Checking potential artifacts

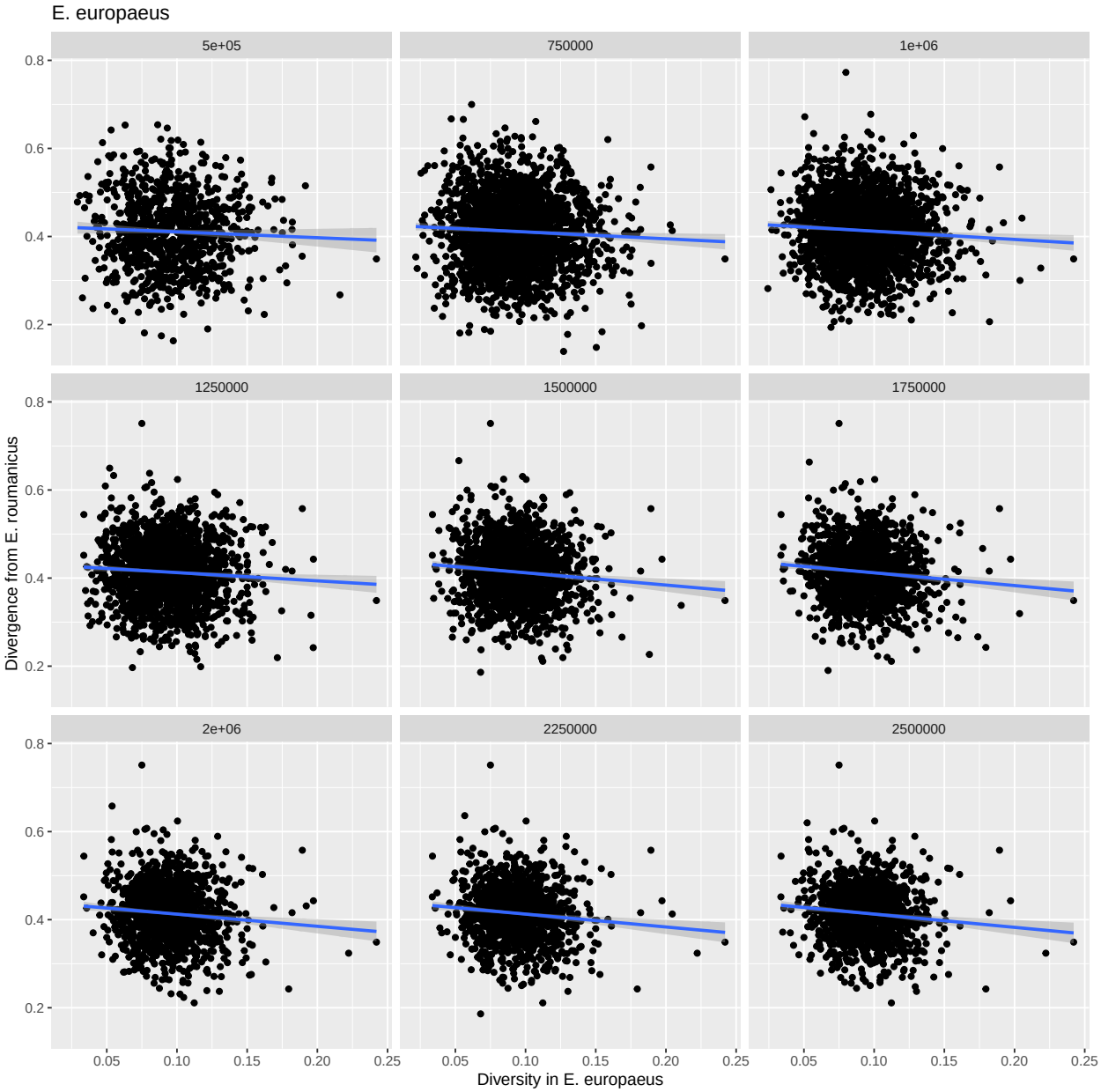
J. Ignacio Lucas Lledó

9/4/2020

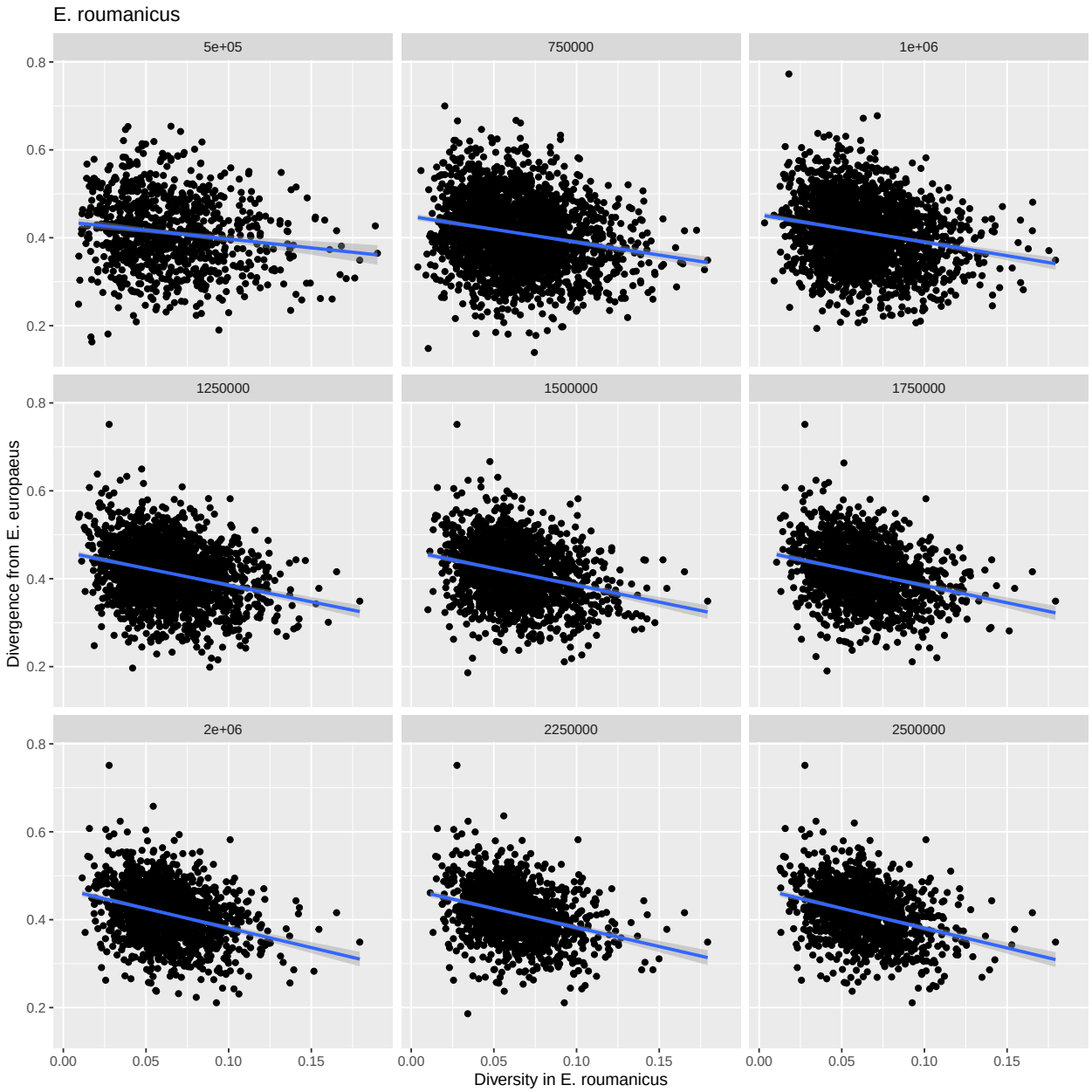
Effect of window size on the correlation between diversity and divergence

In `2020-04-08/windowSize` I run the scripts to obtain estimates of population genetics parameters using a coordinate system for window definition, and using different window sizes. Now, I want to plot the results.

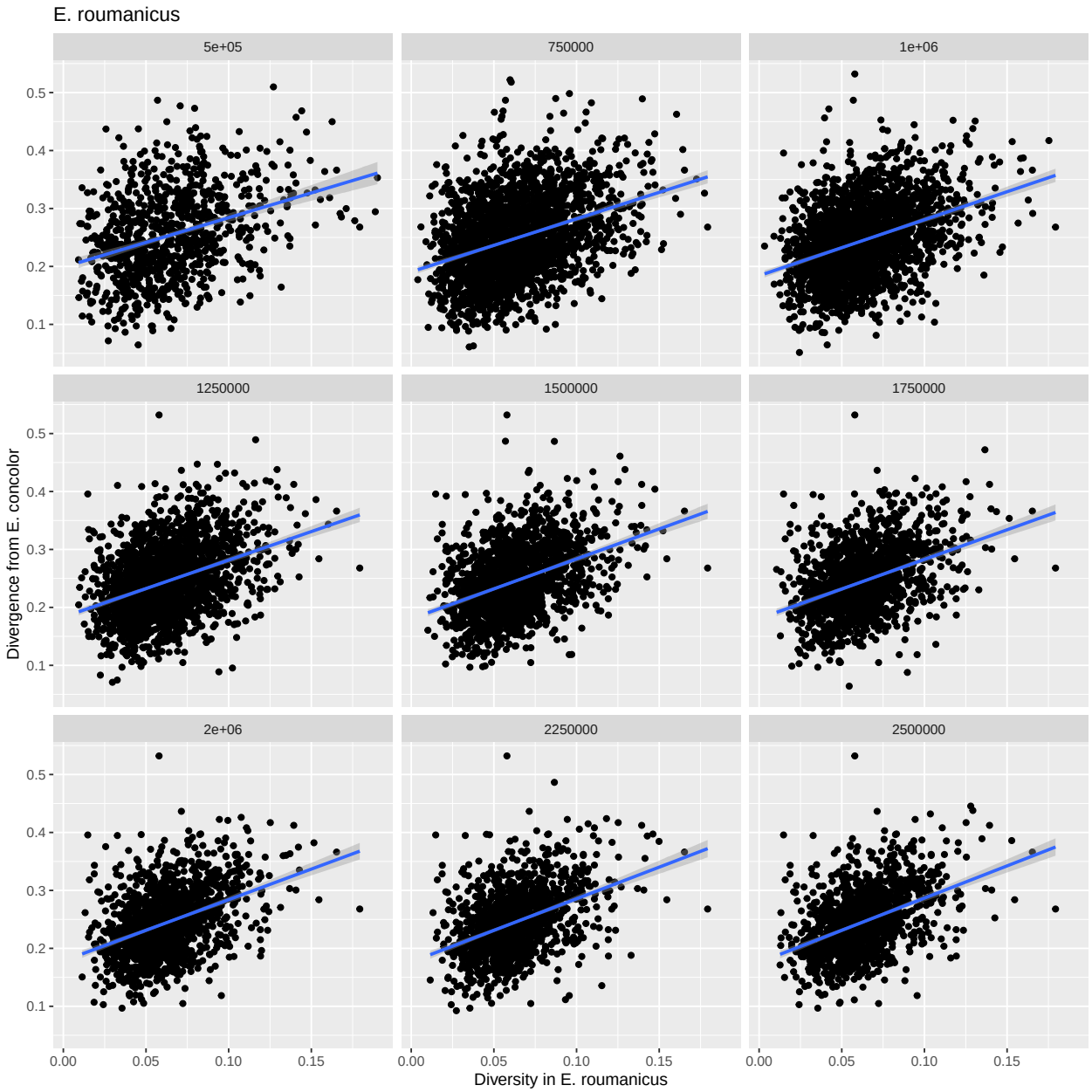
```
popGenStats <- data.frame()
for (W in c(500000, 750000, 1000000, 1250000, 1500000, 1750000, 2000000, 2250000, 2500000)) {
   z <- read.table(sprintf("windowSize/%i.popGenStats.csv", W),
                   header = TRUE, sep = ",",
                   col.names = c('scaffold', 'start', 'end', 'mid', 'sites', 'pi_con',
                                 'pi_rou', 'pi_eur', 'dxy_con_rou', 'dxy_con_eur',
                                 'dxy_rou_eur', 'Fst_con_rou', 'Fst_con_eur', 'Fst_rou_eur'),
                   colClasses = c('NULL', 'NULL', 'NULL', 'NULL', 'integer', 'numeric',
                                  'numeric', 'numeric', 'numeric', 'numeric',
                                  'numeric', 'numeric', 'numeric', 'numeric'))
   z$W <- factor(W, levels = c(500000, 750000, 1000000, 1250000, 1500000, 1750000, 2000000,
                              2250000, 2500000))
   popGenStats <- rbind(popGenStats, z)
}
rm(z)

ggplot(popGenStats, aes(x = pi_eur, y = dxy_rou_eur)) + geom_point() +
  geom_smooth(method='lm') + facet_wrap(~W) + xlab('Diversity in E. europaeus') +
  ylab('Divergence from E. roumanicus') + ggtitle('E. europaeus')
```



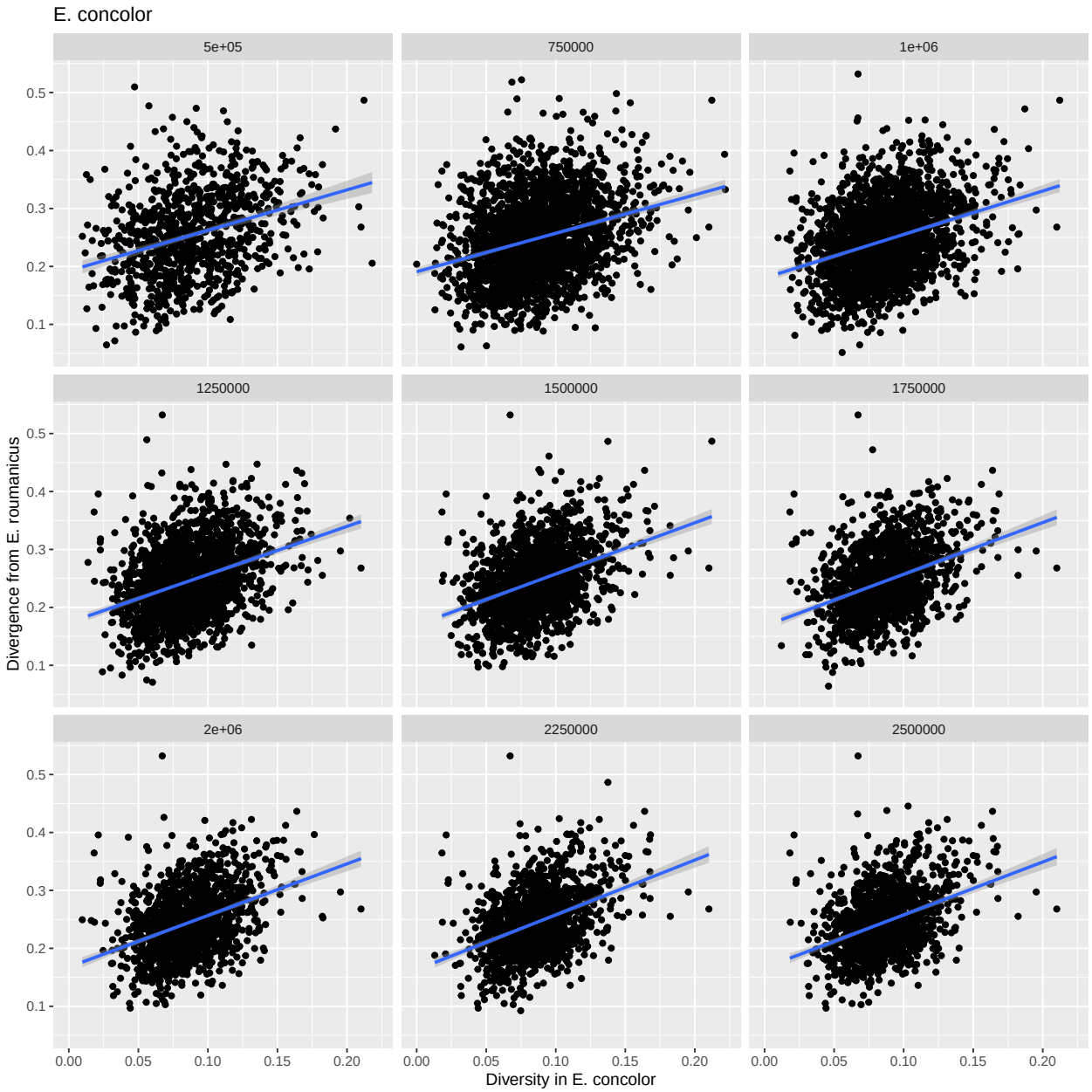

```
ggplot(popGenStats, aes(x = pi_rou, y = dxy_rou_eur)) + geom_point() +
  geom_smooth(method='lm') + facet_wrap(~W) + xlab('Diversity in E. roumanicus') +
  ylab('Divergence from E. europaeus') + ggtitle('E. roumanicus')
```



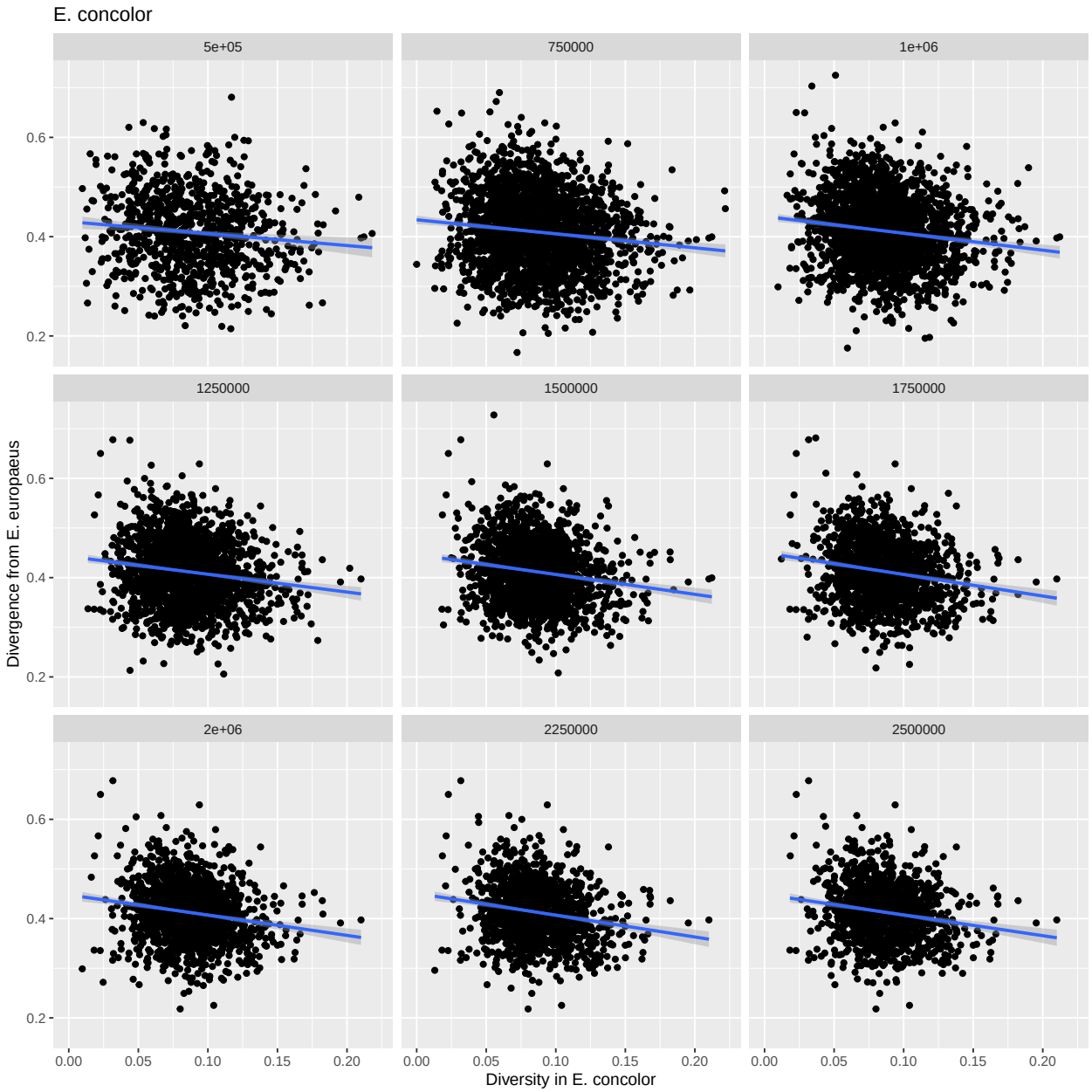

```
ggplot(popGenStats, aes(x = pi_rou, y = dxy_con_rou)) + geom_point() +
  geom_smooth(method='lm') + facet_wrap(~W) + xlab('Diversity in E. roumanicus') +
  ylab('Divergence from E. concolor') + ggtitle('E. roumanicus')
```



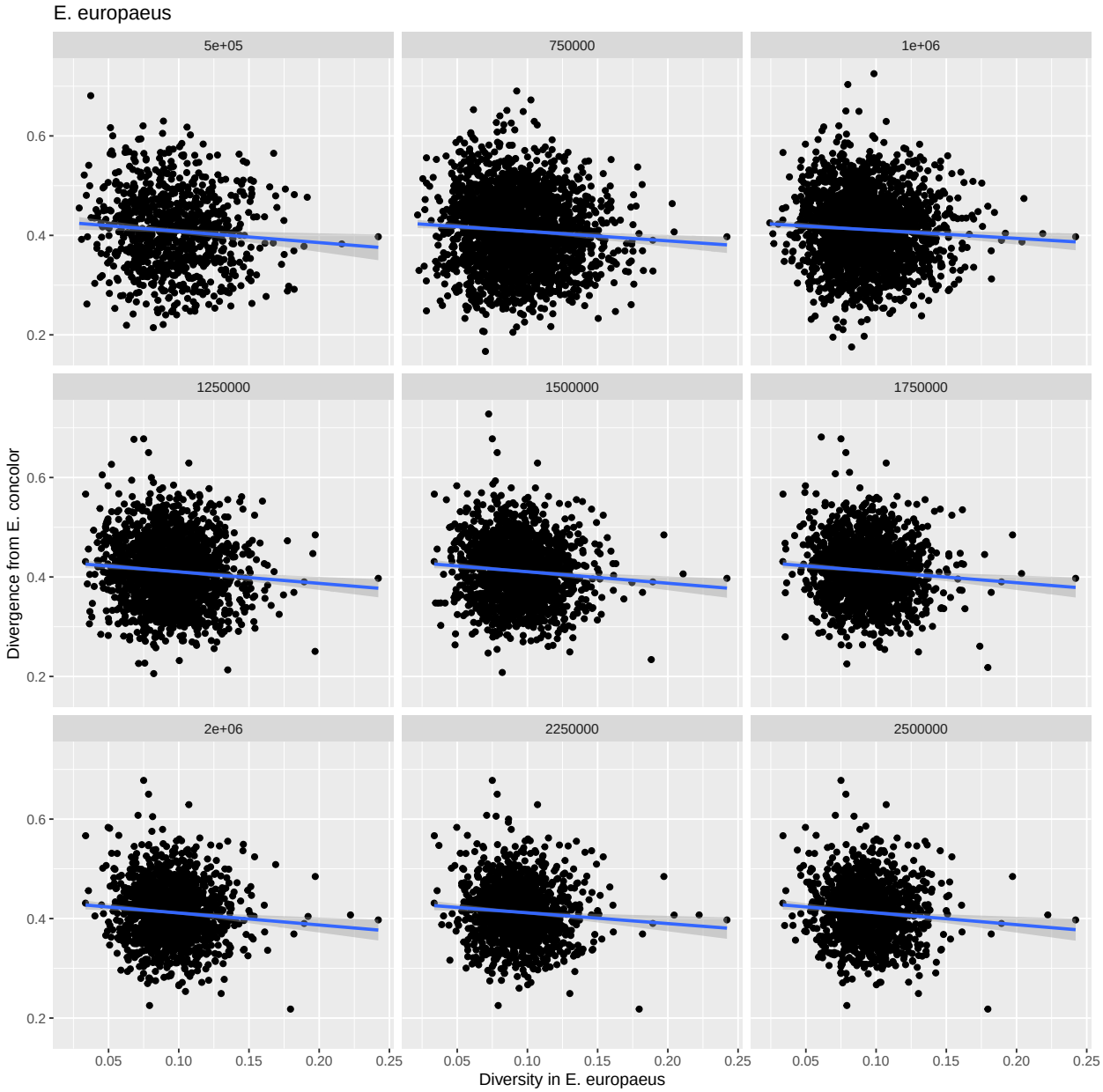

```
ggplot(popGenStats, aes(x = pi_con, y = dxy_con_rou)) + geom_point() +
  geom_smooth(method='lm') + facet_wrap(~W) + xlab('Diversity in E. concolor') +
  ylab('Divergence from E. roumanicus') + ggtitle('E. concolor')
```




```
ggplot(popGenStats, aes(x = pi_con, y = dxy_con_eur)) + geom_point() +
  geom_smooth(method='lm') + facet_wrap(~W) + xlab('Diversity in E. concolor') +
  ylab('Divergence from E. europaeus') + ggtitle('E. concolor')
```

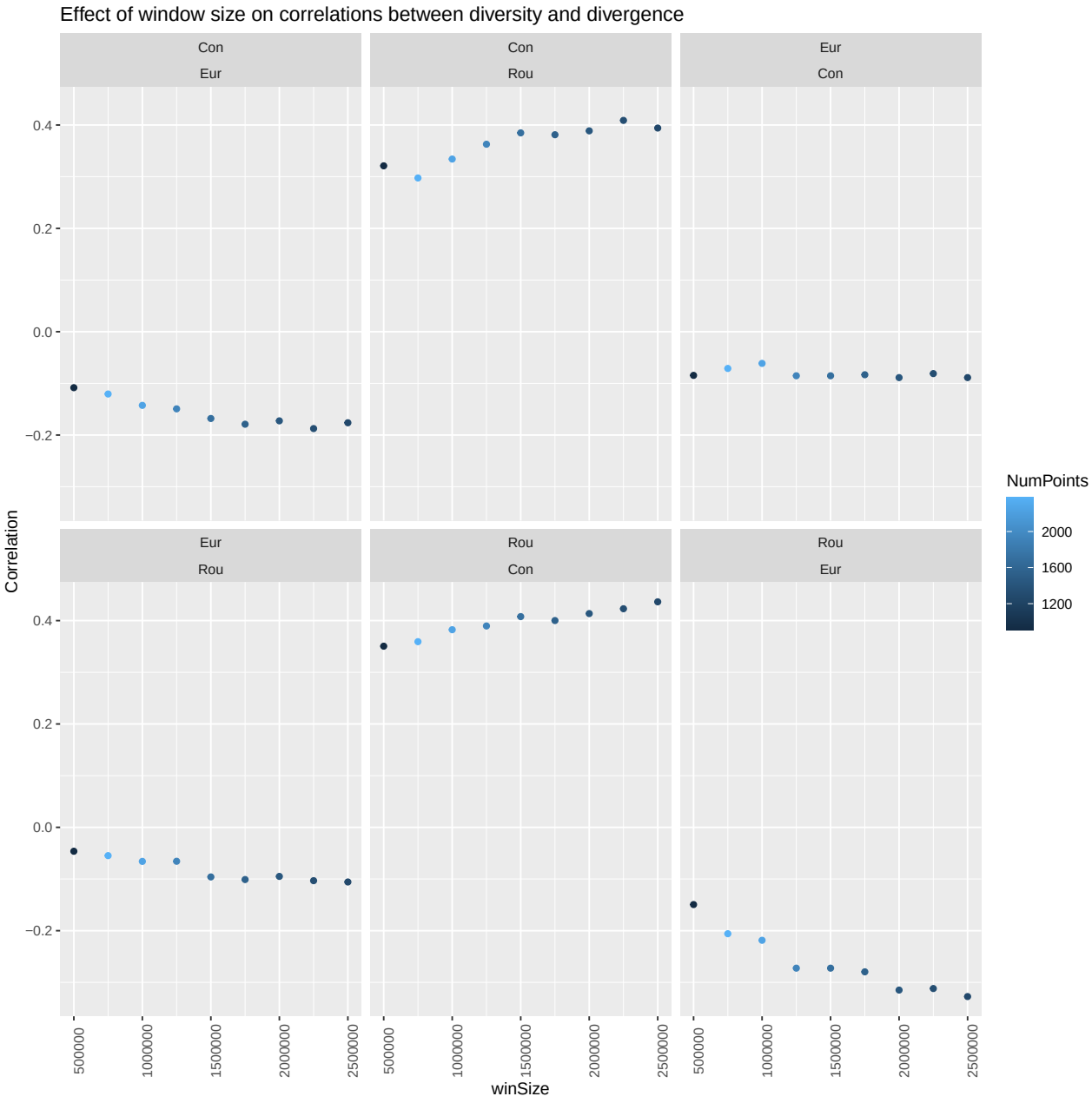



```
ggplot(popGenStats, aes(x = pi_eur, y = dxy_con_eur)) + geom_point() +
  geom_smooth(method='lm') + facet_wrap(~W) + xlab('Diversity in E. europaeus') +
  ylab('Divergence from E. concolor') + ggtitle('E. europaeus')
```




```
X <- data.frame(NumPoints = sapply(split(popGenStats$sites, popGenStats$W), length),
                winSize = seq(from=500000, to=2500000, by=250000))
X$CorEurRou <- sapply(split(popGenStats, popGenStats$W),
                      function(x) cor(x$pi_eur, x$dxy_rou_eur))
X$CorRouEur <- sapply(split(popGenStats, popGenStats$W),
                      function(x) cor(x$pi_rou, x$dxy_rou_eur))
X$CorRouCon <- sapply(split(popGenStats, popGenStats$W),
                      function(x) cor(x$pi_rou, x$dxy_con_rou))
X$CorConRou <- sapply(split(popGenStats, popGenStats$W),
                      function(x) cor(x$pi_con, x$dxy_con_rou))
X$CorConEur <- sapply(split(popGenStats, popGenStats$W),
                      function(x) cor(x$pi_con, x$dxy_con_eur))
X$CorEurCon <- sapply(split(popGenStats, popGenStats$W),
                      function(x) cor(x$pi_eur, x$dxy_con_eur))
```

```
Xlong <- pivot_longer(X, 3:8, names_to = c('Diversity_in', 'Divergence_from'),
                     names_pattern = "Cor(...)(...)", values_to='Correlation')
ggplot(Xlong, aes(x=winSize, y=Correlation, color=NumPoints)) + geom_point() +
  facet_wrap(~ Diversity_in + Divergence_from) +
  ggtitle('Effect of window size on correlations between diversity and divergence') +
  theme(axis.text.x = element_text(angle = 90))
```



Clearly, neither changing the definition of genomic windows to a coordinate system, nor changing the window size removed the patterns that we saw.

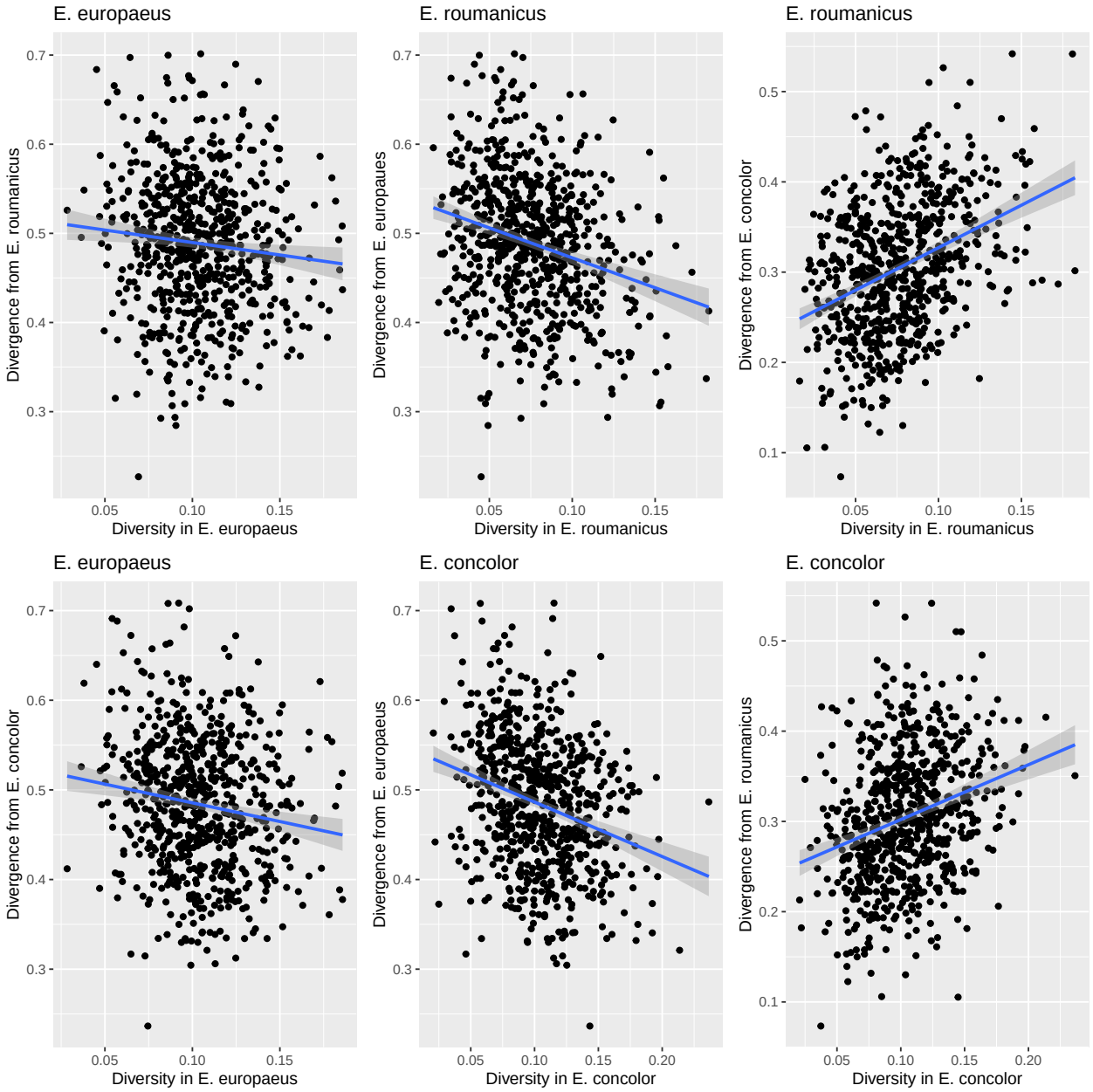
Ascertaintment bias

My second concern was the selection of sites included in the analysis, which is not random with respect to variation and divergence. Downsampling the *E. roumanicus* and *E. europaeus* populations to only five individuals and then removing invariant sites, I balance the power of detection of variable sites among species. I use 100 different sub-samples of those species. Not all replicates have the same number of genomic windows with data, because of the removal of windows without enough variable sites.

First I plot the average values of diversity and divergence among replicates using only genomic windows with data in the 100 replicates (697 windows). Then, using all replicates, I plot the distributions of the relevant correlation coefficients (violin plot).

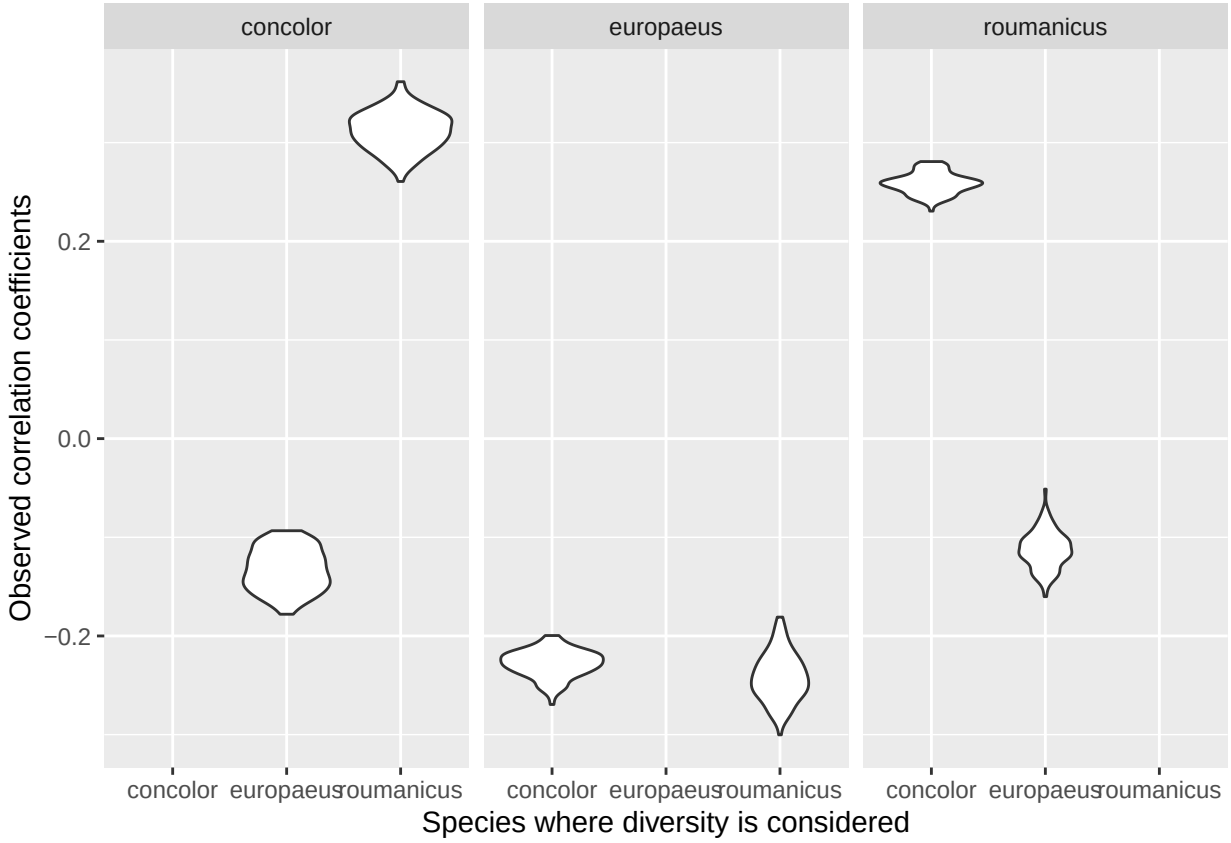
```
Subsamples <- data.frame()
for (Rep in 1:100) {
  z <- read.table(sprintf("subsampling/%03i.popGenStats.csv", Rep),
                  sep = ',', header = TRUE,
                  col.names = c('scaffold','start','end','mid','sites',
                                'pi_con', 'pi_rou', 'pi_eur',
                                'dxy_con_rou', 'dxy_con_eur', 'dxy_rou_eur',
                                'Fst_con_rou', 'Fst_con_eur', 'Fst_rou_eur'),
                  colClasses = c('character', 'integer', 'integer', 'integer', 'integer',
                                 'numeric', 'numeric', 'numeric',
                                 'numeric', 'numeric', 'numeric',
                                 'numeric', 'numeric', 'numeric'))
  z$Rep <- factor(Rep, levels=(1:100))
  Subsamples <- rbind(Subsamples, z)
}
rm(z)

Common <- ddply(Subsamples, .(scaffold, start, end), summarize,
                sites = mean(sites),
                pi_con = mean(pi_con),
                pi_rou = mean(pi_rou),
                pi_eur = mean(pi_eur),
                dxy_con_rou = mean(dxy_con_rou),
                dxy_con_eur = mean(dxy_con_eur),
                dxy_rou_eur = mean(dxy_rou_eur),
                Fst_con_rou = mean(Fst_con_rou),
                Fst_con_eur = mean(Fst_con_eur),
                Fst_rou_eur = mean(Fst_rou_eur),
                NumRep = length(Rep))
Common <- Common[Common$NumRep == 100, -14]

p1 <- ggplot(Common, aes(x=pi_eur, y=dxy_rou_eur)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. europaeus') +
  ylab('Divergence from E. roumanicus') + ggtitle('E. europaeus')
p2 <- ggplot(Common, aes(x=pi_rou, y=dxy_rou_eur)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. roumanicus') +
  ylab('Divergence from E. europaues') + ggtitle('E. roumanicus')
p3 <- ggplot(Common, aes(x=pi_rou, y=dxy_con_rou)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. roumanicus') +
  ylab('Divergence from E. concolor') + ggtitle('E. roumanicus')
p4 <- ggplot(Common, aes(x=pi_con, y=dxy_con_rou)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. concolor') +
```

```
  ylab('Divergence from E. roumanicus') + ggtitle('E. concolor')
p5 <- ggplot(Common, aes(x=pi_con, y=dxy_con_eur)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. concolor') +
  ylab('Divergence from E. europaeus') + ggtitle('E. concolor')
p6 <- ggplot(Common, aes(x=pi_eur, y=dxy_con_eur)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. europaeus') +
  ylab('Divergence from E. concolor') + ggtitle('E. europaeus')
grid.arrange(p1, p2, p3, p6, p5, p4, nrow=2)
```




```
Correlations <- data.frame(Rep = 1:100)
Correlations$europaeus_roumanicus <- sapply(split(Subsamples, Subsamples$Rep),
                                            function(x) cor(x$pi_eur, x$dxy_rou_eur))
Correlations$roumanicus_europaeus <- sapply(split(Subsamples, Subsamples$Rep),
                                            function(x) cor(x$pi_rou, x$dxy_rou_eur))
```

```
Correlations$roumanicus_concolor  <- sapply(split(Subsamples, Subsamples$Rep),
                                            function(x) cor(x$pi_rou, x$dxy_con_rou))
Correlations$concolor_roumanicus  <- sapply(split(Subsamples, Subsamples$Rep),
                                            function(x) cor(x$pi_con, x$dxy_con_rou))
Correlations$concolor_europaeus   <- sapply(split(Subsamples, Subsamples$Rep),
                                            function(x) cor(x$pi_con, x$dxy_con_eur))
Correlations$europaeus_concolor   <- sapply(split(Subsamples, Subsamples$Rep),
                                            function(x) cor(x$pi_eur, x$dxy_con_eur))
CorLong <- pivot_longer(Correlations, 2:7, names_to = c('Diversity_in', 'Divergence_from'),
                        names_sep = '_', values_to = 'Cor')
ggplot(CorLong, aes(x = Diversity_in, y=Cor)) + geom_violin() +
  facet_wrap(~ Divergence_from) + xlab('Species where diversity is considered') +
  ylab('Observed correlation coefficients')
```



Conclusions

I am not sure of how the seleccion of variable sites could bias the relationship between divergence and diversity. The subsampling reduces the number of genomic windows where these statistics get measured, removing some where not enough data is left. The estimates are presumed to vary with the composition of the subsamples. But the observed relationships stay the same. The do not seem to be an artifact of how the windows and sites included in the analysis is done. I could be missing something, but I think we need to look for alternative explanations.

From the plots, I perceive that all comparisons involving *E. europaeus* are unexpectedly non-neutral. One additional potential artifact I can think of is the fact that we are using a reference genome from *E. europaeus*.

This means that the more divergent a sequence is in any of the other species, the less likely it is to get properly aligned. Thus, we could be missing variation in *E. roumanicus* and *E. concolor* precisely where diverence is higher. That would cause negative correlations between diversity in *E. roumanicus* or *E. concolor* and their divergence from *E. europaeus*. But the negative correlation is also significant when using diversity in *E. europaeus*, which does not follow from the argument. In any case, if the reference genome was to blame, we should notice an effect of the minimum sequencing depth. Limiting the analysis to high-coverage sites should reduce the bias, if it exists.