

Vienna – March 19-22, 2003



# Two methods to separate the score distributions of induced and non-induced genes

Stefanie Scheid



MAX-PLANCK-GESELLSCHAFT

Max Planck Institute for Molecular Genetics  
Computational Diagnostics  
Innestrasse 63-73, D-14195 Berlin, Germany  
[stefanie.scheid@molgen.mpg.de](mailto:stefanie.scheid@molgen.mpg.de)

# Microarray Experiment

Two classes of disease status, set of patients for each class.

Question: Are there **differences in gene expression** between classes?

Genes showing differences are called differentially expressed, up/downregulated, induced.

## How do we detect induced genes?

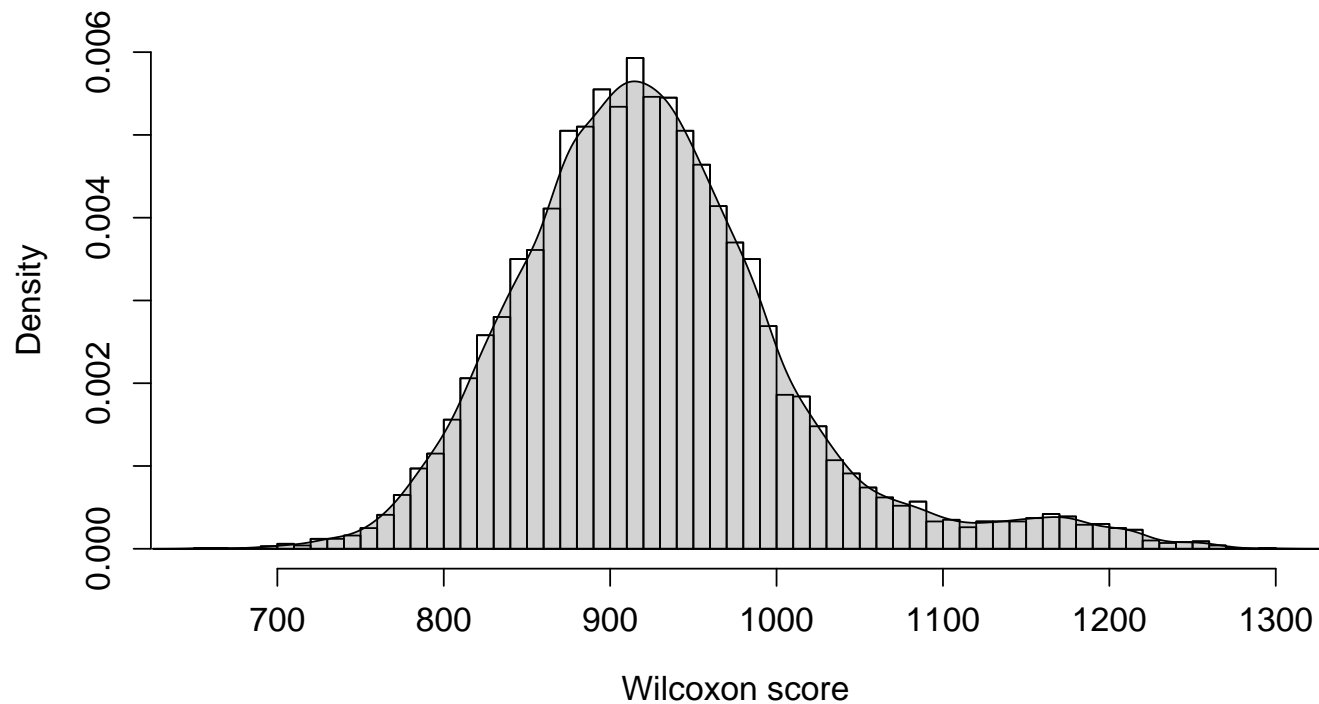
Statistically: Assign score for differential gene expression and test for significance.

Simplify: Test for difference in mean gene expression between classes.

Possible scores: t-test statistic, Wilcoxon ranksum score.

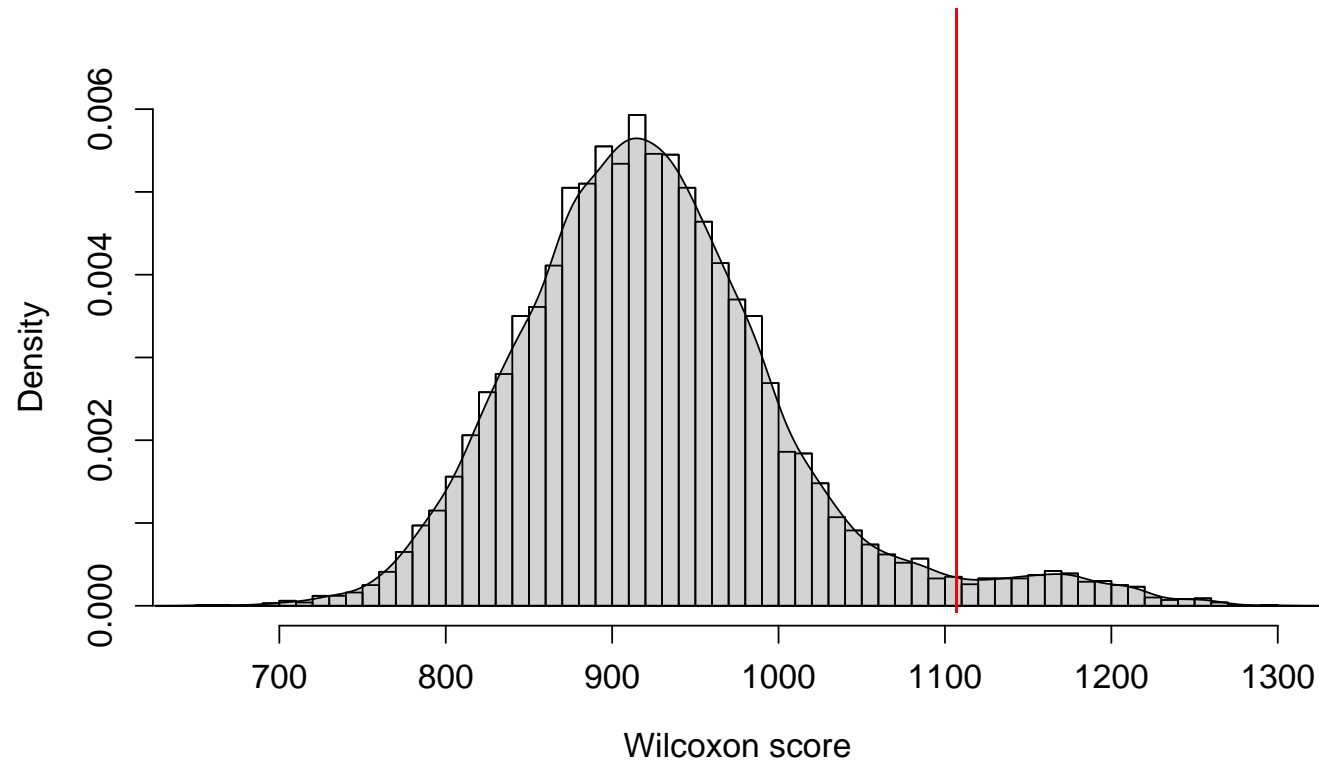
# Score distribution

Overall score distribution is **mixture** of score distributions of induced and non-induced genes:



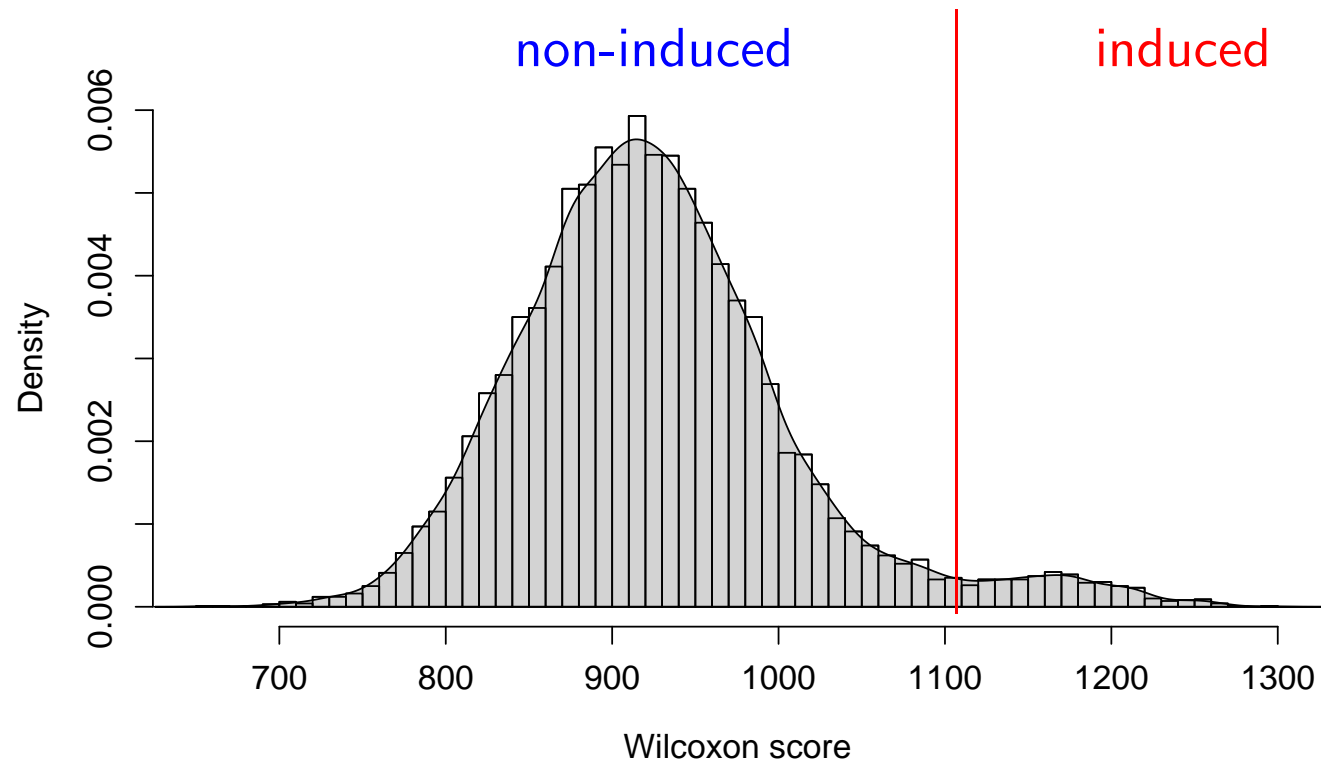
# Score distribution

Overall score distribution is **mixture** of score distributions of induced and non-induced genes:



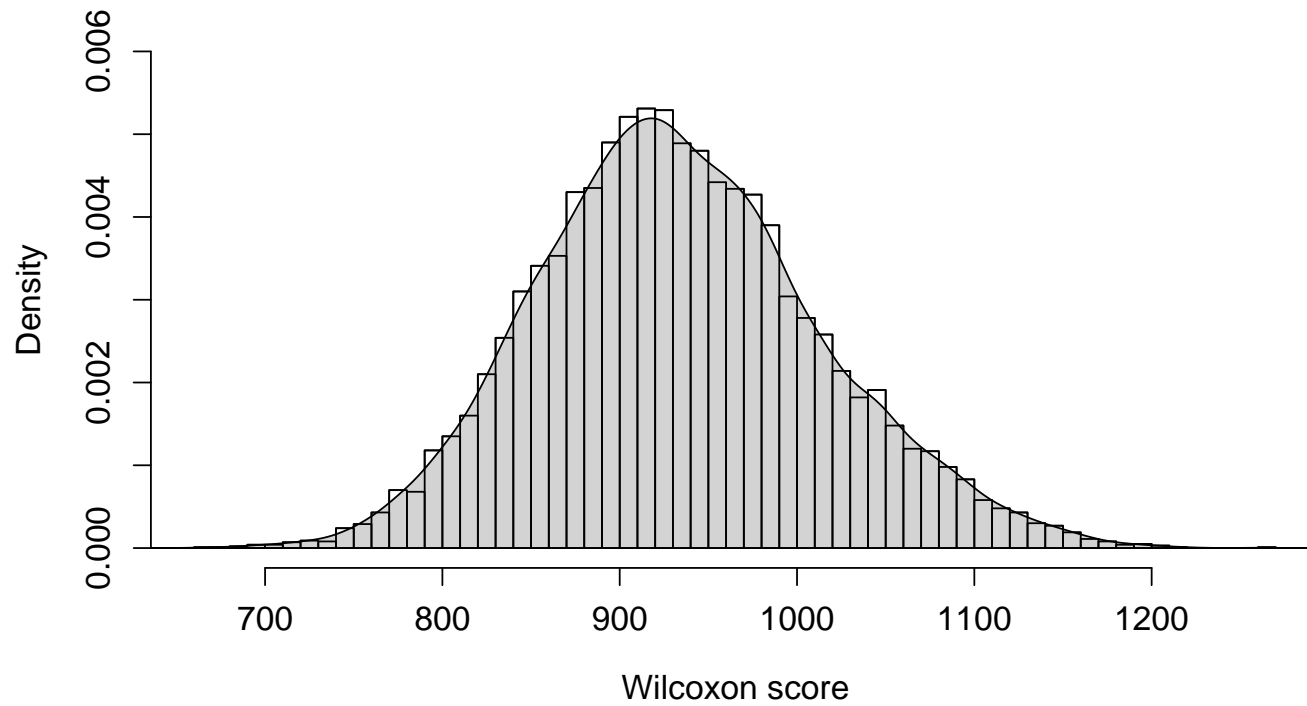
# Score distribution

Overall score distribution is **mixture** of score distributions of induced and non-induced genes:



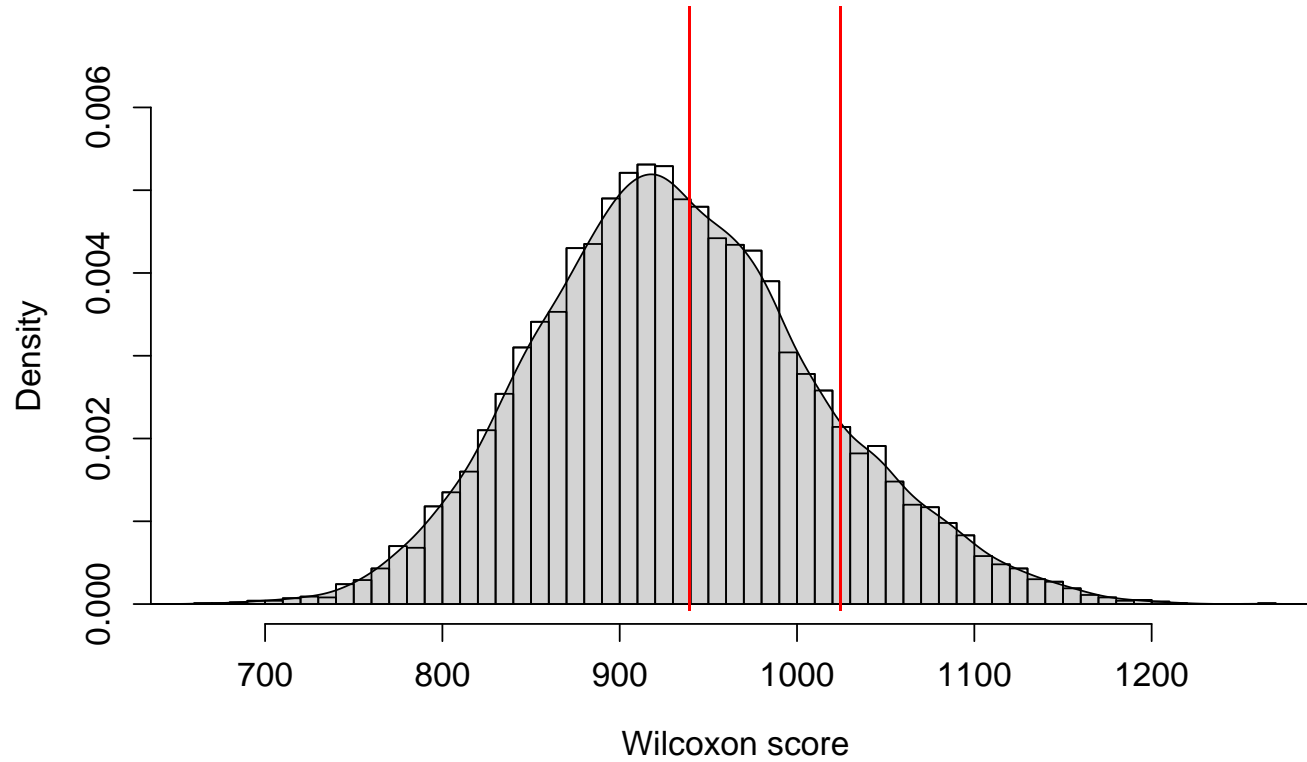
# Score distribution with twilight zone

What if score distributions overlap?



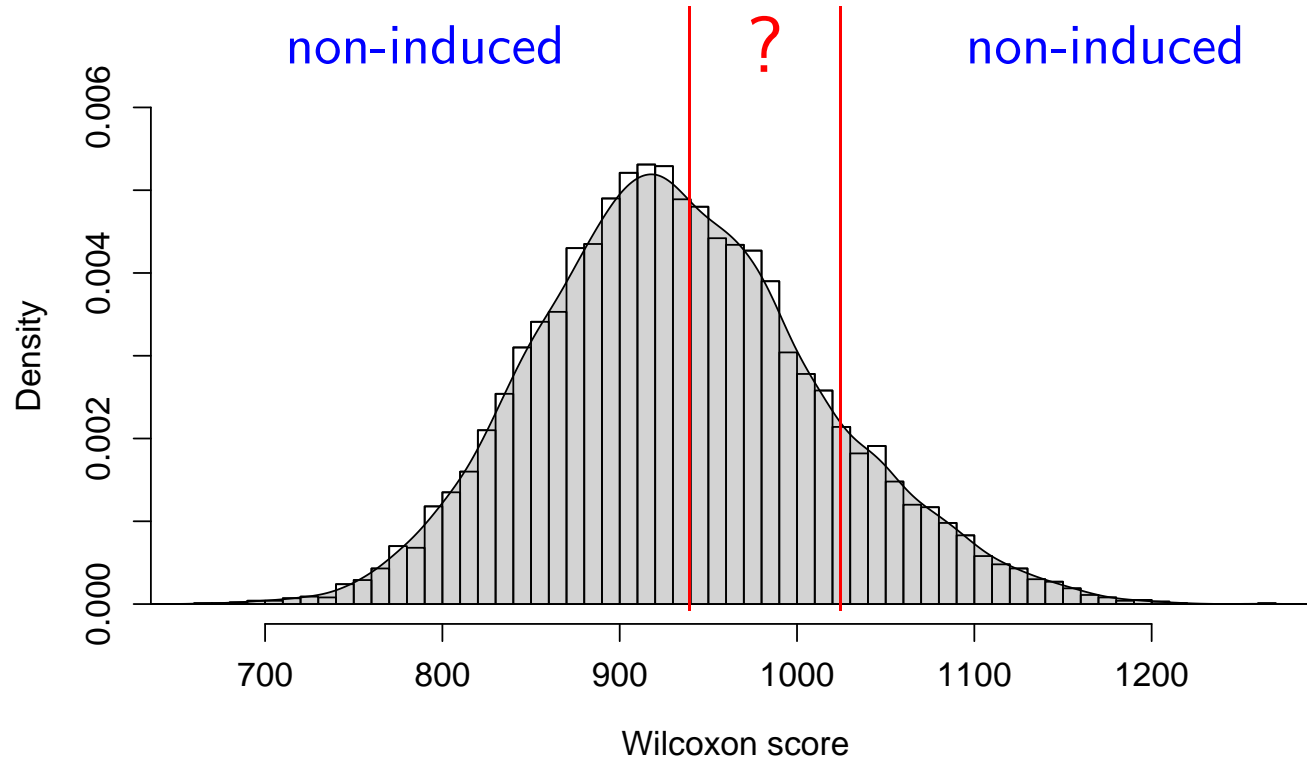
# Score distribution with twilight zone

What if score distributions overlap?



# Score distribution with twilight zone

What if score distributions overlap?



# How can we reconstruct the mixture?

Significance testing causes multiplicity problem.

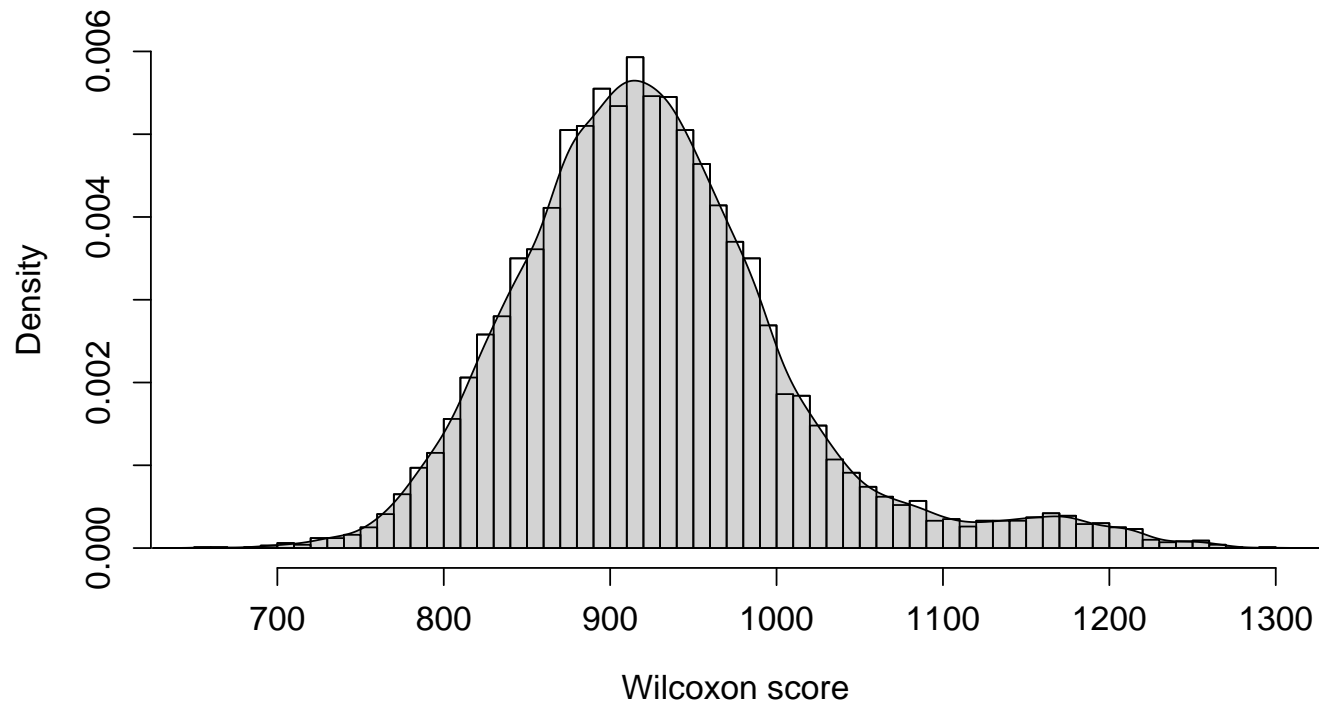
False Discovery Rate (FDR):

Expected number of falsely called induced genes among all genes called induced

= Probability of genes in rejection area to be non-induced.

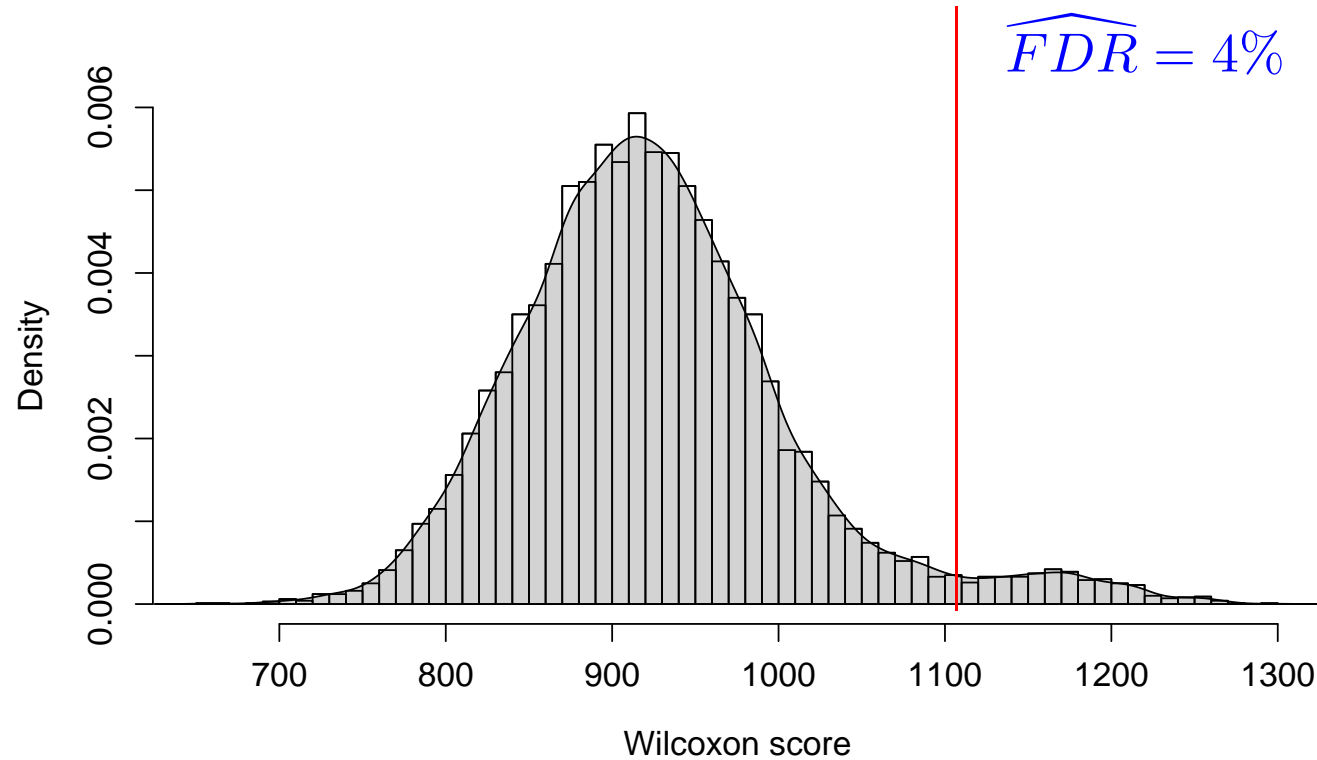
# From extremal FDR to bin-wise FDR

Define rejection area by threshold values and estimate FDR:



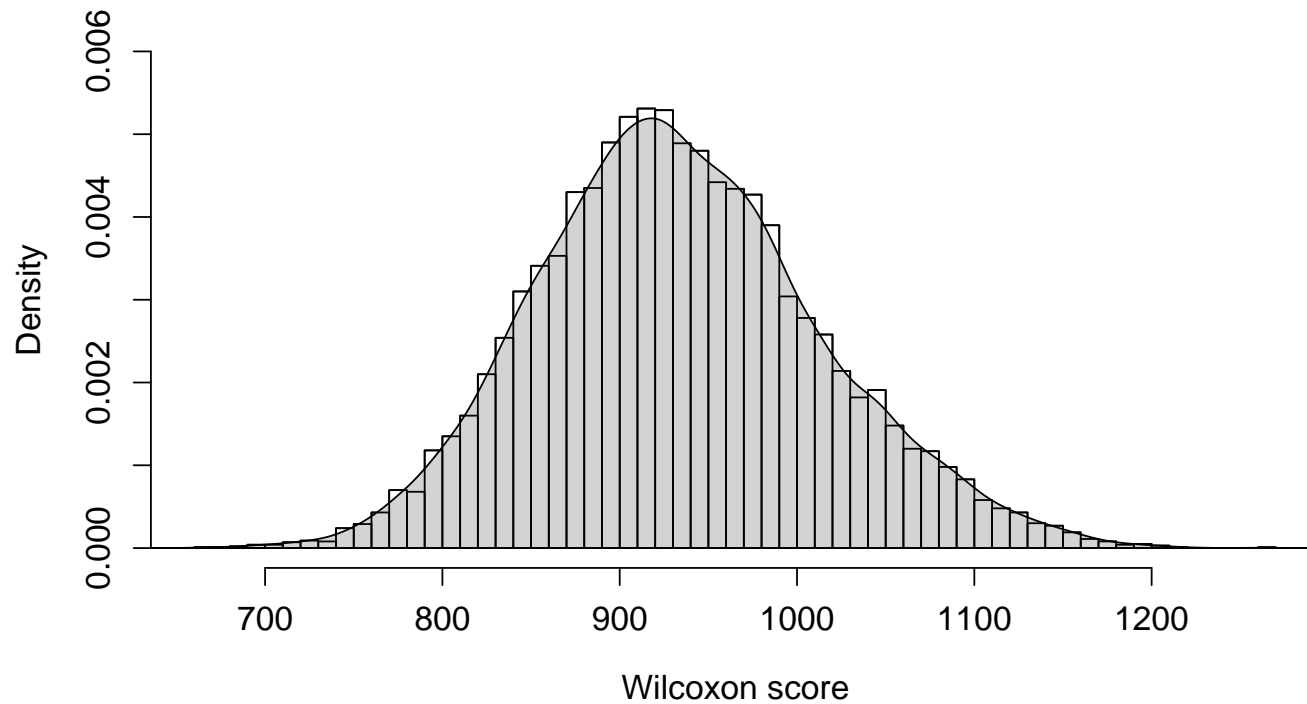
# From extremal FDR to bin-wise FDR

Define rejection area by threshold values and estimate FDR:



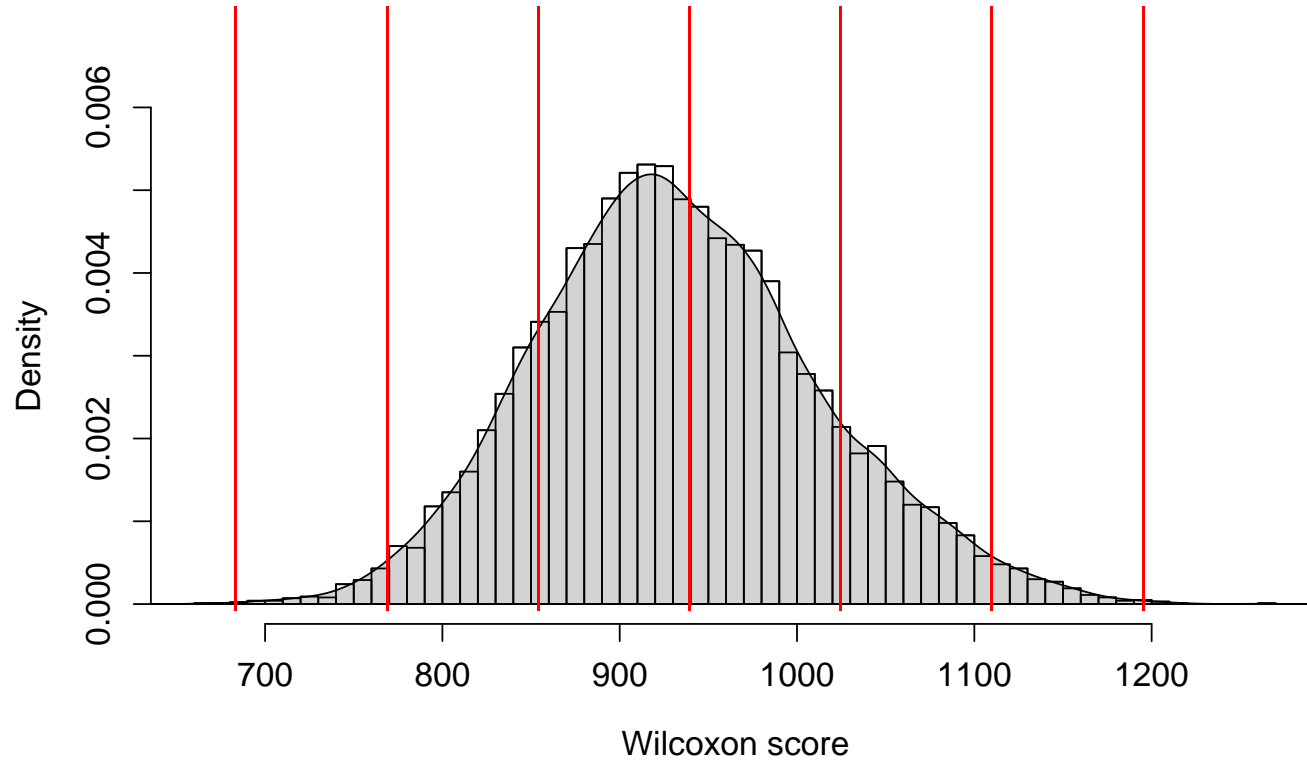
# From extremal FDR to bin-wise FDR

Define rejection area by threshold values and estimate FDR:



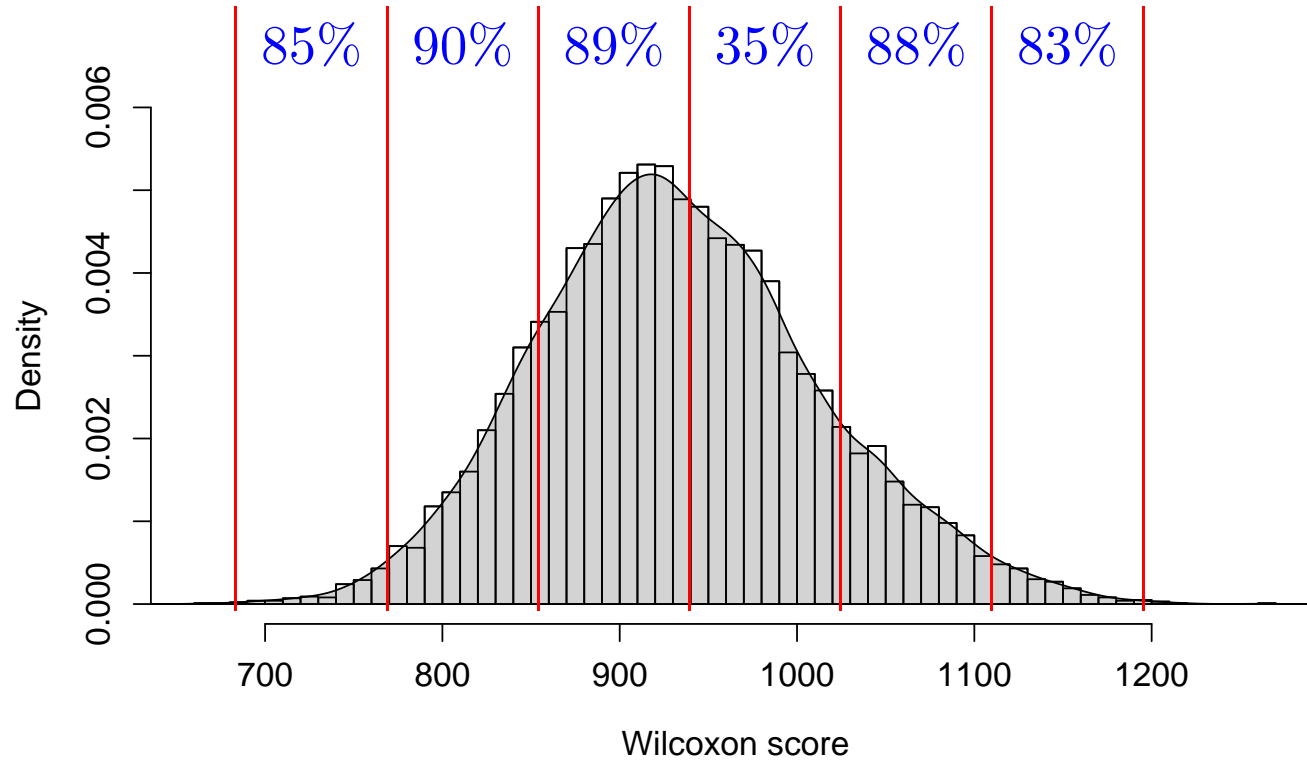
# From extremal FDR to bin-wise FDR

Define rejection area by threshold values and estimate FDR:



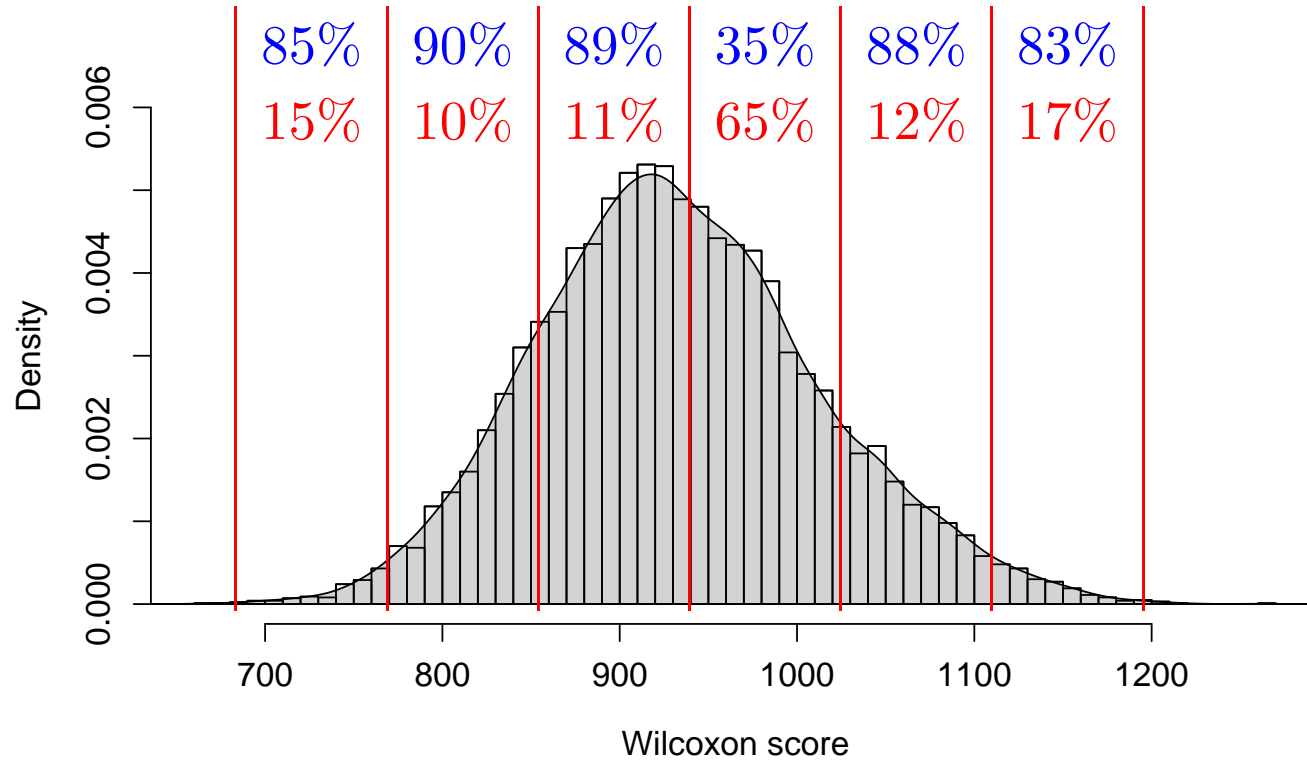
# From extremal FDR to bin-wise FDR

Define rejection area by threshold values and estimate FDR:



# From extremal FDR to bin-wise FDR

Define rejection area by threshold values and estimate FDR:



# Estimating bin-wise FDR

Define useful binning.

Estimate FDR as in Tusher et al. (2001) using class permutation:

$N_i$  = Number of observed scores in bin  $i$

$N_i^k$  = Number of scores of  $k$ th permutation in bin  $i$

For bin  $i$ :

$$\begin{aligned} FDR_i &= Prob(\text{non-induced} | \text{"induced"}) \\ &= \frac{Prob(\text{"induced"} | \text{non-induced})}{Prob(\text{"induced"})} Prob(\text{non-induced}) \end{aligned}$$

For bin  $i$ :


$$\begin{aligned} FDR_i &= Prob(\text{non-induced} | \text{"induced"}) \\ &= \frac{Prob(\text{"induced"} | \text{non-induced})}{Prob(\text{"induced"})} Prob(\text{non-induced}) \end{aligned}$$

$$\widehat{FDR}_i = \frac{\text{median}(N_i^k)}{N_i} Prob(\text{non-induced})$$

For bin  $i$ :

$$\begin{aligned} FDR_i &= Prob(\text{non-induced} | \text{"induced"}) \\ &= \frac{Prob(\text{"induced"} | \text{non-induced})}{Prob(\text{"induced"})} Prob(\text{non-induced}) \end{aligned}$$

$$\widehat{FDR}_i = \frac{\text{median}(N_i^k)}{N_i} Prob(\text{non-induced})$$



requires prior  
knowledge or  
good estimator

Calculate lower and upper quartile of all permutation scores.

$$\widehat{Prob}(\text{non-induced}) = \frac{\text{Number of observed scores in } [q_{.25}, q_{.75}]}{0.5 \cdot \text{Total number of observed scores}}$$

Calculate lower and upper quartile of all permutation scores.

$$\widehat{Prob}(\text{non-induced}) = \frac{\text{Number of observed scores in } [q_{.25}, q_{.75}]}{0.5 \cdot \text{Total number of observed scores}}$$

$$\Rightarrow \widehat{FDR}_i = \frac{\text{median}_k(N_i^k)}{N_i} \widehat{Prob}(\text{non-induced})$$

# Simulation study

Two classes with 30 samples each, 1000 genes, 10 000 permutations, 20 bins.

Two features of (non-induced) gene expression:

1. Characteristic **gene profile** across samples: Draw expression value from lognormal distribution and add individual standard normal error.
2. Correlation due to pathways/coregulation, **“clumpy dependence”**: Add same standard normal error to blocks of 50 genes.

Induce a fraction  $\pi$  of genes in one class with mean offset  $\mu$  from  $N(\mu, \sigma = 0.2)$ .

$\pi = 5, 15, 25, 50\%$

$\mu = 0.5, 0.7$

Induce a fraction  $\pi$  of genes in one class with mean offset  $\mu$  from  $N(\mu, \sigma = 0.2)$ .

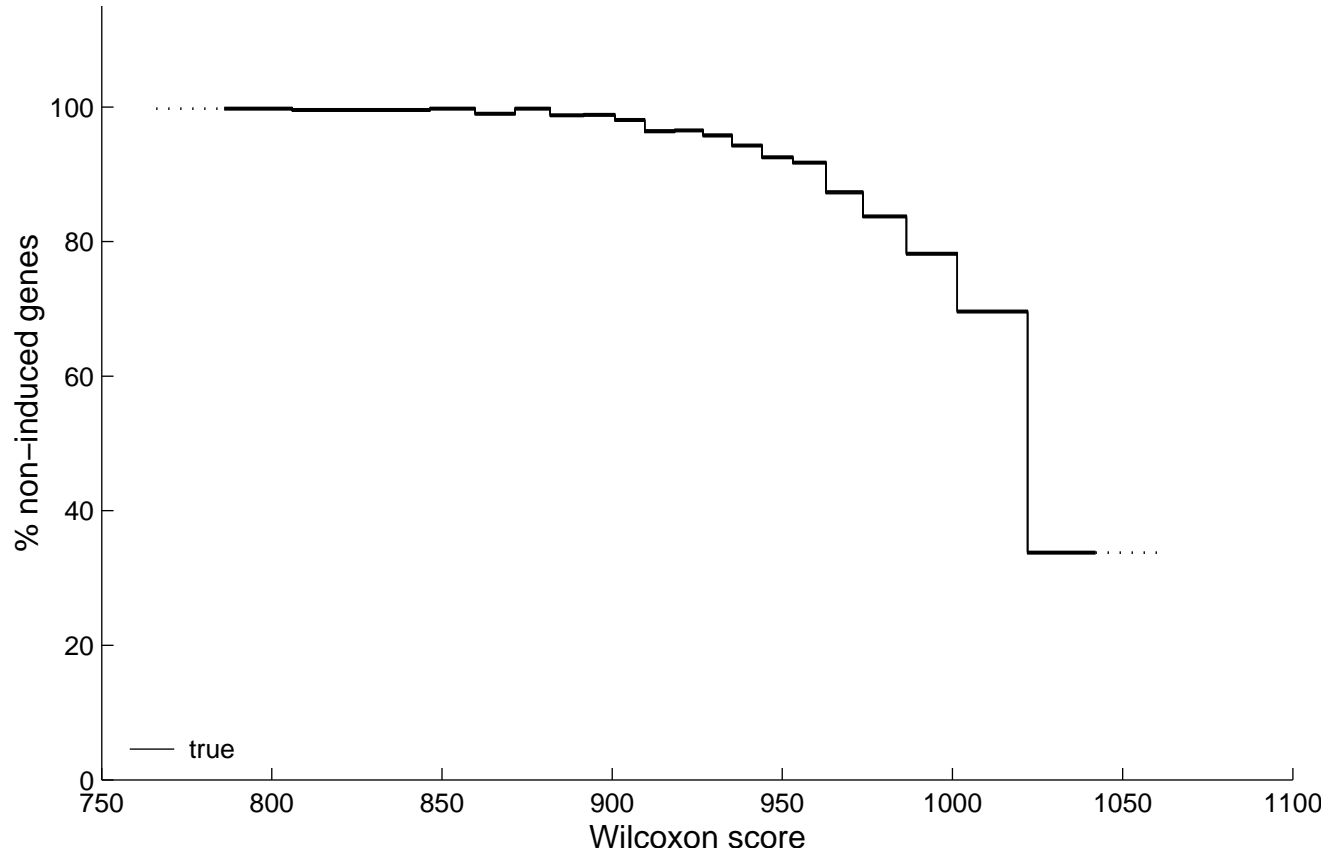
$\pi = 5, 15, 25, 50\%$

$\mu = 0.5, 0.7$

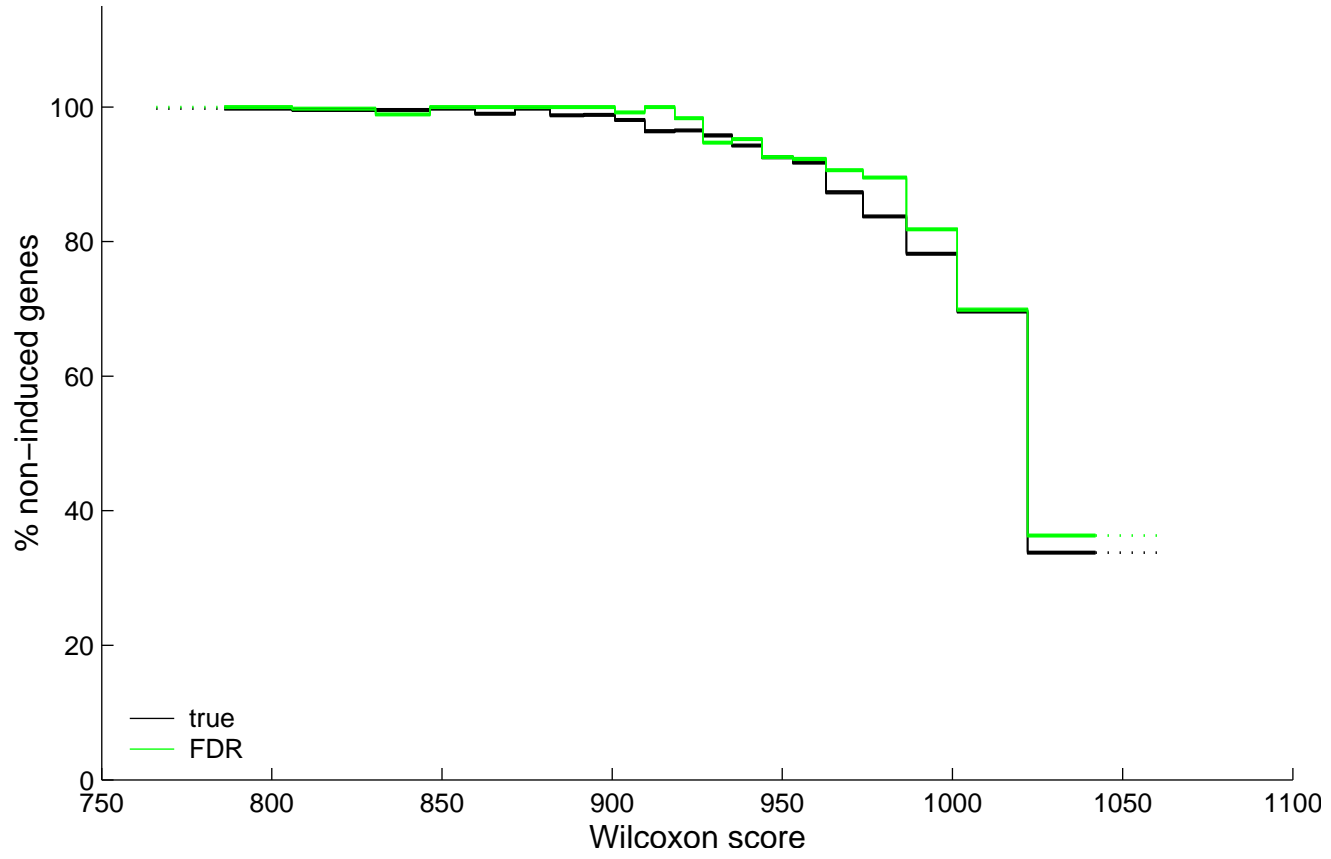
Repeat each parameter combination 20 times  $\Rightarrow$

Averaged  $\widehat{FDR}_i$  and averaged **true proportion** of non-induced genes.

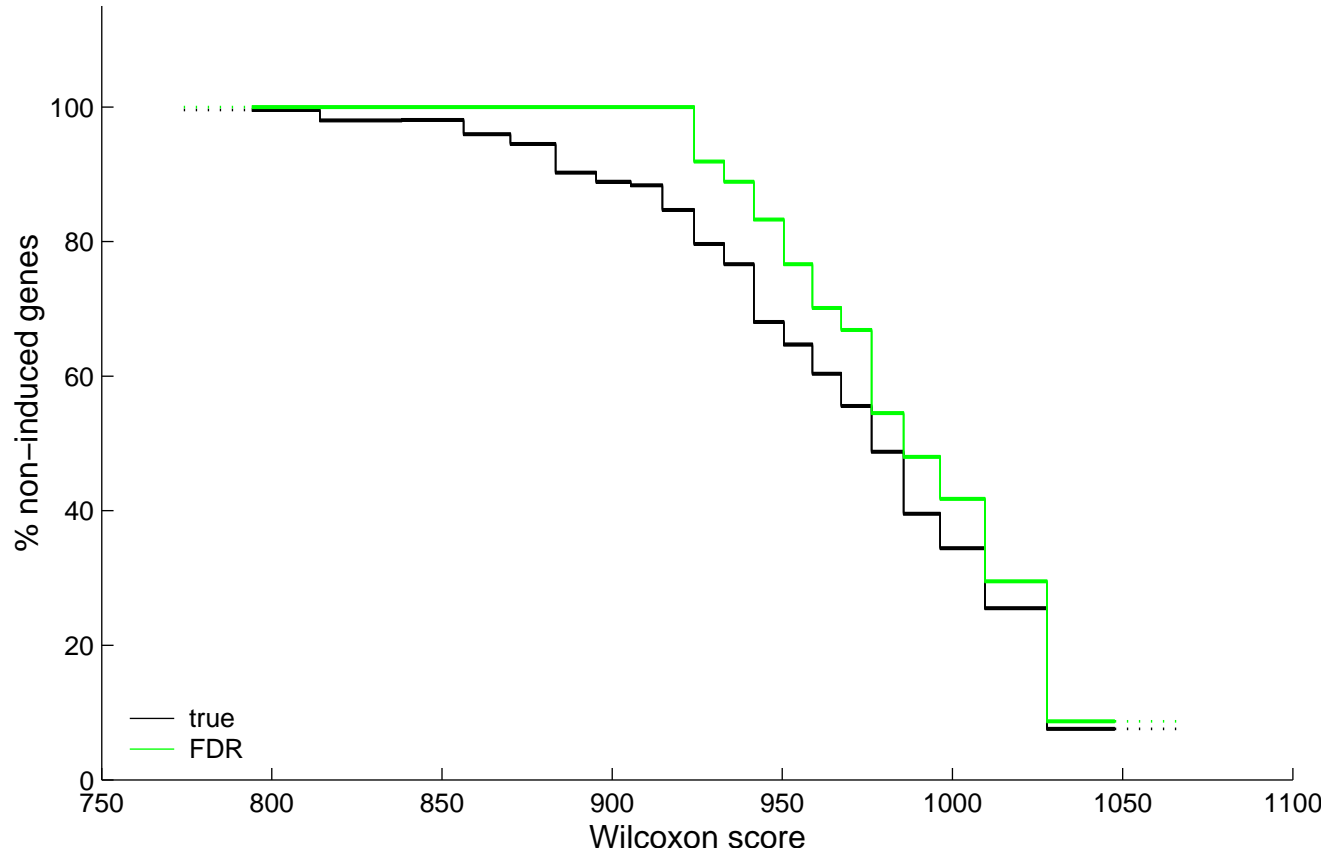
$\pi_{\text{true}} = 15\%$     $\pi_{\text{est.}} = 12.49\%$     $\mu = 0.5$



$\pi_{\text{true}} = 15\%$     $\pi_{\text{est.}} = 12.49\%$     $\mu = 0.5$



$\pi_{\text{true}} = 50\%$     $\pi_{\text{est.}} = 41.18\%$     $\mu = 0.5$



## Successive Exclusion Procedure (SEP)

Exclude **minimal number of genes** such that for remaining genes the observed score distribution is not significantly different from those derived by class permutations.

Gene exclusion operates bin-wise.

Score distribution of remaining genes and score distribution of excluded genes constitute the mixture.

⇒ Bin-wise reconstruction of overall mixture.

## Definitions

$N_i$  = Observed frequency in bin  $B_i$

$N_i^k$  = Frequency in bin  $B_i$  after  $k$ th class permutation

$\Delta_i = N_i - N_i^k$

$s_j$  = Observed score for gene  $j$

$s_j^k$  = Score for gene  $j$  after  $k$ th class permutation

$[b_{.05,i} , b_{.95,i}]$  = Bin-wise lower and upper quantiles of frequencies

$N_i^k$  of *all* permutations

# SEP Algorithm

1. Find set  $\mathcal{S}$  of removable genes:

$$\mathcal{S} = \{j : (s_j \in B_i \text{ with } \Delta_i > 0) \text{ AND } (s_j^k \in B_{i'} \text{ with } \Delta_{i'} < 0)\}$$

If  $|\mathcal{S}| > 0$ , go to 2.

If  $|\mathcal{S}| = 0$ , go to 3.

2. Exclude randomly one gene  $\in \mathcal{S}$  from further consideration.

Recompute  $N_i, N_i^k, \Delta_i$  for remaining scores, go to 1.

3. Find set  $\mathcal{S}$  of removable genes:

$$\mathcal{S} = \{j : s_j \in B_i \text{ with } N_i > b_{.95,i}\}$$

If  $|\mathcal{S}| > 0$ , go to 4.

If  $|\mathcal{S}| = 0$ , go to 5.

4. Exclude randomly one gene  $\in \mathcal{S}$  from further consideration.

Recompute  $N_i, N_i^k, \Delta_i$  for remaining scores, go to 3.

5. Find set  $\mathcal{S}'$  of *excluded* genes  $j'$ :

$$\mathcal{S}' = \{j' : s_{j'} \in B_i \text{ with } N_i < b_{.05,i}\}$$

If  $|\mathcal{S}'| > 0$ , go to 6.

If  $|\mathcal{S}'| = 0$ , end.

6. Include randomly one gene  $\in \mathcal{S}'$  again.

Recompute  $N_i, N_i^k, \Delta_i$  for remaining scores, go to 5.

For all class permutations compute bin frequencies  $N_i^k$  of **remaining scores**.  
Compute observed bin frequencies  $N_i^o$  including **all scores**.

The estimated probability that a gene is induced given its score is contained in bin  $B_i$  is given as:

$$\widehat{Pr}(\text{Gene } j \text{ is induced} | s_j \in B_i) = \frac{\text{median}_k(N_i^k)}{N_i^o}.$$

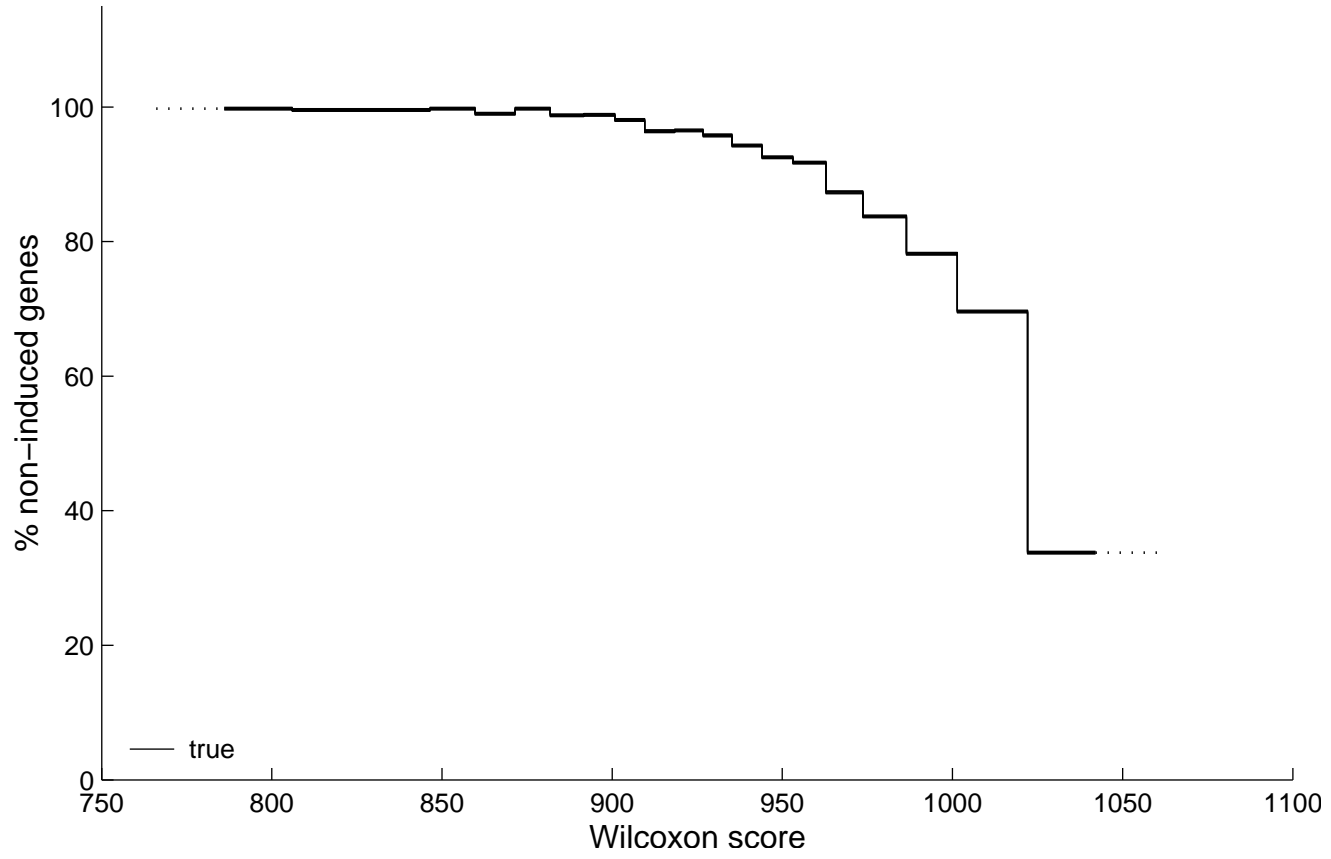
For all class permutations compute bin frequencies  $N_i^k$  of **remaining scores**.  
Compute observed bin frequencies  $N_i^o$  including **all scores**.

The estimated probability that a gene is induced given its score is contained in bin  $B_i$  is given as:

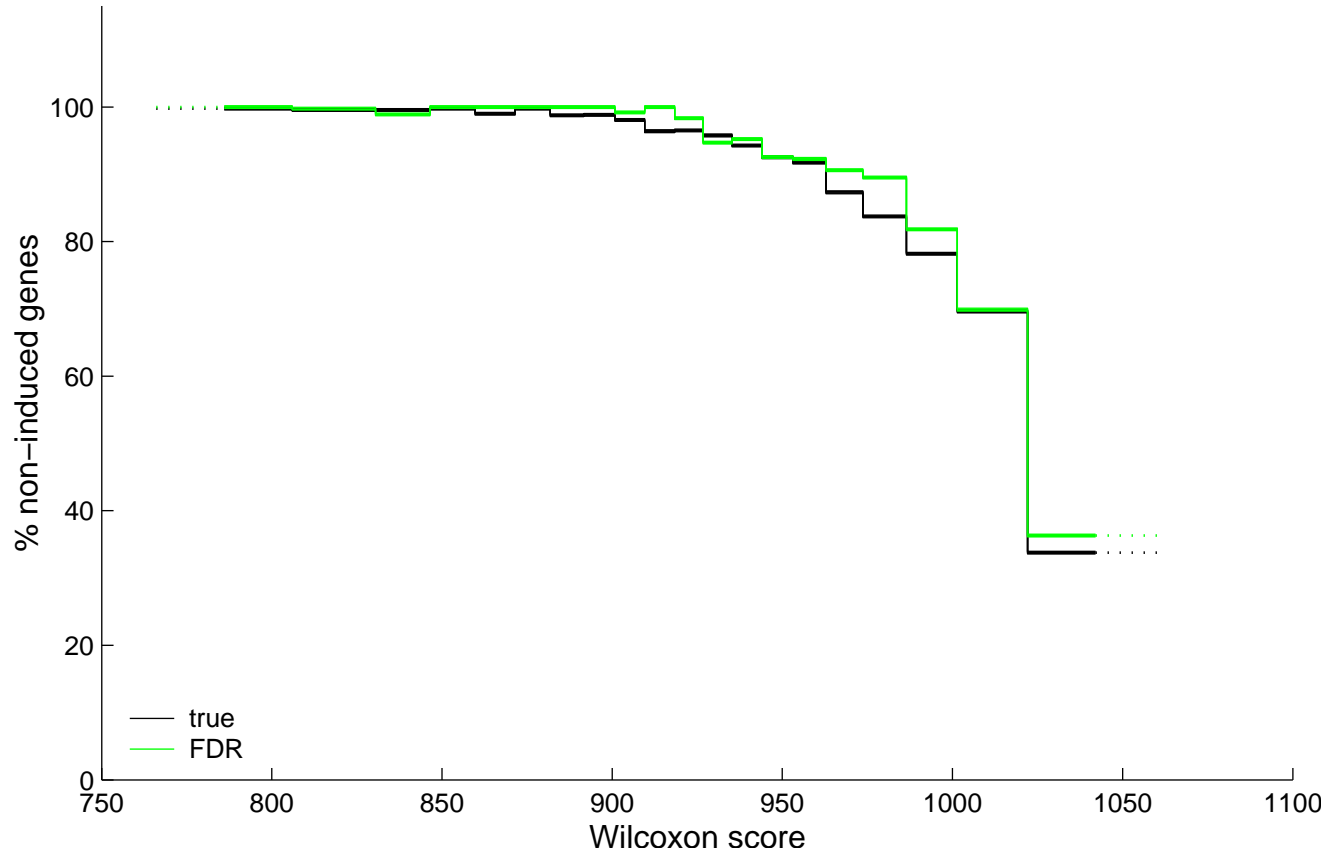
$$\widehat{Pr}(\text{Gene } j \text{ is induced} | s_j \in B_i) = \frac{\text{median}_k(N_i^k)}{N_i^o}.$$

Let's compare **true**, **FDR** and **SEP**.

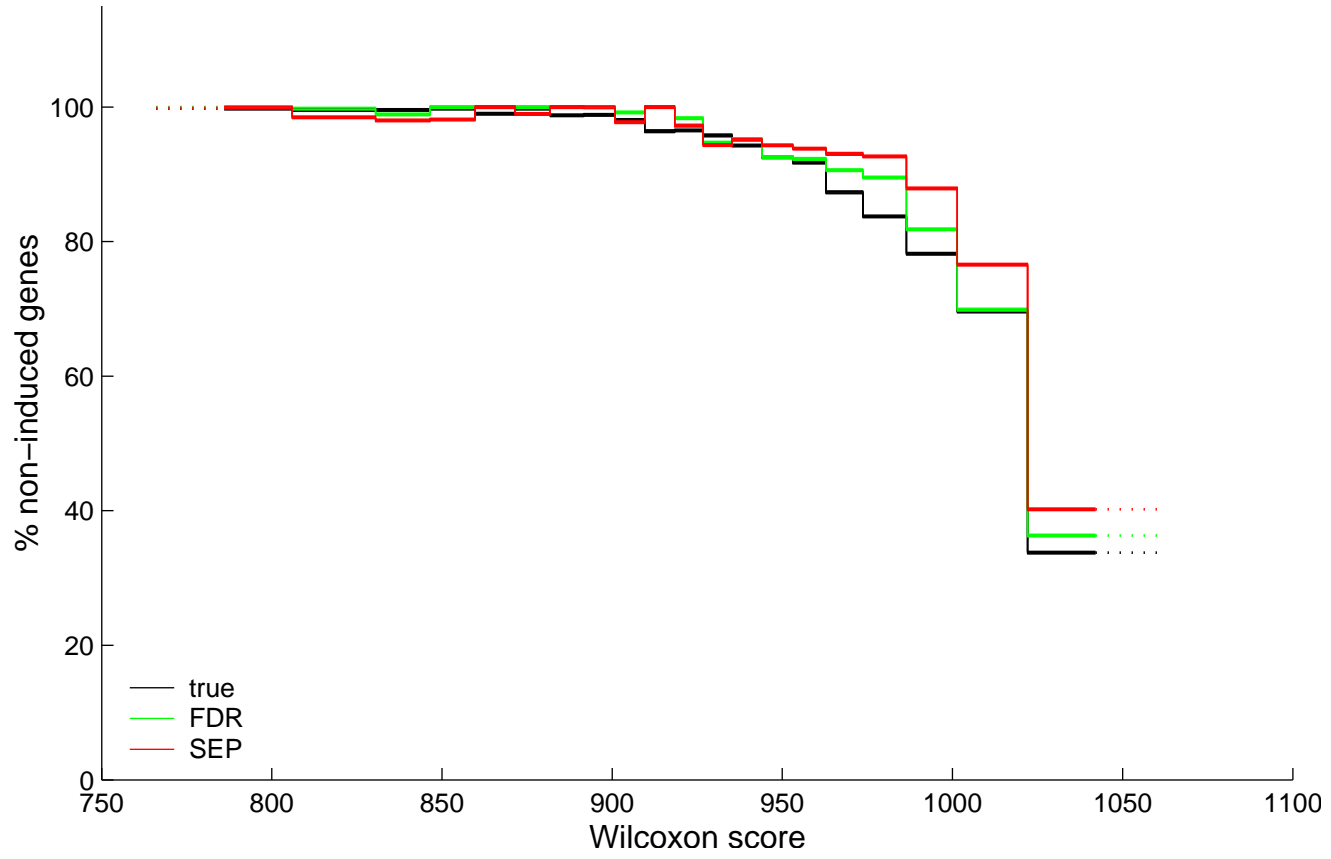
$\pi_{\text{true}} = 15\%$     $\pi_{\text{est.}} = 12.49\%$     $\mu = 0.5$



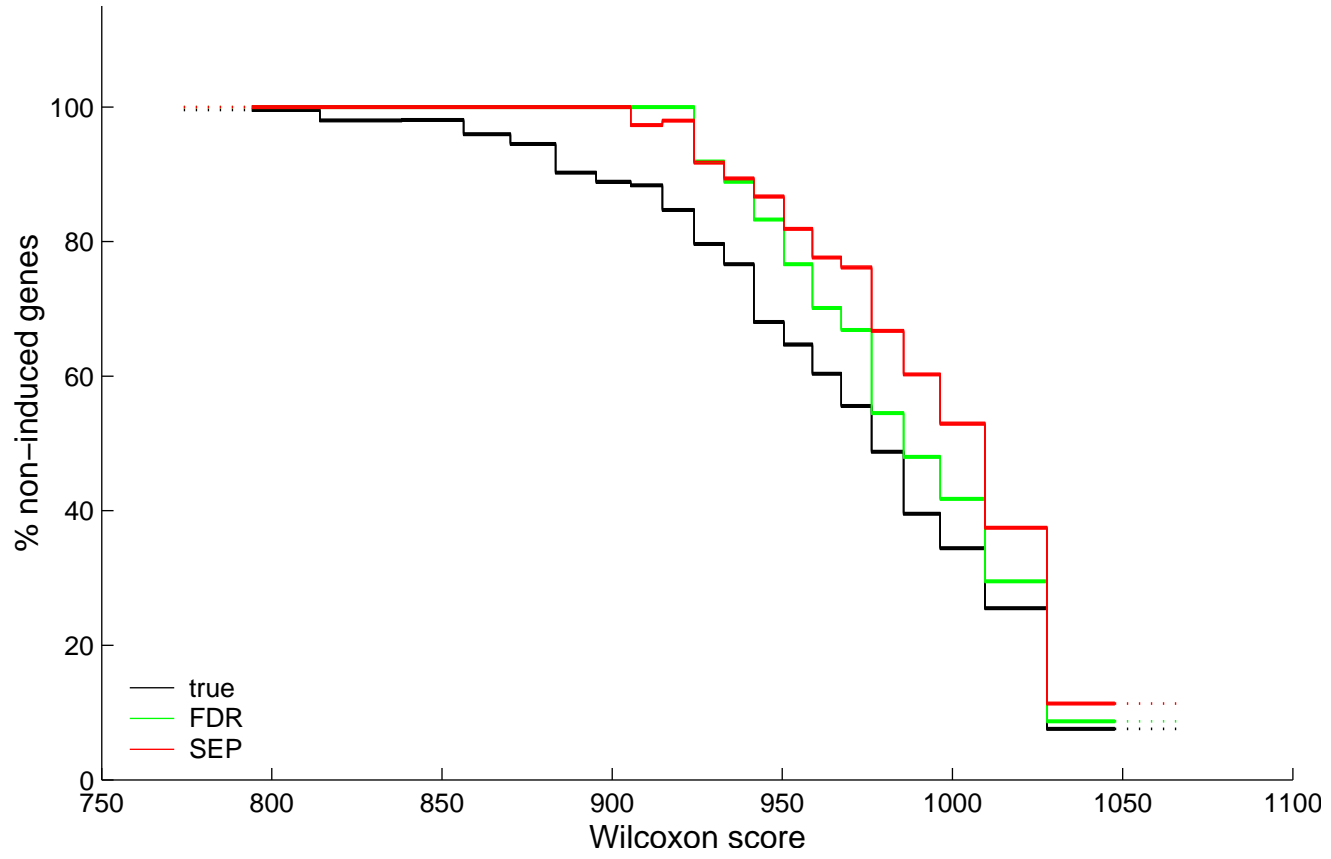
$\pi_{\text{true}} = 15\%$     $\pi_{\text{est.}} = 12.49\%$     $\mu = 0.5$



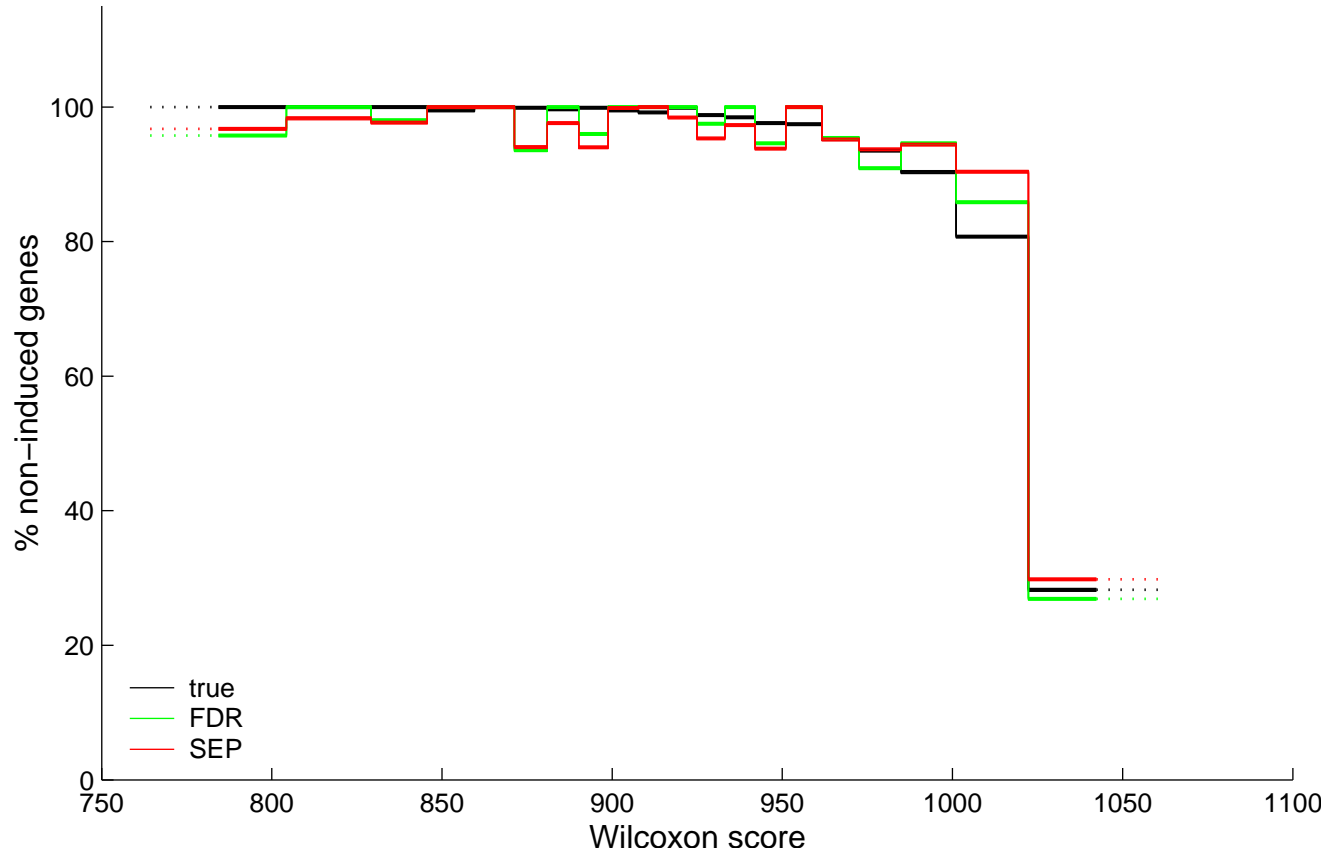
$\pi_{\text{true}} = 15\%$     $\pi_{\text{est.}} = 12.49\%$     $\mu = 0.5$



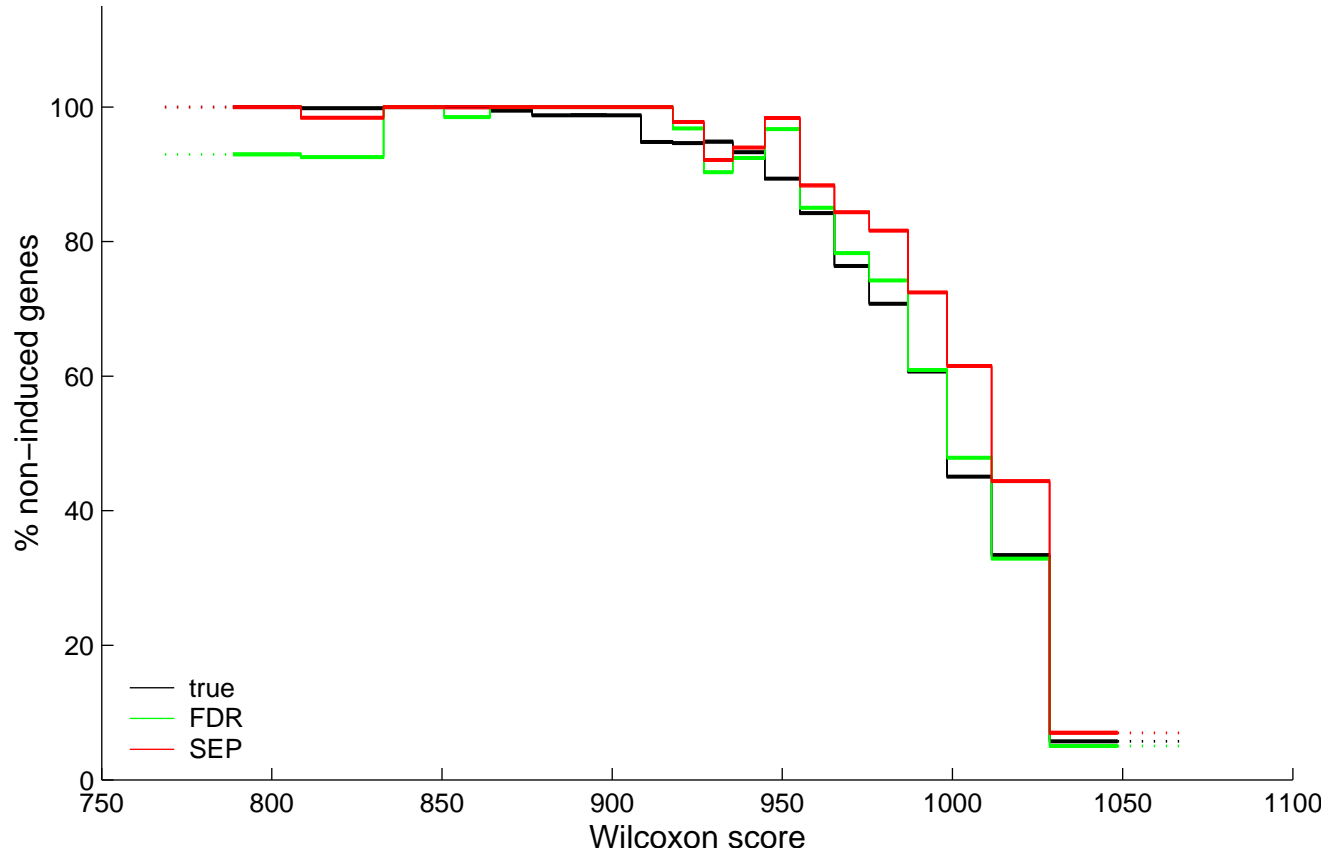
$\pi_{\text{true}} = 50\%$     $\pi_{\text{est.}} = 41.18\%$     $\mu = 0.5$



$\pi_{\text{true}} = 15\%$     $\pi_{\text{est.}} = 14.53\%$     $\mu = 0.7$



$\pi_{\text{true}} = 50\%$     $\pi_{\text{est.}} = 49.37\%$     $\mu = 0.7$



## Discussion

Both methods discover overall shape of true percentage curve.

Bin-wise FDR is in general closer to true percentage.

As expected, SEP is more conservative. Underestimated percentages of induced genes are due to minimal removal idea.

Further research: Use sliding window FDR to achieve smoother estimates.

## FDR references

Benjamini, Y., Hochberg, Y. 1995. **Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing**. Journal of the Royal Statistical Society B, 57(1):289–300.

Storey, J. D., Tibshirani, R. 2001. **Estimating False Discovery Rates Under Dependence, with Applications to DNA Microarrays**, Stanford University: Stanford Technical Report.

Tusher, V. G., Tibshirani, R., Chu, G. 2001. **Significance analysis of microarrays applied to the ionizing radiation response**. Proceedings of the National Academy of Sciences, 98(9):5116–5121.