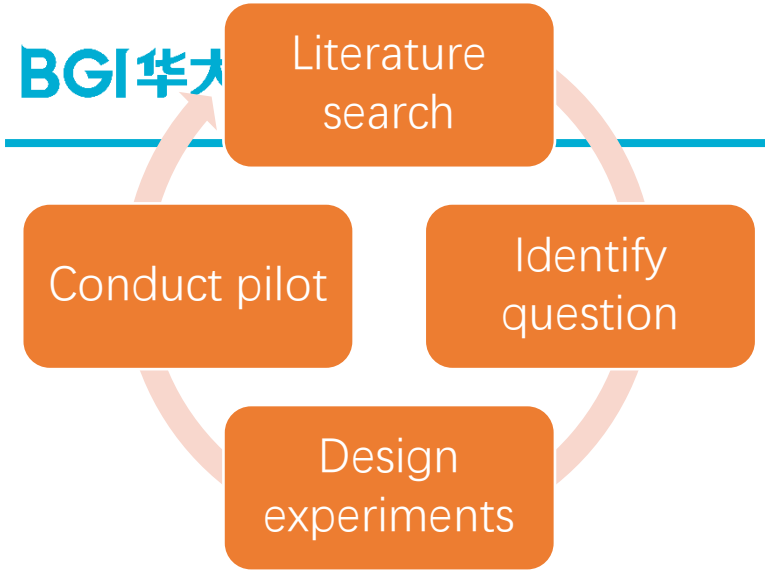
Scientific figure

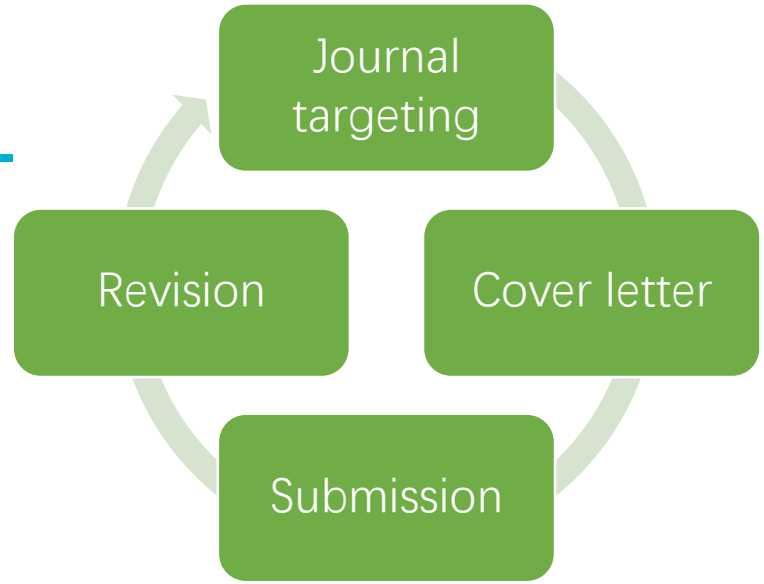
魏桐

3/22/2023

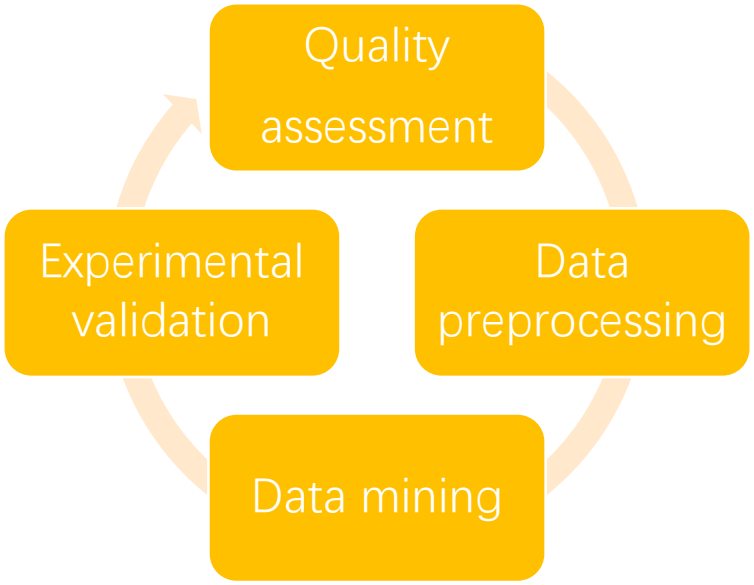
华大生命科学研究院
BGI · research

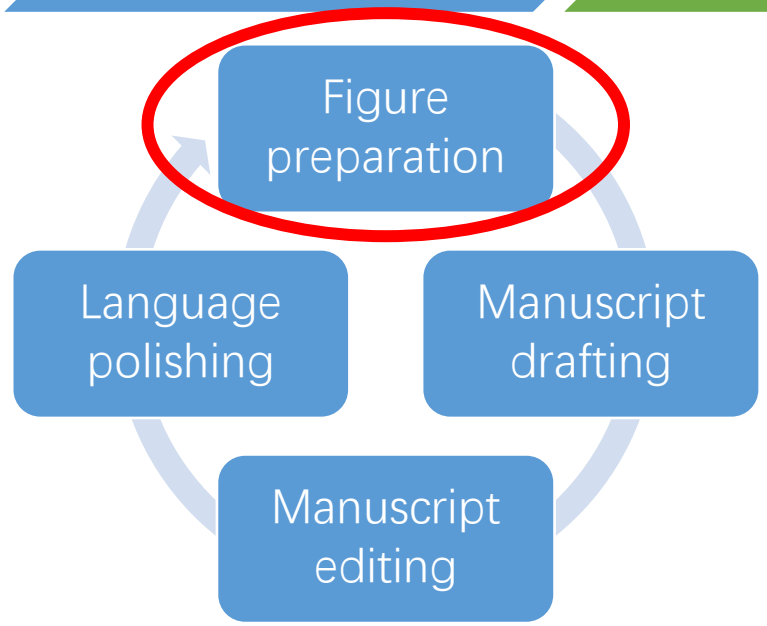
Research cycle





Project design → Data analysis → Manuscript preparation → Submission & revision

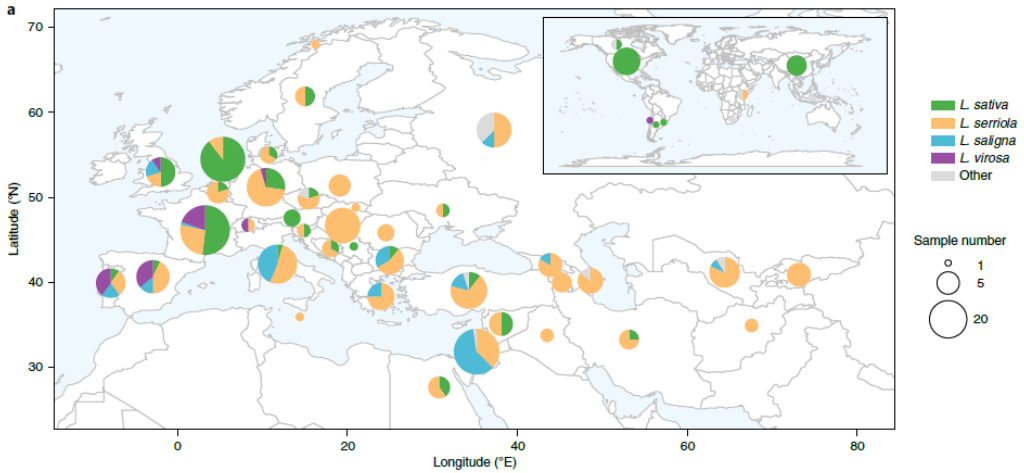




BGI华大

What is scientific figure

Provide **an efficient way to communicate** your research



vs

Plant materials and sequencing. The collection of *Lactuca* SSD lines (http://www.wur.eu/cgnsc002) used in this study comprises a core subset of the regular collection of the Centre for Genetic Resources, the Netherlands (CGN) and includes all crop types of cultivated lettuce and main wild relatives used in plant breeding[32]. The total study set of 445 SSD lines included 131 cultivated lettuce (*L. sativa*) accessions collected worldwide, 201 *L. serriola* accessions, 57 *L. saligna* accessions, 37 *L. virosa* accessions and 19 lines from another eight *Lactuca* species (Supplementary Table 1).

Part I
Plotting by R

华大生命科学研究院
BGI · research

What is R

- R in brief
  - A programming language for statistical computing and graphics
  - Open-source implementation of the S language developed by Ross Ihaka and Robert Gentleman in 1990s
  - A variety of packages from an active community
- Usage
  - Data processing
  - Statistical analysis
  - Scientific figures

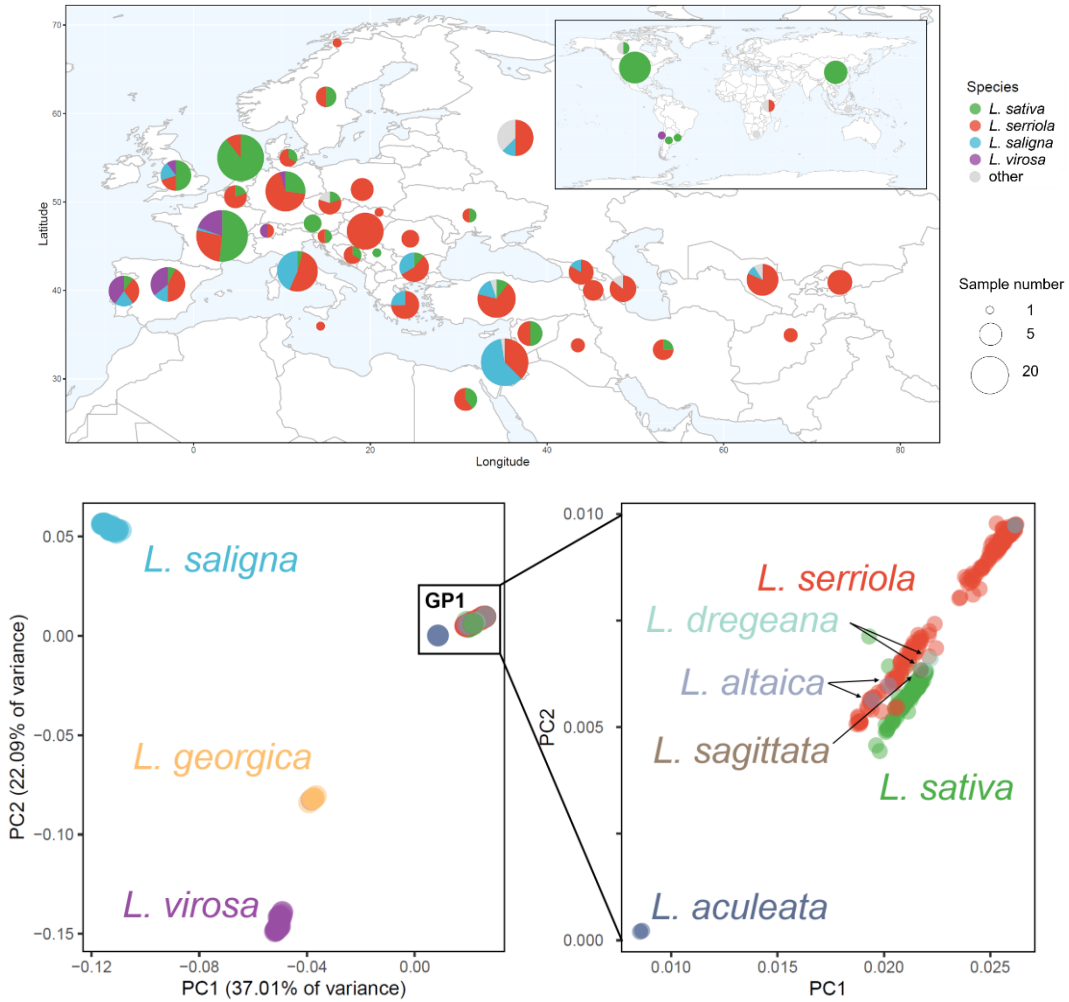
BGI华大

Examples

Statistical analysis

```
> x <- rnorm(10)
> y <- rnorm(10, .5)
> t.test(x, y)

        Welch Two Sample t-test

data:  x and y
t = -0.72454, df = 15.418, p-value = 0.4796
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.1112264  0.5464145
sample estimates:
 mean of x  mean of y
0.02802619 0.31043212

> lm(y ~ x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     0.3013       0.3253
```

Scientific figure



BGI华大

R download

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2018-07-02, Feather Spray) R-3.5.1.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

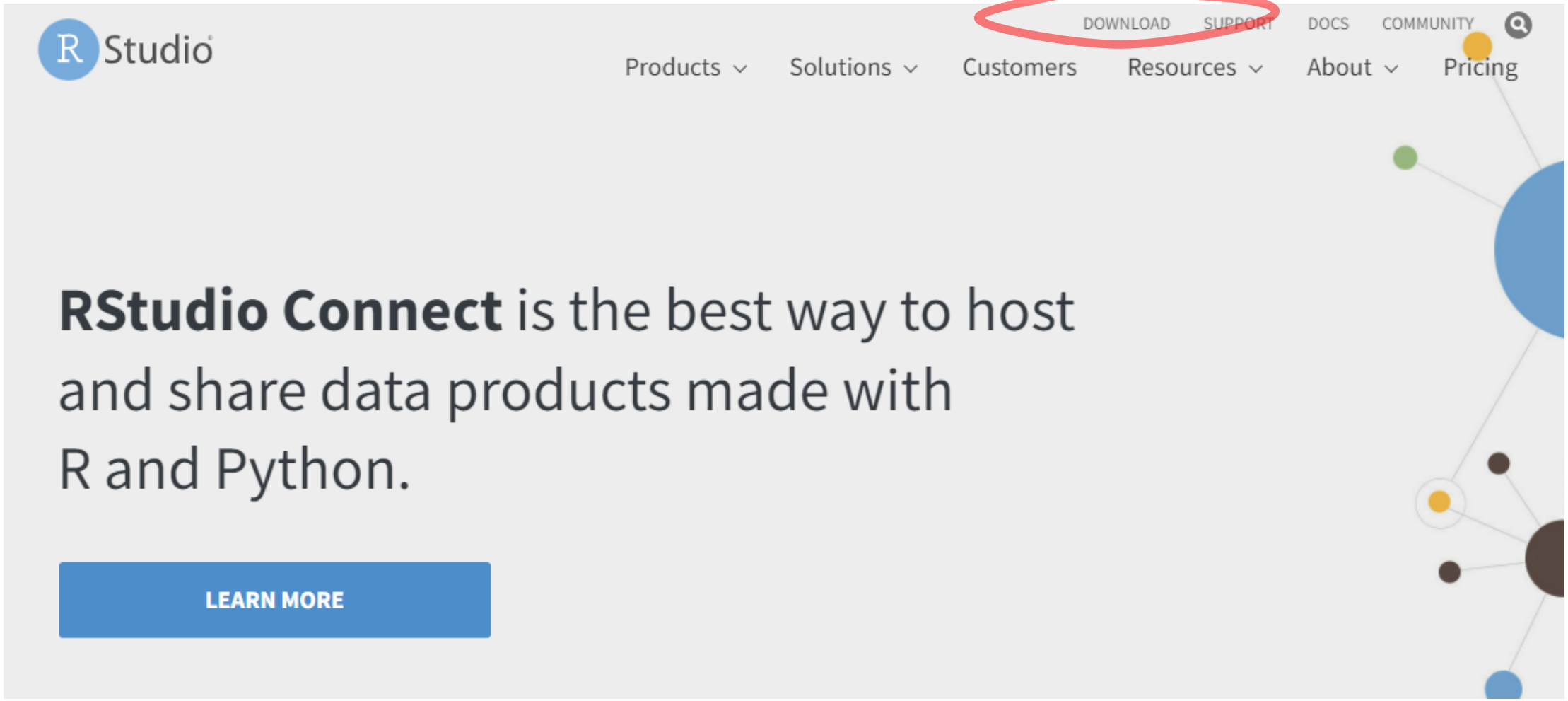
Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

https://cran.r-project.org

BGI华大

R Studio download

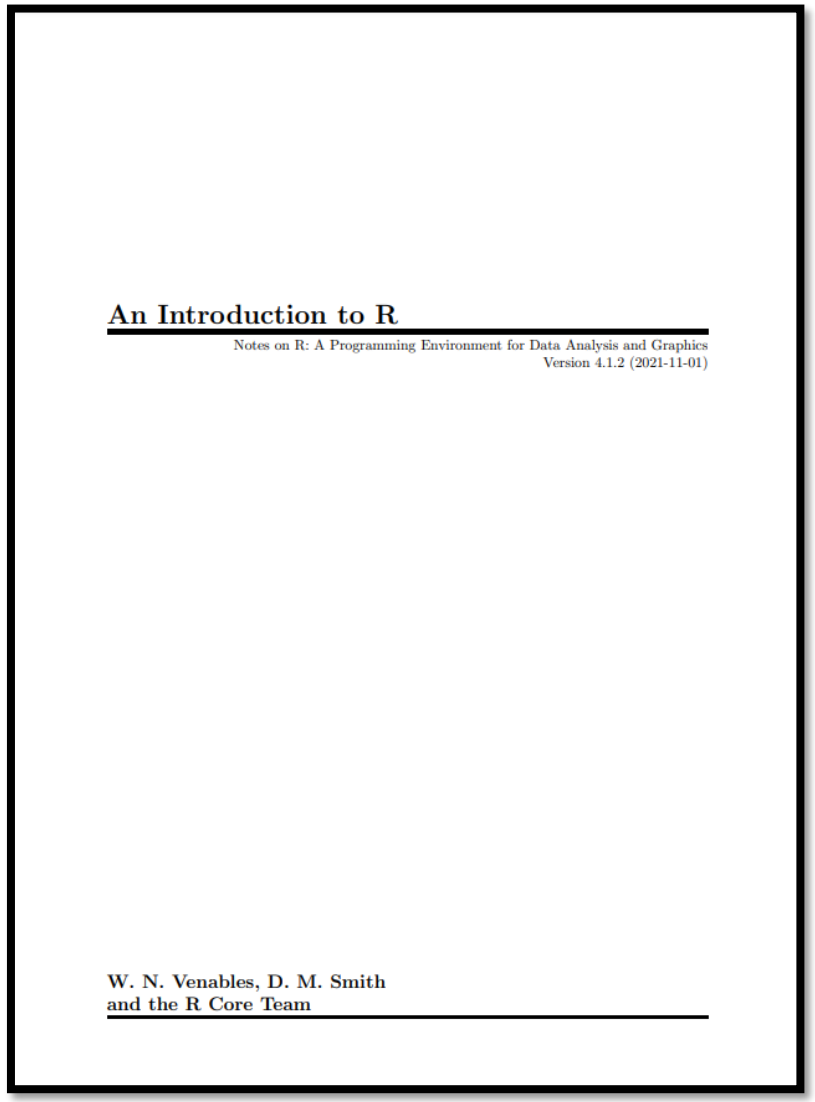


BGI华大

R Studio

BGI华大

Learning material: R intro



An Introduction to R

Notes on R: A Programming Environment for Data Analysis and Graphics
Version 4.1.2 (2021-11-01)

W. N. Venables, D. M. Smith
and the R Core Team

1. Introduction and preliminaries
2. Simple manipulations; numbers and vectors
3. Objects, their modes and attributes
4. Ordered and unordered factors
5. Arrays and matrices
6. Lists and data frames
7. Reading data from files
8. Probability distributions
9. Grouping, loops and conditional execution
10. Writing your own functions
11. Statistical models in R
12. Graphical procedures
13. Packages
14. OS facilities

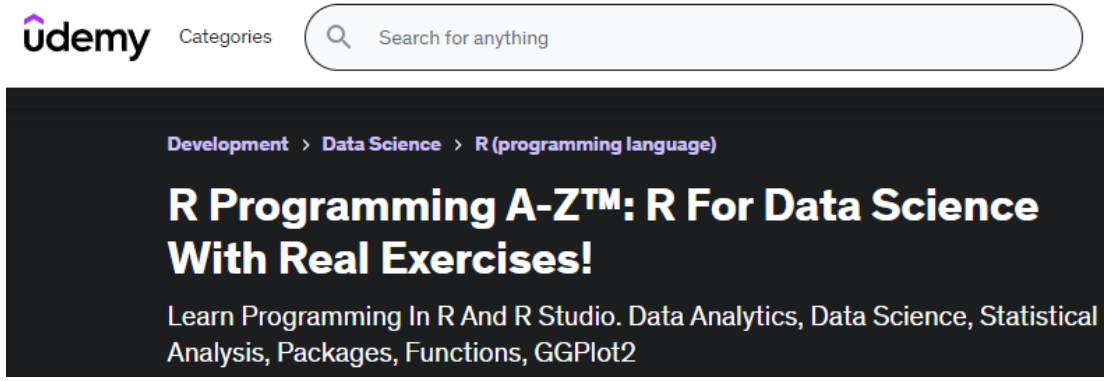
https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

BGI华大

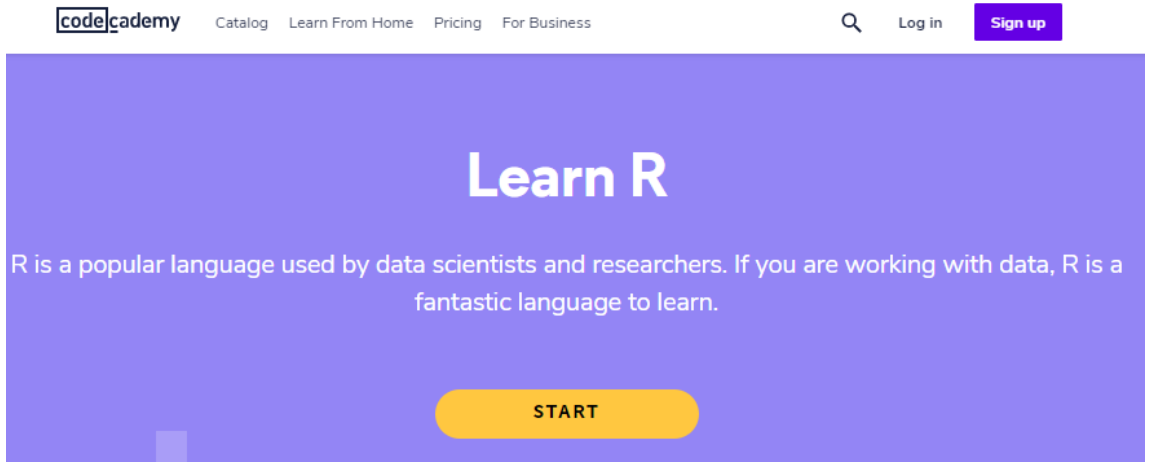
Other online materials

Coursera

Udemy



codecademy



edX

BGI华大

R codes in the script panel

Comment lines

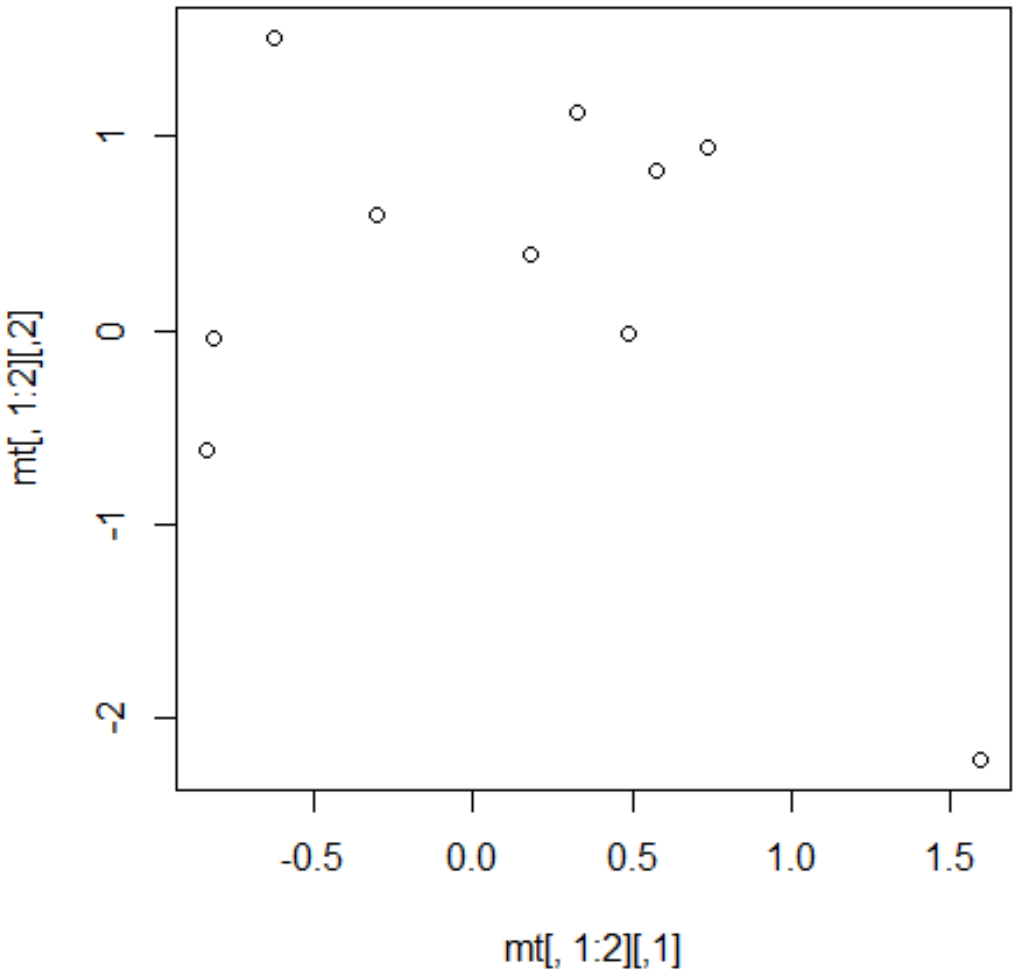
```
# Step 0, read data and metadata
# 0.1, read the count data into R environment
# 0.2, generate a design matrix

dat <- read.table("tophat_counts.txt", header = T, sep = "\t")
rownames(dat) <- dat$Gene_ID
dat <- dat[, -1]
dat.log2 <- log2(dat + 1)
# labeling
sample.ID <- colnames(dat)
group.ID <- sub("_\\d", "", sample.ID)
sample.mat <- sapply(sample.ID, strsplit, split = "_|\\.", simplify = T)
sample.mat <- t(as.data.frame(sample.mat))
colnames(sample.mat) <- c("genotype", "treatment", "time", "replicate")
```

object <- function(arguments)

BGI华大

Function: plot()

```
> set.seed(1)
> mt <- matrix(rnorm(20), ncol = 2)
> mt
             [,1]        [,2]
 [1,] -0.6264538  1.51178117
 [2,]  0.1836433  0.38984324
 [3,] -0.8356286 -0.62124058
 [4,]  1.5952808 -2.21469989
 [5,]  0.3295078  1.12493092
 [6,] -0.8204684 -0.04493361
 [7,]  0.4874291 -0.01619026
 [8,]  0.7383247  0.94383621
 [9,]  0.5757814  0.82122120
[10,] -0.3053884  0.59390132
> plot(mt[, 1:2])
```

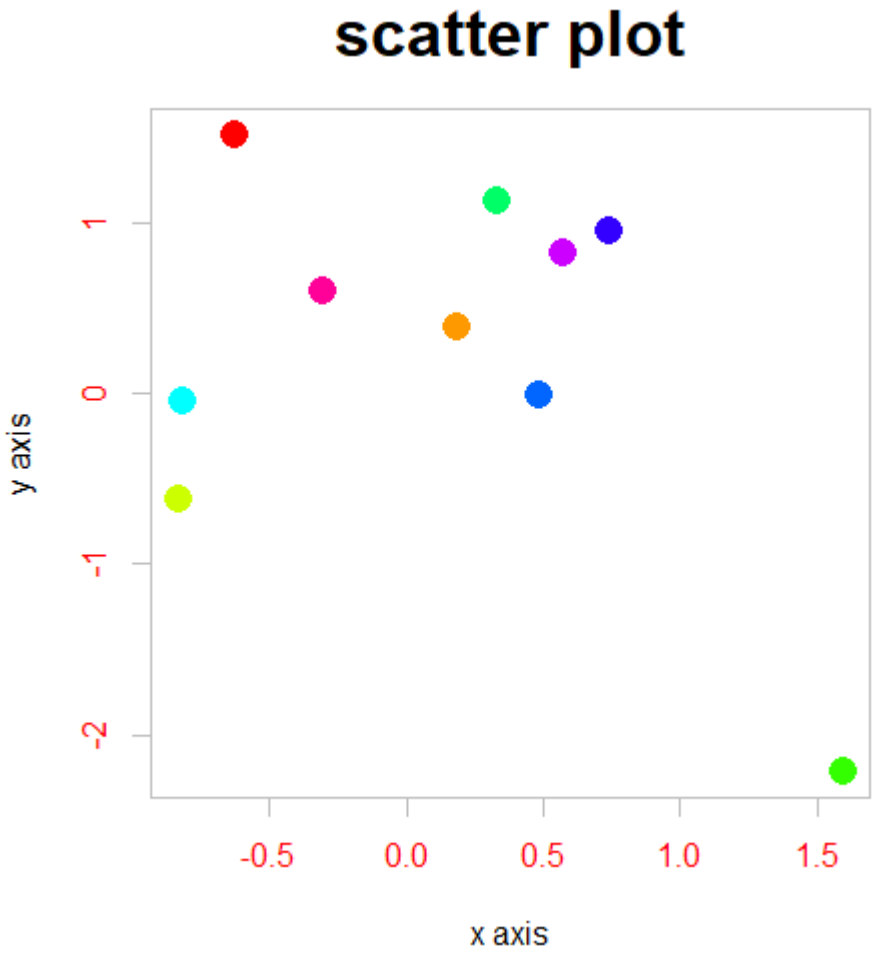


`?graphics::plot` for more information

BGI华大

Function: plot()

```
> plot(mt[, 1:2],
+      pch = 16, cex = 2,
+      col = rainbow(10),
+      xlab = "x axis",
+      ylab = "y axis",
+      main = "scatter plot",
+      cex.main = 2,
+      fg = "grey",
+      col.axis = "red")
```

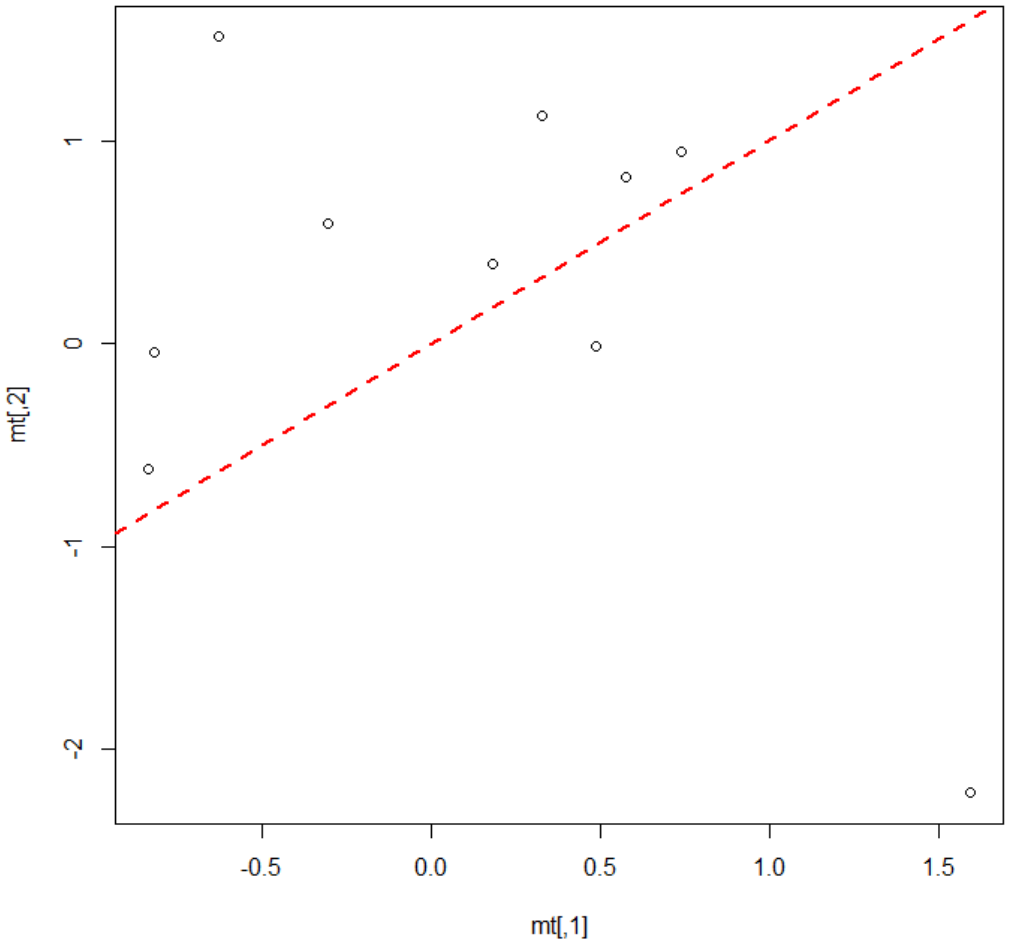


`?graphics::plot` for more information

BGI华大

Function: abline()

```
> plot(mt)
> abline(a = 0, b = 1,
+        col = "red", lwd = 2, lty = 2)
```

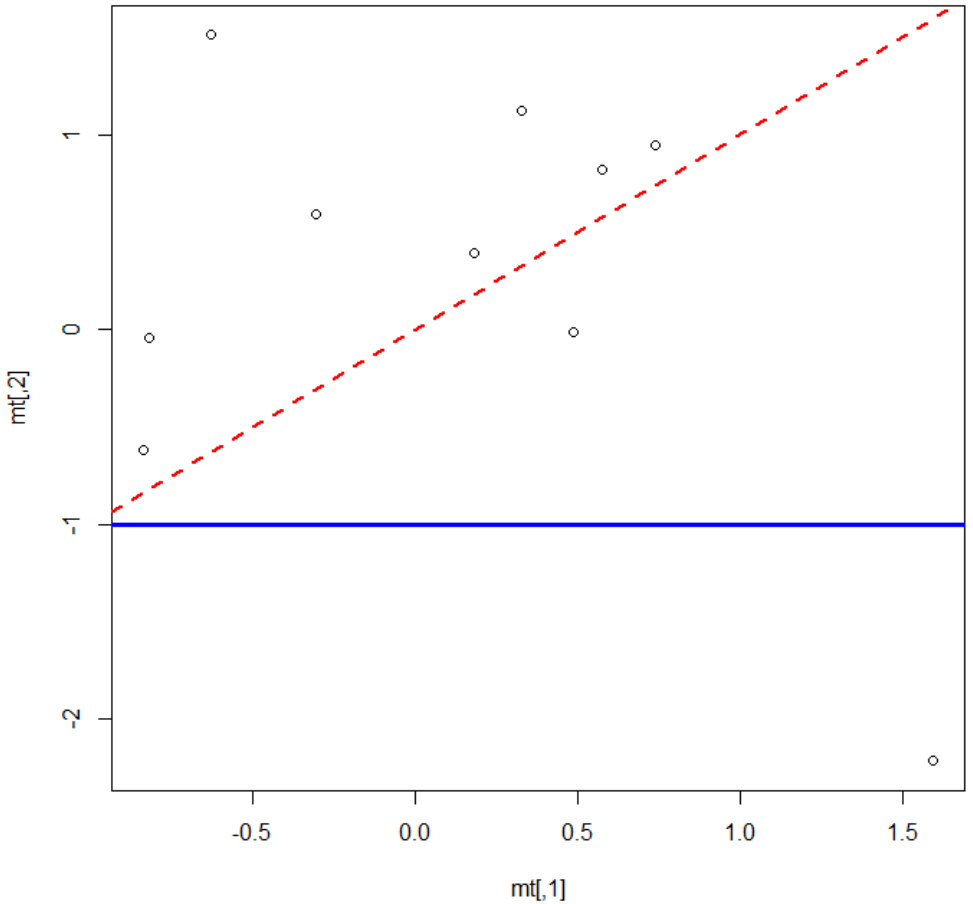


BGI华大

Function: abline()

```
> plot(mt)
> abline(a = 0, b = 1,
+        col = "red", lwd = 2, lty = 2)
```

```
> abline(h = -1, col = "blue", lwd = 3)
```



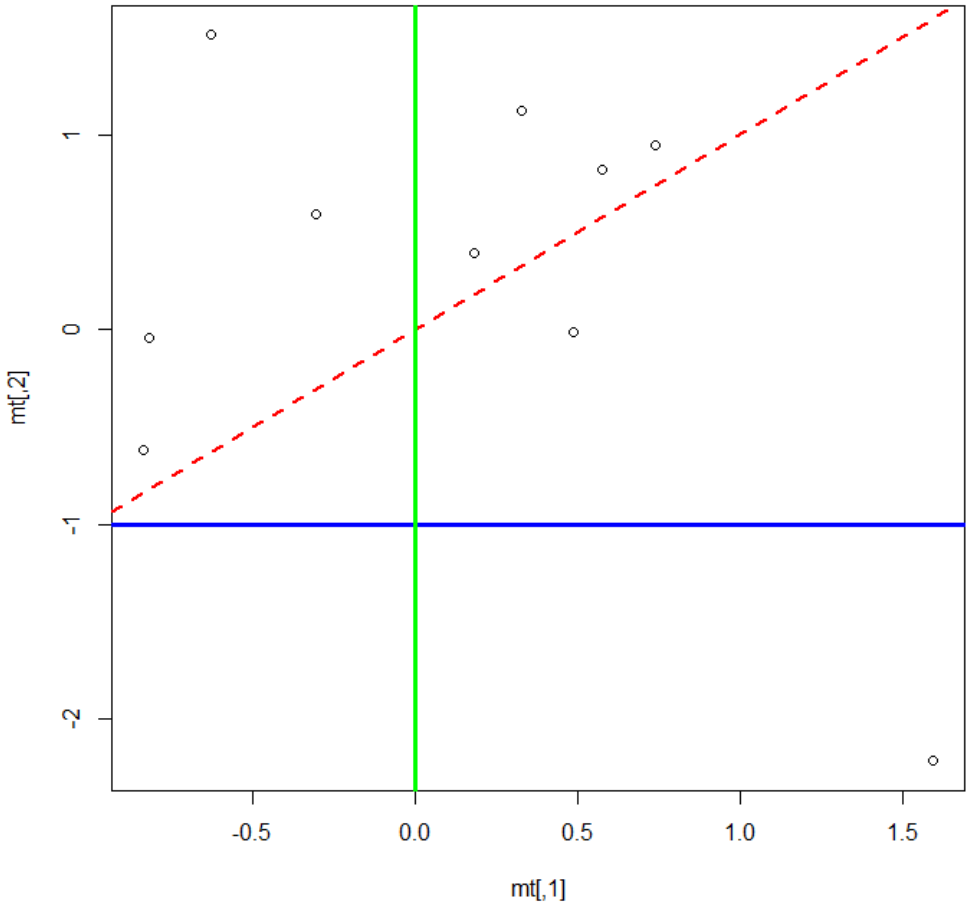
BGI华大

Function: abline()

```
> plot(mt)
> abline(a = 0, b = 1,
+        col = "red", lwd = 2, lty = 2)
```

```
> abline(h = -1, col = "blue", lwd = 3)
```

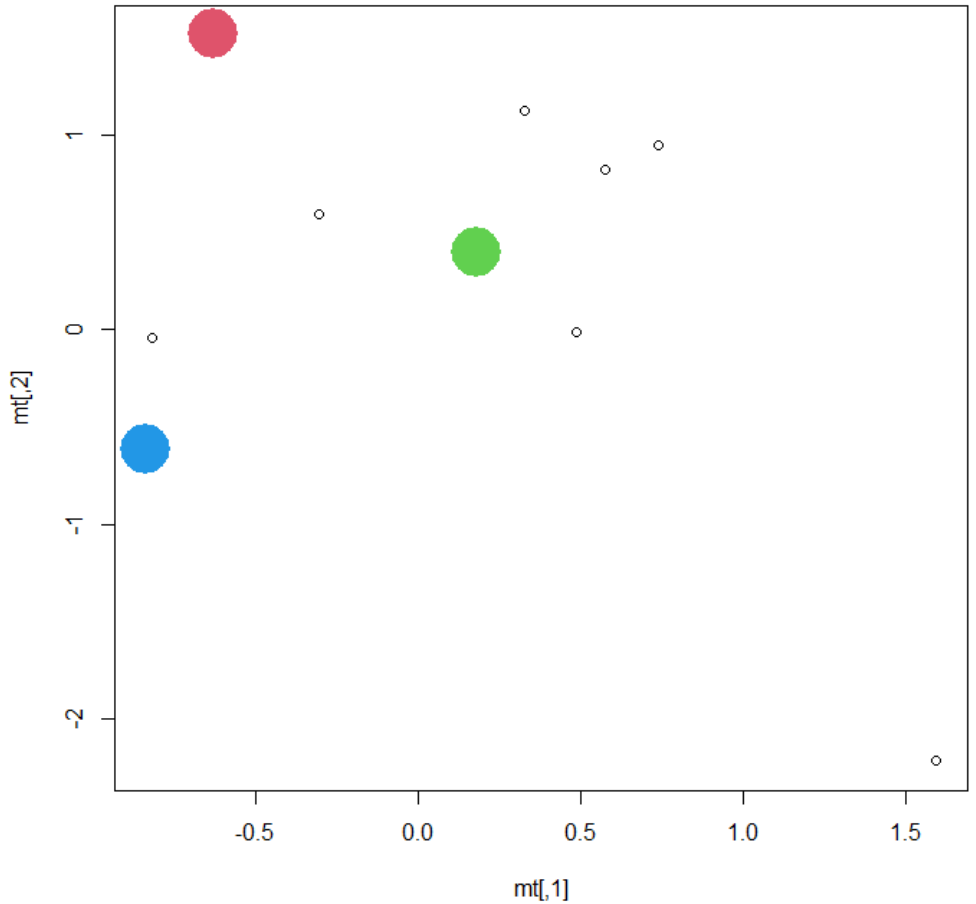
```
> abline(v = 0, col = "green", lwd = 3)
```



BGI华大

Function: points()

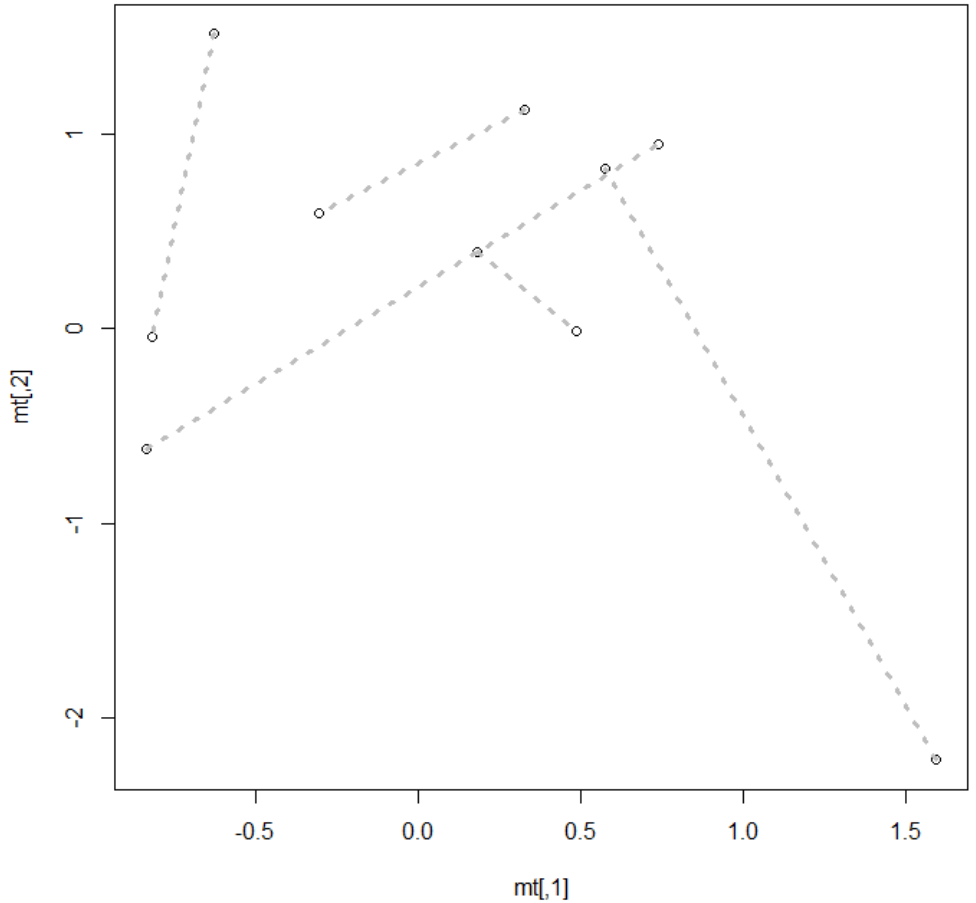
```
> plot(mt)
> points(mt[1:3, ],
+        pch = 16, col = 2:4, cex = 5)
```



BGI华大

Function: segments()

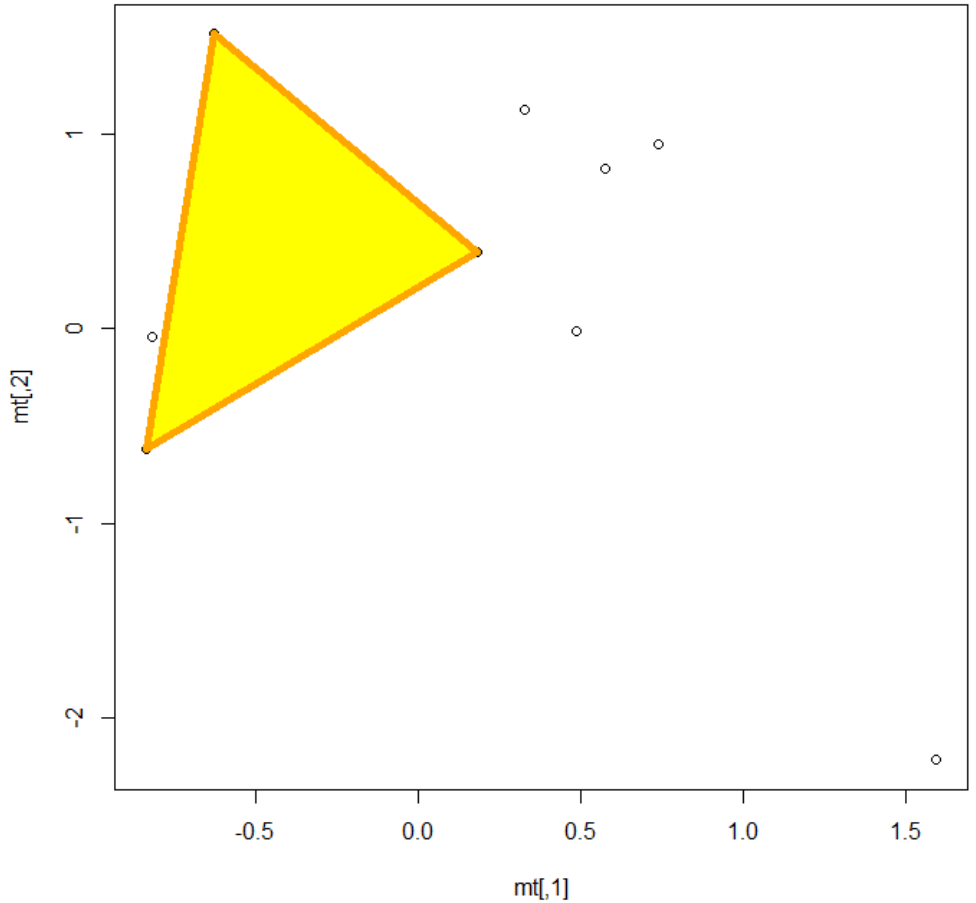
```
> plot(mt)
> segments(x0 = mt[1:5, 1],
+          y0 = mt[1:5, 2],
+          x1 = mt[6:10, 1],
+          y1 = mt[6:10, 2],
+          col = "grey",
+          lwd = 3, lty = 3)
```



BGI华大

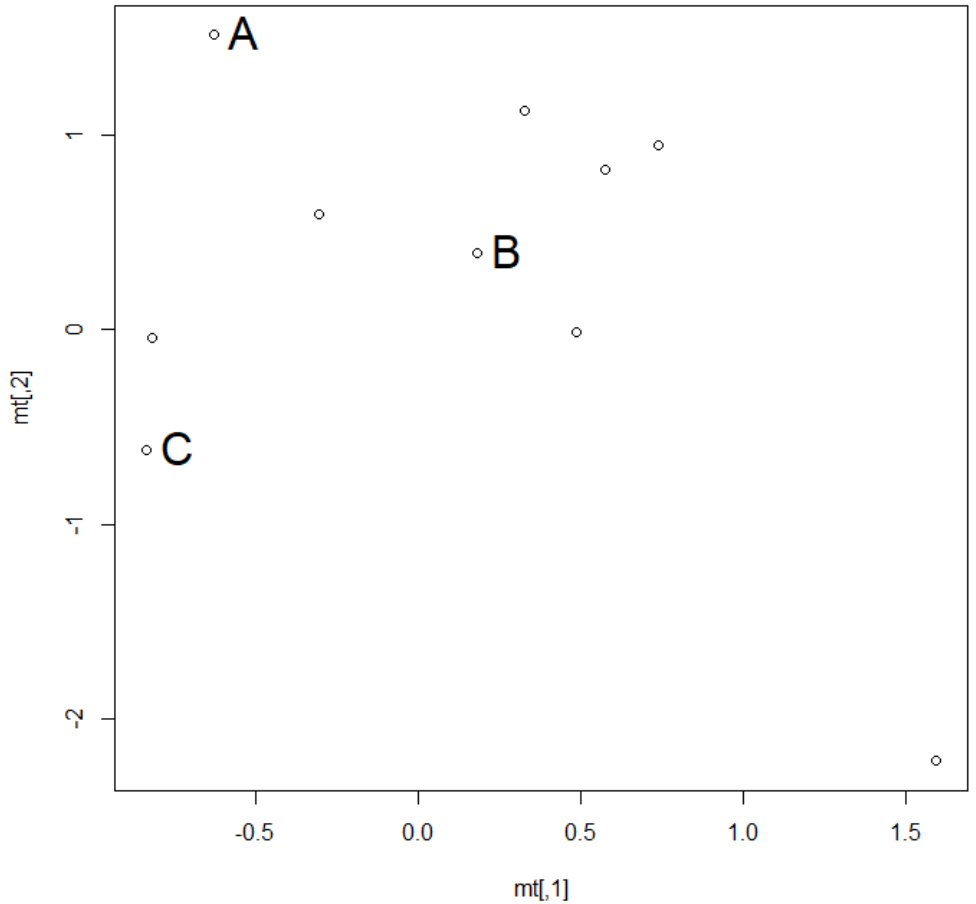
Function: polygon()

```
> plot(mt)
> polygon(mt[1:3, ],
+         col = "yellow",
+         border = "orange",
+         lwd = 5)
```



Function: text()

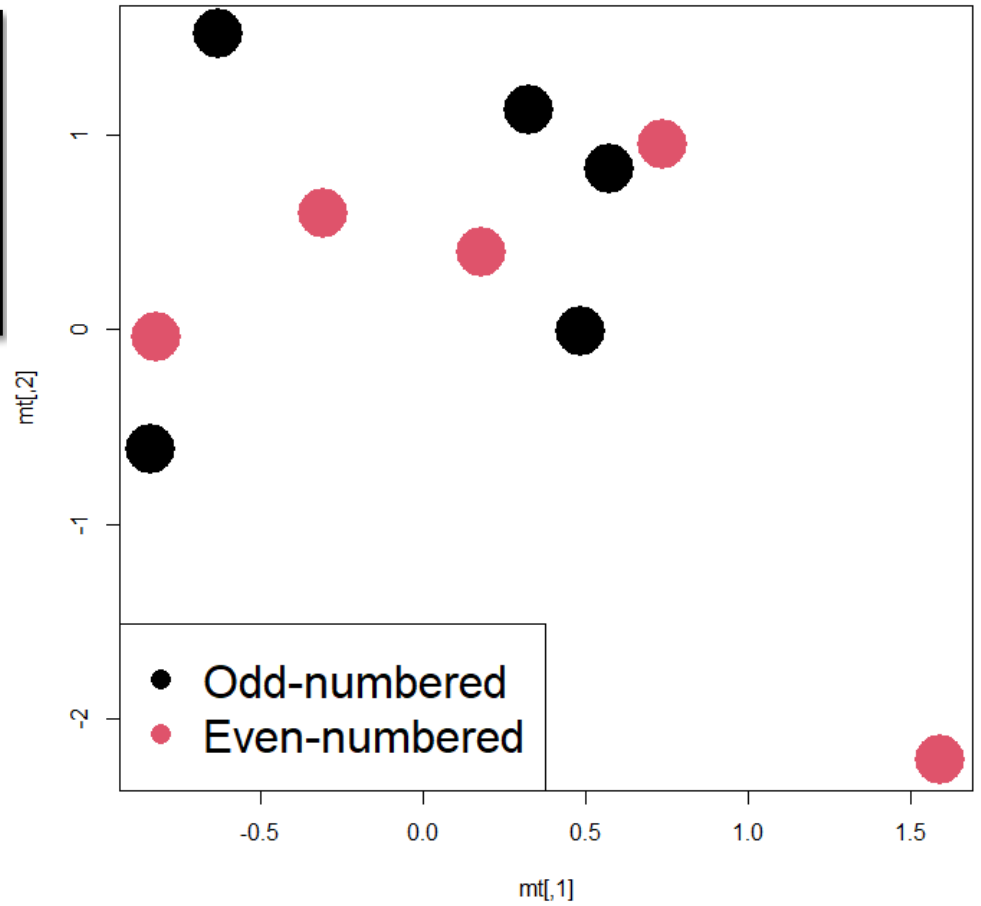
```
> plot(mt)
> text(mt[1:3, ],
+      labels = LETTERS[1:3],
+      cex = 2,
+      pos = 4)
```



BGI华大

Function: legend()

```
> plot(mt, col = 1:2, pch = 16, cex = 5)
> legend("bottomleft",
+        legend = c("Odd-numbered",
+                   "Even-numbered"),
+        col = 1:2, pch = 16, cex = 2)
```

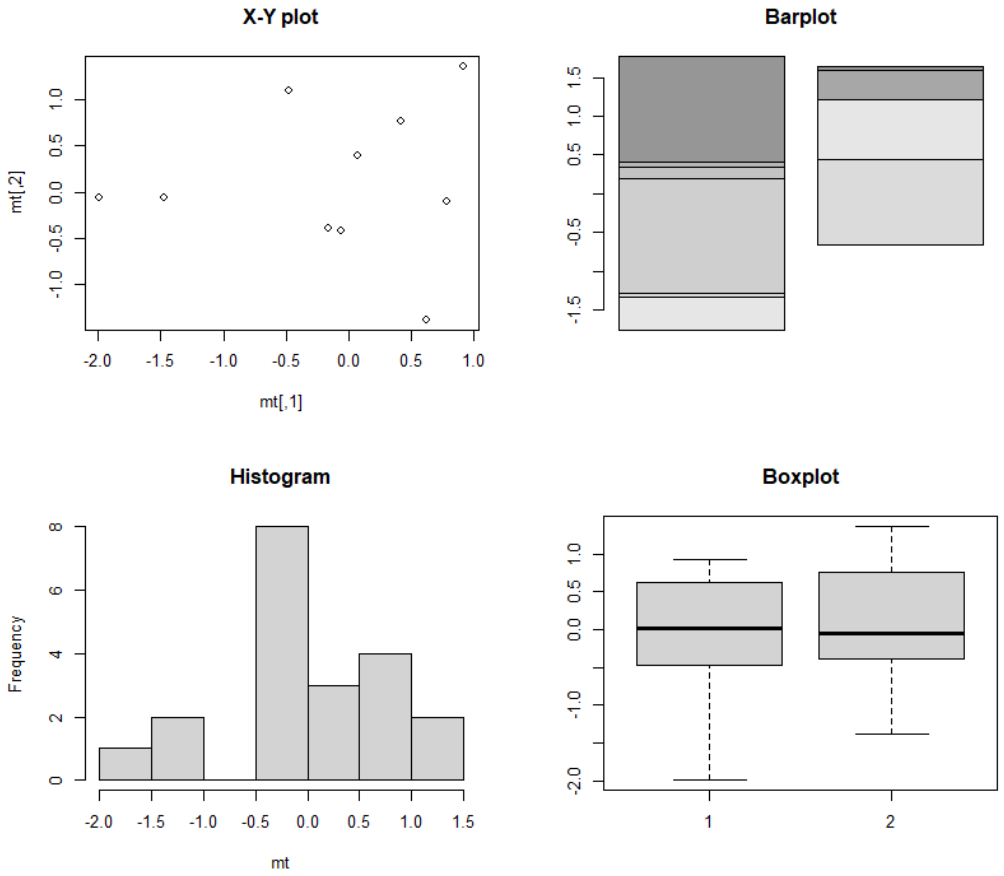


BGI华大

Function: par()

```
> par(mfrow = c(2, 2))
```

```
> plot(mt, main = "X-Y plot")
> barplot(mt, main = "Barplot")
> hist(mt, main = "Histogram")
> boxplot(mt, main = "Boxplot")
```



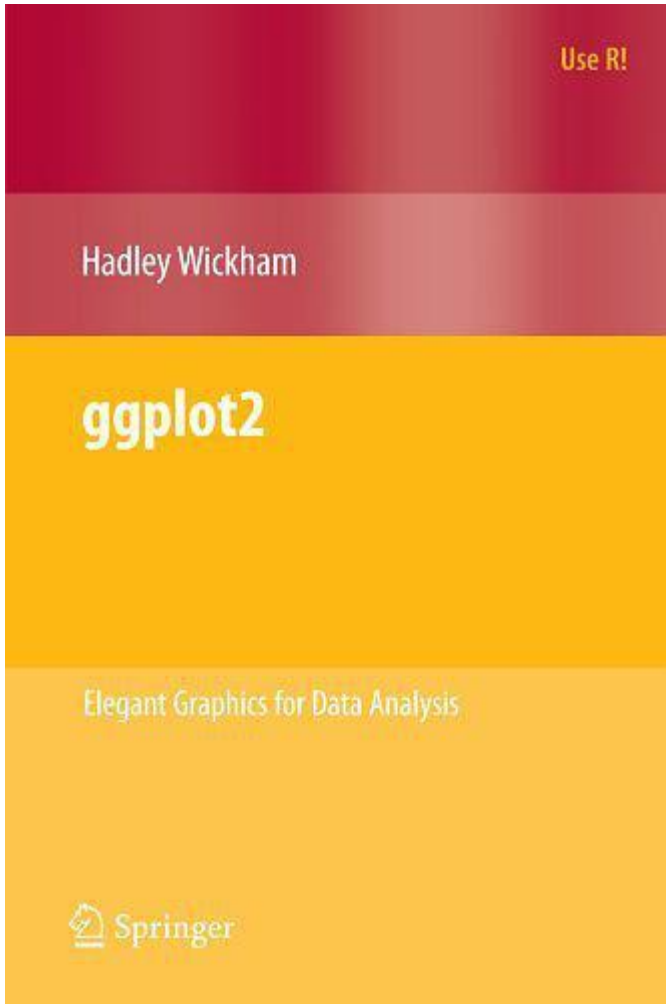
The graphics package

- High-level plotting functions
  - `plot(x, y)`
  - `hist(); barplot(); boxplot()`
- Low-level plotting functions
  - Text: `legend(); title(); text()`
  - Points: `points()`
  - Lines: `lines(); abline()`
  - Shapes: `rect(); polygon()`
- Parameter: `par()`

`help(package = "graphics")` for more information

The gglot2 package

- Layer by layer
- Geometric shapes, functions starting with `geom()`
- Aesthetics including color, arguments `ase()`
- Others: scales, facets, theme, statistical transformation, etc.



R plot types

- Data types: numerical, categorical & ordinal

- 1-D* numerical data: density lines (summary stats)
- 1-D categorical: barplot of counts
- 2-D numerical: scatterplot
- 2-D numerical vs categorical: boxplot/violin
- 2-D categorical: heatmap
- 3-D numerical/categorical: scatterplot w/ other aesthetics
- >3-D numerical/categorical: facet

*D: dimension

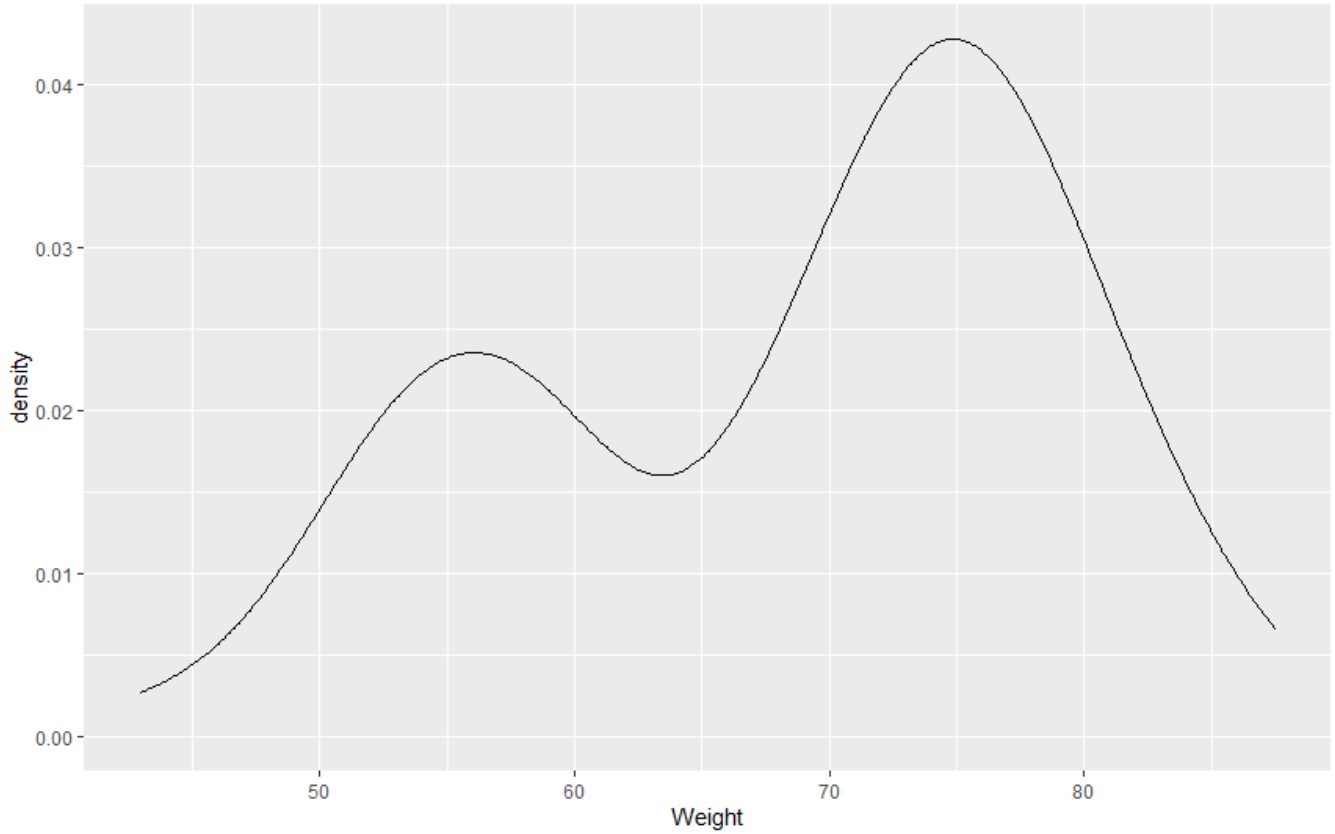
A simulated data frame

```
set.seed(1)
dat <- data.frame(Height = c(rnorm(100, 175, 10),
                             rnorm(50, 165, 10)),
                  Weight = c(rnorm(100, 75, 5),
                             rnorm(50, 55, 5)),
                  Sex = rep(c("Male", "Female"),
                            c(100, 50)),
                  Age = sample(rep(c("Young", "Old"),
                                   each = 75)))
```

```
> head(dat)
    Height   Weight  Sex Age
1 168.7355 77.25094 Male Old
2 176.8364 74.90720 Male Old
3 166.6437 73.40966 Male Old
4 190.9528 70.35319 Male Old
5 178.2951 67.56270 Male Old
6 166.7953 69.62404 Male Old
```

Density lines

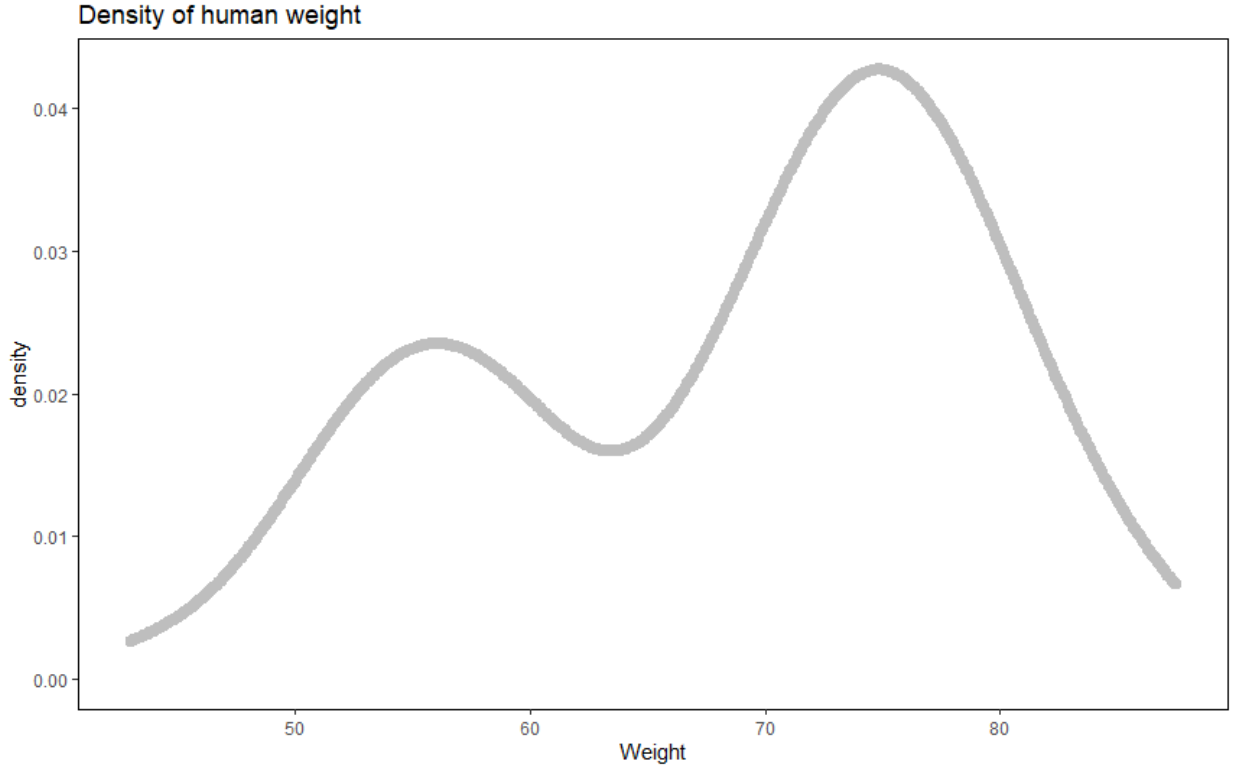
```
library(ggplot2)
g <- ggplot(dat)
g + geom_density(aes(x = Weight)) # density plot
```

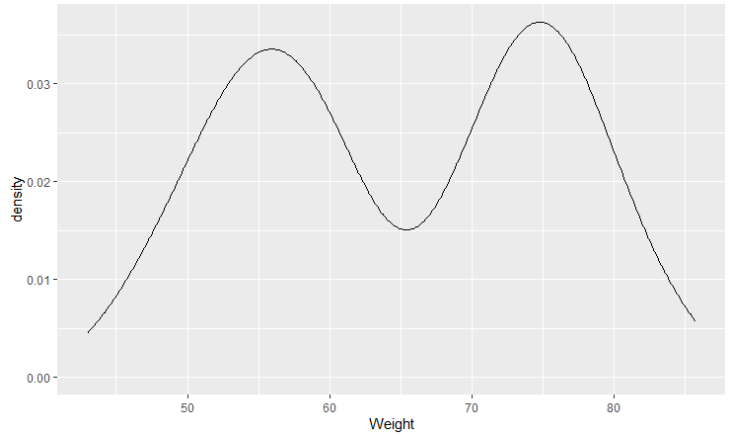


BGI华大

Density lines

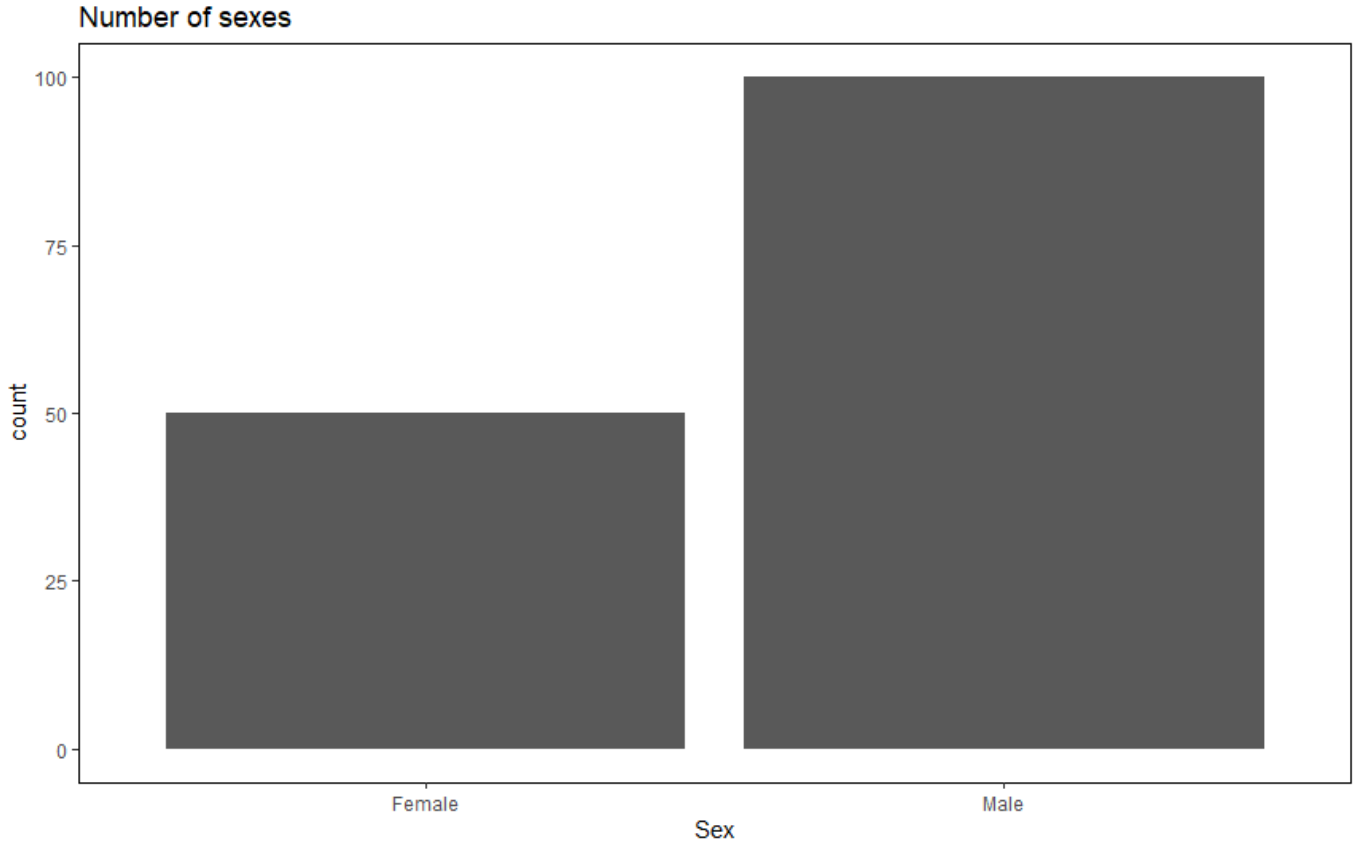
```
g + geom_density(aes(x = Weight),
                 color = "grey", size = 3) +
  labs(title = "Density of human weight") +
  theme(panel.border = element_rect(color = "black", fill = NA),
        panel.background = element rect(fill = "white"))
```





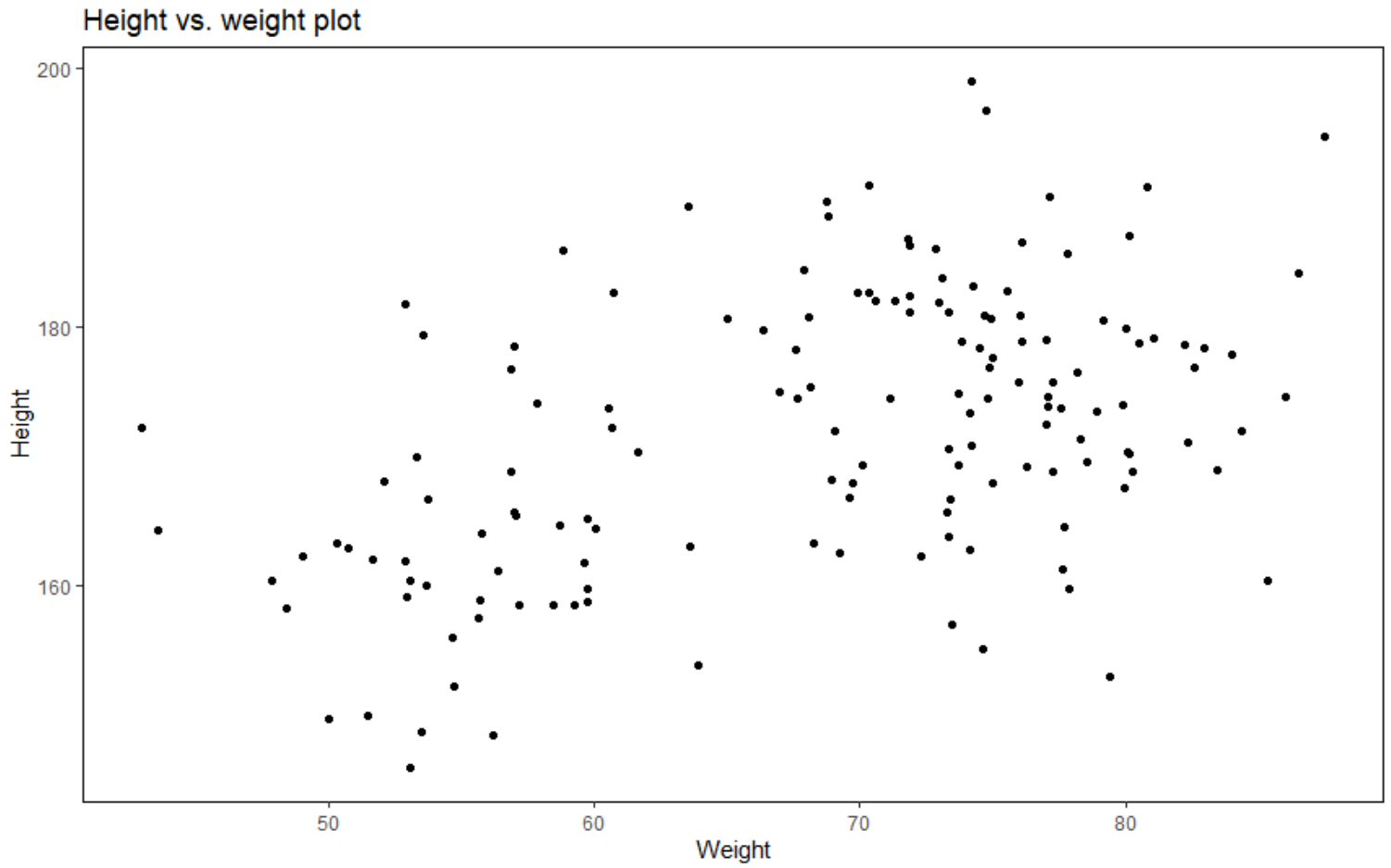
Barplot

```
g <- ggplot(dat) +
  theme(panel.border = element_rect(color = "black", fill = NA),
        panel.background = element_rect(fill = "white"))
g + geom_bar(aes(x = Sex)) + labs(title = "Number of sexes")
```



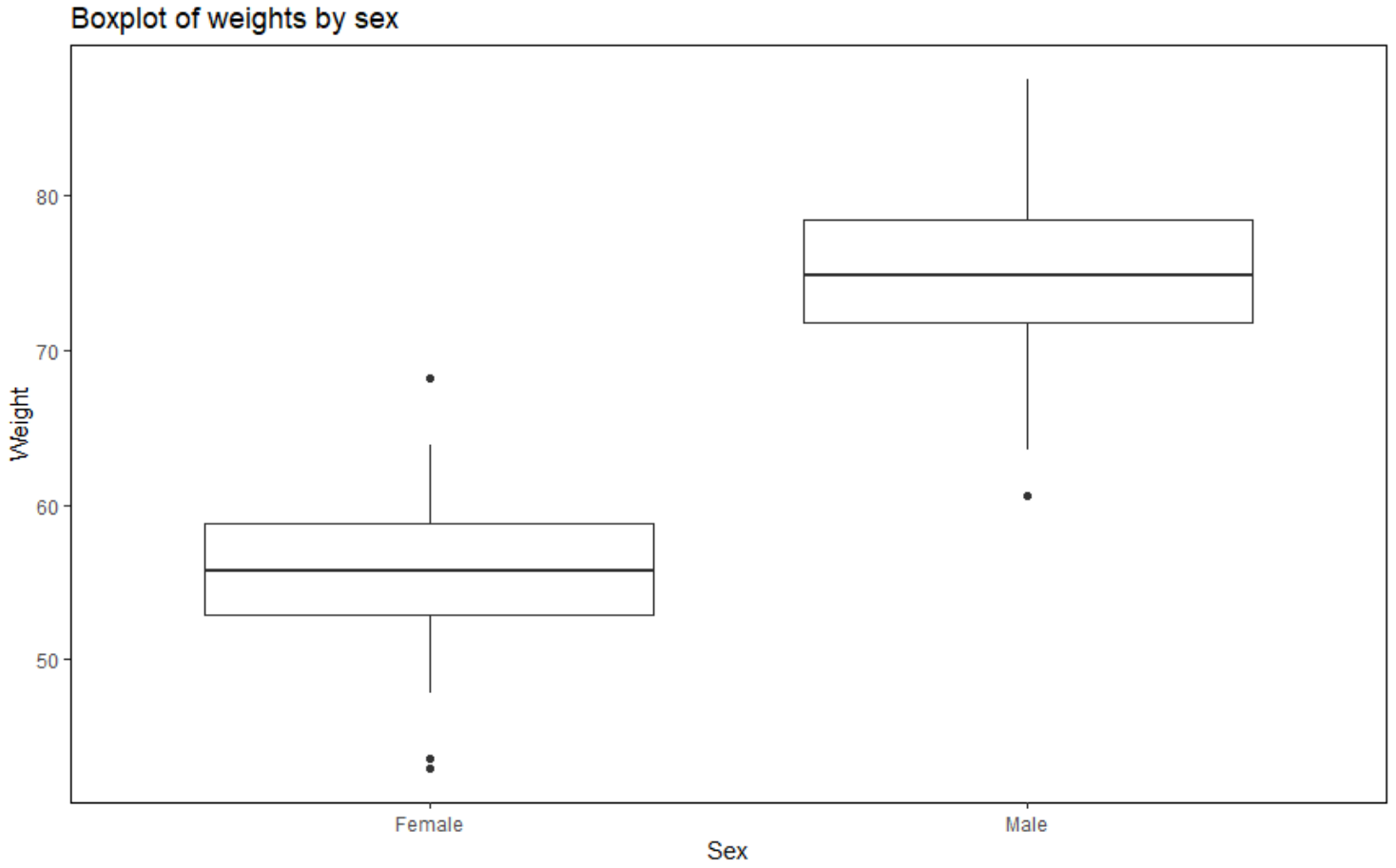
Scatterplot

```
g + geom_point(aes(x = Weight, y = Height)) +
  labs(title = "Height vs. weight plot")
```



Boxplot

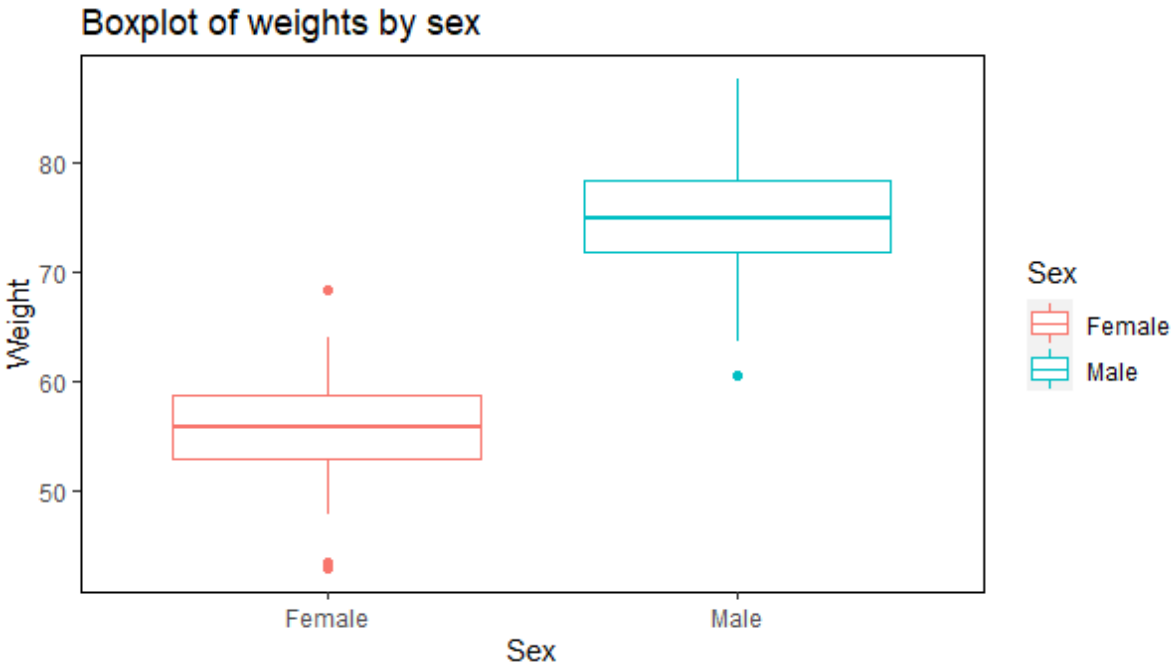
```
g + geom_boxplot(aes(x = Sex, y = Weight)) +
  labs(title = "Boxplot of weights by sex")
```



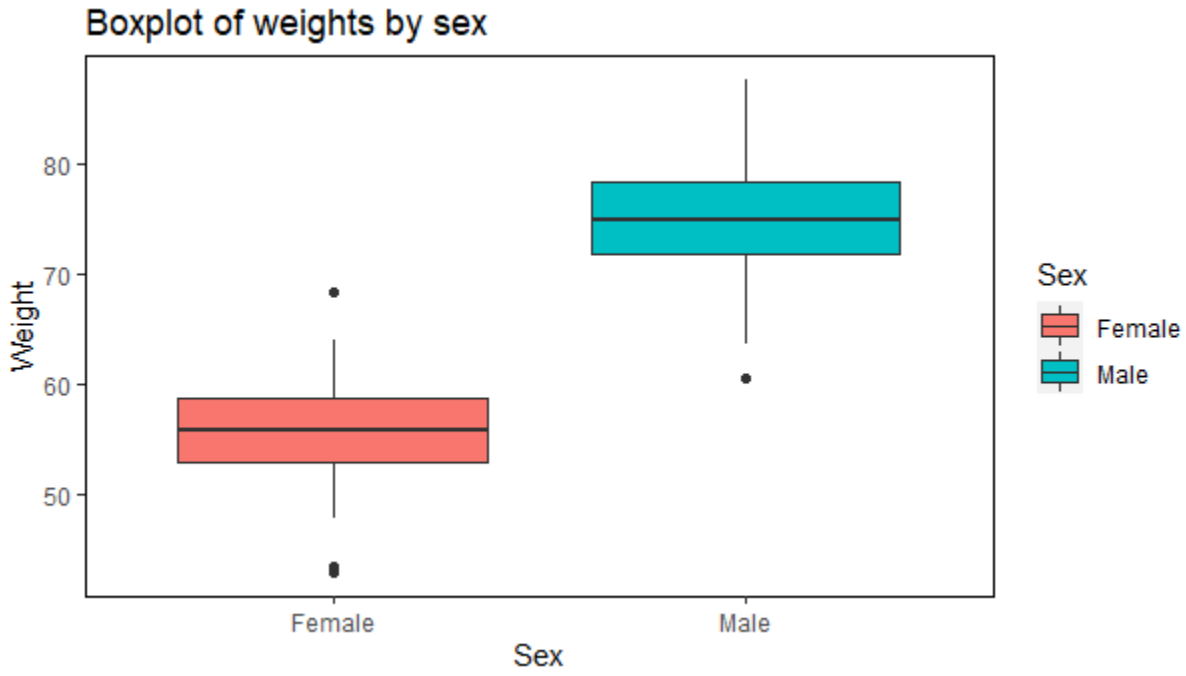
BGI华大

Parameter: color vs. fill

```
g + geom_boxplot(aes(Sex, Weight, color = Sex)) +
  labs(title = "Boxplot of weights by sex")
```

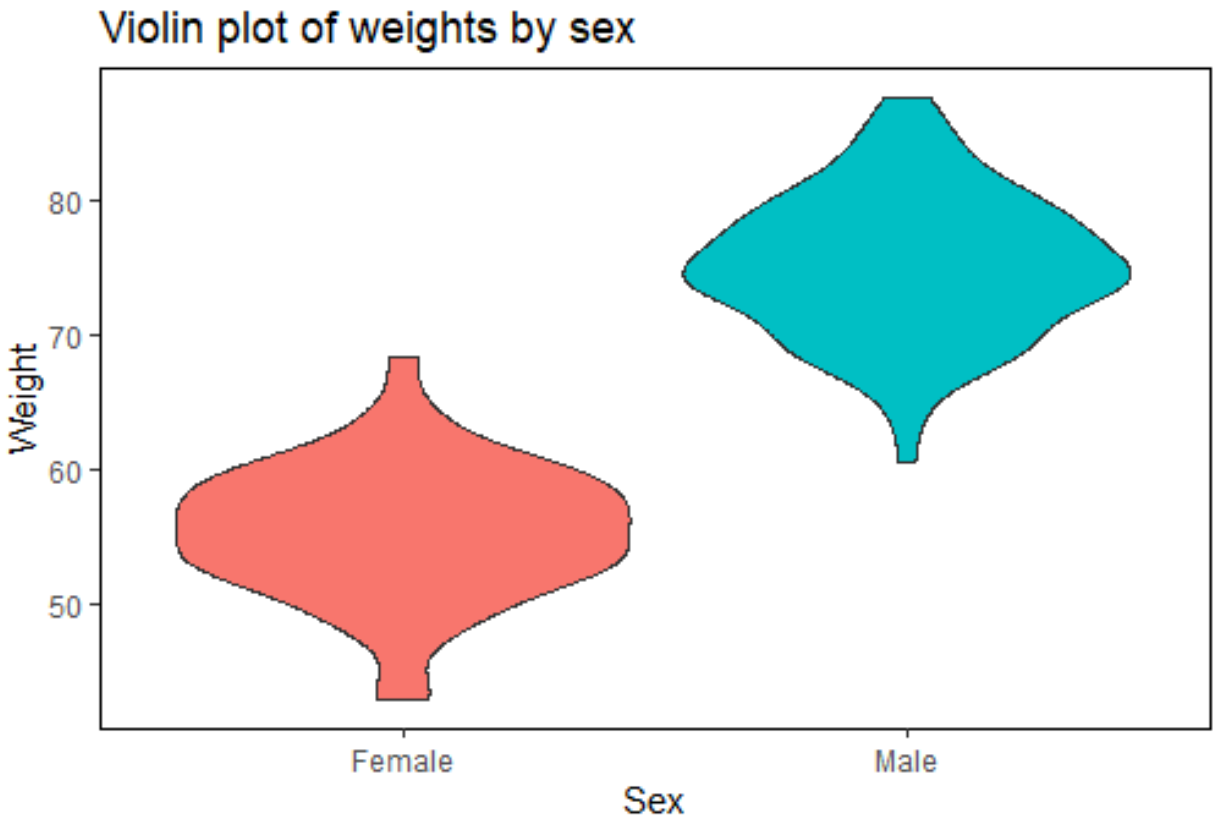



```
g + geom_boxplot(aes(Sex, Weight, fill = Sex)) +
  labs(title = "Boxplot of weights by sex")
```

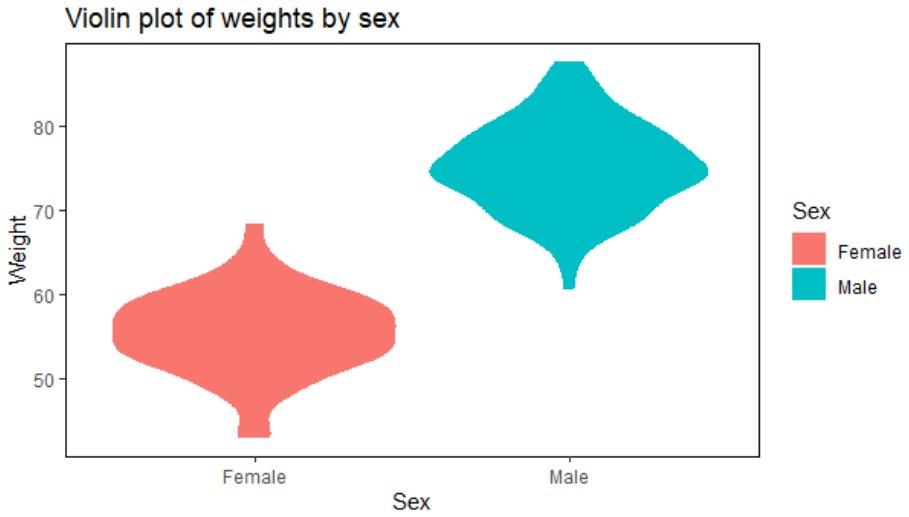


Violin plot

```
g + geom_violin(aes(x = Sex, y = Weight, fill = Sex)) +
  labs(title = "Violin plot of weights by sex")
```



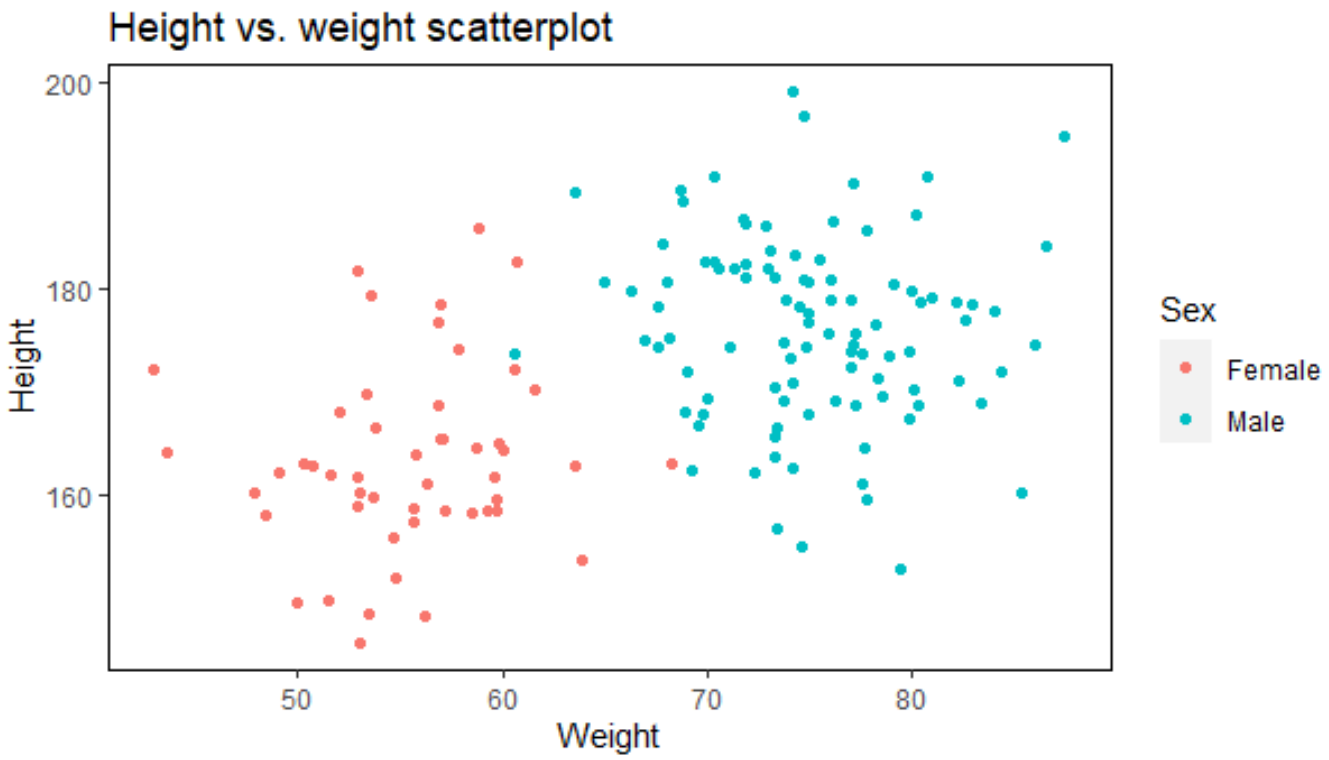

```
geom_violin(…, color = NA)
```



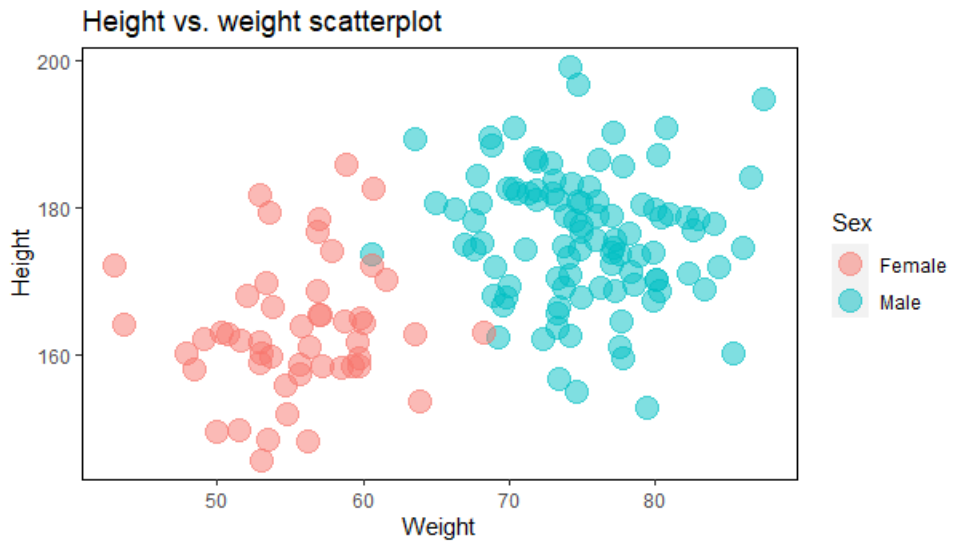
BGI华大

Scatterplot w/ a 3rd variable

```
g + geom_point(aes(x = Weight, y = Height, color = Sex)) +
  labs(title = "Height vs. weight scatterplot")
```

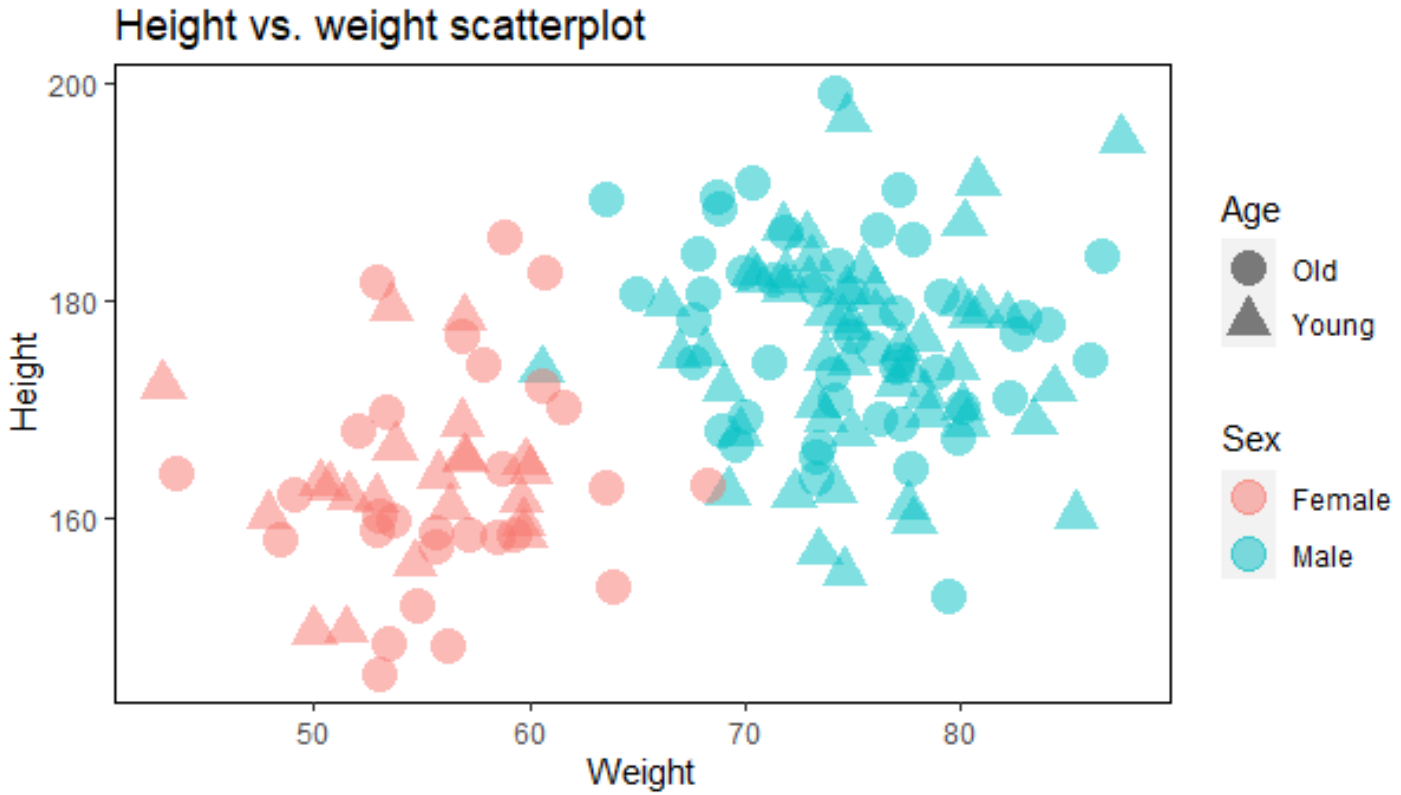



```
geom_point(…, size = 5, alpha = 0.5)
```



4-D data: aesthetics

```
g + geom_point(aes(x = Weight, y = Height,
                   color = Sex, shape = Age),
               size = 5, alpha = .5) +
  labs(title = "Height vs. weight scatterplot")
```

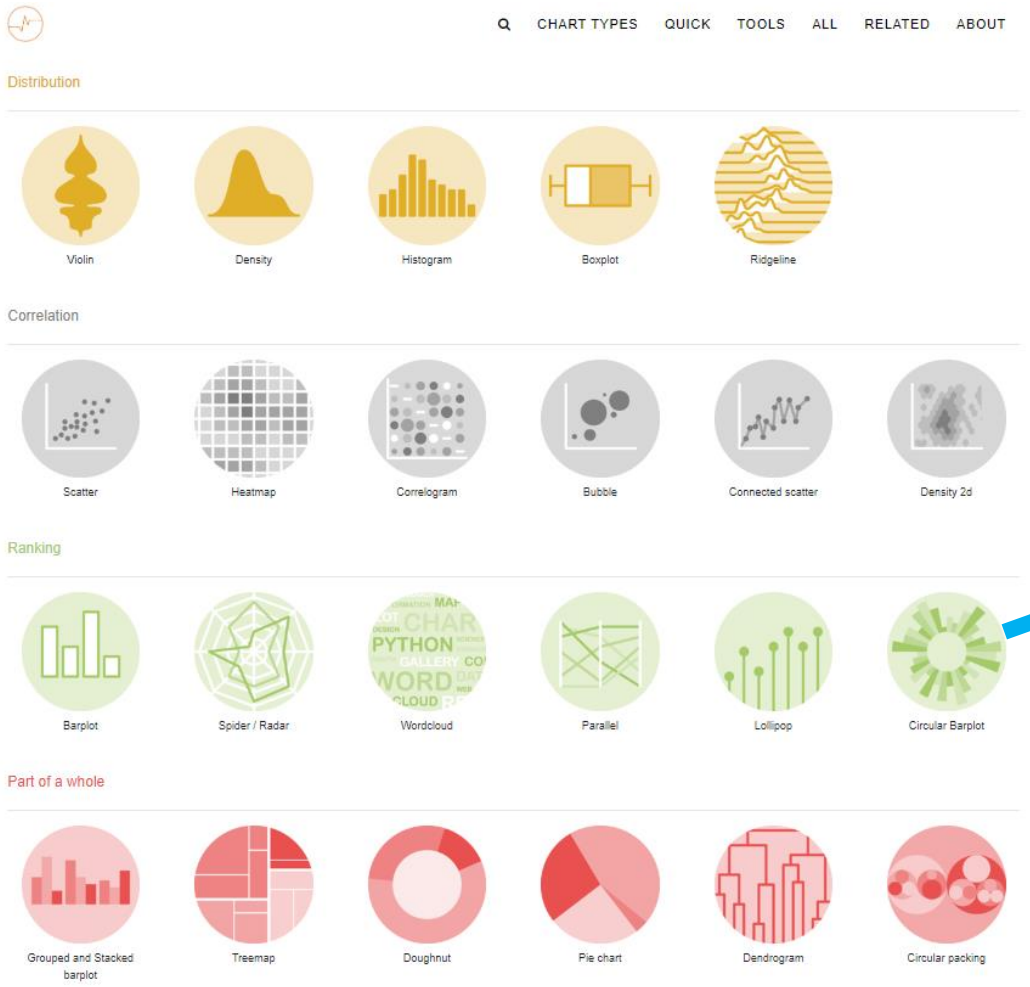


4-D data: facet

```
g + geom_point(aes(x = Weight, y = Height, color = Sex),
                   size = 5, alpha = .5) +
  labs(title = "Height vs. weight scatterplot") +
  facet_wrap(facets = vars(Age))
```

BGI华大

More plots in R gallery

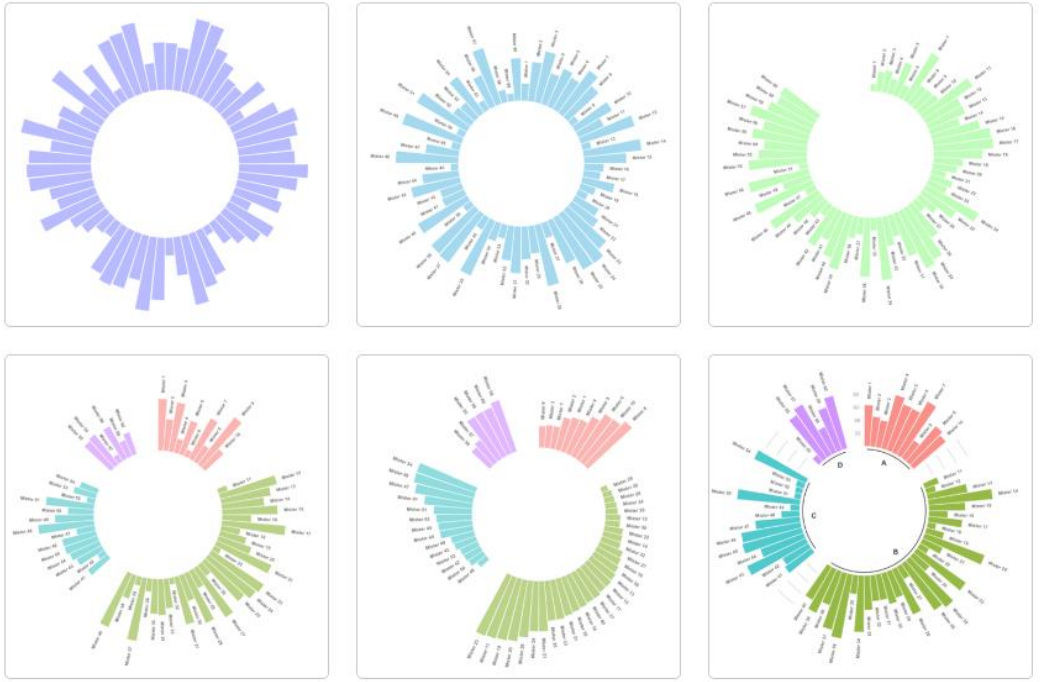


Circular barplot

This is the circular barplot section of the gallery, a variation of the well known barplot. Note that even if visually appealing, circular barplot must be used with care since groups do not share the same Y axis. It is very adapted for cyclical data though. Visit data-to-viz.com for more info.

STEP BY STEP

Here is a set of examples leading to a proper circular barplot, step by step. The first most basic circular barchart shows how to use `coord_polar()` to make the barchart circular. Next examples describe the next steps to get a proper figure: gap between groups, labels and customization.



Part II
A case study

BGI华大

Case I. Lettuce domestication

nature genetics

ARTICLES

https://doi.org/10.1038/s41588-021-00831-0

Check for updates

Whole-genome resequencing of 445 *Lactuca* accessions reveals the domestication history of cultivated lettuce

Tong Wei[1,11], Rob van Treuren[2,11]✉, Xinjiang Liu[1,11], Zhaowu Zhang[1,3], Jiongjiong Chen[4], Yang Liu[1], Shanshan Dong[5], Peinan Sun[4], Ting Yang[1], Tianming Lan[1,6], Xiaogang Wang[7], Zhouquan Xiong[7], Yaqiong Liu[8], Jinpu Wei[8], Haorong Lu[8], Shengping Han[8], Jason C. Chen[8], Xuemei Ni[1], Jian Wang[1,9], Huanming Yang[1,9], Xun Xu[1,10], Hanhui Kuang[4], Theo van Hintum[2], Xin Liu[1]✉ and Huan Liu[1]✉

Lettuce (*Lactuca sativa*) is an important vegetable crop worldwide. Cultivated lettuce is believed to be domesticated from *L. serriola*; however, its origins and domestication history remain to be elucidated. Here, we sequenced a total of 445 *Lactuca* accessions, including major lettuce crop types and wild relative species, and generated a comprehensive map of lettuce genome variations. In-depth analyses of population structure and demography revealed that lettuce was first domesticated near the Caucasus, which was marked by loss of seed shattering. We also identified the genetic architecture of other domestication traits and wild introgressions in major resistance clusters in the lettuce genome. This study provides valuable genomic resources for crop breeding and sheds light on the domestication history of cultivated lettuce.

Nat Genet 2021, 53:752-760

Case I. Lettuce population

- As a vegetable crop
  - A rich source of vitamin K and vitamin A, and a moderate source of folate and iron
  - Consumed worldwide

- As a model Asteraceae plant
  - A short life cycle; easy for transformation
  - Well-maintained germplasms
  - Various agronomic traits
  - Potential bioreactor

var. *longifoliaf*

var. *angustanairish*

BGI华大

Case I. Scientific questions

?

1. **Where** was lettuce domesticated?

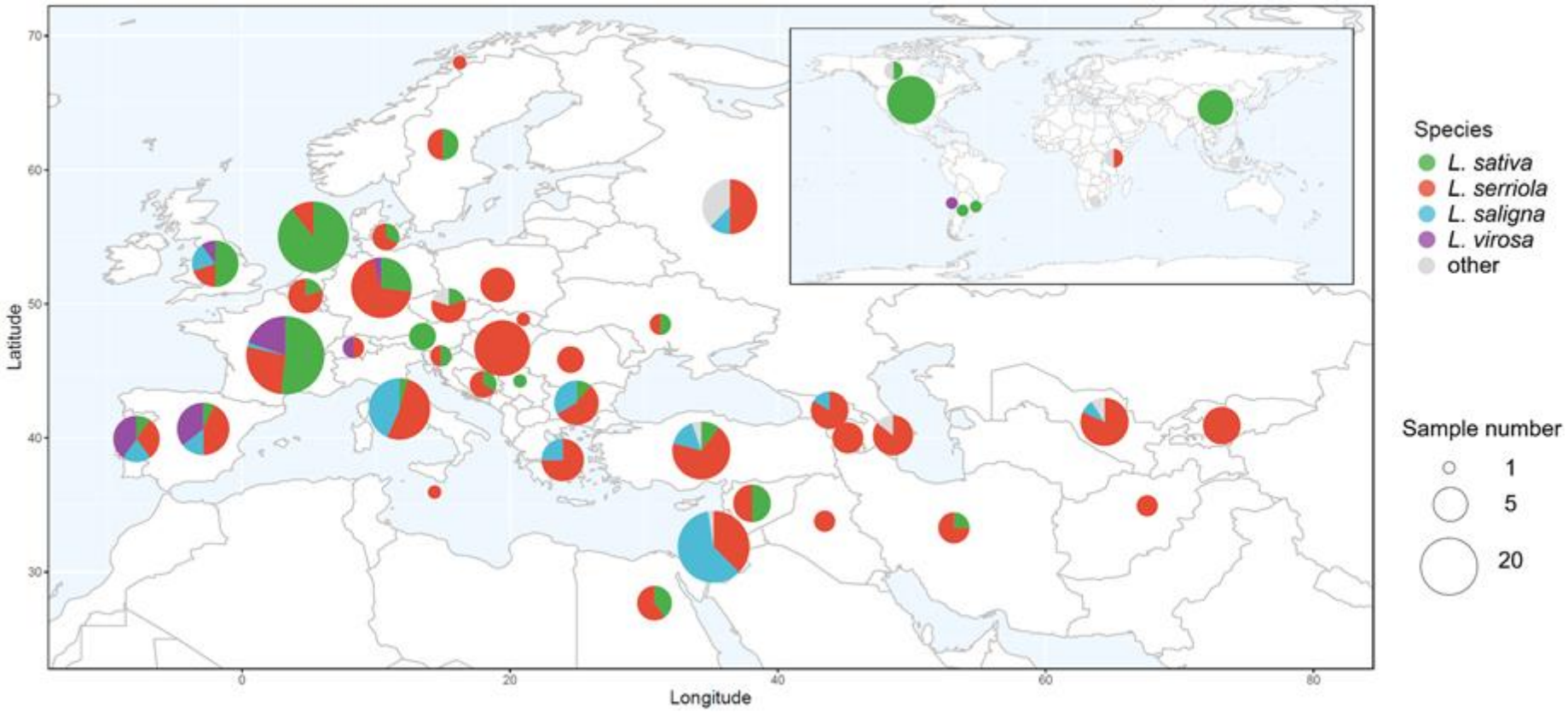
2. **When** was lettuce domesticated?

3. What are the **key events** during lettuce domestication?

4. What are the **genetic determinants** for the domestication traits?

Case I. Global representation in world map

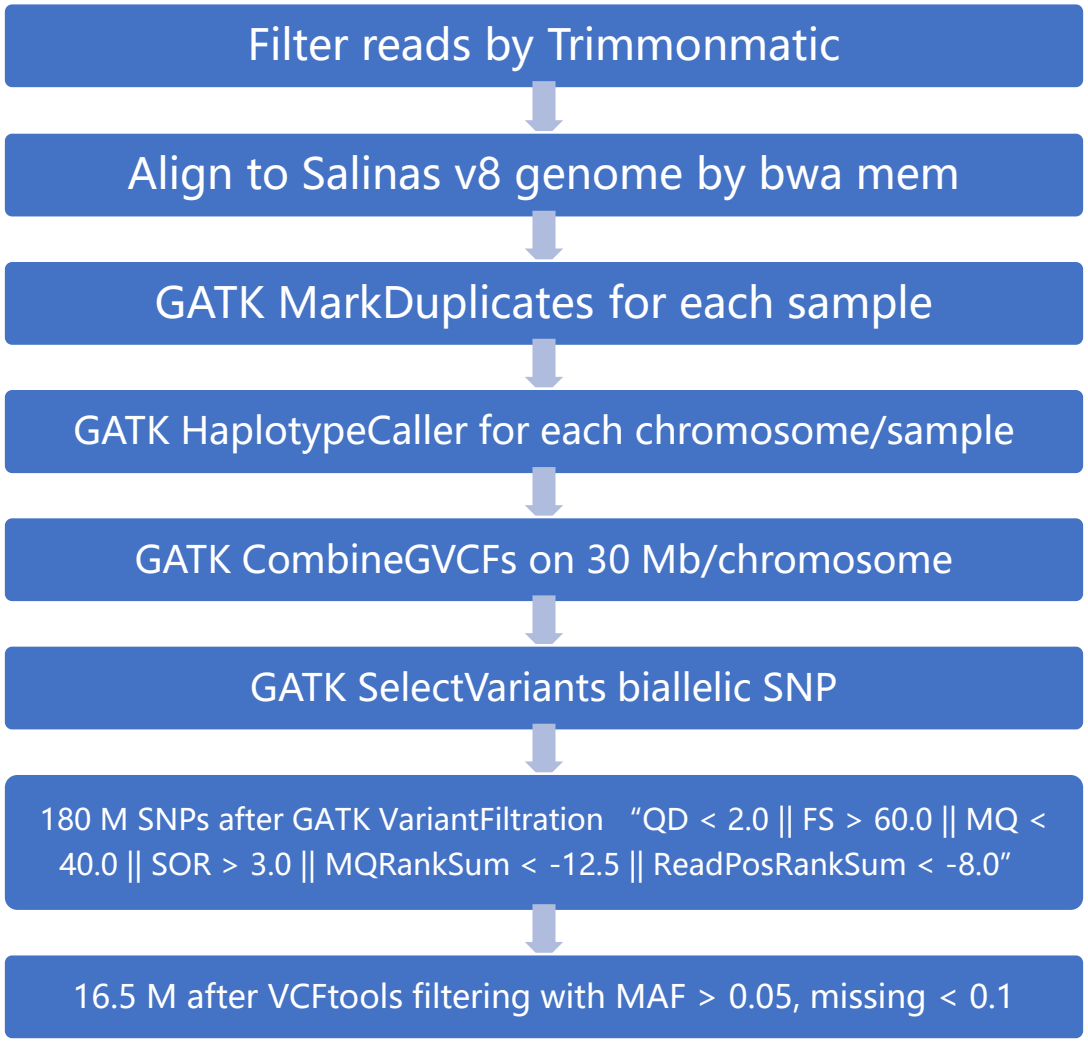
445 lines from 47 countries, mainly composed of GP breeding materials of *L. sativa*, *L. serriola*, *L. saligna* and *L. virosa*



Nat Genet 2021, 53:752-760

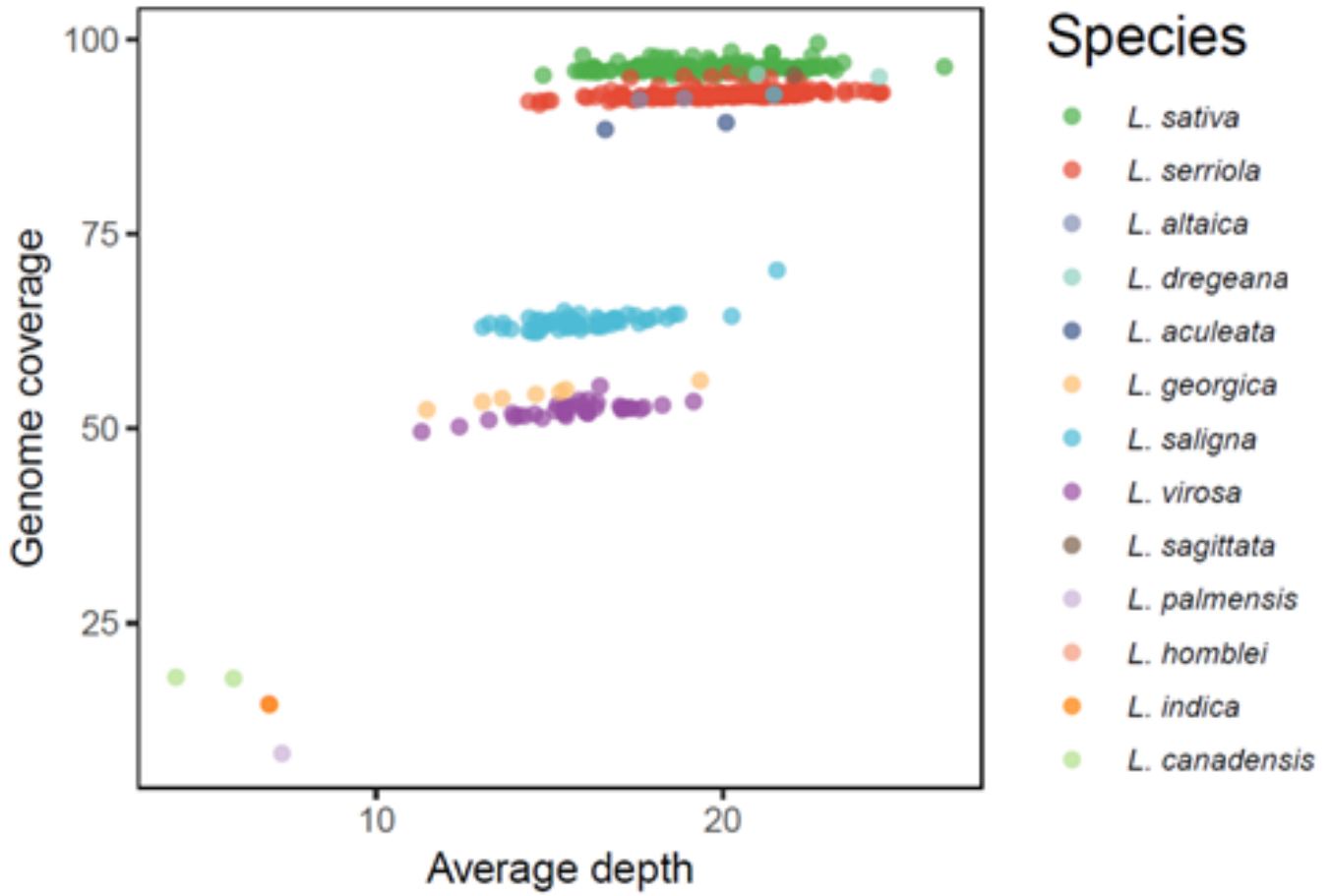
BGI华大

Case I. Pipeline in flow chart



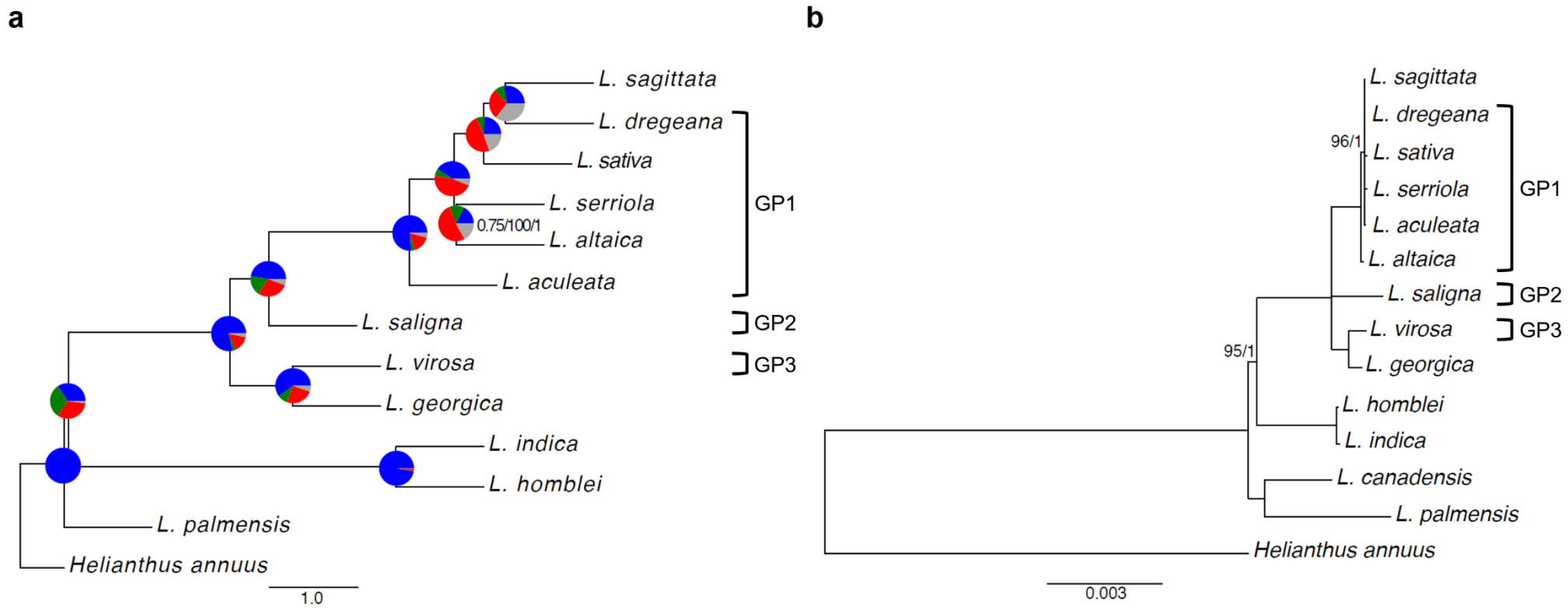
BGI华大

Case I. Read alignment in scatter plot



Nat Genet 2021, 53:752-760

Case I. Phylogeny in tree plot



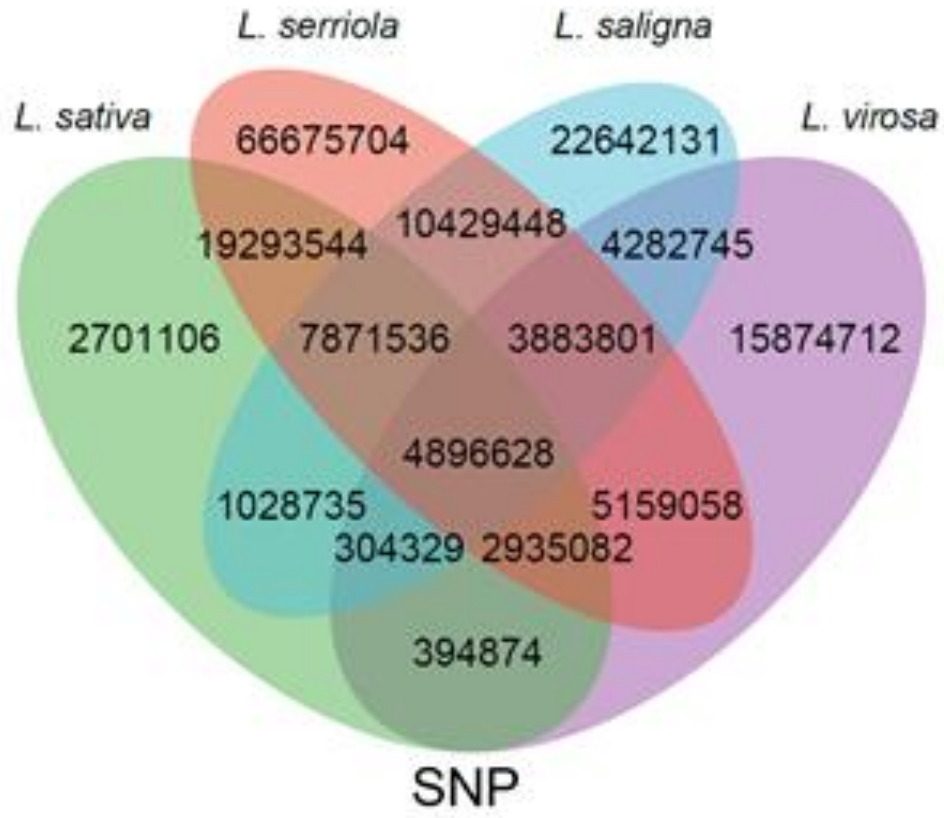
Coalescence-based tree from 4,513 nuclear genes

Concatenation-based tree from 75 plastid genes

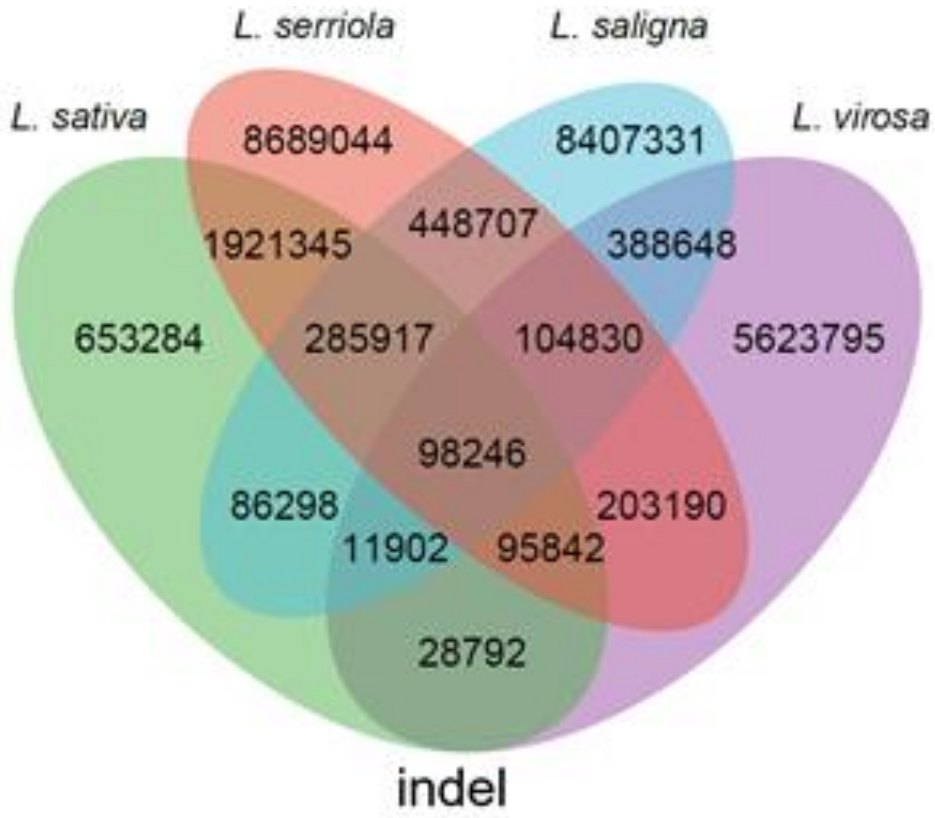
Nat Genet 2021, 53:752-760

Case I. SNP/indel stats in Venn diagram

88.51% SNPs in *L. sativa* are shared with *L. serriola.*



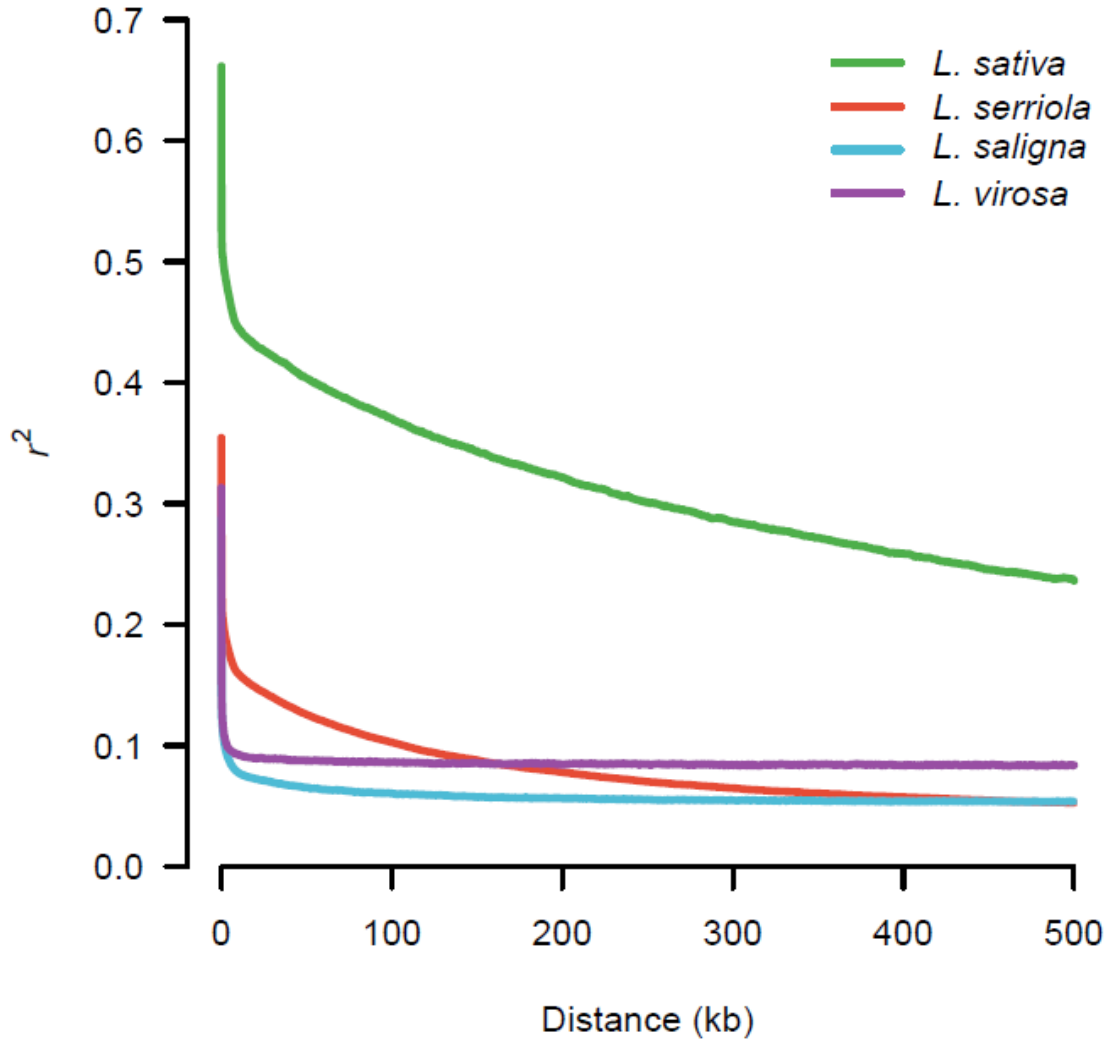
SNP



indel

Nat Genet 2021, 53:752-760

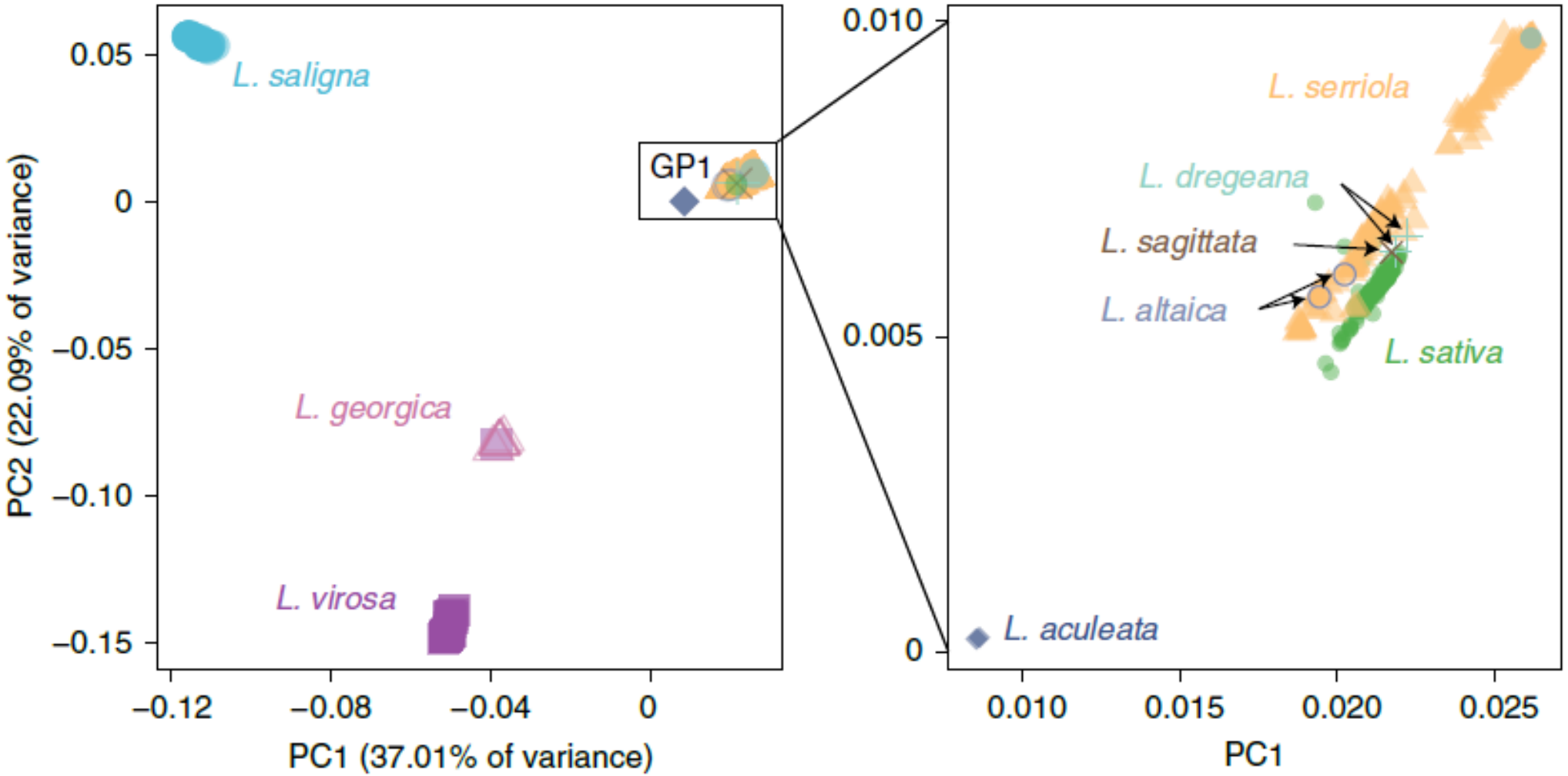
Case I. Linkage disequilibrium in line chart



Nat Genet 2021, 53:752-760

BGI华大

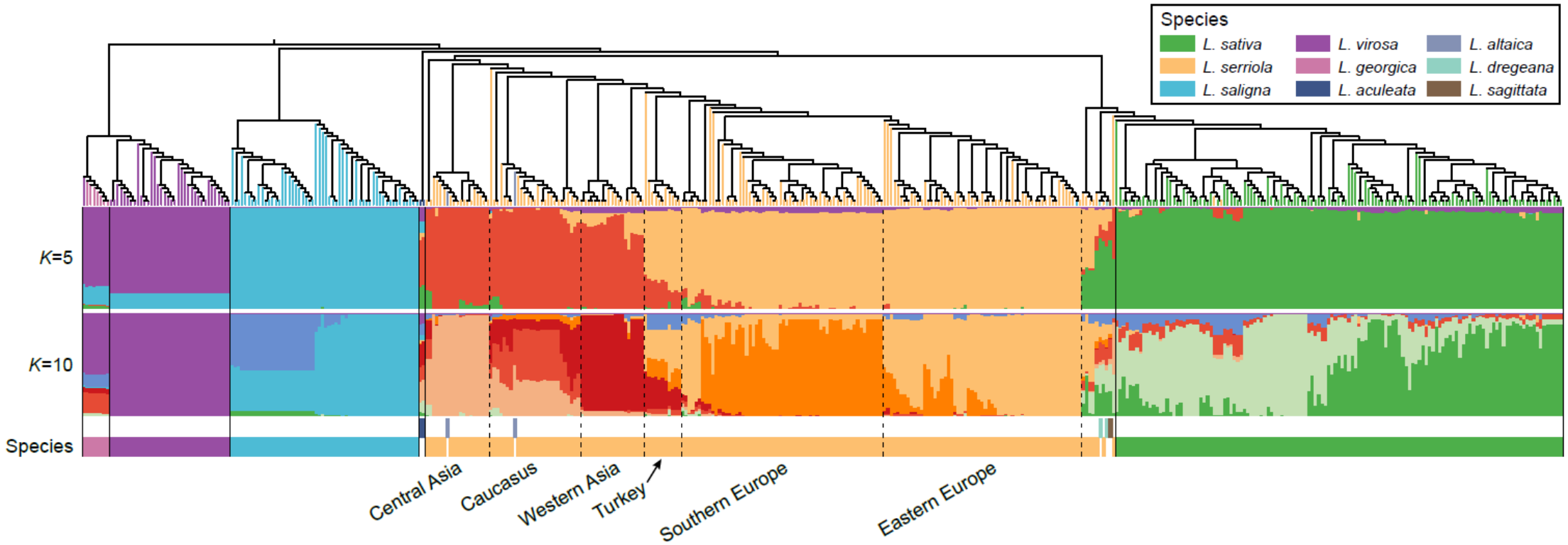
Case I. PCA in scatter plot



Nat Genet 2021, 53:752-760

BGI华大

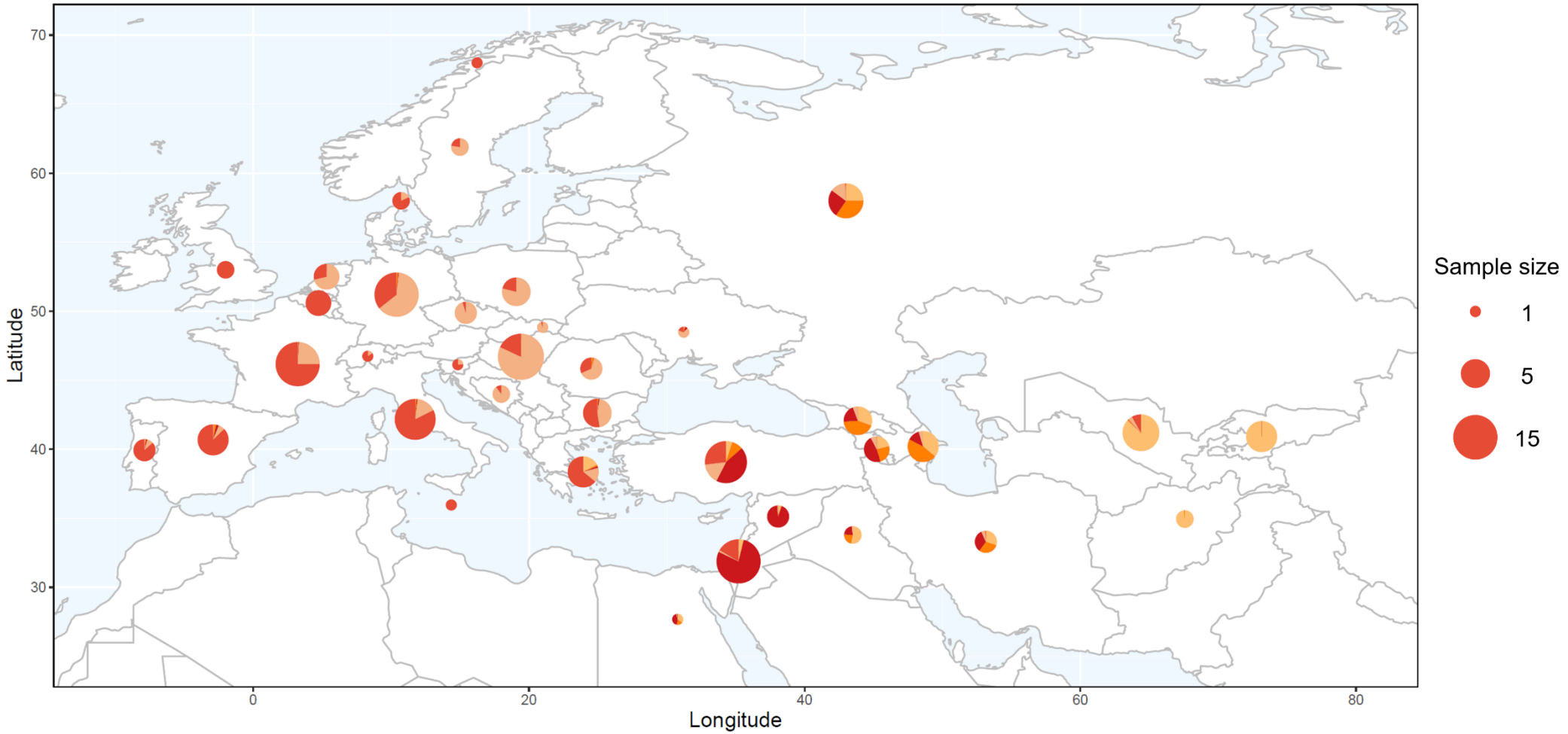
Case I. Structure in barplot



Nat Genet 2021, 53:752-760

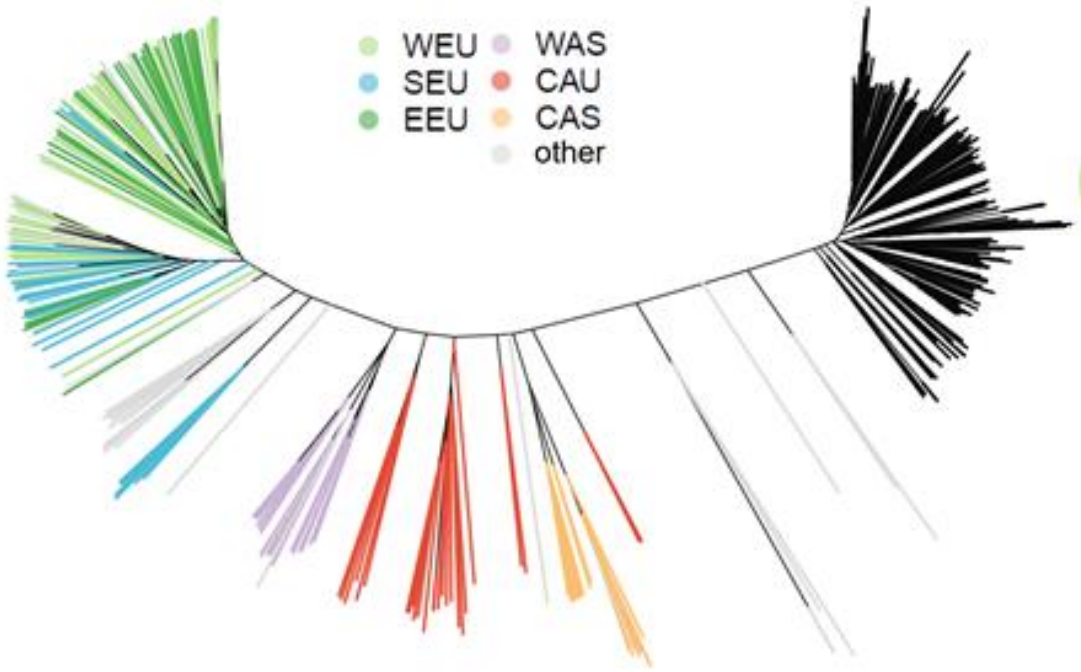
BGI华大

Case I. Genetic composition in piechart



Nat Genet 2021, 53:752-760

Case I. Genetic & geographic grouping

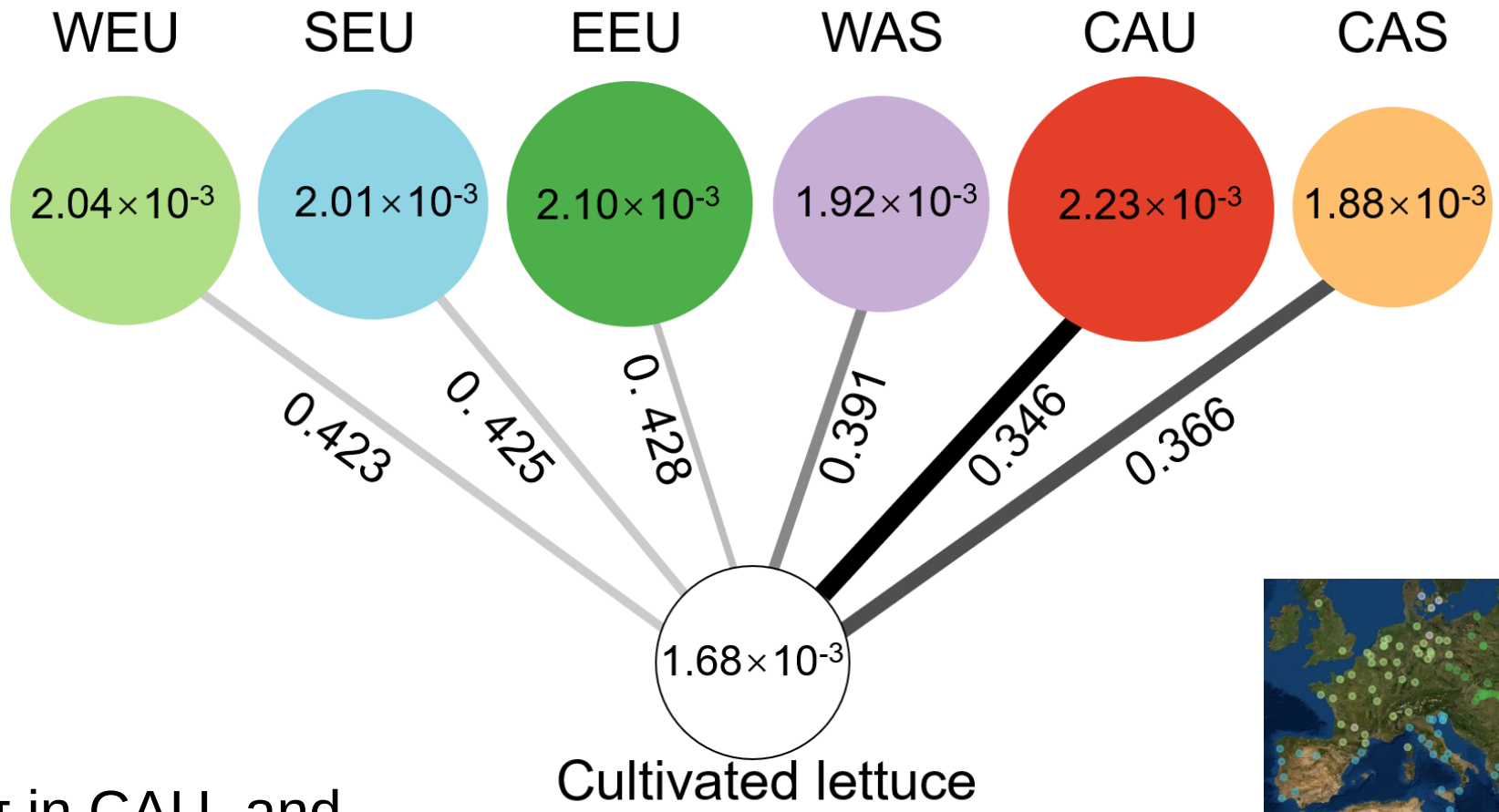


3 European groups
- WEU: western Europe
- SEU: southern Europe
- EEU: eastern Europe

3 Asian groups
- WAS: western Asian
- CAU: Caucasian
- CAS: central Asian

Nat Genet 2021, 53:752-760

Case I. Population statistic in infographics



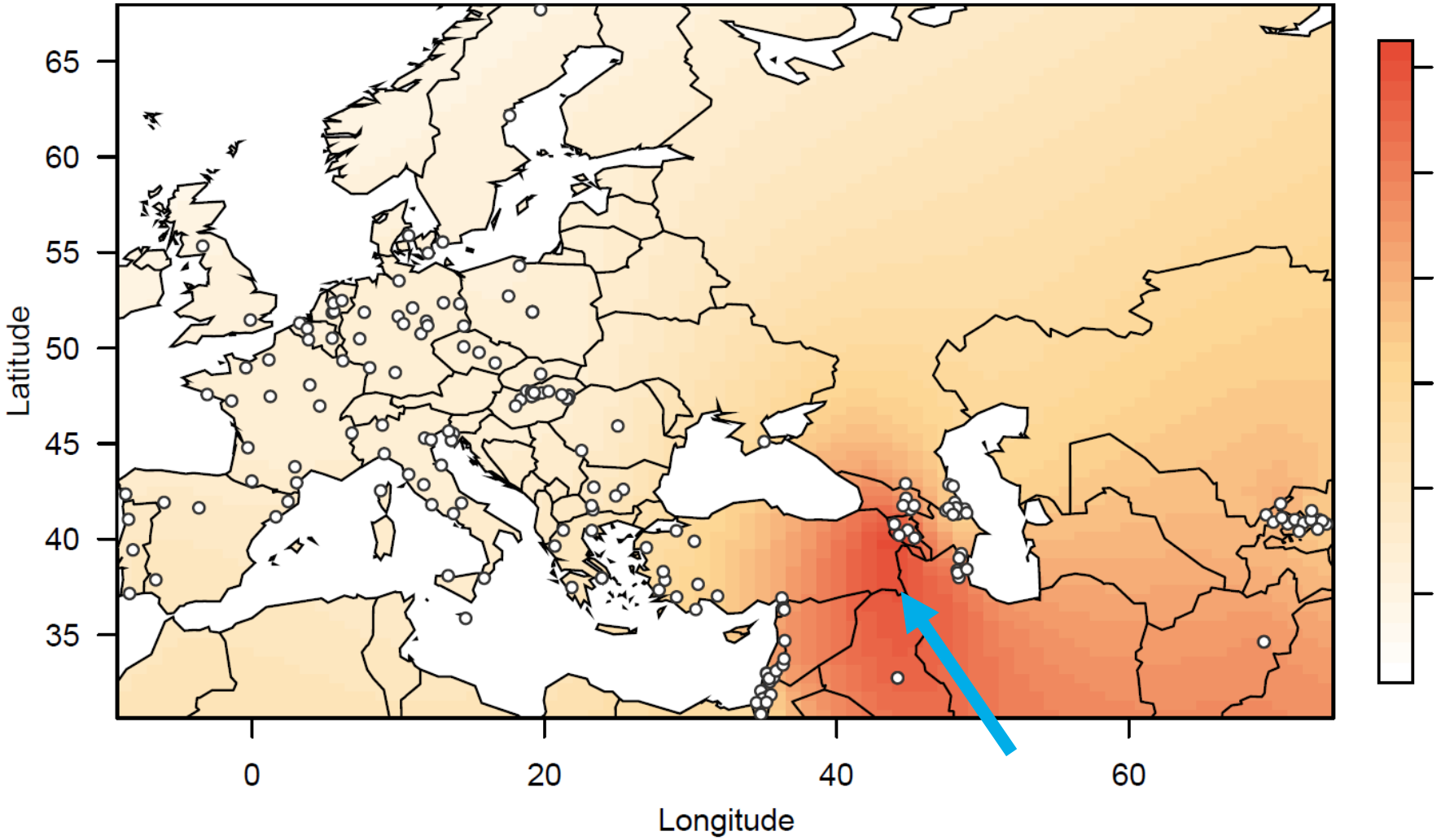
The highest π in CAU, and
the smallest F_{ST} between CAU & lettuce

Nat Genet 2021, 53:752-760

BGI华大

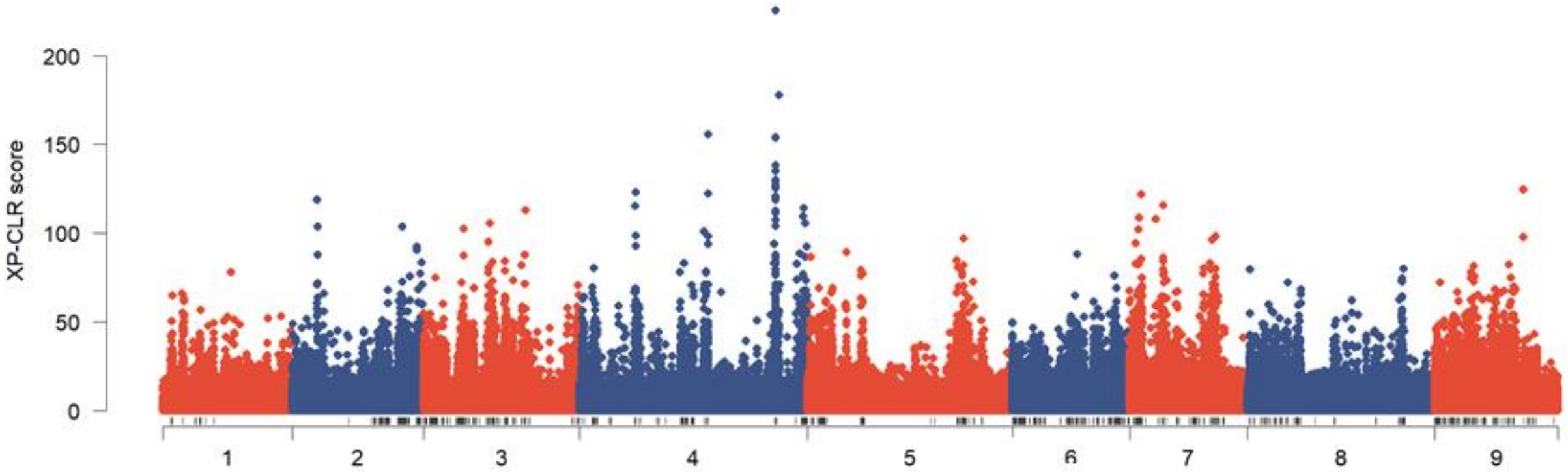
Case I. Singleton heatmap on map

The highest singletons in CAU



Case I. Selective sweeps in Manhattan plot

107.7 Mb and 2,304 genes within 4,089 selective sweeps



Nat Genet 2021, 53:752-760

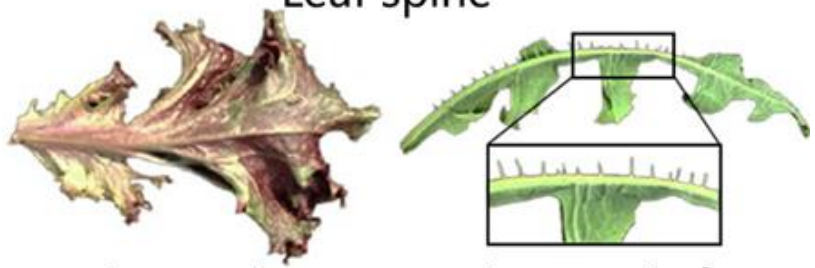
BGI华大

Case I. GWAS in Manhattan plot

Seed shattering

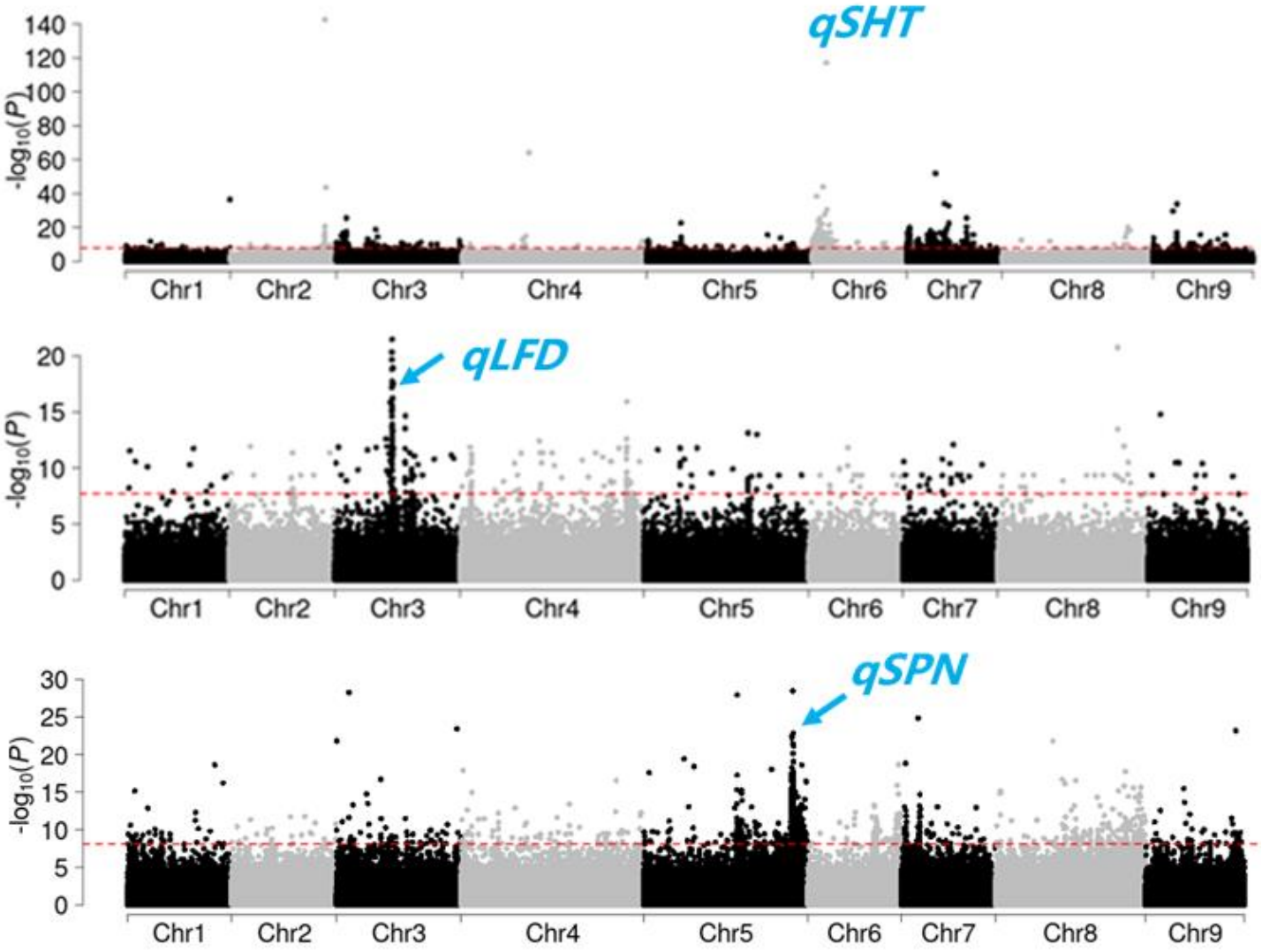
Leaf morphology

Leaf spine



L. sativa

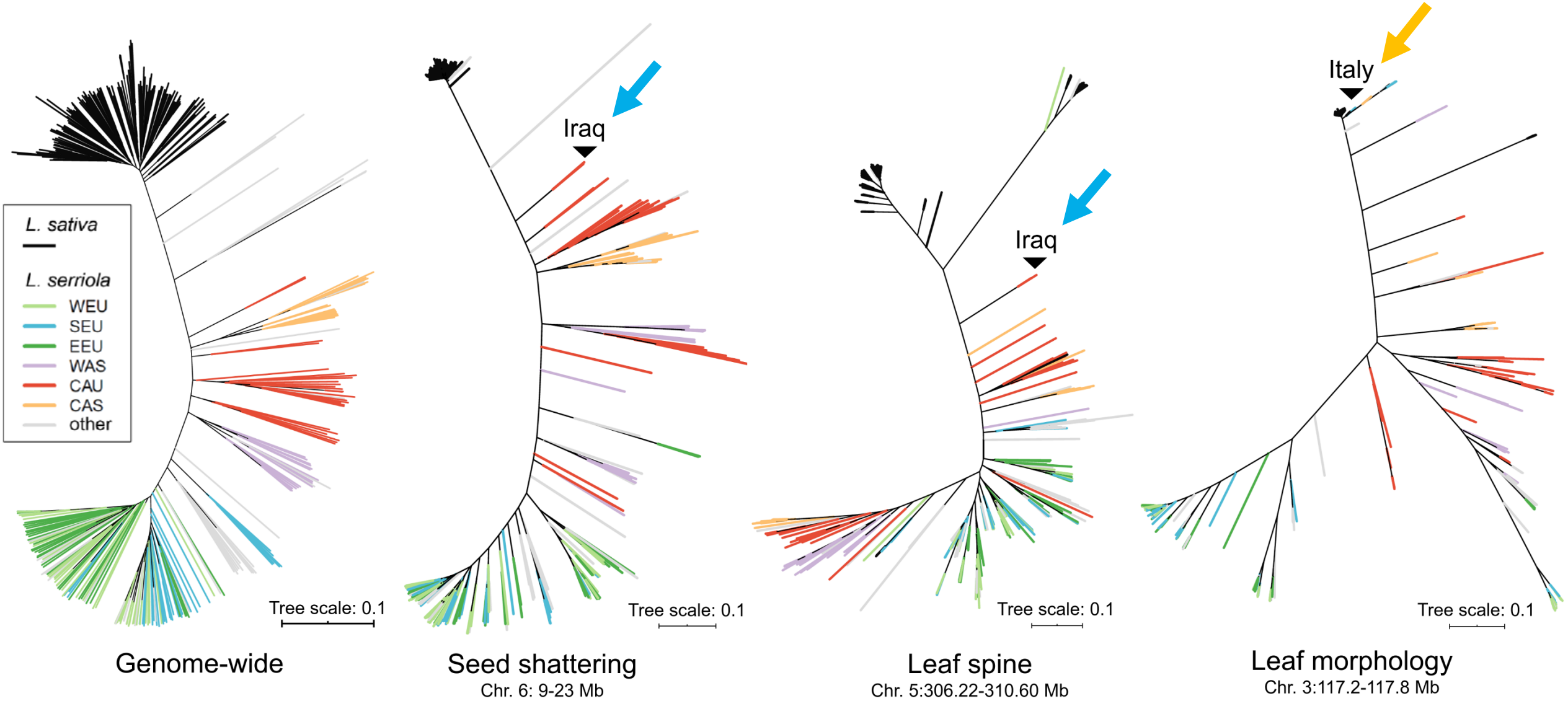
L. serriola



Nat Genet 2021, 53:752-760

BGI华大

Case I. Trait origins deduced in tree plot



Genome-wide

Seed shattering
Chr. 6: 9-23 Mb

Leaf spine
Chr. 5:306.22-310.60 Mb

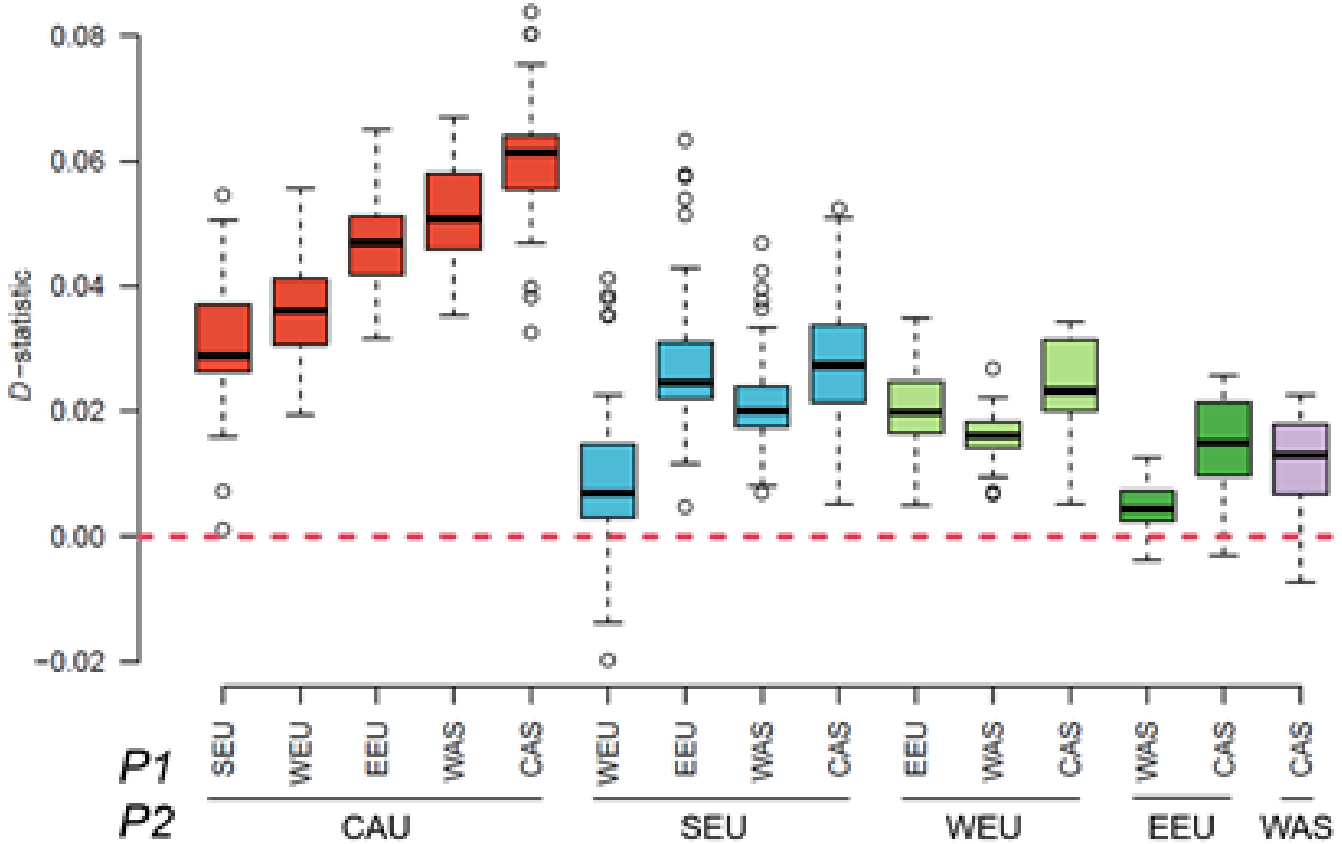
Leaf morphology
Chr. 3:117.2-117.8 Mb

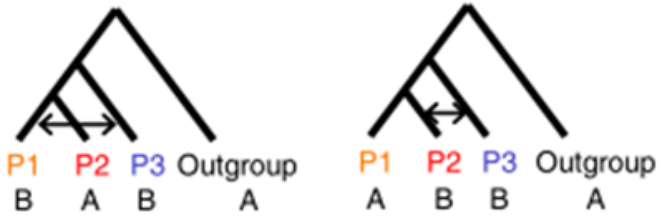
Nat Genet 2021, 53:752-760

BGI华大

Case I. ABBA replicates in boxplot

A close relationship in CAU and SEU

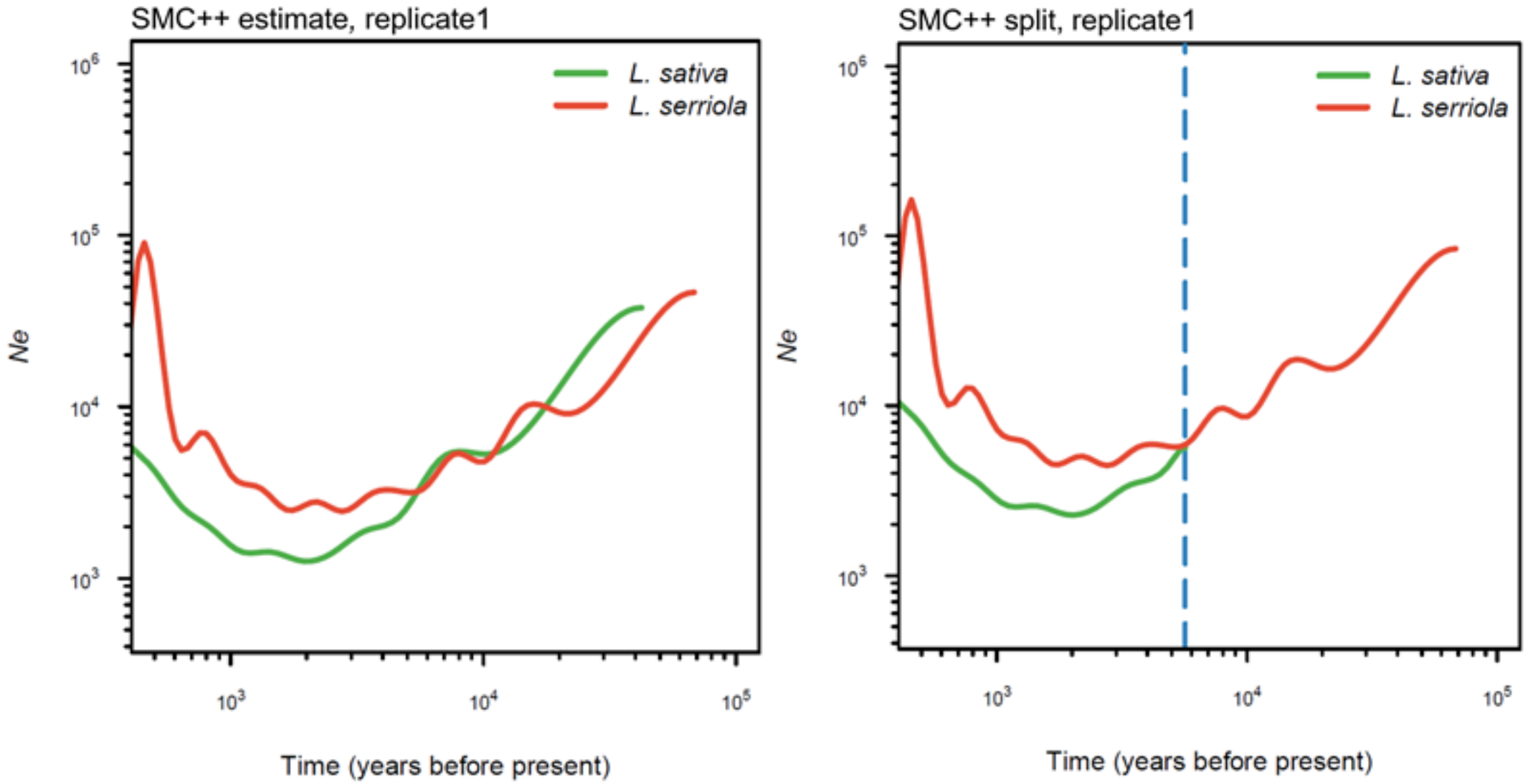




Outgroup: *L. saligna*
P3: *L. sativa*
P1 & P2: *L. serriola* groups

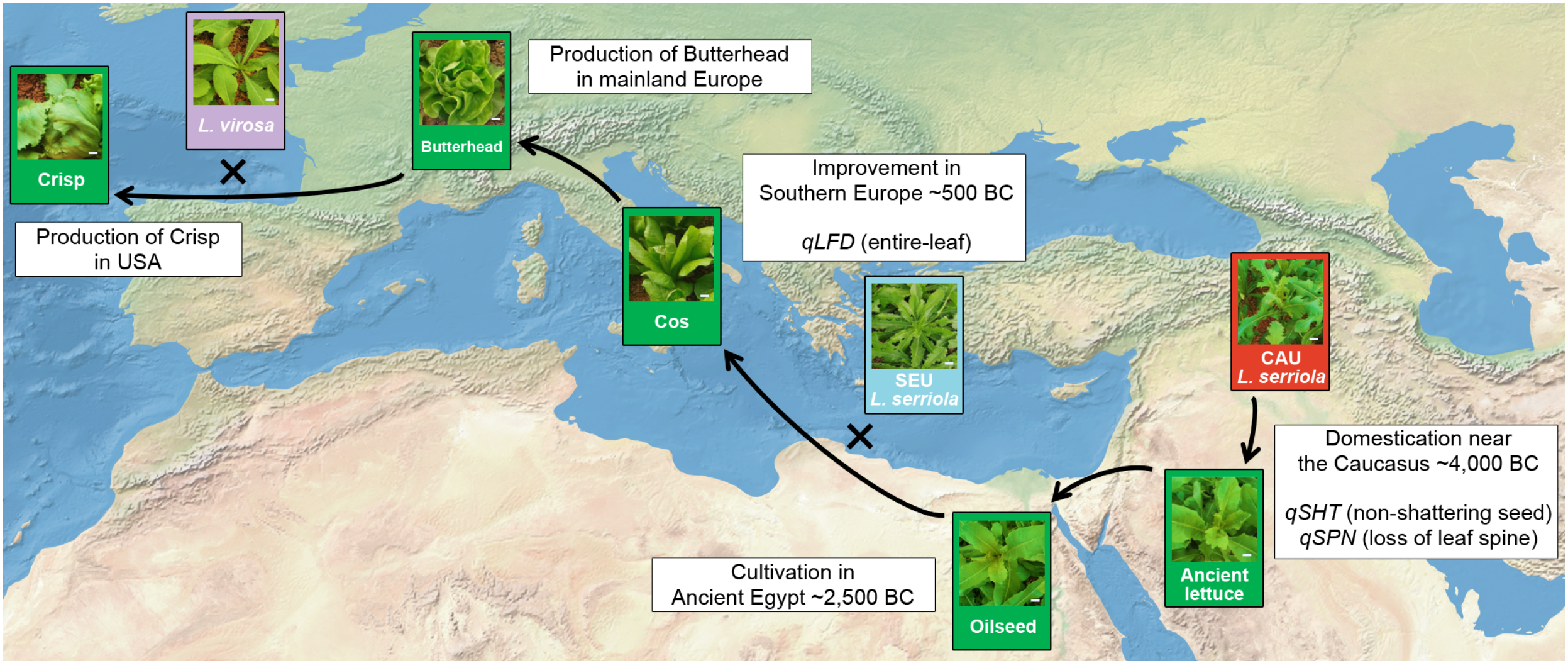
Nat Genet 2021, 53:752-760

Case I. Effective population size in line chart



Nat Genet 2021, 53:752-760

Case I. Domestication history in infographics



Nat Genet 2021, 53:752-760

BGI华大

Case I. Answers to scientific questions

- Q: **Where and when** was lettuce was domesticated?
- A: Lettuce was domesticated **near Caucasus approximately 6,000 years ago**.

- Q: What are the **major events** and traits during lettuce domestication and improvement?
- A: **Non-shattering seeds** marked lettuce domestication; the **entire-leaf trait was introduced later** from a Southern European wild population.

Nat Genet 2021, 53:752-760

Case I. Main figures about population

Fig. 1 | Phylogeny and population structure of cultivated lettuce and its wild relative species.

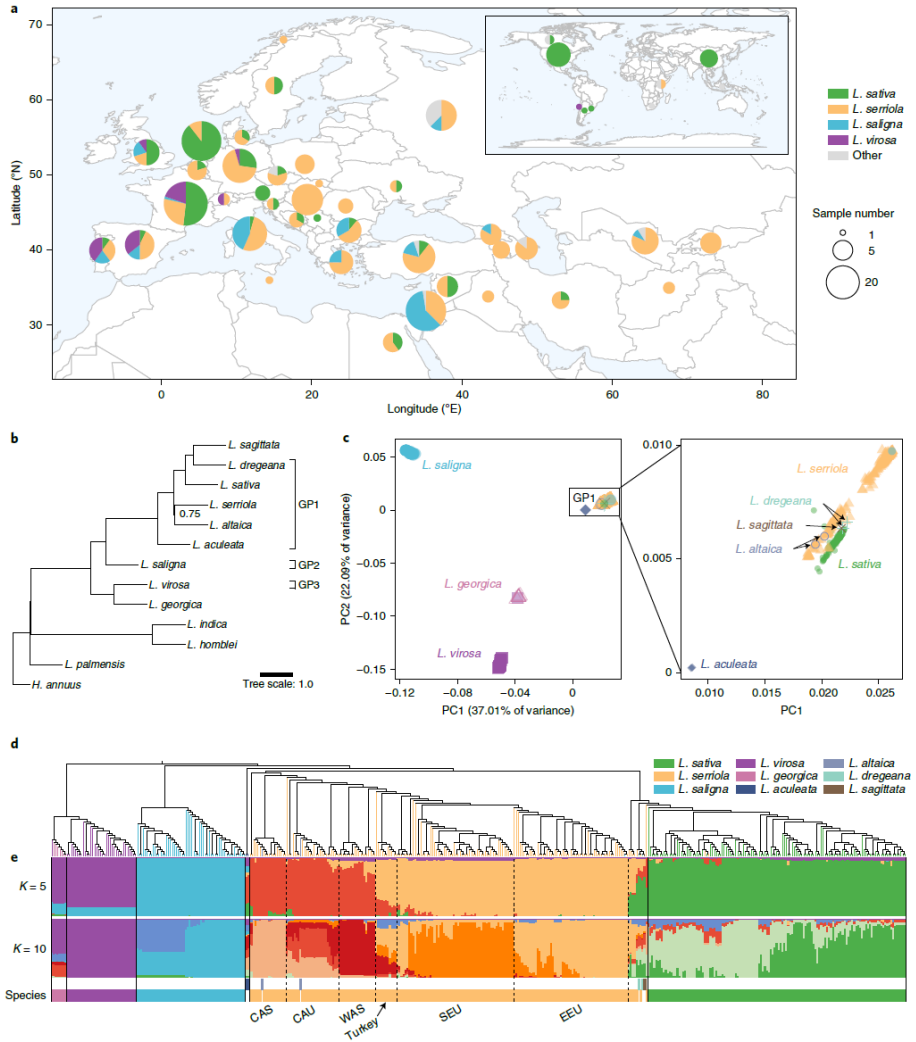
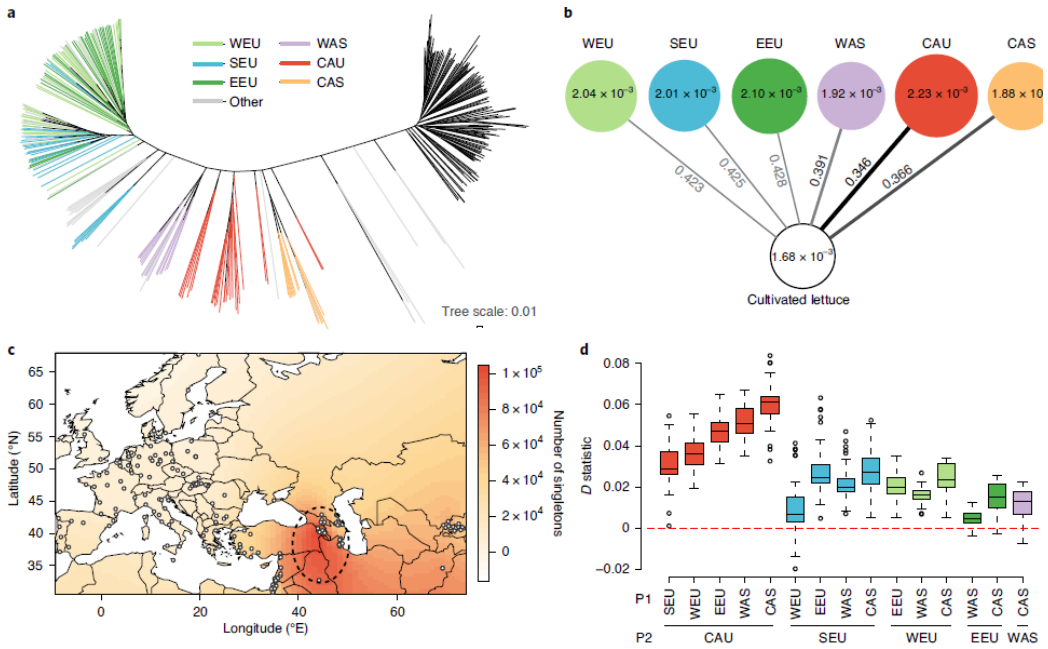


Fig. 2 | Proposed domestication center of cultivated lettuce near the Caucasus.



Case I. Main figures about function

Fig. 3 | Identification of selective sweeps associated with domestication traits in cultivated lettuce.

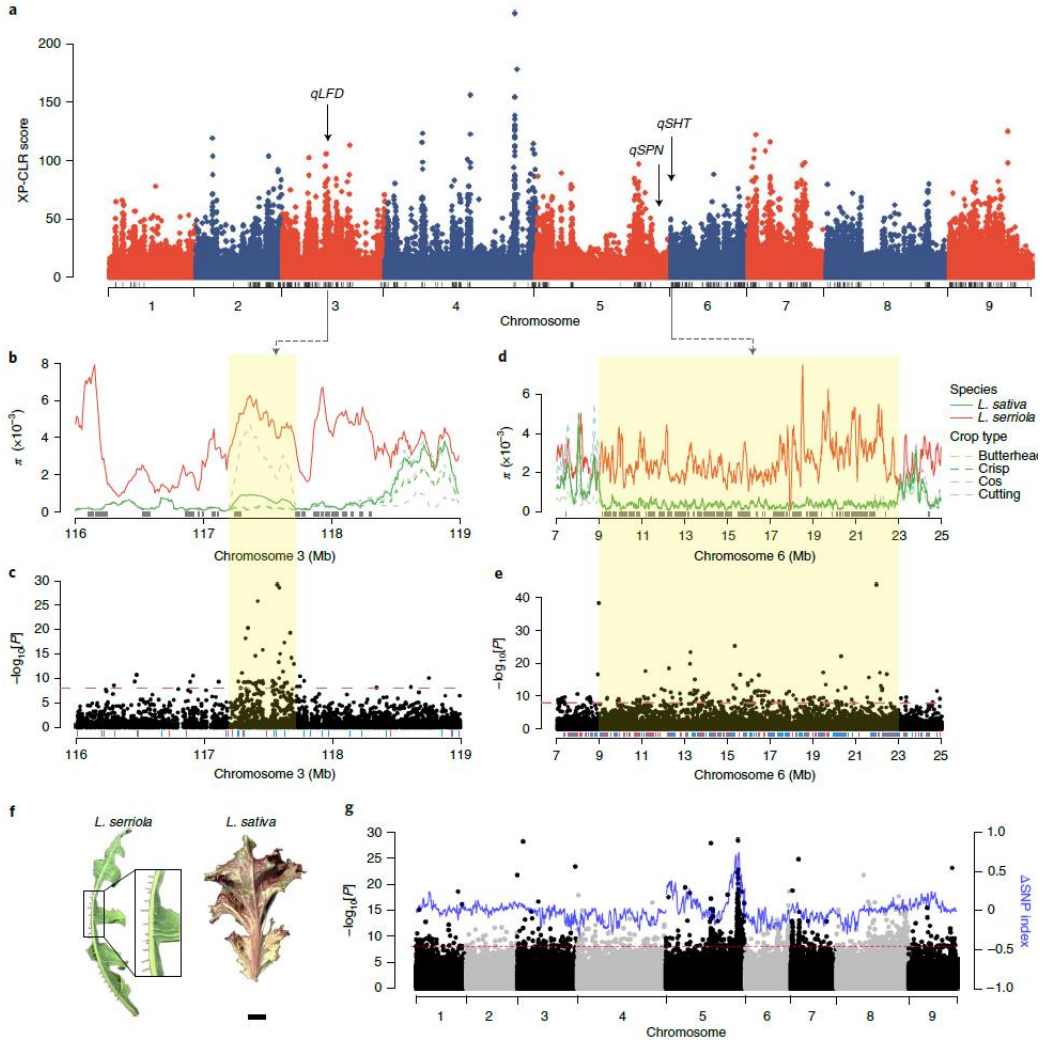
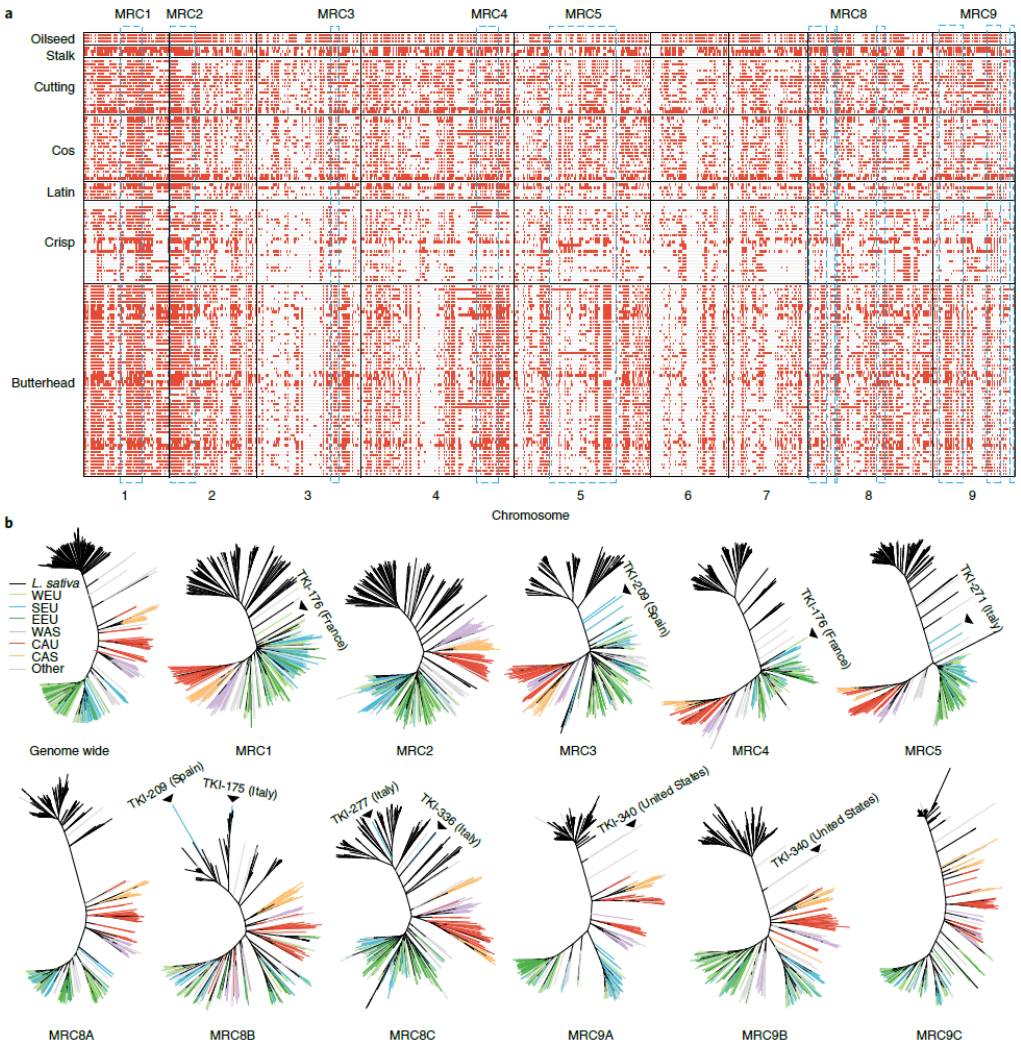
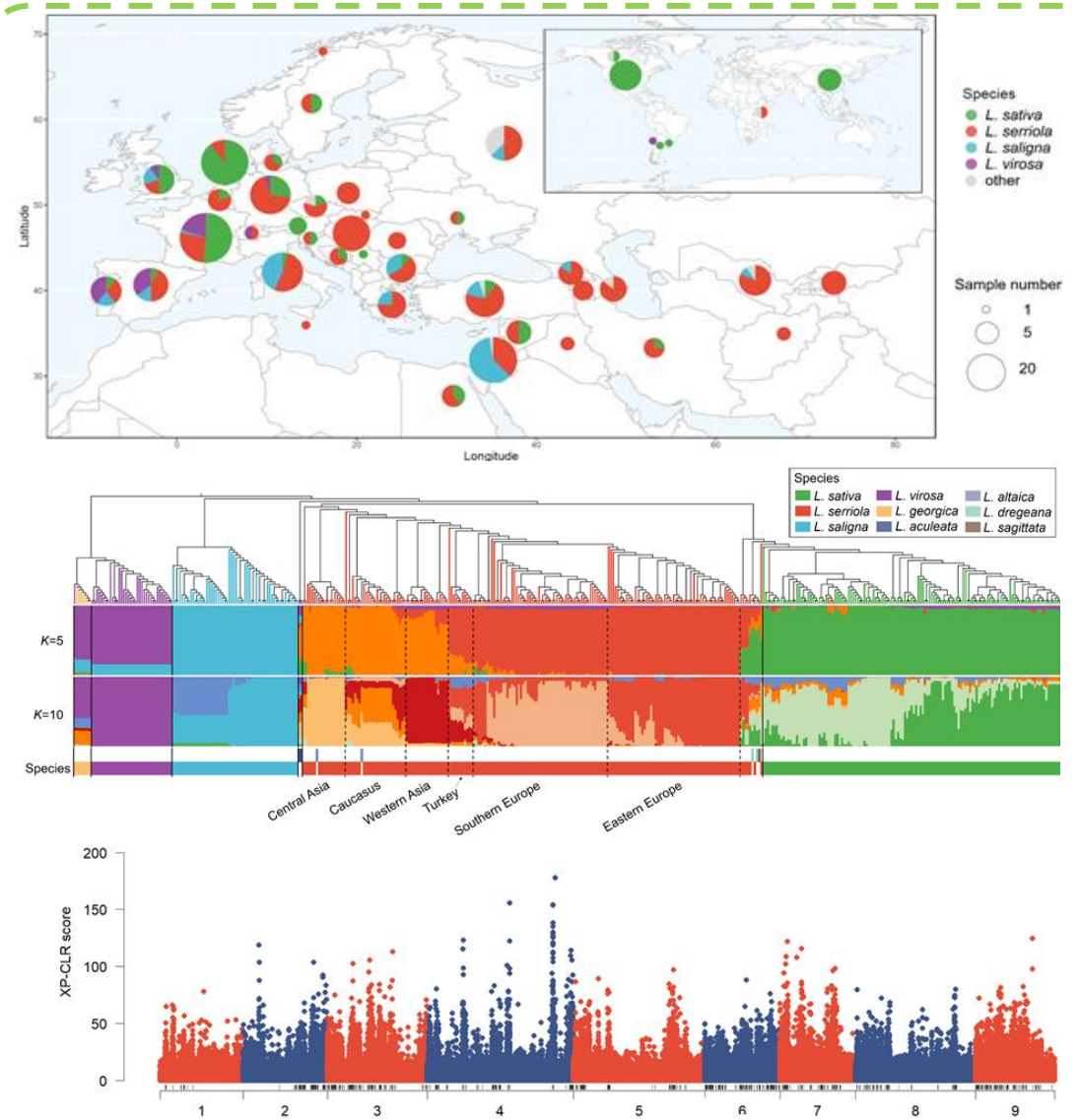


Fig. 4 | Introgressive contribution of *L. serriola* to lettuce resistance breeding.



Case I summary

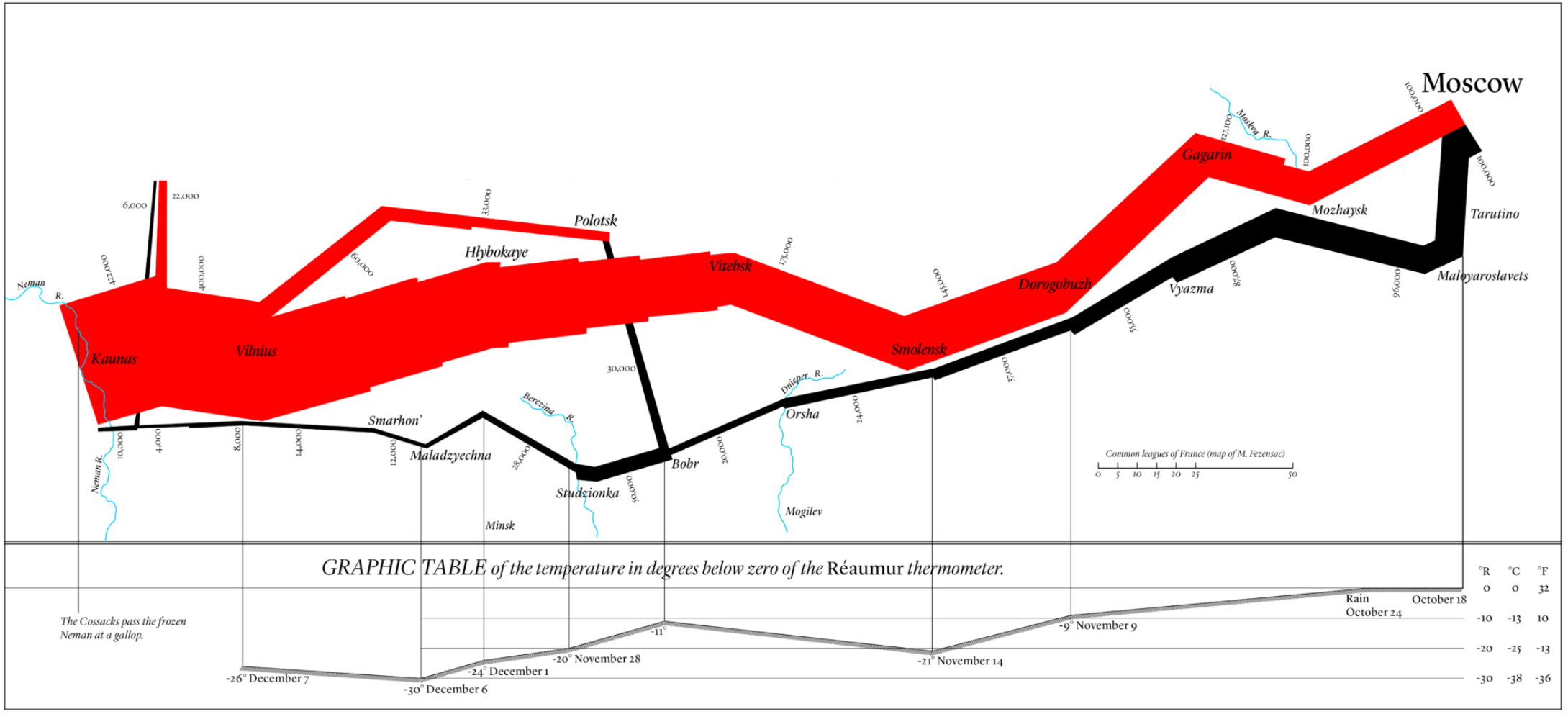
- Population genomics of 445 cultivated and wild lettuce lines.
- Generated a comprehensive variation map.
- Curated GP status.
- Identified selective sweeps and major genetic determinants of domestication traits.
- Revealed lettuce domestication history.



Nat Genet 2021, 53:752-760

BGI华大

May the infographics work for you



Charles Minard’s infographic of Napoleon’s 1812 march on Moscow

Questions?

华大生命科学研究院
BGI · research