Diversity and divergence in E. europaeus, E. roumanicus and E. concolor

J. Ignacio Lucas Lledó

7/4/2020

The data

I read the data in three main data frames:

- `allStats` includes the genome-wide population genetics for the three species and the introgression statistics for the pair *E. roumanicus* and *E. europaeus*.
- `geneStats` has the population genetics statistics for the three species only in genic regions: diversity (π), divergence (d_{xy}), and fixation index (F_{ST}).
- `interStats` has the population genetics statistics for the three species only in intergenic regions.

```
popgen     <- read.table('nonoverlap.PopGenStats.csv',  header=TRUE, sep=',',
                         col.names=c('scaffold', 'start', 'end', 'mid', 'sites',
                                     'pi_rou', 'pi_eur', 'pi_con',
                                     'dxy_rou_eur', 'dxy_rou_con', 'dxy_eur_con',
                                     'Fst_rou_eur', 'Fst_rou_con', 'Fst_eur_con'))
abbababa   <- read.table('nonoverlap.abbababa.csv', header=TRUE, sep=',')
abbababa$fd[abbababa$D <= 0] <- 0.0
abbababa$fdM[abbababa$D <= 0] <- 0.0
allStats   <- merge(popgen, abbababa, by=c('scaffold', 'start', 'end'))
allStats   <- allStats[! is.na(allStats$fd),]
geneStats  <- read.table('genes.PopGenStats.csv', header=TRUE, sep=',',
                         col.names=c('scaffold', 'start', 'end', 'mid', 'sites',
                                     'pi_rou', 'pi_eur', 'pi_con',
                                     'dxy_rou_eur', 'dxy_rou_con', 'dxy_eur_con',
                                     'Fst_rou_eur', 'Fst_rou_con', 'Fst_eur_con'))
interStats <- read.table('inter.PopGenStats.csv', header=TRUE, sep=',',
                         col.names=c('scaffold', 'start', 'end', 'mid', 'sites',
                                     'pi_rou', 'pi_eur', 'pi_con',
                                     'dxy_rou_eur', 'dxy_rou_con', 'dxy_eur_con',
                                     'Fst_rou_eur', 'Fst_rou_con', 'Fst_eur_con'))
```

E. roumanicus and *E. europaeus*

First I model the divergence between *E. roumanicus* and *E. europaeus* as a set of linear relationships with the rellevant genetic diversities and with the measure of genetic introgression f_d. Note that no introgression is detected in about 38% of sites, where $f_d = 0$.

```
m1 <- lm(dxy_rou_eur ~ pi_eur,                data=allStats)
m2 <- lm(dxy_rou_eur ~ pi_rou,                data=allStats)
```

```
m3 <- lm(dxy_rou_eur ~ pi_eur + fd,          data=allStats)
m4 <- lm(dxy_rou_eur ~ pi_rou + pi_eur,      data=allStats)
m5 <- lm(dxy_rou_eur ~ pi_rou + fd,          data=allStats)
m6 <- lm(dxy_rou_eur ~ pi_rou + pi_eur + fd, data=allStats)

summary(m6)
```

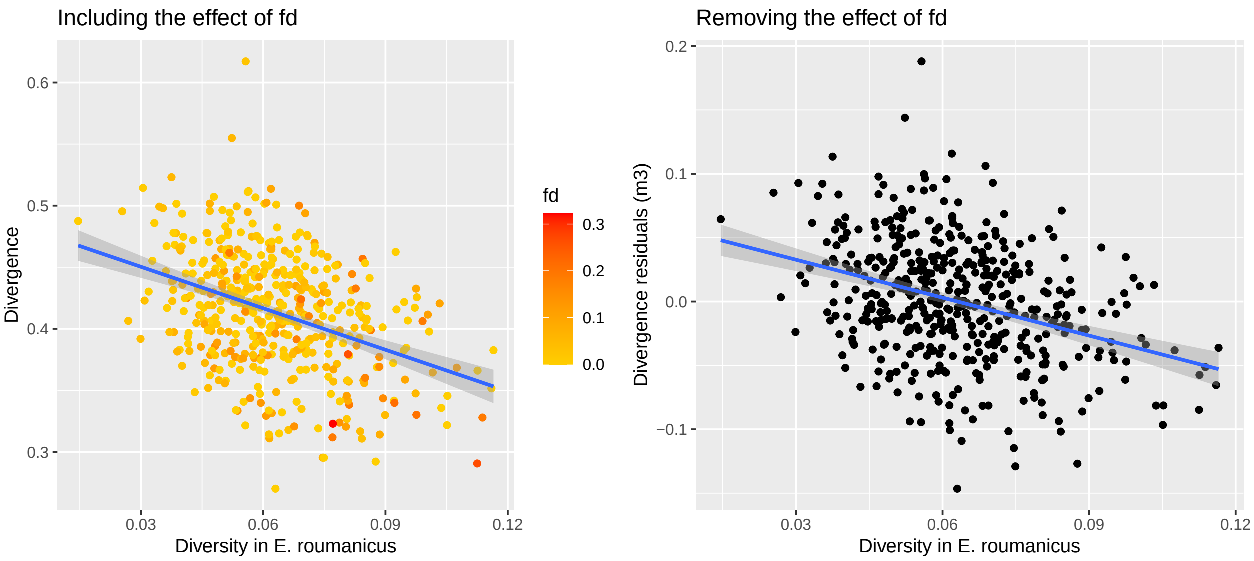
```
##
## Call:
## lm(formula = dxy_rou_eur ~ pi_rou + pi_eur + fd, data = allStats)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.144873 -0.028334  0.001964  0.029477  0.183495
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.50871    0.01467  34.672  < 2e-16 ***
## pi_rou      -1.03027    0.12438  -8.283 1.28e-15 ***
## pi_eur      -0.27617    0.13827  -1.997  0.04637 *
## fd          -0.12919    0.03950  -3.271  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04358 on 466 degrees of freedom
## Multiple R-squared:  0.1772, Adjusted R-squared:  0.1719
## F-statistic: 33.46 on 3 and 466 DF,  p-value: < 2.2e-16
```

```
allStats$res3 <- residuals(m3)
allStats$res5 <- residuals(m5)
```

Previous observation that the divergence between *E. roumanicus* and *E. europaeus* shows a negative correlation with the genetic diversity within *E. roumanicus* is maintained. We can also see that the relationship with the diversity in *E. europaeus* is also negative, although much less important.

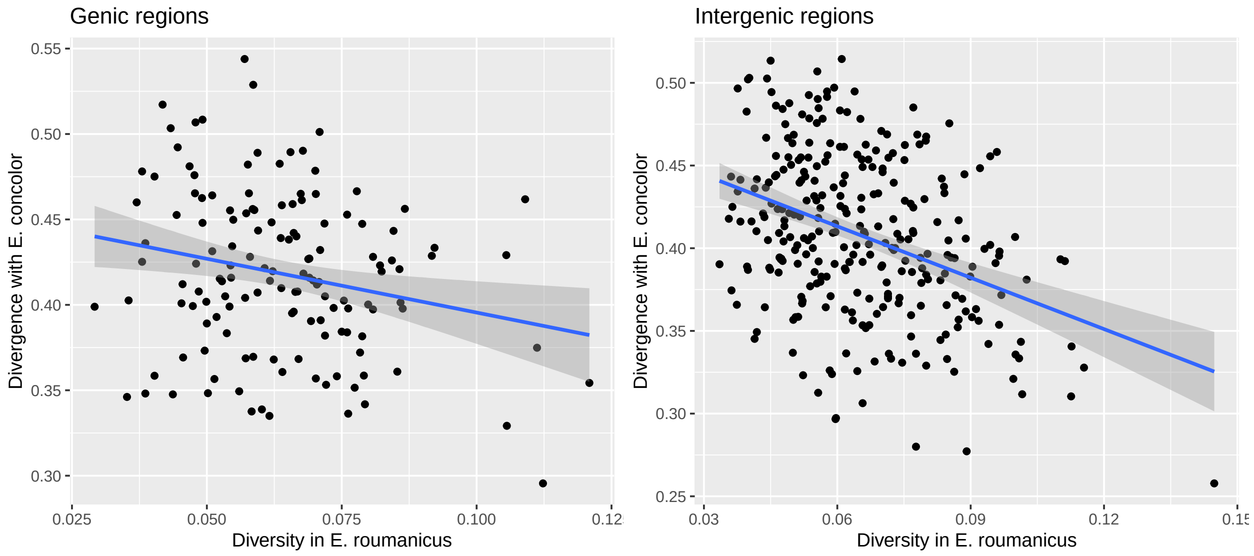
Using the residuals from model 3, we remove the effect of the detected introgression (*fd*), and the effect of diversity in *E. europaeus*. As we can see below, even after removing those effects, a considerable fraction of the variation in divergence among genomic sites could be explained by levels of genetic variation in *E. roumanicus*, and the relationship is negative.

```
p1 <- ggplot(allStats, aes(x=pi_rou, y=dxy_rou_eur, color=fd)) +
  geom_point() + geom_smooth(method='lm') + xlab('Diversity in E. roumanicus') +
  ylab('Divergence') + scale_colour_gradient2(low='yellow',mid='orange',high='red',midpoint=0.1) +
  ggtitle('Including the effect of fd')
p2 <- ggplot(allStats, aes(x=pi_rou, y=res3)) +
  geom_point() + geom_smooth(method='lm') + xlab('Diversity in E. roumanicus') +
  ylab('Divergence residuals (m3)') + ggtitle('Removing the effect of fd')
grid.arrange(p1, p2, nrow=1)
```



The fact that our measure of introgression does not remove completely the negative correlation suggests two interpretations: either our measure of introgression badly underestimates introgression, or there is something else going one. Both options are compatible.

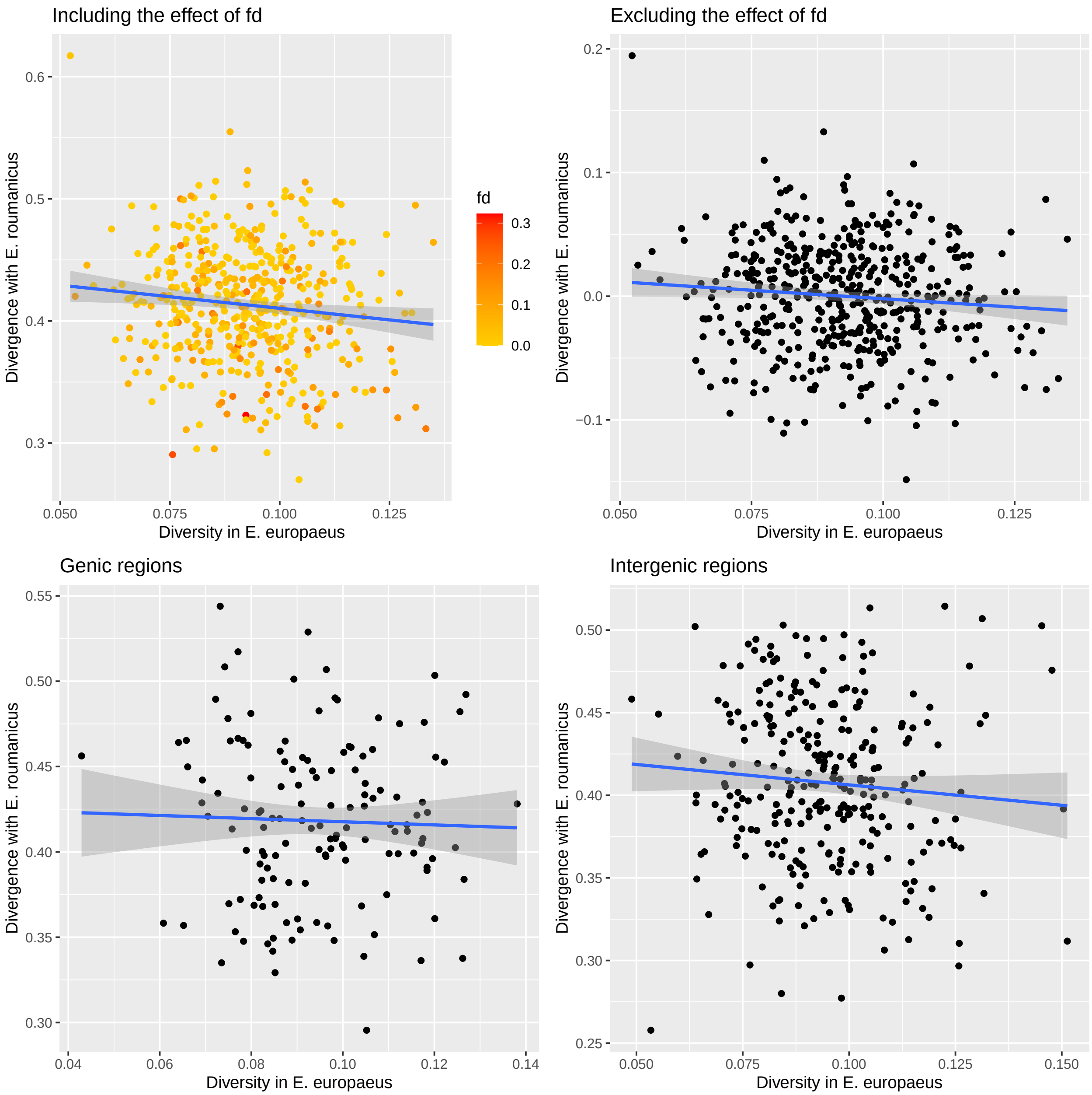
```
p1 <- ggplot(geneStats, aes(x=pi_rou, y=dxy_rou_eur)) +
  geom_point() + geom_smooth(method='lm') + xlab('Diversity in E. roumanicus') +
  ylab('Divergence with E. concolor') + ggtitle('Genic regions')
p2 <- ggplot(interStats, aes(x=pi_rou, y=dxy_rou_eur)) +
  geom_point() + geom_smooth(method='lm') + xlab('Diversity in E. roumanicus') +
  ylab('Divergence with E. concolor') + ggtitle('Intergenic regions')
grid.arrange(p1, p2, nrow=1)
```



The fact that the above negative correlation is stronger in intergenic regions could point at a functional effect. But could also be a consequence of the larger number of data points. I would not give it much importance.

What about the correlation between diversity in *E. europaeus* and its divergence with *E. roumanicus*? The linear models suggested an also negative, but milder relationship.

```
p1 <- ggplot(allStats, aes(x=pi_eur, y=dxy_rou_eur, color=fd)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. europaeus') +
  scale_colour_gradient2(low='yellow', mid='orange', high='red',midpoint=0.1) +
  ylab('Divergence with E. roumanicus') + ggtitle('Including the effect of fd')
p2 <- ggplot(allStats, aes(x=pi_eur, y=res5)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. europaeus') +
  ylab('Divergence with E. roumanicus') + ggtitle('Excluding the effect of fd')
p3 <- ggplot(geneStats, aes(x=pi_eur, y=dxy_rou_eur)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. europaeus') +
  ylab('Divergence with E. roumanicus') + ggtitle('Genic regions')
p4 <- ggplot(interStats, aes(x=pi_eur, y=dxy_rou_eur)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. europaeus') +
  ylab('Divergence with E. roumanicus') + ggtitle('Intergenic regions')
grid.arrange(p1, p2, p3, p4, nrow=2)
```



The lack of correlation is also a departure from the neutral expectations, and requires an explanation. I guess, we could claim introgression as well.

E. roumanicus and *E. concolor*

We can do the same for *E. roumanicus* and *E. concolor*. I include the estimated introgression between *E. roumanicus* and *E. europaeus* in the model, because it could affect positively the divergence between *E. roumanicus* and *E. concolor*.

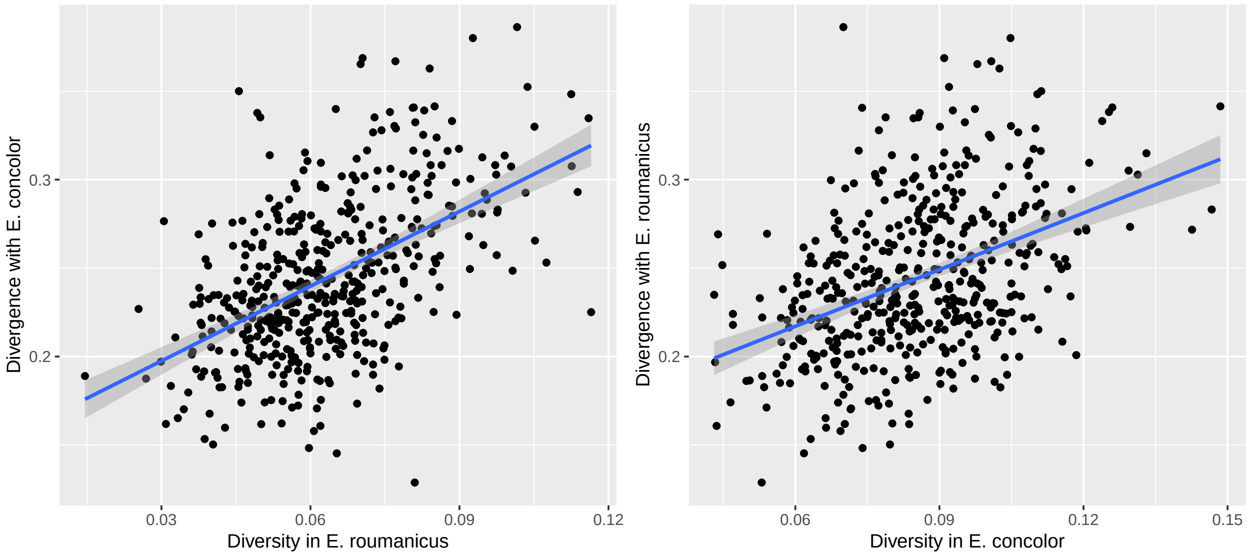
```
m1 <- lm(dxy_rou_con ~ pi_rou,               data=allStats)
m2 <- lm(dxy_rou_con ~ pi_con,               data=allStats)
m3 <- lm(dxy_rou_con ~ pi_rou + pi_con,      data=allStats)
m4 <- lm(dxy_rou_con ~ pi_rou + fd,          data=allStats)
m5 <- lm(dxy_rou_con ~ pi_con + fd,          data=allStats)
m6 <- lm(dxy_rou_con ~ pi_rou + pi_con + fd, data=allStats)

summary(m6)
```

```
##
## Call:
## lm(formula = dxy_rou_con ~ pi_rou + pi_con + fd, data = allStats)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.11388 -0.02340 -0.00253  0.02271  0.11338
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.104053   0.009184  11.330  < 2e-16 ***
## pi_rou      1.141196   0.102867  11.094  < 2e-16 ***
## pi_con      0.745138   0.095527   7.800 4.09e-14 ***
## fd          0.117003   0.032192   3.635 0.000309 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03522 on 466 degrees of freedom
## Multiple R-squared:  0.3818, Adjusted R-squared:  0.3778
## F-statistic: 95.94 on 3 and 466 DF,  p-value: < 2.2e-16
```

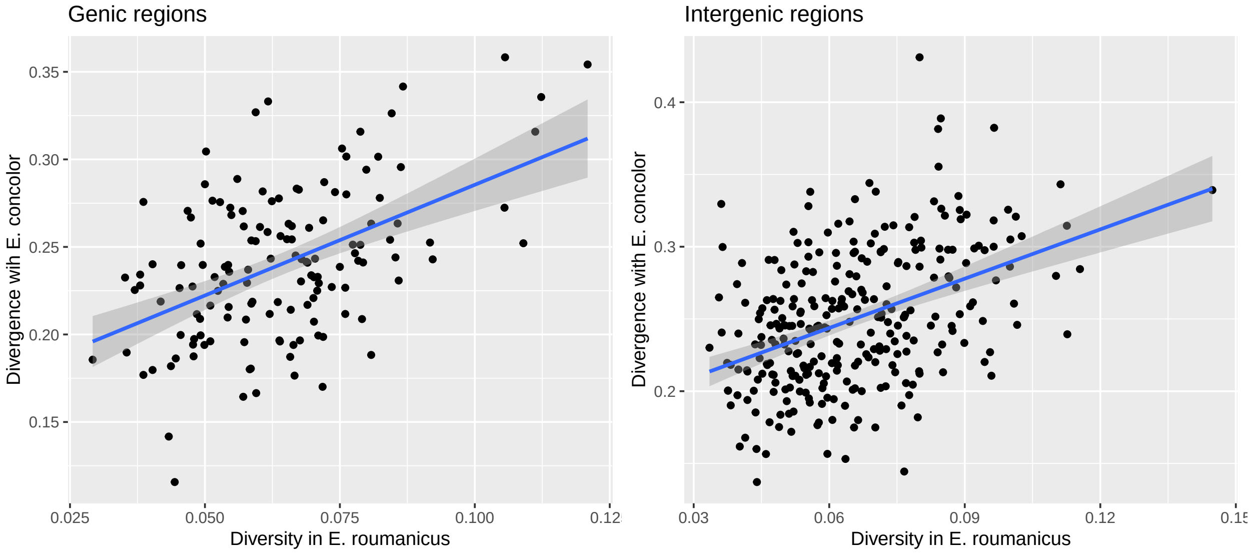
All terms in `m6` are significant and with a positive effect:

```
p1 <- ggplot(allStats, aes(x=pi_rou, y=dxy_rou_con)) +
  geom_point() + geom_smooth(method='lm') + xlab('Diversity in E. roumanicus') +
  ylab('Divergence with E. concolor')
p2 <- ggplot(allStats, aes(x=pi_con, y=dxy_rou_con)) +
  geom_point() + geom_smooth(method='lm') + xlab('Diversity in E. concolor') +
  ylab('Divergence with E. roumanicus')
grid.arrange(p1, p2, nrow=1)
```



The positive relationship is exactly what we expect from the neutral theory.

```r
p1 <- ggplot(geneStats, aes(x=pi_rou, y=dxy_rou_con)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. roumanicus') +
  ylab('Divergence wih E. concolor') + ggtitle('Genic regions')
p2 <- ggplot(interStats, aes(x=pi_rou, y=dxy_rou_con)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. roumanicus') +
  ylab('Divergence with E. concolor') + ggtitle('Intergenic regions')
grid.arrange(p1, p2, nrow=1)
```


Genic regions

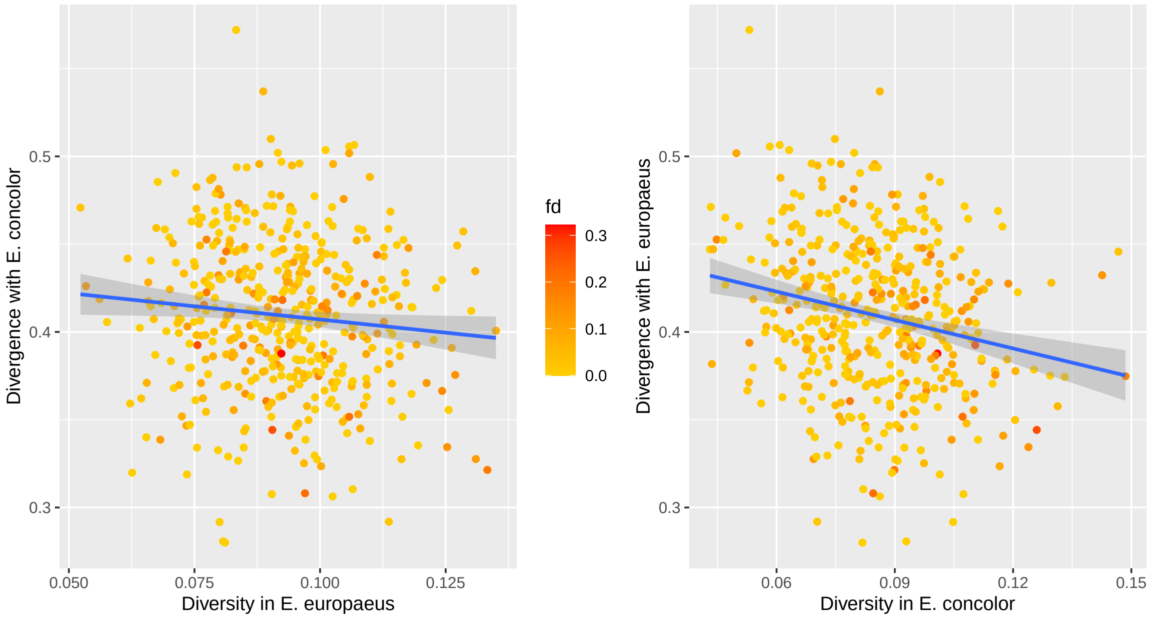
Intergenic regions

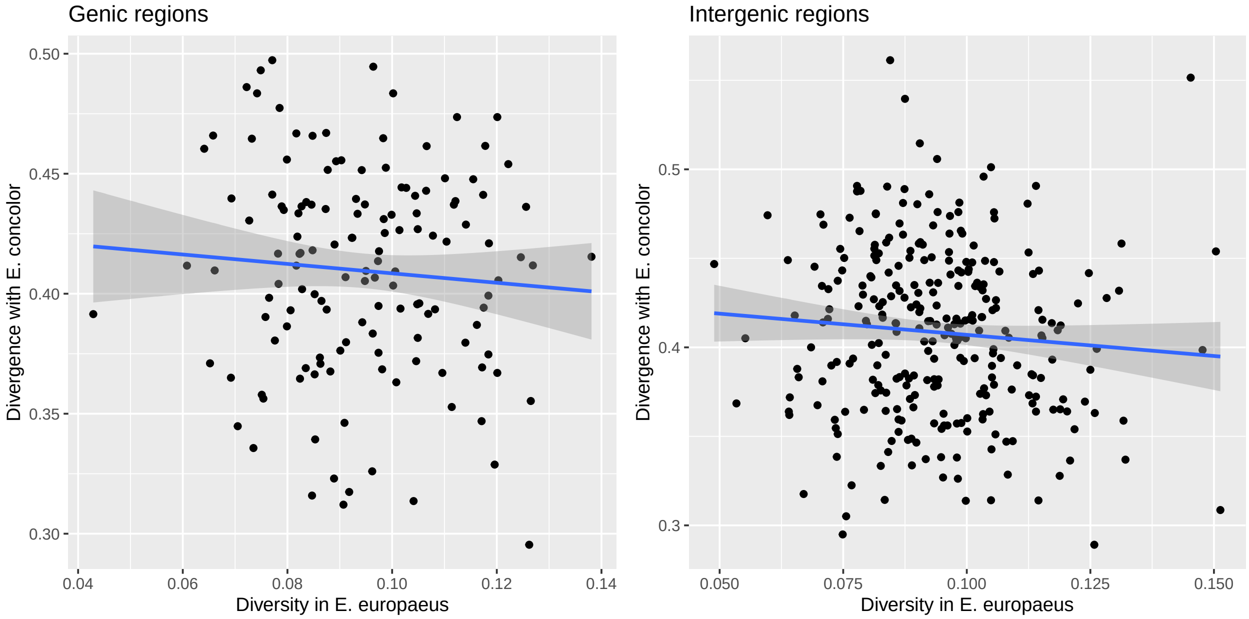
According to this, the divesity seggregating in *E. roumanicus*, even in genic regions, is quite neutral. Recall that we actually sampled very few non-synonymous polymorphisms, and we did not distinguish exons from introns within genic regions. Our functional classification of sites could be too rough to notice the selective constraints.

E. europaeus and *E. concolor*

The analysis would not be complete without the last comparison. Because there is clearly, absolutely no introgression between *E. europaeus* and *E. concolor*, the expected pattern is the same as between *E. roumanicus* and *E. concolor*: a positive relationship between diversity and divergence. I skip the linear models.

```r
p1 <- ggplot(allStats, aes(x=pi_eur, y=dxy_eur_con, color=fd)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. europaeus') +
  scale_colour_gradient2(low='yellow', mid='orange', high='red',midpoint=0.1) +
  ylab('Divergence with E. concolor')
p2 <- ggplot(allStats, aes(x=pi_con, y=dxy_eur_con, color=fd)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. concolor') +
  scale_color_gradient2(low='yellow', mid='orange', high='red', midpoint=0.1) +
  ylab('Divergence with E. europaeus')
p3 <- ggplot(geneStats, aes(x=pi_eur, y=dxy_eur_con)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. europaeus') +
  ylab('Divergence with E. concolor') + ggtitle('Genic regions')
p4 <- ggplot(interStats, aes(x=pi_eur, y=dxy_eur_con)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. europaeus') +
  ylab('Divergence with E. concolor') + ggtitle('Intergenic regions')
grid.arrange(p1, p2, p3, p4, nrow=2)
```

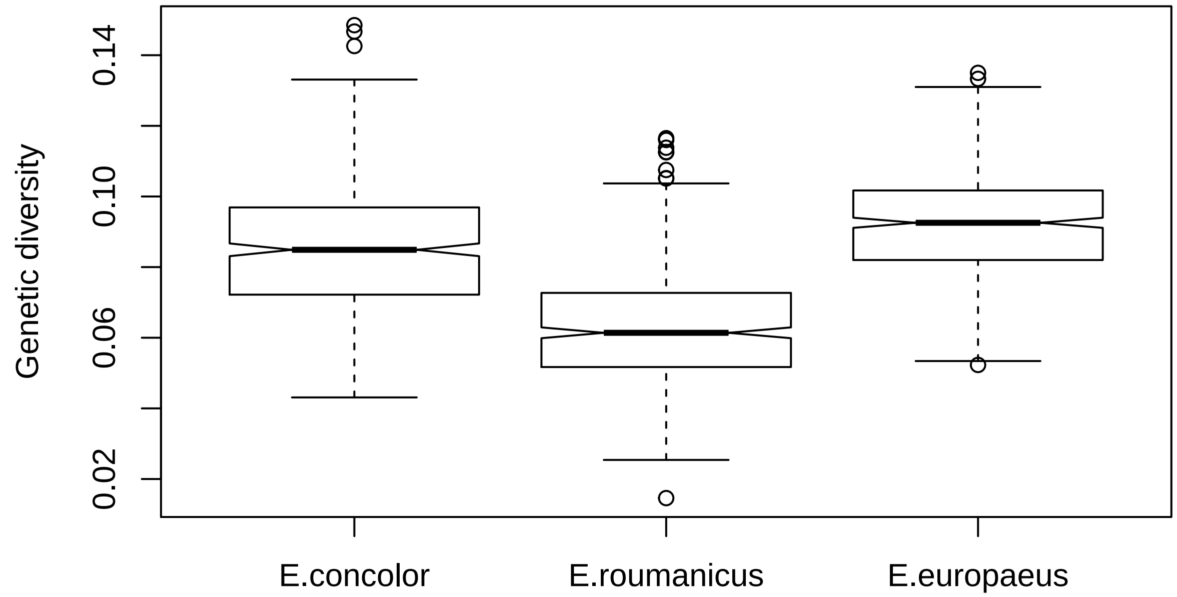




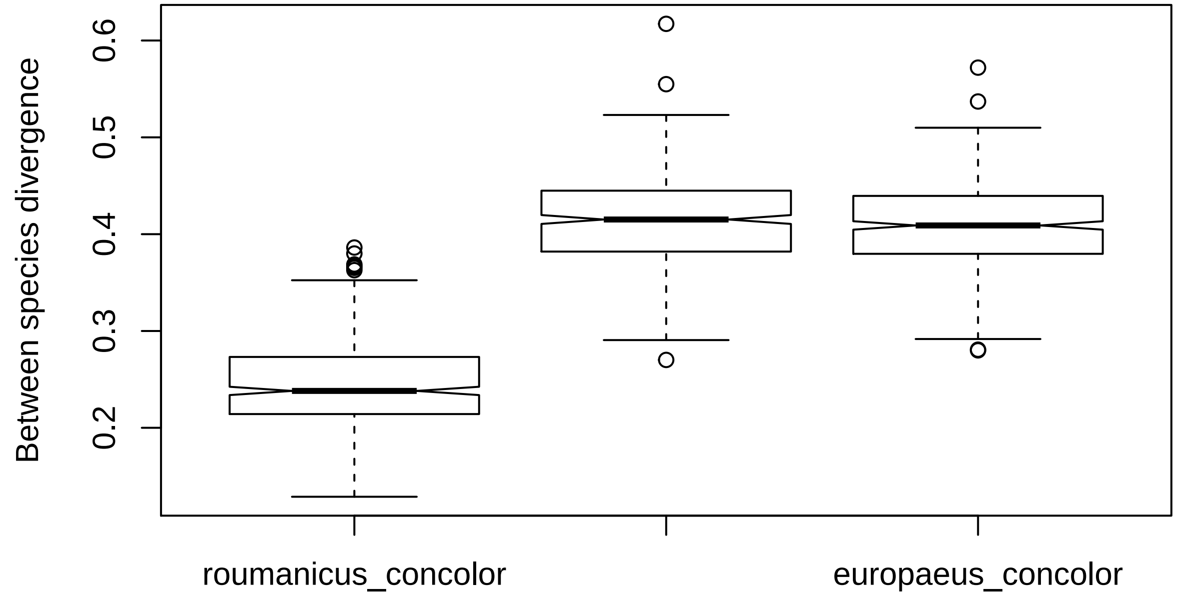
This unexpected lack of correlation makes me think that introgression is not the end of the story. I do not see a way to explain this with introgression. There must be something else going on here. Now, it looks like *E. europaeus* is peculiar. Its main difference may be its population structure. I have not been able yet to imagine a way for the population structure to affect the relationship between divergence and diversity.

Some more plots

```
boxplot(list(E.concolor   = allStats$pi_con,
             E.roumanicus = allStats$pi_rou,
             E.europaeus  = allStats$pi_eur),
        ylab = 'Genetic diversity',
        notch = TRUE)
```



```
boxplot(list(roumanicus_concolor  = allStats$dxy_rou_con,
             roumanicus_europaeus = allStats$dxy_rou_eur,
             europaeus_concolor   = allStats$dxy_eur_con),
        ylab = 'Between species divergence',
        notch = TRUE)
```



Session Information

```
sessionInfo()
```

```
## R version 3.6.3 (2020-02-29)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.4 LTS
##
```

```
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.7.1
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.7.1
##
## locale:
##  [1] LC_CTYPE=ca_ES.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=es_ES.UTF-8        LC_COLLATE=ca_ES.UTF-8
##  [5] LC_MONETARY=es_ES.UTF-8    LC_MESSAGES=ca_ES.UTF-8
##  [7] LC_PAPER=es_ES.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=es_ES.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] gridExtra_2.3 ggplot2_3.3.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.3       pillar_1.4.3     compiler_3.6.3   tools_3.6.3
##  [5] digest_0.6.25    evaluate_0.14    lifecycle_0.2.0  tibble_2.1.3
##  [9] gtable_0.3.0     nlme_3.1-145     lattice_0.20-40  mgcv_1.8-31
## [13] pkgconfig_2.0.3  rlang_0.4.5      Matrix_1.2-18    yaml_2.2.1
## [17] xfun_0.12        withr_2.1.2      dplyr_0.8.5      stringr_1.4.0
## [21] knitr_1.28       grid_3.6.3       tidyselect_1.0.0 glue_1.3.2
## [25] R6_2.4.1         rmarkdown_2.1    purrr_0.3.3      farver_2.0.3
## [29] magrittr_1.5     scales_1.1.0     htmltools_0.4.0  splines_3.6.3
## [33] assertthat_0.2.1 colorspace_1.4-1 labeling_0.3     stringi_1.4.6
## [37] munsell_0.5.0    crayon_1.3.4
```