Занятие 2
Популяционная генетика

Татьяна Татаринова

Так что же такое GWAS?

Исследование полногеномной ассоциации

Ищем SNP…

связанный с фенотипом.

<u>Цель:</u>

Объяснять

Понимание

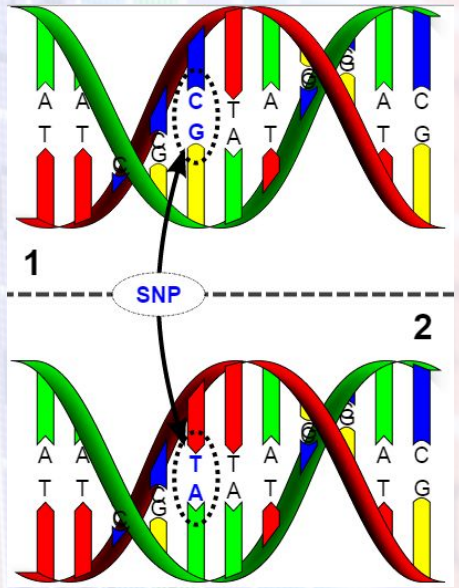
Механизмы

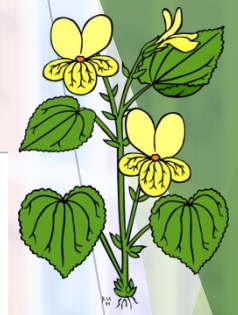
Терапия

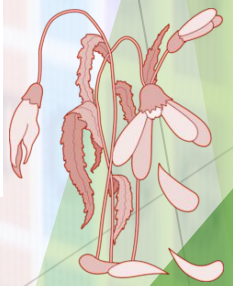
Предсказывать

Вмешательство

Профилактика

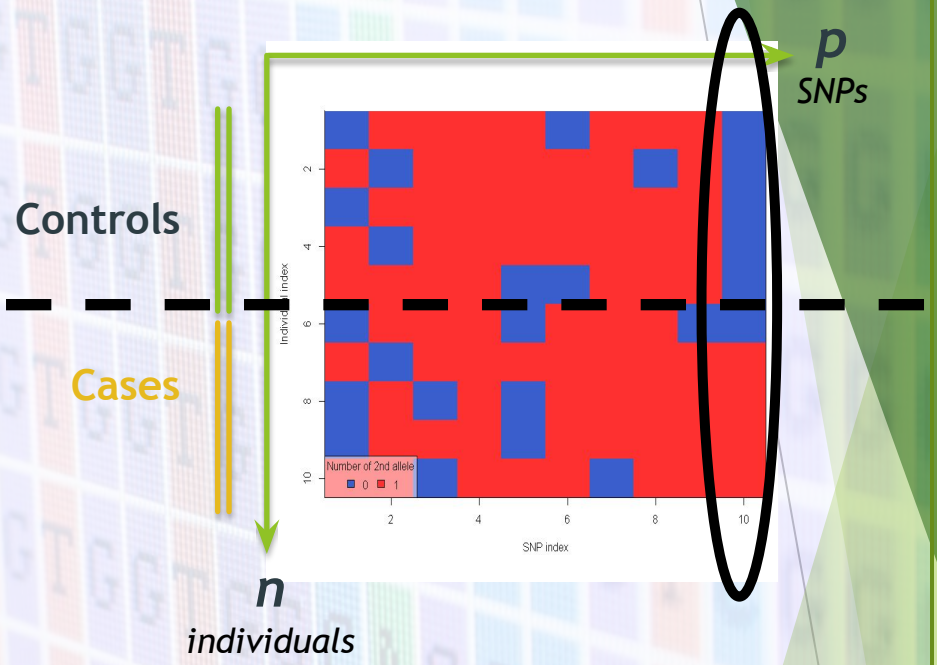


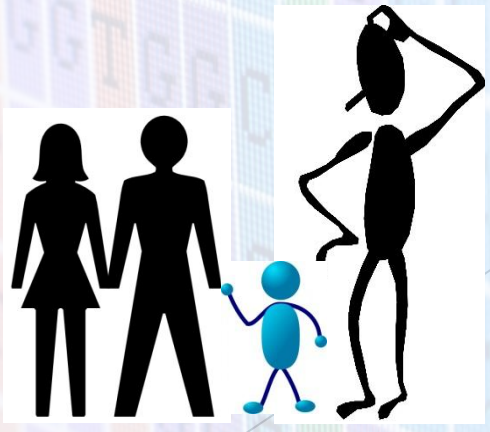




Ассоциация

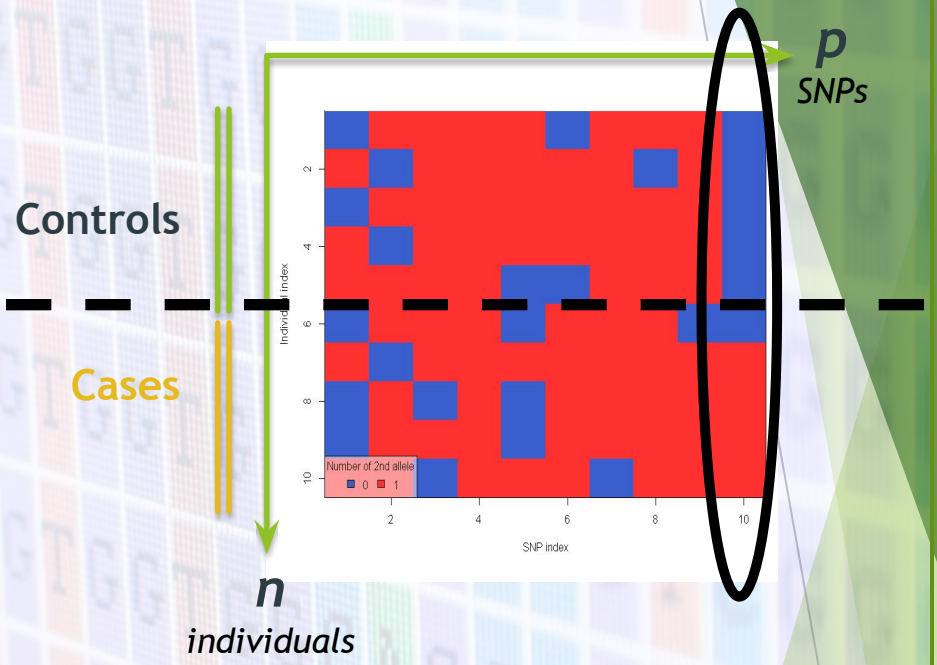
- Определение
- Любая связь между двумя измеренными величинами, которая делает их статистически зависимыми.
- Наследственность
- Доля дисперсии, объясняемая генетикой
- P = G + E + G * E
- Наследственность> 0





Ассоциация

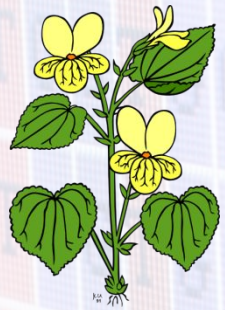
- Определение
- Любая связь между двумя измеренными величинами, которая делает их статистически зависимыми.
- Наследственность
- Доля дисперсии, объясняемая генетикой
- P = G + E + G * E
- Наследственность> 0

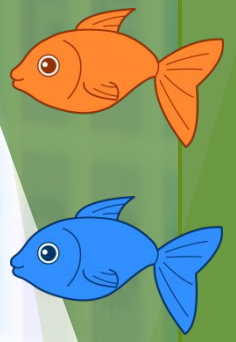


Почему?

Окружающая среда, взаимодействие генов и окружающей среды
Сложные черты, небольшие эффекты, редкие варианты
Уровни экспрессии генов
Методология GWAS?

Дизайн GWAS





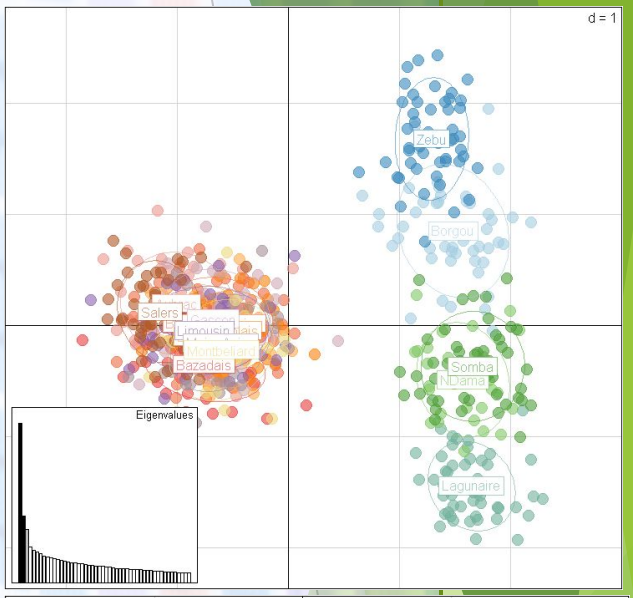
- Кейс-контроль
- Четко определенный «случай»
- Известная наследственность
-

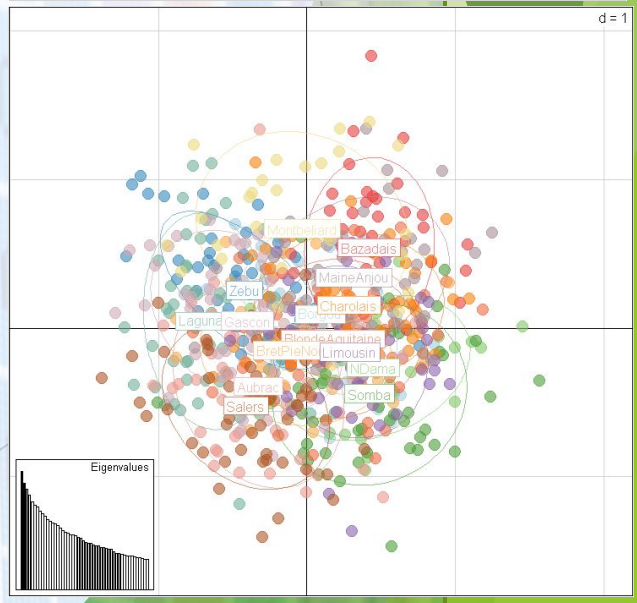
- Вариации
- Количественные фенотипические данные
- Например: Рост, концентрация биомаркеров
- Явные модели
- Например. Доминантный или рецессивный

Стратификация населения II

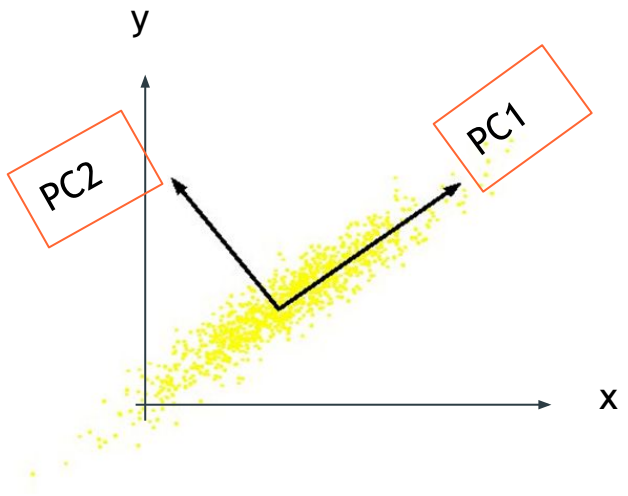
Процесс

- Визуализация
  - Филогенетика
  - PCA
- Коррекция данных
  - Геномный контроль
  - Регрессия по основным компонентам PCA





Что же там, в наших данных? Применим PCA - Метод главных компонент
Для начала рассмотрим простые примеры



Метод основных компонент - подбор новой системы координат, которая позволяет описать наши многомерные данные меньшим количеством переменных.

Почему это возможно: многие из измерений зависят друг от друга. Например, у вас есть данные о легковых машинах, и расход бензина указан как в милях на галлон, так и в километрах на литр. Корреляция в этом случае идеальная.

Поэтому на первом шаге PCA считается матрица корреляций. Потом для этой матрицы вычисляются собственные вектора и собственные значения (линейная алгебра!). Собственный вектор, соответствующий наибольшему собственному значению совпадает с направлением наибольшего изменения.

Новая ось PC1 соответствует направлению наибольшего изменения в данных.

Пример: планеты Солнечной системы

Планета	Расстояние до солнца	Диаметр	Плотность
Меркурий	0.387	4878	5.42
Венера	0.723	12104	5.25
Земля	1	12756	5.52
Марс	1.524	6787	3.94
Юпитер	5.203	142800	1.314
Сатурн	9.539	120660	0.69
Уран	19.18	51118	1.29
Нептун	30.06	49528	1.64
Плутон	39.53	2300	2.03

Как вы думаете, есть какая корреляция между колонками?

Пример: планеты Солнечной системы

Планета	Расстояние до солнца	Диаметр	Плотность
Меркурий	0.387	4878	5.42
Венера	0.723	12104	5.25
Земля	1	12756	5.52
Марс	1.524	6787	3.94
Юпитер	5.203	142800	1.314
Сатурн	9.539	120660	0.69
Уран	19.18	51118	1.29
Нептун	30.06	49528	1.64
Плутон	39.53	2300	2.03

Корреляция между колонками?

	Расстояние до солнца	Диаметр	Плотность
Расстояние до солнца	1.00	-0.05	-0.59
Диаметр	-0.05	1.00	-0.71
Плотность	-0.59	-0.71	1.00

Три измерения явно излишни

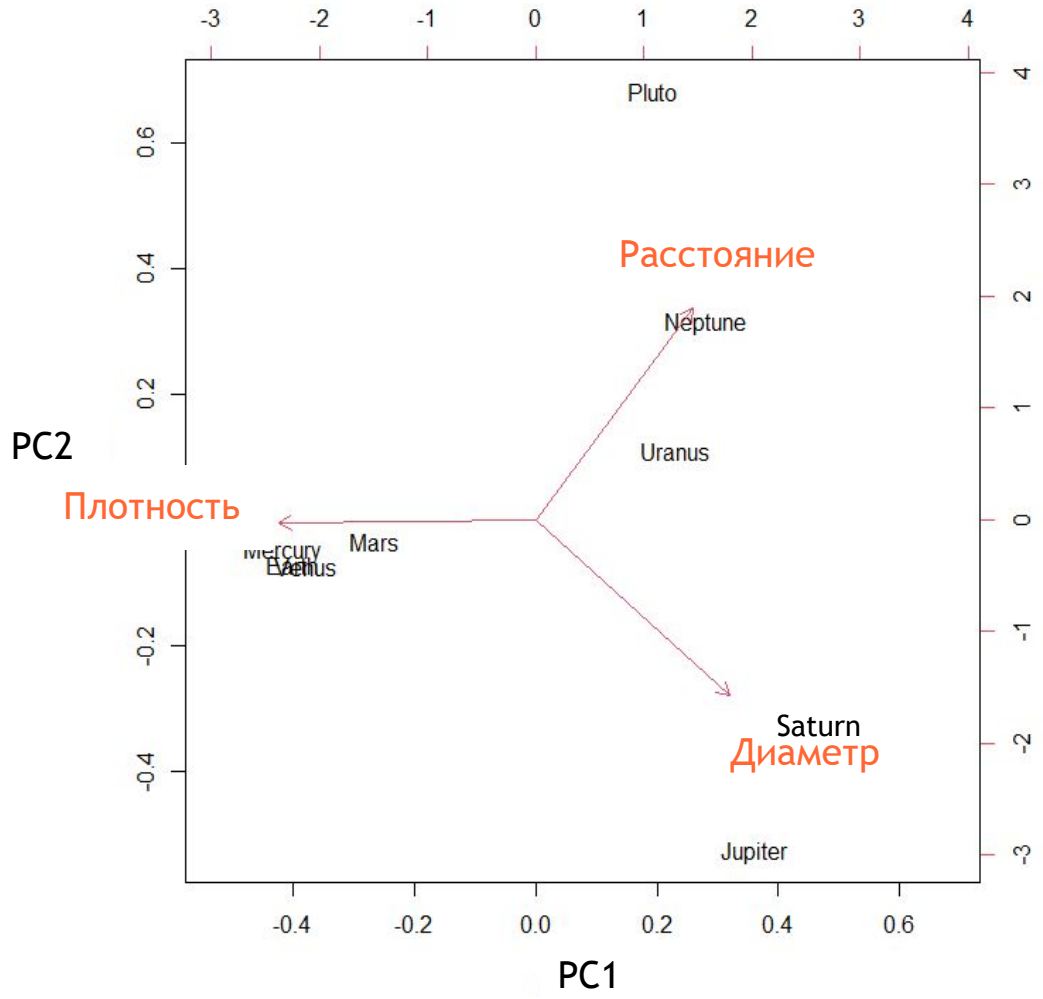
Что показывает PCA?

	PC1	PC2	PC2
Стандартное отклонение	1.379158	1.024481	0.219917
Пропорция дисперсии на компоненту	0.634025	0.349854	0.016121
Совокупная пропорция дисперсии	0.634025	0.983879	1

Сколько компонент оставить? Два подхода

1. Все со стандартными отклонениями больше 1
2. Совокупная пропорция дисперсии больше 0.9

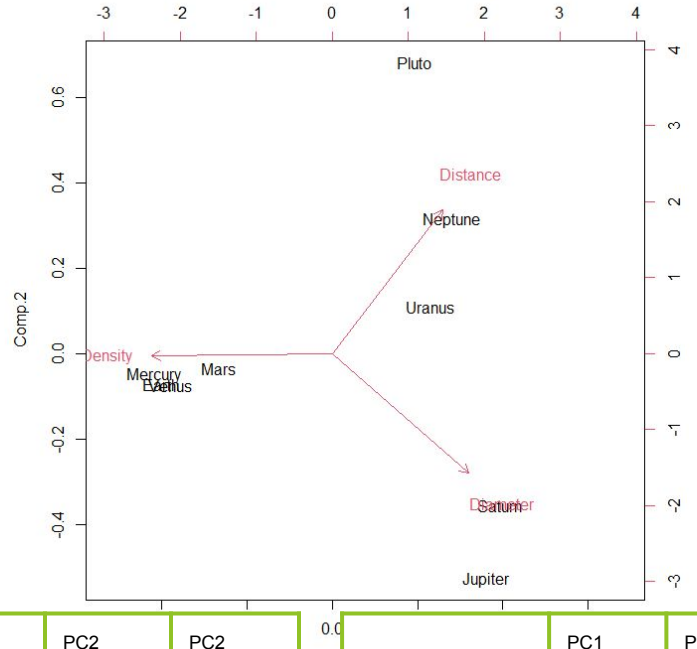
Результаты



	PC1	PC2	PC2
Расстояние	0.44	0.769	0.463
Диаметр	0.541	-0.639	0.547
Плотность	-0.716		0.698

Эффект нормализации: даем равный шанс разным группам измерений

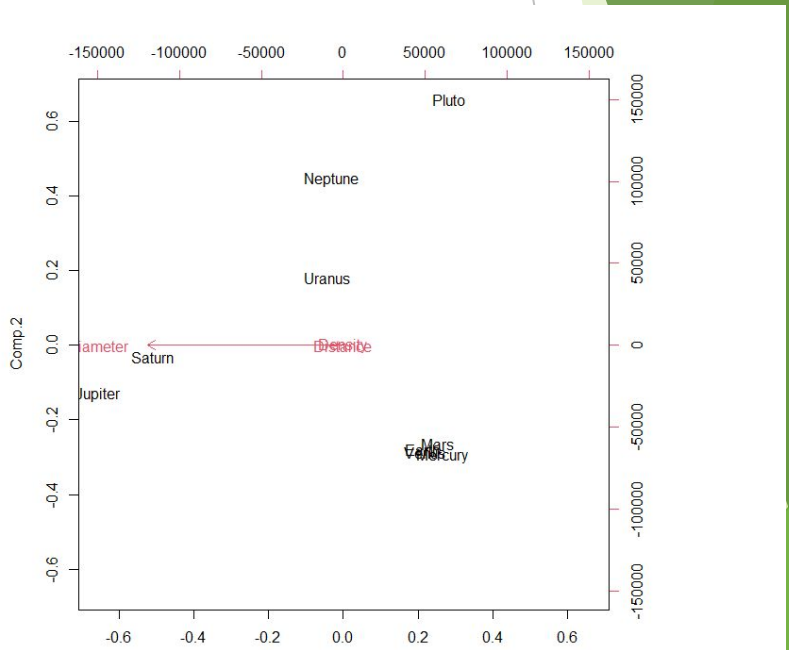
С нормализацией 2 компоненты



	PC1	PC2	PC2
Стандартное отклонение	1.38	1.02	0.22
Пропорция дисперсии	0.63	0.35	0.02
Совокупная пропорция	0.63	0.98	1.00

	PC1	PC2	PC2
Расстояние	0.44	0.769	0.463
Диаметр	0.541	-0.639	0.547
Плотность	-0.716		0.698

Без нормализации 1 главная компонента



	PC1	PC2	PC2
Стандартное отклонение	49846	13.70	0.58
Пропорция дисперсии	1.00	0.00	0.00
Совокупная пропорция	1.00	1.00	1.00

	PC1	PC2	PC2
Расстояние		0.996	
Диаметр	-1		
Плотность			0.996

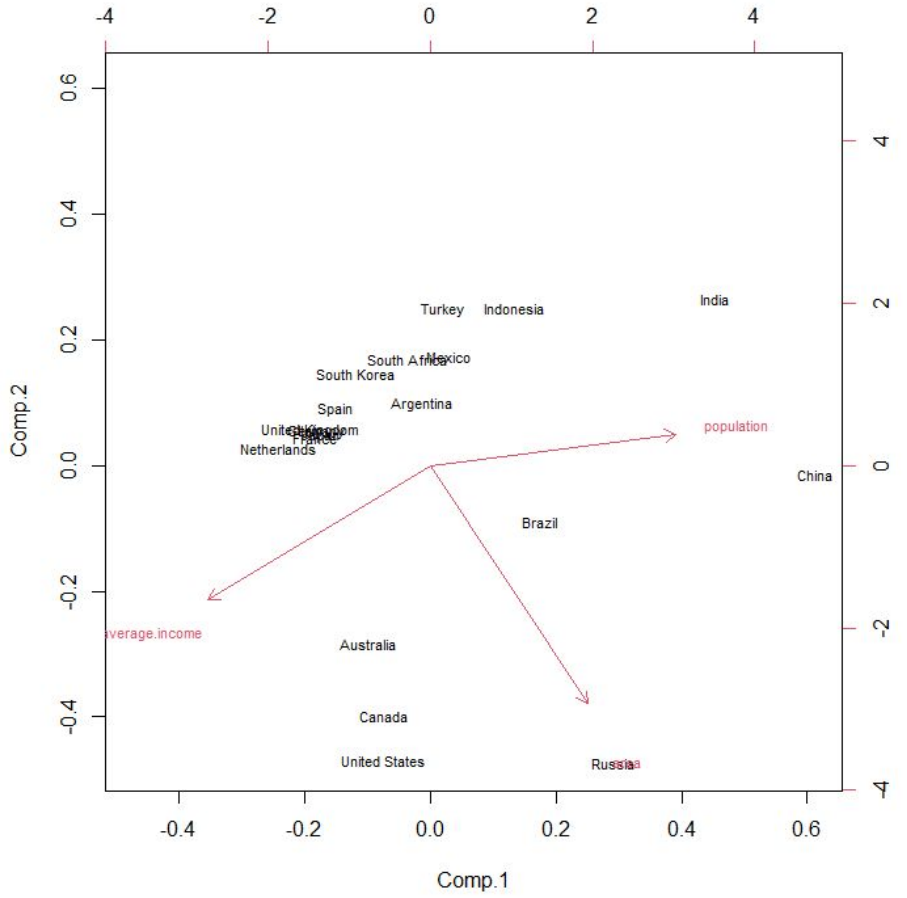
Теперь попробуйте сами

```
# загрузите файл с данными планет
# planets.csv
Planets=read.csv('Planets.csv', row.names = 1); Planets
cor(Planets)
pcaPlanets=princomp(Planets, cor=T)
summary(pcaPlanets)
biplot(pcaPlanets)
loadings(pcaPlanets)
```

Повторим упражнение с данными о странах мира

Данные

Страна	Население, мил	Доход на душу населения, долл	Площадь, км
United States	292.6	38600	9809431
China	1309	4910	9556100
Japan	127.8	27400	377801
India	1084	2850	3203975
Germany	82.75	26600	356955
United Kingdom	59.77	26700	244101
France	60.09	26500	547026
Italy	57.24	26500	301277
Brazil	183.3	7820	8511996
Russia	144.2	8620	17075400
Mexico	105.8	9170	1967183
Canada	32.09	29500	9970610
Spain	40.97	22400	504750
South Korea	48.02	18800	99016
Indonesia	230.9	3090	1948732
Australia	20.07	27900	7682300
South Africa	45.17	12200	1220662
Turkey	71.54	6800	779452
Netherlands	16.32	29800	41640
Argentina	38.64	12300	2780400



Вопросы:

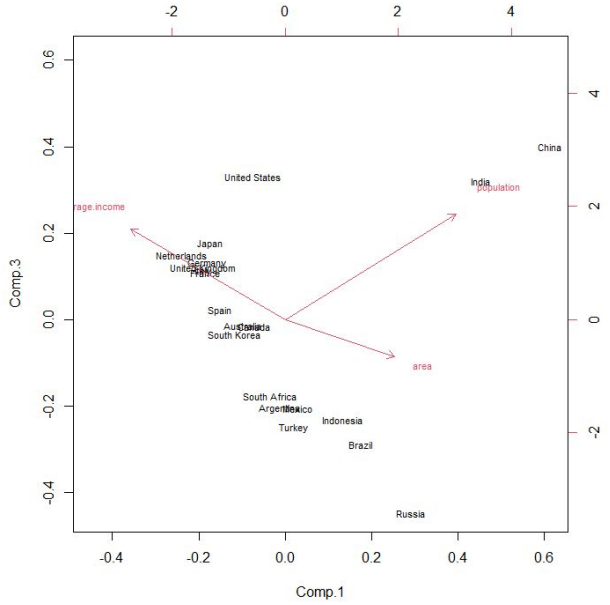
Нужно ли проводить нормализацию?

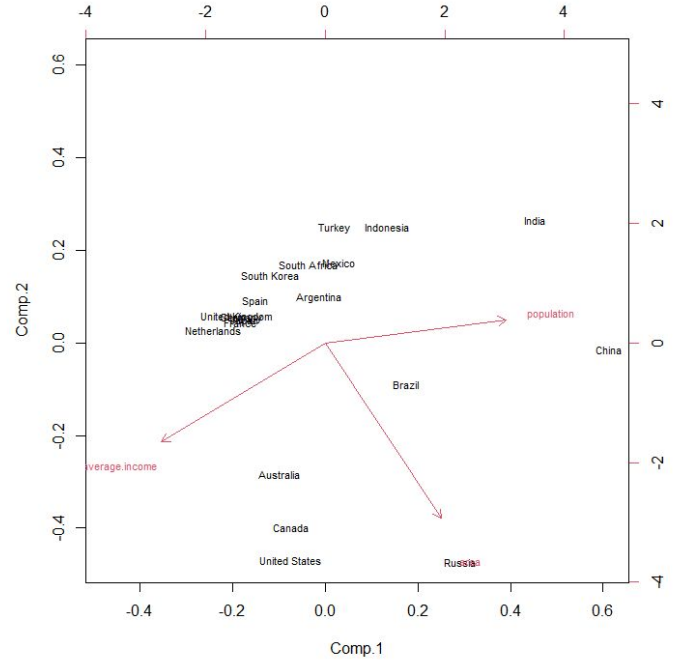
Сколько главных компонент нужно сохранить?

Результаты

Корреляция	Население	Доход на душу	Площадь
Население	1	-0.46375	0.274766
Доход на душу	-0.46375	1	-0.11911
Площадь	0.274766	-0.11911	1

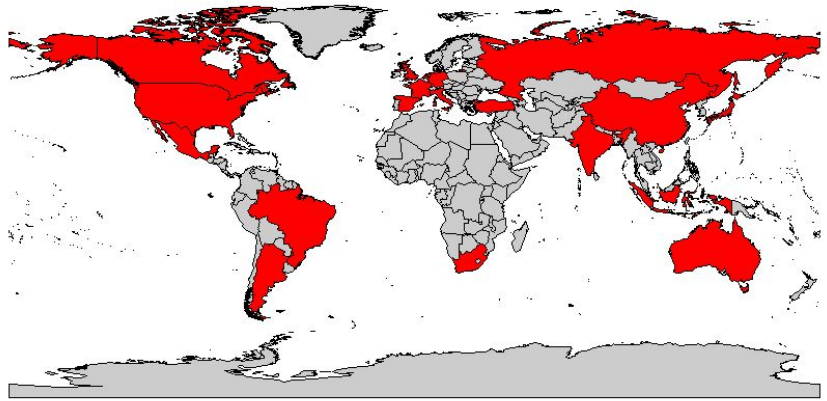
	PC1	PC2	PC3
Стандартное отклонение	1.263472	0.946988	0.711935
Пропорция дисперсии	0.53212	0.298929	0.168951
Совокупная пропорция	0.53212	0.831049	1





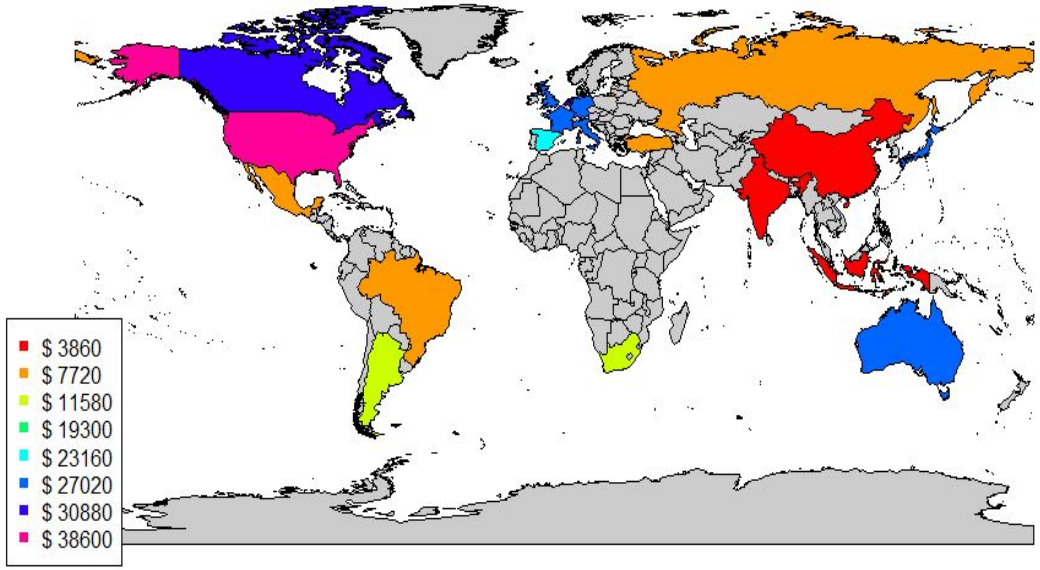
Простая карта

```
install.packages(“maptools”)
library(maptools)
data(wrld_simpl)
myCountries = wrld_simpl@data$NAME %in% row.names(Countries)
plot(wrld_simpl, col = c(gray(.80), "red")[myCountries+1])
```



Не совсем простая карта

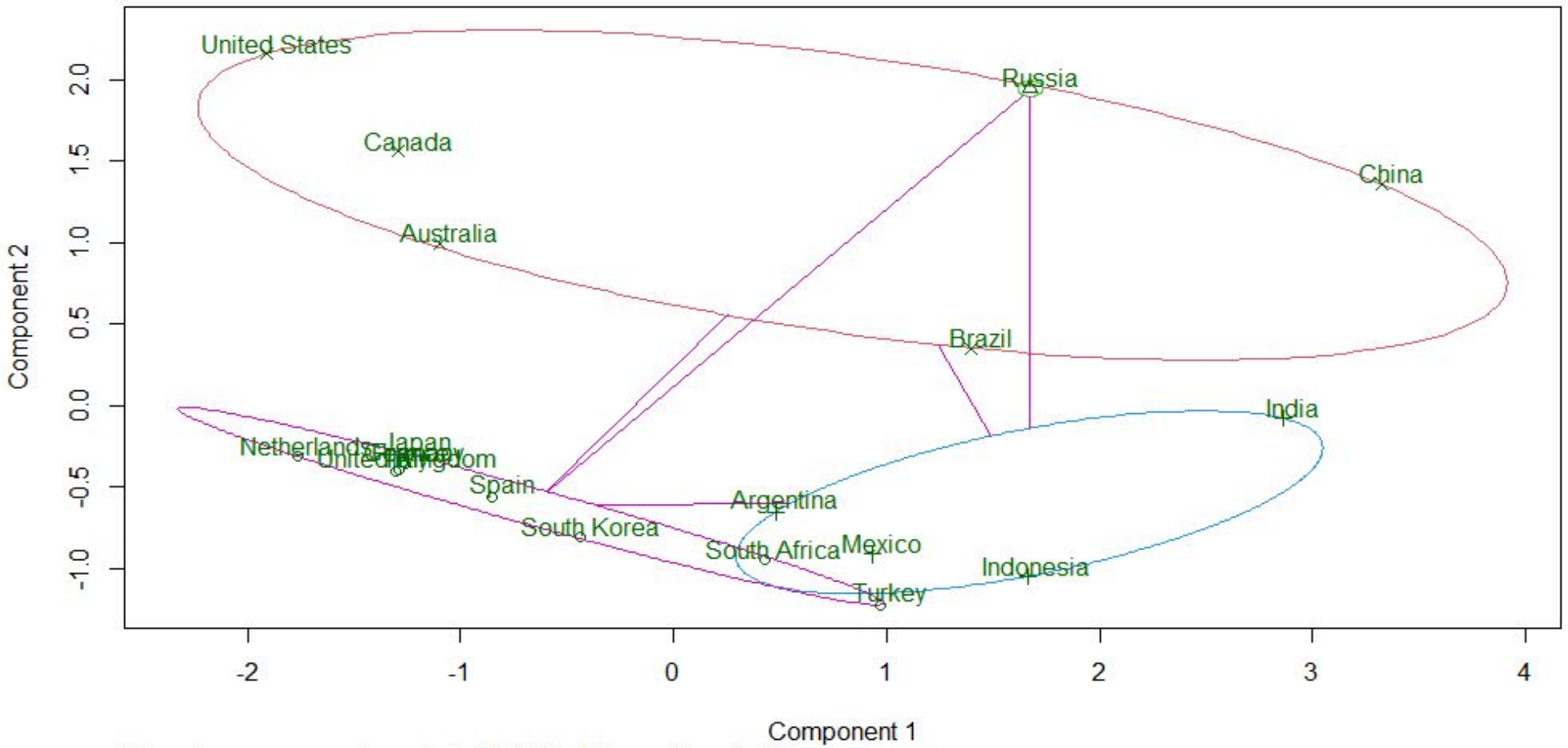
```
MAX_INCOME=max(Countries$average.income)
# добавили колонку
Countries$index=round(10*Countries$average.income/MAX_INCOME); Countries
COL=rainbow(10) # задали цвета
ALL_COUNTRIES=as.data.frame(cbind( as.character(wrld_simpl@data$NAME ), gray(.80)))
head(ALL_COUNTRIES) # создали новый объект
names(ALL_COUNTRIES)=c('Name', "Color")
row.names(ALL_COUNTRIES)=ALL_COUNTRIES$Name
for(i in 1:dim(Countries)[1]){
  ALL_COUNTRIES[row.names(Countries)[i],'Color']=COL[Countries[i,'index']]
}
head(ALL_COUNTRIES)# цвет зависит от дохода
unique(ALL_COUNTRIES$Color)# сколько категорий?
plot(wrld_simpl, col = ALL_COUNTRIES$Color) # нарисовали карту
# добавили легенду
sort(unique(Countries$index))*MAX_INCOME/10, COL[sort(unique(Countries$index))]
legend('bottomleft',
       legend = paste("$",sort(unique(Countries$index))*MAX_INCOME/10),
       col= COL[sort(unique(Countries$index))],  pch=15)
```



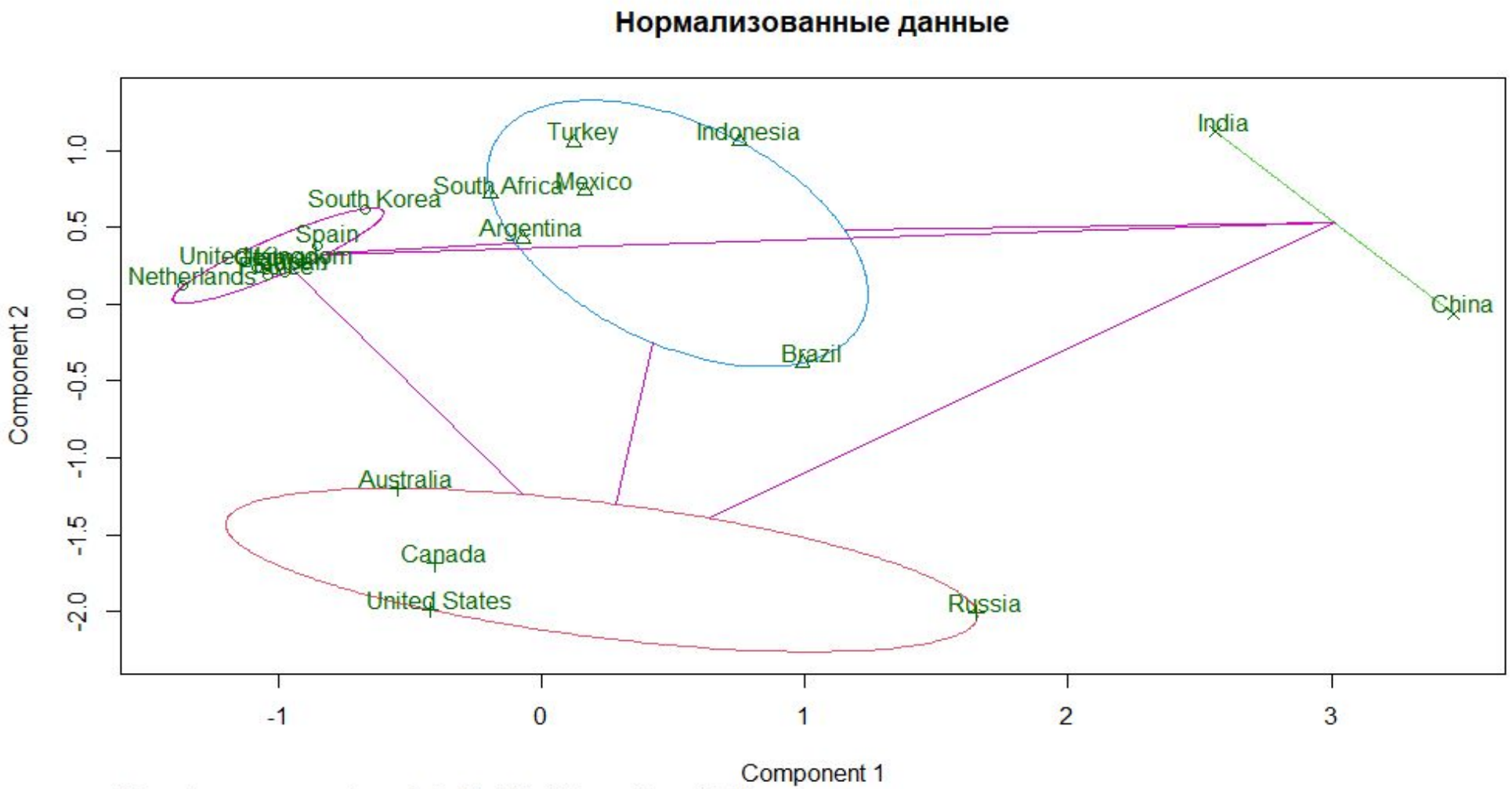
K-MEANS

```
install.packages(“cluster”)
library(cluster)
Countries_clusters=kmeans(Countries[,1:3], centers=4, nstart=25)
clusplot(Countries, Countries_clusters$cluster,labels=3, color=TRUE)

ccs=data.frame(sapply(Countries[,1:3], scale)) ##нормализация
rownames(ccs)=rownames(Countries)
Countries_clusters_scaled=kmeans(ccs, centers=4, nstart=25)
clusplot(ccs, Countries_clusters_scaled$cluster,labels=3, color=TRUE)
```

Сырые данные



These two components explain 84.97 % of the point variability.


Нормализованные данные
Turkey
Indonesia
India
South Africa
Mexico
South Korea
Spain
Argentina
Netherlands
China
Brazil
Australia
Canada
United States
Russia
Component 2
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-1
0
1
2
3
Component 1
These two components explain 83.1 % of the point variability.

Возвращаемся к нуклеотидам Скачайте данные в вашу рабочую директорию

- sativas413.ped
- sativas413.fam
- sativas413.map
- sativas413.pheno
- sativas413.csv

Библиотеки

```
#install packages
install.packages(c("poolr","qqman","BGLR","rrBLUP","DT", "dplyr"))
install.packages(c("rnaturalearth",'rnaturalearthdata','rgeos','ggspatial'))
devtools::install_github("dkahle/ggmap", ref = "tidyup")
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("SNPRelate")
```

Библиотеки

```
library(rrBLUP)
library(BGLR)
library(DT)
library(SNPRelate)
library(dplyr)
library(qqman)
library(poolr)
library(OpenStreetMap)
library(rjson)
```

```
library(rgdal)
library(RgoogleMaps)
library(mapproj)
library(sf)
library(OpenStreetMap)
library(ggplot2)
library(sf)
library(rnaturalearth)
library(rnaturalearthdata)
```

Шаг 1. Подготовка данных SNP в R

```
rm (список = ls ())
setwd («# ваш рабочий каталог, содержащий файл»)
Geno<- read_ped ("sativas413.ped")
head(Geno)
# объединяет данные маркерного аллеля в формате 0, 1, 2 и 3; 2 представляет отсутствующие данные
p = Geno$p; p
n = Geno$n; n
Geno = Geno$x; Geno
# Информация о присоединении
FAM <- read.table("sativas413.fam"); head( FAM )
# Информация о положении снипов на геноме
MAP <-  read.table("sativas413.map"); head( MAP )
# Перекодировать данные в файл ped
Geno[Geno == 2] <- NA # Преобразование отсутствующих данных в NA
Geno[Geno == 0] <- 0 # Преобразование 0 данных в 0
Geno[Geno == 1] <- 1 # Преобразование 1 в 1
Geno[Geno == 3] <- 2 # Преобразование 3 в 2
# Преобразование данных маркера в матрицу, транспонирование и проверка размера
Geno <- matrix (Geno, nrow = p, ncol = n, byrow = TRUE)
Geno<- t (Geno)
dim(Geno)
```

0	00 Homozygote "1"/"1"
1	01 Heterozygote
2	10 Missing genotype
3	11 Homozygote "2"/"2"

Шаг 2. Считайте данные фенотипа в R

```
## прочитать фенотип
ris.pheno <- read.table ("sativas413.pheno",
                         header = TRUE, stringsAsFactors = FALSE, sep = "\t")
# Просмотр первых нескольких столбцов и строк данных
ris.pheno [1: 5, 1: 5]
dim( ris.pheno)
# сравнить с фенотипическим файлом
rownames (Geno) <- FAM$V2; head(Geno)
table(rownames (Geno) == ris.pheno$NSFTVID)
# Теперь давайте извлечем первый фенотип припишем его объекту y
y <- matrix (ris.pheno$Flowering.time.at.Arkansas) # использовать первый фенотип
rownames (y) <- ris.pheno$NSFTVID
index<-!is.na (y)
y <- y [index, 1, drop = FALSE] # 374
Geno <- Geno [index,] # 374 x 36901
table(rownames (Geno) == rownames (y))
```

Шаг 3. Фильтрация данных SNP


```
# Здесь мы будем использовать цикл for, чтобы определить недостающие и преобразовать их в средние значения по столбцам.
for (j in 1: ncol (Geno)) {
  Geno [, j] <- ifelse (is.na (Geno [, j]), mean (Geno [, j], na.rm = TRUE), Geno [, j])
}
# Фильтр редких аллелей. Здесь мы будем убирать аллели с частотой <5% и сохранять объект как Geno1
# Затем мы сопоставим и отбросим маркеры из файла информации о карте и сохраним его как MAP1
p <- colSums (Geno) / (2 * nrow (Geno))
maf <- ifelse (p> 0.5, 1 - p, p)
maf.index <- который (maf <0,05)
Geno1 <- Geno [, -maf.index]
dim(Geno1)
dim(Geno)
# подмножество на основе сохраненных SNP
MAP1 <- MAP [-maf.index,]; dim(MAP1)
```

Шаг 4. Структура популяции

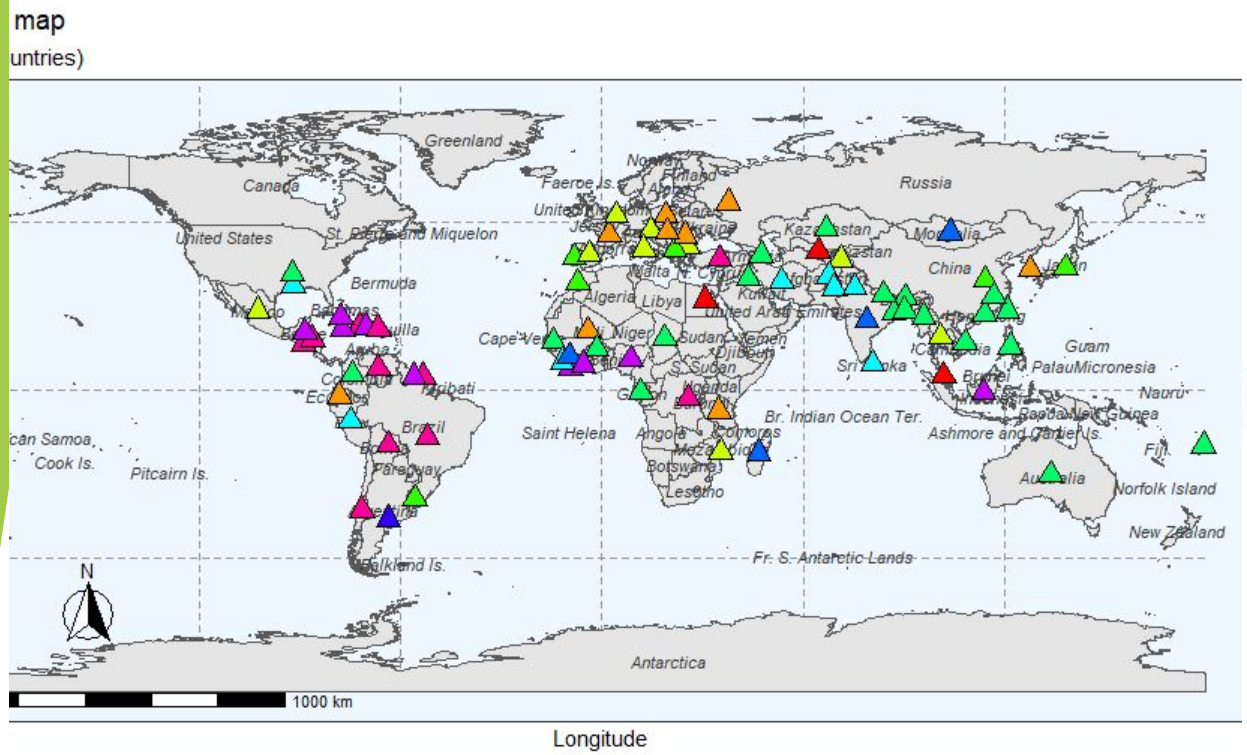
```
# Создать файл геноматрицы и назначить имена строк и столбцов из файлов fam и map
Geno1 <- as.matrix (Geno1)
sample<- row.names (Geno1)
length(sample)
colnames(Geno1) <- MAP1 $ V2
snp.id <- colnames (Geno1)
length(snp.id)
snpgdsCreateGeno ("44k.gds", genmat = Geno1, sample.id = sample, snp.id = snp.id,
                  snp.chromosome = MAP1 $ V1, snp.position = MAP1 $ V4, snpfirstdim = FALSE)
# Теперь откройте файл 44k.gds
geno_44k <- snpgdsOpen ("44k.gds")
snpgdsSummary ("44k.gds")
```

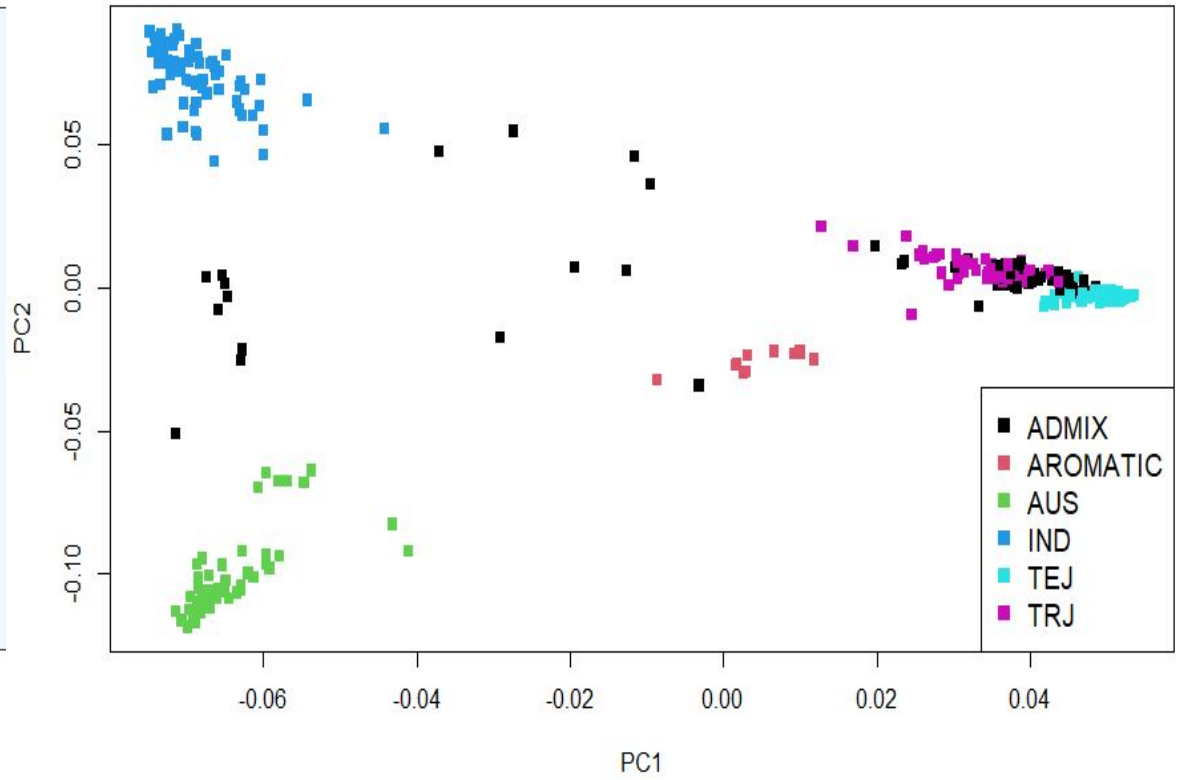
Шаг 5. PCA

```
#PCA анализ
pca <- snpgdsPCA (geno_44k, snp.id = colnames (Geno1))
# график результатов PCA
pca <- data.frame (sample.id = row.names (Geno1), EV1 = pca $ eigenvect [, 1], EV2 = pca $ eigenvect [, 2], EV3 = pca $ eigenvect [, 3], EV4 = pca $ eigenvect [, 4], stringsAsFactors = FALSE)
plot(pca $ EV2, pca $ EV1, xlab = "собственный вектор 3", ylab = "собственный вектор 4")
# добавить информацию о населении на участок
pca_1 <- read.csv ("sativas413.csv", header = TRUE, stringsAsFactors = FALSE)
pca_2 <- pca_1 [match (pca $ sample.id, pca_1 $ NSFTV.ID),]
# Извлечь информацию о населении и добавить выходной файл PCA
pca_population <- cbind (pca_2 $ Sub.population, pca)
colnames (pca_population) [1] <- "Population"
# Постройте и добавьте названия населения
plot(pca_population $ EV1, pca_population $ EV2, xlab = "PC1", ylab = "PC2", col = c (1: 6) [factor (pca_population $ Population)])
legend(x = "right", legend= levels(factor (pca_population $Population)), col = c (1: 6), pch = 1, cex = 0.6)
```

Шаг 5a. Визуализация PCA

```
# Извлечь информацию о местоположении и добавить выходной файл PCA
pca_country <- as.data.frame(cbind(as.numeric(pca$EV1),as.numeric(pca$EV2),as.numeric(pca$EV3),
                                   as.numeric(pca$EV4),as.numeric(pca_2$Latitude),
                                   as.numeric(pca_2$Longitude)));head(pca_country)
names(pca_country)=c('PC1','PC2','PC3','PC4','LAT','LONG')
names(pca_country)
pca_country=na.omit(pca_country)
# корреляция с географией
cor (pca_country [, 1: 6], use= 'pairwise.complete.obs')
#map
map()
# определить цвета для карты
NCOLS = 1+max (unique (round (50 * (pca_country $ PC1 - min(pca_country $ PC1)))))
                          NCOLS
COL = rainbow(NCOLS)
pca_country$ind1=round (50 * (pca_country $ PC1 -min((pca_country $ PC1)))) +1
points(pca_country$LONG,pca_country$LAT, col=COLS[pca_country$ind1], pch=16)
```

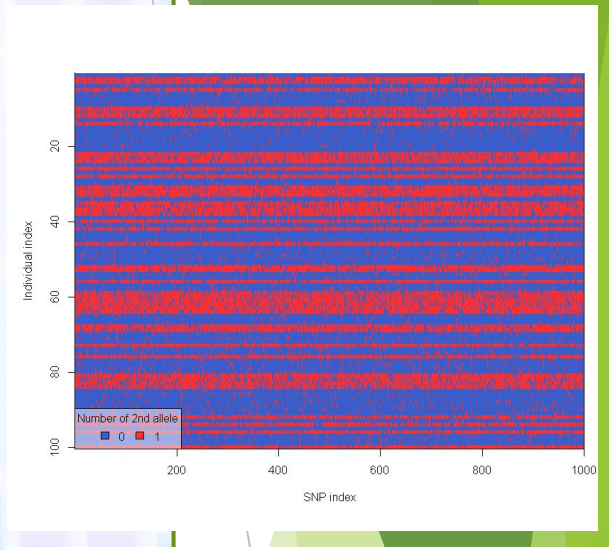


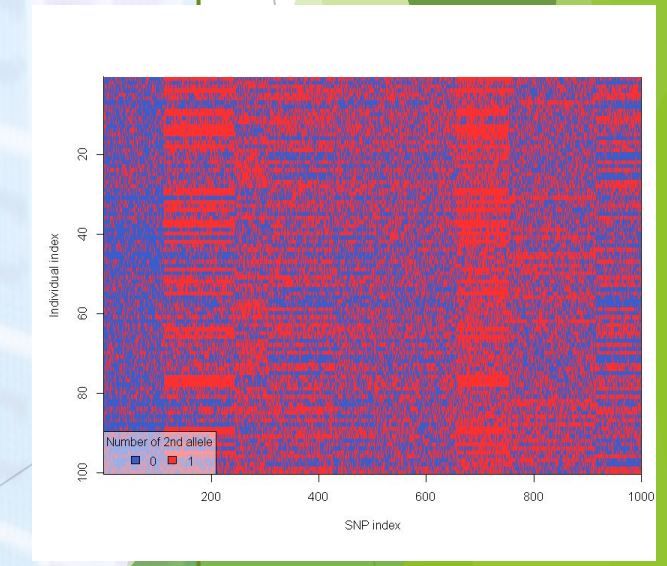


Анализ PCA: цвет по населению и по географии

Нарушение равновесия по сцеплению (LD)

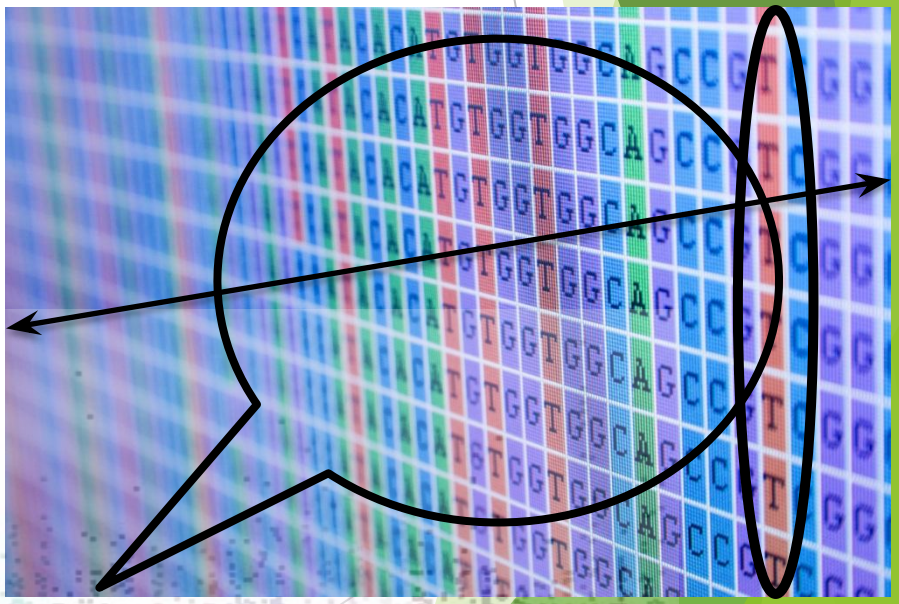
- Аллели в отдельных локусах **зависимы** друг от друга
- Проблема? Да и нет
- Слишком много LD - проблема ▫ шум >> сигнал
- Некоторые (предсказуемые) LD могут быть полезны ▫ позволяет использовать «маркерные» SNP





Методы тестирования ассоциации

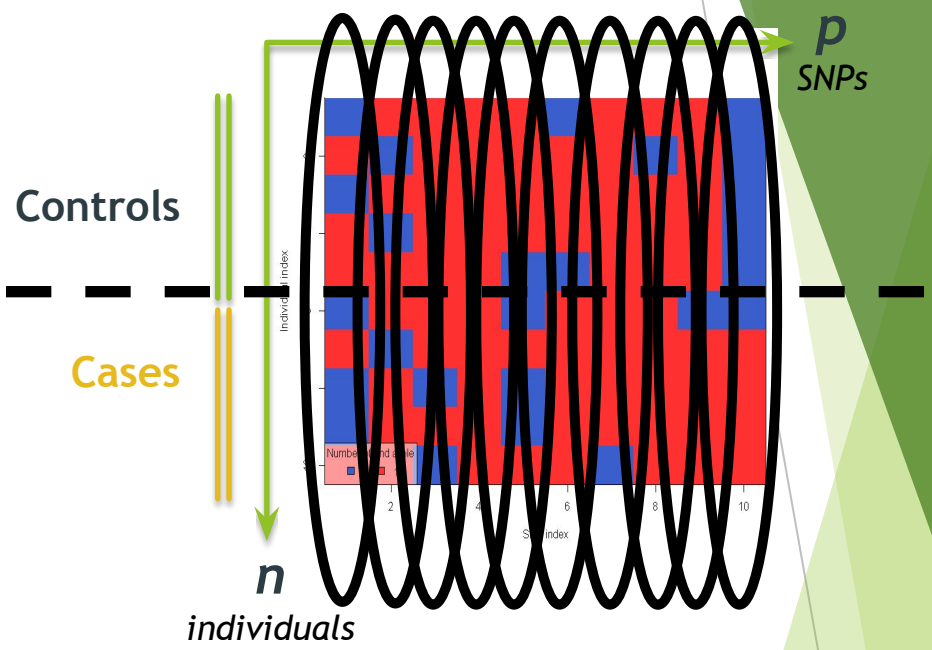
- Стандартный GWAS
  - Одномерные методы
- Использования взаимодействий
  - Многовариантные методы
    - Методы штрафной регрессии (LASSO)
    - Факториальные методы (ФС на основе DAPC)



Одномерные методы

Статистика индивидуальных тестов

Поправка на множественное тестирование



Вариации

Тестирование

Точный критерий Фишера, критерий тенденции Кохрана-Армитиджа, критерий хи-квадрат, дисперсионный анализ

Золотой стандарт - точный тест Фишера

Исправление

Бонферрони

Золотой стандарт - FDR

Одномерный GWAS- сильные и слабые стороны

Сильные стороны

Простота

Вычислительно быстро

Консервативный

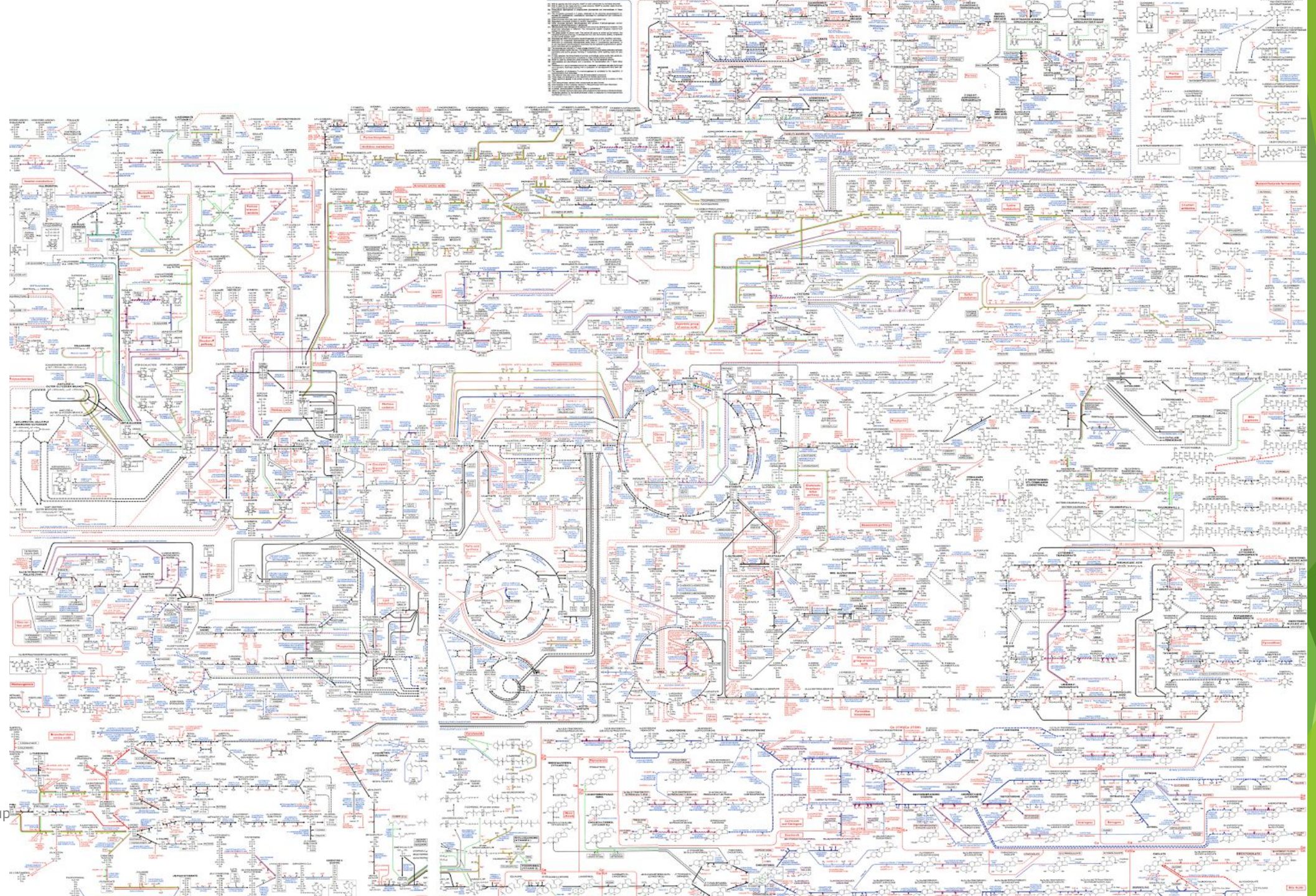
Легко интерпретировать

Недостатки

Многомерная система, одномерная структура

Размер эффекта отдельных SNP может быть слишком маленьким.

Предельные эффекты отдельных SNP ≠ комбинированные эффекты



Тестир

Многовариантные методы

Штрафная регрессия

Штрафная регрессия LASSO

Эластичная сетка

Регрессия хребта (ридж Регрессия)

Байесовские подходы

Байесовское разбиение

Байесовская логистическая регрессия с выбором переменной случайного поиска

Сопоставление байесовской ассоциации эпистаза

Факторные методы

Sparse-PCA

Метод многофакторного уменьшения размерности

Контролируемый PCA

ФС на основе DAPC (snpzip)

Отношение шансов-на основе MDR

Тестирование на ассоциацию

Нейронные сети

Нейронные сети, оптимизированные для генетического программирования

Параметрический убывающий метод

Логические деревья

Выбор логической функции

Монте-Карло Логическая регрессия

Логическая регрессия

Модифицированное программирование экспрессии генов логической регрессии

Установить связь подход

Генетическое программирование для ассоциативных исследований

Непараметрические методы

Случайные леса

Ограниченный метод разделения

Комбинаторный метод разбиения

Ридж регрессия (Метод регуляризации Тихонова) с использованием пакета rrBLUP

Ридж регрессия BLUP (Meuwissen et al. 2001)

https://rdrr.io/cran/rrBLUP/man/GWAS.html

Модель маркерных эффектов

Y - вектор данных

μ - общее среднее

1 - вектор единиц

Xi - матрица расчета

g - генетический эффект i-го маркера

e - ошибка

GWAS $y = \mu 1 + \sum Xi\, gi + e$

$$y_i = \beta_0 + \beta_1 w_i + \beta_2 x_i + \varepsilon_i$$

Design matrix

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} = \begin{bmatrix} 1 & w_1 & x_1 \\ 1 & w_2 & x_2 \\ 1 & w_3 & x_3 \\ 1 & w_4 & x_4 \\ 1 & w_5 & x_5 \\ 1 & w_6 & x_6 \\ 1 & w_7 & x_7 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \end{bmatrix}$$

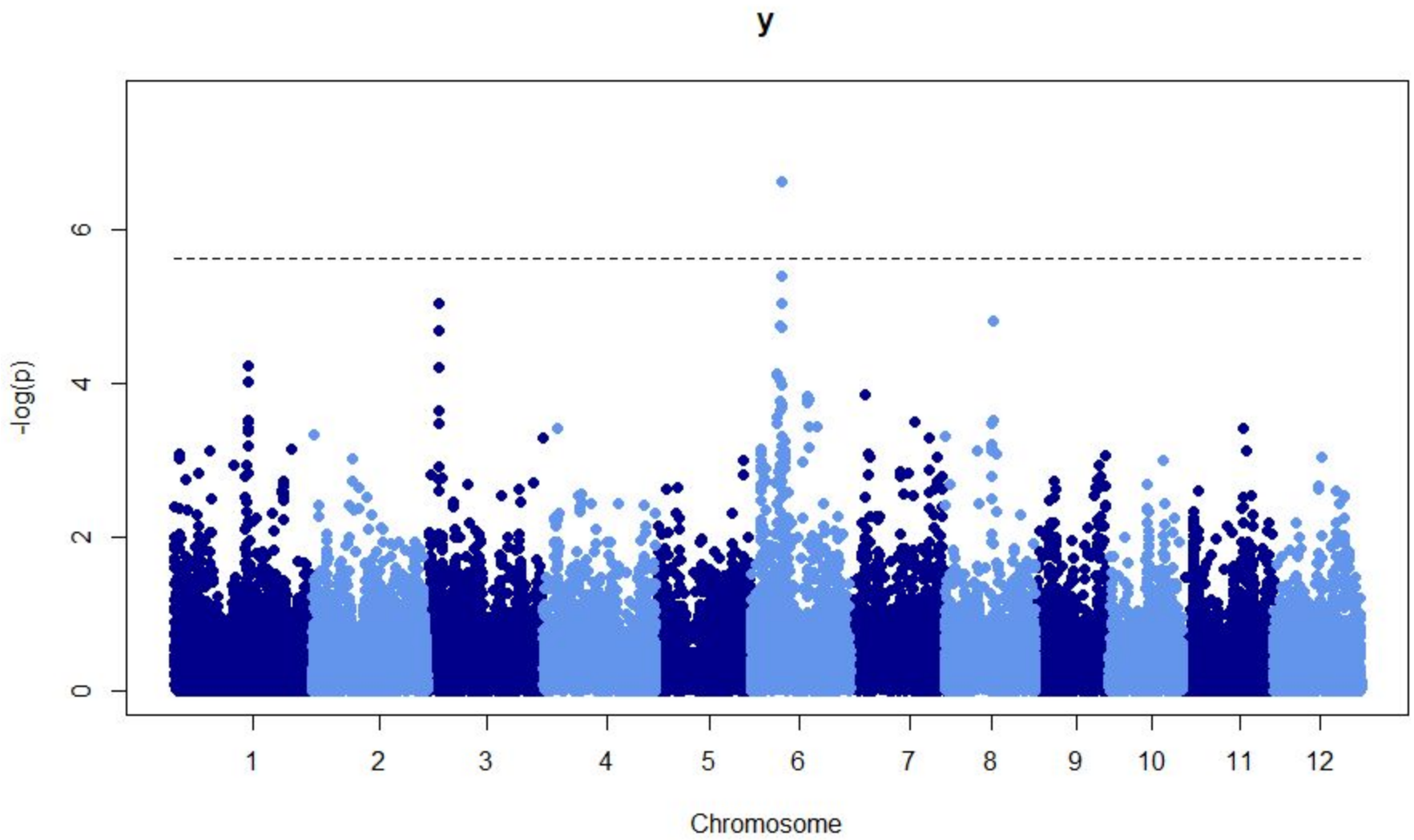
Ридж регрессия - это способ создания модели, когда количество переменных-предикторов в наборе превышает количество наблюдений (ситуация с недостаточным объемом данных).

Код R: установите и загрузите необходимые библиотеки

```
install.packages (c («poolr», «qqman», «BGLR», «rrBLUP», «DT», «SNPRelate», «dplyr», «RgoogleMaps», «ggmap», «mapproj», «sf», "OpenStreetMap", "инструменты разработчика"))
install.packages (c ("rnaturalearth", 'rnaturalearthdata', 'rgeos', 'ggspatial'))
library(rrBLUP)
library(BGLR)
library(DT)
library(SNPRelate)
library(rgeos)
library(ggspatial)
library(qqman)
library(пуллер)
library(OpenStreetMap)
library(rjson)
library(rgdal)
library(RgoogleMaps)
library(mapproj)
library(sf)
library(dplyr)
Библиотека (инструменты разработчика)
devtools :: install_github ("dkahle / ggmap",
   ref = "tidyup")
```

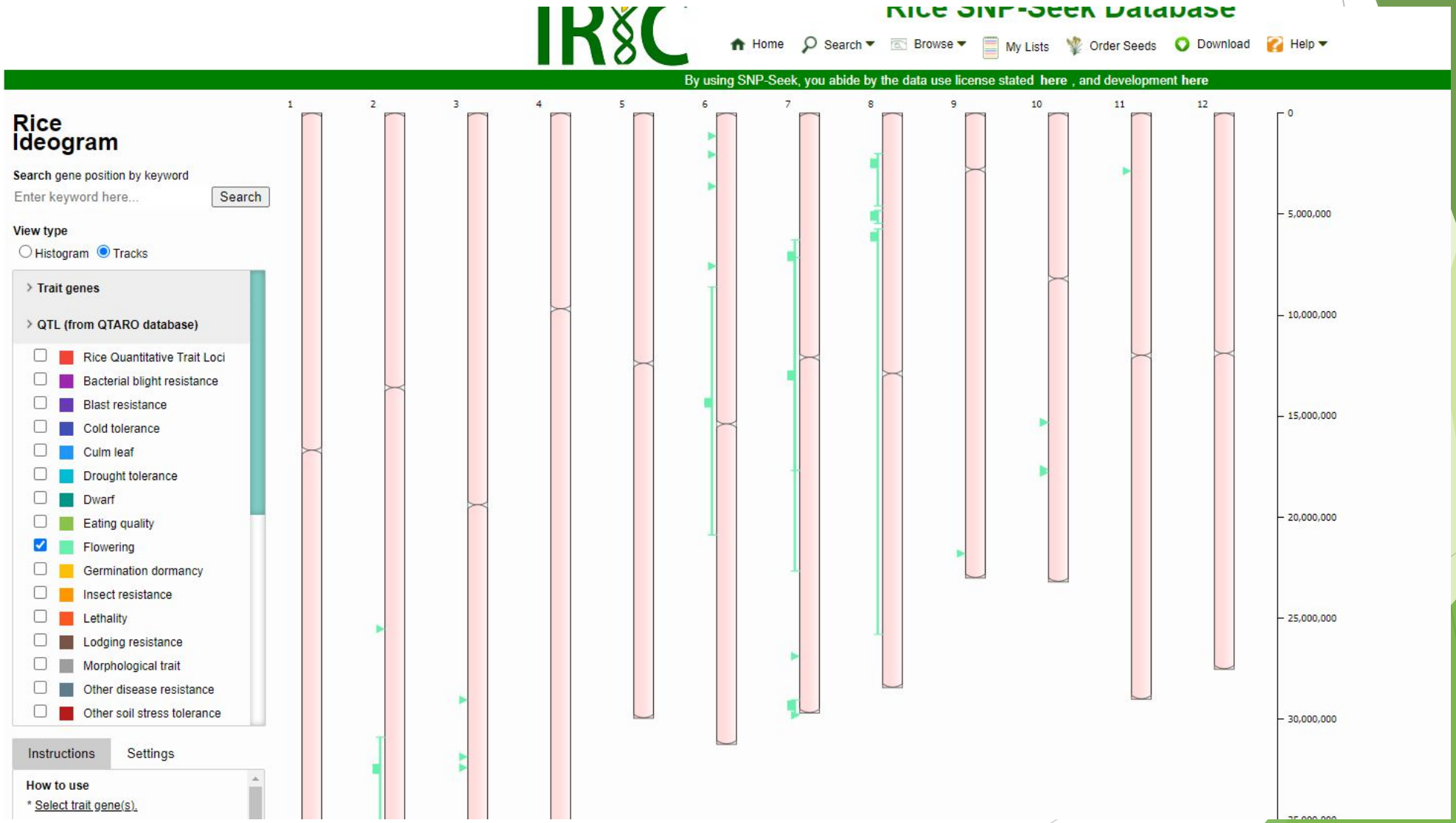
GWAS

```
# создаем файл geno для анализа GWAS пакета rrBLUP

geno_final <- data.frame (marker = MAP1 [, 2], chrom = MAP1 [, 1], pos = MAP1 [, 4], t (Geno1 - 1), check.names = FALSE)

dim(Geno1)

# создаем фенофайл

pheno_final <- data.frame (NSFTV_ID = rownames (y), y = y)

# Запустить анализ GWAS

myGWAS <- GWAS (pheno_final, geno_final, min.MAF = 0,05, P3D = TRUE, plot = TRUE)
```


y
-log(p)
0
2
4
6
1
2
3
4
5
6
7
8
9
10
11
12
Chromosome

https://snp-seek.irri.org/

Использовать поиск по SNP



Есть много новых инструментов

Однако, если вы освоите старый добрый rrBLUP, это поможет понять другие.

Теперь упражнения

Посмотрите на доступные фенотипы, выберите один и повторите упражнения.

"Flowering.time.at.Aberdeen" "Flowering.time.at.Faridpur" "Flowering.time.at.Aberdeen" "FT.ratio.of.Arkansas.Aberdeen" "FT.ratio.of.Faridpur.Aberdeen" " Culm.habit "" Лист. Опушение Flag.leaf.length "" Flag.leaf.width "" Awn.presence "

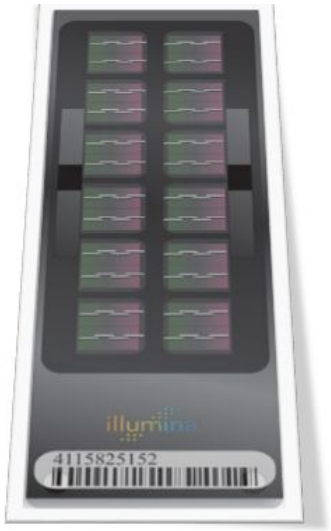
«Число метелок на растение» «Высота растения» «Длина метелки» «Первичное число ветвей» «Число семян на одну ветвь» «Цветки на одну ветвь» «Метелка. Плодородие» Seed.length "" Seed.width "" Seed.volume "" Seed.surface.area "" Brown.rice.seed.length "" Brown.rice.seed.width "" Brown.rice.surface.area "" Brown .rice.volume "" Seed.length.width.ratio "" Brown.rice.length.width.ratio "" Seed.color "" Pericarp.color "" Straighthead.suseptability "" Blast.resistance "

"Amylose.content" "Alkali.spreading.value" "Protein.content"

"Year07Flowering.time.at.Arkansas" "Year06Flowering.time.at.Arkansas"

Подход к прогнозированию географической структуры населения (GPS)

Сделать вывод о происхождении человека из полногеномной коллекции маркеров, информативных о происхождении.



От SNP к добавке

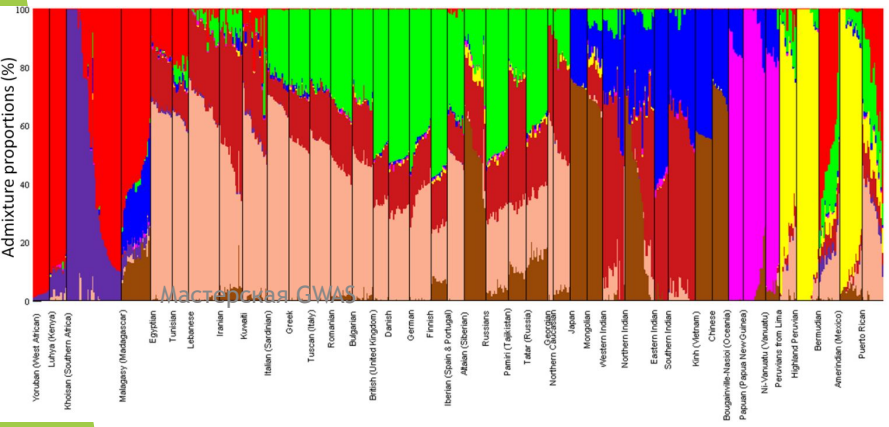
Чтобы сделать вывод о структуре популяции на основе данных генотипа, необходимо сначала уменьшить размерность набора данных из-за тысяч SNP, которые он включает.

Тысячи SNP

```
HGDP00985 HGDP00985 0 0 1 6 T T G G T C A G T C G G G G G A C A C C G G C A G G C C A G C C A G C C C C C
HGDP01094 HGDP01094 0 0 1 6 C T G G T C A G T C G G G G G A C A C C G G C C G G C C A G T C A G C C C C C
HGDP00982 HGDP00982 0 0 1 6 C C G G T C A G C C G G G G G G C C C C G G A A A G C C A G T C G G C C C C C
HGDP00911 HGDP00911 0 0 1 6 C T G G C C G G C C G G G G A A A A C C G G C A A G A C A G T C G G C A C C T
HGDP01202 HGDP01202 0 0 1 6 C C G G C C G G T C A G G G G G C C C C G G C C A A C C G G T C G G C C C C C
```

ДОБАВКА

	к северо-востоку Азиатский	Средиземноморье	Южноафриканский	Юго-Западная Азия	Коренной американец	Океанический	Юговосточная Азия	Северный Европейский	К югу от Сахары Африканский
HGDP00985	0,5253	0,0202	0	0,2222	0,0404	0,0101	0,0101	0,1717	0
HGDP01094	0,04	0,04	0	0,03	0,83	0	0,01	0,05	0
HGDP00982	0,0102	0,1531	0,0306	0,0714	0,0408	0	0,0102	0,2041	0,4796



Пропорции примесей в географически смежных популяциях, таких как итальянцы и греки, а также в популяциях с похожей историей, таких как британцы и немцы, одинаковы.

NATURE COMMUNICATIONS | ARTICLE OPEN

Geographic population structure analysis of worldwide human populations infers their biogeographical origins

Eran Elhaik, Tatiana Tatarinova, Dmitri Chebotarev, Ignazio S. Piras, Carla Maria Calò, Antonella De Montis, Manuela Atzori, Monica Marini, Sergio Tofanelli, Paolo Francalacci, Luca Pagani, Chris Tyler-Smith, Yali Xue, Francesco Cucca, Theodore G. Schurr, Jill B. Gaieski, Carlalynne Melendez, Miguel G. Vilar, Amanda C. Owings, Rocío Gómez + *et al.*

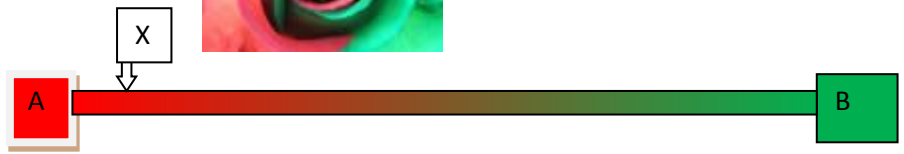
Affiliations | Contributions | Corresponding author

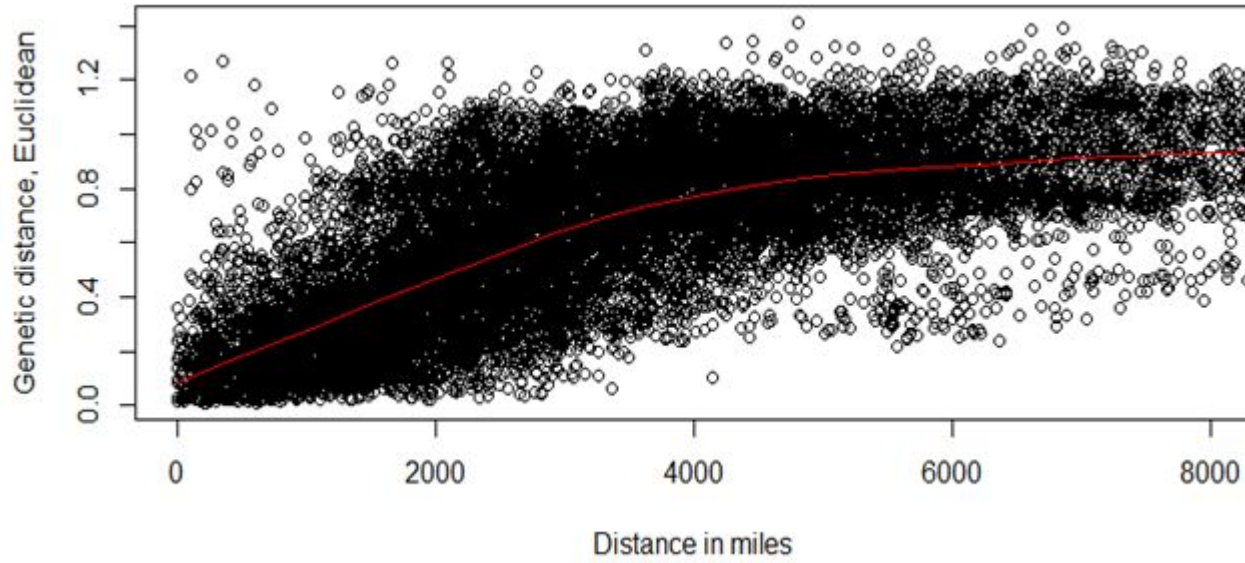
Nature Communications **5**, Article number: 3513 | doi:10.1038/ncomms4513

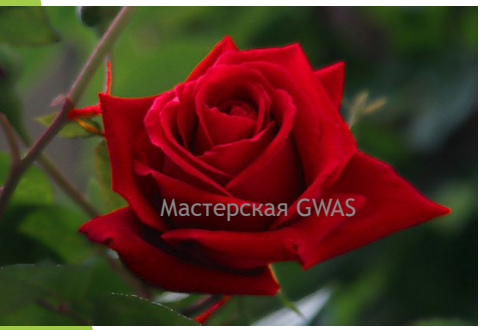
Received 17 April 2013 | Accepted 26 February 2014 | Published 29 April 2014

Прогноз био-происхождения

- Зная связь между географическими и генетическими расстояниями, можно ли определить географическое происхождение человека с известным генотипом?
-
- Мы решил ать простой подход



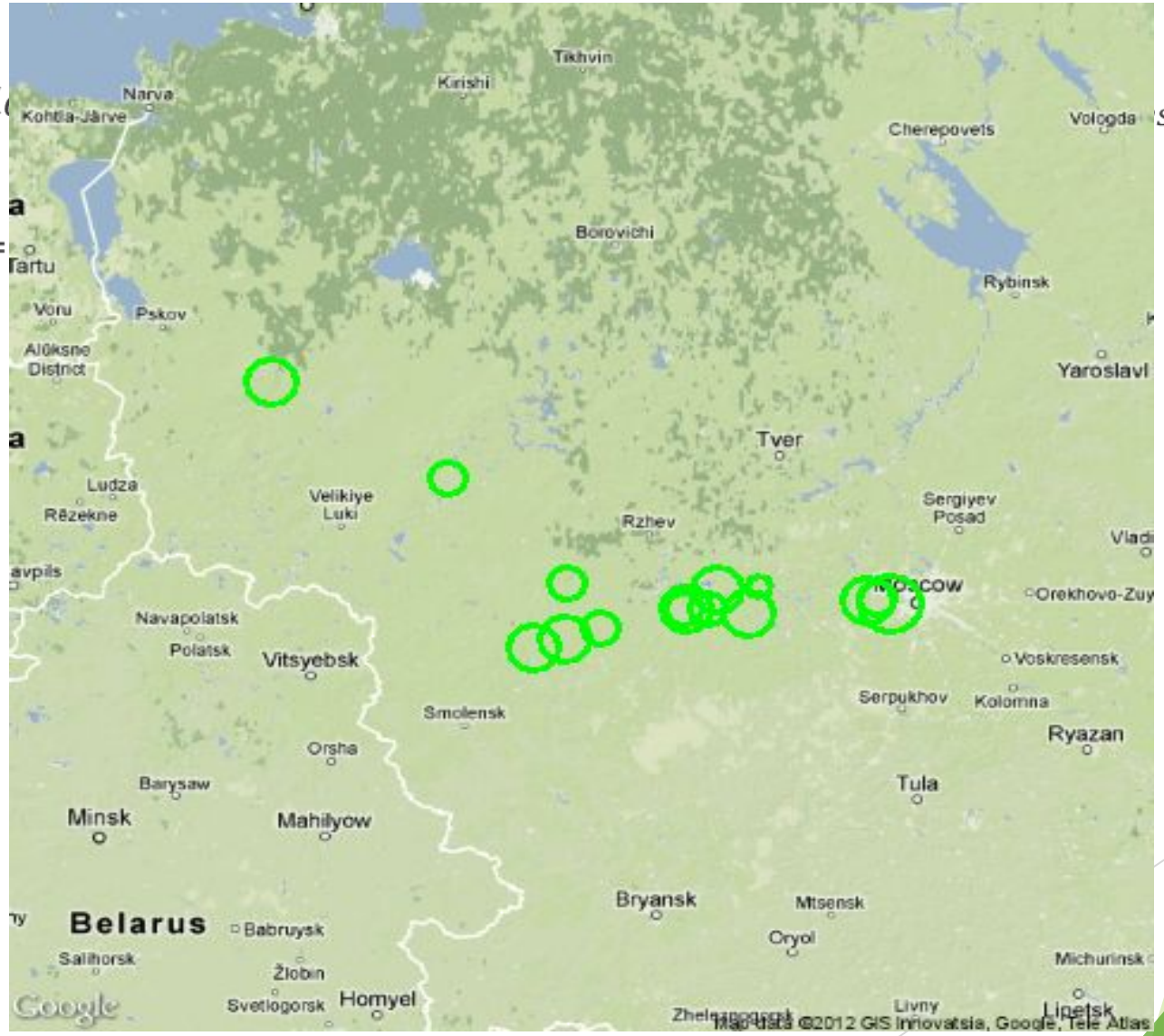


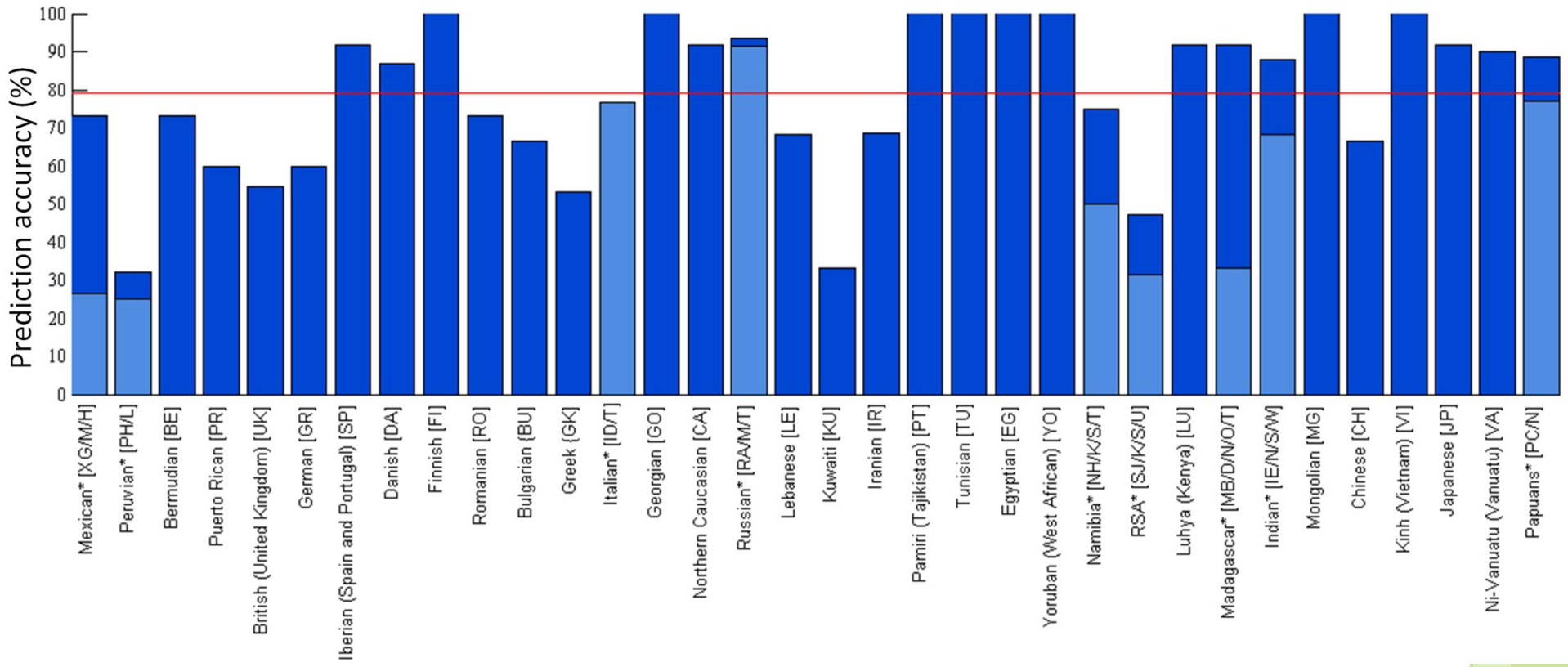


Неизвестные образцы

- $Positi$ $st)$

where w_m =





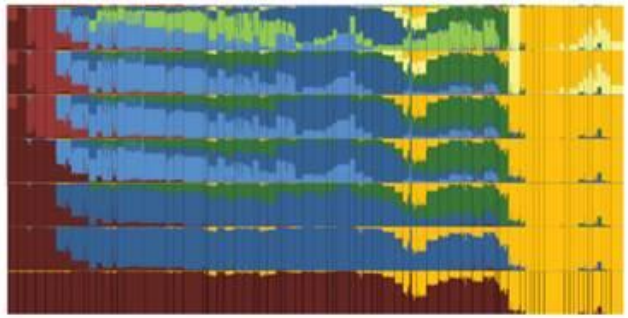
- GPS точно назначен
- ~ 100% всех людей в свои континентальные регионы
- 80% всех людей в страну происхождения
- 60% всех людей в своем внутреннем регионе страны

Моделирующая добавка

The GPS customization scheme

1. Assemble a diverse dataset with at least 500 markers genotyped in >2 populations with >2 individuals in each population.

2. Run ADMIXTURE with various *Ks*. Choose the *k* that represents the most biologically correct divisions, rather than noise.



3. Generate *k putative ancestral populations*.

4. Associate your *k putative ancestral populations* with geographic regions (e.g., North Africans, South Europeans).

5. Create the *reference population dataset* by keeping only the <u>unmixed</u> populations with known geographical from the your original dataset.

6. Calculate the relationship between genetic and geographic distances.

7. Run GPS for the *reference population dataset* and test the assignment accuracy. High specificity and sensitivity indicate a successful design.

8. Apply all GPS tools for individuals of unknown origins

1001 Genomes - Каталог генетической изменчивости Arabidopsis thaliana.

Мы взяли 450 инбредных штаммов A. thaliana из 22 стран, секвенированные и проанализированные Институтом биологии развития Max-Plank и компанией Monsanto из проекта 1001 Genomes (http://1001genomes.org/projects/MPICWang2013/)

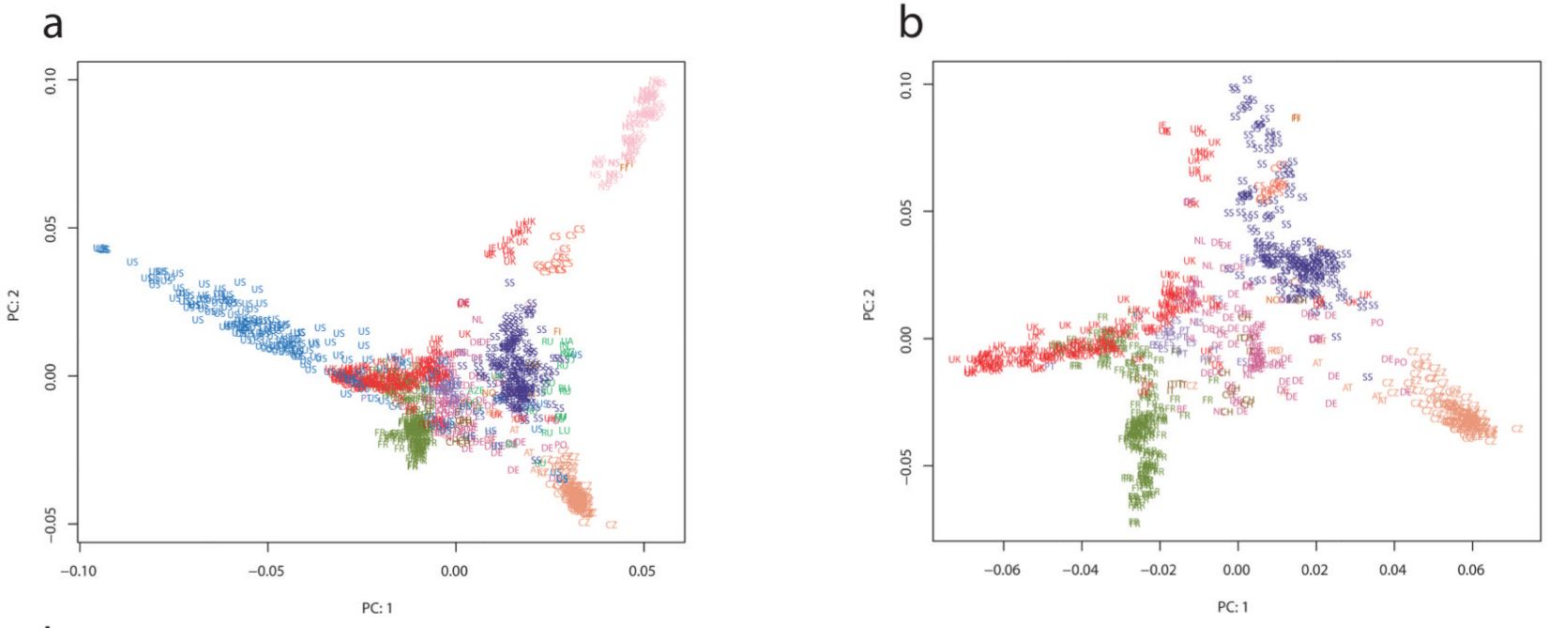
Образцы из 22 стран

Массив SNP генотипа 250K

Координаты точки сбора

Мастерская GWAS Антон Елисеев / Кристина Кривоносова

Структура населения



- Мэтью В. Хортон и др. Полногеномные паттерны генетической изменчивости во всем мире образцов A. thaliana из панели RegMap. Nat. Genet. 44 (2)

Фильтрация SNP

Мы взяли файлы вариантов, доступные на http://1001genomes.org/data/MPI/MPICWang2013/releases/current/.

отфильтрованы инделы и неаутосомные варианты. Обнаружено около 8 млн вариантов.

Мы выбрали 80 000 SNP из исходного набора данных и создали файл PLINK.

Используя PLINK, мы удалили SNP с частотой минорного аллеля <0,05.

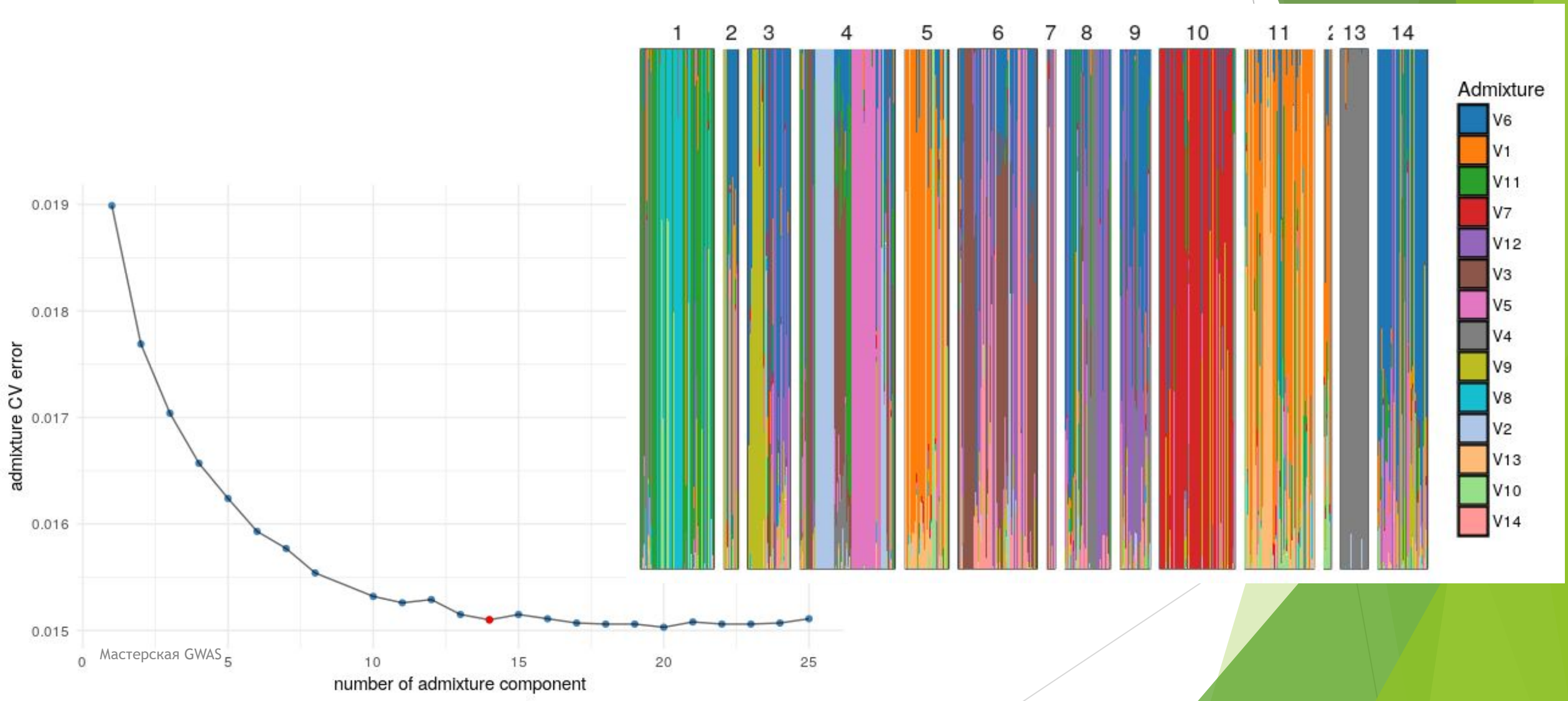
и далее удалил LD с r2> 0,2, используя окно размером 50

Количество оставшихся SNP - 40 518.

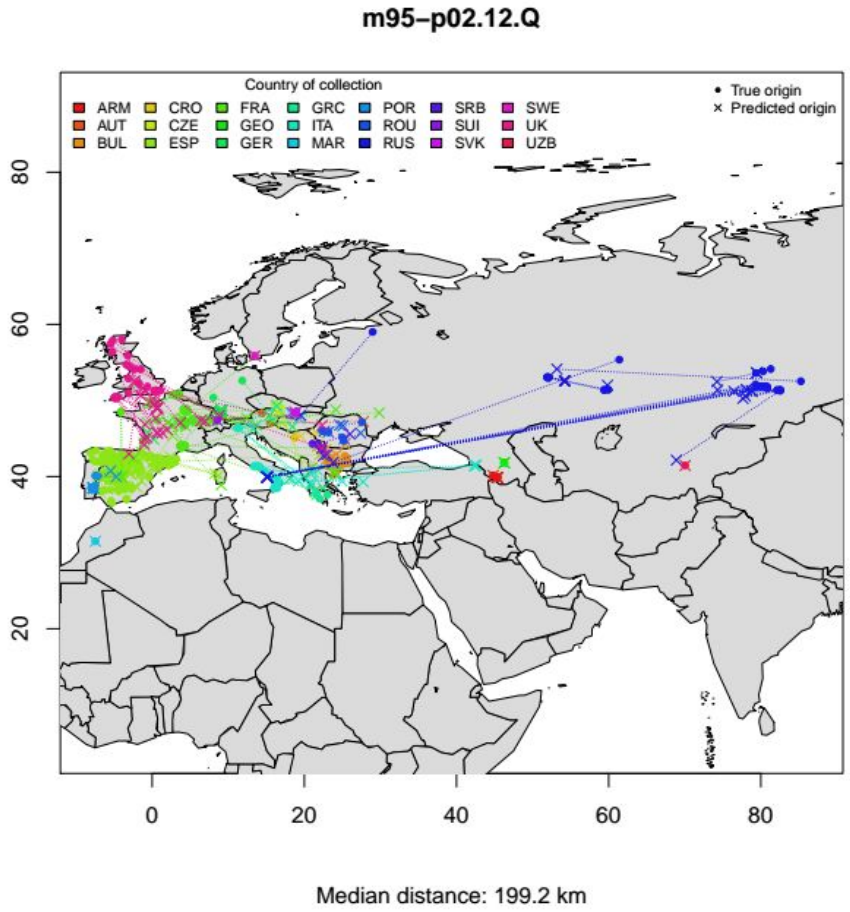
ДОБАВКА

- Мы выполнили несколько анализов ADMIXTURE с числом предковых популяций K от 3 до 20, а затем для последующих анализов было выбрано K = 14, так как это давало наименьшее медианное расстояние при проверке исключения по одному.
- Затем мы разделили некоторые из исходных популяций на 41 генетически однородную субпопуляцию с помощью K-средних.

Примесь


1 2 3 4 5 6 7 8 9 10 11 12 13 14
Admixture
V6
V1
V11
V7
V12
V3
V5
V4
V9
V8
V2
V13
V10
V14
admixture CV error
0.019
0.018
0.017
0.016
0.015
0 5 10 15 20 25
number of admixture component

Проверка GPS по одному разу



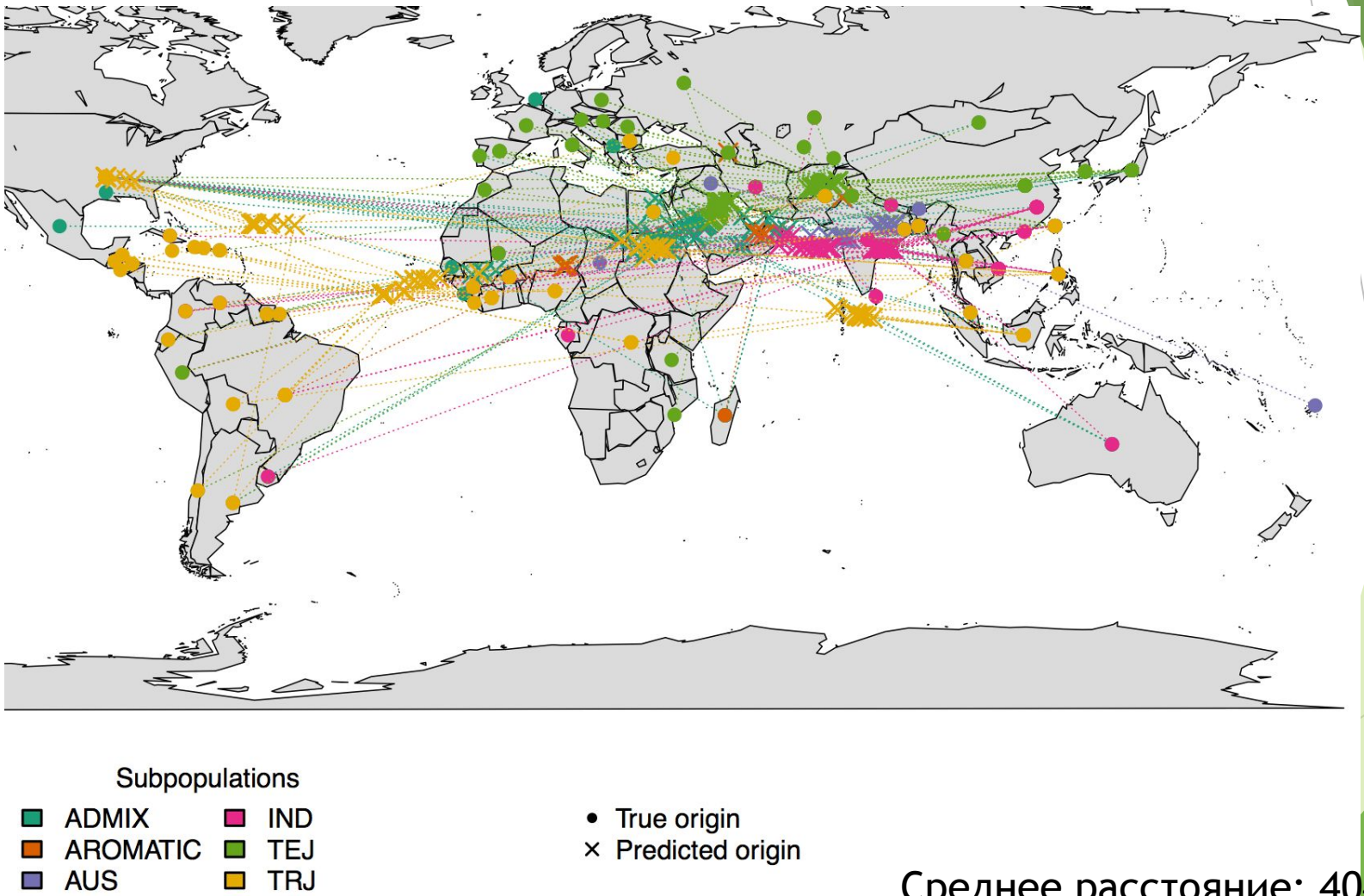
- Процент популяций, которые точно нанесены на карту 60%
- Среднее расстояние до правильного населения 200 км

Исходная гипотеза

Географическое положение Arabidopsis должно быть связано с холодоустойчивостью. Более холодный климат в Северном полушарии наблюдается в более высоких широтах. Растения могут адаптироваться к более холодному климату, регулируя время цветения. Следовательно, мы ожидаем увидеть связь с генами, контролирующими время и широту цветения.

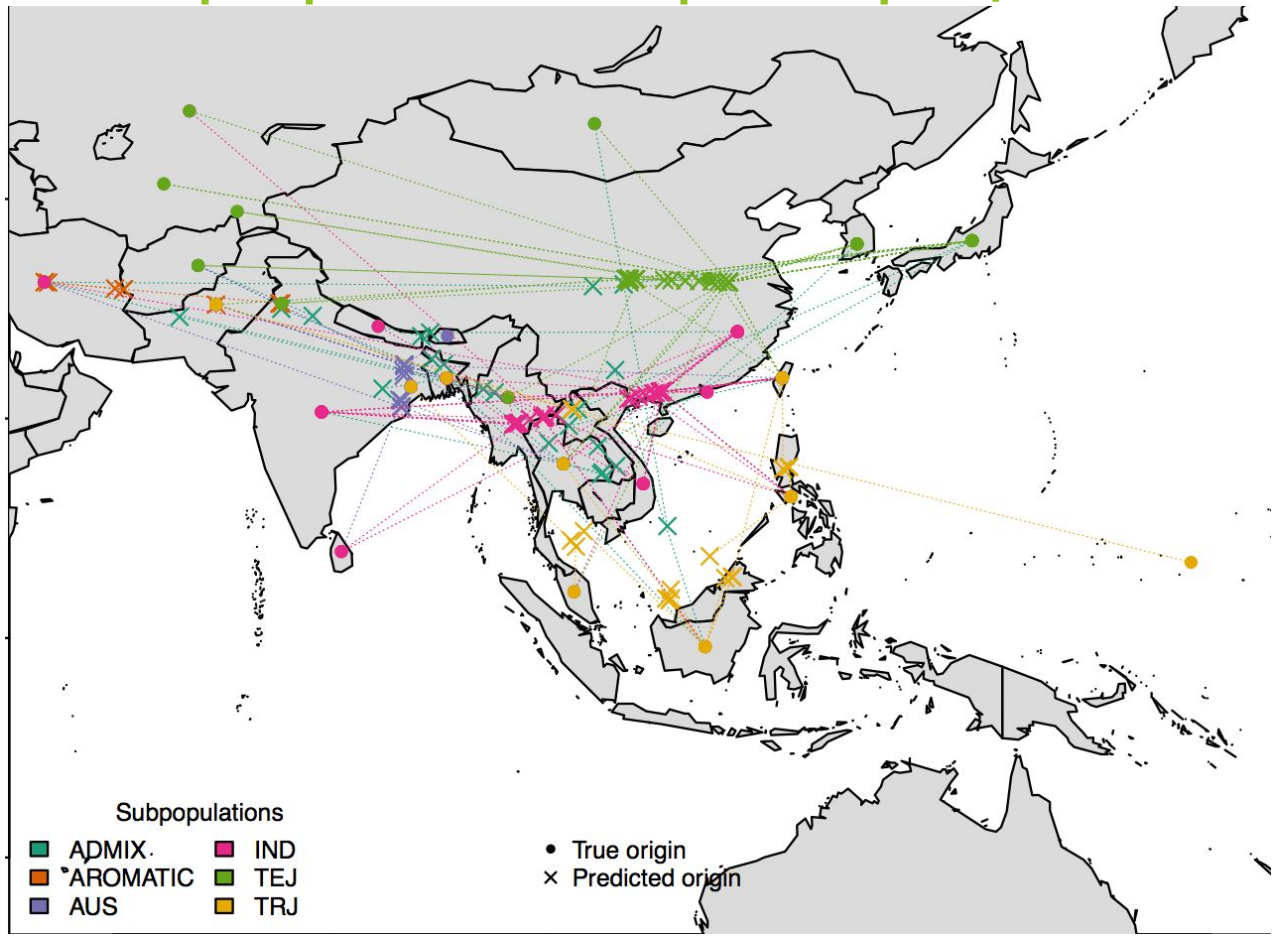
GPS-анализ O. sativa

Точность для риса



Среднее расстояние: 4043 км

Прогнозирование GPS после SNP и географической фильтрации

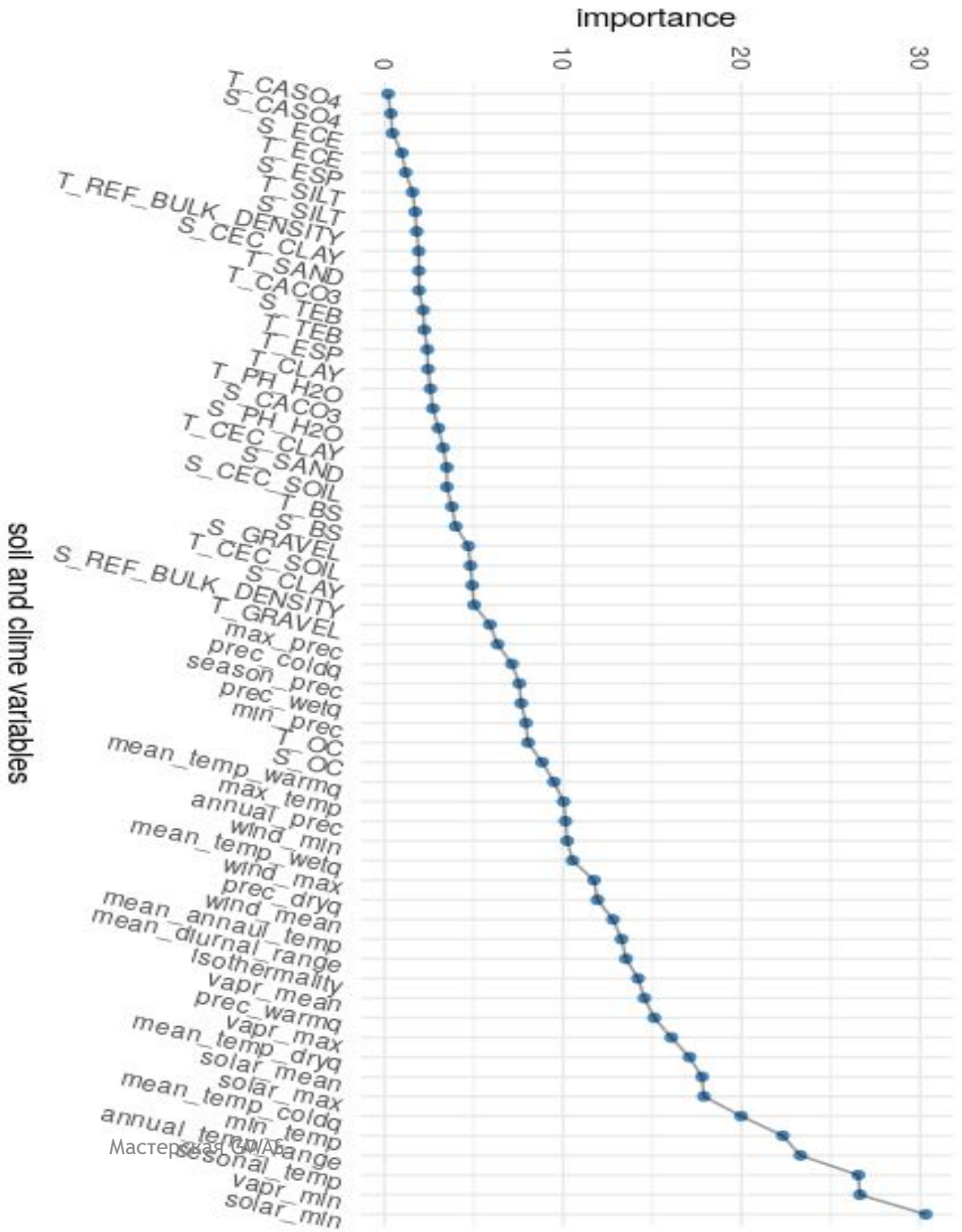


Среднее расстояние: 1141 км

Почему не работает с рисом?

Рис не может выбрать свою вторую половинку, а арабидопсис может.

Наборы данных: WorldClime и SoilDB

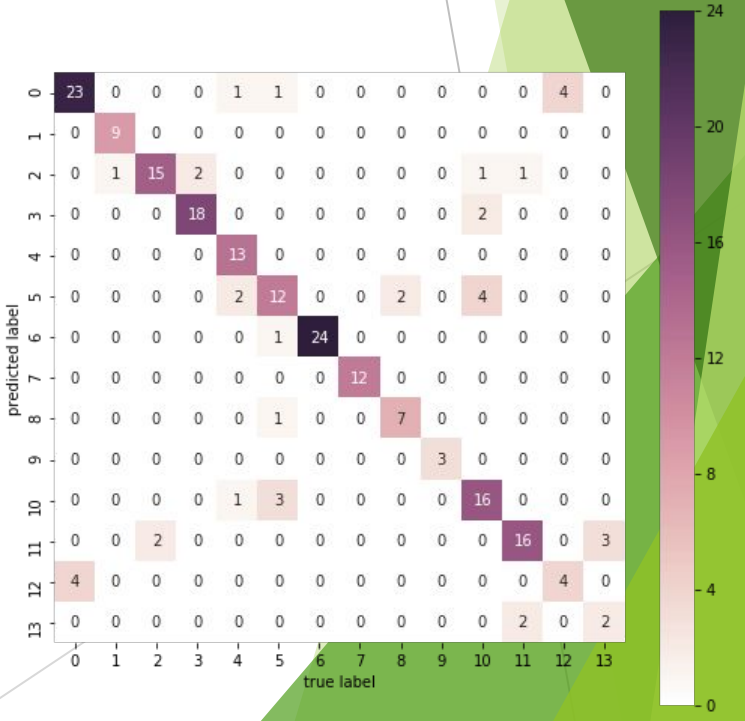


Компоненты примеси больше зависят от климата, чем от параметров почвы.

Климат и почва не являются независимыми (p-значение теста Мантеля 0,03)

Затем спрогнозируйте класс компонента примеси по климату и почве.

Точность 0,85



Часть 1 заключение

Мы умеем моделировать дикие виды - не так уж и много с одомашненными

Климат и почва влияют на генетику (да!)

M. truncatula - модельное бобовое растение.

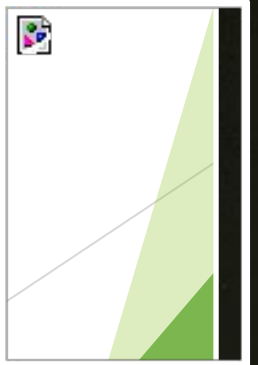
- Семейство бобовых (бобовых) растений
- Возможность установления клубенькового симбиоза
- Высокая синтения к бобовым культурам

- Небольшой диплоидный геном (500 МБ) секвенирован Tang et al., 2014 (BMC Genomics)

- Многочисленные геномные ресурсы
- RIL популяции
- Коллекции мутантов
- Большие базы данных EST и RNASEQ
- Ресурсы HAPMAP Branca et al., 2011 (PNAS)
- 288 геномов уже секвенированы

- Инструменты биотехнологии для валидации
- Большое биоразнообразие Gentzbittel et al., 2015 (Front. Pl. Sci)

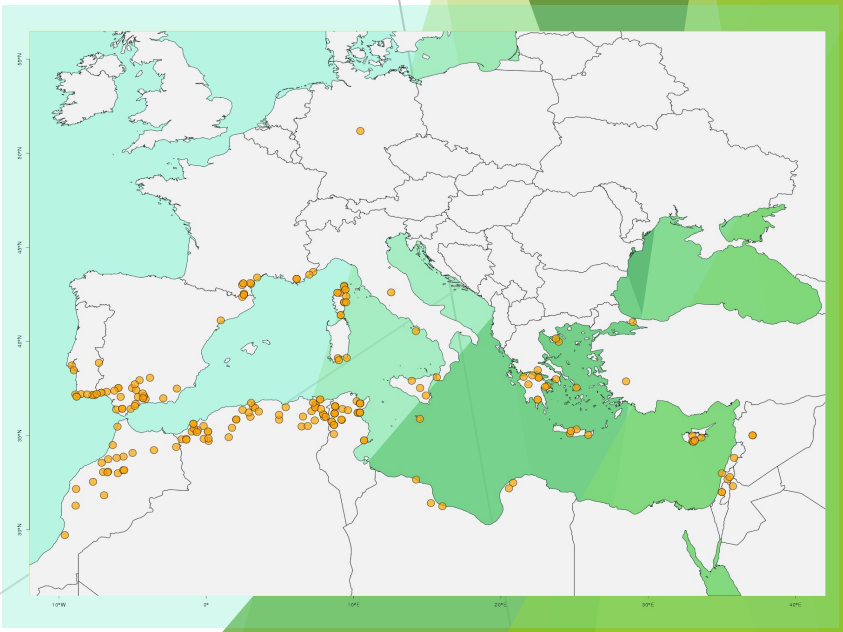
Medicago truncatula



Проект NSF HapMap

Мастерская GWAS

Текущее состояние: 288 последовательных образцов.
Итого: 16.516.721 SNP
30 строк при> 20X
Остальные -> 5X



Бранка и др., 2011 г.

Бобовые - это основные сельскохозяйственные виды, используемые в питании человека и животных.

Зерновые бобовые

- Соя
- Нута
- Фасоль
- Горох
- Арахис
- Чечевица

Высокая сельскохозяйственная ценность

Кормовые бобовые

- Люцерна
- Клевер

Фото: Б. Жулье.

Мастерская GWAS

Зерновые Бобовые очень богаты белком (горох: в 3 раза больше, чем злаки)

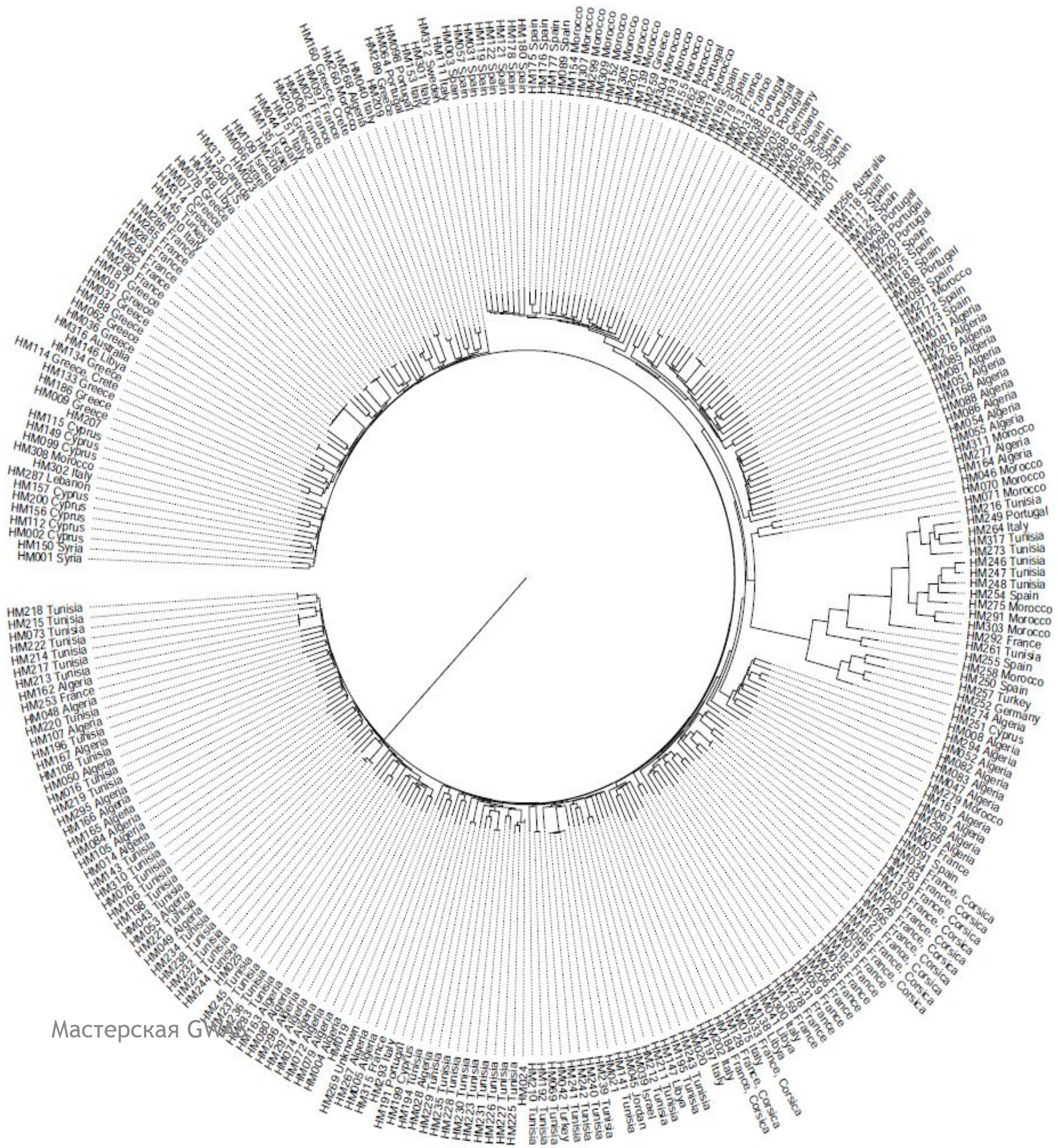
Данные HapMap для M. truncatula

Текущее состояние: 288 последовательных образцов.

 # SNP

Chr1: 1,508,346
Chr2: 1.964.419
Chr3: 2.472.365
Chr4: 2.225.537
Chr5: 3.251.290
Chr6: 1.172.980
Chr7: 2.262.094
Chr8: 1.659.690

Итого: 16.516.721
~ 40% SNP в кодирующих регионах



18 сестринских таксонов

Вот данные, доступные для 288 образцов Medicago, которые уже были секвенированы, что привело к общему количеству более 16,5 миллионов SNP. Среди них 18 образцов, которые сильно отличались от других, были переклассифицированы в некоторые сестринские таксоны, несмотря на полное ботаническое

M. truncatula спонтанно встречается по всему Средиземноморскому бассейну.

Фото: J-M Prospéri, INRA Montpellier, Франция.

Фото: А. Абдельгерфи, INA, Алжир, Алжир

Фото: Л. Генцбиттель, ENSAT, Тулуза, Франция.

Фото: J-M Prospéri, INRA Montpellier, Франция.

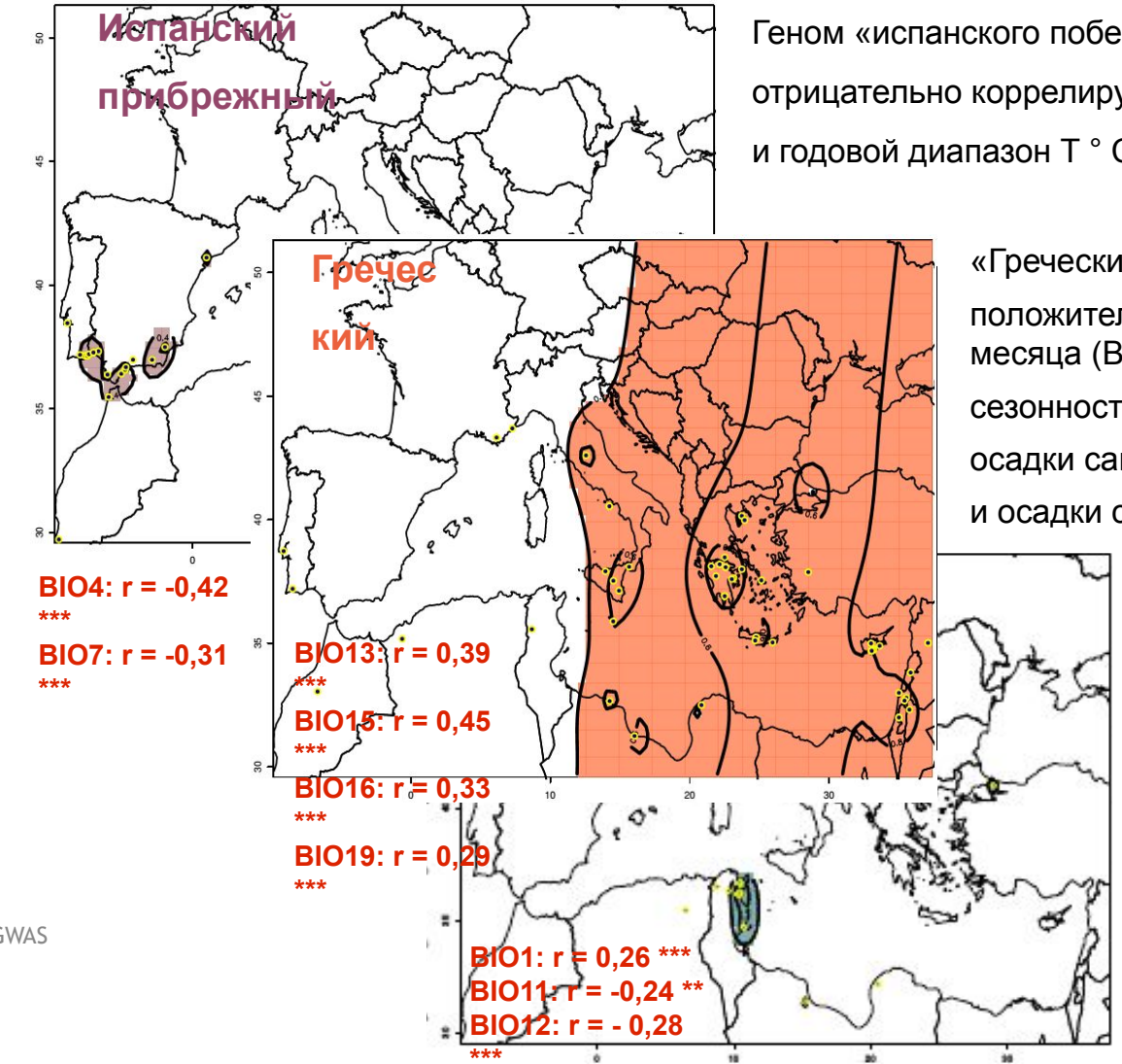
Фото: Э. Ауани, CBBC, Тунис

Фото: Л. Генцбиттель, ENSAT, Тулуза, Франция.

Доступно несколько коллекций: DZ, TN, FR, AU, US,…. на тысячи образцов

Некоторые наследственные геномы могут быть адаптированы к местным биоклиматическим

Мы искали предполагаемую связь с 19 биоклиматическими переменными, как определено WorldClim (http://www.worldclim.org).



Геном «испанского побережья»

отрицательно коррелирует с сезонностью T ° C (BIO4),

и годовой диапазон T ° C (BIO7).

«Греческий» геном

положительно коррелирует с осадками самого влажного месяца (BIO13),

сезонность осадков (BIO15),

осадки самого влажного квартала (BIO16)

и осадки самого холодного квартала (BIO19).

Геном «Южного побережья Туниса»

положительно коррелирует со средней годовой T ° C (BIO1) и средней T ° C более холодного квартала (BIO11)

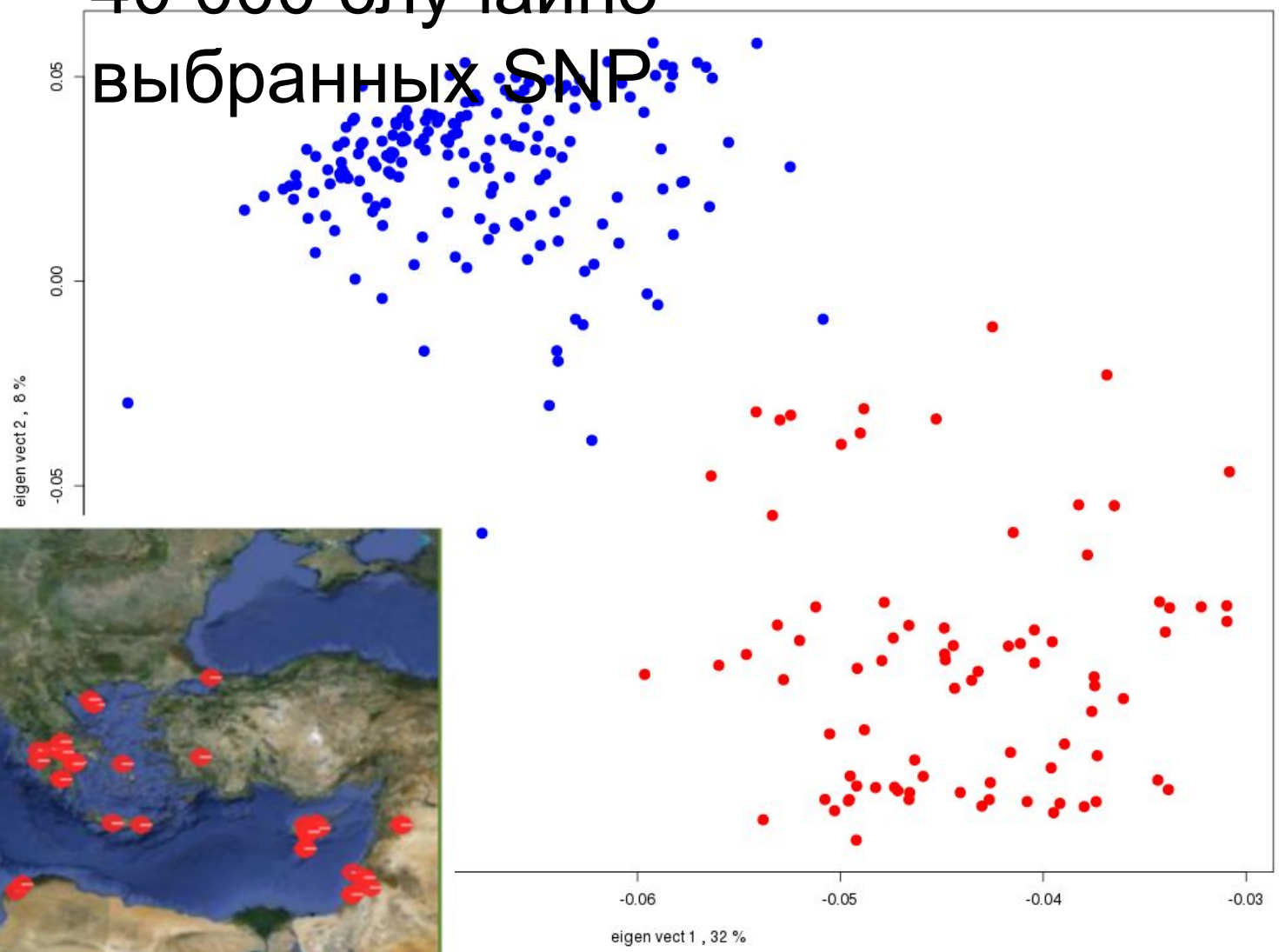
и отрицательно коррелировал с годовым количеством осадков (BIO12).

Геном «Северного побережья Туниса» не имеет существенной корреляции с какими-либо биоклиматическими переменными.

PCA определяет две субпопуляции Medicago среди 262 образцов.

40 000 случайно выбранных SNP

В соответствии с результатами Bonhomme et al. 2014 г.



Уточненная популяционная структура видов Medicago с использованием алгоритмов Admixture и GI

840K randomly selected SNPs (~100,000 per chromosome)

↓

Admixture analysis

↓

'leave-one-out' test for assignation of accessions to actual populations, *using inferred ancestral genomes proportions*.

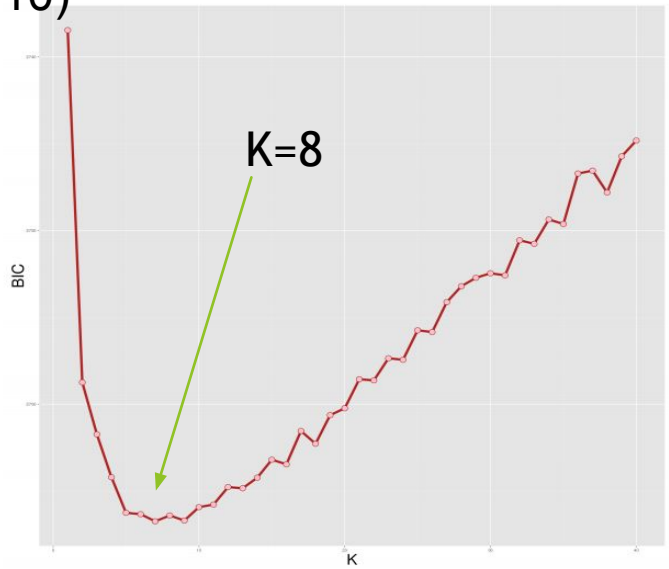
↓

Choice of # putative ancestral populations

Мастерская GWAS

Gentzbittel et al., Genome Biol

Fig. 1: BIC as a function of increasing values of K, using Discriminant Analysis of Principal Components (DAPC) applied on the 840K SNP dataset. (Jombart et al 2010)



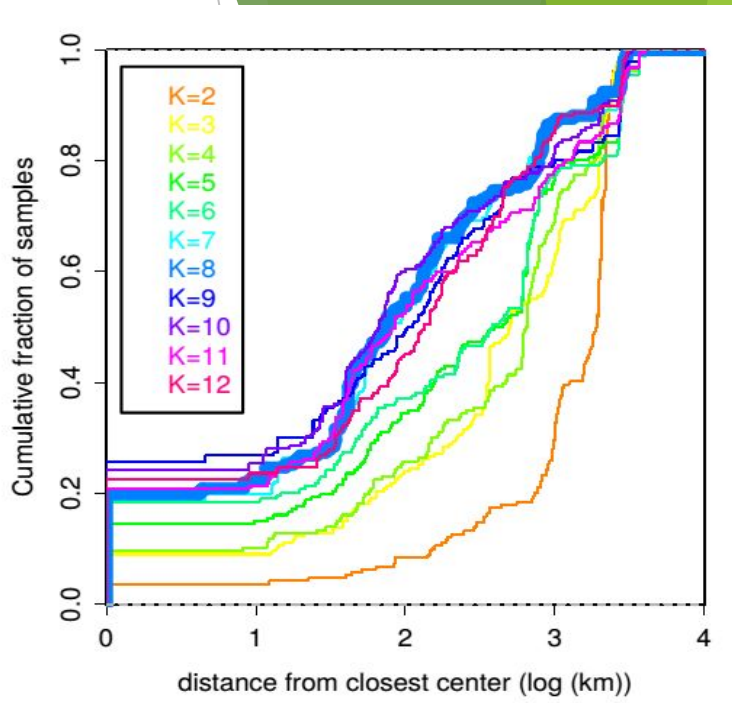


Fig. 2: Predicted distance from true origin for each *M. truncatula* accession using the leave-one-out procedure, for increasing values of K.

Medicago, вероятно, будет иметь 8 наследственных геномов.

Уточненная популяционная структура видов Medicago с использованием алгоритмов Admixture и GPS

840.000 randomly selected SNPs

↓

Admixture analysis

↓

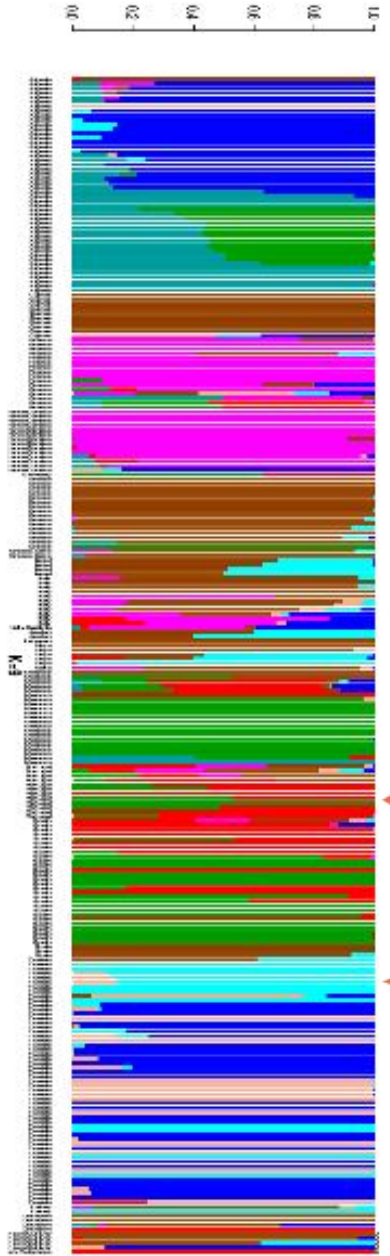
'leave-one-out' test for assignation of accessions to actual populations, *using inferred ancestral genomes proportions*.

↓

Choice of # putative ancestral populations

↓

Mtr is likely to be structured in 8 ancestral genomes



Frequent ancestral genome

Complex admixture pattern

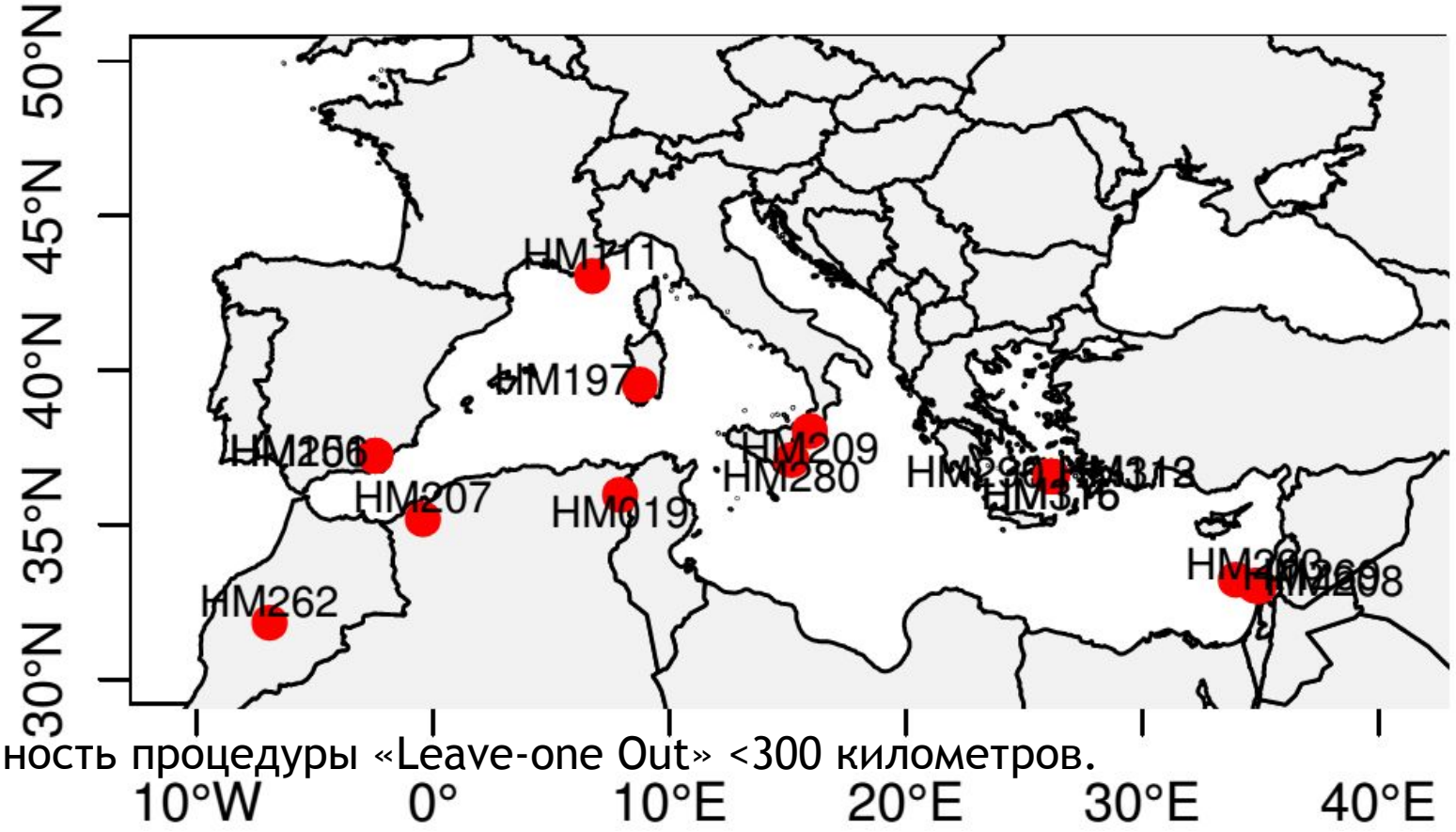
Very low Admixture pattern

Rare ancestral genome

An 'ancestral genome' is characterized by its pattern of allele frequencies

Gentzbittel et al., Genome biol. ,

Предсказано местонахождение 17 неизвестных образцов, включая эталонный геном Jemalong-A17.



Точность процедуры «Leave-one Out» <300 километров.

Линейная связь между географическими и генетическими расстояниями при d <950 км
(R = 0,8, значение p = 10-4, критерий Мантеля)

Gentzbittel et al., Genome biol. ,

Уточненная популяционная структура видов Medicago с использованием алгоритмов Admixture и GPS

840.000 randomly selected SNPs

↓

Admixture analysis

↓

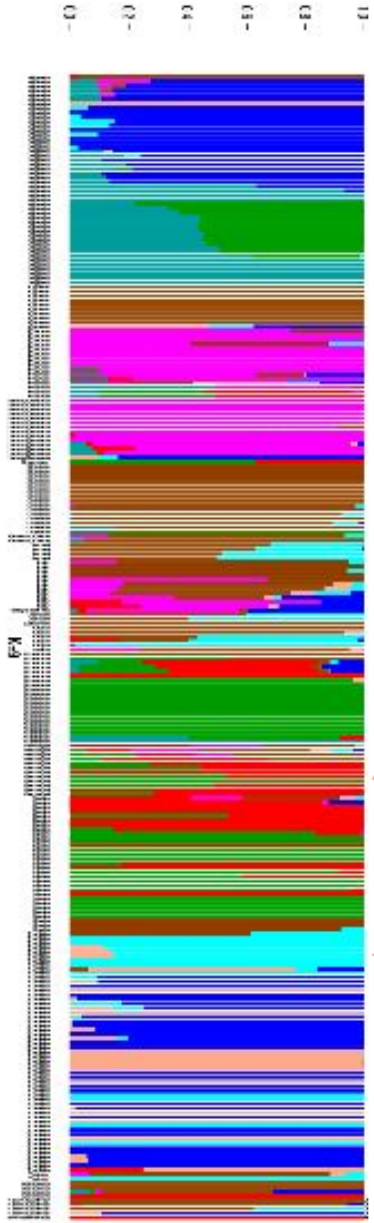
'leave-one-out' test for assignation of accessions to actual populations, *using inferred ancestral genomes proportions*.

↓

Choice of # putative ancestral populations

↓

Mtr is likely to be structured in 8 ancestral genomes



Is this population structure computed from genetic diversity data consistent with:

-geographical distribution?
-Bioclimatic covariables?
-Other biological covariables?

Frequent ancestral genome

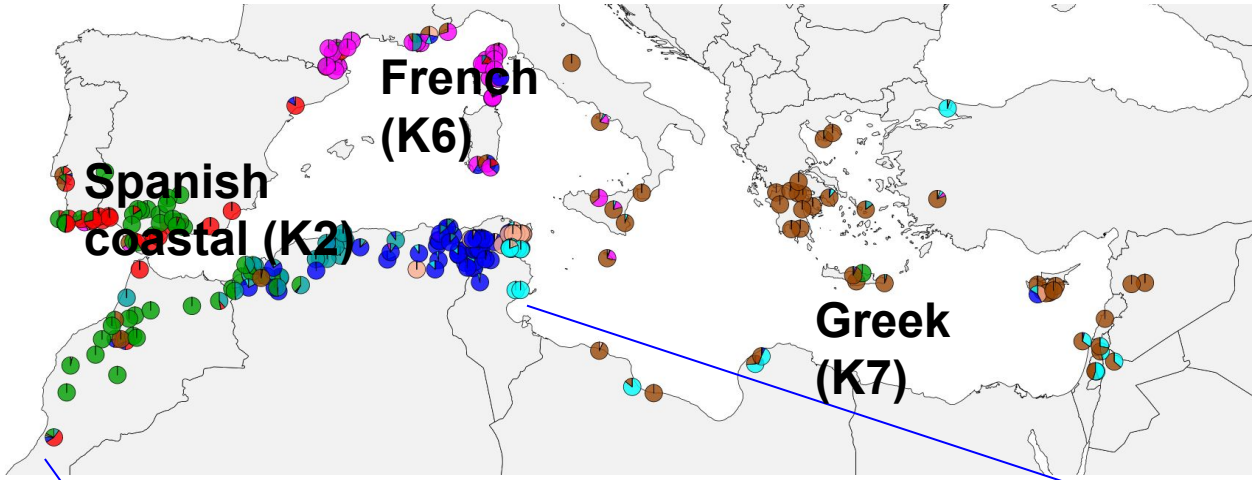
Complex admixture pattern

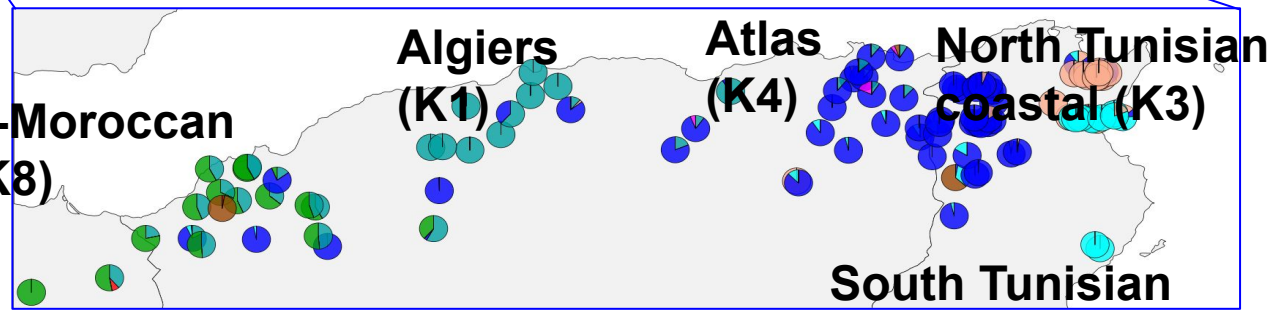
Very low Admixture pattern

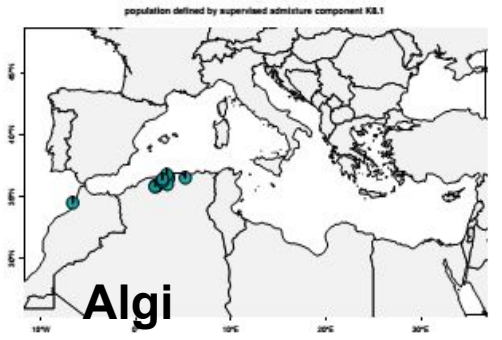
Rare ancestral genome

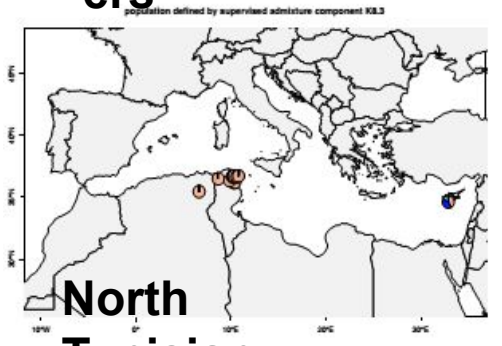
An 'ancestral genome' is characterized by its pattern of allele frequencies

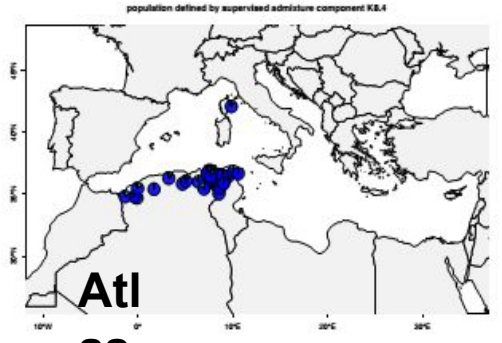
Geographic distribution of the putative 8 ancestral genomes of *M. truncatula*

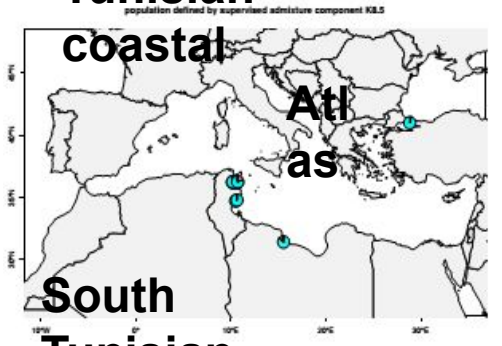


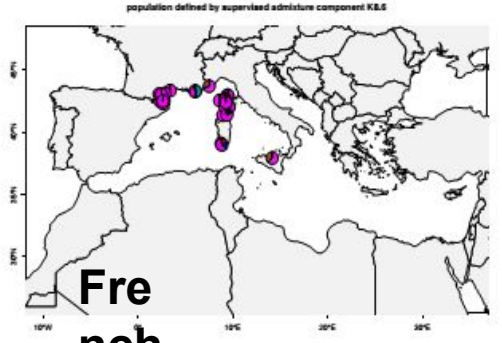


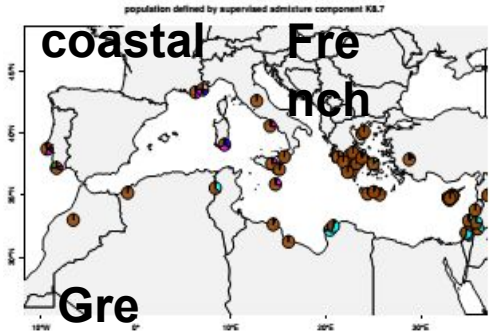












5 out of 8 populations are

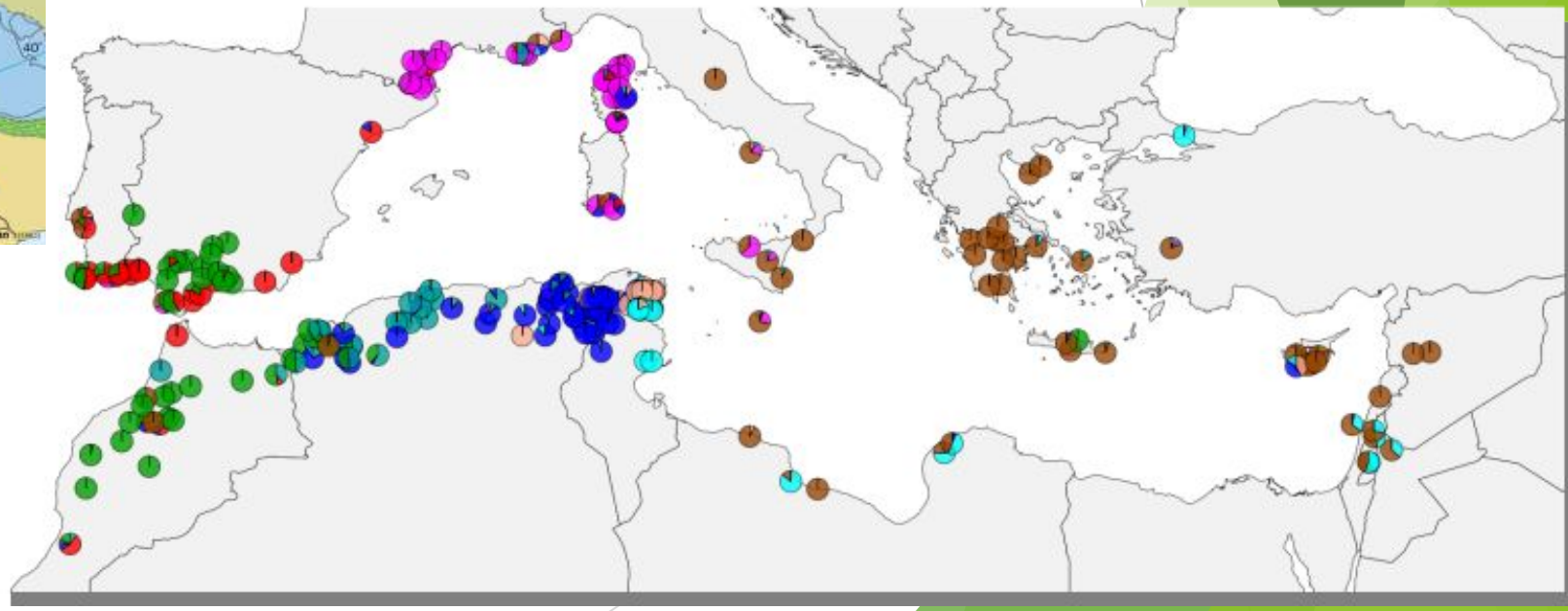
Maghreb as a putative center of diversity for M. truncatula species

Gentzbittel et al., Genome biol. ,

Last glaciations & main glacial refugia may explain *M. truncatula* population stru

-20.000 BP

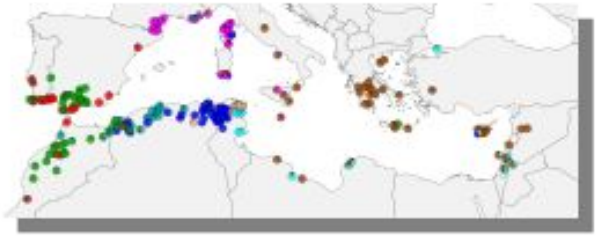
Figure 2. Mediterranean basin hot-spots. 1: Canaries and Madeiran archipelagos. 2: High and Middle Atlas Mountains. 3: Baetic-Rifan complex. 4: Maritime and Ligurian Alps. 5: Tyrrhenian islands. 6: Southern and Central Greece. 7: Crete. 8: Anatolia and Cyprus. 9: Syria–Lebanon–Israel. 10: Mediterranean Cyrenaic. The thick line defines the limits of the Mediterranean area.

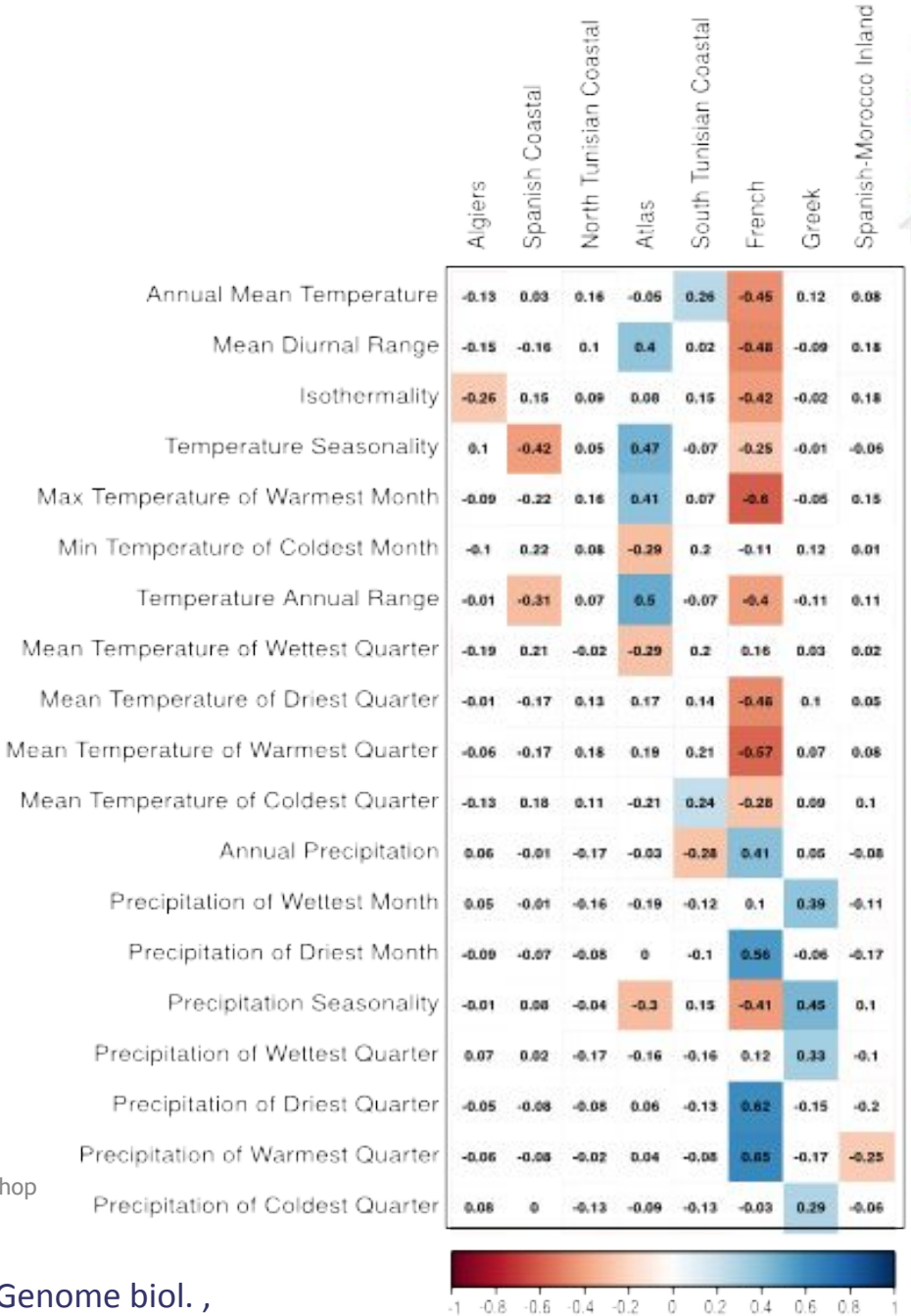


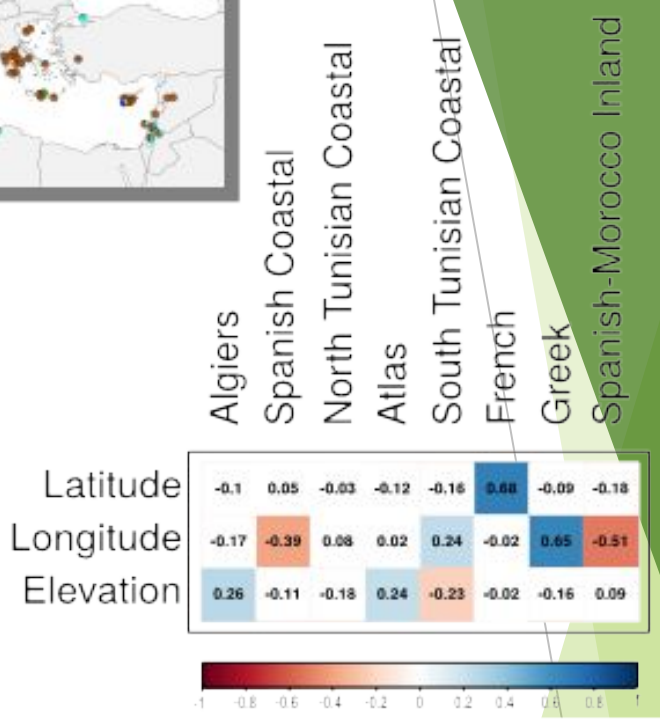
HOT-SPOTS ANALYSIS FOR CONSERVATION OF PLANT BIODIVERSITY IN THE MEDITERRANEAN BASIN[1]

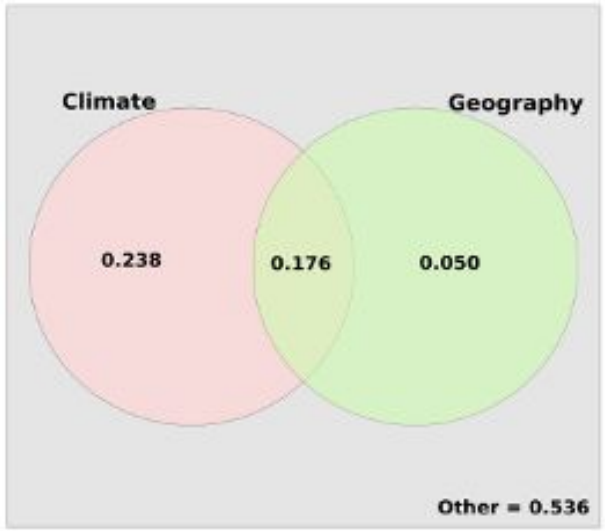
Frédéric Médail[2] and Pierre Quézel[2]

Admixture components are correlated with current conditions





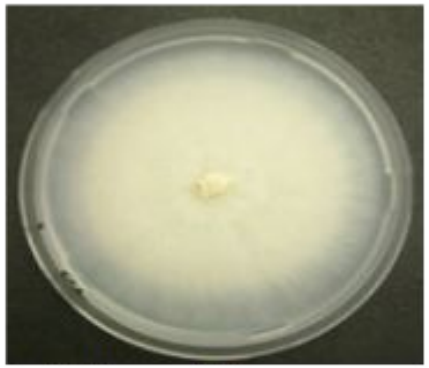




Gentzbittel et al., Genome biol. ,

Verticillium sp. are soil-borne fungal pathogens for Legumes

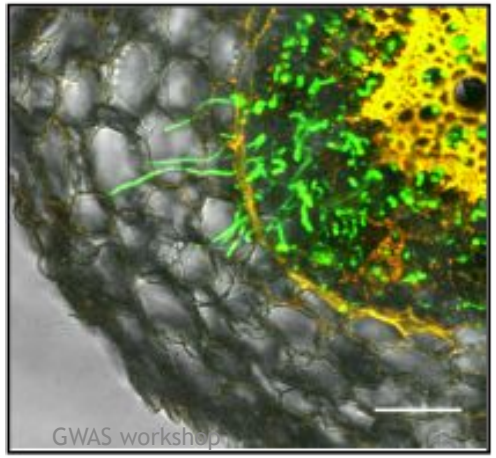
- Soil-borne pathogens: Threats difficult to control
- Wilt diseases in more than 300 plant species

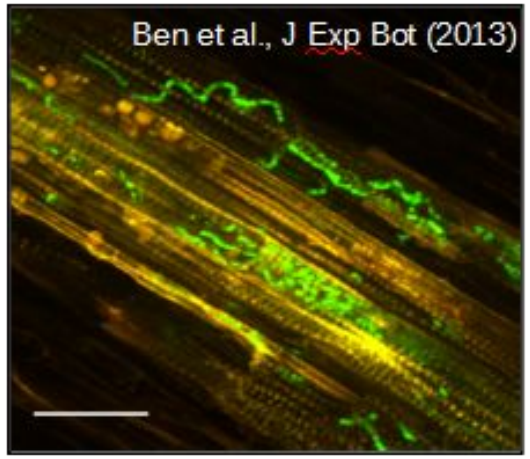


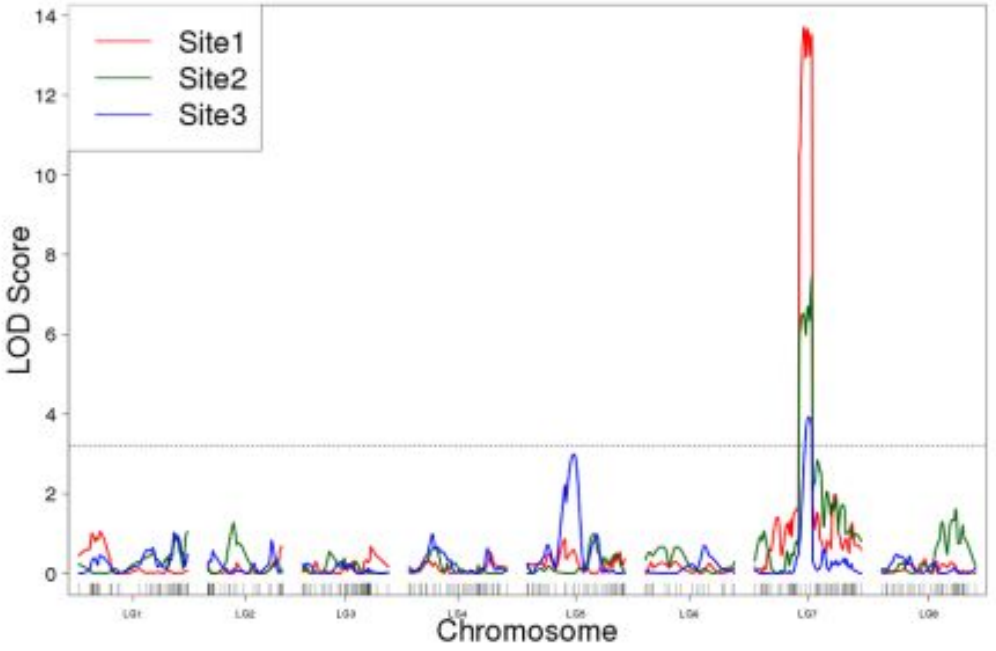
Verticillium alfalfae

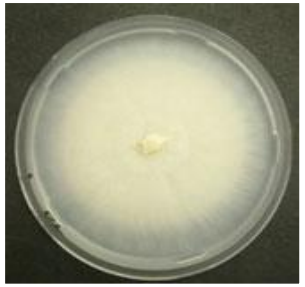
Verticillium wilt in Medicago truncatula

Verticillium wilt in alfalfa

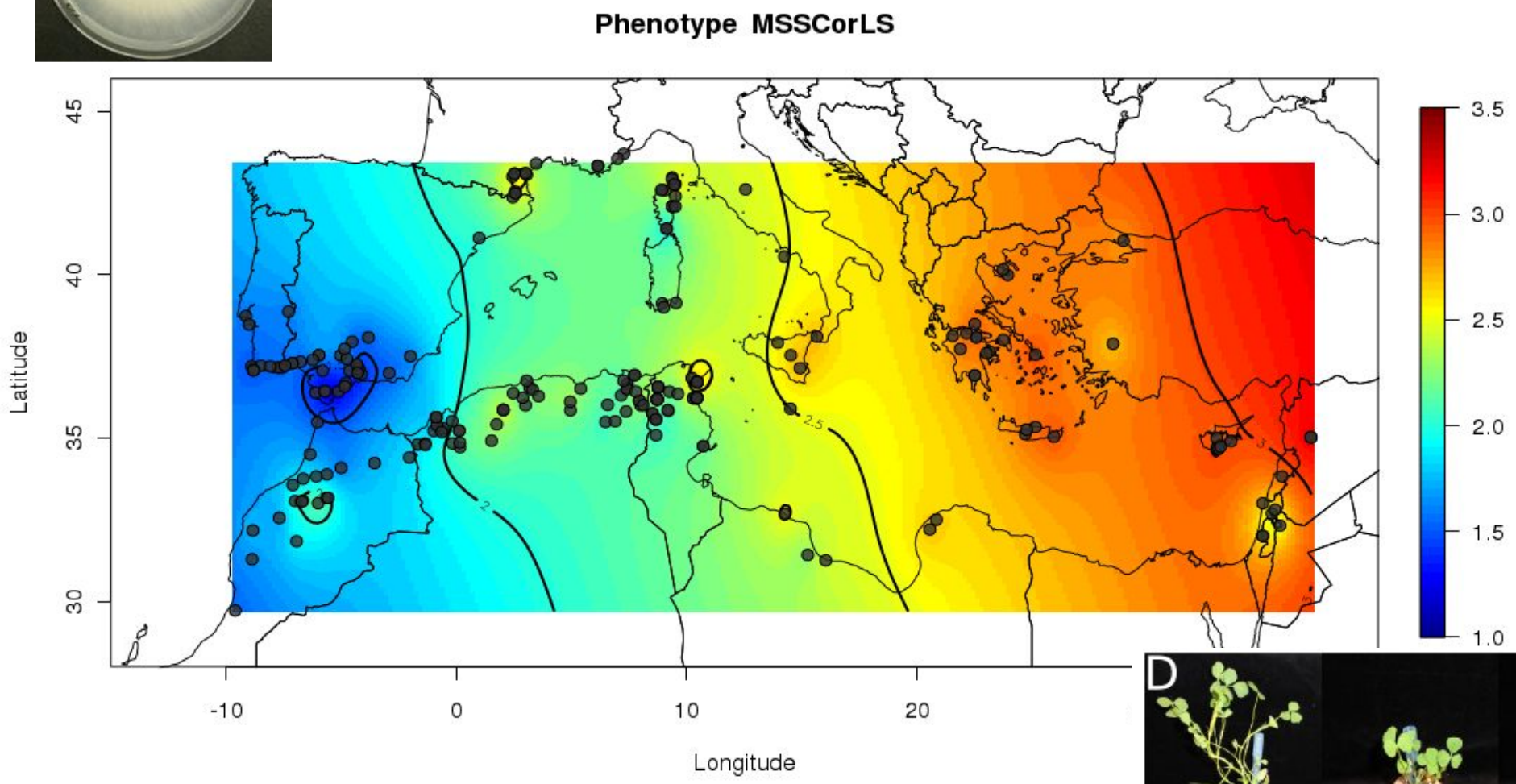




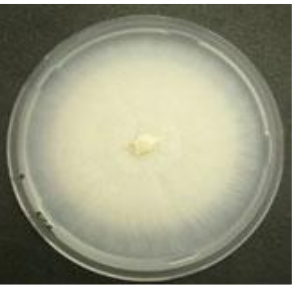




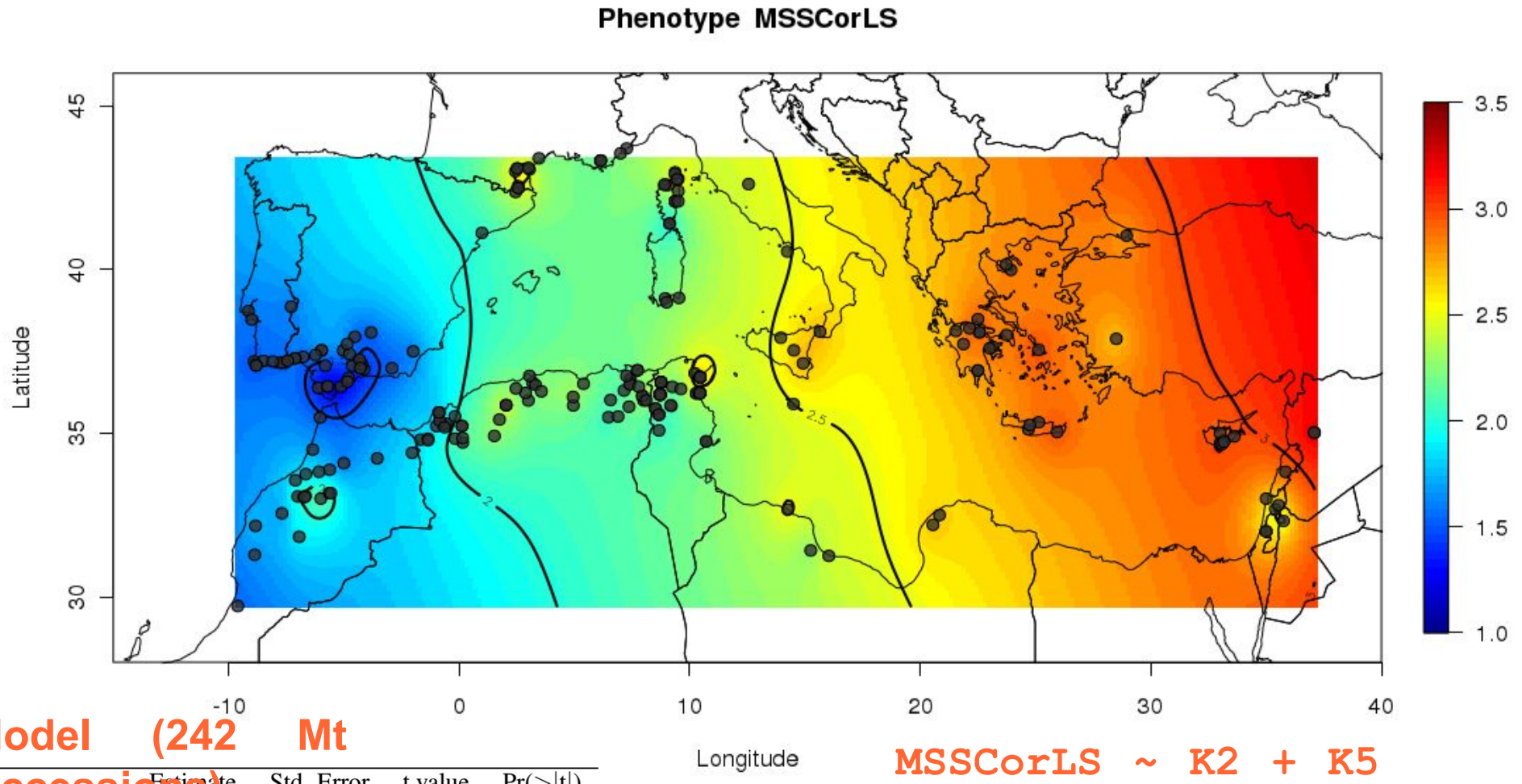
Ancestral genomes may present different levels of resistance to Verticillium wilt



V. alfalfae symptoms on susceptible *M.truncatula* (Ben *et al*. 2013)



Ancestral genomes present different levels of resistance to Verticillium wilt



Linear Model (242 Mt HAPMAP accessions)

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	2.4541	0.0861	28.50	0.0000
Spanish Coastal	-1.4147	0.2290	-6.18	0.0000
South Tunisian Coastal	-0.7222	0.2367	-3.05	0.0026
Greek	0.6017	0.1636	3.68	0.0003
Spanish-Moroccan Inland	-0.8518	0.1804	-4.72	0.0000

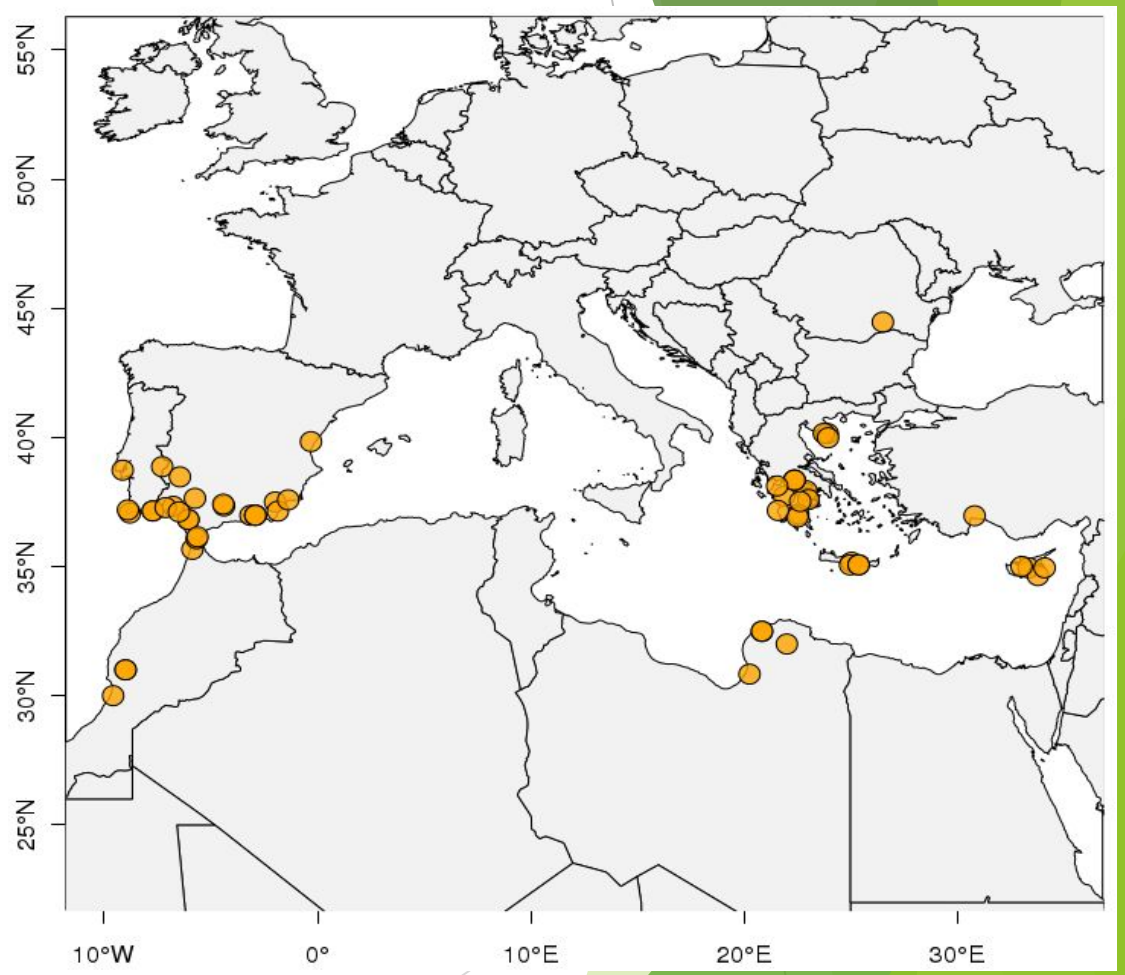
MSSCorLS ~ K2 + K5 + K7 + K8

r^2=0.31, *P-value* < 2.2×10^{-16}

Gentzbittel et al., Genome biol. ,

Can phenotype be predicted using WhoGem?

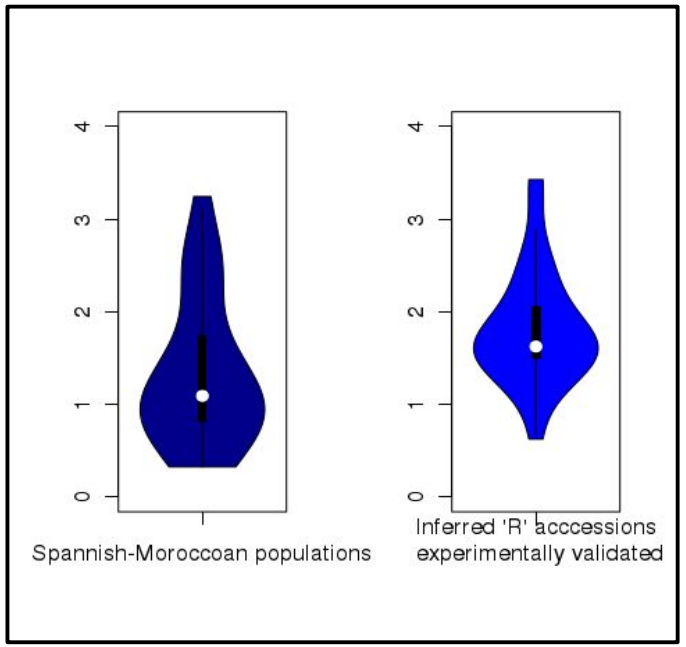
- Use plant model (*Medicago truncatula*) to test applicability of admixture framework to predict phenotypes.
- GPS was used to predict provenance of 59 unknown *M. truncatula* accessions, and MSS was inferred based on the similarity of admixture profiles to the training set accessions using WhoGem model.
- *M. truncatula* was infected with *V. alfalfae* and wilting symptoms recorded.
- MSS (Maximum symptom score) ranges from 0 (healthy) to 4 (dead), 2 threshold.



V. alfalfae symptoms on susceptible *M.truncatula (*Ben *et al*. 2013)

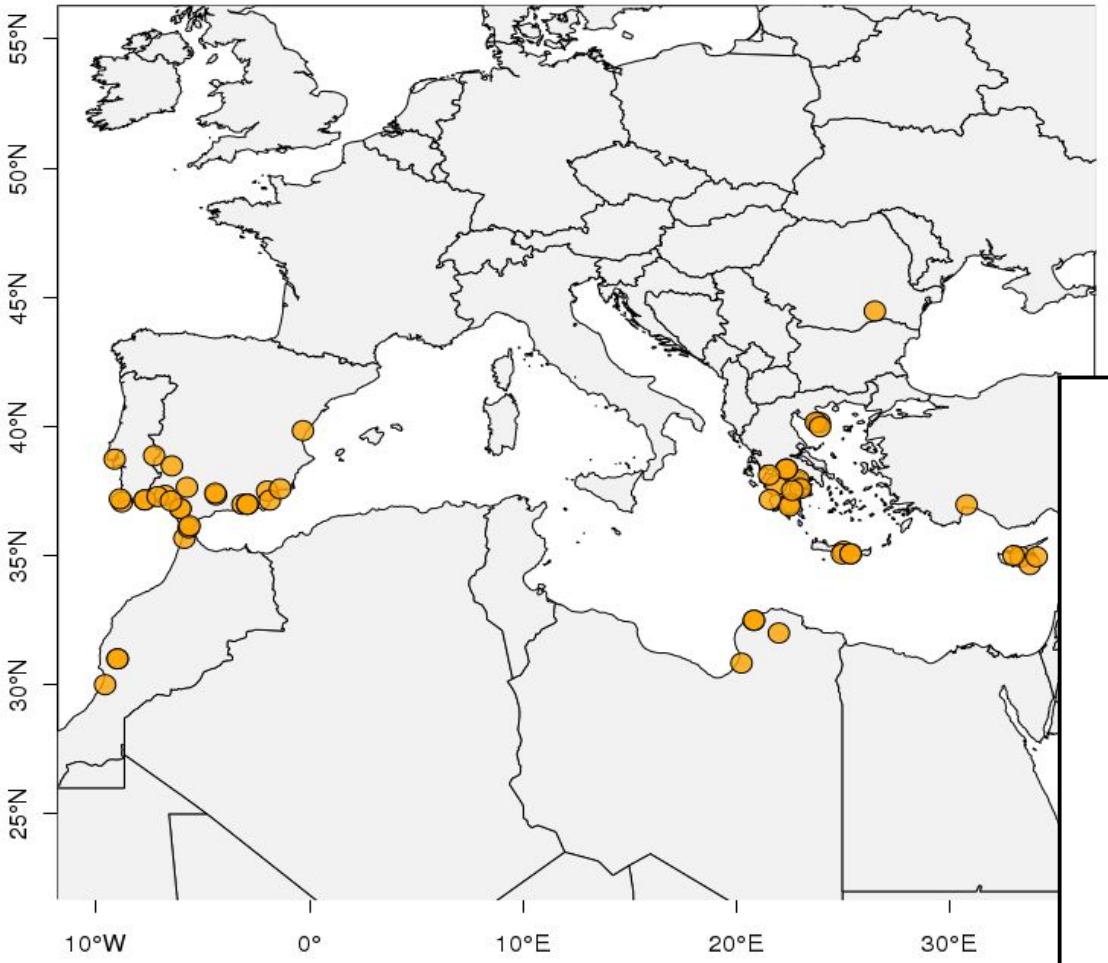
We thank Jean-Marie Prospéri (INRA Montpellier, France) for

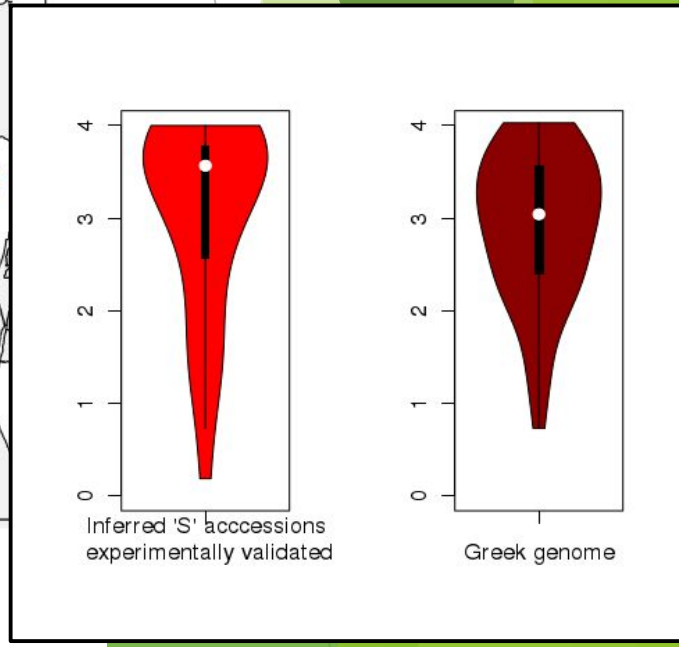
Experimental validation of predicted resistance levels

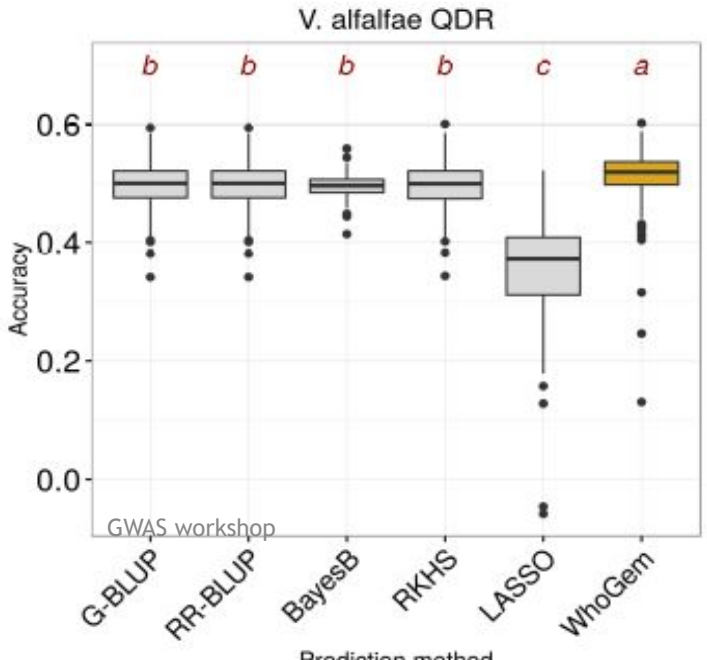


21/29 samples that are predicted to be resistant were resistant
24/30 samples predicted to be susceptible were susceptible

Chi-square *P-value* = $1.5.10^{-4}$

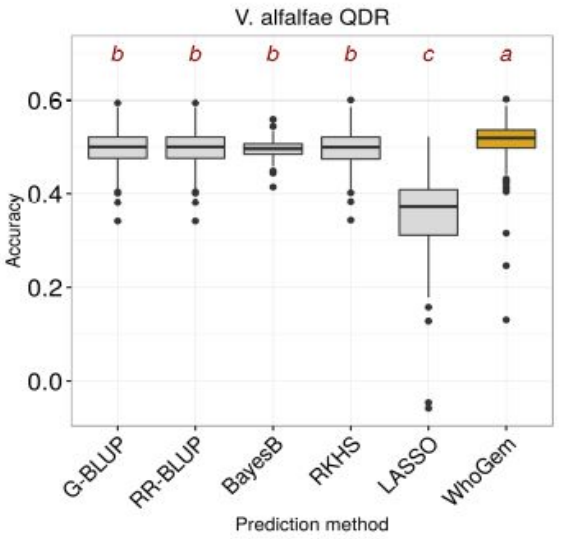


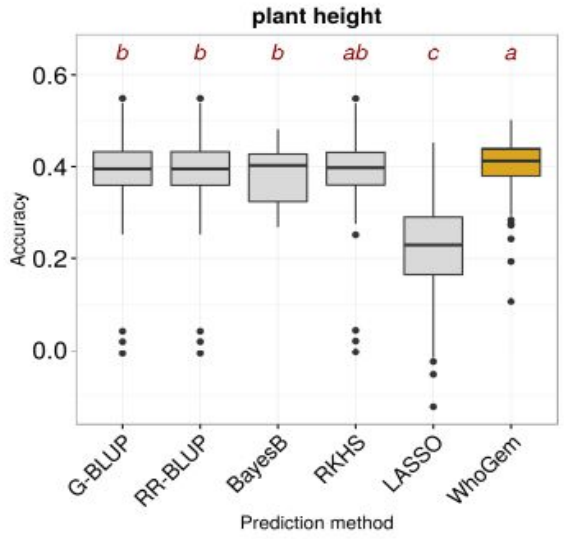


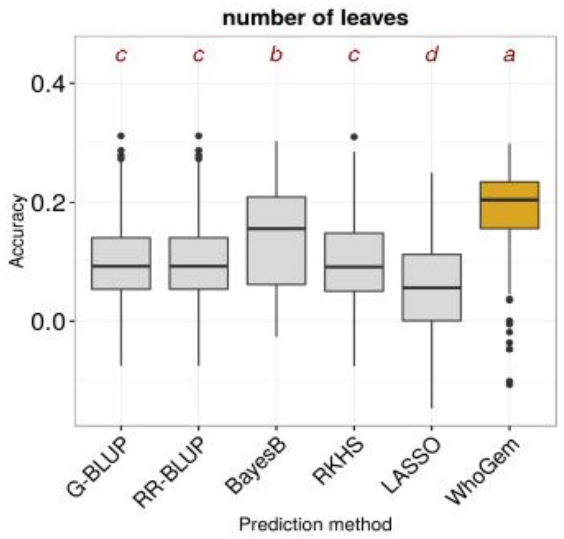


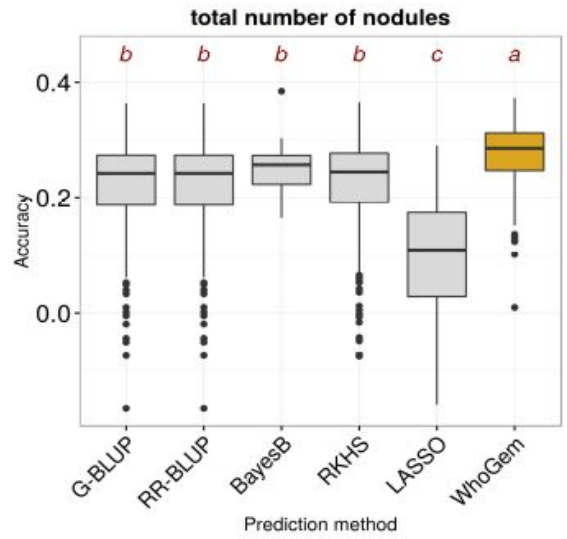
Gentzbittel et al., Genome biol. ,

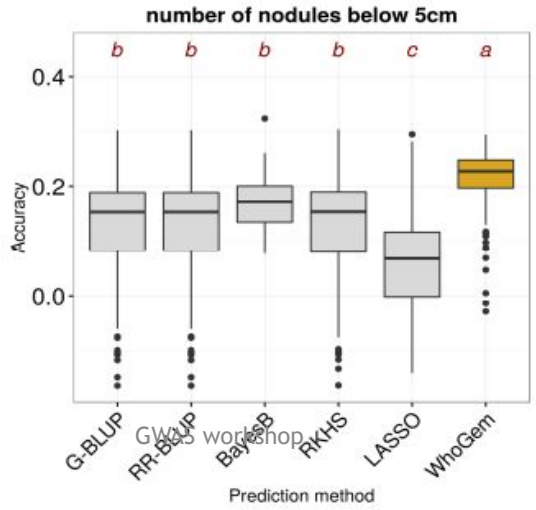
WhoGem models can be significant predictors of quantitative functional traits in plants

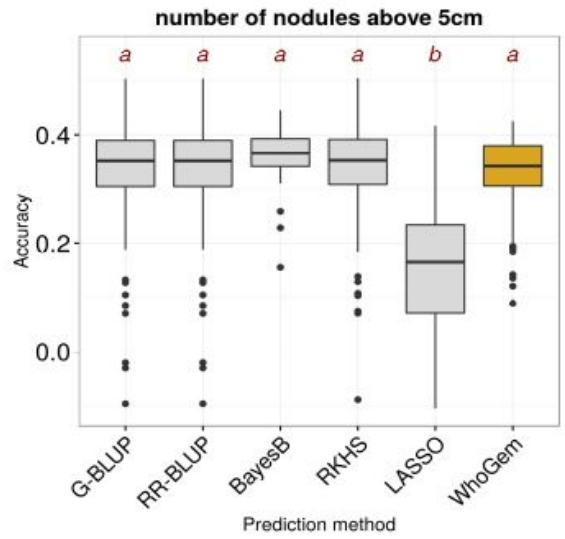


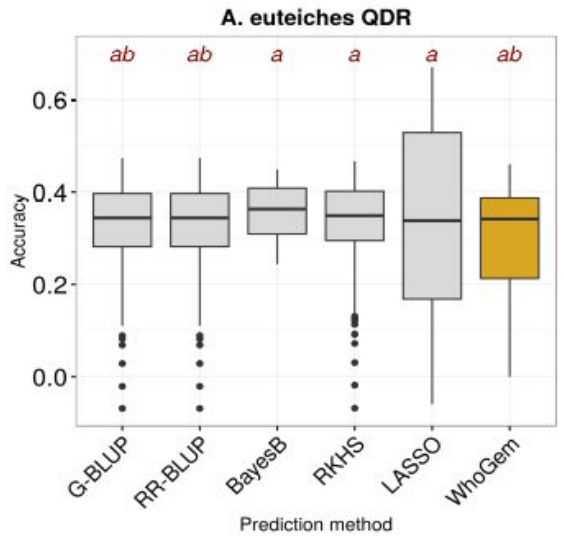












Comparisons of accuracy of prediction for several quantitative traits of adaptive relevance in *M. truncatula* using 5 Genomic Selection algorithms and our WhoGEM method.
Accuracy is computed using 50 runs of five-fold cross-validation for all methods.

Phenotypic data from:
Mazurier et al, in preparation
Stanton-Geddes et al., 2013
Bonhomme et al., 2013

Gentzbittel et al., Genome biol. ,

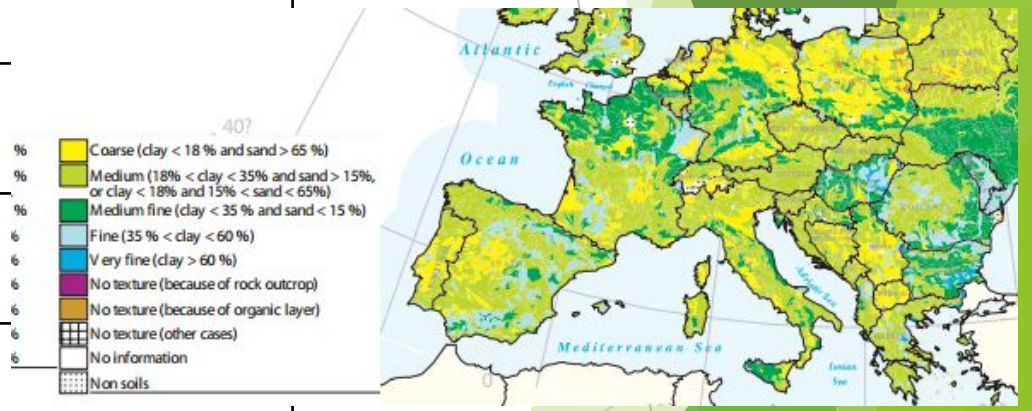
What is missing?

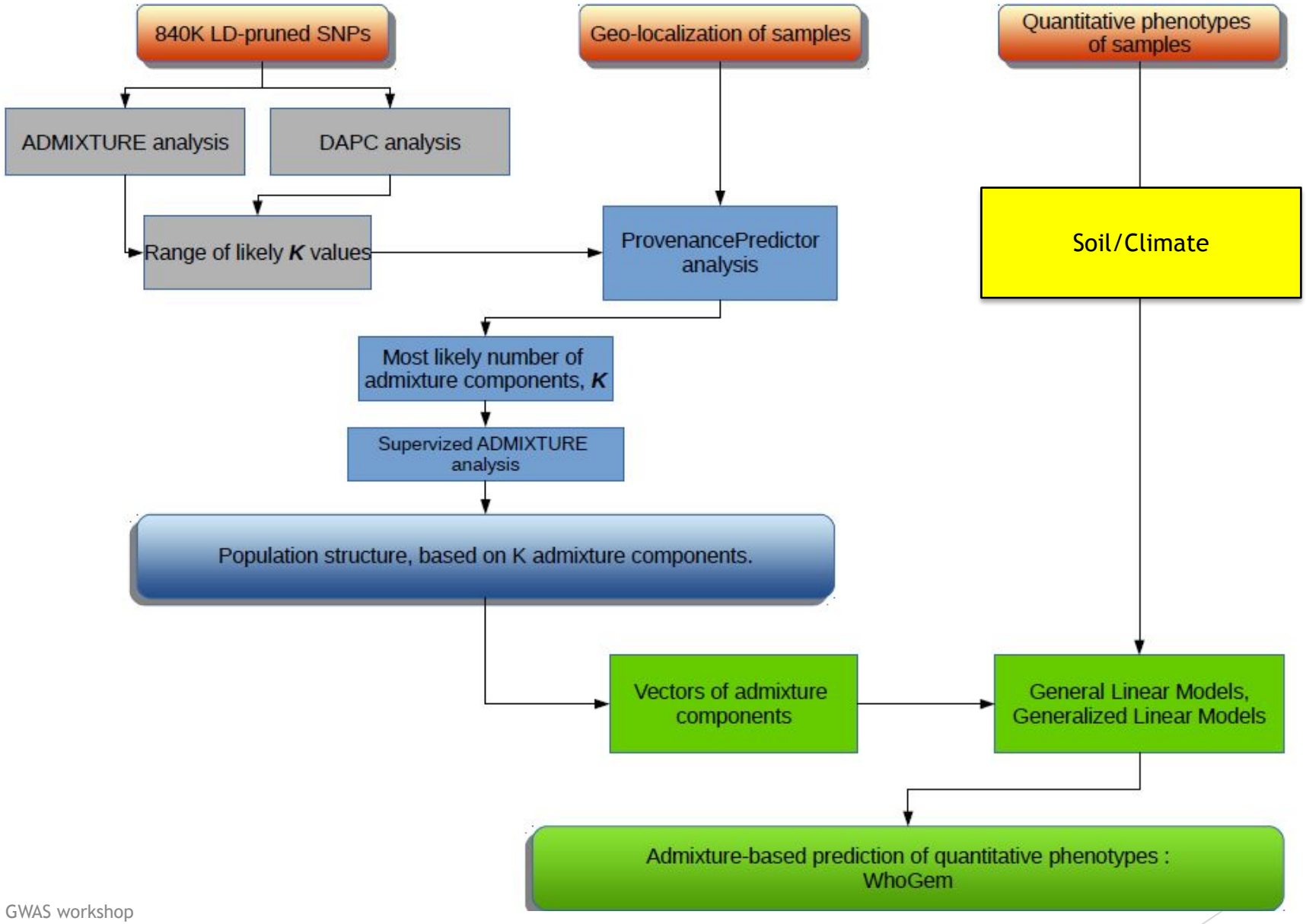
Distance between two points should not be just geometric distance. Add:

1. Mode of propagation (e.g. Medicago seeds stick to goats)
2. Soil
3. Climate
4. Geography

	MAX correlation	MIN correlation	Positively correlated with	Negatively correlated with
K1	0.42	-0.31	Isothermality	Precipitation of Wettest Quarter
K2	0.42	-0.46	Precipitation in February	Mean Diurnal Range (Mean of monthly (max temp - min temp))
K3	0.40	-0.25	Precipitation in August	Precipitation in March
K4	0.67	-0.40	Precipitation in May / June	Min. temperature in November
K5	0.34	-0.31	Temperature Annual Range	Precipitation in March
K6	0.74	-0.31	Latitude	Precipitation of Wettest Quarter
K7	0.59	-0.51	Temperature Seasonality/Max Temperature in July	Isothermality
K8	0.63	-0.59	Precipitation in December/ January / Coldest Quarter	Longitude

	MAX correlation	MIN correlation	Positively correlated with	Negatively correlated with
K1	0.16	-0.10	material	limiting use factor
K2	0.32	-0.24	Usage/Slope	texture, grain size
K3	0.13	-0.21	limiting use factor	erodibility
K4	0.27	-0.20	soil crusting texture factor	slope
K5	0.15	-0.24	root obstacle depth	erodibility
K6	0.28	-0.52	material	subsurface
K7	0.46	-0.21	slope	Crusting
K8	0.56	-0.27	Texture	slope




840K LD-pruned SNPs
Geo-localization of samples
Quantitative phenotypes of samples
ADMIXTURE analysis
DAPC analysis
Range of likely K values
ProvenancePredictor analysis
Soil/Climate
Most likely number of admixture components, K
Supervized ADMIXTURE analysis
Population structure, based on K admixture components.
Vectors of admixture components
General Linear Models, Generalized Linear Models
Admixture-based prediction of quantitative phenotypes : WhoGem

Part 2 conclusions

- Large proportion of phenotypic variation between individuals may be best explained by population admixture.
- Variation in genome admixture proportion explains most of phenotypic variation for quantitative functional traits.
- We experimentally confirm the prediction of differences in quantitative disease resistance levels in the wild model legume *Medicago truncatula.*
- Admixture components were found to be significantly related to climate and geography, also positive selection at the species level might not explain current adaptation.
- Phenotypes can be predicted using genome-wide patterns of admixture, when incorporating covariates such as individuals' provenance.
- This insight contributes to the understanding of adaptation, and can accelerate plant and animal breeding, and biomedical research programs.

This methodology may serve as a basis for analyses in other plant species and for other functional quantitative
traits of interest.

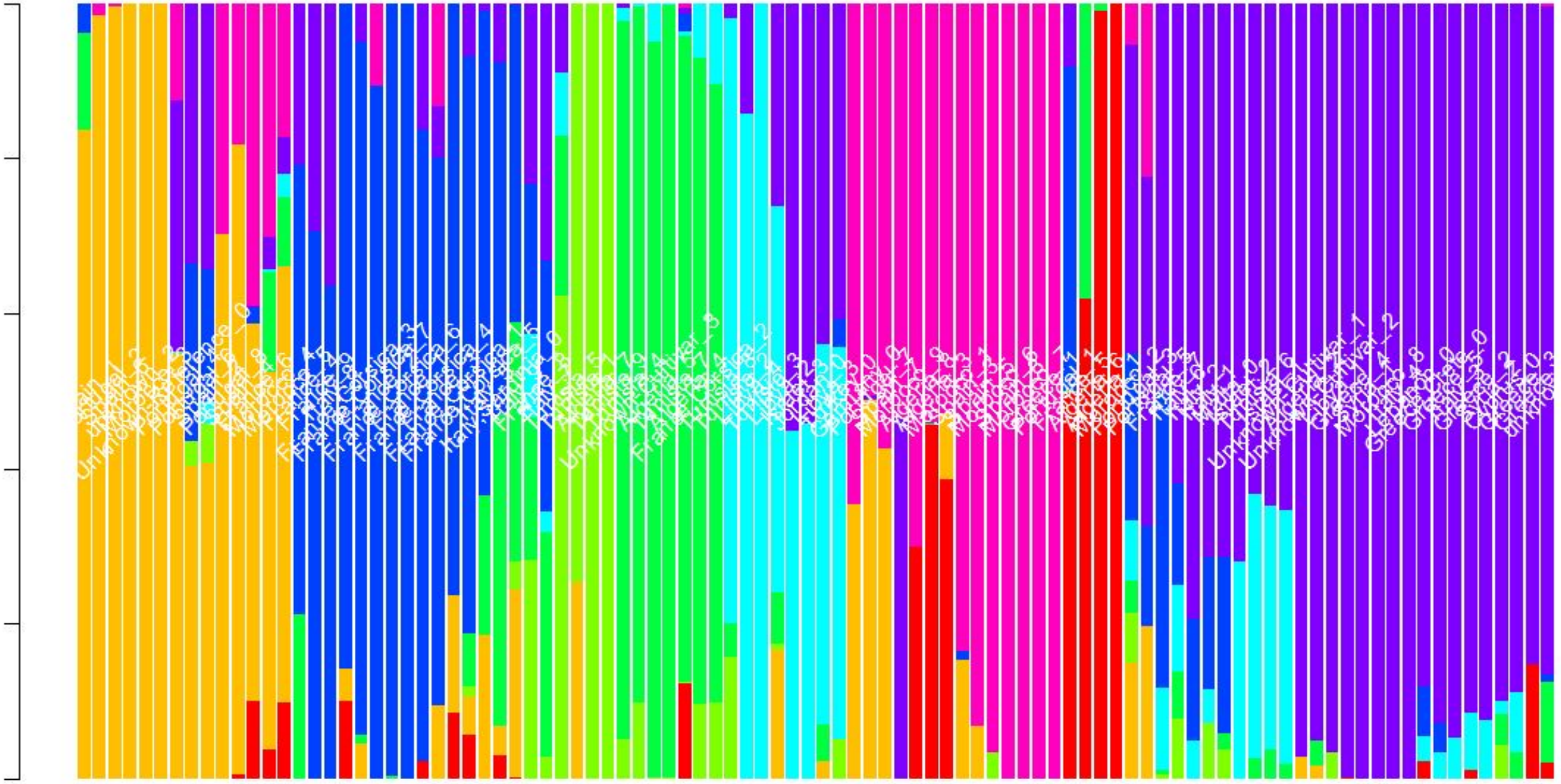
Practice

Admixture.csv

Pheno.csv

```
ADM=read.csv("Admixture.csv",row.names = 1)
# Instruction to name bars:
barN
  ret
  for
    if
      r
  }
  retu
}
# Siz
appe
par(r
x=ba
)
text(
```

Clustered vectors



Merge datasets

```
ADM=read.csv('Admixture.csv',row.names = 1);    TEST=subset(ADM, is.na(ADM$Lat))

  REF=subset(ADM, !is.na(ADM$Lat));    Phen=read.csv('pheno.csv',row.names = 1); head(Phen)

  #merge reference adm with pheno

  AMDPH=na.omit(merge(REF, Phen, by='row.names'))
```

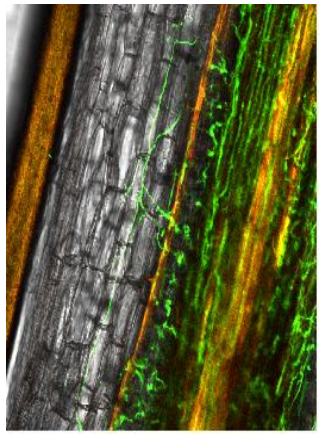
DATA=AMDPH[,c(4:6,8:15,17)]; cor(DATA)

	Lat	Long	Elev	K1	K2	K3	K4	K5	K6	K7	K8	MSS
Lat	1	-0.1208	-0.08527	-0.0795	0.031188	-0.05421	-0.15353	-0.20474	0.759062	-0.13096	-0.15013	0.031638
Long	-0.1208	1	-0.19809	-0.15237	-0.40834	0.075826	0.004852	0.248445	-0.04453	0.703125	-0.54288	0.39841
Elev	-0.08527	-0.19809	1	0.31451	-0.07847	-0.18658	0.268594	-0.23606	0.001009	-0.18933	0.069882	-0.004
K1	-0.0795	-0.15237	0.31451	1	-0.10838	-0.1018	-0.11645	-0.12181	-0.09978	-0.15409	0.025256	0.030009
K2	0.031188	-0.40834	-0.07847	-0.10838	1	-0.09725	-0.19296	-0.13028	-0.10436	-0.16469	-0.02408	-0.35086
K3	-0.05421	0.075826	-0.18658	-0.1018	-0.09725	1	-0.12727	-0.06539	-0.11848	-0.14095	-0.15113	0.152299
K4	-0.15353	0.004852	0.268594	-0.11645	-0.19296	-0.12727	1	-0.14567	-0.19199	-0.26904	-0.27647	0.056788
K5	-0.20474	0.248445	-0.23606	-0.12181	-0.13028	-0.06539	-0.14567	1	-0.14277	-0.04562	-0.18434	-0.09248
K6	0.759062	-0.04453	0.001009	-0.09978	-0.10436	-0.11848	-0.19199	-0.14277	1	-0.14562	-0.19251	0.086721
K7	-0.13096	0.703125	-0.18933	-0.15409	-0.16469	-0.14095	-0.26904	-0.04562	-0.14562	1	-0.25092	0.343465
K8	-0.15013	-0.54288	0.069882	0.025256	-0.02408	-0.15113	-0.27647	-0.18434	-0.19251	-0.25092	1	-0.30013
MSS	0.031638	0.39841	-0.004	0.030009	-0.35086	0.152299	0.056788	-0.09248	0.086721	0.343465	-0.30013	1

```
##GPS
 M=25;
for(i in 1:dim(TEST)[1]){
    Y=TEST[i,7:14];Y
    DIST=c()
    for(j in 1:dim(AMDPH)[1]){
      Z=AMDPH[j,8:15];Z
      d=sum((Y-Z)^2);d
      DIST=c(DIST,d)
    }
    I=order(DIST)[1:M];I
    Ph=AMDPH$MaxSymptomScore[I];Ph
    Dist=DIST[I]+1.E-15;Dist
    w=min(Dist)/Dist;w
    predPh=sum(w*Ph)/sum(w);predPh
    predLat=sum(w*AMDPH$Lat[I])/sum(w);predLat;     predLon=sum(w*AMDPH$Long[I])/sum(w);predLon
    output=paste(row.names(TEST[i,]),TEST$Pop[i],TEST$Country[i],predPh,predLat,predLon, sep=',')
    write.table(output,"GPS_RES.csv",append = T, row.names = F, col.names = F, quote=T)
 }
```

Sample	Pop	Country	MSS	Lat	Lon
HM112	SA28097	Cyprus	2.972435	36.44321	26.17896
HM115	Cyprus_C	Cyprus	2.972435	36.44321	26.17896
HM313	PI660408SSD	Cyprus	2.972435	36.44321	26.17896
HM111	SA27192	Italy	2.840499	36.70259	22.93486
HM293	PI660411SSD	Italy	2.609747	36.17789	21.94247
HM197	SA12455	Italy/Sardinia	2.011748	39.83869	7.174536
HM262	PI564941SSD	Morocco	1.728855	36.17965	-5.39374
HM280	D1.2.3	Syria	3.185377	38.69052	23.28353
HM256	PI442895SSD	unknown	0.966927	37.095	-5.132
HM269	PI660470SSD	unknown	2.699733	36.65048	20.22743
HM290	PI577640SSD	unknown	2.972435	36.44321	26.17896
HM316	PI660421SSD	unknown	2.972435	36.44321	26.17896
HM019	Borung	Unknown-Cultivar	2.887806	36.3135	8.496228
HM207	Caliph	Unknown-Cultivar	3.182311	36.43172	24.90595
HM208	Paraggio	Unknown-Cultivar	2.134387	33.83131	26.18263
HM209	Sephi	Unknown-Cultivar	2.972435	36.44321	26.17896
HM101	A17_Varma	Unknown-Reference	0.966927	37.095	-5.132

```
#GLM
library(lmtest); library(dplyr); library(car); library(rcompanion)
#set up formulas
formula=formula(MaxSymptomScore~K1+K2+K3+K4+K5+K6+K7+K8)
formula0=MaxSymptomScore ~ 1
#initialize models
model.null = glm(formula0,        data=AMDPH); summary(model.null)
model.full = glm(formula, data=AMDPH); summary(model.full )
#perform stepwise reduction
S=step(model.null,    scope = list(upper=model.full),    direction="both",   test="Chisq",      data=AMDPH)
summary(S)
#final model
model.final = glm(S$formula, data=AMDPH);model.final
summary(model.final)
AMDPH$predy = predict.glm(model.final, newdata = AMDPH,    type="response" )
```



Collaborations and funding

Laurent Gentzbittel, Ecolab, Toulouse, France/Skoltech, Moscow, Russia

Nevin Young's lab, Univ. of Minnesota, USA

Sergey Nuzdhin's lab, USC, USA

Chris Town's lab, JCVI, USA

Mounawer Badri & Naceur Djebali, CBBC, Tunisia

Jean-Marie Prospéri, INRA Mauguio, France

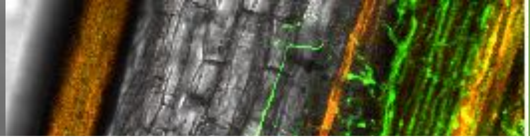
Nick Alexandrov & Dmitro Chebotarov, IRRI, Philippines

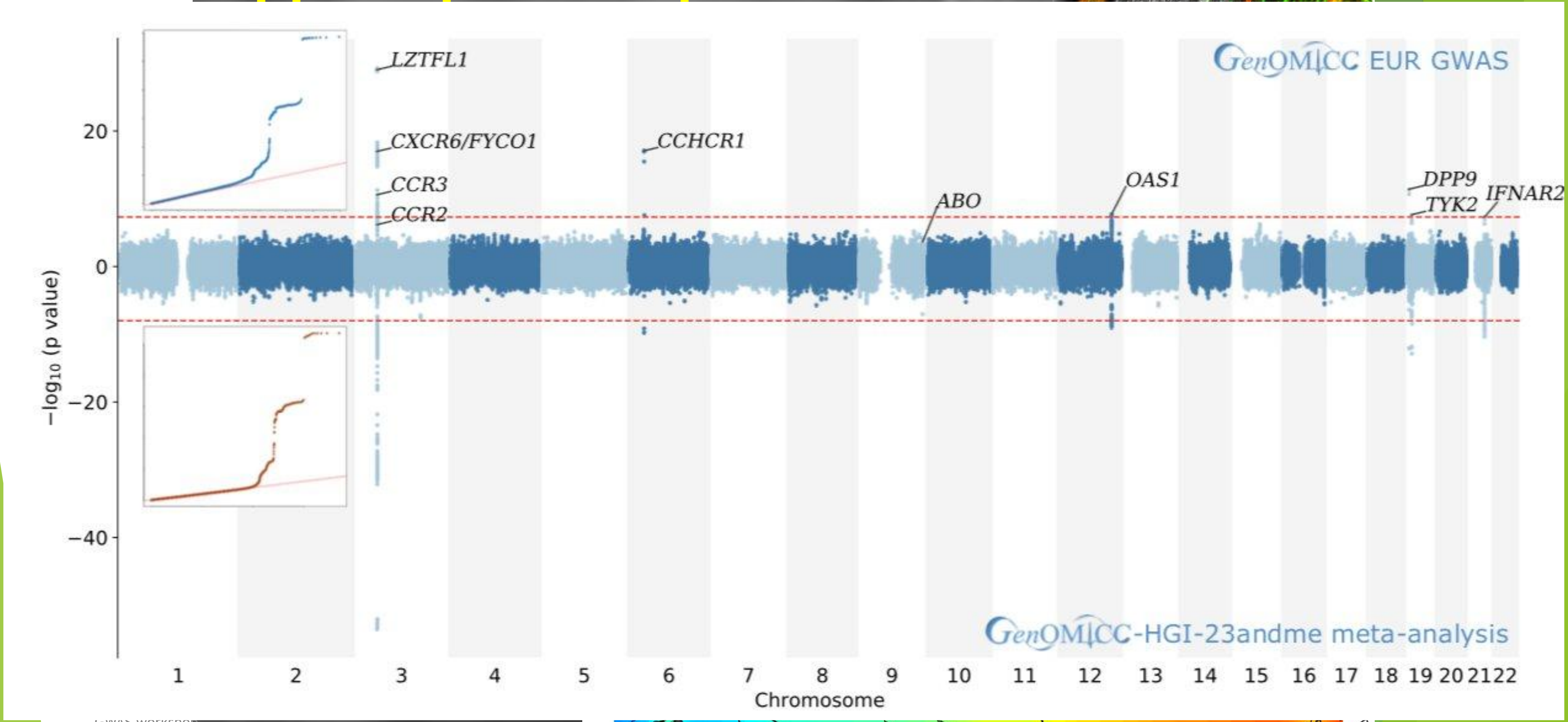
NSF GWAS workshop

UNIVERSITY OF SOUTHERN CALIFORNIA 1880

INP TOULOUSE ENSAT

ИНСТИТУТ БИОИНФОРМАТИКИ




GenOMICC EUR GWAS
LZTFL1
CXCR6/FYCO1
CCR3
CCR2
CCHCR1
ABO
OAS1
DPP9
TYK2
IFNAR2
20
0
−20
−40
−log10 (p value)
GenOMICC-HGI-23andme meta-analysis
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome