Week 2

- Last week, we talked broadly about the statistical difficulties inherent in an omics project

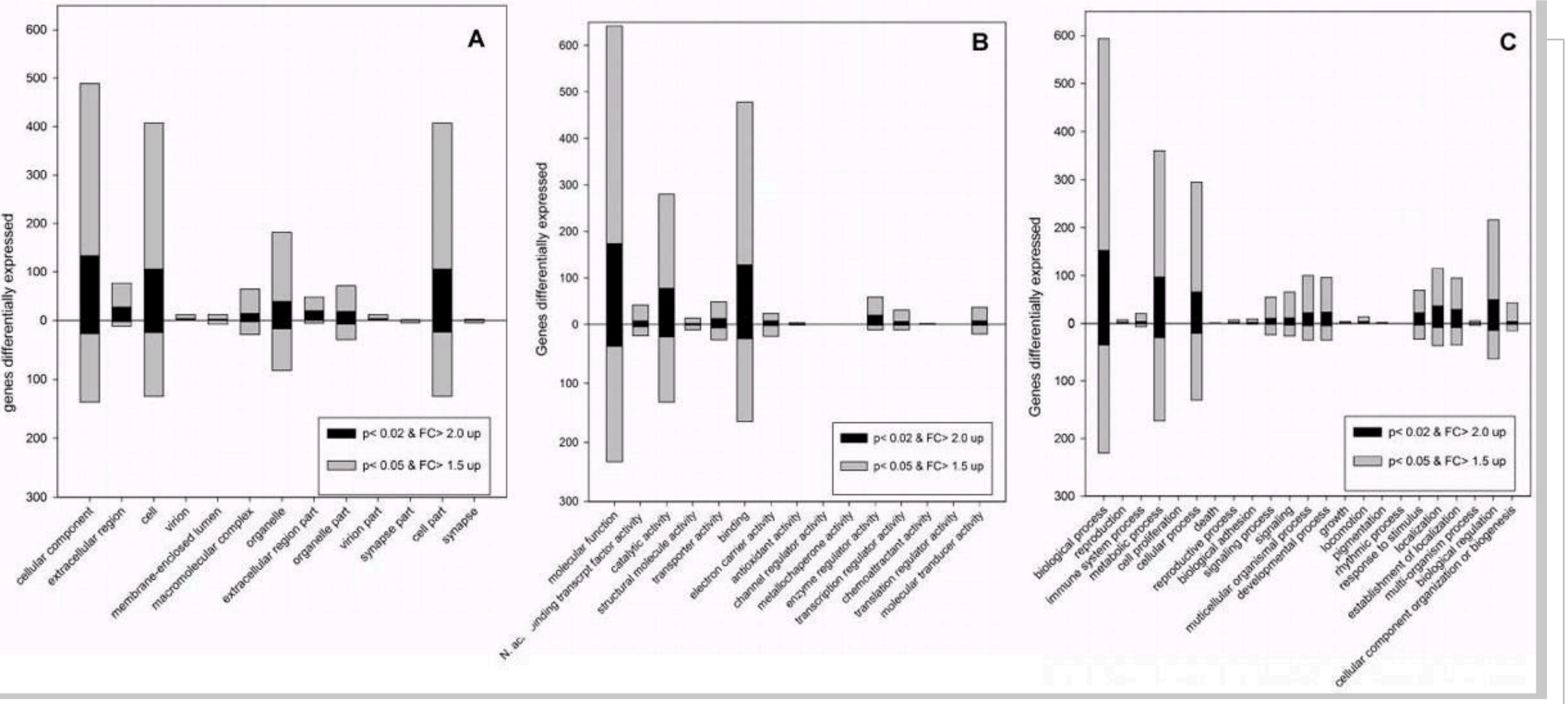
- This week, we will look at how to do a two-class comparison to generate a list of differentially expressed genes. . . and why a two-class comparison isn’t always the best way to interrogate transcriptomic data

Differentially Expressed Genes

- Researchers default to t-test for two class-comparisons or a fold-change cut-off

- What might be some the problems using a student’s T-test? Or a fold-change cut-off of two? 1.5?

Differentially Expressed Genes



BMC Bioinformatics. 2012; 13(Suppl 2): S11.
Published online 2012 Mar 13. doi: 10.1186/1471-2105-13-S2-S11

PMCID: PMC3305783
PMID: 22536862

Fold change and p-value cutoffs significantly alter microarray interpretations

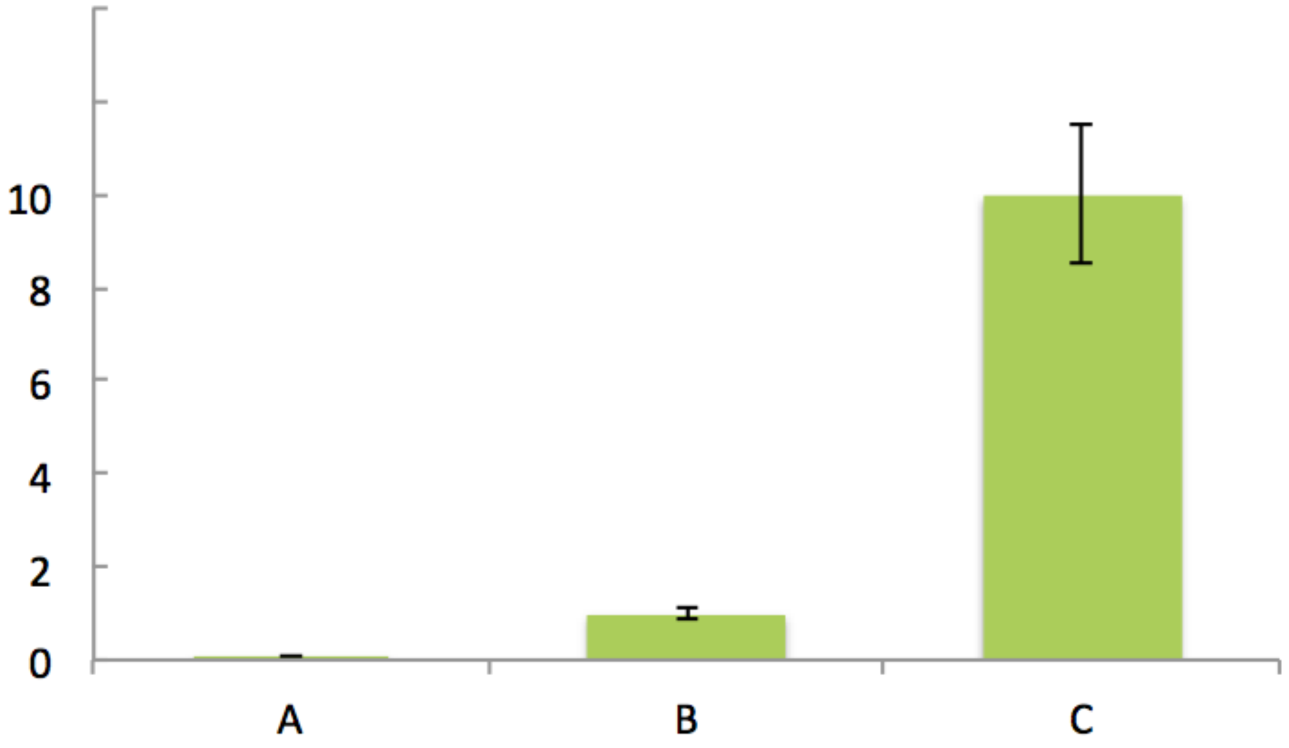
Mark R Dalman,✉[1] Anthony Deeter,[2] Gayathri Nimishakavi,[2] and Zhong-Hui Duan[2]

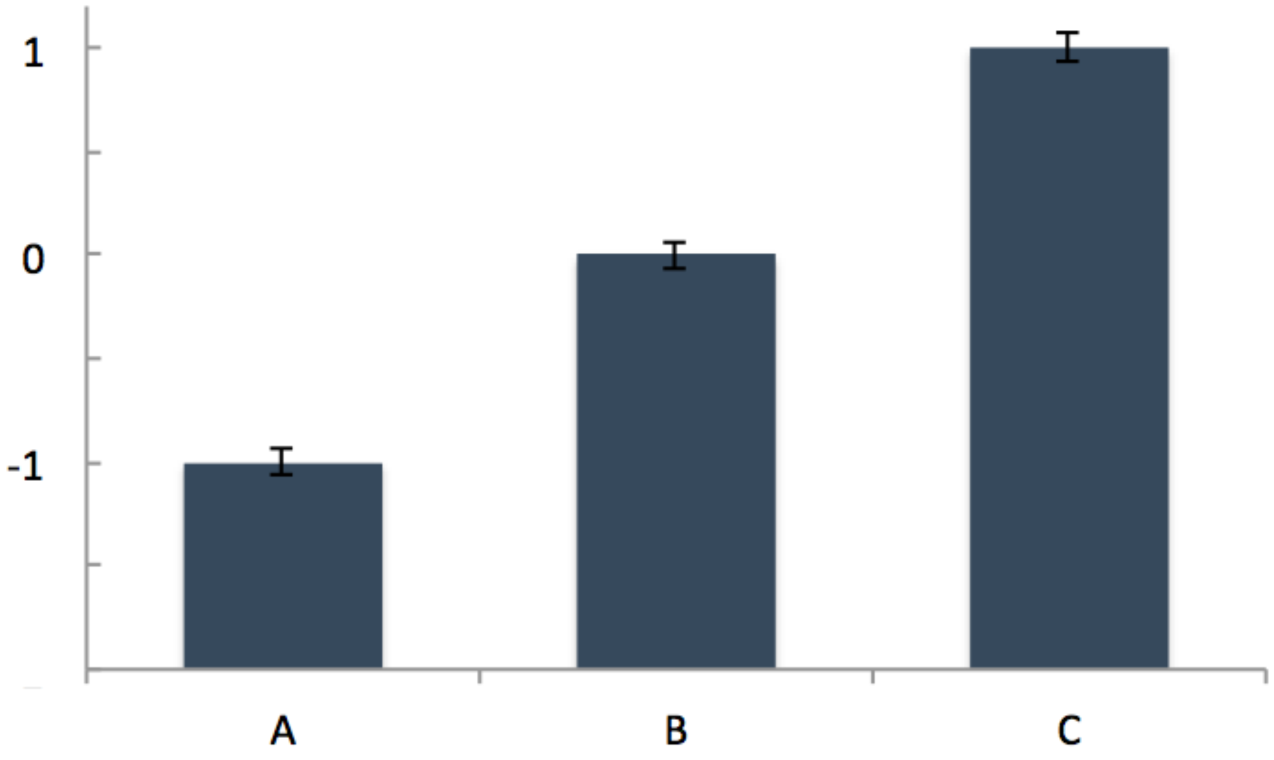
Differentially Expressed Genes

- Arbitrary fold-change cut-offs are acceptable. . . but only as dimensionality reduction
- T-tests assume:
    - Normal distribution - are there any reasons you would expect genes NOT to be normally distributed?
    - Independence - do genes vary independently?
    - A reliable estimate of variability
    - Question: Can you get statistical significance from a microarray or an RNASeq study with two replicates? Why do you need a reliable estimate of variability?

Always Log Transform Your Data

- Always log transform your data
- Gene expression data are heavily skewed - half of the genes typically have a fold-change between 0 and 1, and the other half between 1 and infinity
- **Never** use a parametric test on non-transformed data



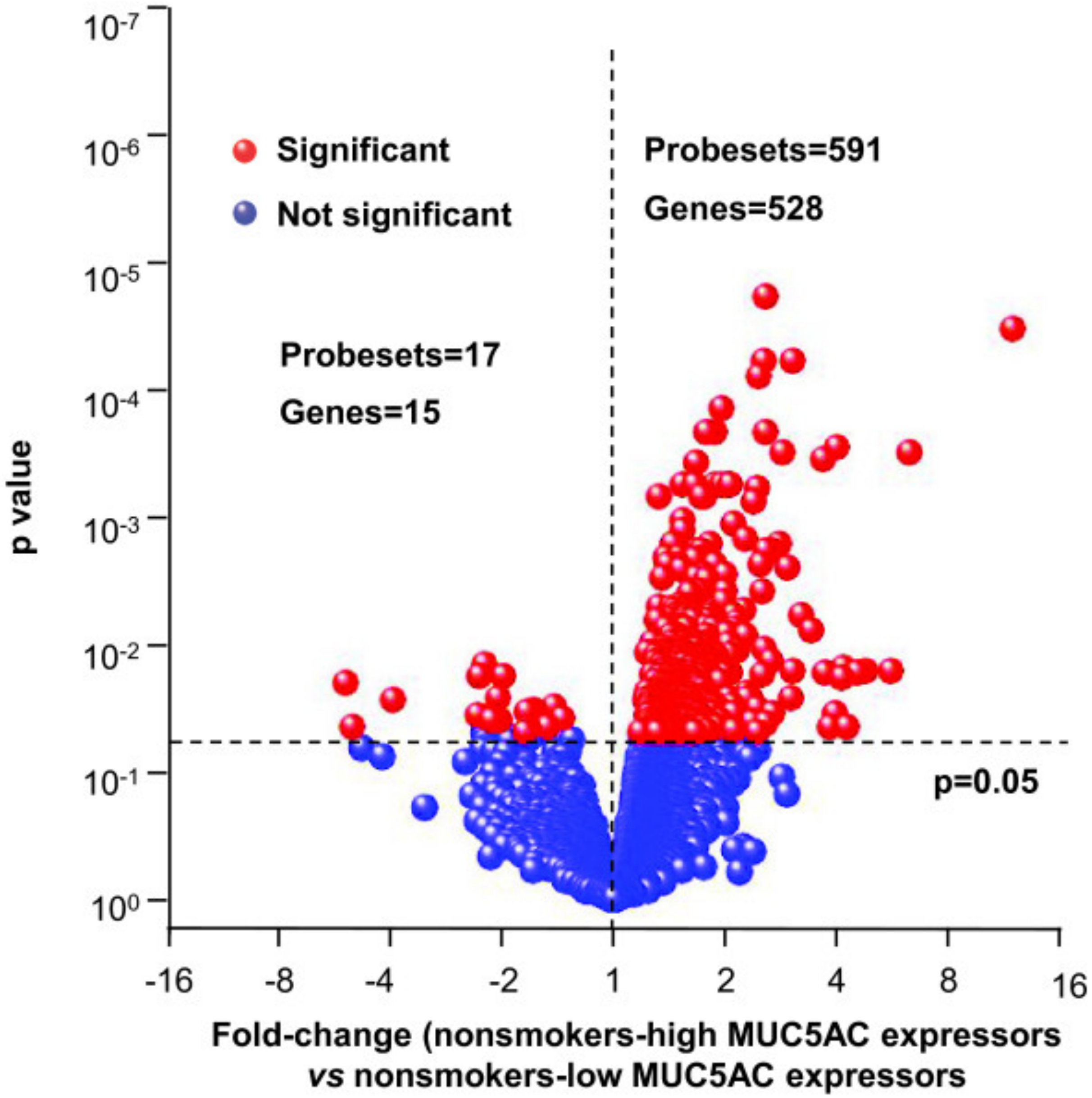


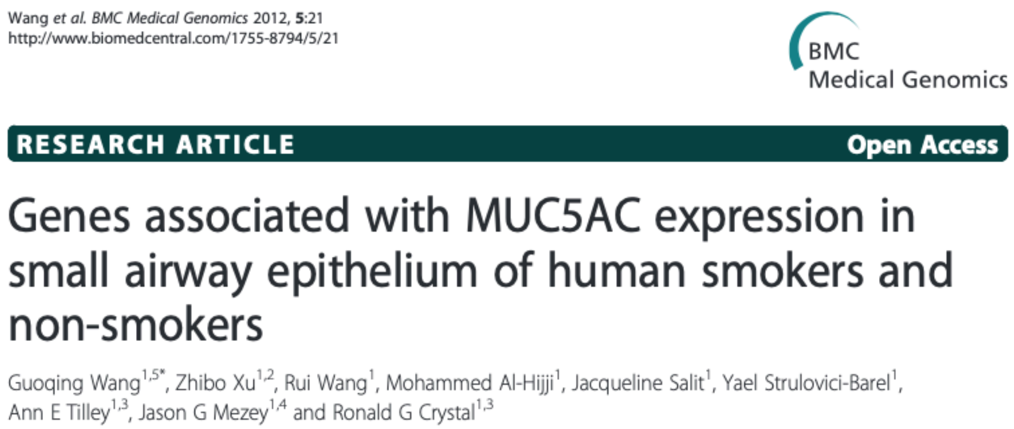
T-test

- Estimating variability is difficult with very few samples
- *Limma* avoids this problem by (1) estimating the average variance of all genes as the *expected* variance and (2) this information is used in a Bayesian estimate of the variability o a given gene/transcript
- Available as a package in R as well as implemented in GEO2R

Volcano Plots

- Volcano plots allow you to quickly to look at both biological and statistical significance





Wang *et al. BMC Medical Genomics* 2012, **5**:21
http://www.biomedcentral.com/1755-8794/5/21

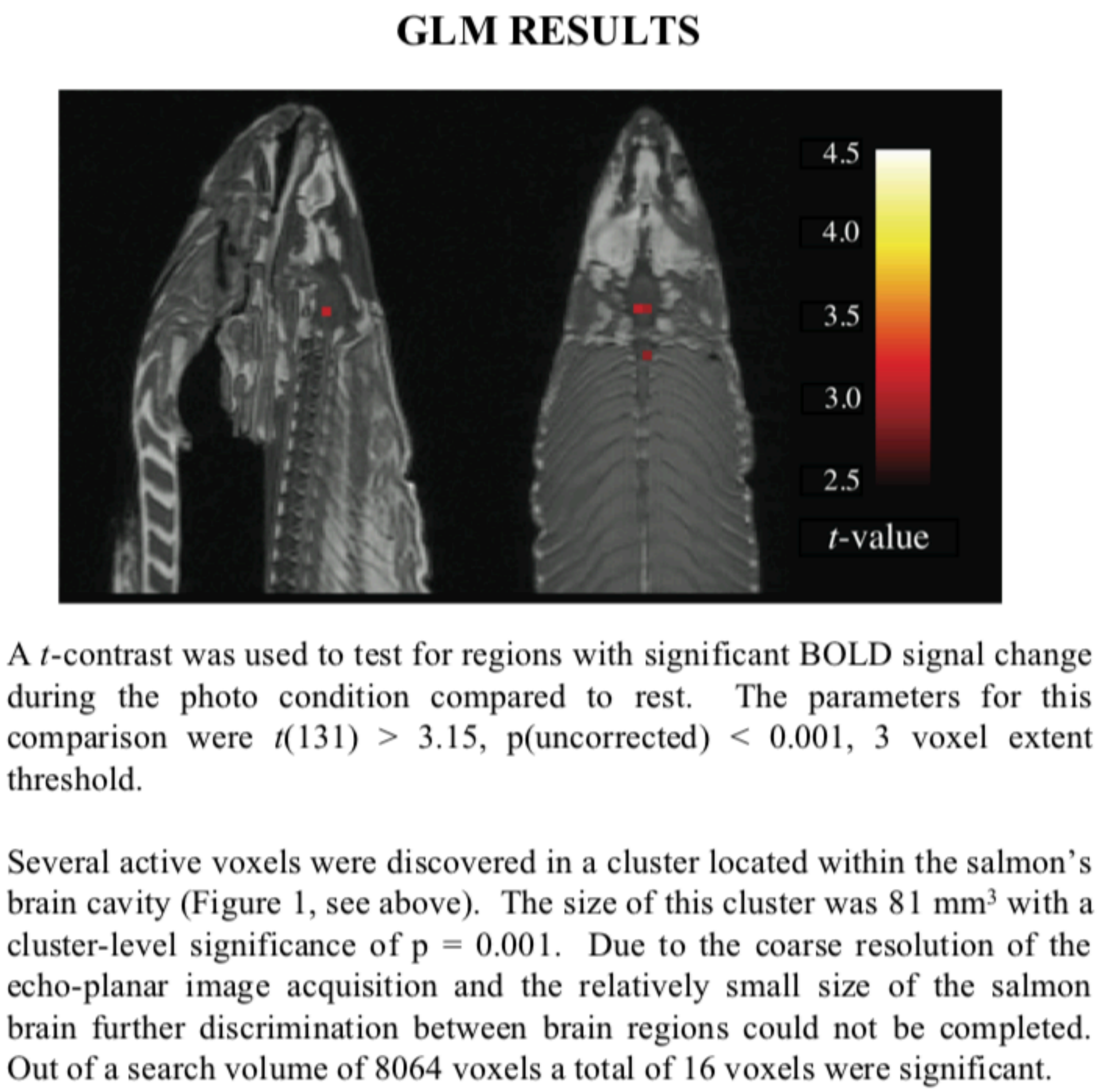
BMC
Medical Genomics

RESEARCH ARTICLE **Open Access**

Genes associated with MUC5AC expression in small airway epithelium of human smokers and non-smokers

Guoqing Wang[1,5*], Zhibo Xu[1,2], Rui Wang[1], Mohammed Al-Hijji[1], Jacqueline Salit[1], Yael Strulovici-Barel[1], Ann E Tilley[1,3], Jason G Mezey[1,4] and Ronald G Crystal[1,3]

Bennet CM, Baird AA, Miller MB, and Wolford GL. (2009). Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction. Journal of Serendipitous and Unexpected Results 1(1):1-5.



GLM RESULTS

A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

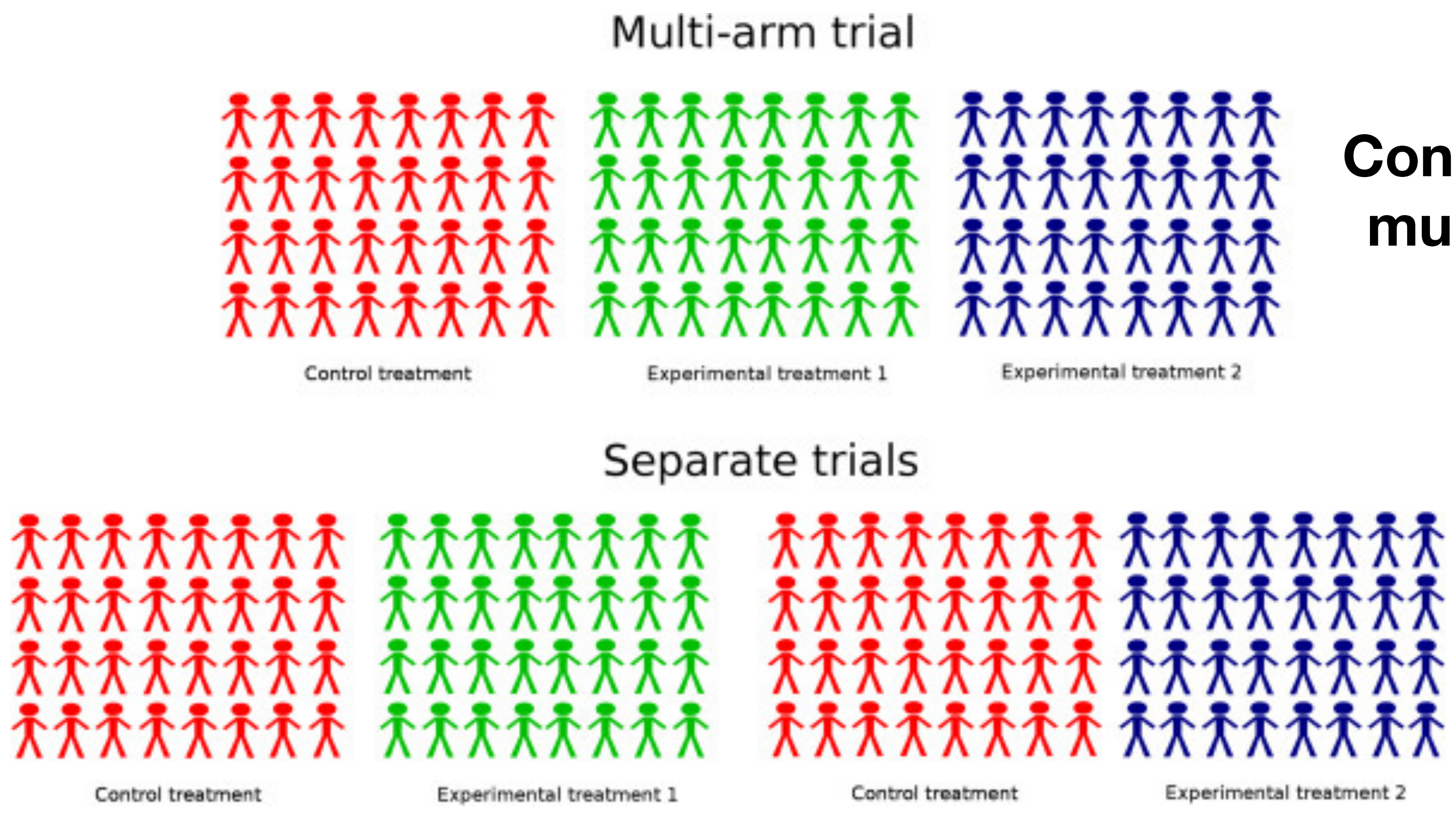
Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

Trials. 2014; 15: 364.
Published online 2014 Sep 17. doi: 10.1186/1745-6215-15-364

PMCID: PMC4177585
PMID: 25230772

Correcting for multiple-testing in multi-arm trials: is it necessary and is it done?

James M S Wason, Lynne Stecher, and Adrian P Mander



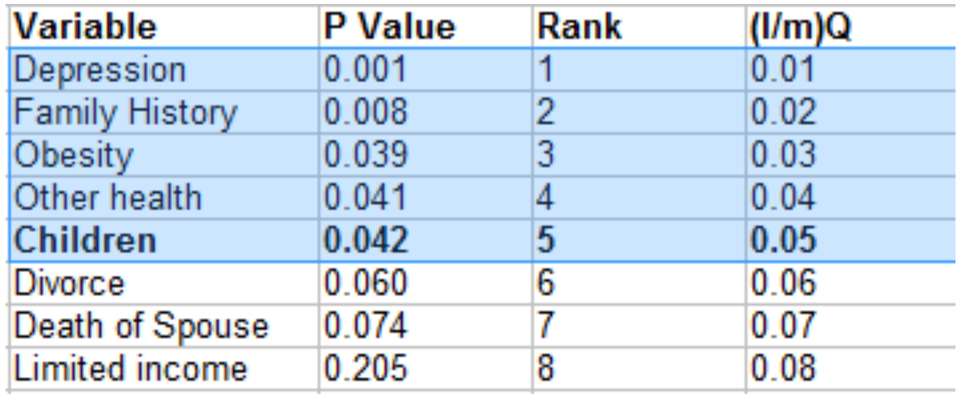
Conclusion: Less than 50 percent of multi-arm clinical trials correct for multiple hypothesis testing.

Bonferroni

- Simplest approach
- Very conservative - you simple divide your p-value by the number of tests
- Bonferroni test compares the **Family-wise error rate** (FWER) -> assuming all the variables have identical distribution in the two groups, what is the probability that you really have some significant differences
- Rarely used in trancriptomics

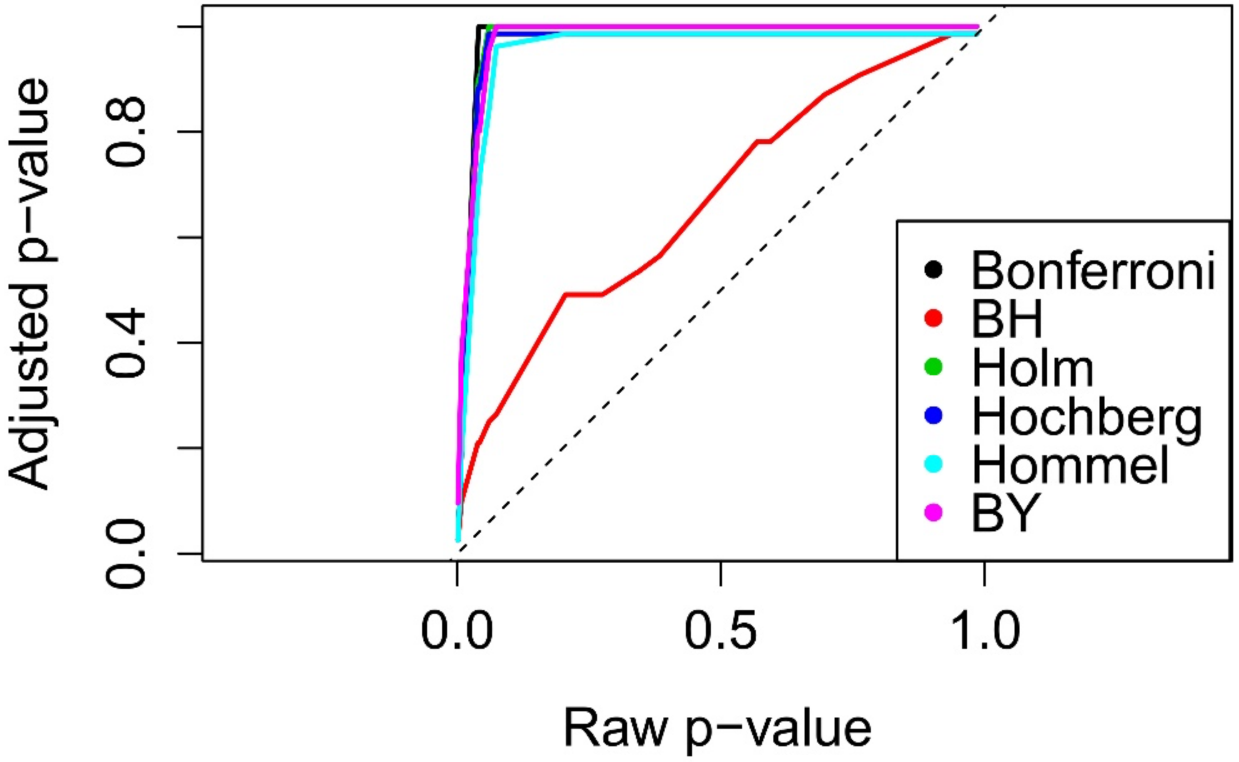
Benjamini-Hochberg

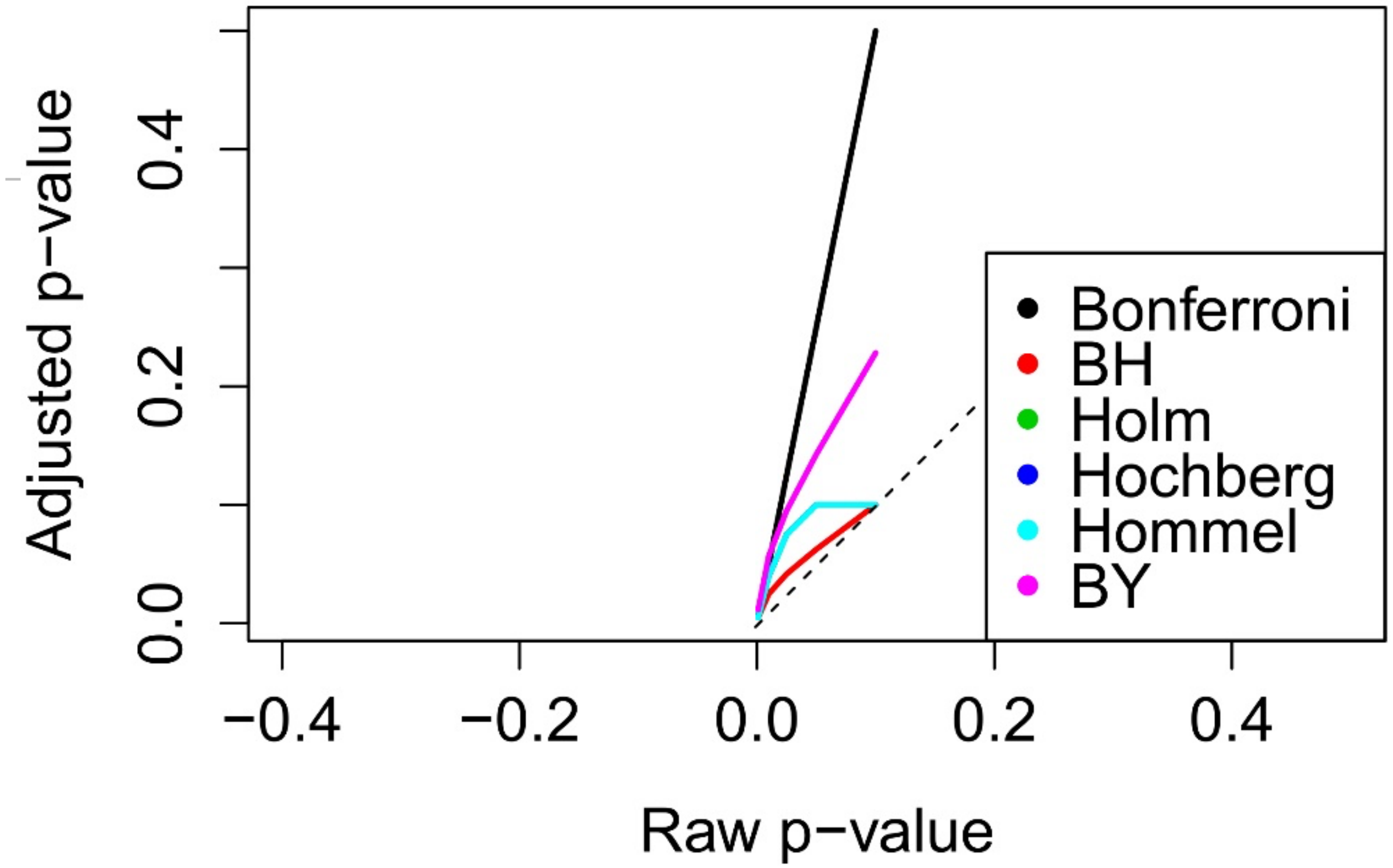
- Benjamini-Hochberg correction controls the **False discovery rate (FDR)** - expected proportion of false positives among the variables for which you claim the existence of a difference.
- For example, if with FDR controlled to 5%, 20 tests are positive, "in average" only 1 of these tests will be a false positive.Consists
  - Put the individual p-values in ascending order.
  - Assign ranks to the p-values.
  - BH critical value (i/m)Q, i = the individual p-value's rank, m = total number of tests, Q = the false discovery rate
  - Highest uncorrected p-value below critical value is cut-off



Variable	P Value	Rank	(I/m)Q
Depression	0.001	1	0.01
Family History	0.008	2	0.02
Obesity	0.039	3	0.03
Other health	0.041	4	0.04
Children	**0.042**	**5**	**0.05**
Divorce	0.060	6	0.06
Death of Spouse	0.074	7	0.07
Limited income	0.205	8	0.08

https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2015/10
bh2.pngto.datasciencecentral.com/wp-content/uploads/2015/10/bh2.png

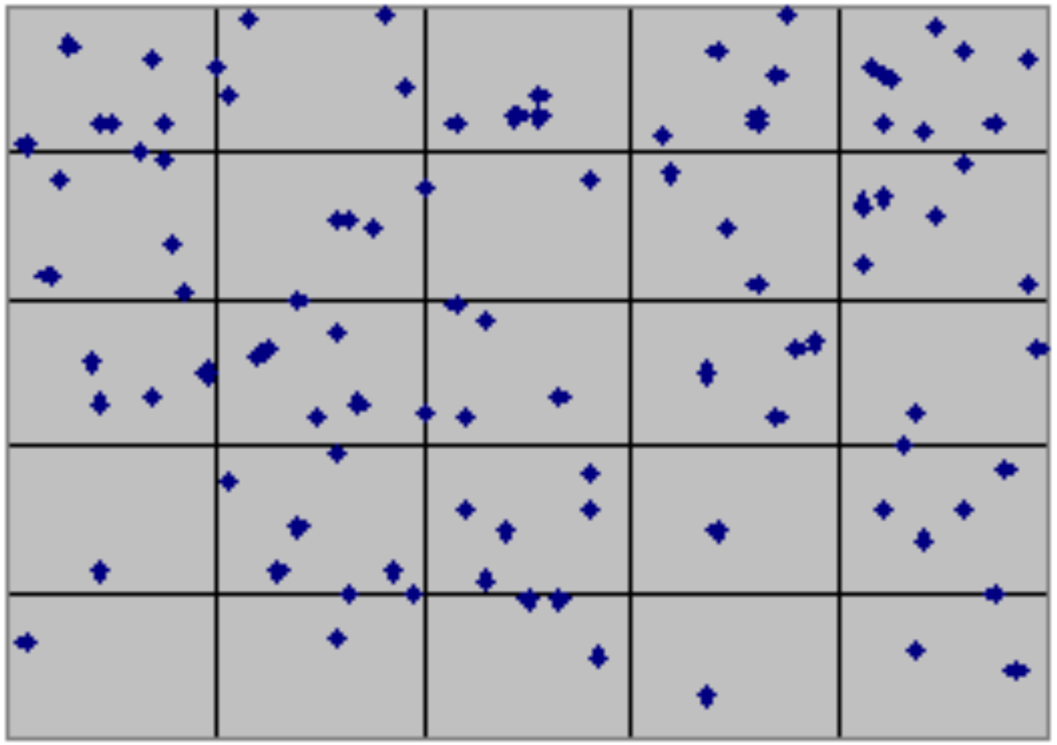


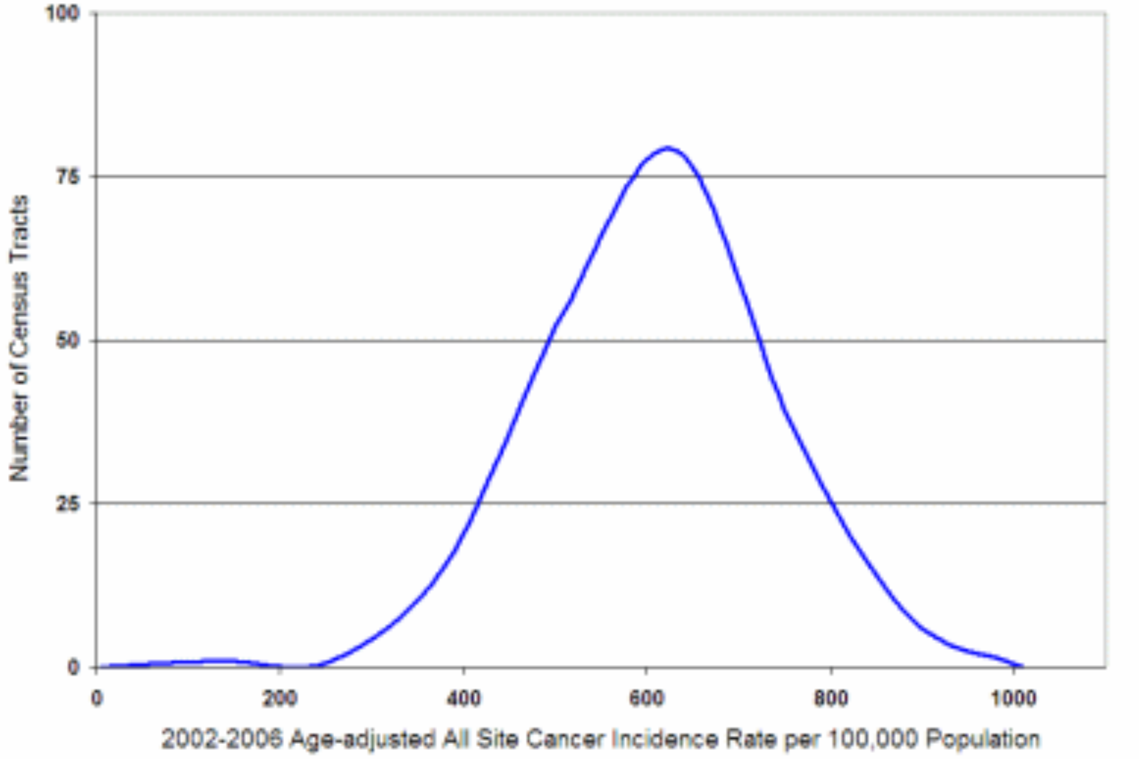


Regardless of method chosen - there is always a trade-off.

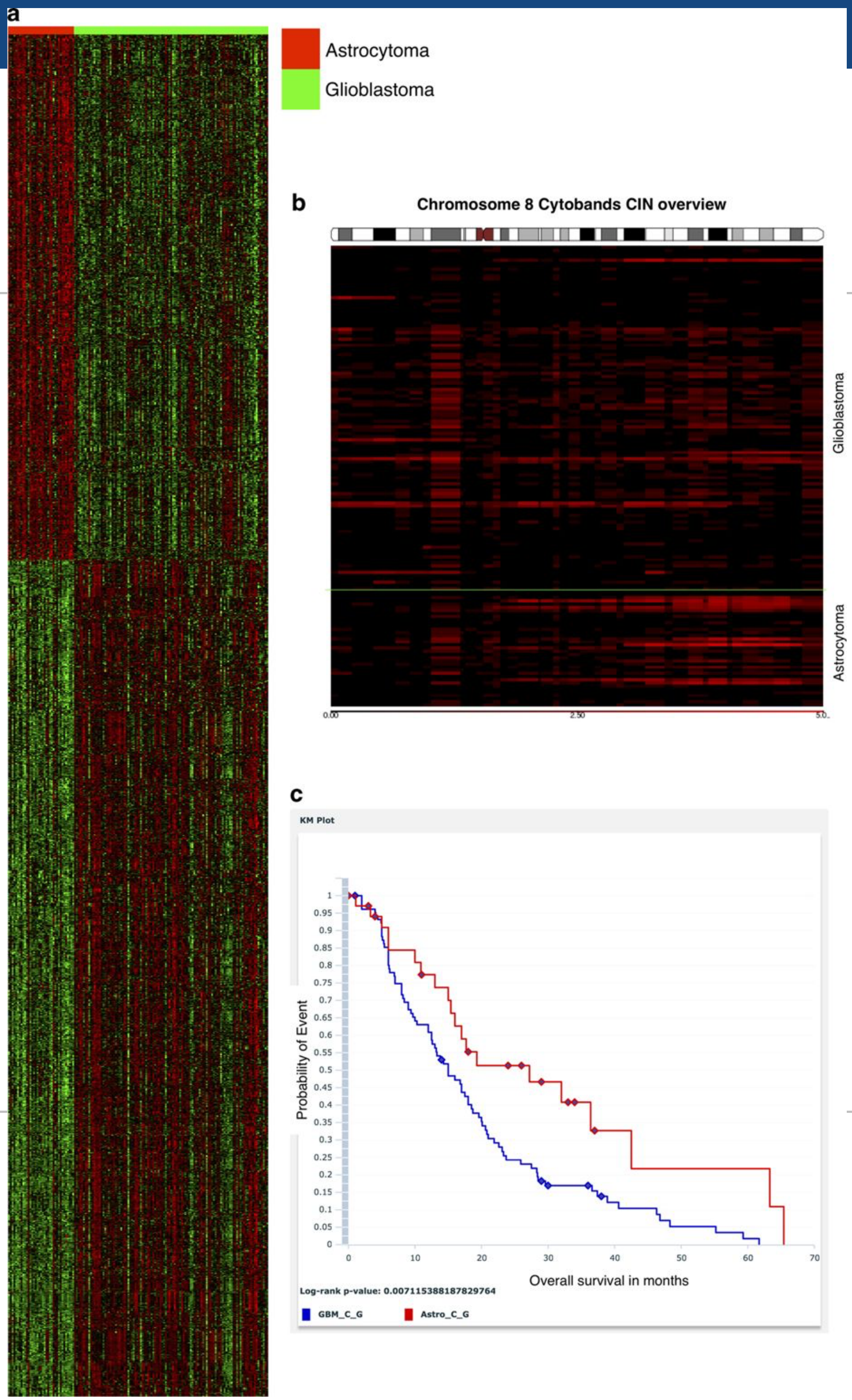
Increasing the dimensionality can reduce your power: If you are looking for a small number of genes, you simply *cannot* see it with 30,000 comparisons.

- How do you detect meaningful clusters?

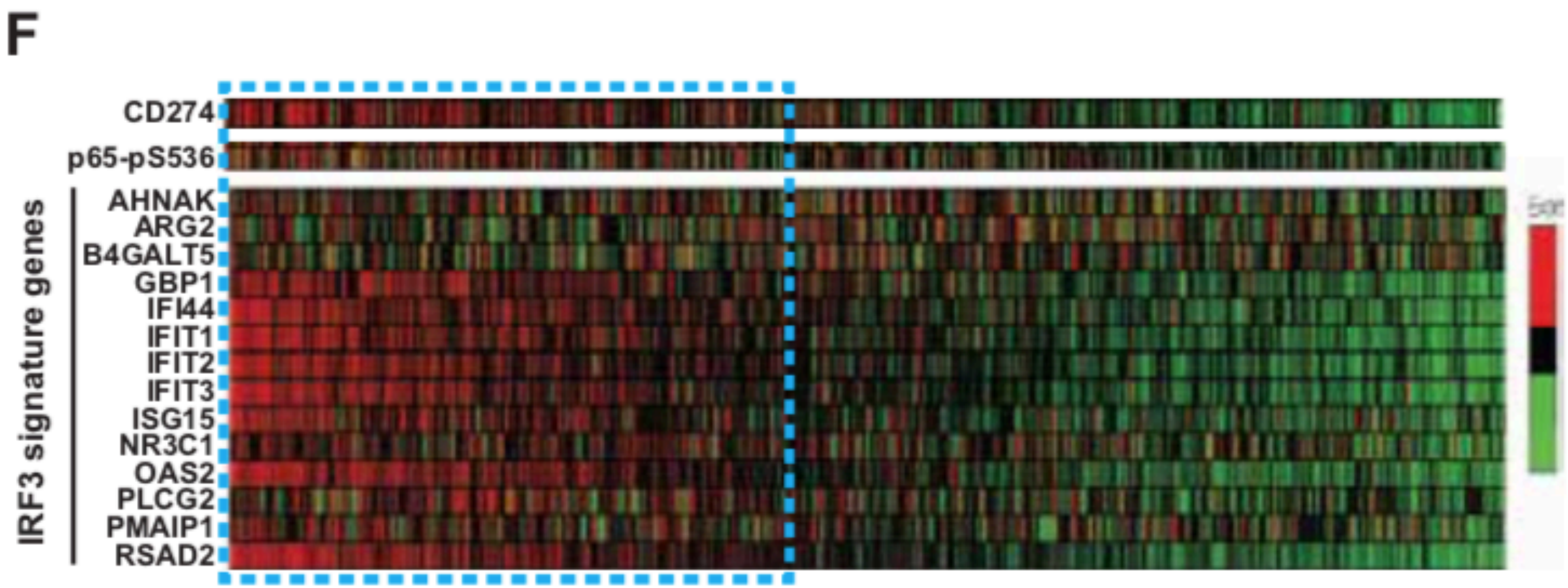




GOOD HEATMAP vs BAD HEATMAPS



GOOD HEATMAP vs BADHEATMAP



Thinking beyond the t-test

- Dose-response and time-course data should be analyzed in different ways
- Usually you should avoid ANOVA
- More powerful approach is to look for linear trends in the data

Thinking beyond the t-test

- Gene-by-gene test of statistical significance is not always the best way to analyze data
- A list of genes DE genes is not, in and of itself, that informative.
- It’s also not reproducible

GEMS (Gene Expression Metasignatures), a Web Resource for Querying Meta-analysis of Expression Microarray Datasets: 17β-Estradiol in MCF-7 Cells

Scott A. Ochsner, David L. Steffen, Susan G. Hilsenbeck, Edward S. Chen, Christopher Watkins and Neil J. McKenna

DOI: 10.1158/0008-5472.CAN-08-3492 Published January 2009

Table 2.

Genes with a combined *q* value of <0.05 identified by the meta-analysis

- NOTE: Genes are binned according to a number of different individual dataset FC criteria, ranging from no FC criteria to FC of ≥2 in all underlying datasets. Full gene lis provided in Supplementary Table T2. NA, not applicable.

# of independent datasets with FC of >2.0	Meta-analysis	
	Early	Late
—	2,313	4,144
1	526	1,213
2	140	516
3	67	321
4	20	118
5	6	29
6	NA	5
7	NA	0

Thinking beyond the t-test

- List of differential genes leave a lot of information on the table - for example:
  - If you have 20 genes with a fold change of + 1.5 on the DNA damage repair pathway, but only 7 of them had an FDR < .05 - do the other 13 genes have something important to tell you? Why or why not?
  - Gene expression data has a lot of information which can be exploited with techniques from machine learning

Machine-learning vs statistics

- Classical statistics asks - is the difference in these two groups there by chance?
- Machine-learning asks - what is the pattern in this data telling me? Can include
    - **Class-discovery** - Can I use transcriptomics to **classify** cancer vs non-cancer tissue?
    - **Network approaches** - what is the **correlation** in genes telling me about **regulation**?

Group Projects

- Use GEO2R to generate a list of differentially expressed genes
- Correct for multiple hypothesis testing Bonferonni and BH, compare the list of genes between uncorrected, Bonferonni, and BH
- Do you think this list of genes is worth further study? Be prepared to discuss.
- Think about whether you want to continue with this data set!

Project

- Analyze a data set for differentially expressed genes, look for pathway differences, and explore possible regulatory mechanisms.
- Synthesize the results and compare to the published conclusions

Remember. . .

“Statistics means never having to say you’re certain!”