Hi all,

I hope you are all doing well. I'm sending this around to update you all on the progress I've made on the mouse population genomics project. This document is intended to keep you all in the loop as to what I've been up to and is not intended to be a full methods or results write-up. If you have questions, comments or suggestions I would be very glad to hear them.

Attached are a number of figures and a table showing key results that I have obtained. Colour coding is by taxa. All results are for non-CpG-prone sites only.

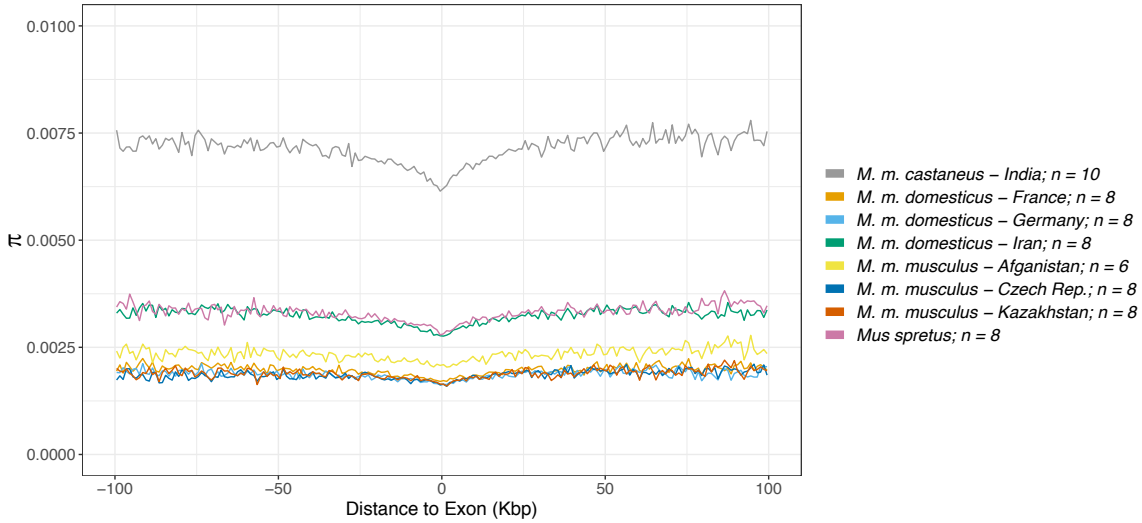
Troughs in diversity around functional elements – Physical distance

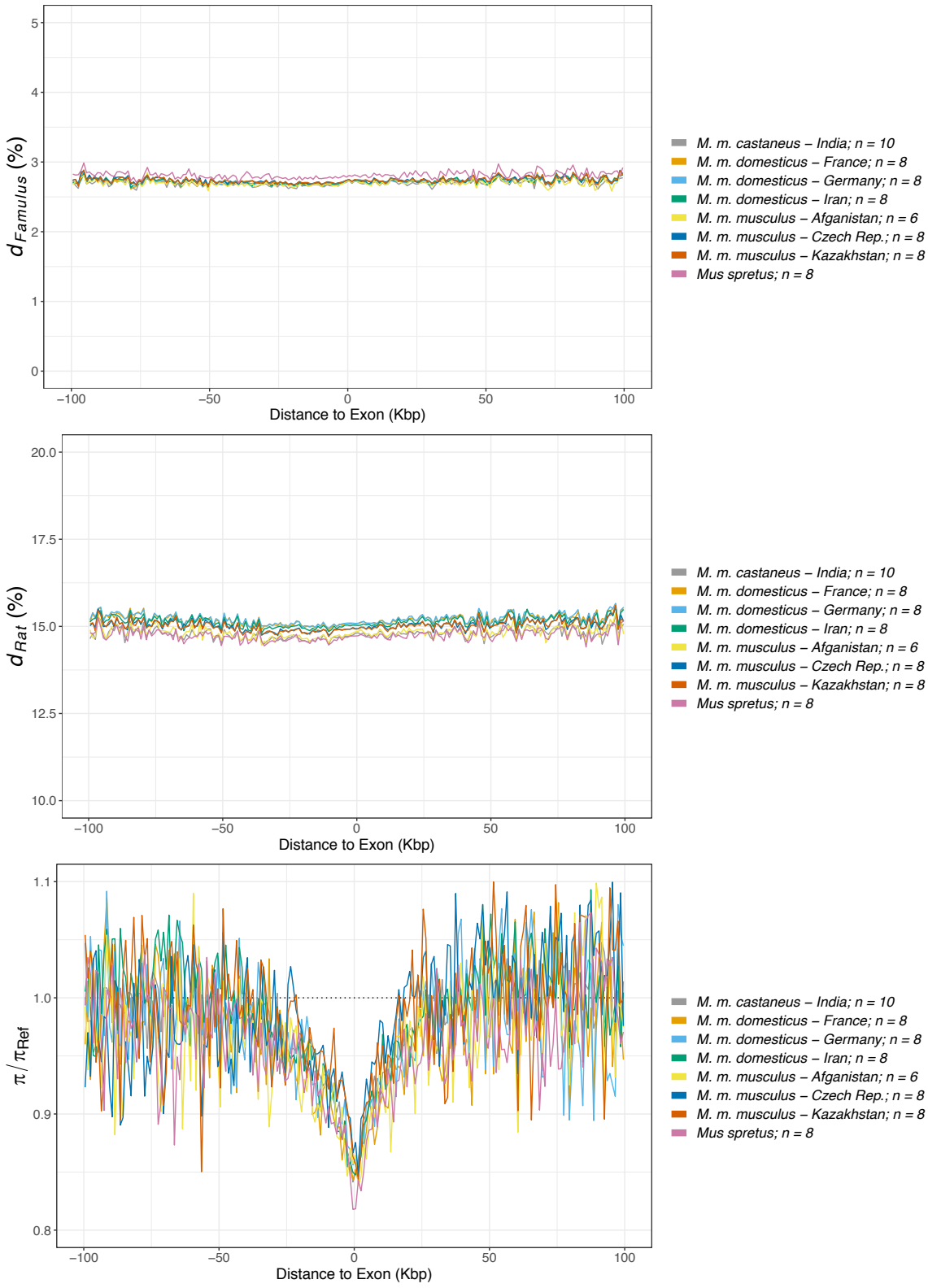
I've plotted diversity, divergence and the ratio of the two around exons and CNEs in bins of 1000bp and 100bp respectively. I've used either *Mus famulus* or *Rattus norvegicus* as outgroups when plotting these data (outgroups are specified in the plots). The first thing that you will notice is that there are dramatic differences in diversity between the taxa. This is not surprising and is well known. Secondly, you'll note that all taxa show a trough in diversity, the magnitudes of which differ according to the level of diversity in a given population. In the case of exons, divergence is relatively invariant around the selected sites, but in the flanks of CNEs, there is a reduction in between-species divergence, similar to what Halligan et al (2013) observed.

From discussions with Peter and Dan over the years, we think that the trough in divergence comes about because the edges of some of the CNEs identified with phastCons (which fits an hidden Markov model to phylogenetic data) may be missed and so sites that are truly under functional constraint may be included in the analysis. However, it's worth noting that the depth of the trough in divergence is not as deep as Dan had found, suggesting that the more up-to-date alignment that Rory used to identify CNEs has reduced the extent of this problem.

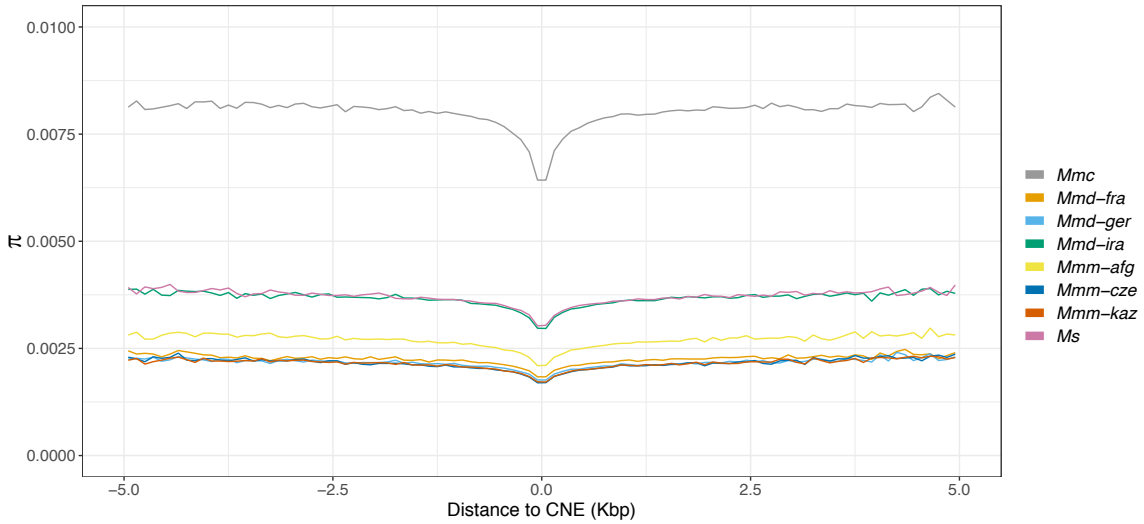
I have plotted the relative reduction in the π/divergence ratio for each of the sub-species. To calculate relative reductions, I divide π/divergence by the mean value from regions far from functional elements. You can see that the relative reductions in diversity are remarkably similar between taxa, consistent to what Eva Deinum found when comparing Rats and *Mus musculus castaneus*.

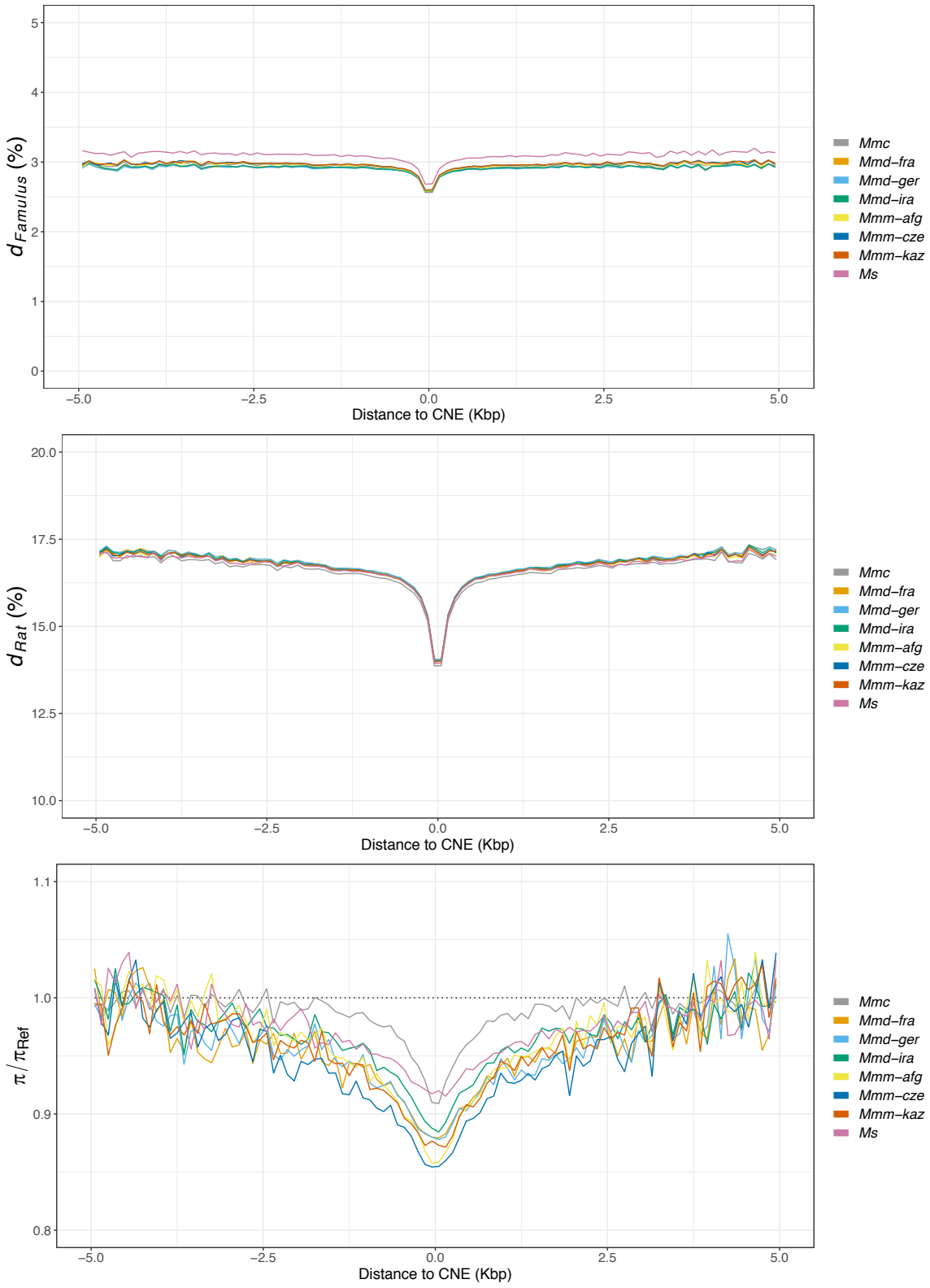
Genetic diversity and divergence surrounding exons (physical distance)




$d_{Famulus}$ (%)
Distance to Exon (Kbp)
d_{Rat} (%)
Distance to Exon (Kbp)
π / π_{Ref}
Distance to Exon (Kbp)
M. m. castaneus – India; n = 10
M. m. domesticus – France; n = 8
M. m. domesticus – Germany; n = 8
M. m. domesticus – Iran; n = 8
M. m. musculus – Afganistan; n = 6
M. m. musculus – Czech Rep.; n = 8
M. m. musculus – Kazakhstan; n = 8
Mus spretus; n = 8

Genetic diversity and divergence surrounding CNEs (physical distance)




$d_{Famulus}$ (%)
5
4
3
2
1
0
−5.0
−2.5
0.0
2.5
5.0
Distance to CNE (Kbp)
Mmc
Mmd−fra
Mmd−ger
Mmd−ira
Mmm−afg
Mmm−cze
Mmm−kaz
Ms
d_{Rat} (%)
20.0
17.5
15.0
12.5
10.0
−5.0
−2.5
0.0
2.5
5.0
Distance to CNE (Kbp)
Mmc
Mmd−fra
Mmd−ger
Mmd−ira
Mmm−afg
Mmm−cze
Mmm−kaz
Ms
π / π_{Ref}
1.1
1.0
0.9
0.8
−5.0
−2.5
0.0
2.5
5.0
Distance to CNE (Kbp)
Mmc
Mmd−fra
Mmd−ger
Mmd−ira
Mmm−afg
Mmm−cze
Mmm−kaz
Ms

Troughs in diversity around functional elements –Genetic distance

I have used the LD-based recombination maps for *M. m. castaneus* to estimate the reductions in diversity on the scale of genetic distance. For each analysis window, I estimate the genetic distance to focal elements on the basis of the genetic map and then bin analysis windows based on those distances. I used an LD-based map, constructed by Ben using the mm10 variant calls for *Mus musculus castaneus*. As we've probably all discussed, this is not an ideal approach because LD-based maps may very well be confounded with the effects of selection at linked sites. However, LD-based maps do potentially capture recombination rate variation at the scale of approximately the inverse of diversity. I am in the process of estimating the genetic distances using the pedigree-based recombination map given in Liu et al (2014 – Genetics). That study used a large genotyping array to estimate recombination rates from a total of ~21,000 crossing-over events, and give recombination rate variation down to finer scales than the pedigree-based map I had used previously (the Cox map).

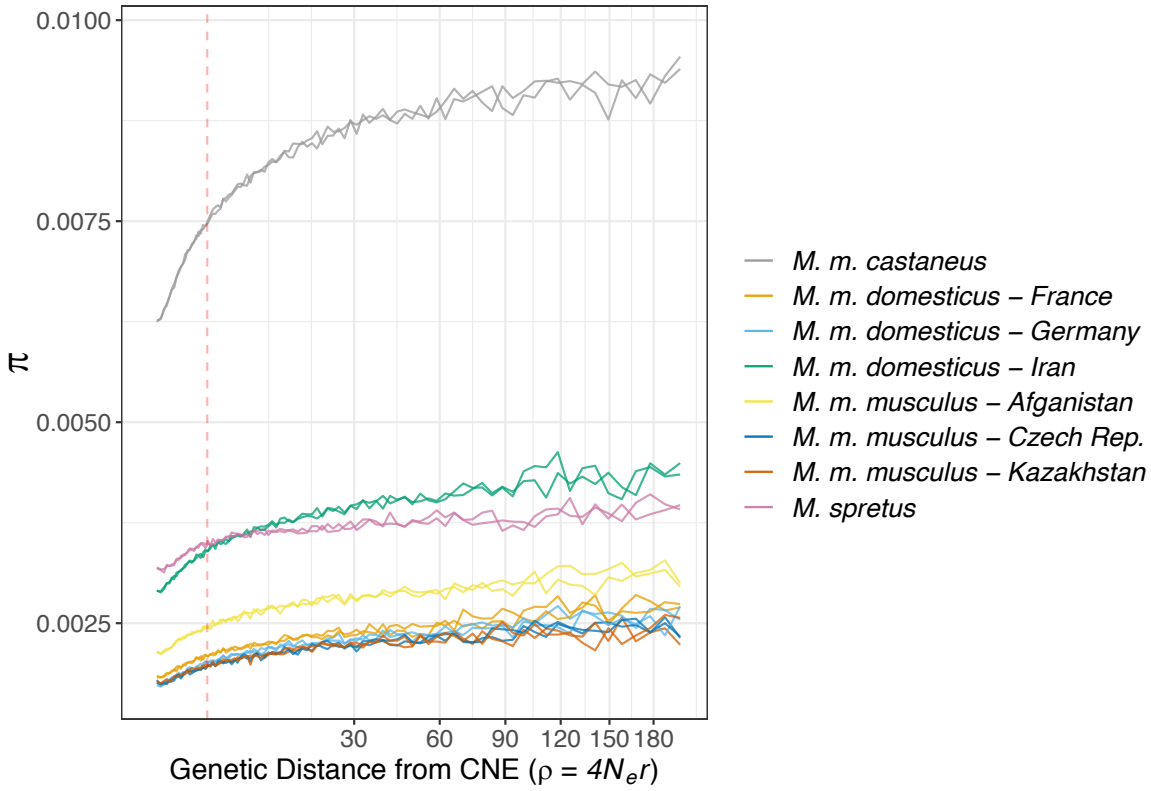
Ben generated recombination rate maps for each of the different mouse taxa that I'm analysing here but found that for all except *M. m. castaneus,* LDhelmet produced maps that were of dubious quality. Obviously, we have no ground truth for the recombination map, but by looking at the correlation of the LD-based maps with pedigree-based maps, he found that only the *castaneus* map was similar. Note that the pedigree-based maps were generated using mouse lines predominantly derived from *Mus musculus domesticus*, so it seems reasonable to assume that the LD-based maps estimated from wild-caught *M. m. domesticus* data should be more similar to the pedigree-based maps than any other mouse taxa. Since they are not, and since genetic diversity and the number of individuals sampled was greatest for *M. m. castaneus*, one interpretation is that the LD-based mapping is underpowered for the other taxa. Based on that argument, I decided to calculate genetic distances using only the *castaneus* map. Note that the results I present here use the raw $4N_e r$ values from LDhelmet. An alternative approach would be to scale estimated distances by the ratio of nucleotide diversity in *castaneus* with that of the focal taxa, to reflect differences in *Ne* between the mouse groups.

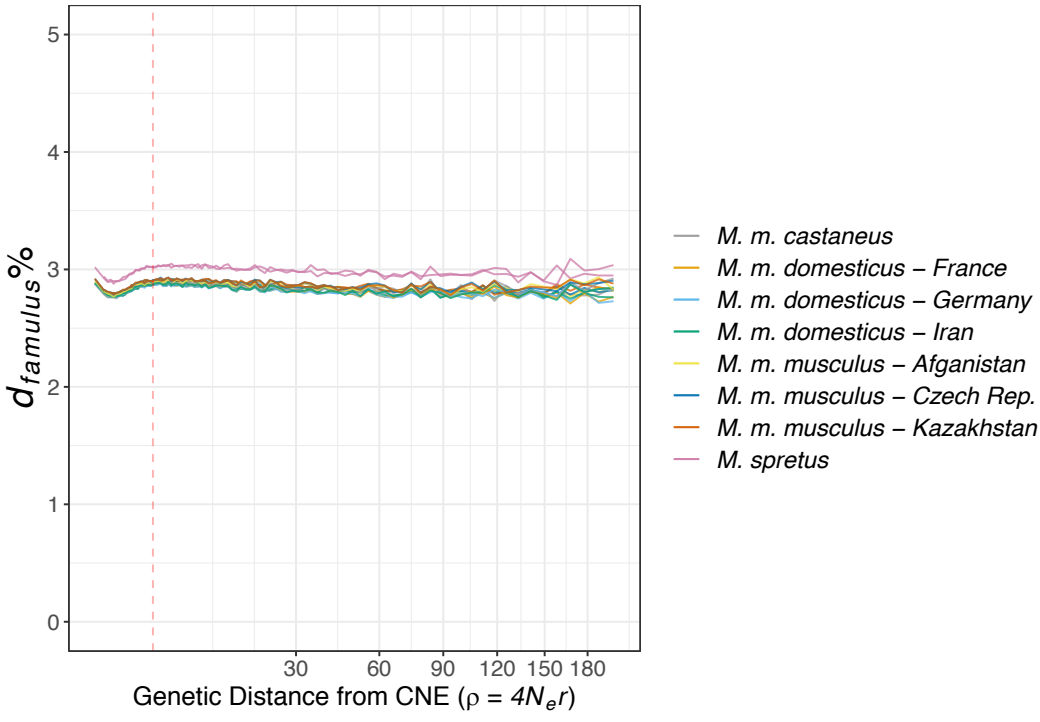
As with physical distances, there are troughs in genetic diversity around both exons and CNEs on the scale of genetic distance. As above, this is not at all surprising, but it is worth mentioning. In the regions surrounding exons, genetic divergence to both Rat and *M. famulus* are relatively invariant. In the case of CNEs, the trough in d_{rat} around the elements is apparent, but unlike for physical distance is it clear to see that the reduction in diversity extends beyond the trough in divergence.

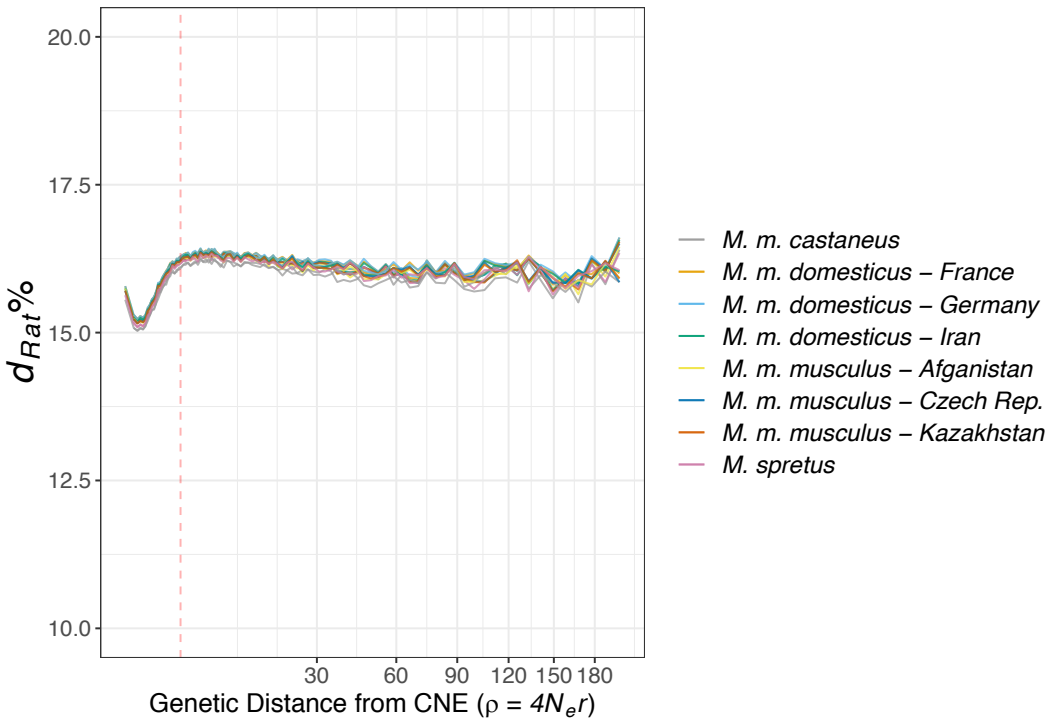
There is a striking similarity for the shape of the troughs in relative π around both exons and CNEs. This is expected if all *Mus musculus* taxa have similar DFEs and recombination rate variation, and if selected mutations typically have selective effects $s > 1/Ne$. *Mus spretus* exhibits a shallower reduction in relative diversity than *M. musculus* taxa, and it will be interesting to see how estimates of positive selection compare between the groups. I think that this is a nice finding, and that a version of these figures will probably be central to the paper. The next step is to ask whether background selection makes a similar contribution to these troughs for each of the taxa analysed (see below).

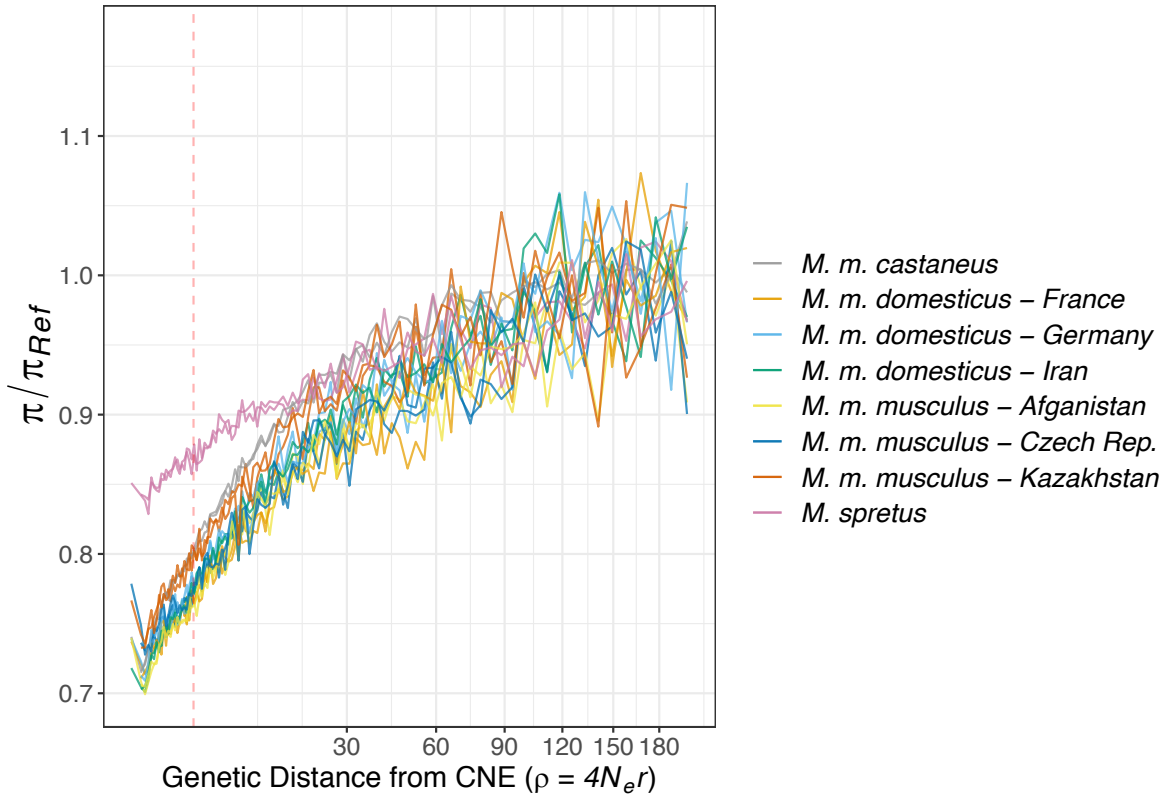
As with physical distance, the troughs in diversity on the scale of genetic distance are wider and deeper for exons than they are for CNEs. This is consistent with stronger selective effects on mutations occurring in exons, but I will reign in my speculation as I will be fitting sweep models shortly.

Genetic diversity and divergence surrounding CNEs (genetic distance)



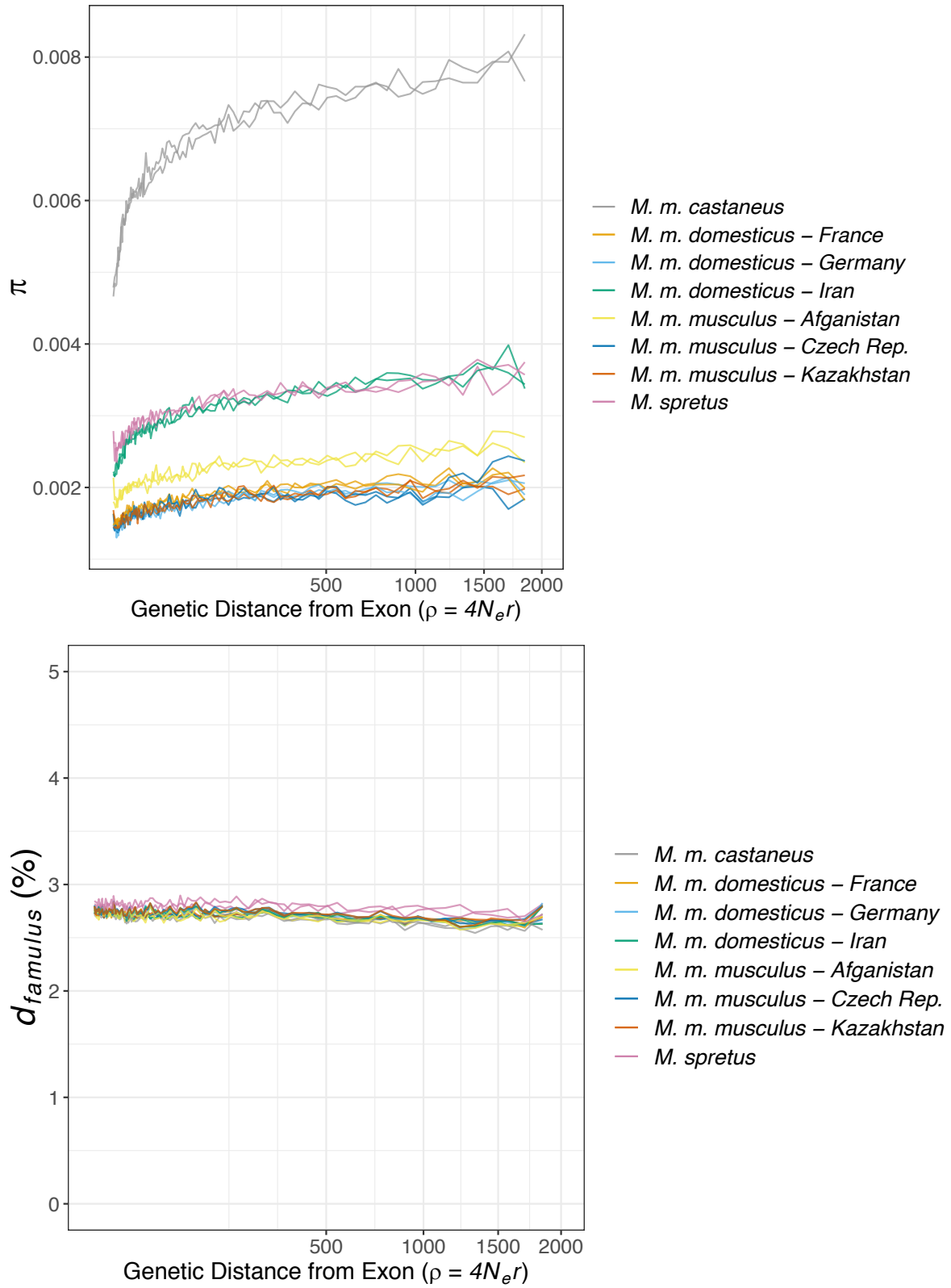


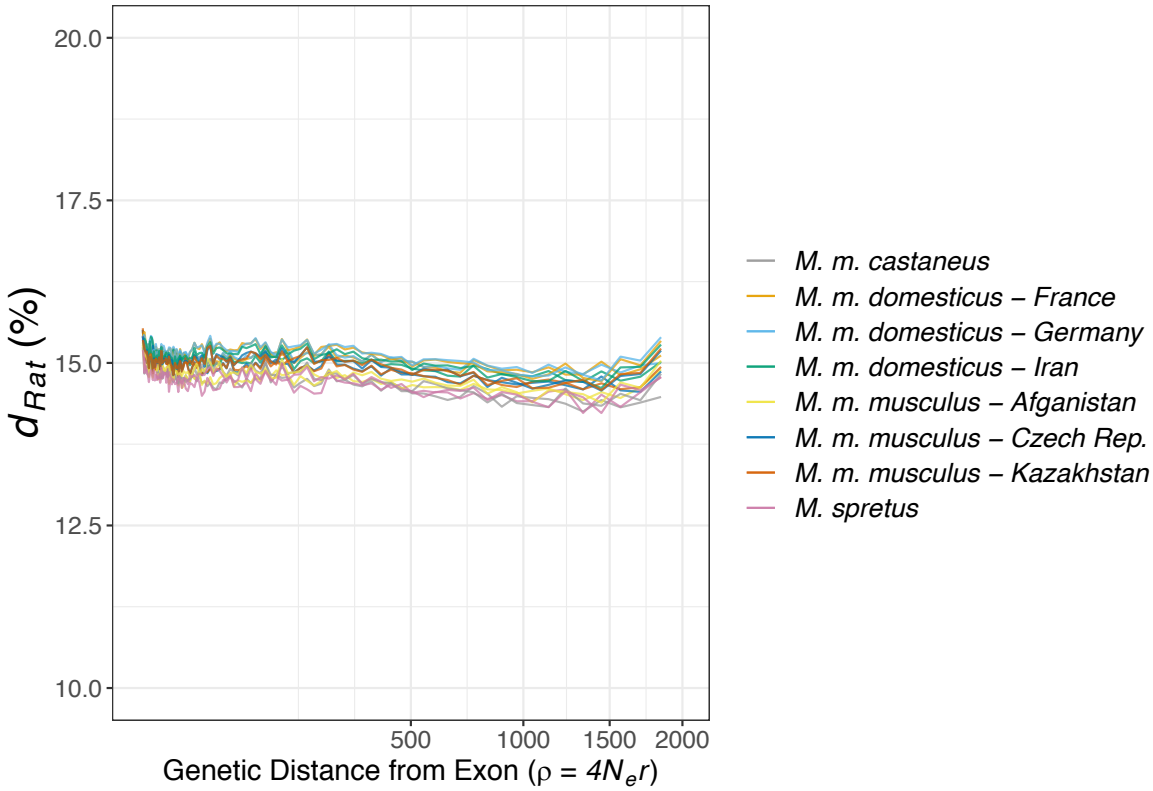


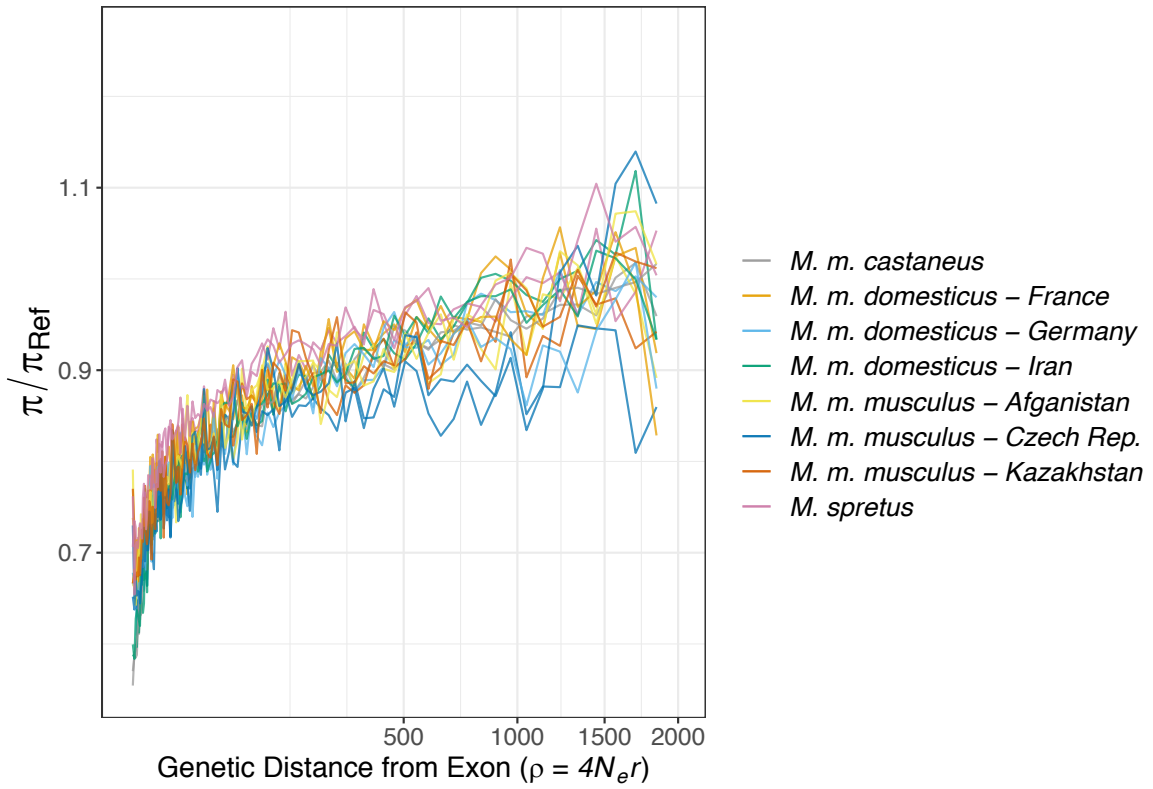


In each figure the mean diversity for non-CpG sites in bins of 100bp is plotted in regions surrounding CNEs. The dashed vertical line marks the edge of the trough in divergence observed around CNEs.

Genetic diversity and divergence surrounding protein-coding exons (genetic distance)







In the above figures, for the sake of plotting, genetic diversity up and down stream are plotted on a positive axis.

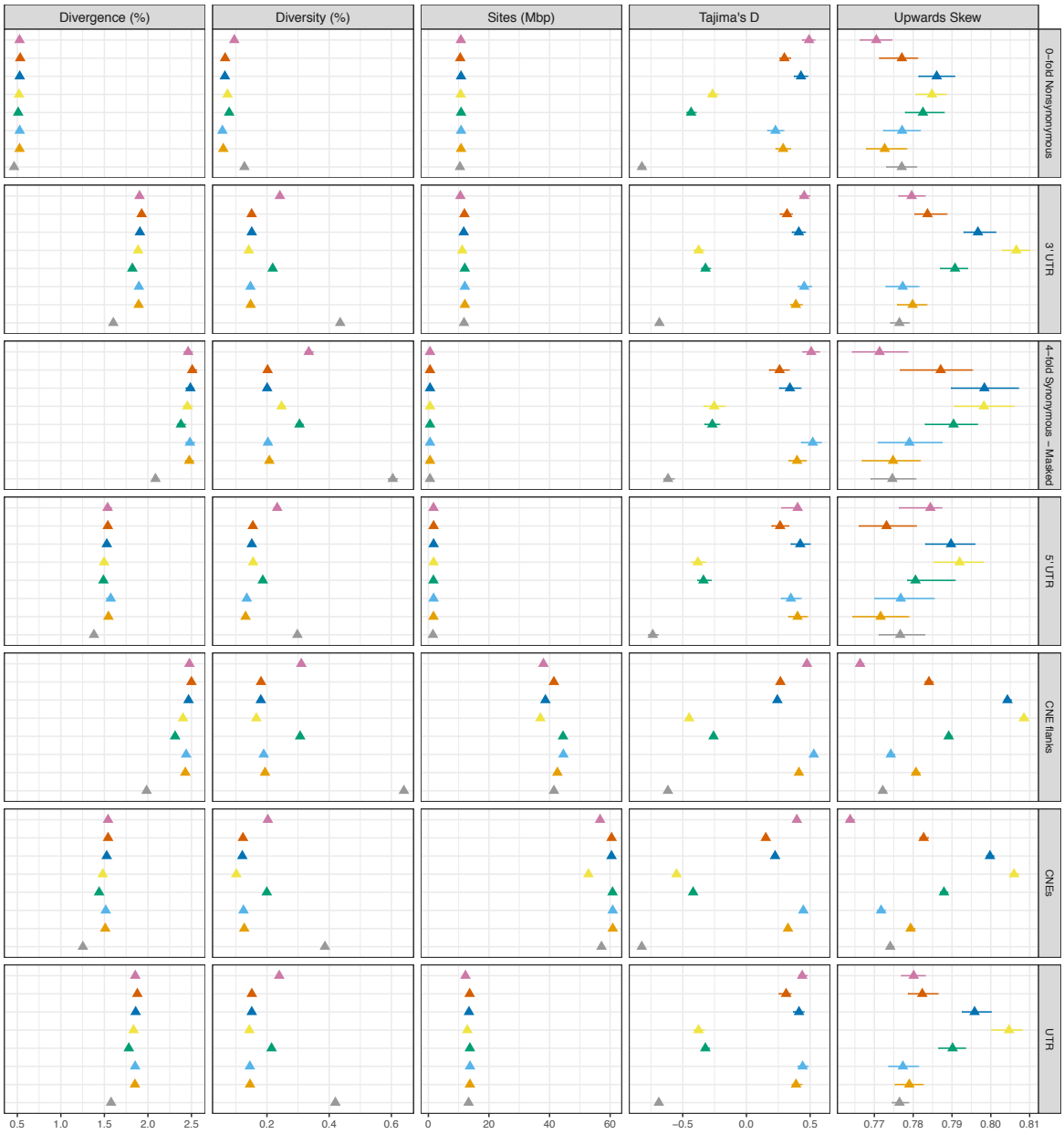
Summary statistics for different elements in the mouse genome (Figures 5 and 6)

Figure 5 shows summary statistics for several classes of sites in the genome: 4-fold synonymous (masked and unmasked – see below) and 0-fold nonsynonymous sites, UTRs and the set of CNEs that Rory identified along with sites flanking those elements. The bars on the plots represent the 95% interval of 100 bootstrap replicates.

The patterns that can be seen in these summary are consistent with purifying selection acting in each of the 'selected' sites, as putatively neutral sites (4-fold and CNE-flanks) exhibit higher diversity and greater values of Tajima's D. There are clear differences in both Tajima's D and diversity for the different taxa, indicating that there have been processes that have shaped nucleotide variability above and beyond selection at linked sites. In particular, it seems that population size change is a plausible explanation for the differences observed.

In the attached figures, you will note that there are two classes of 4-fold synonymous sites, masked and unmasked. Savisaar and Hurst (2018 - Genome Research) showed that synonymous sites within exonic splice enhancers (ESEs) exhibit functional constraint and that when inferring the DFE for deleterious mutations, including 4-fold sites within ESEs can result in downwardly biased estimates of the strength of selection. Motivated by that finding, I identified a set of putative ESEs for coding regions in the mouse genome. Rosina Savisaar kindly sent me a list of ESE motifs and I used these to query coding-sequences in the mouse genome. About two-thirds of non-CpG-prone 4-fold synonymous sites were within putative ESEs. I refer to the dataset without these sites as the Masked data.

The figure below shows summary statistics for each of the categories of sites analysed. The horizontal bars represent the 95% range from 100 bootstrap replicates. 'Upward skew' is a measure of how much the uSFS is weighted towards high frequency variants. 'Divergence' is the taxon specific divergence estimated using *est-sfs*.



M. spretus – n = 8
M. m. musculus – Kazakhstan – n = 8
M. m. musculus – Czech Rep. – n = 8
M. m. musculus – Afghanistan – n = 6
M. m. domesticus – Iran – n = 8
M. m. domesticus – Germany – n = 8
M. m. domesticus – France – n = 8
M. m. castaneus – n = 10

Estimating the DFE and the proportion of fixed differences between species attributable to positive selection (Figure 7)

For several classes of sites in the mouse genome (0-fold nonsynonymous, 4-fold synonymous , UTRs, CNEs and CNE-flanking regions), I have inferred the uSFS with Peter's est-sfs software (using *Mus famulus* and *Mus pahari* as outgroups with the Kimura 2-parameter mutation model).

For each of the selected classes of sites (0-fold sites, UTRs and CNEs), I have inferred the uSFS and used it to estimate the DFE using the program *polyDFE* (Tataru et al 2016). *polyDFE* is conceptually similar to DFE-alpha, but doesn't require the fitting of a model of demographic history so runs fairly quickly. p*olyDFE* can model the full DFE (i.e. a distribution containing both harmful and beneficial mutations) or can be constrained to just the deleterious portion and in addition, you can choose whether or not to model divergence. I chose to estimate the DFE in the absence of divergence data. For UTRs and 0-fold sites, I used 4-fold sites as the neutral reference class (see below). For CNEs, I used CNE-flanking sites. For each class of sites, I modelled the deleterious portion of the DFE as a gamma distribution, allowing the program to fit a discrete class of advantageous mutations. Tataru et al showed in their study that doing this results in more accurate inference of the deleterious portion of the DFE.

For each of the site classes analysed, DFE estimates broadly agree with previously published estimates. For 0-fold sites, the majority of mutations are strongly disadvantageous. For UTRs and CNEs, there is a larger contribution of mildly deleterious mutations.

When you run the *polyDFE* program, it takes divergence data from the uSFS and calculates ALPHA, whilst accounting for the contribution that deleterious mutations may make to divergence. As I have used *Mus famulus* and *Mus pahari* to infer the uSFS, under Peter's est-sfs framework, the inferred divergence in the uSFS represents alleles which have become fixed after the split of *Mus famulus* and *Mus XXX*, where *XXX* is the particular taxa in question. If you compare the ALPHA estimates for *M. m. castaneus* that I'm showing here with those estimated for *M. m. castaneus* by Halligan et al (2013) you'll notice that they're quite different. There are several reasons for this. Firstly, this is because Dan used divergence from Rat rather than *castaneus*-specific divergence and secondly, masking ESEs results in a higher amount of synonymous site divergence but doesn't affect synonymous site diversity by the same amount, so will affect estimates of ALPHA for both 0-fold sites and UTRs.

The table on the following page contains estimates of parameters of the DFE for harmful mutations obtained using polyDFE. Each parameter estimate is presented with the upper and lower bounds of a 95% interval obtained from 100 bootstrap replicates.
The shape (β) mean of the gamma distribution in units of $2N_es$ are given.

		β			Mean $2N_es$			α		
			lower	upper		lower	upper		lower	upper
0-fold	Mmc	0.186	0.156	0.978	-50044.6	-90194.6	-699.711	0.544	0.378	0.991
	Mmd_fra	0.202	0.181	7.492	-724.3	-4174.6	-257.365	-0.009	-0.014	1.000
	Mmd_ger	0.246	0.213	8.162	-233.5	-6352.0	-127.492	0.002	-0.004	1.000
	Mmd_ira	0.133	0.126	0.164	-26929.2	-49820.7	-5764.614	0.013	-0.001	0.099
	Mmm_afg	0.190	0.166	0.215	-1153.7	-2984.7	-519.955	-0.006	-0.015	0.006
	Mmm_cze	0.237	0.204	8.168	-263.5	-6464.8	-115.086	-0.021	-0.026	1.000
	Mmm_kaz	0.224	0.197	0.255	-361.6	-830.5	-184.119	-0.017	-0.025	-0.008
	Ms	0.155	0.130	0.174	-5146.4	-25100.8	-1862.848	-0.015	-0.025	-0.002
CNEs	Mmc	0.111	0.104	0.132	-93.0	-110.5	-67.901	0.209	0.190	0.279
	Mmd_fra	0.061	0.042	0.080	-277.4	-10597.2	-93.753	0.005	0.002	0.064
	Mmd_ger	0.065	0.058	0.071	-202.3	-436.7	-123.641	0.017	0.002	0.030
	Mmd_ira	0.065	0.061	0.109	-200.1	-304.6	-38.605	0.013	0.008	0.100
	Mmm_afg	0.088	0.047	0.098	-158.7	-14280.9	-107.595	0.121	0.095	0.139
	Mmm_cze	0.064	0.060	0.069	-200.4	-320.0	-130.916	0.003	0.000	0.009
	Mmm_kaz	0.091	0.086	0.096	-33.3	-41.8	-26.571	0.010	0.000	0.013
	Ms	0.054	0.036	0.065	-710.3	-50000.0	-453.570	0.007	0.002	0.035
UTRs	Mmc	0.035	0.033	0.068	-15205.9	-16220.5	-493.635	0.163	0.131	0.299
	Mmd_fra	0.035	0.029	0.045	-14712.2	-42962.2	-4821.992	0.169	0.083	0.357
	Mmd_ger	0.060	0.026	0.111	-85.1	-16899.0	-13.751	0.104	0.029	0.204
	Mmd_ira	0.033	0.029	0.038	-15310.4	-15660.3	-5473.827	0.134	0.088	0.174
	Mmm_afg	0.046	0.041	0.105	-78909.6	-79209.9	-203.927	0.312	0.248	0.373
	Mmm_cze	0.027	0.023	0.040	-34112.6	-47024.0	-3901.827	0.113	0.009	0.333
	Mmm_kaz	0.024	0.021	0.032	-78731.9	-83950.7	-39031.738	0.069	0.014	0.126
	Ms	0.031	0.027	0.037	-15263.6	-20060.2	-9297.723	0.120	0.072	0.241

Things still to do (hopefully they won't take as long as the above steps!)

There are two things that I have left to do which, when finished, will completely replicate the earlier version of this study (i.e. the third chapter of my thesis). First, I need to run simulations to obtain the effects of background selection for each of the taxa analysed. To do this, I will use the DFE estimates I obtained for each of the site classes. Second, with the simulated BGS effects, I'll fit a model of hard sweeps to the troughs in diversity and estimate selection parameters compatible with the observations.