Exploiting the EMERALD mixture design for model based microarray platform comparisons by Bayesian inference of technical and biological variance components

Thomas Tüchler[1] Florian Klinglmüller[2] Peter Sykacek[1]
David P. Kreil[1]

[1] WWTF Chair of Bioinformatics, BOKU University, 1190 Vienna, Austria.
[2] Medical Statistics and Informatics, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria

5 December 2008

Evaluation approaches

- Spike-based approaches
  Knowing concentration of particular RNA species
- Non spike-based approaches
  Knowing features about whole samples

Evaluation approaches

- Spike-based approaches
  Knowing concentration of particular RNA species
- Non spike-based approaches
  Knowing features about whole samples

$\longrightarrow$ Ability to detect subtle biological differences

The EMERALD dataset

- 6 rats (genetically different)
- 4 titrations of two tissues
  (liver to kidney 4:0, 3:1, 1:3, 0:4)
- 3 technical replicates per sample
- 3 Platforms
  (Affymetrix, Agilent, Illumina)
- 216 microarrays

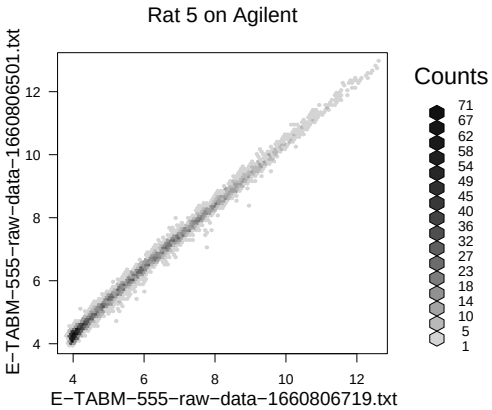
→ Biological *vs.* technical variance

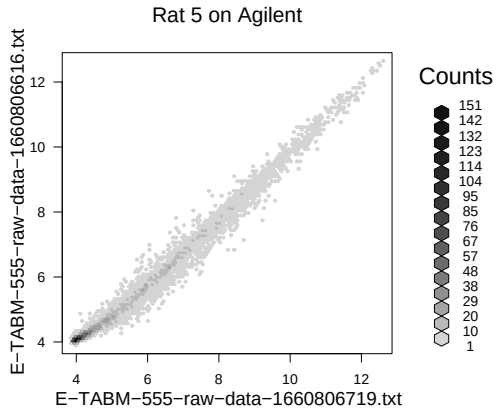
Outline

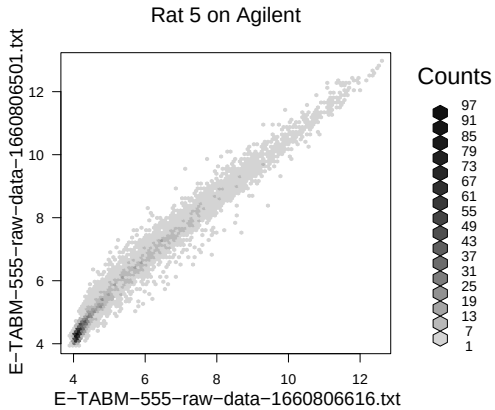
- Preprocessing
  - Comparable measurements
  - Clean data
- Modelling
  - Exploiting mixture design
- Investigating
  - Biological *vs.* technical variance
  - Impact of signal intensity and Normalization

Exploring the data

All platforms are affected by outlier slides



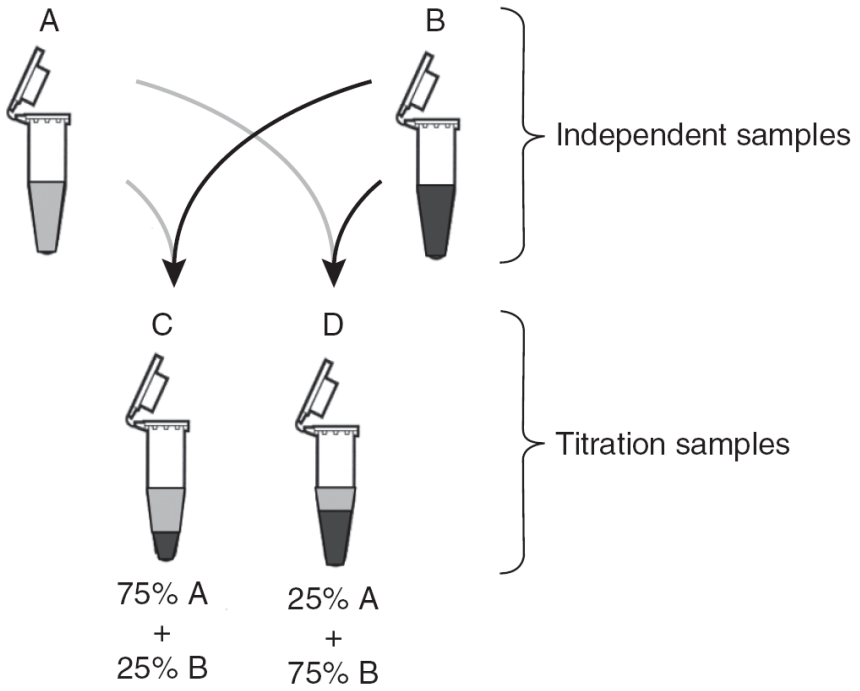




- Affymetrix: 3B-3
- Agilent: third replicate series (hybridization names '. . .-3') conducted by operator 'A' (high 'AmpLabelingInputMass')
- Illumina: low cRNA yield (3 of 6), plate location 3 (4 of 6) and hybridization date 10/04/08 (5 of 6)

Modelling the mixture factor

Mixing **total RNA**



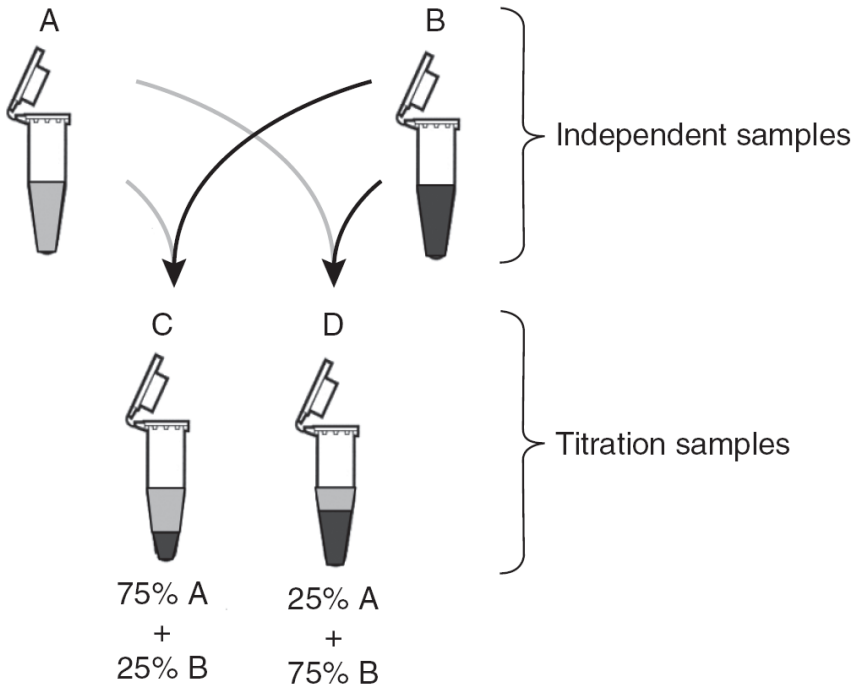
(from Shippy, 2006)

Measuring **mRNA**

$$
\begin{aligned}
C_{\text{mRNA}} &= 0.75 \cdot a \cdot A_{\text{totRNA}} + \\
&\quad 0.25 \cdot b \cdot B_{\text{totRNA}} \\
\rho &= b/a
\end{aligned}
$$

Modelling the mixture factor

Mixing **total RNA**



(from Shippy, 2006)

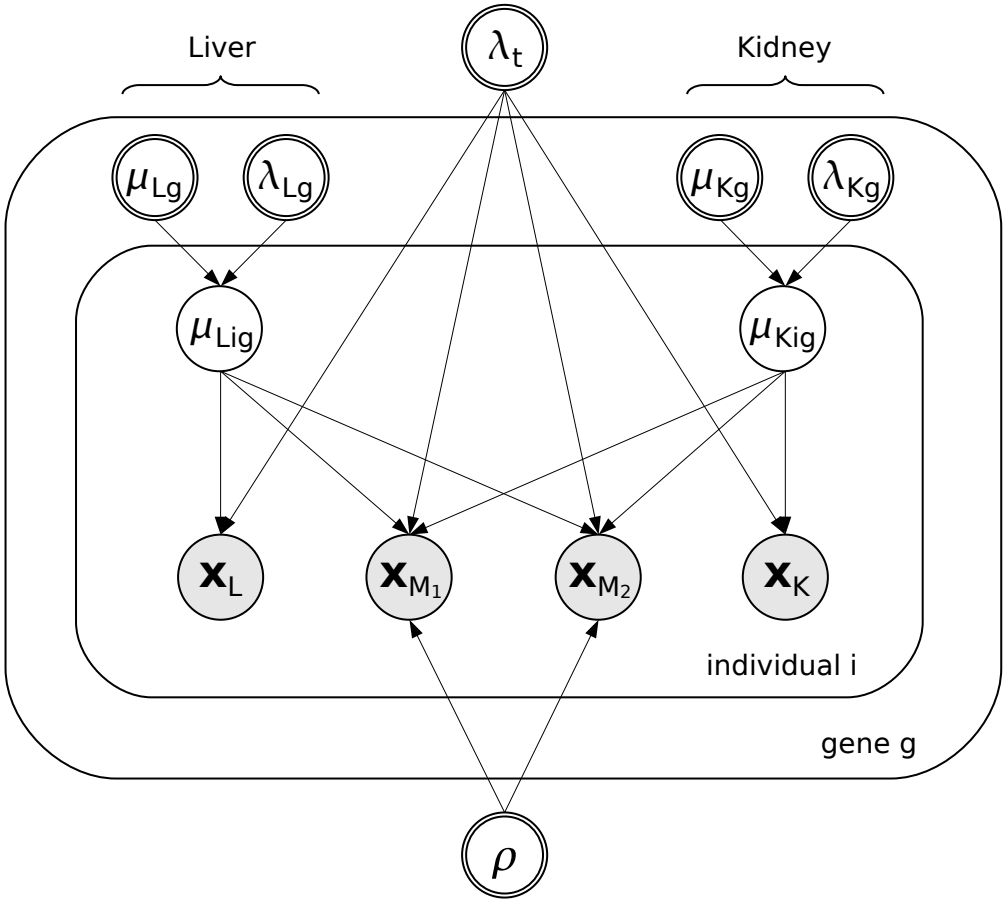
Measuring **mRNA**

$$
\begin{aligned}
C_{\text{mRNA}} &= 0.75 \cdot a \cdot A_{\text{totRNA}} + \\
&\quad 0.25 \cdot b \cdot B_{\text{totRNA}} \\
\rho &= b/a
\end{aligned}
$$

$\rho = 2/3$	A	B	C	D
Liver	100	82	33	0
Kidney	0	18	67	100

$\rightarrow$ Titration $\neq$ Mixture ratio

Modelling the EMERALD data



Directed Acyclic Graph (DAG)

$$
\begin{aligned}
\mu_{Lig} &\sim \mathcal{N}\left(\mu_{Lg}, \lambda_{Lg}^{-1}\right) \\
\mu_{Kig} &\sim \mathcal{N}\left(\mu_{Kg}, \lambda_{Kg}^{-1}\right) \\
\mathbf{x}_{Lig} &\sim \mathcal{N}\left(\mu_{Lig}, \lambda_t^{-1}\right) \\
\mathbf{x}_{Kig} &\sim \mathcal{N}\left(\mu_{Kig}, \lambda_t^{-1}\right) \\
\mathbf{m}_{ig} &= f\left(\mu_{Lig}, \mu_{Kig}, \rho\right) \\
\mathbf{x}_{M_1 ig} &\sim \mathcal{N}\left(m_{1ig}, \lambda_t^{-1}\right) \\
\mathbf{x}_{M_2 ig} &\sim \mathcal{N}\left(m_{2ig}, \lambda_t^{-1}\right)
\end{aligned}
$$

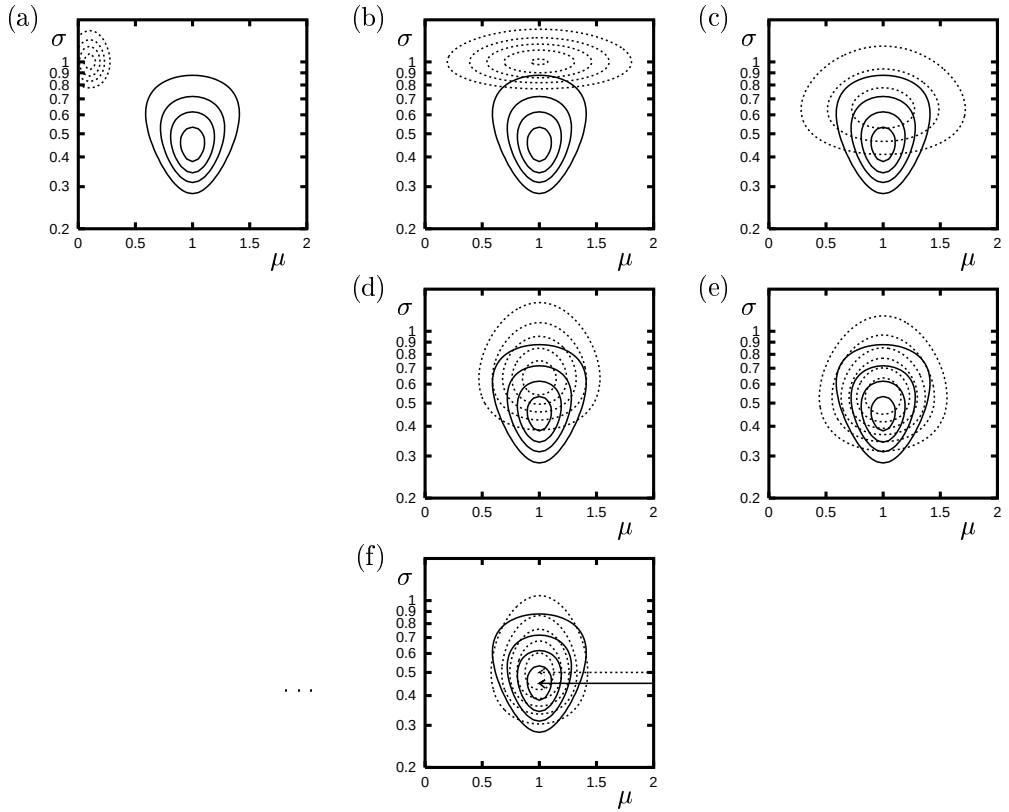
Inference using Variational Bayes

Approximate true posterior $\mathrm{p}(\theta|x)$ by factorized $Q(\theta)$:

$$Q(\theta) = \prod_i Q_i(\theta_i)$$

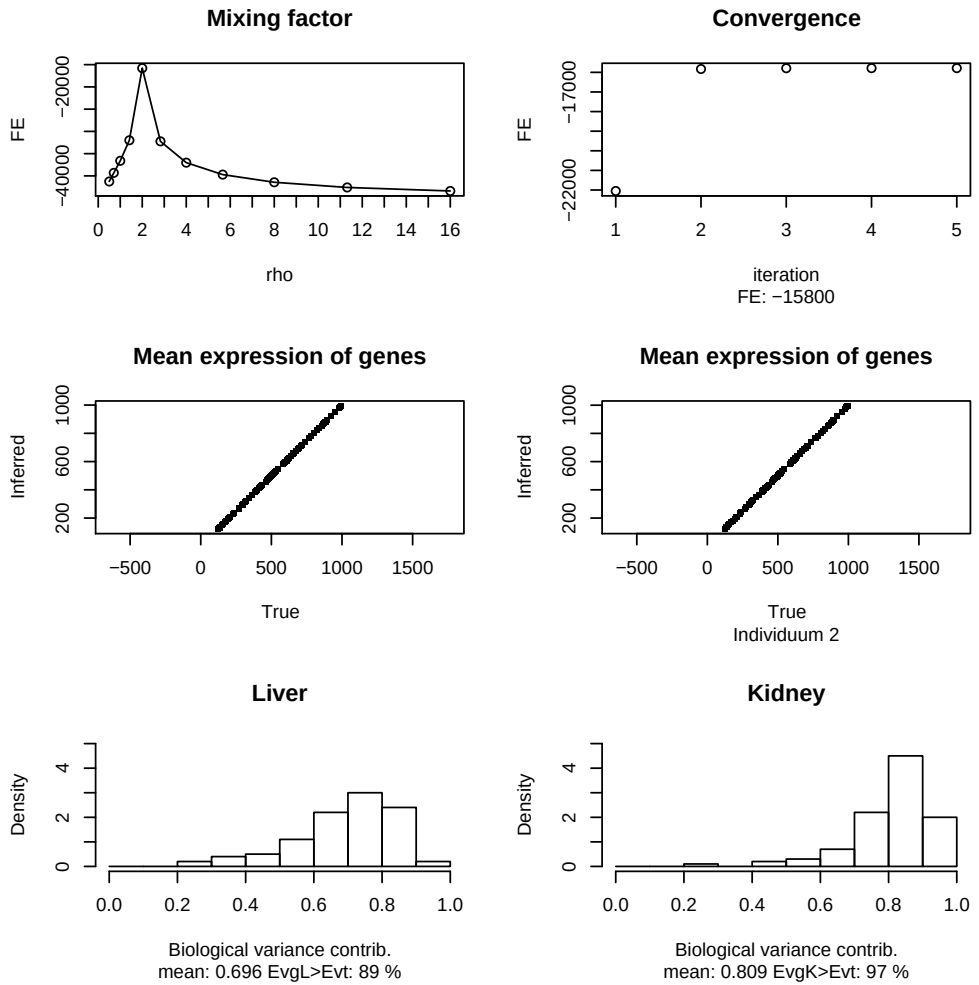
Optimizing *free energy* between $\mathrm{p}(\theta|x)$ and $Q(\theta)$:

$$FE = -\int Q(\theta) \ln \frac{Q(\theta)}{\mathrm{p}(x,\theta)} \, d\theta$$



Updates for μ and σ of a Gaussian (from MacKay, 2003)

Validating the implementation



Simulation validating retrieval and conv.

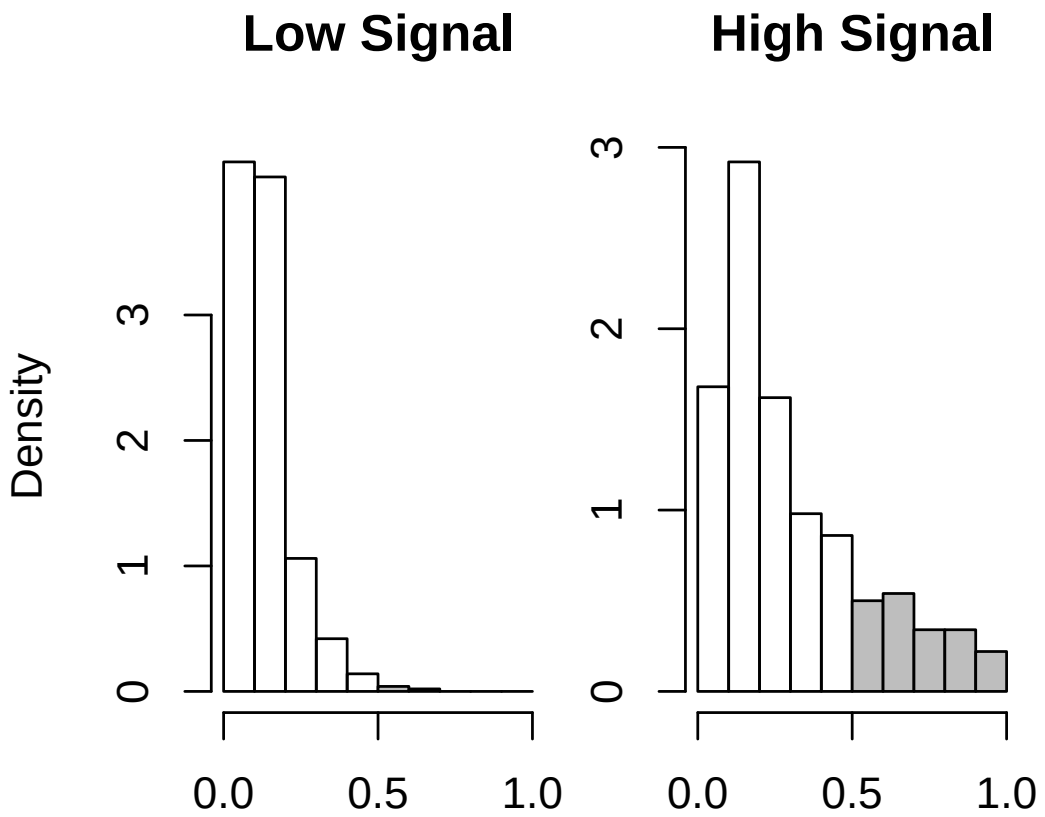
Pros

- Full posterior distributions
- Fast convergence
- Monitor convergence
- Model comparison yard stick

Cons

- Approximation

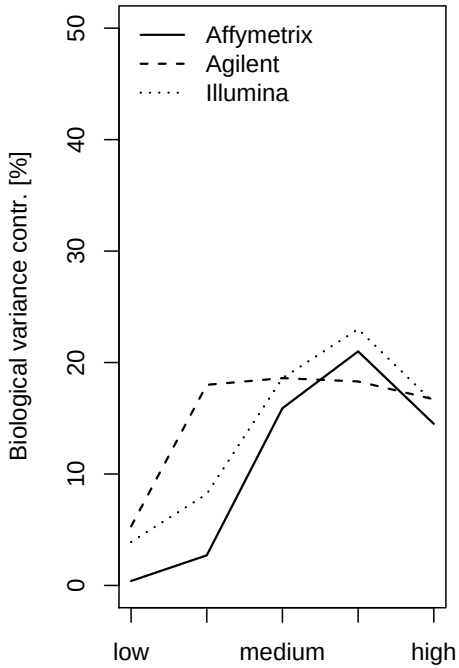
Results within platform and intensity



$$0 < \frac{\text{var}_{bio}}{\text{var}_{bio} + \text{var}_{tec}} < 1$$

- Biological variance detectable
- Biological $<$ technical variance
- More mRNA in kidney samples; $\rho > 1$ (*cf.* Liggett, 2008)

Results across platforms and intensities

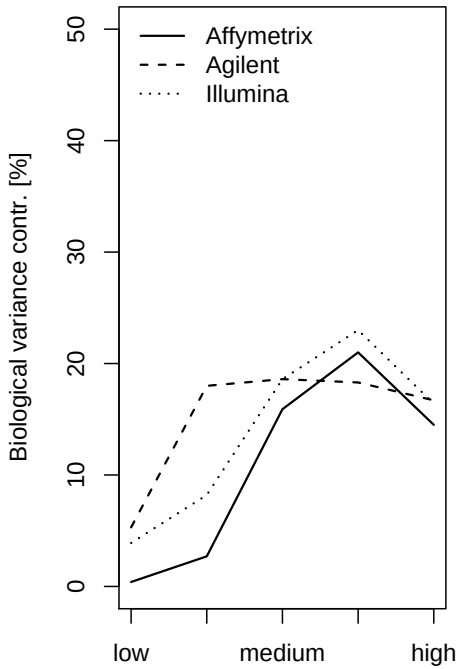


Baseline

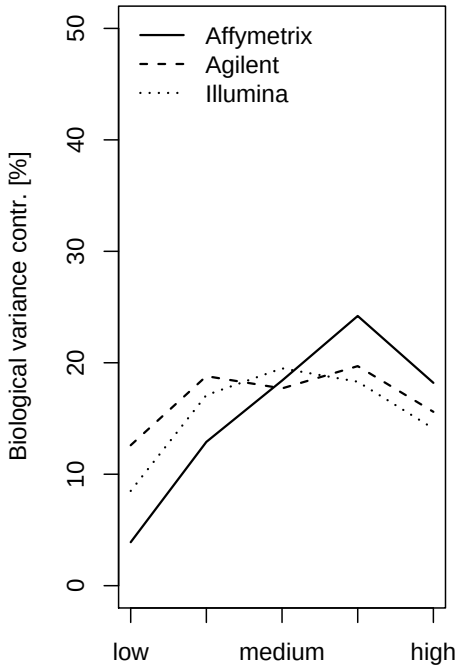
Percentage of genes with biological $>$ technical variance

- Platform and intensity dependent
- Normalization dependent

Results across platforms and intensities



Baseline

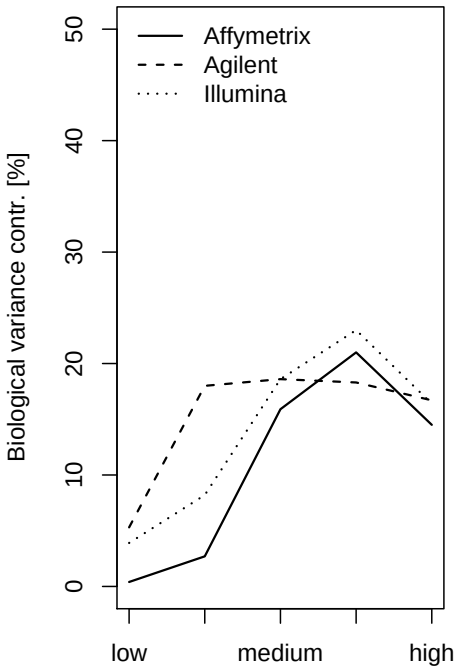


Quantile

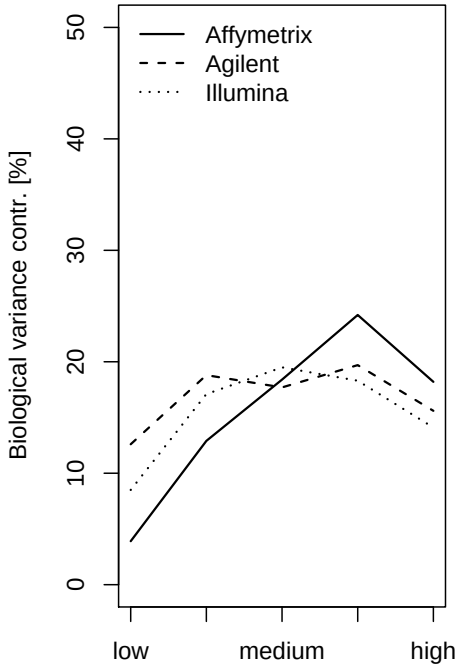
Percentage of genes with biological $>$ technical variance

- Platform and intensity dependent
- Normalization dependent

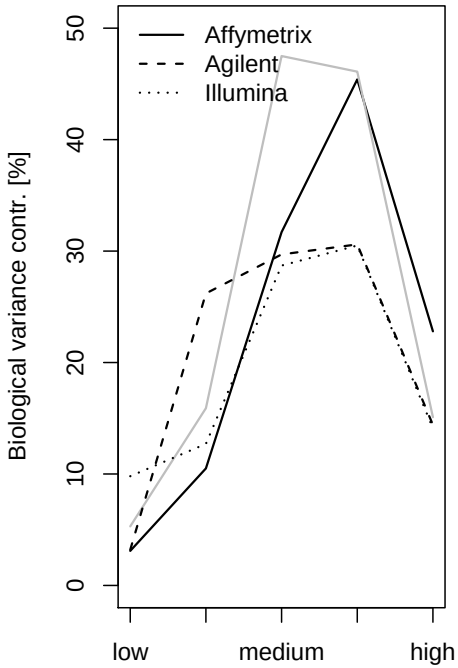
Results across platforms and intensities



Baseline



Quantile



VSN

Percentage of genes with biological $>$ technical variance

- Platform and intensity dependent
- Normalization dependent

Conclusions and Outlook

Methods

- Detecting biological differences as performance measure
- VB model exploiting mixture design

Conclusions and Outlook

Methods

- Detecting biological differences as performance measure
- VB model exploiting mixture design

  $\longrightarrow$ Efficient Bayesian inference for genome size data

Conclusions and Outlook

Methods

- Detecting biological differences as performance measure
- VB model exploiting mixture design
  $\longrightarrow$ Efficient Bayesian inference for genome size data

EMERALD dataset

- Biological variance detectable
- Platform and intensity dependence

Conclusions and Outlook

Methods

- Detecting biological differences as performance measure
- VB model exploiting mixture design
  $\longrightarrow$ Efficient Bayesian inference for genome size data

EMERALD dataset

- Biological variance detectable
- Platform and intensity dependence

Wiener Wissenschafts-, Forschungs- und Technologiefonds
GEN-AU GENOME RESEARCH IN AUSTRIA
AUSTRIAN RESEARCH CENTERS
Baxter
ACBT
BOKU