PostGWAS Intermediate Bioinformatics Online Course

Fine-mapping and visualisation course

Scott Hazelhurst and Jean-Tristan Brandenburg

Sydney Brenner Institute for Molecular Bioscience (Wits University) and H3ABioNet

H3ABioNet
Pan African Bioinformatics Network for H3Africa

SDIMB
SYDNEY BRENNER INSTITUTE
FOR MOLECULAR BIOSCIENCE

Course organisation

3 sessions :

1. Visualisation
2. Fine-Mapping
3. Consolidation of 2 previous session

Part. 1 : Visualisation

Visualisation of GWAS: Learning Objectives

- Overview of figures used in GWAS in different level.
- Understand when to used, how to read and analyze figures.
- Detected bias in your GWAS result using figures.

Visualisation of GWAS : Learning Outcomes

- Used various platform or software to plot figures :
  - Manhantan plot
  - QQ plot
  - regional plot
  - forest plot.
  - Phenotype distribution by genotype
- understand how to interpret and used figures.

Introduction

Visualisation of GWAS

- Different scales :

Visualisation of GWAS

- Different scales :
  - Global visualisation : your summary statistics using QQ-plot or Manhattan.

Visualisation of GWAS

- Different scales :
  - Global visualisation : your summary statistics using QQ-plot or Manhattan.
  - Regional plot : zoom of specific regions, annotation can be add, defined putative lead SNPs.

Visualisation of GWAS

- Different scales :
  - Global visualisation : your summary statistics using QQ-plot or Manhattan.
  - Regional plot : zoom of specific regions, annotation can be add, defined putative lead SNPs.
  - Positions : analyse Phenotype / genotypes, cluster of genotype from array calling, forest plot from meta analysis.

Visualisation of GWAS

- Different scales :
  - Global visualisation : your summary statistics using QQ-plot or Manhattan.
  - Regional plot : zoom of specific regions, annotation can be add, defined putative lead SNPs.
  - Positions : analyse Phenotype / genotypes, cluster of genotype from array calling, forest plot from meta analysis.
- Why :

Visualisation of GWAS

- Different scales :
  - Global visualisation : your summary statistics using QQ-plot or Manhattan.
  - Regional plot : zoom of specific regions, annotation can be add, defined putative lead SNPs.
  - Positions : analyse Phenotype / genotypes, cluster of genotype from array calling, forest plot from meta analysis.
- Why :
  - Identify and understand bias at different scales.

Visualisation of GWAS

- Different scales :
  - Global visualisation : your summary statistics using QQ-plot or Manhattan.
  - Regional plot : zoom of specific regions, annotation can be add, defined putative lead SNPs.
  - Positions : analyse Phenotype / genotypes, cluster of genotype from array calling, forest plot from meta analysis.
- Why :
  - Identify and understand bias at different scales.
  - Identify positions of interest with possible annotation.

Visualisation of GWAS

- Different scales :
  - Global visualisation : your summary statistics using QQ-plot or Manhattan.
  - Regional plot : zoom of specific regions, annotation can be add, defined putative lead SNPs.
  - Positions : analyse Phenotype / genotypes, cluster of genotype from array calling, forest plot from meta analysis.
- Why :
  - Identify and understand bias at different scales.
  - Identify positions of interest with possible annotation.
  - Figures to publish.

Visualisation of GWAS

- Different scales :
  - Global visualisation : your summary statistics using QQ-plot or Manhattan.
  - Regional plot : zoom of specific regions, annotation can be add, defined putative lead SNPs.
  - Positions : analyse Phenotype / genotypes, cluster of genotype from array calling, forest plot from meta analysis.
- Why :
  - Identify and understand bias at different scales.
  - Identify positions of interest with possible annotation.
  - Figures to publish.
- Data used :

Visualisation of GWAS

- Different scales :
  - Global visualisation : your summary statistics using QQ-plot or Manhattan.
  - Regional plot : zoom of specific regions, annotation can be add, defined putative lead SNPs.
  - Positions : analyse Phenotype / genotypes, cluster of genotype from array calling, forest plot from meta analysis.
- Why :
  - Identify and understand bias at different scales.
  - Identify positions of interest with possible annotation.
  - Figures to publish.
- Data used :
  - Result of summary statistics

Visualisation of GWAS

- Different scales :
  - Global visualisation : your summary statistics using QQ-plot or Manhattan.
  - Regional plot : zoom of specific regions, annotation can be add, defined putative lead SNPs.
  - Positions : analyse Phenotype / genotypes, cluster of genotype from array calling, forest plot from meta analysis.
- Why :
  - Identify and understand bias at different scales.
  - Identify positions of interest with possible annotation.
  - Figures to publish.
- Data used :
  - Result of summary statistics
  - Annotation using external data set.

Visualisation of GWAS

- Different scales :
  - Global visualisation : your summary statistics using QQ-plot or Manhattan.
  - Regional plot : zoom of specific regions, annotation can be add, defined putative lead SNPs.
  - Positions : analyse Phenotype / genotypes, cluster of genotype from array calling, forest plot from meta analysis.
- Why :
  - Identify and understand bias at different scales.
  - Identify positions of interest with possible annotation.
  - Figures to publish.
- Data used :
  - Result of summary statistics
  - Annotation using external data set.
  - Linkage disequilibrium using your genotype data or external data set

Visualisation of GWAS

- Different scales :
  - Global visualisation : your summary statistics using QQ-plot or Manhattan.
  - Regional plot : zoom of specific regions, annotation can be add, defined putative lead SNPs.
  - Positions : analyse Phenotype / genotypes, cluster of genotype from array calling, forest plot from meta analysis.
- Why :
  - Identify and understand bias at different scales.
  - Identify positions of interest with possible annotation.
  - Figures to publish.
- Data used :
  - Result of summary statistics
  - Annotation using external data set.
  - Linkage disequilibrium using your genotype data or external data set
  - Genotype data : LD computation, individual information

Visualisation of GWAS

- Different scales :
  - Global visualisation : your summary statistics using QQ-plot or Manhattan.
  - Regional plot : zoom of specific regions, annotation can be add, defined putative lead SNPs.
  - Positions : analyse Phenotype / genotypes, cluster of genotype from array calling, forest plot from meta analysis.
- Why :
  - Identify and understand bias at different scales.
  - Identify positions of interest with possible annotation.
  - Figures to publish.
- Data used :
  - Result of summary statistics
  - Annotation using external data set.
  - Linkage disequilibrium using your genotype data or external data set
  - Genotype data : LD computation, individual information
  - Raw data of sequencing from array.

Approaches : Softwares, platforms

- R, python using "base" function

Approaches : Softwares, platforms

- R, python using "base" function
  - highly configurable : can build your own plot, used your own genetics data, own database

Approaches : Softwares, platforms

- R, python using "base" function
  - highly configurable : can build your own plot, used your own genetics data, own database
  - running on server, computer (need some time high resource)

Approaches : Softwares, platforms

- R, python using "base" function
  - highly configurable : can build your own plot, used your own genetics data, own database
  - running on server, computer (need some time high resource)
  - need a good level bio informatics and time consuming

Approaches : Softwares, platforms

- R, python using "base" function
  - highly configurable : can build your own plot, used your own genetics data, own database
  - running on server, computer (need some time high resource)
  - need a good level bio informatics and time consuming
  - statics.

Approaches : Softwares, platforms

- R, python using "base" function
  - highly configurable : can build your own plot, used your own genetics data, own database
  - running on server, computer (need some time high resource)
  - need a good level bio informatics and time consuming
  - statics.
  - ex : ggplot2, plot.

Approaches : Softwares, platforms

- R, python using "base" function
  - highly configurable : can build your own plot, used your own genetics data, own database
  - running on server, computer (need some time high resource)
  - need a good level bio informatics and time consuming
  - statics.
  - ex : ggplot2, plot.
- R-library, python library for :

Approaches : Softwares, platforms

- R, python using "base" function
  - highly configurable : can build your own plot, used your own genetics data, own database
  - running on server, computer (need some time high resource)
  - need a good level bio informatics and time consuming
  - statics.
  - ex : ggplot2, plot.
- R-library, python library for :
  - Less configurable : limited to options defined by function but own genotype.

Approaches : Softwares, platforms

- R, python using "base" function
  - highly configurable : can build your own plot, used your own genetics data, own database
  - running on server, computer (need some time high resource)
  - need a good level bio informatics and time consuming
  - statics.
  - ex : ggplot2, plot.
- R-library, python library for :
  - Less configurable : limited to options defined by function but own genotype.
  - Run on server, computer (need some time high resource)

Approaches : Softwares, platforms

- R, python using "base" function
  - highly configurable : can build your own plot, used your own genetics data, own database
  - running on server, computer (need some time high resource)
  - need a good level bio informatics and time consuming
  - statics.
  - ex : ggplot2, plot.
- R-library, python library for :
  - Less configurable : limited to options defined by function but own genotype.
  - Run on server, computer (need some time high resource)
  - need a middle level of bio informatics.

Approaches : Softwares, platforms

- R, python using "base" function
  - highly configurable : can build your own plot, used your own genetics data, own database
  - running on server, computer (need some time high resource)
  - need a good level bio informatics and time consuming
  - statics.
  - ex : ggplot2, plot.
- R-library, python library for :
  - Less configurable : limited to options defined by function but own genotype.
  - Run on server, computer (need some time high resource)
  - need a middle level of bio informatics.
  - ex : R with fastman, qqman, forestplot, Manhattan Plot

Approaches : Softwares, platforms

- R, python using "base" function
  - highly configurable : can build your own plot, used your own genetics data, own database
  - running on server, computer (need some time high resource)
  - need a good level bio informatics and time consuming
  - statics.
  - ex : ggplot2, plot.
- R-library, python library for :
  - Less configurable : limited to options defined by function but own genotype.
  - Run on server, computer (need some time high resource)
  - need a middle level of bio informatics.
  - ex : R with fastman, qqman, forestplot, Manhattan Plot
- Web interface:

Approaches : Softwares, platforms

- R, python using "base" function
  - highly configurable : can build your own plot, used your own genetics data, own database
  - running on server, computer (need some time high resource)
  - need a good level bio informatics and time consuming
  - statics.
  - ex : ggplot2, plot.
- R-library, python library for :
  - Less configurable : limited to options defined by function but own genotype.
  - Run on server, computer (need some time high resource)
  - need a middle level of bio informatics.
  - ex : R with fastman, qqman, forestplot, Manhattan Plot
- Web interface:
  - Less configurable : limited to options defined on interface, ex: cannot used your own genotype, linkage disequilibrium.

Approaches : Softwares, platforms

- R, python using "base" function
  - highly configurable : can build your own plot, used your own genetics data, own database
  - running on server, computer (need some time high resource)
  - need a good level bio informatics and time consuming
  - statics.
  - ex : ggplot2, plot.
- R-library, python library for :
  - Less configurable : limited to options defined by function but own genotype.
  - Run on server, computer (need some time high resource)
  - need a middle level of bio informatics.
  - ex : R with fastman, qqman, forestplot, Manhattan Plot
- Web interface:
  - Less configurable : limited to options defined on interface, ex: cannot used your own genotype, linkage disequilibrium.
  - Run on external server, computer doesn't need a lot of resource, transfer of big file.

Approaches : Softwares, platforms

- R, python using "base" function
  - highly configurable : can build your own plot, used your own genetics data, own database
  - running on server, computer (need some time high resource)
  - need a good level bio informatics and time consuming
  - statics.
  - ex : ggplot2, plot.
- R-library, python library for :
  - Less configurable : limited to options defined by function but own genotype.
  - Run on server, computer (need some time high resource)
  - need a middle level of bio informatics.
  - ex : R with fastman, qqman, forestplot, Manhattan Plot
- Web interface:
  - Less configurable : limited to options defined on interface, ex: cannot used your own genotype, linkage disequilibrium.
  - Run on external server, computer doesn't need a lot of resource, transfer of big file.
  - doesn't need a specific bioinformatics level.

Approaches : Softwares, platforms

- R, python using "base" function
  - highly configurable : can build your own plot, used your own genetics data, own database
  - running on server, computer (need some time high resource)
  - need a good level bio informatics and time consuming
  - statics.
  - ex : ggplot2, plot.
- R-library, python library for :
  - Less configurable : limited to options defined by function but own genotype.
  - Run on server, computer (need some time high resource)
  - need a middle level of bio informatics.
  - ex : R with fastman, qqman, forestplot, Manhattan Plot
- Web interface:
  - Less configurable : limited to options defined on interface, ex: cannot used your own genotype, linkage disequilibrium.
  - Run on external server, computer doesn't need a lot of resource, transfer of big file.
  - doesn't need a specific bioinformatics level.
  - example : FUMA, locuszoomV2

Little reminder

Summary statistics

Summary statistics :

- first line :header descriptive of each column.

```
chr rs n_miss allele1 af beta se p_wald
1 1:920733:T:C 0 T 0.098 -1.454952e+00 1.977754e+00 4.622846e-01
```

Summary statistics

Summary statistics :

- first line :header descriptive of each column.
- Each line represent result at one position of association.

```
chr rs n_miss allele1 af beta se p_wald
1 1:920733:T:C 0 T 0.098 -1.454952e+00 1.977754e+00 4.622846e-01
```

Summary statistics

Summary statistics :

- first line :header descriptive of each column.
- Each line represent result at one position of association.
- some column found on files :

```
chr rs n_miss allele1 af beta se p_wald
1 1:920733:T:C 0 T 0.098 -1.454952e+00 1.977754e+00 4.622846e-01
```

Summary statistics

Summary statistics :

- first line :header descriptive of each column.
- Each line represent result at one position of association.
- some column found on files :
  - Chromosome : CHR, chro, chrom

```
chr rs n_miss allele1 af beta se p_wald
1 1:920733:T:C 0 T 0.098 -1.454952e+00 1.977754e+00 4.622846e-01
```

Summary statistics

Summary statistics :

- first line :header descriptive of each column.
- Each line represent result at one position of association.
- some column found on files :
  - Chromosome : CHR, chro, chrom
  - Positions : bp BP POS

```
chr rs n_miss allele1 af beta se p_wald
1 1:920733:T:C 0 T 0.098 -1.454952e+00 1.977754e+00 4.622846e-01
```

Summary statistics

Summary statistics :

- first line :header descriptive of each column.
- Each line represent result at one position of association.
- some column found on files :
  - Chromosome : CHR, chro, chrom
  - Positions : bp BP POS
  - Unique id of locus : rsid rs, SNP

```
chr rs n_miss allele1 af beta se p_wald
1 1:920733:T:C 0 T 0.098 -1.454952e+00 1.977754e+00 4.622846e-01
```

Summary statistics

Summary statistics :

- first line :header descriptive of each column.
- Each line represent result at one position of association.
- some column found on files :
  - Chromosome : CHR, chro, chrom
  - Positions : bp BP POS
  - Unique id of locus : rsid rs, SNP
  - Effect allele : allele 1, a1..

```
chr rs n_miss allele1 af beta se p_wald
1 1:920733:T:C 0 T 0.098 -1.454952e+00 1.977754e+00 4.622846e-01
```

Summary statistics

Summary statistics :

- first line :header descriptive of each column.
- Each line represent result at one position of association.
- some column found on files :
  - Chromosome : CHR, chro, chrom
  - Positions : bp BP POS
  - Unique id of locus : rsid rs, SNP
  - Effect allele : allele 1, a1..
  - Non effect allele : allele 0, allele 2, a0..

```
chr rs n_miss allele1 af beta se p_wald
1 1:920733:T:C 0 T 0.098 -1.454952e+00 1.977754e+00 4.622846e-01
```

Summary statistics

Summary statistics :

- first line :header descriptive of each column.
- Each line represent result at one position of association.
- some column found on files :
  - Chromosome : CHR, chro, chrom
  - Positions : bp BP POS
  - Unique id of locus : rsid rs, SNP
  - Effect allele : allele 1, a1..
  - Non effect allele : allele 0, allele 2, a0..
  - Effect of genotype on phenotype, usually if $\beta > 0$, mean that allele1 increase phenotype compared to individual carry allele0 : beta, B

```
chr rs n_miss allele1 af beta se p_wald
1 1:920733:T:C 0 T 0.098 -1.454952e+00 1.977754e+00 4.622846e-01
```

Summary statistics

Summary statistics :

- first line :header descriptive of each column.
- Each line represent result at one position of association.
- some column found on files :
  - Chromosome : CHR, chro, chrom
  - Positions : bp BP POS
  - Unique id of locus : rsid rs, SNP
  - Effect allele : allele 1, a1..
  - Non effect allele : allele 0, allele 2, a0..
  - Effect of genotype on phenotype, usually if $\beta > 0$, mean that allele1 increase phenotype compared to individual carry allele0 : beta, B
  - Standard error of effect : se, SE, standard_error...

```
chr rs n_miss allele1 af beta se p_wald
1 1:920733:T:C 0 T 0.098 -1.454952e+00 1.977754e+00 4.622846e-01
```

Summary statistics

Summary statistics :

- first line :header descriptive of each column.
- Each line represent result at one position of association.
- some column found on files :
  - Chromosome : CHR, chro, chrom
  - Positions : bp BP POS
  - Unique id of locus : rsid rs, SNP
  - Effect allele : allele 1, a1..
  - Non effect allele : allele 0, allele 2, a0..
  - Effect of genotype on phenotype, usually if $\beta > 0$, mean that allele1 increase phenotype compared to individual carry allele0 : beta, B
  - Standard error of effect : se, SE, standard_error...
  - P value : the probability of obtaining results at least as extreme as the observed result : P, p_wald

```
chr rs n_miss allele1 af beta se p_wald
1 1:920733:T:C 0 T 0.098 -1.454952e+00 1.977754e+00 4.622846e-01
```

Summary statistics

Summary statistics :

- first line :header descriptive of each column.
- Each line represent result at one position of association.
- some column found on files :
  - Chromosome : CHR, chro, chrom
  - Positions : bp BP POS
  - Unique id of locus : rsid rs, SNP
  - Effect allele : allele 1, a1..
  - Non effect allele : allele 0, allele 2, a0..
  - Effect of genotype on phenotype, usually if $\beta > 0$, mean that allele1 increase phenotype compared to individual carry allele0 : beta, B
  - Standard error of effect : se, SE, standard_error...
  - P value : the probability of obtaining results at least as extreme as the observed result : P, p_wald
  - Individual number - not in all summary statistics - : N, n_total_sum.

```
chr rs n_miss allele1 af beta se p_wald
1 1:920733:T:C 0 T 0.098 -1.454952e+00 1.977754e+00 4.622846e-01
```

Summary statistics

Summary statistics :

- first line :header descriptive of each column.
- Each line represent result at one position of association.
- some column found on files :
  - Chromosome : CHR, chro, chrom
  - Positions : bp BP POS
  - Unique id of locus : rsid rs, SNP
  - Effect allele : allele 1, a1..
  - Non effect allele : allele 0, allele 2, a0..
  - Effect of genotype on phenotype, usually if $\beta > 0$, mean that allele1 increase phenotype compared to individual carry allele0 : beta, B
  - Standard error of effect : se, SE, standard_error...
  - P value : the probability of obtaining results at least as extreme as the observed result : P, p_wald
  - Individual number - not in all summary statistics - : N, n_total_sum.
  - Others : INFO - imputation score, N case / control...

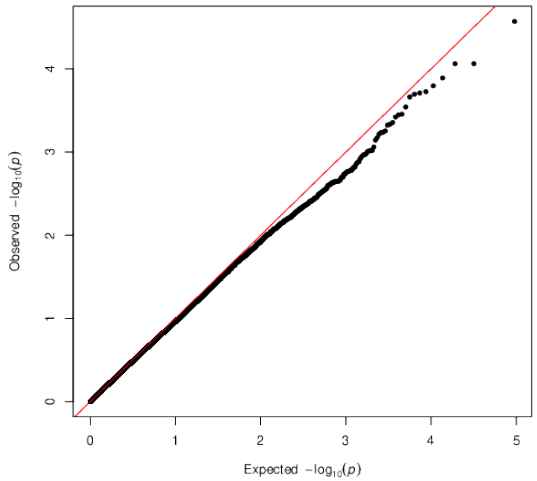
```
chr rs n_miss allele1 af beta se p_wald
1 1:920733:T:C 0 T 0.098 -1.454952e+00 1.977754e+00 4.622846e-01
```

QQ-Plot

Global visualisation : QQ-plot

Quantile Quantile plot

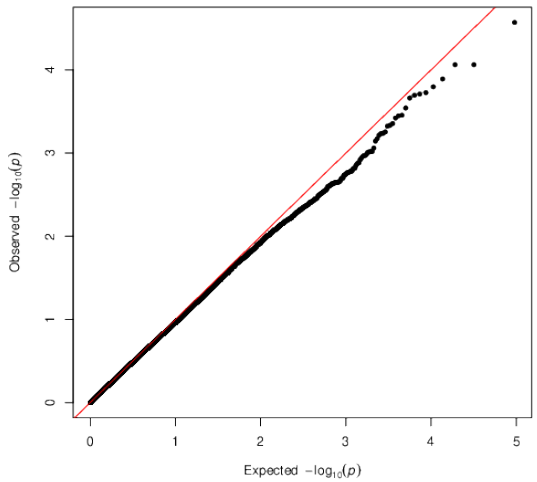
- The QQ plot is a graphical representation of the deviation of the observed P values from the null hypothesis.



Global visualisation : QQ-plot

Quantile Quantile plot

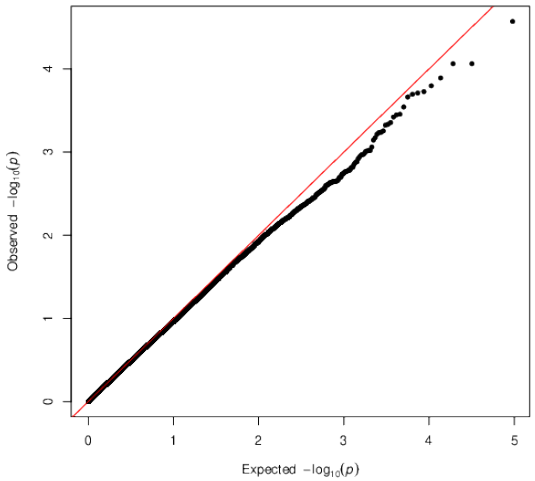
- The QQ plot is a graphical representation of the deviation of the observed P values from the null hypothesis.
- Purpose : Determine if there are a likely a large number of false positive



Global visualisation : QQ-plot

Quantile Quantile plot

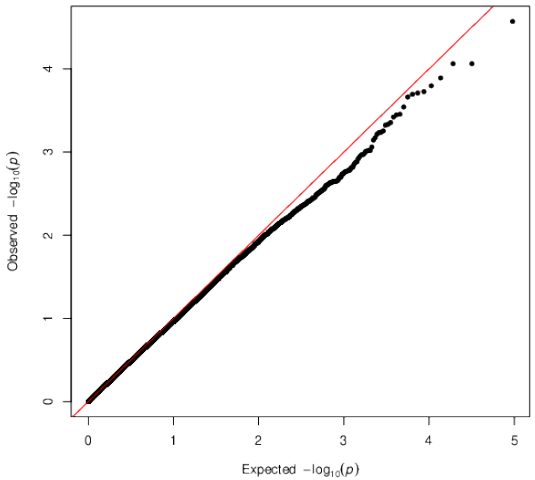
- The QQ plot is a graphical representation of the deviation of the observed P values from the null hypothesis.
- Purpose : Determine if there are a likely a large number of false positive
- Logic : Main of SNPs are not associated with your phenotype only a very, very small number of SNPs should be associated with our trait.



Global visualisation : QQ-plot

Quantile Quantile plot

- The QQ plot is a graphical representation of the deviation of the observed P values from the null hypothesis.
- Purpose : Determine if there are a likely a large number of false positive
- Logic : Main of SNPs are not associated with your phenotype only a very, very small number of SNPs should be associated with our trait.
- Figure : P-values for each SNP are sorted from largest to smallest and plotted against expected values.



Global visualisation : QQ-plot

- P-value distribution should follow a random law, for 100 tests, P-value

source : https://jnmaloof.github.io/BIS180L_web/slides/11_QQPlots.html#1

Global visualisation : QQ-plot

- P-value distribution should follow a random law, for 100 tests, P-value
  - We expect 1 test (1%) to have a p-value of $<= 0.01$

source : `https://jnmaloof.github.io/BIS180L_web/slides/11_QQPlots.html#1`

Global visualisation : QQ-plot

- P-value distribution should follow a random law, for 100 tests, P-value
  - We expect 1 test (1%) to have a p-value of $<= 0.01$
  - We expect 5 test (5%) to have a p-value of $<= 0.05$

source : `https://jnmaloof.github.io/BIS180L_web/slides/11_QQPlots.html#1`

Global visualisation : QQ-plot

- P-value distribution should follow a random law, for 100 tests, P-value
  - We expect 1 test (1%) to have a p-value of $<= 0.01$
  - We expect 5 test (5%) to have a p-value of $<= 0.05$
  - We expect 50 test (50%) to have a p-value of $<= 0.50$

source : `https://jnmaloof.github.io/BIS180L_web/slides/11_QQPlots.html#1`

Global visualisation : QQ-plot

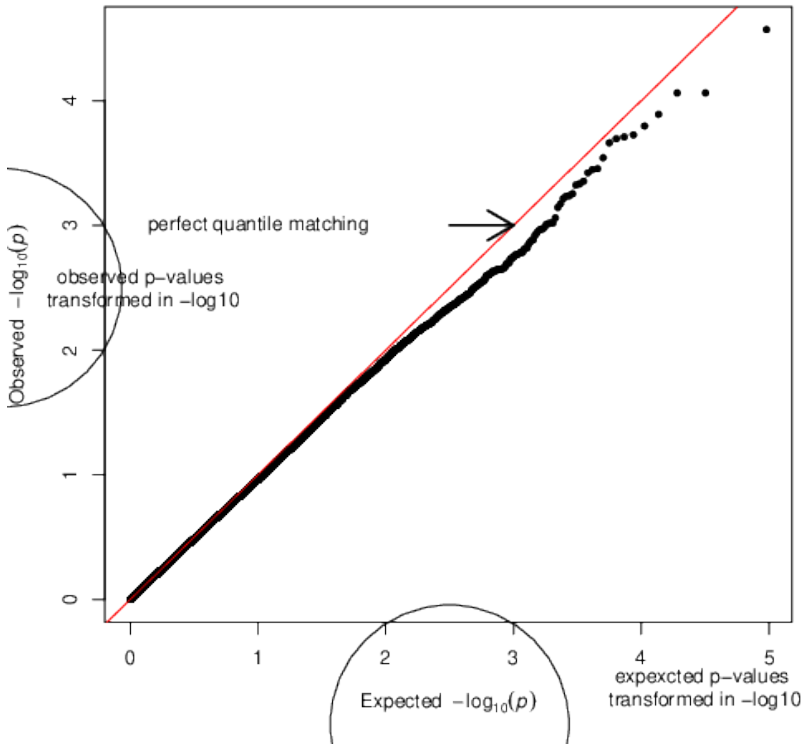
- P-value distribution should follow a random law, for 100 tests, P-value
  - We expect 1 test (1%) to have a p-value of $<= 0.01$
  - We expect 5 test (5%) to have a p-value of $<= 0.05$
  - We expect 50 test (50%) to have a p-value of $<= 0.50$
- We can computed expected p-value using quantile distribution for each SNPs using quantile.

source : `https://jnmaloof.github.io/BIS180L_web/slides/11_QQPlots.html#1`

Global visualisation : QQ-plot

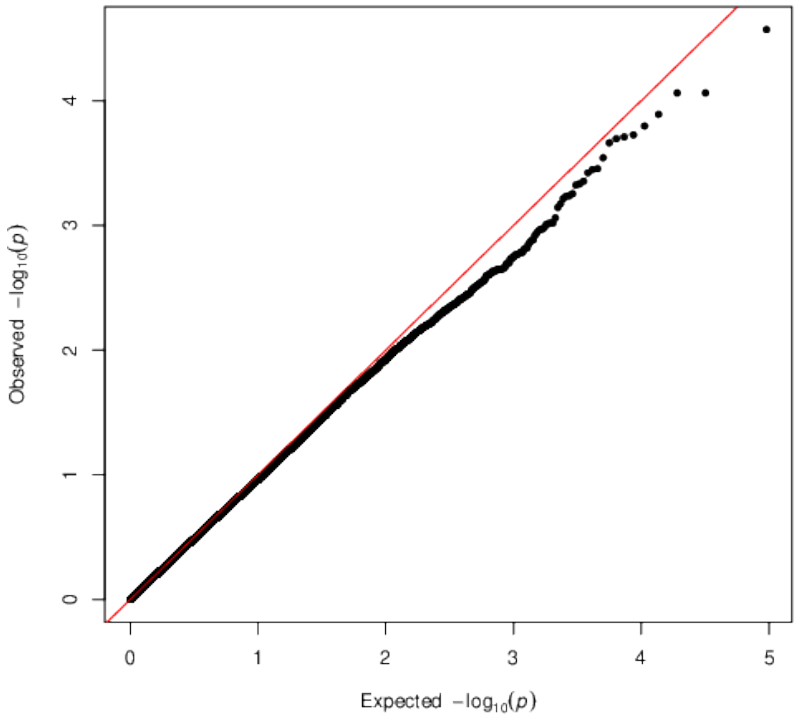
Quantile Quantile plot

- defined in x axis by expected p-value transformed using -log10
- defined in x axis by observed p-value using -log10
- red line represent perfect quantile matching



Global visualisation : QQ-plot

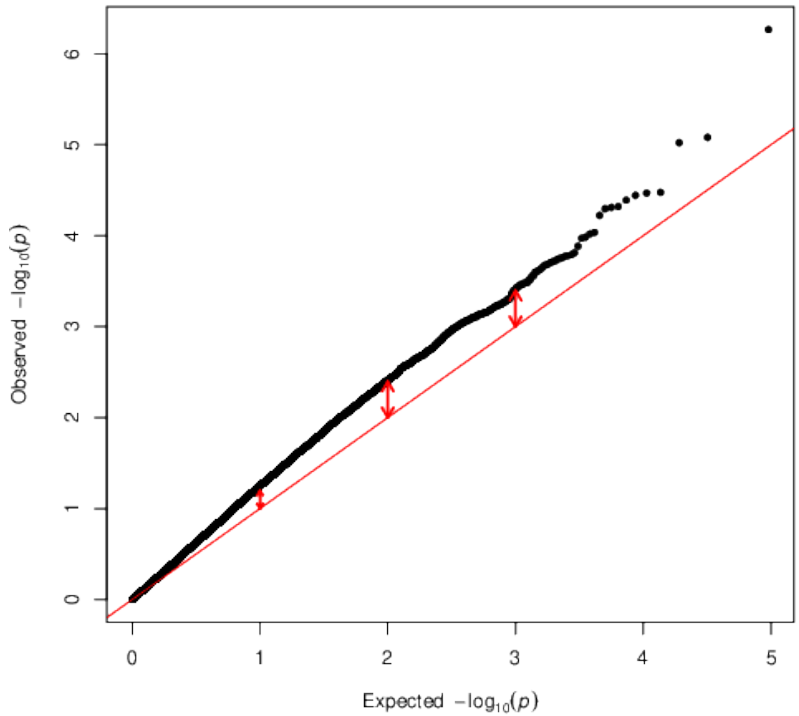
- example QQ-plot where points fit with perfect quantile matching : fit with red line
- no significant positions less than $5x10E-8$ $(-log10 > 7.3)$



Global visualisation : QQ-plot

QQ plot showed a biais compared to red line :

- inflation of p-value compared to null model
- cause :
  - genetics structure
  - allele frequency are not control and lot of low maf < 0.01 %



Global visualisation : QQ-plot and inflation factors

- The genomic inflation factor estimates the amount of inflation by comparing observed test statistics across all genetic variants to those expected under the hypothesis of no effect.

$$\chi^2 = q_\chi^2(1 - P, 1)$$

with P p-value and q_χ^2 quantile function of χ^2 with 1 degrees of freedom

$$\lambda = median(\chi^2)/0.456$$

Global visualisation : QQ-plot and inflation factors

- The genomic inflation factor estimates the amount of inflation by comparing observed test statistics across all genetic variants to those expected under the hypothesis of no effect.
- The genomic inflation factor λ is defined as the ratio of the median of the empirically observed distribution of the test statistic to the expected median, thus quantifying the extent of the bulk inflation and the excess false positive rate.

$$\chi^2 = q_\chi^2(1 - P, 1)$$

with P p-value and q_χ^2 quantile function of χ^2 with 1 degrees of freedom

$$\lambda = median(\chi^2)/0.456$$

Global visualisation : QQ-plot and inflation factors

- The genomic inflation factor estimates the amount of inflation by comparing observed test statistics across all genetic variants to those expected under the hypothesis of no effect.
- The genomic inflation factor λ is defined as the ratio of the median of the empirically observed distribution of the test statistic to the expected median, thus quantifying the extent of the bulk inflation and the excess false positive rate.
- commonly accepted that $\lambda <$ 1.1 are acceptable

$$\chi^2 = q_\chi^2(1 - P, 1)$$

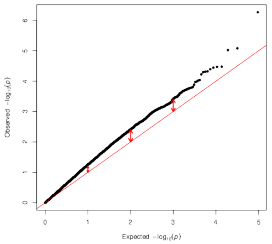
with P p-value and q_χ^2 quantile function of χ^2 with 1 degrees of freedom

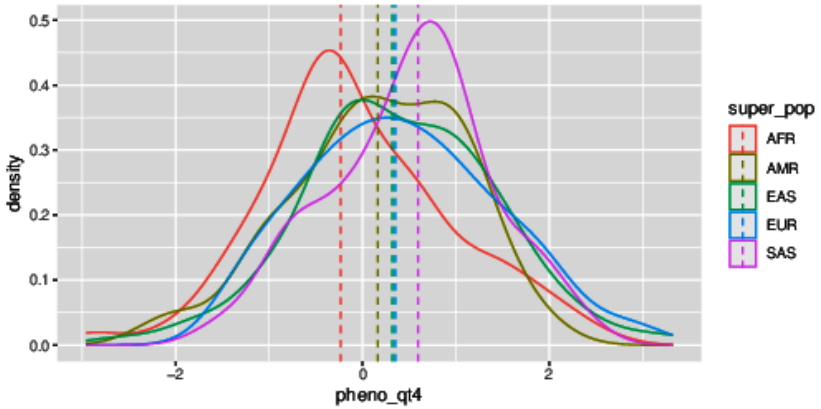
$$\lambda = median(\chi^2)/0.456$$

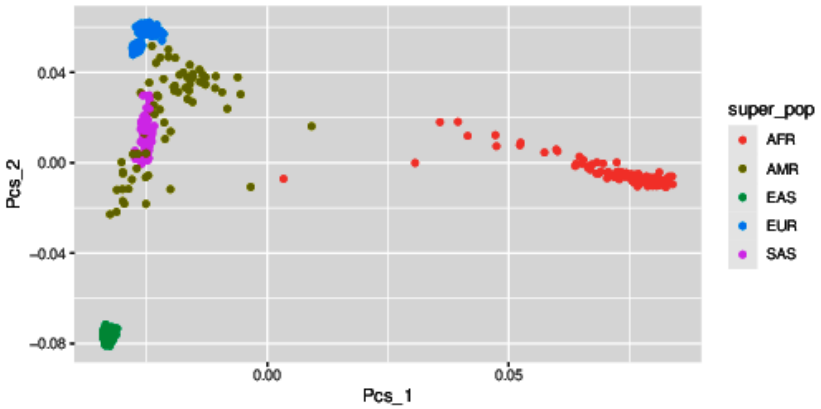
Global visualisation : QQ-plot and inflation factors

QQ plot showed a bias :

- Inflation factors is 1.4
- Median of phenotype showed a relation between genetics structure and phenotype distribution.
- explanation :
  - genetics generated from 1000 genome with 5 populations, phenotypes simulated using a bias using population
  - GWAS done using plink without structure correction
- correction of bias :
  - running GWAS using a linear mixed models using relatedness or Principal Component
  - correct P-value using $P_c = P/\lambda$

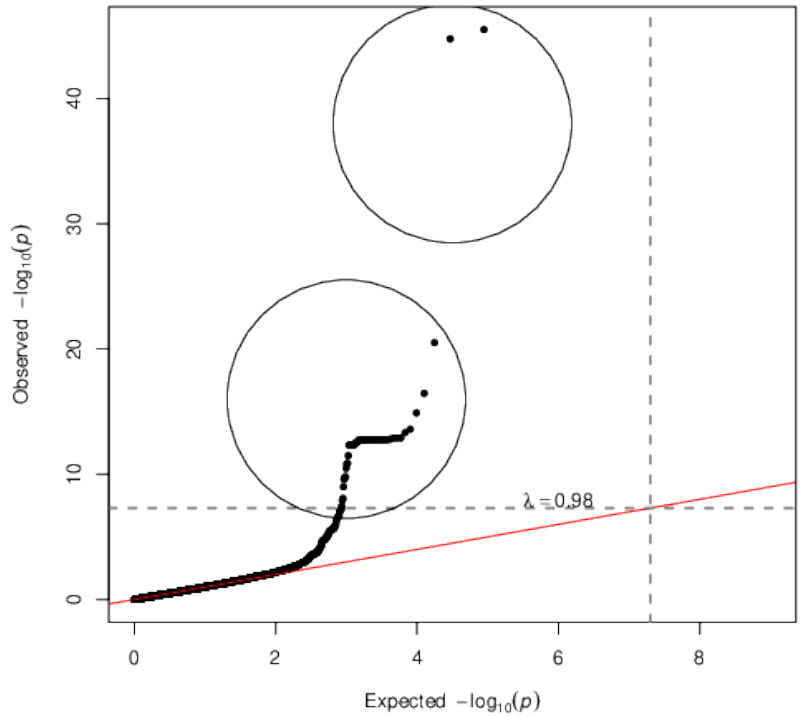






Global visualisation : QQ - plot

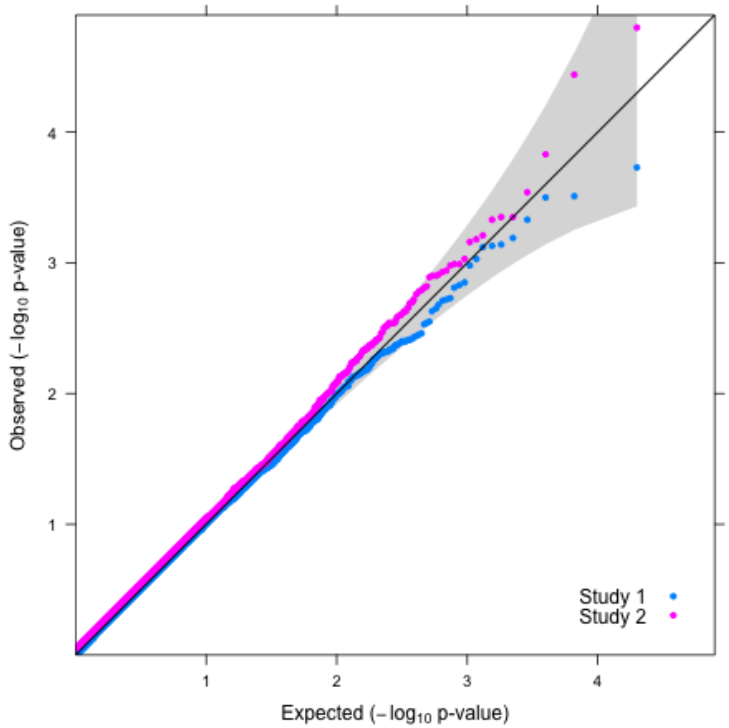
QQ plot indicated also if you have a some significant SNPs



Global visualisation : QQ - plot

Resources in R : qqman, fastman R-library

- A Fancier QQ Plot by Matthew Flickinger : `https://genome.sph.umich.edu/wiki/Code_Sample:_Generating_QQ_Plots_in_R`

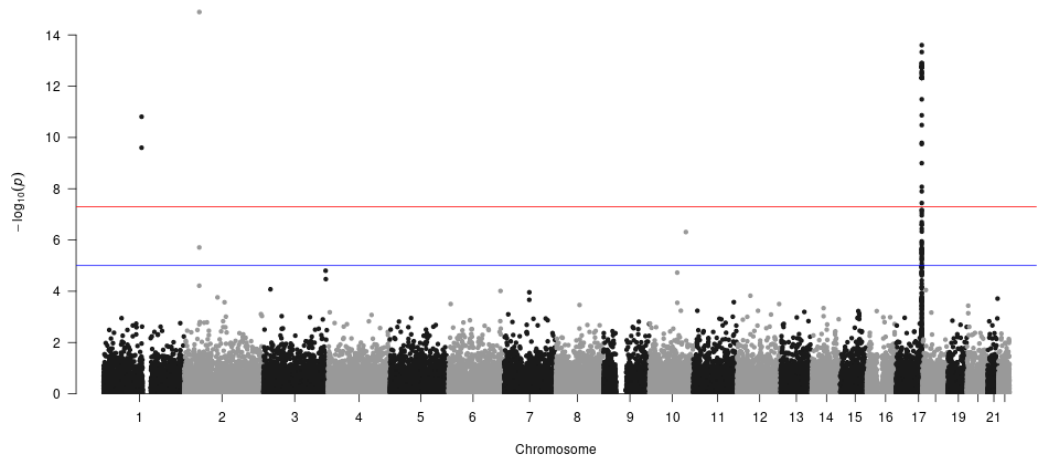


Resource using interface web : FUMA, LocusZoom V2

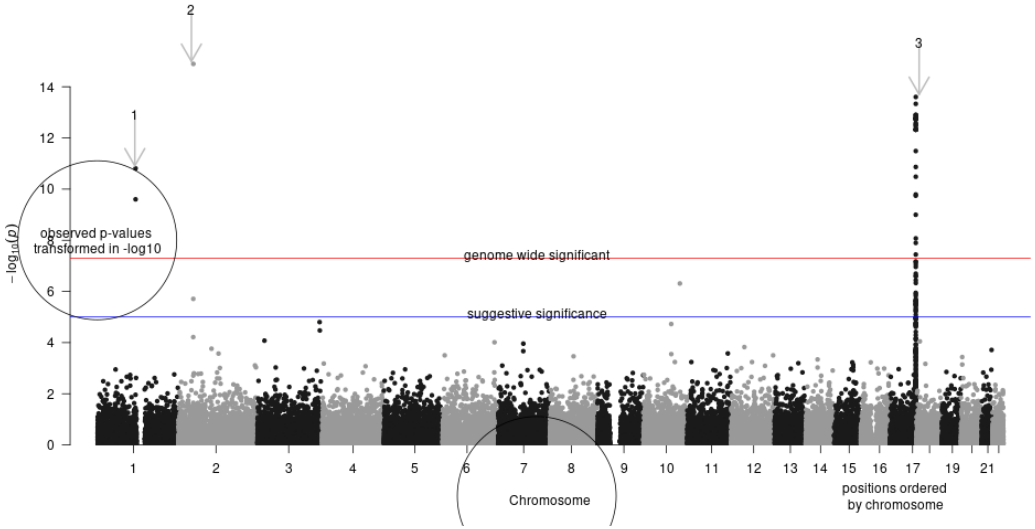
Manhattan plot

Global visualisation : Manhattan plot

Manhattan plots represent the P values of the entire GWAS on a genomic scale. The P values are represented in genomic order by chromosome and position on the chromosome (x axis). The value on the y axis represents the $-log10$ of the P value.



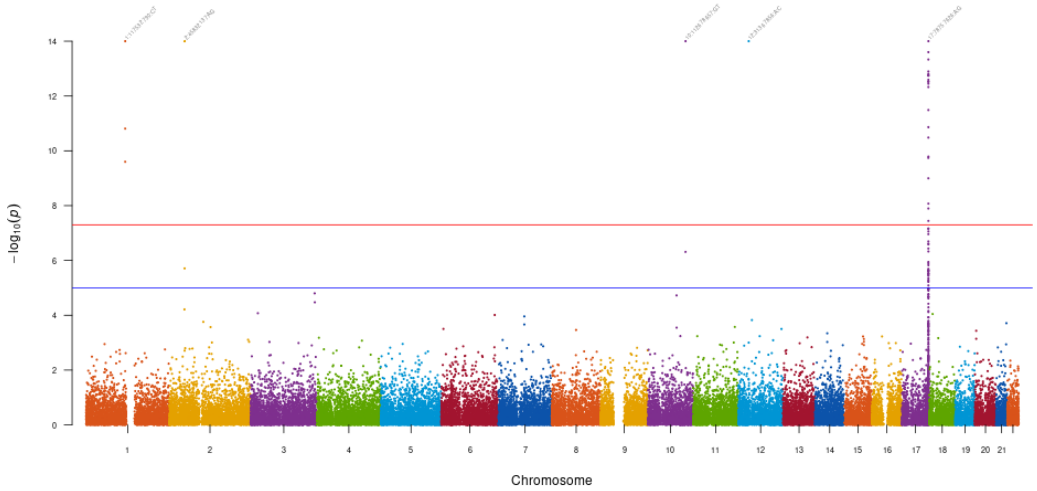
Global visualisation : Manhattan plot



- x axis : result ordered by chromosome and positions
- y axis : $-log10(P)$
- horizontal red line : significance at $5x10^{-8}$
- horizontal blue line : suggestive significant at 10^{-5}
- arrows 1,2 showed low support of result (need to do a zoom of region to confirm)
- arrows 3, "tower" high support of result
- example build using qqman library

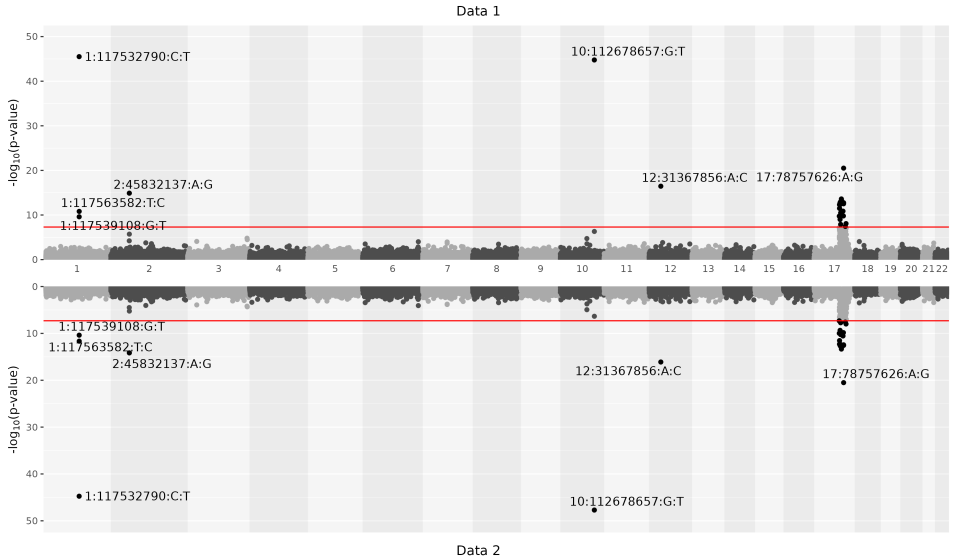
Global visualisation: Manhattan plot - other example

Using fastman library, give more custom option as annotation of lead snps.



Global visualisation: Manhattan plot - other example

Using hudson library,allowed to used mirror Manhattan plot with two data set and annotations of lead snps.



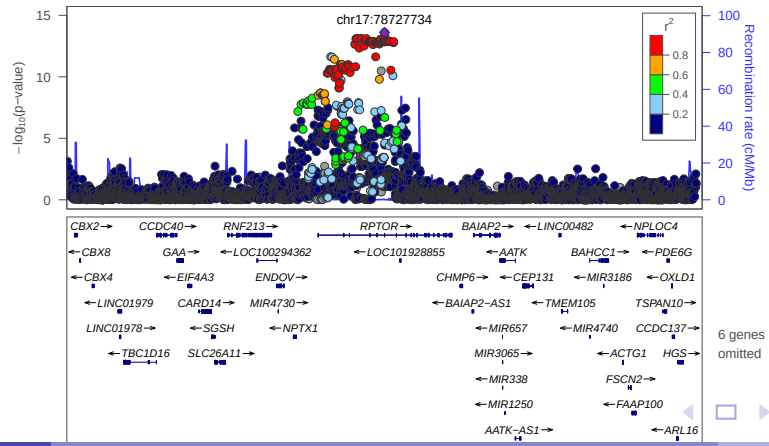
Resource using interface web : fuma, locus zoom V2.

Regional plot

Local visualisation: Regional Association Plot

Pre-defined area of the genome extracted where you plot p-value with annotation, ld...

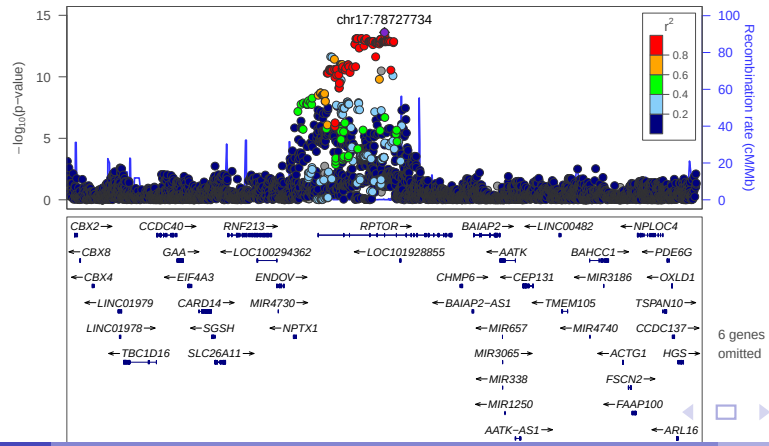
- A regional association plot is essentially a zoomed-in Manhattan plot.



Local visualisation: Regional Association Plot

Pre-defined area of the genome extracted where you plot p-value with annotation, ld...

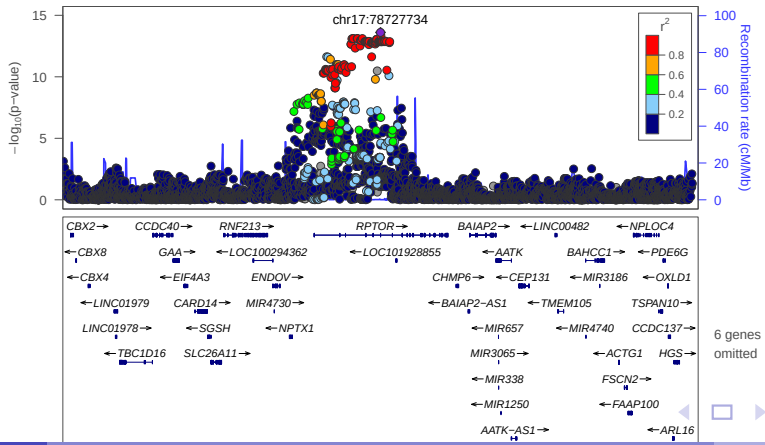
- A regional association plot is essentially a zoomed-in Manhattan plot.
- Allowing the researcher to look at associations in a specific region of the genome



Local visualisation: Regional Association Plot

Pre-defined area of the genome extracted where you plot p-value with annotation, ld...

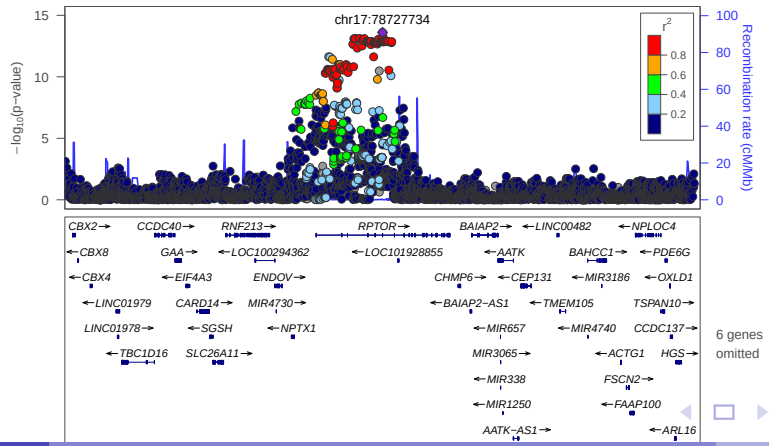
- A regional association plot is essentially a zoomed-in Manhattan plot.
- Allowing the researcher to look at associations in a specific region of the genome
- Analyse how is support lead SNP by other SNPs using LD between Lead SNPs and snps in windows



Local visualisation: Regional Association Plot

Pre-defined area of the genome extracted where you plot p-value with annotation, ld...

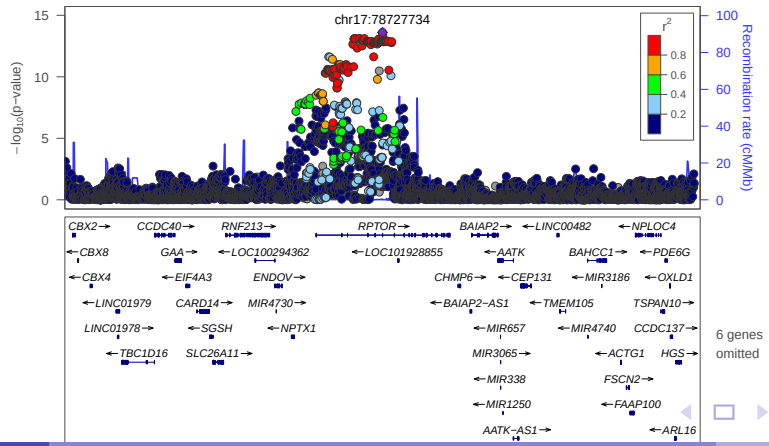
- A regional association plot is essentially a zoomed-in Manhattan plot.
- Allowing the researcher to look at associations in a specific region of the genome
- Analyse how is support lead SNP by other SNPs using LD between Lead SNPs and snps in windows
- To generated your regional you can use :



Local visualisation: Regional Association Plot

Pre-defined area of the genome extracted where you plot p-value with annotation, ld...

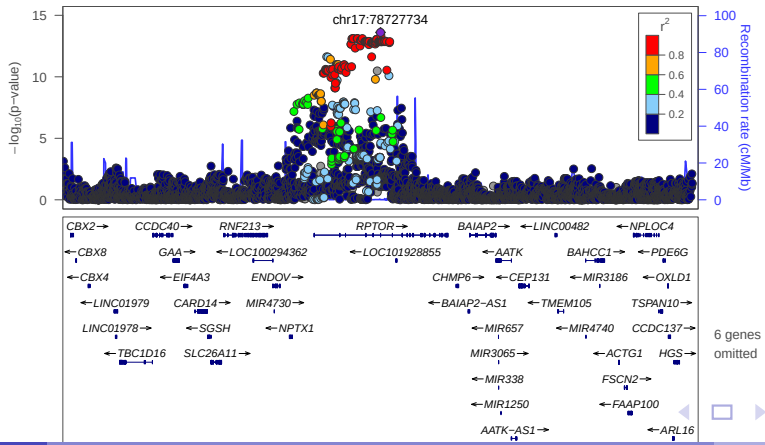
- A regional association plot is essentially a zoomed-in Manhattan plot.
- Allowing the researcher to look at associations in a specific region of the genome
- Analyse how is support lead SNP by other SNPs using LD between Lead SNPs and snps in windows
- To generated your regional you can use :
  - locus zoom stand alone : statics, R, python 2.0



Local visualisation: Regional Association Plot

Pre-defined area of the genome extracted where you plot p-value with annotation, ld...

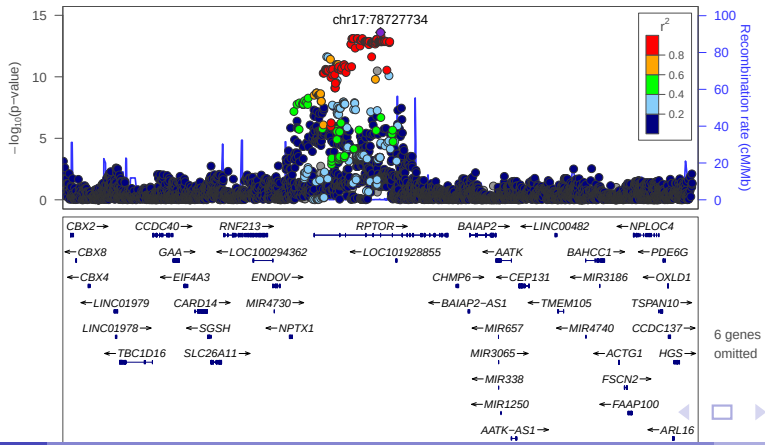
- A regional association plot is essentially a zoomed-in Manhattan plot.
- Allowing the researcher to look at associations in a specific region of the genome
- Analyse how is support lead SNP by other SNPs using LD between Lead SNPs and snps in windows
- To generated your regional you can use :
  - locus zoom stand alone : statics, R, python 2.0
  - locus zoom v2 : java or graphic interface, interactive



Local visualisation: Regional Association Plot

Pre-defined area of the genome extracted where you plot p-value with annotation, ld...

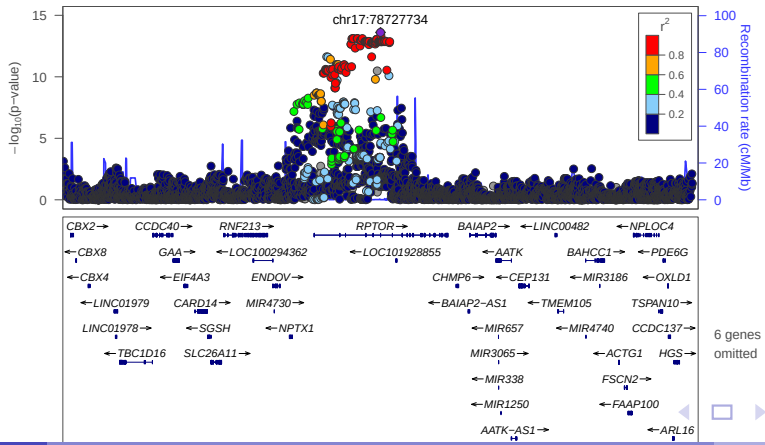
- A regional association plot is essentially a zoomed-in Manhattan plot.
- Allowing the researcher to look at associations in a specific region of the genome
- Analyse how is support lead SNP by other SNPs using LD between Lead SNPs and snps in windows
- To generated your regional you can use :
  - locus zoom stand alone : statics, R, python 2.0
  - locus zoom v2 : java or graphic interface, interactive
  - FUMA : website, graphic interface, interactive



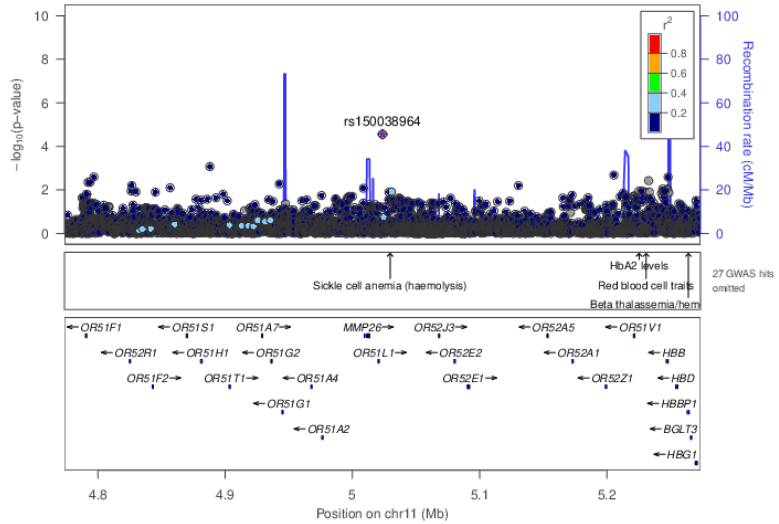
Local visualisation: Regional Association Plot

Pre-defined area of the genome extracted where you plot p-value with annotation, ld...

- A regional association plot is essentially a zoomed-in Manhattan plot.
- Allowing the researcher to look at associations in a specific region of the genome
- Analyse how is support lead SNP by other SNPs using LD between Lead SNPs and snps in windows
- To generated your regional you can use :
  - locus zoom stand alone : statics, R, python 2.0
  - locus zoom v2 : java or graphic interface, interactive
  - FUMA : website, graphic interface, interactive
  - various R and python library : QC_GWAS, RACER...



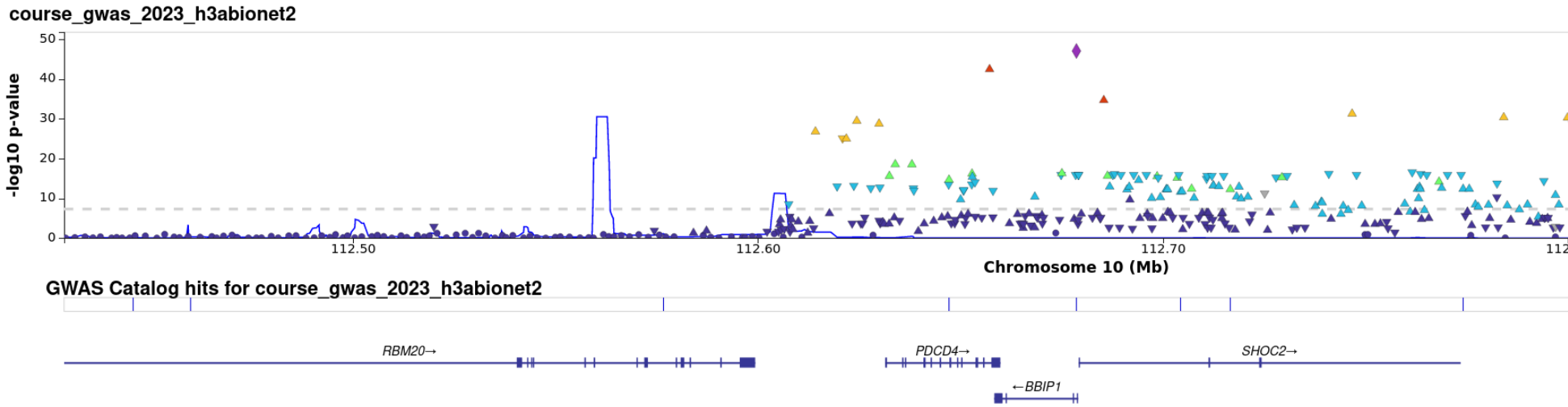
Local visualisation : LocusZoom example



If you lead SNPs has a low support : few SNPs has a low p-value on the region, you need to check for instance, frequency of your positions, plot phenotypes compared to genotype; if positions had been genotyped need to check cluster of luminosity compared to genotyping.

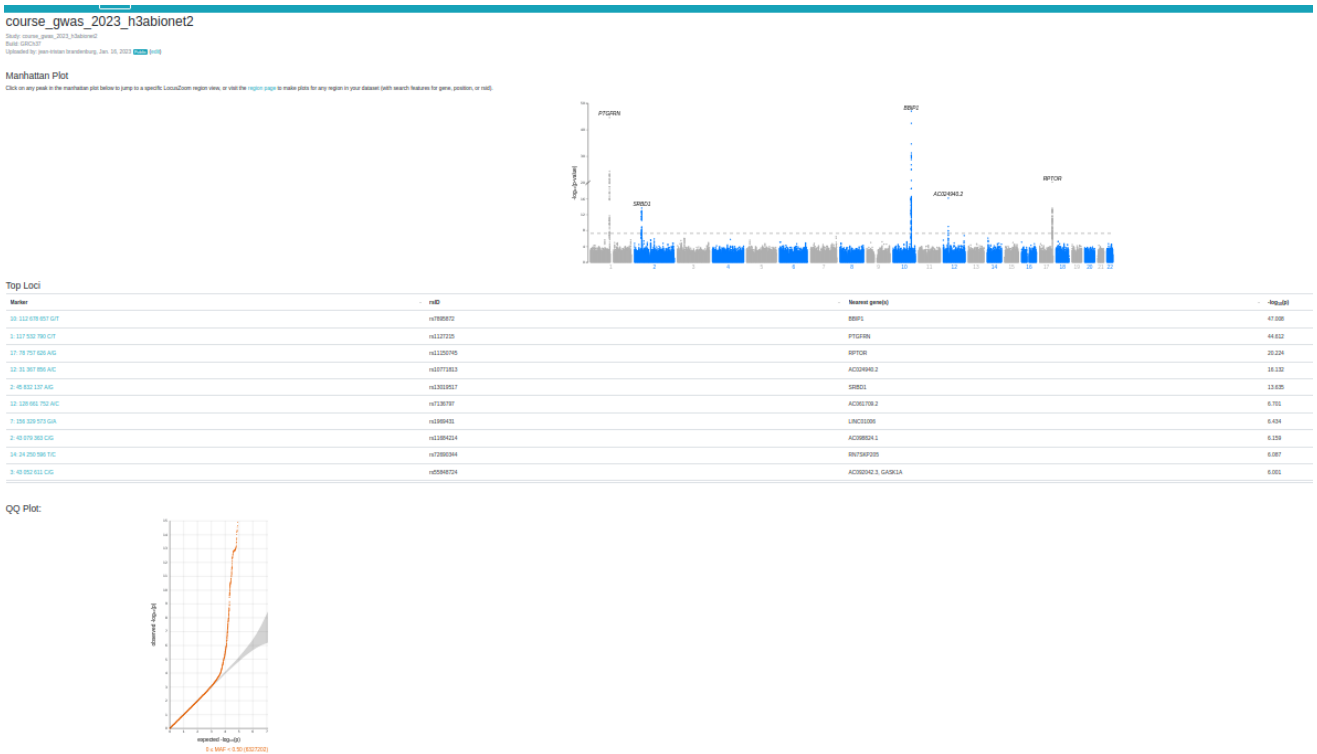
Local visualisation: Locus Zoom V2

Locus zoom v2 : web interface. using local visualisation gave information relative to gene in region, GWAS catalog and Linkage disequilibrium between lead SNPs and SNPs in the region. LD used come from external data set : 1000 Genome project



Local visualisation: Locus Zoom V2

Locus zoom V2 give an easy interactive using way to visualise and explore data, build also QQ plot, best result, Manhattan plot and regional plot, but less option to customise

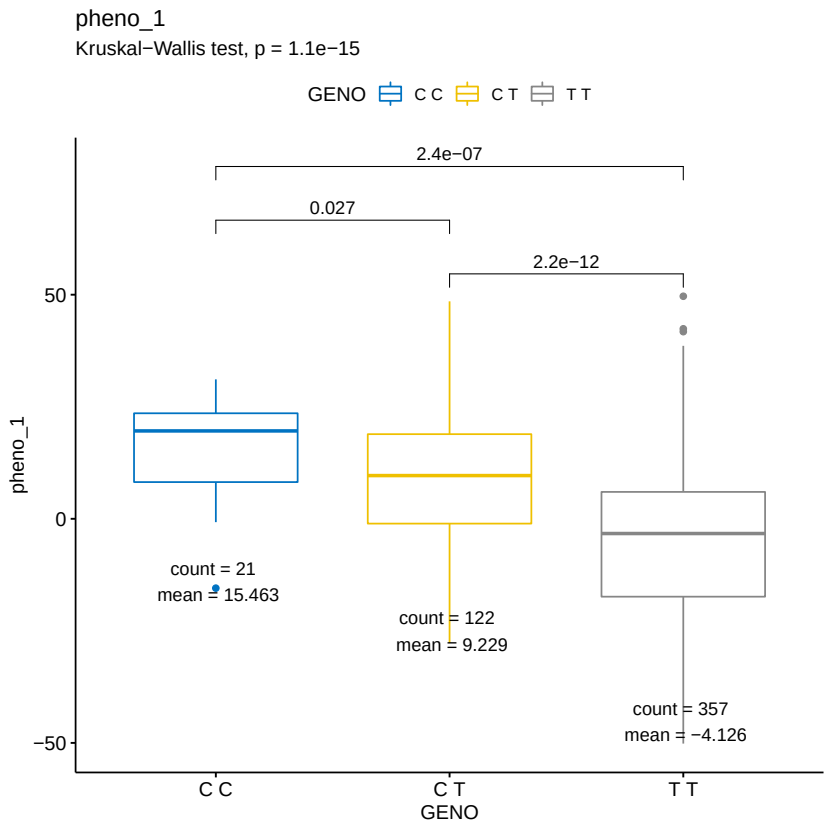


see example : `https://my.locuszoom.org/gwas/91333/`

Positions visualisation

Positions visualisation : distribution phenotypes by genotype

Plot distribution of your phenotype by genotype : give and indication how your variable follow your genotype : follow an additive models? are they properly distributed between genotypes?



Position visualisation : Cluster plot

genotype calling with array:

- steps consist to transform image (raw data) in genotype using intensity, each positions individuals defined by two level of intensity.
- Algorithm to defined genotype used intensity transformed at each locus and each individual.
- error of cluster, or level intensity can have impact of false positive on your GWAS result.
- Positions had been genotyped and are significant can be checked.
- Plot of cluster plot will depend array : illumina, Affymetrix and algorithm /software used (i.e. different output).

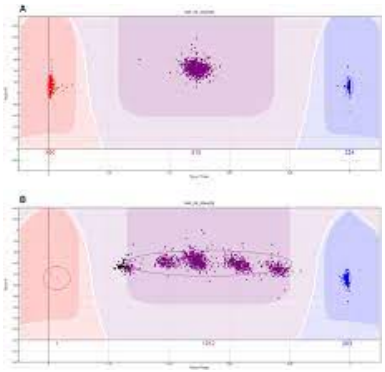


Figure: source : Bianco, 2020, plos one

Meta analysis

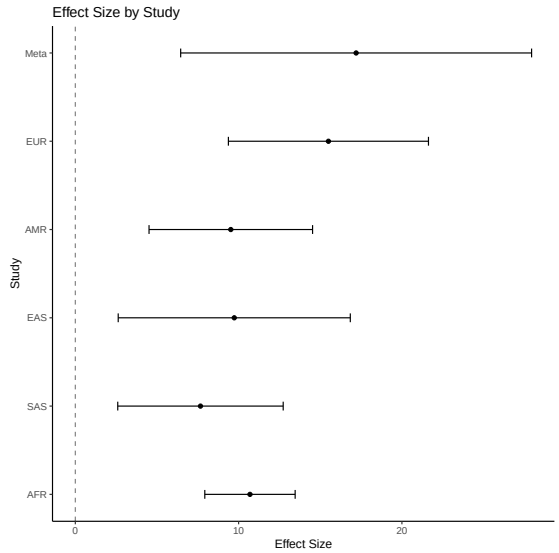
- Meta-analysis is a set of methods that allows the quantitative combination of data from multiple studies.
- Meta-analysis gave your a beta, se and p-value.
- QQ-plot, Manhattan plot and Regional Association plot can be done on summary statistics
- for regional plot or Fine mapping: you need to used adapted data to computed Linkage disequilibrium.
- forest plot is a representation at one position gave you distribution of your various beta and se

Meta Analysis in GWAS

Meta analysis : forest plot

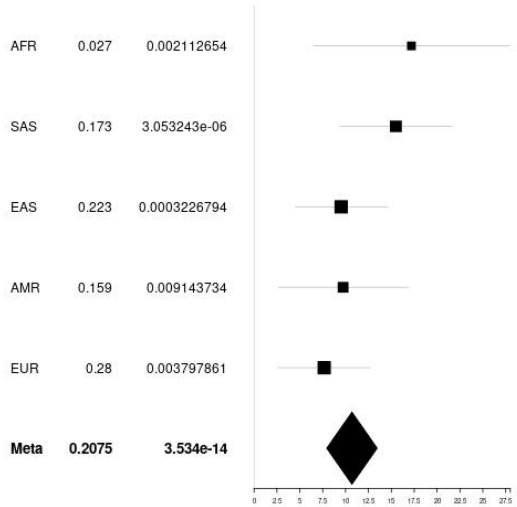
A forest plot,is a graphical display of estimated results from a number of scientific association at one positions running same phenotype from different population

- give your indication how different summary follow same trend.
  - Effect size.
  - standard error.
  - highlight meta result.



Meta analysis : forest plot, other example

Used forestplot library for R - 3.6, we add information as Allele frequency and P-value



Conclusion

Conclusion

Visualisation of result :

- Visualisation in GWAS is a step essential and complementary of statistics method to :
  - understand your data, extract data of interest
  - research bias
  - publications
- Softwares and approaches are more or less easy and more or less configurable.

Conclusion : resources

Softwares / library	Scale	Data	Platforms
qqman / hudson / fastman	QQ - plot or/and Manhattan plot	Summary statistics / annotation	R - library
CGT-VL / Z-browse/ Fuma / locus zoom V2	QQ - plot or/and Manhattan plot / regional plot	Summary statistics	web interface
Locus Zoom Stand alone	Regional plot	Summary statistics / ld	R/python script
R / forestplot	Positions / meta analyse	Summary statistics	R
boxplot / ggplot2	Positions :pheno geno	phenotype / genotype	R
gwaslab	QQ - plot or/and Manhattan plot / regional plot	Summary statistics	python
Assocplots	Manhattan / QQ plot	Summary statistics	python, Interactive, statics
Genesis tool	POPULATION structure	genotype	java / interative

Thank you for your attention

contact : jean-tristan.brandenburg@wits.ac.za

Global visualisation : QQ - plot in practice

```
# import library
library(data.table) ## allow to read big file using fread function
data_sumstat<-fread('assoc/result.plink') # reading summary statistics
# identify colums contains p-value
head(data_sumstat, 2)
# dimension of data
dim(data_sumstat)
# extraction of column contains p_value (P for plink or p_wald for gemma)
pvector<-data_sumstat$P
# clean na value, null value and allow finite vale
pvector <- pvector[!is.na(pvector) & !is.nan(pvector) & !is.null(pvector) & is.finite(
    pvector) & pvector < 1 & pvector > 0]
```

Global visualisation : QQ - plot in practice

```
# sort and transform p -value
o = -log10(sort(pvector, decreasing = FALSE))
# computed value expected, pppoints probability points for the continuous sample quantile
e = -log10(ppoints(length(pvector)))
# plot of expexted points and observed in y-axis
plot(e,o, xlab='Expected', ylab='Observed')
# add red line
abline(0, 1, col = "red")
```

Global visualisation : QQ - plot in pratice

using qqman library (qq2.r)

```
# import library
library(data.table) ## allow to read big file using fread function
library(qqman)
data_sumstat<-fread('assoc/result.plink') # reading summary statistics
## save your plot
png("qq.png") ## function to save can be png, tiff, pdf, avoid used pdf all points will
     be plot and file can reach a big size.
qq(data_sumstat$P) ## implemented same line than before.
dev.off() # save plot end
```

Global visualisation : QQ - plot in practice

computed genomic inflation factor (λ)

```
# import library
library(data.table) ## allow to read big file using fread function
data_sumstat<-fread('assoc/result.plink') # reading summary statistics

chisq <- qchisq(1-data_sumstat$P[!is.na(data_sumstat$P)],1) # transform p in chisq using
    density
lambda<-median(chisq)/qchisq(0.5,1) # computed lambda
```

Global visualisation : QQ - plot

Other resources

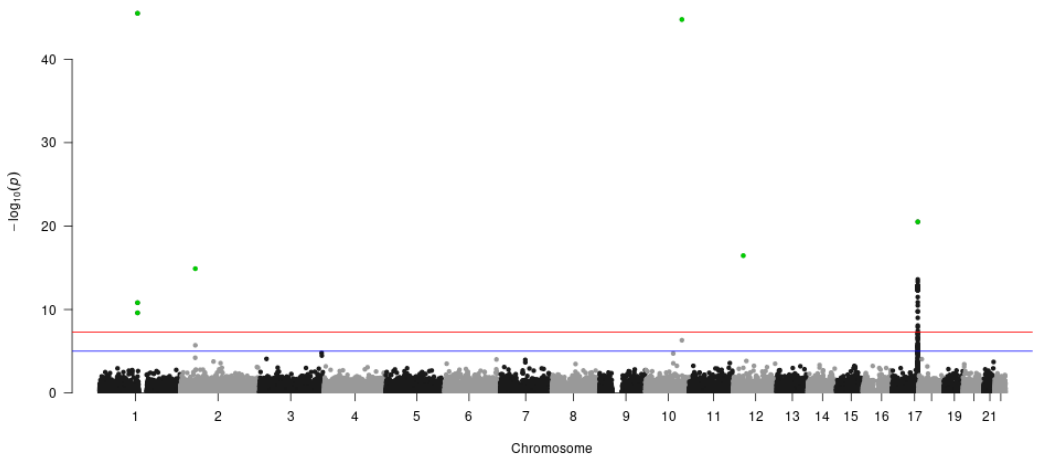
- fastman (R-library) : https://github.com/kaustubhad/fastman

```
# import library
library(data.table) ## allow to read big file using fread function
#devtools::install_github('kaustubhad/fastman')
library(fastman)
data_sumstat<-fread('assoc/result.plink') # reading summary statistics
# identify colums contains p-value
head(data_sumstat, 2)
# dimension of data
fastqq (data_sumstat$P)
```

Global visualisation : Manhattan plot - exercise

using qqman library. manhattan. qqman offer few custom option

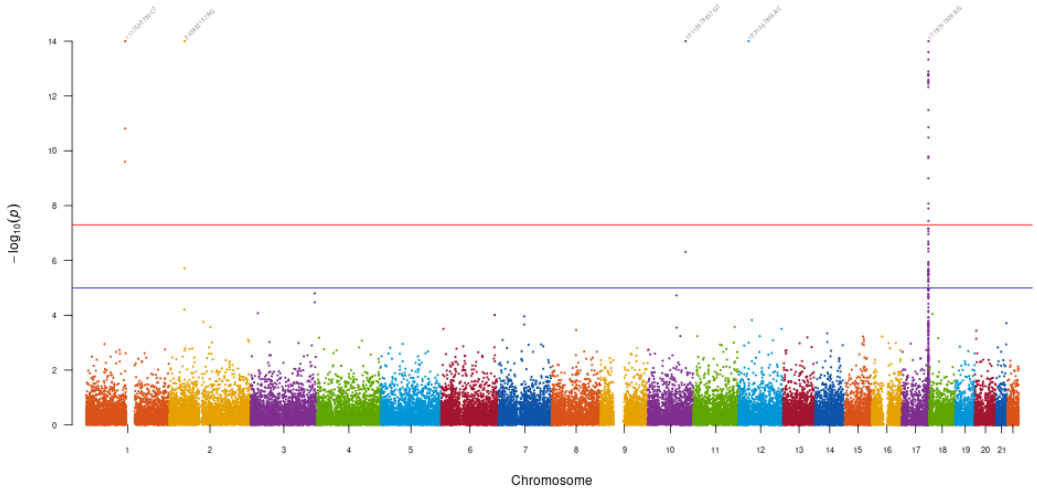
```
# import library
library(qqman)
library(data.table)
library(data.table) ## allow to read big file using fread function
data_sumstat<-fread('assoc/asso_phenoqt2.gemma') # reading summary statistics
#dentify colums contains p-value
head(data_sumstat, 2)
# dimension of data
dim(data_sumstat)
png('figure/man_qqman.png',  width = 480*2, height = 480)
## used mamhattan function of qq with argument chro header, bp header, p header and snp
    header
manhattan(datagemma,chr = "chr",bp = "ps",p = "p_wald", snp='rs')
dev.off()
```



Global visualisation : Manhattan plot - other example

using fastman library. give more custom option as annotation of lead SNPs.

```
# import library
library(data.table) ## allow to read big file using fread function
#devtools::install_github('kaustubhad/fastman')
library(fastman)
data_sumstat<-fread('assoc/asso_phenoqt2.gemma') # reading summary statistics
# identify colums contains p-value
head(data_sumstat, 2)
# dimension of data
png('qq_fastman.png')
fastman(data_sumstat,  chr = "chr", bp = "ps", p = "p_wald", snp="rs",annotatePval=5E-8)
dev.off()
```



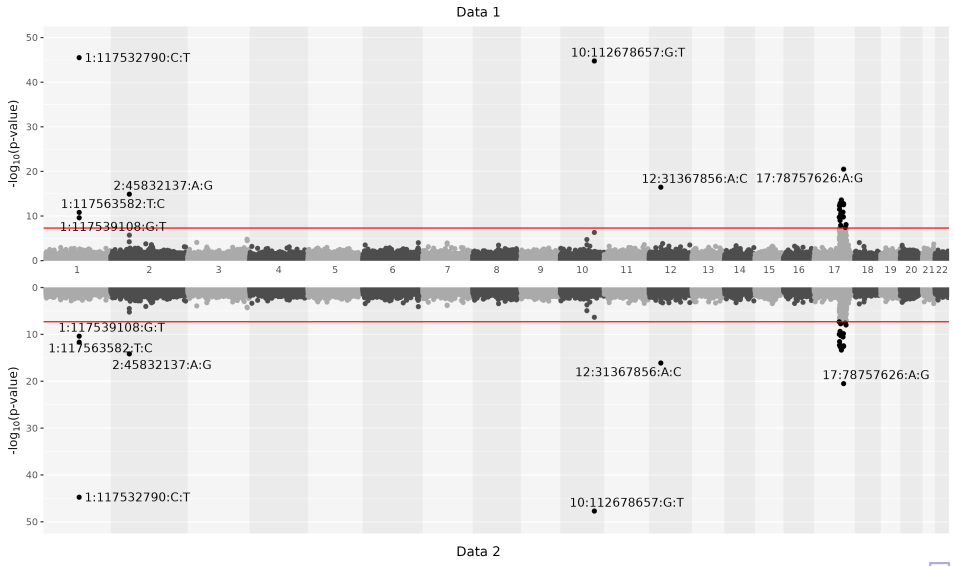
Position visualisation : Cluster plot

```
Index,Name,Chr,Position,AA Freq,AB Freq,BB Freq,Call Freq,Minor Freq,10% GC,50% GC
1,200610-1,MT,2757,0.9935668,0.00,0.006433167,0.997149,0.006433167,0.2633592,0.2637034
2,200610-10,MT,6753,0.9957234,0.00,0.00427655,1,0.00427655,0.3727026,0.3727026
3,200610-100,MT,15173,0.02068474,0.00,0.9793153,0.9992872,0.02068474,0.2195367,0.2195367
```

Global visualisation : Manhattan plot - exercises

using hudson library library. give mirror Manhattan plot, annotation of snps

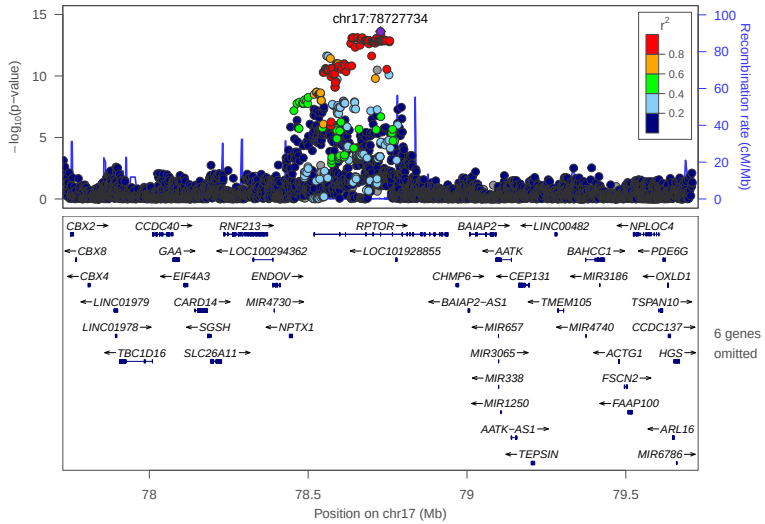
```
# import library
library(data.table) ## allow to read big file using fread function
library(hudson)
data_sumstat_1<-fread('assoc/asso_phenoqt1.gemma') # reading summary statistics
data_sumstat_2<-fread('assoc/asso_phenoqt2.gemma') # reading summary statistics
# data frame, must contain SNP, CHR, POS, pvalue, optional Shape
# select of column of interrest and change name, order is important.
data_sumstat_1<-data_sumstat_1[,c('rs', "chr",'ps', 'p_wald')];names(data_sumstat_1)<-c("
    SNP","CHR","POS", "pvalue")
data_sumstat_2<-data_sumstat_2[,c('rs', "chr",'ps', 'p_wald')];names(data_sumstat_2)<-c("
    SNP","CHR","POS", "pvalue")
# Create a basic plot with Bonferroni lines and highlighting using the toy gwas datasets
gmirror(top=data_sumstat_2, bottom=data_sumstat_1, tline=5*10**-8, bline=5*10**-8,
    toptitle="Data 1", bottomtitle = "Data 2", highlight_p = c(5*10**-8, 5*10**-8),
    annotate_p = c(5*10**-8, 5*10**-8), highlighter="black" ,file="man_hudson")
```



Local visualisation : LocusZoom stand alone

Locus zoom give gene info, Id with lead snps. Option for region, can be rsid or region of interrest. we selected chr start and end.

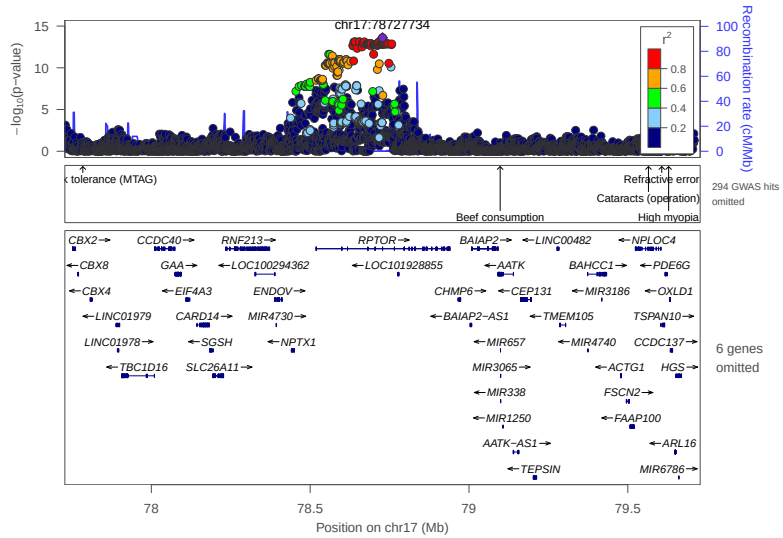
```
#we format file in epact
echo -e "#CHROM\tBEGIN\tEND\tMARKER_ID\tPVALUE" > data/data_regplot.epact
sed '1d' data/data_regplot.gemma |awk '{print $1"\t"$3"\t"$3"\t"$2"\t"$12}' >> data/
    data_regplot.epact
# we run locuszoom using database from locus zoom
locuszoom/bin/locuszoom --build hg19 --epacts  data/data_regplot.epact --source 1000
    G_Nov2014 --pop EUR --chr 17 --start 77727734 --end 79727734
# we run locuszoom using database from locus zoom add locus zoom
```



Local visualisation : LocusZoom stand alone

We add computed LD around lead SNPs and format file for locus zoom and add option for GWAS catalog. locus zoom offer a range of option that can be highly customise.

```
## plot using LD
#we computed LD
plink -bfile data/chr7_77727734_79727734      --ld-window-kb 10000 --ld-snp  17:78727734:T:C  --ld-window 99999  --ld-window-r2 0 -out data/chr7_77727734_79727734 --r2
# format plink ld for locus zoom
echo -e "snp1\tsnp2\tdprime\trsquare" > data/chr7_77727734_79727734_lz.ld
# we format for locuszoom LD and  defined snps as "chr"$chro:bp
sed '1d' data/chr7_77727734_79727734.ld | awk '{print "chr"$1":"$2"\tchr"$1":"$5"\tNA\t"$7}' >> data/chr7_77727734_79727734_lz.ld
locuszoom/bin/locuszoom --build hg19 --epacts  data/data_regplot.epact --chr 17 --start 77727734 --end 79727734 --gwas-cat whole-cat_significant-only --ld data/chr7_77727734_79727734_lz.ld -p gcld
```



Position visualisation : Cluster plot

```
MT 2755 200610-1 A G . . ... 0/0:1:0.263703:-0.000443399:0.0505827:2.85271:0.268731:3.12144:0.0597945:24607:115
```

Meta analysis : forest plot

format and build data set using different summary statistic for a specific position

```
library(data.table)
DataAFR<-fread('data/AFR_pheno.gemma');
DataEUR<-fread('data/EUR_pheno.gemma')
DataAMR<-fread('data/AMR_pheno.gemma')
DataEAS<-fread('data/EAS_pheno.gemma')
DataSAS<-fread('data/SAS_pheno.gemma')
DataMeta<-fread('data/metal_res_metal.format')
chr=17;bp<-7872773
DataAFR_chrbp<-DataAFR[DataAFR$chr==chr & DataAFR$ps==bp,c('allele1', 'allele0', 'af','beta','se','p_wald')];names(DataAFR_chrbp)<-c('A1','A2', 'AF', 'mean','SE', 'P')
DataSAS_chrbp<-DataSAS[DataSAS$chr==chr & DataSAS$ps==bp,c('allele1', 'allele0', 'af','beta','se','p_wald')];names(DataSAS_chrbp)<-c('A1','A2', 'AF', 'mean','SE', 'P')
DataEAS_chrbp<-DataEAS[DataEAS$chr==chr & DataEAS$ps==bp,c('allele1', 'allele0', 'af','beta','se','p_wald')];names(DataEAS_chrbp)<-c('A1','A2', 'AF', 'mean','SE', 'P')
DataAMR_chrbp<-DataAMR[DataAMR$chr==chr & DataAMR$ps==bp,c('allele1', 'allele0', 'af','beta','se','p_wald')];names(DataAMR_chrbp)<-c('A1','A2', 'AF', 'mean','SE', 'P')
DataEUR_chrbp<-DataEUR[DataEUR$chr==chr & DataEUR$ps==bp,c('allele1', 'allele0', 'af','beta','se','p_wald')];names(DataEUR_chrbp)<-c('A1','A2', 'AF', 'mean','SE', 'P')
DataMeta_chrbp<-DataMeta[DataMeta$CHRO==chr & DataMeta$BP==bp,c('Allele1', 'Allele2', 'Freq1','Effect','StdErr','P-value')];names(DataMeta_chrbp)<-c('A1','A2', 'AF', 'mean','SE', 'P')
## DataMeta doesn't have same ref / alt we swithch beta / af
DataMeta_chrbp$mean<- - DataMeta_chrbp$mean;DataMeta_chrbp$AF <- 1 - DataMeta_chrbp$AF
Data_reschrbp<-rbind(DataAFR_chrbp,DataSAS_chrbp,DataEAS_chrbp, DataAMR_chrbp,DataEUR_chrbp,DataMeta_chrbp)
Data_reschrbp<-cbind(Pop=c('AFR','SAS', 'EAS','AMR', 'EUR', 'Meta'), Data_reschrbp, is.summary=c(F,F,F,F,F,T))
Data_reschrbp$index<-nrow(Data_reschrbp):1
# we computed lower and upper using 95%
Data_reschrbp$lower<- Data_reschrbp$mean-1.96 * Data_reschrbp$SE
Data_reschrbp$upper<- Data_reschrbp$mean+1.96 * Data_reschrbp$SE
```

Positions visualisation : distribution phenotypes by genotype

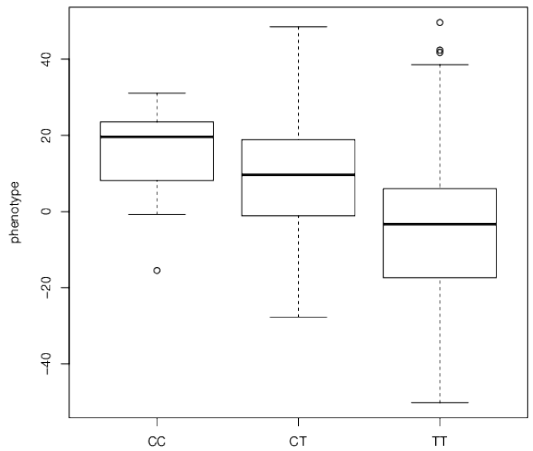
Extract genotype from your plink file in ped file using plink

```
plinkbase=data/chr7_77727734_79727734
filers=data/17_78727734.bed
echo "17        78727734        78727734        17:78727734" > $filers
# format in ped file
out=data/17_78727734bp
plink -bfile $plinkbase --extract range $filers  --recode tab --out $out
```

Positions visualisation : distribution phenotypes by genotype

using R to plot distribution phenotypes by genotype

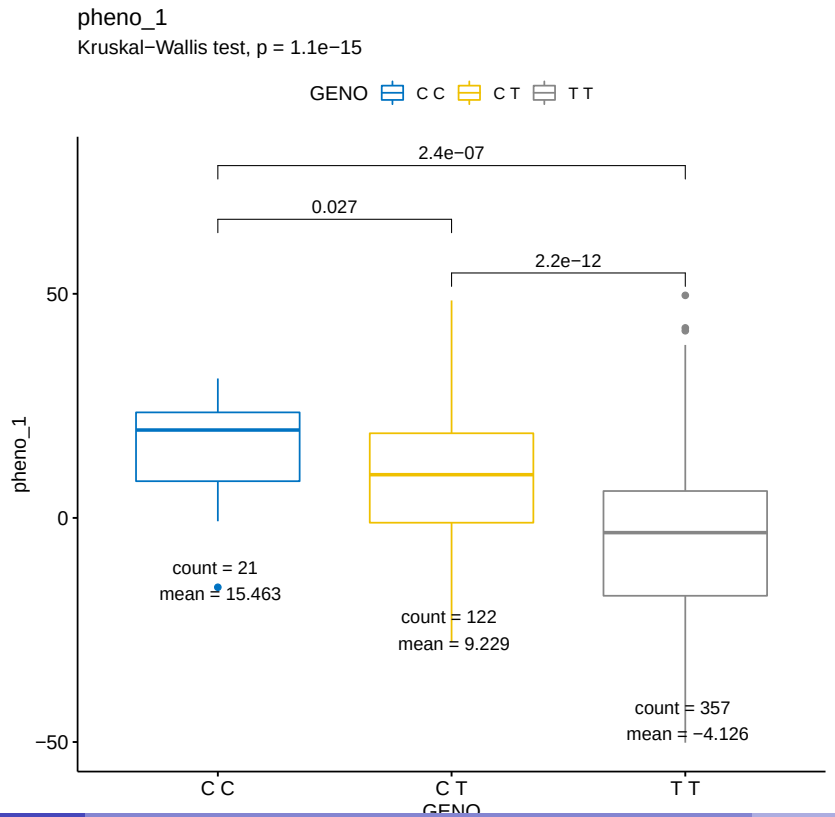
```
## reading pheno
pheno<-read.table('data/KGPH3abionetsub_pheno_qt.pheno', header=T)
## reading geno from plink output
geno<-read.table('data/17_78727734bp.ped')
geno<-geno[,c(1,2,7,8)]
names(geno)<-c('FID','IID', 'G1','G2')
geno$genot<-paste(geno$G1,geno$G2,sep='')
# name geno header=
genopheno<-merge(pheno , geno, by=c(1,2),all=F)
# boxplot by genotype
boxplot(pheno_1~genot,data=genopheno, xlab='', ylab='phenotype')
```



Position visualisation : distribution phenotypes by genotype

Improved figure using "ggpubr" and "gridExtra", script from h3agwas/annotation/ workflow
possibility to add co-variable with plot of residuals. possibility to do a GxE plot.

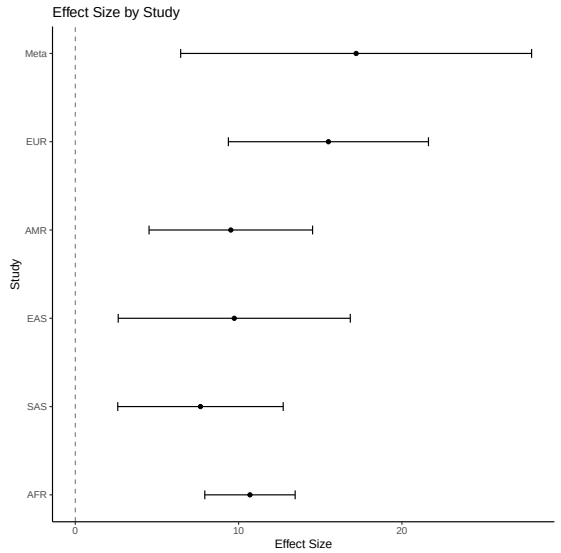
```
data=data/KGPH3abionetsub_pheno_qt.pheno
outpdf=h3agwas_geno.pdf
pheno=pheno_1
./an_plotboxplot.r --ped $out".ped" --data $data --out $outpdf --pheno $pheno
```



Meta analysis : forest plot

Used ggplot2 to build your own forestplot

```
library(ggplot2)
ggplot(data=Data_reschrbp, aes(y=index, x=mean, xmin=lower, xmax=upper)) +
  geom_point() +
  geom_errorbarh(height=.1) +
  scale_y_continuous(breaks=1:nrow(Data_reschrbp), labels=Data_reschrbp$Pop) +
  labs(title='Effect Size by Study', x='Effect Size', y = 'Study') +
  geom_vline(xintercept=0, color='black', linetype='dashed', alpha=.5) +
  theme_classic()
  ## save using ggsave
ggsave("meta_ggplot.pdf")
```



source : https://www.r-bloggers.com/2022/09/forest-plot-in-r-quick-guide/

Meta analysis : forest plot, other example

Used forestplot library for R - 3.6, we add information as Allele frequency and P-value

```
1 library(forestplot)
2 jpeg('forestplot_lib_meta.jpeg')
3 forestplot(Data_reschrbp_2[,c('Pop','AF','P')], mean=Data_reschrbp_2$mean, lower=Data_
      reschrbp_2$lower, upper=Data_reschrbp_2$upper, is.summary=Data_reschrbp$is.summary)
4 dev.off()
```