Assessment of reference bias

J. Ignacio Lucas Lledó

10/4/2020

```
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(gridExtra)
library(plyr)
POPGENSTATS_FILE <- '../2020-04-08/windowSize/750000.popGenStats.csv'
```

The argument is: if highly divergent loci were more difficult to map, we should see a negative relationship between divergence from the reference (*E. europaeus*) and depth of coverage. Because of a lower depth of coverage, it would then be possible that diversity measures were downwardly biased in highly divergent loci. Because coverage of depth would mediate this effect of the reference genome, we can assess it and correct it statistically, if we know the average depth of coverage in every window.

In this folder I use estimates of diversity and depth obtained for every site with vcftools. Then I calculate my own estimates of divergence. Then, I bin the results in genomic windows, using window definitions from previous analysis, to allow for comparison. Neither my own estimates of divergence (averaged per window) nor vcftools' estimates of diversity (π, also averaged) are exactly the same as the window-wise values obtained with Simon H. Martin's python scripts, but they are very correlated (not shown). And there is a good reason for the discrepancy. Simon creates fake haplotypes, and estimates expected numbers of differences among them. I treat each site separately. I didn't check (would not be difficult), but I suppose that's the reason for the difference.

```
AF_con <- read.table('concolor.frq', header=FALSE, skip=1,
                     col.names = c('CHROM','POS', 'N_ALLELES', 'N_CHR', 'FREQ1', 'FREQ2'))
AF_rou <- read.table('roumanicus.frq', header=FALSE, skip=1,
                     col.names = c('CHROM', 'POS', 'N_ALLELES', 'N_CHR', 'FREQ1', 'FREQ2'))
AF_eur <- read.table('europaeus.frq', header=FALSE, skip=1,
                     col.names = c('CHROM', 'POS', 'N_ALLELES', 'N_CHR', 'FREQ1', 'FREQ2'))
PI_con <- read.table('concolor.sites.pi', header=TRUE)
PI_rou <- read.table('roumanicus.sites.pi', header=TRUE)
PI_eur <- read.table('europaeus.sites.pi', header=TRUE)
DP_con <- read.table('concolor.ldepth', header=TRUE)
DP_rou <- read.table('roumanicus.ldepth', header=TRUE)
DP_eur <- read.table('europaeus.ldepth', header=TRUE)
FST_con_rou <- read.table('concolor_roumanicus.weir.fst', header=TRUE)
FST_con_eur <- read.table('concolor_europaeus.weir.fst', header=TRUE)
FST_rou_eur <- read.table('roumanicus_europaeus.weir.fst', header=TRUE)
names(FST_con_rou)[3] <- 'FST'
names(FST_con_eur)[3] <- 'FST'
names(FST_rou_eur)[3] <- 'FST'
FST_con_rou[! (is.nan(FST_con_rou$FST)) & FST_con_rou$FST < 0, 'FST'] <- 0
FST_con_eur[! (is.nan(FST_con_eur$FST)) & FST_con_eur$FST < 0, 'FST'] <- 0
FST_rou_eur[! (is.nan(FST_rou_eur$FST)) & FST_rou_eur$FST < 0, 'FST'] <- 0
```

```
all <- data.frame(
  CHROM  = AF_con$CHROM,
  POS    = AF_con$POS,
  N_con  = AF_con$N_CHR,
  N_rou  = AF_rou$N_CHR,
  N_eur  = AF_eur$N_CHR,
  AF_con = AF_con$FREQ1,
  AF_rou = AF_rou$FREQ1,
  AF_eur = AF_eur$FREQ1,
  DP_con = DP_con$SUM_DEPTH,
  DP_rou = DP_rou$SUM_DEPTH,
  DP_eur = DP_eur$SUM_DEPTH,
  PI_con = PI_con$PI,
  PI_rou = PI_rou$PI,
  PI_eur = PI_eur$PI,
  FST_con_rou = FST_con_rou$FST,
  FST_con_eur = FST_con_eur$FST,
  FST_rou_eur = FST_rou_eur$FST
)
rm(AF_con, AF_rou, AF_eur, DP_con, DP_rou, DP_eur,
   PI_con, PI_rou, PI_eur, FST_con_rou, FST_con_eur, FST_rou_eur)
```

I read one of the results in `2020-04-08` in `PGS` (population genetic statistics) to re-use the coordinates. Then, I assign to every position the number of the window they belong to as factor `W`.

```
PGS <- read.table(POPGENSTATS_FILE, 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('character', 'integer', 'integer', 'integer',
                                 'integer', 'numeric',
                                 'numeric', 'numeric', 'numeric', 'numeric',
                                 'numeric', 'numeric', 'numeric', 'numeric'))

all$W <- factor(NA, levels=1:length(PGS$scaffold))
for (w in 1:length(PGS$scaffold)) {
  all[all$CHROM == PGS$scaffold[w] & all$POS %in% PGS$start[w]:PGS$end[w], 'W'] <- w
}
```

Below I implement my measure of divergence. For every site, it represents the proportion of comparisons between alleles of different populations that are different. Then, I remove a small portion of sites with very high depths of coverage, which were skewing the distribution of depths.

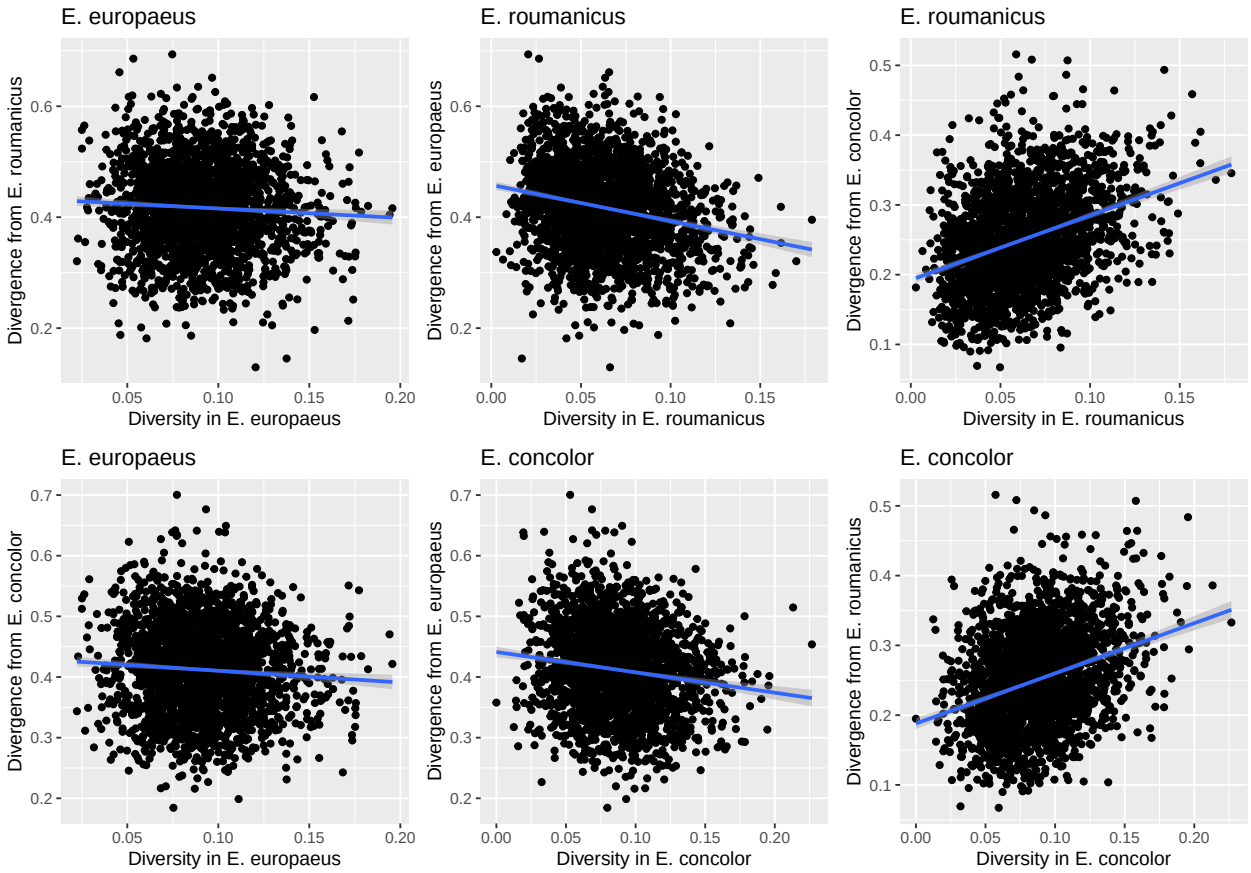
```
dxy <- function(Nx, Ny, Fx, Fy) {
  D <- (Nx * Fx * Ny * (1 - Fy) + Nx * (1 - Fx) * Ny * Fy) / (Nx * Ny)
  return(D)
}

all$dxy_con_rou <- dxy(all$N_con, all$N_rou, all$AF_con, all$AF_rou)
all$dxy_con_eur <- dxy(all$N_con, all$N_eur, all$AF_con, all$AF_eur)
all$dxy_rou_eur <- dxy(all$N_rou, all$N_eur, all$AF_rou, all$AF_eur)

# Depth is very skewed. I remove outliers:
f <- all$DP_con <= quantile(all$DP_con, 0.996) &
  all$DP_rou <= quantile(all$DP_rou, 0.996) &
```

```
  all$DP_eur <= quantile(all$DP_eur, 0.996)
all <- all[f,]
```

Now, I summarize the variables by window and remove sites not assigned (they belong to windows excluded for having too low numbers of sites) and I remove 17 windows that are left with less than 30 sites after having removed the depth outliers.

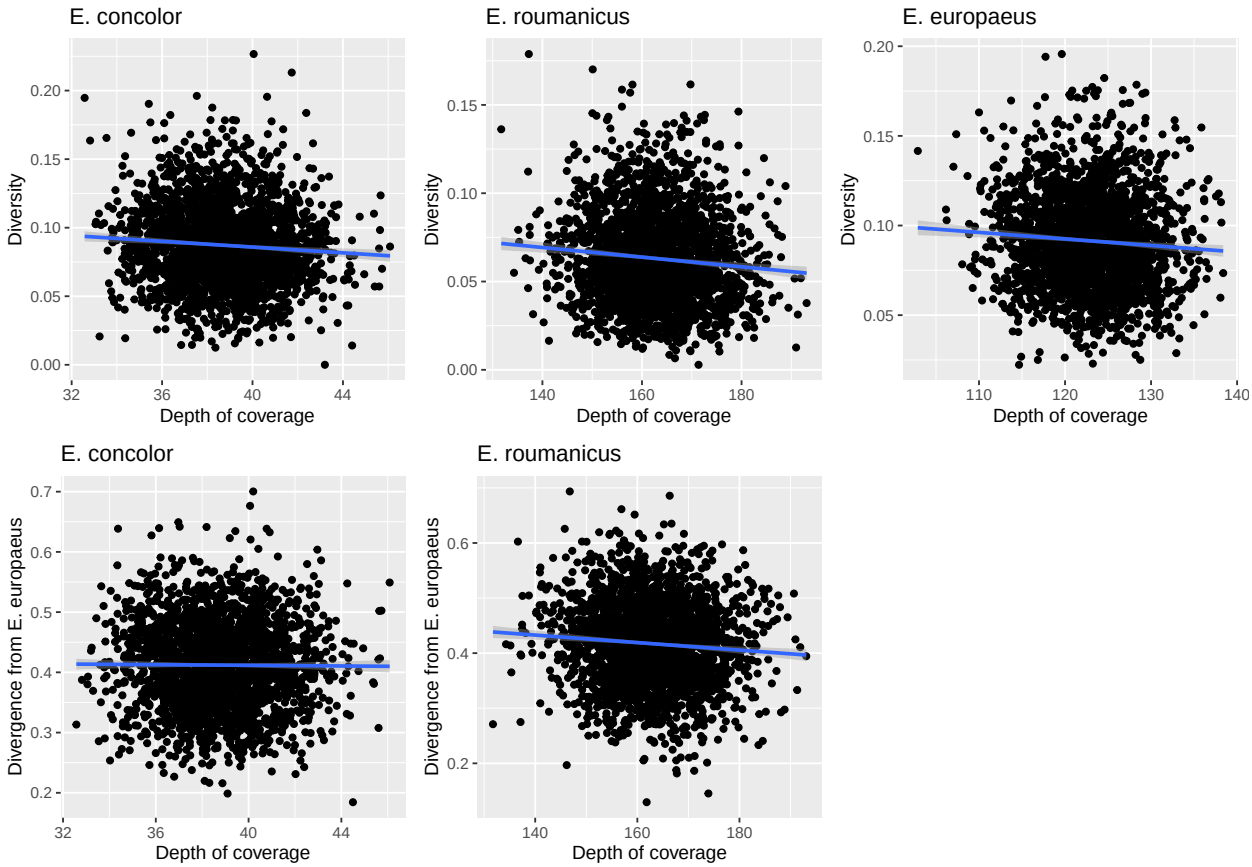
```
Wall <- ddply(all, .(W), summarize,
              sites = length(POS),
              pi_con = mean(PI_con, na.rm = TRUE),
              pi_rou = mean(PI_rou, na.rm = TRUE),
              pi_eur = mean(PI_eur, na.rm = TRUE),
              dxy_con_rou = mean(dxy_con_rou, na.rm = TRUE),
              dxy_con_eur = mean(dxy_con_eur, na.rm = TRUE),
              dxy_rou_eur = mean(dxy_rou_eur, na.rm = TRUE),
              Fst_con_rou = mean(FST_con_rou, na.rm = TRUE),
              Fst_con_eur = mean(FST_con_eur, na.rm = TRUE),
              Fst_rou_eur = mean(FST_rou_eur, na.rm = TRUE),
              N_con = mean(N_con),
              N_rou = mean(N_rou),
              N_eur = mean(N_eur),
              AF_con = mean(AF_con),
              AF_rou = mean(AF_rou),
              AF_eur = mean(AF_eur),
              DP_con = mean(DP_con),
              DP_rou = mean(DP_rou),
              DP_eur = mean(DP_eur))
Wall <- Wall[! is.na(Wall$W),]
Wall <- Wall[Wall$sites >= 30, ]
```

```
p1 <- ggplot(Wall, 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(Wall, aes(x=pi_rou, y=dxy_rou_eur)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. roumanicus') +
  ylab('Divergence from E. europaeus') + ggtitle('E. roumanicus')
p3 <- ggplot(Wall, 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(Wall, 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(Wall, 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(Wall, 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)
```



Even using statistics calculated in a different way, and removing depth outliers, the curious relationships that we observed before are still there. Let's see if depth has anything to do. Recall, what I suspected is that the more divergent and diverse sites would get lower coverage.

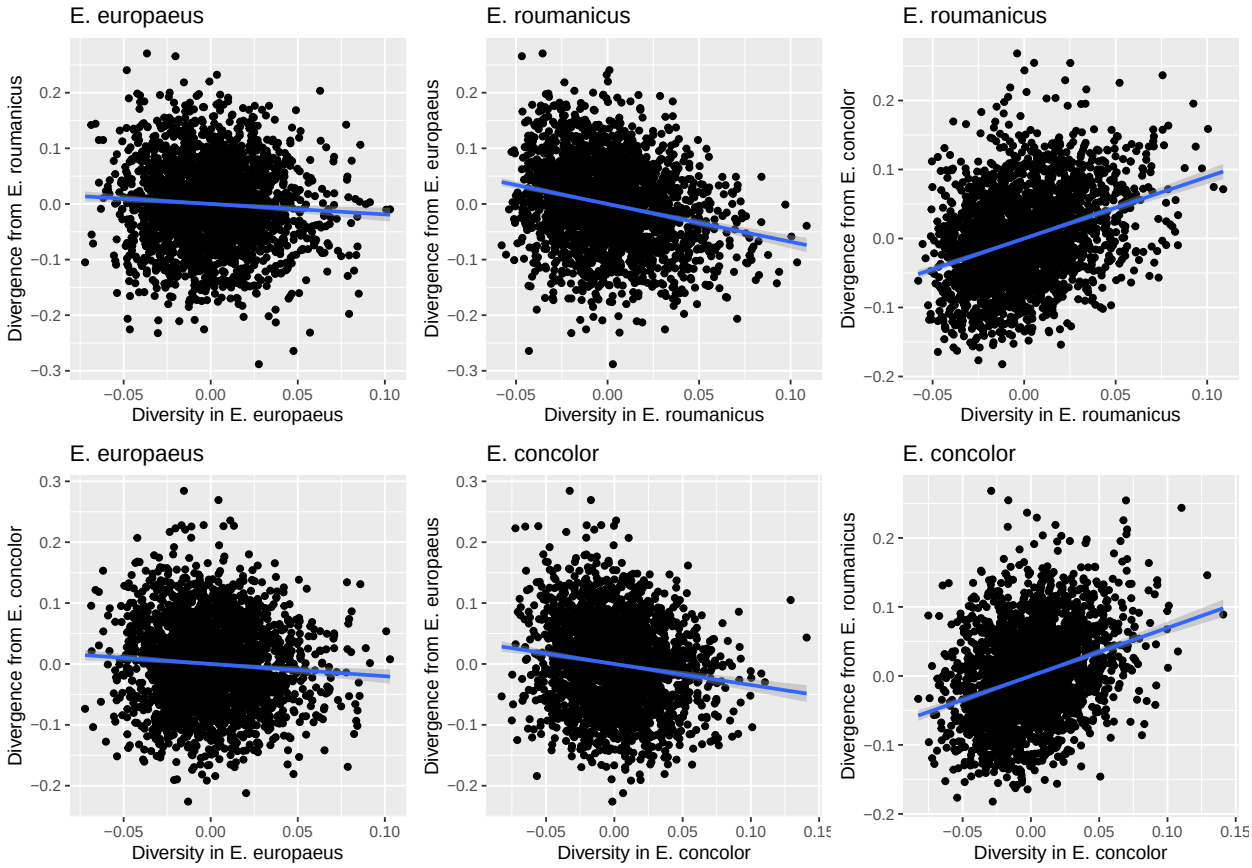
```
p1 <- ggplot(Wall, aes(x=DP_con, y=pi_con)) + geom_point() +
  geom_smooth(method='lm') + xlab('Depth of coverage') +
  ylab('Diversity') + ggtitle('E. concolor')
p2 <- ggplot(Wall, aes(x=DP_rou, y=pi_rou)) + geom_point() +
  geom_smooth(method='lm') + xlab('Depth of coverage') +
  ylab('Diversity') + ggtitle('E. roumanicus')
p3 <- ggplot(Wall, aes(x=DP_eur, y=pi_eur)) + geom_point() +
  geom_smooth(method='lm') + xlab('Depth of coverage') +
  ylab('Diversity') + ggtitle('E. europaeus')
p4 <- ggplot(Wall, aes(x=DP_con, y=dxy_con_eur)) + geom_point() +
  geom_smooth(method='lm') + xlab('Depth of coverage') +
  ylab('Divergence from E. europaeus') + ggtitle('E. concolor')
p5 <- ggplot(Wall, aes(x=DP_rou, y=dxy_rou_eur)) + geom_point() +
  geom_smooth(method='lm') + xlab('Depth of coverage') +
  ylab('Divergence from E. europaeus') + ggtitle('E. roumanicus')
grid.arrange(p1,p2, p3, p4, p5, nrow=2)
```



Well, in some cases, a slightly negative relationship is observed, as expected. But the relationship is so weak that I can't expect much from it. Let's remove it, anyways from all variables, and see if the patterns remain.

```
Wall$res_pi_con <- residuals(lm(Wall$pi_con ~ Wall$DP_con))
Wall$res_pi_rou <- residuals(lm(Wall$pi_rou ~ Wall$DP_rou))
Wall$res_pi_eur <- residuals(lm(Wall$pi_eur ~ Wall$DP_eur))
Wall$res_dxy_con_rou <- residuals(lm(Wall$dxy_con_rou ~ Wall$DP_con + Wall$DP_rou))
Wall$res_dxy_con_eur <- residuals(lm(Wall$dxy_con_eur ~ Wall$DP_con + Wall$DP_eur))
Wall$res_dxy_rou_eur <- residuals(lm(Wall$dxy_rou_eur ~ Wall$DP_rou + Wall$DP_eur))

p1 <- ggplot(Wall, aes(x=res_pi_eur, y=res_dxy_rou_eur)) + geom_point() +
  geom_smooth(method='lm')  + xlab('Diversity in E. europaeus') +
  ylab('Divergence from E. roumanicus') + ggtitle('E. europaeus')
p2 <- ggplot(Wall, aes(x=res_pi_rou, y=res_dxy_rou_eur)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. roumanicus') +
  ylab('Divergence from E. europaeus') + ggtitle('E. roumanicus')
p3 <- ggplot(Wall, aes(x=res_pi_rou, y=res_dxy_con_rou)) + geom_point() +
  geom_smooth(method='lm') + xlab('Diversity in E. roumanicus') +
  ylab('Divergence from E. concolor')  + ggtitle('E. roumanicus')
p4 <- ggplot(Wall, aes(x=res_pi_con, y=res_dxy_con_rou)) + geom_point() +
  geom_smooth(method='lm')   + xlab('Diversity in E. concolor') +
  ylab('Divergence from E. roumanicus') +  ggtitle('E. concolor')
p5 <- ggplot(Wall, aes(x=res_pi_con, y=res_dxy_con_eur)) + geom_point() +
  geom_smooth(method='lm')  +  xlab('Diversity in E. concolor') +
  ylab('Divergence from E. europaeus')  +  ggtitle('E. concolor')
p6 <- ggplot(Wall, aes(x=res_pi_eur, y=res_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)
```



Since the patterns remain, I conclude that the depth of coverage biases potentially introduced by the reference bias is not an artifactual cause of the observed relationships. I can't think of any other artifact. Let's look for a biological explanation.

```
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=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
```

```
##
## other attached packages:
## [1] plyr_1.8.6    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    farver_2.0.3     purrr_0.3.3
## [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
```