GWAS and PLINK PRACTICAL

Lucia Colodro-Conde and Katrina Grasby

Contents

- Basic association tests
- Logistic regression
- Linear regression
- Plotting results

Data are freely available from:

ARTICLE

Genetic Control of Human Brain Transcript Expression in Alzheimer Disease

Jennifer A. Webster,[1,2,3,16] J. Raphael Gibbs,[4,5,16] Jennifer Clarke,[6] Monika Ray,[7] Weixiong Zhang,[7,8] Peter Holmans,[9] Kristen Rohrer,[4] Alice Zhao,[4] Lauren Marlowe,[4] Mona Kaleem,[4] Donald S. McCorquodale III,[10] Cindy Cuello,[10] Doris Leung,[4] Leslie Bryden,[4] Priti Nath,[4] Victoria L. Zismann,[1,2] Keta Joshipura,[1,2] Matthew J. Huentelman,[1,2] Diane Hu-Lince,[1,2] Keith D. Coon,[1,2,11] David W. Craig,[1,2] John V. Pearson,[1,2] NACC-Neuropathology Group,[12] Christopher B. Heward,[13,17] Eric M. Reiman,[1,2,14] Dietrich Stephan,[1,2,14] John Hardy,[4,5] and Amanda J. Myers[10,15,*]

http://labs.med.miami.edu/myers/LFuN/LFUN/DATA.html

Note these are genotyped data of 358 unrelated participants and we have limited the analyses to the autosomal chromosomes.

0. Set up

We will use the text editor, the terminal, R Studio, and the browser.

Exercises: Monday_pract2_Exercises.txt

Related documentation:

- PLINK 1.9: https://www.cog-genomics.org/plink/1.9/assoc
- PLINK 1: http://zzz.bwh.harvard.edu/plink/anal.shtml

Basic Association Test

Case-control

- Each individual contributes two counts to 2x2 table.
- Test of association

	Cases	Controls	Total
G	n_{1A}	n_{1U}	$n_{1\cdot}$
T	n_{0A}	n_{0U}	$n_{0\cdot}$
Total	$n_{\cdot A}$	$n_{\cdot U}$	$n_{\cdot\cdot}$

$$X^2 = \sum_{i=0,1} \sum_{j=A,U} \frac{\left(n_{ij} - E\left[n_{ij}\right]\right)^2}{E\left[n_{ij}\right]}$$

where $E\left[n_{ij}\right] = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}$

$$OR = \frac{n_{1A} n_{0U}}{n_{1U} n_{0A}}$$

- X^2 has χ^2 distribution with 1 degrees of freedom under null hypothesis.

-- assoc

It works with case/control and continuous phenotypes.

Case-control (1df chi-square test, outputs assoc)

PLINK will recognise this is a case/control analysis because the phenotype just has:
1 (for controls),
2 (for cases),
and 0/-9/non-numeric (for missing).

Continuous (regression, outputs qassoc)

If the file had more values than 0/-9/non-numeric, 1, and 2, PLINK will recognise the phenotype as continuous.

Note the code for cases, controls, and missing data can be changed using different flags.

This is an Alzheimer's disease case-control sample

- Go to the case-control folder
- How many cases and controls do you have?
- Is there any missing data?

This is an Alzheimer's disease case-control sample

- Go to the case-control folder
- How many cases and controls do you have?
- Is there any missing data?

```
awk '{print $6}' adclean.cc.fam | sort | uniq -c
```

170 cases and 182 controls, no missing data

Exercise 1. Association test for a binary trait

```
plink
--bfile adclean.cc
--assoc
--out 1_adclean.cc
```

Exercise 1. Association test for a binary trait

1_adclean.cc.log

```
Options in effect:
  --assoc
  --bfile adclean.cc
  --out 1_adclean.cc

64148 MB RAM detected; reserving 32074 MB for main workspace.
297237 variants loaded from .bim file.
352 people (188 males, 164 females) loaded from .fam.
352 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked.
Before main variant filters, 352 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.985941.
297237 variants and 352 people pass filters and QC.
Among remaining phenotypes, 170 are cases and 182 are controls.
```

1_adclean.cc.assoc

CHR	SNP	BP	A1	F_A	F_U	A2	CHISQ	P	OR
1	rs3094315	752566	C	0.1824	0.1425	T	2.046	0.1526	1.343
1	rs4040617	779322	G	0.1235	0.09669	A	1.294	0.2554	1.317
1	rs4075116	1003629	G	0.3029	0.2956	A	0.04531	0.8314	1.036
1	rs9442385	1097335	T	0.05952	0.09341	G	2.818	0.09324	0.6143
1	rs10907175	1130727	C	0.08434	0.0678	A	0.6691	0.4134	1.266
1	rs6603781	1158631	T	0.1596	0.1096	C	3.721	0.05375	1.544
1	rs11260562	1165310	T	0.04412	0.04945	C	0.1119	0.738	0.8872
1	rs6685064	1211292	T	0.05422	0.06319	C	0.252	0.6157	0.8499
1	rs307378	1268847	T	0.02679	0.0221	G	0.1611	0.6882	1.218

.assoc, .assoc.fisher (case/control association allelic test report)

Produced by --assoc acting on a case/control phenotype.

A text file with a header line, and then one line per variant typically with the following 9-10 fields:

CHR	Chromosome code
SNP	Variant identifier
BP	Base-pair coordinate
A1	Allele 1 (usually minor)
F_A	Allele 1 frequency among cases
F_U	Allele 1 frequency among controls
A2	Allele 2
CHISQ	Allelic test chi-square statistic. *Not present with 'fisher'/'fisher-midp' modifier.*
P	Allelic test p-value
OR	odds(allele 1 | case) / odds(allele 1 | control)

If the 'counts' modifier is present, the 5th and 6th fields are replaced with:

C_A	Allele 1 count among cases
C_U	Allele 1 count among controls

If --ci 0.*xy* has also been specified, there are three additional fields at the end:

SE	Standard error of odds ratio estimate
L*xy*	Bottom of *xy*% symmetric approx. confidence interval for odds ratio
H*xy*	Top of *xy*% approx. confidence interval for odds ratio

Always check which allele is the effect allele!

https://www.cog-genomics.org/plink/1.9/formats#assoc

(you will rarely use an association test)

Logistic and linear regression

Allowing the inclusion of covariates

How many covariates does the file adpc.txt have? What are they?

How many covariates does the file adpc.txt have? What are they?

```
FID IID PC1 PC2 PC3 PC4
WGAAD 10 0.0550949 0.0507711 0.00845787 -0.00116914
WGAAD 15 0.0470604 0.0474843 0.00315769 0.00810905
WGAAD 18 0.0564277 0.0471303 0.00803162 -0.00242266
WGAAD 20 0.0564051 0.0436962 0.00419304 -0.007482
WGAAD 24 0.0540288 0.0477145 0.00711973 -0.00223988
WGAAD 25 0.0475798 0.0504094 0.00207224 0.00637812
WGAAD 28 0.0570727 0.0493075 0.00609508 -0.00164
WGAAD 29 0.054579 0.0496459 0.00995307 -0.00327564
WGAAD 31 0.0552207 0.0516809 0.00705046 -0.00485599
```

They are principal components of genetic ancestry – *more on this on Tuesday*

Exercise 2.1. Logistic regression with PC1-PC4 as covariates

```
plink
--bfile adclean.cc
--logistic
--covar adpc.txt
--out 2.1_adclean.cc
```

`--freq` can be added to create a separate file with the MAF (only founders) and this can be merged with the results

`--beta` can be added to obtain regression coefficients instead of odds ratios

Exercise 2.1. Logistic regression with PC1-PC4 as covariates

2.1_adclean.cc.log

```
Options in effect:
  --bfile adclean.cc
  --covar adpc.txt
  --logistic
  --out 2.1_adclean.cc


Random number seed: 1551416198
64148 MB RAM detected; reserving 32074 MB for main workspace.
297237 variants loaded from .bim file.
352 people (188 males, 164 females) loaded from .fam.
352 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked.
--covar: 4 covariates loaded.
Before main variant filters, 352 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.985941.
297237 variants and 352 people pass filters and QC.
Among remaining phenotypes, 170 are cases and 182 are controls.
Writing logistic model association results to 2.1_adclean.cc.assoc.logistic ..
```

Exercise 2.1. Logistic regression with PC1-PC4 as covariates

2.1_adclean.cc.assoc.logistic

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P
1	rs3094315	752566	C	ADD	349	1.352	1.437	0.1508
1	rs3094315	752566	C	PC1	349	4.426e-50	-1.866	0.06207
1	rs3094315	752566	C	PC2	349	8.759e+17	0.7017	0.4828
1	rs3094315	752566	C	PC3	349	2.97e+21	0.8452	0.398
1	rs3094315	752566	C	PC4	349	9.61e-25	-1.017	0.3094
1	rs4040617	779322	G	ADD	351	1.276	1.047	0.2953
1	rs4040617	779322	G	PC1	351	8.682e-51	-1.892	0.05843
1	rs4040617	779322	G	PC2	351	2.841e+21	0.8415	0.4001
1	rs4040617	779322	G	PC3	351	2.764e+17	0.687	0.4921

Exercise 2.1. Logistic regression with PC1-PC4 as covariates

.assoc.linear, .assoc.logistic (multi-covariate association analysis report)

Produced by --linear/--logistic.

A text file with a header line, and **T** lines per variant typically with the following nine fields (where **T** is normally the number of terms, but the 'genotypic' and 'hethom' modifiers and the --tests flag can change this):

CHR	Chromosome code. *Not present with 'no-snp' modifier.*
SNP	Variant identifier. *Not present with 'no-snp'.*
BP	Base-pair coordinate. *Not present with 'no-snp'.*
A1	Allele 1 (usually minor). *Not present with 'no-snp'.*
TEST	Test identifier
NMISS	Number of observations (nonmissing genotype, phenotype, and covariates)
'BETA'/'OR'	Regression coefficient (--linear, '--logistic beta') or odds ratio (--logistic without 'beta')
STAT	T-statistic
P	Asymptotic p-value for t-statistic

If --ci 0.*xy* has also been specified, the following three fields are inserted before 'STAT':

SE	Standard error of beta (log-odds) estimate
L*xy*	Bottom of *xy*% symmetric approx. confidence interval
H*xy*	Top of *xy*% approx. confidence interval

Refer to the PLINK 1.07 documentation for more details.

https://www.cog-genomics.org/plink/1.9/formats#assoc_linear

Exercise 2.2. Logistic regression with PC1-PC4 as covariates hiding the covariates)

Exercise 2.2. Logistic regression with PC1-PC4 as covariates hiding the covariates)

https://www.cog-genomics.org/plink/1.9/assoc#linear

Exercise 2.2. Logistic regression with PC1-PC4 as covariates hiding the covariates)

```
plink
--bfile adclean.cc
--logistic hide-covar
--covar adpc.txt
--out 2.2_adclean.cc
```

Exercise 2.2. Logistic regression with PC1-PC4 as covariates hiding the covariates)

2.2_adclean.cc.log

```
Options in effect:
  --bfile adclean.cc
  --covar adpc.txt
  --logistic hide-covar
  --out 2.2_adclean.cc


Random number seed: 1551713462
64148 MB RAM detected; reserving 32074 MB for main workspace.
297237 variants loaded from .bim file.
352 people (188 males, 164 females) loaded from .fam.
352 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked.
--covar: 4 covariates loaded.
Before main variant filters, 352 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.985941.
297237 variants and 352 people pass filters and QC.
Among remaining phenotypes, 170 are cases and 182 are controls.
Writing logistic model association results to 2.2_adclean.cc.assoc.logistic ...
```

Exercise 2.2. Logistic regression with PC1-PC4 as covariates hiding the covariates)

2.2_adclean.cc.assoc.logistic

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P
1	rs3094315	752566	C	ADD	349	1.352	1.437	0.1508
1	rs4040617	779322	G	ADD	351	1.276	1.047	0.2953
1	rs4075116	1003629	G	ADD	351	1.013	0.08192	0.9347
1	rs9442385	1097335	T	ADD	350	0.6388	-1.587	0.1125
1	rs10907175	1130727	C	ADD	343	1.353	1.063	0.2879
1	rs6603781	1158631	T	ADD	344	1.496	1.817	0.06918
1	rs11260562	1165310	T	ADD	352	0.8613	-0.4002	0.689
1	rs6685064	1211292	T	ADD	348	0.8395	-0.529	0.5968
1	rs307378	1268847	T	ADD	349	1.212	0.4058	0.6849

Let’s plot the results

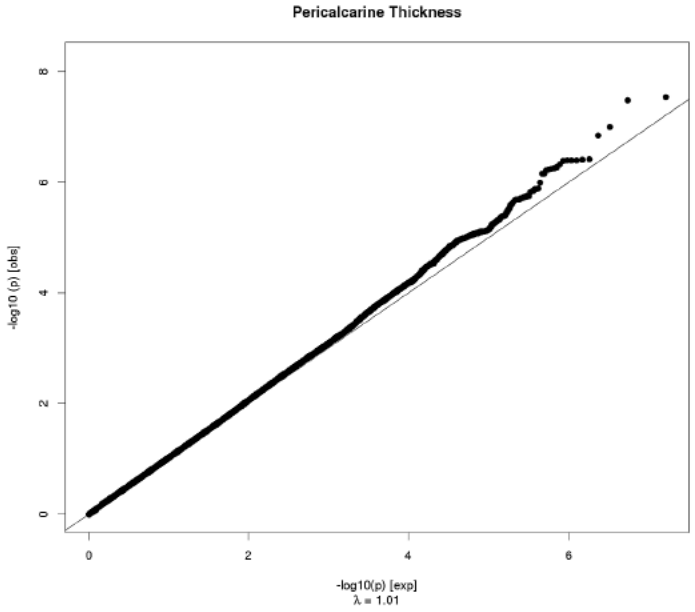
QQ plots, Manhattan plots, and regional plots

QQ plot

- Checks the overall distribution of test statistics or –log10 p-values with the expectation under the null hypothesis of no association (the diagonal line shows where the points should fall under the null).
- Evaluates systematic bias and inflation (undetected sample duplications, unknown familial relationships, gross population stratification, problems in QC…).

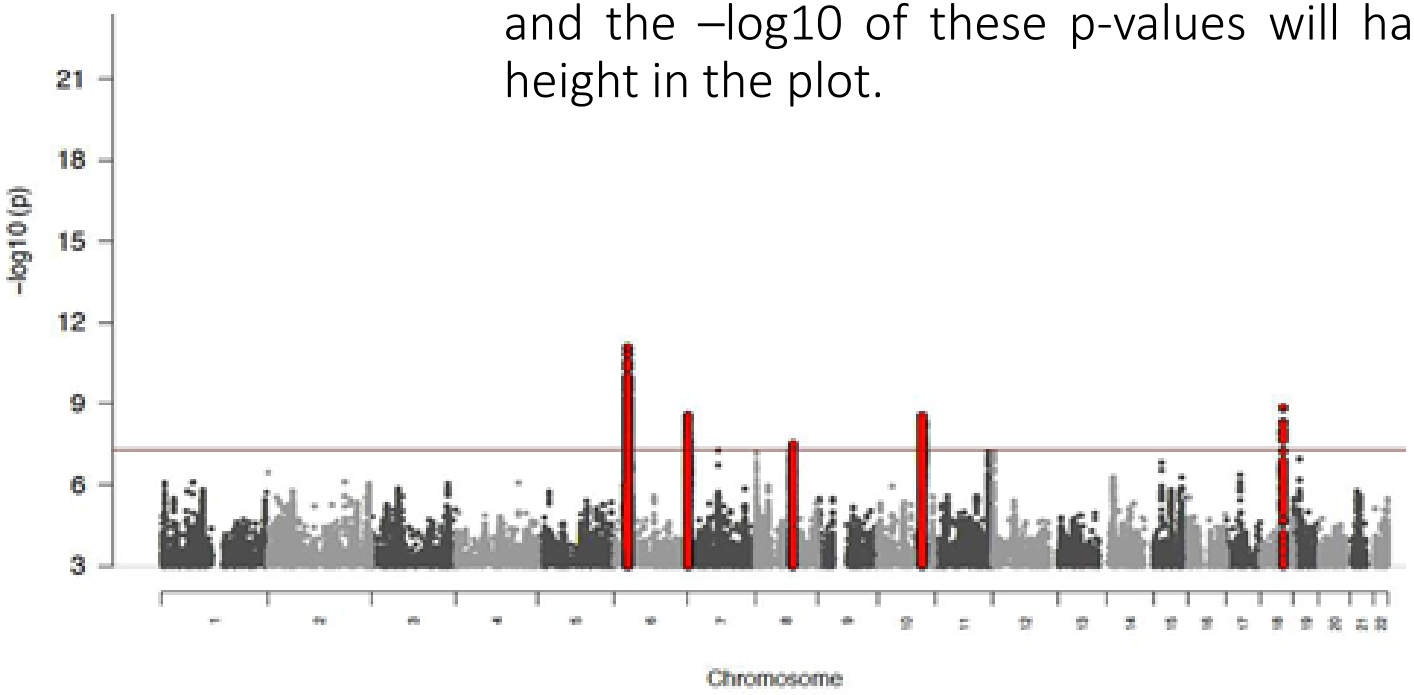
Lambda

- It is the ratio of the median of the empirically observed distribution of the test statistic to the expected median.
- It quantifies the extent of the bulk inflation and the excess false positive rate.
  - The expected median of the chi-square distribution with one degree of freedom is 0.455.
  - lambda=median(chi^2)/0.455
- It should be close to 1.

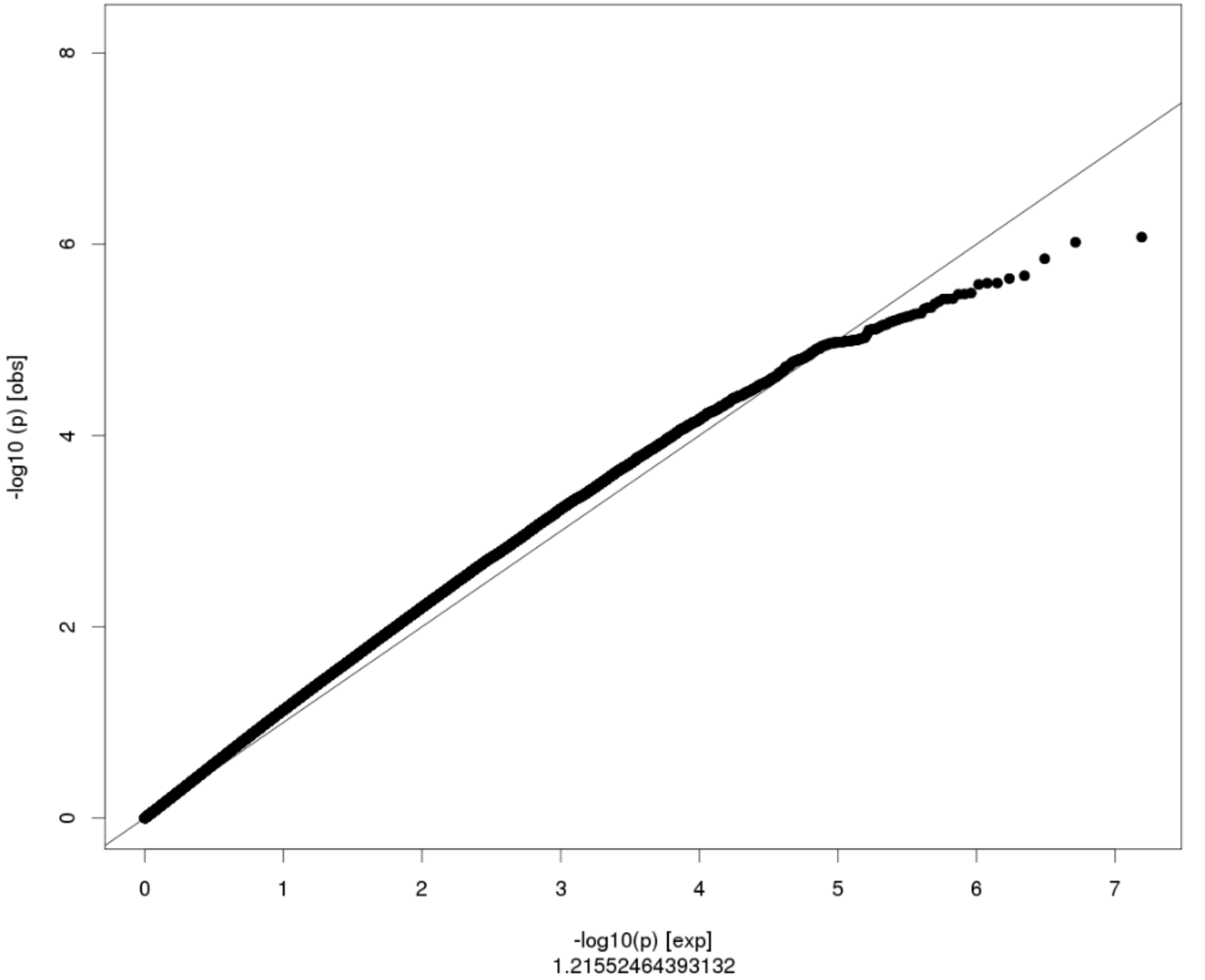


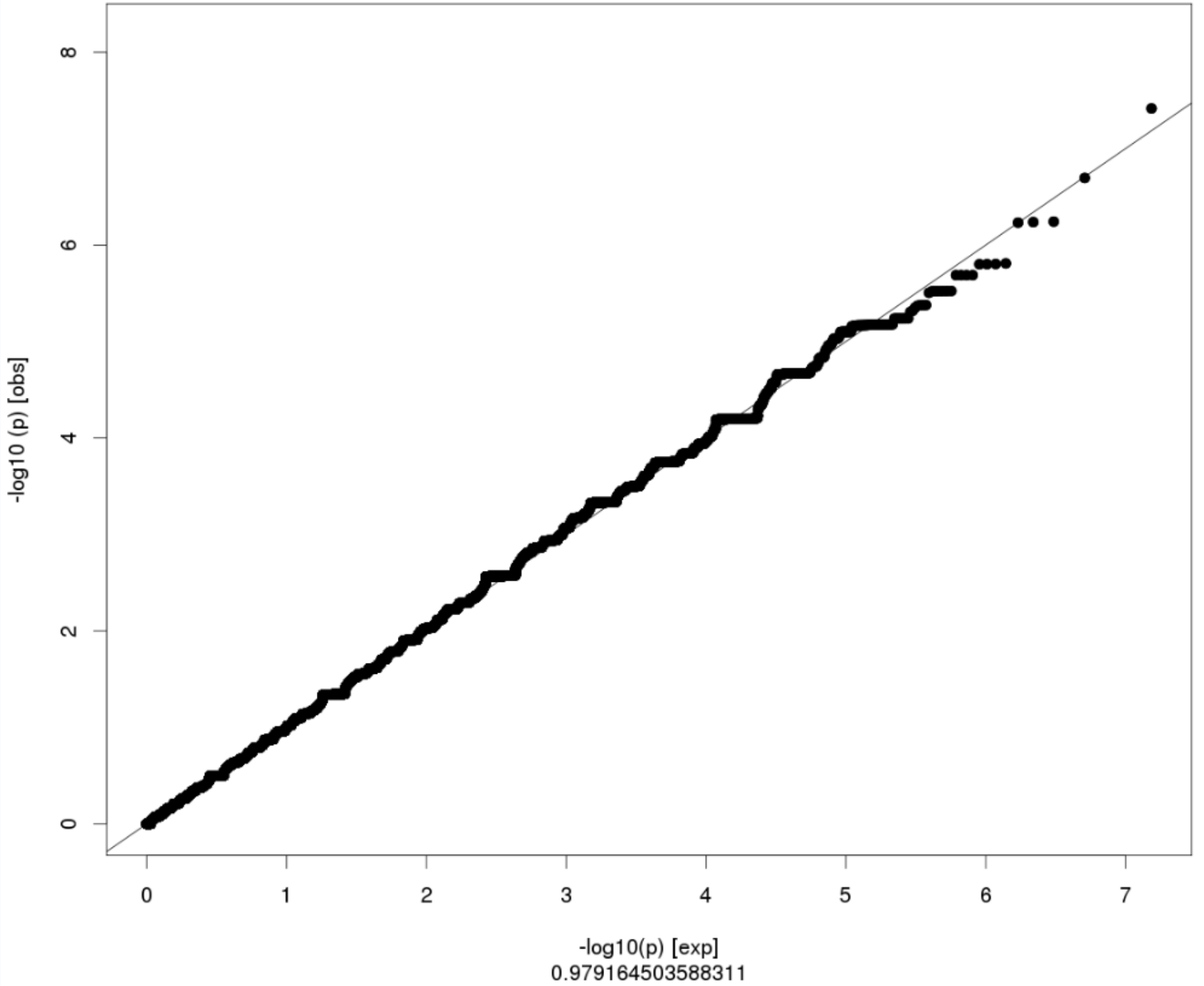
Manhattan plot

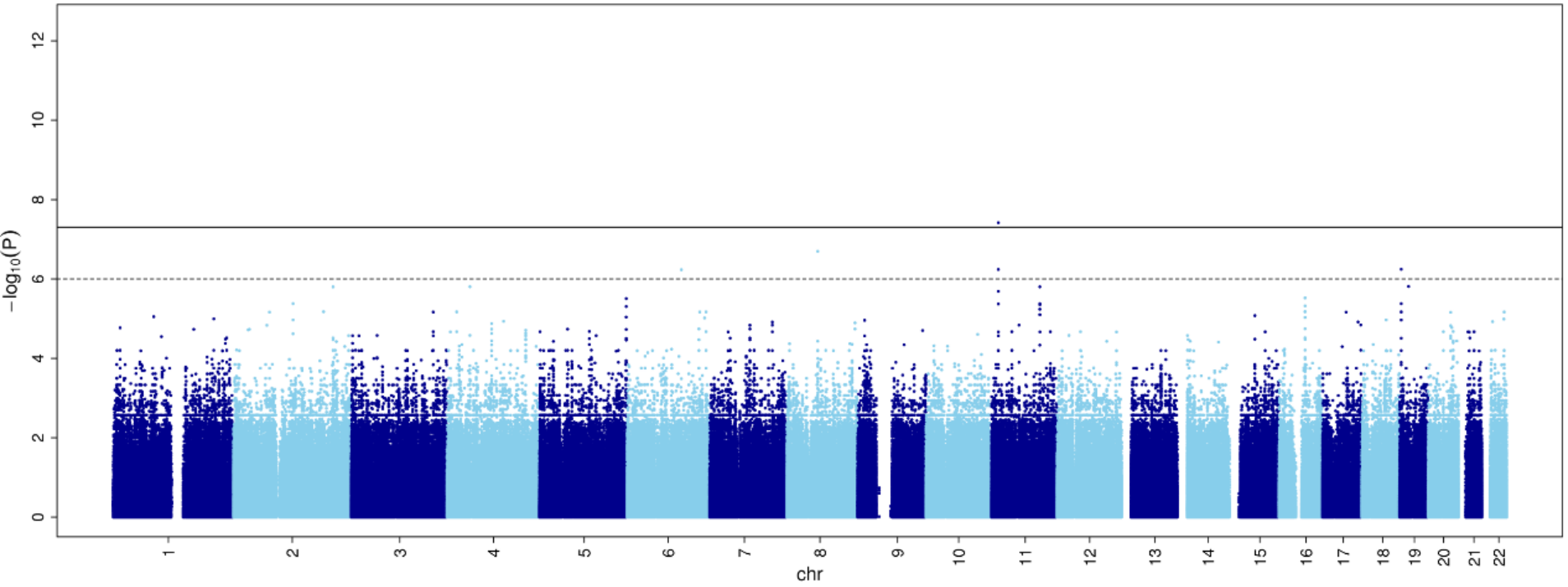
- Plots the -log10 of the association p-value for each SNP against the genomic coordinates.
- The strongest associations will have the smallest p-values and the –log10 of these p-values will have the highest height in the plot.

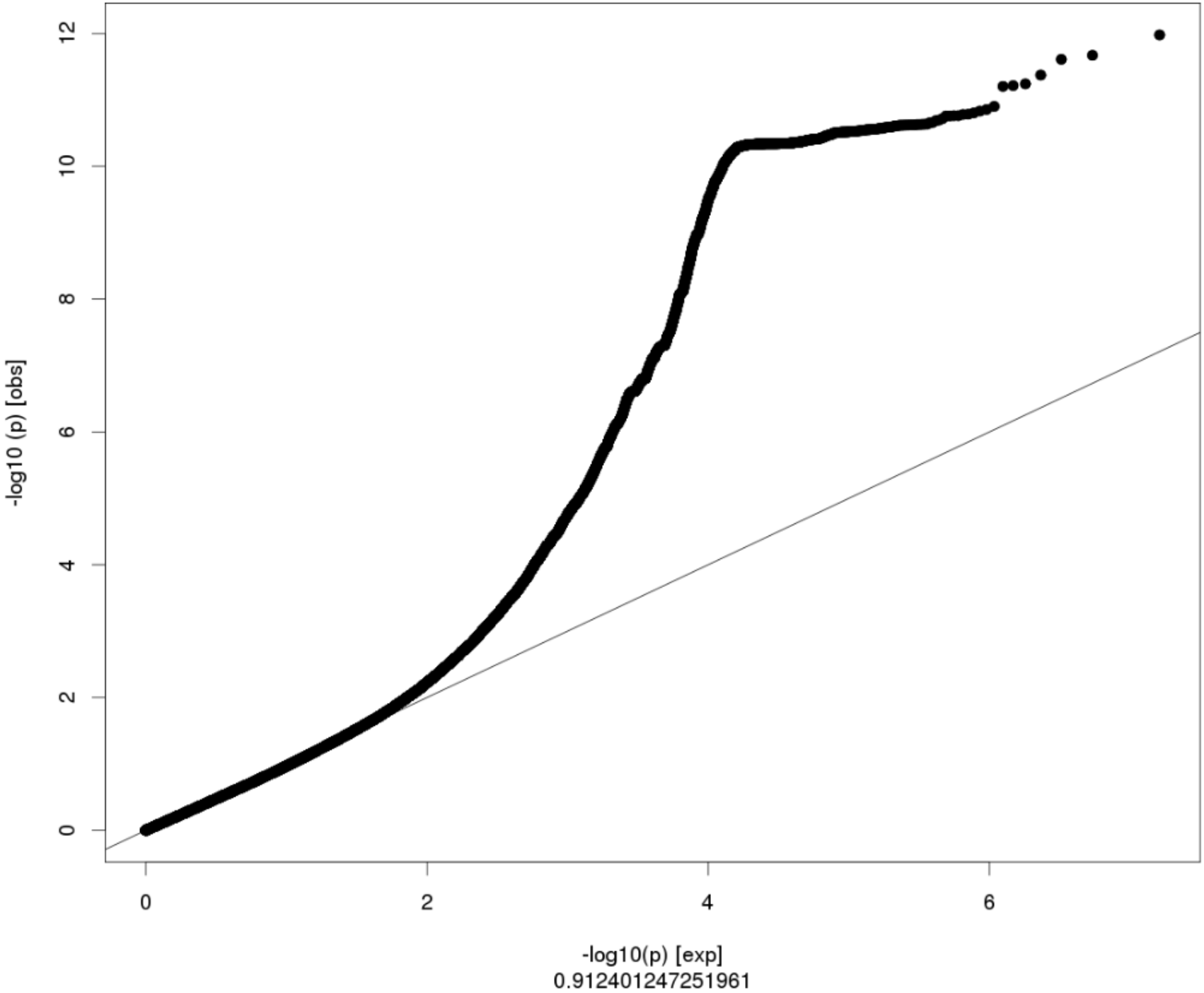


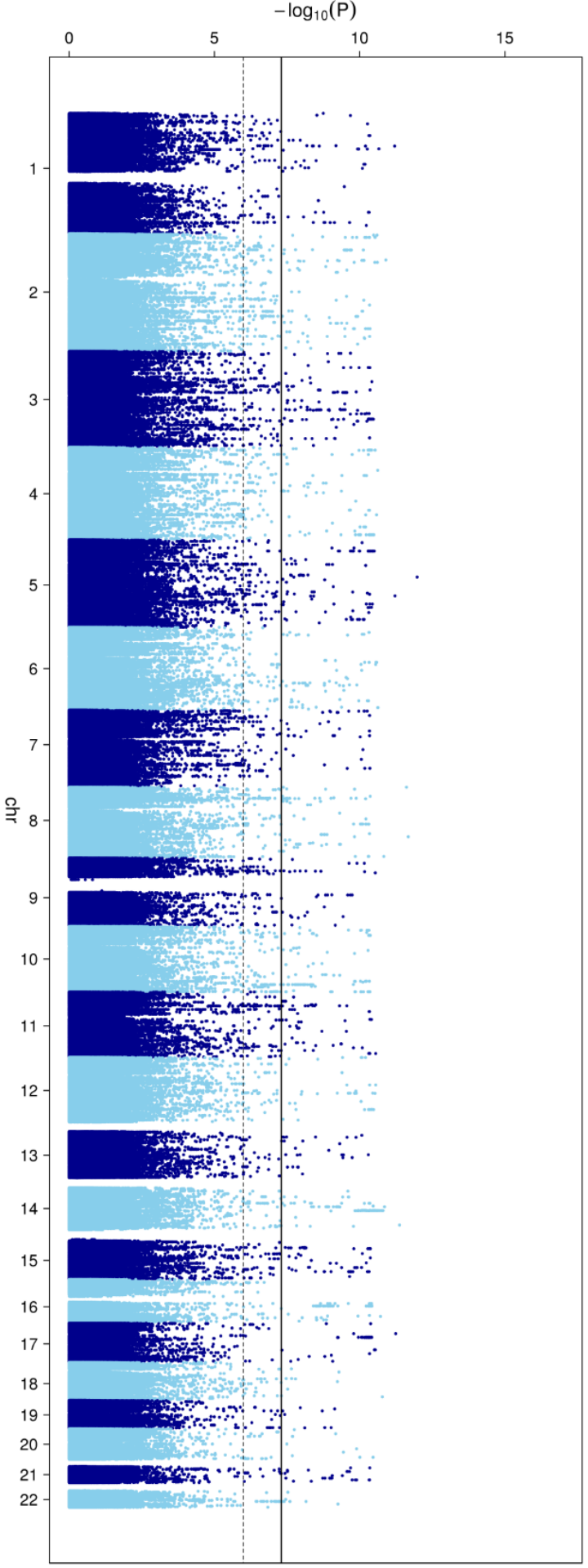
Some examples of problematic QQ and Manhattan plots


-log10 (p) [obs]
0
2
4
6
8
0
1
2
3
4
5
6
7
-log10(p) [exp]
1.21552464393132


-log10 (p) [obs]
0
2
4
6
8
0
1
2
3
4
5
6
7
-log10(p) [exp]
0.979164503588311


−log10(P)
0
2
4
6
8
10
12
chr
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22


-log10 (p) [obs]
0
2
4
6
8
10
12
0
2
4
6
-log10(p) [exp]
0.912401247251961


−log10(P)
0
5
10
15
chr
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

qqman package in R to build Manhattan and QQ plots

Usage

```
manhattan(x, chr = "CHR", bp = "BP", p = "P", snp = "SNP",
  col = c("gray10", "gray60"), chrlabs = NULL,
  suggestiveline = -log10(1e-05), genomewideline = -log10(5e-08),
  highlight = NULL, logp = TRUE, annotatePval = NULL,
  annotateTop = TRUE, ...)
```

Usage

```
qq(pvector, ...)
```

`x`	A data.frame with columns "BP," "CHR," "P," and optionally, "SNP."
`chr`	A string denoting the column name for the chromosome. Defaults to PLINK's "CHR." Said column must be numeric. If you have X, Y, or MT chromosomes, be sure to renumber these 23, 24, 25, etc.
`bp`	A string denoting the column name for the chromosomal position. Defaults to PLINK's "BP." Said column must be numeric.
`p`	A string denoting the column name for the p-value. Defaults to PLINK's "P." Said column must be numeric.
`snp`	A string denoting the column name for the SNP name (rs number). Defaults to PLINK's "SNP." Said column should be a character.
`col`	A character vector indicating which colors to alternate.
`chrlabs`	A character vector equal to the number of chromosomes specifying the chromosome labels (e.g., `c(1:22, "X", "Y", "MT")`).
`suggestiveline`	Where to draw a "suggestive" line. Default -log10(1e-5). Set to FALSE to disable.
`genomewideline`	Where to draw a "genome-wide sigificant" line. Default -log10(5e-8). Set to FALSE to disable.
`highlight`	A character vector of SNPs in your dataset to highlight. These SNPs should all be in your dataset.
`logp`	If TRUE, the -log10 of the p-value is plotted. It isn't very useful to plot raw p-values, but plotting the raw value could be useful for other genome-wide plots, for example, peak heights, bayes factors, test statistics, other "scores," etc.
`annotatePval`	If set, SNPs below this p-value will be annotated on the plot.
`annotateTop`	If TRUE, only annotates the top hit on each chromosome that is below the annotatePval threshold.
`...`	Arguments passed on to other plot/points functions

Exercise 2.3. Logistic regression with PC1-PC4 as covariates – plot the results

Use the script Rscript_qqMan.R

The script provided is ready to work with a file with no headers that will contain the chromosome, base-pair, and p-value of each SNP.

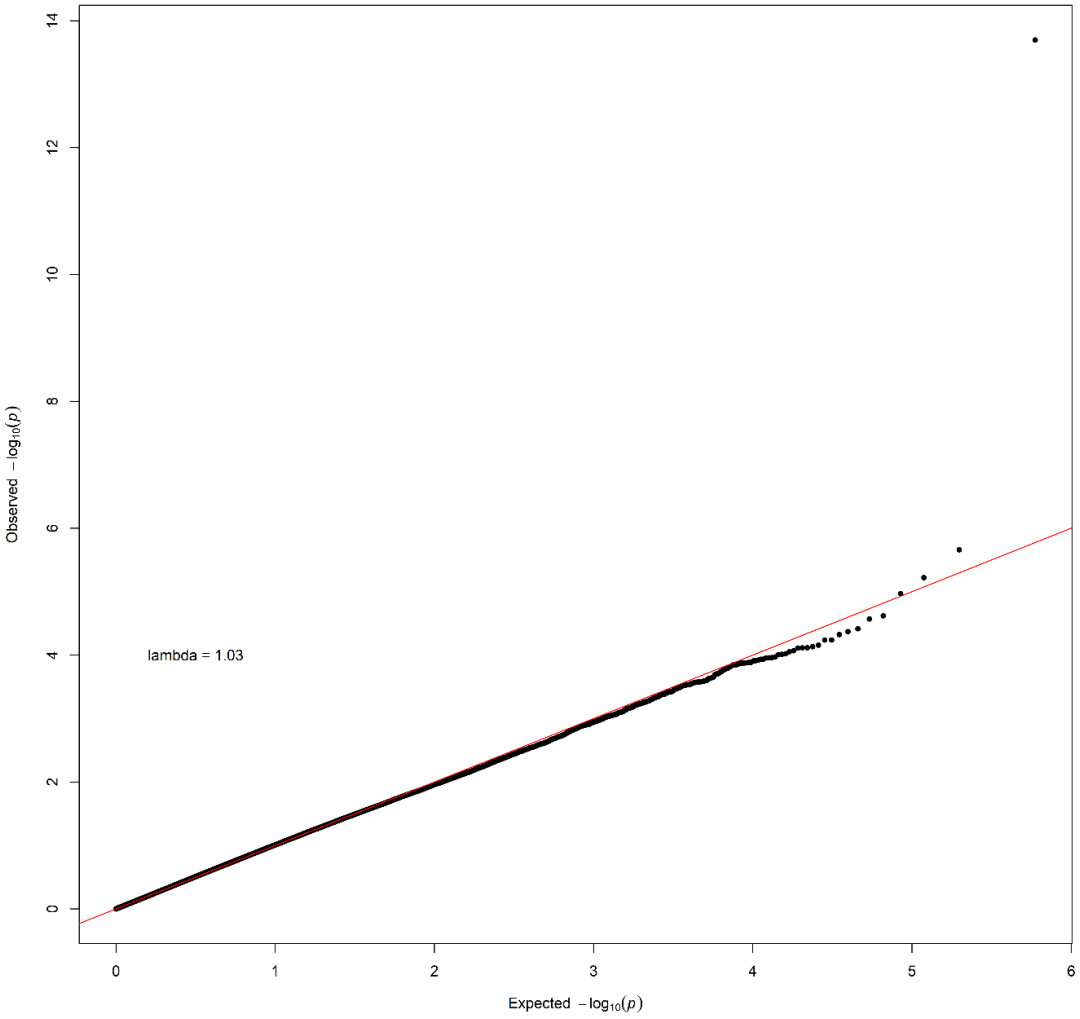
To prepare the file, do:

```
awk '{if (NR>1) print $1,$3,$9}'
2.2_adclean.cc.assoc.logistic |
grep –v NA
 > plot_adclean.logistic.txt
```

plot_adclean.logistic

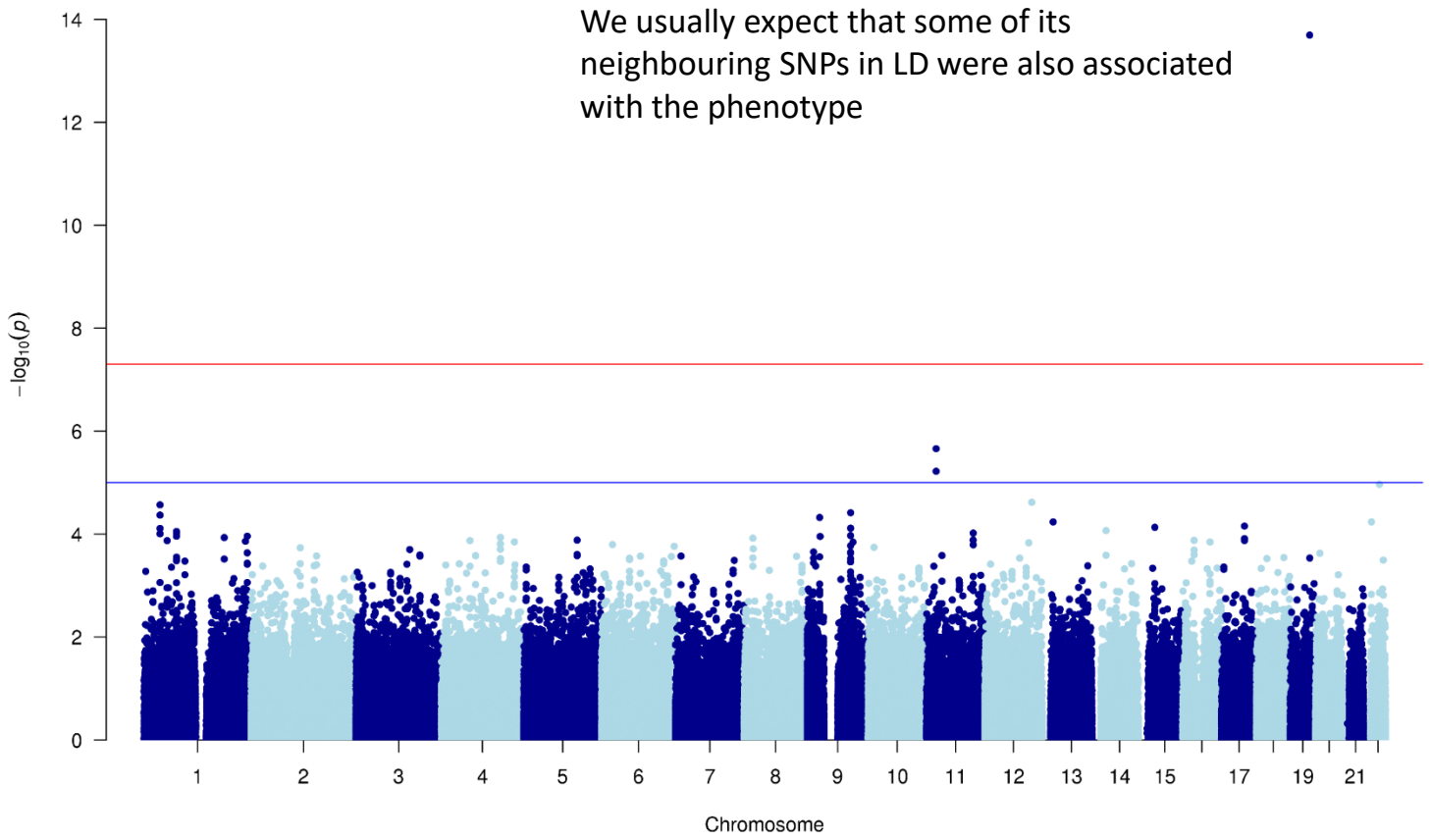
```
1 752566 0.1508
1 779322 0.2953
1 1003629 0.9347
1 1097335 0.1125
1 1130727 0.2879
```

QQ plot



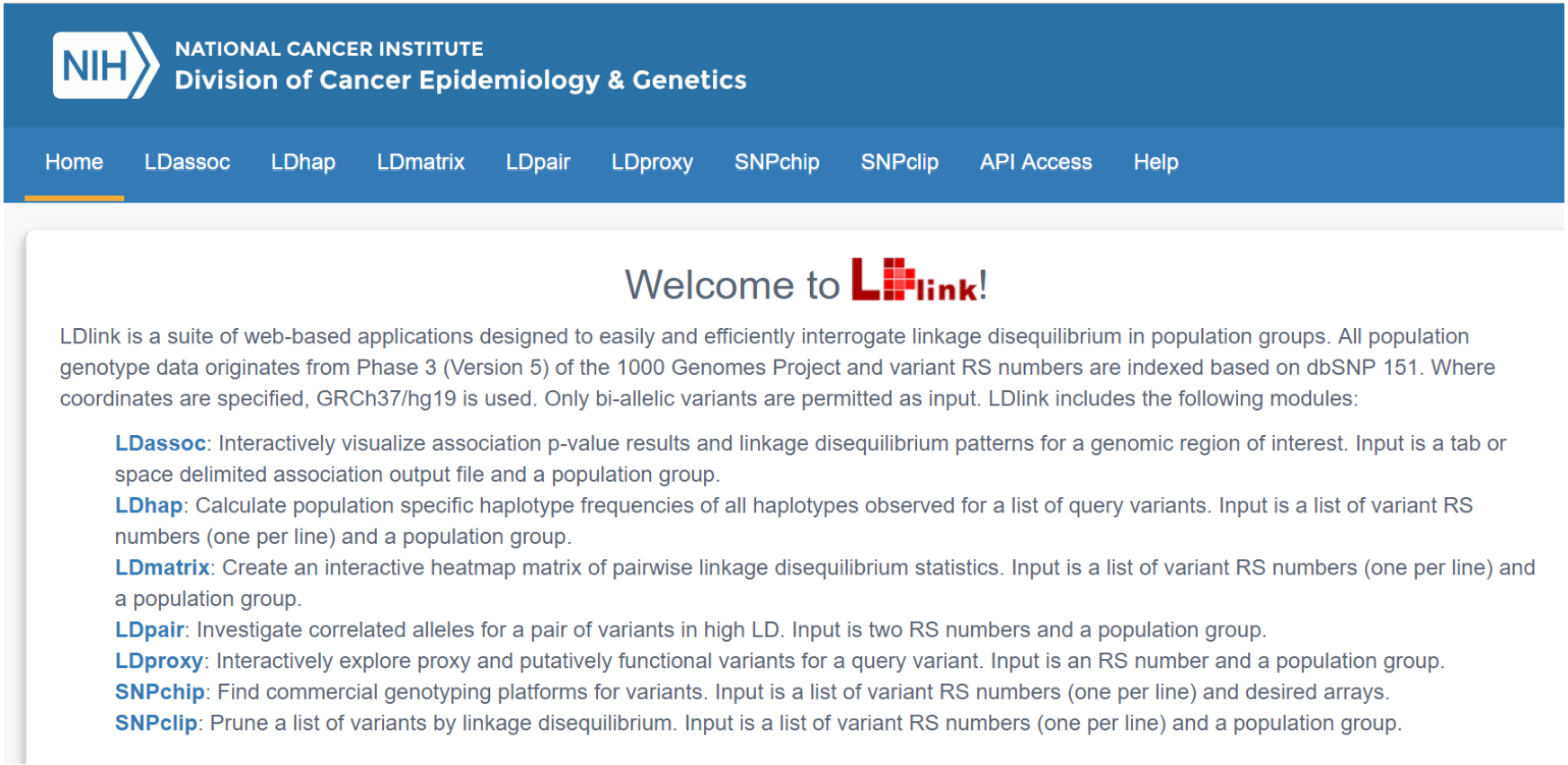
Logistic regression with PC1-PC4 as covariates

Manhattan plot



Logistic regression with PC1-PC4 as covariates

Regional plot



https://ldlink.nci.nih.gov/

Exercise 2.3. Logistic regression with PC1-PC4 as covariates – plot the results

To prepare the file, do:

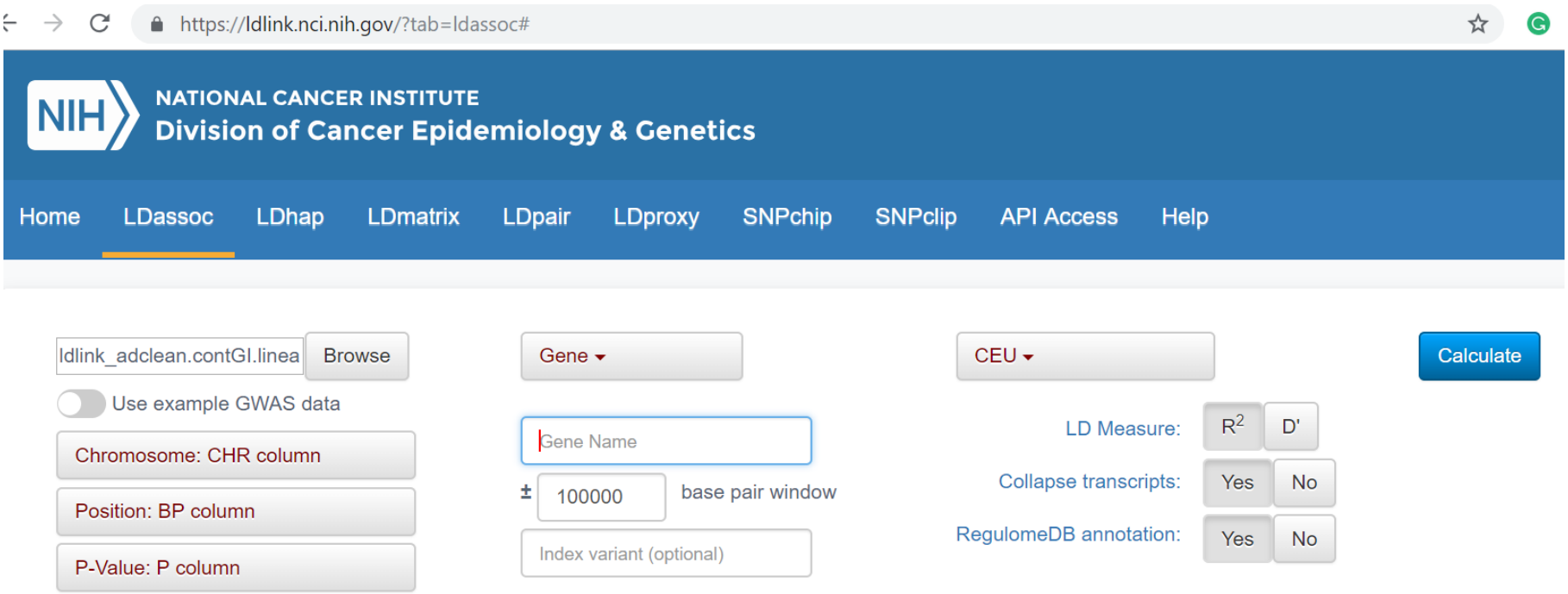
```
awk '{if (NR==1 || $1==19) print
$1,$3,$2,$9}'
2.2_adclean.cc.assoc.logistic |
grep -v NA >
ld19.adclean.cc.logistic.txt
```

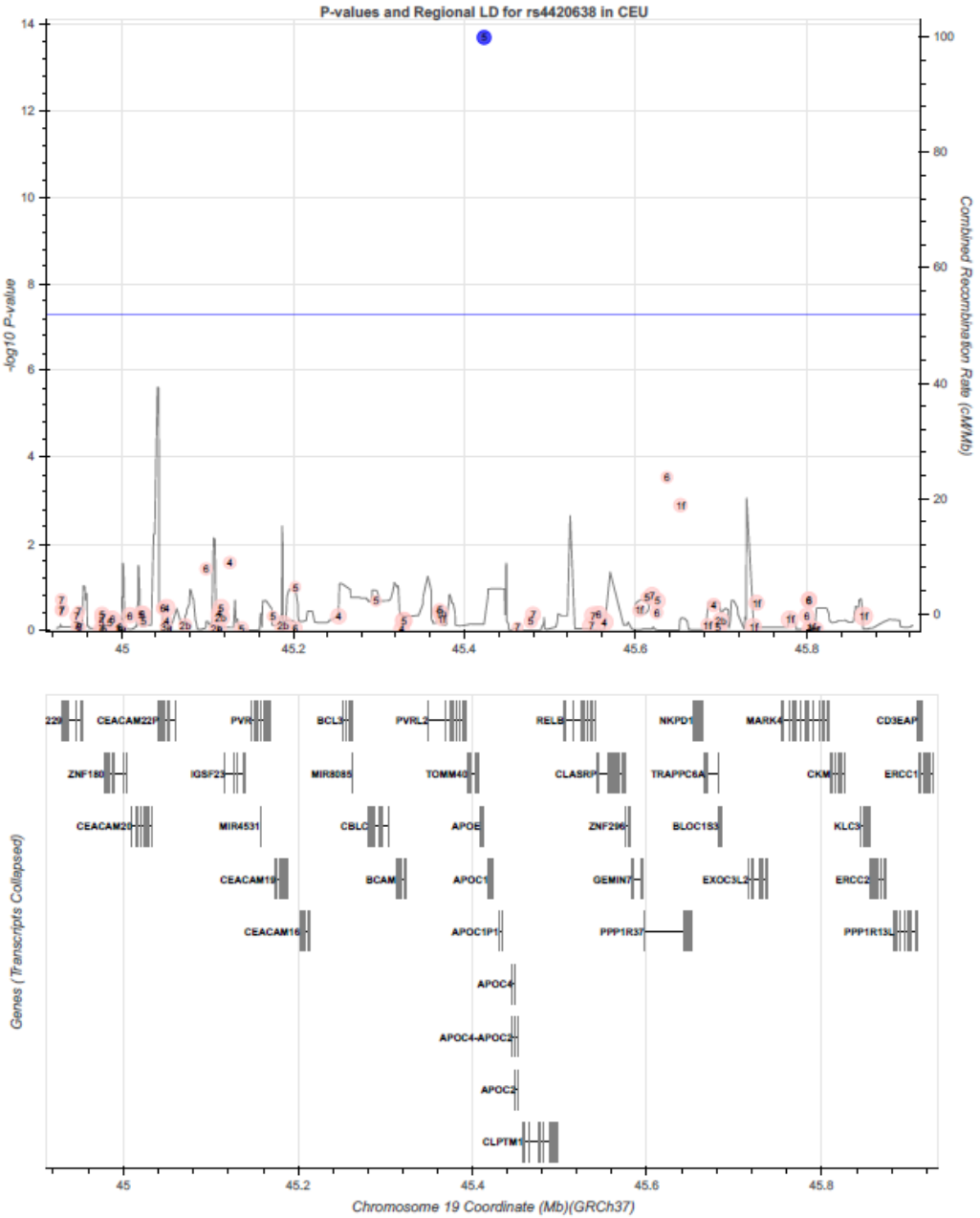
CHR BP SNP P
19 261033 rs8105536 0.6398
19 277776 rs11084928 0.1609
19 293934 rs1106581 0.8085
19 301639 rs4897940 0.7669

To find the name of the SNP, do:

```
sort -k4 -r
ld19.adclean.cc.logistic.txt | head
```

rs4420638


https://ldlink.nci.nih.gov/?tab=ldassoc#
NIH
NATIONAL CANCER INSTITUTE
Division of Cancer Epidemiology & Genetics
Home
LDassoc
LDhap
LDmatrix
LDpair
LDproxy
SNPchip
SNPclip
API Access
Help
ldlink_adclean.contGI.linea
Browse
Use example GWAS data
Chromosome: CHR column
Position: BP column
P-Value: P column
Gene
Gene Name
± 100000 base pair window
Index variant (optional)
CEU
LD Measure: R2 D'
Collapse transcripts: Yes No
RegulomeDB annotation: Yes No
Calculate


P-values and Regional LD for rs4420638 in CEU
-log10 P-value
Combined Recombination Rate (cM/Mb)
Genes (Transcripts Collapsed)
Chromosome 19 Coordinate (Mb)(GRCh37)
CEACAM22P
PVR
BCL3
PVRL2
RELB
NKPD1
MARK4
CD3EAP
ZNF180
IGSF23
MIR8085
TOMM40
CLASRP
TRAPPC6A
CKM
ERCC1
CEACAM20
MIR4531
CBLC
APOE
ZNF296
BLOC1S3
KLC3
CEACAM19
BCAM
APOC1
GEMIN7
EXOC3L2
ERCC2
CEACAM16
APOC1P1
PPP1R37
PPP1R13L
APOC4
APOC4-APOC2
APOC2
CLPTM1

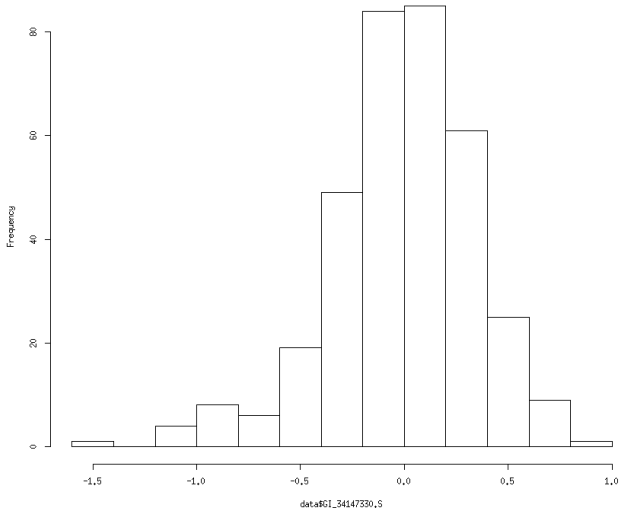
Exercise 3.1. Linear regression with PC1-PC4 as covariates, (hiding the covariates and using the --pheno command)

- Go to the continuous folder: **`cd ../continuous`**
- Have a look to the file adclean.cont.txt

Exercise 3.1. Linear regression with PC1-PC4 as covariates, (hiding the covariates and using the --pheno command)

- Go to the continuous folder: **`cd ../continuous`**
- Have a look to the file adclean.cont.txt

GI_34147330-S is a transcript probe (gene expression)



n mean sd median trimmed mad min max range skew kurtosis se
352 0 0.35 0.01 0.01 0.31 -1.43 0.8 2.24 -0.62 1.11 0.02

Exercise 3.1. Linear regression with PC1-PC4 as covariates, (hiding the covariates and using the --pheno command)

```
plink
--bfile adclean.cont
--linear hide-covar
--pheno adclean.cont.txt
--covar adpc.txt
--out 3.1_adclean.cont
```

The --pheno option allows for the specification of alternative (one or more) phenotypes.
When using the --pheno option, the original PED or FAM file must still contain a phenotype Column 6.
When using --linear, adding the option --standard-beta will first standard the phenotype (mean 0, unit variance), so the resulting coefficients will be standardized.

Exercise 3.1. Linear regression with PC1-PC4 as covariates, (hiding the covariates and using the --pheno command)

3.1_adclean.cont.log

```
Options in effect:
  --bfile adclean.cont
  --covar adpc.txt
  --linear hide-covar
  --out 3.1_adclean.cont
  --pheno adclean.cont.txt

297237 variants loaded from .bim file.
352 people (188 males, 164 females) loaded from .fam.
352 phenotype values present after --pheno.
Using 1 thread.
Warning: This run includes BLAS/LAPACK linear algebra operations which
currently disregard the --threads limit.  If this is problematic, you may
want
to recompile against single-threaded BLAS/LAPACK.
--covar: 4 covariates loaded.
Before main variant filters, 352 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.985941.
297237 variants and 352 people pass filters and QC.
Phenotype data is quantitative.
Writing linear model association results to 3.1_adclean.cont.assoc.linear
```

Exercise 3.1. Linear regression with PC1-PC4 as covariates, (hiding the covariates and using the --pheno command)

3.1_adclean.cont.assoc.linear

CHR	SNP	BP	A1	TEST	NMISS	BETA	STAT	P
1	rs3094315	752566	C	ADD	349	-0.007604	-0.1616	0.8717
1	rs4040617	779322	G	ADD	351	0.01192	0.228	0.8198
1	rs4075116	1003629	G	ADD	351	0.01296	0.3529	0.7244
1	rs9442385	1097335	T	ADD	350	0.0141	0.231	0.8174
1	rs10907175	1130727	C	ADD	343	-0.05706	-0.8978	0.3699
1	rs6603781	1158631	T	ADD	344	-0.04046	-0.8216	0.4119
1	rs11260562	1165310	T	ADD	352	-0.05212	-0.6232	0.5336
1	rs6685064	1211292	T	ADD	348	-0.07093	-0.9571	0.3392
1	rs307378	1268847	T	ADD	349	0.07568	0.7111	0.4775

Exercise 3.2. Linear regression with PC1-PC4 as covariates – plot the results

Exercise 3.2. Linear regression with PC1-PC4 as covariates – plot the results

Use the script Rscript_qqMan.R

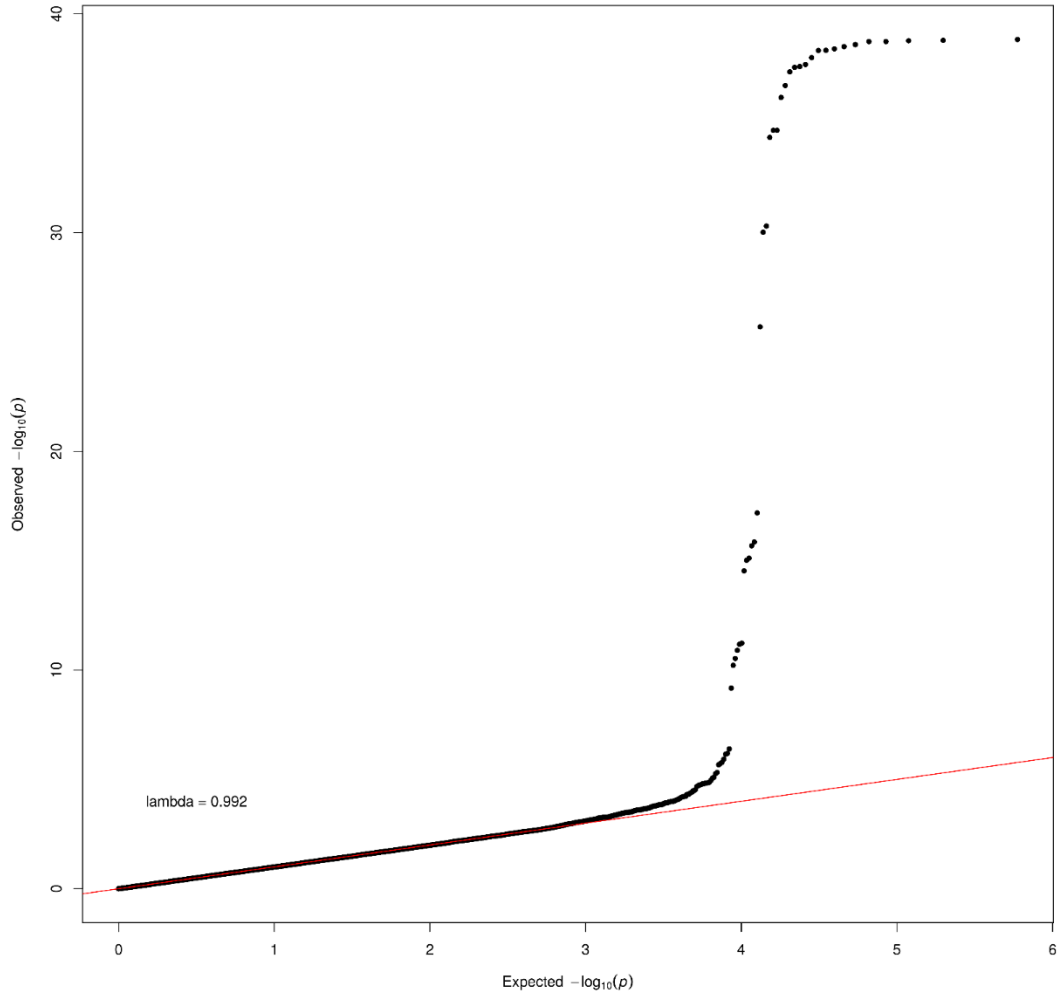
The script provided is ready to work with a file with no headers that will contain the chromosome, base-pair, and p-value of each SNP.

To prepare the file, do:

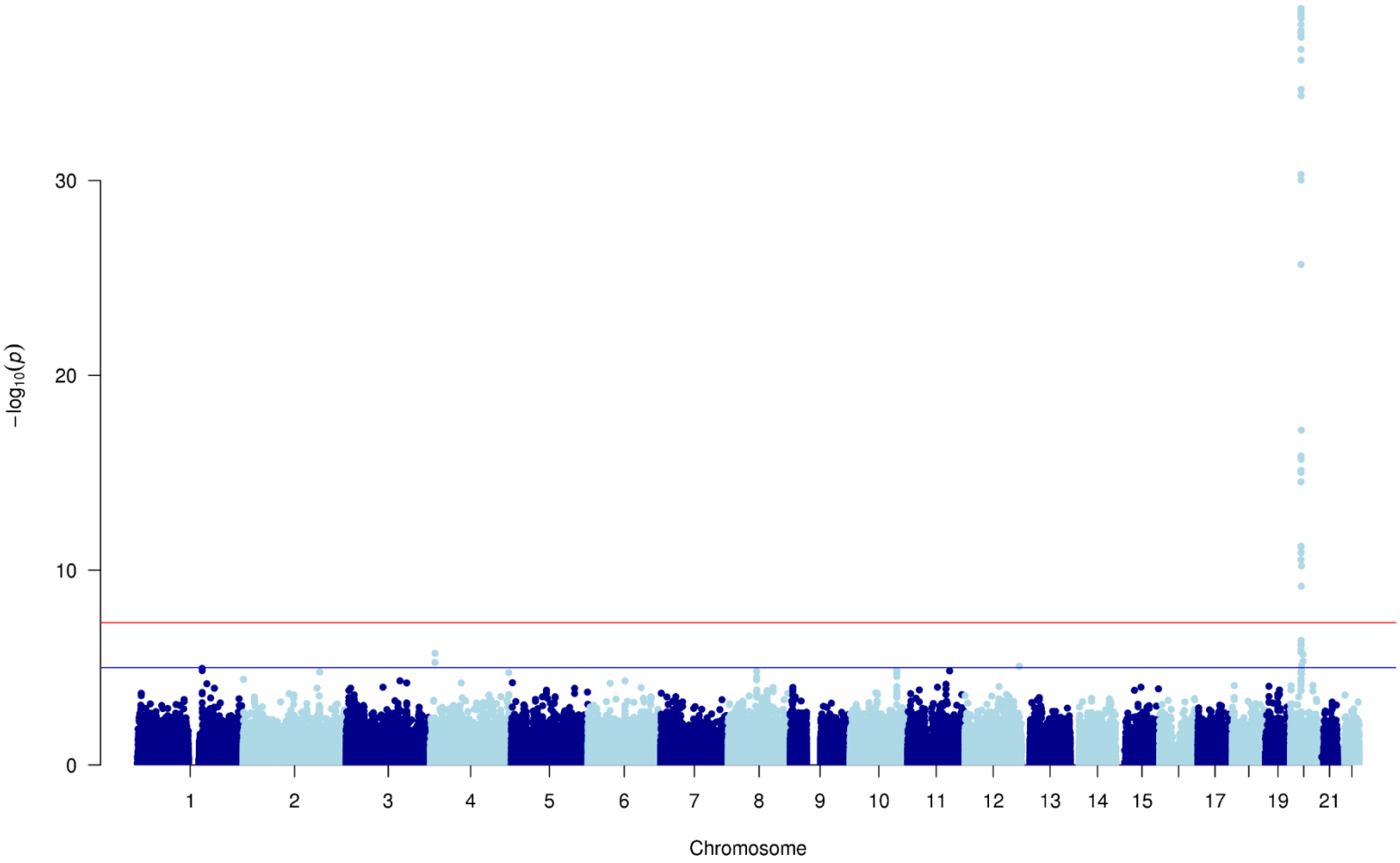
```
awk '{if (NR>1) print $1,$3,$9}'
3.1_adclean.cont.assoc.linear | grep –v NA
 > plot.adclean.cont.linear.txt
```

plot.adclean.cont.linear.txt

```
1 752566 0.1176
1 779322 0.1008
1 1003629 0.5088
1 1097335 0.7746
1 1130727 0.885
```



Linear regression with PC1-PC4 as covariates



Linear regression with PC1-PC4 as covariates

Exercise 2.3. Logistic regression with PC1-PC4 as covariates – plot the results

To prepare the file, do:

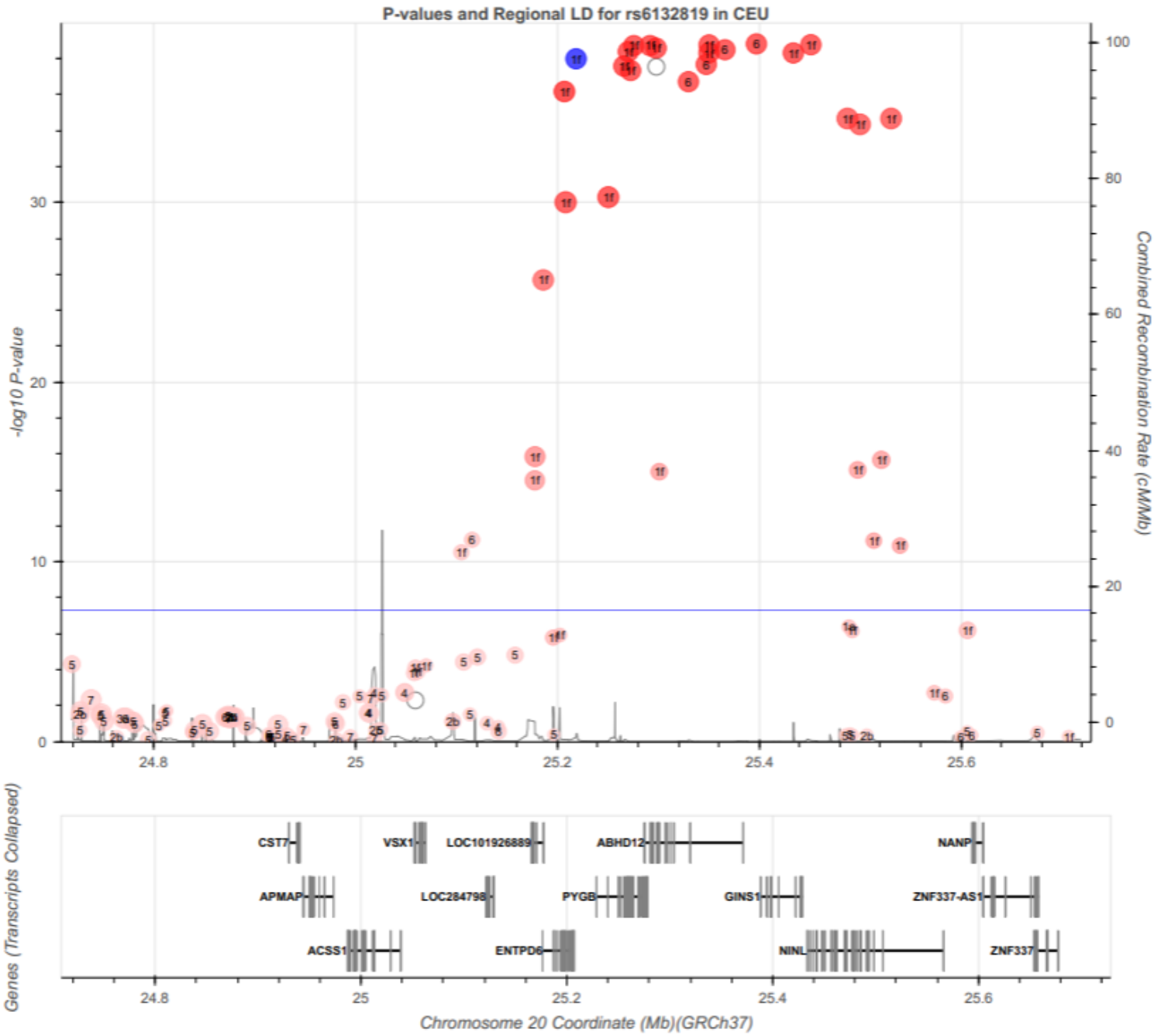
```
awk '{if (NR==1 || $1==19) print
$1,$3,$2,$9}'
2.2_adclean.cont.assoc.linear |
grep -v NA >
ld20.adclean.cont.linear.txt
```

CHR BP SNP P
20 69408 rs17685809 0.05855
20 109272 rs6038037 0.006237
20 134476 rs6055084 0.1093
20 138125 rs2298108 0.3534
20 138148 rs2298109 0.3534
20 138460 rs6077288 0.3176

To find the name of the SNP, do:

```
sort -k4 -r
ld20.adclean.cont.linear.txt | head
```

rs6132819


P-values and Regional LD for rs6132819 in CEU
-log10 P-value
Combined Recombination Rate (cM/Mb)
0
10
20
30
0
20
40
60
80
100
24.8
25
25.2
25.4
25.6
Genes (Transcripts Collapsed)
CST7
VSX1
LOC101926889
ABHD12
NANP
APMAP
LOC284798
PYGB
GINS1
ZNF337-AS1
ACSS1
ENTPD6
NINL
ZNF337
Chromosome 20 Coordinate (Mb)(GRCh37)

Exercise 3.3. Linear regression with PC1 as covariate, (hiding the covariates and using the --pheno command)

```
plink
--bfile adclean.cont
--linear  hide-covar
--pheno adclean.cont.txt
--covar adpc.txt
--covar-name PC1
--out 3.3_adclean.cont
```

Exercise 3.3. Linear regression with PC1 as covariate, (hiding the covariates and using the --pheno command)

3.3_adclean.cont.log

```
Options in effect:
  --bfile adclean.cont
  --covar adpc.txt
  --covar-name PC1
  --linear hide-covar
  --out 3.3_adclean.cont
  --pheno adclean.cont.txt

64148 MB RAM detected; reserving 32074 MB for main workspace.
297237 variants loaded from .bim file.
352 people (188 males, 164 females) loaded from .fam.
352 phenotype values present after --pheno.
Using 1 thread.
Warning: This run includes BLAS/LAPACK linear algebra operations which
currently disregard the --threads limit.  If this is problematic, you may want
to recompile against single-threaded BLAS/LAPACK.
--covar: 1 out of 4 covariates loaded.
Before main variant filters, 352 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.985941.
297237 variants and 352 people pass filters and QC.
Phenotype data is quantitative.
Writing linear model association results to 3.3_adclean.cont.assoc.linear ...
```

Exercise 3.3. Linear regression with PC1 as covariate, (hiding the covariates and using the --pheno command)

3.3_adclean.cont.assoc.linear

CHR	SNP	BP	A1	TEST	NMISS	BETA	STAT	P
1	rs3094315	752566	C	ADD	349	-0.0549	-1.517	0.1303
1	rs4040617	779322	G	ADD	351	-0.06181	-1.538	0.1249
1	rs4075116	1003629	G	ADD	351	-0.01604	-0.5652	0.5723
1	rs9442385	1097335	T	ADD	350	-0.0161	-0.3403	0.7338
1	rs10907175	1130727	C	ADD	343	0.001747	0.03535	0.9718
1	rs6603781	1158631	T	ADD	344	-0.03843	-1.028	0.3046
1	rs11260562	1165310	T	ADD	352	-0.07412	-1.155	0.2488
1	rs6685064	1211292	T	ADD	348	0.0268	0.473	0.6365
1	rs307378	1268847	T	ADD	349	0.07615	0.9301	0.353