A Flexible Probe Level Approach to Improving the Quality and Relevance of Affymetrix Microarray Data

Chris Harbron

Discovery Statistics

AstraZeneca

Non-Clinical Statistics Conference,

Leuven, September 2008

AstraZeneca

BxMed Statistics

Microarrays

- Enable measurements of the levels of gene expression of many thousands of genes simultaneously
- Provides an detailed description of the biology at a molecular level

AstraZeneca

Uses Of Gene Expression In The Pharmaceutical Industry

Drug Discovery | Drug Development | Marketed Products

- Identification of drug targets
- Understanding Modes Of Action
- Personalised Medicine
- Understanding Drug Safety
- Biomarkers For Early Assessment Of Efficacy
- Support For Existing & Identifying New Indications

AstraZeneca

Statistics

Microarrays

- Best thing about microarrays:
- Analyse 1000s of genes simultaneously
- Won't miss anything

- Worst thing about microarrays:
- Analyse 1000s of genes simultaneously
- Can end up missing the interesting results in a mass of false positives

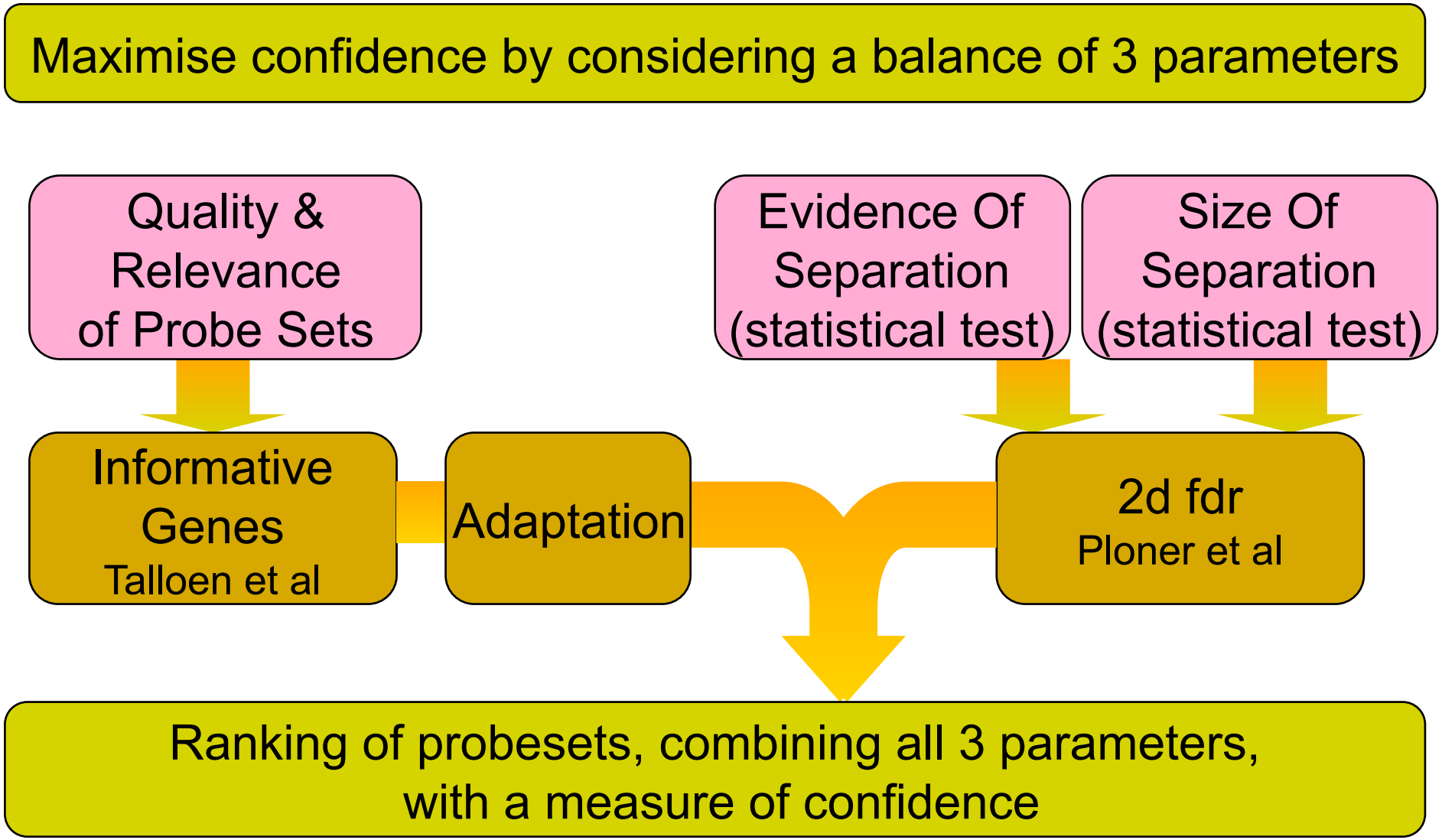
AstraZeneca

BxMed Statistics

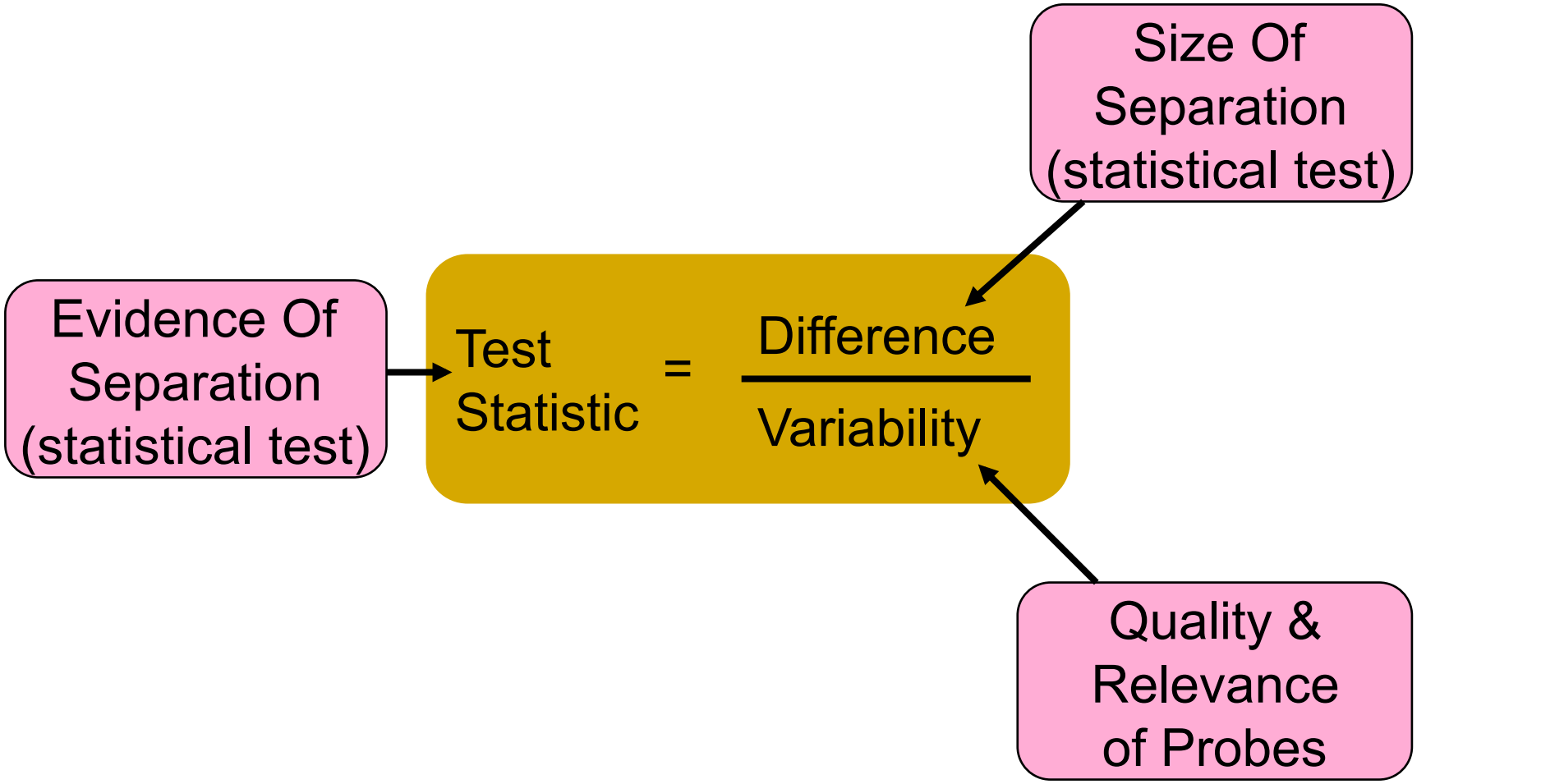
Reducing False Positives : Filtering

- Often people try and reduce the false positives issue by pre-filtering the genes before analysis
  - Present / Absent calls, Variability, Minimum / average expression level
- And by subsequently selecting arbitrary cut-offs post-analysis
  - p-value & fold change
- Lots of arbitrary choices
- May miss things – some properties may not directly translate across platforms and species
- Present / Absent calls based on differences between PM & MM
  - Assumes no signal in MM which we know to be untrue.
  - Also affected by GC content of middle base
  - Arbitrary cut-off from significance test

3d fdr


Maximise confidence by considering a balance of 3 parameters
Quality & Relevance of Probe Sets
Evidence Of Separation (statistical test)
Size Of Separation (statistical test)
Informative Genes
Talloen et al
Adaptation
2d fdr
Ploner et al
Ranking of probesets, combining all 3 parameters, with a measure of confidence

3 Correlated Criteria



AstraZeneca

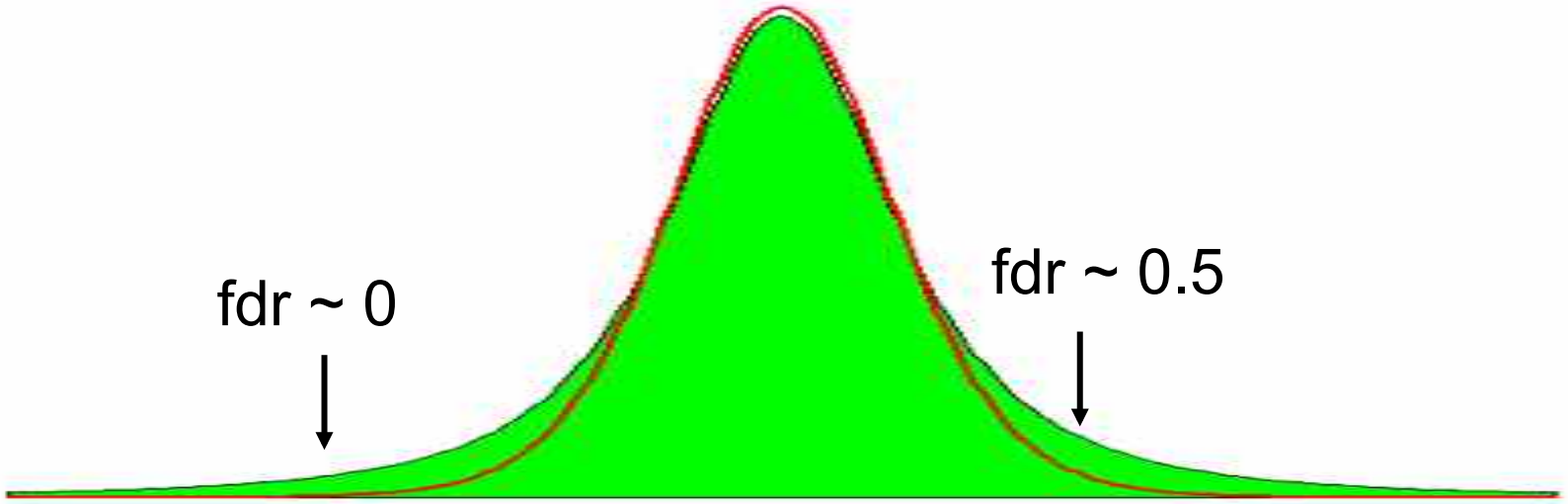
BxMed Statistics

Assessing False Positives
Local False Discovery Rate (fdr)

Expected proportion of genes with observed statistic Z=z which are false positives

$$fdr(z) = \pi_0 \frac{f_0(z)}{f(z)}$$

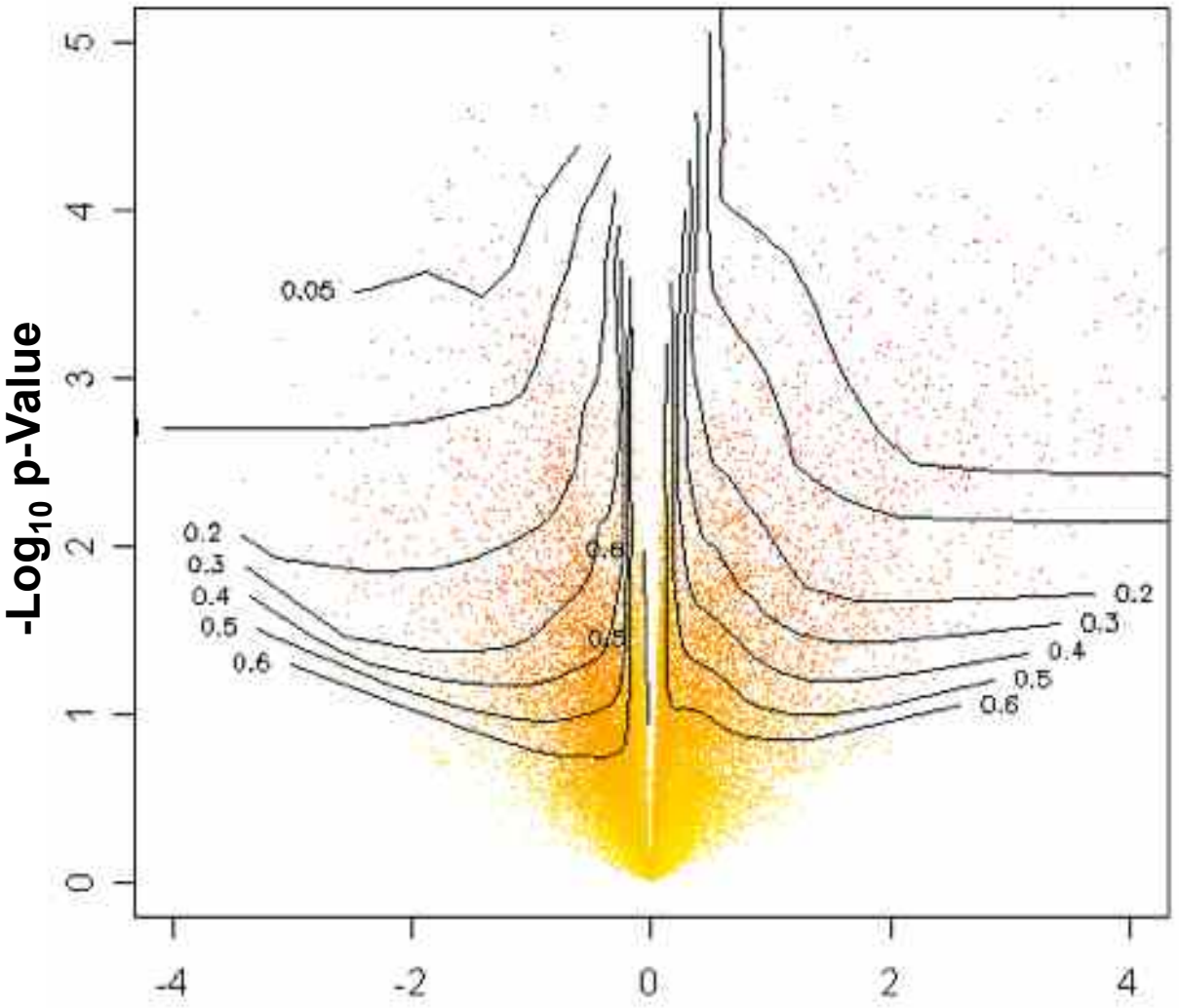
= Proportion of truly non-DE genes x (Density for non-DE genes / Observed Density)



Distinct from, but related to, global FDR

2d fdr

Ploner et al Bioinformatics 2006



Log Fold Change – Difference Between Groups

Extends concept of fdr to joint distribution of two statistics

$$\pi_0 \frac{f_0(z_1, z_2)}{f(z_1, z_2)}$$

Calculates likelihood of being of each probeset being a false positive based on a combination of significance and difference

Informative / Non-Informative Calls & The PCPV Statistic

I/NI Calls - Talloen et al, Bioinformatics 2007

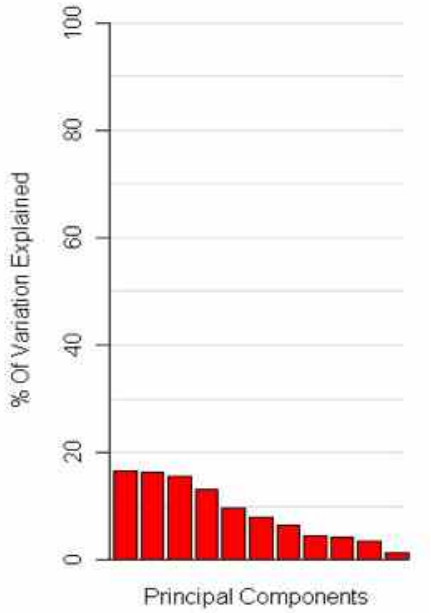
- –Makes use of the multiple probes in an Affymetrix probeset
- –Bayesian estimate of a signal to noise ratio
- –If a probeset is informative, then the same pattern should be seen within all the probes within the probeset
- –Binary classification

PCPV statistic uses similar concept

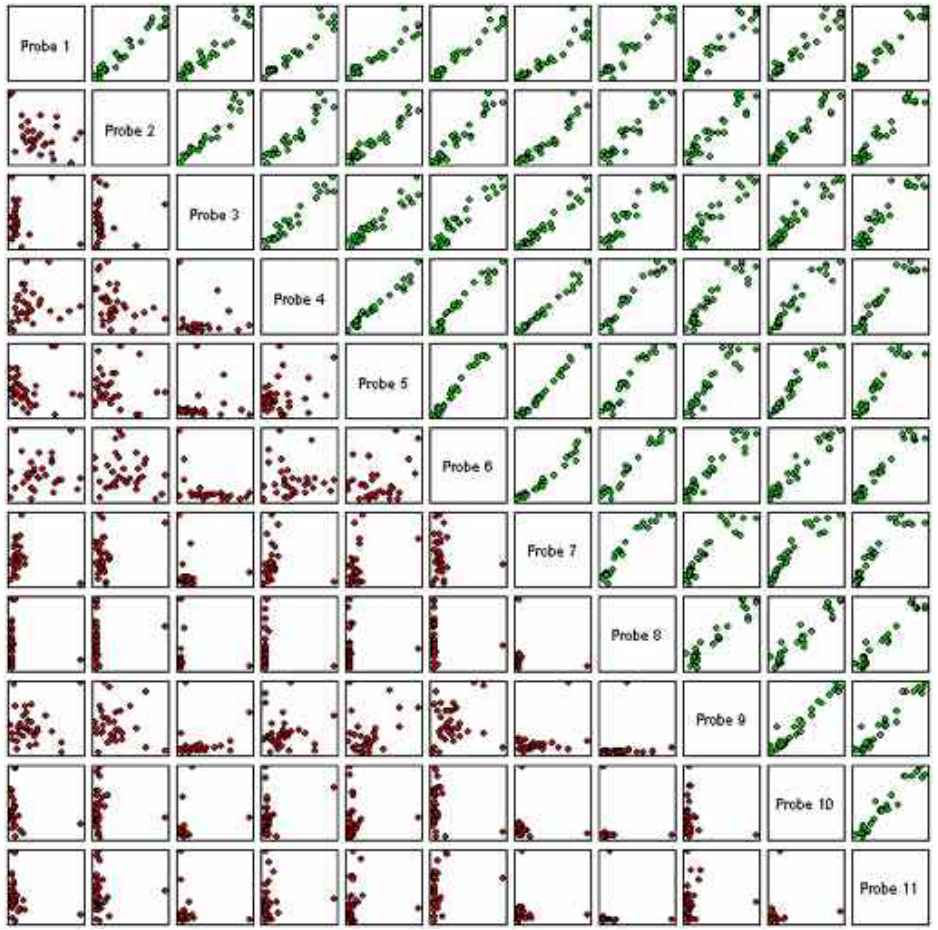
- –Percentage of total variation in probe intensity explained in the first principal component
- –Continuous measure of information

Informative / Non-Informative Calls Relationship To PCPV

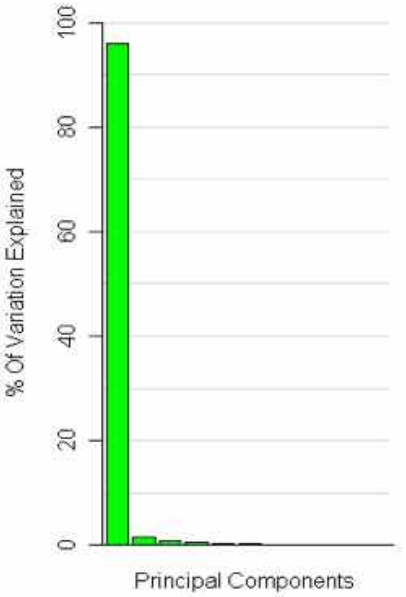
Non-Informative Probe Set



Low PCPV Statistic



Informative Probe Set



High PCPV Statistic

AstraZeneca

Statistics

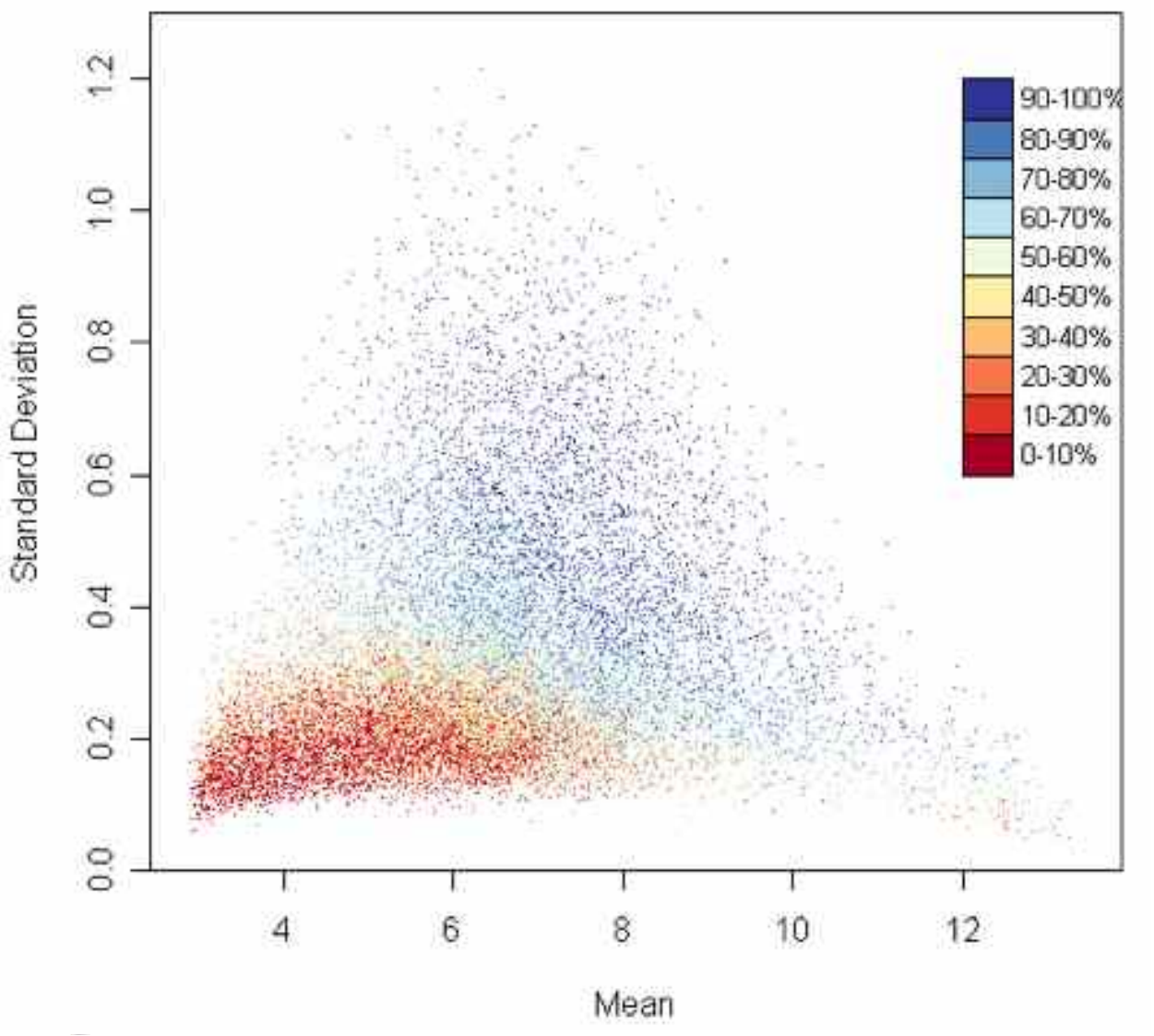
Informative / Non-Informative Calls & The PCPV Statistic

- If a probeset had a low PCPV statistic, i.e. its constituent probes are non-correlated, then either:
  - It's just measuring noise, i.e. there's no differences between the samples
    - Low levels of expression dominated by noise
    - No variation in expression between samples
  - It's an unreliable set of probes
- Either way, it's not very interesting
- Doesn't necessarily follow that the gene is interesting in the sense of changing with what we are interested in, e.g. treatment

AstraZeneca

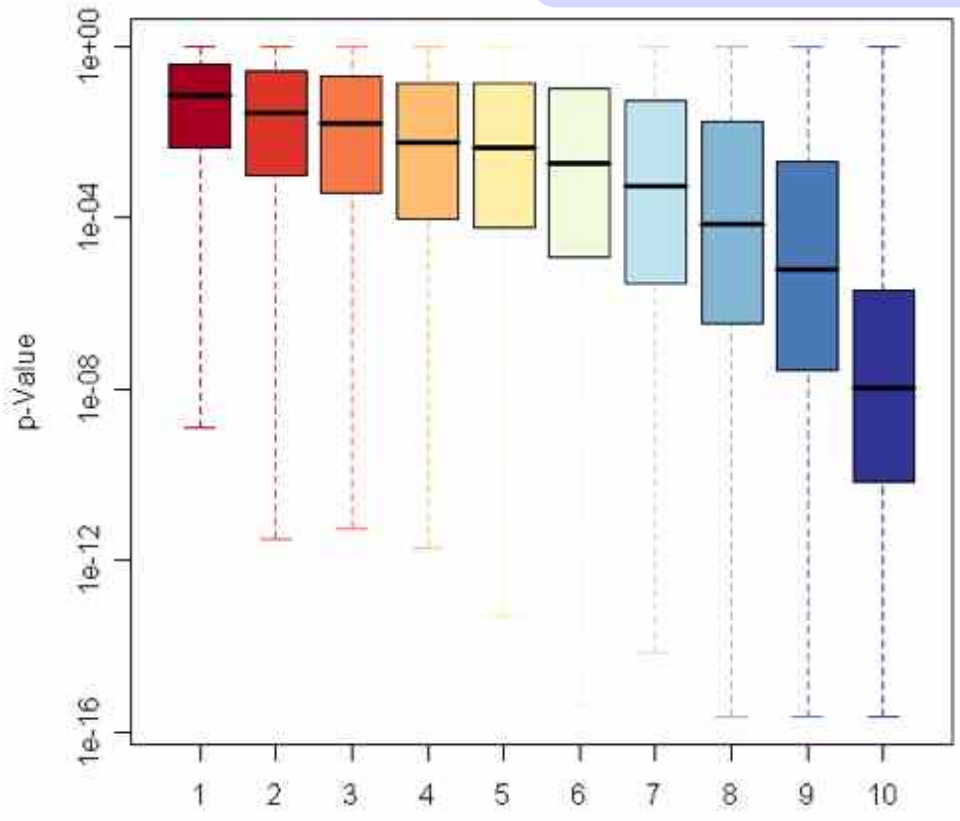
BioMed Statistics

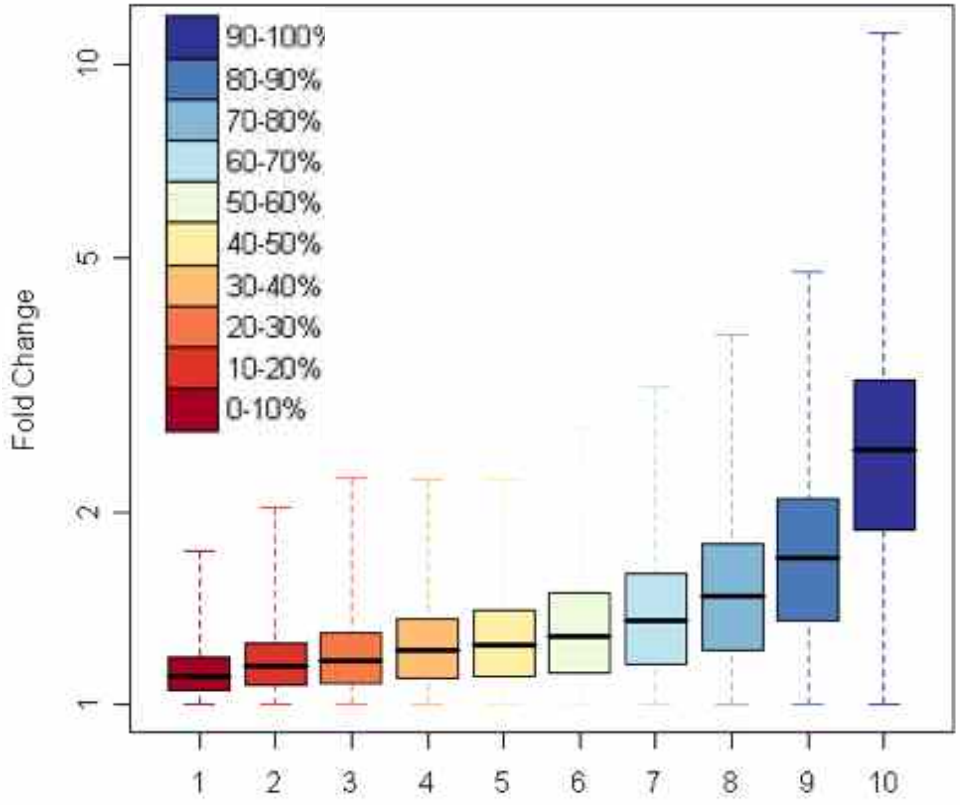
Higher PCPV Statistics Have More Interesting Profiles



Probes With Higher PCPV Statistics Tend To Be More Interesting

But not exclusively so

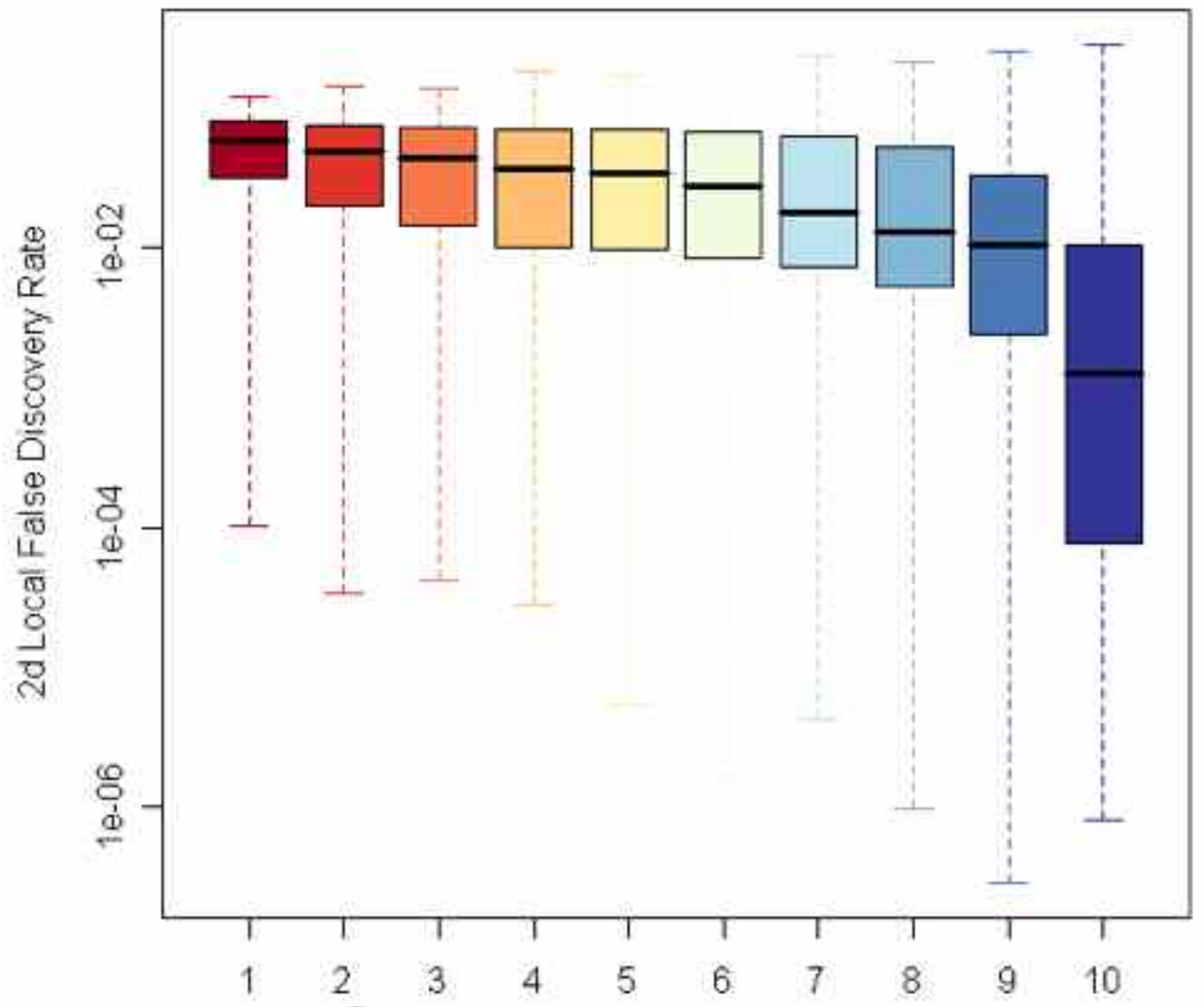




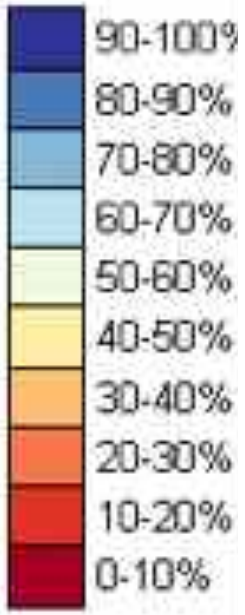
AstraZeneca

Statistics

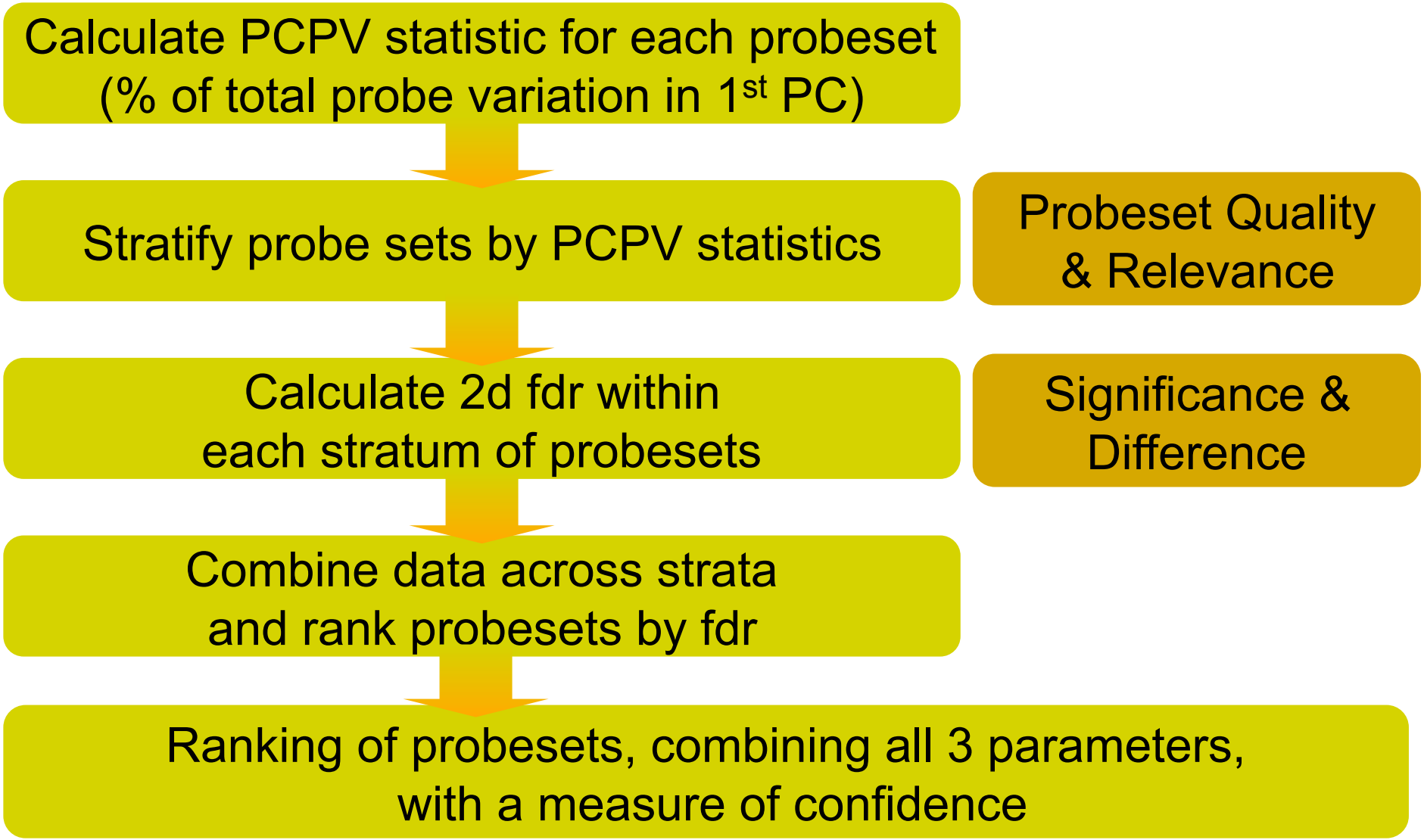
Probes With Higher PCPV Statistics Tend To Be More Interesting



But not exclusively so



3d fdr
Stratified PCPV



3d fdr
Stratified PCPV

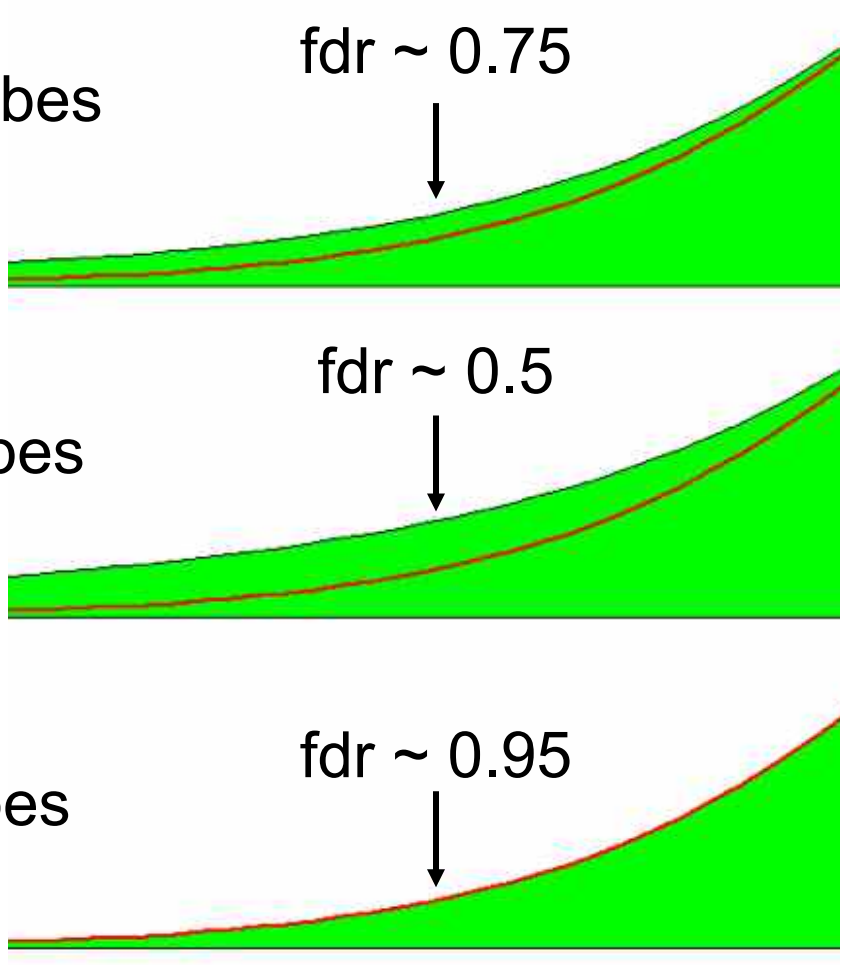
Entire Set Of Probes

=

High Quality Probes

+

Low Quality Probes



Expected distribution of non-DE genes

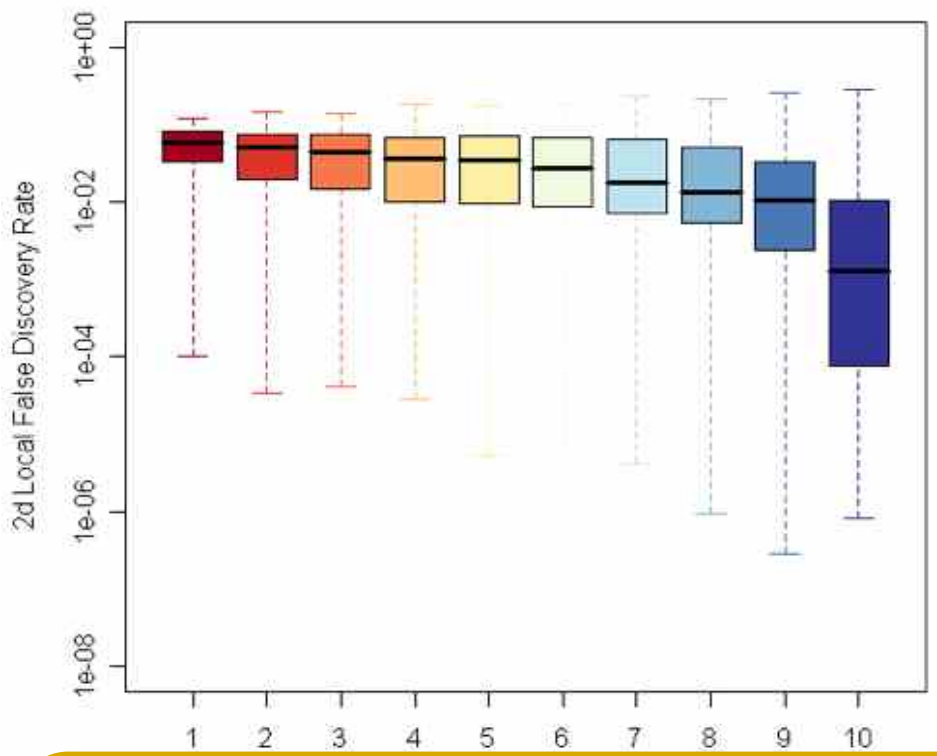
Observed distribution

AstraZeneca

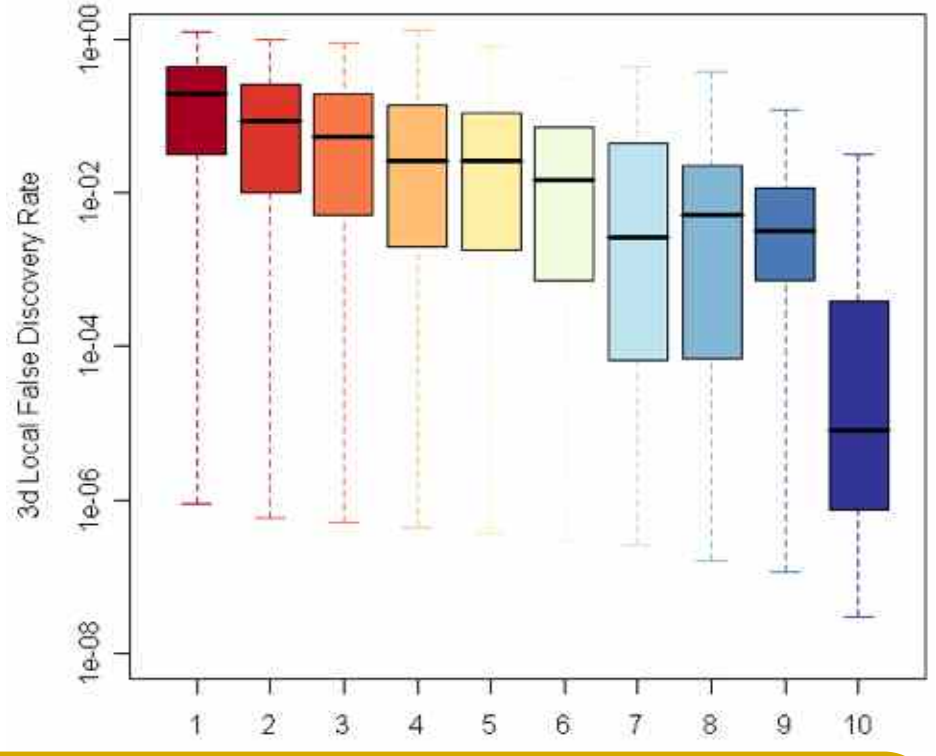
BioMed Statistics

3d fdr Results

2d fdr

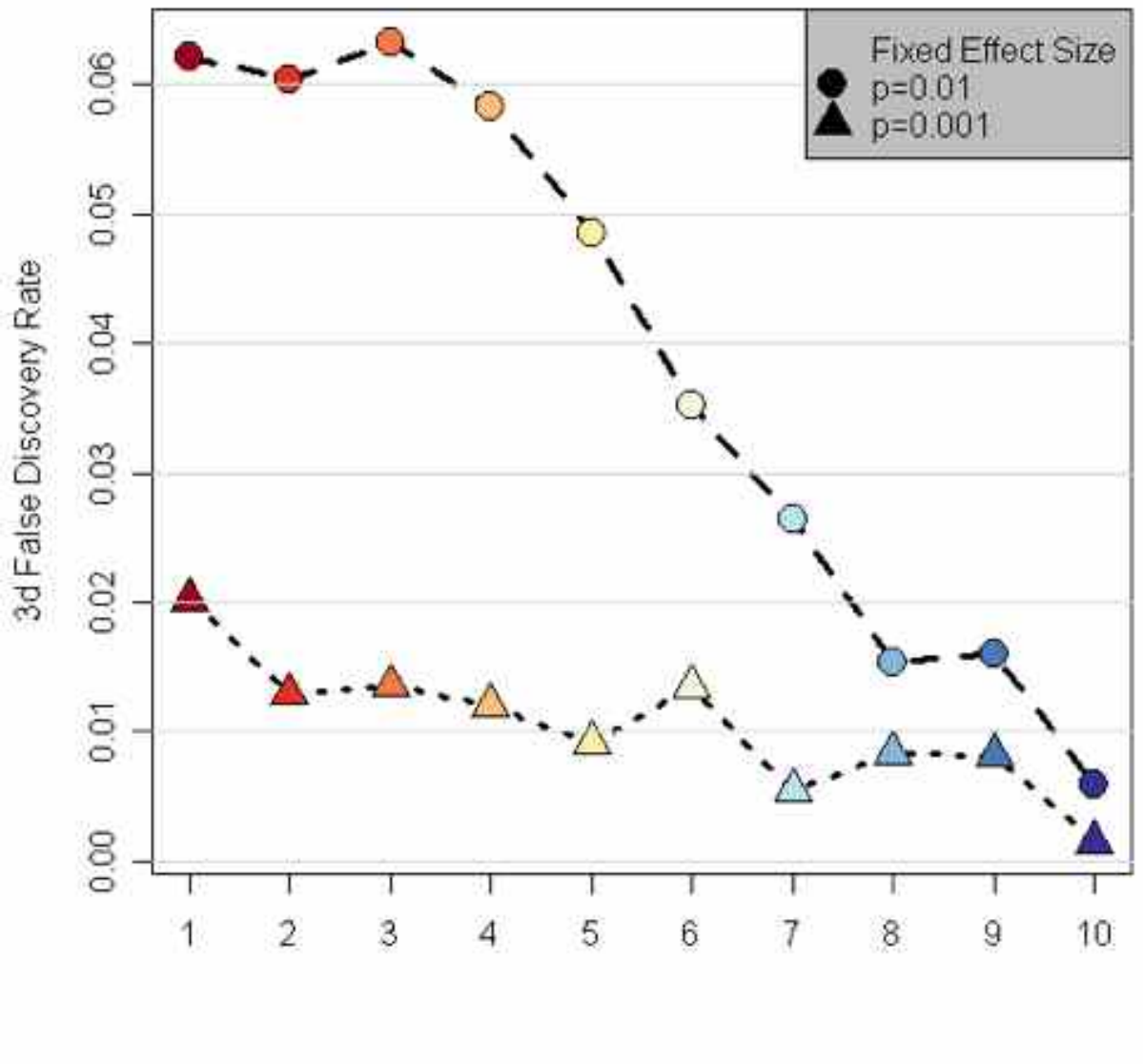


3d fdr

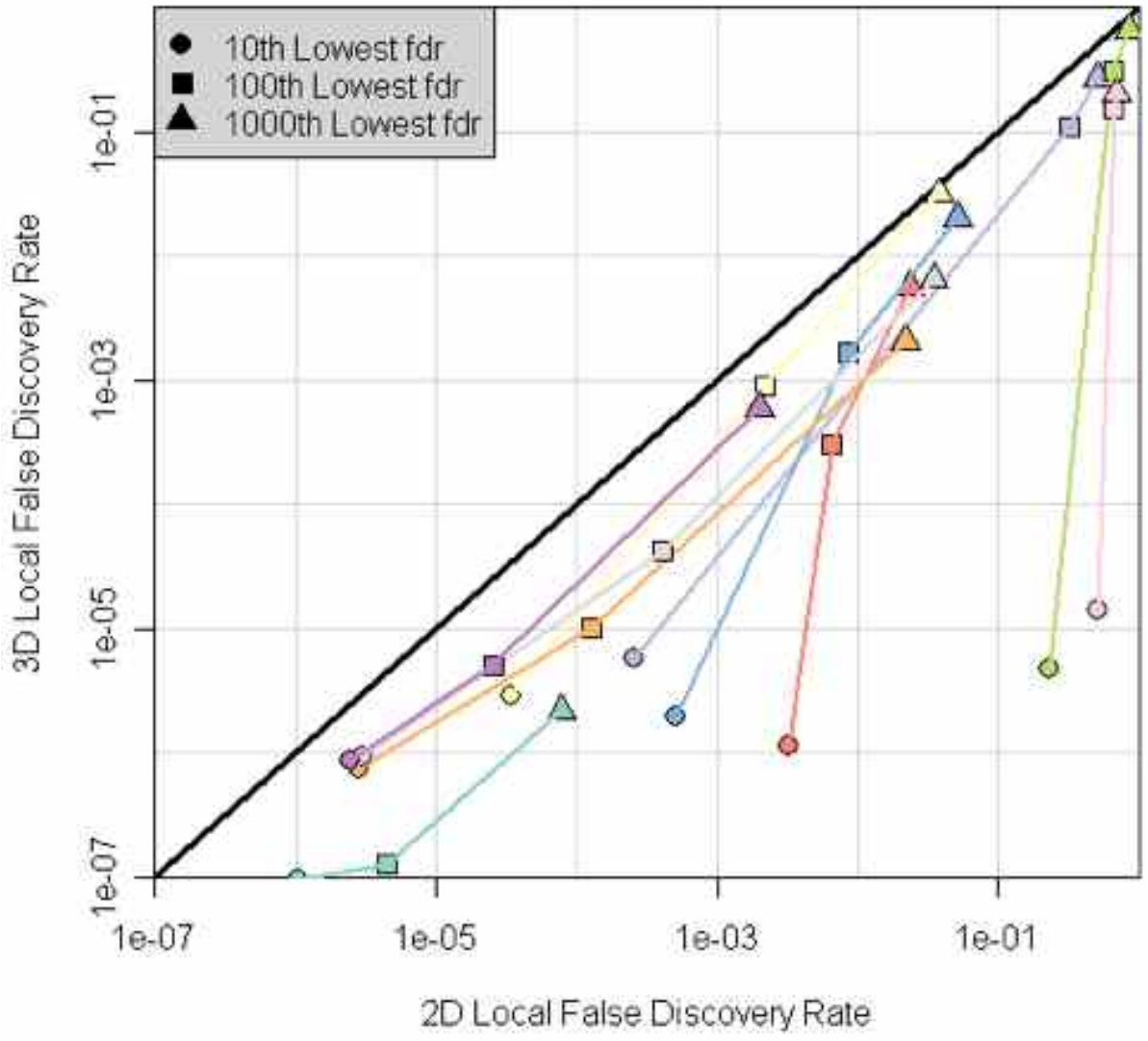


Increase in confidence (lower fdr) for high relevance probesets
Decrease in confidence (higher fdr) for lower relevance probesets
High confidence probesets (low fdr) enriched, but not exclusively, from higher relevance probesets

3d fdr Results



Applicable Over Different DataSets



Selected 10 datasets with available covariate information at random from GEO

Consistently able to detect genes with more confidence using 3d fdr approach

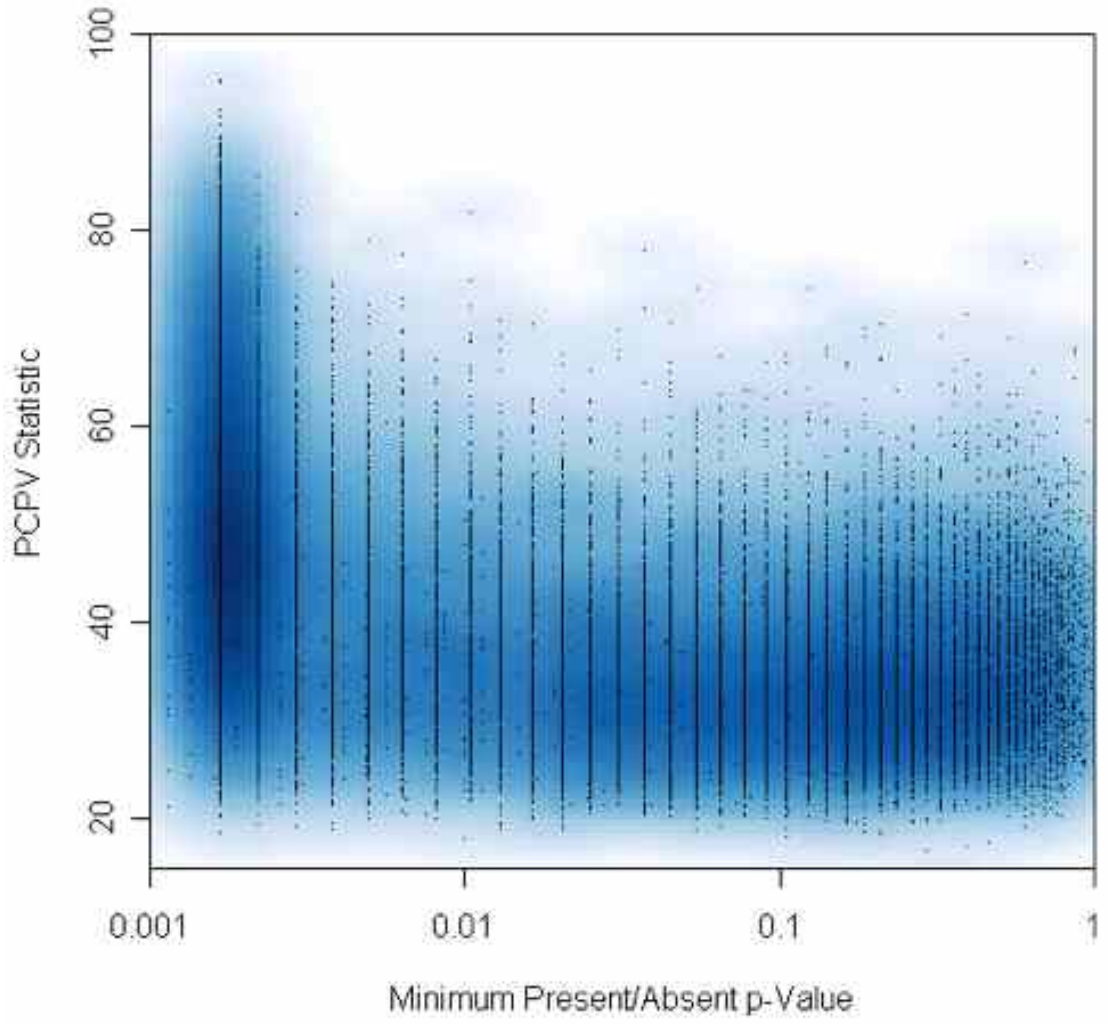
Statistics

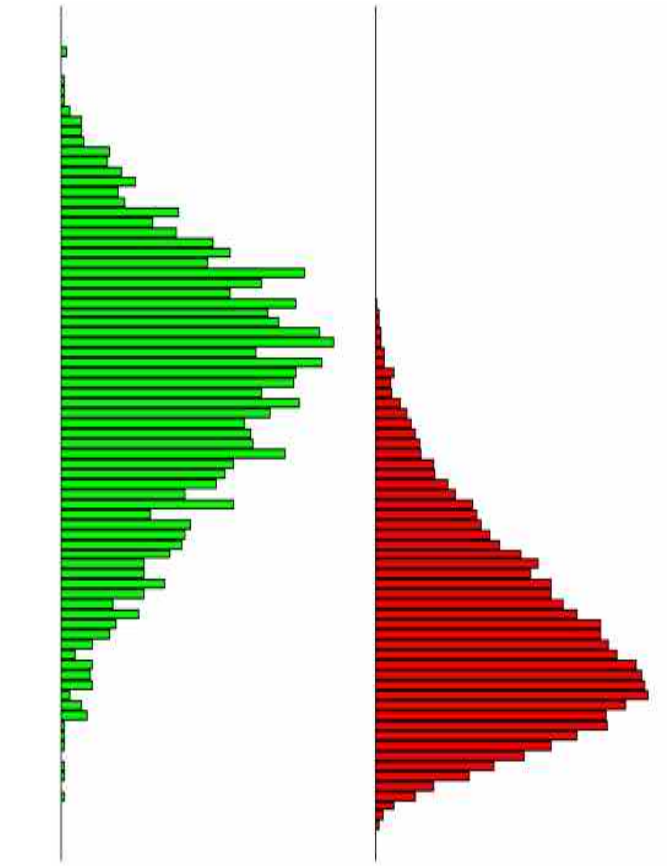
Summary

- Single ordering of genes combining different properties on a rational basis
- A gene which is outstanding on one parameter, but not others could still be selected for further investigation
  - Will get missed with standard “and” selection
- Removes arbitrary filtering decisions
- Tried a robust PCA (as RMA fitting is a robust method – median polish)
  - Little change
- Shown for a 2-group t-test – easily extended to ANOVA or regression situation or any other test statistic

Back Up Slides

Relationship Of PCPV to Other Quality Filters





Informative ProbeSets

Non-Informative ProbeSets