Basic stats about teosinte data

Samantha Melissa Pacheco Gómez

2026-02-26

Introducción

Vamos a analizar lo más básico de la metadata de teosintes del artículo "Genomic diversity and population structure of teosinte (Zea spp.) and its conservation implications" (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0291944), descargados de: https://doi.org/10.5061/dryad.2547d7wxp

```
Data =
↪  read.csv2("/home/sam/Documents/sur_ecoevo_lab/data/teosinte/archivos/data_teosinte.csv",
   header = TRUE, sep = ",", as.is = FALSE)  #cargar datos
```

summary stats

```
str(Data)
```

```
## 'data.frame': 3604 obs. of 16 variables: ## $
DNASample_code: Factor w/ 3603 levels "1_11","1_12",..: 208
209 210 207 211 212 213 214 215 216 ... ## $ Library_plate
: Factor w/ 44 levels "P_Teo_01_b","P_Teo_03_b",..: 39 26
26 26 26 26 26 26 26 ... ## $ POB_CODE : Factor w/ 74
levels "AMEC","BAPX",..: 68 68 68 68 68 68 68 68 68 68 ...
## $ POB_NUMBER : int 66 66 66 66 66 66 66 66 66 66 ... ##
$ Accession : Factor w/ 276 levels
"CIM10003","CIM11083",..: 110 110 110 110 110 110 110 110
110 110 ... ## $ Race : Factor w/ 6 levels
"","Balsas","Chalco",..: 1 1 1 1 1 1 1 1 1 1 ... ## $ Taxon
: Factor w/ 7 levels "Zea diploperennis",..: 1 1 1 1 1 1 1
1 1 1 ... ## $ Locality : Factor w/ 272 levels
"1_5_km_S_El_Limón",..: 207 207 207 207 207 207 207 207 207
207 ... ## $ Municipality : Factor w/ 129 levels
"Agua_Blanca",..: 36 36 36 36 36 36 36 36 36 36 ... ## $
State : Factor w/ 19 levels "Chihuahua","Chinandega",..: 16
16 16 16 16 16 16 16 16 16 ... ## $ Country : Factor w/ 3
levels "Guatemala","México",..: 2 2 2 2 2 2 2 2 2 2 ... ##
$ Altitude : int 1393 1393 1393 1393 1393 1393 1393 1393
1393 1393 ... ## $ Latitude : Factor w/ 274 levels
"12.89","12.89583333",..: 266 266 266 266 266 266 266 266
266 266 ... ## $ Longitude : Factor w/ 275 levels
"-100.0061111",..: 170 170 170 170 170 170 170 170 170 170
```

```
... ## $ Sampling_date : int 2010 2010 2010 2010 2010 2010
2010 2010 2010 2010 ... ## $ Sample_name : Factor w/ 3604
levels "AMEC_289_1","AMEC_289_10",..: 3323 3324 3325 3322
3326 3327 3328 3329 3330 3331 ...
```

```
summary(Data)
```

```
##  DNASample_code              Library_plate     POB_CODE
##  57_13  :   2   P_Teo2_Ex16_lib_8118:  93   CHPU   : 124
##  1_11   :   1   P_Teo2_Ex11_lib_8007:  92   BGUA   : 122
##  1_12   :   1   P_Teo3_21_lib_8116  :  92   BZAC   : 122
##  1_13   :   1   P_Teo_07_b          :  91   BVPU   : 117
##  1_14   :   1   P_Teo_08_b          :  91   AMEC   : 115
##  1_15   :   1   P_Teo_10_b          :  91   BTEJ   : 105
##  (Other):3597   (Other)             :3054   (Other):2899
##    POB_NUMBER          Accession              Race
##  Min.   : 1.00   JLHNM-661   :  30              : 386
##  1st Qu.:16.00   JSG-JMHC-626:  30   Balsas     :1793
##  Median :35.00   RMM-VHRO-241:  29   Chalco     : 722
##  Mean   :34.07   CIM27480    :  28   Durango    :  87
##  3rd Qu.:50.00   JSG-JMHC-632:  28   Mesa Central: 563
##  Max.   :74.00   JSG-LCL-565 :  28   Nobogame   :  53
##                  (Other)     :3431
##                             Taxon
##  Zea diploperennis              : 136
##  Zea luxurians                  : 111
##  Zea mays ssp. huehuetenangensis:  38
##  Zea mays ssp. mexicana         :1425
##  Zea mays ssp. parviglumis      :1793
##  Zea nicaraguensis              :  20
##  Zea perennis                   :  81
##                                      Locality
##  San_Nicolás_Buenos_Aires                :  30
##  Zacatlancillo_K80_Carr_51_cerro_El_Chivo:  30
##  Arroyo_Tambor                           :  29
##  El_Tepopote_Huista_La_Cofradía          :  29
##  Calabacillas_Potrero_Michoacanes_Tuitán :  28
##  El_Estanco                              :  28
##  (Other)                                 :3430
##              Municipality        State
##  Teloloapan        : 123   Michoacán:1067
##  Villa_Purificación: 117   México   : 809
##  Tejupilco         :  96   Jalisco  : 516
##  Guachinango       :  95   Guerrero : 458
##  Huajicori         :  91   Puebla   : 153
##  Tzitzio           :  89   Nayarit  : 106
##  (Other)           :2993   (Other)  : 495
##       Country        Altitude          Latitude
##  Guatemala: 120   Min.   :   9   19.63527778:  40
##  México   :3464   1st Qu.:1000   18.41666667:  30
##  Nicaragua:  20   Median :1437   19.17138889:  30
##                   Mean   :1518   19.75861111:  29
##                   3rd Qu.:1951   16.32388889:  28
```

```
##                   Max.   :2770   19.2175    :  28
##                                  (Other)    :3419
##         Longitude    Sampling_date       Sample_name
##  -97.55722222:  30   Min.   :1968   AMEC_289_1 :   1
##  -99.9675    :  30   1st Qu.:2005   AMEC_289_10:   1
##  -104.85     :  29   Median :2007   AMEC_289_11:   1
##  -101.3172222:  28   Mean   :2006   AMEC_289_12:   1
##  -101.9      :  28   3rd Qu.:2007   AMEC_289_13:   1
##  -103.5797222:  28   Max.   :2013   AMEC_289_14:   1
##  (Other)     :3431   NA's   :13     (Other)    :3598
```

```
colnames(Data)
```

```
##  [1] "DNASample_code" "Library_plate"  "POB_CODE"
##  [4] "POB_NUMBER"     "Accession"      "Race"
##  [7] "Taxon"          "Locality"       "Municipality"
## [10] "State"          "Country"        "Altitude"
## [13] "Latitude"       "Longitude"      "Sampling_date"
## [16] "Sample_name"
```

Preguntas básicas

¿Cuántas variantes por población?

Primero, vemos cuántos individuos hay por cada población solo por curiosidad.

```
# individuos por población (POB_CODE)
ind_por_pob <- sort(table(Data$POB_CODE), decreasing = TRUE)
ind_por_pob
```

```
##
## CHPU BGUA BZAC BVPU AMEC BTEJ CHAP CHAL BEJU ZDNA BTZI CHUR
##  124  122  122  117  115  105  100   97   94   91   89   82
## BTAR TOLU BPCH BHUE CHTX MCUI BNOC BMAN BIXC BCOL BMAZ BOTZ
##   79   78   76   74   72   69   68   64   57   55   54   54
## NOBO MORO BTAC YURI INDA BCAR MZAM ZDJA DURA BJUA BOAX NDUR
##   53   51   50   50   49   46   45   45   44   43   43   43
## ZLJA ZPMI BTAL ZPJA BTCT BTEL BTUZ HUEH BHUI BZUL CHDF TARI
##   42   41   40   40   39   39   39   38   37   37   37   34
## CHMI SJER ZLOX BAPX MPUR BOLI CHGO ZLAB PENJ MORE TLAX HUAN
##   32   30   29   28   28   27   27   27   26   24   24   23
## BVBR MVIJ NICA BTIQ MALI PONC BMOR BQUE MDOB SNRA ZLJU BSAU
##   20   20   20   17   15   15   14   14   14   14   13   12
## COJU BRED
##    5    3
```

```
length(ind_por_pob)  # número de poblaciones (en la metadata son 74 niveles)
```

```
## [1] 74
```

Para ver las variantes por población, primero unimos la metadata con los genotipos para saber a qué población pertenece cada ID

```
fam <-
↪  read.table("/home/sam/Documents/sur_ecoevo_lab/data/teosinte/archivos/T3604_33929_all.fam",
    header = FALSE, stringsAsFactors = FALSE)
colnames(fam) <- c("FID", "IID", "PID", "MID", "SEX", "PHENO")

# unir con metadata: IID == Sample_name
meta_fam <- merge(fam, Data, by.x = "IID", by.y = "Sample_name",
    all.x = TRUE)

# checar cuántos match hubo
sum(!is.na(meta_fam$POB_CODE))
```

```
## [1] 3604
```

Ahora hay que contar SNPs polimórficos por población con genotipos

```
library(SNPRelate, lib.loc = "/home/sam/R/x86_64-pc-linux-gnu-library/4.5")

snpgdsBED2GDS(bed.fn =
↪  "/home/sam/Documents/sur_ecoevo_lab/data/teosinte/archivos/T3604_33929_all.bed",
    bim.fn =
    ↪  "/home/sam/Documents/sur_ecoevo_lab/data/teosinte/archivos/T3604_33929_all.bim",
    fam.fn =
    ↪  "/home/sam/Documents/sur_ecoevo_lab/data/teosinte/archivos/T3604_33929_all.fam",
    out.gdsfn = "teosinte.gds")

genofile <- snpgdsOpen("teosinte.gds")
```

```
# vector población por individuo en el mismo orden del GDS
# (usa meta_fam, pero asegurándote que esté en el mismo
# orden que sample.id)
samp <- read.gdsn(index.gdsn(genofile, "sample.id"))
pop <- meta_fam$POB_CODE[match(samp, meta_fam$IID)]

# contar SNPs polimórficos por población
pops <- sort(unique(pop[!is.na(pop)]))

res <- data.frame(POB_CODE = pops, n_poly = NA, n_snps = NA,
    n_ind = NA)

for (i in seq_along(pops)) {
    p <- pops[i]
    idx <- which(pop == p)
    af <- snpgdsSNPRateFreq(genofile, sample.id = samp[idx])
    # af$MinorFreq = MAF por SNP dentro de la población
    res$n_poly[i] <- sum(af$MinorFreq > 0, na.rm = TRUE)
    res$n_snps[i] <- length(af$MinorFreq)
    res$n_ind[i] <- length(idx)
}

res[order(res$n_poly, decreasing = TRUE), ]
```

```
##    POB_CODE n_poly n_snps n_ind
## 33     BZAC  33580  33929   122
## 26     BTEJ  33383  33929   105
## 30     BTZI  33254  33929    89
## 9      BIXC  33134  33929    57
## 18     BPCH  33066  33929    76
## 17     BOTZ  32937  33929    54
## 10     BJUA  32647  33929    43
## 36     CHAP  32573  33929   100
## 14     BNOC  32563  33929    68
## 34     BZUL  32536  33929    37
## 4      BCOL  32478  33929    55
## 6      BGUA  32384  33929   122
## 1      AMEC  32273  33929   115
## 27     BTEL  32270  33929    39
## 23     BTAL  32201  33929    40
## 12     BMAZ  32146  33929    54
## 66     YURI  32140  33929    50
## 7      BHUE  32108  33929    74
## 35     CHAL  32088  33929    97
## 52     MORO  31956  33929    51
## 65     TOLU  31805  33929    78
## 22     BTAC  31790  33929    50
## 49     MCUI  31658  33929    69
## 40     CHPU  31647  33929   124
## 3      BCAR  31603  33929    46
## 31     BVBR  31265  33929    20
## 47     INDA  31054  33929    49
## 29     BTUZ  30942  33929    39
## 16     BOLI  30895  33929    27
## 41     CHTX  30841  33929    72
## 42     CHUR  30804  33929    82
## 55     MZAM  30724  33929    45
## 63     TARI  30400  33929    34
## 2      BAPX  29920  33929    28
## 5      BEJU  29733  33929    94
## 39     CHMI  29643  33929    32
## 11     BMAN  29378  33929    64
## 53     MPUR  29346  33929    28
## 24     BTAR  29173  33929    79
## 28     BTIQ  29102  33929    17
## 8      BHUI  29031  33929    37
## 37     CHDF  28689  33929    37
## 51     MORE  28053  33929    24
## 44     DURA  27909  33929    44
## 58     NOBO  27876  33929    53
## 38     CHGO  27848  33929    27
## 59     PENJ  27378  33929    26
## 45     HUAN  27301  33929    23
## 25     BTCT  27048  33929    39
## 54     MVIJ  26679  33929    20
## 61     SJER  25347  33929    30
## 64     TLAX  24873  33929    24
## 32     BVPU  24351  33929   117
```

```
## 50     MDOB  23299  33929    14
## 13     BMOR  22997  33929    14
## 15     BOAX  22752  33929    43
## 48     MALI  22326  33929    15
## 60     PONC  22196  33929    15
## 62     SNRA  21586  33929    14
## 46     HUEH  21462  33929    38
## 56     NDUR  19948  33929    43
## 19     BQUE  19679  33929    14
## 21     BSAU  16291  33929    12
## 20     BRED  15714  33929     3
## 43     COJU  15007  33929     5
## 74     ZPMI  13875  33929    41
## 73     ZPJA  12947  33929    40
## 67     ZDJA  11647  33929    45
## 69     ZLAB   8346  33929    27
## 70     ZLJA   7912  33929    42
## 68     ZDNA   6636  33929    91
## 71     ZLJU   6613  33929    13
## 57     NICA   5006  33929    20
## 72     ZLOX   3172  33929    29
```

```
snpgdsClose(genofile)
```

¿Geolocalización de cada individuo?

Sí nos proporcionan la ubicación geográfica de cada individuo, a continuación está una clasificación por población.

```
Data$Latitude <- as.numeric(as.character(Data$Latitude))
Data$Longitude <- as.numeric(as.character(Data$Longitude))

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
geo_pop <- Data %>%
    group_by(POB_CODE) %>%
    summarise(n_ind = n(), lat = mean(Latitude, na.rm = TRUE),
        lon = mean(Longitude, na.rm = TRUE), alt_mean = mean(Altitude,
            na.rm = TRUE), lat_min = min(Latitude, na.rm = TRUE),
        lat_max = max(Latitude, na.rm = TRUE), lon_min = min(Longitude,
```

```
            na.rm = TRUE), lon_max = max(Longitude, na.rm = TRUE),
        .groups = "drop") %>%
    arrange(POB_CODE)

geo_pop
```

```
## # A tibble: 74 x 9
##    POB_CODE n_ind   lat    lon alt_mean lat_min lat_max
##    <fct>    <int> <dbl>  <dbl>    <dbl>   <dbl>   <dbl>
##  1 AMEC       115  19.1  -98.8    2491.    19.1    19.2
##  2 BAPX        28  18.2  -99.9    1093.    18.2    18.2
##  3 BCAR        46  18.9 -101.      774.    18.9    19.0
##  4 BCOL        55  17.4  -99.3    1011.    17.4    17.5
##  5 BEJU        94  19.9 -104.     1066.    19.9    20.0
##  6 BGUA       122  20.8 -105.     1095.    20.6    20.8
##  7 BHUE        74  18.8 -101.      631     18.7    18.9
##  8 BHUI        37  18.3  -99.3    1099.    18.3    18.3
##  9 BIXC        57  18.5  -99.8    1712.    18.5    18.5
## 10 BJUA        43  19.3 -100.     1221.    19.2    19.3
## # i 64 more rows
## # i 2 more variables: lon_min <dbl>, lon_max <dbl>
```

Este otro código fue sugerido por chat sobre "poblaciones inconsistentes" y se refiere a poblaciones muy dispersas que podrían estar mal etiquetados o ser una mezcla de poblaciones diferentes, tener un error en la metadata o tener coordenadas incorrectas. Pero yo creo que no, que simplemente no conoce México.

```
geo_pop %>%
    mutate(lat_range = lat_max - lat_min, lon_range = lon_max -
        lon_min) %>%
    arrange(desc(lat_range + lon_range)) %>%
    head(15)
```

```
## # A tibble: 15 x 11
##    POB_CODE n_ind   lat    lon alt_mean lat_min lat_max
##    <fct>    <int> <dbl>  <dbl>    <dbl>   <dbl>   <dbl>
##  1 CHMI        32  19.6 -102.     2168.    19.4    19.7
##  2 BMAZ        54  17.4  -99.4     924.    17.3    17.5
##  3 MZAM        45  20.1 -102.     1592.    20.0    20.2
##  4 BTCT        39  17.0  -99.4     523.    17.0    17.2
##  5 BGUA       122  20.8 -105.     1095.    20.6    20.8
##  6 CHPU       124  19.1  -97.5    2490.    19.0    19.2
##  7 BEJU        94  19.9 -104.     1066.    19.9    20.0
##  8 TOLU        78  19.3  -99.7    2660.    19.3    19.4
##  9 BCOL        55  17.4  -99.3    1011.    17.4    17.5
## 10 BTZI        89  19.5 -101.     1038.    19.3    19.6
## 11 BNOC        68  19.1 -101.      857.    19.0    19.2
## 12 CHTX        72  19.5  -98.9    2271.    19.4    19.5
## 13 CHDF        37  19.2  -99.1    2509.    19.1    19.3
## 14 INDA        49  19.8 -101.     1895.    19.7    19.9
## 15 BTUZ        39  19.1 -101.      680.    19.0    19.2
## # i 4 more variables: lon_min <dbl>, lon_max <dbl>,
## #   lat_range <dbl>, lon_range <dbl>
```

Fecha específica de recolección o gradiente de varias décadas

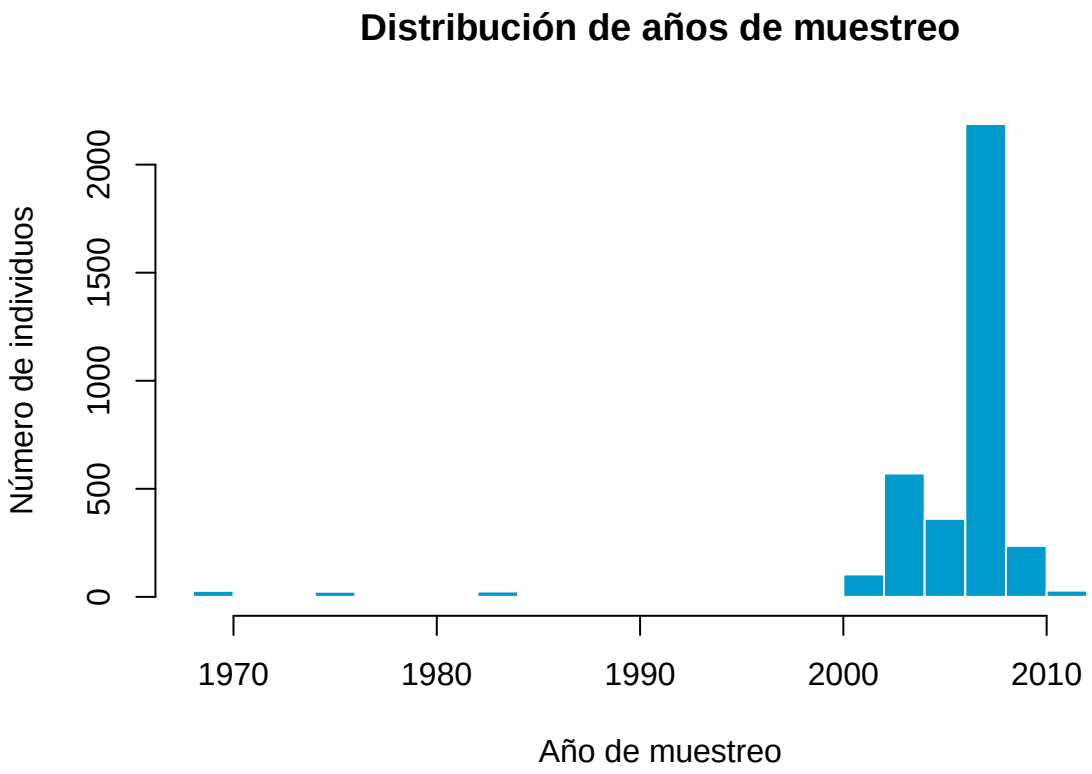
```
table(Data$Sampling_date) |>
    sort(decreasing = TRUE)
```

```
##
## 2007 2005 2004 2003 2008 2009 2010 2002 2012 1976 1968 1969
## 1979  351  309  262  209  119  117  104   29   24   14   14
## 1984 1983 2006 1978 2013 1991
##   13   12   10    9    9    7
```

```
range(Data$Sampling_date, na.rm = TRUE)
```

```
## [1] 1968 2013
```

```
hist(Data$Sampling_date, breaks = 20, col = "deepskyblue3", border = "white",
    main = "Distribución de años de muestreo", xlab = "Año de muestreo",
    ylab = "Número de individuos")
```



¿altitud, temperatura?

No contamos con variables climáticas, habría que checar WorldClim.

Pero en efecto, tenemos altitud, latitud y longitud de cada individuo y podemos verlo en la ubicación geográfica.

¿posición genómica aproximada?

```
bim <-
↪  read.table("/home/sam/Documents/sur_ecoevo_lab/data/teosinte/archivos/T3604_33929_all.bim",
    header = FALSE, stringsAsFactors = FALSE)
colnames(bim) <- c("CHR", "SNP", "CM", "BP", "A1", "A2")

table(bim$CHR)  # SNPs por cromosoma
```

```
##
##    1    2    3    4    5    6    7    8    9   10
## 5486 4151 3935 2997 4028 2722 3080 2896 2414 2220
```

```
summary(bim$BP)  # resumen de posiciones
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##     10232  34595261 137391485 120528242 181126820 301168145
```

```
head(bim)
```

```
##   CHR        SNP CM      BP A1 A2
## 1   1  S1_992727 -9  992727  C  T
## 2   1 S1_1005413 -9 1005413  G  C
## 3   1 S1_1763292 -9 1763292  C  T
## 4   1 S1_1763397 -9 1763397  G  A
## 5   1 S1_1780412 -9 1780412  T  G
## 6   1 S1_1780418 -9 1780418  G  A
```

```
library(dplyr)
bim %>%
    group_by(CHR) %>%
    summarise(n_snps = n(), min_BP = min(BP), max_BP = max(BP),
        .groups = "drop") %>%
    arrange(CHR)
```

```
## # A tibble: 10 x 4
##      CHR n_snps  min_BP    max_BP
##    <int>  <int>   <int>     <int>
##  1     1   5486  992727 301168145
##  2     2   4151   10232 236973204
##  3     3   3935 1261965 232044134
##  4     4   2997   30681 241030367
##  5     5   4028  536169 217608792
##  6     6   2722  272264 169083700
##  7     7   3080   32638 176550961
##  8     8   2896   96346 175698433
##  9     9   2414  318789 156422039
## 10    10   2220  629096 149597102
```

¿ quién es la referencia?

draft ZeaGBSv2.7.

Notas finales

Se utilizó ChatGPT para la realización de estos códigos de análisis de los archivos bed, bim ,fam.

Faltaría revisar y conectar las variables climáticas a cada individuo o población segun su ubicación geográfica y año de recolección.

Duda: ¿es suficiente el año de recolección para asumir cómo estuvo el clima en la recolección partiendo de que en distintas estaciones o momentos climáticos del año se observa algo distinto?

¿Las coordenadas de altitud, longitud y latitud son por población o por individuo?