Johnson & Johnson
PHARMACEUTICAL RESEARCH
& DEVELOPMENT

FARMS: a probabilistic latent variable model for summarizing Affymetrix array data at probe level

Djork-Arné Clevert, Sepp Hochreiter
Institute of Bioinformatics, Johannes Kepler University Linz
Willem Talloen, An De Bond, Hinrich Göhlmann
Johnson & Johnson Pharmaceutical Research & Development, a division of
Janssen Pharmaceutica n.v., Beerse, Belgium

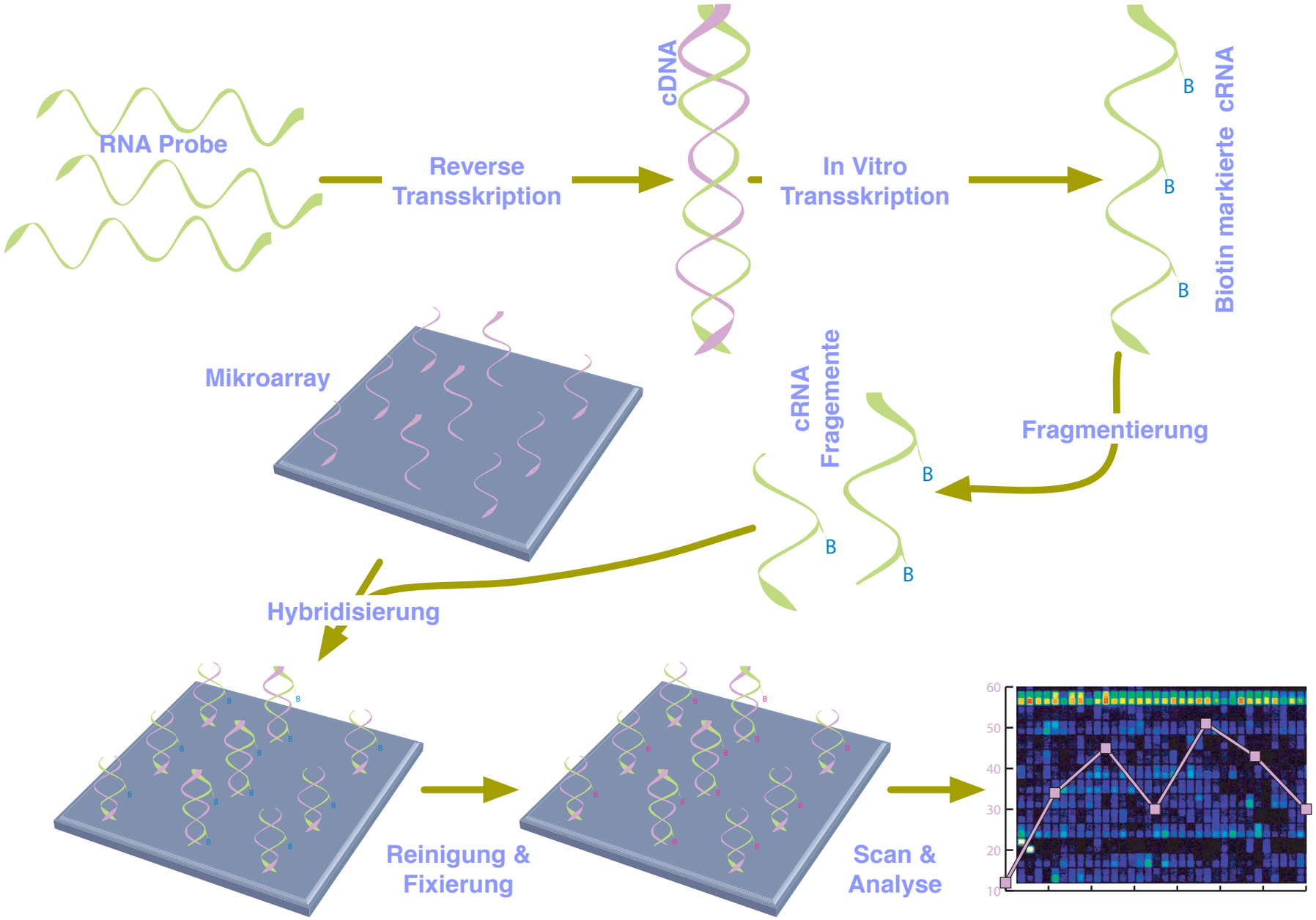
Overview

- Introduction
- Microarray technology
- Model & assumption
- Data sets & experiments
- Results
- FARMS I/NI-Calls
- Results
- Conclusion

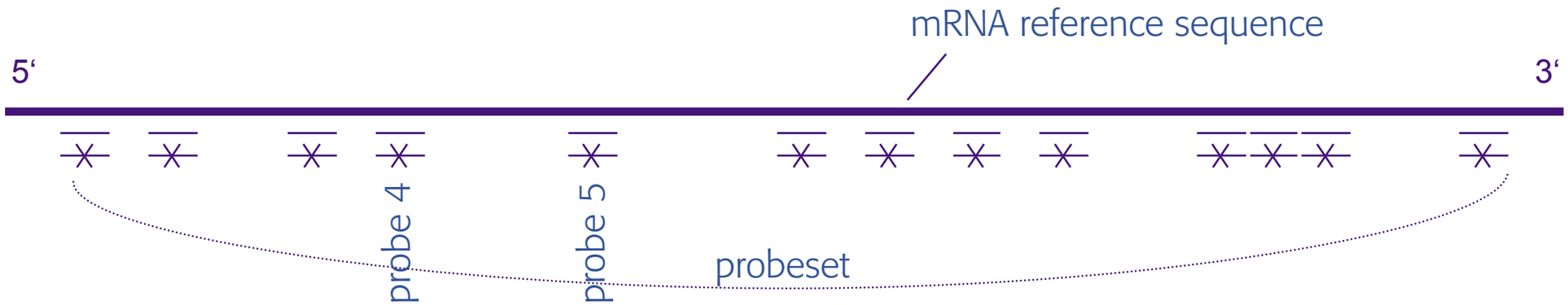
Microarrays

- Microarrays measure simultaneously cellular concentrations of thousands of mRNAs
- mRNA concentration ~ activity of a gene
- Activity of a gene = expression level
- Basis for the functional genome analysis

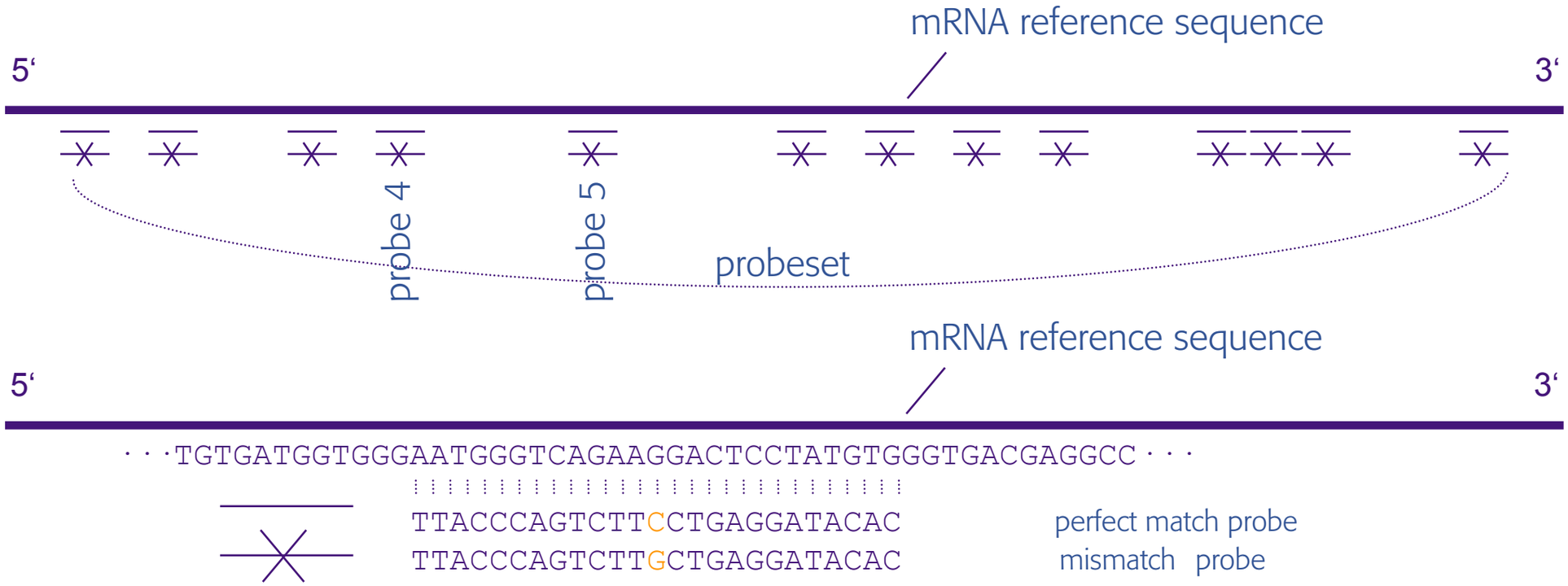
Affymetrix technology



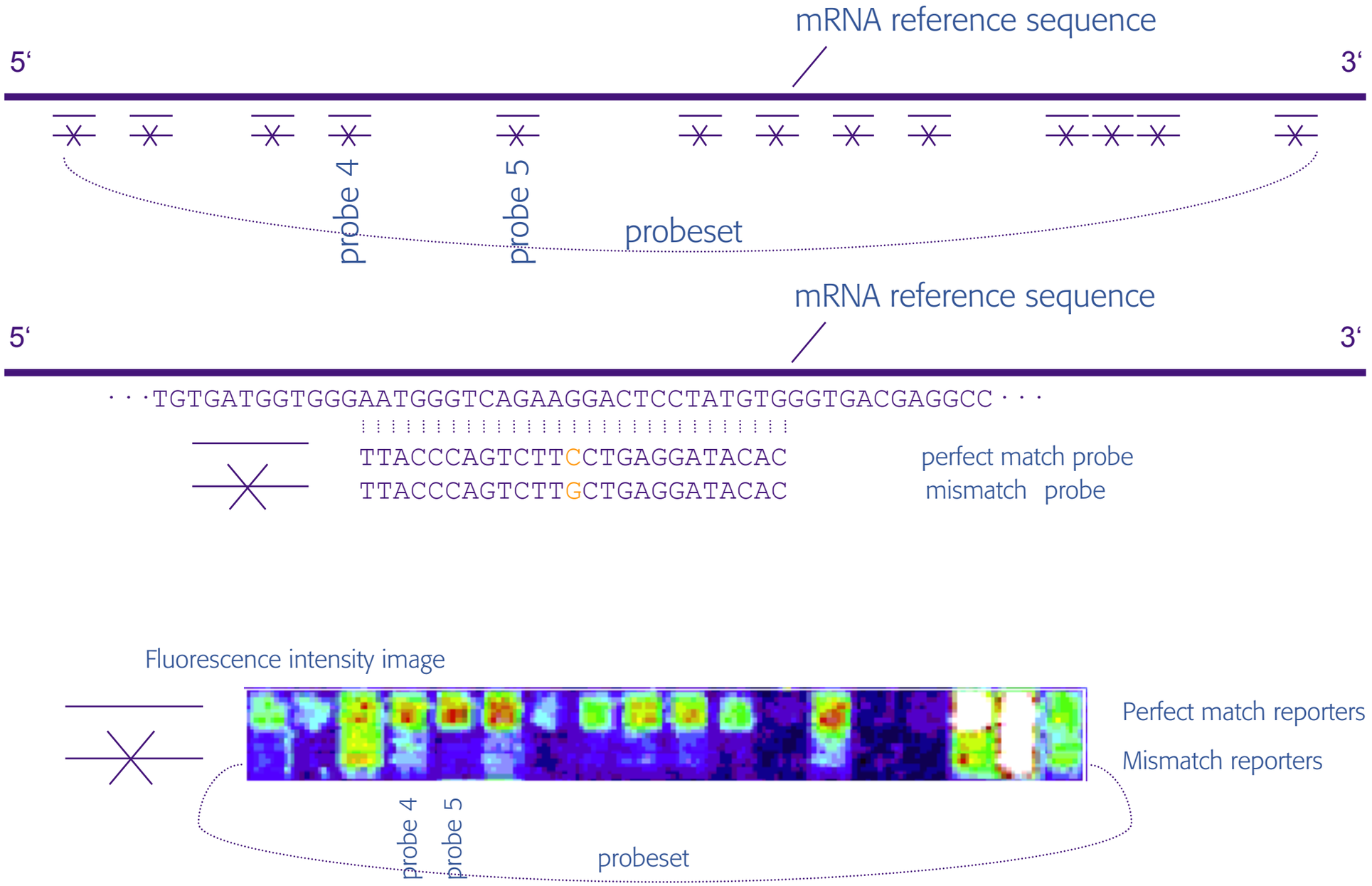
Microarray design



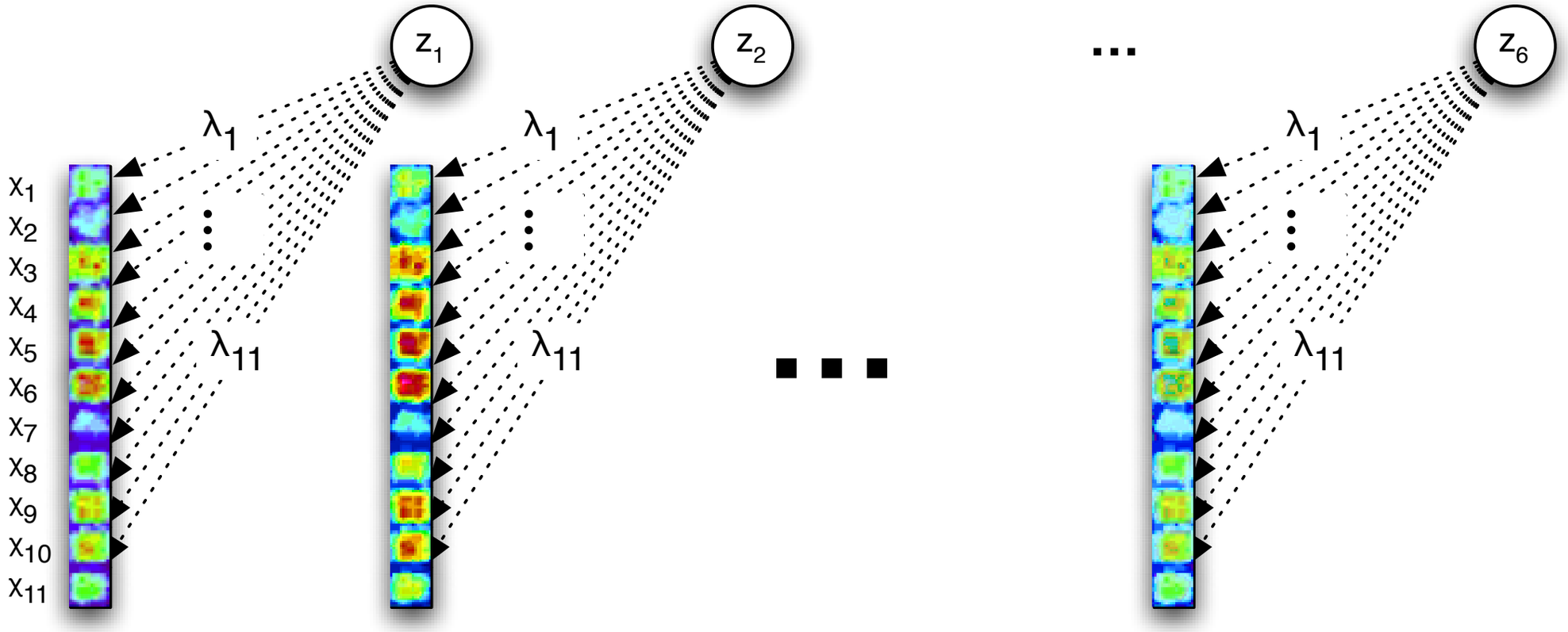
Microarray design



Microarray design



Example:
one PM-probe set and six arrays



$$x = \lambda z + \epsilon$$

Factor analysis

- Generative model:

$$\boldsymbol{x} = \boldsymbol{\lambda} z \ + \ \boldsymbol{\epsilon}$$

where

$$\boldsymbol{x}, \boldsymbol{\lambda} \ \in \ \mathbb{R}^n, \ z \sim \mathcal{N}(0,1), \ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$$

From this it follows that:

$$\boldsymbol{x} \sim \mathcal{N}\left(\mathbf{0} \ , \ \boldsymbol{\lambda}\boldsymbol{\lambda}^T \ + \ \boldsymbol{\Psi}\right)$$

  - parameter estimation with EM-algorithm
- z models the correlation between the data elements
- $\boldsymbol{\epsilon}$ accounts for the independent noise in the data

Prior knowledge

- Increasing mRNA concentration leads to a larger signals
  - ➡ negative values of λ are not plausible
- Observed variance in the data is often low
  - ➡ high values of λ are unlikely
- Most genes from a chip are non-relevant
  - ➡ most genes with a $\lambda \approx$ zero

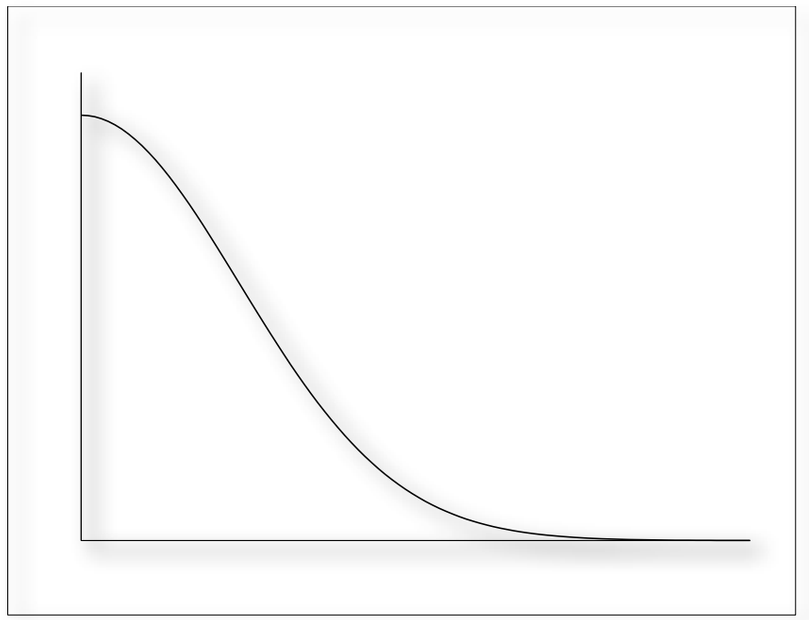
Bayesian posterior & prior

- Bayesian posterior:

$$p(\boldsymbol{\lambda}, \boldsymbol{\Psi} \mid \{\boldsymbol{x}\}) \;\propto\; p(\{\boldsymbol{x}\} \mid \boldsymbol{\lambda}, \boldsymbol{\Psi})\; p(\boldsymbol{\lambda}, \boldsymbol{\Psi})$$

- Prior distribution:
  - rectified Gaussian

$$p(\boldsymbol{\lambda}, \boldsymbol{\Psi}) \;=\; p(\boldsymbol{\lambda})$$

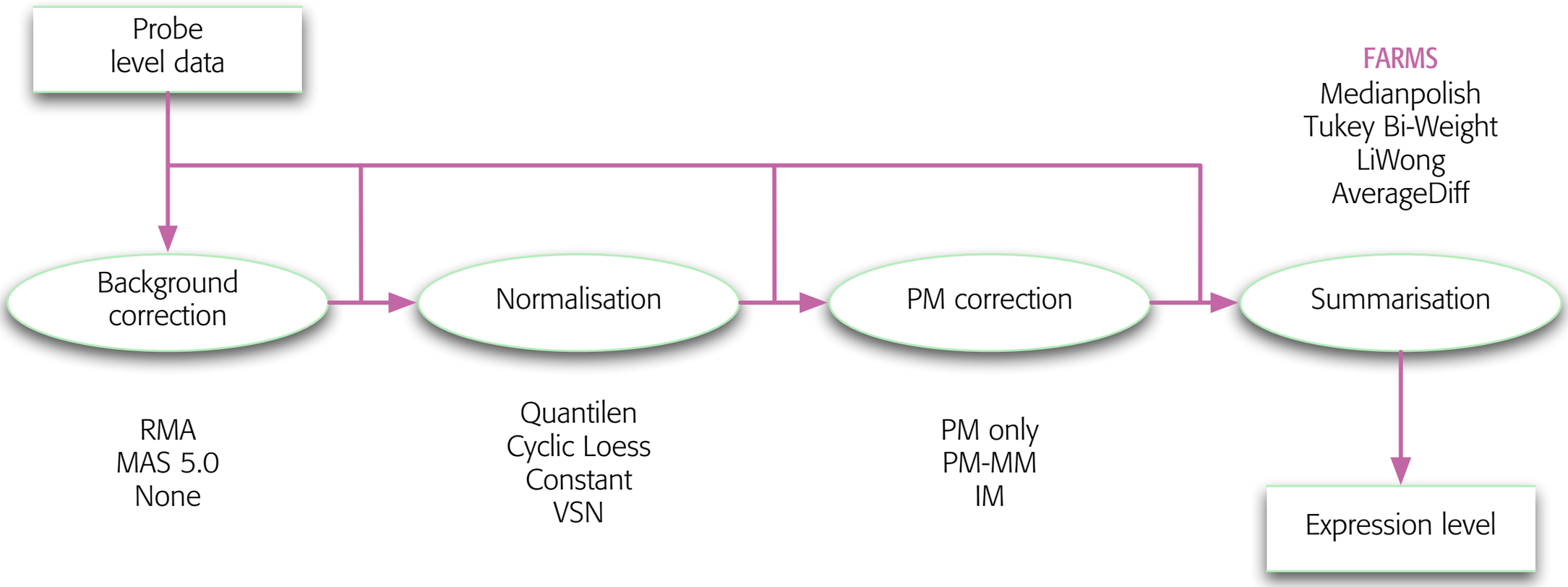


Data sets

- Affymetrix spiked-in data set „A“
  - 59 arrays HGU95A_v2
    - 14 artificially entered cDNA fragments
    - 0, 0.25, 0.5, 1, 2, 4, 8, … , 1024 pM

- Affymetrix spiked-in data set „B“
  - 42 arrays HGU133A
    - 42 artificially entered cDNA fragments
    - 0, 0.0125, 0.25, 0.5, 1, … , 512 pM

Preprocessing chain



Results

Affycomp II Benchmark (AUC - area under the curve):

	Intensity	FARMS	RMA	GCRMA	MAS 5.0	MBEI
HGU133	Low	**0.94**	0.51	0.62	0.07	0.21
	Med	**0.99**	0.91	0.94	0.00	0.43
	High	**1.00**	0.64	0.59	0.00	0.16
	Mean	**0.95**	0.60	0.69	0.05	0.26
HGU95	Low	**0.91**	0.57	0.45	0.09	-
	Med	**1.00**	0.91	0.91	0.00	-
	High	**0.98**	0.96	0.92	0.00	-
	Mean	**0.93**	0.65	0.57	0.06	-

Computational costs for processing 60 arrays:

	FARMS	RMA	MAS 5.0	MBEI
Computational time [s]	**92**	384	851	591

Analysis of microarray data

- Problem of multiple testing and over-fitting
  - Because of the high dimensionality of data
  - Because of the technology (noise)
  - Because of the biology, most genes are non-informative

➡ Informative pre-filtering is desired

- Using array information to filter genes
  - A/P calls: excluding probe sets that are always absent

Internal consistency

- The correlation of intensities between probes of the same probe set across chips
  - When intensities are high or low for all probes in an individual chip there needs to be a strong correlation
- Strong correlation → consistency
  - This means that all fragments of a gene tell the same story

Internal consistency

Informative gene

Dots represent individual chips

Non-informative gene

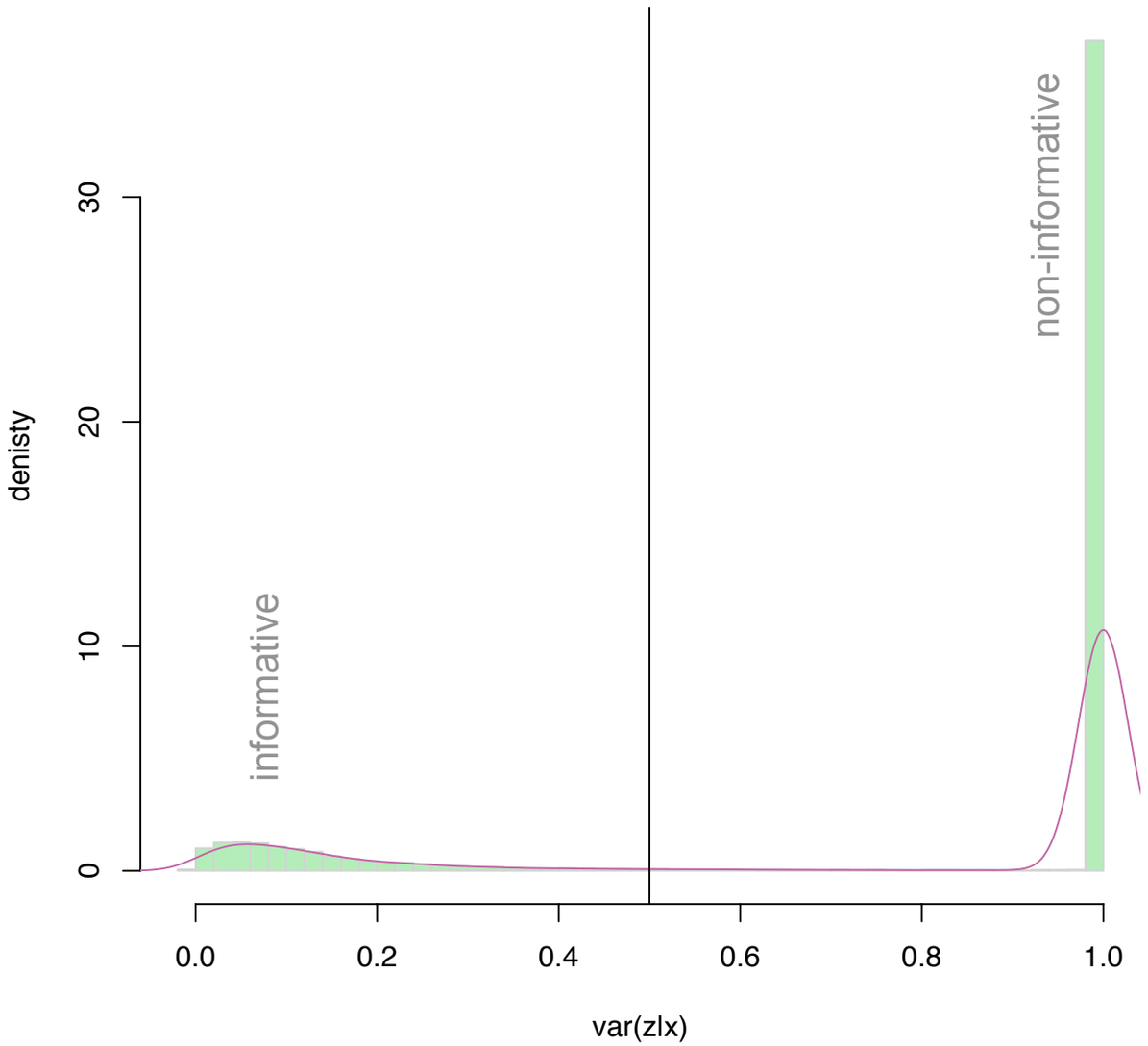
Background: I/NI-call

- Variance of the extracted factor z given the data:

$$\mathrm{var}\left(z \mid \boldsymbol{x}\right) = \left(1 + \boldsymbol{\lambda}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\lambda}\right)^{-1}$$

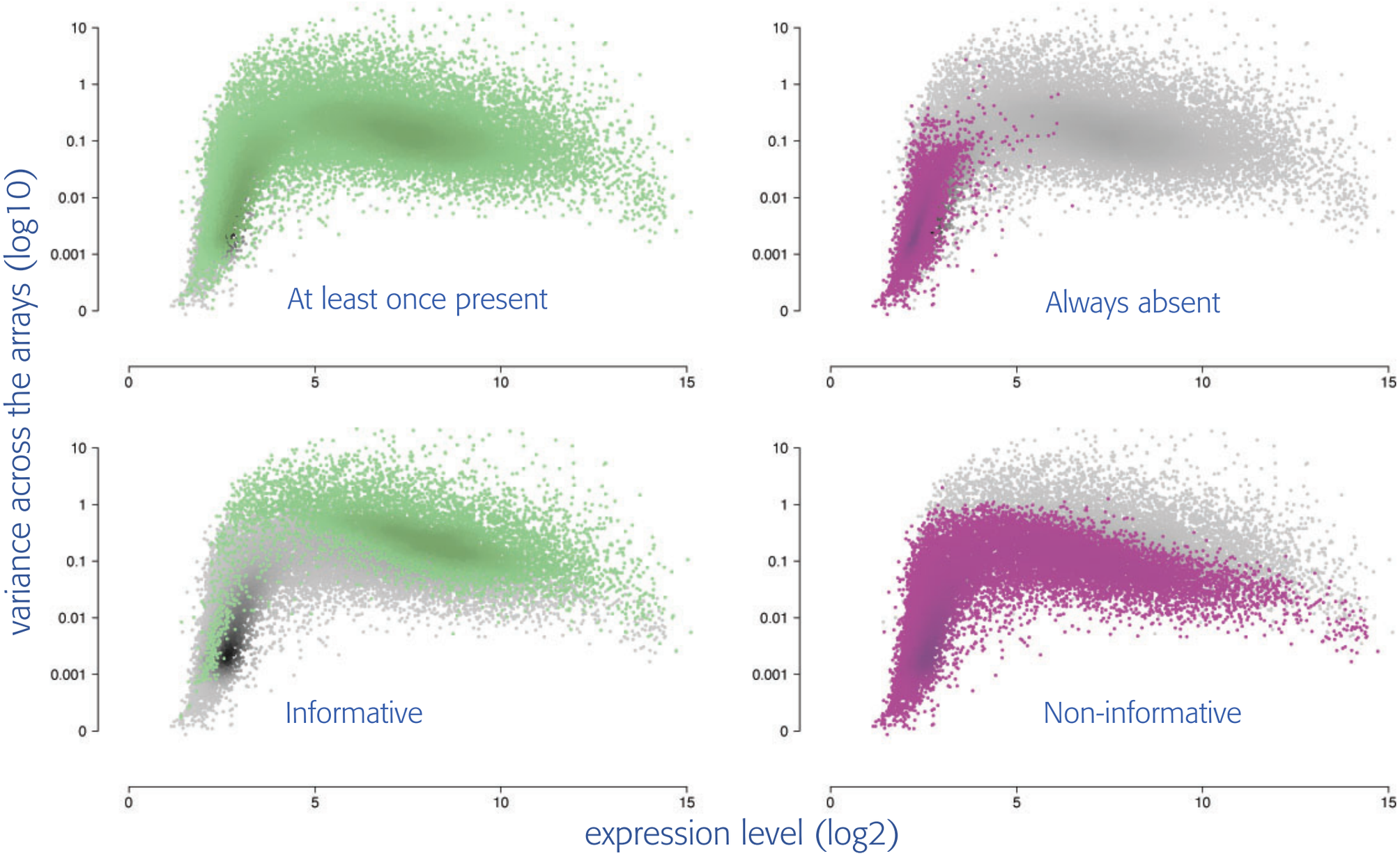
  - provides a measure of how much variation in the probe set data x is explained by the factor z
  - value between [0-1]
    - var(z|x) = 0 → data can be completely explained by z
    - var(z|x) = 1 → data cannot be explained by z
    - var(z|x) = 0.5 → signal-to-noise-ratio = 1
  - criterion for unsupervised feature selection

I/NI-calls in action



- clear bimodal distribution of var(z|x)
- distinct modes for Non-Inf. and Inf. genes

I/NI-calls vs. A/P-calls



Results I/NI-calls

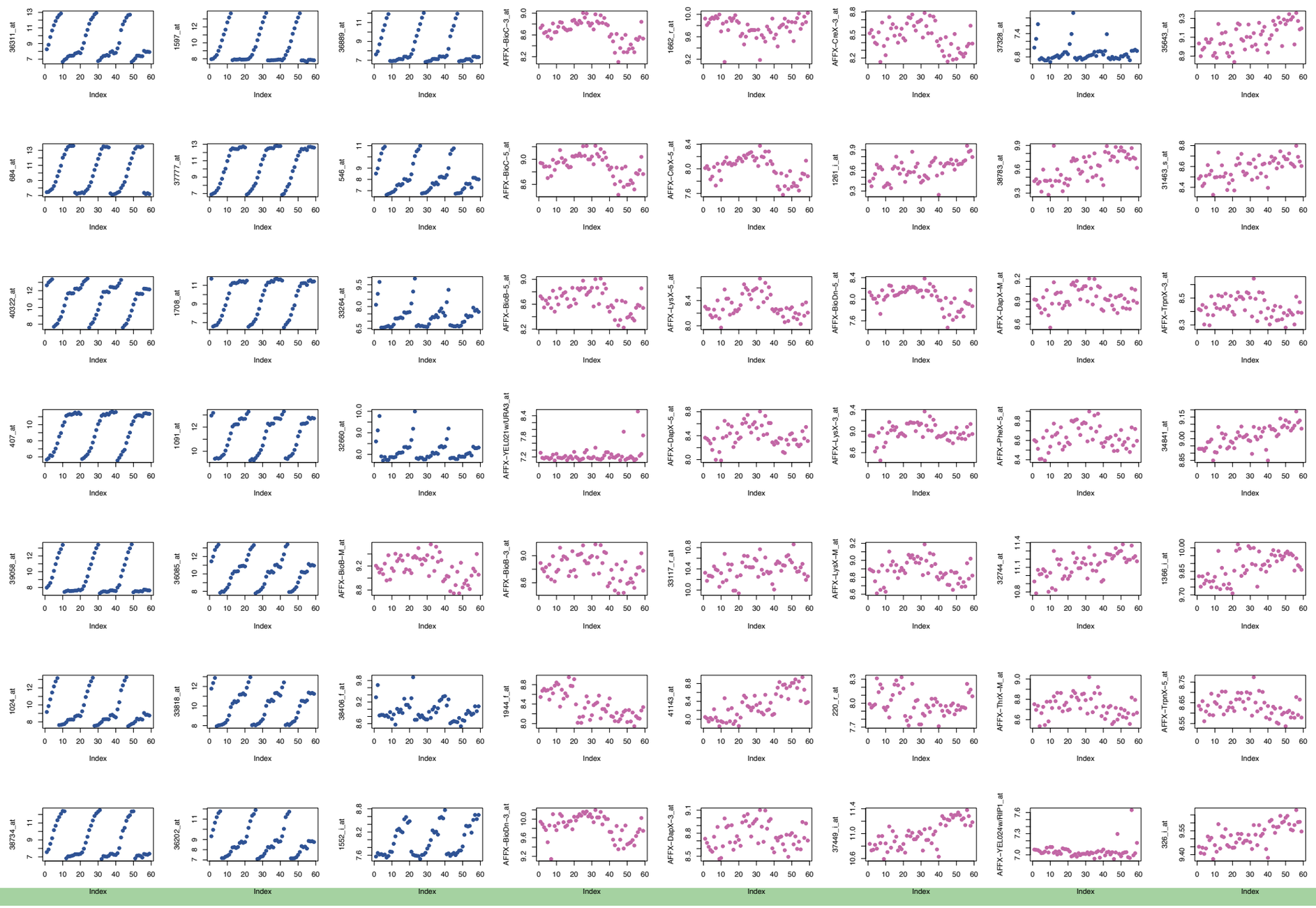
- On average: 84 (±1.5)% exclusion rate
  - applied on 30 real life studies
  - A/P calls excluded only 33 (±1)%
- Validation on spiked-in data

	Informative	Non-informative	Exclusion rate	Detected Spiked-ins	Detected Pseudo Spiked-ins
HGU133A	81	22219	99.63%	42/42	28/28*
HGU95_V2	56	12570	99.56%	14/14	5/5**

* McGee et al. 2006
** Wolfinger and Chu 2002; Cope et al. 2004

81probe sets with I/NI-call

56 probe sets with I/NI-call



Conclusion

- FARMS summarization outperforms all Affycomp II competitors (57) in terms of sensitivity and specificity (AUC)
- I/NI calls offers a critical contribution to the curse of high-dimensionality in the analysis of microarray data
  - I/NI calls filters informative genes in a statistically sound and objective manner
  - The smaller gene set contains less false positives

Further information

- Talloen W, Clevert DA, Hochreiter S, Amaratunga D, Bijnens L, Kass S and Göhlmann H: **I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data.** Bioinformatics 2007 Advance Access published on October 5, 2007.
- Hochreiter S, Clevert DA, Obermayer K: **A new summarization method for Affymetrix probe level data**. Bioinformatics 2006, 22: 943-949.
- FARMS homepage:
  - http://www.bioinf.jku.at/software/farms/farms.html
- Affycomp II benchmark:
  - http://affycomp.biostat.jhsph.edu/