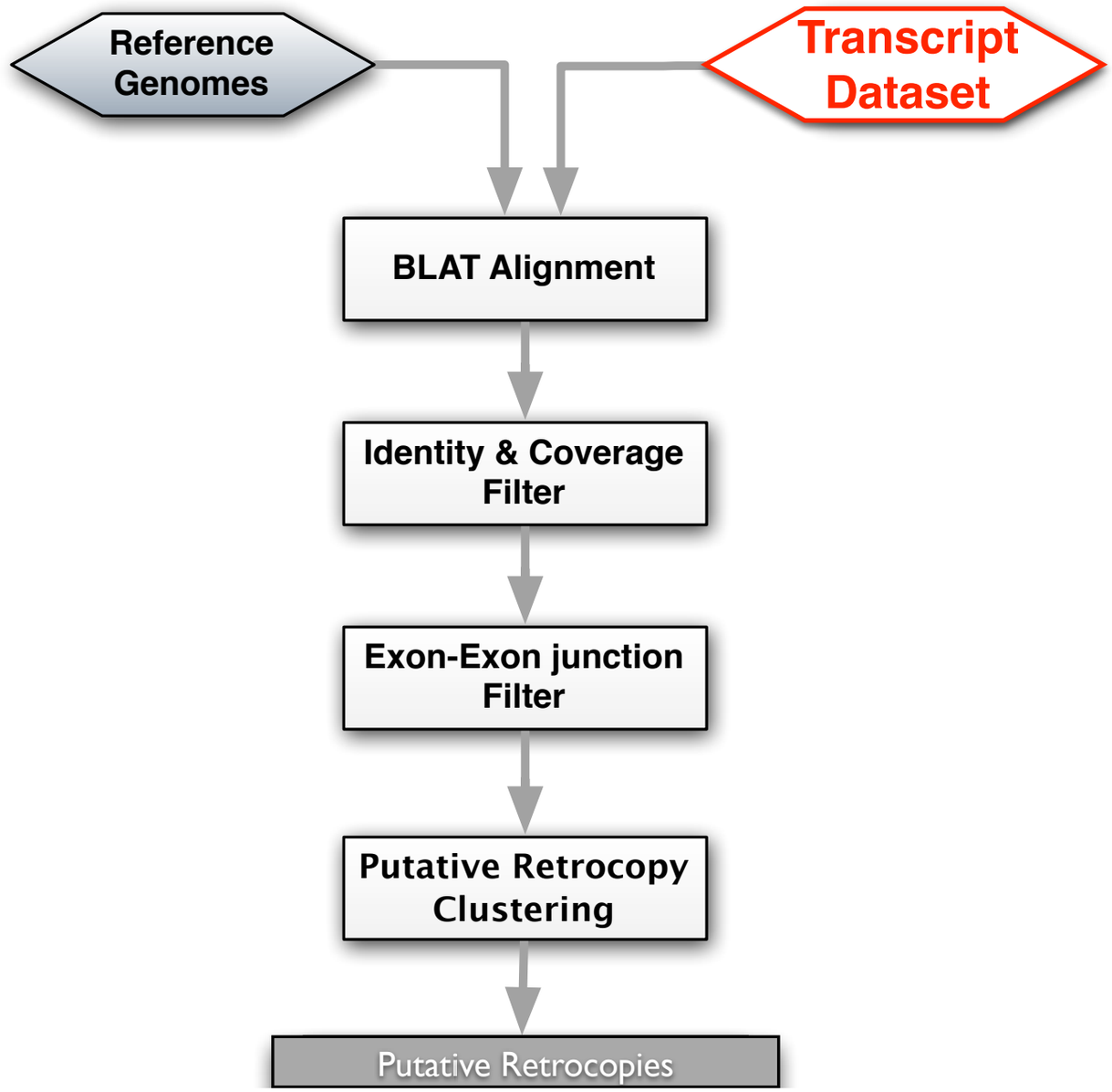
- **REL 1509 assembly** - fasta sequences & repeat masked DNA sequences
- **Gene annotation:**
    - cgp_consensus_gene_set: Contains both genePred and GTF files for the consensus between the comparative gene predictions produced by Stefanie and the TMR consensus set
    - augustus_gene_set: Contains just the Augustus predictions from UCSC
- **Pseudogenes liftoever:**
    - transmap_pseudogenes: contains the genePred and liftover pseudogenes from the GENCODE M8
- **Protein coding sequence:**
    - GENCODE M8 - Ensembl 83

Annotation pipelines

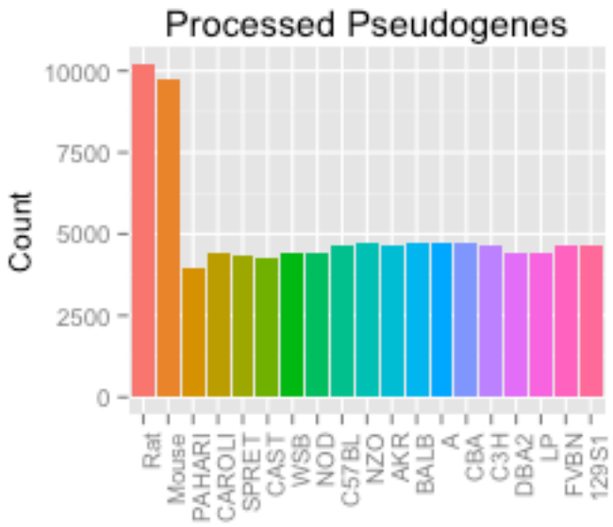
PseudoPipe

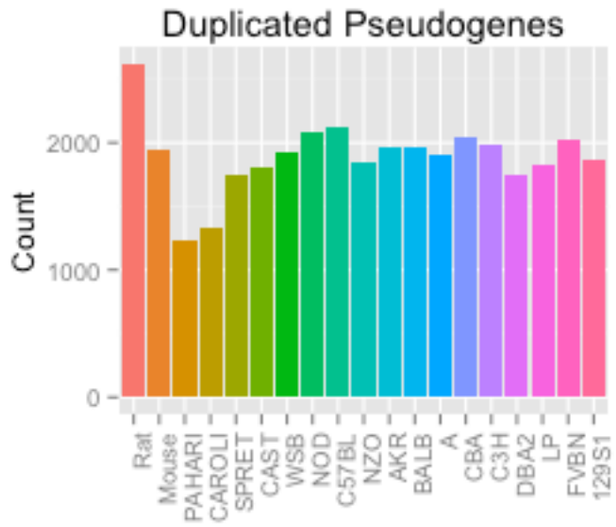


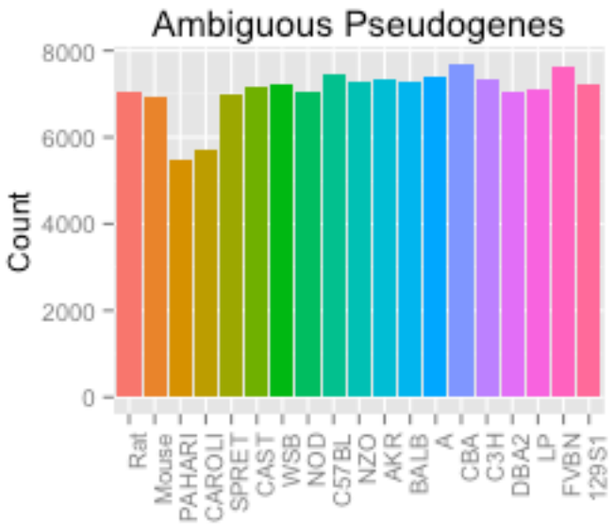
RCPedia

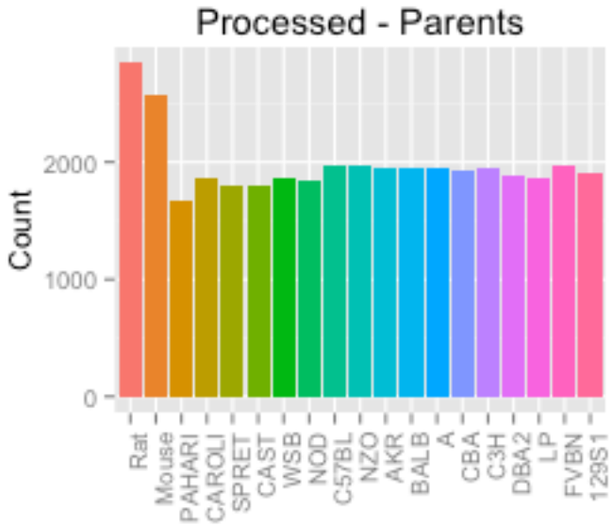


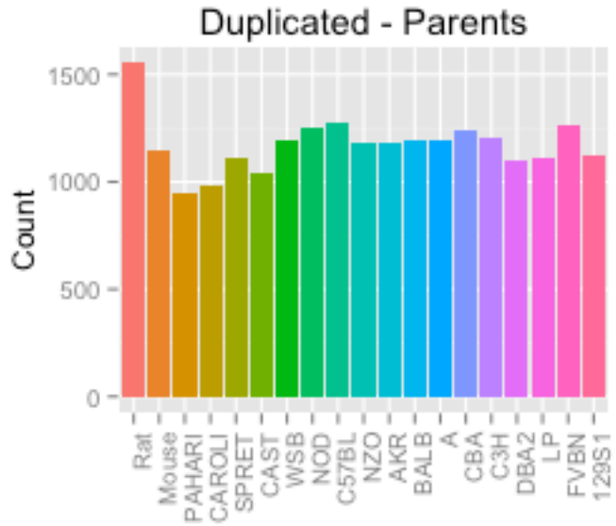
Pseudopipe predictions

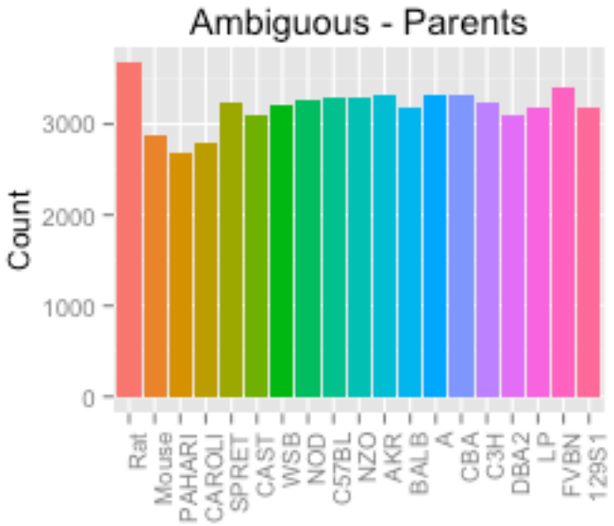




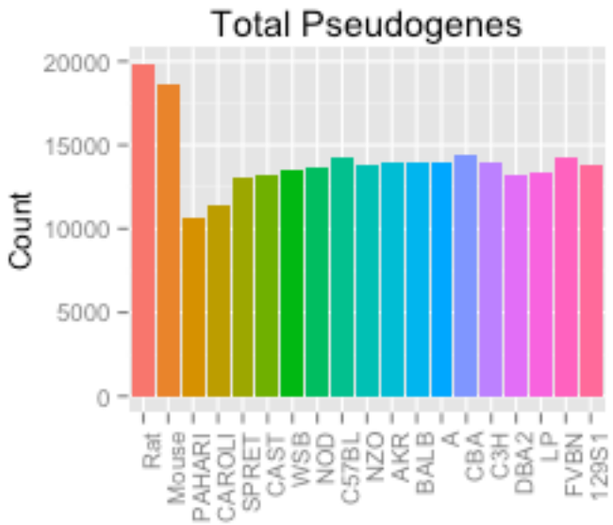


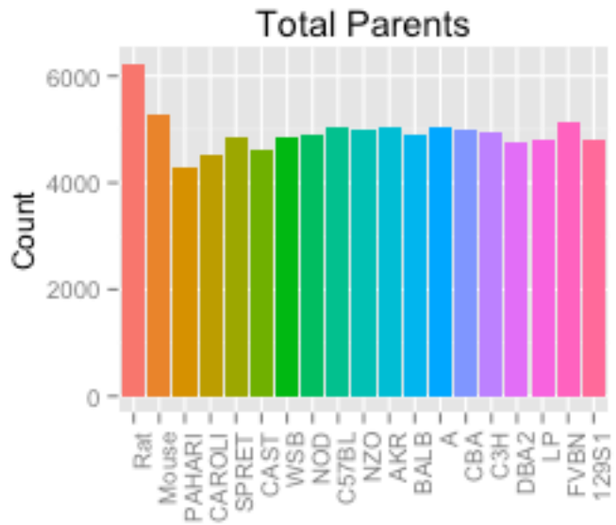


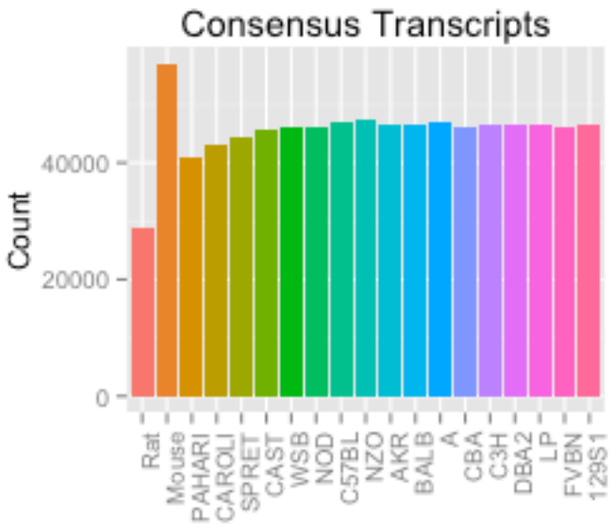




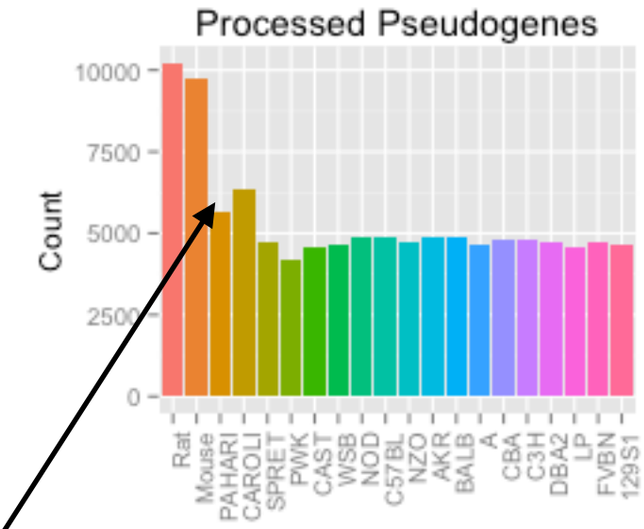
CGP consensus v01

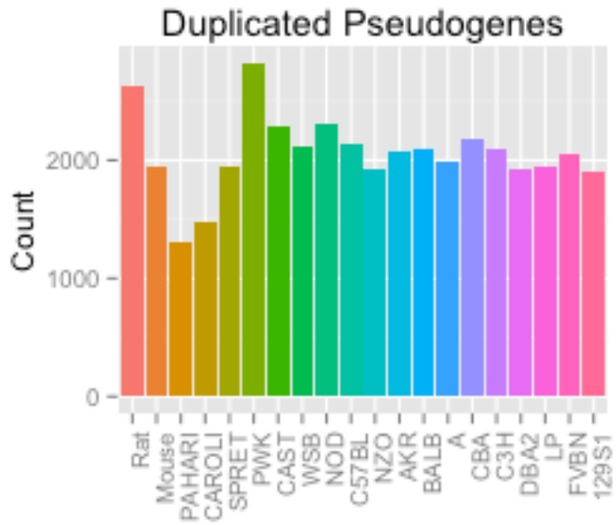


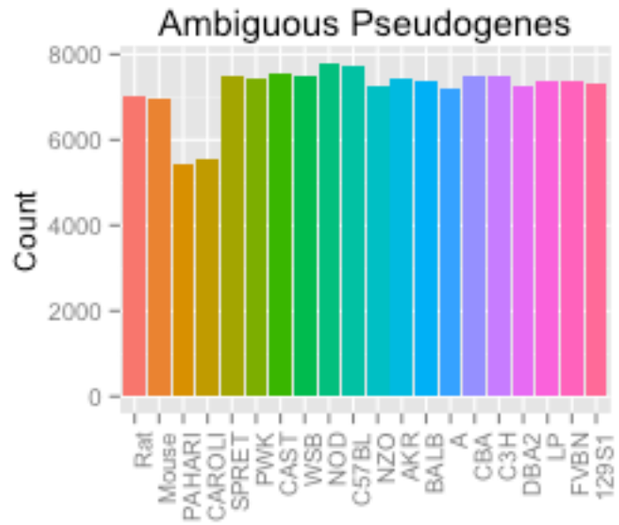




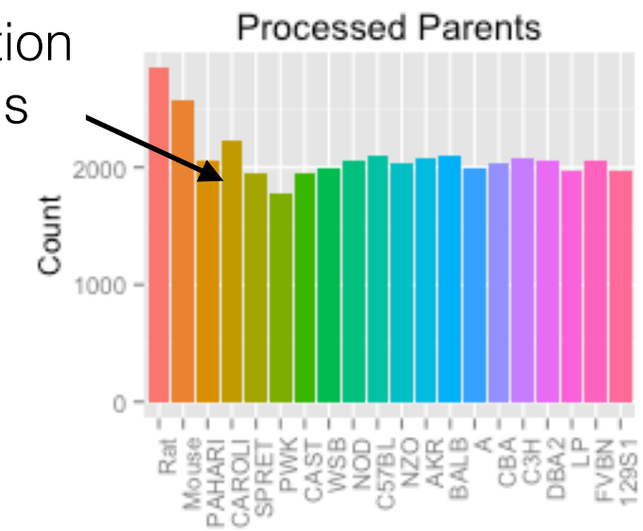
Pseudopipe predictions

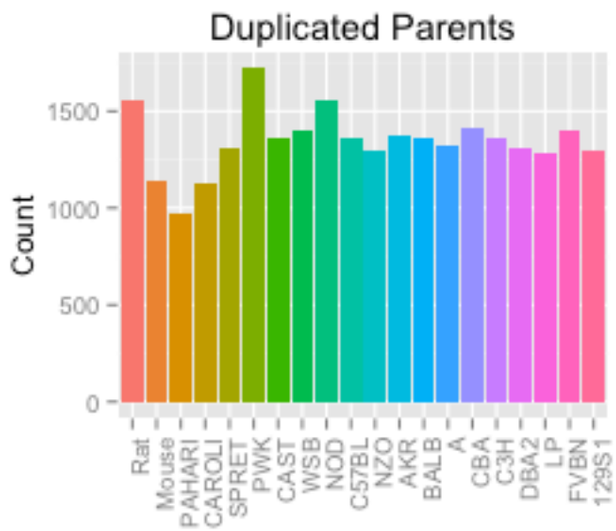


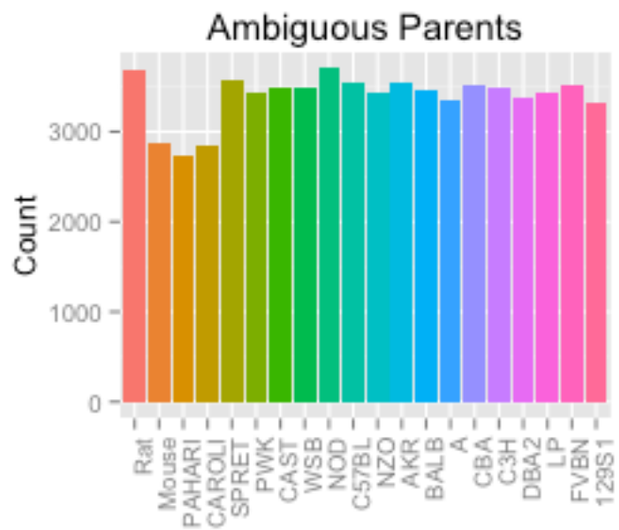




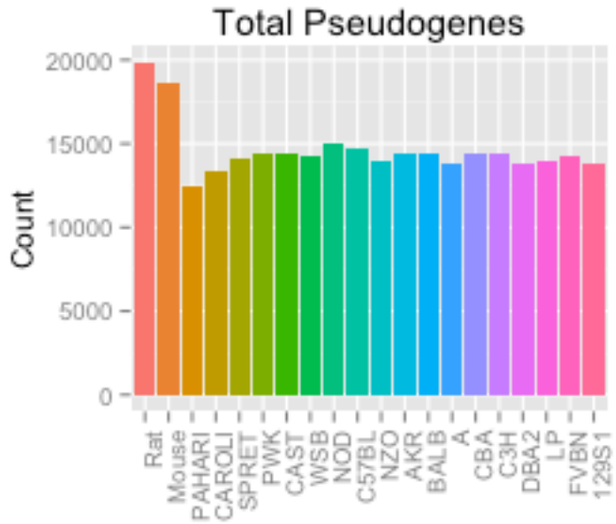
* improving the annotation for more distant strains

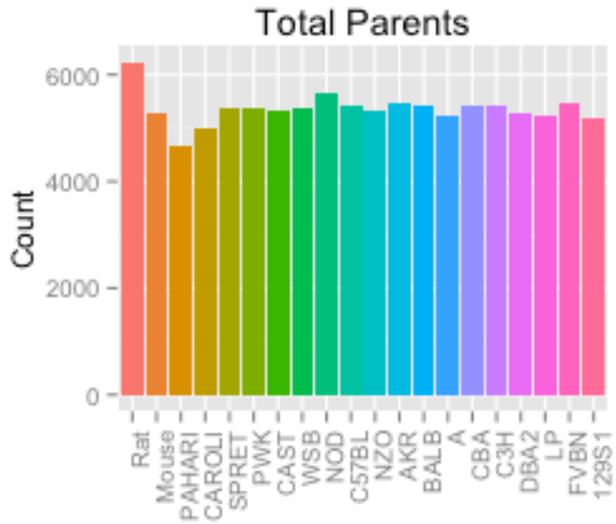


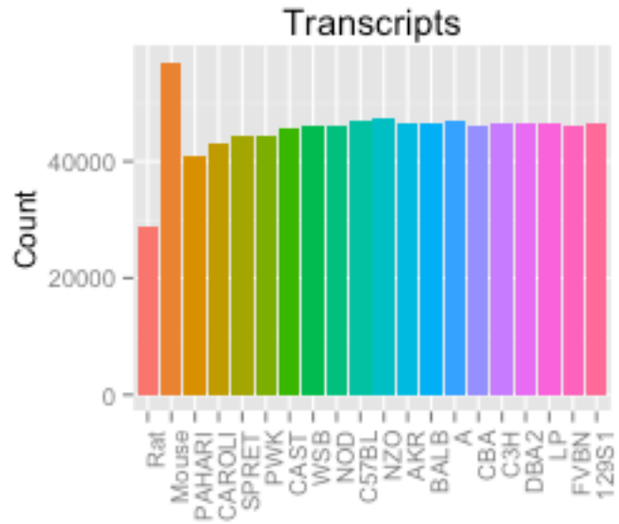




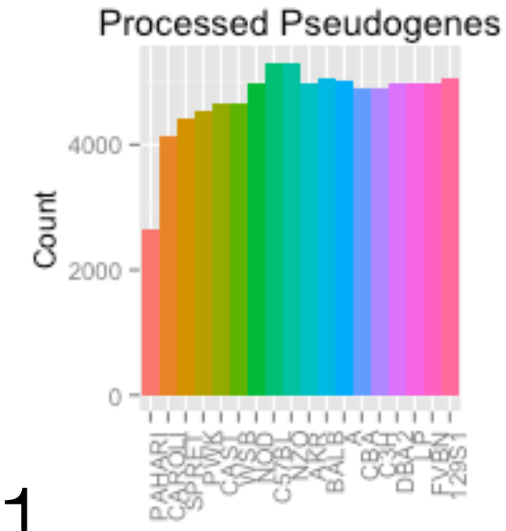
CGP consensus v02

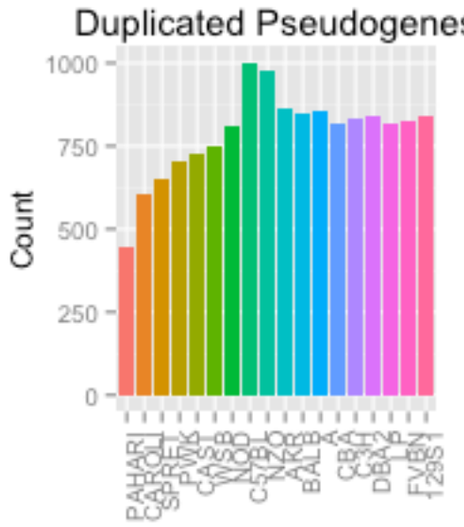


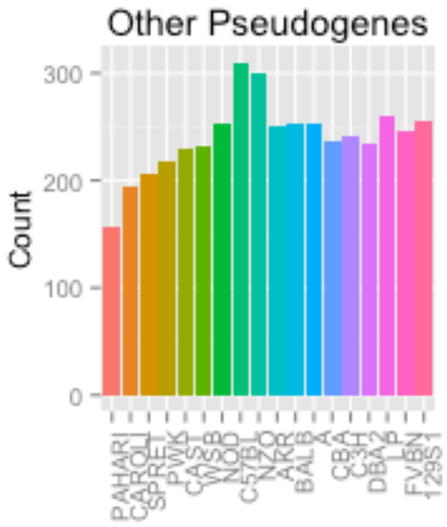


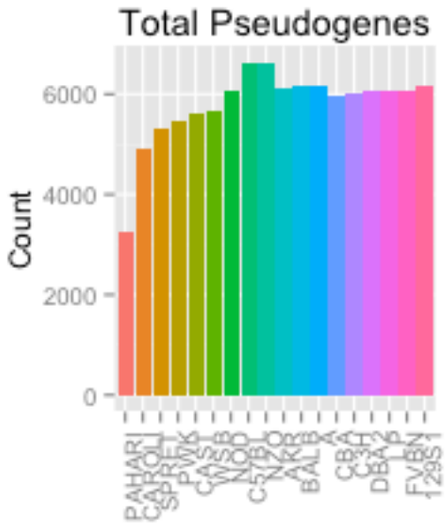


Transmap v01

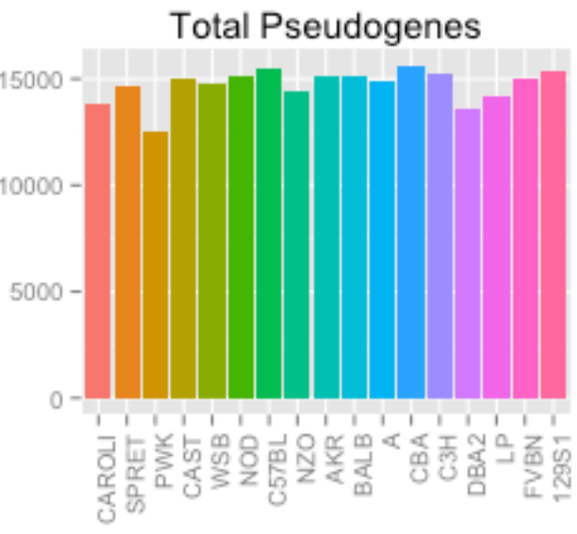


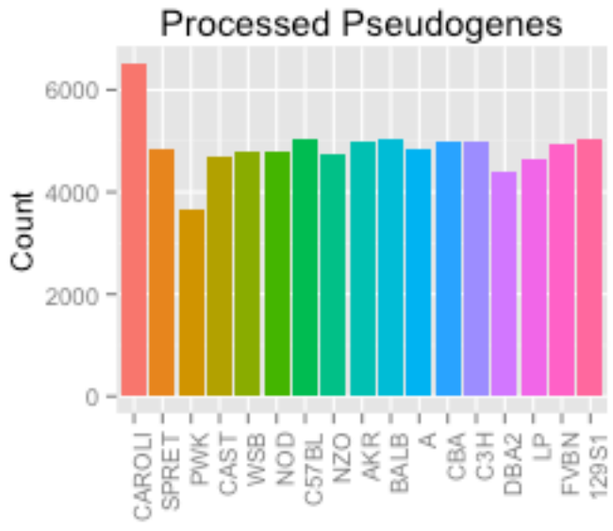


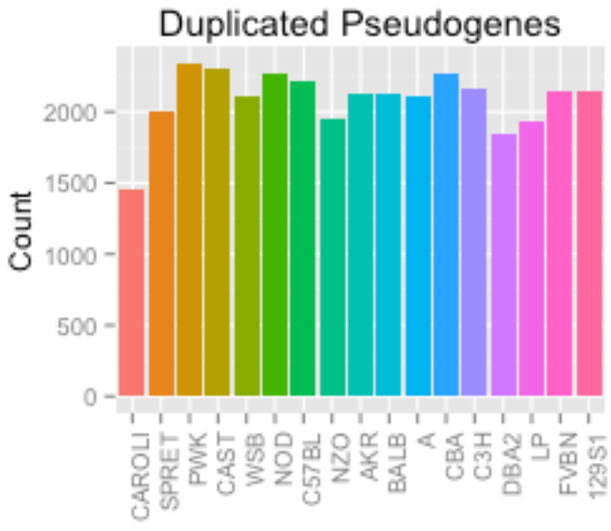


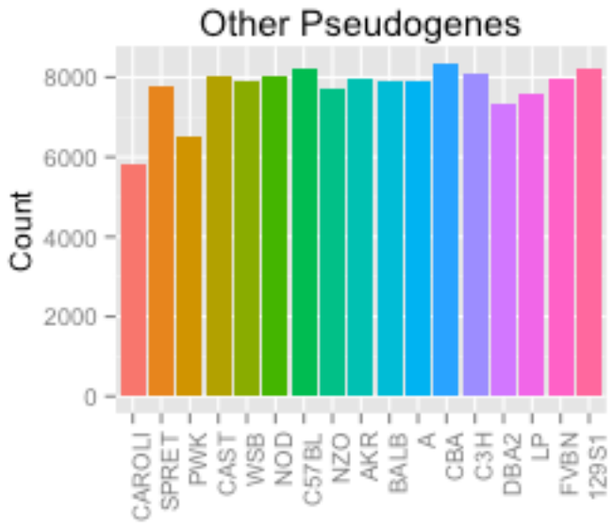


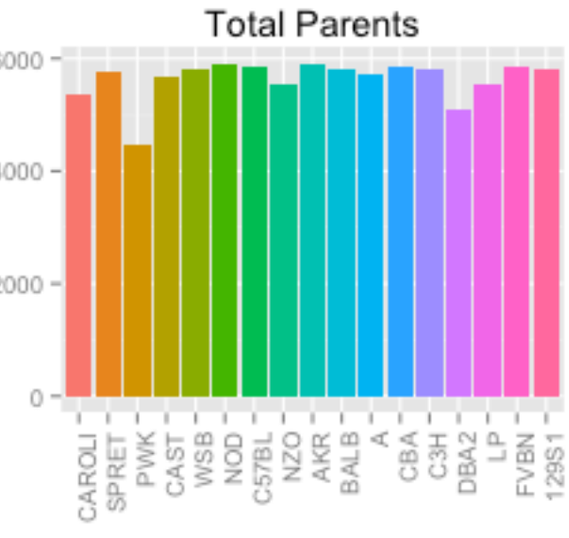
Augustus v01

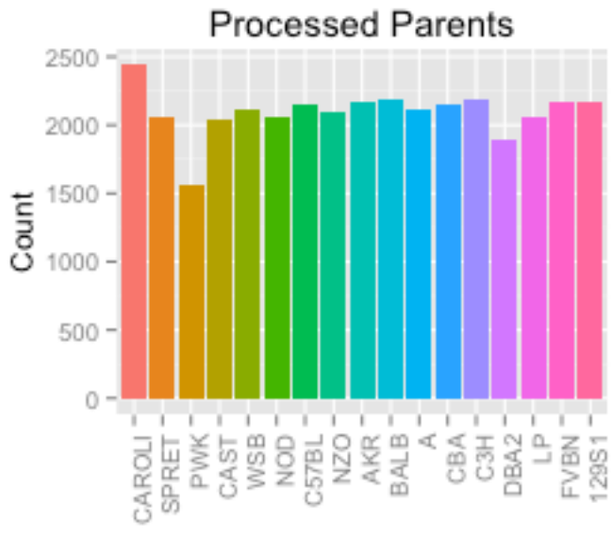


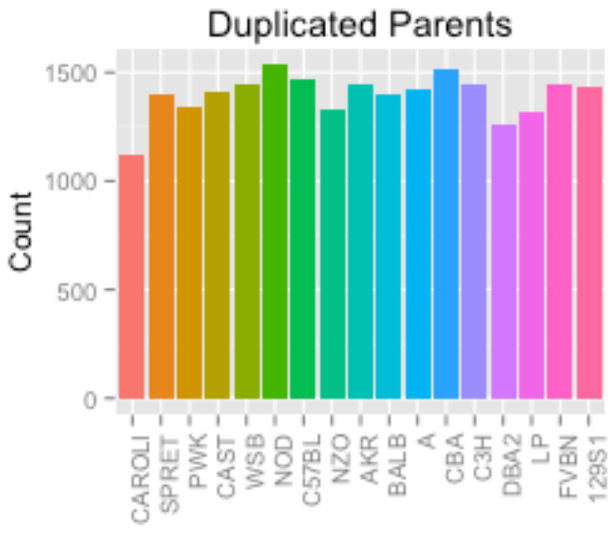


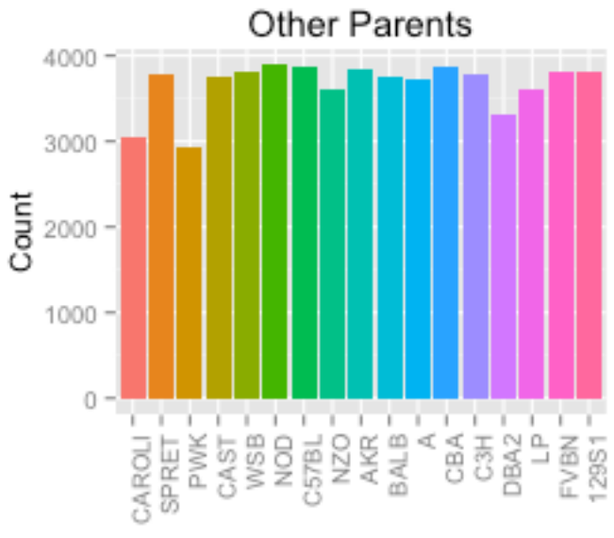










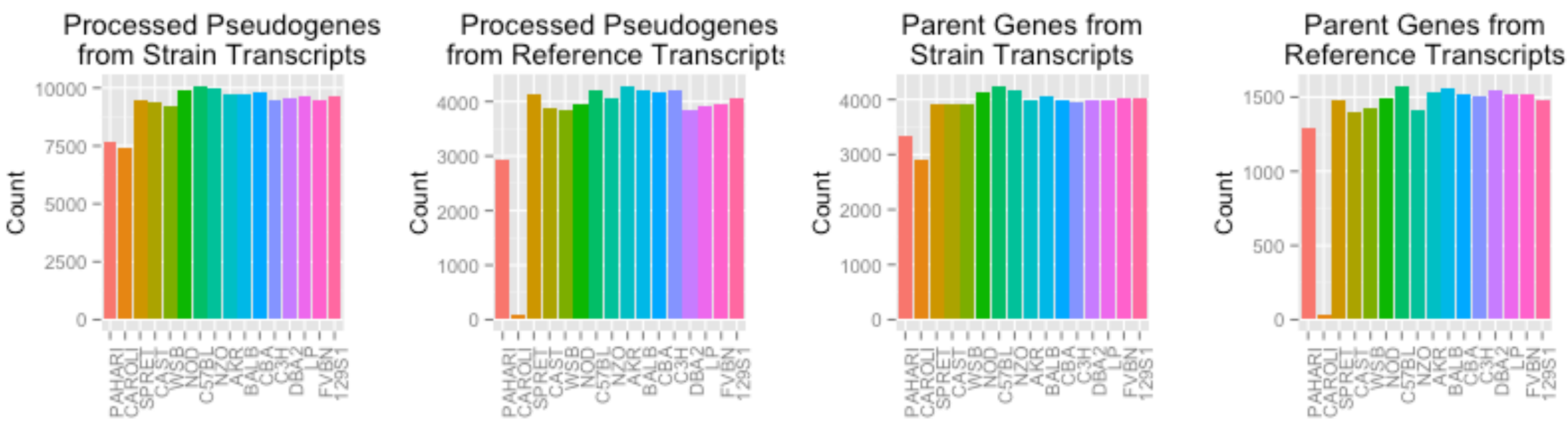


PseudoPipe annotation consistency

	CGPv02-v01	CGPv02-MM8liftover	CGPv01-MM8liftover	Aug-MM8Liftover
129S1	12593	4692	4749	5022
A	12877	4639	4742	4985
AKR	13359	4805	4600	4899
BALB	13341	4971	4796	5109
C3H	13311	5036	4910	5202
C57BL	13902	5292	5221	5516
CAROLI	10260	3988	3306	4020
CAST	12389	4465	4107	4585
CBA	13469	4706	4637	4876
DBA2	12461	4947	4553	4685
FVBN	13294	4864	4825	5110
LP	12741	4715	4572	4786
NOD	12896	4856	4539	4909
NZO	13433	5106	5075	5181
PAHARI	9370	2326	2002	544
PWK	11933	5772	4877	4877
SPRET	12080	4111	3691	4173
WSB	12080	4830	4636	4919

RCPedia annotation CGPv02

- Uses input strain annotated transcripts and mouse reference genome GENCODE MM8 annotated transcripts



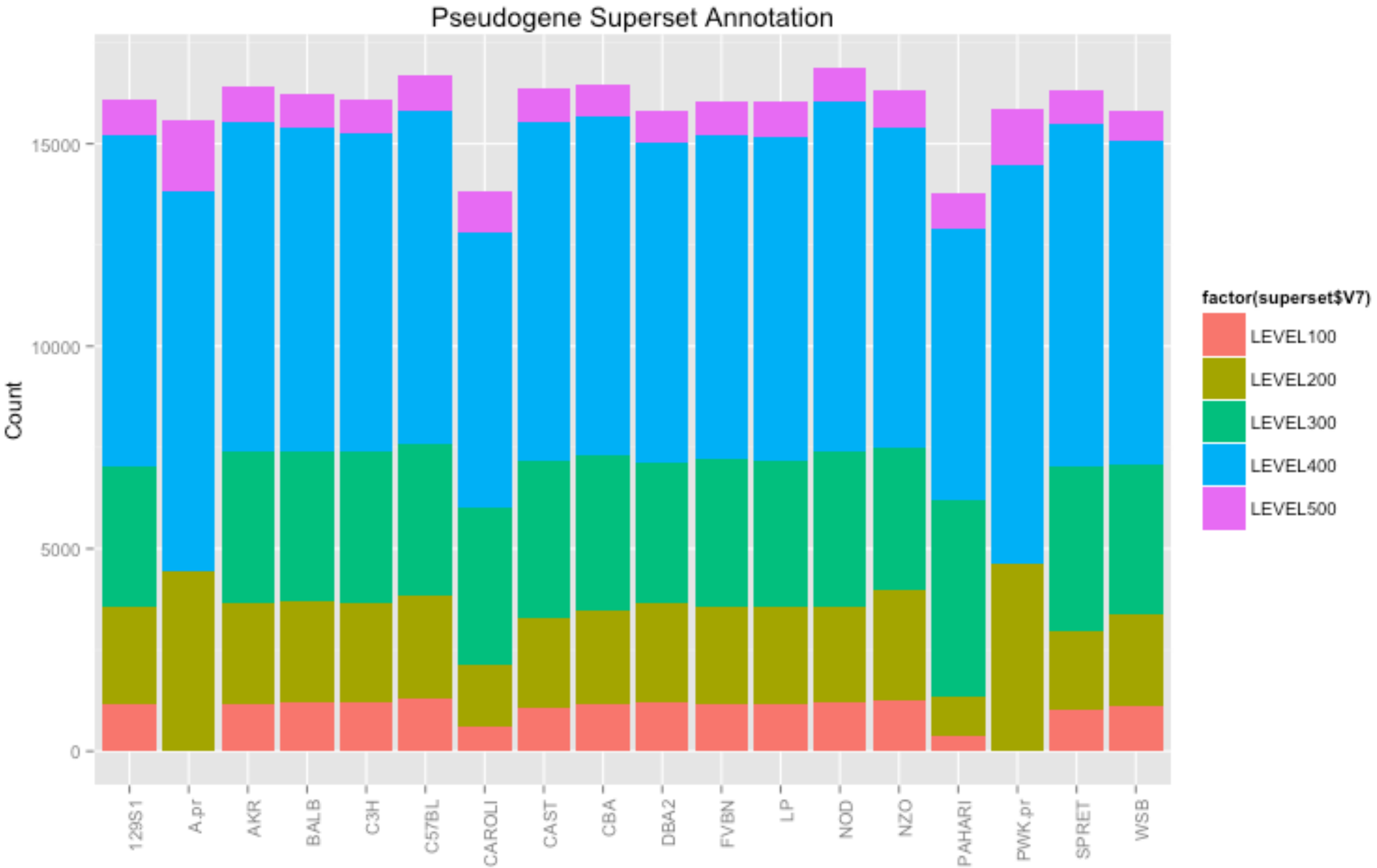
The pseudogene annotation is highly dependent on the accuracy of the input protein coding annotation

~ CGPv02 input protein coding annotation ~

Pseudopipe misses the pseudogenes resulting from strain specific transcripts

Pseudogene Super set

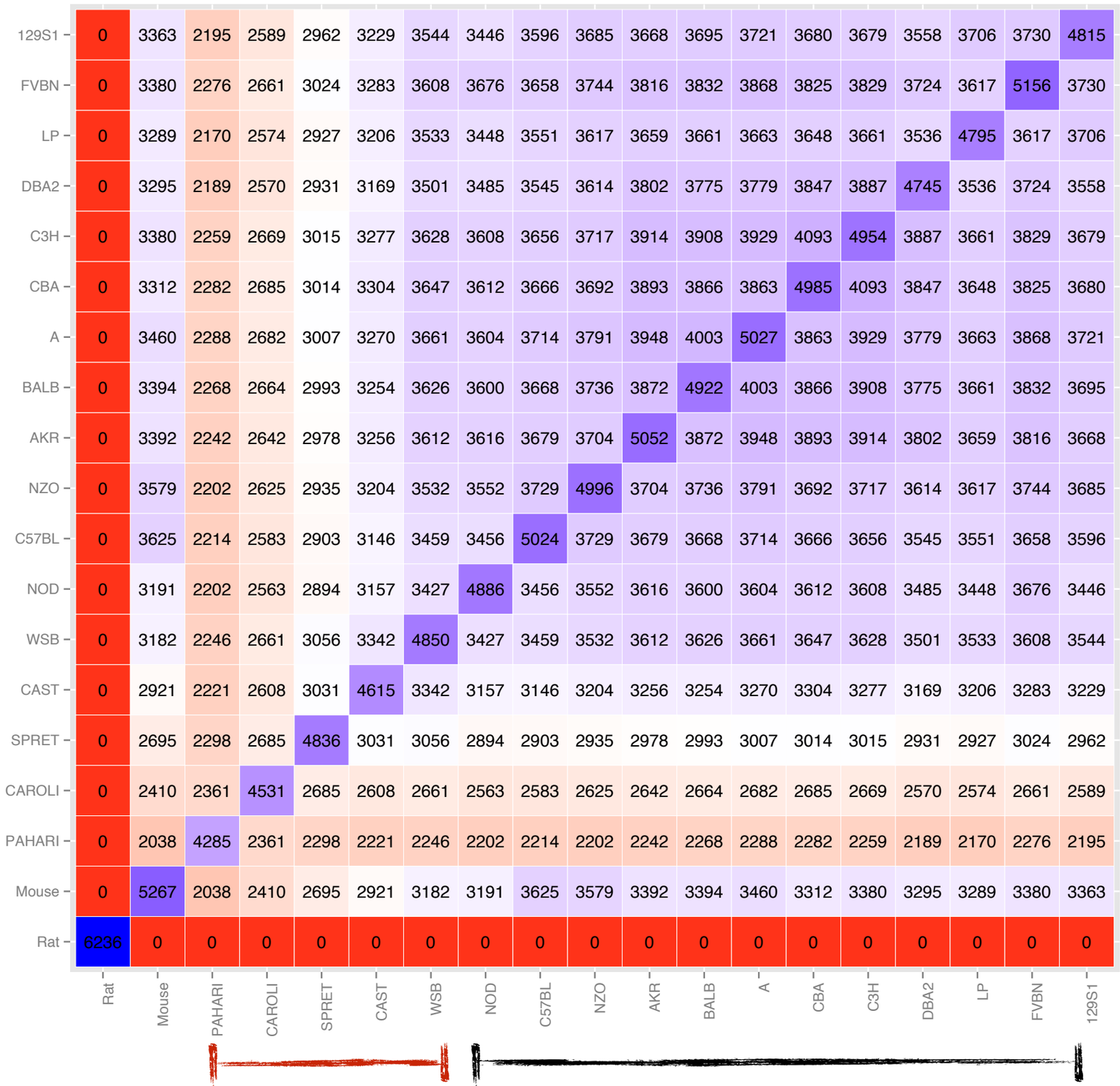
- Using the union of RCPedia & Pseudopipe as well as overlapping with the liftover from manually annotated Mouse reference pseudogene sets.
    - LEVEL100 - pseudogenes identified by both pipelines and by manual annotation
    - LEVEL200 - pseudogenes identified by only 1 pipeline and by manual annotation
    - LEVEL300 - pseudogenes identified by both automatic pipelines
    - LEVEL400 - pseudogenes identified by only an automatic annotation pipeline
    - LEVEL500 - pseudogenes annotated only using the liftover of manually curated reference genome



* uniform distribution of the annotation levels
* CAROLI and PAHARI have lower numbers of conserved pseudogenes with respect to the reference genome

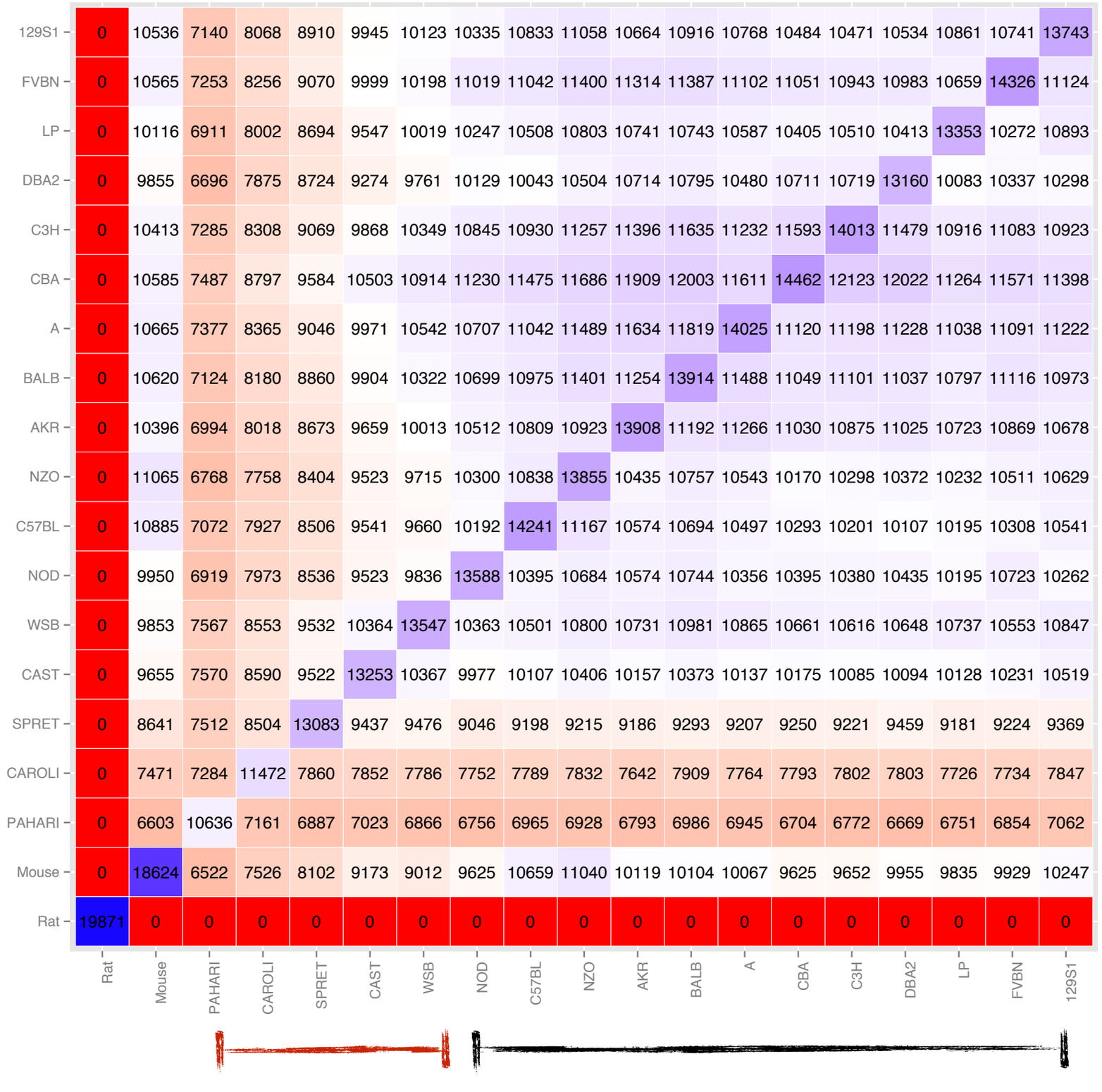
ANNOTATION ANALYSIS

Parent gene conservation



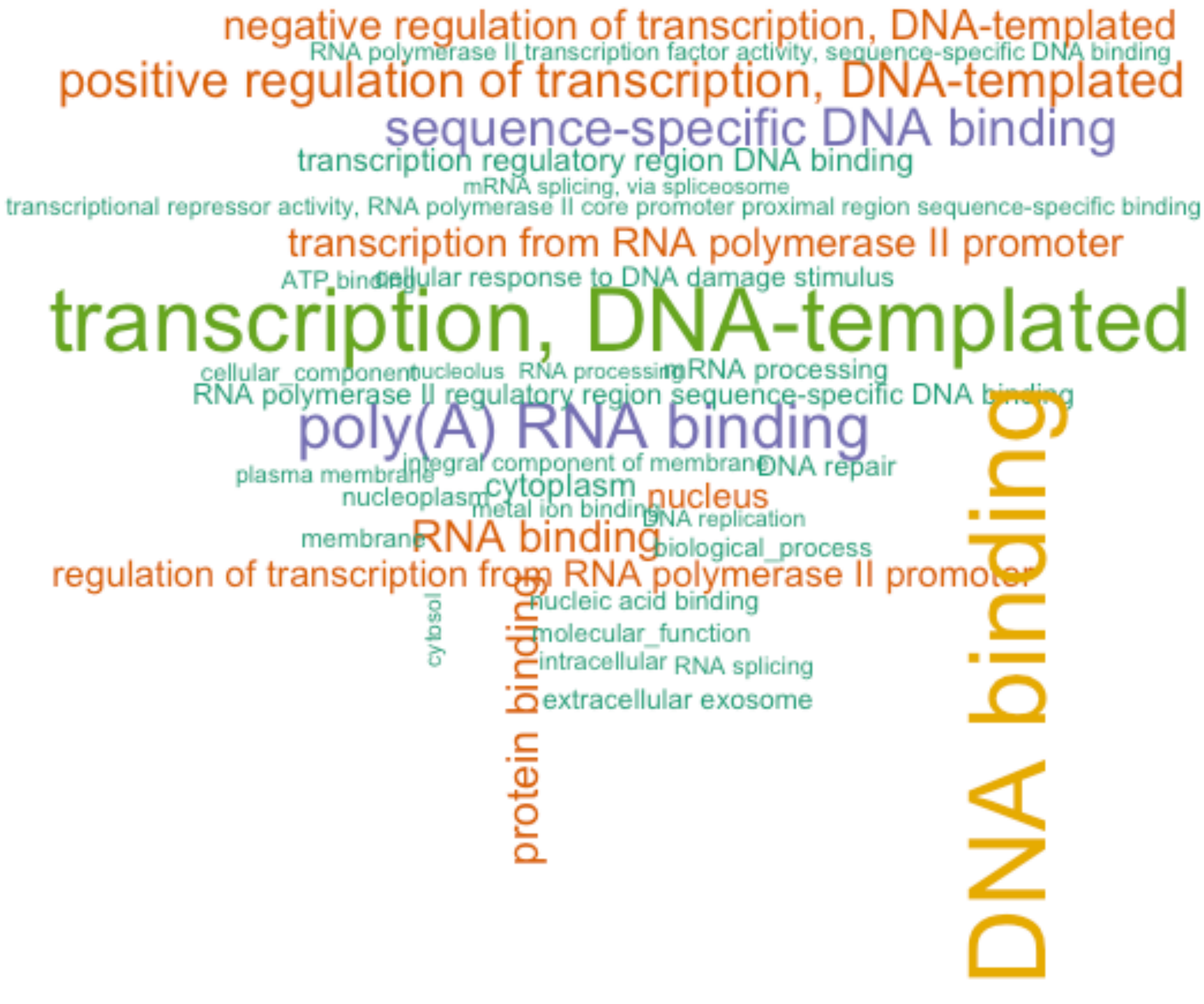
60% of the parent genes are shared between any two strains.

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

What is the biological function of pseudogene parents ?



Top families

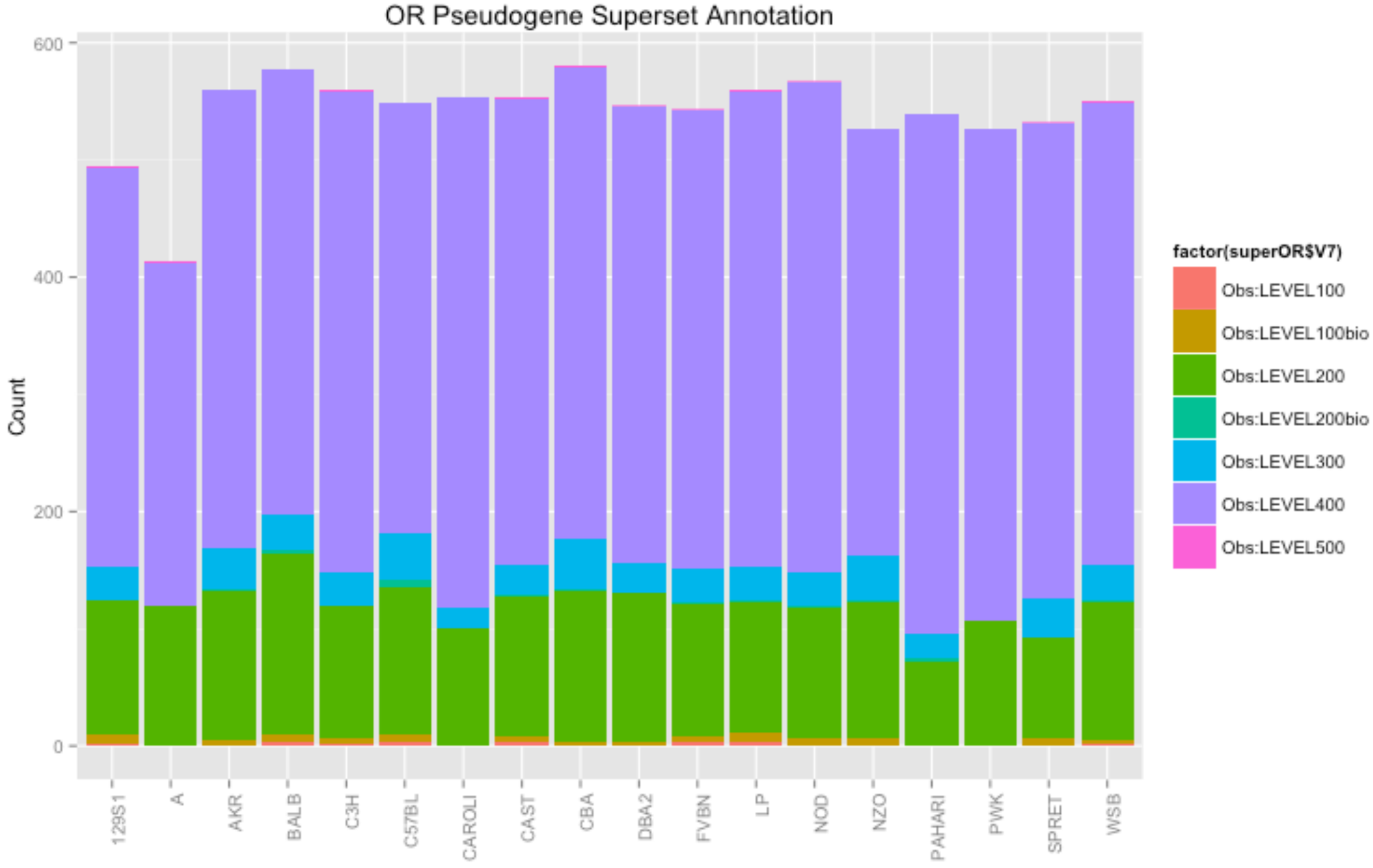
Mouse		PAHARI		CAROLI		SPRET		CAST		WSB	
660	ZnF	437	GPCR	461	GPCR	523	GPCR	463	GPCR	524	GPCR
486	Struct	252	ZnF	241	RRM	337	ZnF	271	ZnF	312	ZnF
480	Trm112p	228	RRM	237	ZnF	290	RRM	270	RRM	259	RRM
467	TF	176	TF	157	Ribo	247	TF	224	TF	222	ANF
434	GpDhC	153	Ribo	152	EGF	222	Kinase	224	ANF	220	GPCR
429	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

Case study: Olfactory Receptors

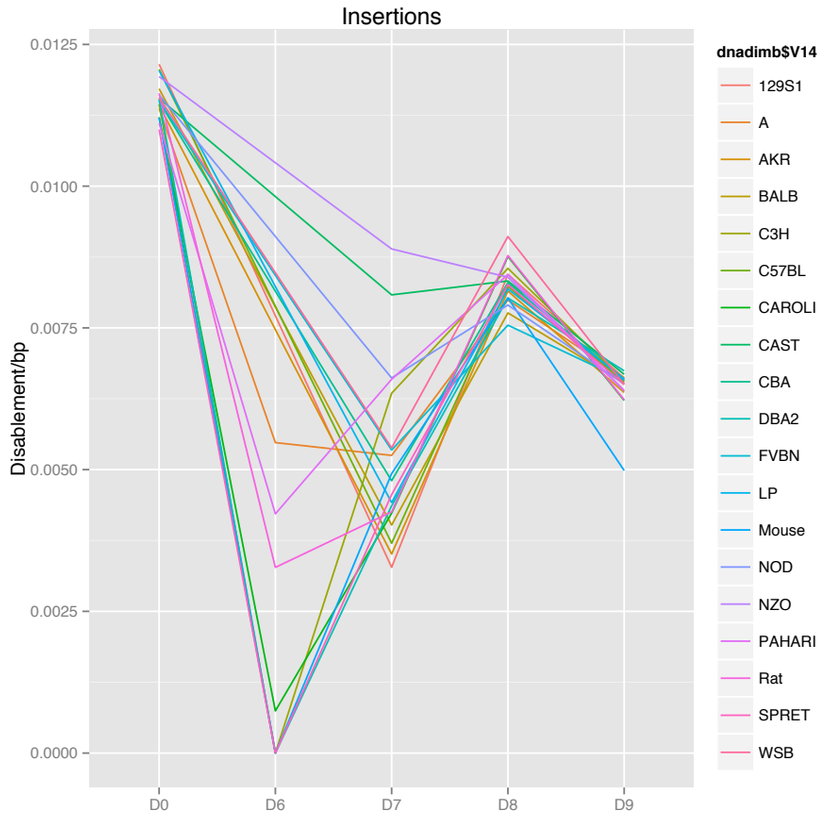
- 1294 transcripts & 1106 OR genes in mouse reference genome

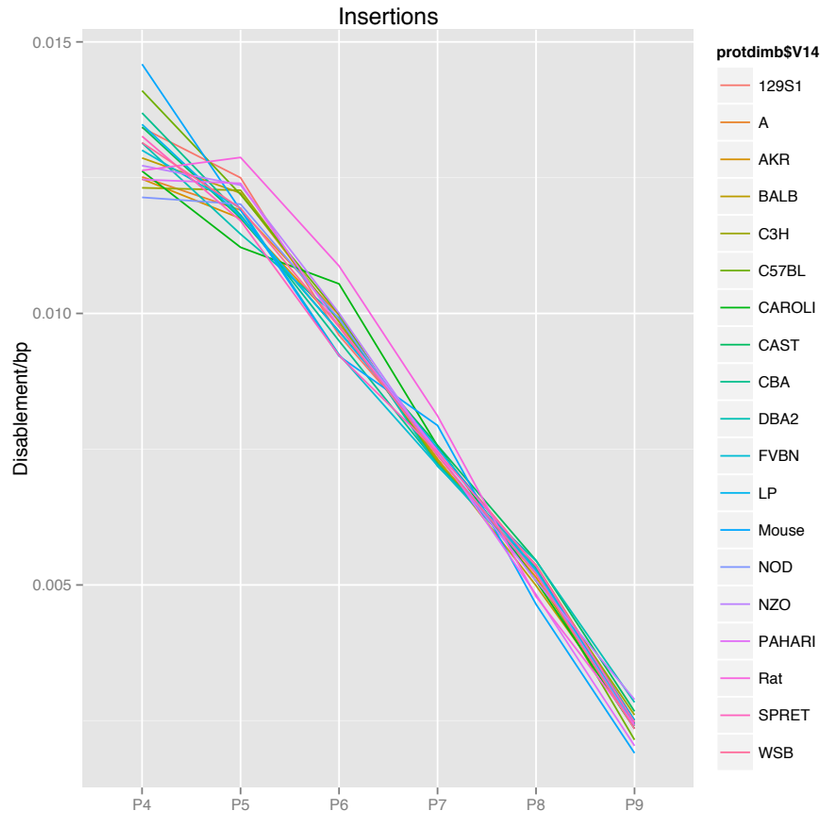


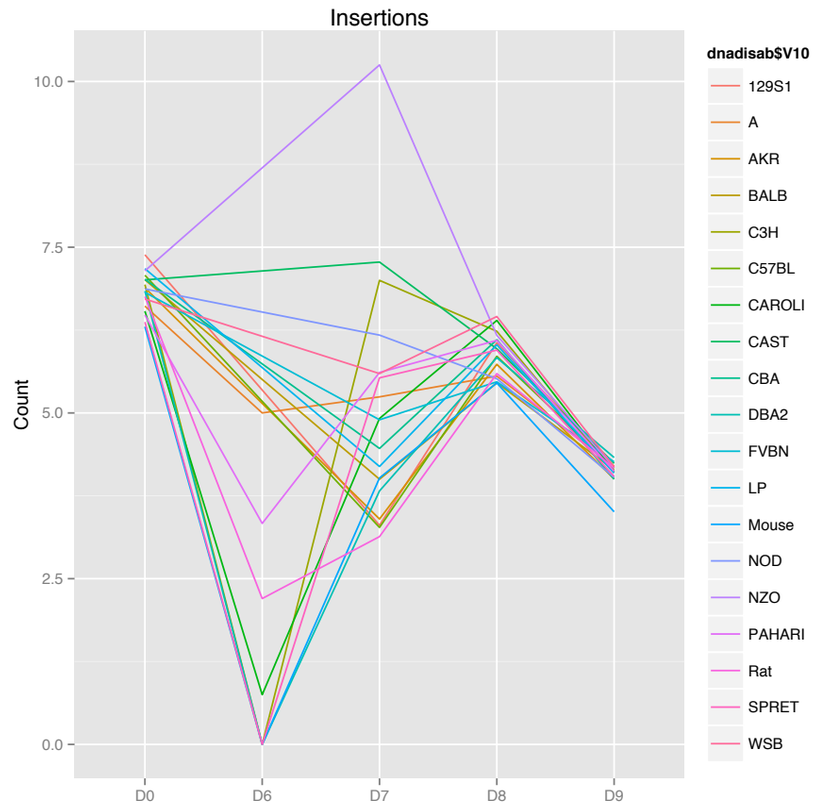
OR pseudogenes are conserved across all strains

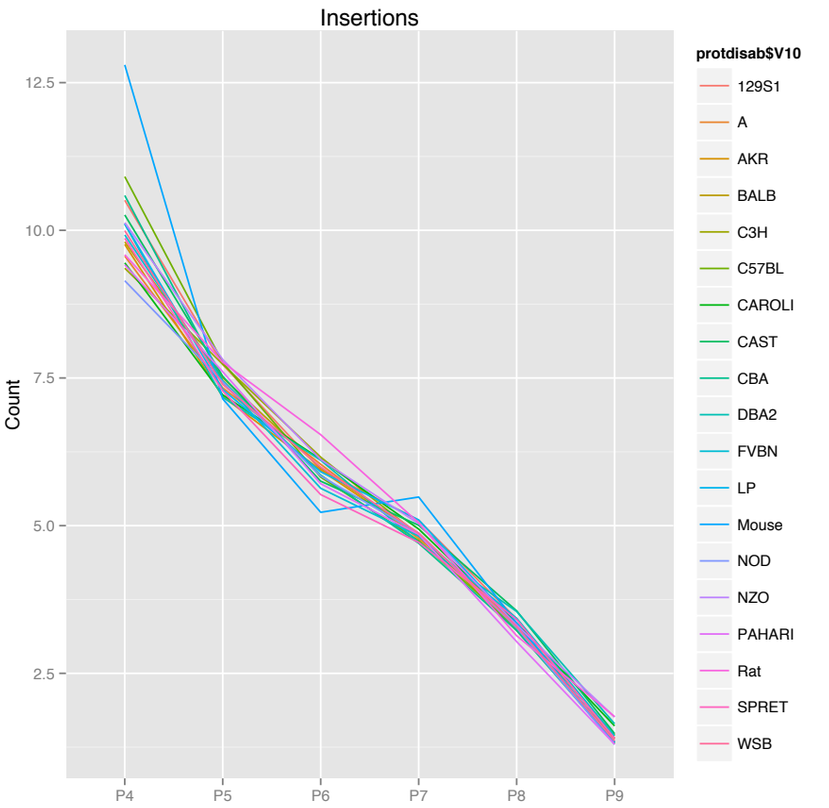
PSEUDOGENE SEQUENCE ANALYSIS

Average disablements distribution as function of sequence similarity to parent genes



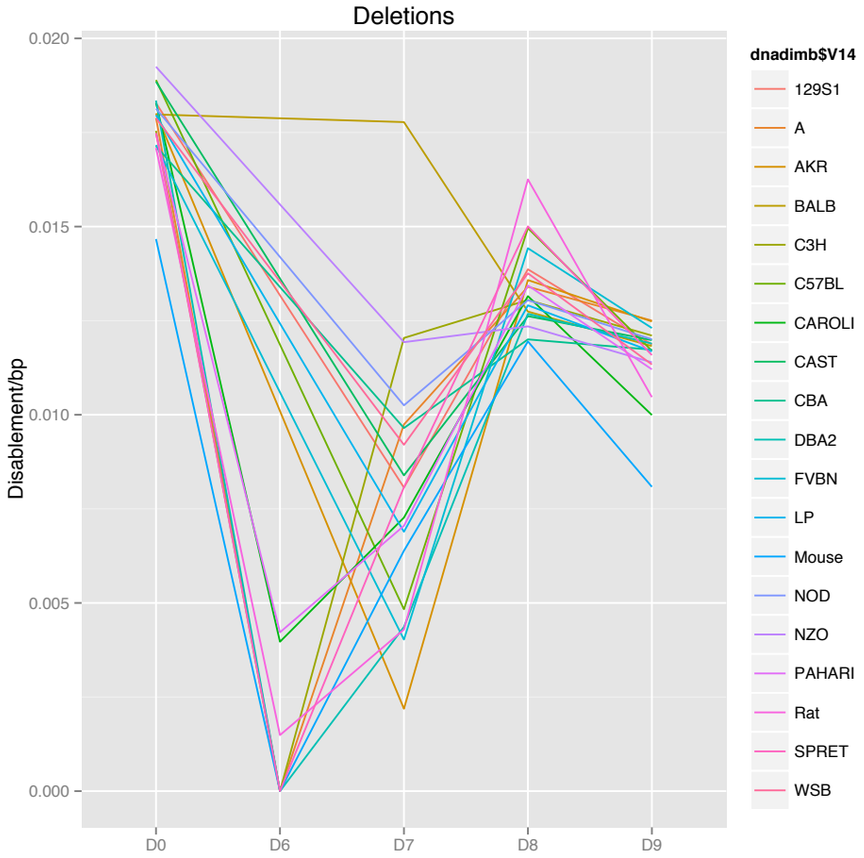


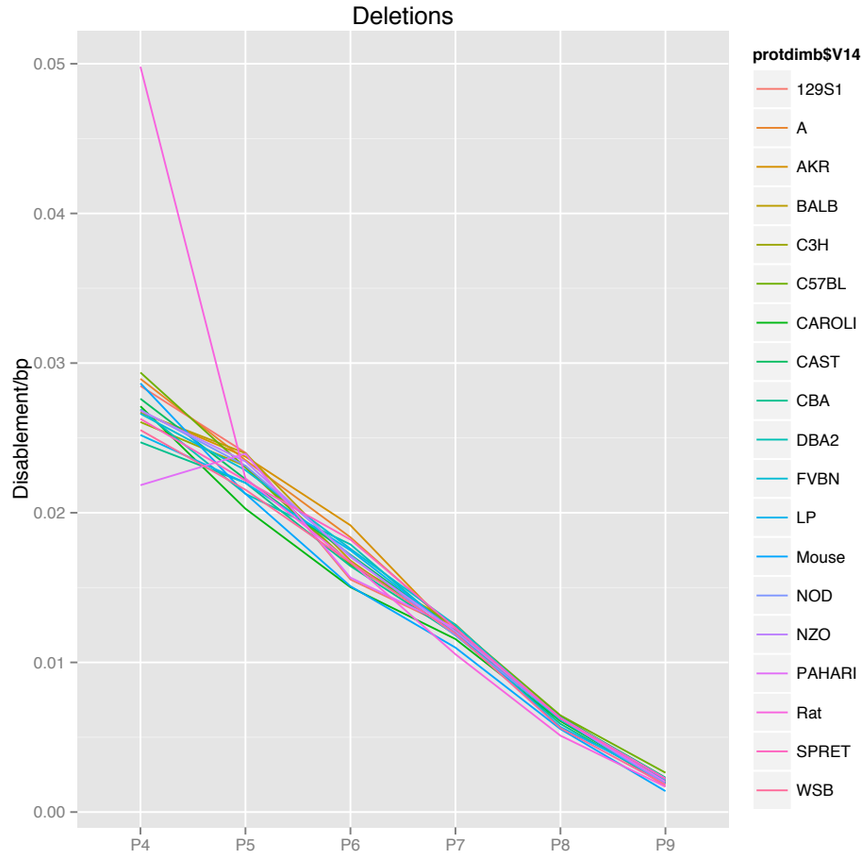


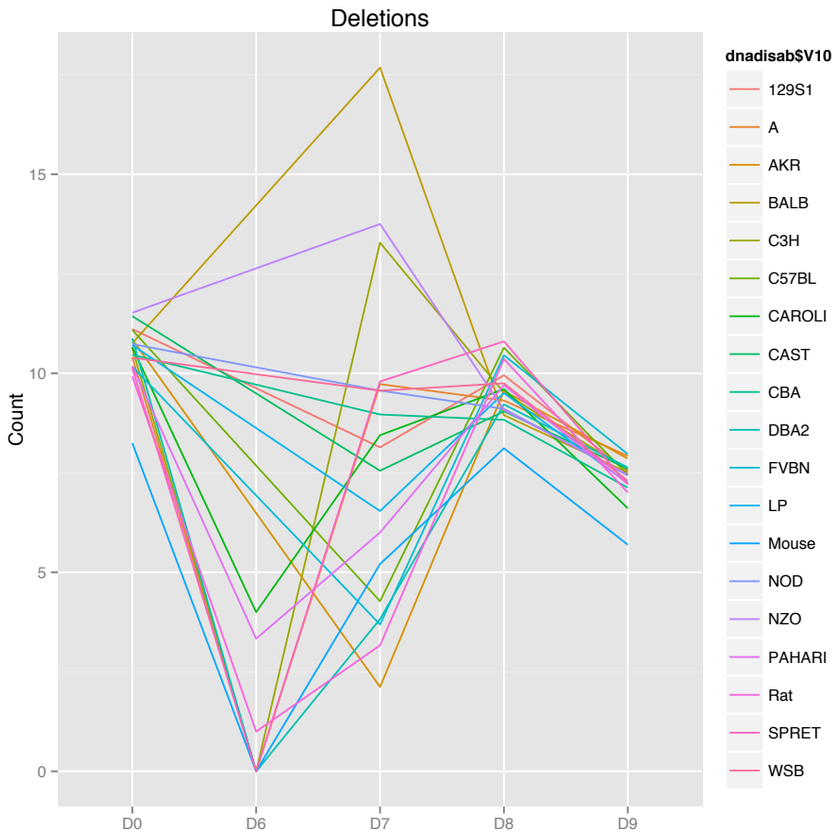


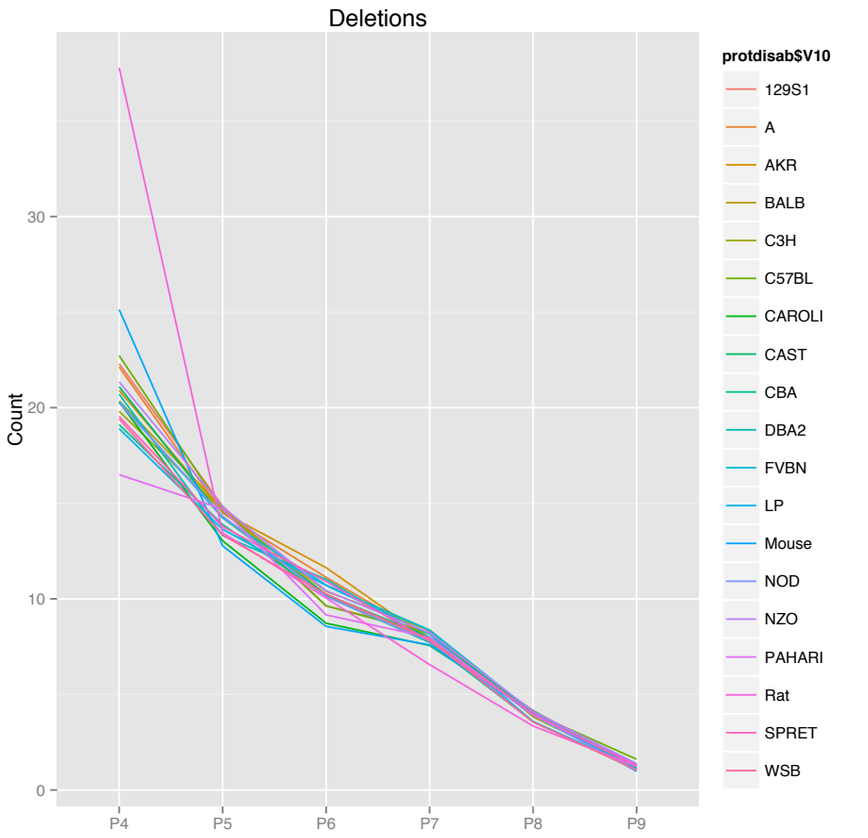
DNA similarity

AA similarity



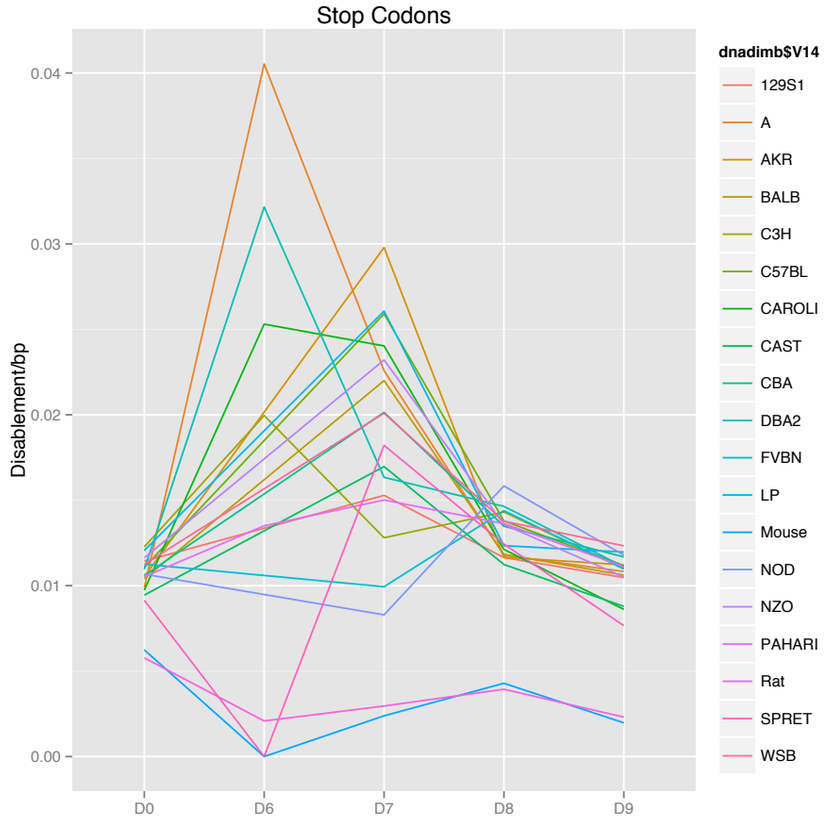


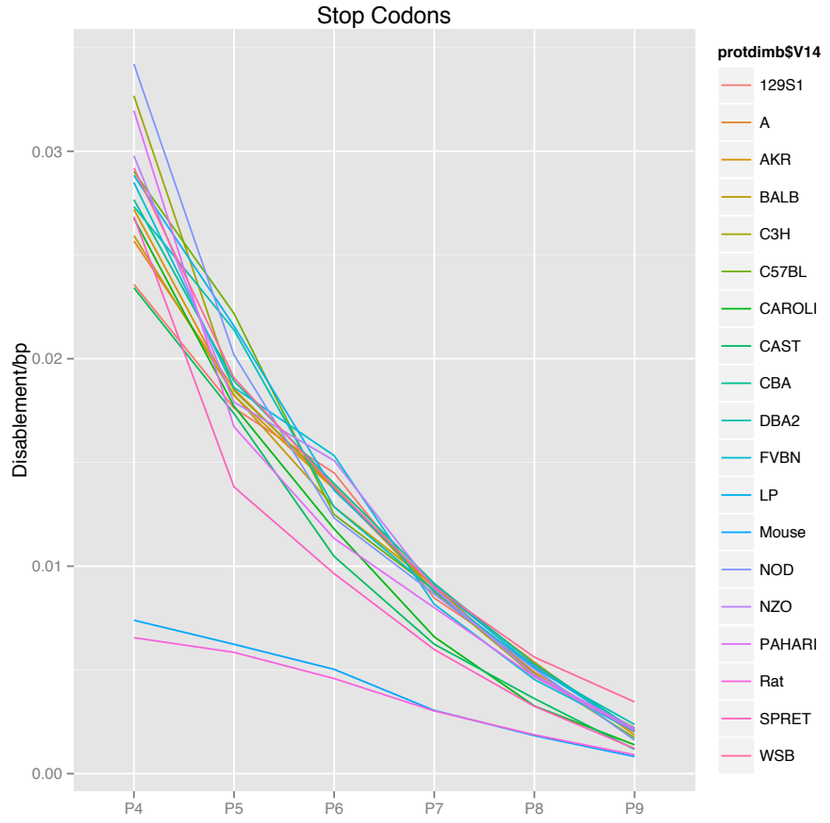


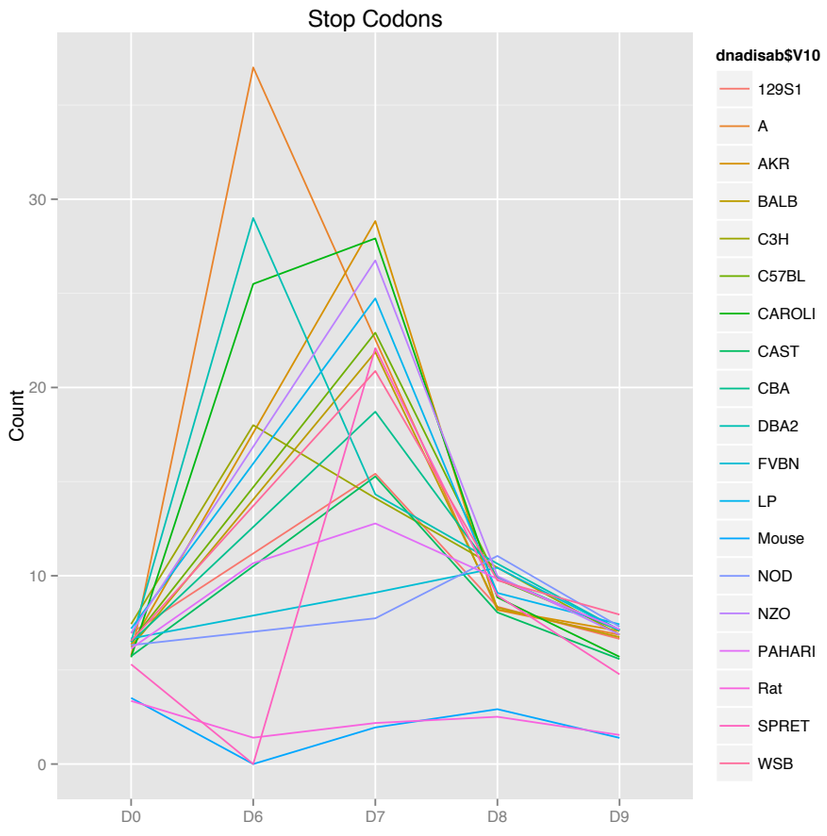


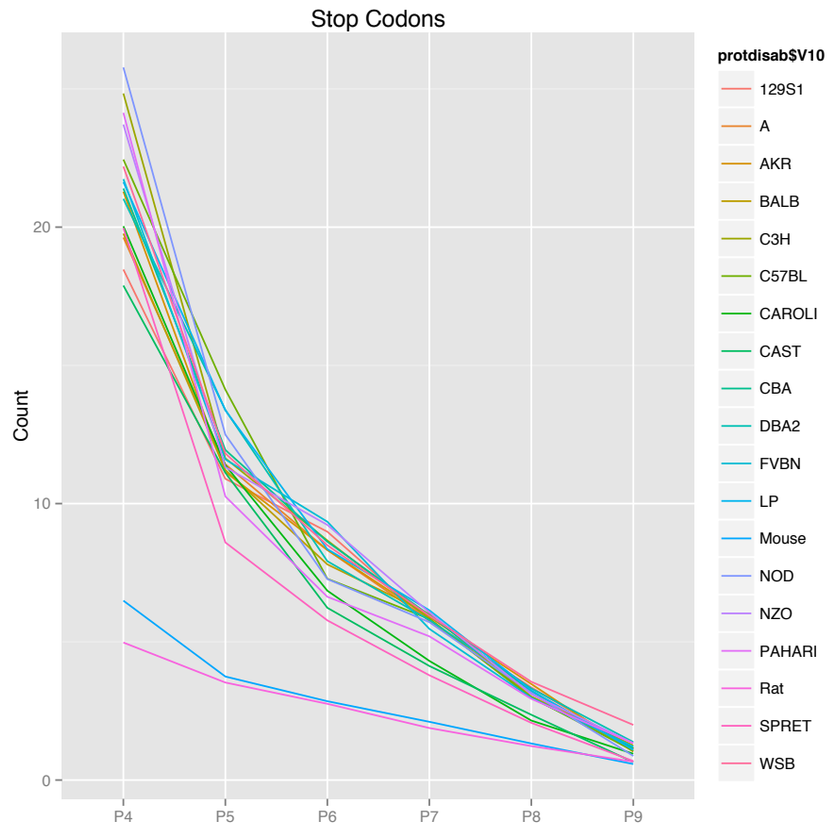
DNA similarity

AA similarity



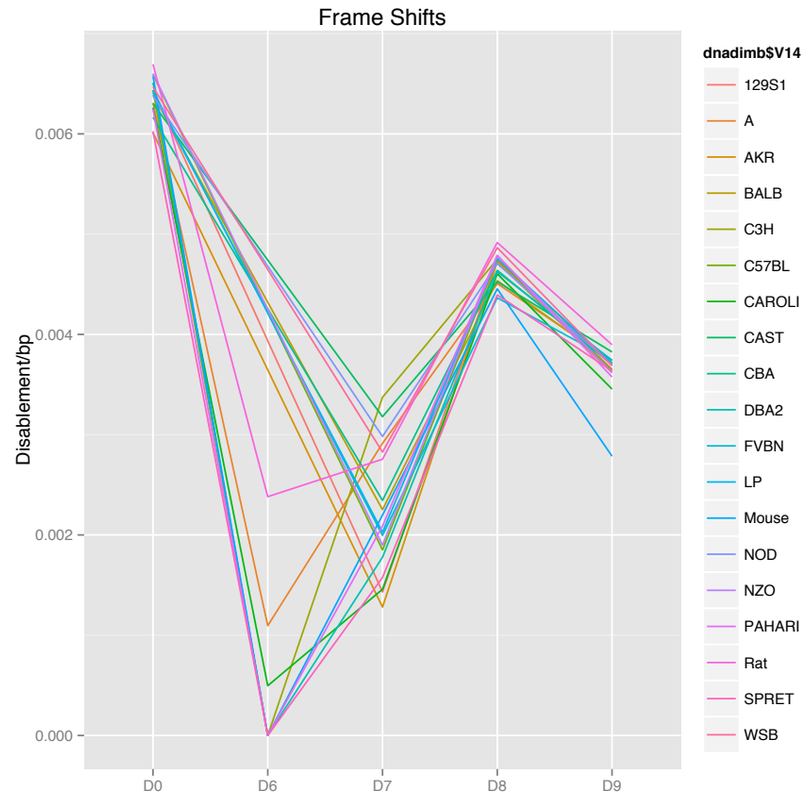


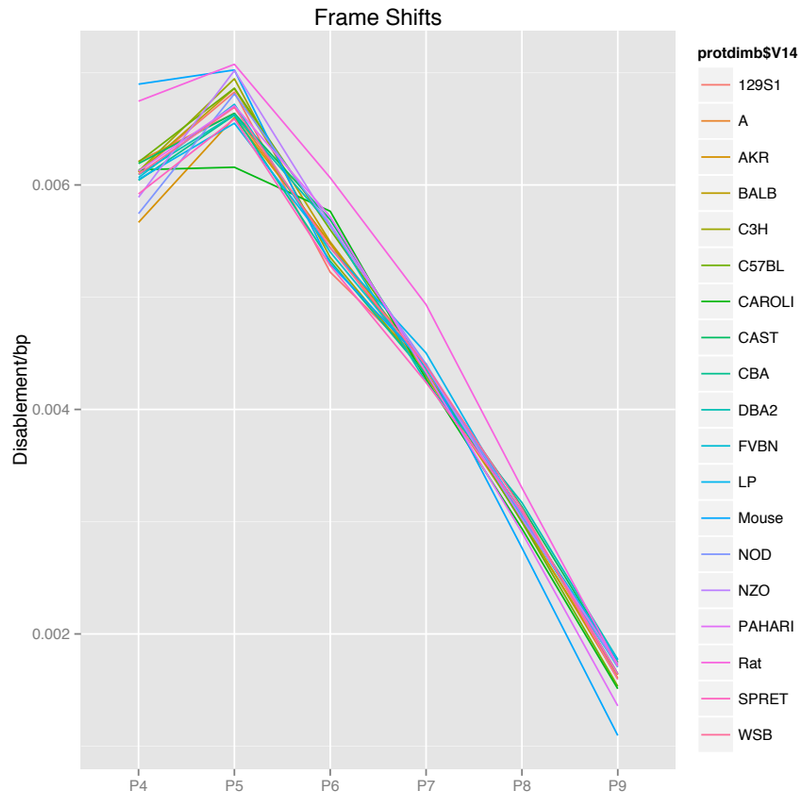


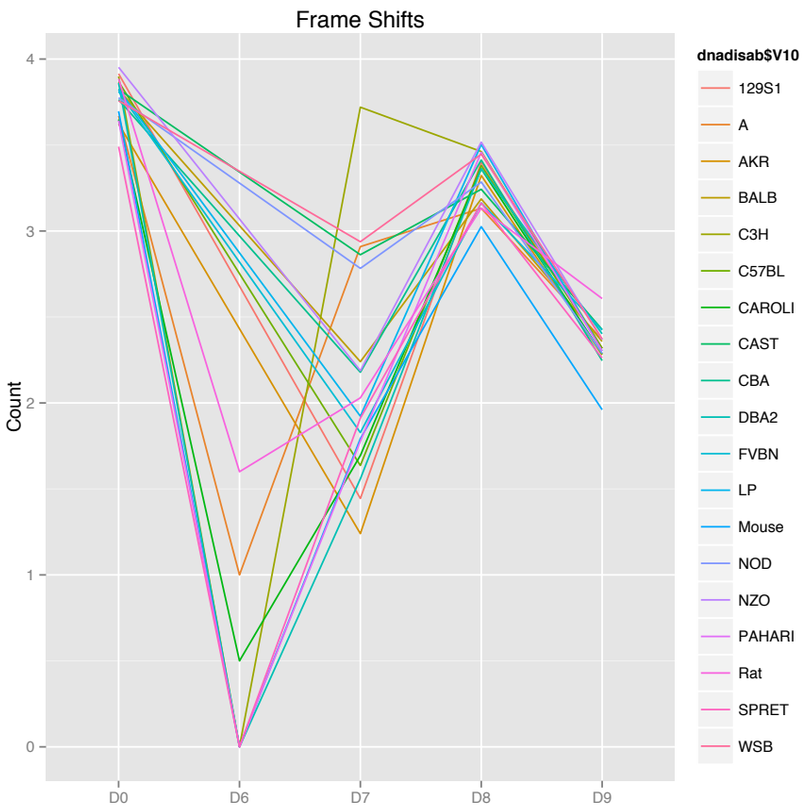


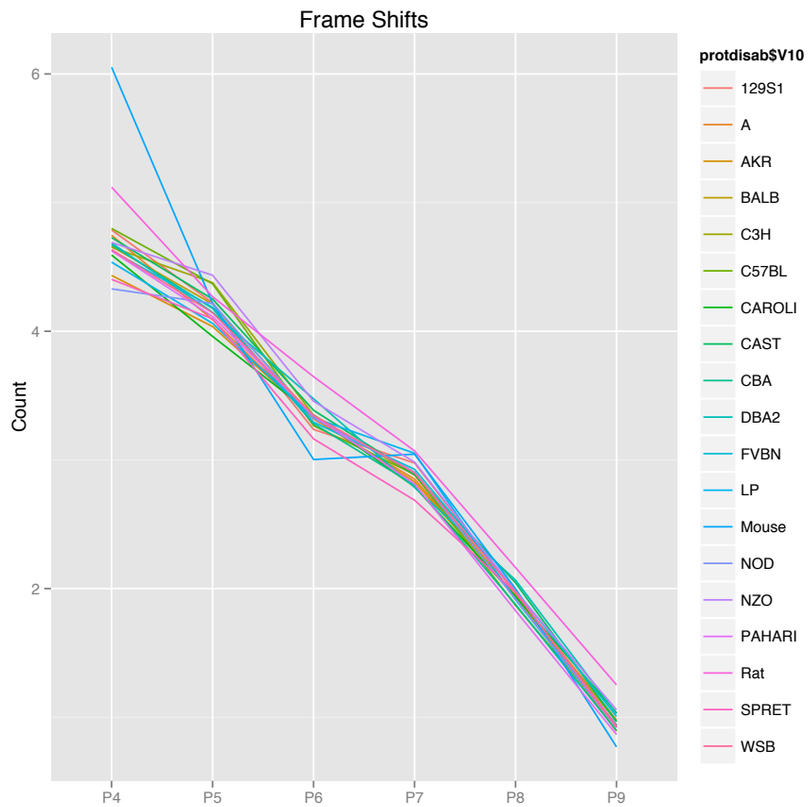
DNA similarity

AA similarity







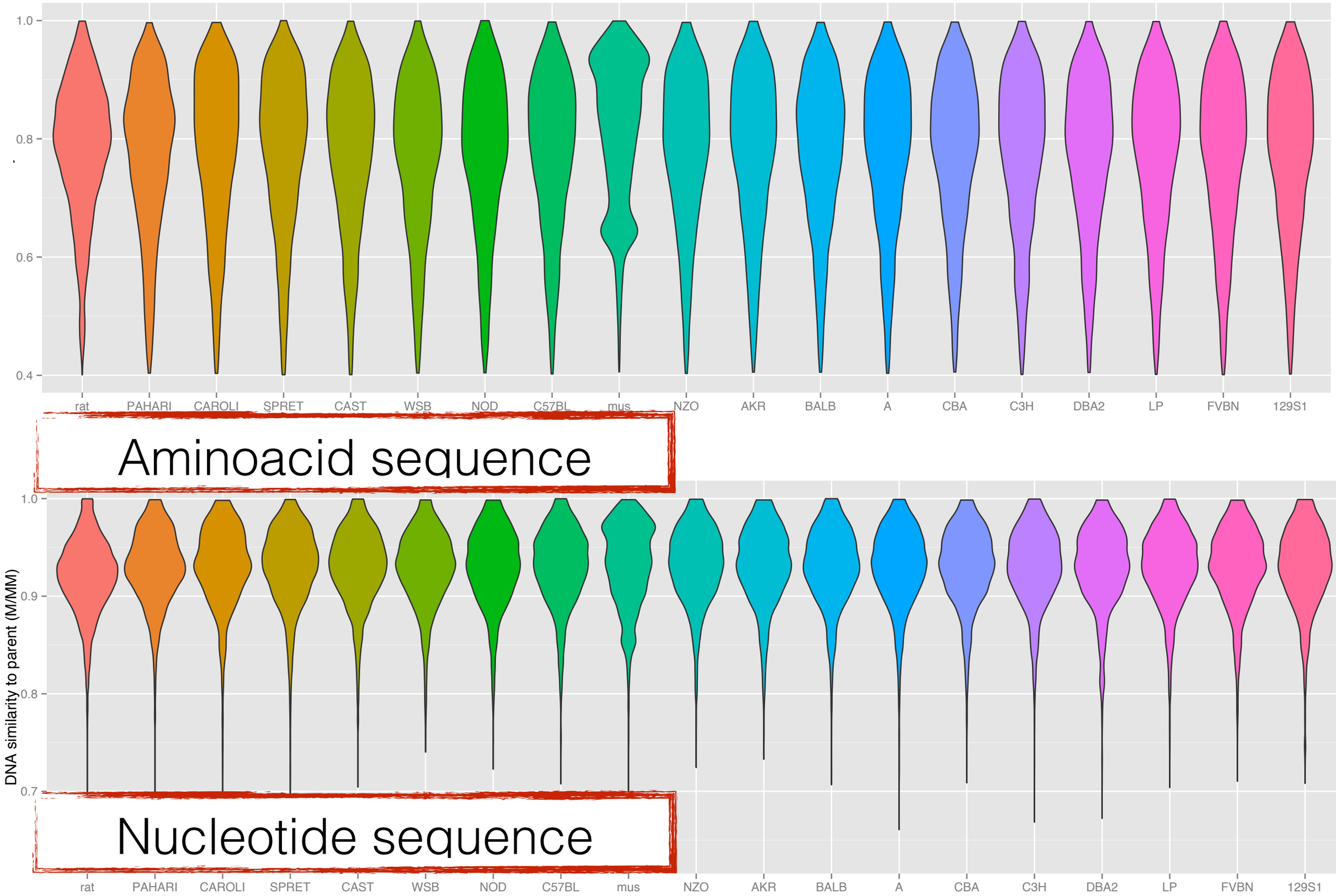


DNA similarity

AA similarity

PSEUDOGENES IN MOUSE STRAINS

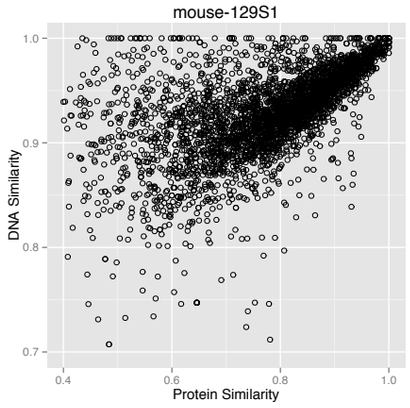
Pseudogene similarity to parents

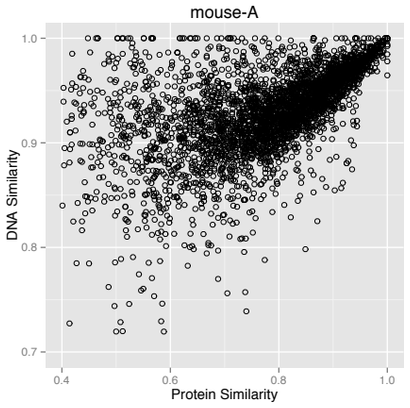


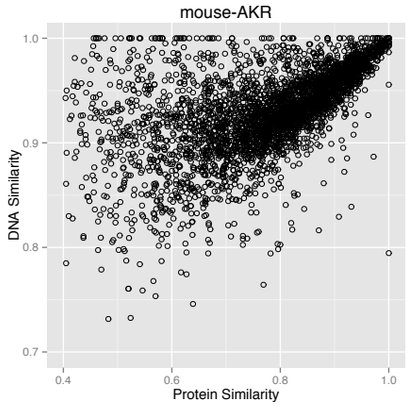
Aminoacid sequence

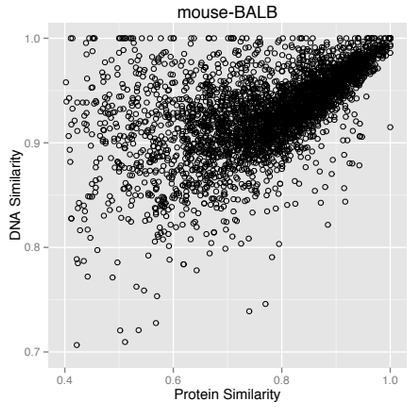
Nucleotide sequence

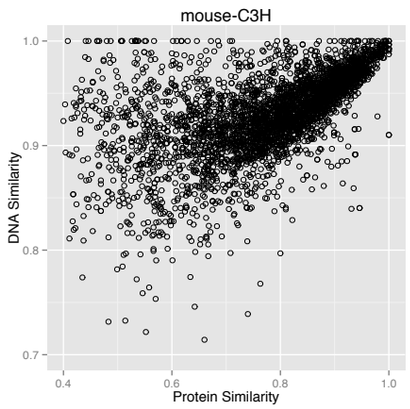
Comparing the protein sequence similarity vs DNA sequence similarity

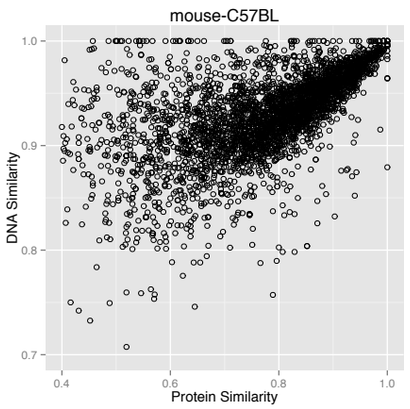


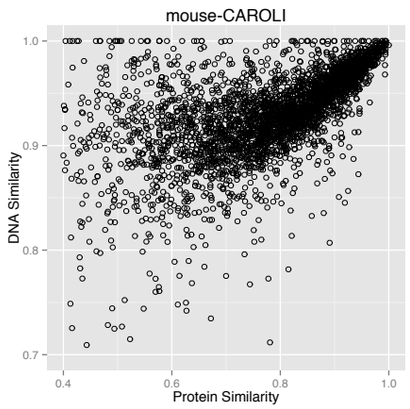


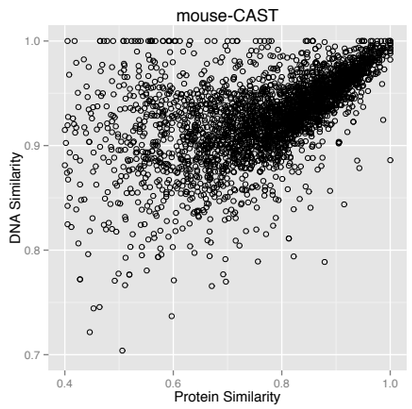


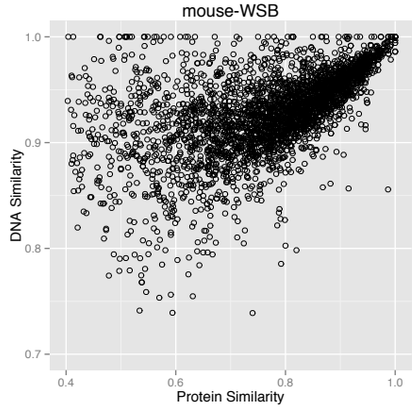


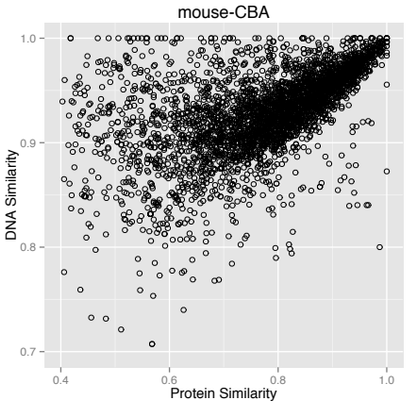


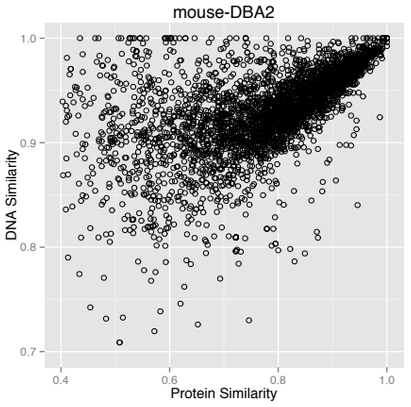


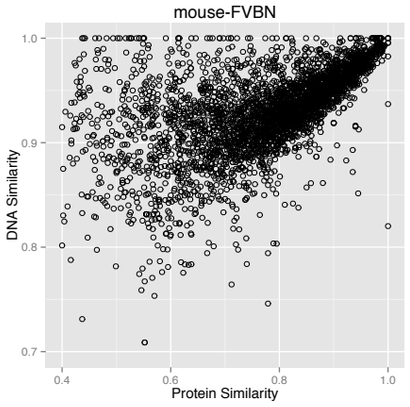


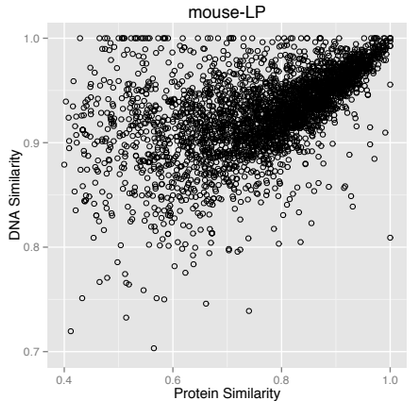


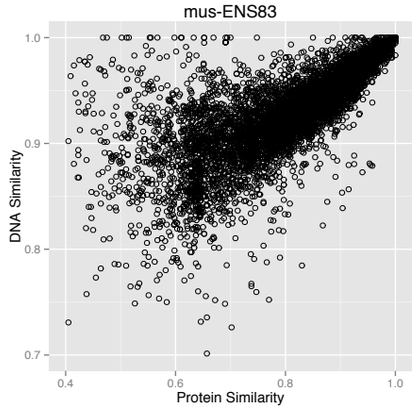


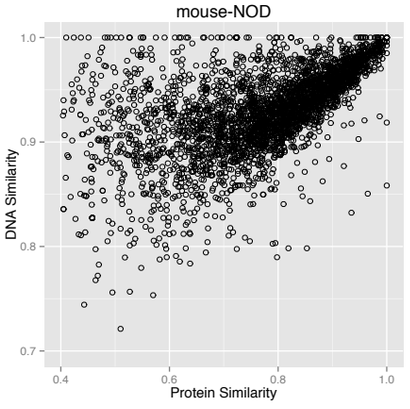


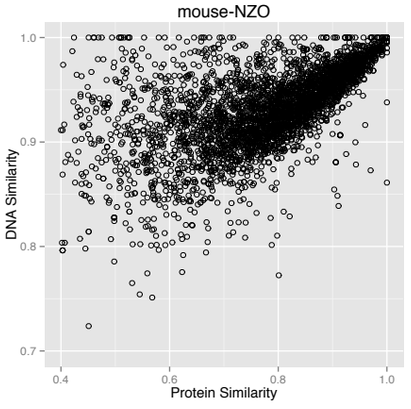


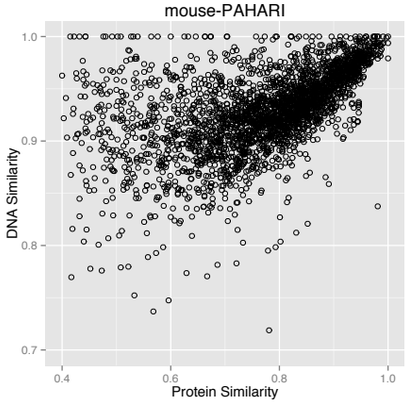


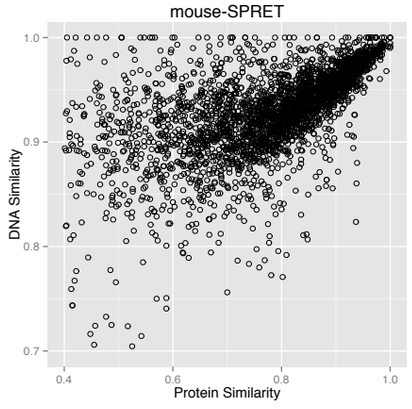


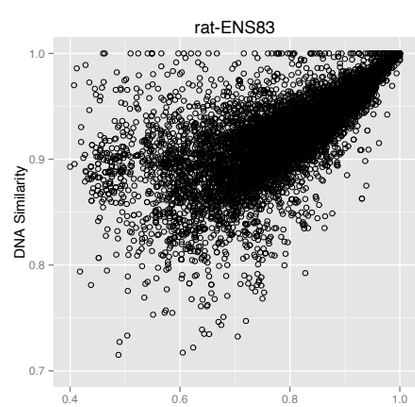




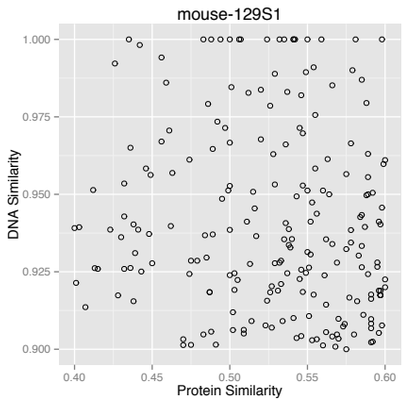


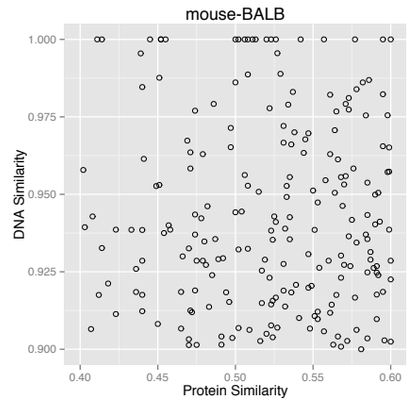


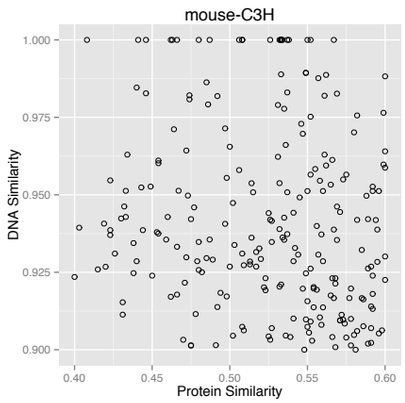


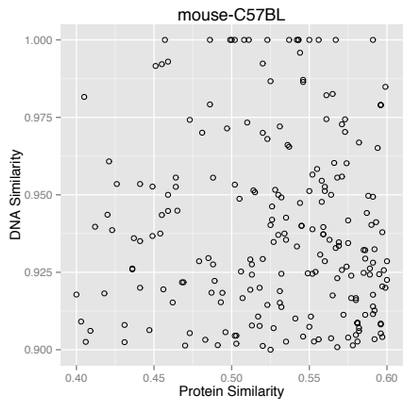


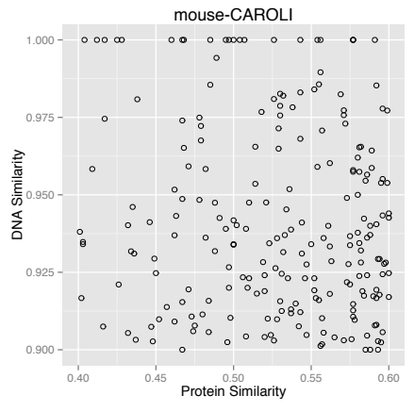
Zoom in at DNA sim >0.9 and protein sim <0.6

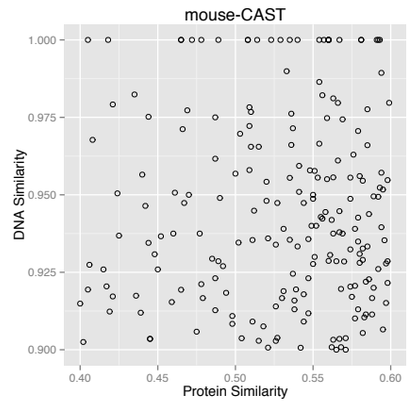


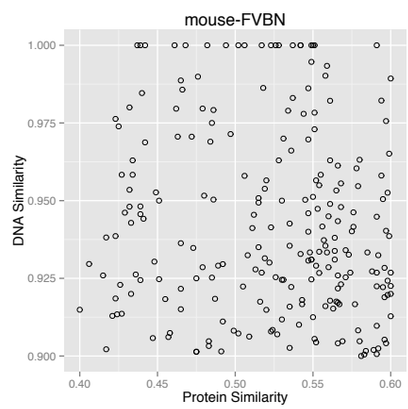


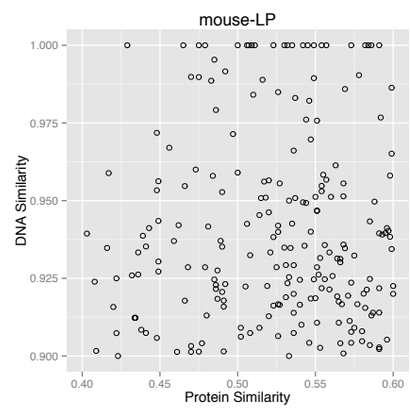


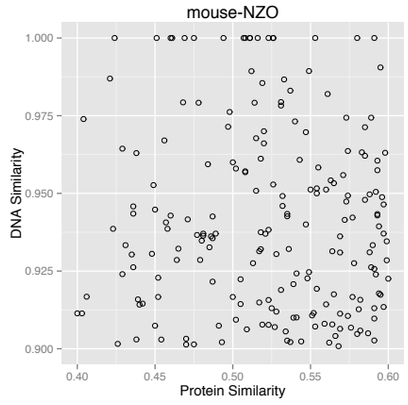


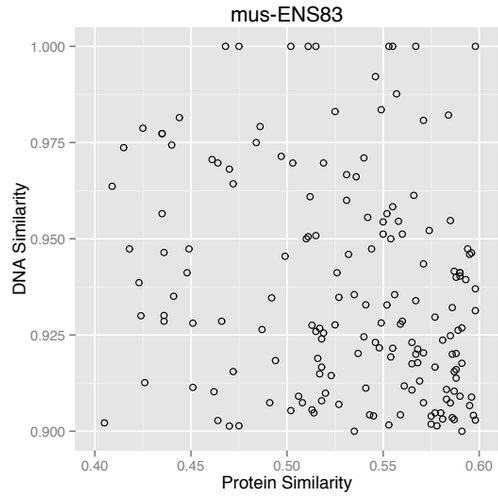


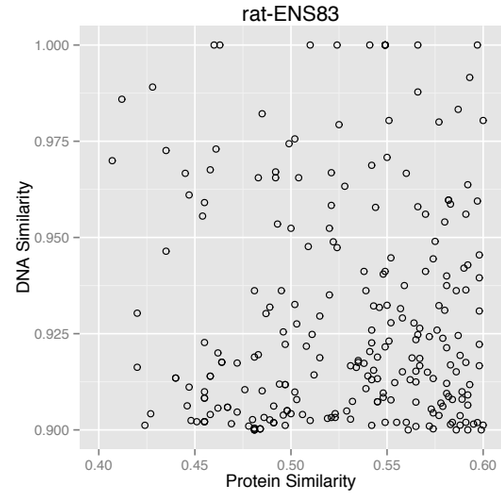












Discrepancies in sequence similarity

- 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

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:

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

3 - it’s a unitary pseudogene / LOF event

WORK IN PROGRESS

Pseudogene Activity

- Evaluate the expression levels of pseudogenes in various tissues and test activity conservation o across strains

- Identify LOF and GOF events using orthologous pseudogene information

Thanks

Mark Gerstein

Fabio Navarro

Ian Fiddes

Thomas Keane