GenABEL: an R package for Genome Wide Association Analysis

Archana Bhardwaj

Outline

- R : conditional statements : if, else and for loop
- GeneABEL
- Genetic data QC
- GWA association analysis

Arithemetic operations

- Square roots, base-10 logarithm, and exponentiation can be done straightforwardly with R

```
> sqrt(5)
 [1] 2.236068
> log10(2.24)
[1] 0.350248
 > exp(0.35)
[1] 1.419068
```

- The arithmetic operations and functions can be nested:

```
> exp(log10(sqrt(2 + 3)))
 [1] 1.418337
```

Conditional execution : *if* statement

```
> x = 0.1
 if( x < 0.2)
{ x <- x + 1
cat("increment that number!\n")
 }
Increment that number!

>x
[1] 1.1
```

Conditional execution : *if* statement and the corresponding *else* statement

```
> x = 2.0
> if ( x < 0.2)
{
 x <- x + 1
cat("increment that number!\n")
} else
 { x <- x - 1
cat("nah, make it smaller.\n");
} nah,make it smaller.

> x
[1] 1
```

Conditional execution : *for loop*

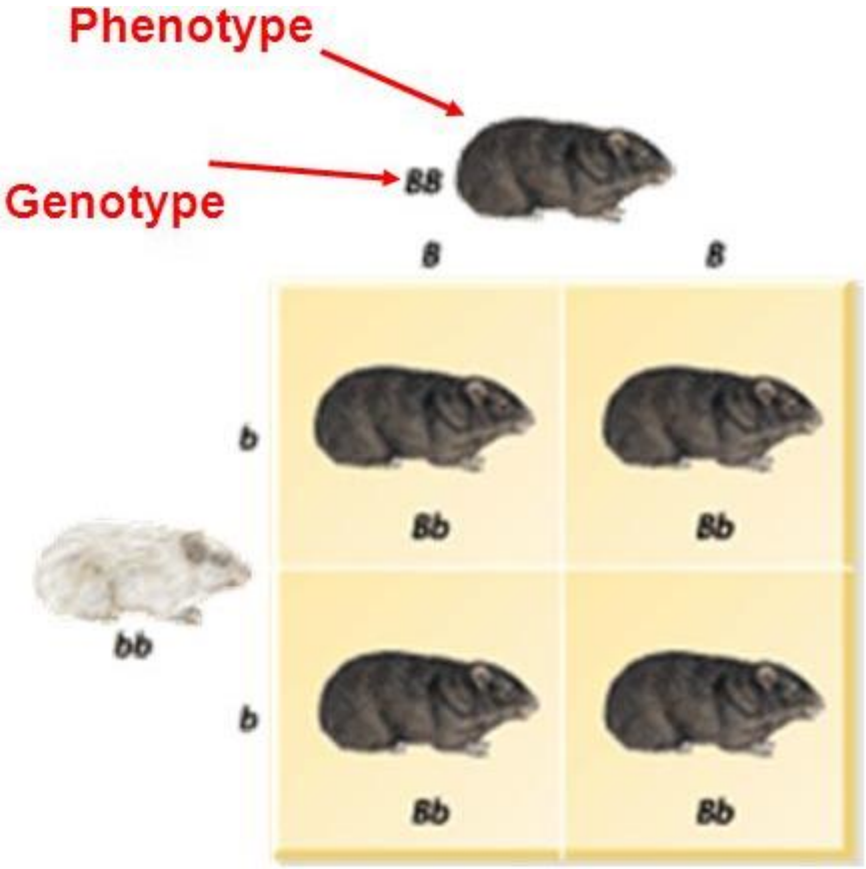
```
➢ for (list in seq(0,1,by=0.3))
 { cat(list,"\n"); } 0 0.3 0.6 0.9

> x <- c(1,2,4,8,16)
for (loop in x)
{ cat("value of loop: ",loop,"\n");
}
 value of loop: 1
value of loop: 2
value of loop: 4
value of loop: 8
value of loop: 16
```

Introduction

- **GenABEL is an R library developed to facilitate Genome-Wide Association analysis of binary and quantitative traits.**
- **Features of GenABEL :**
  - **specific facilities for storage and manipulation of large data**
  - **QC**
  - **Maximum Likelihood estimation of linear, logistic and Cox regression on Genome-wide scale**
  - **Specific functions to analyze and display the results**

Relationship between Genotypes and Phenotypes

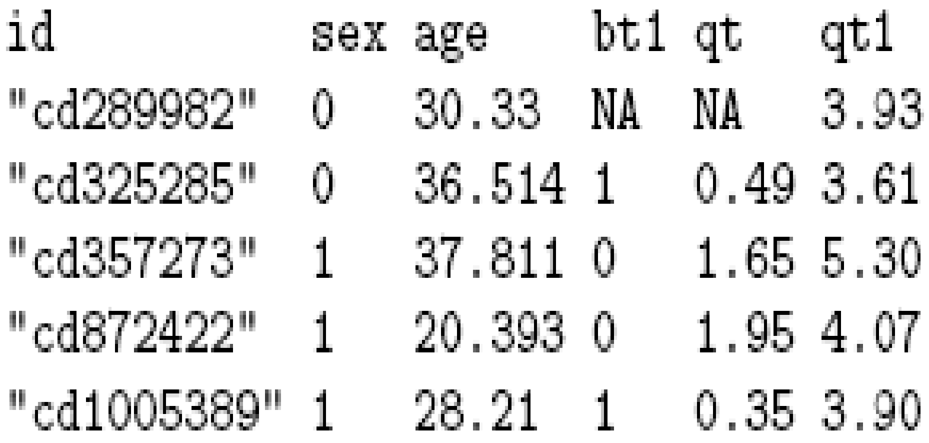
- **Genotype:** Indicates the alleles that the organism has inherited regarding a particular trait.
- **Phenotype:** The actual visible trait of the organism.



Importing data to GenABEL(1)

➢ **Need a phenotypic and genotypic data**

- **Example of a phenotype file :**

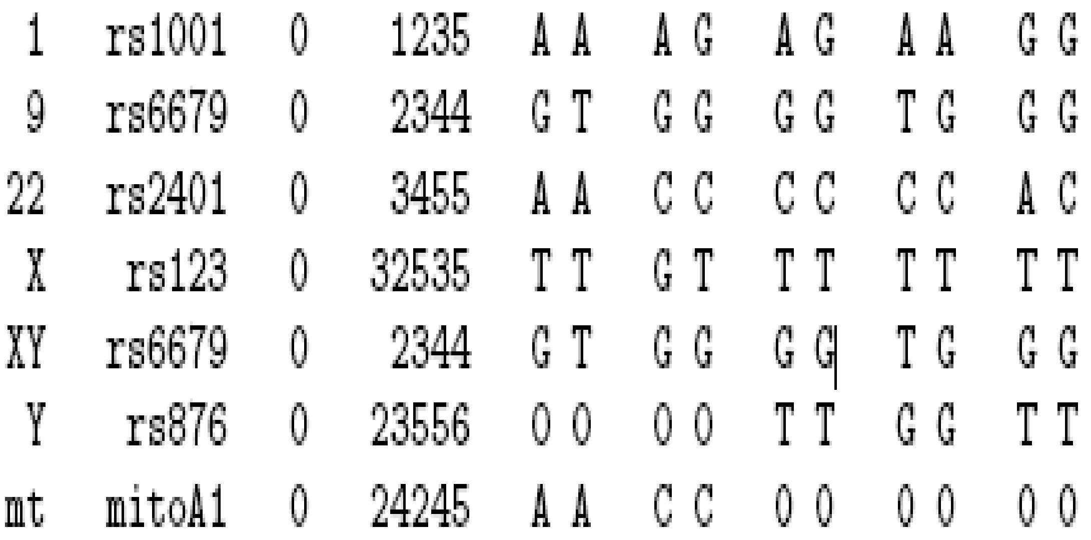


```
id          sex age    bt1 qt   qt1
"cd289982"  0   30.33  NA  NA   3.93
"cd325285"  0   36.514 1   0.49 3.61
"cd357273"  1   37.811 0   1.65 5.30
"cd872422"  1   20.393 0   1.95 4.07
"cd1005389" 1   28.21  1   0.35 3.90
```

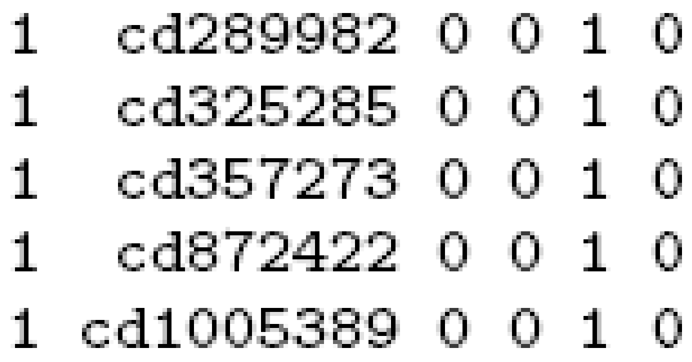
Definition: A phenotype is the composite of an organism's observable characteristics or traits, such as its morphology, development, biochemical or physiological properties, behavior, and products of behavior. For example : eyes color, height

Importing data to GenABEL(2)

- Need a phenotypic and genotypic data
  - Example of a genotypic data



```
1   rs1001  0   1235  A A  A G  A G  A A  G G
9   rs6679  0   2344  G T  G G  G G  T G  G G
22  rs2401  0   3455  A A  C C  C C  C C  A C
X    rs123  0  32535  T T  G T  T T  T T  T T
XY  rs6679  0   2344  G T  G G  G G  T G  G G
Y    rs876  0  23556  0 0  0 0  T T  G G  T T
mt  mitoA1  0  24245  A A  C C  0 0  0 0  0 0
```



```
1   cd289982 0 0 1 0
1   cd325285 0 0 1 0
1   cd357273 0 0 1 0
1   cd872422 0 0 1 0
1 cd1005389 0 0 1 0
```

GWAS main philosophy

- **GWAS = Genome Wide Association Studies**
- **IDEA = GWAS involve scan for large number of genetic markers across the whole genome of many individuals to find specific genetic variations associated with the disease and/or other phenotype**
- **Find the genetic variation(s) that contribute(s) and explain(s) complex diseases**

GWAS visually

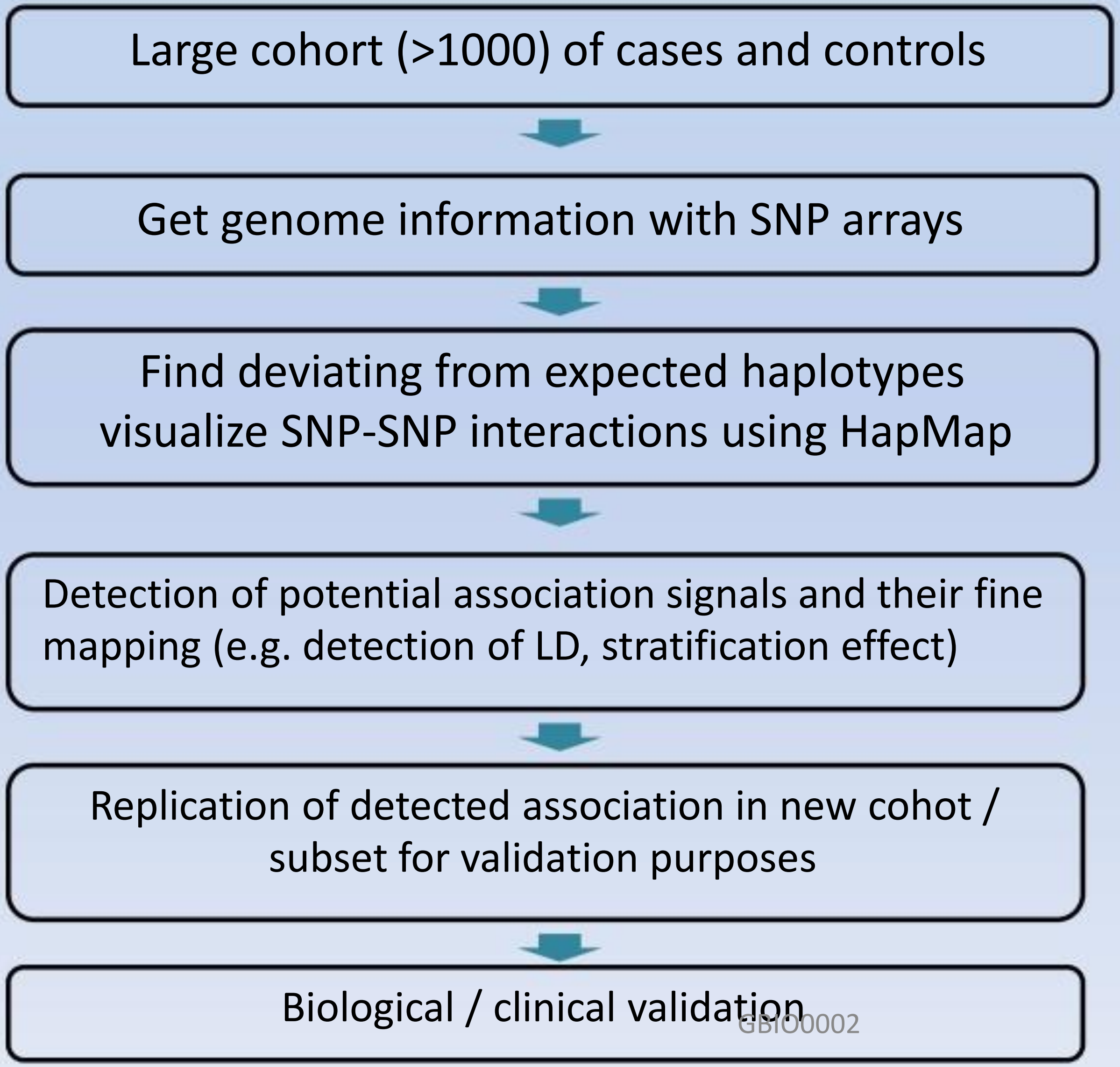
- GWAS tries to uncover links between genetic basis of the disease

- Which set of SNPs explain the phenotype?

Genotype	Phenotype
ATGCAGTT	control
TTGCAGTT	control
CTGCAGTT	control
ATGCGGTT	case
TTGCGGTT	case
CTGCCGTT	case

SNP

GWAS workflow



Large cohort (>1000) of cases and controls

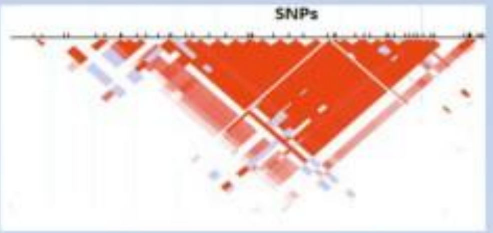
Get genome information with SNP arrays

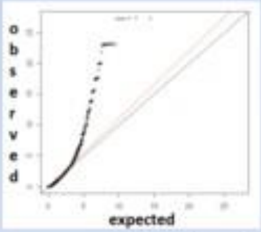
Find deviating from expected haplotypes
visualize SNP-SNP interactions using HapMap

Detection of potential association signals and their fine mapping (e.g. detection of LD, stratification effect)

Replication of detected association in new cohot / subset for validation purposes

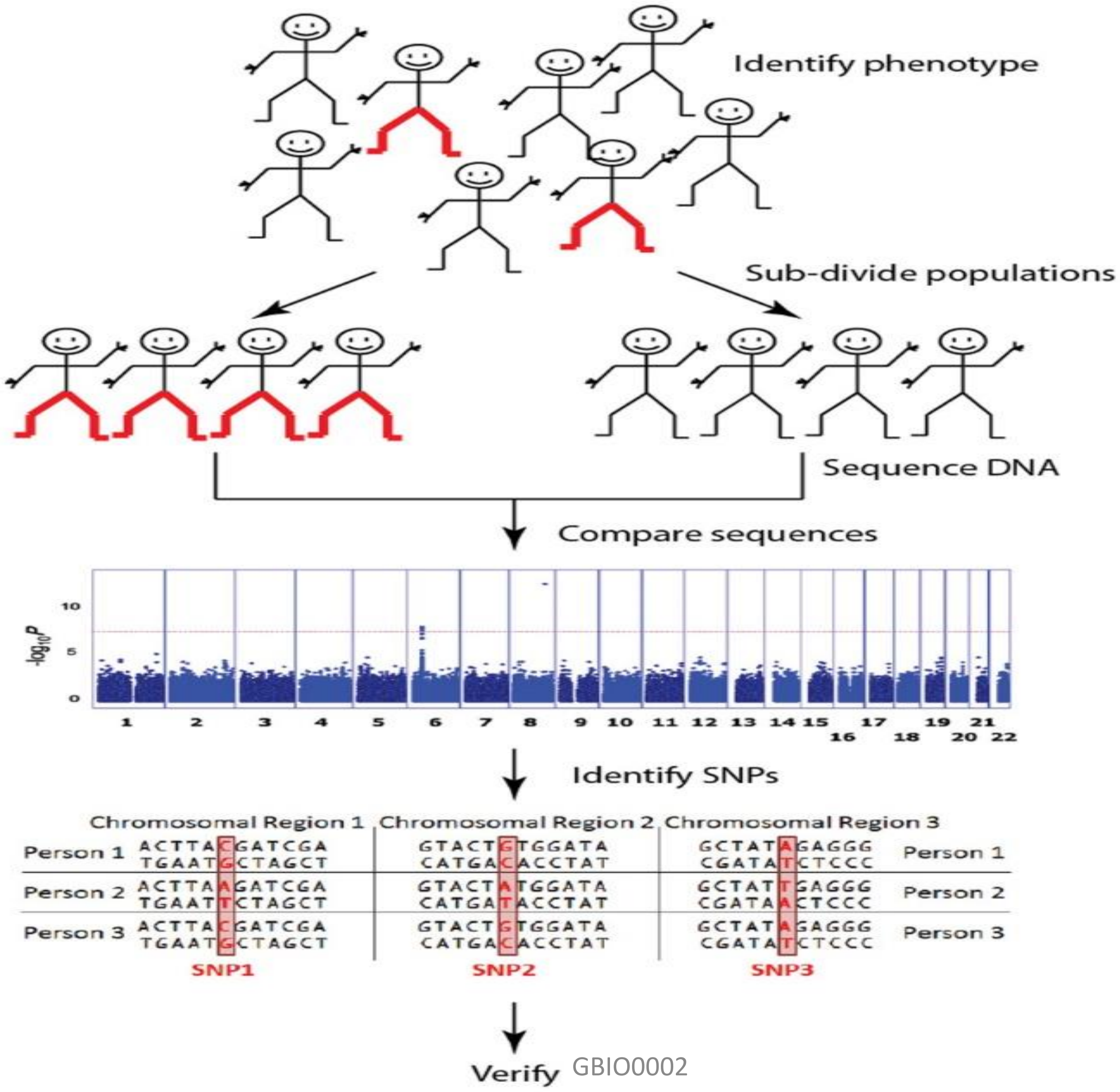
Biological / clinical validation





	AT	AG	Total
cases observed	35	65	100
contorls observed	125	25	100
Totals	160	90	200

GWAS workflow



	Chromosomal Region 1	Chromosomal Region 2	Chromosomal Region 3	
Person 1	ACTTACGATCGA TGAATGCTAGCT	GTACTGTGGATA CATGACACCTAT	GCTATAGAGGG CGATATCTCCC	Person 1
Person 2	ACTTAAGATCGA TGAATTCTAGCT	GTACTATGGATA CATGATACCTAT	GCTATTGAGGG CGATAACTCCC	Person 2
Person 3	ACTTACGATCGA TGAATGCTAGCT	GTACTGTGGATA CATGACACCTAT	GCTATAGAGGG CGATATCTCCC	Person 3
	SNP1	SNP2	SNP3	

Which tools to use to do GWAS workflows?

How to find SNP-Disease associations?

Common tools

- Some of the popular tools
  - SVS Golden Helix (data filtering and normalization)

    - Is commercial software providing ease of use compared to other free solutions requiring use of numerous libraries

    - Has unique feature on CNV Analysis

    - Manual: http://doc.goldenhelix.com/SVS/latest/
  - Biofilter (pre-selection of SNPs using database info)
  - GenABEL library implemented in R (http://www.genabel.org/)

Introduction to GenABEL (1/2)

- This library allows to do complete GWAS workflow

- GWAS data and corresponding attributes (SNPs, phenotype, sex, etc.) are stored in data object gwas.data-class

- The object attributes could be accessed with @

  - phenotype data: gwaa_object@phdata

  - number of people in study: gwaa_object@gtdata@nids

Introduction to GenABEL(2/2)

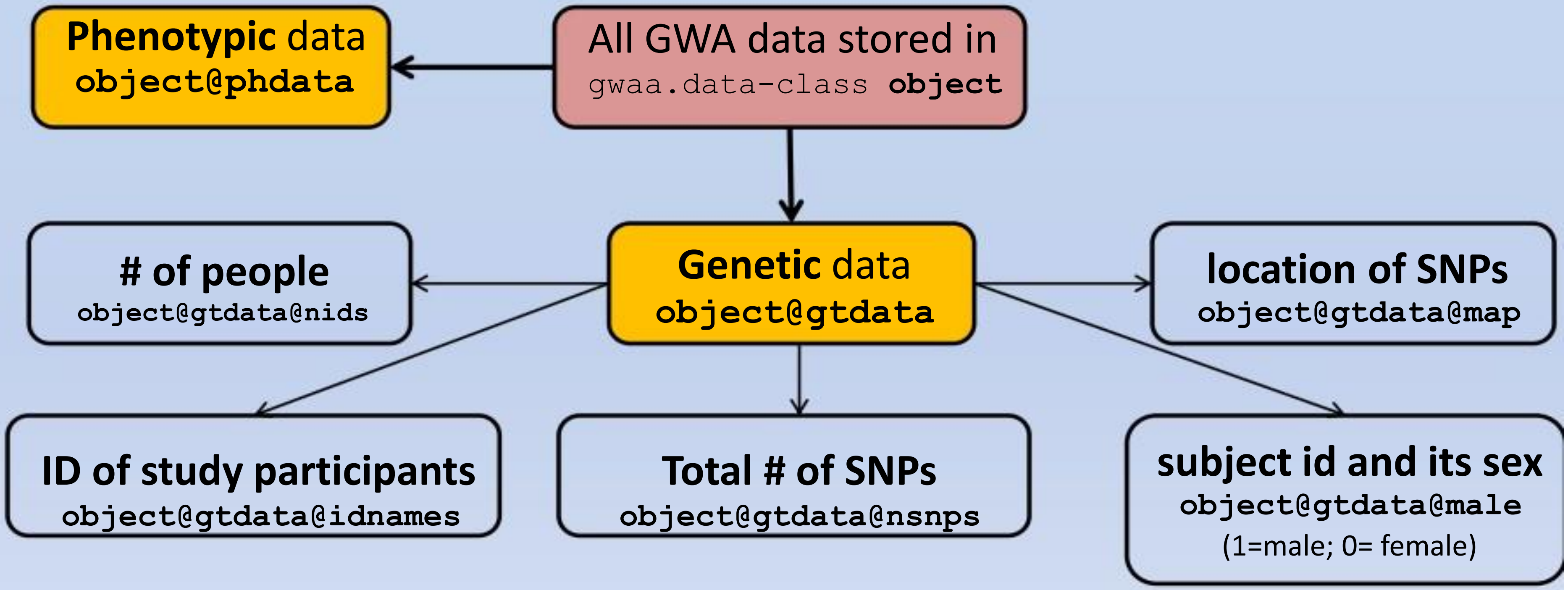
- number of SNPs: gwaa_object@gtdata@nsnps

- SNP names: gwaa_object@gtdata@snpnames

- Chromosome labels:gwaa_object@gtdata@chromosome

- SNPs map positions: gwaa_object@gtdata@map

GeneABEL object structure



Phenotypic data
object@phdata

All GWA data stored in
gwaa.data-class **object**

Genetic data
object@gtdata

# of people
object@gtdata@nids

location of SNPs
object@gtdata@map

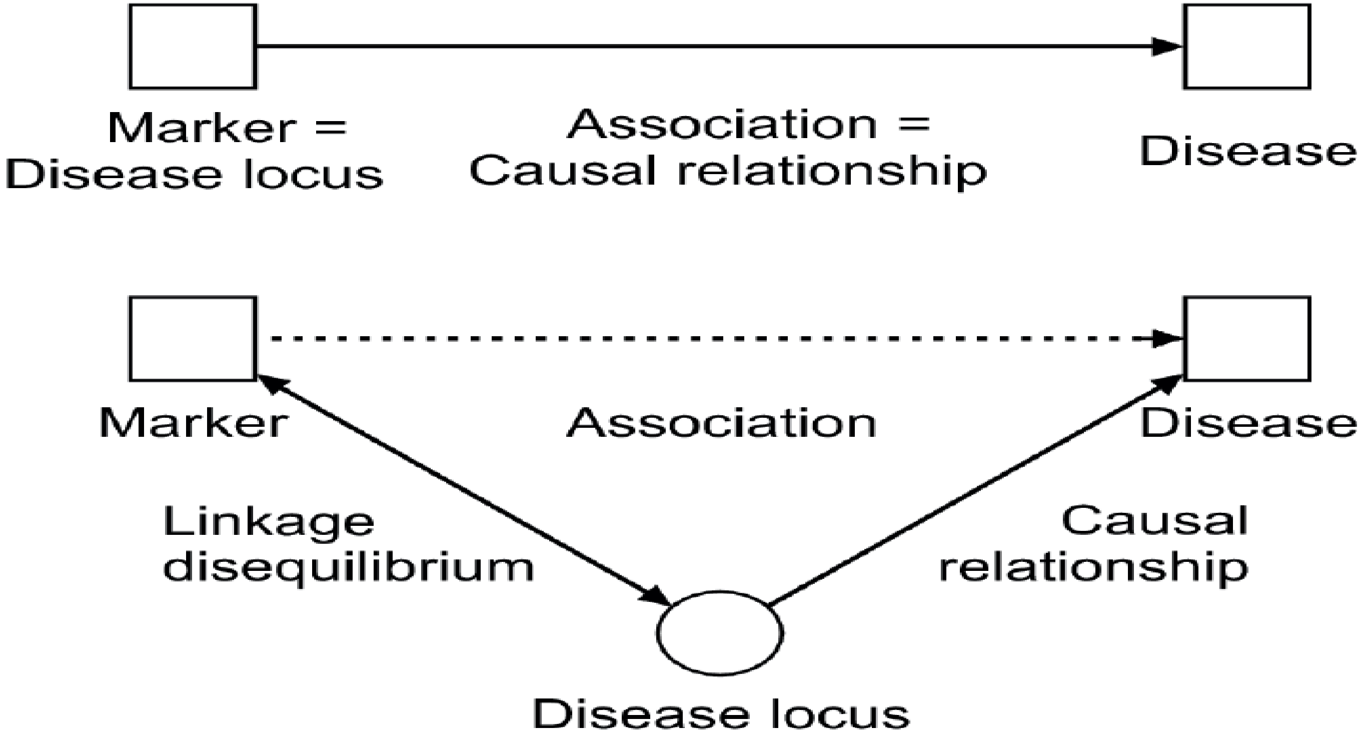
ID of study participants
object@gtdata@idnames

Total # of SNPs
object@gtdata@nsnps

subject id and its sex
object@gtdata@male
(1=male; 0= female)

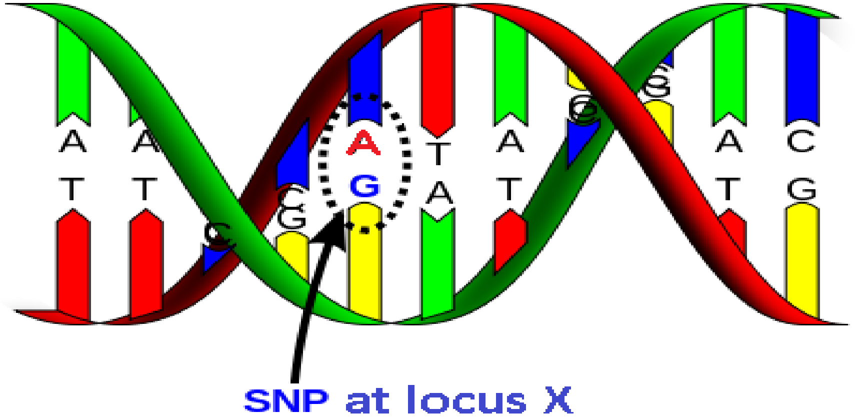
The main question of GWA studies

- What is the **causal model** underlying genetic **association**?



Important genetic terms

- Given position in the genome (i.e. locus) has several associated alleles (A and G) which produce genotypes r_A/r_G

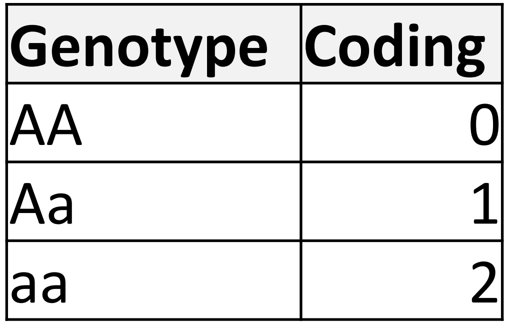


- Haplotypes

  - Combination of alleles at different loci

Genotype coding

- **For given bi-allelic marker/SNP/loci there could be total of 3 possible genotypes given alleles A and a**

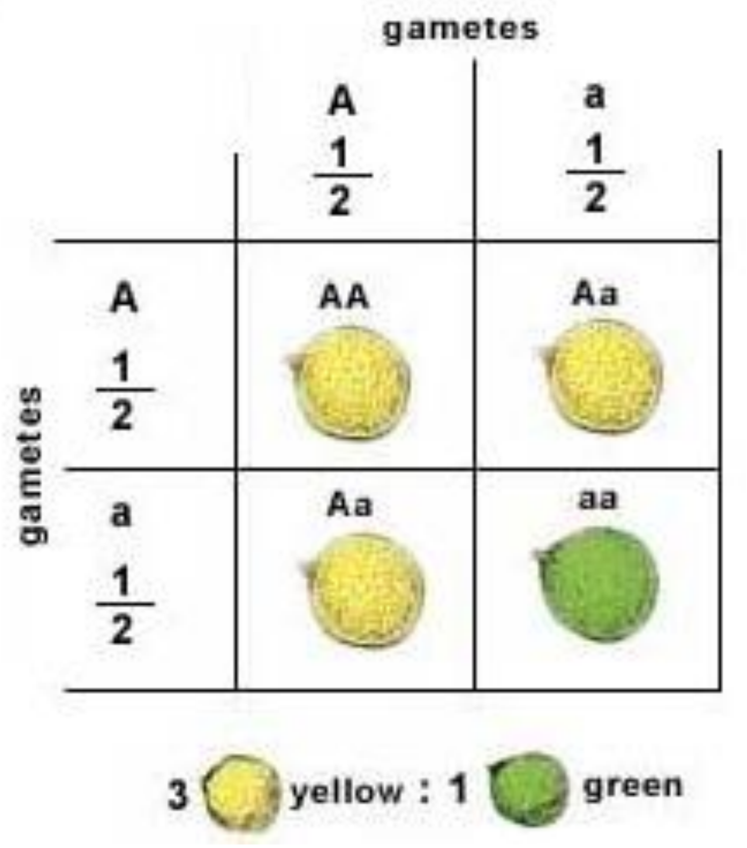


Genotype	Coding
AA	0
Aa	1
aa	2

Note: A is major allele and a is minor

Genetic allelic dominance

- Dominance describes relationship of two alleles (A and a) in relation to final phenotype
  - if one allele (e.g. A) “masks” the effect of other (e.g. a) it is said to be <u>dominant</u> and masked allele <u>recessive</u>
  - Here the dominant allele A gives pea a yellow color

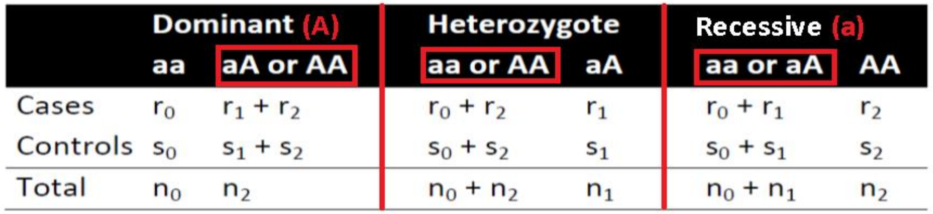


Source: http://nissemann.blogspot.be/2009_04_01_archive.html

Homozygote dominant:	AA
Heterozygote:	Aa
Homozygote recessive:	aa

Genotype genetic based models

- Hypothesis (H_o): the genetic effects of AA and Aa are the same (A is dominant allele)
- Hypothesis (H_o): the genetic effects of aa and AA are the same
- Hypothesis (H_o): the genetic effects of aa and Aa are the same (a is recessive allele)



	Dominant (A)		Heterozygote		Recessive (a)	
	aa	aA or AA	aa or AA	aA	aa or aA	AA
Cases	r_0	$r_1 + r_2$	$r_0 + r_2$	r_1	$r_0 + r_1$	r_2
Controls	s_0	$s_1 + s_2$	$s_0 + s_2$	s_1	$s_0 + s_1$	s_2
Total	n_0	n_2	$n_0 + n_2$	n_1	$n_0 + n_1$	n_2

- $\chi^2_{dom} = n \cdot \frac{\left(r_0(s_1 + s_2) - (r_1 + r_2)s_0\right)^2}{r \cdot s \cdot n_0 \cdot (n_1 + n_2)}$ Dominant
- $\chi^2_{het} = n \cdot \frac{\left(r_1(s_0 + s_2) - (r_0 + r_2)s_1\right)^2}{r \cdot s \cdot n_1 \cdot (n_0 + n_2)}$ Heterozygous
- $\chi^2_{rec} = n \cdot \frac{\left((r_0 + r_1)s_2 - r_2(s_0 + s_1)\right)^2}{r \cdot s \cdot (n_0 + n_1) \cdot n_2}$ Recessive

Working on Linux? : GenABEL Installation (1/2)

- Download source code from website https://cran.r-project.org/src/contrib/Archive/GenABEL/

- Download most updated version 1.8

- Go to link and download dependent packages

  https://mran.microsoft.com/snapshot/2016-10-02/web/packages/GenABEL.data/index.html

GenABEL Installation (2/2)

- **First install the dependency packages**

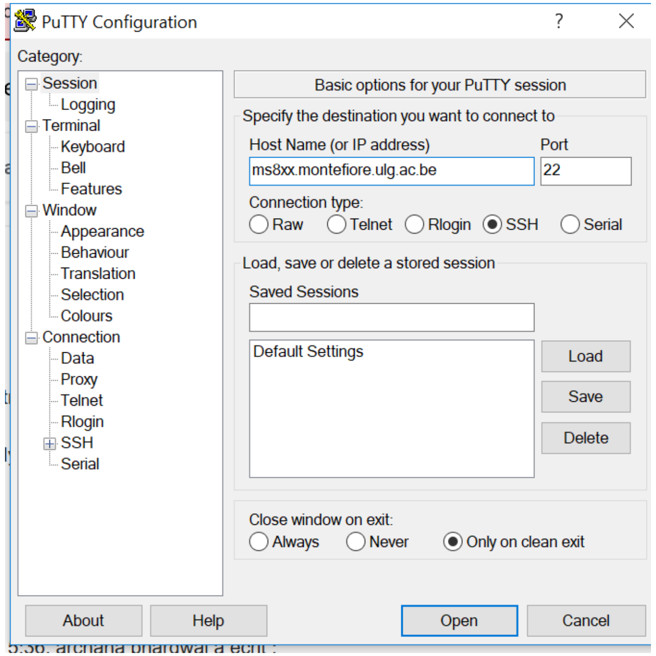
```
               > install.packages('genetics")
              > install.packages('haplo.stats")
```

- **Do installation by running following command**

```
        > install.packages("/dirpath/GenABEL.data_1.0.0.tar.gz",
repos=NULL, type="source")
        > install.packages('dirpath/GenABEL_1.8-0.tar.gz', repos=NULL,
type="source')
```

Working on Windows?

- ➢ **Download Putty from https://putty.org/**
- ➢ **Step 1 : Login into ms8xx.montefiore.ulg.ac.be (01 <= xx <= 18)**



PuTTY Configuration

Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address): ms8xx.montefiore.ulg.ac.be

Port: 22

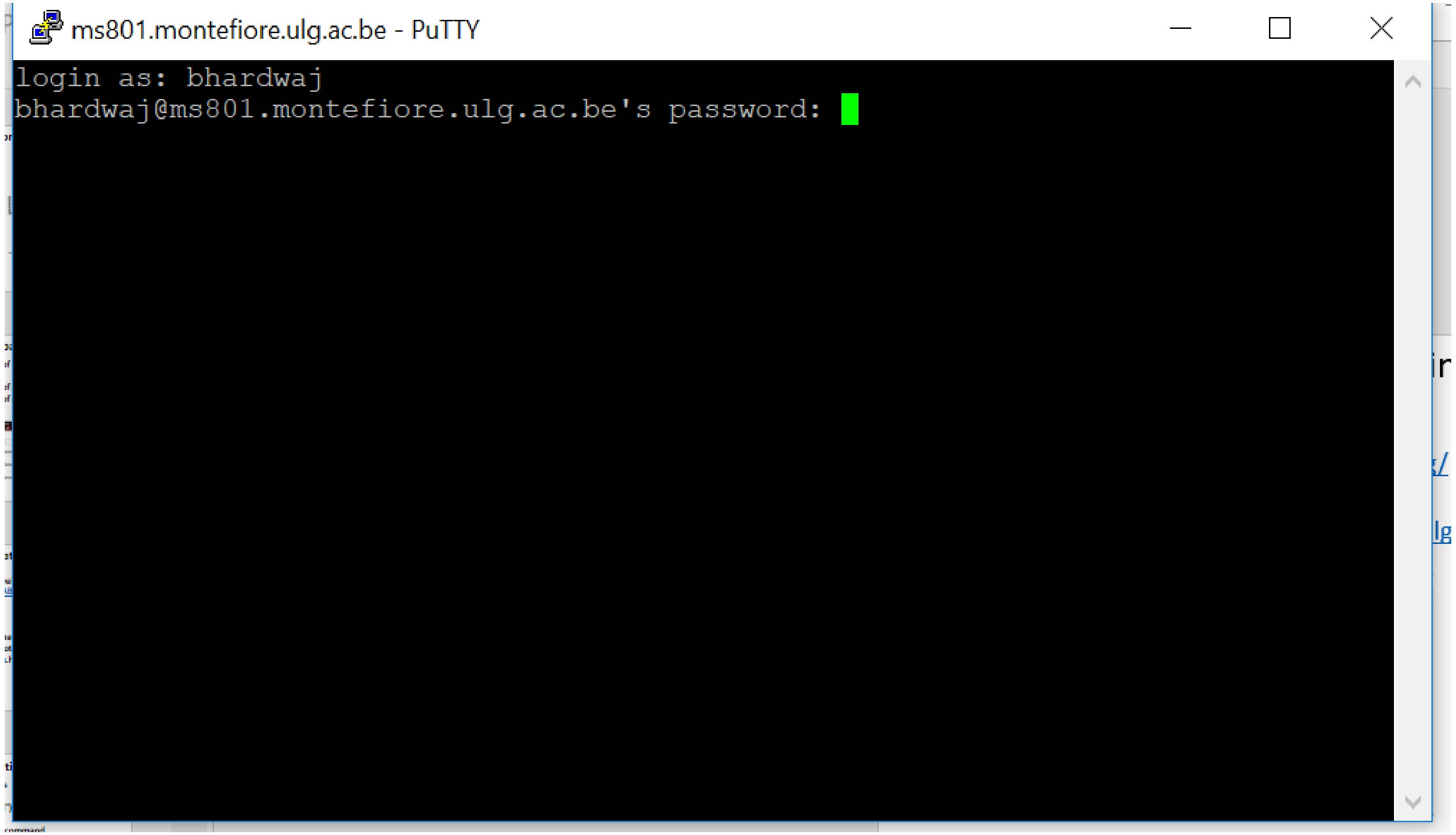
Connection type: ○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

Load, save or delete a stored session

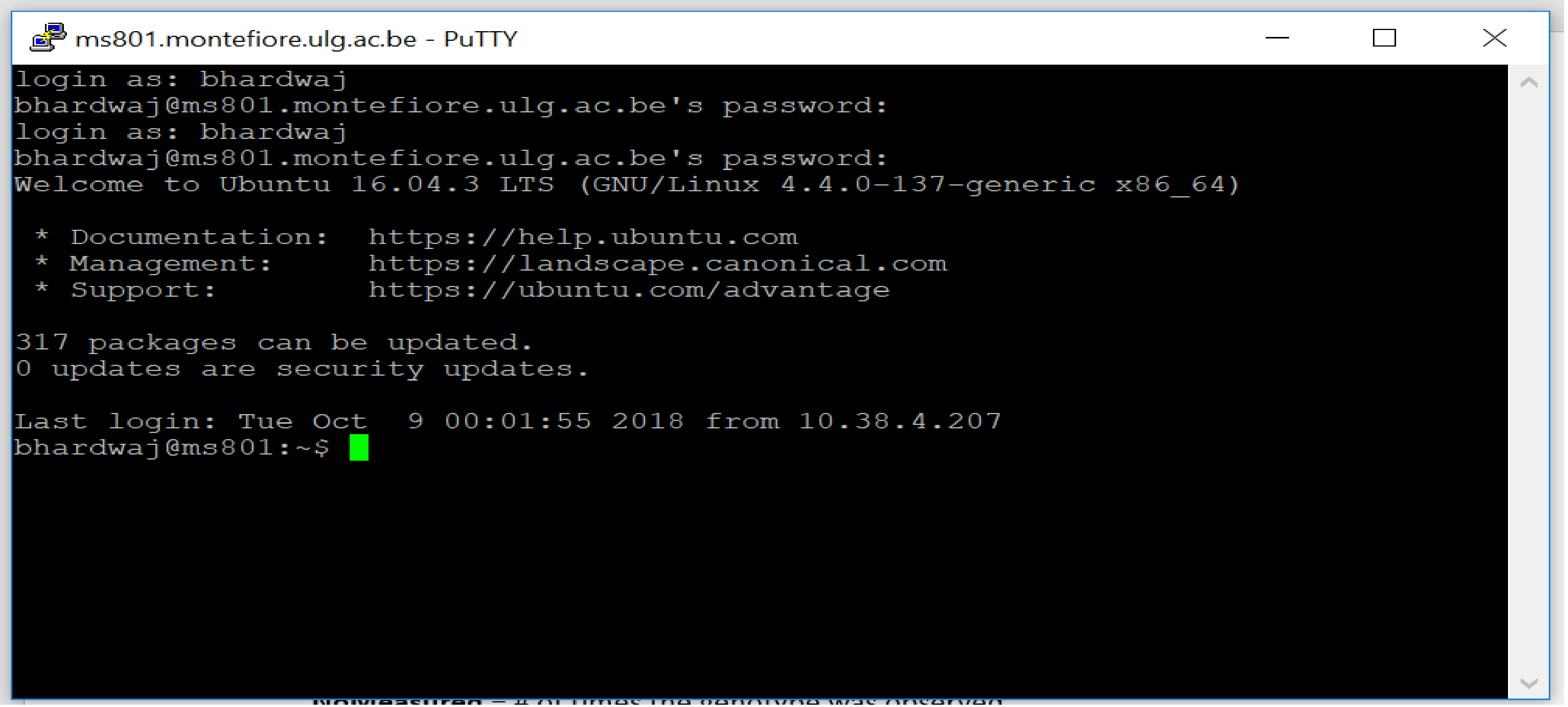
Saved Sessions: Default Settings

Close window on exit: ○ Always ○ Never ● Only on clean exit

- Step 2 : Enter the username and password into ms801.montefiore.ulg.ac.be (01 <= xx <= 18)



```
login as: bhardwaj
bhardwaj@ms801.montefiore.ulg.ac.be's password: 
```



```
login as: bhardwaj
bhardwaj@ms801.montefiore.ulg.ac.be's password:
login as: bhardwaj
bhardwaj@ms801.montefiore.ulg.ac.be's password:
Welcome to Ubuntu 16.04.3 LTS (GNU/Linux 4.4.0-137-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

317 packages can be updated.
0 updates are security updates.

Last login: Tue Oct  9 00:01:55 2018 from 10.38.4.207
bhardwaj@ms801:~$ 
```

Connnected !!

Hands on GenABEL

➢ Load package

```
> library("GenABEL")

> data(ge03d2ex)                        #loads data(i.e. from GWAS) on type 2 diabetes
 > summary(ge03d2ex)[1:5,]              #view first 5 SNP data  (genotypic data)
```

```
          Chromosome  Position   Strand  A1 A2 NoMeasured CallRate    Q.2       P.11 P.12  P.22  Pexact        Fmax        Plrt
rs1646456          1       653        +   C  G        135 0.9926471 0.33333333   57   66   12 0.3323747 -0.10000000 0.2404314
rs4435802          1      5291        +   C  A        134 0.9852941 0.07462687  114   20    0 1.0000000 -0.08064516 0.2038385
rs946364           1      8533        -   T  C        134 0.9852941 0.27611940   68   58    8 0.3949055 -0.08275286 0.3302839
rs299251           1     10737        +   A  G        135 0.9926471 0.04444444  123   12    0 1.0000000 -0.04651163 0.4549295
rs2456488          1     11779        +   G  C        135 0.9926471 0.34814815   59   58   18 0.5698988  0.05343327 0.5360019
```

>= 0,98 means good genotyping

```
 > summary(ge03d2ex@phdata)  #view phenotypic data
```

```
 id                 sex              age              dm2              height           weight           diet              bmi
 Length:136         Min.   :0.0000   Min.   :23.84    Min.   :0.0000   Min.   :150.2    Min.   : 46.63   Min.   :0.00000   Min.   :17.30
 Class :character   1st Qu.:0.0000   1st Qu.:38.33    1st Qu.:0.0000   1st Qu.:161.5    1st Qu.: 69.02   1st Qu.:0.00000   1st Qu.:24.56
 Mode  :character   Median :1.0000   Median :48.71    Median :1.0000   Median :169.4    Median : 81.15   Median :0.00000   Median :28.35
                    Mean   :0.5294   Mean   :49.07    Mean   :0.6324   Mean   :169.4    Mean   : 87.40   Mean   :0.05882   Mean   :30.30
                    3rd Qu.:1.0000   3rd Qu.:58.57    3rd Qu.:1.0000   3rd Qu.:175.9    3rd Qu.:102.79   3rd Qu.:0.00000   3rd Qu.:35.69
                    Max.   :1.0000   Max.   :81.57    Max.   :1.0000   Max.   :191.8    Max.   :161.24   Max.   :1.00000   Max.   :59.83
                                                                       NA's   :1        NA's   :1                          NA's   :1
```

A1 A2 = allele 1 and 2

Strand = DNA strand + or -

NoMeasured = # of times the genotype was observed

Pexact = P-value of the exact test for HWE

Fmax = estimate of deviation from HWE, allowing meta-analysis

Position = genomic position (bp)

CallRate = allelic frequency expressed as a ratio

Exploring phenotypic data

- **See aging phenotype data in compressed form**

```
> descriptives.trait(ge03d2ex)
        No    Mean     SD
id     136      NA     NA
sex    136   0.529  0.501
age    136  49.069 12.926
dm2    136   0.632  0.484
height 135 169.440  9.814
weight 135  87.397 25.510
diet   136   0.059  0.236
bmi    135  30.301  8.082
```

- **Extract all sexes of all individuals**

```
> ge03d2ex@phdata$sex # accessing sex column of the data frame with $
1 0 1 0 0 1 1 0 0 1 0 0 1 1 0 0 0 1 1 1 1 1 1 1 1 1 0 1 0 1 0 1 1 0 0 0 1 1 1 1 0 1 1
1 0 1 1 1 1 0 1 1 1 1 0 1 0 1 0 0 1 0 1 1 0 1 0 0 1 0 1 1 1 1 1 0 1 0 0 1 0 1 0 0 0 0 0
0 0 0 1 0 0 0 1 1 0 1 1 0 1 0 1 0 1 1 0 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 1 0 1 1 1 0 1 0
0 1 1 0
```

- **Sorting data by binary attribute (e.g. sex)**

```
> descriptives.trait(ge03d2ex, by=ge03d2ex@phdata$sex)

 No(by.var=0)    Mean      SD No(by.var=1)     Mean     SD   Ptt    Pkw  Pexact
id              64      NA      NA          72      NA     NA    NA    NA      NA
sex             64      NA      NA          72      NA     NA    NA    NA      NA
age             64  46.942  12.479          72  50.959 13.107 0.070 0.081      NA
dm2             64   0.547   0.502          72   0.708  0.458 0.053 0.052   0.074
height          64 162.680   6.819          71 175.534  7.943 0.000 0.000      NA
weight          64  78.605  26.908          71  95.322 21.441 0.000 0.000      NA
diet            64   0.109   0.315          72   0.014  0.118 0.025 0.019   0.026
bmi             64  29.604   9.506          71  30.930  6.547 0.352 0.040      NA
```

- **How many people are included in the study?**

```
> nids(ge03d2ex)
```

- **How many of these are males?**

```
> sum(male(ge03d2ex))
```

- **How many are females?**

```
> nids(ge03d2ex) - sum(male(ge03d2ex) )
```

- **What is male proportion?**

```
> sum(male(ge03d2ex)) / nids(ge03d2ex)
```

Exploring genotypic data / statistics

```
>descriptives.marker(ge03d2ex)
```

```
$`Minor allele frequency distribution`
     X<=0.01 0.01<X<=0.05 0.05<X<=0.1 0.1<X<=0.2    X>0.2
No   146.000      684.000     711.000     904.000 1555.000
Prop   0.036        0.171       0.178       0.226    0.389

$`Cumulative distr. of number of SNPs out of HWE, at different alpha`
     X<=1e-04 X<=0.001 X<=0.01 X<=0.05 all X
No     46.000   71.000 125.000 275.000  4000
Prop    0.012    0.018   0.031   0.069     1

$`Distribution of proportion of successful genotypes (per person)`
     X<=0.9 0.9<X<=0.95 0.95<X<=0.98 0.98<X<=0.99 X>0.99
No    1.000           0            0      135.000      0
Prop  0.007           0            0        0.993      0

$`Distribution of proportion of successful genotypes (per SNP)`
     X<=0.9 0.9<X<=0.95 0.95<X<=0.98 0.98<X<=0.99    X>0.99
No   37.000       6.000      996.000     1177.000 1784.000
Prop  0.009       0.002        0.249        0.294    0.446

$`Mean heterozygosity for a SNP`
[1] 0.2582298

$`Standard deviation of the mean heterozygosity for a SNP`
[1] 0.1592255

$`Mean heterozygosity for a person`
[1] 0.2476507

$`Standard deviation of mean heterozygosity for a person`
[1] 0.04291038
```

Number of copies minor/rare alleles out of total number (i.e. 4000)

Total number of SNPs

Proportion of missing values (subject id9049 has missing info on "sex" "age" "dm2" "height" "weight" "diet" "bmi")

Number of SNPs that successfully were able to identify sample genotype (i.e. call rate). E.g. in this case 98% SNPs were able to identify / explain more than 96% genotypes

Only 25% SNPs are heterozygous (i.e. have different types of alleles)

Assessing quality of the raw data (1)

- **Test for Hardy-Weinberg equilibrium given set of SNPs**

  - in controls (i.e. dm2 = 0)

```
> dim(ge03d2ex@gtdata)
[1]  136 4000
> ge03d2ex@phdata$dm2
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
> summary(ge03d2ex@gtdata[(ge03d2ex@phdata$dm2 == 0),])
          Chromosome Position Strand A1 A2 NoMeasured CallRate        Q.2 P.11 P.12 P.22    Pexact        Fmax      Plrt
rs1646456          1      653      +  C  G         50     1.00 0.34000000   23   20    7 0.5275140  0.10873440 0.4448755
rs4435802          1     5291      +  C  A         48     0.96 0.04166667   44    4    0 1.0000000 -0.04347826 0.6766092
rs946364           1     8533      -  T  C         50     1.00 0.33000000   22   23    5 1.0000000 -0.04025328 0.7750907
rs299251           1    10737      +  A  G         50     1.00 0.06000000   44    6    0 1.0000000 -0.06382979 0.5358747
rs2456488          1    11779      +  G  C         50     1.00 0.37000000   21   21    8 0.5450202  0.09909910 0.4849983
#extract the exact HWE test P-values into separate vector "Pexact"
> Pexact0<-summary(ge03d2ex@gtdata[(ge03d2ex@phdata$dm2 == 0),])[,"Pexact"]
# perform chi square test on the Pexact values and calculate λ (inflation factor).
# If λ=1.0 no inflation or diflation of test statistic (i.e. no stratification effect)
> estlambda(Pexact0, plot=TRUE)
$estimate
[1] 1.029817
$se
[1] 0.002185684
```

Inflation of test statistic (`Pexact`) is seen, see stratification effect

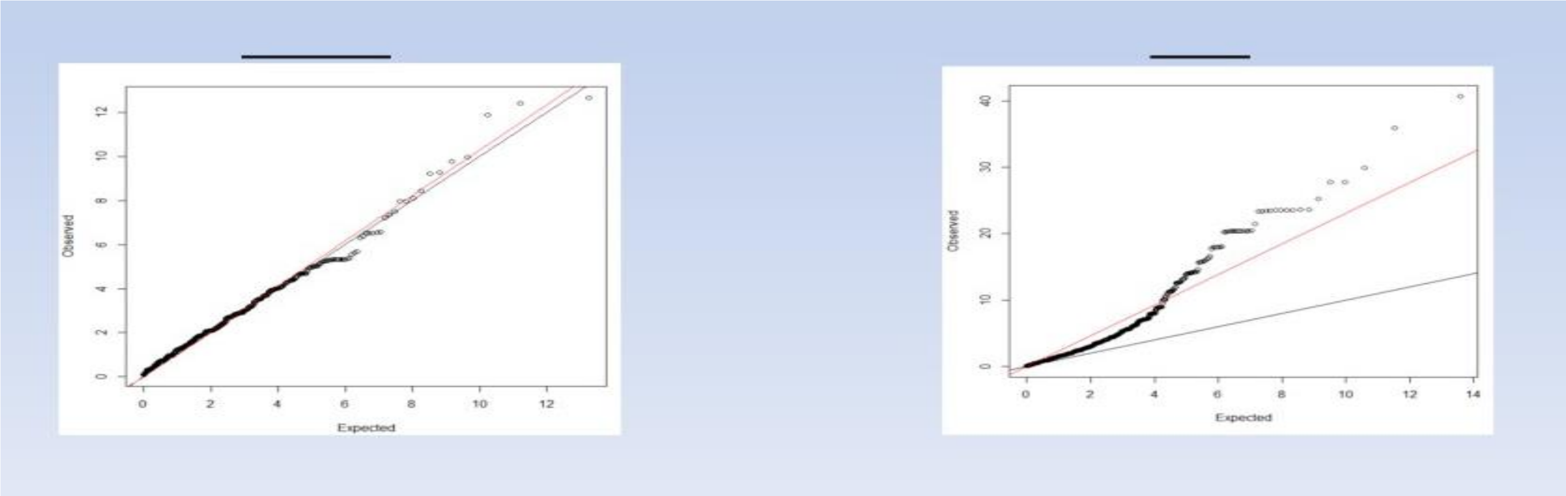
Assessing quality of the raw data (2)

1. Test for Hardy-Weinberg equilibrium on given set of SNPs

   - in cases (i.e. dm2 = 1)

```
Pexact1<-summary(ge03d2ex@gtdata[(ge03d2ex@phdata$dm2 == 1),])[,"Pexact"]
estlambda(Pexact1, plot=TRUE)
$estimate
[1] 2.304846
$se
[1] 0.01319398
```

Controls (raw data) **Cases (raw data)**



Q-Q plots:

The red line shows the fitted slope to the data. The **black line** represents the theoretical slope without any stratification

Save the Plots

- **Save plots on system**

```
> pdf(‘filename’)
> plot(qTest, df="Pc1df")
> dev.off()
```

- **Check the output**

- **Draw QQ plot for (srdta) for cases and controls and check the output**

Loci Association Analysis

- Let’s apply a simple method that uses both *mixed model* and *regression* to find statistically significant associations between trait (presence/absence of diabetes type 2) and loci (SNPs)

```
> qTest = qtscore(dm2,ge03d2ex,trait="binomial")
> descriptives.scan(qTest, sort="Pc1df")
          Chromosome Position Strand A1 A2   N       effB    se_effB  chi2.1df         P1df     effAB     effBB  chi2.2df         P2df        Pc1df
rs1719133          1  4495479      +  T  A 136 0.33729339 0.09282784 13.202591 0.0002795623 0.4004237  0.000000 14.729116 0.0006333052 0.0003504258
rs2975760          3 10518480      +  A  T 134 3.80380024 1.05172986 13.080580 0.0002983731 3.4545455 10.000000 13.547345 0.0011434877 0.0003732694
rs7418878          1  2808520      +  A  T 136 3.08123060 0.93431795 10.875745 0.0009743183 3.6051282  4.871795 12.181064 0.0022642036 0.0011762545
rs5308595          3 10543128      -  C  G 133 3.98254950 1.21582875 10.729452 0.0010544366 3.3171429       Inf 10.766439 0.0045930101 0.0012699705
rs4804634          1  2807417      +  C  G 132 0.43411456 0.13400290 10.494949 0.0011970132 0.5240642  0.173913 11.200767 0.0036964462 0.0014362332
```

effAB / effBB = odds ratio of each possible genotype combination

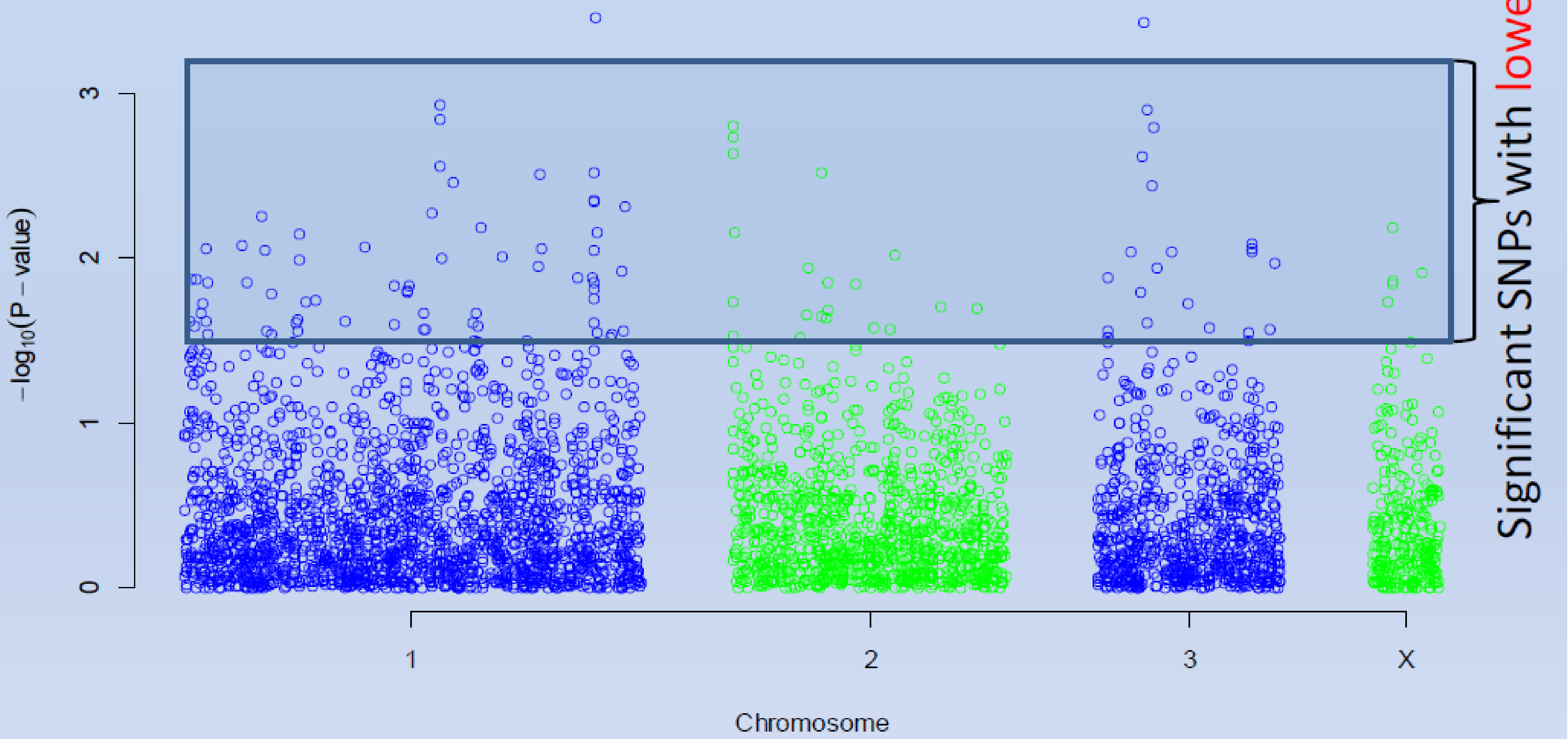
P1df / P2df = probability values from GWA analysis

Pc1df = probability values corrected for inflation factor (stratification effects) at 1 degree of freedom

# plot the lambda corrected values and visualize the Manhattan plot

```
plot(qTest, df="Pc1df")
```

Manhattan plot of raw data



Significant SNPs with lowest P

- To visualize to SNPs associated to the trait use

```
> descriptives.scan(qTest)
```

To see all the results covering all SNPs

```
> results(qTest)
```

- load the (srdta) in R

- Find statistically significant associations between “sex” trait and loci (SNPs) one by one in srdta

- plot the lambda corrected values and visualize the Manhattan plot

Empirical resembling with `qtscore`

- The previous GWA ran only once
- Lets re-run 500 times the same test with random resembling of the data

  - This method is more rigorous

  - Empirical distribution of P-values are obtained

```
> qTest.E <- qtscore(dm2,ge03d2ex,times=500)

> descriptives.scan(qTest.E,sort="Pc1df")
```

```
          Chromosome Position Strand A1 A2   N       effB    se_effB  chi2.1df  P1df Pc1df      effAB      effBB  chi2.2df P2df
rs1719133          1  4495479      +  T  A 136 -0.2652064 0.07298850 13.202591 0.458 0.540 -0.2080882 -0.7375000 14.729116   NA
rs2975760          3 10518480      +  A  T 134  0.2340655 0.06471782 13.080580 0.478 0.558 0.2755102  0.4090909 13.547345   NA
rs7418878          1  2808520      +  A  T 136  0.2089098 0.06334746 10.875745 0.862 0.912 0.2807405  0.3268398 12.181064   NA
rs5308595          3 10543128      -  C  G 133  0.2445516 0.07465893 10.729452 0.890 0.924 0.2564832  0.4623656 10.766439   NA
```

- None of the top SNPs hits the $P<0.05$ significance!

- None of the "P2df" values pass threshold (i.e. all = NA)

Quality Control (QC) of GWA data

- **Since we suspect irregularities in our data we will do**

  - Simple QC assessment without HWE threshold

  - Clean data

- Will run check.marker()QC function

```
> QCresults <- check.marker(ge03d2ex,p.level=0)
> summary(QCresults)
$`Per-SNP fails statistics`
          NoCall NoMAF NoHWE Redundant Xsnpfail
NoCall        42     0     0         0        0
NoMAF         NA   384     0         0        0
NoHWE         NA    NA     0         0        0
Redundant     NA    NA    NA         0        0
Xsnpfail      NA    NA    NA        NA        1

$`Per-person fails statistics`
            IDnoCall HetFail IBSFail isfemale ismale isXXY otherSexErr
IDnoCall           1       0       0        0      0     0           0
HetFail           NA       3       0        0      0     0           0
IBSFail           NA      NA       2        0      0     0           0
isfemale          NA      NA      NA        2      0     0           0
ismale            NA      NA      NA       NA      0     0           0
isXXY             NA      NA      NA       NA     NA     0           0
otherSexErr       NA      NA      NA       NA     NA    NA           0
```

Number of SNPs with call rate lower than of 0.95

# of SNPs with MAF < 5/(2* # of SNPs)

Checking QC results

➢ Check how many subjects PASSED the QC test

```
> names(QCresults)
"nofreq"      "nocall"      "nohwe"       "Xmrkfail"    "hetfail"     "idnocall"    "ibsfail"
"isfemale"    "ismale"      "otherSexErr" "snpok"  "idok"         "call"
> QCresults$idok
"id199"  "id300"  "id403"  "id415"  "id666"  "id689"  "id765"  "id830"  "id908"  "id980"
"id994"  "id1193" "id1423" "id1505" "id1737" "id1827" "id1841" "id2068" "id2094" "id2097"
"id2151" "id2317" "id2618" "id2842" "id2894" "id2985" … "id6934`
> length(QCresults$idok)
[1] 128
```

➢ Check how many SNPs PASSED the QC test

```
> length(QCresults$snpok)
  [1] 3573

#see the SNP ids that passed the QC test
QCresults$idok
```

Select “cleaned” data

- Selection of data from original object **BOTH**:

*by particular individual *by set of SNPs

```
> ge03d2ex["id199", "rs1646456"]
```

```
      id     sex        age dm2    height    weight diet       bmi
      id199    1 59.22872    1 163.9123 80.40746    0 29.92768
@nids = 1
@nsnps = 1
@nbytes = 1
@idnames = id199
@snpnames = rs1646456
@chromosome = 1
@coding =  11
@strand =  01
@map = 653
@male = 1
@gtps = 80
```

- Give vectors of QC passed SNPs and individuals

```
>Cdata <- ge03d2ex[QCresults$idok, QCresults$snpok]
```

Check the quality of overall QC data

```
>descriptives.marker(Cdata)
$`Minor allele frequency distribution`
     X<=0.01 0.01<X<=0.05 0.05<X<=0.1 0.1<X<=0.2    X>0.2
No         0      508.000     677.000     873.000 1515.000
Prop       0        0.142       0.189       0.244    0.424

$`Cumulative distr. of number of SNPs out of HWE, at different alpha`
     X<=1e-04 X<=0.001 X<=0.01 X<=0.05 all X
No     44.000   66.000 117.000 239.000  3573
Prop    0.012    0.018   0.033   0.067     1

$`Distribution of proportion of successful genotypes (per person)`
     X<=0.9 0.9<X<=0.95 0.95<X<=0.98 0.98<X<=0.99 X>0.99
No        0           0            0       65.000 63.000
Prop      0           0            0        0.508  0.492

$`Distribution of proportion of successful genotypes (per SNP)`
     X<=0.9 0.9<X<=0.95 0.95<X<=0.98 0.98<X<=0.99    X>0.99
No        0           0      458.000      814.000 2301.000
Prop      0           0        0.128        0.228    0.644

$`Mean heterozygosity for a SNP`
[1] 0.2787418

$`Standard deviation of the mean heterozygosity for a SNP`
[1] 0.1497257

$`Mean heterozygosity for a person`
[1] 0.26521

$`Standard deviation of mean heterozygosity for a person`
[1] 0.01888496
```

Better but still lots of HWE outliers. Population structure not accounted for?

The **lambda** did not improved significantly also indicating that the QC data still has factors not accounted for

```
estlambda(summary(Cdata)[,"Pexact"])

$estimate
[1] 2.150531
```

The genotype data has much higher call rates since individuals with NA values eliminated in the range of > 99%

Which sub-group causing deviation from HWE?

```
> descriptives.marker(Cdata[Cdata@phdata$dm2==0,])[2]
```

```
$`Cumulative distr. of number of SNPs out of HWE, at
different alpha`
     X<=1e-04 X<=0.001 X<=0.01 X<=0.05 all X
No          0        0   7.000  91.000  3573
Prop        0        0   0.002   0.025     1
```

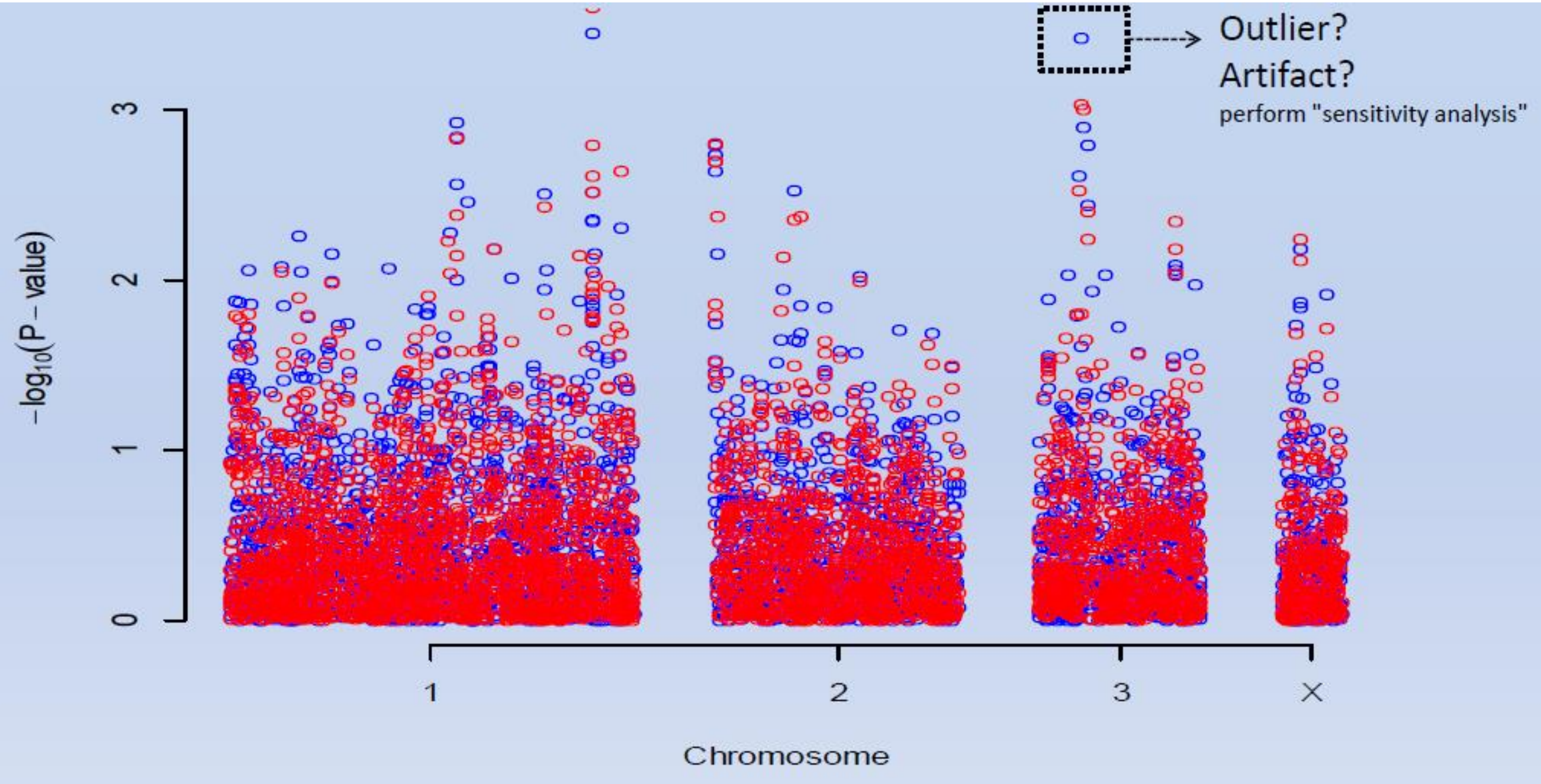
controls

```
> descriptives.marker(Cdata[Cdata@phdata$dm2==1,])[2]
$`Cumulative distr. of number of SNPs out of HWE, at
different alpha`
     X<=1e-04 X<=0.001 X<=0.01 X<=0.05 all X
No     46.000    70.00127.000228.000    3573
Prop    0.013     0.02   0.036   0.064     1
```

cases

As expected, the cases display the greatest deviation from HWE

Compare raw and cleaned data



Outlier?
Artifact?
perform "sensitivity analysis"

```
  > plot(qTest, df="Pc1df", col="blue")
> qTest_QC = qtscore(dm2,Cdata,trait="binomial")
> add.plot(qTest_QC , df="Pc1df", col="red")
```

Note that the cleaned data values are all lower this is due to lower λ value

Finding population structure (1)

- ➢ The data seems to have clear population substructure that we should account for in order to do sensible data analysis
- ➢ Need to detect individuals that are “genetic outliers” compared to the rest using SNP data

- compute matrix of genetic kinship between subjects of this study

```
Cdata.gkin <- ibs(Cdata[,autosomal(Cdata)],weight="freq")
Cdata.gkin[1:5,1:5]
```

```
                id199          id300        id403        id415        id666
id199     0.494427766 3255.00000000 3253.00000000 3241.00000000 3257.0000000
id300 -0.011754266       0.49360296 3261.00000000 3250.00000000 3264.0000000
id403 -0.012253378      -0.01262949     0.50541775 3247.00000000 3262.0000000
id415 -0.001812109       0.01388179    -0.02515438    0.53008236 3251.0000000
id666 -0.018745051      -0.02127344     0.02083723   -0.02014175    0.5306584
```

- The numbers below the diagonal show the genomic estimate of kinship ('genome-wide IBD'),
- The numbers on the diagonal correspond to 0.5 plus the genomic homozigosity
- The numbers above the diagonal tell how many SNPs were typed successfully for both subjects

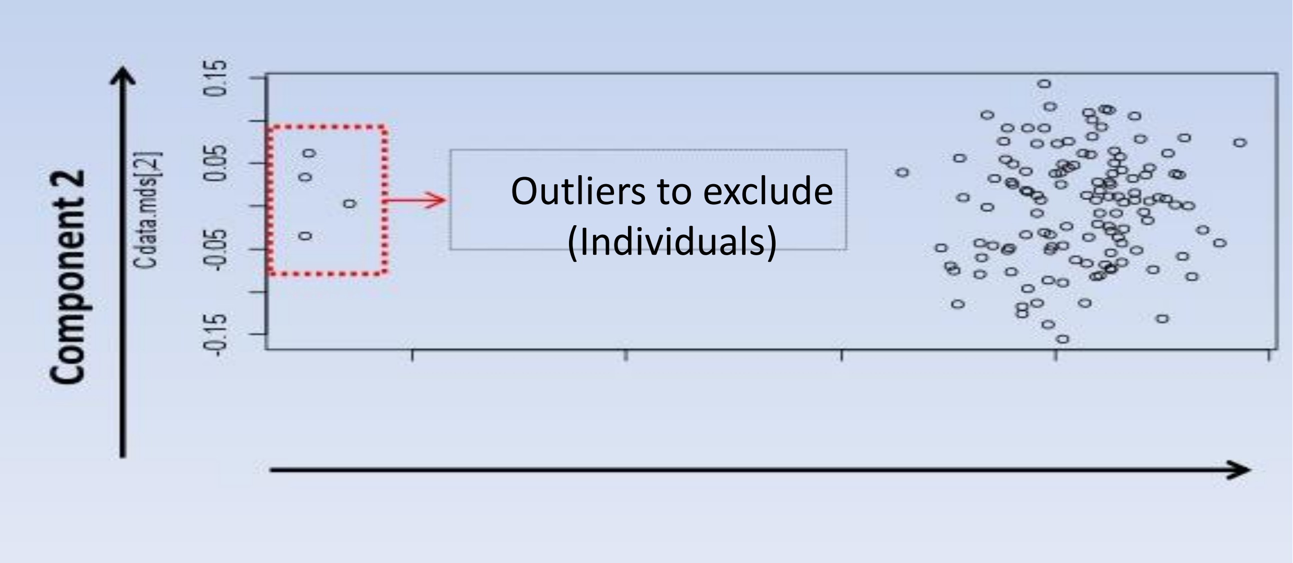
Finding population structure (2)

2. Compute distance matrix from previous

```
Cdata.dist <- as.dist(0.5-Cdata.gkin)
```

3. Do Classical Multidimensional Scaling (PCA) and visualize results

```
Cdata.mds <- cmdscale(Cdata.dist)
plot(Cdata.mds)
```



Component 1

- The PCA fitted the genetic distances along the 2 components
- Points are individuals
- There are clearly two clusters
- Need to select all individuals from biggest cluster

Second round of QC

- **Select the ids of individuals from each cluster**

  - Can use k-means since clusters are well defined

```
> kmeans.res <- kmeans(Cdata.mds,centers=2,nstart=1000)
>cluster1 <- names(which(kmeans.res$cluster==1))
>cluster2 <- names(which(kmeans.res$cluster==2))
```

- **Select another clean dataset using data for individuals in cluster #2 (the largest)**

```
> Cdata2 <- Cdata[cluster1,]
```

- **Perform QC on new data**

```
> QCdata2 <- check.marker(Cdata2, hweids=(phdata(Cdata2)$dm == 0), fdr = 0.2)
```

Second round QC results

- ➢ Visualize the QC results and make conclusions

```
> summary(QCdata2)
$`Per-SNP fails statistics`
          NoCall NoMAF NoHWE Redundant Xsnpfail
NoCall         0     0     0         0        0
NoMAF         NA    40     0         0        0
NoHWE         NA    NA     0         0        0
Redundant     NA    NA    NA         0        0
Xsnpfail      NA    NA    NA        NA        0

$`Per-person fails statistics`
            IDnoCall HetFail IBSFail isfemale ismale isXXY otherSexErr
IDnoCall           0       0       0        0      0     0           0
HetFail           NA       0       0        0      0     0           0
IBSFail           NA      NA       0        0      0     0           0
isfemale          NA      NA      NA        0      0     0           0
ismale            NA      NA      NA       NA      0     0           0
isXXY             NA      NA      NA       NA     NA     0           0
otherSexErr       NA      NA      NA       NA     NA    NA           0
```

- **All markers passed the HWE test**
- **40 markers did not pass the MAF QC test and need to be removed**
- **No phenotypic errors**

- ➢ Clean the dataset again excluding those SNPs

```
Cdata2 <- Cdata2[QCdata2$idok, QCdata2$snpok]
```

Final QC test before GWA

- ➢ Final check on cases and controls QC data

```
> descriptives.marker(Cdata2[phdata(Cdata2)$dm2==1,])[2]
$`Cumulative distr. of number of SNPs out of HWE, at different alpha`
     X<=1e-04 X<=0.001 X<=0.01 X<=0.05 all X
No          0        1  17.000  79.000  3533
Prop        0        0   0.005   0.022     1
```

cases

```
> descriptives.marker(Cdata2[phdata(Cdata2)$dm2==0,])[2]
$`Cumulative distr. of number of SNPs out of HWE, at different alpha`
     X<=1e-04 X<=0.001 X<=0.01 X<=0.05 all X
No          0        0   7.000  91.000  3533
Prop        0        0   0.002   0.026     1
```

controls

- **Finally most of markers are within the HWE at $\alpha < 0.05$**
- **Still controls have marker distribution that better follows HWE**

Perform GWA on new data

- **Perform the mixture model regression analysis**

```
> Cdata2.qt <- qtscore(Cdata2@phdata$dm2, Cdata2,trait="binomial")

 > descriptives.scan(Cdata2.qt,sort="Pc1df")
```

```
Summary for top 10 results, sorted by Pc1df
          Chromosome Position Strand A1 A2   N      effB    se_effB  chi2.1df         P1df     effAB     effBB  chi2.2df         P2df        Pc1df
rs1719133          1  4495479      +  T  A 124 0.3167801 0.08614528 13.522368 0.0002357368 0.3740771 0.0000000 14.677906 0.0006497303 0.0003048399
rs4804634          1  2807417      +  C  G 121 0.4119844 0.12480696 10.896423 0.0009635013 0.6315789 0.1739130 12.375590 0.0020543516 0.0011885463
rs8835506          2  6010852      +  A  T 121 3.5378209 1.08954331 10.543448 0.0011660066 4.0185185 4.0185185 12.605556 0.0018312105 0.0014292471
rs4534929          1  4474374      +  C  G 123 0.4547151 0.14160410 10.311626 0.0013219476 0.4830918 0.1739130 10.510272 0.0052206352 0.0016136479
rs1013473          1  4487262      +  A  T 124 2.7839368 0.86860745 10.272393 0.0013503553 3.0495868 5.8441558 10.926296 0.0042401869 0.0016471605
rs3925525          2  6008501      +  C  G 124 3.2807631 1.03380675 10.070964 0.0015062424 3.6923077 4.0000000 11.765985 0.0027864347 0.0018306610
rs3224311          2  6009769      +  G  C 124 3.2807631 1.03380675 10.070964 0.0015062424 3.6923077 4.0000000 11.765985 0.0027864347 0.0018306610
rs2975760          3 10518480      +  A  T 123 3.1802120 1.00916993  9.930784 0.0016253728 3.0000000 8.0000000 10.172522 0.0061810866 0.0019704699
rs2521089          3 10487652      -  T  C 123 2.7298775 0.87761175  9.675679 0.0018672326 3.0147059 5.0000000 10.543296 0.0051351403 0.0022533033
rs1048031          1  4485591      +  G  T 122 0.4510793 0.14548378  9.613391 0.0019316360 0.4844720 0.1714286 9.965696 0.0068545128 0.0023284084
```

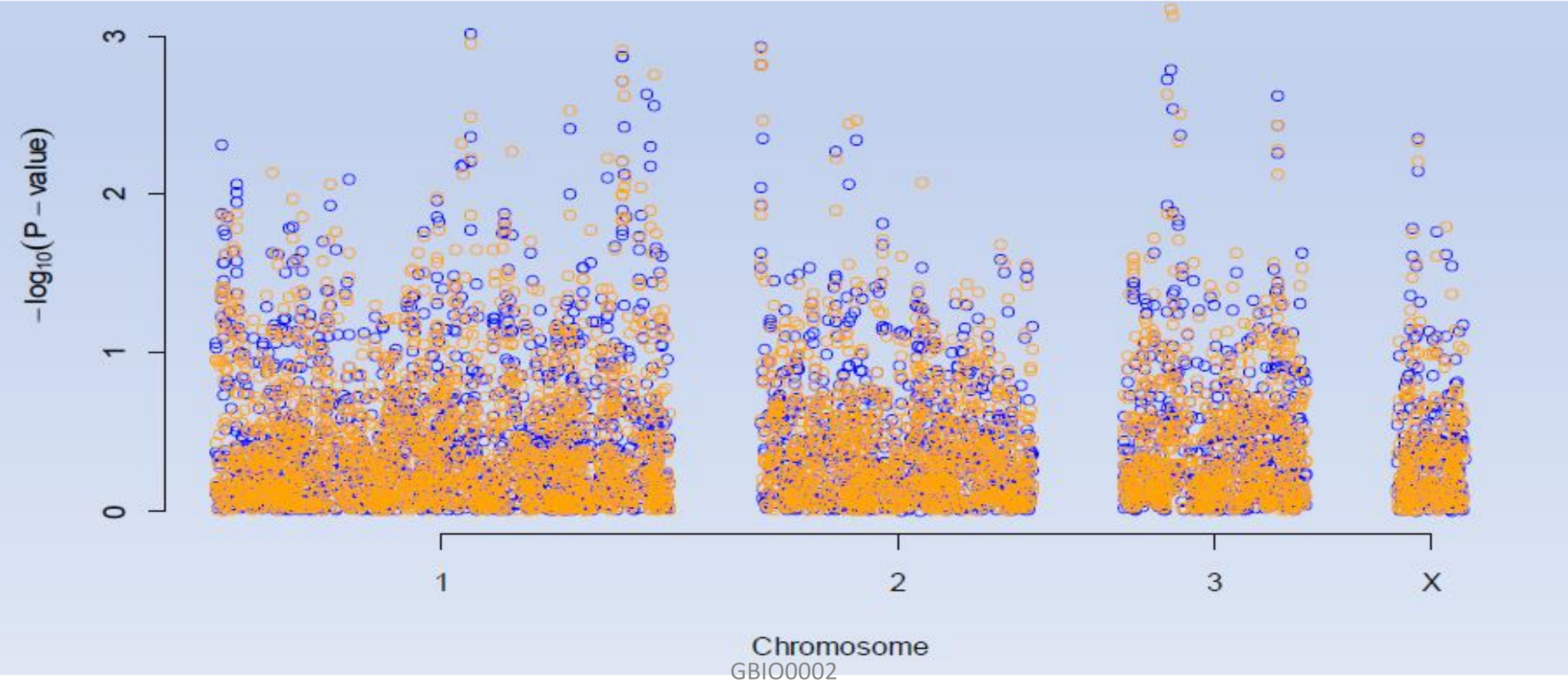
- **Compare results to the previous QC round 1 results**

Manhattan plots

- Compare results QC round 1 vs QC round 2

```
> plot(Cdata2.qt)
> add.plot(qTest_QC, col="orange")
```

Round 1 QC
Round 2 QC
(accounted for population stratification effects)



Perform more rigorous GWA test computing GW (empirical) significance

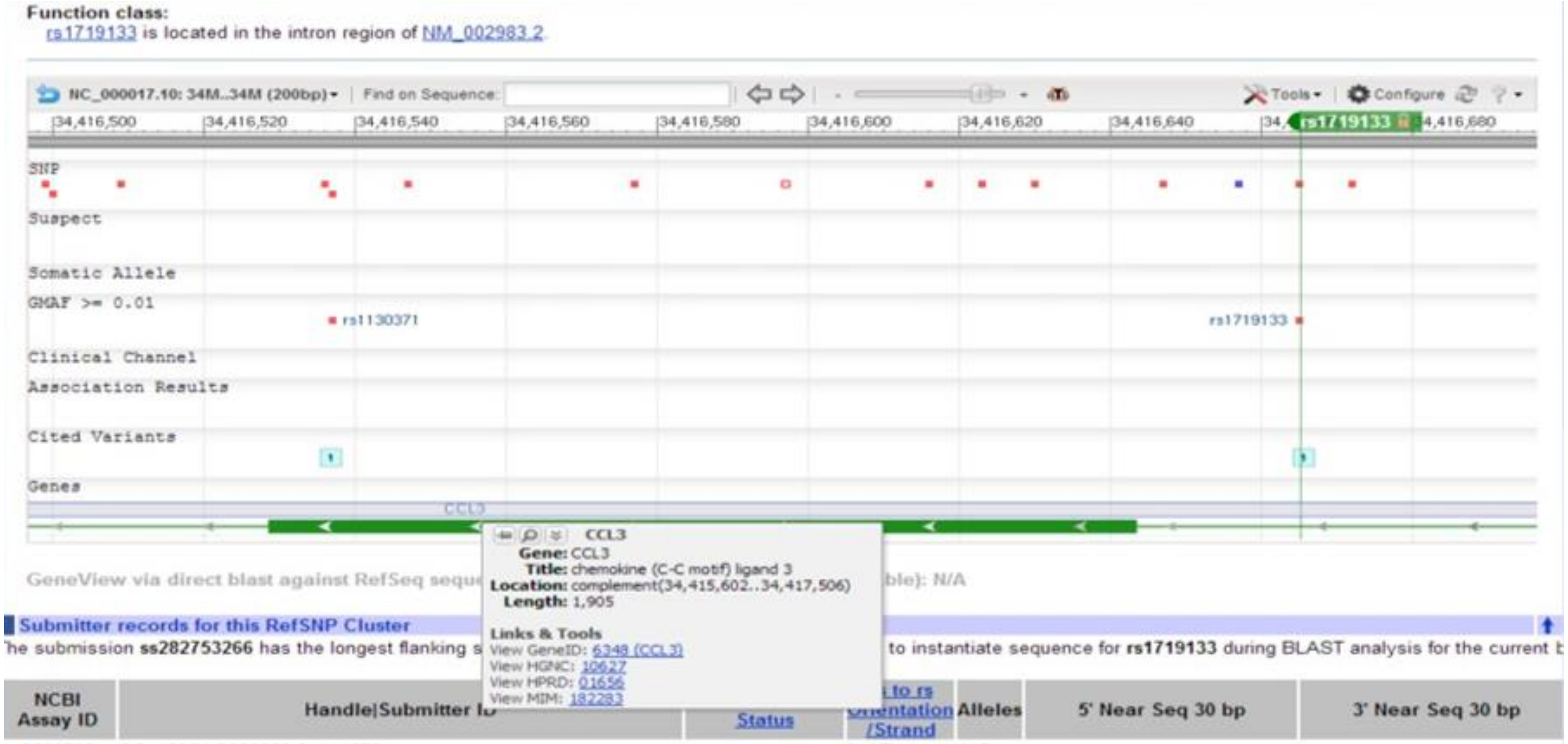
- Will perform the GWA analysis 500 times obtaining GWA statistics

```
> Cdata2.qte <- qtscore(Cdata2@phdata$dm2, times=500, Cdata2,trait="binomial"
> descriptives.scan(Cdata2.qte,sort="Pc1df")
Summary for top 10 results, sorted by Pc1df
          Chromosome Position Strand A1 A2   N      effB    se_effB  chi2.1df  P1df Pc1df      effAB      effBB  chi2.2df  P2df
rs1719133          1  4495479      +  T  A 124 0.3167801 0.08614528 13.522368 0.346 0.402 0.3740771 0.0000000 14.677906 0.566
rs4804634          1  2807417      +  C  G 121 0.4119844 0.12480696 10.896423 0.844 0.894 0.6315789 0.1739130 12.375590 0.944
rs8835506          2  6010852      +  A  T 121 3.5378209 1.08954331 10.543448 0.886 0.938 4.0185185 4.0185185 12.605556 0.920
rs4534929          1  4474374      +  C  G 123 0.4547151 0.14160410 10.311626 0.922 0.946 0.4830918 0.1739130 10.510272 1.000
rs1013473          1  4487262      +  A  T 124 2.7839368 0.86860745 10.272393 0.924 0.952 3.0495868 5.8441558 10.926296 1.000
rs3925525          2  6008501      +  C  G 124 3.2807631 1.03380675 10.070964 0.942 0.960 3.6923077 4.0000000 11.765985 0.986
rs3224311          2  6009769      +  G  C 124 3.2807631 1.03380675 10.070964 0.942 0.960 3.6923077 4.0000000 11.765985 0.986
rs2975760          3 10518480      +  A  T 123 3.1802120 1.00916993  9.930784 0.948 0.966 3.0000000 8.0000000 10.172522 1.000
rs2521089          3 10487652      -  T  C 123 2.7298775 0.87761175  9.675679 0.964 0.978 3.0147059 5.0000000 10.543296 1.000
rs1048031          1  4485591      +  G  T 122 0.4510793 0.14548378  9.613391 0.966 0.982 0.4844720 0.1714286  9.965696 1.000
```

- Results had improved for the P2df statistic, but none of them fall under GW significance level of 0.05
  - `rs1719133` does not pass the significance tests but is the one with the best level of association compared to other SNPs

Biological interpretations

- For illustration purposes let’s extract information on the rs1719133 from dbSNP
- Seems to target CCL3 - pro-inflamatory cytokine. The CCL2 was implicated in T1D [1]



Function class:
rs1719133 is located in the intron region of NM_002983.2.

CCL3
Gene: CCL3
Title: chemokine (C-C motif) ligand 3
Location: complement(34,415,602..34,417,506)
Length: 1,905

Links & Tools
View GeneID: 6348 (CCL3)
View HGNC: 10627
View HPRD: 01656
View MIM: 182283

GeneView via direct blast against RefSeq seque ble): N/A

Submitter records for this RefSNP Cluster

he submission **ss282753266** has the longest flanking s to instantiate sequence for **rs1719133** during BLAST analysis for the current b

NCBI Assay ID	Handle|Submitter ID	Status	to rs Orientation /Strand	Alleles	5' Near Seq 30 bp	3' Near Seq 30 bp

Conclusions

- GWA studies are popular these days mainly due to high throughput technology development such as genotyping chips (i.e. SNP arrays) and sequences
- Analysis of GW data requires several steps of quality control in order to draw conclusions
- GenABEL provides tools to perform GWAs and automate some of the steps

References

[1] Ruili Guan et al. Chemokine (C-C Motif) Ligand 2 (CCL2) in Sera of Patients with Type 1 Diabetes and Diabetic Complications. PLoS ONE 6(4): e17822

[2] Yurii Aulchenko, GenABEL tutorial http://www.genabel.org/sites/default/files/pdfs/GenABEL-tutorial.pdf

[3] GenABEL project developers, GenABEL: genome-wide SNP association analysis 2012, R package version 1.7-2

[4] Geraldine M Clarke et.al. Basic statistical analysis in genetic case-control studies. Nat Protoc. 2011 February ; 6(2): 121–133

[5] Lobo, I. Same genetic mutation, different genetic disease phenotype. Nature Education 2008, 1(1)