Differential expression

Wolfgang Huber

Robert Gentleman

Anja von Heydebreck

Florian Hahne

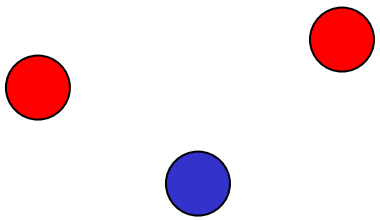
p >> n

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: Hyperplanes

Problem: If you find a perfectly separating hyperplane, it doesn't mean anything. It is not surprising. It is not a significant finding. You will 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.

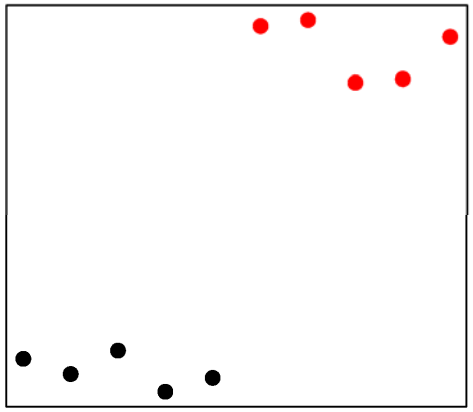
Two major approaches:

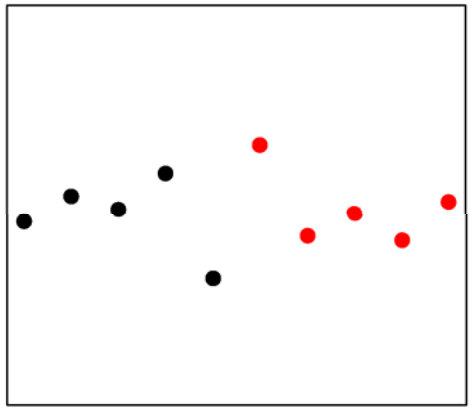
- only the hyperplanes perpendicular to one of the n coordinate axis ⇒ gene-by-gene discrimination, gene-by-gene hypothesis testing.
- any other reasonable, not too complex set of hypersurfaces ⇒ machine learning

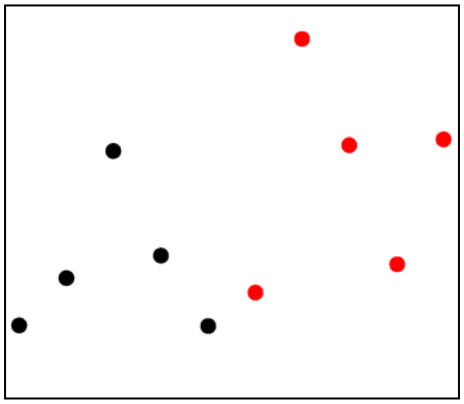
The question

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

The gene-by-gene approach:

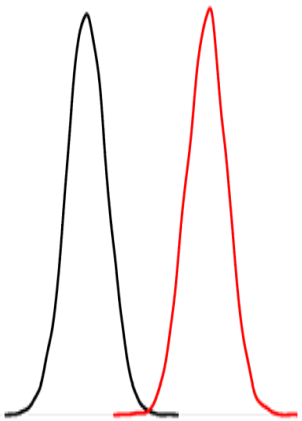


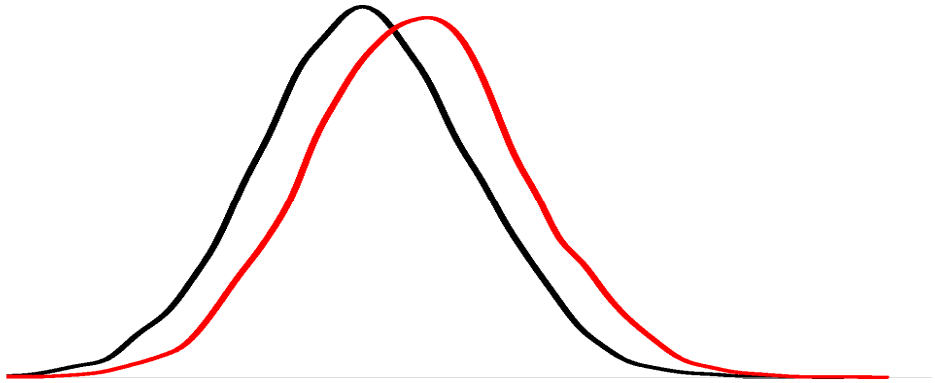




The question

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





Fold change vs p-value

Problem: there are two basic selection strategies that are widely used

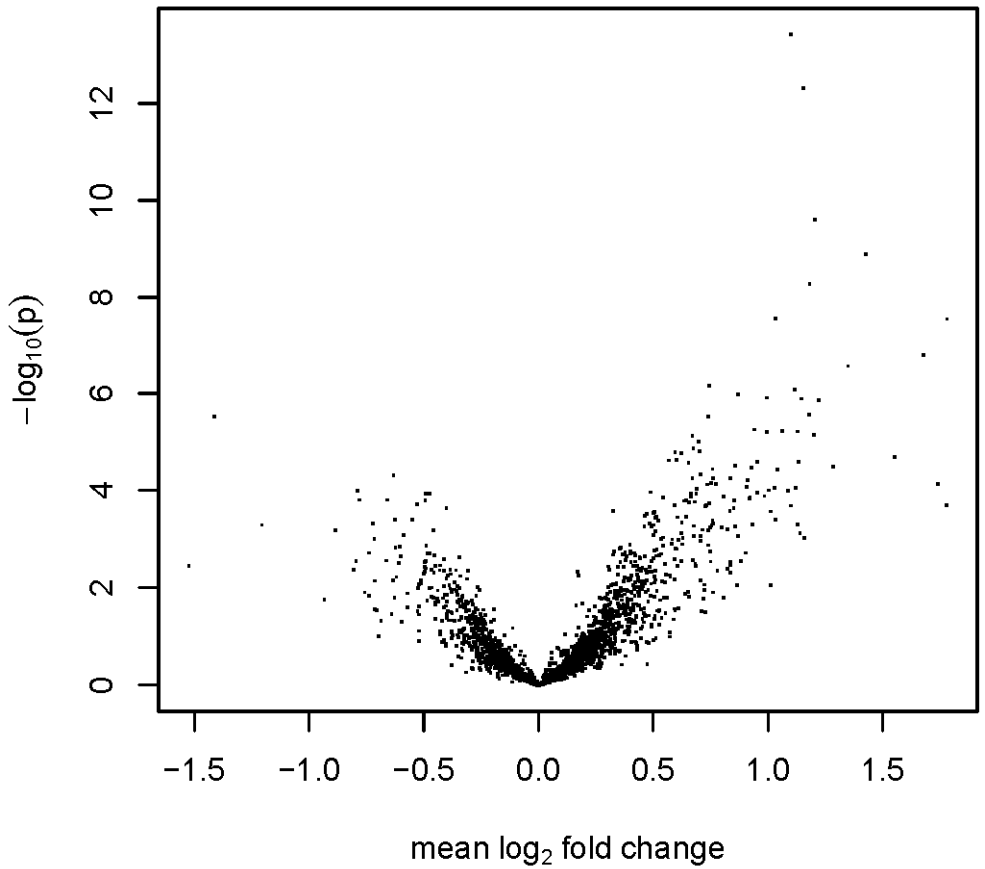
Fold change (effect size):
if the size of the effect (for two sample comparisons we often call this the fold-change) is sufficiently large; often values like 1.5 or 2.0 are used

p-value:

- genes are deemed to be interesting if the observed p-value is suitably small

Fold change vs p-value

Volcano plot:



Modeling Considerations

Problem: with few arrays you are unwilling to make parametric assumptions about gene expression values

Nonparametric assumption: the use of a permutation test, or similar non-parametric tool is tempting

But: such assumptions reduce the power and hence the ability to discriminate. When you do not have much data (many samples) you need a model to help make inference.

Aggregation across genes: one of the basic strategies used is to aggregate information across genes

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.

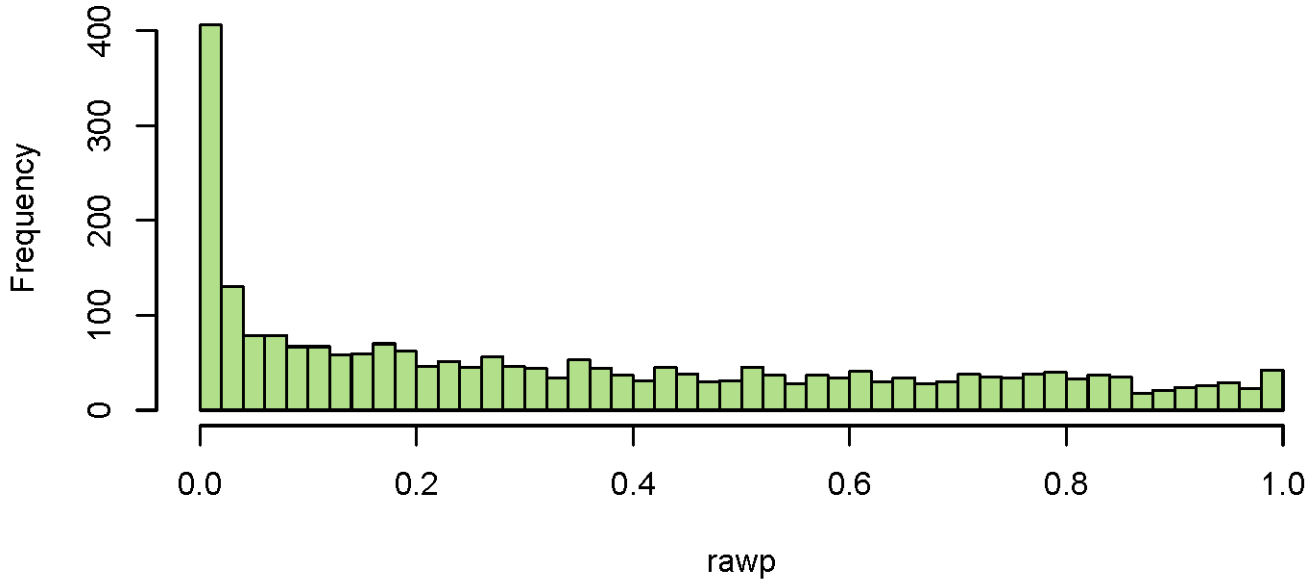
t-test

Test for differences in means between two groups given the variability within each group

$$\frac{\textit{difference between group means}}{\textit{variability of groups}}$$

$$\frac{\overline{X}_1 - \overline{X}_2}{SE\left(\overline{X}_1 - \overline{X}_2\right)}$$

distribution of p-values



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
Kendziorski et al. (Ebarrays) 2003
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

Moderated / Bayesian t-tests

solution: in most cases, an overall estimate of the variance, s_o^2, is computed

- then for each gene, an estimate of the per gene variance, s_g^2, is computed

- the variance used is a weighted average of s_o^2 and s_g^2

- the actual method of estimating the overall variance and the method of averaging is slightly different in different contexts

Moderated / Bayesian t-tests

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

But for smaller data sets the differences will be larger.

To test how these two procedures might compare in practice we devise the following simulation (our problem here is the lack of a gold standard data set).

We will declare the 109 genes with a FDR below 0.05 (on the whole set of samples) as truly differentially expressed genes.

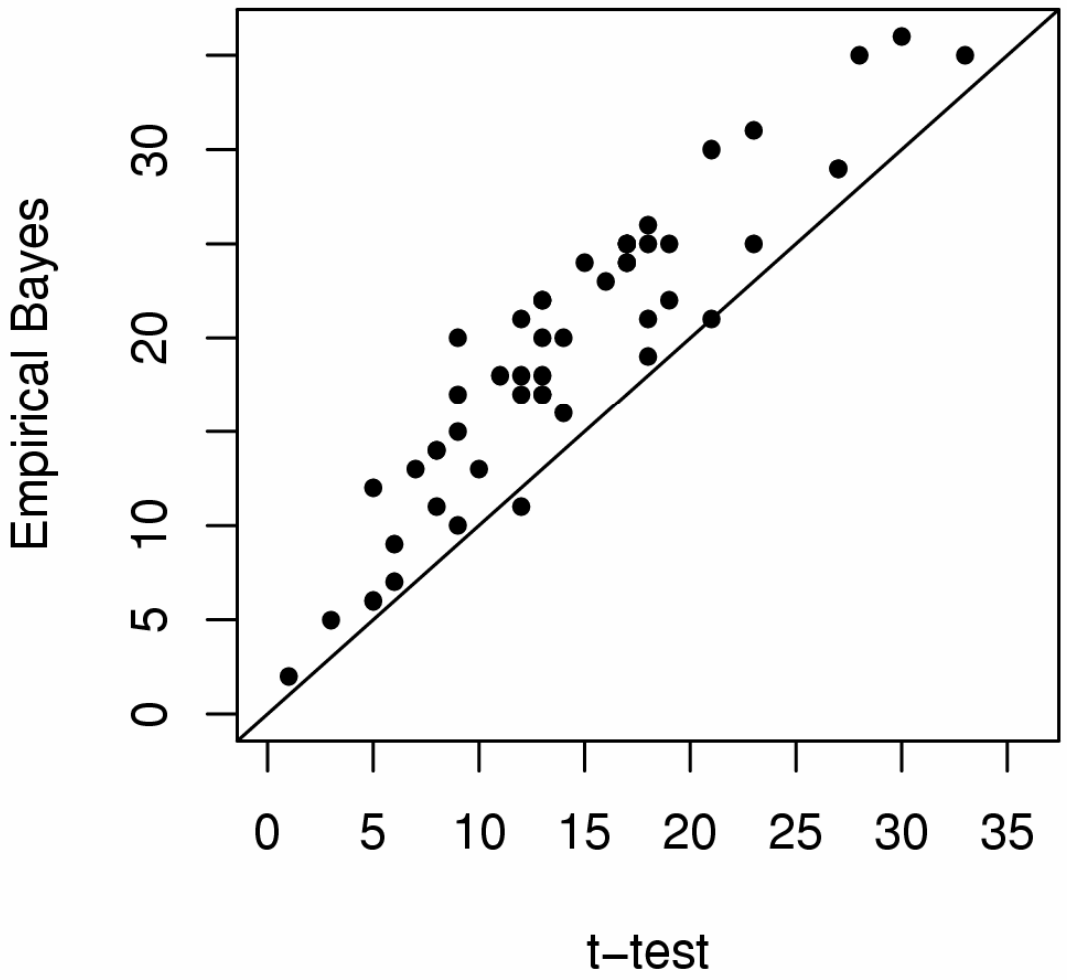
Moderated / Bayesian t-tests

We then sample from this data set 8 arrays, 4 from each of the two phenotypes of interest.

For each sample we use both a t-test and a moderated t-test to determine differentially expressed genes.

In the next picture we compare the number of truly differentially expressed genes selected, by each method, on each run.

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.

p-value corrections

problem: we have made very many tests and the resulting p-values are difficult to interpret

band-aid: statisticians have turned p-value corrections into a growth industry - but they are really more of a band-aid than a solution

solution: test fewer, more directed hypotheses - you will still need to correct, but the amount of correction needed will be much smaller

p-value corrections

methodology: there are now more methods than we could ever consider

- but basically what they all do is to reduce the critical value used to determine whether or not to reject

- since the truly false hypotheses tend to have smaller p-values, this adjustment enriches those rejected for those that are truly false

- but among the casualties are those hypotheses that are truly false, but which did not obtain an extraordinarily small p-value

→ trade-off between sensitivity and specificity

p-value corrections

software: the multtest package (by K. Pollard, Y. Ge and S. Dudoit) provides a wide variety of p-value correction methods

- multtest provides a variety of t- and f-tests, including robust versions of each test
- Single-step and step-down minP and maxT methods can be used to control the chosen type I error rate
- options for error rate control include FWER, gFWER, TPPFP FDR
- check the vignette and other package documentation for more deatils

FWER

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

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

This is a large loss of power!

FDR

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

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

Data Reduction

Problem: most of the genes do not show differences in expression across the arrays

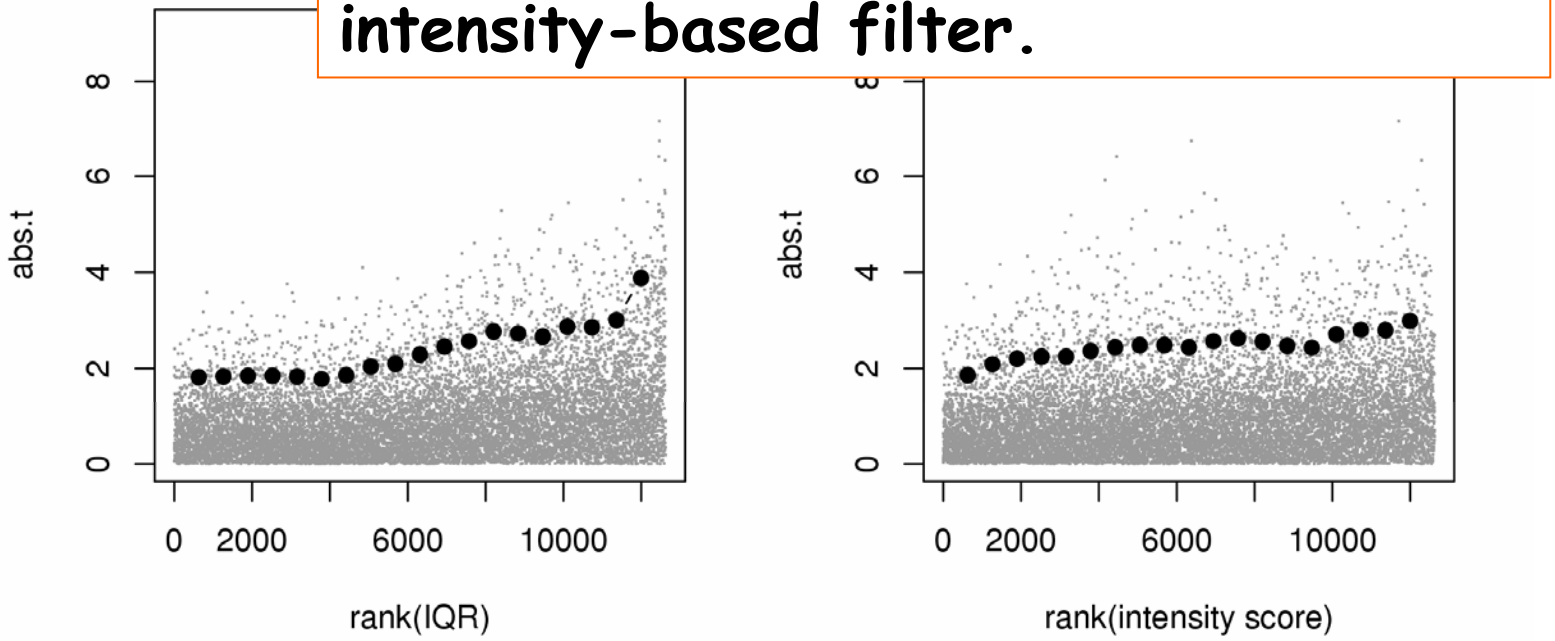
- you should consider a reduction in the set of gene/probes that are under consideration
- not all genes are expressed in all tissues
- one of the basic assumptions of normalization is that most of the genes have not changed expression levels across conditions
- these observations argue in favor of reducing the set of genes
- we recommend using some form of non-specific filtering

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.



Variability Filtering

Problem: as we have noted earlier, the expression estimate itself does not tell us about mRNA abundance

-we noted that only within-gene between array comparisons are valid

- filtering on absolute expression values (say removing those below 100) is falling into that same trap - absolute numbers do not tell us about the true mRNA abundance

- you are probably better off filtering genes by some measure of the variability (MAD, IQR, etc) across arrays

- genes that show no variation across the conditions measured are not interesting

Top 5 (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"
```

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

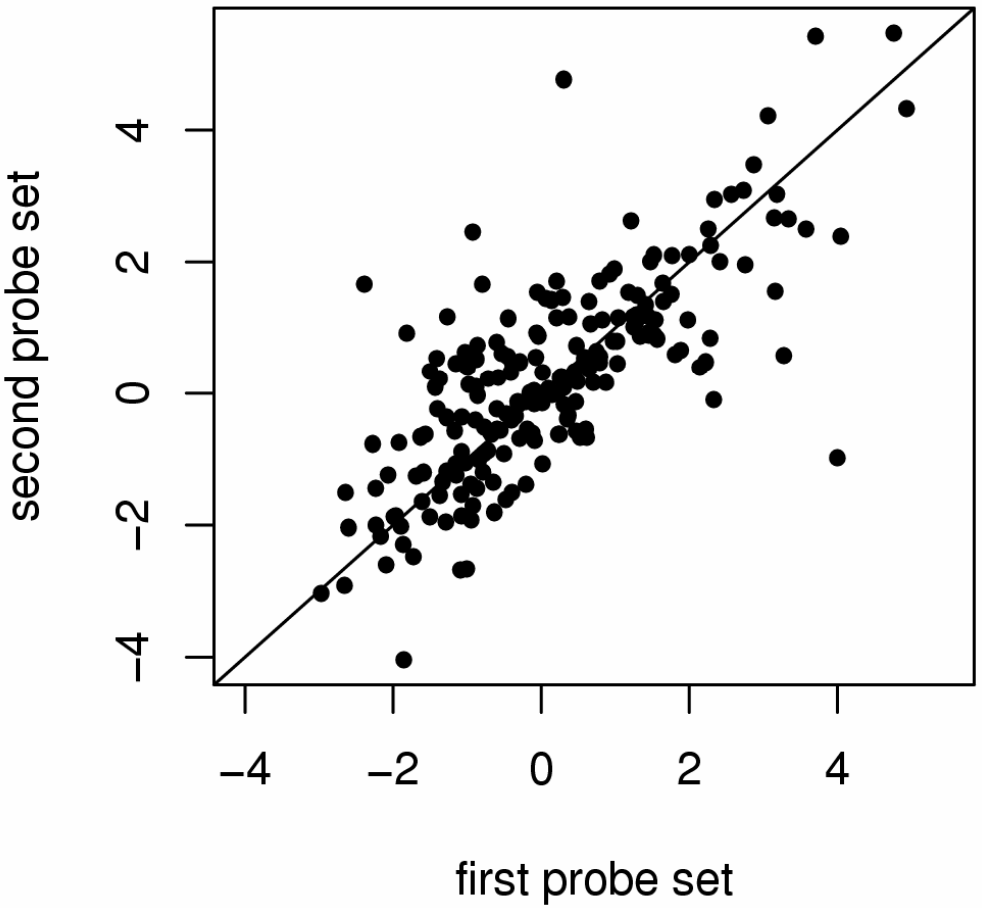
Of the 2263 LocusLink IDs that have more than one probe set identified with them, in 509 cases the nonspecific filtering step selected some, but not all corresponding probe sets.

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:

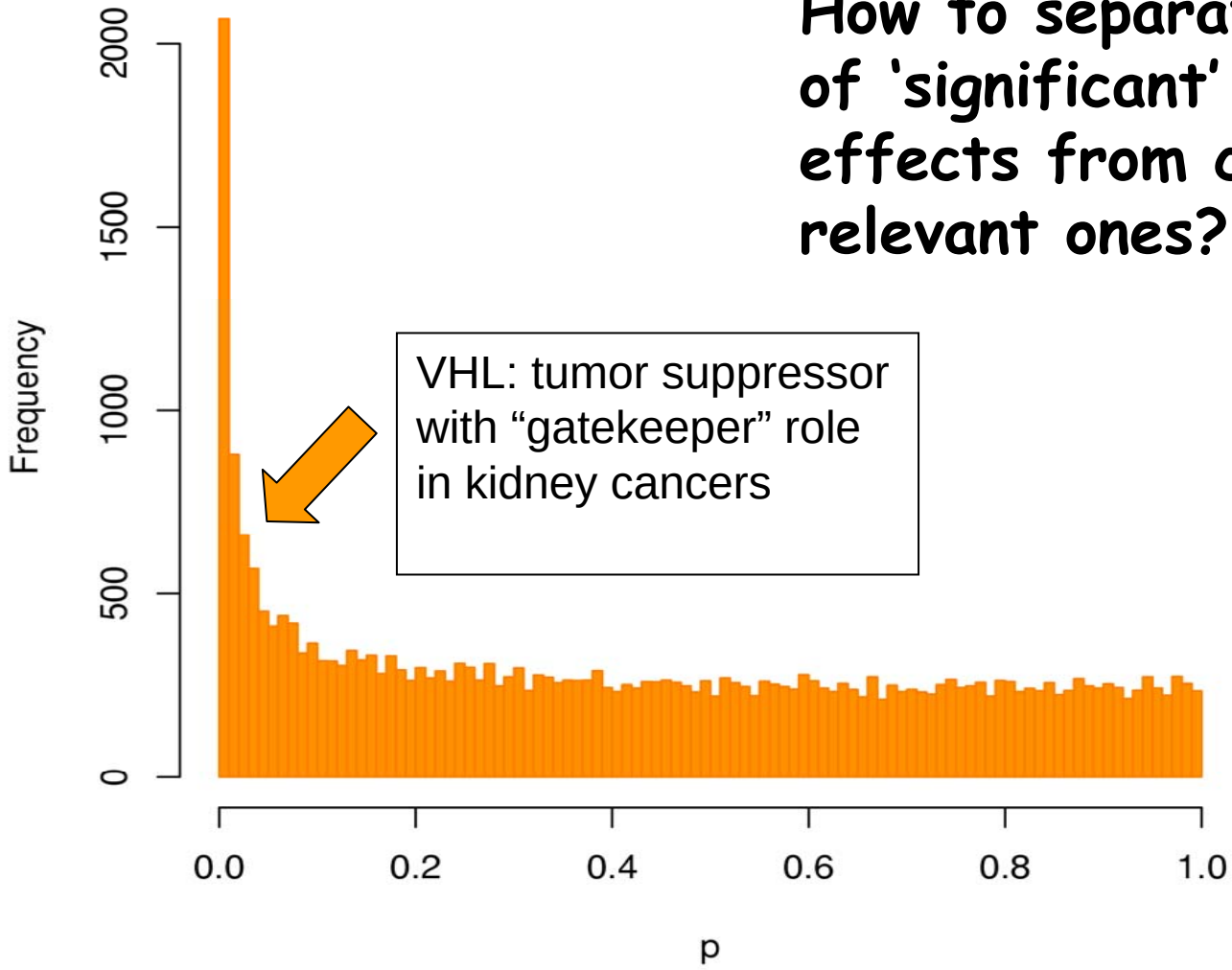
```
> ABL1PS <- names(which(lls == ABL1LL))
> t.ABL1 <- mt.maxT(exprs(eset)[ABL1PS, ],
                    classlabel = cl, B = 1e+05)
> p.ABL1 <- t.ABL1$rawp[order(t.ABL1$index)]
> names(p.ABL1) <- ABL1PS
> p.ABL1 <- sort(signif(p.ABL1, 2))
> p.ABL1
1636_g_at 1635_at 39730_at 1656_s_at 32974_at 32975_g_at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

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?

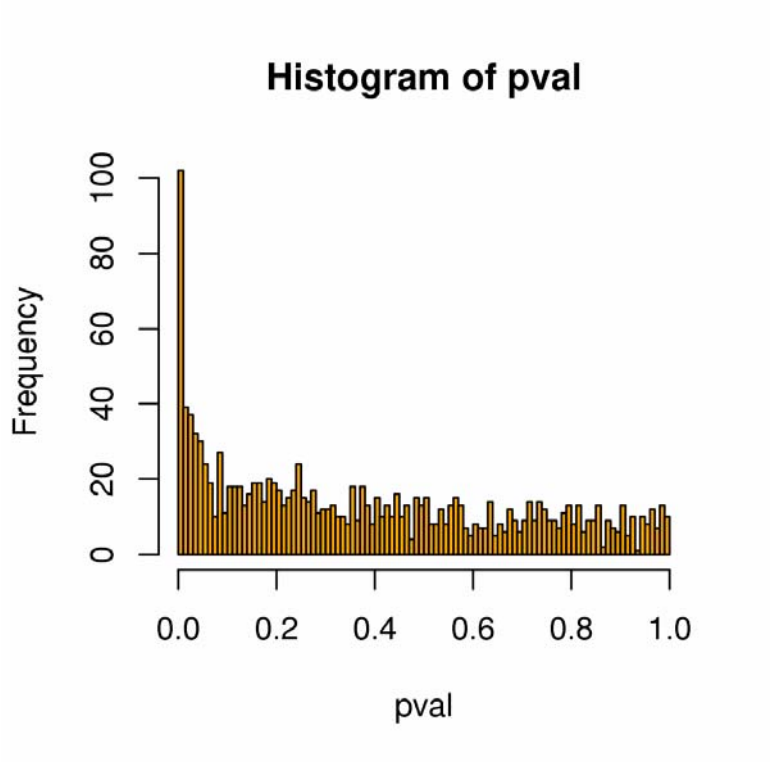
```
> ll  <- getLL(geneNames(esetSub), "hgu95av2")
> chr <- getCHR(geneNames(esetSub), "hgu95av2")
> chromosomes <- unique(chr[!is.na(chr)])

> 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 …
```

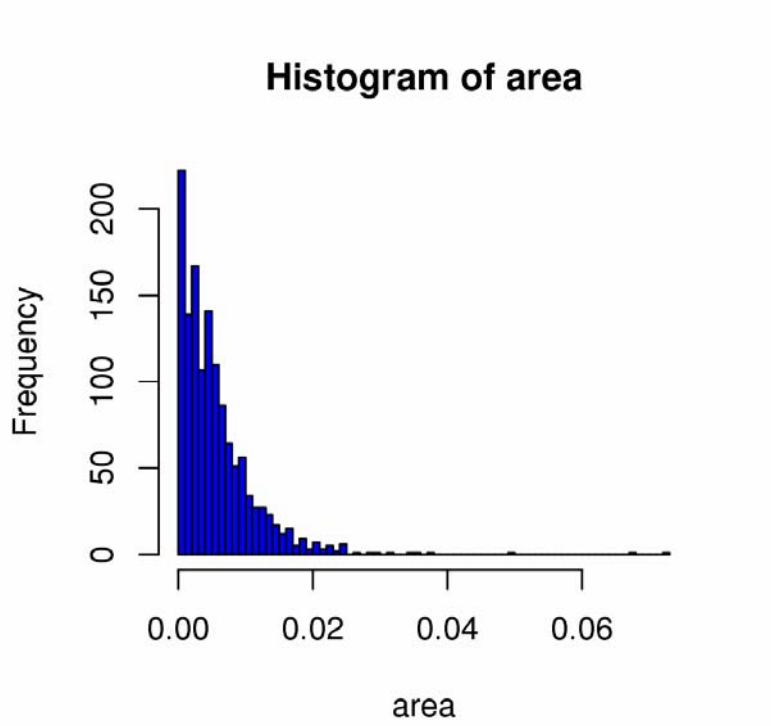
Discrimination scores - ROC curve analysis

.Call("Axel Benners Talk")

Discrimination scores - ROC curve analysis

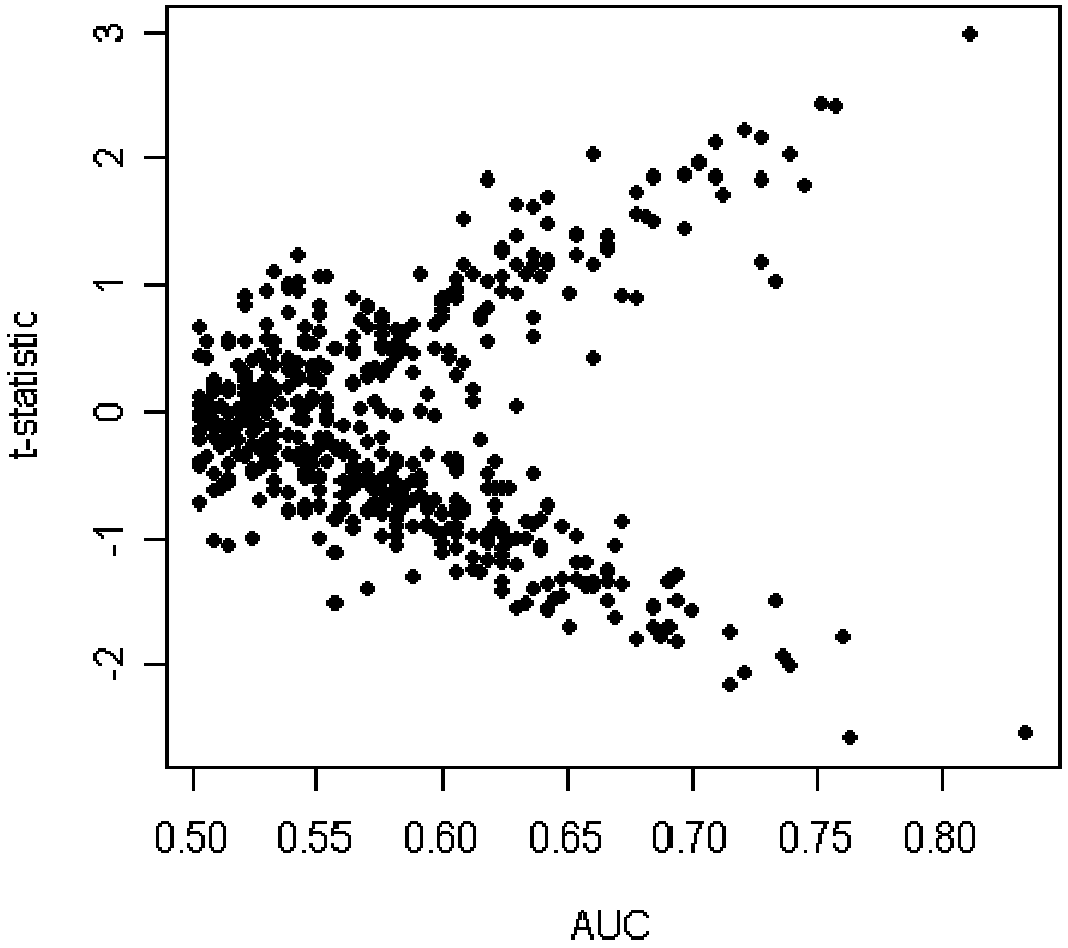


t-test

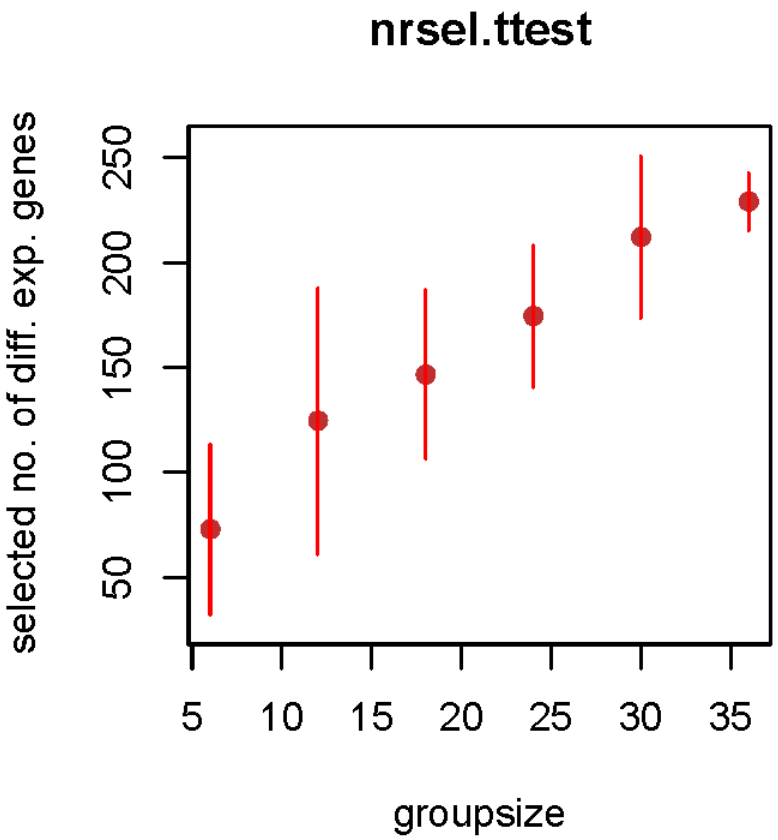


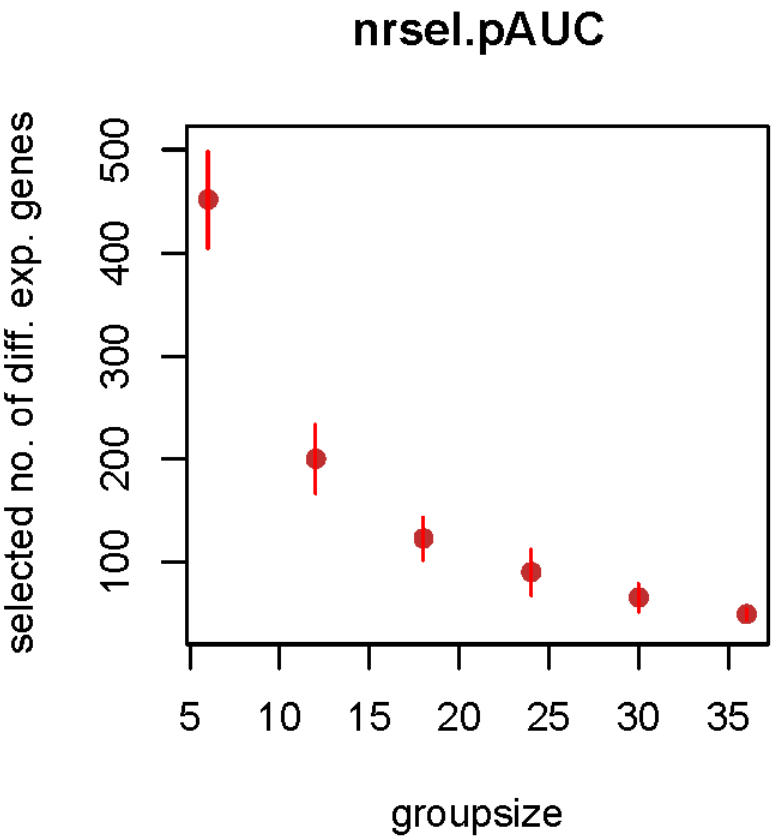
pAUC

Discrimination scores - ROC curve analysis



Discrimination scores - ROC curve analysis





Conclusion

- Testing all genes on the chip one 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!