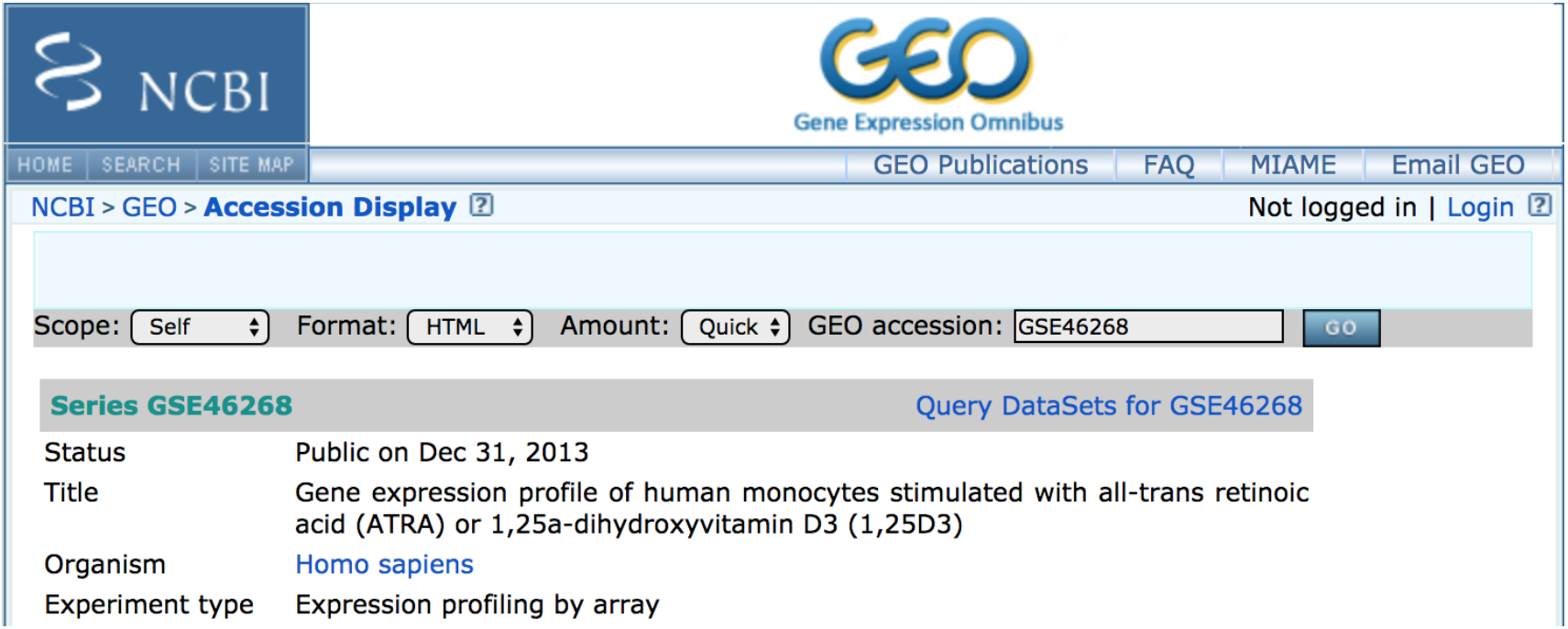
Processing NCBI Geo Databases using R

Bioinformatica
Febrero-Junio 2020

GEO Database

Data type	Description
GEO Platform (GPL)	These files describe a particular type of microarray. They are annotation files.
GEO Sample (GSM)	Files that contain all the data from the use of a single chip. For each gene there will be multiple scores including the main one, held in the VALUE column.
GEO Series (GSE)	Lists of GSM files that together form a single experiment.
GEO Dataset (GDS)	These are curated files that hold a summarized combination of a GSE file and its GSM files. They contain normalized expression levels for each gene from each sample (i.e. just the VALUE field from the GSM file).

GEO Database



Overall design — Monocytes derived from four independent healthy blood donors that were stimulated with control (CTRL), ATRA or 1,25D3 at 10-8M for 18 hours.

Contributor(s) — Wheelwright M, Kim EW, DeLeon A, Krutzik SR, Liu PT

Citation(s) — Wheelwright M, Kim EW, Inkeles MS, De Leon A et al. All-trans retinoic acid-triggered antimicrobial activity against Mycobacterium tuberculosis is dependent on NPC2. *J Immunol* 2014 Mar 1;192(5):2280-2290. PMID: 24501203

GEO Database

Platforms (1)	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
Samples (12) Less...	GSM1127890 monocyte from donor 1 stimulated with CTRL
	GSM1127891 monocyte from donor 2 stimulated with CTRL
	GSM1127892 monocyte from donor 3 stimulated with CTRL
	GSM1127893 monocyte from donor 4 stimulated with CTRL
	GSM1127894 monocyte from donor 1 stimulated with ATRA
	GSM1127895 monocyte from donor 2 stimulated with ATRA
	GSM1127896 monocyte from donor 3 stimulated with ATRA
	GSM1127897 monocyte from donor 4 stimulated with ATRA
	GSM1127898 monocyte from donor 1 stimulated with 1,25D
	GSM1127899 monocyte from donor 2 stimulated with 1,25D
	GSM1127900 monocyte from donor 3 stimulated with 1,25D
	GSM1127901 monocyte from donor 4 stimulated with 1,25D

Sample GSM1127890 Query DataSets for GSM1127890

Status	Public on Dec 31, 2013
Title	monocyte from donor 1 stimulated with CTRL
Sample type	RNA

Sample GSM1127894 Query DataSets for GSM1127894

Status	Public on Dec 31, 2013
Title	monocyte from donor 1 stimulated with ATRA
Sample type	RNA

Sample GSM1127898 Query DataSets for GSM1127898

Status	Public on Dec 31, 2013
Title	monocyte from donor 1 stimulated with 1,25D
Sample type	RNA

Bioconductor

```
source("http://bioconductor.org/biocLite.R")
biocLite("GEOquery")
biocLite("limma")
biocLite("Biobase")
biocLite("affy")
```

Bioconductor

Analysis & comprehension of high-throughput genomic data

- 15 years old; 1211 packages; widely used
- Sequencing (RNAseq, ChIPseq, variants, copy number, ...), microarrays, flow cytometery, proteomics, ...
- http://bioconductor.org, https://support.bioconductor.org

Themes

- Interoperable – classes to work with genome-scale data, shared (where possible!) across packages

Bioconductor

Bioconductor: *GenomicRanges*

```
> gr = exons(TxDb.Hsapiens.UCSC.hg19.knownGene); gr
GRanges with 289969 ranges and 1 metadata column:
            seqnames              ranges strand |   exon_id
               <Rle>           <IRanges>  <Rle> | <integer>
        [1]     chr1       [11874, 12227]      + |         1
        [2]     chr1       [12595, 12721]      + |         2
        [3]     chr1       [12613, 12721]      + |         3
        ...      ...                  ...    ... ...       ...
[289967]     chrY [59358329, 59359508]      - |    277748
[289968]     chrY [59360007, 59360115]      - |    277749
[289969]     chrY [59360501, 59360854]      - |    277750
---
seqinfo: 93 sequences (1 circular) from hg19 genome
```

GRanges
- `length(gr); gr[1:5]`
- `seqnames(gr)`
- `start(gr)`
- `end(gr)`
- `width(gr)`
- `strand(gr)`

DataFrame
- `mcols(gr)`
- `gr$exon_id`

Seqinfo
- `seqlevels(gr)`
- `seqlengths(gr)`
- `genome(gr)`

- Data: aligned reads, called peaks, SNP locations, CNVs, ...
- Annotation: gene models, variants, regulatory regions, ...
- `findOverlaps()`, `nearest()`, and many other useful range-based operations.

GEOquery

The command `getGEO()` is a function from the packages `GEOquery` (S. Davis and Meltzer 2007) that can download data directly from the GEO database http://www.ncbi.nlm.nih.gov/geo/ .

```
getGEO(GEO = NULL, filename = NULL, destdir = tempdir(), GSElimits=NULL,
GSEMatrix=TRUE,AnnotGPL=FALSE,getGPL=TRUE)
```

Format name	Format
SOFT	Simple Omnibus Format in Text.
MINiML	(MIAME Notation in Markup Language - XML format
Matrix	spreadsheet containing the final, normalized values that are comparable across rows and Samples

gset <- getGEO ("GSE46268" , GSEMatrix =TRUE, destdir="/Users/emartinez/Clases/Bioinformatica/NCBI_GEO")

En caso de error, consultar este link.
https://support.bioconductor.org/p/105121/

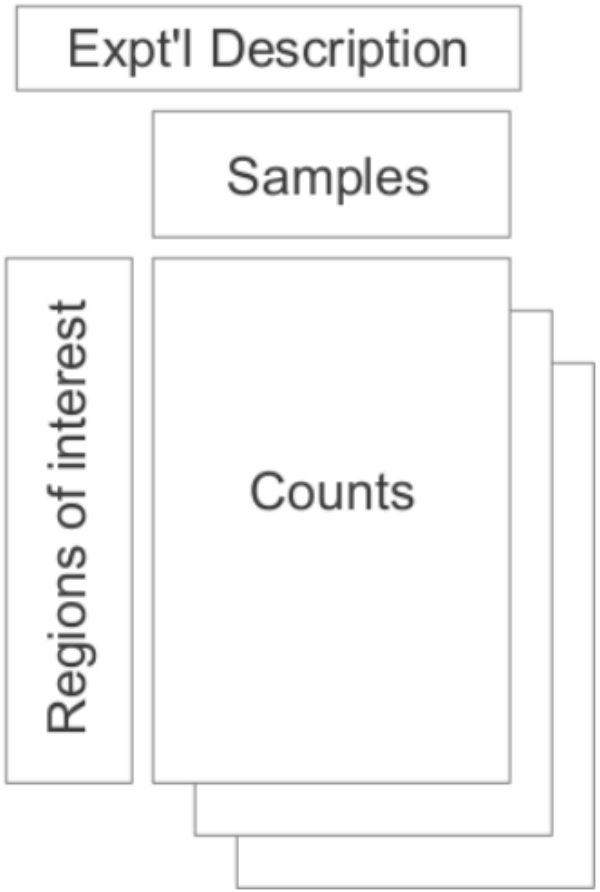
GEOquery

```
> gset
$GSE46268_series_matrix.txt.gz
ExpressionSet (storageMode: lockedEnvironment)
assayData: 54675 features, 12 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: GSM1127890 GSM1127891 ... GSM1127901 (12 total)
  varLabels: title geo_accession ... data_row_count (38 total)
  varMetadata: labelDescription
featureData
  featureNames: 1007_s_at 1053_at ... AFFX-TrpnX-M_at (54675 total)
  fvarLabels: ID GB_ACC ... Gene Ontology Molecular Function (16 total)
  fvarMetadata: Column Description labelDescription
experimentData: use 'experimentData(object)'
Annotation: GPL570
> class(gset)
[1] "list"
> names(gset)
[1] "GSE46268_series_matrix.txt.gz"
```

gset <- gset[[1]]

Bioconductor

Bioconductor: *SummarizedExperiment*



Regions of interest × samples

- `assay()` – `matrix`, e.g., counts of reads overlapping regions of interest.
- `rowData()` – regions of interest as `GRanges` or `GRangesList`
- `colData()` – `DataFrame` describing samples.

GEOquery

```
class(gset)
```

```
[1] "ExpressionSet"
attr(,"package")
[1] "Biobase"
```

```
length(gset)
```

```
[1] 1
```

```
slotNames(gset)
```

```
[1] "experimentData"    "assayData"         "phenoData"
[4] "featureData"       "annotation"        "protocolData"
[7] ".__classVersion__"
```

```
> class(pData(phenoData(gset)))
[1] "data.frame"
> class(gset@phenoData@data)
[1] "data.frame"
> dim(pData(phenoData(gset)))
[1] 12 38
> dim(gset@phenoData@data)
[1] 12 38
```

GEOquery: phenoData

```
> colnames(pData(phenoData(gset)))
 [1] "title"                 "geo_accession"           "status"                  "submission_date"         "last_update_date"        "type"
 [7] "channel_count"         "source_name_ch1"         "organism_ch1"            "characteristics_ch1"     "characteristics_ch1.1"   "characteristics_ch1.2"
[13] "characteristics_ch1.3" "treatment_protocol_ch1"  "growth_protocol_ch1"     "molecule_ch1"            "extract_protocol_ch1"    "label_ch1"
[19] "label_protocol_ch1"    "taxid_ch1"               "hyb_protocol"            "scan_protocol"           "data_processing"         "platform_id"
[25] "contact_name"          "contact_email"           "contact_phone"           "contact_laboratory"      "contact_department"      "contact_institute"
[31] "contact_address"       "contact_city"            "contact_state"           "contact_zip/postal_code" "contact_country"         "supplementary_file"
[37] "supplementary_file.1"  "data_row_count"
```

```
pData(phenoData(gset))[ , c(12,13)]
```

	characteristics_ch1.2	characteristics_ch1.3
GSM1127890	individual: donor 1	treatment: control
GSM1127891	individual: donor 2	treatment: control
GSM1127892	individual: donor 3	treatment: control
GSM1127893	individual: donor 4	treatment: control
GSM1127894	individual: donor 1	treatment: all-trans retinoic acid
GSM1127895	individual: donor 2	treatment: all-trans retinoic acid
GSM1127896	individual: donor 3	treatment: all-trans retinoic acid
GSM1127897	individual: donor 4	treatment: all-trans retinoic acid
GSM1127898	individual: donor 1	treatment: 1,25a-dihydroxyvitamin D3
GSM1127899	individual: donor 2	treatment: 1,25a-dihydroxyvitamin D3
GSM1127900	individual: donor 3	treatment: 1,25a-dihydroxyvitamin D3
GSM1127901	individual: donor 4	treatment: 1,25a-dihydroxyvitamin D3

GEOquery: feature names

```
> dim(gset@featureData@data)
[1] 54675    16
```

```
> gset@featureData@data$ENTREZ_GENE_ID[1:10]
 [1] 780 /// 100616237  5982              3310              7849              2978              7318 /// 100847079 7067
[10] 1571
21880 Levels: 100131755 10406 11078 11099 11202 112597 /// 101930489 113235 113277 114609 116842 117144 121340 123207 124540
> gset@featureData@data$Gene Symbol
Error: unexpected symbol in "gset@featureData@data$Gene Symbol"
```

```
> fvarLabels(gset)
 [1] "ID"                              "GB_ACC"                          "SPOT_ID"                         "Species Scientific Name"
 [5] "Annotation Date"                 "Sequence Type"                   "Sequence Source"                 "Target Description"
 [9] "Representative Public ID"        "Gene Title"                      "Gene Symbol"                     "ENTREZ_GENE_ID"
[13] "RefSeq Transcript ID"            "Gene Ontology Biological Process" "Gene Ontology Cellular Component" "Gene Ontology Molecular Function"
> fvarLabels(gset) <- make.names(fvarLabels(gset))
> fvarLabels(gset)
 [1] "ID"                              "GB_ACC"                          "SPOT_ID"                         "Species.Scientific.Name"
 [5] "Annotation.Date"                 "Sequence.Type"                   "Sequence.Source"                 "Target.Description"
 [9] "Representative.Public.ID"        "Gene.Title"                      "Gene.Symbol"                     "ENTREZ_GENE_ID"
[13] "RefSeq.Transcript.ID"            "Gene.Ontology.Biological.Process" "Gene.Ontology.Cellular.Component" "Gene.Ontology.Molecular.Function"
> gset@featureData@data$Gene.Symbol[1:10]
 [1] DDR1 /// MIR4640 RFC2            HSPA6            PAX8             GUCA1A           MIR5193 /// UBA7 THRA            PTPN21           CCL5
23521 Levels: ADAM32 AFG3L1P AK9 ALG10 ARMCX4 ATP6V1E2 BEST4 C15orf40 C19orf26 C4orf33 CATSPER1 CCDC11 CCDC185 CCDC65 CCL5 CENPBD1 CILP2 CNOT7 CORO6
```

GEOquery: Make groups

```
> pData(phenoData(gset))[13]
                      characteristics_ch1.3
GSM1127890                 treatment: control
GSM1127891                 treatment: control
GSM1127892                 treatment: control
GSM1127893                 treatment: control
GSM1127894   treatment: all-trans retinoic acid
GSM1127895   treatment: all-trans retinoic acid
GSM1127896   treatment: all-trans retinoic acid
GSM1127897   treatment: all-trans retinoic acid
GSM1127898 treatment: 1,25a-dihydroxyvitamin D3
GSM1127899 treatment: 1,25a-dihydroxyvitamin D3
GSM1127900 treatment: 1,25a-dihydroxyvitamin D3
GSM1127901 treatment: 1,25a-dihydroxyvitamin D3
> class(pData(phenoData(gset))[13])
[1] "data.frame"
> pData(phenoData(gset))[13][,1]
                              V2                               V3                               V4                               V5
               treatment: control               treatment: control               treatment: control               treatment: control
                              V6                               V7                               V8                               V9
  treatment: all-trans retinoic acid   treatment: all-trans retinoic acid   treatment: all-trans retinoic acid   treatment: all-trans retinoic acid
                             V10                              V11                              V12                              V13
treatment: 1,25a-dihydroxyvitamin D3 treatment: 1,25a-dihydroxyvitamin D3 treatment: 1,25a-dihydroxyvitamin D3 treatment: 1,25a-dihydroxyvitamin D3
Levels: treatment: 1,25a-dihydroxyvitamin D3 treatment: all-trans retinoic acid treatment: control
> class(pData(phenoData(gset))[13][,1])
[1] "factor"
> as.character(pData(phenoData(gset))[13][,1])
 [1] "treatment: control"                "treatment: control"                "treatment: control"                "treatment: control"
 [5] "treatment: all-trans retinoic acid"   "treatment: all-trans retinoic acid"   "treatment: all-trans retinoic acid"   "treatment: all-trans retinoic acid"
 [9] "treatment: 1,25a-dihydroxyvitamin D3" "treatment: 1,25a-dihydroxyvitamin D3" "treatment: 1,25a-dihydroxyvitamin D3" "treatment: 1,25a-dihydroxyvitamin D3"
> gsub(pattern="treatment: ", replacement="", as.character(pData(phenoData(gset))[13][,1]))
 [1] "control"                "control"                "control"                "control"                "all-trans retinoic acid"   "all-trans retinoic acid"
 [7] "all-trans retinoic acid"   "all-trans retinoic acid"   "1,25a-dihydroxyvitamin D3" "1,25a-dihydroxyvitamin D3" "1,25a-dihydroxyvitamin D3" "1,25a-dihydroxyvitamin D3"
```

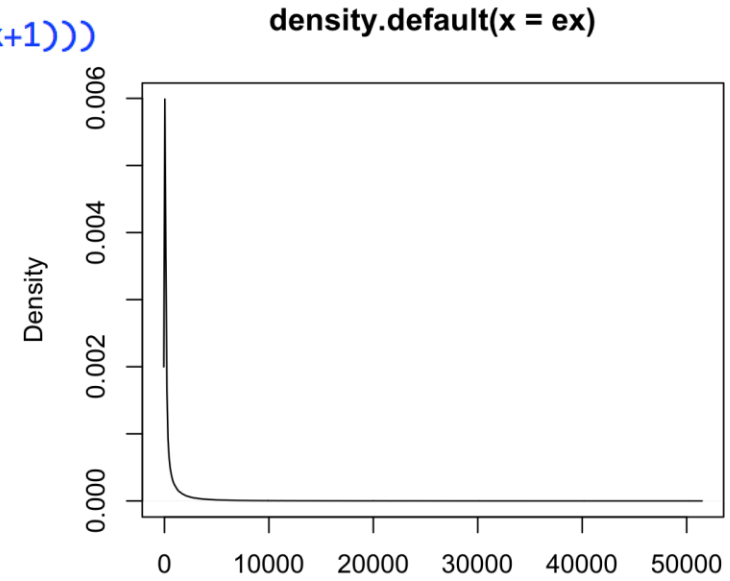
GEOquery: Expression Data

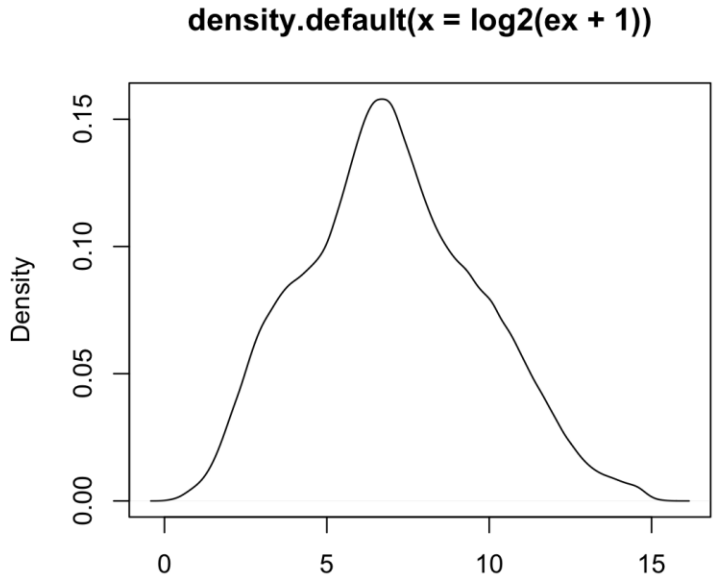
```
> dim(exprs(gset))
[1] 54675    12
> exprs(gset)[1:10,]
          GSM1127890 GSM1127891 GSM1127892 GSM1127893 GSM1127894 GSM1127895 GSM1127896 GSM1127897 GSM1127898 GSM1127899 GSM1127900 GSM1127901
1007_s_at   625.3410   668.4860   861.0220   801.1640   823.6010   613.4630   861.5760  616.08300   471.2490   588.2740  751.40300   555.9900
1053_at     607.3380   714.6880   506.8760   750.2360   716.4290   565.8000   659.0040  671.96100   419.0740   540.5410  465.02500   463.0000
117_at     2553.5700  1700.8800  1253.9900  1550.6200  1818.8000  1091.6300  1057.5200 1668.04000  1116.6500   677.5040  703.72400   979.8570
121_at     1612.6800  1523.6100  1526.7600  1732.1500  1165.3000  1258.7700  1234.6900 1234.59000   992.8480  1430.4100 1267.15000  1330.9300
1255_g_at    67.4127    36.7925    37.2724    51.8988    40.1649    39.6933    36.7995  69.31450    26.1009    41.8341   42.94730    26.1911
1294_at     581.0890   524.4970   695.8520   637.0150   664.0800   498.6110   781.0160  668.63500   438.7350   419.1830  640.87000   490.2770
1316_at     365.3030   403.2550   232.9740   275.0810   229.5450   386.0970   243.1830  273.82500   286.1880   349.2800  307.37700   248.2060
1320_at      43.6584     7.9691    33.9928    16.9462    27.6283    13.9847    10.3745   7.90585    10.7775    17.1919    7.15019    16.0987
1405_i_at  2750.7100  9637.4600  8840.6900  2555.7300  2919.9300  9589.3700 10121.6000 2794.71000  2335.0900  6793.6300 9007.02000  2158.1000
1431_at      38.3734    54.3899    36.3182    37.6997    37.5305    33.2139    55.6216  71.28740    31.6305    77.6095   51.95260    31.8953
```

ex <- exprs (gset)

ex_log2 <- log2(ex+1)

```
> par(mfrow=c(1,2))
> plot(density(ex))
> plot(density(log2(ex+1)))
```





T-test

```
> x <- rnorm(10)
> x
 [1] -1.6766891  1.1420621  0.2532277 -0.7291988 -1.0431424 -0.6642873
 [7] -1.2064824 -0.4233017 -0.7779013  0.5675905
> t.test(x)

    One Sample t-test

data:  x
t = -1.6674, df = 9, p-value = 0.1298
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1.0741948  0.1625703
sample estimates:
 mean of x
-0.4558123
> x <- rnorm(10, mean=10, sd=1)
> x
 [1]  9.311877 11.156537  9.770292  9.700711  8.969781  9.257209 10.460785
 [8]  9.990371 10.267516 11.694930
> t.test(x)

    One Sample t-test

data:  x
t = 36.892, df = 9, p-value = 3.915e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  9.441259 10.674742
sample estimates:
mean of x
   10.058
```

T-test

```
> x <- rnorm(10, mean=10, sd=1)
> x
 [1]  9.958552  9.712997  9.108839 10.897602  9.411299 10.016481  9.435284
 [8] 10.610639  9.274237 10.324511
> y <- rnorm(10)
> y
 [1] -1.1057192 -1.2173110  0.8897431  1.0182269 -0.4183636 -2.1034508
 [7]  0.2707008 -1.2580893 -0.4934180 -1.2186306
> t.test(x,y)

    Welch Two Sample t-test

data:  x and y
t = 27.967, df = 14.517, p-value = 4.917e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  9.640791 11.236560
sample estimates:
 mean of x  mean of y
 9.8750442 -0.5636312
```

T-test

```
> xy_ttest <- t.test(x,y)
> xy_ttest

    Welch Two Sample t-test

data:  x and y
t = 27.967, df = 14.517, p-value = 4.917e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  9.640791 11.236560
sample estimates:
 mean of x  mean of y
 9.8750442 -0.5636312

> xy_ttest$statistic
       t
27.96667
> xy_ttest$p.value
[1] 4.916691e-14
```

T-test

```
> compare_groups <- gsub(pattern="treatment: ", replacement="", as.character(pData(phenoData(gset))[13][,1]))
> compare_groups
 [1] "control"                 "control"                 "control"                 "control"                 "all-trans retinoic acid"   "all-trans retinoic acid"
 [7] "all-trans retinoic acid" "all-trans retinoic acid" "1,25a-dihydroxyvitamin D3" "1,25a-dihydroxyvitamin D3" "1,25a-dihydroxyvitamin D3" "1,25a-dihydroxyvitamin D3"
```

```
> vd_ctrl_ttest <- apply(ex_log2, 1, function(x) {
+ aux <- t.test(x[which(compare_groups == "1,25a-dihydroxyvitamin D3")], x[which(compare_groups == "control")]);
+ aux$p.value
+ })
> length(vd_ctrl_ttest)
[1] 54675

> vd_ctrl_ttest[1:10]
 1007_s_at    1053_at     117_at     121_at  1255_g_at    1294_at    1316_at    1320_at  1405_i_at    1431_at
0.11453297 0.03271019 0.01206182 0.04419045 0.14230442 0.11761575 0.72877886 0.25241449 0.75746942 0.72358841
> sort(vd_ctrl_ttest)[1:10]
   226099_at  223993_s_at  210838_s_at  215111_s_at    206504_at    210058_at  225353_s_at    226950_at    244352_at    209696_at
5.158774e-08 2.573393e-07 4.246613e-07 6.086988e-07 6.552216e-07 7.784512e-07 9.224370e-07 1.049815e-06 1.675303e-06 3.018344e-06
> order(vd_ctrl_ttest)[1:10]
 [1] 35356 33268 20219 24406 15951 19464 34611 36206 53603 19105
> cbind(compare_groups, ex_log2[35356,])
          compare_groups
GSM1127890 "control"                   "10.3775359136266"
GSM1127891 "control"                   "10.4244392620533"
GSM1127892 "control"                   "10.5851597307016"
GSM1127893 "control"                   "10.3449172427325"
GSM1127894 "all-trans retinoic acid"   "10.2293836354257"
GSM1127895 "all-trans retinoic acid"   "10.6133843113419"
GSM1127896 "all-trans retinoic acid"   "10.4198127416242"
GSM1127897 "all-trans retinoic acid"   "10.2895463215015"
GSM1127898 "1,25a-dihydroxyvitamin D3" "13.1984941536391"
GSM1127899 "1,25a-dihydroxyvitamin D3" "13.2357823970245"
GSM1127900 "1,25a-dihydroxyvitamin D3" "13.3416994962817"
GSM1127901 "1,25a-dihydroxyvitamin D3" "13.0476300620057"
```

T-test

```
> cbind(compare_groups, ex_log2[35356,])
           compare_groups
GSM1127890 "control"                    "10.3775359136266"
GSM1127891 "control"                    "10.4244392620533"
GSM1127892 "control"                    "10.5851597307016"
GSM1127893 "control"                    "10.3449172427325"
GSM1127894 "all-trans retinoic acid"    "10.2293836354257"
GSM1127895 "all-trans retinoic acid"    "10.6133843113419"
GSM1127896 "all-trans retinoic acid"    "10.4198127416242"
GSM1127897 "all-trans retinoic acid"    "10.2895463215015"
GSM1127898 "1,25a-dihydroxyvitamin D3" "13.1984941536391"
GSM1127899 "1,25a-dihydroxyvitamin D3" "13.2357823970245"
GSM1127900 "1,25a-dihydroxyvitamin D3" "13.3416994962817"
GSM1127901 "1,25a-dihydroxyvitamin D3" "13.0476300620057"
```

```
> t.test(ex_log2[35356,which(compare_groups == "control")],
ex_log2[35356,which(compare_groups == "all-trans retinoic acid")])

    Welch Two Sample t-test

data:  ex_log2[35356, which(compare_groups == "control")] and ex_log2[35356,
which(compare_groups == "all-trans retinoic acid")]
t = 0.44846, df = 5.0426, p-value = 0.6724
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2121959  0.3021585
sample estimates:
mean of x mean of y
 10.43301  10.38803
```

Wilcoxon Test

```
> x
 [1]  9.958552  9.712997  9.108839 10.897602  9.411299 10.016481  9.435284 10.610639  9.274237 10.324511
> y
 [1] -1.1057192 -1.2173110  0.8897431  1.0182269 -0.4183636 -2.1034508  0.2707008 -1.2580893 -0.4934180 -1.2186306
> wilcox.test(x)

    Wilcoxon signed rank test

data:  x
V = 55, p-value = 0.001953
alternative hypothesis: true location is not equal to 0
```

```
> wilcox.test(x, y)

    Wilcoxon rank sum test

data:  x and y
W = 100, p-value = 1.083e-05
alternative hypothesis: true location shift is not equal to 0
```

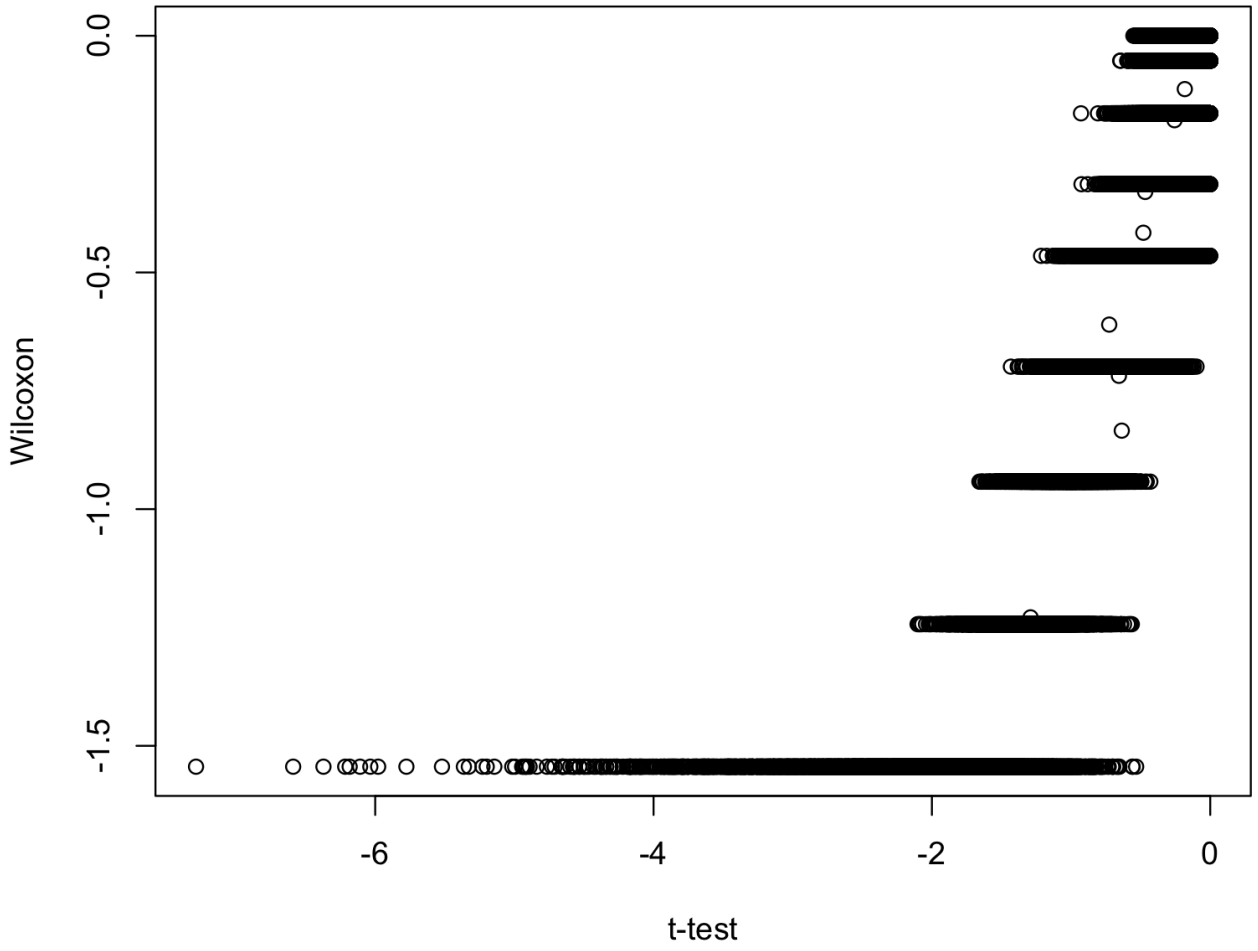
```
> xy_wilcoxon <- wilcox.test(x, y)
> xy_wilcoxon$statistic
  W
100
> xy_wilcoxon$p.value
[1] 1.082509e-05
```

Wilcoxon Test

```
> vd_ctrl_wilcoxon <- apply(ex_log2, 1, function(x) {
+ aux <- wilcox.test(x[which(compare_groups == "1,25a-dihydroxyvitamin D3")], x[which(compare_groups == "control")]);
+ aux$p.value
+ })
> length(vd_ctrl_wilcoxon)
[1] 54675

> vd_ctrl_wilcoxon[1:10]
 1007_s_at    1053_at     117_at     121_at  1255_g_at    1294_at    1316_at    1320_at  1405_i_at    1431_at
0.11428571 0.05714286 0.02857143 0.02857143 0.34285714 0.20000000 0.88571429 0.34285714 0.48571429 0.88571429
> sort(vd_ctrl_wilcoxon)[1:10]
     117_at      121_at     1487_at 1552257_a_at   1552280_at   1552291_at   1552340_at   1552359_at   1552365_at 1552419_s_at
 0.02857143  0.02857143  0.02857143   0.02857143   0.02857143   0.02857143   0.02857143   0.02857143   0.02857143   0.02857143
> order(vd_ctrl_wilcoxon)[1:10]
 [1]   3   4  12  15  30  37  71  79  83 120
> cbind(compare_groups, ex_log2[3,])
           compare_groups
GSM1127890 "control"                   "11.3188647534861"
GSM1127891 "control"                   "10.7329136004708"
GSM1127892 "control"                   "10.2934601532143"
GSM1127893 "control"                   "10.599559561748"
GSM1127894 "all-trans retinoic acid"   "10.829564188458"
GSM1127895 "all-trans retinoic acid"   "10.0935892249807"
GSM1127896 "all-trans retinoic acid"   "10.0478328129856"
GSM1127897 "all-trans retinoic acid"   "10.7048028152225"
GSM1127898 "1,25a-dihydroxyvitamin D3" "10.1262527534162"
GSM1127899 "1,25a-dihydroxyvitamin D3" "9.40621351059668"
GSM1127900 "1,25a-dihydroxyvitamin D3" "9.46091453694006"
GSM1127901 "1,25a-dihydroxyvitamin D3" "9.93789900977802"
```

```
> sort(vd_ctrl_ttest)[1:5]
   226099_at  223993_s_at  210838_s_at  215111_s_at    206504_at
5.158774e-08 2.573393e-07 4.246613e-07 6.086988e-07 6.552216e-07
> order(vd_ctrl_ttest)[1:5]
[1] 35356 33268 20219 24406 15951
> vd_ctrl_wilcoxon[order(vd_ctrl_ttest)[1:5]]
  226099_at 223993_s_at 210838_s_at 215111_s_at   206504_at
 0.02857143  0.02857143  0.02857143  0.02857143  0.02857143
```

T-test vs Wilcoxon

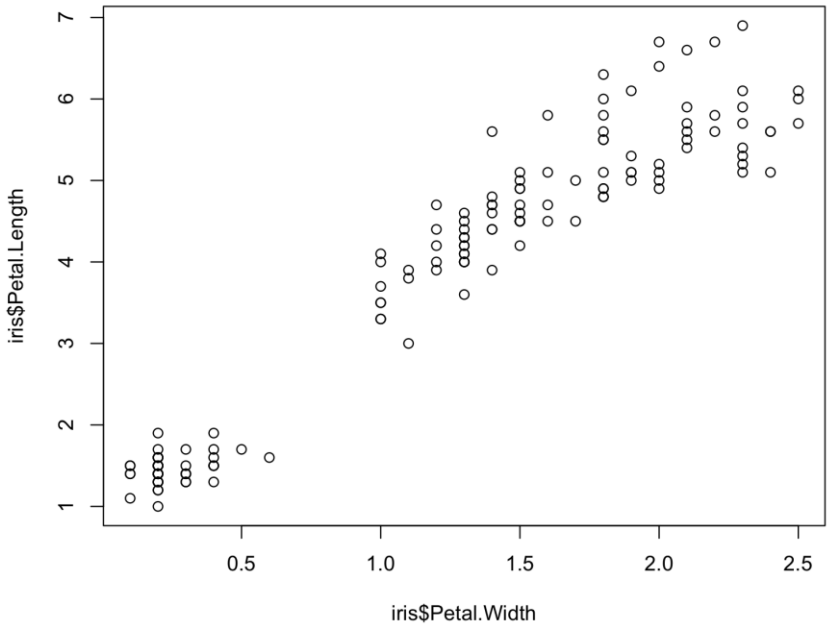



```
plot(log10(vd_ctrl_ttest), log10(vd_ctrl_wilcoxon), xlab="t-test", ylab="Wilcoxon")
```

Correlation

```
> library(psych)
> data(iris)
```

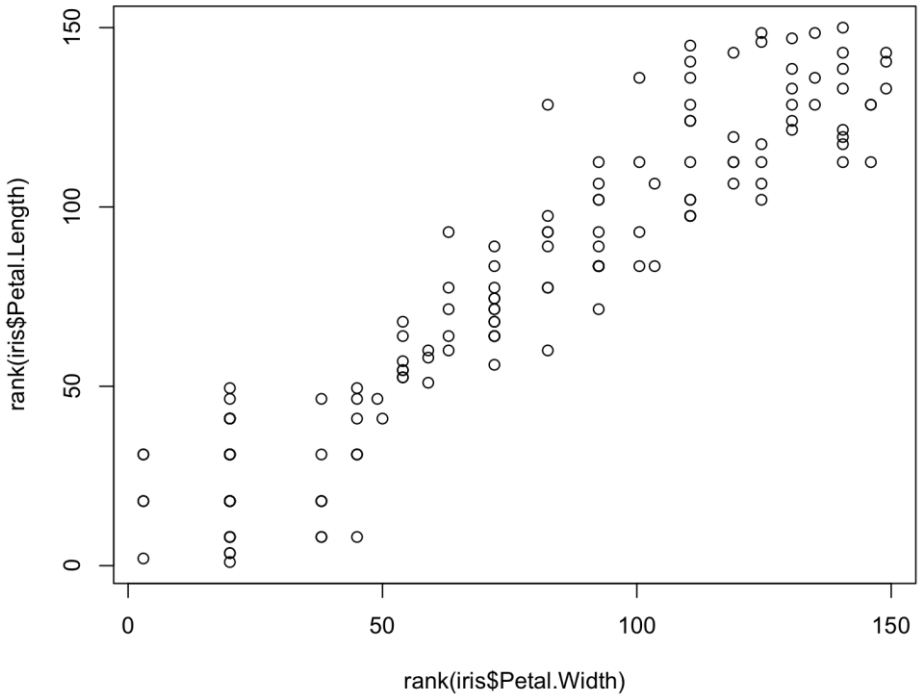
```
> cor(iris$Petal.Width, iris$Petal.Length)
[1] 0.9628654
> cor.test(iris$Petal.Width, iris$Petal.Length)

    Pearson's product-moment correlation

data:  iris$Petal.Width and iris$Petal.Length
t = 43.387, df = 148, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9490525 0.9729853
sample estimates:
      cor
0.9628654

> cor.test(iris$Petal.Width, iris$Petal.Length)$estimate
      cor
0.9628654
> cor.test(iris$Petal.Width, iris$Petal.Length)$p.value
[1] 4.675004e-86
```




```
> plot(iris$Petal.Width, iris$Petal.Length)
```

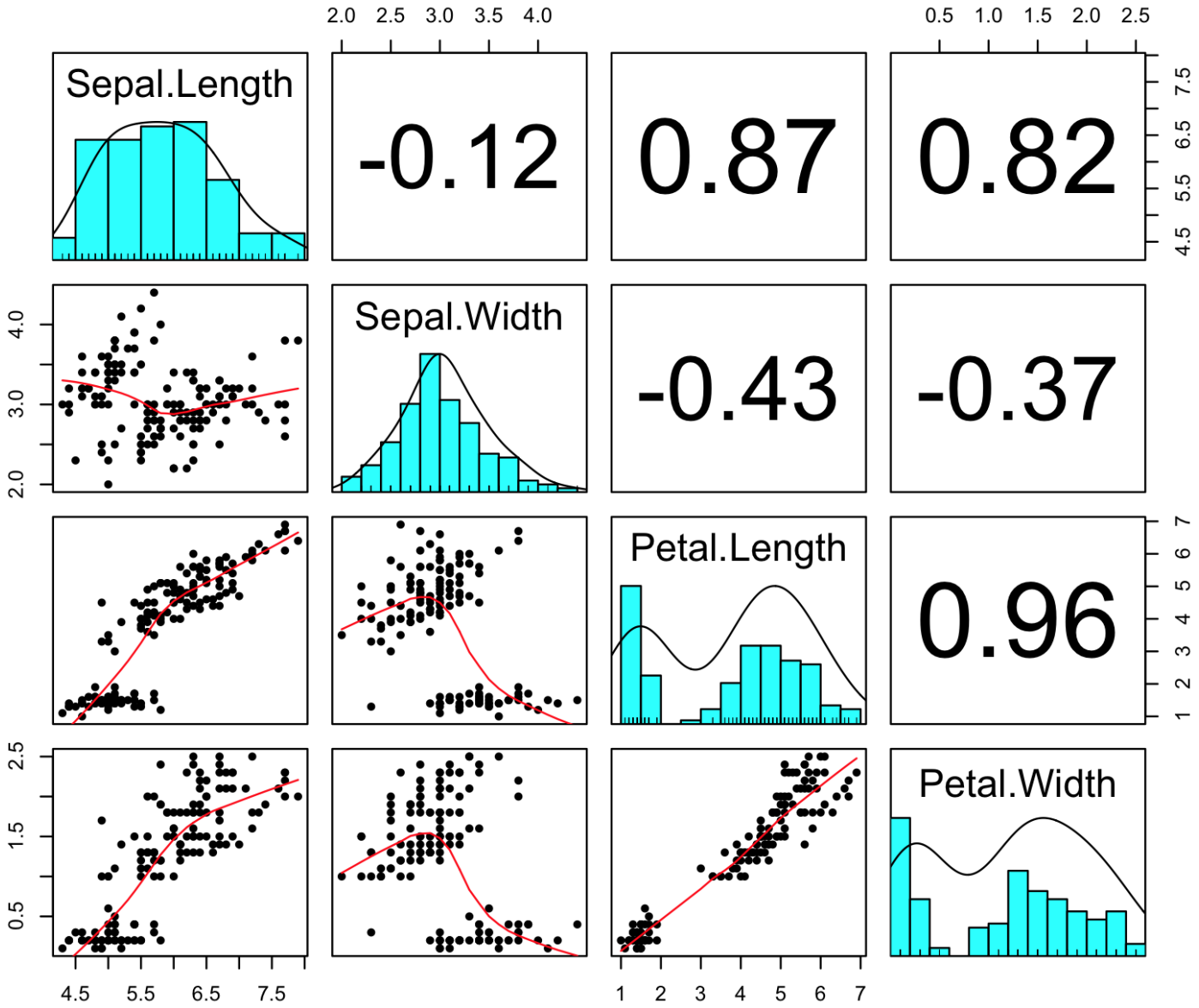
Correlation

```
> plot(rank(iris$Petal.Width), rank(iris$Petal.Length))
> cor(iris$Petal.Width, iris$Petal.Length, method="spearman")
[1] 0.9376668
> cor.test(iris$Petal.Width, iris$Petal.Length, method="spearman")

    Spearman's rank correlation rho

data:  iris$Petal.Width and iris$Petal.Length
S = 35061, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9376668
```



Correlation

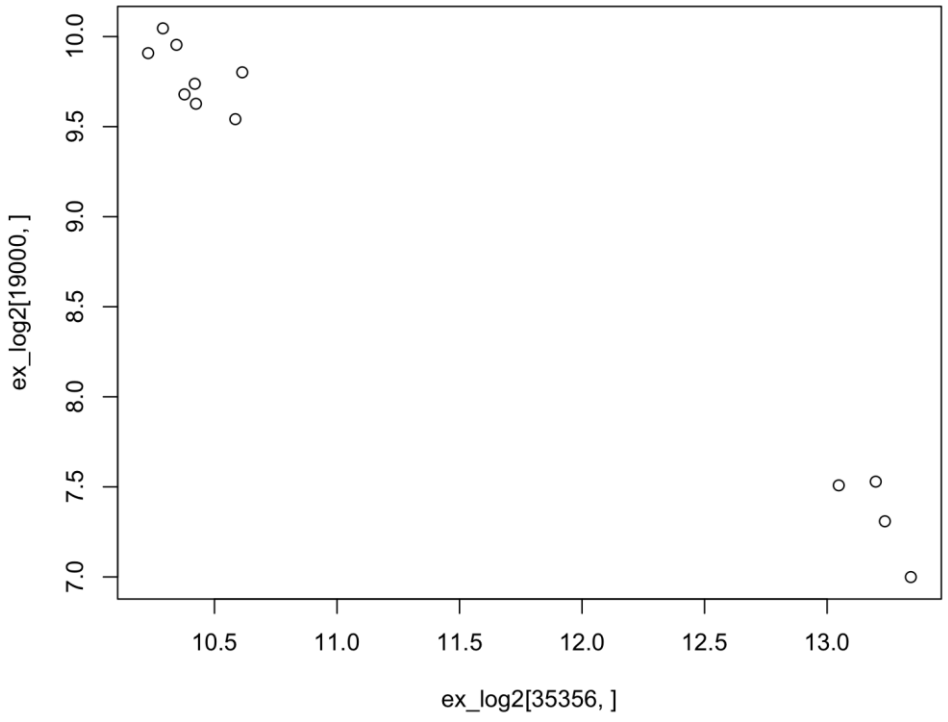
```
> pairs.panels(x=iris[,-5], method="pearson", ellipses=F)
```



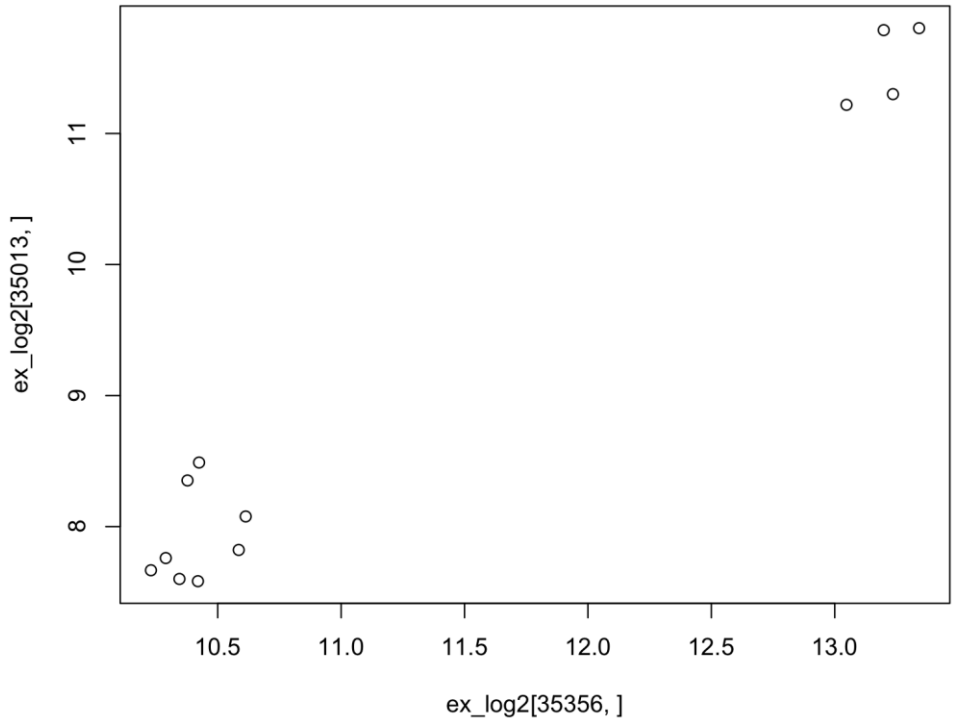
Correlation

```
> cor_35356 <- cor(ex_log2[35356,], t(ex_log2))
> names(cor_35356) <- rownames(ex_log2)
> sort(cor_35356)[1:10]
209589_s_at 225353_s_at   209588_at   228273_at 211165_x_at   223615_at   204150_at 218045_x_at   206206_at   203473_at
 -0.9936453  -0.9895084  -0.9875440  -0.9835309  -0.9820767  -0.9776556  -0.9730968  -0.9719871  -0.9712928  -0.9703368
> order(cor_35356)[1:10]
 [1] 19000 34611 18999 37528 20528 32892 13598 27331 15653 12921
> sort(cor_35356, decreasing=T)[1:10]
  226099_at   225755_at 223993_s_at   238638_at   206504_at   228667_at 201300_s_at 203716_s_at 200798_x_at   210058_at
  1.0000000   0.9861506   0.9838890   0.9805354   0.9785763   0.9784463   0.9778126   0.9776461   0.9776349   0.9776094
> order(cor_35356, decreasing=T)[1:10]
 [1] 35356 35013 33268 47888 15951 37922 10749 13164 10247 19464
```

```
> plot(ex_log2[35356,], ex_log2[19000,])
```




```
> plot(ex_log2[35356,], ex_log2[35013,])
```



Correlation

```
> cbind(compare_groups, ex_log2[35356,], ex_log2[35013,])
          compare_groups
GSM1127890 "control"                    "10.3775359136266" "8.35225092936301"
GSM1127891 "control"                    "10.4244392620533" "8.48905717094573"
GSM1127892 "control"                    "10.5851597307016" "7.82195889014769"
GSM1127893 "control"                    "10.3449172427325" "7.60061171069061"
GSM1127894 "all-trans retinoic acid"    "10.2293836354257" "7.66687728474809"
GSM1127895 "all-trans retinoic acid"    "10.6133843113419" "8.07744062948797"
GSM1127896 "all-trans retinoic acid"    "10.4198127416242" "7.58343634433843"
GSM1127897 "all-trans retinoic acid"    "10.2895463215015" "7.76018767359572"
GSM1127898 "1,25a-dihydroxyvitamin D3" "13.1984941536391" "11.7885959960983"
GSM1127899 "1,25a-dihydroxyvitamin D3" "13.2357823970245" "11.300152331024"
GSM1127900 "1,25a-dihydroxyvitamin D3" "13.3416994962817" "11.8037113844204"
GSM1127901 "1,25a-dihydroxyvitamin D3" "13.0476300620057" "11.2184845305792"
> cbind(compare_groups, ex_log2[35356,], ex_log2[19000,])
          compare_groups
GSM1127890 "control"                    "10.3775359136266" "9.67900674751115"
GSM1127891 "control"                    "10.4244392620533" "9.62698843900318"
GSM1127892 "control"                    "10.5851597307016" "9.54140835871482"
GSM1127893 "control"                    "10.3449172427325" "9.95392868890556"
GSM1127894 "all-trans retinoic acid"    "10.2293836354257" "9.90785357365131"
GSM1127895 "all-trans retinoic acid"    "10.6133843113419" "9.80138826333245"
GSM1127896 "all-trans retinoic acid"    "10.4198127416242" "9.7378236297982"
GSM1127897 "all-trans retinoic acid"    "10.2895463215015" "10.045568563291"
GSM1127898 "1,25a-dihydroxyvitamin D3" "13.1984941536391" "7.5290791106569"
GSM1127899 "1,25a-dihydroxyvitamin D3" "13.2357823970245" "7.3089214518831"
GSM1127900 "1,25a-dihydroxyvitamin D3" "13.3416994962817" "6.99914314541559"
GSM1127901 "1,25a-dihydroxyvitamin D3" "13.0476300620057" "7.50835734005231"
```

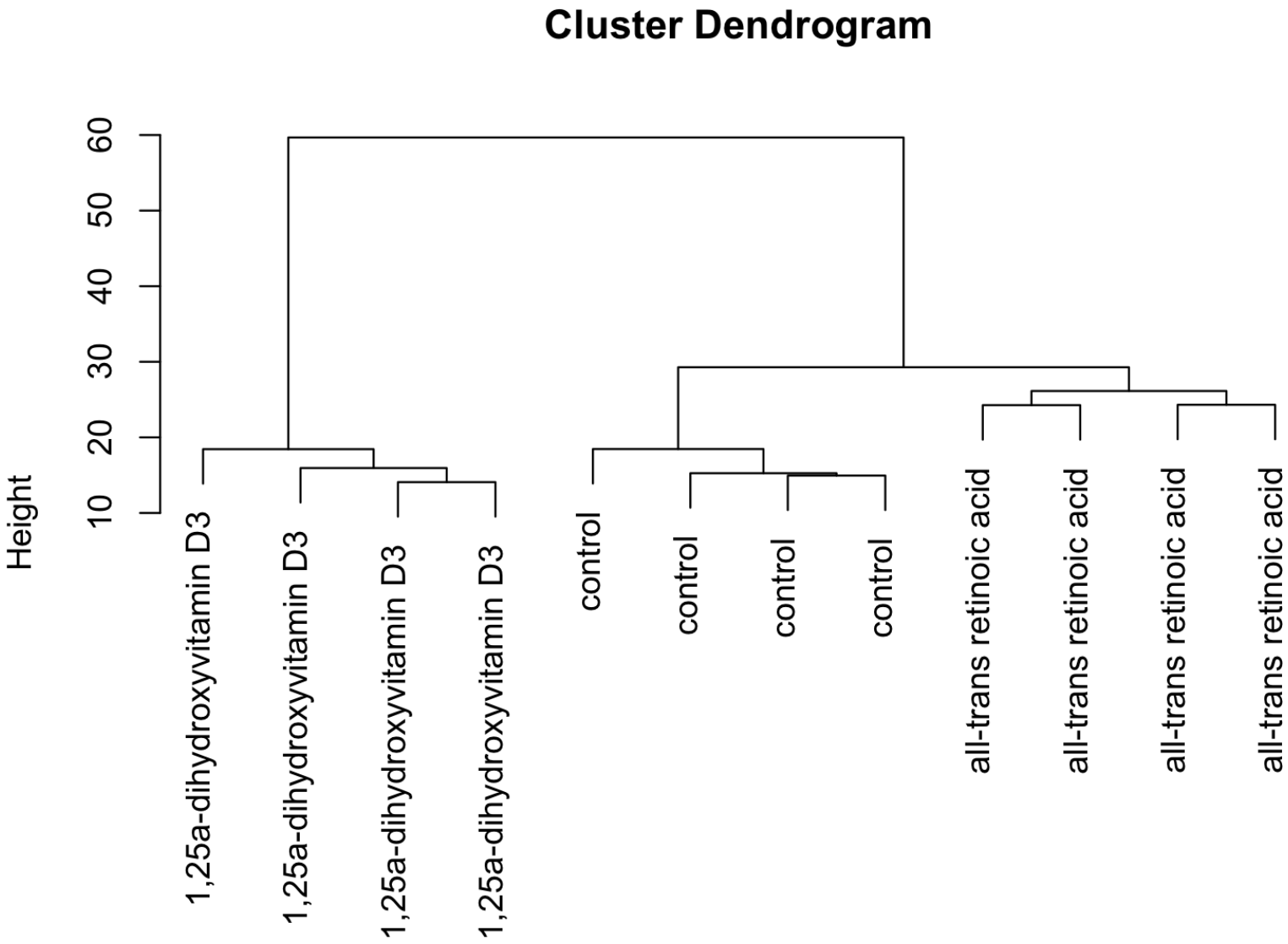
Clustering

```
> vd_ctrl_ttest_top_1500 <- ex_log2[order(vd_ctrl_ttest)[1:1500],]
> dim(vd_ctrl_ttest_top_1500)
[1] 1500   12
> vd_ctrl_ttest_dist <- dist(t(vd_ctrl_ttest_top_1500))
```

```
> vd_ctrl_ttest_dist
          GSM1127890 GSM1127891 GSM1127892 GSM1127893 GSM1127894 GSM1127895 GSM1127896 GSM1127897 GSM1127898 GSM1127899 GSM1127900
GSM1127891   15.40054
GSM1127892   19.59723   18.39105
GSM1127893   14.94540   15.11868   17.38061
GSM1127894   28.92634   28.24972   30.40264   29.63161
GSM1127895   30.41764   26.60963   28.10557   29.22762   26.96922
GSM1127896   31.94930   29.74703   27.61916   30.15118   26.50387   24.26265
GSM1127897   30.01917   28.82748   30.24082   28.30808   24.31203   25.80513   25.26257
GSM1127898   61.34943   59.37979   60.89674   61.62962   60.44575   57.17620   57.25206   61.13861
GSM1127899   60.95672   57.78773   59.19408   60.58241   60.30047   54.88019   55.64745   60.59318   17.08427
GSM1127900   60.33049   57.79084   57.92565   60.38541   60.00952   55.54599   54.62053   60.15104   18.50467   18.21063
GSM1127901   63.82337   61.08086   62.55200   63.34621   62.84723   58.73572   58.46601   62.70937   14.07787   14.80261   18.60290
```

Clustering

```
> vd_ctrl_ttest_hh <- hclust(vd_ctrl_ttest_dist, method="average")
> plot(vd_ctrl_ttest_hh, labels=compare_groups)
```



PCA

```
> vd_ctrl_ttest_pca <- prcomp(t(vd_ctrl_ttest_top_1500))
> summary(vd_ctrl_ttest_pca)
Importance of components:
                          PC1     PC2    PC3     PC4    PC5     PC6     PC7     PC8     PC9    PC10    PC11       PC12
Standard deviation     27.756 9.54009 6.3367 5.47256 5.2200 4.14146 3.66978 3.42764 2.94061 2.84080 2.57158 2.049e-14
Proportion of Variance  0.752 0.08884 0.0392 0.02923 0.0266 0.01674 0.01315 0.01147 0.00844 0.00788 0.00646 0.000e+00
Cumulative Proportion   0.752 0.84084 0.8800 0.90927 0.9359 0.95261 0.96576 0.97723 0.98567 0.99354 1.00000 1.000e+00
```

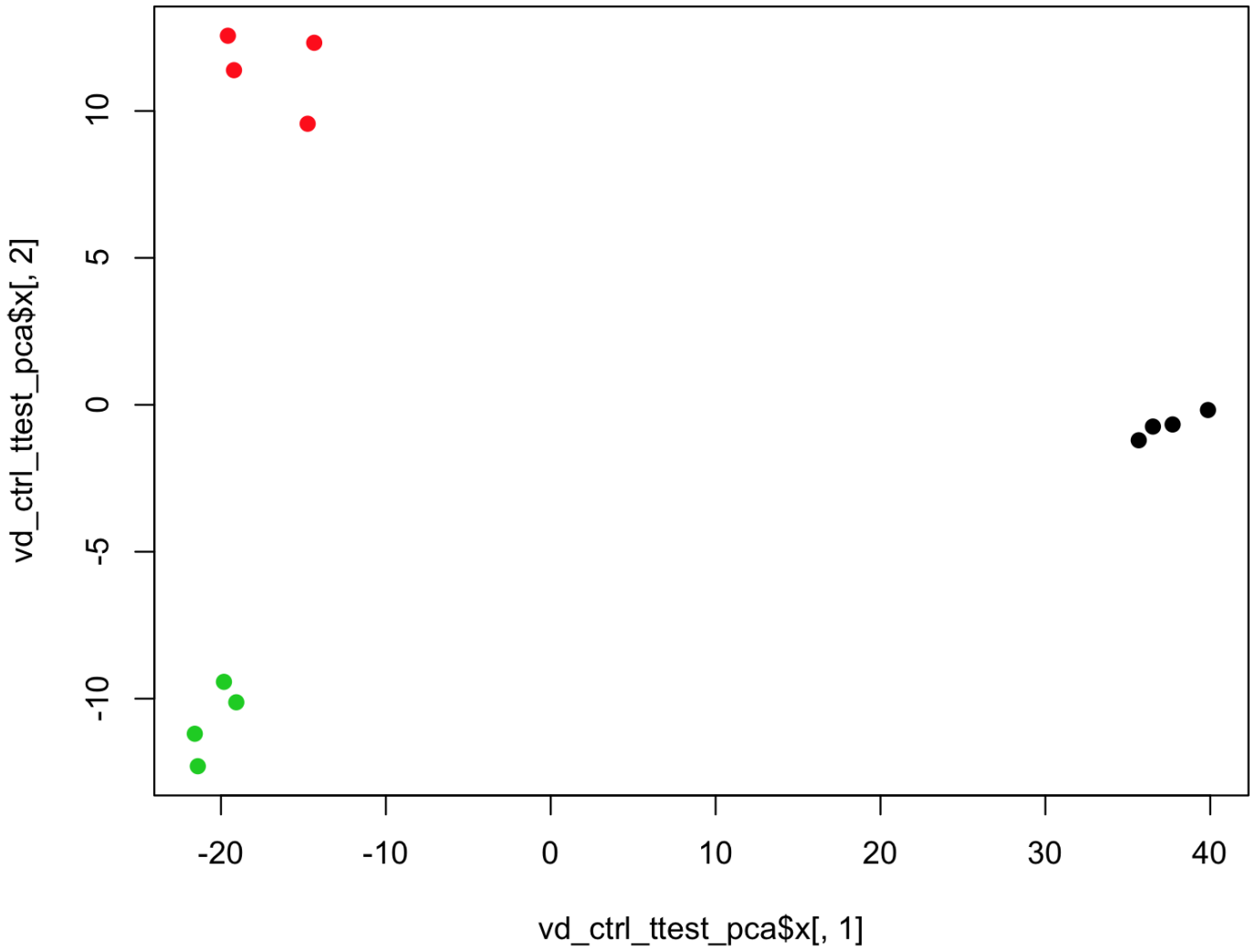
```
> summary(vd_ctrl_ttest_pca)$importance
                          PC1      PC2      PC3      PC4     PC5      PC6      PC7     PC8      PC9     PC10     PC11          PC12
Standard deviation     27.75573 9.540095 6.336669 5.472558 5.22004 4.141464 3.669784 3.42764 2.940614 2.840799 2.571582 2.048999e-14
Proportion of Variance  0.75200 0.088840 0.039200 0.029230 0.02660 0.016740 0.013150 0.01147 0.008440 0.007880 0.006460 0.000000e+00
Cumulative Proportion   0.75200 0.840840 0.880040 0.909270 0.93587 0.952610 0.965760 0.97723 0.985670 0.993540 1.000000 1.000000e+00
> summary(vd_ctrl_ttest_pca)$importance[1:2,]
                          PC1      PC2      PC3      PC4     PC5      PC6      PC7     PC8      PC9     PC10     PC11          PC12
Standard deviation     27.75573 9.540095 6.336669 5.472558 5.22004 4.141464 3.669784 3.42764 2.940614 2.840799 2.571582 2.048999e-14
Proportion of Variance  0.75200 0.088840 0.039200 0.029230 0.02660 0.016740 0.013150 0.01147 0.008440 0.007880 0.006460 0.000000e+00
```

PCA

```
> vd_ctrl_ttest_pca$x
                PC1         PC2         PC3         PC4         PC5         PC6         PC7         PC8         PC9         PC10         PC11         PC12
GSM1127890 -21.40536 -12.2984711   4.4866801 -0.06780603   0.6361824   0.9952461 -6.1432637   0.3577492   5.0562073 -0.8894908   3.02627656 2.272661e-14
GSM1127891 -19.07138 -10.1234127   0.6040836  3.52779222   0.5178268   1.2818710   0.4006862   6.2747186 -3.0840168  5.1051293 -1.31729827 2.300070e-14
GSM1127892 -19.82093  -9.4273756  -6.9752992 -4.18764442   1.5048247  -4.7437713   5.2935369  -5.0597790   1.4593356  2.6113129   0.59447151 1.974430e-14
GSM1127893 -21.58795 -11.1939690   1.0889133 -0.59382537  -3.1897864   3.0140295   1.3590132  -1.6503637  -3.7513188 -5.9414390  -2.28083563 2.181415e-14
GSM1127894 -19.21152  11.3879851  11.4165524 -1.26821008  10.5596977  -1.4551263   2.6285767   0.1567658  -0.5939558 -0.9086806   0.19481017 2.395653e-14
GSM1127895 -14.74423   9.5657578  -7.9217632 13.39202660   1.3511647  -1.5757667  -2.3379703  -2.1614904  -0.8035990 -0.8000179   0.70509574 1.476076e-14
GSM1127896 -14.34604  12.3206650  -9.7689404 -8.91106679   1.6556460   6.5177442  -2.2848970   1.1217855   0.2421664  0.4084463  -0.36906010 2.564355e-14
GSM1127897 -19.58177  12.5603238   5.6943581 -1.41592078 -12.7600075  -3.2849905   0.9275911   0.7667581   1.2085423  0.7908481  -0.29319316 1.883129e-14
GSM1127898  37.71810  -0.6682558   5.1969437  0.09629615   0.2000325   1.6380046  -4.0388684  -5.4867931  -0.2036936  2.9921749  -4.07660910 1.637839e-14
GSM1127899  36.52456  -0.7395023  -2.0278291  3.86288204   0.4488021   2.3570747   5.2865688   3.3601127   5.3631407 -1.8222272  -2.05679782 5.519890e-15
GSM1127900  35.66376  -1.2066587  -3.8694253 -4.66354802   1.4172268  -8.5802896  -3.4660800   3.4819160  -1.9226880 -2.1998358  -0.01646368 6.124441e-15
GSM1127901  39.86277  -0.1770866   2.0757258  0.22902449  -2.3416098   3.8359743   2.3751066  -1.1613797  -2.9701202  0.6537800   5.88960380 2.678413e-15
> predict(vd_ctrl_ttest_pca)
                PC1         PC2         PC3         PC4         PC5         PC6         PC7         PC8         PC9         PC10         PC11         PC12
GSM1127890 -21.40536 -12.2984711   4.4866801 -0.06780603   0.6361824   0.9952461 -6.1432637   0.3577492   5.0562073 -0.8894908   3.02627656 2.272661e-14
GSM1127891 -19.07138 -10.1234127   0.6040836  3.52779222   0.5178268   1.2818710   0.4006862   6.2747186 -3.0840168  5.1051293 -1.31729827 2.300070e-14
GSM1127892 -19.82093  -9.4273756  -6.9752992 -4.18764442   1.5048247  -4.7437713   5.2935369  -5.0597790   1.4593356  2.6113129   0.59447151 1.974430e-14
GSM1127893 -21.58795 -11.1939690   1.0889133 -0.59382537  -3.1897864   3.0140295   1.3590132  -1.6503637  -3.7513188 -5.9414390  -2.28083563 2.181415e-14
GSM1127894 -19.21152  11.3879851  11.4165524 -1.26821008  10.5596977  -1.4551263   2.6285767   0.1567658  -0.5939558 -0.9086806   0.19481017 2.395653e-14
GSM1127895 -14.74423   9.5657578  -7.9217632 13.39202660   1.3511647  -1.5757667  -2.3379703  -2.1614904  -0.8035990 -0.8000179   0.70509574 1.476076e-14
GSM1127896 -14.34604  12.3206650  -9.7689404 -8.91106679   1.6556460   6.5177442  -2.2848970   1.1217855   0.2421664  0.4084463  -0.36906010 2.564355e-14
GSM1127897 -19.58177  12.5603238   5.6943581 -1.41592078 -12.7600075  -3.2849905   0.9275911   0.7667581   1.2085423  0.7908481  -0.29319316 1.883129e-14
GSM1127898  37.71810  -0.6682558   5.1969437  0.09629615   0.2000325   1.6380046  -4.0388684  -5.4867931  -0.2036936  2.9921749  -4.07660910 1.637839e-14
GSM1127899  36.52456  -0.7395023  -2.0278291  3.86288204   0.4488021   2.3570747   5.2865688   3.3601127   5.3631407 -1.8222272  -2.05679782 5.519890e-15
GSM1127900  35.66376  -1.2066587  -3.8694253 -4.66354802   1.4172268  -8.5802896  -3.4660800   3.4819160  -1.9226880 -2.1998358  -0.01646368 6.124441e-15
GSM1127901  39.86277  -0.1770866   2.0757258  0.22902449  -2.3416098   3.8359743   2.3751066  -1.1613797  -2.9701202  0.6537800   5.88960380 2.678413e-15
```

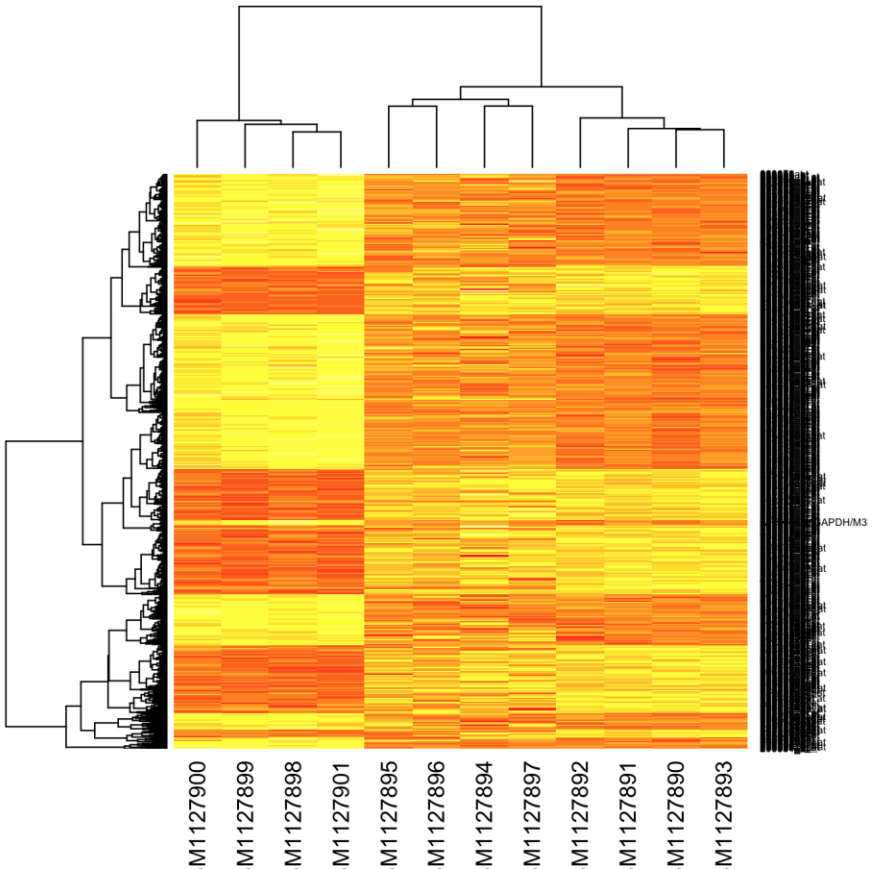
PCA

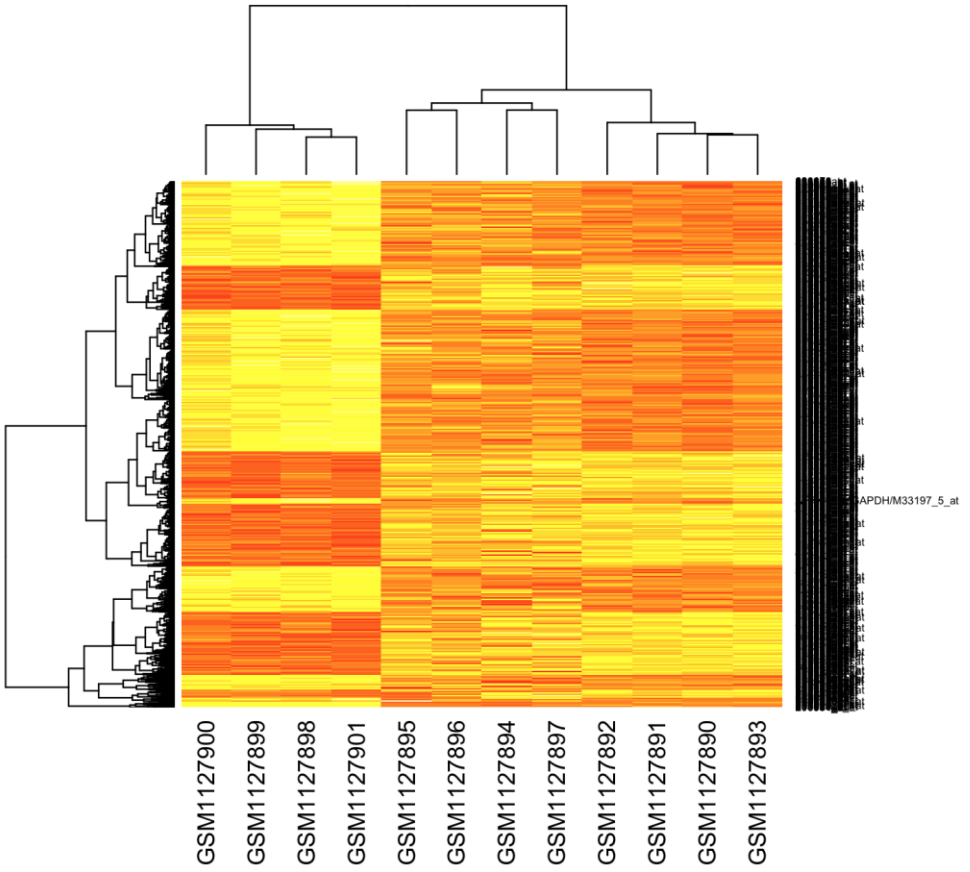
```
> plot(vd_ctrl_ttest_pca$x[,1], vd_ctrl_ttest_pca$x[,2], pch=19, col=as.factor(compare_groups))
> as.factor(compare_groups)
 [1] control                   control                   control                   control                   all-trans retinoic acid   all-trans retinoic acid
 [7] all-trans retinoic acid   all-trans retinoic acid   1,25a-dihydroxyvitamin D3 1,25a-dihydroxyvitamin D3 1,25a-dihydroxyvitamin D3 1,25a-dihydroxyvitamin D3
Levels: 1,25a-dihydroxyvitamin D3 all-trans retinoic acid control
```



Heatmaps

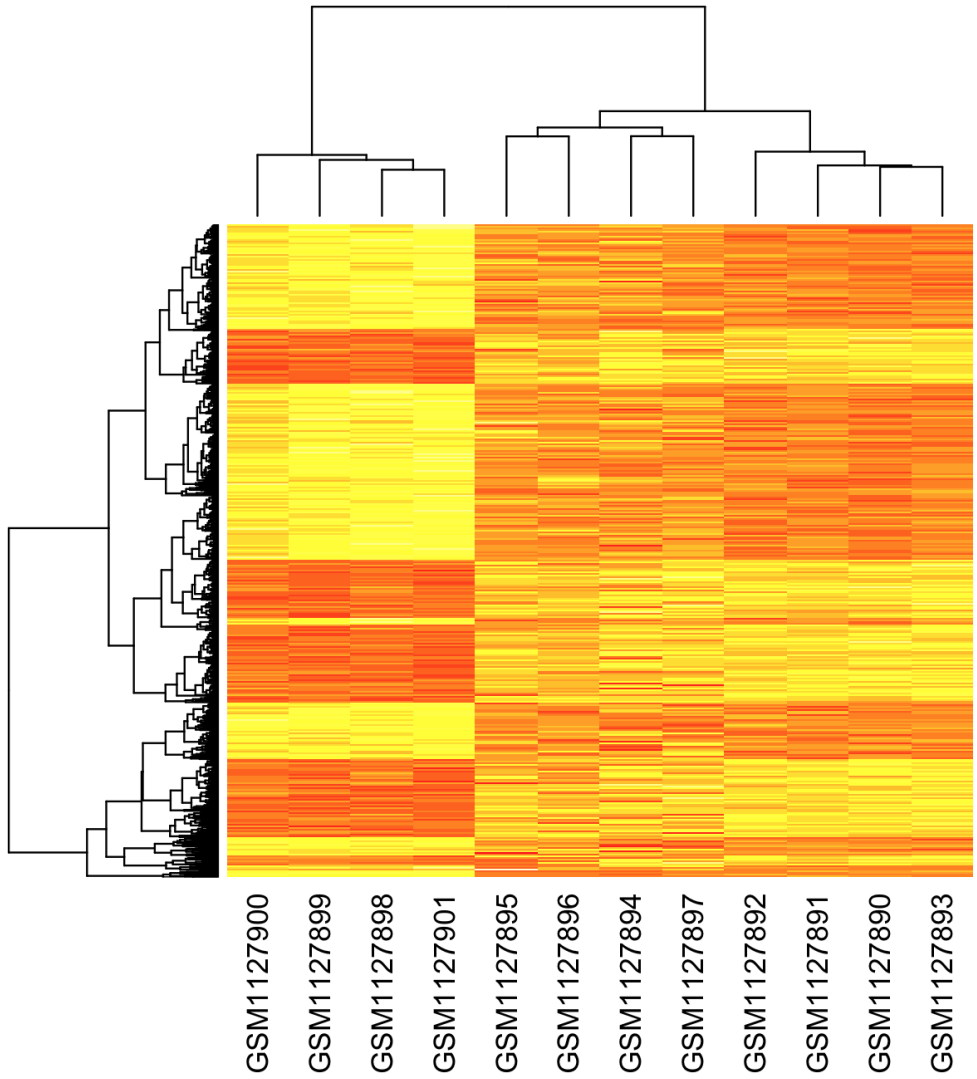
```
> library(ComplexHeatmap)
Loading required package: grid
> library(circlize)
> heatmap(vd_ctrl_ttest_top_1500)
> heatmap(vd_ctrl_ttest_top_1500, margins=c(8,5))
```





Heatmaps

```
> heatmap(vd_ctrl_ttest_top_1500, margins=c(8,5), labRow="")
```




```
> cbind(compare_groups, colnames(vd_ctrl_ttest_top_1500))
      compare_groups
 [1,] "control"                    "GSM1127890"
 [2,] "control"                    "GSM1127891"
 [3,] "control"                    "GSM1127892"
 [4,] "control"                    "GSM1127893"
 [5,] "all-trans retinoic acid"    "GSM1127894"
 [6,] "all-trans retinoic acid"    "GSM1127895"
 [7,] "all-trans retinoic acid"    "GSM1127896"
 [8,] "all-trans retinoic acid"    "GSM1127897"
 [9,] "1,25a-dihydroxyvitamin D3"  "GSM1127898"
[10,] "1,25a-dihydroxyvitamin D3"  "GSM1127899"
[11,] "1,25a-dihydroxyvitamin D3"  "GSM1127900"
[12,] "1,25a-dihydroxyvitamin D3"  "GSM1127901"
```