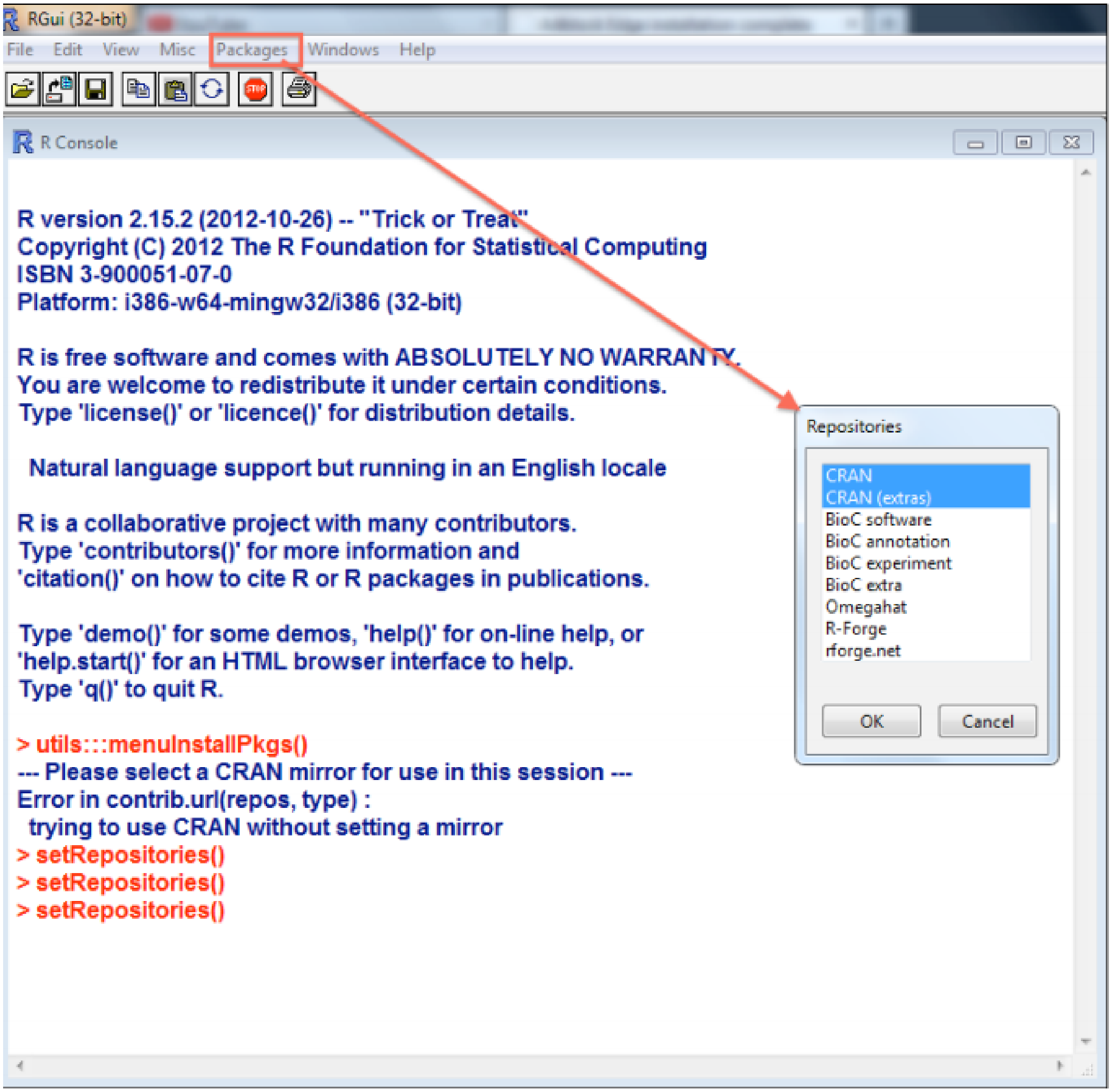
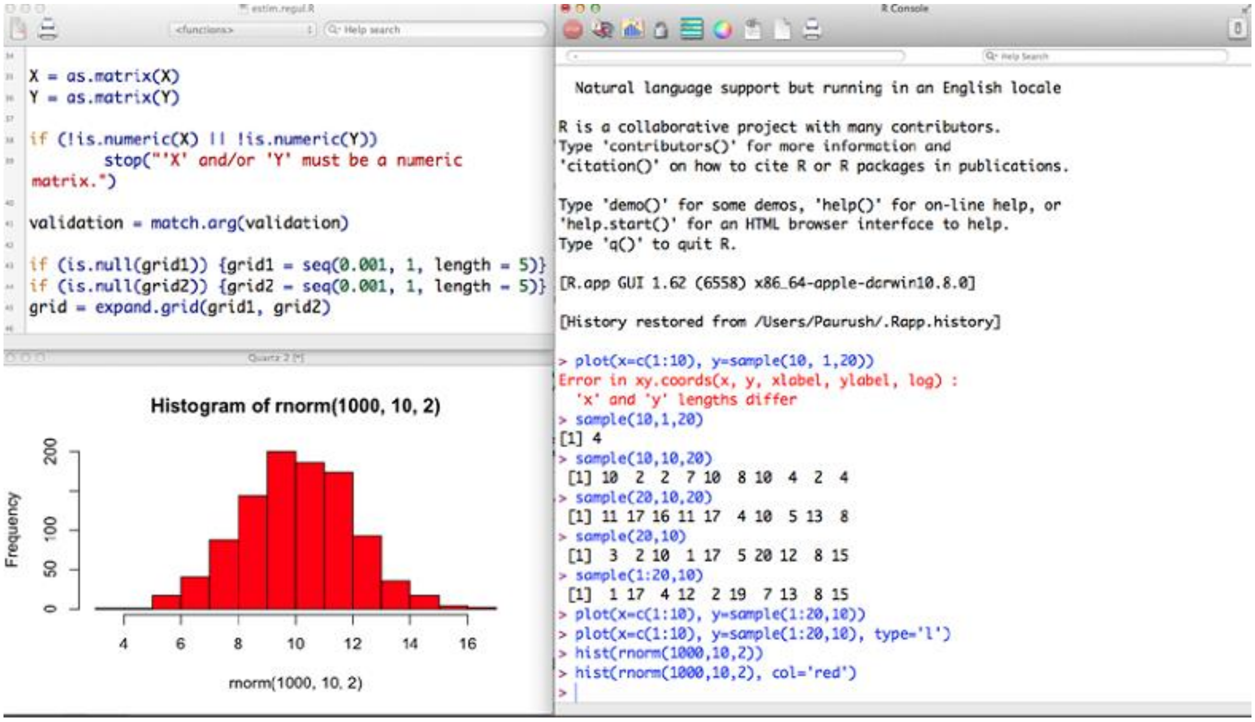
Chapter 1: Starting Bioinformatics with R

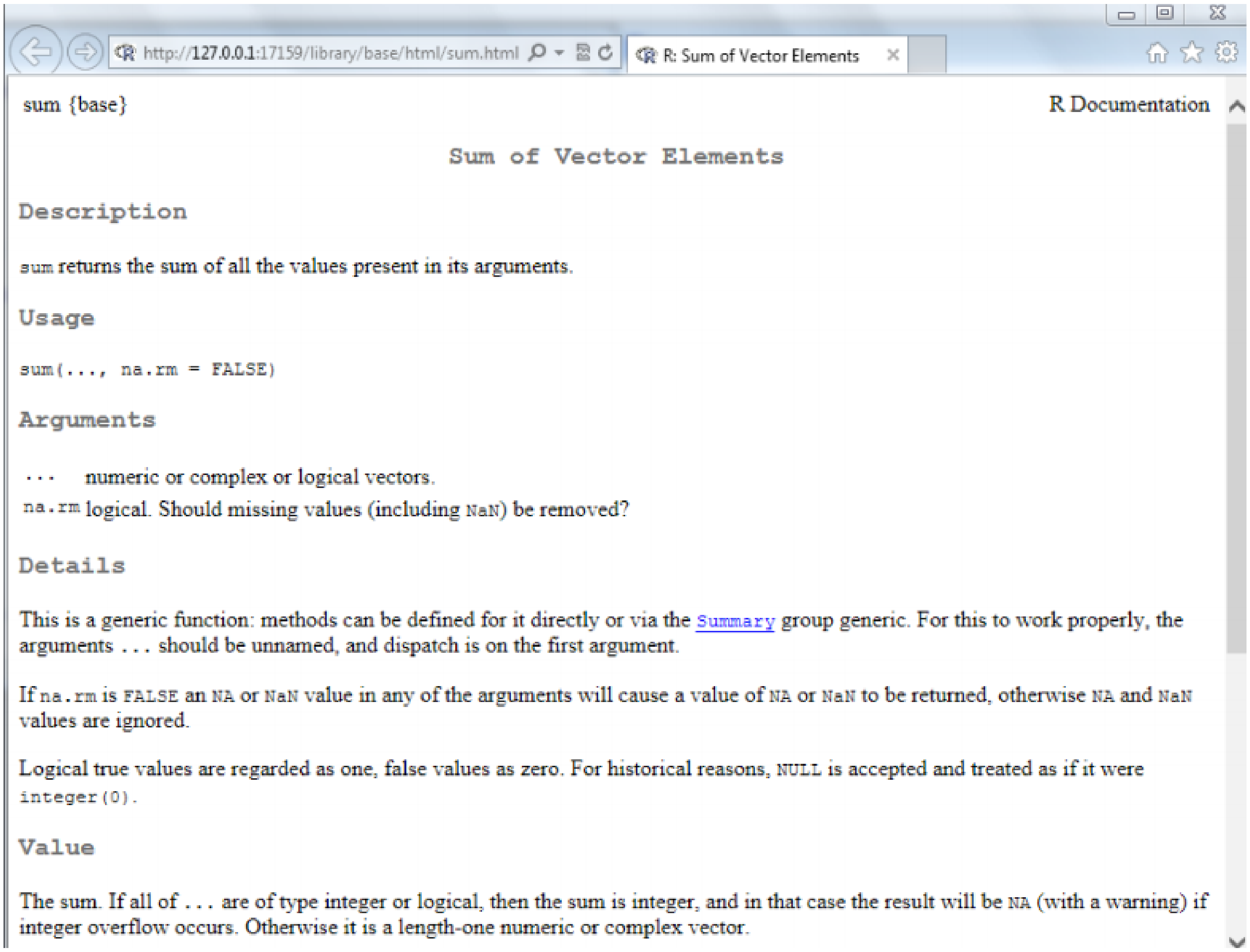


```
R version 2.15.2 (2012-10-26) -- "Trick or Treat"
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> utils:::menuInstallPkgs()
--- Please select a CRAN mirror for use in this session ---
Error in contrib.url(repos, type) :
  trying to use CRAN without setting a mirror
> setRepositories()
> setRepositories()
> setRepositories()
```



```r
X = as.matrix(X)
Y = as.matrix(Y)

if (!is.numeric(X) || !is.numeric(Y))
        stop("'X' and/or 'Y' must be a numeric 
matrix.")

validation = match.arg(validation)

if (is.null(grid1)) {grid1 = seq(0.001, 1, length = 5)}
if (is.null(grid2)) {grid2 = seq(0.001, 1, length = 5)}
grid = expand.grid(grid1, grid2)
```

```
  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.62 (6558) x86_64-apple-darwin10.8.0]

[History restored from /Users/Paurush/.Rapp.history]

> plot(x=c(1:10), y=sample(10, 1,20))
Error in xy.coords(x, y, xlabel, ylabel, log) : 
  'x' and 'y' lengths differ
> sample(10,1,20)
[1] 4
> sample(10,10,20)
 [1] 10  2  2  7 10  8 10  4  2  4
> sample(20,10,20)
 [1] 11 17 16 11 17  4 10  5 13  8
> sample(20,10)
 [1]  3  2 10  1 17  5 20 12  8 15
> sample(1:20,10)
 [1]  1 17  4 12  2 19  7 13  8 15
> plot(x=c(1:10), y=sample(1:20,10))
> plot(x=c(1:10), y=sample(1:20,10), type='l')
> hist(rnorm(1000,10,2))
> hist(rnorm(1000,10,2), col='red')
> 
```



sum {base}　　　R Documentation

Sum of Vector Elements

Description

`sum` returns the sum of all the values present in its arguments.

Usage

```
sum(..., na.rm = FALSE)
```

Arguments

`...`	numeric or complex or logical vectors.
`na.rm`	logical. Should missing values (including `NaN`) be removed?

Details

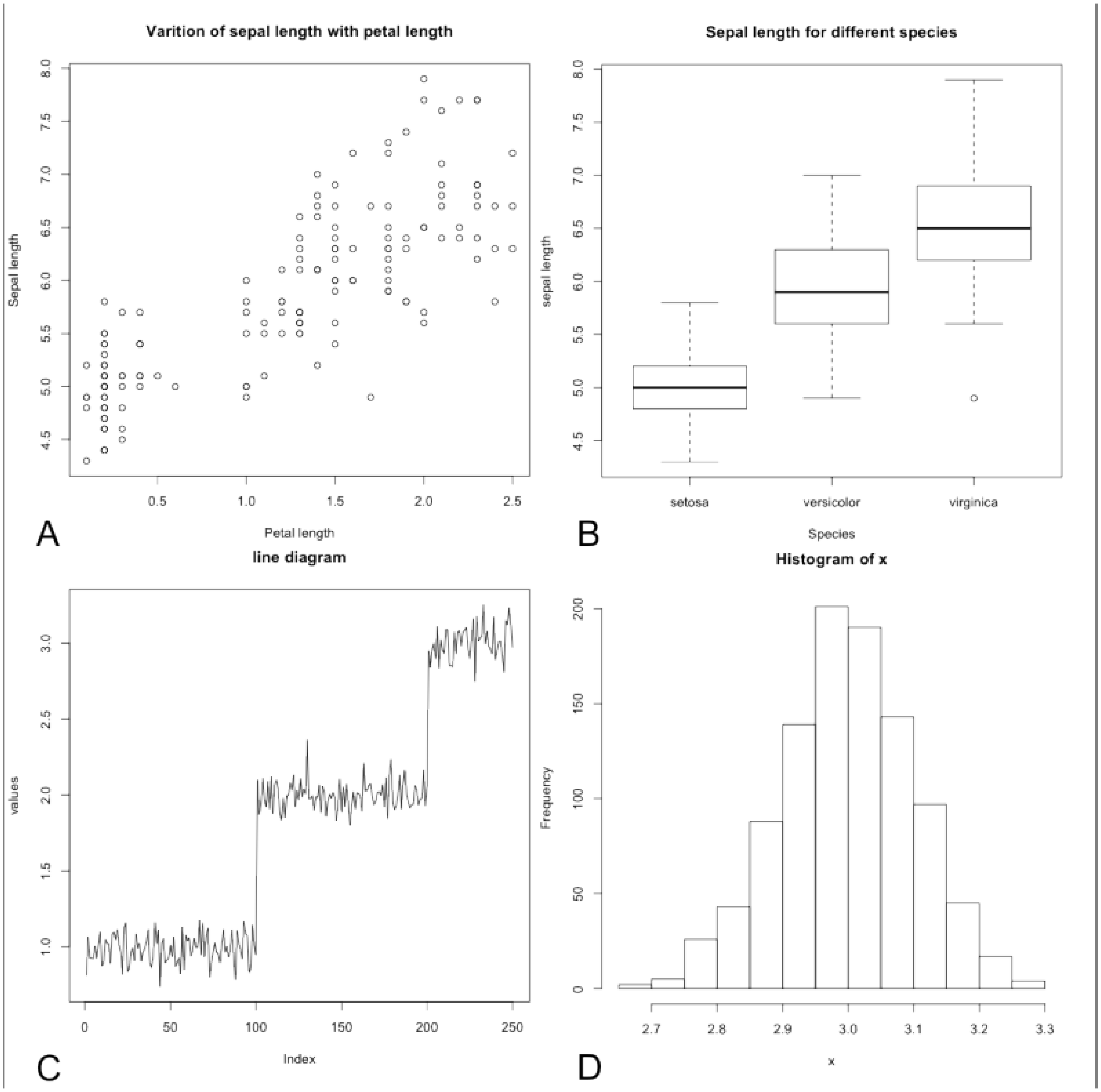
This is a generic function: methods can be defined for it directly or via the `Summary` group generic. For this to work properly, the arguments `...` should be unnamed, and dispatch is on the first argument.

If `na.rm` is `FALSE` an `NA` or `NaN` value in any of the arguments will cause a value of `NA` or `NaN` to be returned, otherwise `NA` and `NaN` values are ignored.

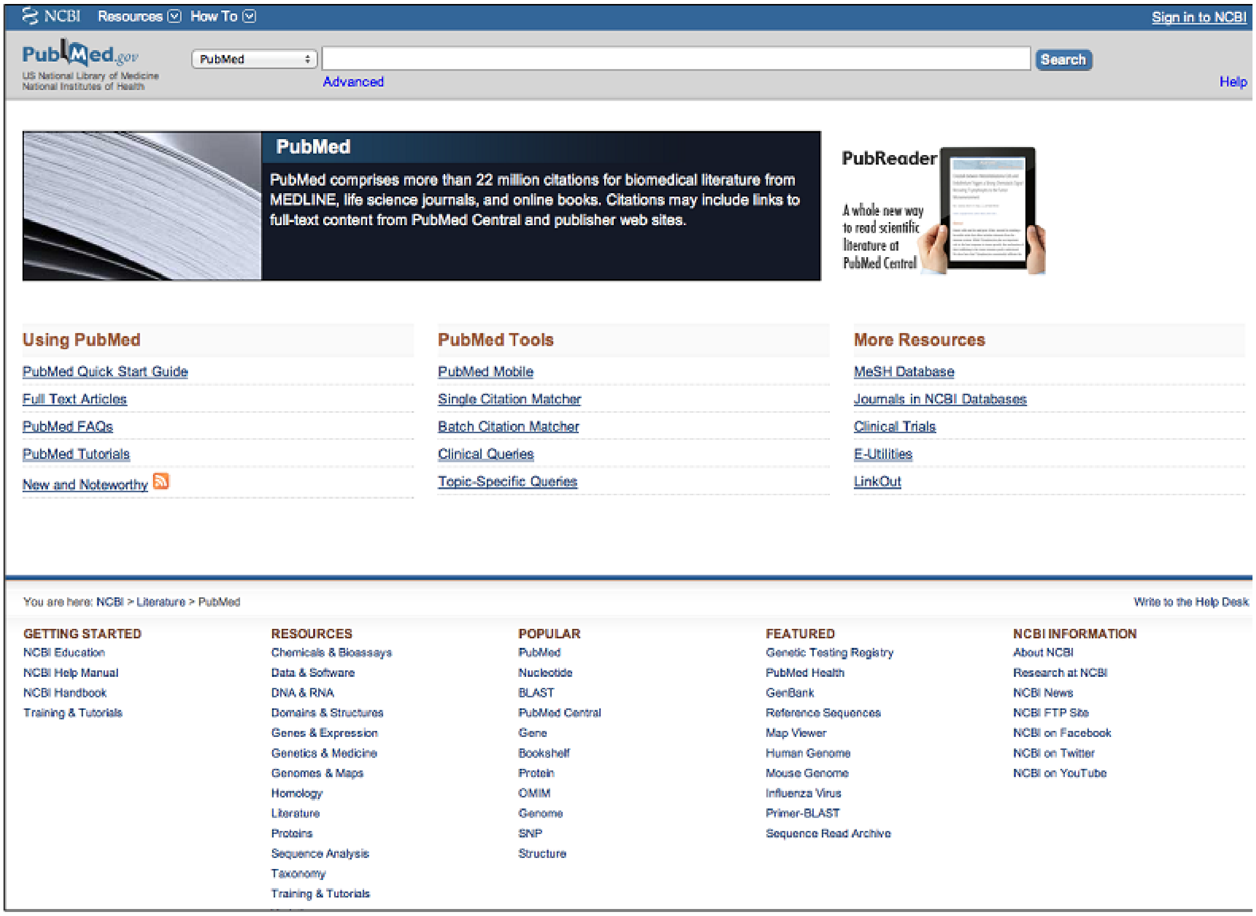
Logical true values are regarded as one, false values as zero. For historical reasons, `NULL` is accepted and treated as if it were `integer(0)`.

Value

The sum. If all of `...` are of type integer or logical, then the sum is integer, and in that case the result will be `NA` (with a warning) if integer overflow occurs. Otherwise it is a length-one numeric or complex vector.



Chapter 2: Introduction to Bioconductor


NCBI Resources How To

Sign in to NCBI

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

Search

Advanced

Help

PubMed

PubMed comprises more than 22 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

PubReader

A whole new way to read scientific literature at PubMed Central

Using PubMed

- PubMed Quick Start Guide
- Full Text Articles
- PubMed FAQs
- PubMed Tutorials
- New and Noteworthy

PubMed Tools

- PubMed Mobile
- Single Citation Matcher
- Batch Citation Matcher
- Clinical Queries
- Topic-Specific Queries

More Resources

- MeSH Database
- Journals in NCBI Databases
- Clinical Trials
- E-Utilities
- LinkOut

You are here: NCBI > Literature > PubMed

Write to the Help Desk

GETTING STARTED

- NCBI Education
- NCBI Help Manual
- NCBI Handbook
- Training & Tutorials

RESOURCES

- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials

POPULAR

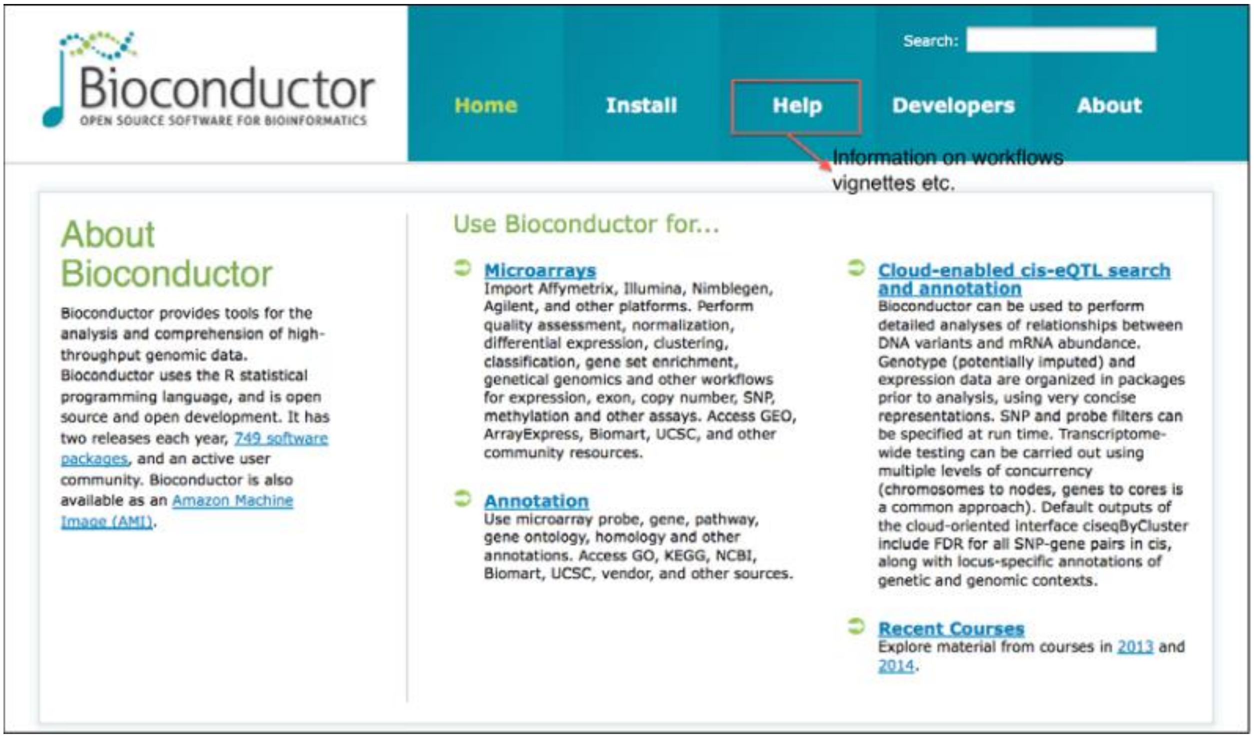
- PubMed
- Nucleotide
- BLAST
- PubMed Central
- Gene
- Bookshelf
- Protein
- OMIM
- Genome
- SNP
- Structure

FEATURED

- Genetic Testing Registry
- PubMed Health
- GenBank
- Reference Sequences
- Map Viewer
- Human Genome
- Mouse Genome
- Influenza Virus
- Primer-BLAST
- Sequence Read Archive

NCBI INFORMATION

- About NCBI
- Research at NCBI
- NCBI News
- NCBI FTP Site
- NCBI on Facebook
- NCBI on Twitter
- NCBI on YouTube


Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

Home Install Help Developers About

Information on workflows
vignettes etc.

About Bioconductor

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 749 software packages, and an active user community. Bioconductor is also available as an Amazon Machine Image (AMI).

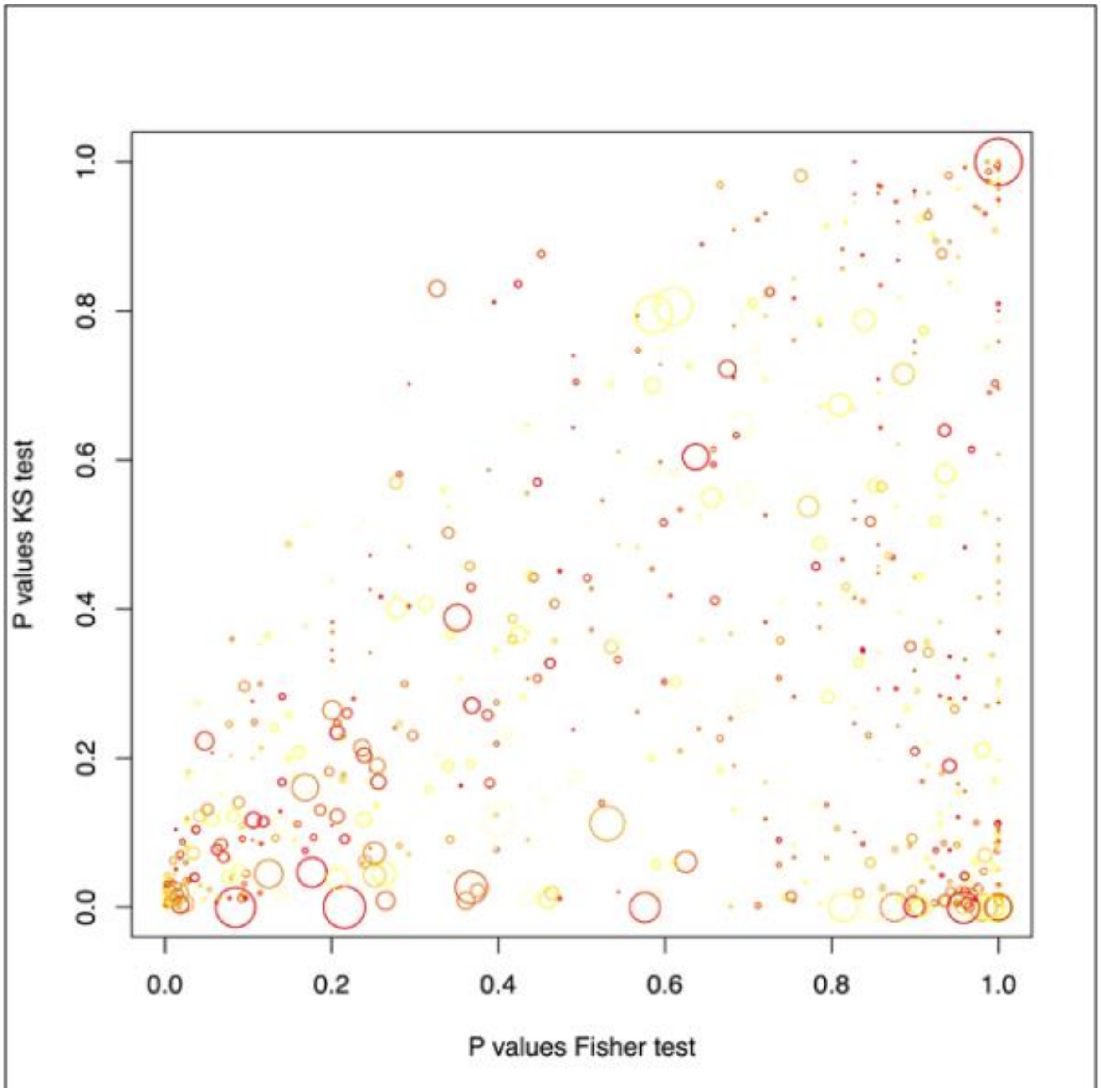
Use Bioconductor for...

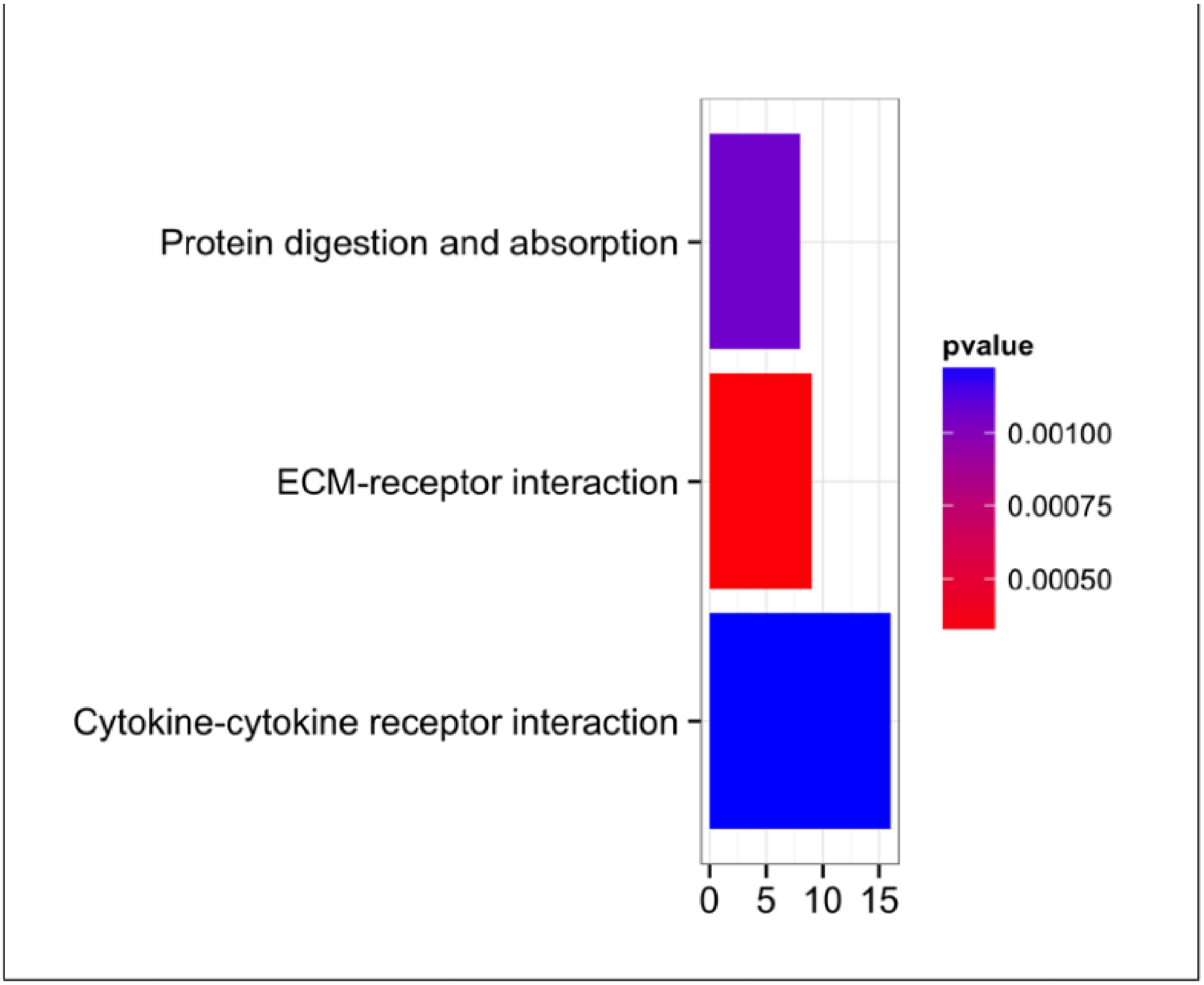
Microarrays
Import Affymetrix, Illumina, Nimblegen, Agilent, and other platforms. Perform quality assessment, normalization, differential expression, clustering, classification, gene set enrichment, genetical genomics and other workflows for expression, exon, copy number, SNP, methylation and other assays. Access GEO, ArrayExpress, Biomart, UCSC, and other community resources.

Annotation
Use microarray probe, gene, pathway, gene ontology, homology and other annotations. Access GO, KEGG, NCBI, Biomart, UCSC, vendor, and other sources.

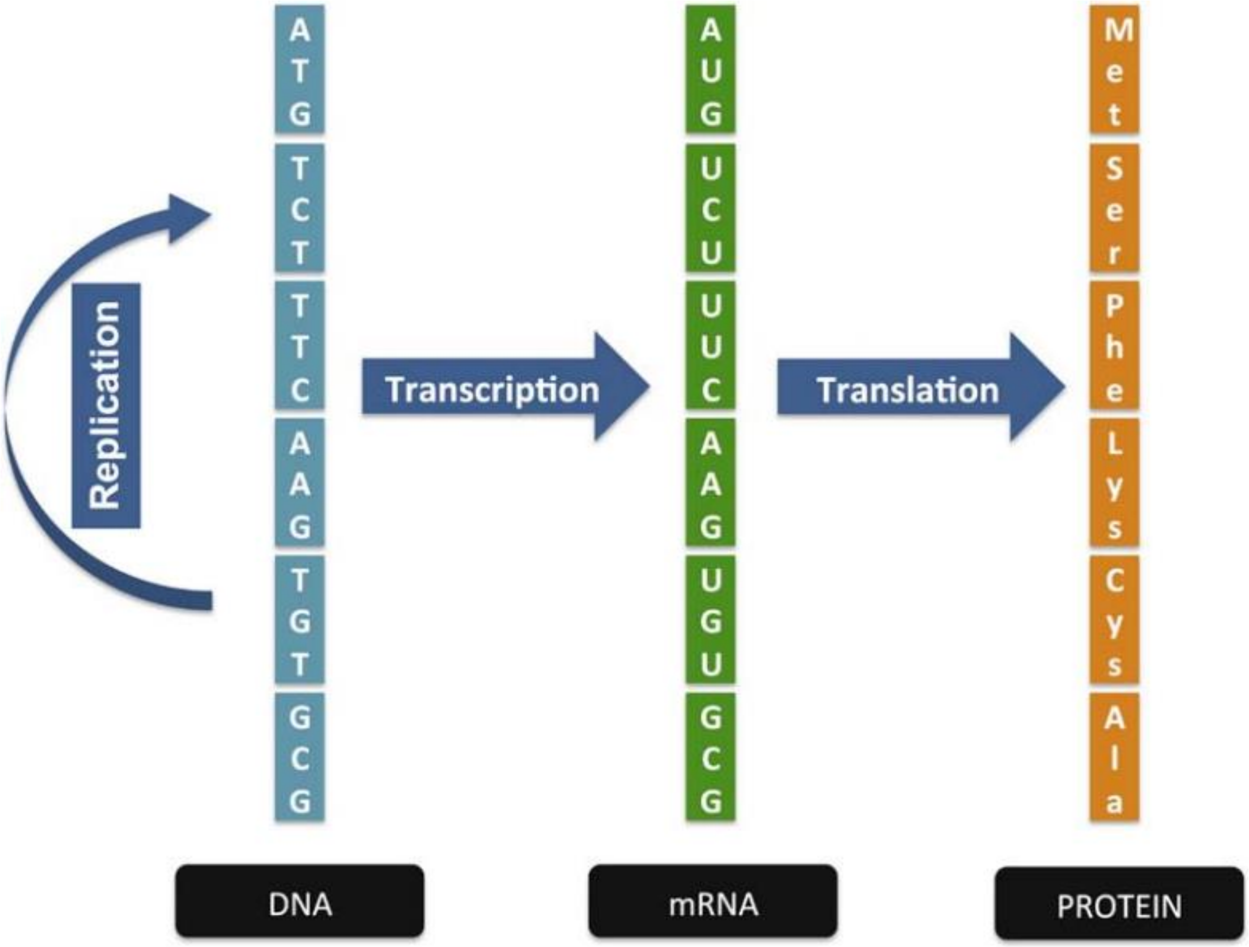
Cloud-enabled cis-eQTL search and annotation
Bioconductor can be used to perform detailed analyses of relationships between DNA variants and mRNA abundance. Genotype (potentially imputed) and expression data are organized in packages prior to analysis, using very concise representations. SNP and probe filters can be specified at run time. Transcriptome-wide testing can be carried out using multiple levels of concurrency (chromosomes to nodes, genes to cores is a common approach). Default outputs of the cloud-oriented interface ciseqByCluster include FDR for all SNP-gene pairs in cis, along with locus-specific annotations of genetic and genomic contexts.

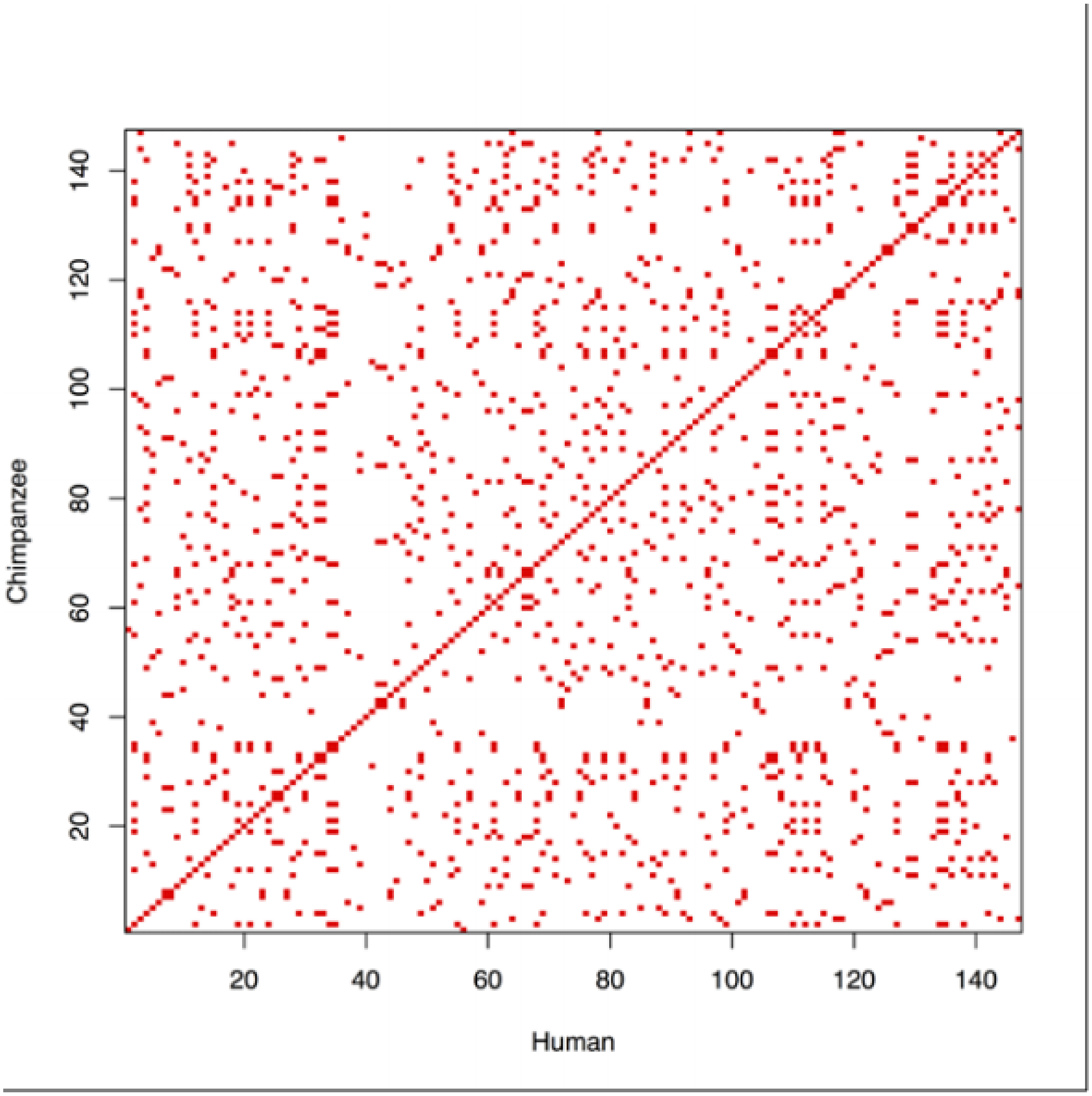
Recent Courses
Explore material from courses in 2013 and 2014.


P values KS test
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
P values Fisher test

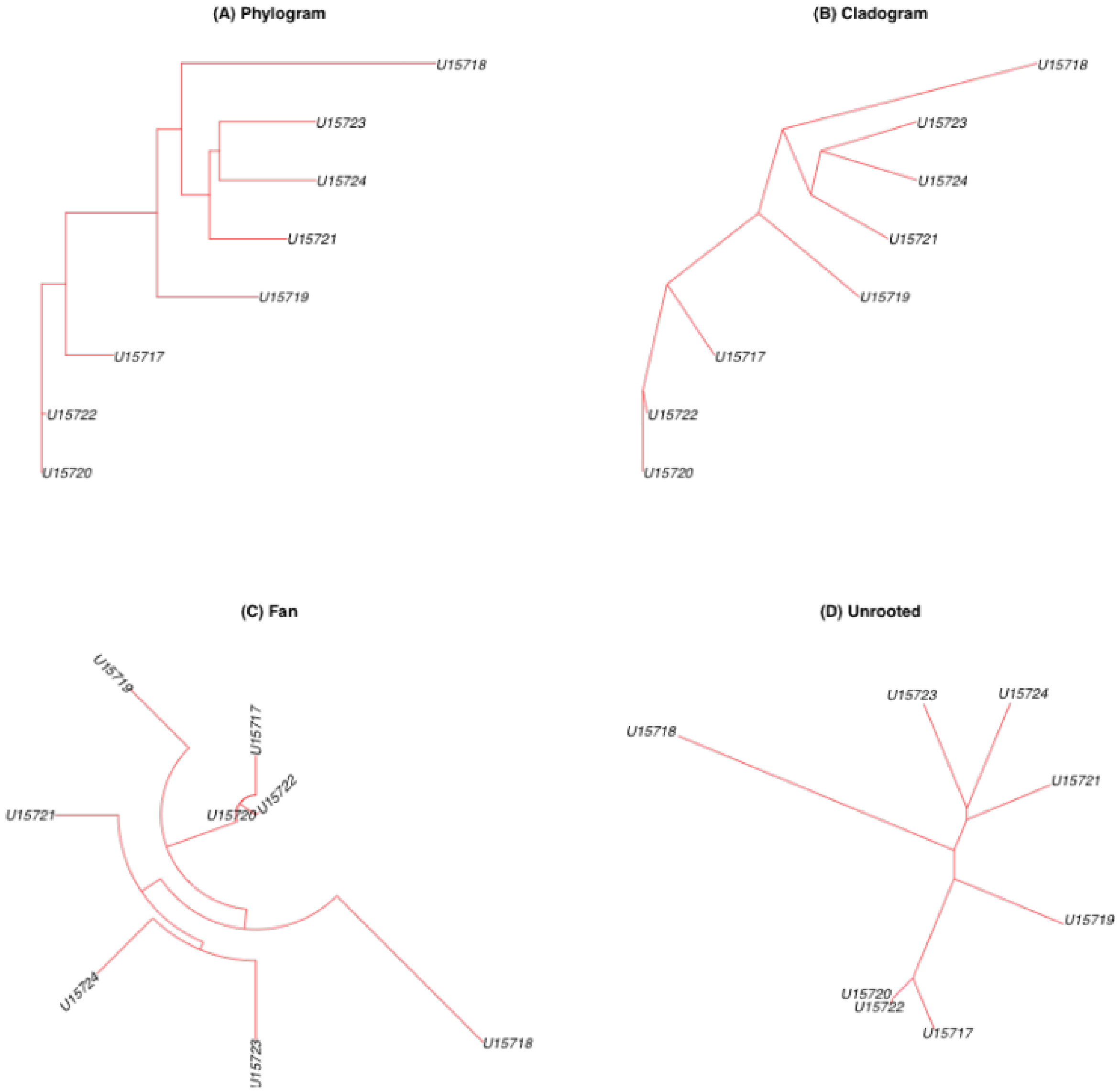


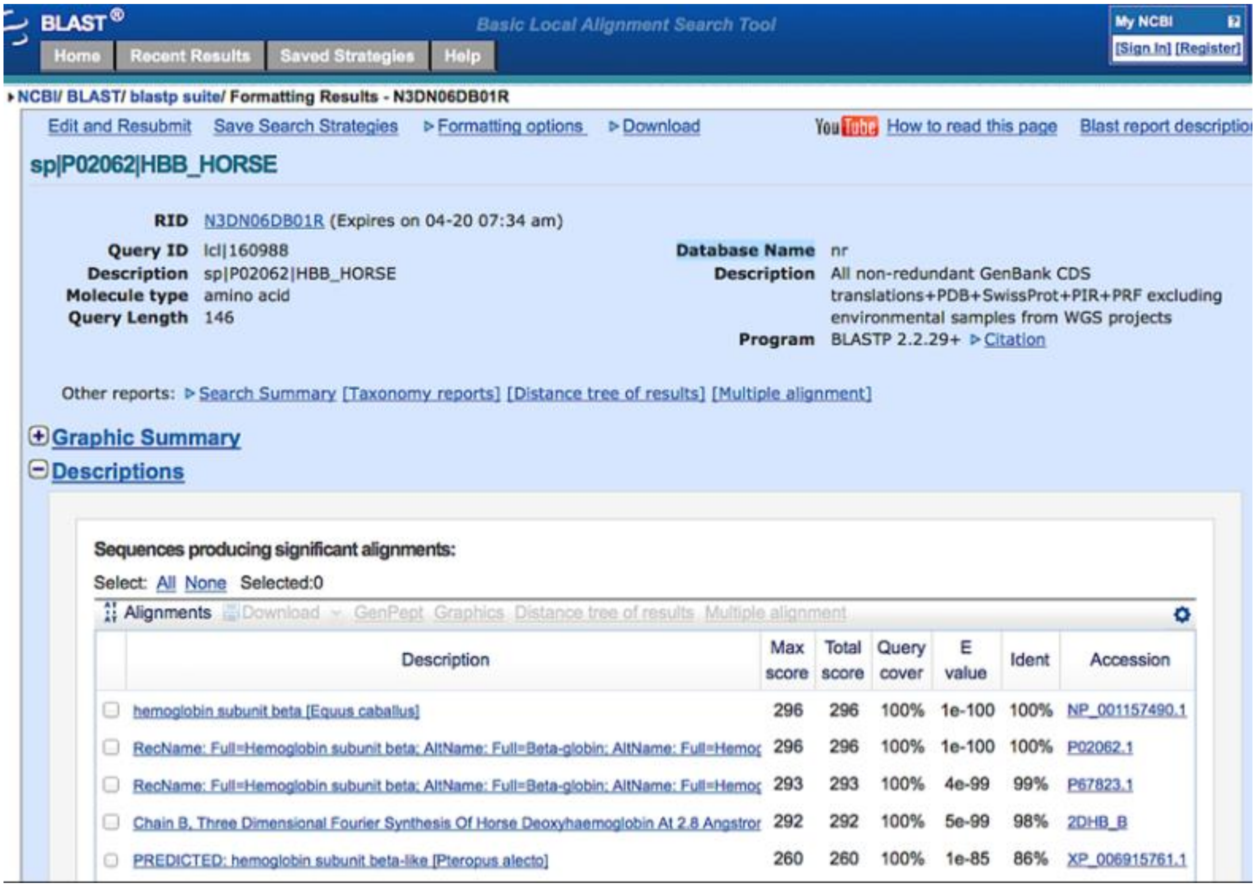
Chapter 3: Sequence Analysis with R


Replication
A T G
T C T
T T C
A A G
T G T
G C G
Transcription
A U G
U C U
U U C
A A G
U G U
G C G
Translation
Met
Ser
Phe
Lys
Cys
Ala
DNA
mRNA
PROTEIN



Chimpanzee
20
40
60
80
100
120
140
Human
20
40
60
80
100
120
140


(A) Phylogram
U15718
U15723
U15724
U15721
U15719
U15717
U15722
U15720
(B) Cladogram
U15718
U15723
U15724
U15721
U15719
U15717
U15722
U15720
(C) Fan
U15719
U15717
U15722
U15720
U15721
U15724
U15723
U15718
(D) Unrooted
U15723
U15724
U15718
U15721
U15719
U15720
U15722
U15717



▸NCBI/ BLAST/ blastp suite/ Formatting Results - N3DN06DB01R

sp|P02062|HBB_HORSE

RID N3DN06DB01R (Expires on 04-20 07:34 am)

Query ID lcl|160988
Description sp|P02062|HBB_HORSE
Molecule type amino acid
Query Length 146

Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.2.29+ ▷Citation

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]

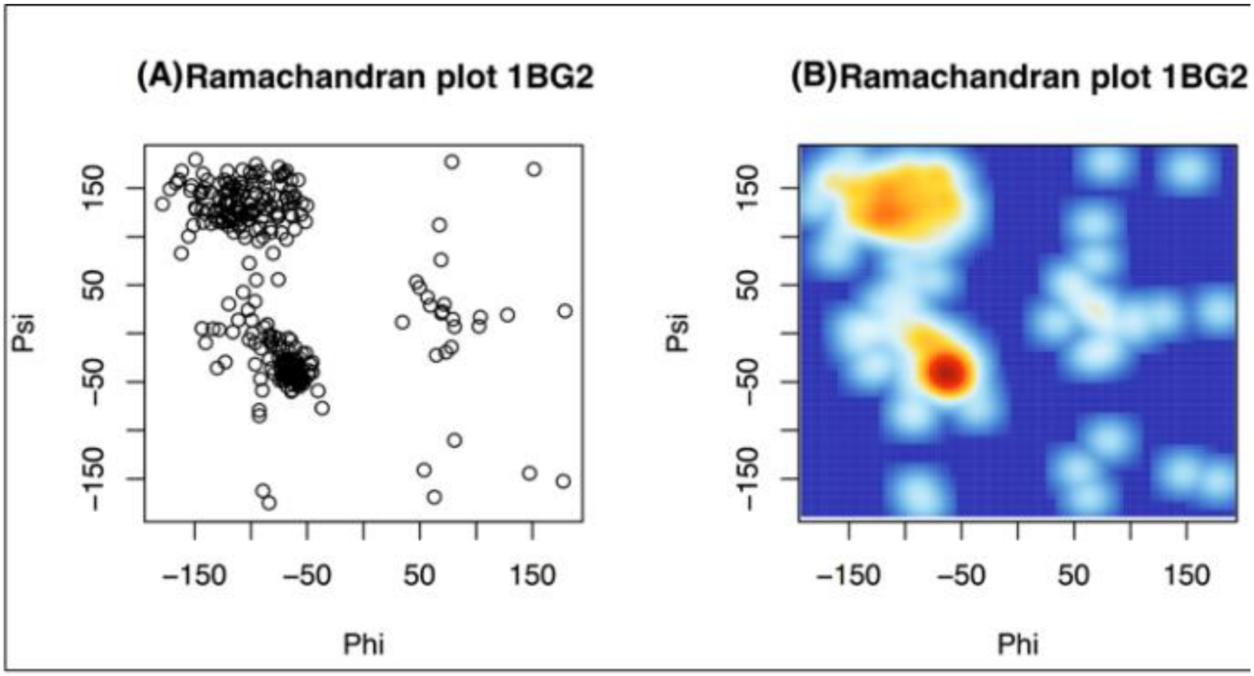
Sequences producing significant alignments:

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	hemoglobin subunit beta [Equus caballus]	296	296	100%	1e-100	100%	NP_001157490.1
☐	RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Hemog	296	296	100%	1e-100	100%	P02062.1
☐	RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Hemog	293	293	100%	4e-99	99%	P67823.1
☐	Chain B, Three Dimensional Fourier Synthesis Of Horse Deoxyhaemoglobin At 2.8 Angstror	292	292	100%	5e-99	98%	2DHB_B
☐	PREDICTED: hemoglobin subunit beta-like [Pteropus alecto]	260	260	100%	1e-85	86%	XP_006915761.1

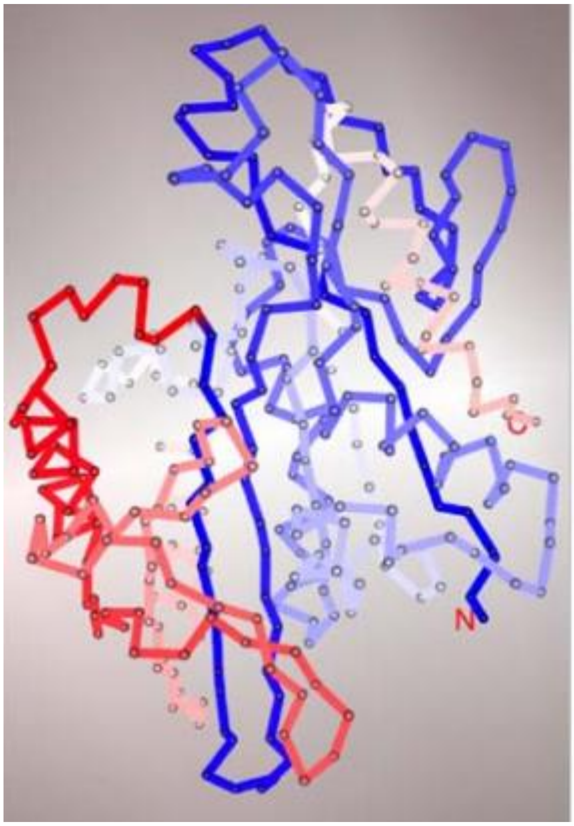
Chapter 4: Protein Structure Analysis with R

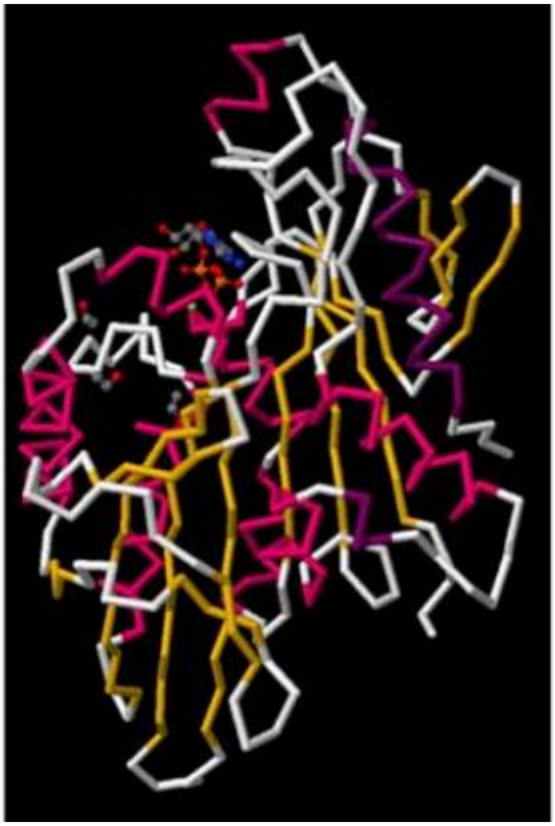
```
-YAFDRVFQSSTSQEQVYNDCAKKIVKDVLEGYNGTIFAYGQTSSGKTHTMEGKLHDPEGMG
TFTFDAVYDQTSCNYGIFQASFKPLIDAVLEGFNSTIFAYGQTGAGKTWTMGGNKEEP---G
-YAFDRVFQSSTSQEQVYNDCAKKIVKDVLEGYNGTIFAYGQTSSGKTHTMEGKLHDPEGMG
```



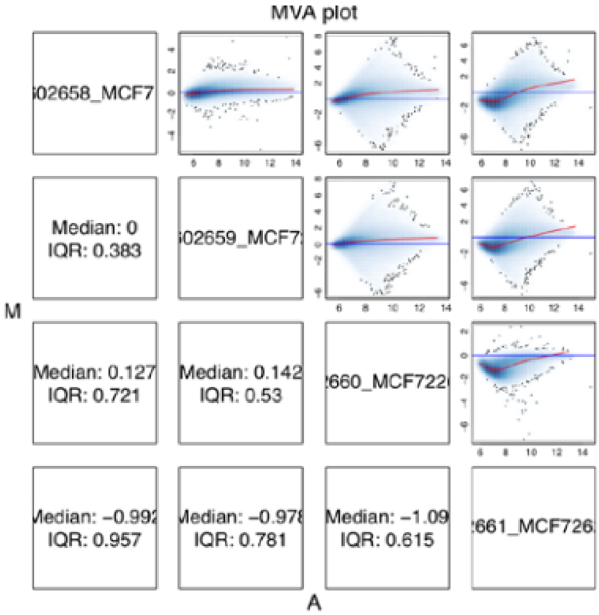
(A)Ramachandran plot 1BG2

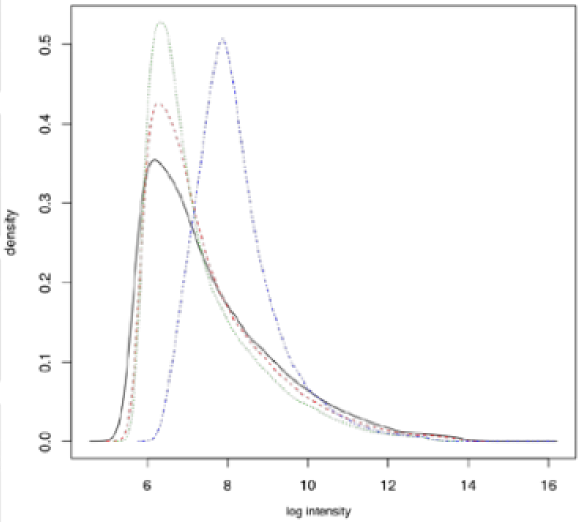
(B)Ramachandran plot 1BG2

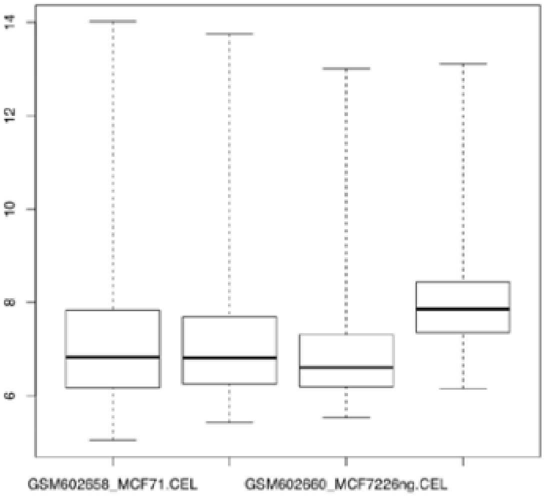


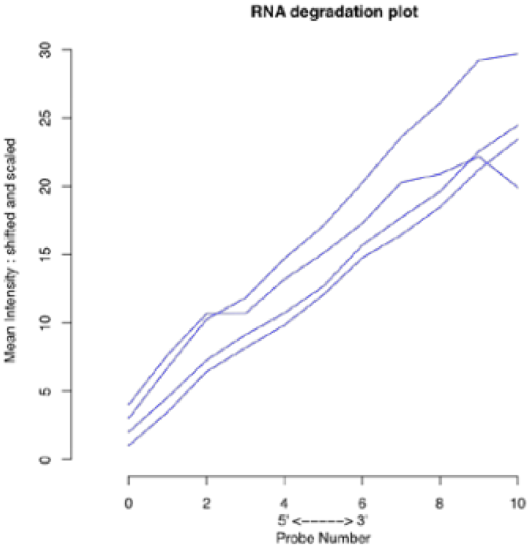


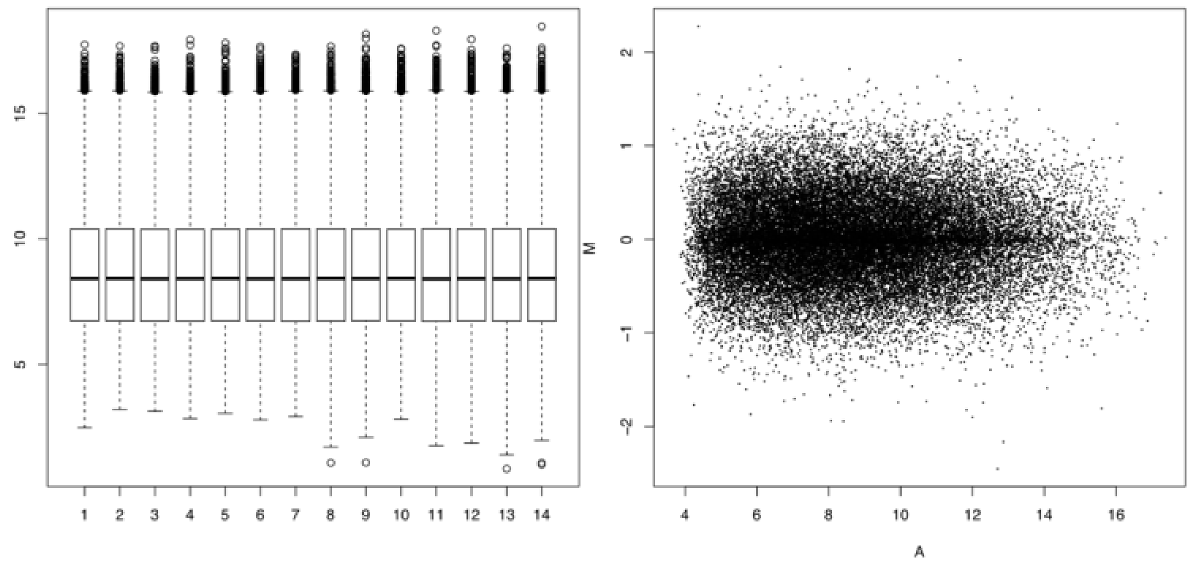
$$d_{ij} = mean \, | \, I_{ia} - I_{ja} \, |$$



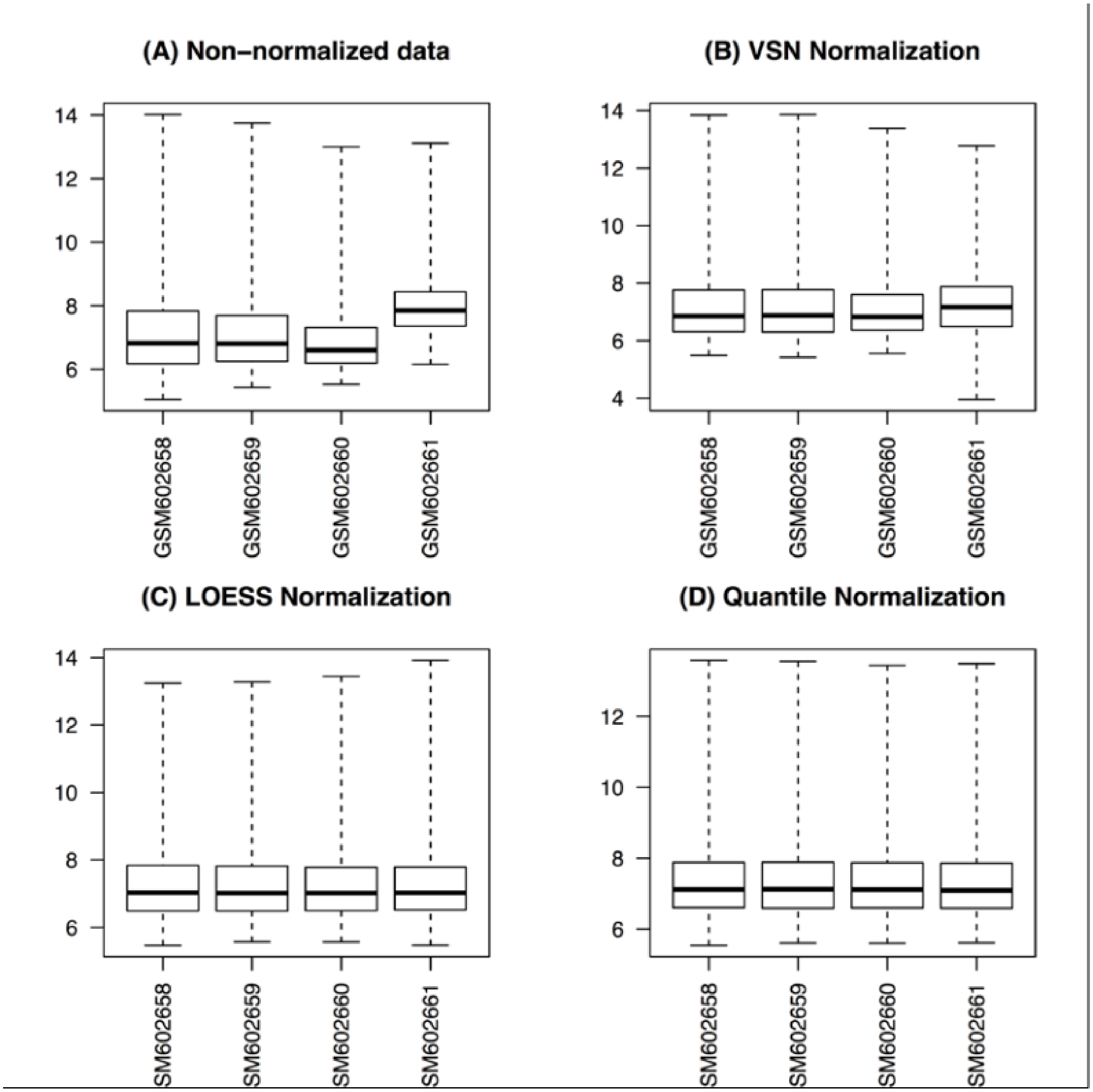








15
10
5
1
2
3
4
5
6
7
8
9
10
11
12
13
14
M
2
1
0
−1
−2
4
6
8
10
12
14
16
A

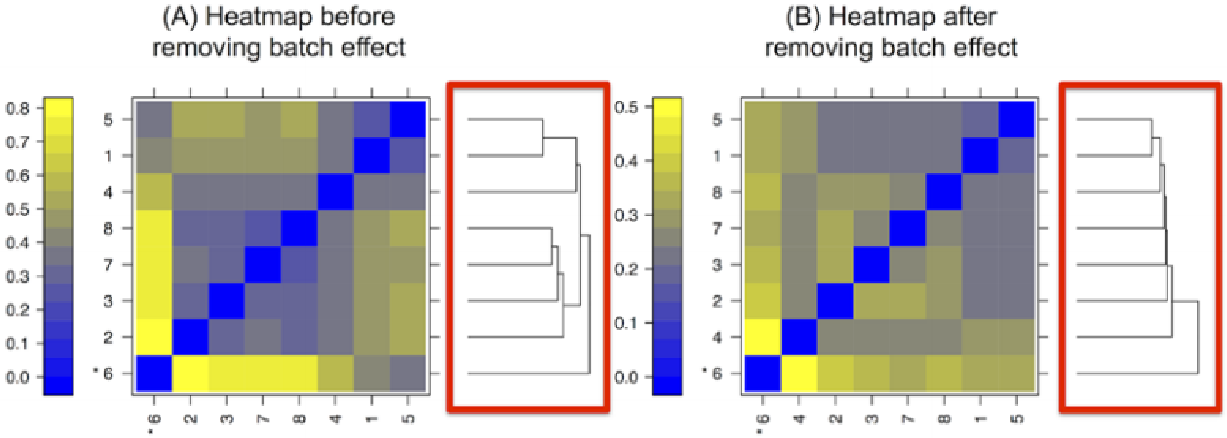


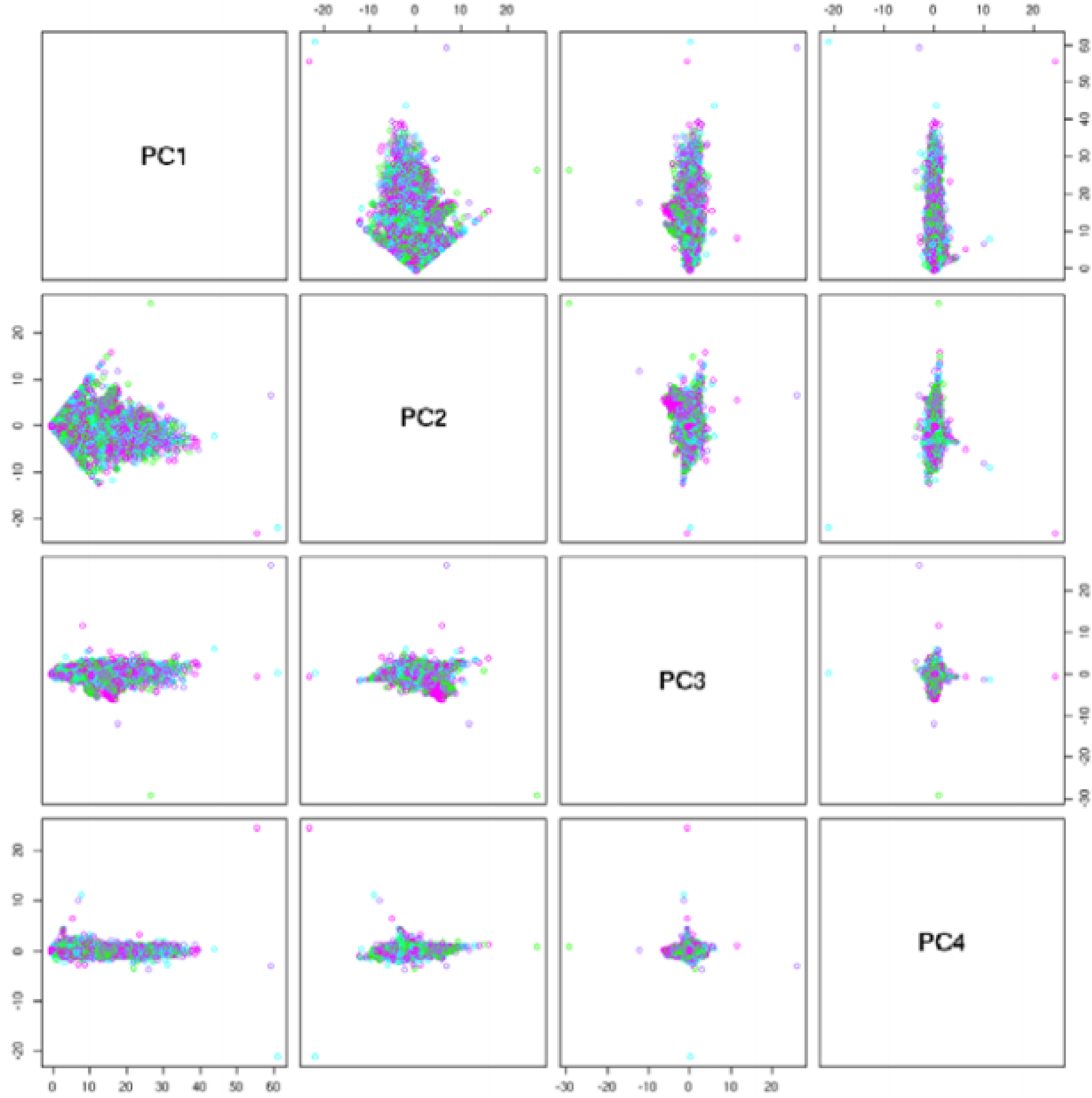
(A) Non-normalized data

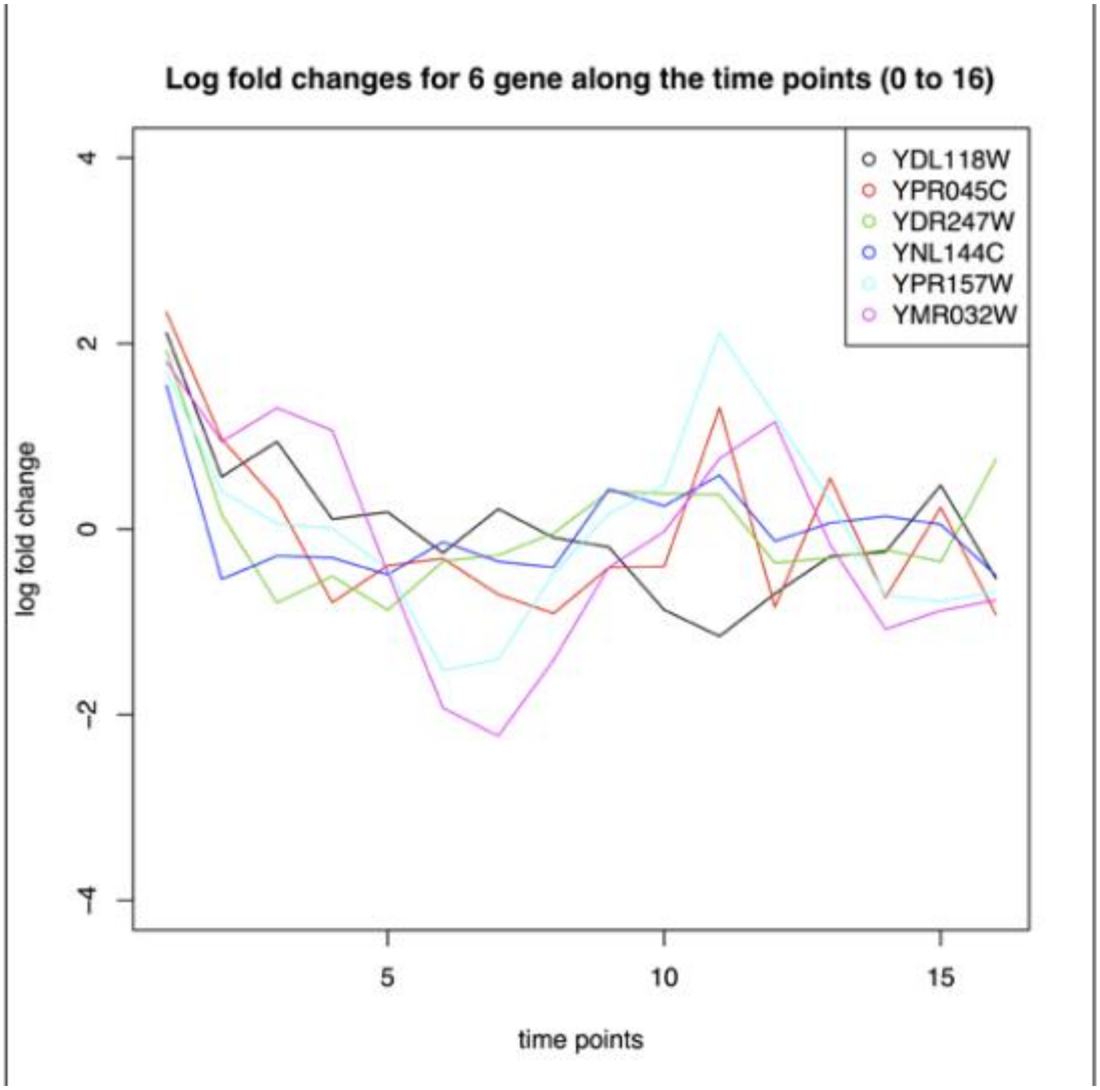
(B) VSN Normalization

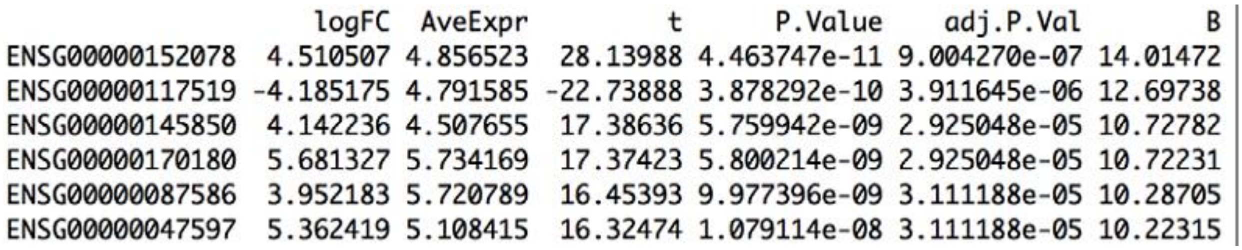
(C) LOESS Normalization

(D) Quantile Normalization

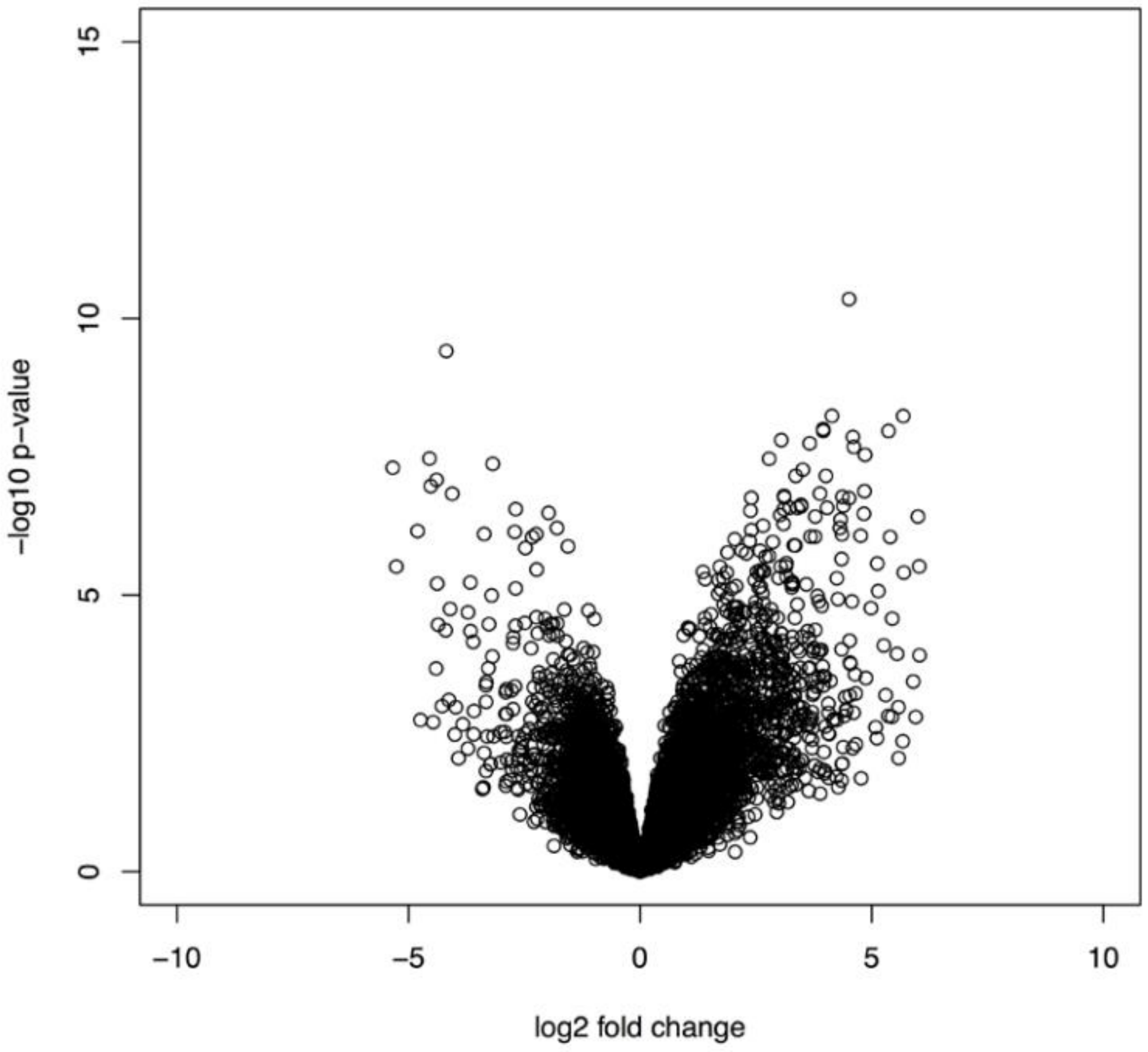

(A) Heatmap before removing batch effect
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
5
1
4
8
7
3
2
* 6
* 6
2
3
7
8
4
1
5
(B) Heatmap after removing batch effect
0.5
0.4
0.3
0.2
0.1
0.0
5
1
8
7
3
2
4
* 6
* 6
4
2
3
7
8
1
5

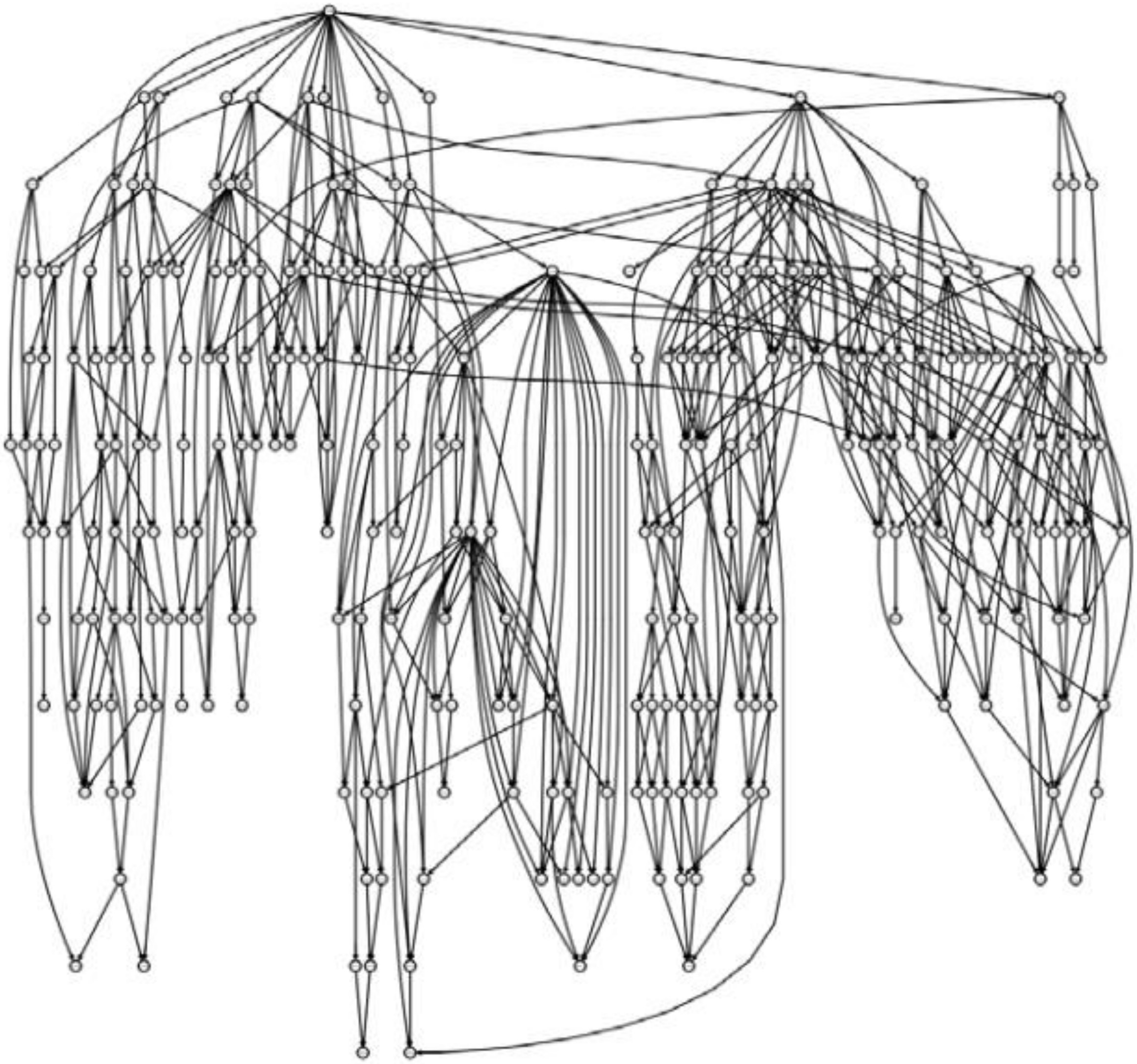

PC1
PC2
PC3
PC4
-20
-10
0
10
20
60
50
40
30
20
10
0
-30
-20
-10
0
10
20
30
40
50
60

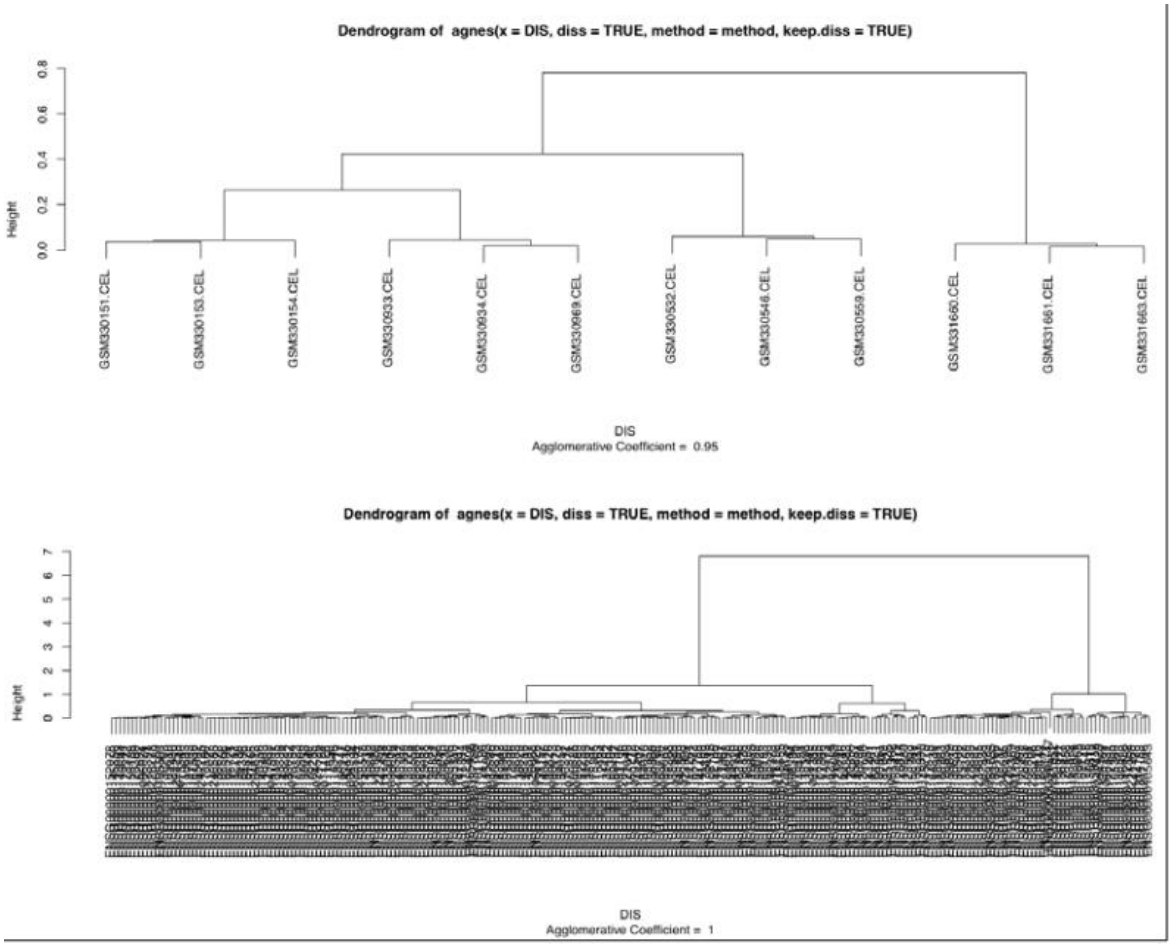


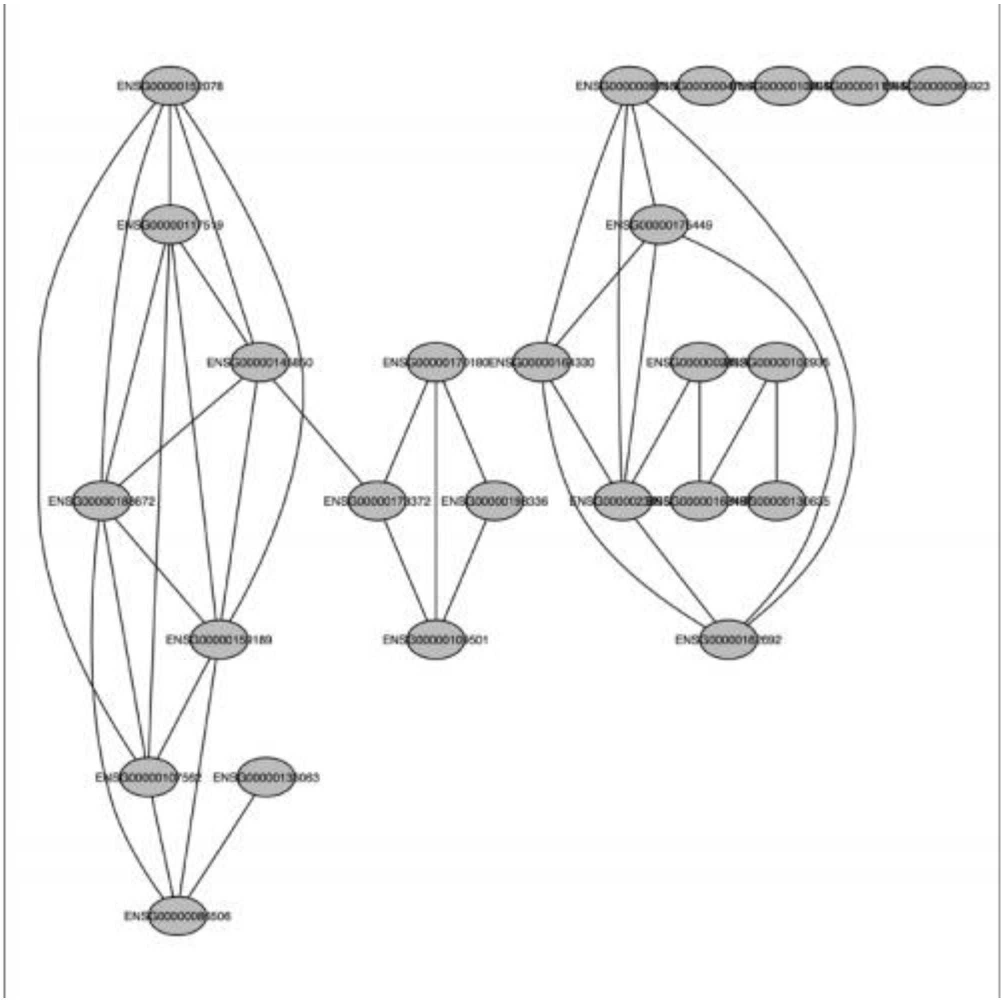


```
                    logFC  AveExpr         t      P.Value    adj.P.Val        B
ENSG00000152078  4.510507 4.856523  28.13988 4.463747e-11 9.004270e-07 14.01472
ENSG00000117519 -4.185175 4.791585 -22.73888 3.878292e-10 3.911645e-06 12.69738
ENSG00000145850  4.142236 4.507655  17.38636 5.759942e-09 2.925048e-05 10.72782
ENSG00000170180  5.681327 5.734169  17.37423 5.800214e-09 2.925048e-05 10.72231
ENSG00000087586  3.952183 5.720789  16.45393 9.977396e-09 3.111188e-05 10.28705
ENSG00000047597  5.362419 5.108415  16.32474 1.079114e-08 3.111188e-05 10.22315
```

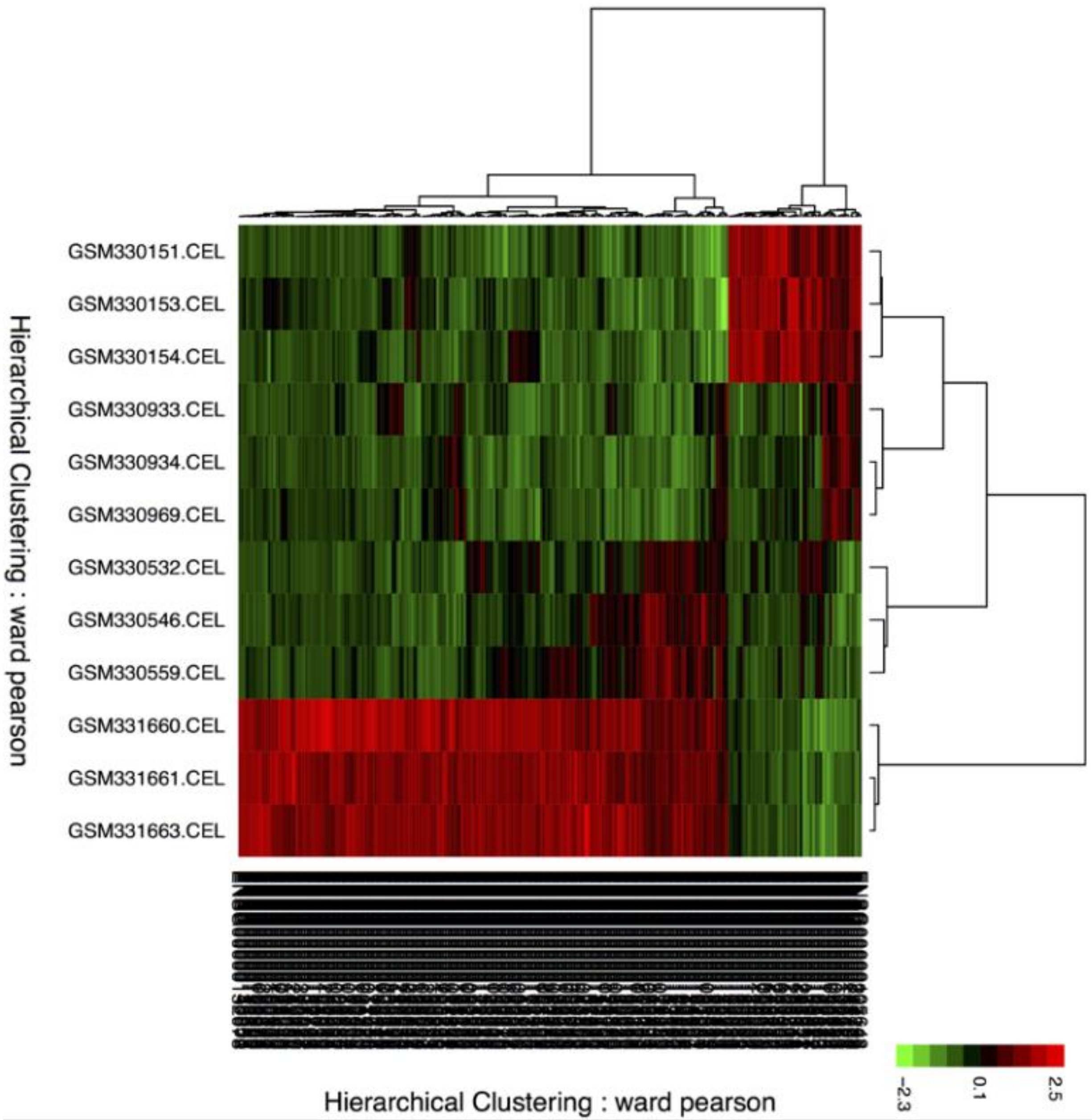

15
10
5
0
-log10 p-value
-10
-5
0
5
10
log2 fold change

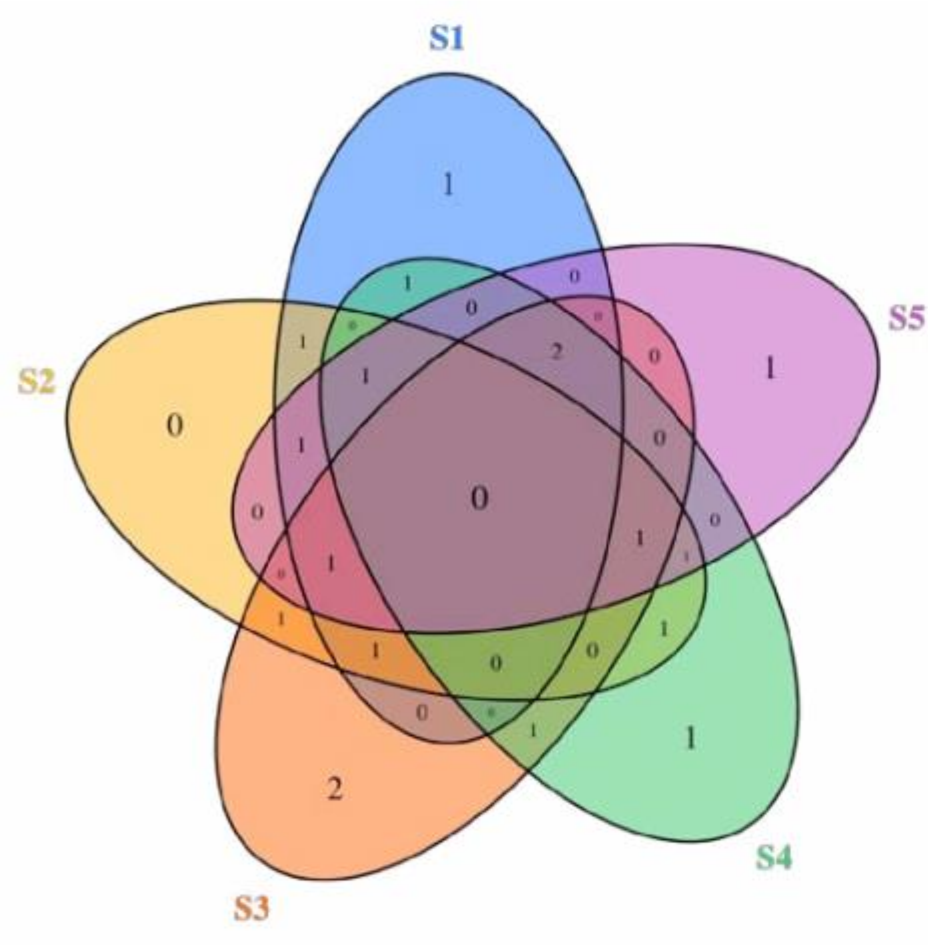


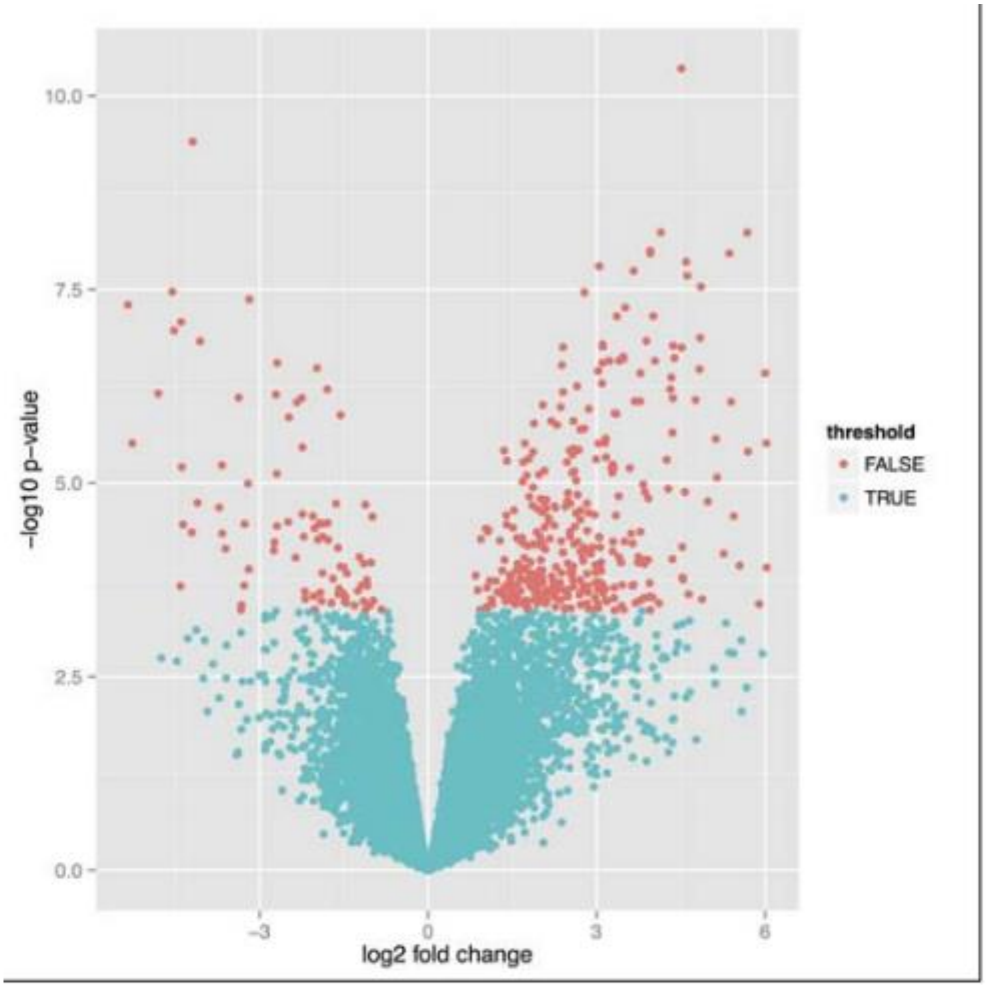

Dendrogram of agnes(x = DIS, diss = TRUE, method = method, keep.diss = TRUE)
Height
0.0
0.2
0.4
0.6
0.8
GSM330151.CEL
GSM330153.CEL
GSM330154.CEL
GSM330933.CEL
GSM330934.CEL
GSM330969.CEL
GSM330532.CEL
GSM330546.CEL
GSM330559.CEL
GSM331660.CEL
GSM331661.CEL
GSM331663.CEL
DIS
Agglomerative Coefficient = 0.95
Dendrogram of agnes(x = DIS, diss = TRUE, method = method, keep.diss = TRUE)
Height
0
1
2
3
4
5
6
7
DIS
Agglomerative Coefficient = 1



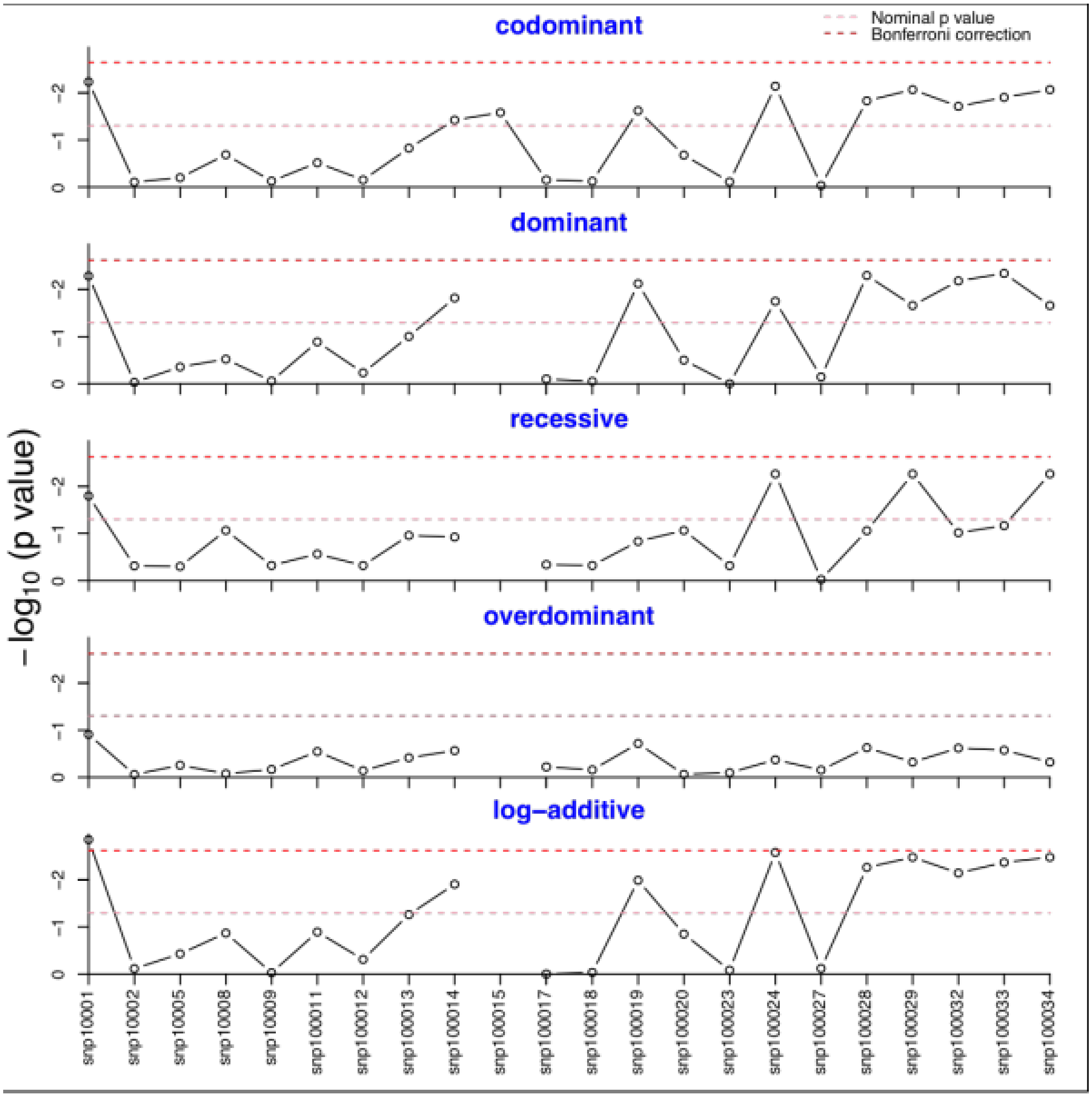
ENSG00000152078
ENSG00000117519
ENSG00000146850
ENSG00000170180
ENSG00000164330
ENSG00000188672
ENSG00000173372
ENSG00000198336
ENSG00000159189
ENSG00000100501
ENSG00000107562
ENSG00000139063
ENSG00000086506
ENSG00000175449
ENSG00000162692
ENSG00000130635
ENSG00000066923

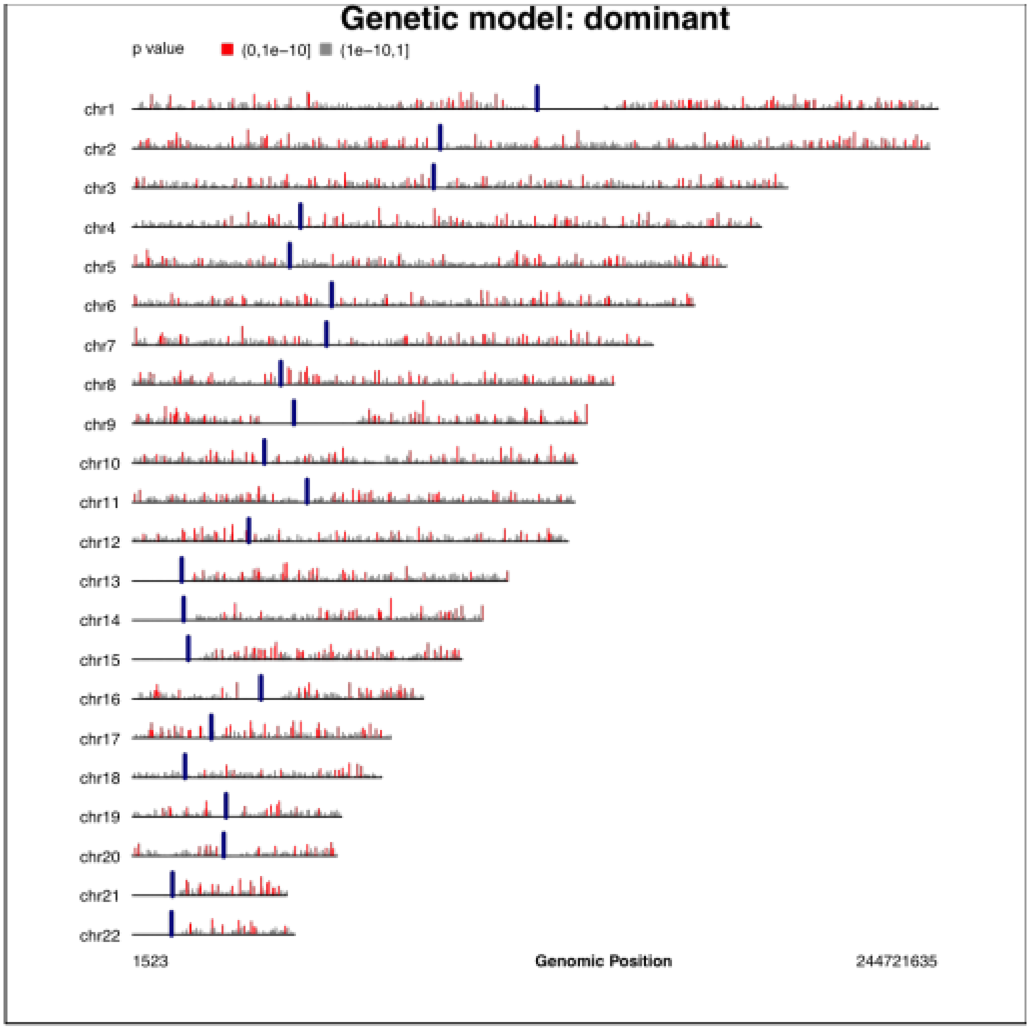

GSM330151.CEL
GSM330153.CEL
GSM330154.CEL
GSM330933.CEL
GSM330934.CEL
GSM330969.CEL
GSM330532.CEL
GSM330546.CEL
GSM330559.CEL
GSM331660.CEL
GSM331661.CEL
GSM331663.CEL
Hierarchical Clustering : ward pearson
−2.3
0.1
2.5
Hierarchical Clustering : ward pearson

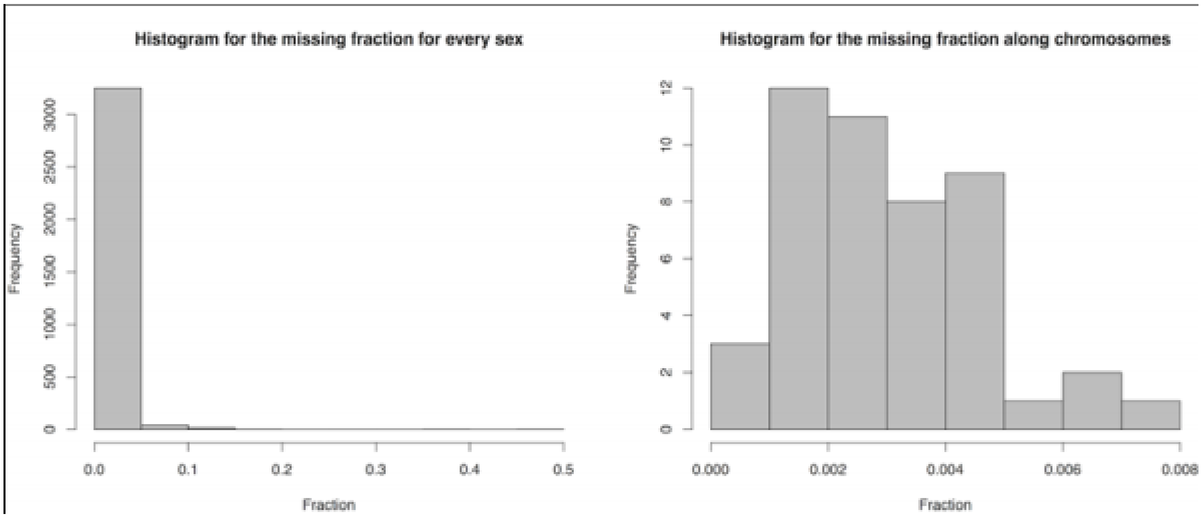

S1
1
1
0
0
0
S5
1
0
0
2
0
1
S2
1
0
1
0
0
0
1
1
0
1
0
1
1
1
0
0
1
0
0
1
1
2
S4
S3



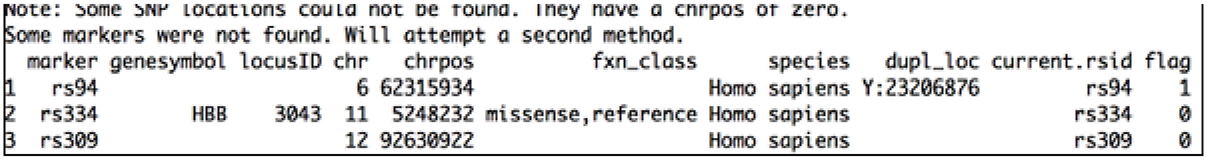
Chapter 6: Analyzing GWAS Data


codominant
dominant
recessive
overdominant
log−additive
Nominal p value
Bonferroni correction
− log10 (p value)
0
−1
−2
snp10001
snp10002
snp10005
snp10008
snp10009
snp100011
snp100012
snp100013
snp100014
snp100015
snp100017
snp100018
snp100019
snp100020
snp100023
snp100024
snp100027
snp100028
snp100029
snp100032
snp100033
snp100034

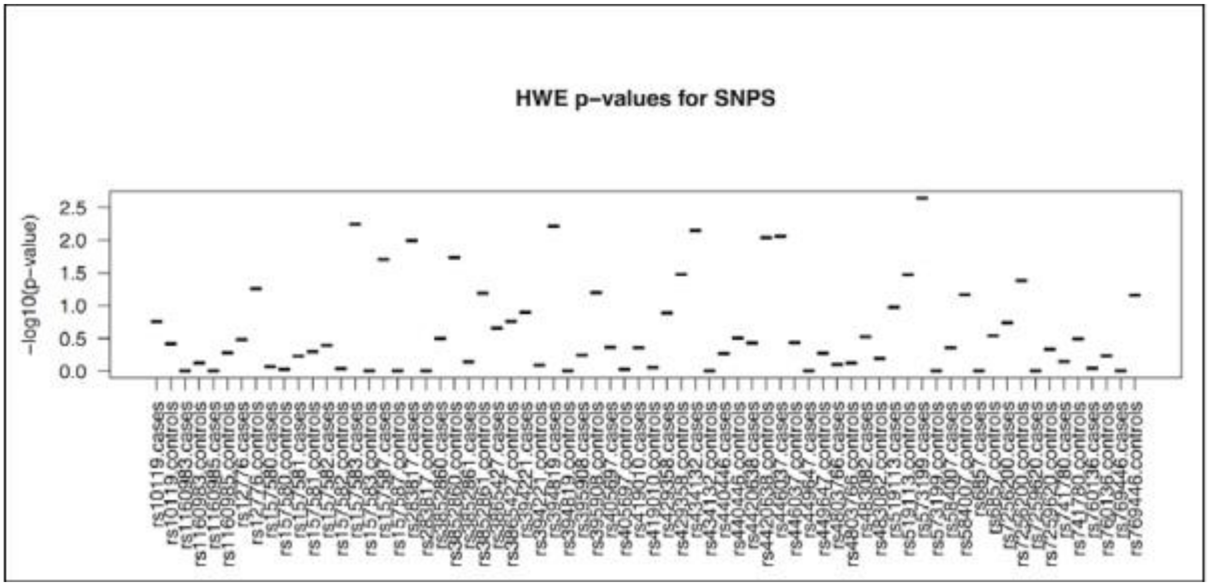

Genetic model: dominant
p value
🟥 (0,1e−10]
▨ (1e−10,1]
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
1523
Genomic Position
244721635

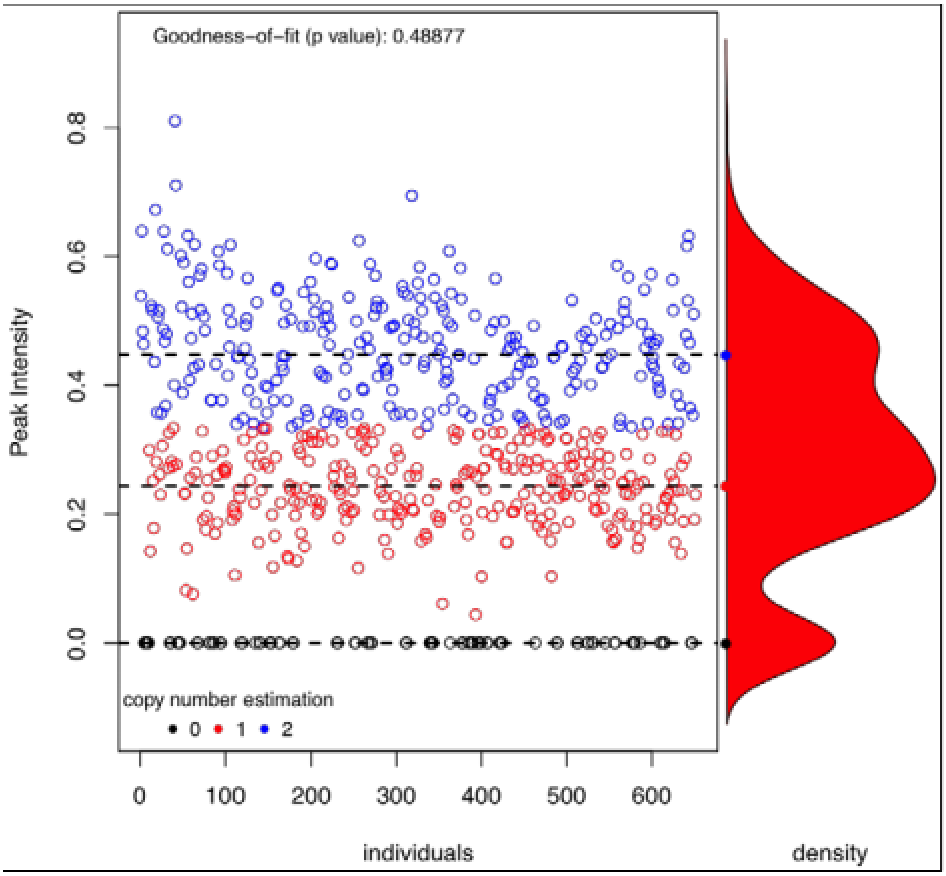

Histogram for the missing fraction for every sex
Frequency
0
500
1000
1500
2000
2500
3000
0.0
0.1
0.2
0.3
0.4
0.5
Fraction
Histogram for the missing fraction along chromosomes
Frequency
0
2
4
6
8
10
12
0.000
0.002
0.004
0.006
0.008
Fraction

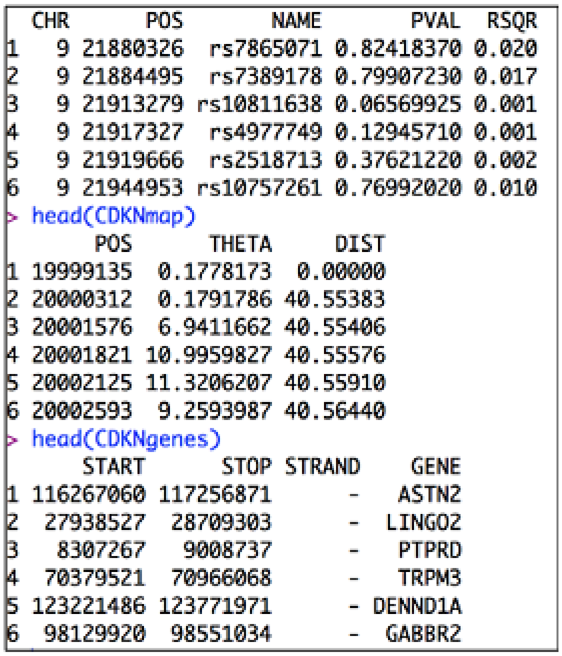
```
  marker variation
3   rs94         Y
2  rs334         N
1  rs309         Y

fiveprime
3
atcacaaaagaagtgaatatgccctgccccaccttaactgatgacattccaccacaaaagaagtgtaaatgg
ccggtccttgccttaagtgatgacattaccttgtgaaagtcctttcc
2
GCAATTTGTACTGATGGTATGGGGCCAAGAGATATATCTTAGAGGGAGGGCTGAGGGTTTGAAGTCCAACTC
CTAAGCCAGTGCCAGAAGAGCCAAGGACAGGTACGGCTGTCATCACTTAGACCTCACCCTGTGGAGCCACAC
CCTAGGGTTGGCCAATCTACTCCCAGGAGCAGGGAGGGCAGGAGCCAGGGCTGGGCATAAAAGTCAGGGCAG
AGCCATCTATTGCTTACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGC
ATCTGACTCCTG
1
tatttttagtagcgacagggtctcaccatgttggccaagctggtcttgaa

threeprime
3
ggctcatcctggctcaaaaagcacccccactgagcaccttgtgacccccactcctgcccaccagagaacaaa
cccccttttgactgtaattttcctttacgtacccaaatcctataaaacG
2
GGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
GTTGGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCATGTGGAGACAGAGAAGACTC
TTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGGTGGTCT
ACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTA
AGGTGAAGGCTC
1
tcctgatctcaggtgatccacccgcctcggcctcccaaagtgctgggatt
  flag
3    0
2    0
1    0
```



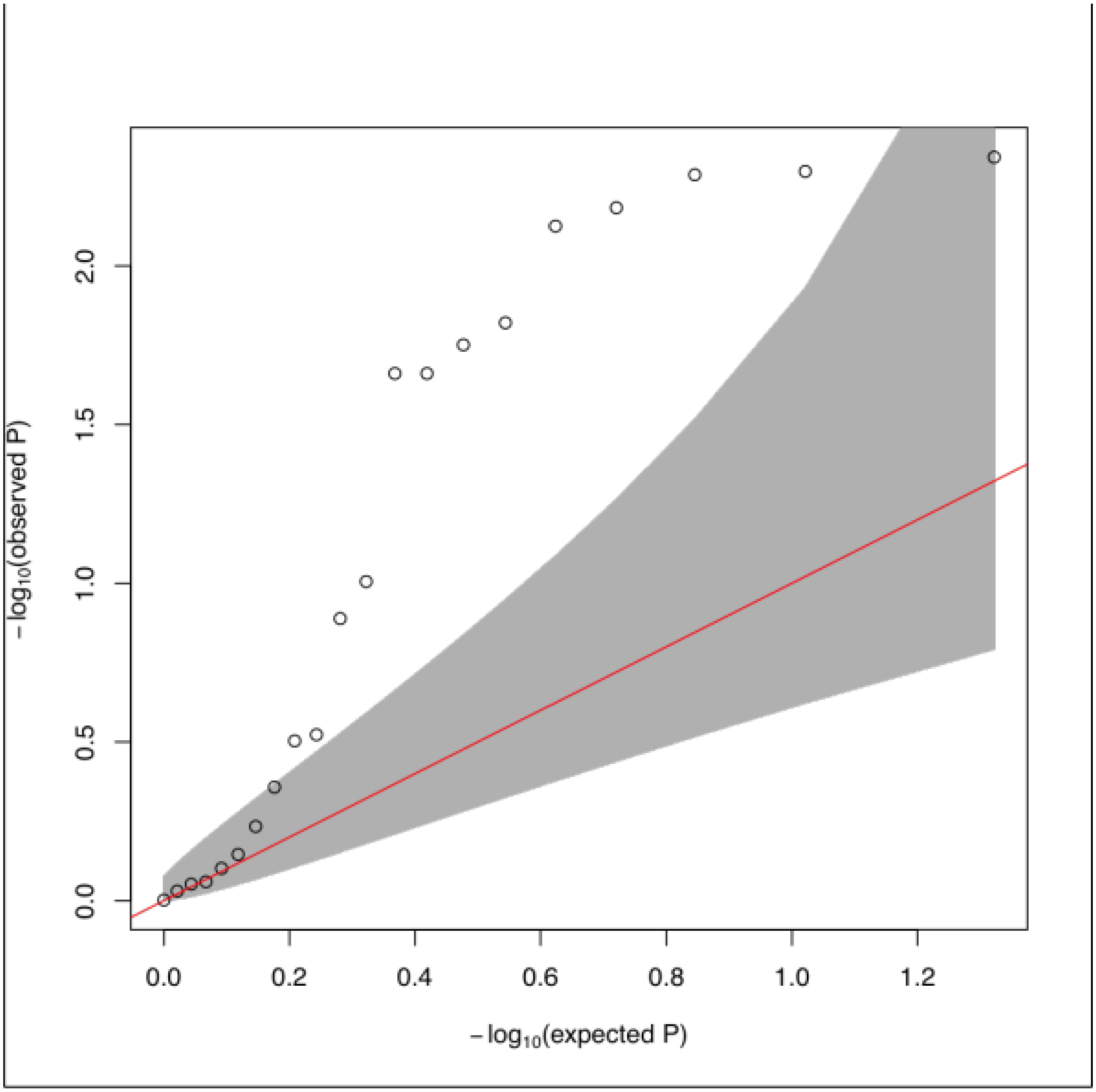
```
Note: Some SNP locations could not be found. They have a chrpos of zero.
Some markers were not found. Will attempt a second method.
  marker genesymbol locusID chr   chrpos           fxn_class      species   dupl_loc current.rsid flag
1   rs94                      6 62315934                     Homo sapiens Y:23206876         rs94    1
2  rs334        HBB    3043  11  5248232 missense,reference Homo sapiens                    rs334    0
3  rs309                     12 92630922                     Homo sapiens                    rs309    0
```



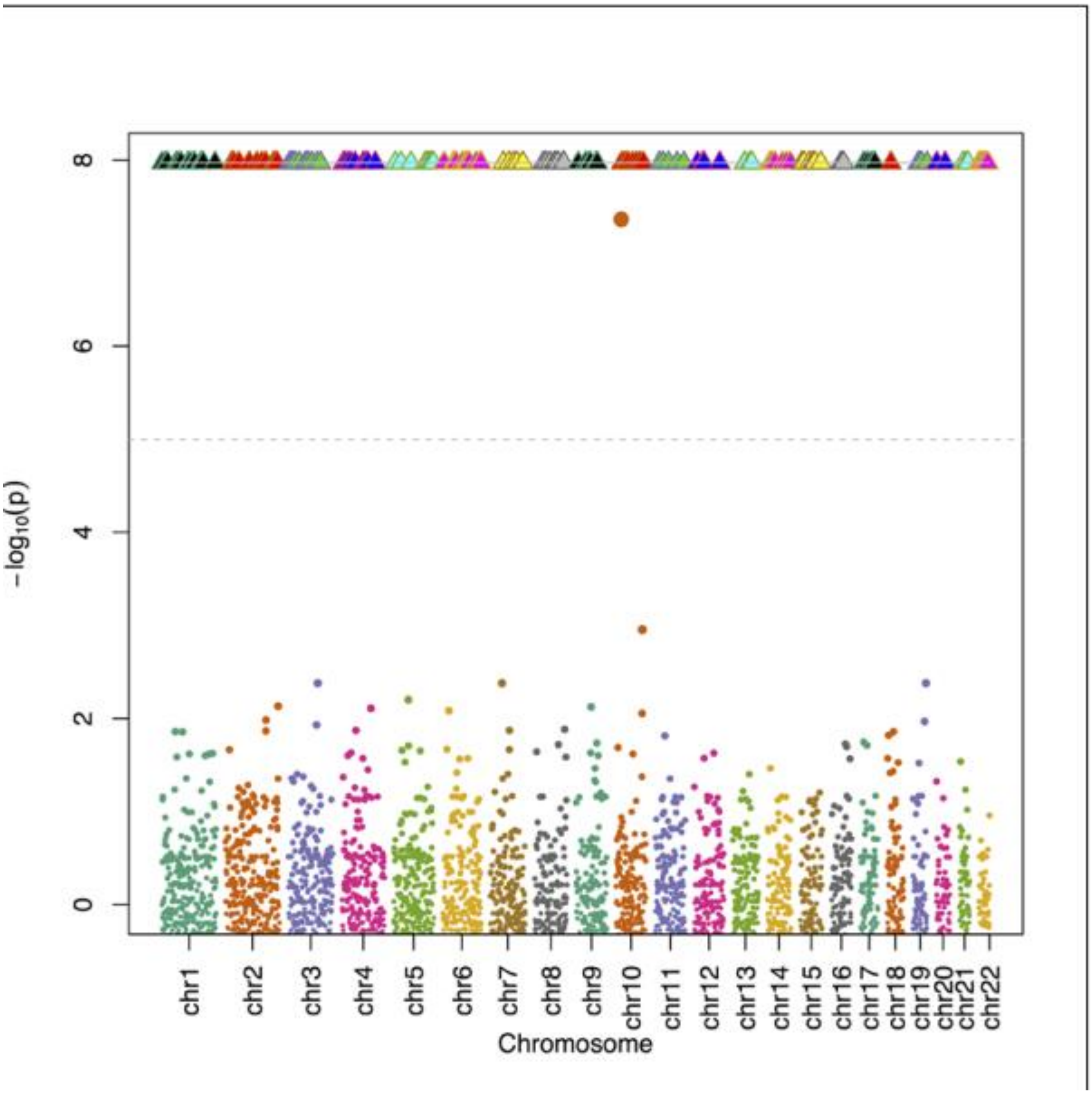


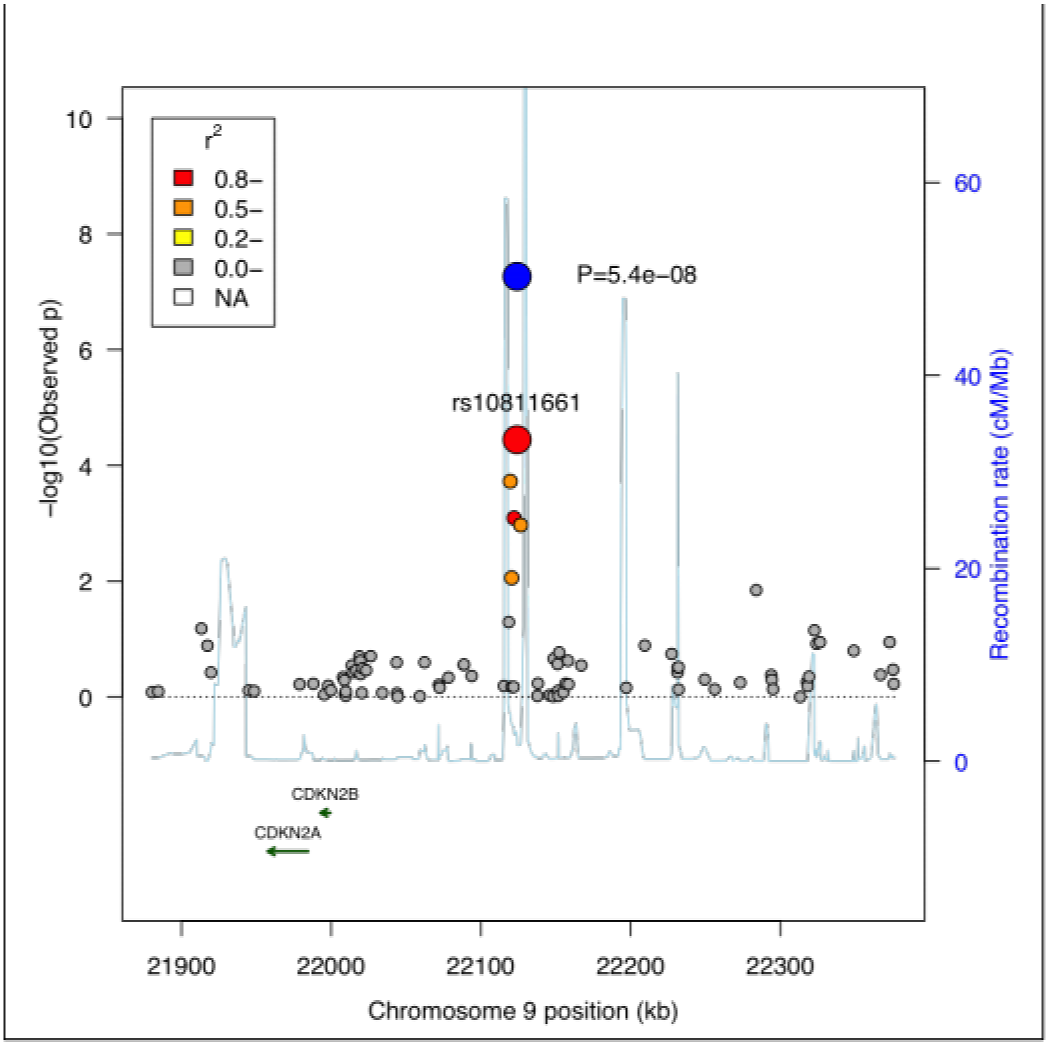


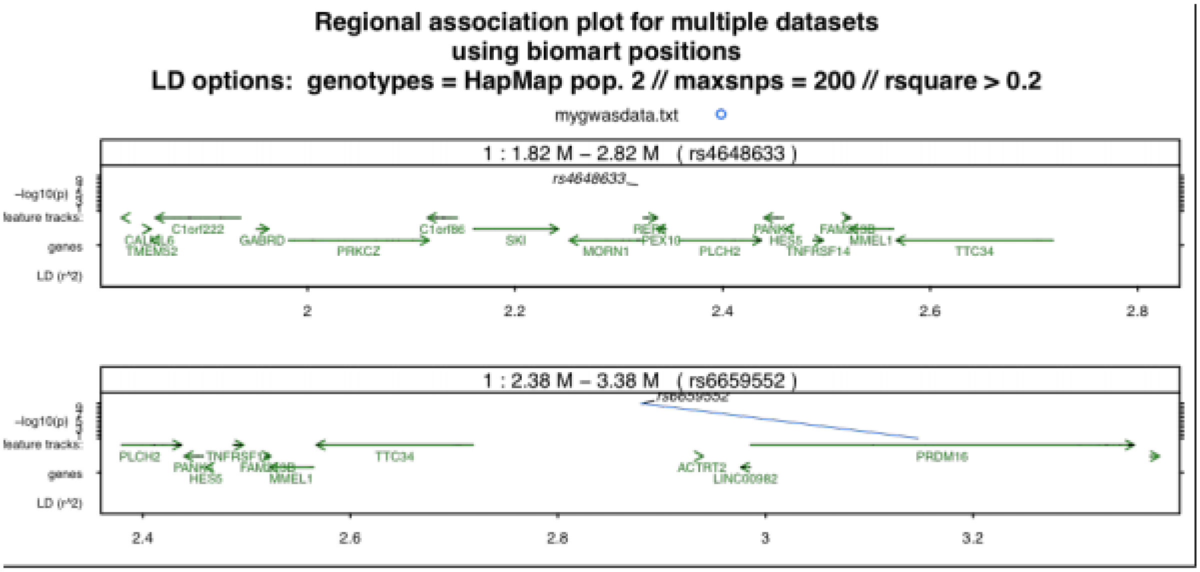
```
  CHR      POS       NAME       PVAL  RSQR
1   9 21880326  rs7865071 0.82418370 0.020
2   9 21884495  rs7389178 0.79907230 0.017
3   9 21913279 rs10811638 0.06569925 0.001
4   9 21917327  rs4977749 0.12945710 0.001
5   9 21919666  rs2518713 0.37621220 0.002
6   9 21944953 rs10757261 0.76992020 0.010
> head(CDKNmap)
       POS      THETA     DIST
1 19999135  0.1778173  0.00000
2 20000312  0.1791786 40.55383
3 20001576  6.9411662 40.55406
4 20001821 10.9959827 40.55576
5 20002125 11.3206207 40.55910
6 20002593  9.2593987 40.56440
> head(CDKNgenes)
      START      STOP STRAND    GENE
1 116267060 117256871      -   ASTN2
2  27938527  28709303      -  LINGO2
3   8307267   9008737      -   PTPRD
4  70379521  70966068      -   TRPM3
5 123221486 123771971      - DENND1A
6  98129920  98551034      -  GABBR2
```

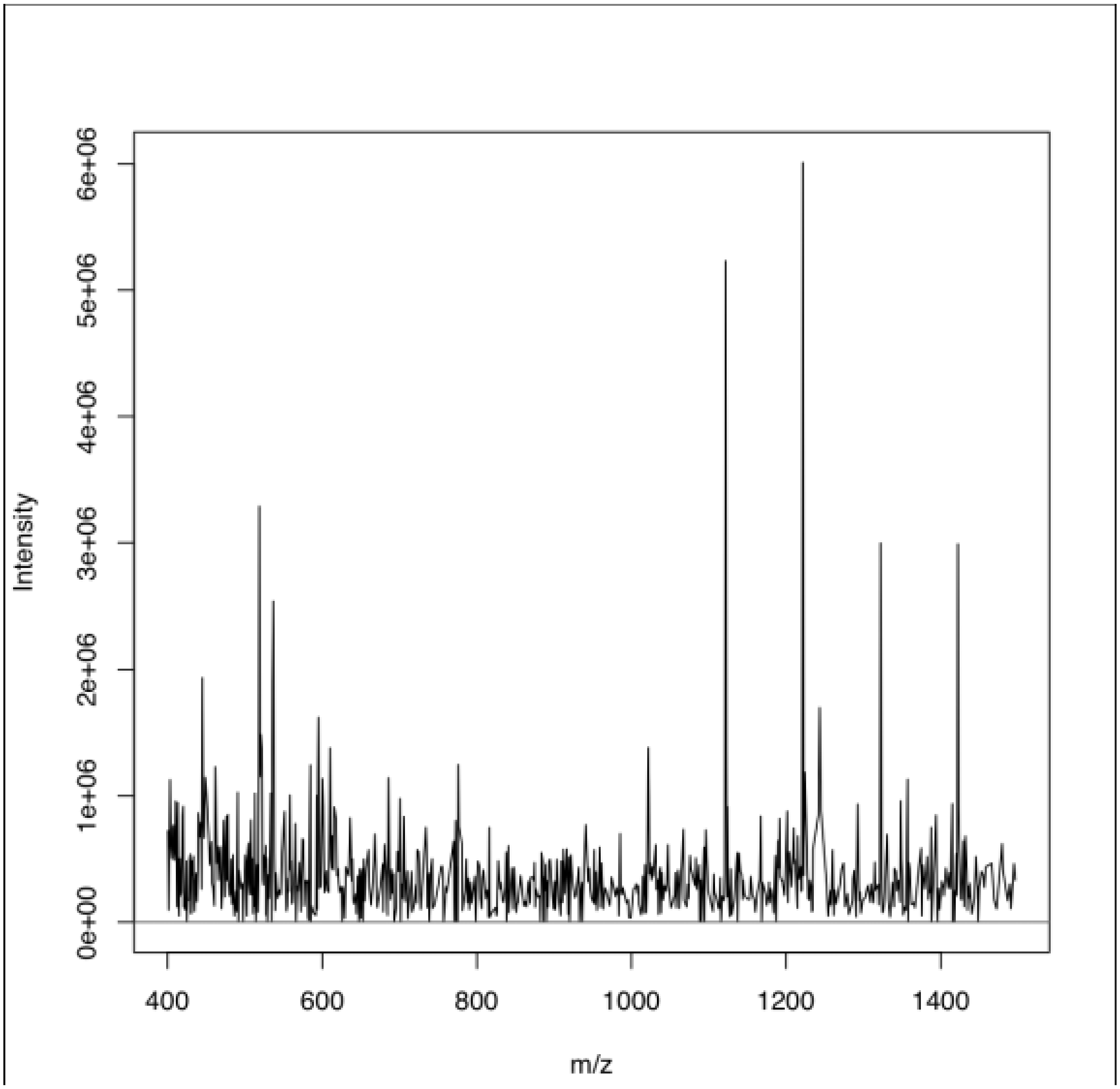


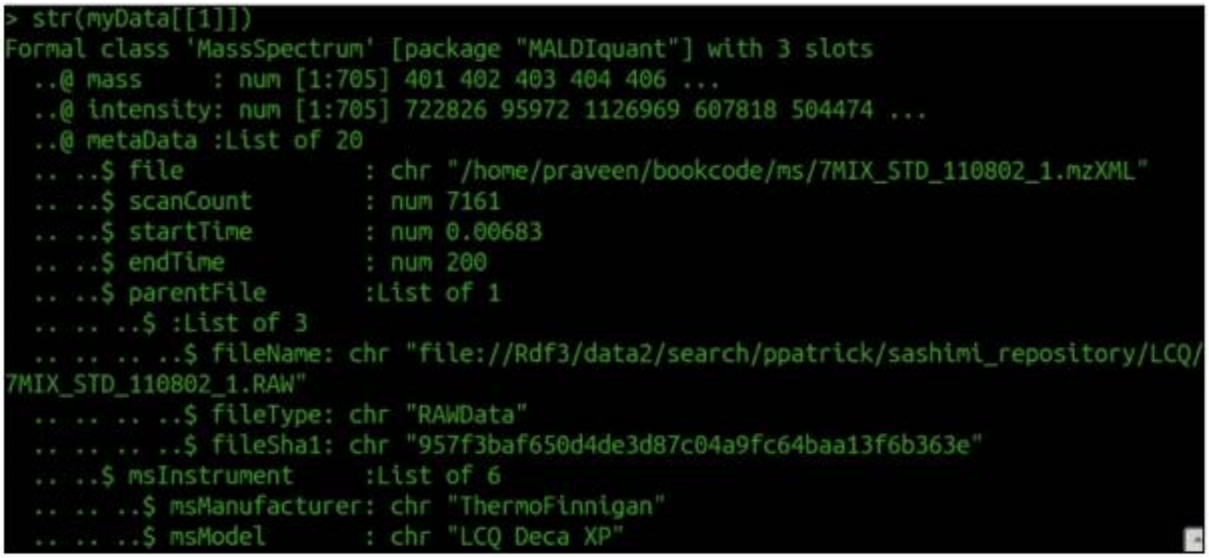
2.0
1.5
1.0
0.5
0.0
$-\log_{10}$(observed P)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
$-\log_{10}$(expected P)


−log₁₀(p)
0
2
4
6
8
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
Chromosome

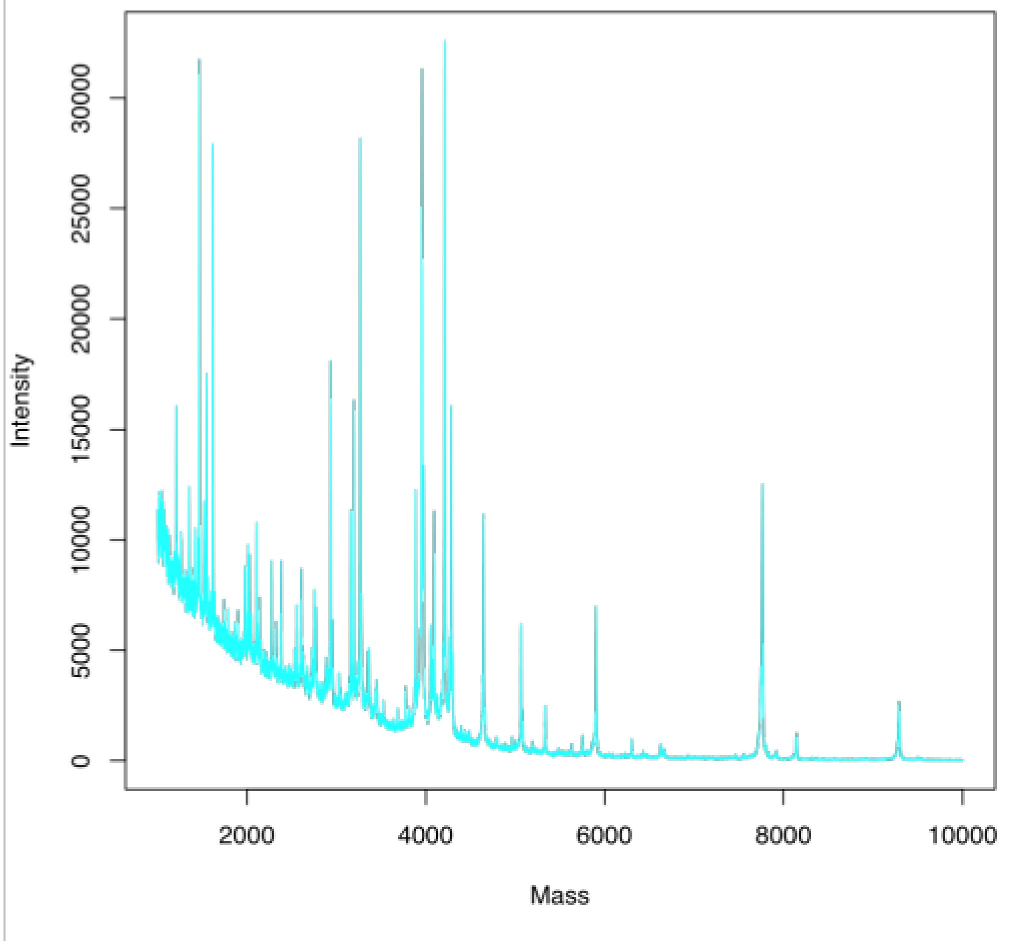

r^2
0.8–
0.5–
0.2–
0.0–
NA
P=5.4e−08
rs10811661
CDKN2B
CDKN2A
−log10(Observed p)
0
2
4
6
8
10
Recombination rate (cM/Mb)
0
20
40
60
21900
22000
22100
22200
22300
Chromosome 9 position (kb)

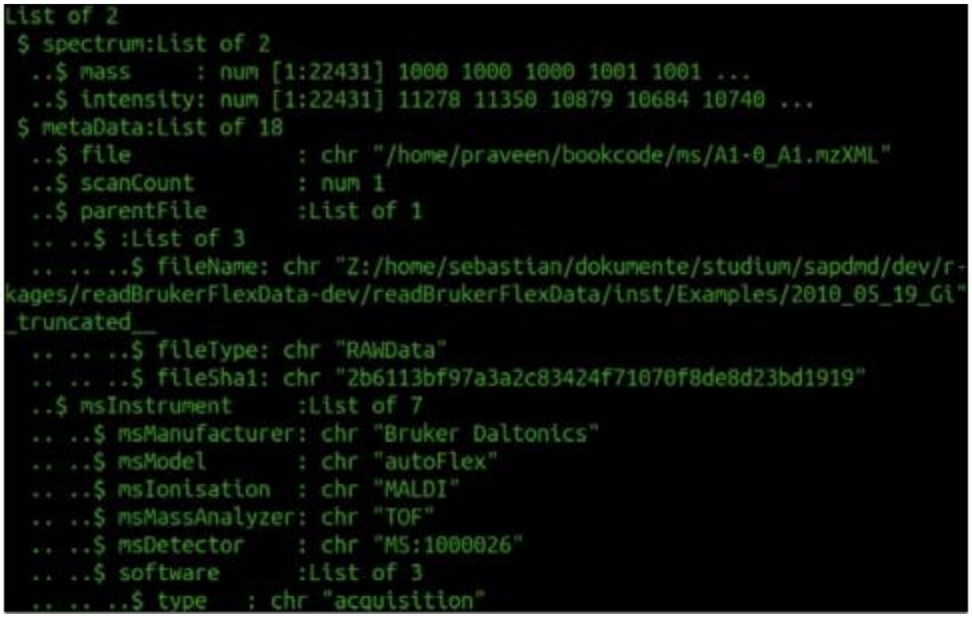

Regional association plot for multiple datasets
using biomart positions
LD options: genotypes = HapMap pop. 2 // maxsnps = 200 // rsquare > 0.2
mygwasdata.txt
1 : 1.82 M – 2.82 M (rs4648633)
rs4648633
−log10(p)
feature tracks:
genes
LD (r^2)
C1orf222
CALML6
TMEM52
GABRD
PRKCZ
C1orf86
SKI
MORN1
RER1
PEX10
PLCH2
PANK4
HES5
TNFRSF14
FAM213B
MMEL1
TTC34
2
2.2
2.4
2.6
2.8
1 : 2.38 M – 3.38 M (rs6659552)
rs6659552
−log10(p)
feature tracks:
genes
LD (r^2)
PLCH2
PANK4
HES5
TNFRSF14
FAM213B
MMEL1
TTC34
ACTRT2
LINC00982
PRDM16
2.4
2.6
2.8
3
3.2


Intensity
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
6e+06
400
600
800
1000
1200
1400
m/z

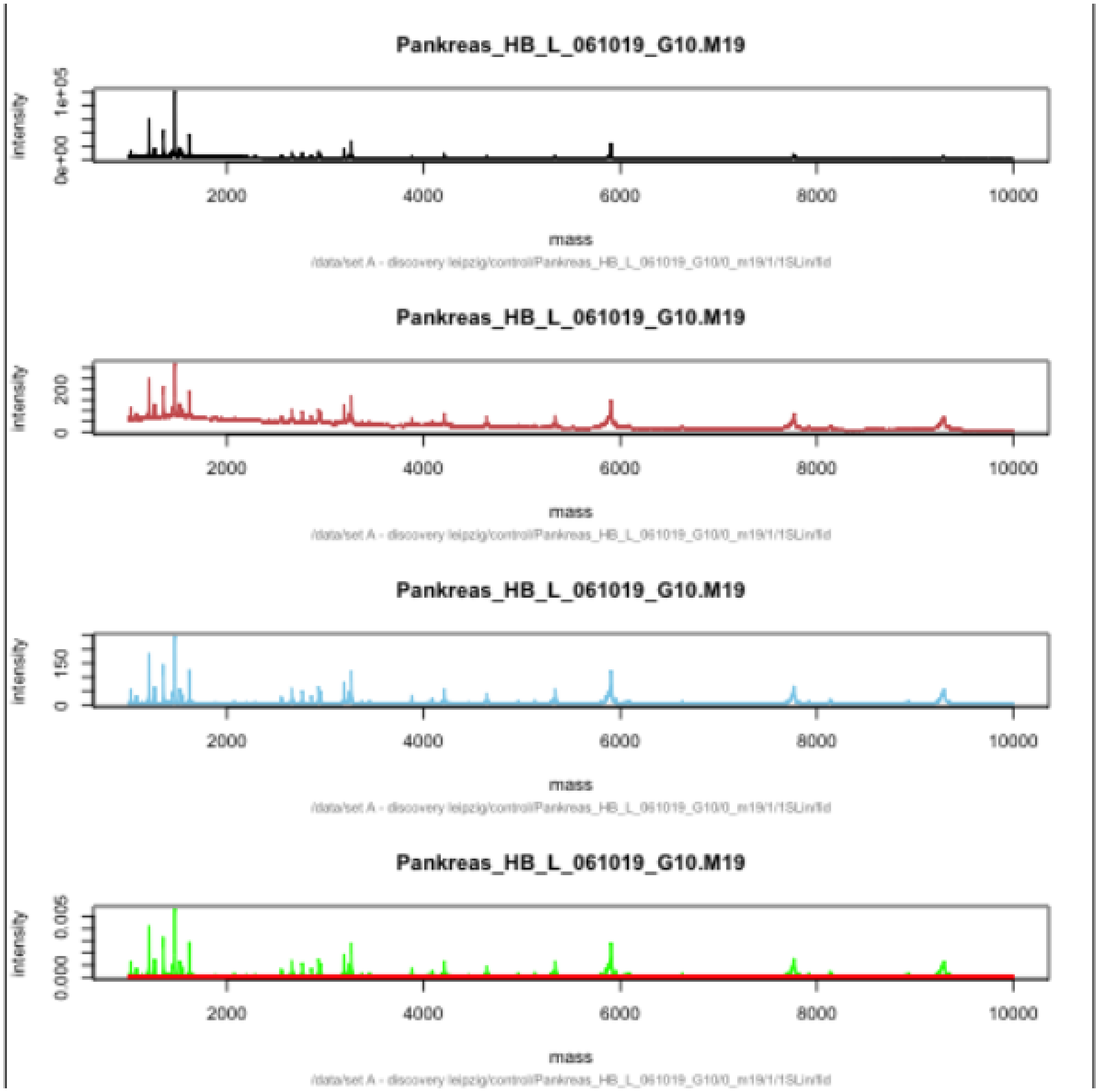


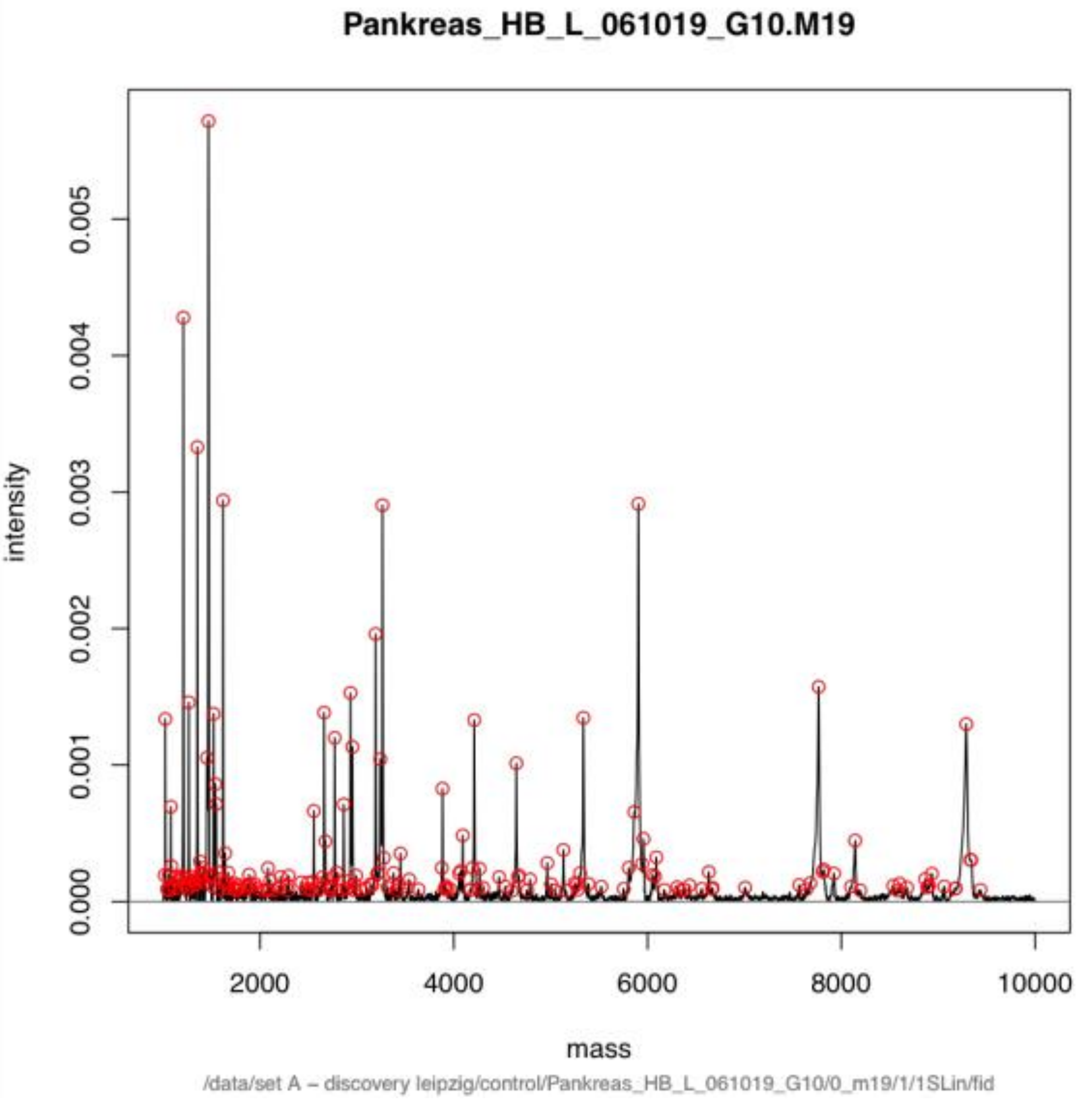
```
> str(myData[[1]])
Formal class 'MassSpectrum' [package "MALDIquant"] with 3 slots
  ..@ mass     : num [1:705] 401 402 403 404 406 ...
  ..@ intensity: num [1:705] 722826 95972 1126969 607818 504474 ...
  ..@ metaData :List of 20
  .. ..$ file               : chr "/home/praveen/bookcode/ms/7MIX_STD_110802_1.mzXML"
  .. ..$ scanCount          : num 7161
  .. ..$ startTime          : num 0.00683
  .. ..$ endTime            : num 200
  .. ..$ parentFile         :List of 1
  .. .. ..$ :List of 3
  .. .. .. ..$ fileName: chr "file://Rdf3/data2/search/ppatrick/sashimi_repository/LCQ/
7MIX_STD_110802_1.RAW"
  .. .. .. ..$ fileType: chr "RAWData"
  .. .. .. ..$ fileSha1: chr "957f3baf650d4de3d87c04a9fc64baa13f6b363e"
  .. ..$ msInstrument       :List of 6
  .. .. ..$ msManufacturer: chr "ThermoFinnigan"
  .. .. ..$ msModel       : chr "LCQ Deca XP"
```





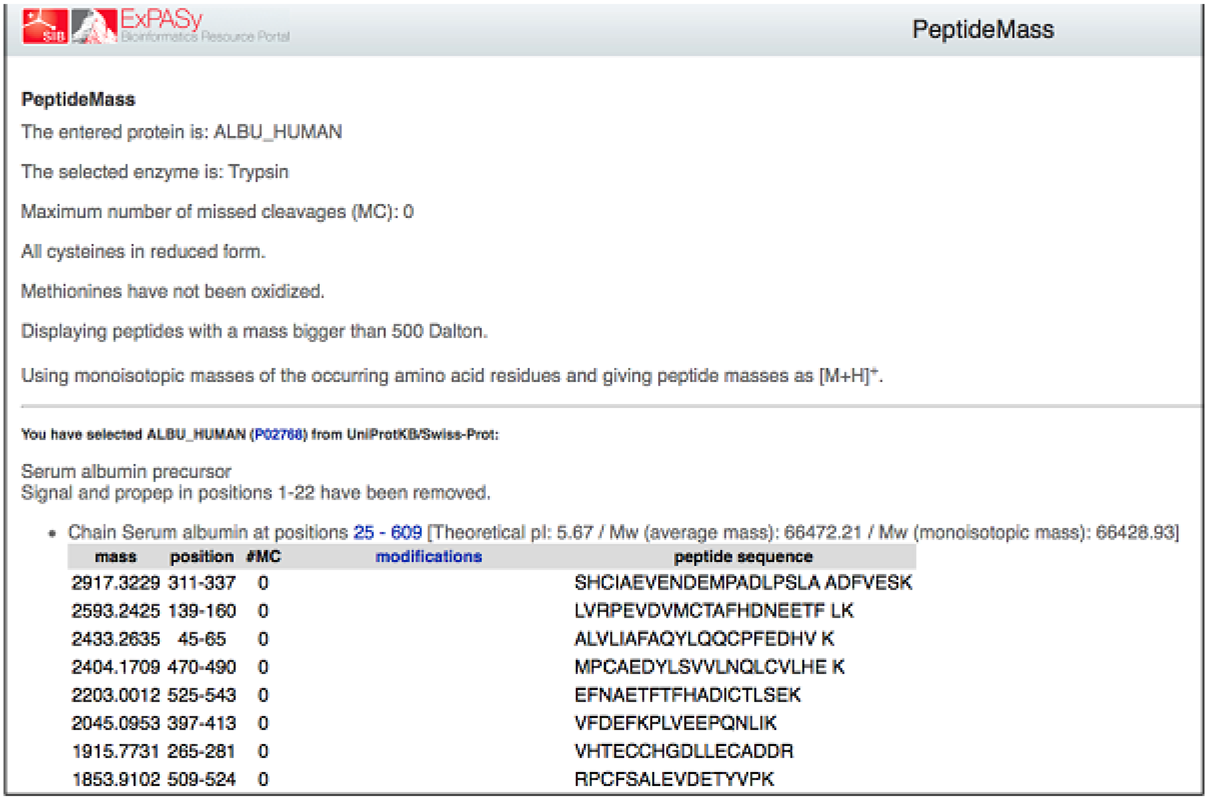
```
List of 2
 $ spectrum:List of 2
  ..$ mass     : num [1:22431] 1000 1000 1000 1001 1001 ...
  ..$ intensity: num [1:22431] 11278 11350 10879 10684 10740 ...
 $ metaData:List of 18
  ..$ file              : chr "/home/praveen/bookcode/ms/A1-0_A1.mzXML"
  ..$ scanCount         : num 1
  ..$ parentFile        :List of 1
  .. ..$ :List of 3
  .. .. ..$ fileName: chr "Z:/home/sebastian/dokumente/studium/sapdmd/dev/r-
kages/readBrukerFlexData-dev/readBrukerFlexData/inst/Examples/2010_05_19_Gi"
_truncated__
  .. .. ..$ fileType: chr "RAWData"
  .. .. ..$ fileSha1: chr "2b6113bf97a3a2c83424f71070f8de8d23bd1919"
  ..$ msInstrument      :List of 7
  .. ..$ msManufacturer: chr "Bruker Daltonics"
  .. ..$ msModel       : chr "autoFlex"
  .. ..$ msIonisation  : chr "MALDI"
  .. ..$ msMassAnalyzer: chr "TOF"
  .. ..$ msDetector    : chr "MS:1000026"
  .. ..$ software      :List of 3
  .. .. ..$ type   : chr "acquisition"
```


Pankreas_HB_L_061019_G10.M19
intensity
0e+00
1e+05
2000
4000
6000
8000
10000
mass
/data/set A - discovery leipzig/control/Pankreas_HB_L_061019_G10/0_m19/1/1SLin/fid
Pankreas_HB_L_061019_G10.M19
intensity
0
200
2000
4000
6000
8000
10000
mass
/data/set A - discovery leipzig/control/Pankreas_HB_L_061019_G10/0_m19/1/1SLin/fid
Pankreas_HB_L_061019_G10.M19
intensity
0
150
2000
4000
6000
8000
10000
mass
/data/set A - discovery leipzig/control/Pankreas_HB_L_061019_G10/0_m19/1/1SLin/fid
Pankreas_HB_L_061019_G10.M19
intensity
0.000
0.005
2000
4000
6000
8000
10000
mass
/data/set A - discovery leipzig/control/Pankreas_HB_L_061019_G10/0_m19/1/1SLin/fid



Pankreas_HB_L_061019_G10.M19

/data/set A – discovery leipzig/control/Pankreas_HB_L_061019_G10/0_m19/1/1SLin/fid



PeptideMass

The entered protein is: ALBU_HUMAN

The selected enzyme is: Trypsin

Maximum number of missed cleavages (MC): 0

All cysteines in reduced form.

Methionines have not been oxidized.

Displaying peptides with a mass bigger than 500 Dalton.

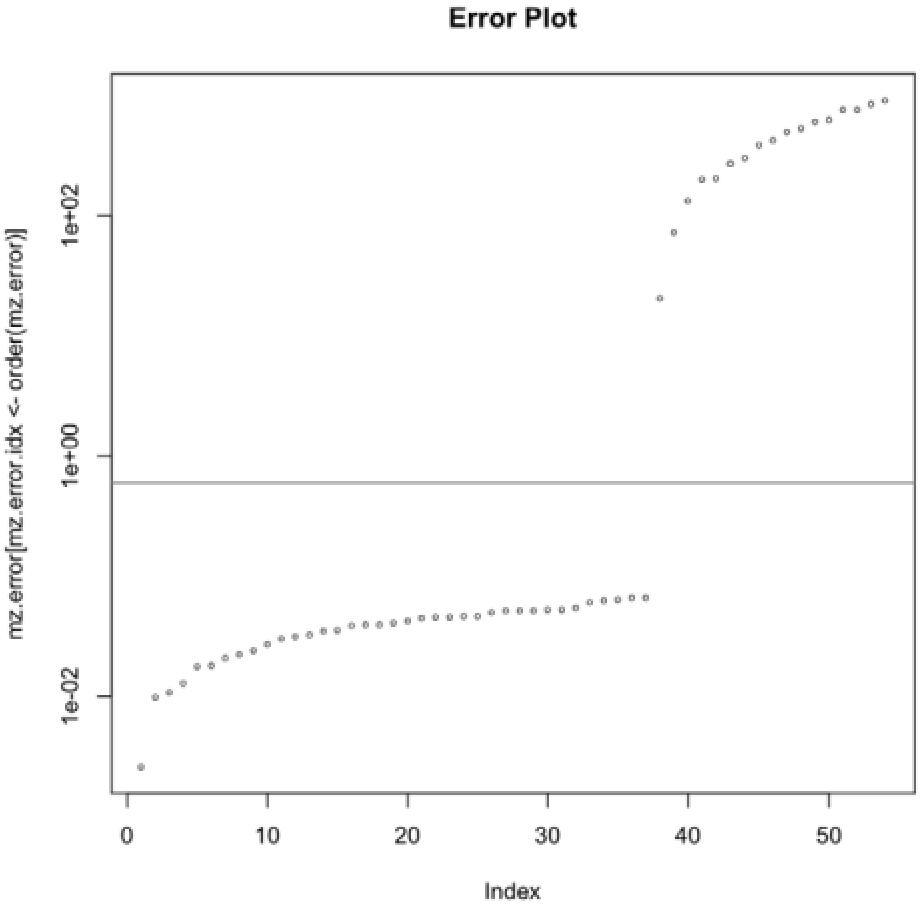
Using monoisotopic masses of the occurring amino acid residues and giving peptide masses as $[M+H]^+$.

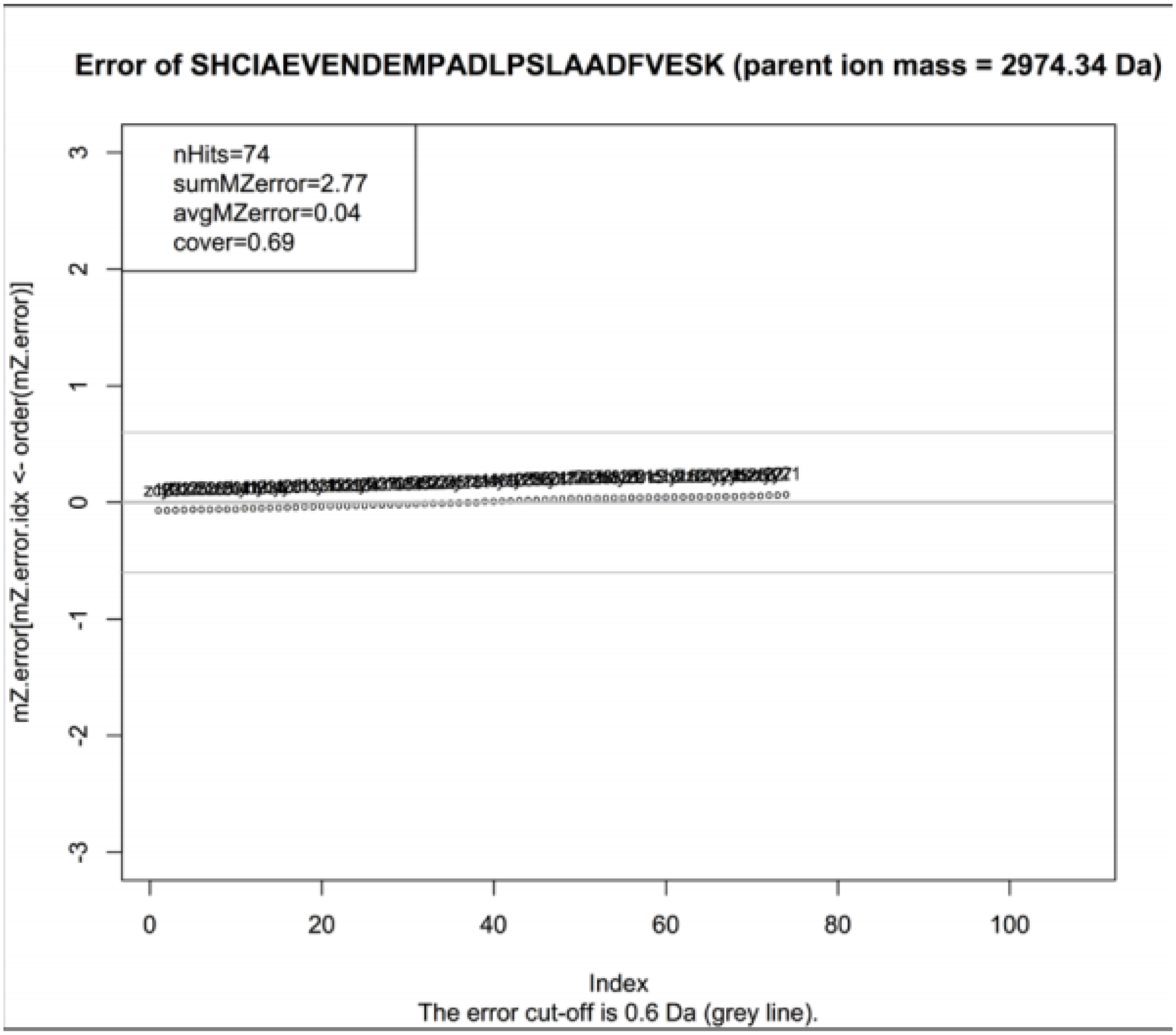
You have selected ALBU_HUMAN (P02768) from UniProtKB/Swiss-Prot:

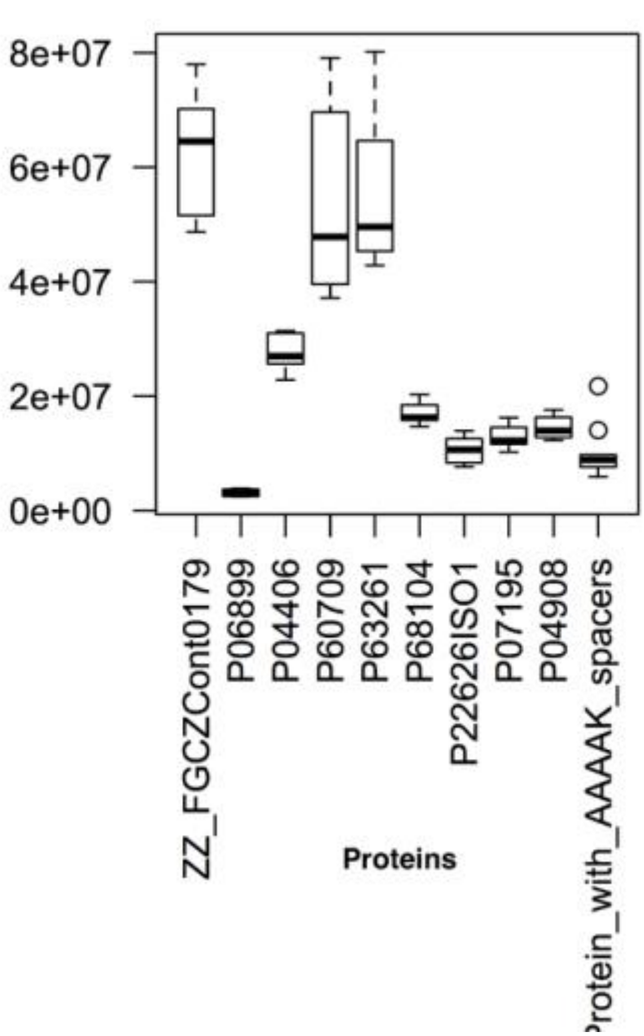
Serum albumin precursor
Signal and propep in positions 1-22 have been removed.

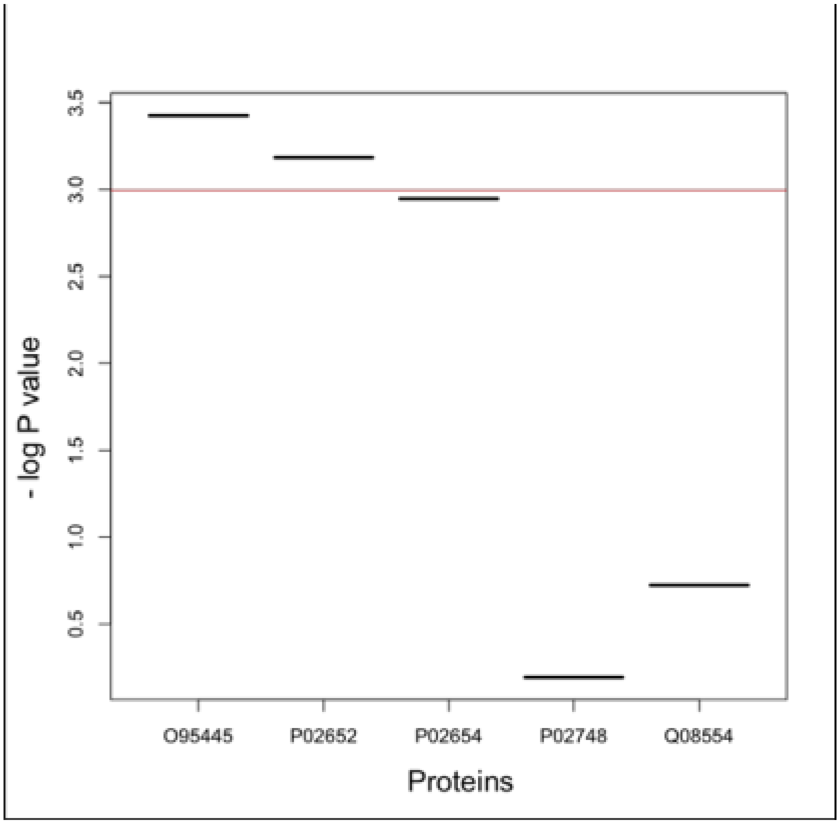
- Chain Serum albumin at positions 25 - 609 [Theoretical pI: 5.67 / Mw (average mass): 66472.21 / Mw (monoisotopic mass): 66428.93]

mass	position	#MC	modifications	peptide sequence
2917.3229	311-337	0		SHCIAEVENDEMPADLPSLA ADFVESK
2593.2425	139-160	0		LVRPEVDVMCTAFHDNEETF LK
2433.2635	45-65	0		ALVLIAFAQYLQQCPFEDHV K
2404.1709	470-490	0		MPCAEDYLSVVLNQLCVLHE K
2203.0012	525-543	0		EFNAETFTFHADICTLSEK
2045.0953	397-413	0		VFDEFKPLVEEPQNLIK
1915.7731	265-281	0		VHTECCHGDLLECADDR
1853.9102	509-524	0		RPCFSALEVDETYVPK

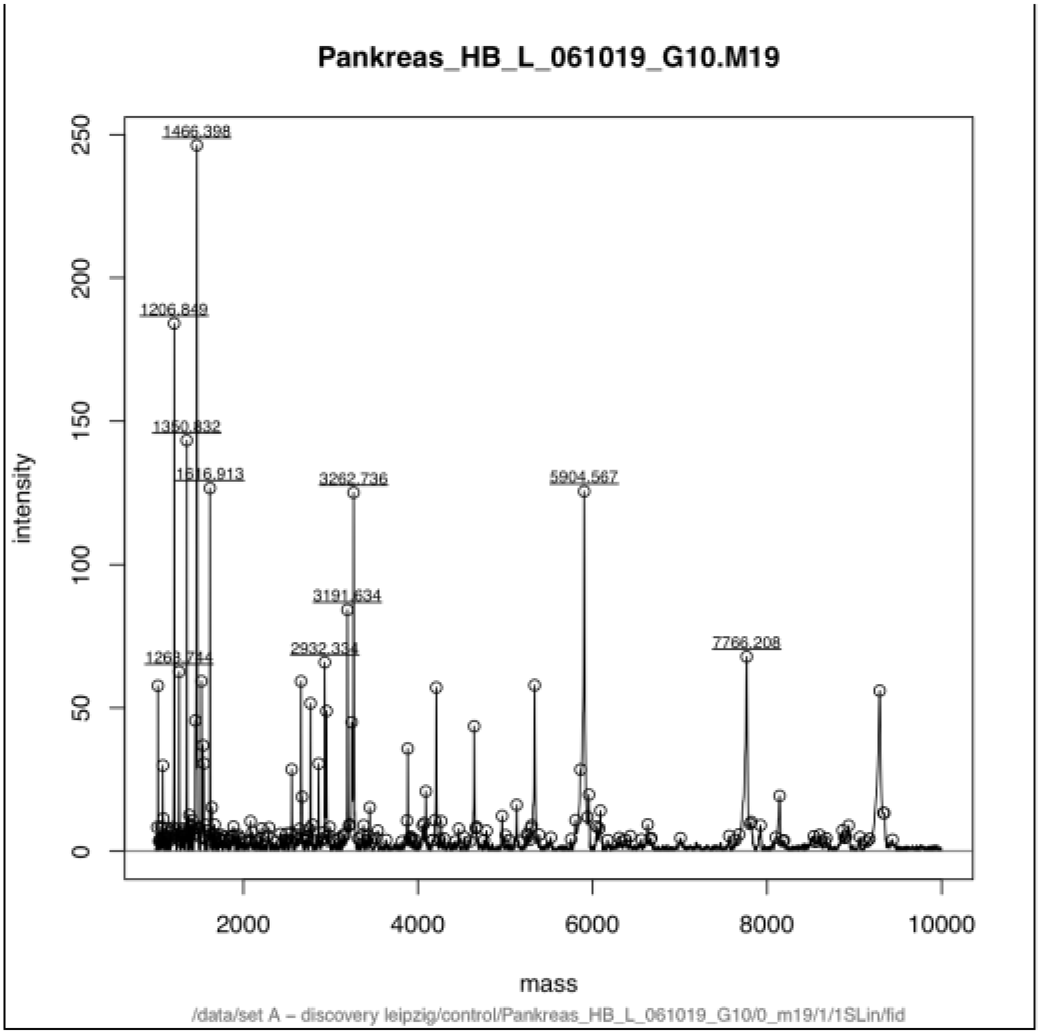

Error Plot
mz.error[mz.error.idx <- order(mz.error)]
1e+02
1e+00
1e-02
0
10
20
30
40
50
Index


Error of SHCIAEVENDEMPADLPSLAADFVESK (parent ion mass = 2974.34 Da)
nHits=74
sumMZerror=2.77
avgMZerror=0.04
cover=0.69
mZ.error[mZ.error.idx <- order(mZ.error)]
3
2
1
0
-1
-2
-3
0
20
40
60
80
100
Index
The error cut-off is 0.6 Da (grey line).

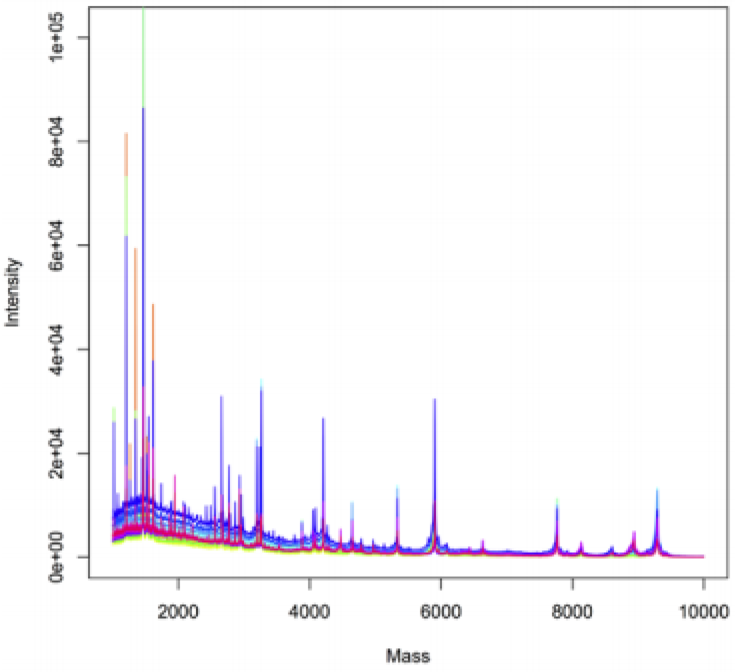

8e+07
6e+07
4e+07
2e+07
0e+00
ZZ_FGCZCont0179
P06899
P04406
P60709
P63261
P68104
P22626ISO1
P07195
P04908
Protein_with_AAAAK_spacers
Proteins

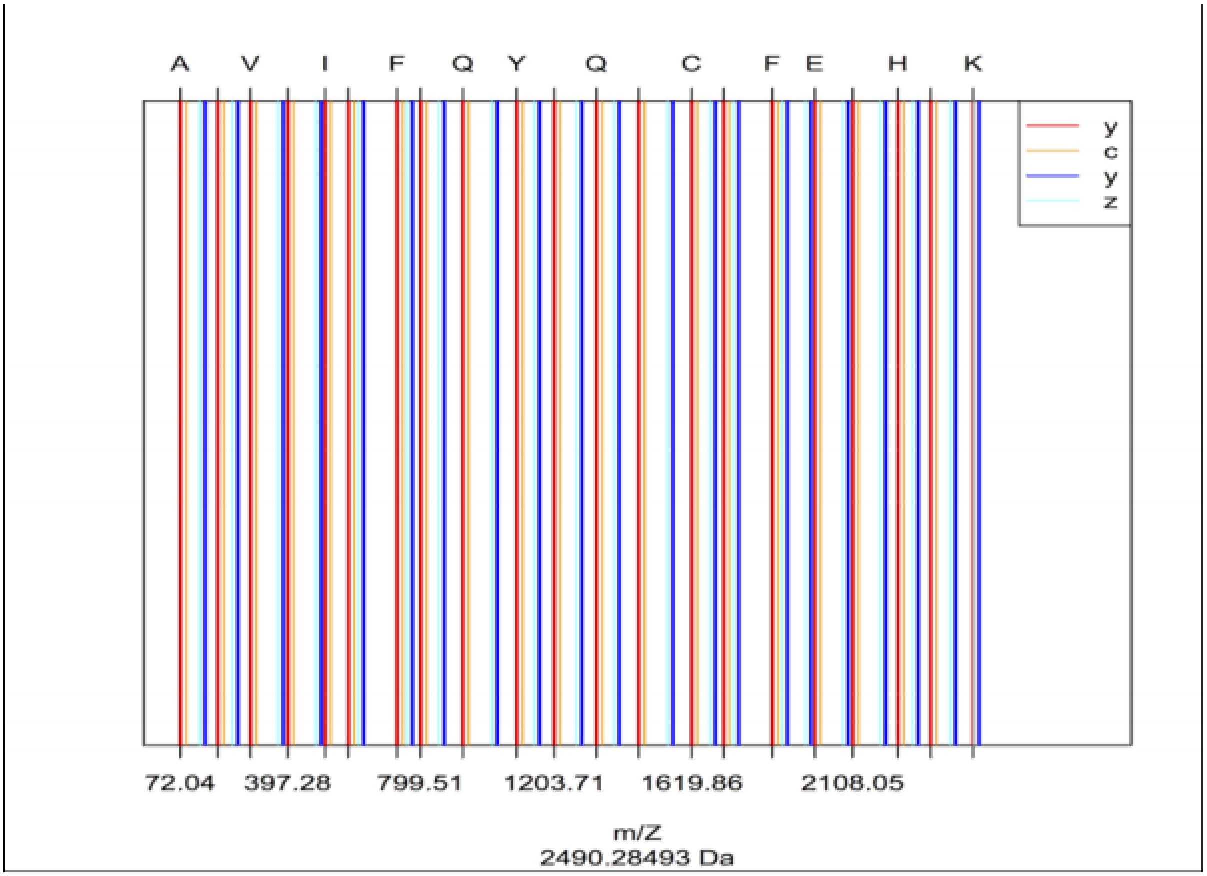


- log P value
3.5
3.0
2.5
2.0
1.5
1.0
0.5
O95445
P02652
P02654
P02748
Q08554
Proteins

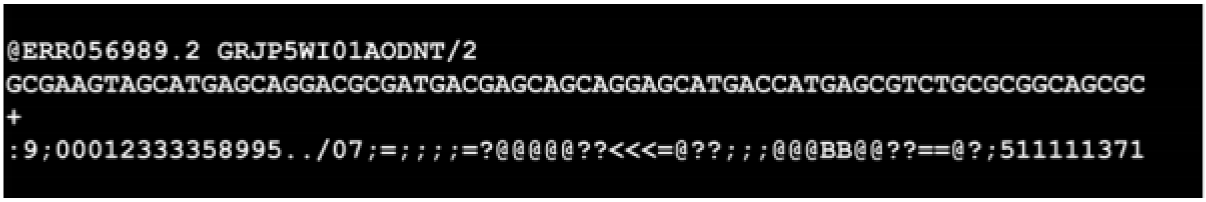

Pankreas_HB_L_061019_G10.M19

/data/set A – discovery leipzig/control/Pankreas_HB_L_061019_G10/0_m19/1/1SLin/fid


Intensity
1e+05
8e+04
6e+04
4e+04
2e+04
0e+00
2000
4000
6000
8000
10000
Mass

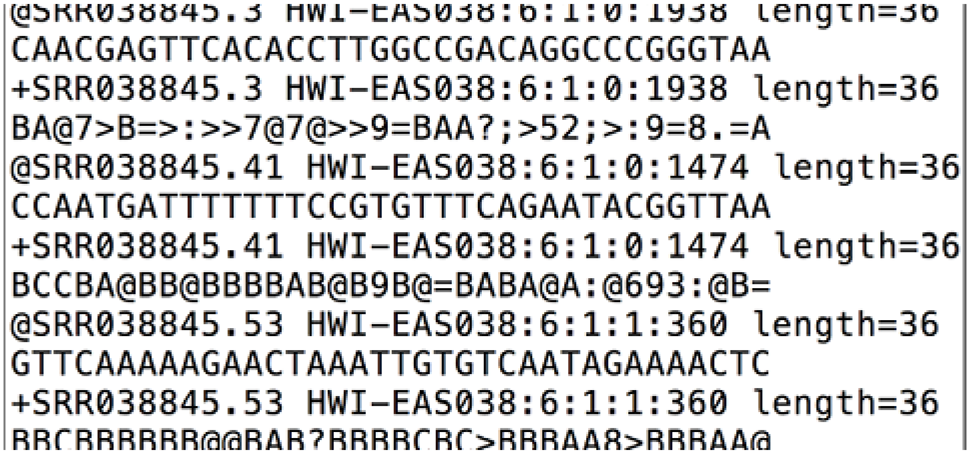


Chapter 8: Analyzing NGS Data

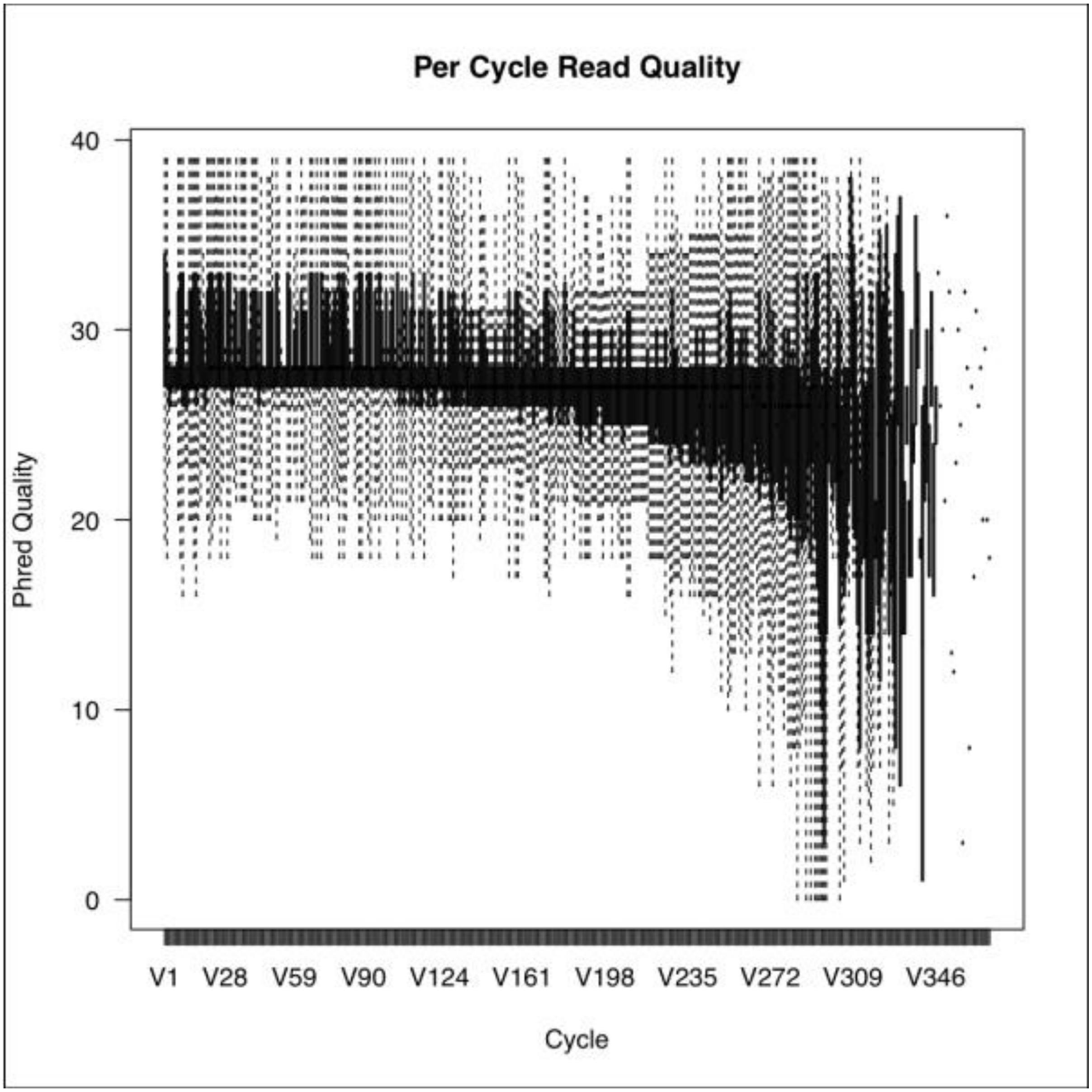


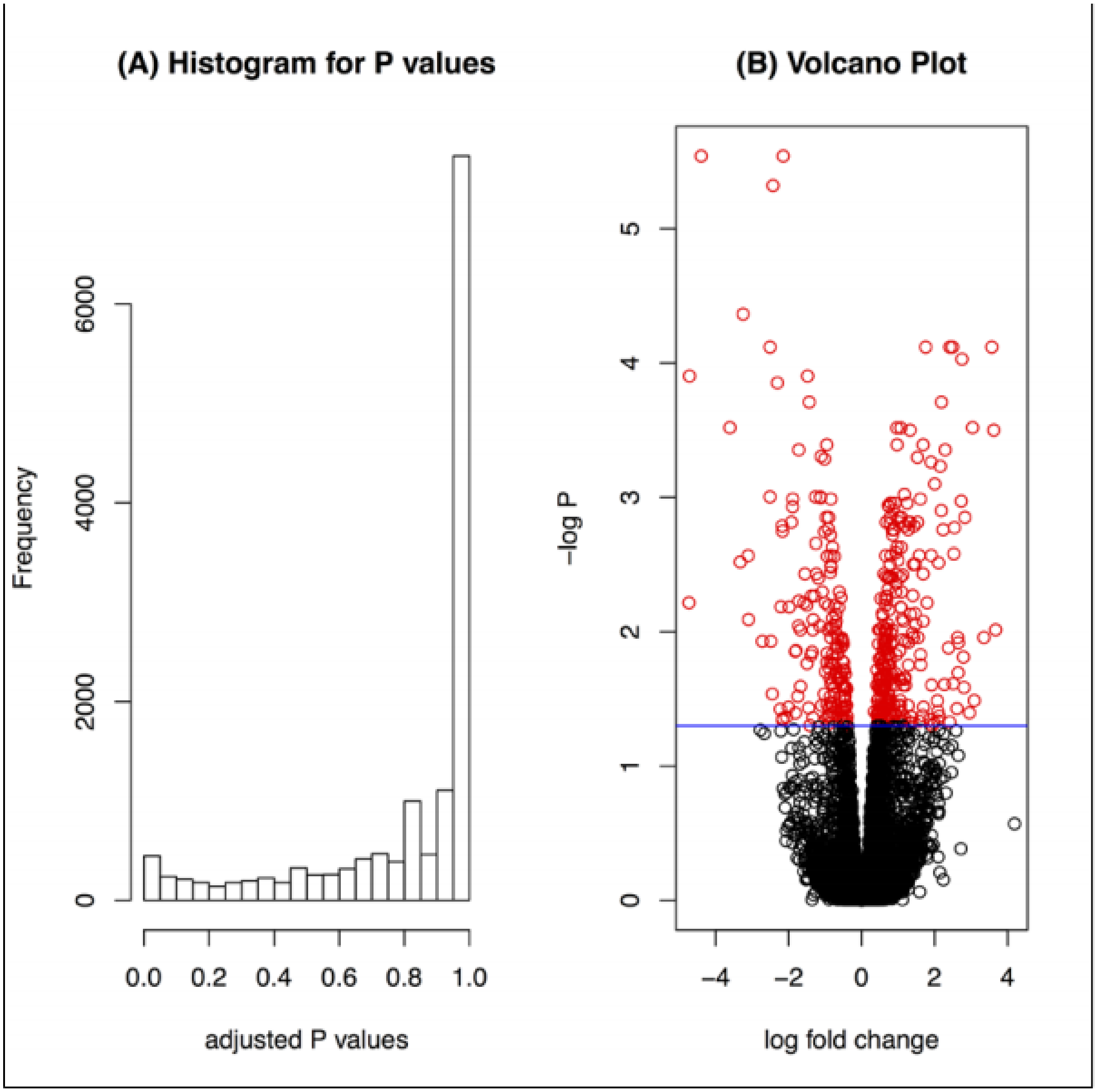
```
@ERR056989.2 GRJP5WI01AODNT/2
GCGAAGTAGCATGAGCAGGACGCGATGACGAGCAGCAGGAGCATGACCATGAGCGTCTGCGCGGCAGCGC
+
:9;00012333358995../07;=;;;;=?@@@@@??<<<=@??;;;@@@BB@@??==@?;511111371
```

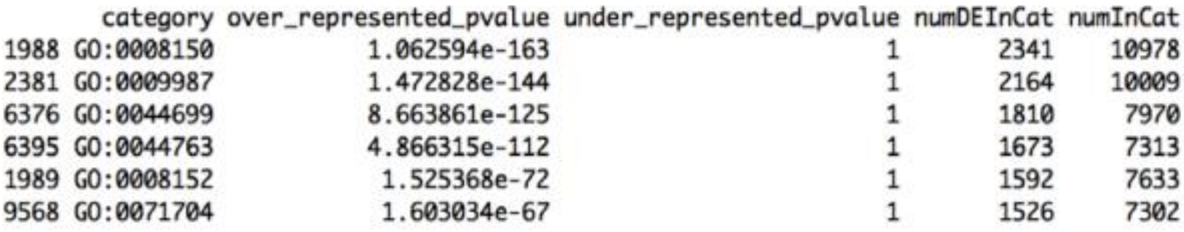
$$Q = -10\ log_{10} P$$



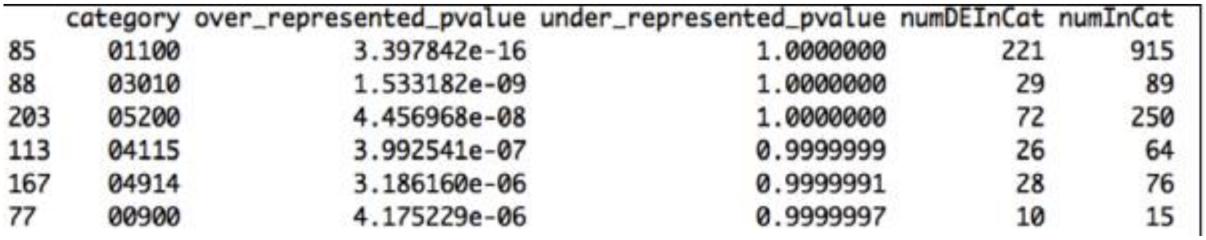
```
@SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
CAACGAGTTCACACCTTGGCCGACAGGCCCGGGTAA
+SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
BA@7>B=>:>>7@7@>>9=BAA?;>52;>:9=8.=A
@SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
CCAATGATTTTTTTCCGTGTTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBAB@B9B@=BABA@A:@693:@B=
@SRR038845.53 HWI-EAS038:6:1:1:360 length=36
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAACTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
BBCBBBBBB@@BAB?BBBBCBC>BBBAA8>BBBAA@
```


Per Cycle Read Quality
Phred Quality
40
30
20
10
0
V1
V28
V59
V90
V124
V161
V198
V235
V272
V309
V346
Cycle

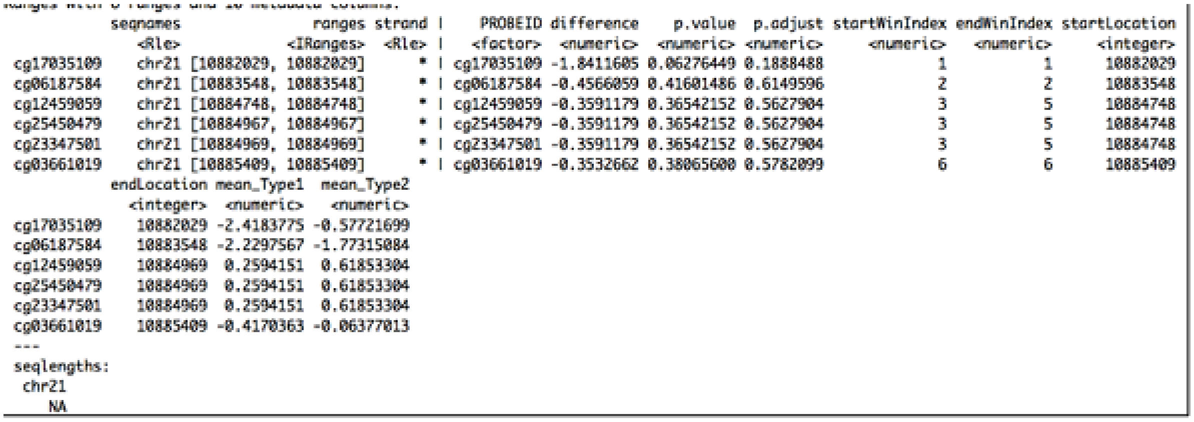




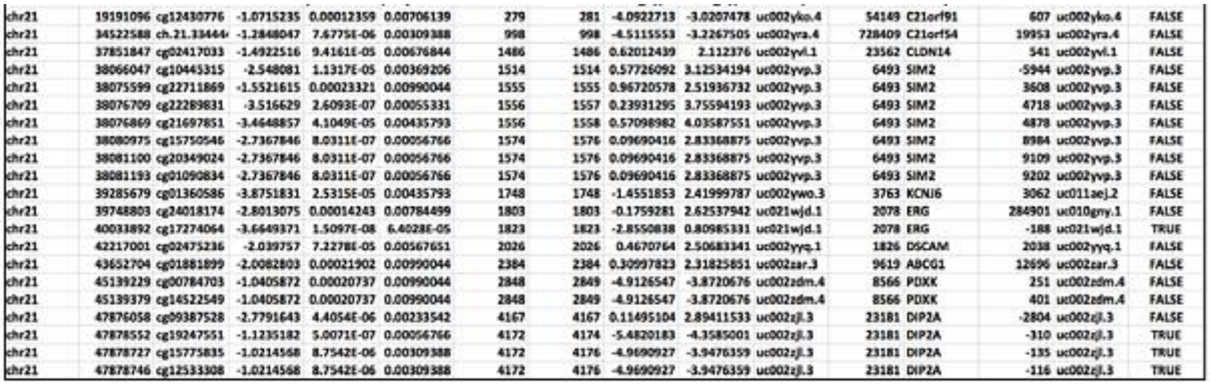
```
        category over_represented_pvalue under_represented_pvalue numDEInCat numInCat
1988 GO:0008150            1.062594e-163                        1       2341    10978
2381 GO:0009987            1.472828e-144                        1       2164    10009
6376 GO:0044699            8.663861e-125                        1       1810     7970
6395 GO:0044763            4.866315e-112                        1       1673     7313
1989 GO:0008152             1.525368e-72                        1       1592     7633
9568 GO:0071704             1.603034e-67                        1       1526     7302
```



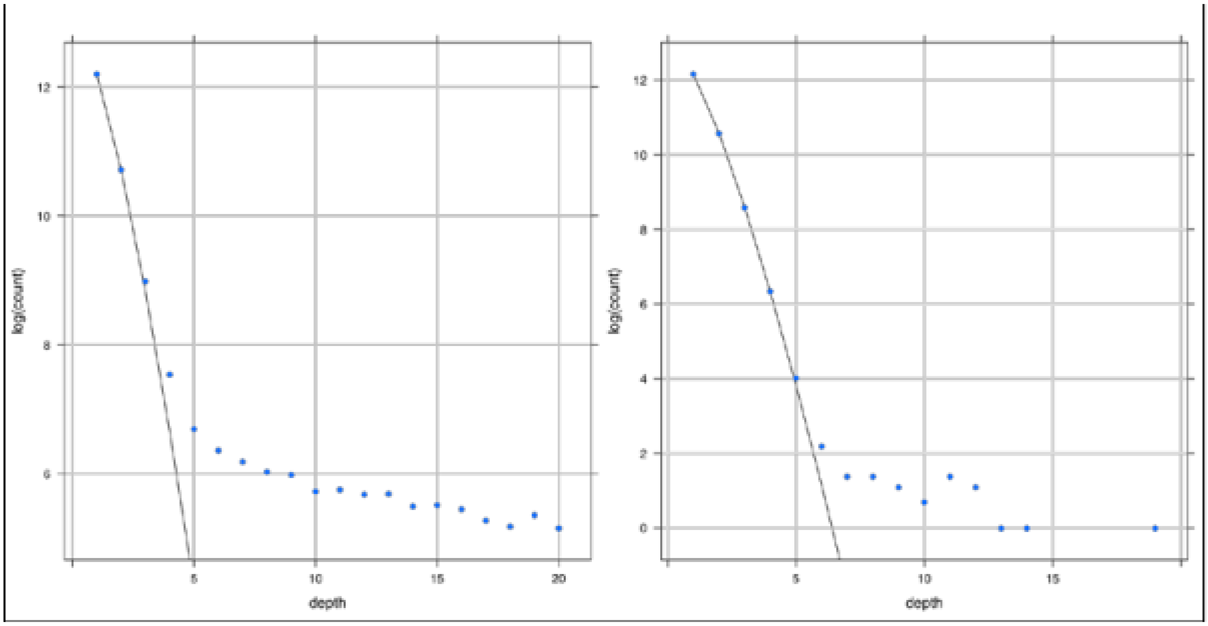
```
    category over_represented_pvalue under_represented_pvalue numDEInCat numInCat
85     01100            3.397842e-16                1.0000000        221      915
88     03010            1.533182e-09                1.0000000         29       89
203    05200            4.456968e-08                1.0000000         72      250
113    04115            3.992541e-07                0.9999999         26       64
167    04914            3.186160e-06                0.9999991         28       76
77     00900            4.175229e-06                0.9999997         10       15
```

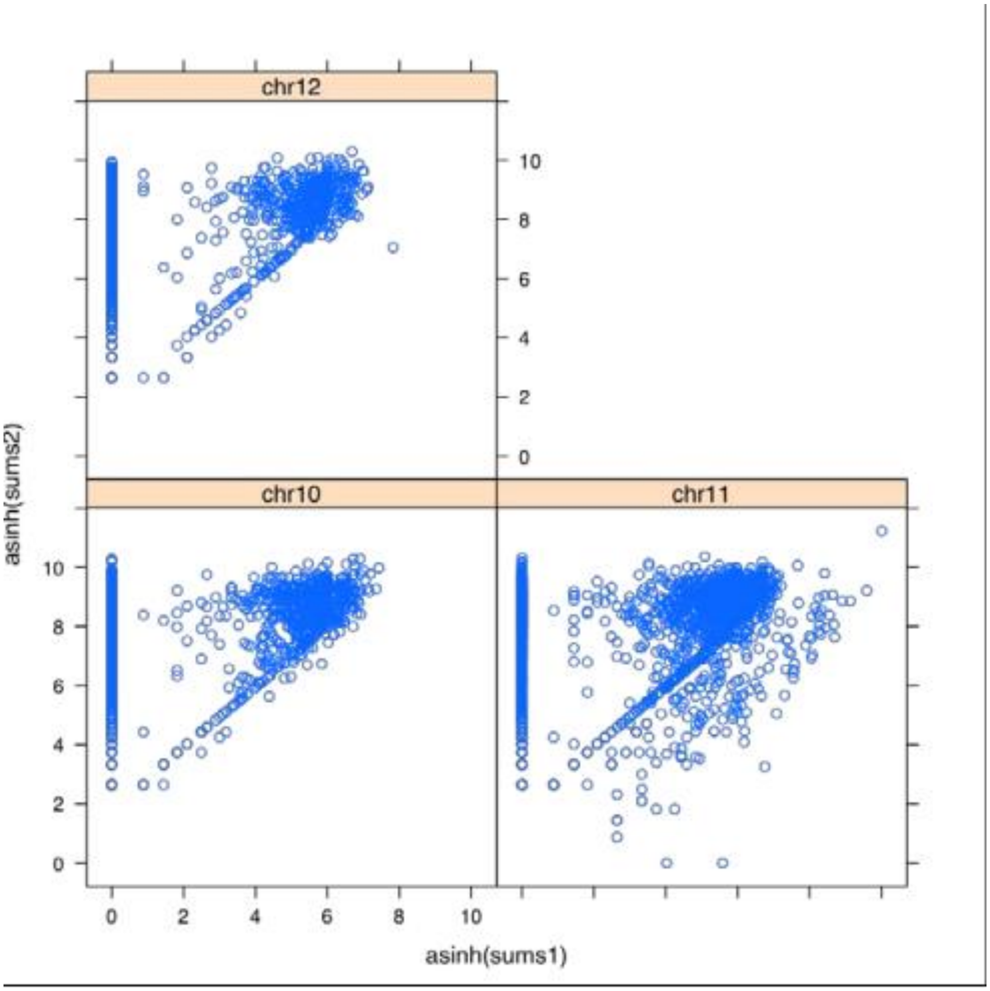


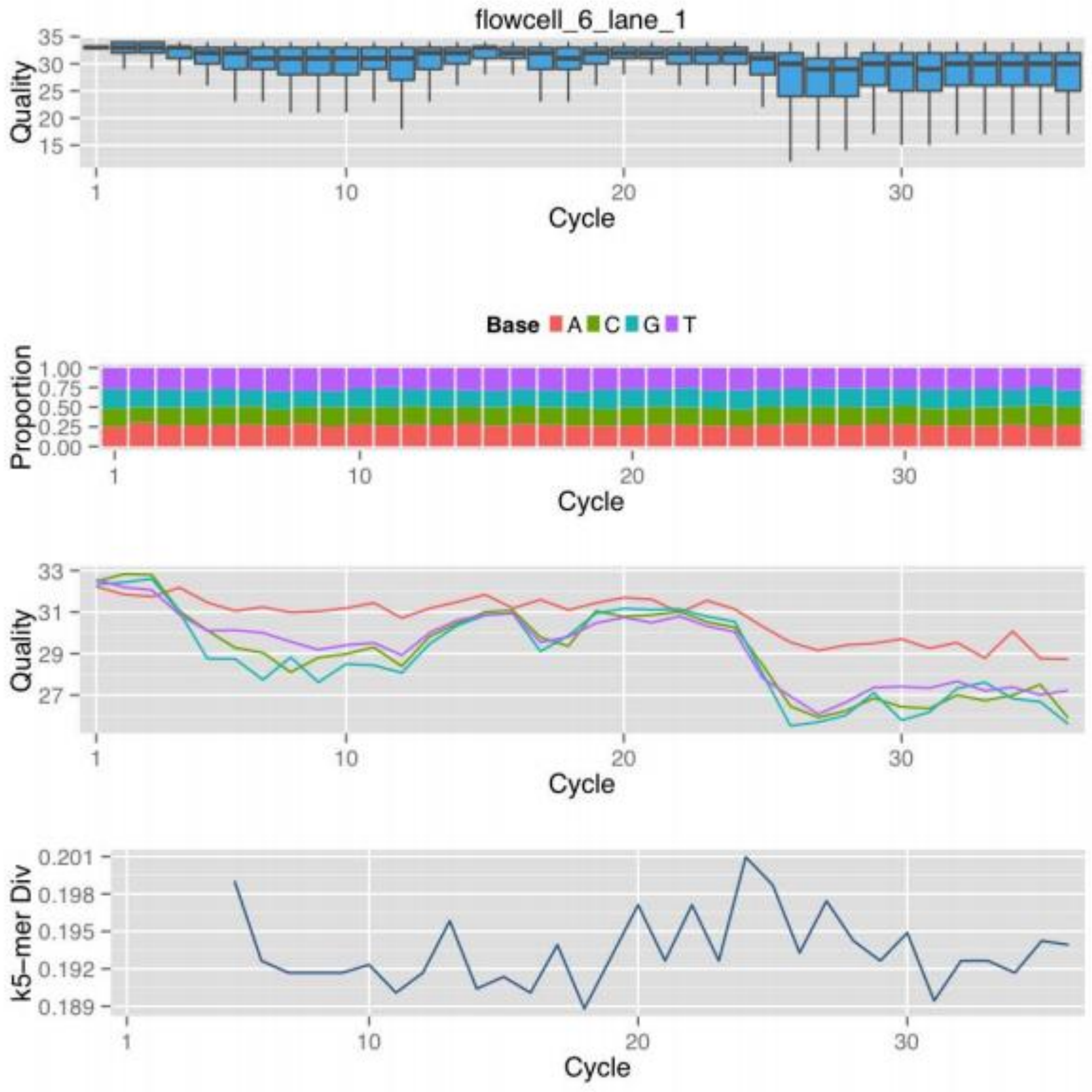
```
             seqnames                 ranges strand |    PROBEID difference    p.value  p.adjust startWinIndex endWinIndex startLocation
                <Rle>              <IRanges>  <Rle> |   <factor>  <numeric>  <numeric> <numeric>     <numeric>   <numeric>     <integer>
  cg17035109    chr21 [10882029, 10882029]      * | cg17035109 -1.8411605 0.06276449 0.1888488             1           1      10882029
  cg06187584    chr21 [10883548, 10883548]      * | cg06187584 -0.4566059 0.41601486 0.6149596             2           2      10883548
  cg12459059    chr21 [10884748, 10884748]      * | cg12459059 -0.3591179 0.36542152 0.5627904             3           5      10884748
  cg25450479    chr21 [10884967, 10884967]      * | cg25450479 -0.3591179 0.36542152 0.5627904             3           5      10884748
  cg23347501    chr21 [10884969, 10884969]      * | cg23347501 -0.3591179 0.36542152 0.5627904             3           5      10884748
  cg03661019    chr21 [10885409, 10885409]      * | cg03661019 -0.3532662 0.38065600 0.5782099             6           6      10885409
             endLocation mean_Type1  mean_Type2
               <integer>  <numeric>   <numeric>
  cg17035109    10882029 -2.4183775 -0.57721699
  cg06187584    10883548 -2.2297567 -1.77315084
  cg12459059    10884969  0.2594151  0.61853304
  cg25450479    10884969  0.2594151  0.61853304
  cg23347501    10884969  0.2594151  0.61853304
  cg03661019    10885409 -0.4170363 -0.06377013
  ---
  seqlengths:
   chr21
      NA
```

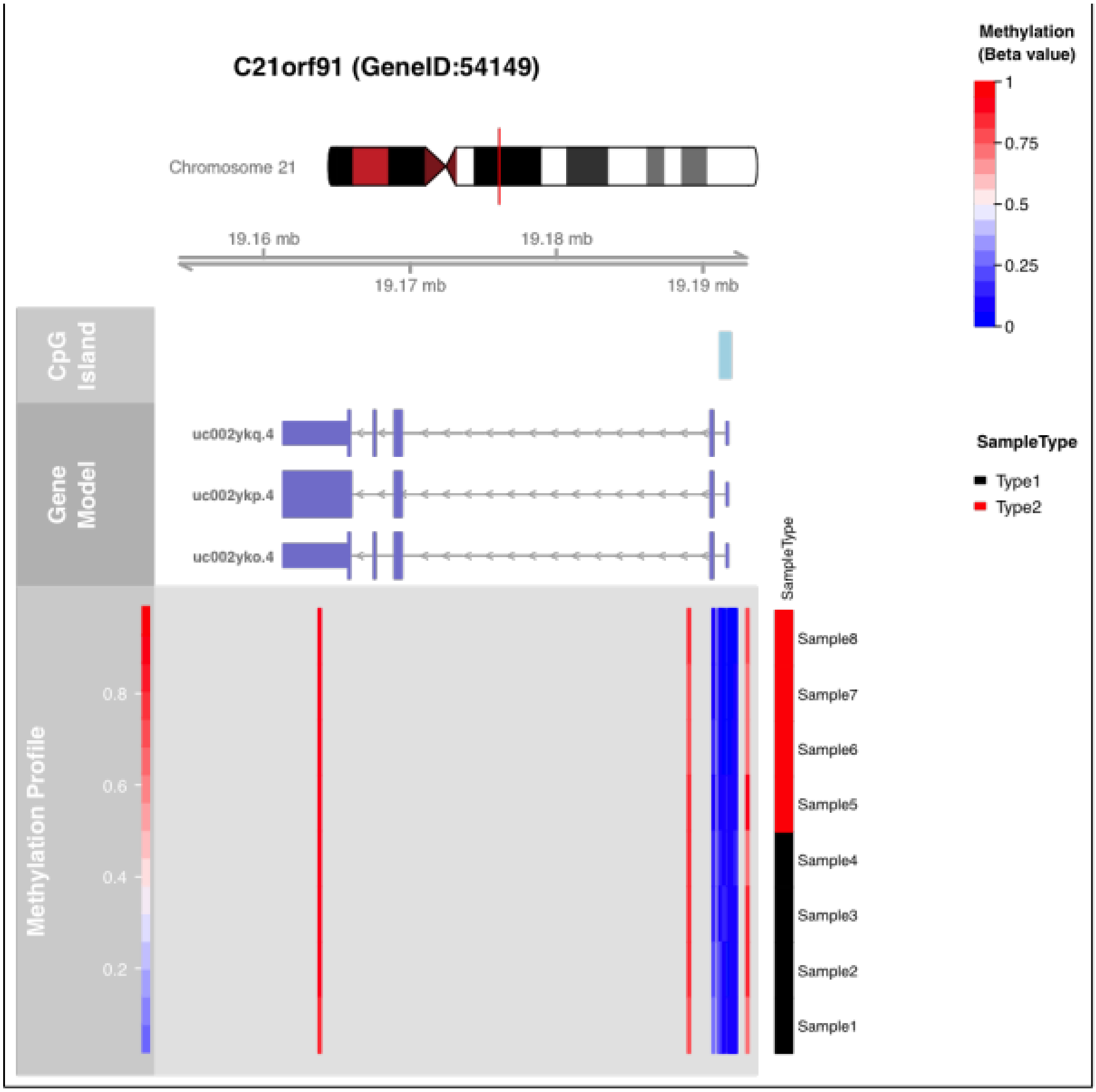


chr21	19191096	cg12430776	-1.0715235	0.00012359	0.00706139	279	281	-4.0922713	-3.0207478	uc002yko.4	54149	C21orf91	607	uc002yko.4	FALSE
chr21	34522588	ch.21.33444	-1.2848047	7.6775E-06	0.00309388	998	998	-4.5115553	-3.2267505	uc002yra.4	728409	C21orf54	19953	uc002yra.4	FALSE
chr21	37851847	cg02417033	-1.4922516	9.4161E-05	0.00676844	1486	1486	0.62012439	2.112376	uc002yvl.1	23562	CLDN14	541	uc002yvl.1	FALSE
chr21	38066047	cg10445315	-2.548081	1.1317E-05	0.00369206	1514	1514	0.57726092	3.12534194	uc002yvp.3	6493	SIM2	-5944	uc002yvp.3	FALSE
chr21	38075599	cg22711869	-1.5521615	0.00023321	0.00990044	1555	1555	0.96720578	2.51936732	uc002yvp.3	6493	SIM2	3608	uc002yvp.3	FALSE
chr21	38076709	cg22289831	-3.516629	2.6093E-07	0.00055331	1556	1557	0.23931295	3.75594193	uc002yvp.3	6493	SIM2	4718	uc002yvp.3	FALSE
chr21	38076869	cg21697851	-3.4648857	4.1049E-05	0.00435793	1556	1558	0.57098982	4.03587551	uc002yvp.3	6493	SIM2	4878	uc002yvp.3	FALSE
chr21	38080975	cg15750546	-2.7367846	8.0311E-07	0.00056766	1574	1576	0.09690416	2.83368875	uc002yvp.3	6493	SIM2	8984	uc002yvp.3	FALSE
chr21	38081100	cg20349024	-2.7367846	8.0311E-07	0.00056766	1574	1576	0.09690416	2.83368875	uc002yvp.3	6493	SIM2	9109	uc002yvp.3	FALSE
chr21	38081193	cg01090834	-2.7367846	8.0311E-07	0.00056766	1574	1576	0.09690416	2.83368875	uc002yvp.3	6493	SIM2	9202	uc002yvp.3	FALSE
chr21	39285679	cg01360586	-3.8751831	2.5315E-05	0.00435793	1748	1748	-1.4551853	2.41999787	uc002ywo.3	3763	KCNJ6	3062	uc011aej.2	FALSE
chr21	39748803	cg24018174	-2.8013075	0.00014243	0.00784499	1803	1803	-0.1759281	2.62537942	uc021wjd.1	2078	ERG	284901	uc010gny.1	FALSE
chr21	40033892	cg17274064	-3.6649371	1.5097E-08	6.4028E-05	1823	1823	-2.8550838	0.80985331	uc021wjd.1	2078	ERG	-188	uc021wjd.1	TRUE
chr21	42217001	cg02475236	-2.039757	7.2278E-05	0.00567651	2026	2026	0.4670764	2.50683341	uc002yyq.1	1826	DSCAM	2038	uc002yyq.1	FALSE
chr21	43652704	cg01881899	-2.0082803	0.00021902	0.00990044	2384	2384	0.30997823	2.31825851	uc002zar.3	9619	ABCG1	12696	uc002zar.3	FALSE
chr21	45139229	cg00784703	-1.0405872	0.00020737	0.00990044	2848	2849	-4.9126547	-3.8720676	uc002zdm.4	8566	PDXK	251	uc002zdm.4	FALSE
chr21	45139379	cg14522549	-1.0405872	0.00020737	0.00990044	2848	2849	-4.9126547	-3.8720676	uc002zdm.4	8566	PDXK	401	uc002zdm.4	FALSE
chr21	47876058	cg09387528	-2.7791643	4.4054E-06	0.00233542	4167	4167	0.11495104	2.89411533	uc002zjl.3	23181	DIP2A	-2804	uc002zjl.3	FALSE
chr21	47878552	cg19247551	-1.1235182	5.0071E-07	0.00056766	4172	4174	-5.4820183	-4.3585001	uc002zjl.3	23181	DIP2A	-310	uc002zjl.3	TRUE
chr21	47878727	cg15775835	-1.0214568	8.7542E-06	0.00309388	4172	4176	-4.9690927	-3.9476359	uc002zjl.3	23181	DIP2A	-135	uc002zjl.3	TRUE
chr21	47878746	cg12533308	-1.0214568	8.7542E-06	0.00309388	4172	4176	-4.9690927	-3.9476359	uc002zjl.3	23181	DIP2A	-116	uc002zjl.3	TRUE

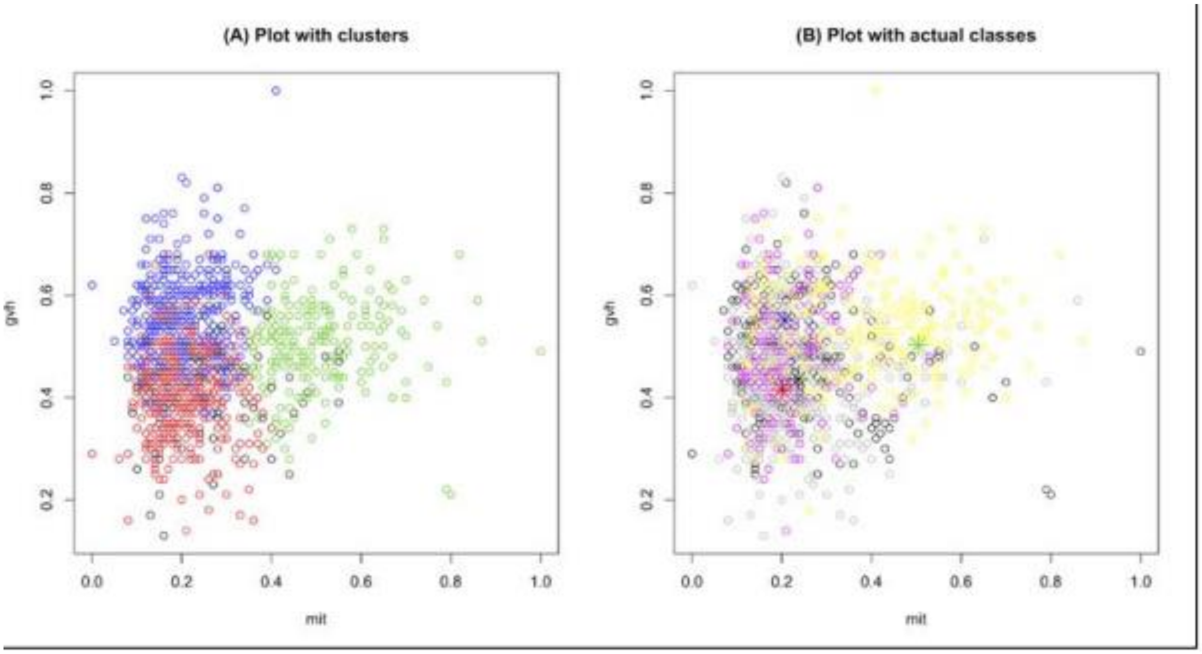

log(count)
depth
12
10
8
6
5
10
15
20
log(count)
depth
12
10
8
6
4
2
0
5
10
15


chr12
chr10
chr11
asinh(sums2)
asinh(sums1)
0
2
4
6
8
10

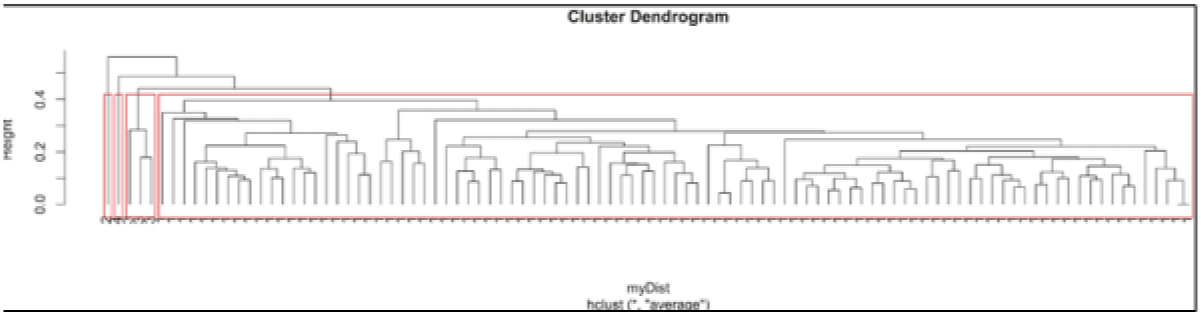

flowcell_6_lane_1
Quality
35
30
25
20
15
1
10
20
30
Cycle
Base A C G T
Proportion
1.00
0.75
0.50
0.25
0.00
1
10
20
30
Cycle
Quality
33
31
29
27
1
10
20
30
Cycle
k5–mer Div
0.201
0.198
0.195
0.192
0.189
1
10
20
30
Cycle

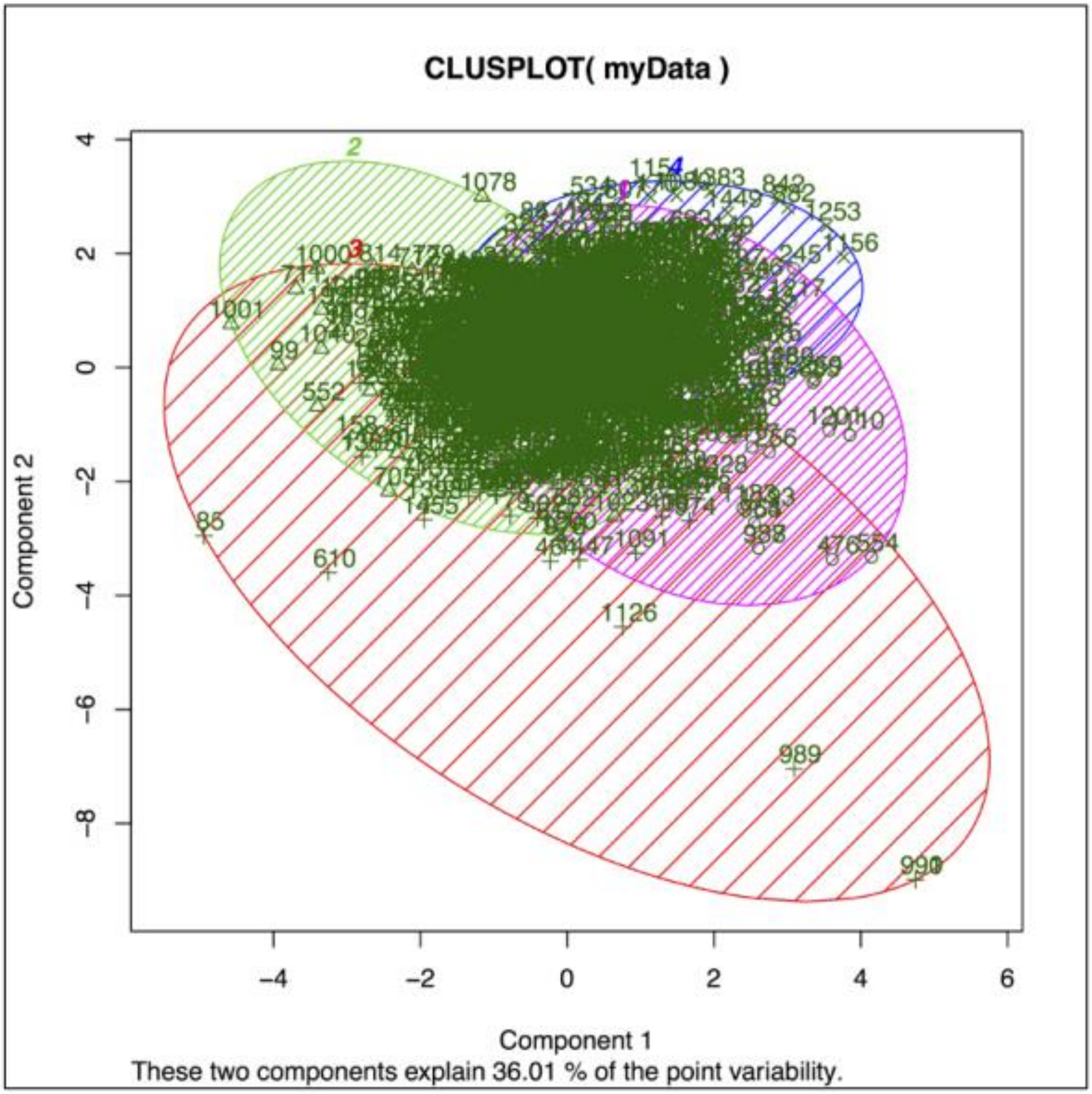


Chapter 9: Machine Learning in Bioinformatics

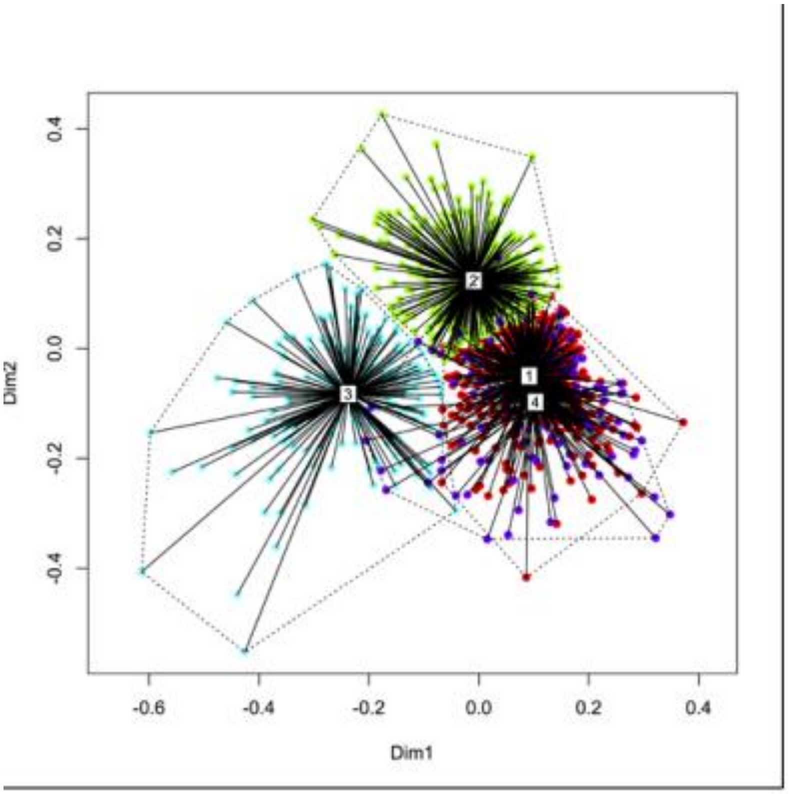

(A) Plot with clusters

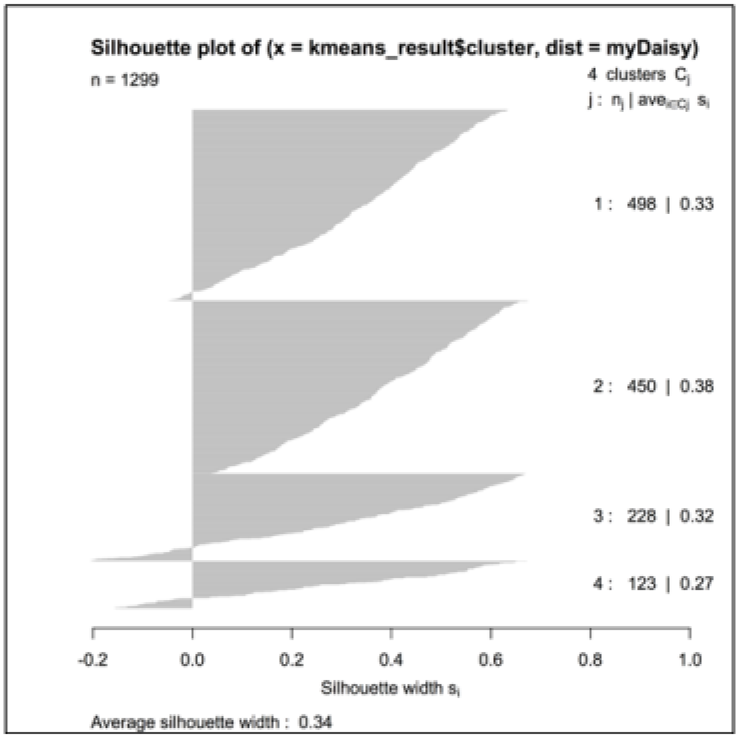
(B) Plot with actual classes

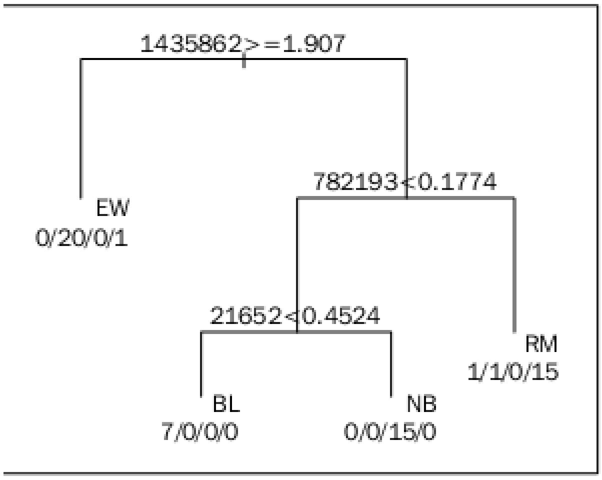

Cluster Dendrogram

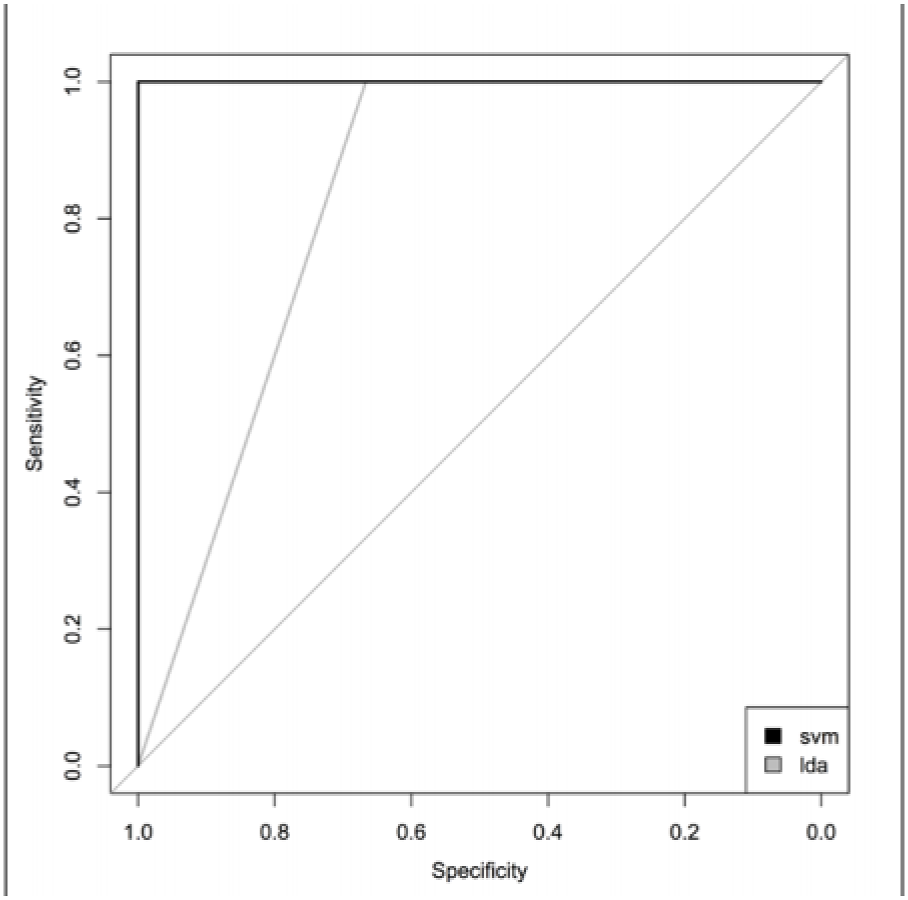


These two components explain 36.01 % of the point variability.


Dim2
0.4
0.2
0.0
-0.2
-0.4
Dim1
-0.6
-0.4
-0.2
0.0
0.2
0.4
1
2
3
4


Silhouette plot of (x = kmeans_result$cluster, dist = myDaisy)
n = 1299
4 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 498 | 0.33
2 : 450 | 0.38
3 : 228 | 0.32
4 : 123 | 0.27
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.34


1435862>=1.907
782193<0.1774
EW
0/20/0/1
21652<0.4524
RM
1/1/0/15
BL
7/0/0/0
NB
0/0/15/0


Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
■ svm
■ lda