Differential expression

Wolfgang Huber

Robert Gentleman

Anja von Heydebreck

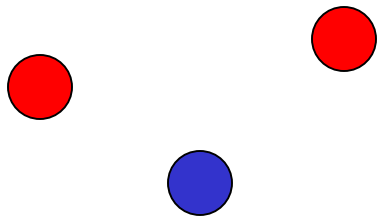
p ≫ n

Goal: find statistically significant associations of biological conditions or phenotypes with gene expression.

Consider the two class problem. Data: n (≈10…100) points in a p-dimensional (≈5000…30000) space.

Problem: There are infinitely many ways to separate the space into two regions by a hyperplane such that the two groups are perfectly separated.

This is a simple geometrical fact and holds as long as n<p!



▶ p □ n

Problem: If I find such a perfectly separating hyperplane, it doesn't mean anything. It is not surprising. It is not a significant finding. I would always find it, no matter how random the data are!

Answer: regularization

Rather than searching in the huge space of all hyperplanes in n-1 dimensional space, restrict ourselves to a much smaller space.

Two major approaches:

- only the hyperplanes perpendicular to one of the n coordinate axis ⇒ *gene-by-gene* discrimination, *gene-by-gene* hypothesis testing.

- any other reasonable, not too complex set of hypersurfaces ⇒ machine learning

Gene by gene tests

t-test

Wilcoxon

F-test / more complex linear models

Cox-regression

Problem:

Treating each gene independently of each other wastes information – many properties may be shared among genes. E.g. their within-group variability.

Moderated / Bayesian t-tests

Rather than estimating within-group variability (denominator of t-test) over and over again for each gene, pool the information from many similar genes

Baldi, Long 2001
Tusher et al. (SAM) 2001
Lönnstedt and Speed 2002
Smyth (limma) 2004

Advantages:
-eliminate occurrence of accidentally large values t-statistic due to accidentally small within-group variance
-effectively introduce a 'fold-change' criterion

Example data

79 samples of acute lymphoblastic leukemia (ALL)
B-cell lymphocytes
37 samples with BCR/ABL fusion (t(9;22)) and 42 without.

Chiaretti et al. (Ritz lab, DFCI)

```
>library(ALL)
>Data(ALL)
```

Nonspecific filtering

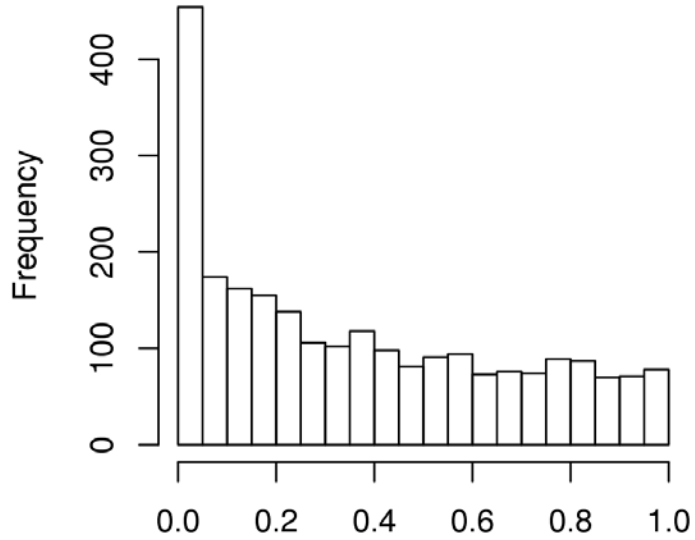
```
> library(genefilter)
> f1 <- pOverA(0.25, log2(100))
> f2 <- function(x) (IQR(x) > 0.5)
> ff <- filterfun(f1, f2)
> selected <- genefilter(eset, ff)
> sum(selected)

[1] 2391

> esetSub <- eset[selected, ]
```

gene-by-gene t-test

```
> library(multtest)
> cl <- as.numeric(esetSub$mol == "BCR/ABL")
> resT <- mt.maxT(exprs(esetSub),
     classlabel = cl, B = 1e+05)
> ord <- order(resT$index)
> rawp <- resT$rawp[ord]
> names(rawp) <- geneNames(esetSub)
```



FWER

Family wise error rate: Probability of at least one false positive.

```
> sum(resT$adjp<0.05)
[1] 18
```

This would imply large loss of power!

Top 3

```
> gnames <- mget(geneNames(esetSub),
                         env = hgu95av2SYMBOL)

> top5 <- resT$index[1:5]

> unlist(gnames[top5])

1636_g_at 39730_at 1635_at 40202_at 37027_at
   "ABL1"   "ABL1"  "ABL1"  "BTEB1"  "AHNAK"
```

FDR

False Discovery Rate: E[FP/(FP+TP)]

```
> res <- mt.rawp2adjp(rawp, proc = "BH")
> sum(res$adjp[, "BH"] < 0.05)
[1] 109
```

Multiple probe sets per gene

```
> library(annotate)
> library(hgu95av2)
> lls <- unlist(contents(hgu95av2LOCUSID))
> tab <- table(table(lls))
```

Multiplicity	1	2	3	4	5	6	7	8	9
No. LocusLink IDs	6756	1581	0498	117	030	17	11	8	1

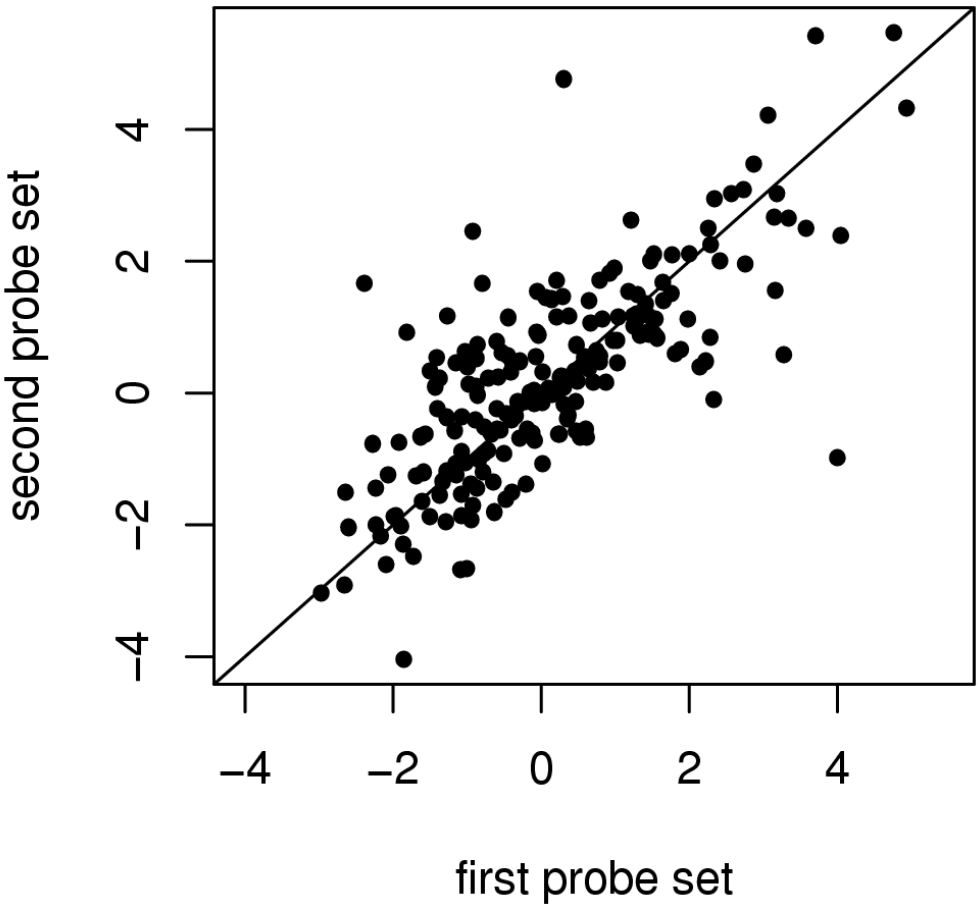
Of the 2263 LocusLink IDs that have more than one probe set identified with them, in 509 cases the nonspecific filtering step selected some, but not all corresponding probe sets.

Multiple probe sets per gene

The three top-scoring probe sets all represented the *ABL1* gene. But there are 5 more probe sets on the chip that also represent the *ABL1* gene, none of which passed our filtering step. The permutation p-values of all eight probe sets are:

```
> ABL1PS <- names(which(lls == ABL1LL))
> t.ABL1 <- mt.maxT(exprs(eset)[ABL1PS, ],
                    classlabel = cl, B = 1e+05)
> p.ABL1 <- t.ABL1$rawp[order(t.ABL1$index)]
> names(p.ABL1) <- ABL1PS
> p.ABL1 <- sort(signif(p.ABL1, 2))
> p.ABL1
1636_g_at 1635_at 39730_at 1656_s_at 32974_at 32975_g_at 2041_i_at
  0.00001 0.00001  0.00001     0.058     0.23       0.53      0.59
2040_s_at
     0.76
```

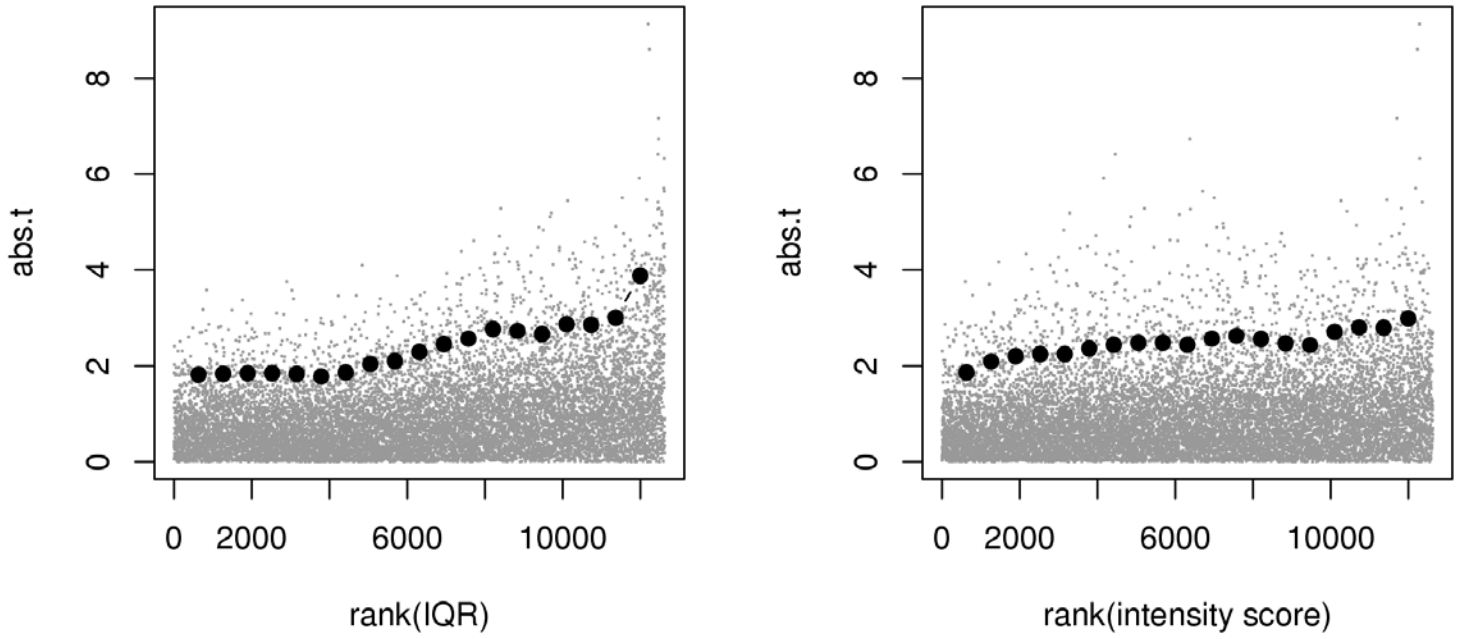
Multiple probe sets per gene



Comparison between t-statistics of 203 pairs of probe sets with same Locuslink IDs.

The relation between prefiltering and multiple testing

```
## Variability based filtering
> IQRs <- esApply(eset, 1, IQR)
## Intensity based filtering
> intensityscore <- esApply(eset, 1, function(x)
                            quantile(x, 0.75))
> abs.t <- abs(mt.teststat(exprs(eset),
                    classlabel = cl))
```

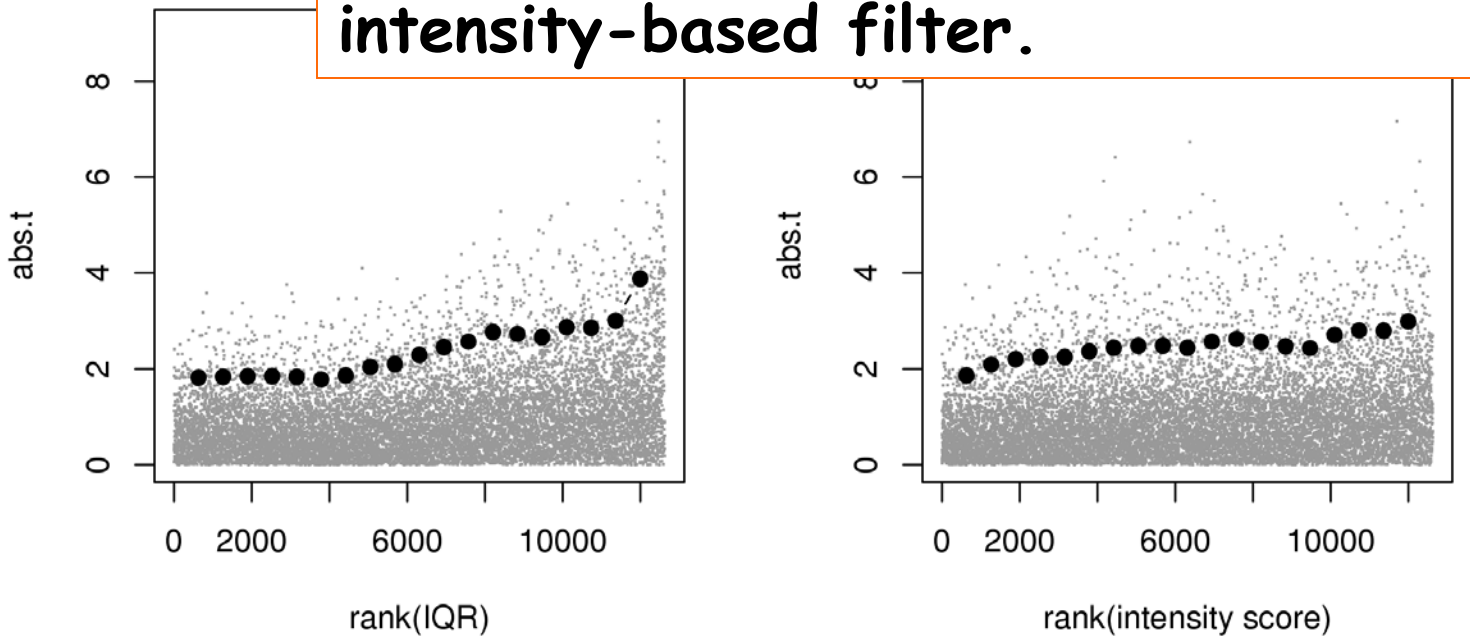


The relation between prefiltering and multiple testing

```
## Variability based filtering
> IQRs <- esApply(eset, 1, IQR)
## Intensity based filtering
> intensityscore

> abs.t <- abs(mt
```

Gene selection by IQR leads to a higher concentration of differentially expressed genes. Less so for intensity-based filter.



Moderated / Bayesian t-tests

With 79 samples, there is no big difference between ordinary and the moderated t-statistic.

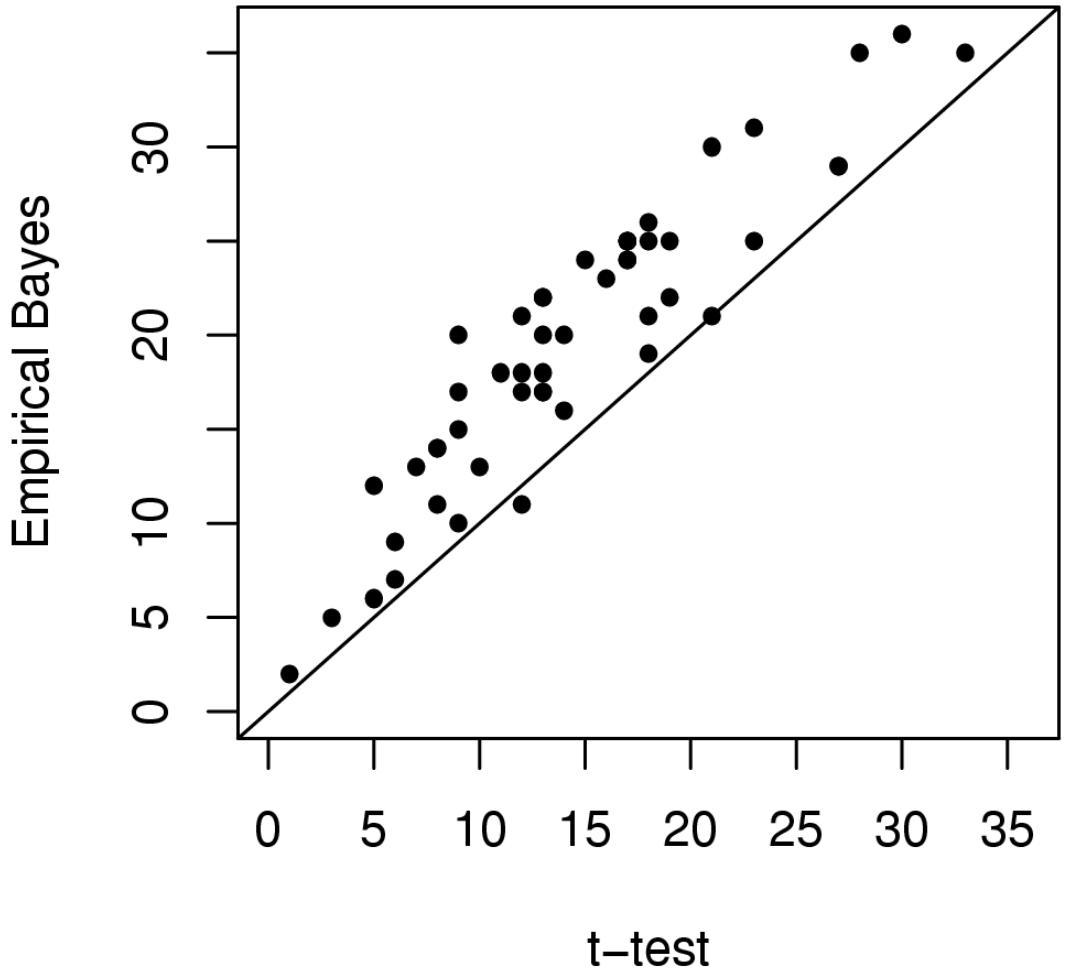
For illustration, look at the behavior of the different approaches for small sample sizes: We repeatedly draw random small sets of arrays from each of the two groups and apply different statistics for differential expression.

The results are compared to those of the analysis of the whole data set. As an approximation, we declare the 109 genes with a FDR below 0.05 (on the whole set of samples) as truly differentially expressed genes.

Moderated t-test

```
> groupsize <- 4
> design <- cbind(c(1, 1, 1, 1, 1, 1, 1, 1),
                  c(0, 0, 0, 0, 1, 1, 1, 1))
> g1 <- sample(which(esetSub$mol == "NEG"), groupsize)
> g2 <- sample(which(esetSub$mol == "BCR/ABL"),
                                          groupsize)
> subset <- c(g1, g2)
> fit <- lm.series(exprs(esetSub)[, subset], design)
> eb <- ebayes(fit)
> tsub <- mt.teststat(exprs(esetSub)[, subset],
                          classlabel = cl[subset],
                          test = "t.equalvar")
> rawpsub <- 2 * (1 - pt(abs(tsub), df=2*groupsize-2))
```

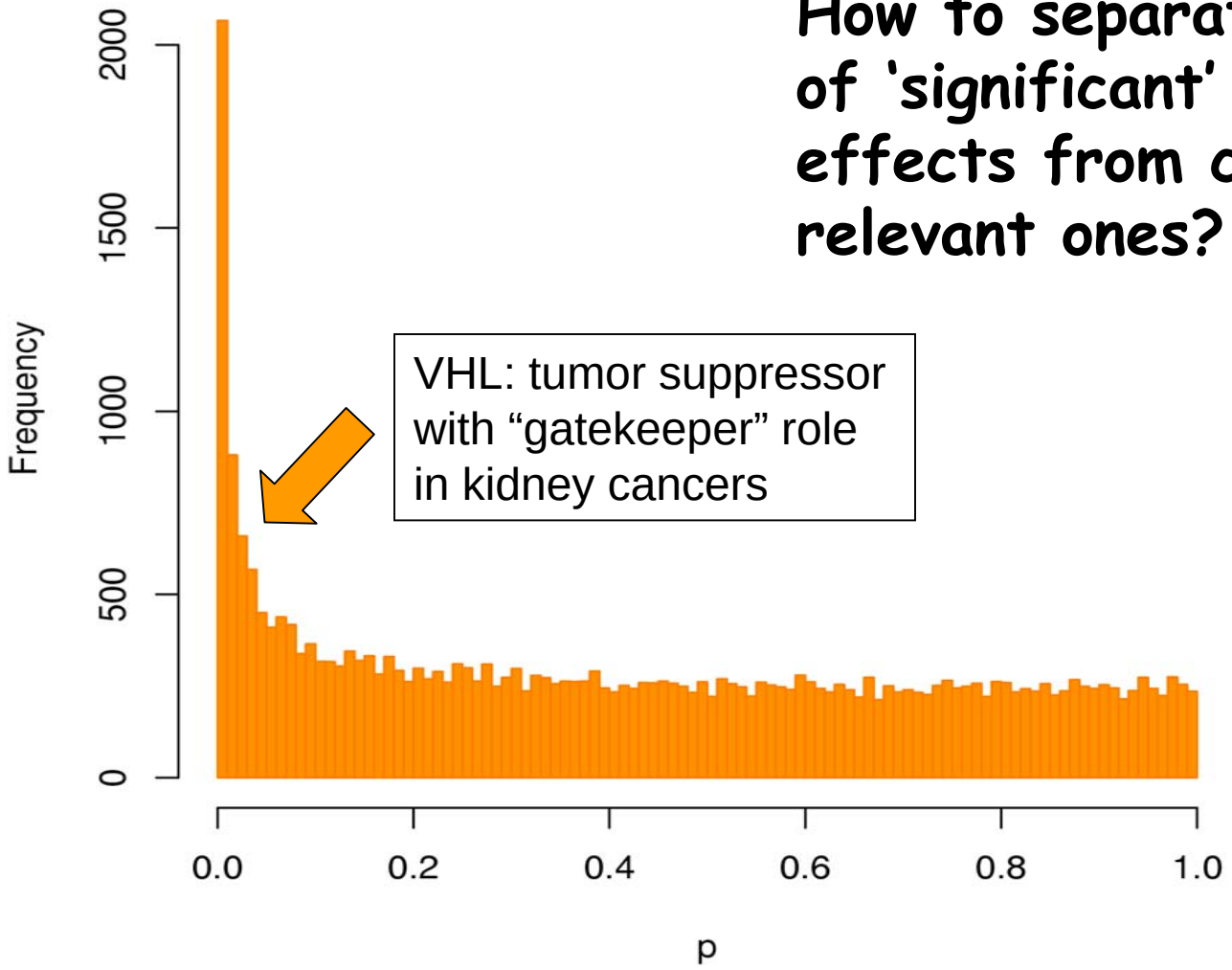
Moderated t-tests



Number of true positives among the top 100 genes selected by the t-test and a test based on a moderated t-statistic, as implemented in the limma package.

Drowning by numbers

How to separate a flood of 'significant' secondary effects from causally relevant ones?



Boer et al. Genome Res. 2001: kidney tumor/normal profiling study

Asking specific questions - using metadata

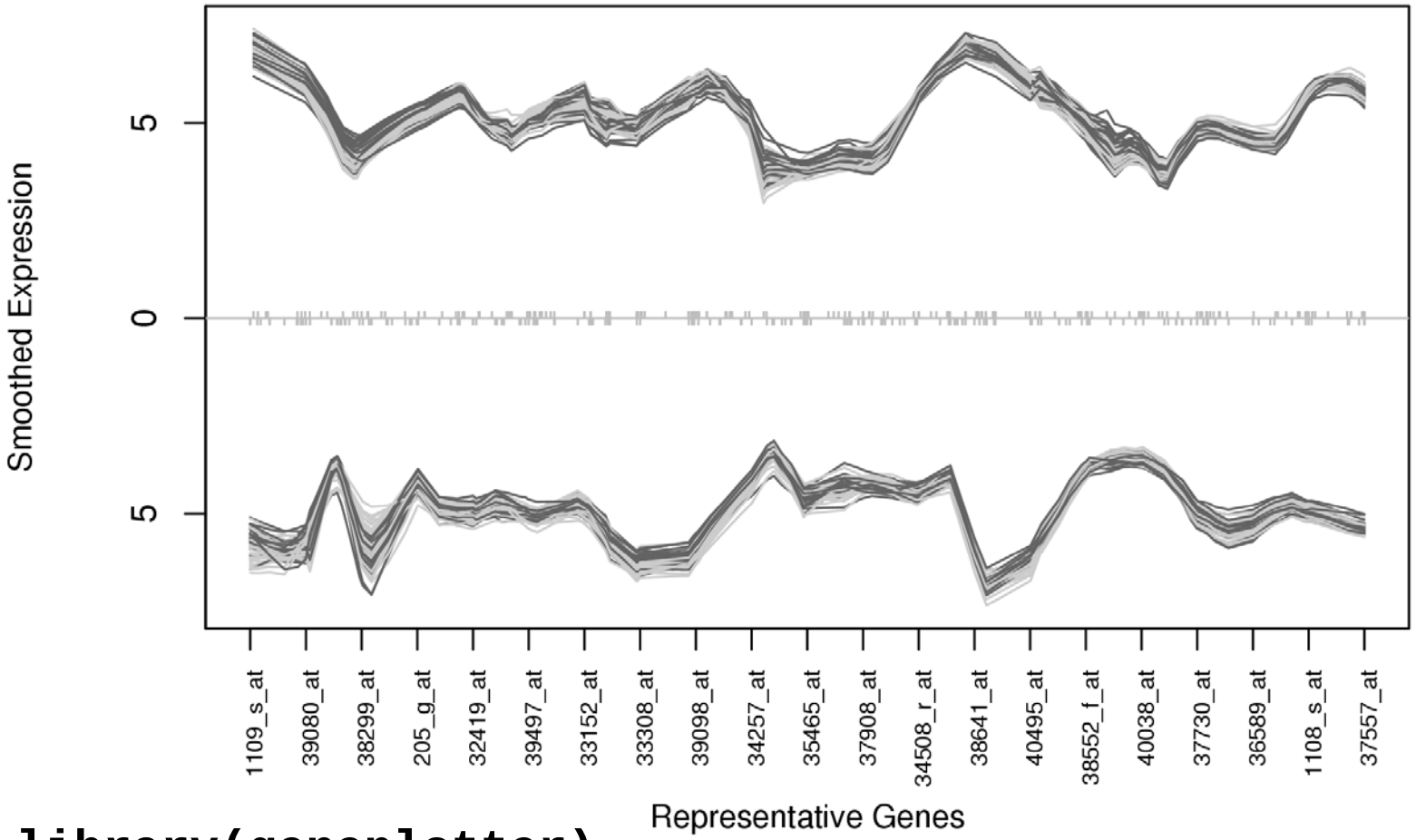
Chromosomal location

Consider all genes with unadjusted p<0.1 (median p if several probe sets per gene). Fisher-test for each chromosome: are there disproportionally many differentially expressed genes on the chromosome?

```
> ll  <- getLL(geneNames(esetSub), "hgu95av2")
> chr <- getCHR(geneNames(esetSub), "hgu95av2")
> chromosomes <- unique(chr[!is.na(chr)])

> ll.pval <- exp(tapply(log(rawp), ll, median))
> ll.chr  <- tapply(chr, ll, unique)
> ll.diff <- (ll.pval < 0.1)
> p.chr   <- sapply(chromosomes, function(x) {
      fisher.test(factor(ll.chr == x),
                  as.factor(ll.diff))$p.value})
> sort(p.chr)
     7     17      X      8     15     21      3      Y      6     12      4 …
0.0086 0.1100 0.1500 0.2000 0.2300 0.3000 0.3000 0.3300 0.3800 0.5100 0.5600 …
```

Chromosome 7




```
> library(geneplotter)
> ms1 <- Makesense(exprs(eset), "hgu95av2")
> plotChr("7", ms1)
```

GO

All genes: 2391 probes from unspecific filtering step.
Go Analysis: 32 that were annotated with "tyrosine kinase activity"

	40480_s_at	2039_s_at	36643_at	2057_g_at
GO analysis	0.00002	0.00025	0.02146	0.07481
All Genes	0.00095	0.01407	0.46938	0.82884

Pathways

In a related disease, chronic myeloid leukemia, BCR/ABL induces loss of adhesion to fibronectin and the marrow stroma.

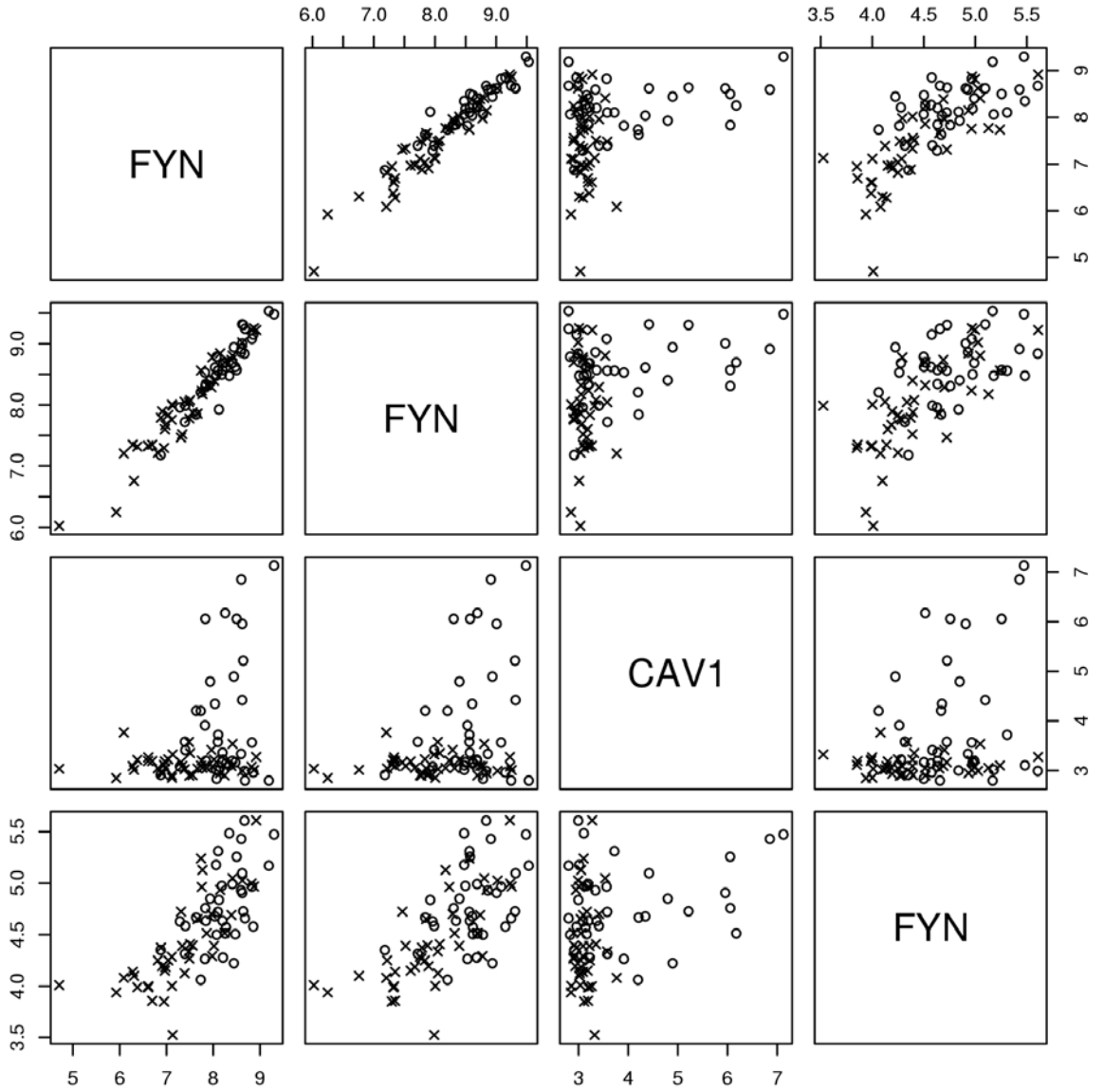
Suggests that there may also be differences between the BCR/ABL + and - samples with respect to expression of genes in the integrin-mediated cell adhesion pathway.

A version of this pathway was obtained from Kyoto Encyclopedia of Genes and Genomes (KEGG) as pathway 04510 (package "KEGG")

114 probes, 71 unique LocusLink Ids

4 differentially expressed (3 FYN, 1 CAV1)
2 FYNs were also selected previously, but not CAV1

Pathways



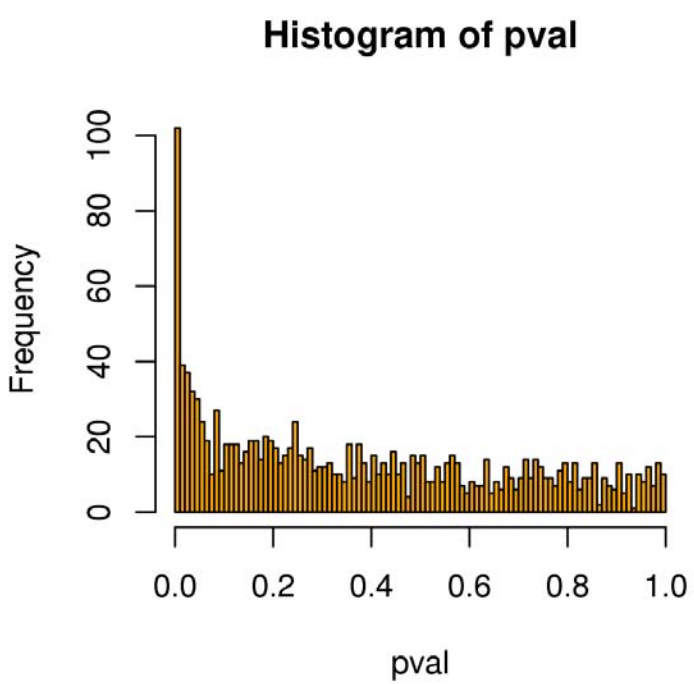
O BCR/ABL+

X BCR/ABL-

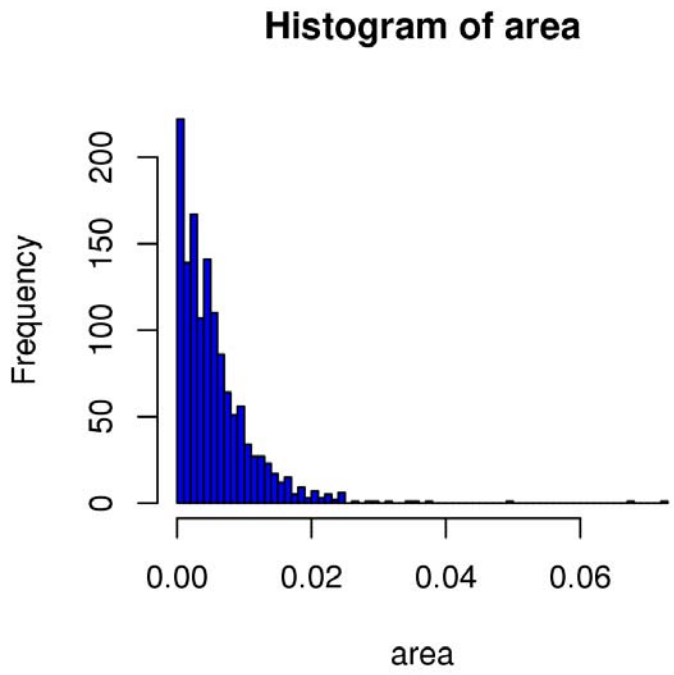
Discrimination scores - ROC curve analysis

.Call("Axel Benners Talk")

Discrimination scores - ROC curve analysis

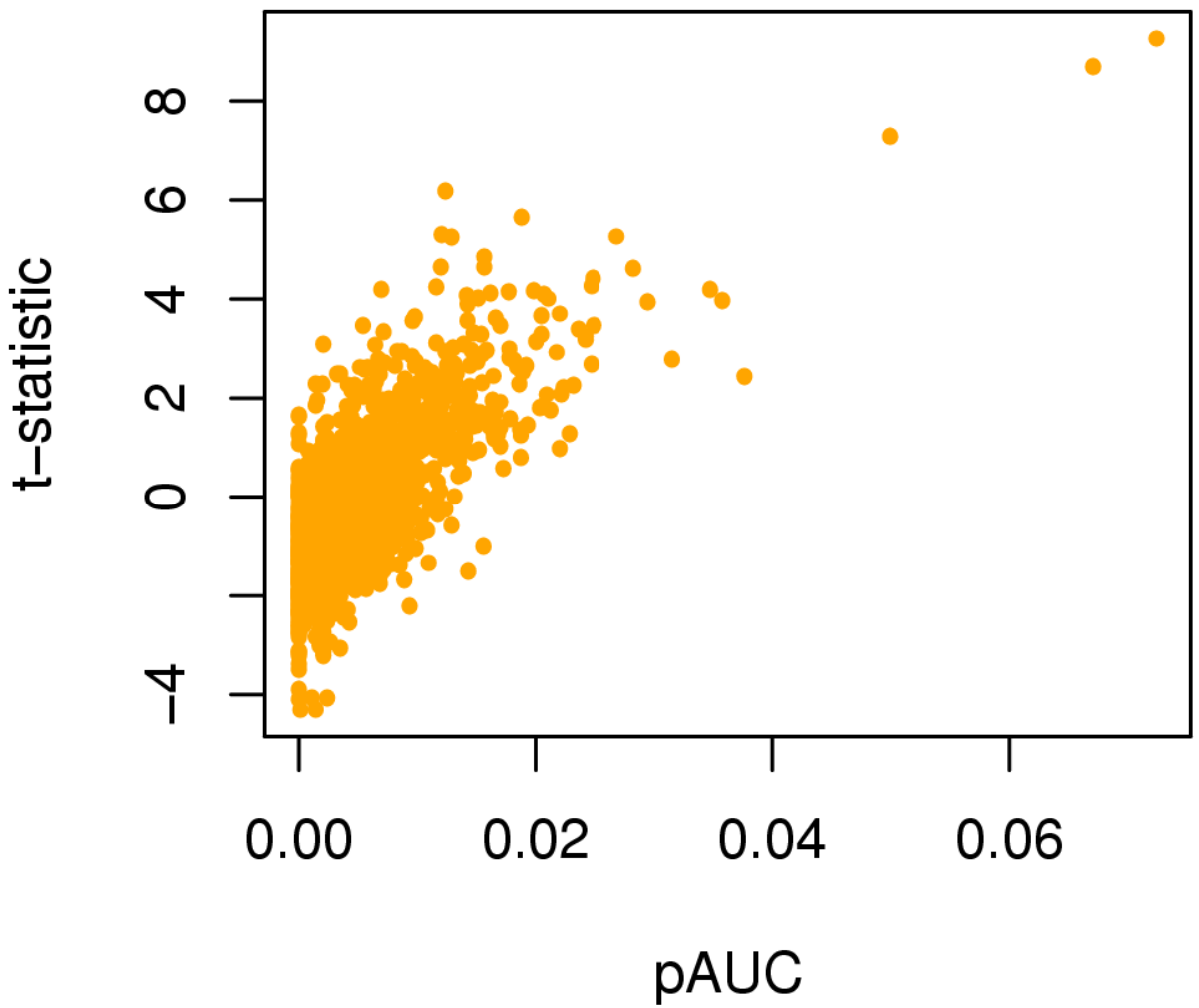


t-test

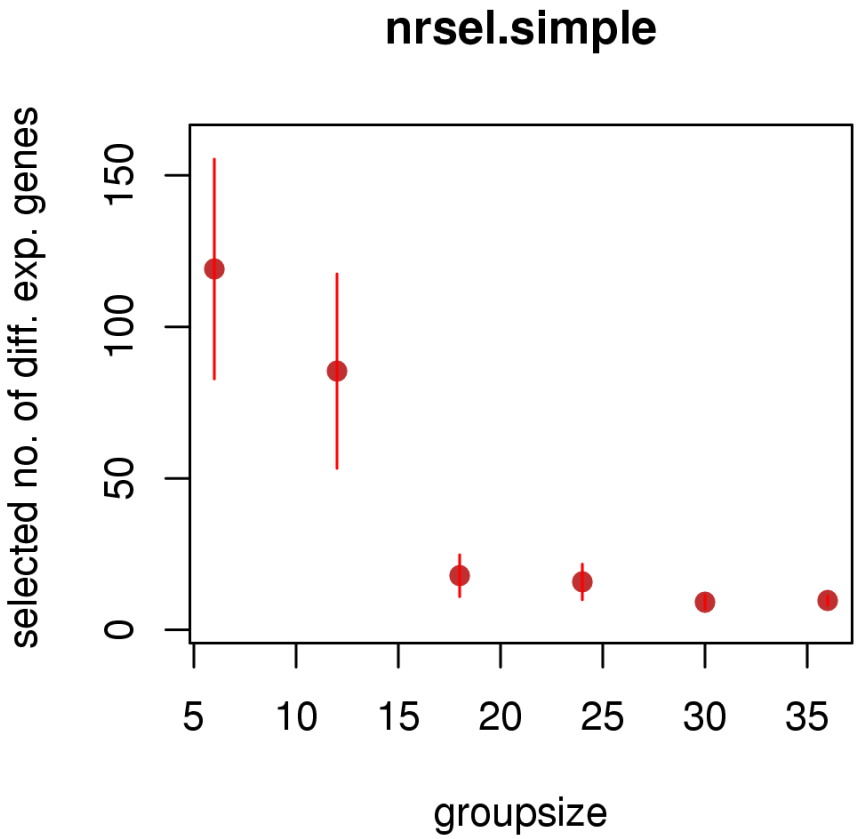


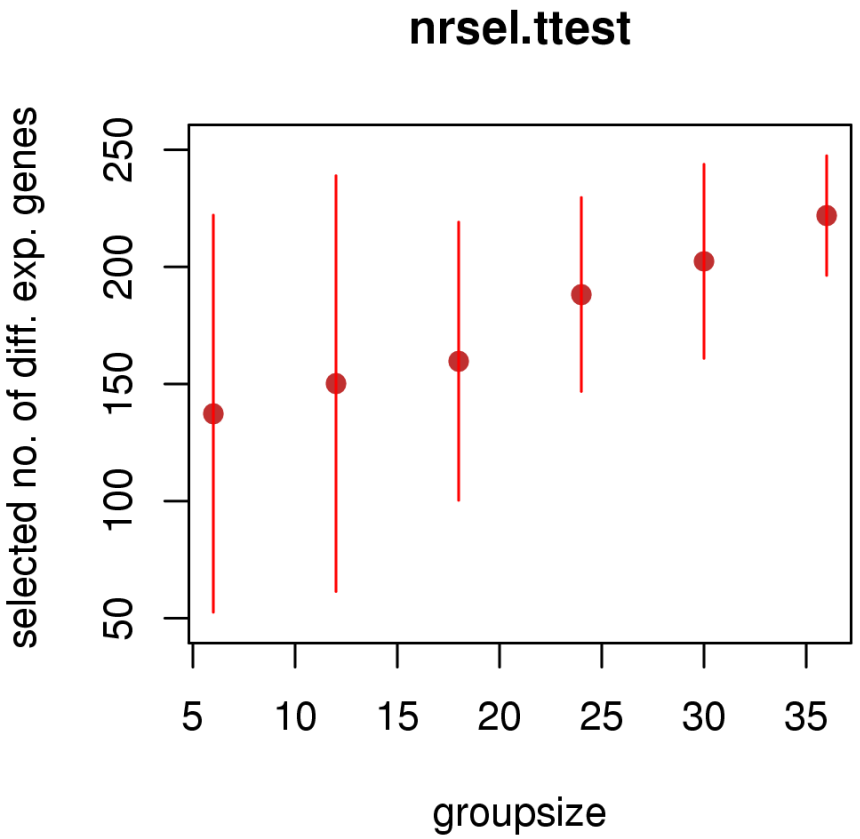
pAUC

Discrimination scores - ROC curve analysis



Discrimination scores - ROC curve analysis





Discrimination scores - ROC curve analysis

See the vignette "tvsroc.Rnw"

Conclusion

- Testing all genes on the chip on after the other and correcting for multiplicity is a band-aid, not a good solution.
- Large Loss of power
- Biologically most relevant need not be statistically most significant (VHL/kidney!)
- Drowning in numbers (secondary effects)
- Bioconductor offers a lot of infrastructure to use metadata and directed hypotheses on genes - use it!