MEDICAL
UNIVERSITY
OF VIENNA

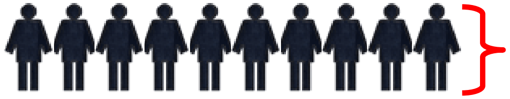
Stopping rules for sequential trials in high-dimensional data

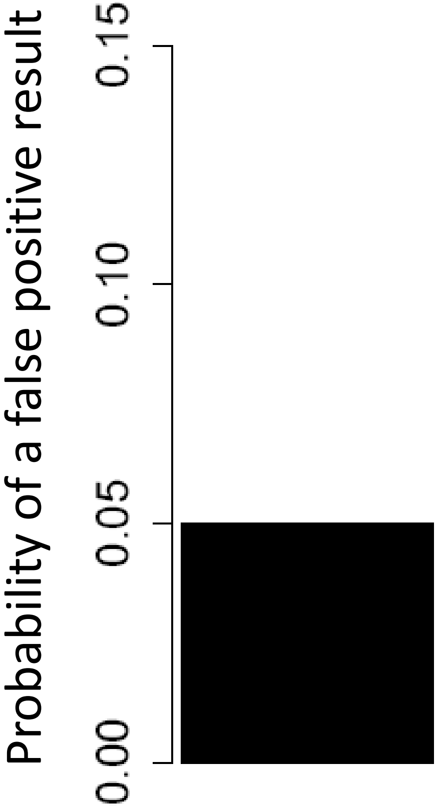
Sonja Zehetmayer,
Alexandra Graf, and Martin Posch

Center for Medical Statistics, Informatics and Intelligent Systems
Medical University of Vienna

Supported by FWF - Funds Nr. T401 and P23167

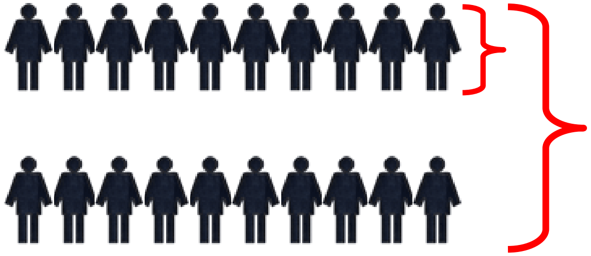
Hunting for significance inflates the probability of a false positive result

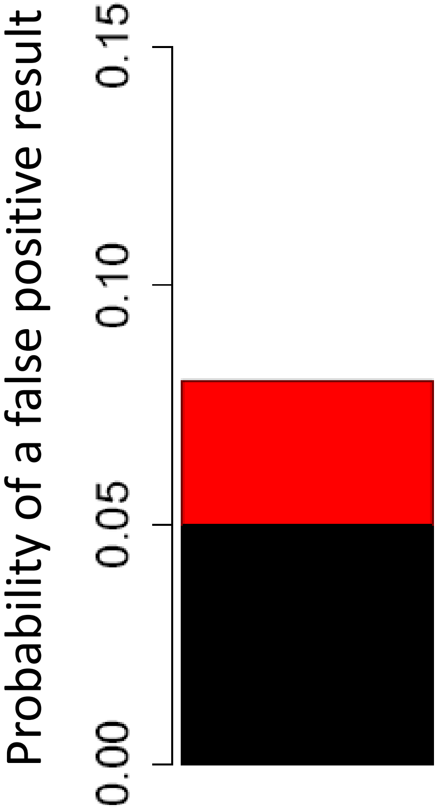




$\alpha = 0.05$

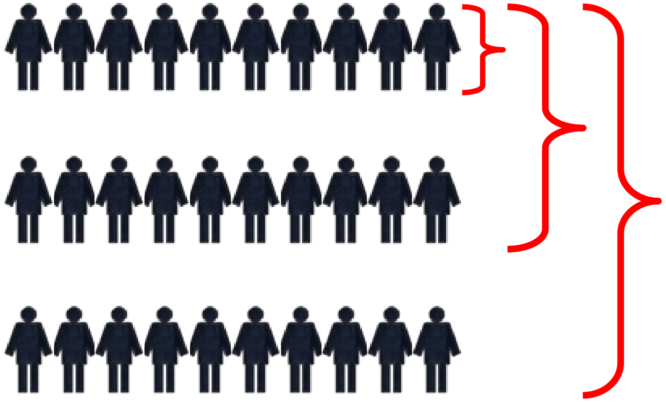
Hunting for significance inflates the probability of a false positive result

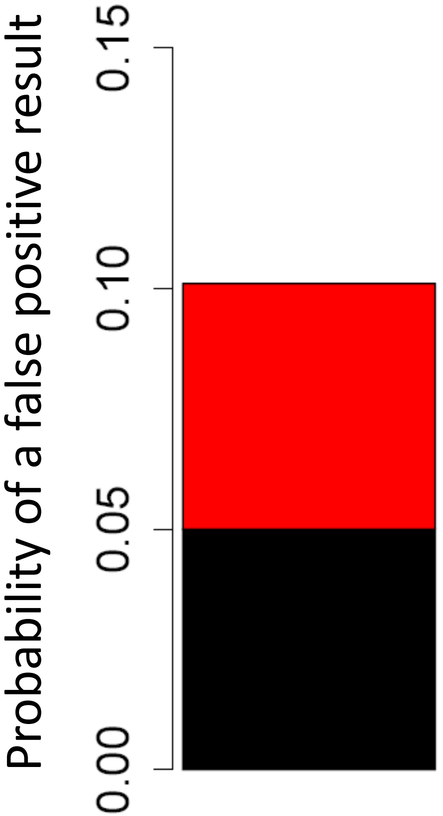




$\alpha = 0.05$

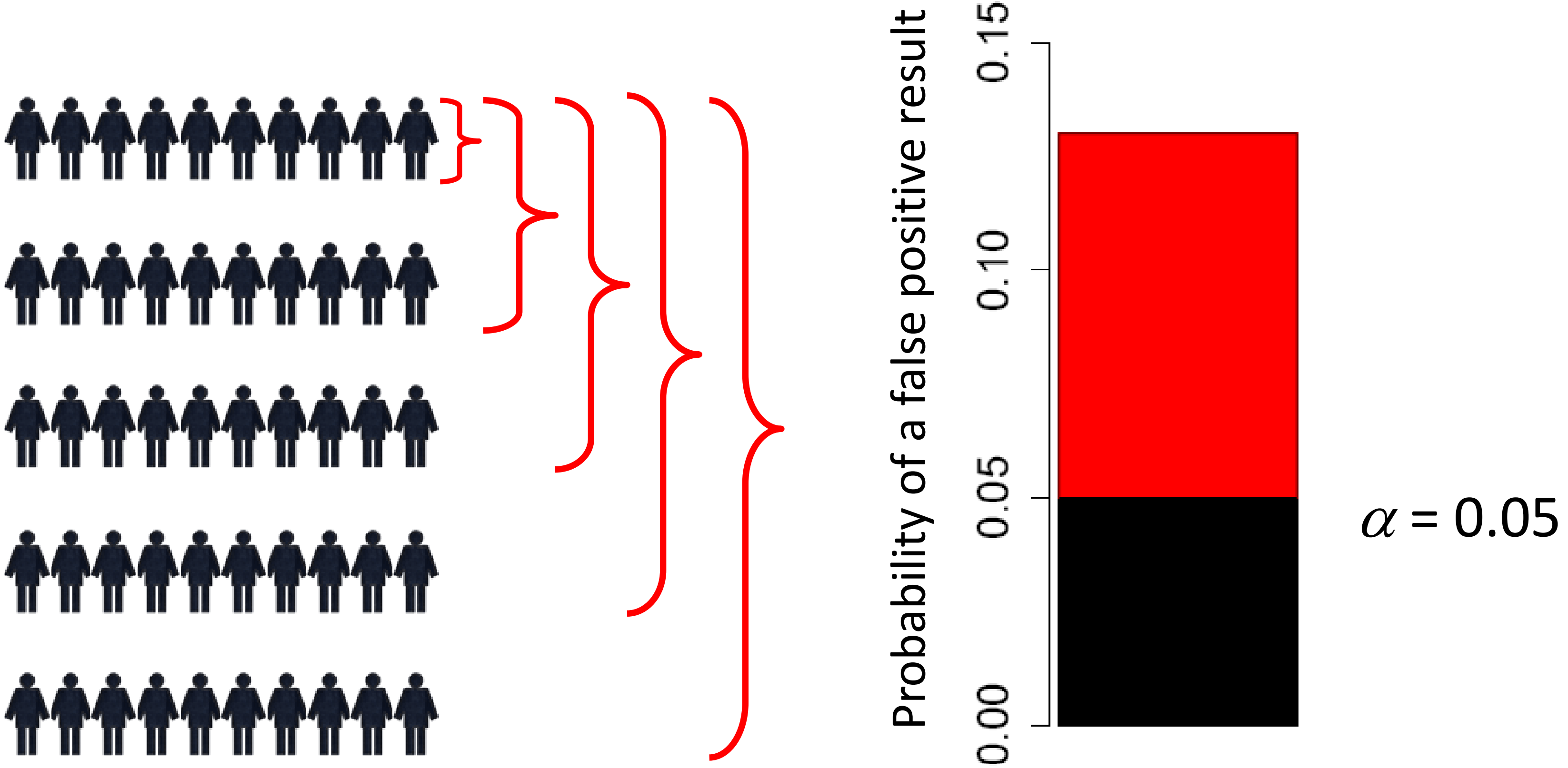
Hunting for significance inflates the probability of a false positive result





α = 0.05

Hunting for significance inflates the probability of a false positive result



Conclusion I

- Testing a single hypothesis repeatedly at several interim analyses at level α ("*Hunting for significance*"), increases the probability of a false positive result.
- Solution: Group sequential tests: adjust α

What about very many hypotheses?

Many hypotheses

- *m* hypotheses (genes), e.g., microarray study

$$H_{0i}: \mu_i = 0 \quad \text{versus} \quad H_{1i}: \mu_i \neq 0, \qquad i=1,\ldots,m$$

The False Discovery Rate (FDR)

Benjamini and Hochberg, 1995

$$\boxed{FDR = E\left(\frac{V}{\max\{R,1\}}\right)}$$

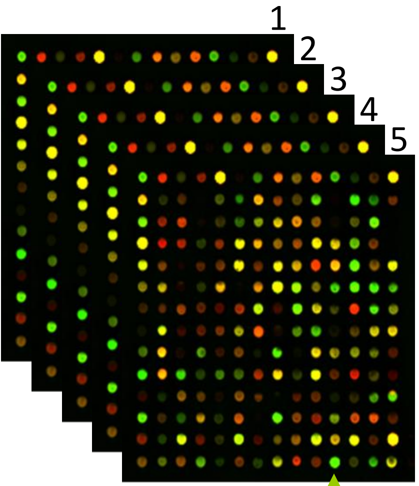
V : number of erroneously rejected null hypotheses

R : number of rejected null hypotheses

FDR of the experiment is controlled according to Benjamini and Hochberg (1995)

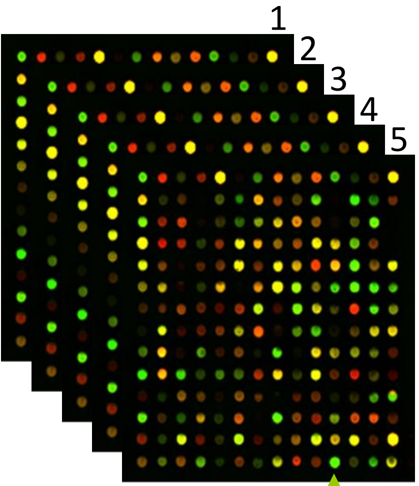
- Order the individual p-values $p_{(1)} \leq \ldots \leq p_{(m)}$
- $d = \text{argmax}_{\,i} \{p_{(i)} \leq i\alpha/m\}$
- Reject all hypotheses with p-values $p_{(1)} \ldots\ p_{(d)}$

This is a conservative procedure for controlling the FDR if the test statistics are independent or positively dependent (Benjamini and Yekutieli, 2001)



Analysis controlling the FDR at level α

1 spot $\hat{=}$ 1 hypothesis



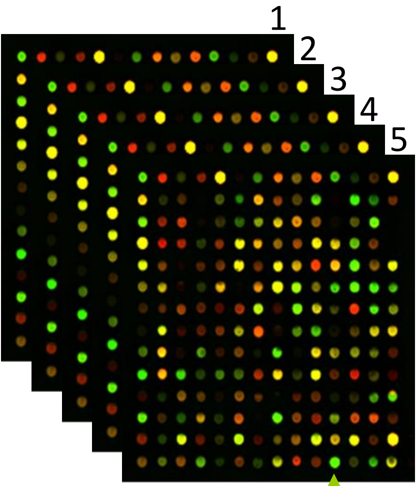
1 spot ≙ 1 hypothesis

Analysis controlling the FDR at level α

Stop the experiment.

Reject all significant hypotheses.

Retain all others.



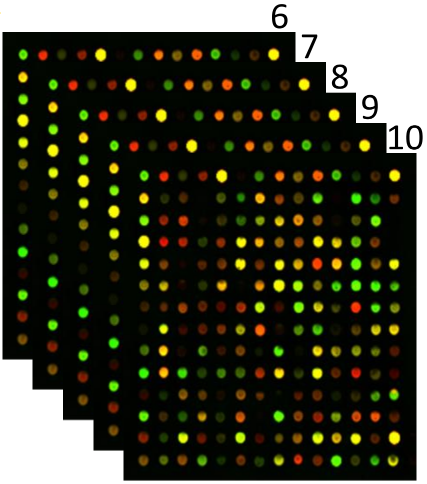
1 spot $\hat{=}$ 1 hypothesis

Analysis controlling the FDR at level α

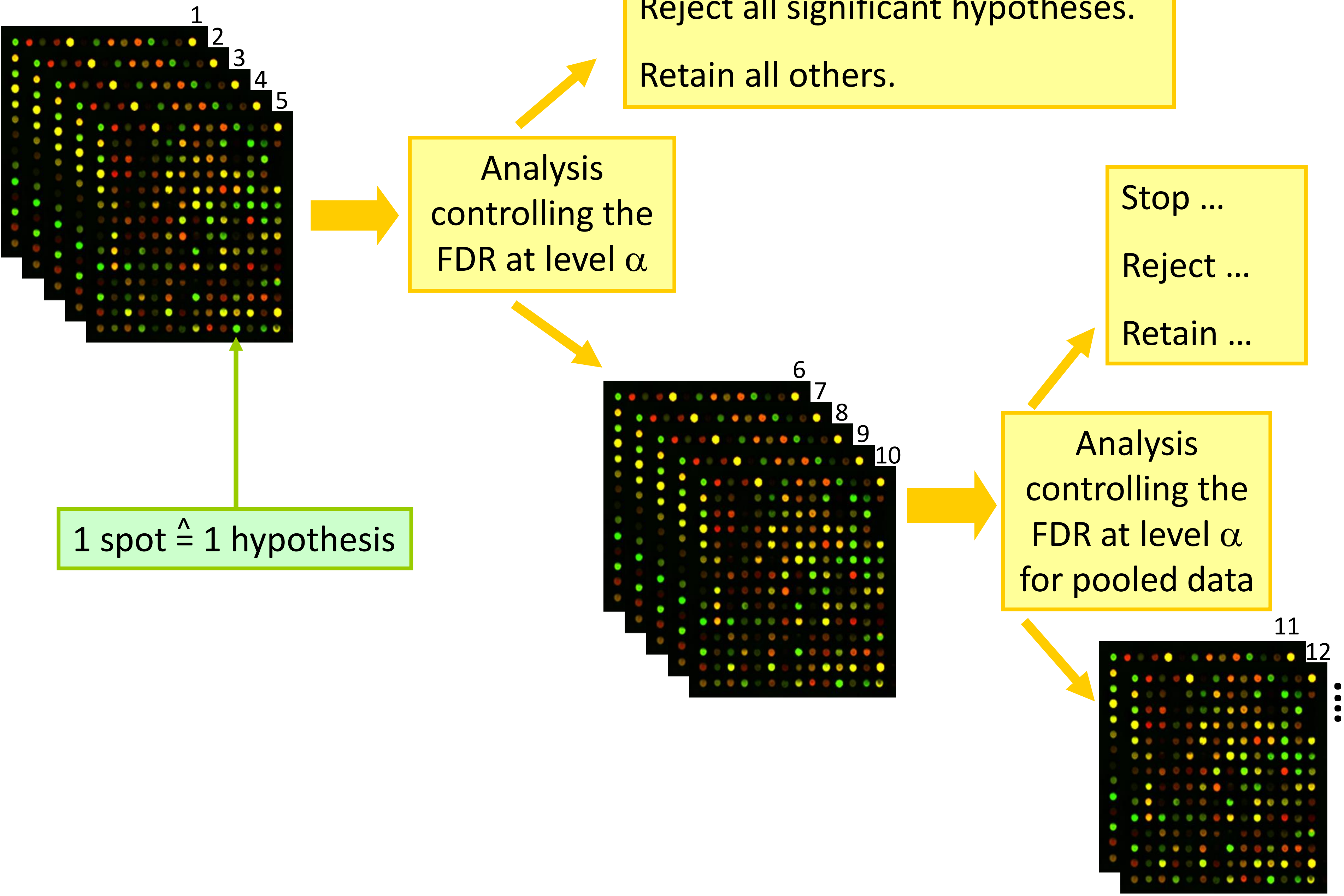
Stop the experiment.

Reject all significant hypotheses.

Retain all others.



Analysis controlling the FDR at level α for pooled data



1 spot $\hat{=}$ 1 hypothesis

Analysis controlling the FDR at level α

Stop the experiment.

Reject all significant hypotheses.

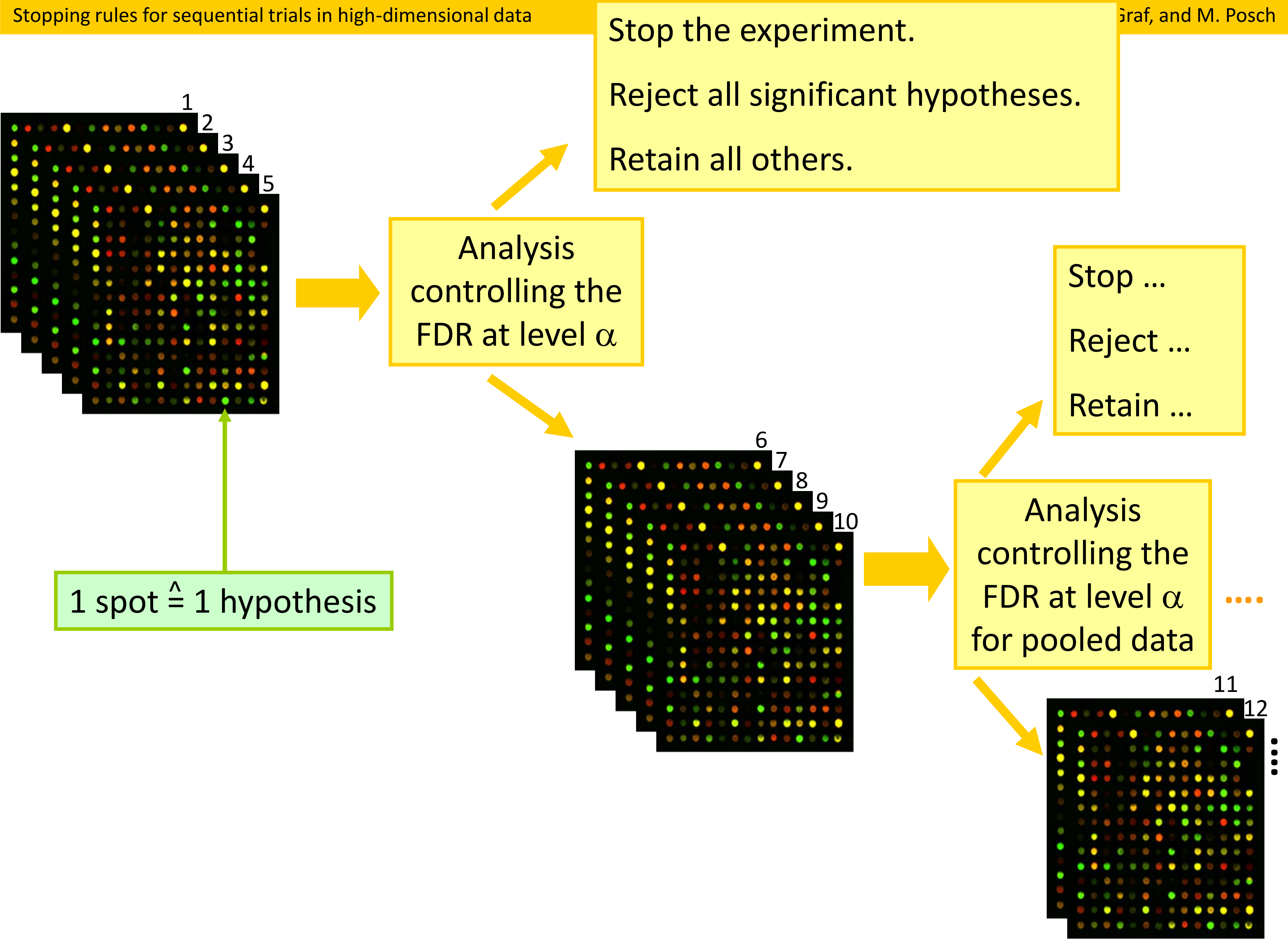
Retain all others.

Analysis controlling the FDR at level α for pooled data

Stop …

Reject …

Retain …



Stop the experiment.

Reject all significant hypotheses.

Retain all others.

Analysis controlling the FDR at level α

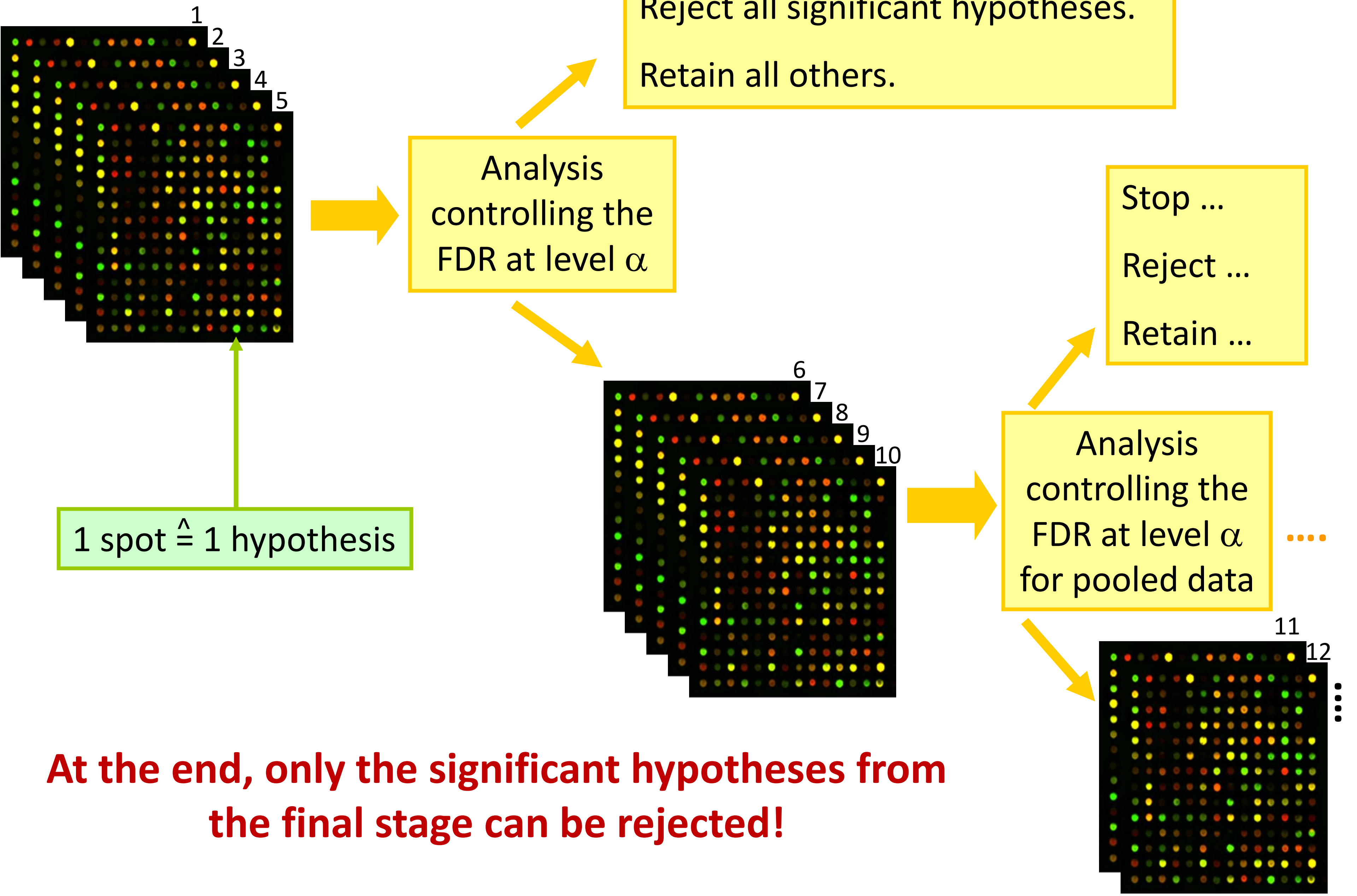
1 spot $\hat{=}$ 1 hypothesis

Stop …

Reject …

Retain …

Analysis controlling the FDR at level α for pooled data



Stop the experiment.

Reject all significant hypotheses.

Retain all others.

Analysis controlling the FDR at level α

Stop …

Reject …

Retain …

Analysis controlling the FDR at level α for pooled data

1 spot $\hat{=}$ 1 hypothesis

At the end, only the significant hypotheses from the final stage can be rejected!

What is the effect of unadjusted repeated analyses on the FDR?

What is the effect of unadjusted repeated analyses on the FDR?

Depends on the number of true null hypotheses m_0:

- In case of $\boldsymbol{m_0/m<1}$:
  For $m \rightarrow \infty$, the FDR is controlled asymptotically regardless of the stopping stage (under suitable assumptions).

- In case of $\boldsymbol{m_0/m=1}$ (global H_0):
  A constraint on the stopping rule has to be imposed:
  Stop early only if at least a certain number *s(m)* of hypotheses can be rejected.
  Then early stopping hardly occurs.

Then the FDR is controlled asymptotically
(Posch, Zehetmayer, Bauer, 2009)

Stopping the experiment

Stopping for futility

- Futility boundary $\alpha_1 > \alpha$

Early rejection

- Proportion of rejected H0
- Δ Proportion of rejected H0
- False Negative Rate
- Δ False Negative Rate
- False Non Discovery Rate
- Concordance

(and at least *s(m)* hypotheses can be rejected)

Stop as soon as the FNR is < 20%

e.g., Zehetmayer & Posch (2010)

- Multiple Type II Error
- Expected proportion of not-rejected true alternative hypotheses among all true alternative hypotheses

$$\boxed{FNR = E\left(1 - \frac{R - V}{m - m_0}\right)}$$

- *R*: # of rejections
- *V:* # of false rejections
- *m:* # of hypotheses
- m_0: # of true null hypotheses

In each stage *k* the *FNR* is estimated from the data

- γ: critical value from the FDR-controlling procedure
- The p-values corresponding to the true null hypotheses are uniformly distributed.

$$FNR_k = E\left(1 - \frac{R_k - V_k}{m - m_0}\right) = 1 - \frac{E(R_k) - m_0\gamma_k}{m - m_0}$$

- $\widehat{m}_{0k}$: estimator for m_0
- $R_k(\gamma) = \# \{p_{ik} < \gamma_k\}$

$$\boxed{\widehat{FNR}_k = 1 - \frac{R_k(\gamma_k) - \hat{m}_{0k}\gamma_k}{m - \hat{m}_{0k}}}$$

Stop as soon as $\Delta FNR < 0.05$

- ΔFNR is based on the increment of the stagewise FNR:

$$\Delta FNR_k = FNR_k - FNR_{k-1}$$

with $FNR_0=1$.

- In each stage ΔFNR is estimated as described before:

$$\Delta\widehat{FNR}_k = \widehat{FNR}_k - \widehat{FNR}_{k-1}$$

Stop as soon as the concordance of the rejected hypotheses from stage to stage > 0.9

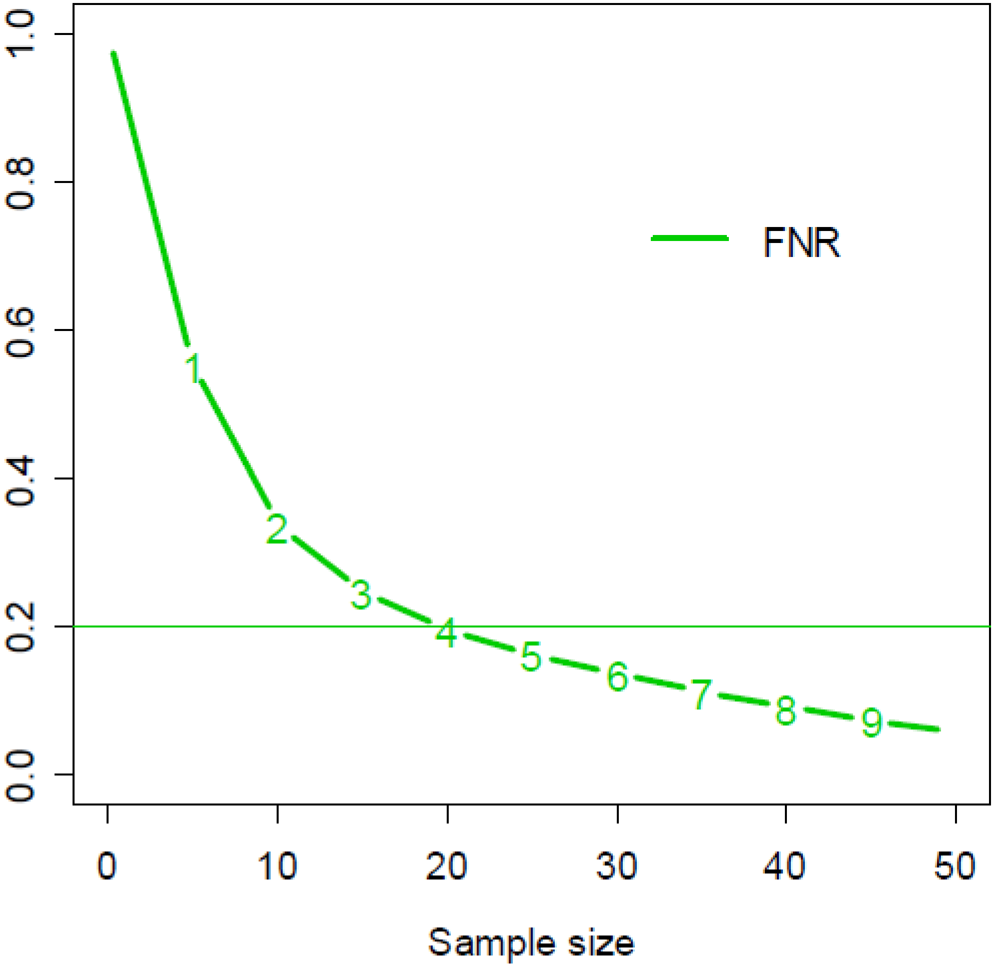
- Concordance (CO) measures the proportion of significant genes in stage *k* which were also significant in stage *k-1*:

$$\mathrm{CO}_k = \sum_i (H_{ir_{k-1}} H_{ir_k}) / \sum_i H_{ir_k}$$

where H_{ir_k} = 1 if hypothesis *i* was significant in stage *k* and 0 else with CO_1=0.

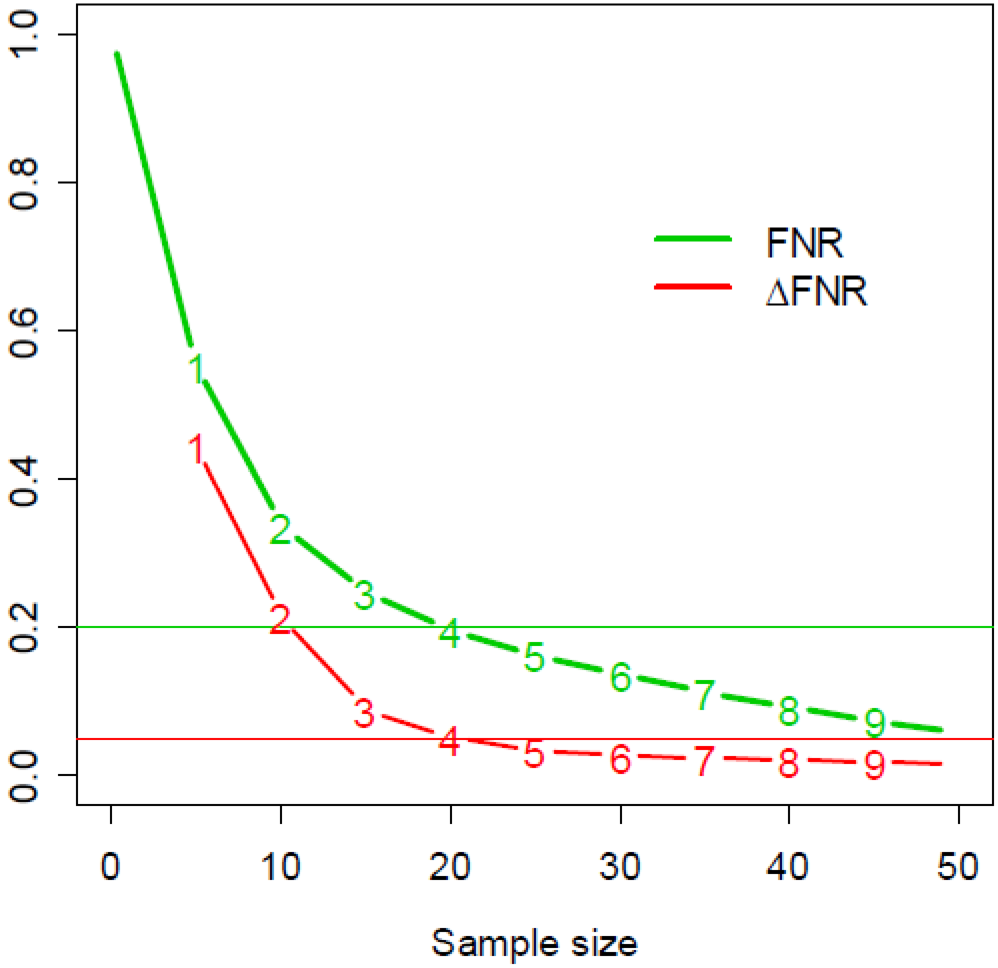
Example: m_0/m=0.9, μ/σ=0.5

True *FNR* for different sample sizes: Theoretical curve



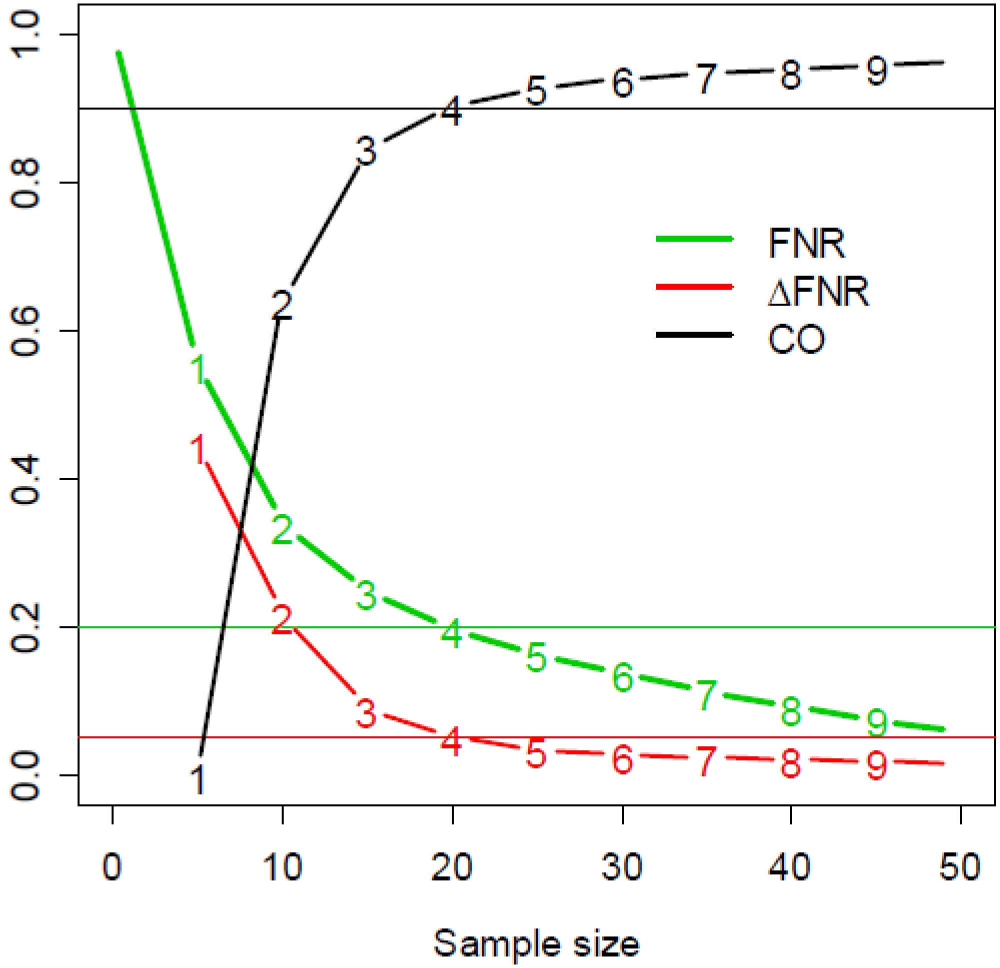
Example: m_0/m=0.9, μ/σ=0.5

True Δ*FNR* for different sample sizes: Theoretical curve



Example: m_0/m=0.9, μ/σ=0.5

True CO for different sample sizes: Theoretical curve



Simulation study (50000 runs)

The setting:

- m=5000 / 50000
- m_0/m=0.9, μ/σ=0.5
- 10 stages with stage-wise sample sizes of 5
- z-tests, α = 0.05
- Stopping rules: FNR<0.2, ΔFNR<0.05, CO>0.9, *s(m)*>9

Simulation study (50000 runs)

The setting:

- m=5000 / 50000
- m_0/m=0.9, μ/σ=0.5
- 10 stages with stage-wise sample sizes of 5
- z-tests, α = 0.05
- Stopping rules: FNR<0.2, ΔFNR<0.05, CO>0.9, *s(m)*>9

Independent data

The FDR is controlled at level α = 0.05 for the 3 considered stopping criteria.

Simulation study (50000 runs)

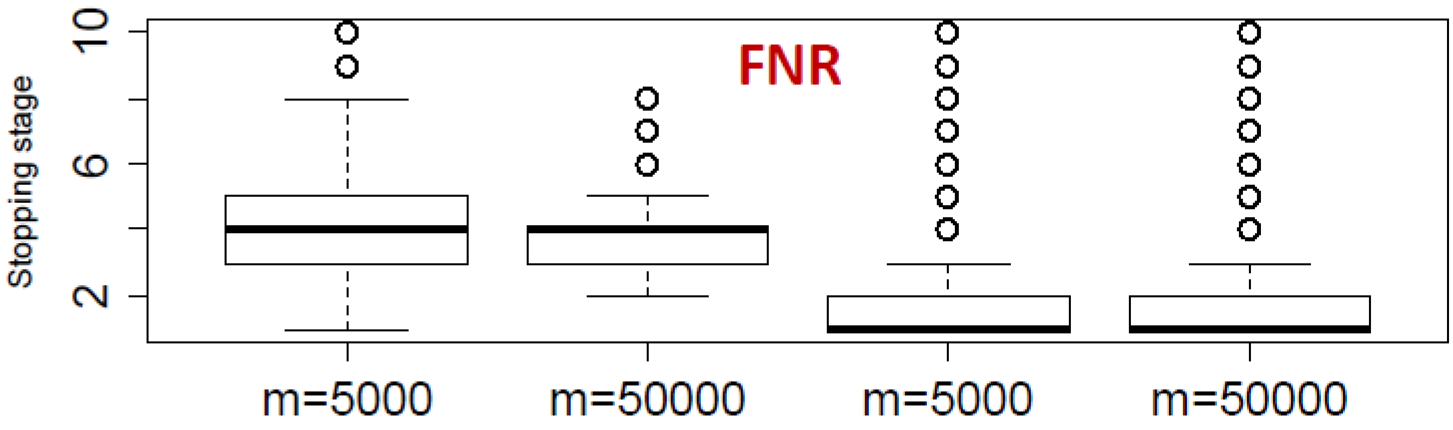
The setting:

- m=5000 / 50000
- m_0/m=0.9, μ/σ=0.5
- 10 stages with stage-wise sample sizes of 5
- z-tests, $\alpha = 0.05$
- Stopping rules: FNR<0.2, ΔFNR<0.05, CO>0.9, $s(m)$>9

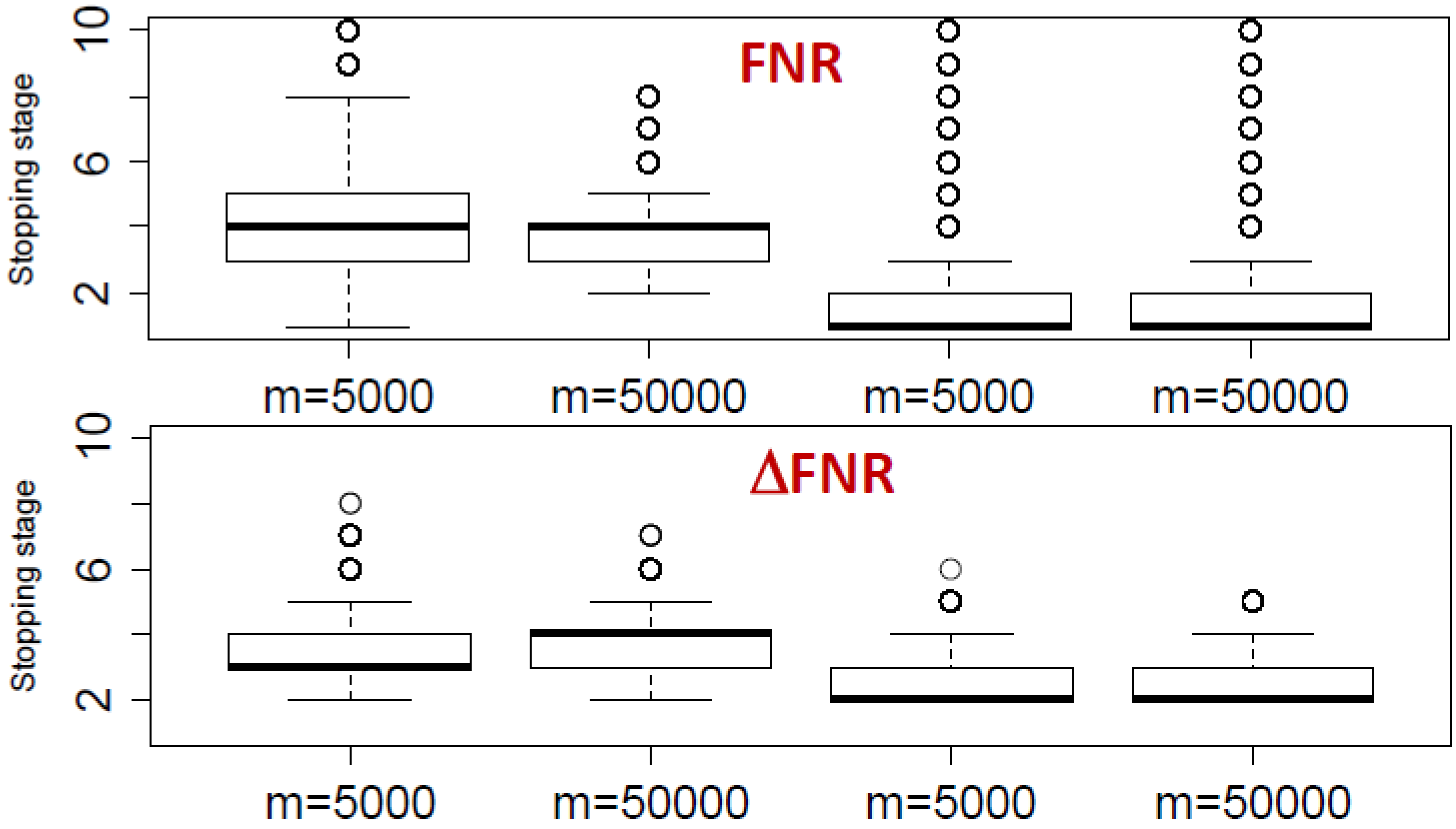
Independent data	**Equi-correlated data ($\rho = 0.5$)**
The FDR is controlled at level $\alpha = 0.05$ for the 3 considered stopping criteria.	The FDR is controlled at level $\alpha = 0.05$ for the 3 considered stopping criteria.

Independent data

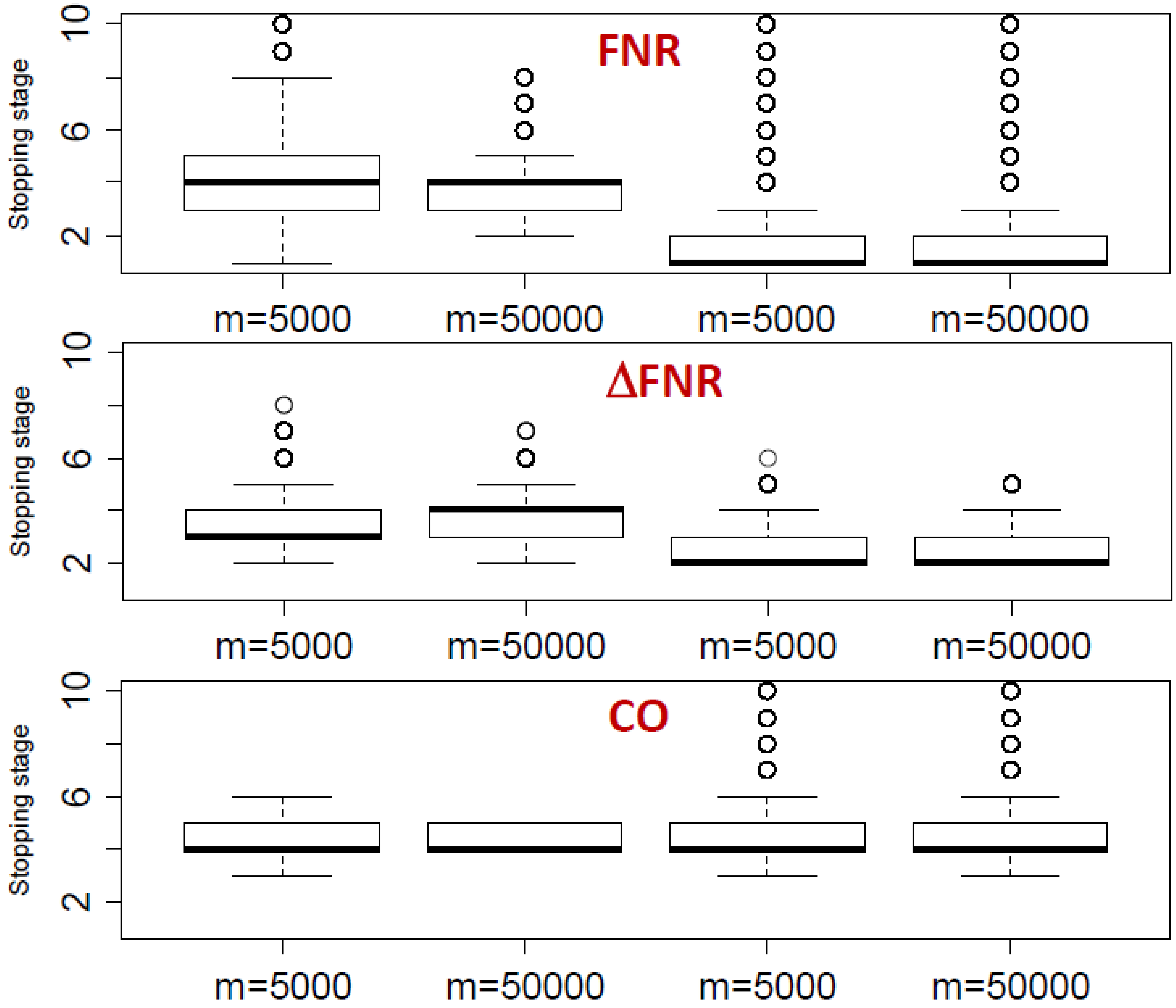
Equi-correlated data



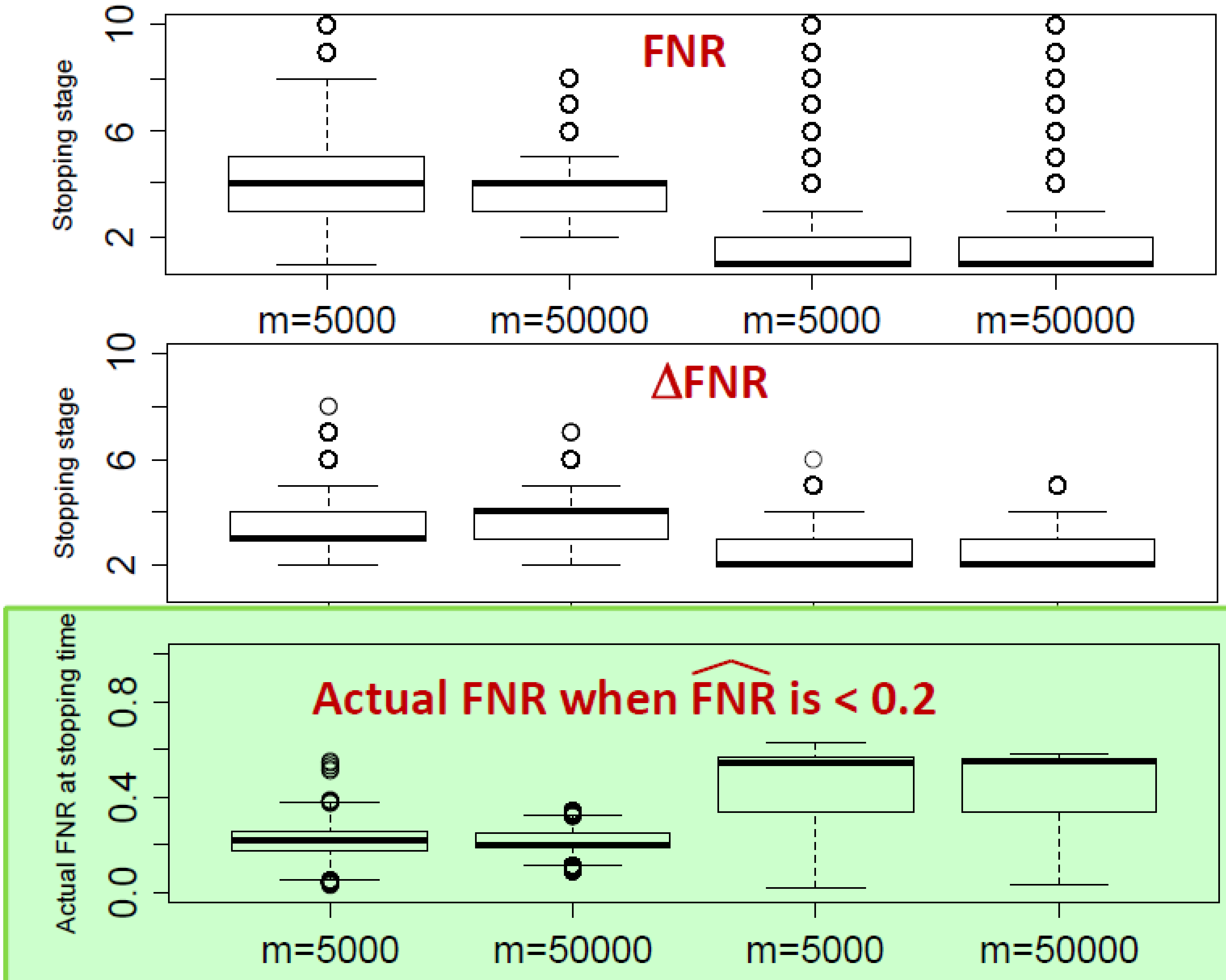
Independent data	Equi-correlated data



Independent data	Equi-correlated data



Independent data **Equi-correlated data**



The Family Wise Error Rate

- Replace the BH procedure by the Bonferroni test
- If no multiplicity adjustment for the repeated looks is applied, the FWER may be inflated (Armitage,1969)
- If stopping rules are applied, that are asymptotically deterministic, the sequential procedure controls the FWER
  - Reason: The sequential procedure degenerates to a fixed sample size procedure
- For the considered stopping rules and scenarios the FWER is controlled at level $\alpha = 0.05$.

Outlook

- Muralidharan (2010) considered an empirical bayes mixture method for effect size estimation (mean values and standard deviations)
- We try to apply the estimated values for a power estimation.

$$\text{Power}(\text{reject}\,|\,\text{effect sizes} > \Delta)$$

Discussion

Is it necessary to adjust for the number of looks?

- If the number of hypotheses is very large, multiple analyses hardly inflate the error rate.

Is this the solution to the sequential problem?

There are limitations

- Result applies only for large *m*
- Convergence rate depends on m_0/m and the alternative
- Appropriate stopping rules
- Increment - Rules seem to work better – however the performance depends on the stage-wise sample size

Selected References

- Armitage P, McPherson CK, Rowe BC (1969) J R Stat Soc Ser B.
- Benjamini Y, Hochberg Y (1995) J R Stat Soc Ser B.
- Marot G, Mayer CD (2009) SAGMB.
- Muralidharan (2010) Annals of Applied Statistics
- Pawitan et al. (2005) Bioinformatics.
- Posch M, Zehetmayer S, Bauer P (2009) Jasa.
- Storey JD, Taylor JE, Siegmund D (2004), J R Stat Soc Ser B.
- Zehetmayer S, Posch M (2010) Bioinformatics.