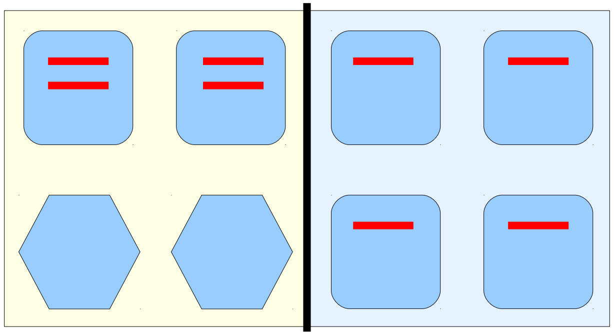
Workshop Linköping, 2011-12-05, Can Systems Biology Aid Personalized Medication?

Screening for molecular signatures in heterogeneous tissue and in pooled samples



Dirk Repsilber

FBN Dummerstorf, Germany

Bioinformatics/Biomathematics @ Genetics and Biometry

Screening for molecular signatures in heterogeneous tissue and in pooled samples

biomarker – definition

- “characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention” (1)
- measurable & differentially regulated ?!

(1) Biomarkers definitions Workgroup, Clin. Pharmacol. Ther. 69, 2001

biomarker – definition

- “characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention” (1)

- measurable & differentially regulated ?!

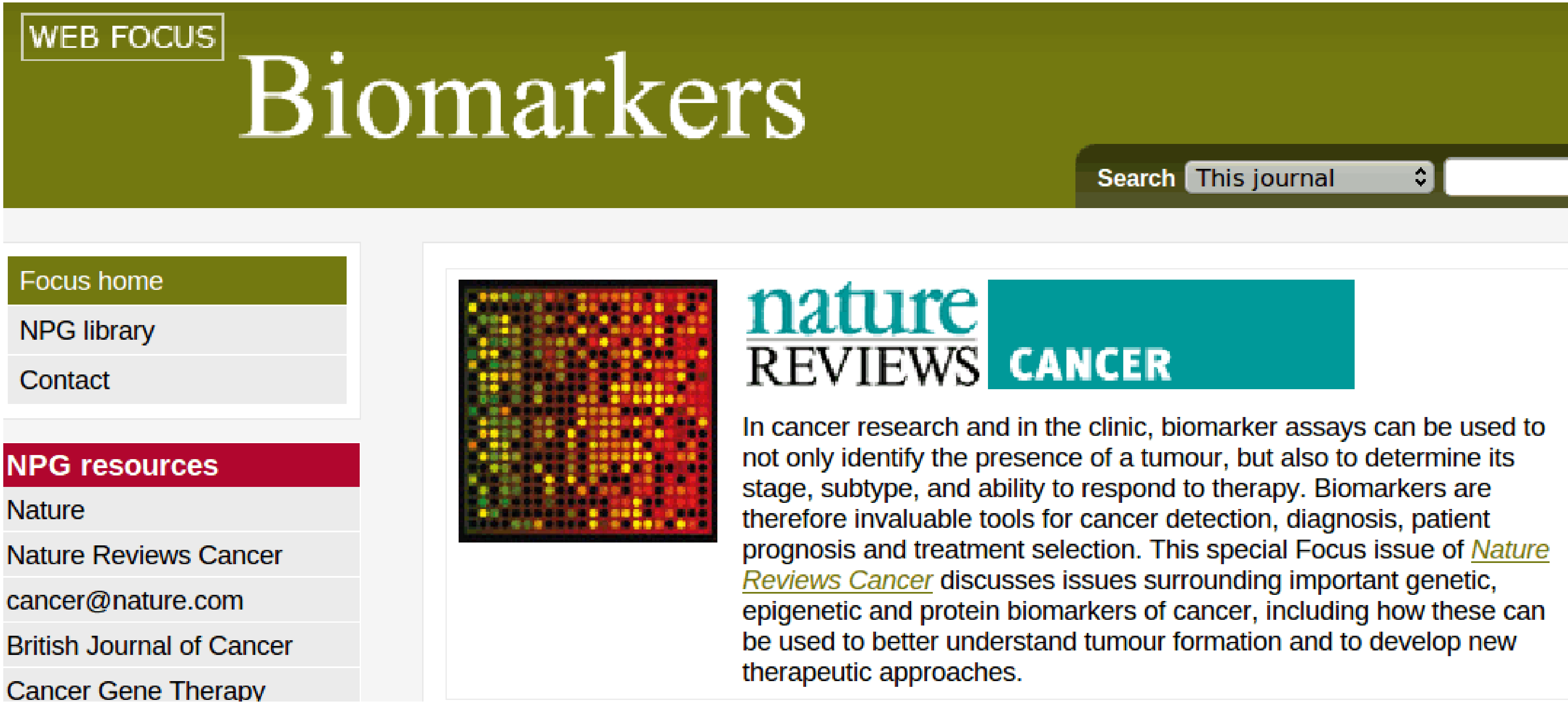
  + valid (defined end-point & study population) (2)

  + reproducible, accurate and unbiased

  + generalizable to new samples

  + easy accessible samples (e.g. blood)

(1) Biomarkers definitions Workgroup, Clin. Pharmacol. Ther. 69, 2001
(2) Wacholder, S. et al., Am J Epidemiol 135, 1992



WEB FOCUS

Biomarkers

Search This journal

- Focus home
- NPG library
- Contact

NPG resources

- Nature
- Nature Reviews Cancer
- cancer@nature.com
- British Journal of Cancer
- Cancer Gene Therapy

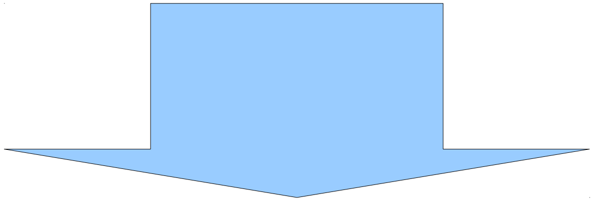
nature REVIEWS CANCER

In cancer research and in the clinic, biomarker assays can be used to not only identify the presence of a tumour, but also to determine its stage, subtype, and ability to respond to therapy. Biomarkers are therefore invaluable tools for cancer detection, diagnosis, patient prognosis and treatment selection. This special Focus issue of *Nature Reviews Cancer* discusses issues surrounding important genetic, epigenetic and protein biomarkers of cancer, including how these can be used to better understand tumour formation and to develop new therapeutic approaches.

nature reviews cancer, Feb. 2006

biomarker – applications

- disease detection
- diagnosis: stage, subtype
- treatment selection and monitoring
- prognosis

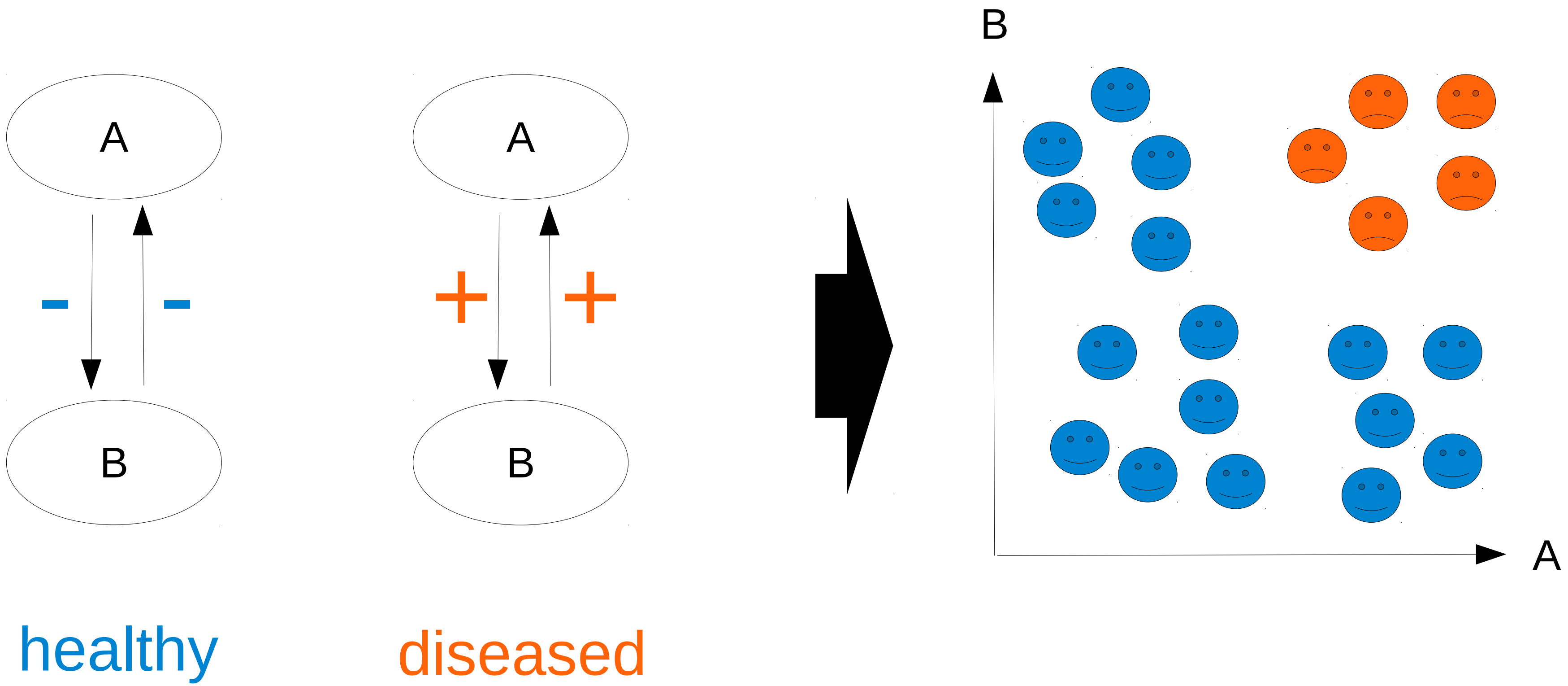


personalized medicine

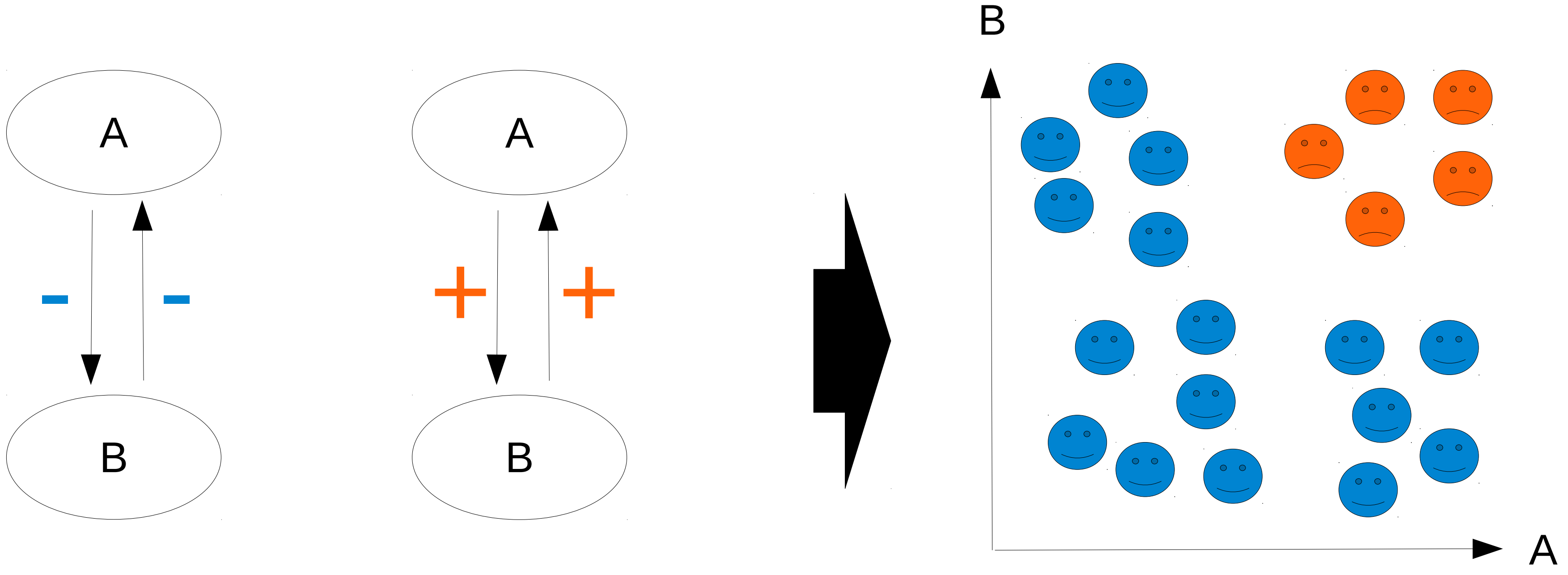
nature reviews cancer, Feb. 2006

biomarker – screening

sometimes: no univariate “profiles”



sometimes: no univariate “profiles”

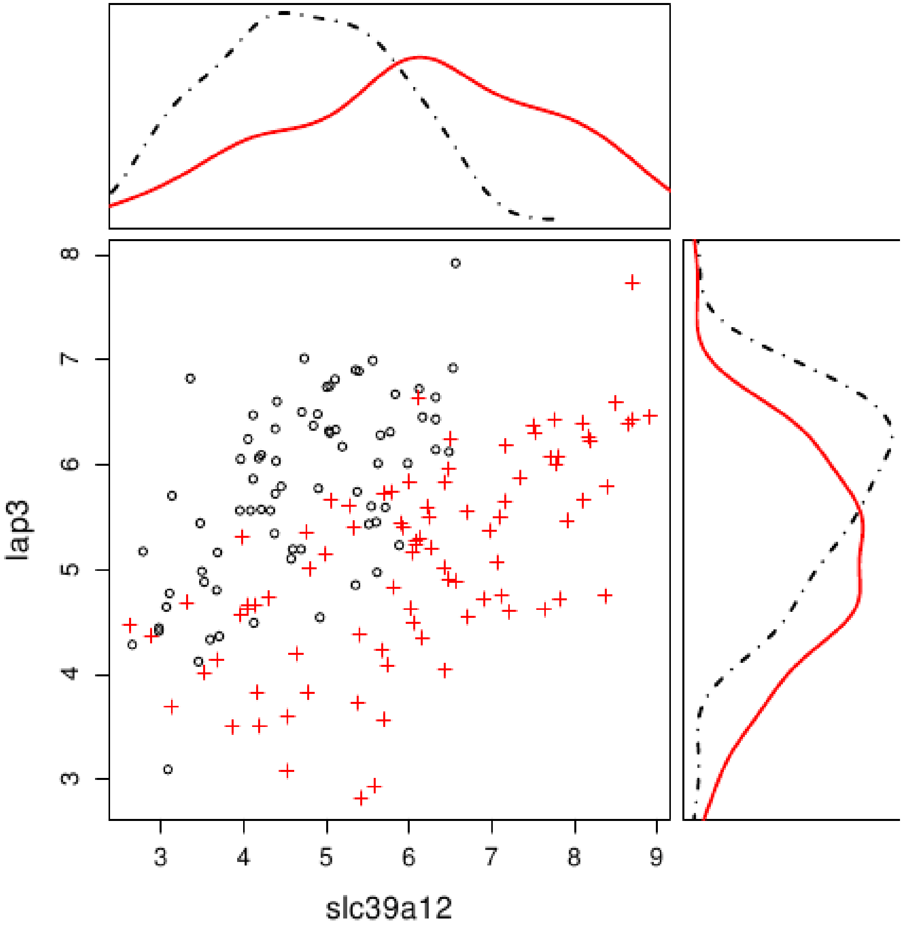


healthy

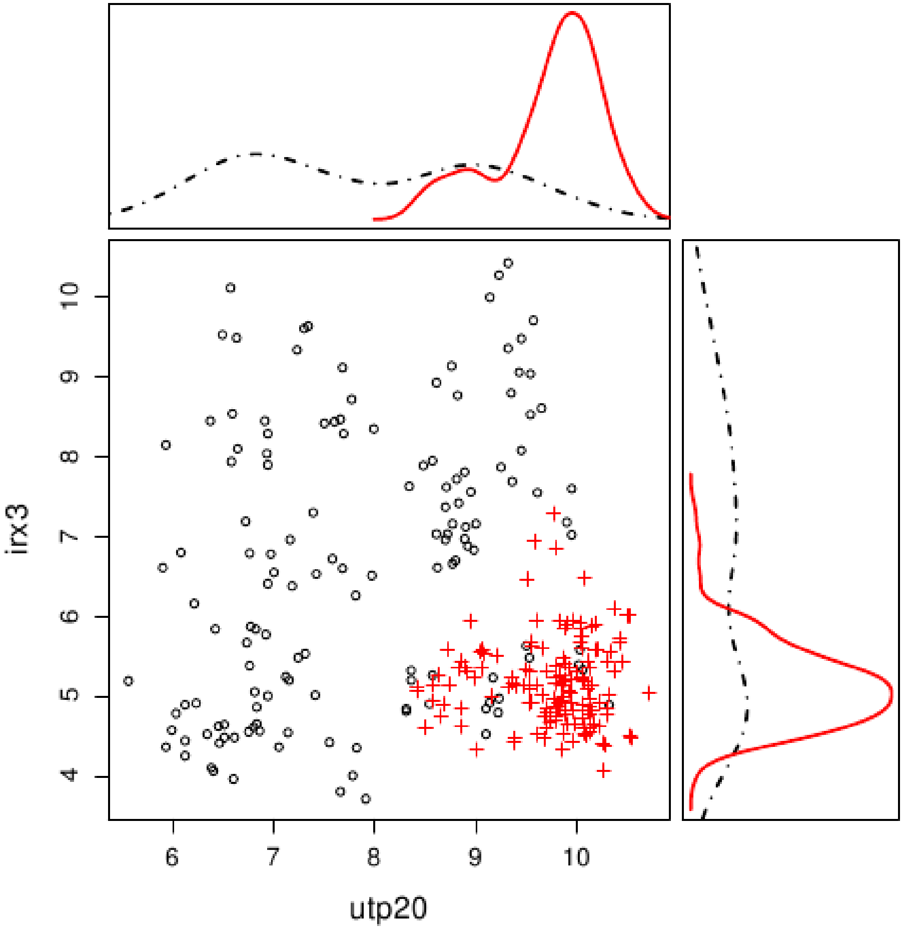
diseased

→ “biosignature”

biosignatures – 2D examples



mouse cell culture: pluripotent vs non-pluripotent



human brain tissue: Alzheimer disease vs healthy

cooperation: Lena Scheubert & Georg Fuellen, IMIBA, Rostock University

biosignatures – in the clinics

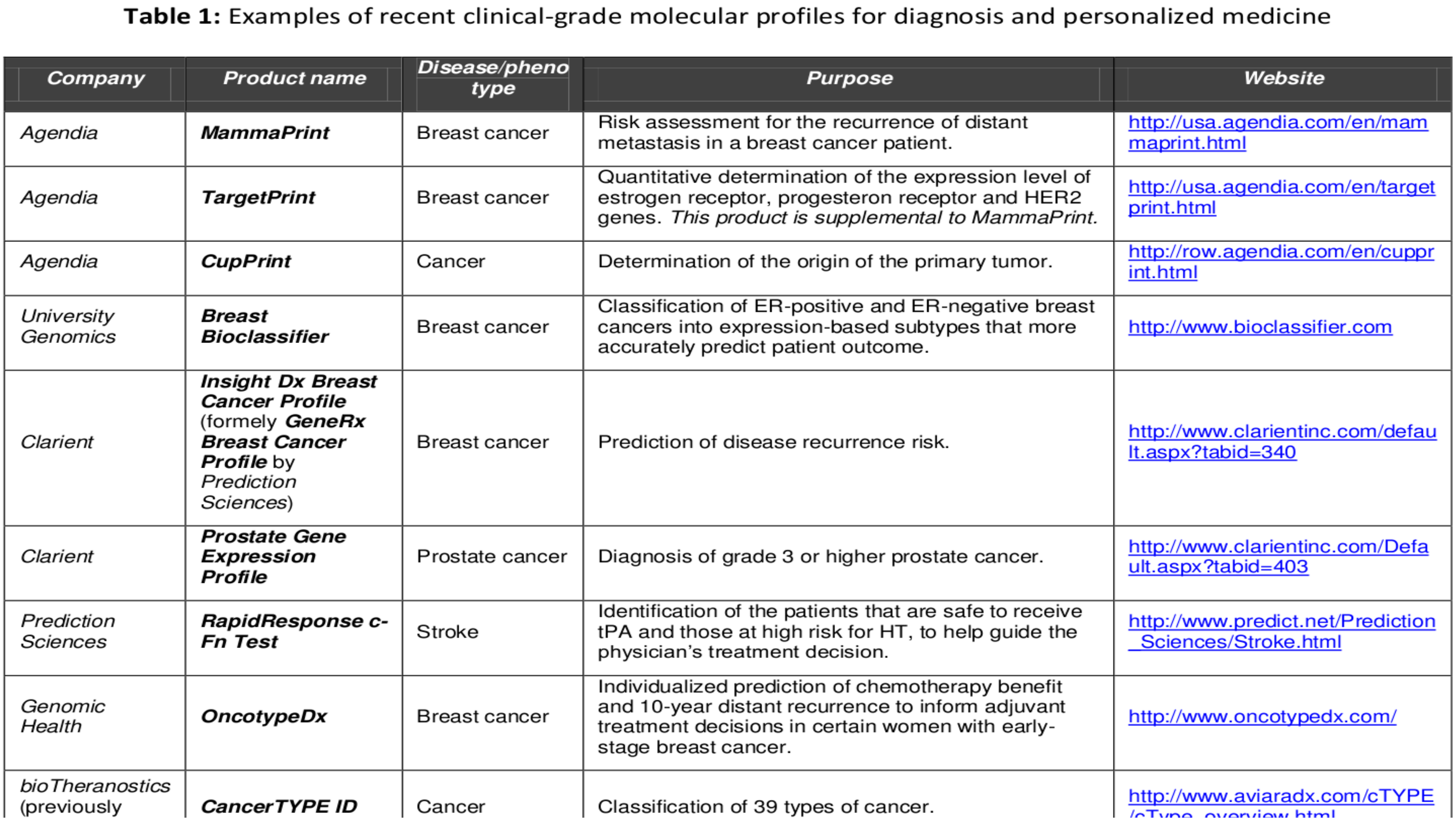


Table 1: Examples of recent clinical-grade molecular profiles for diagnosis and personalized medicine

Company	Product name	Disease/phenotype	Purpose	Website
Agendia	***MammaPrint***	Breast cancer	Risk assessment for the recurrence of distant metastasis in a breast cancer patient.	http://usa.agendia.com/en/mammaprint.html
Agendia	***TargetPrint***	Breast cancer	Quantitative determination of the expression level of estrogen receptor, progesteron receptor and HER2 genes. *This product is supplemental to MammaPrint.*	http://usa.agendia.com/en/targetprint.html
Agendia	***CupPrint***	Cancer	Determination of the origin of the primary tumor.	http://row.agendia.com/en/cupprint.html
University Genomics	***Breast Bioclassifier***	Breast cancer	Classification of ER-positive and ER-negative breast cancers into expression-based subtypes that more accurately predict patient outcome.	http://www.bioclassifier.com
Clarient	***Insight Dx Breast Cancer Profile*** (formely ***GeneRx Breast Cancer Profile*** by *Prediction Sciences*)	Breast cancer	Prediction of disease recurrence risk.	http://www.clarientinc.com/default.aspx?tabid=340
Clarient	***Prostate Gene Expression Profile***	Prostate cancer	Diagnosis of grade 3 or higher prostate cancer.	http://www.clarientinc.com/Default.aspx?tabid=403
Prediction Sciences	***RapidResponse c-Fn Test***	Stroke	Identification of the patients that are safe to receive tPA and those at high risk for HT, to help guide the physician's treatment decision.	http://www.predict.net/Prediction_Sciences/Stroke.html
Genomic Health	***OncotypeDx***	Breast cancer	Individualized prediction of chemotherapy benefit and 10-year distant recurrence to inform adjuvant treatment decisions in certain women with early-stage breast cancer.	http://www.oncotypedx.com/
bioTheranostics (previously	***CancerTYPE ID***	Cancer	Classification of 39 types of cancer.	http://www.aviaradx.com/cTYPE/cType_overview.html

Aliferis et al., 2010

biomarker/biosignature – problems

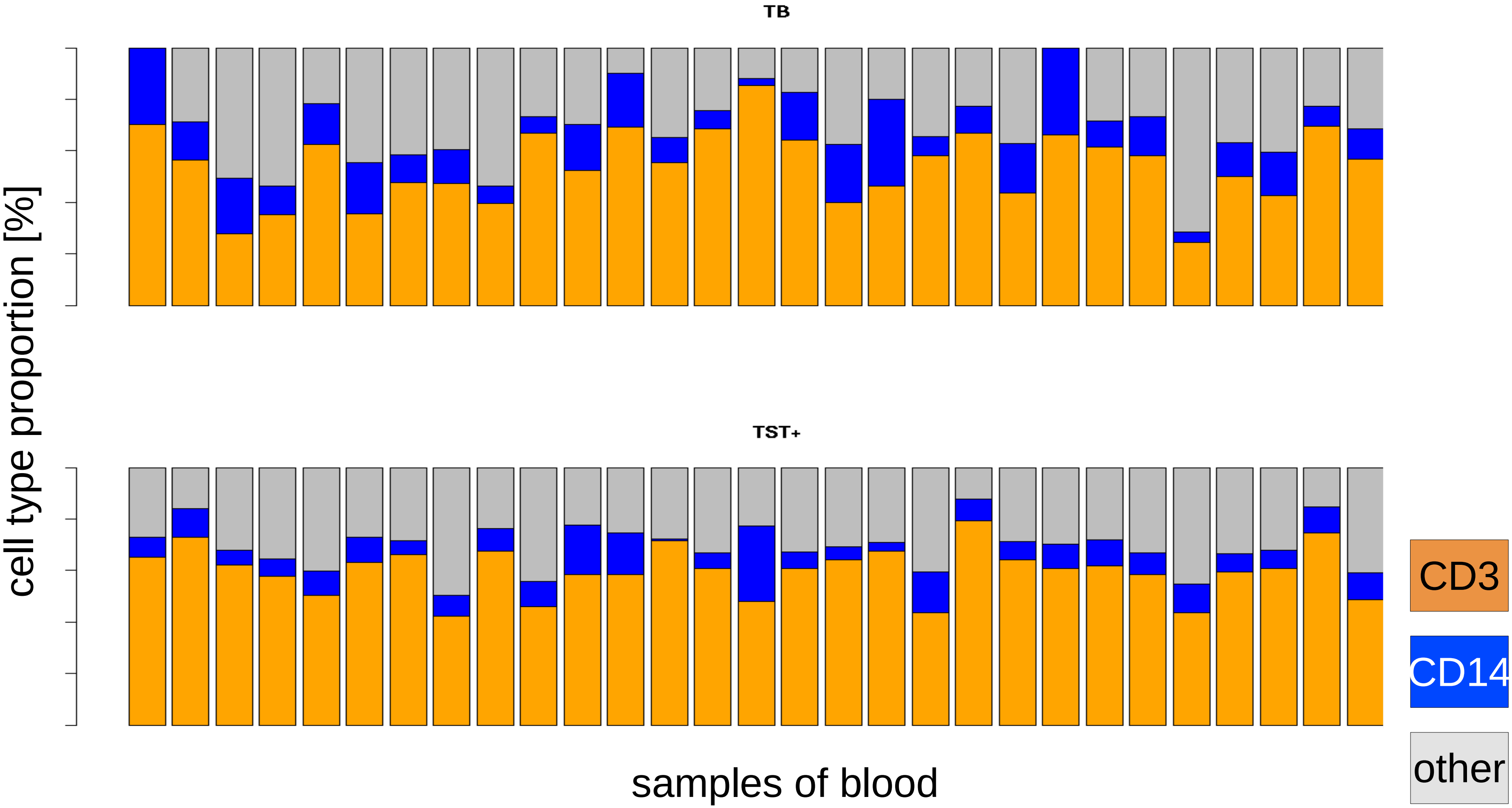
- part I: heterogeneous tissues (= mixtures of cell types)
- part II: pooled sample designs (= mixtures of individual samples)

biomarker/biosignature – problems

- part I: heterogeneous tissues
  (= mixtures of cell types)

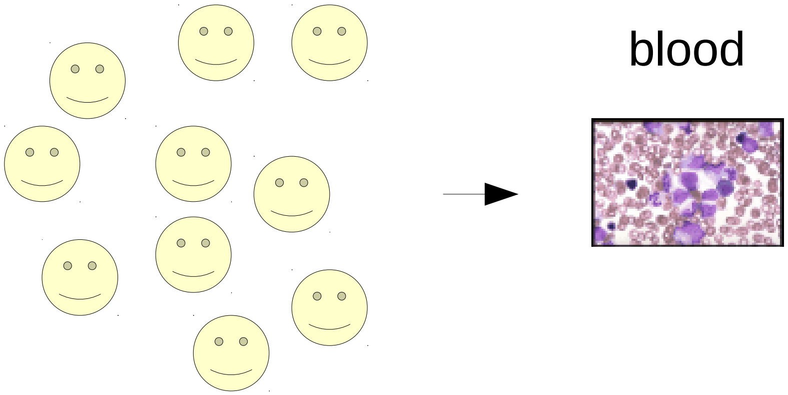
- part II: pooled sample designs
  (= mixtures of individual samples)

blood as heterogeneous tissue: sample heterogeneity



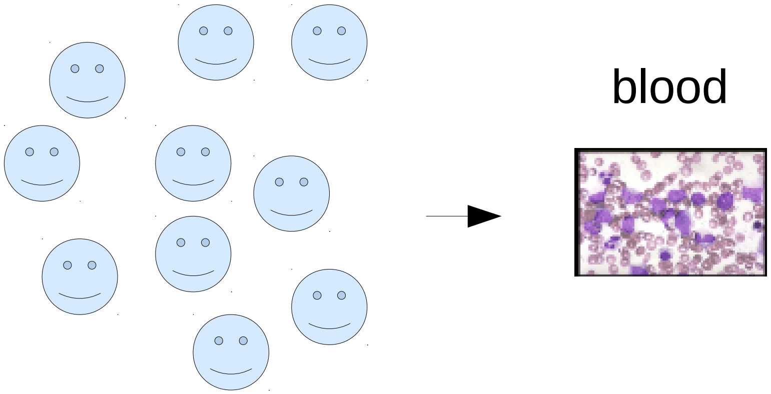
case study

control patients



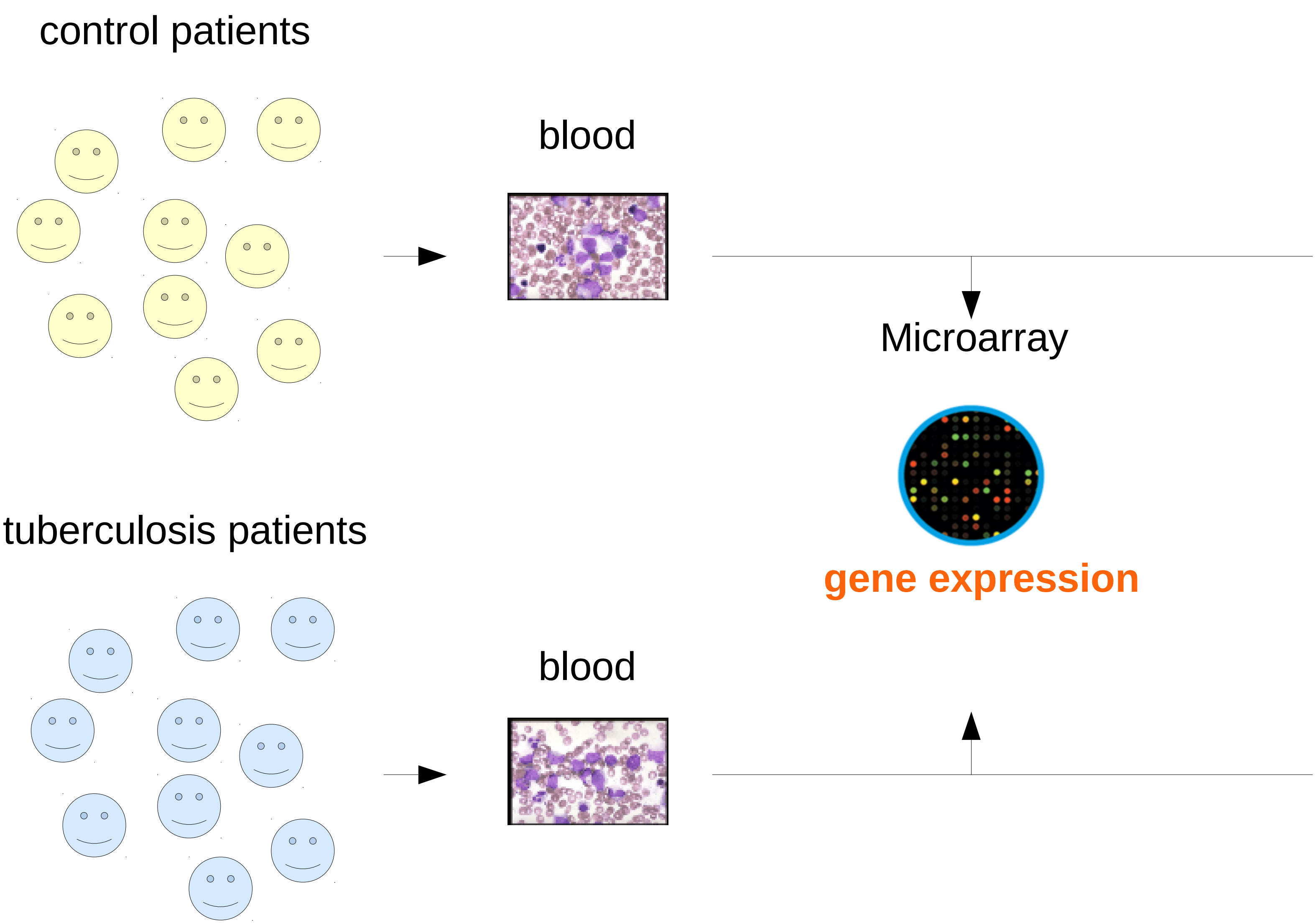
blood

tuberculosis patients



blood

case study



control patients

blood

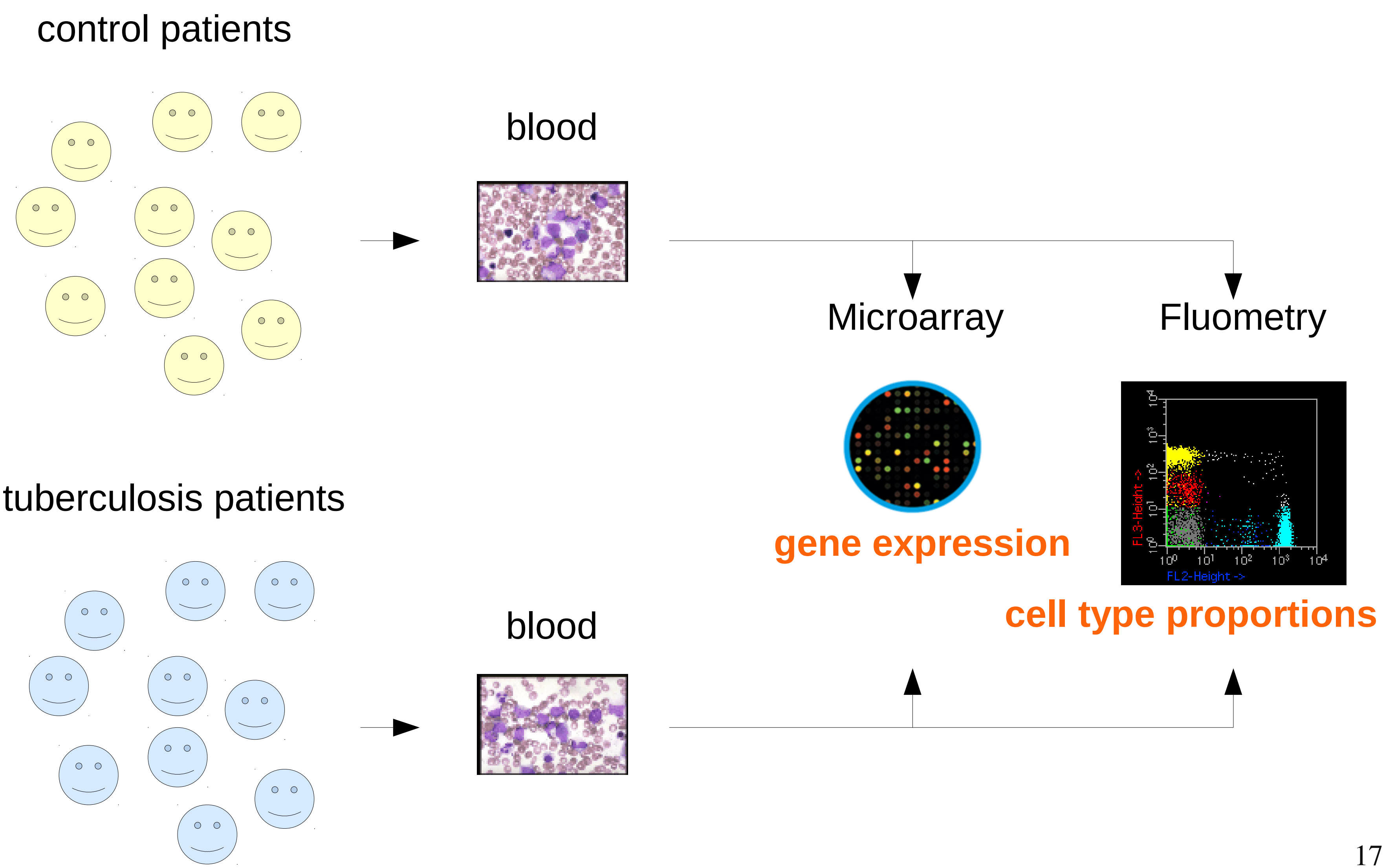
Microarray

gene expression

tuberculosis patients

blood

case study



control patients

blood

Microarray

Fluometry

gene expression

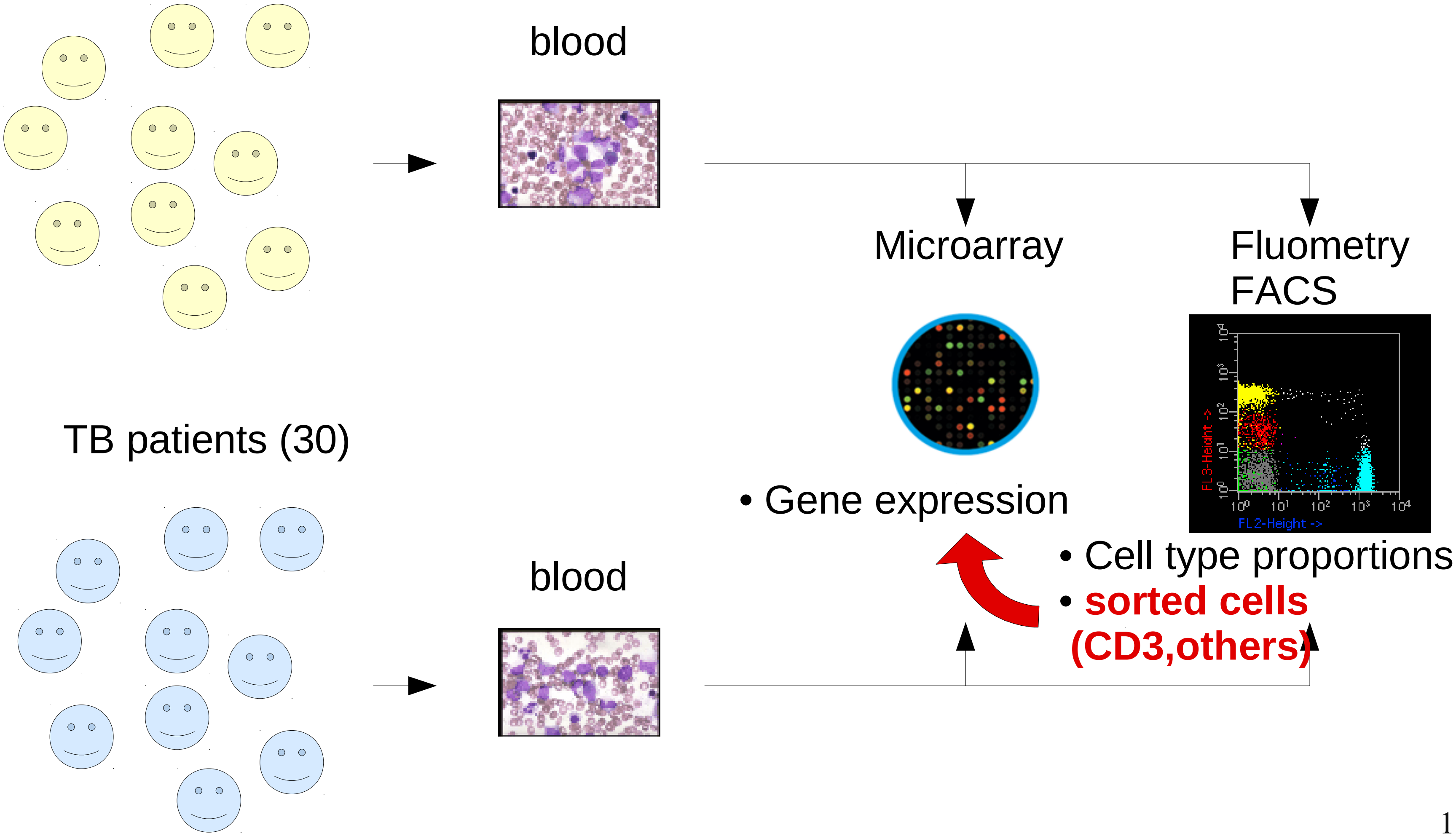
cell type proportions

tuberculosis patients

blood

experimental study

healthy contacts (30)



blood

Microarray

Fluometry
FACS

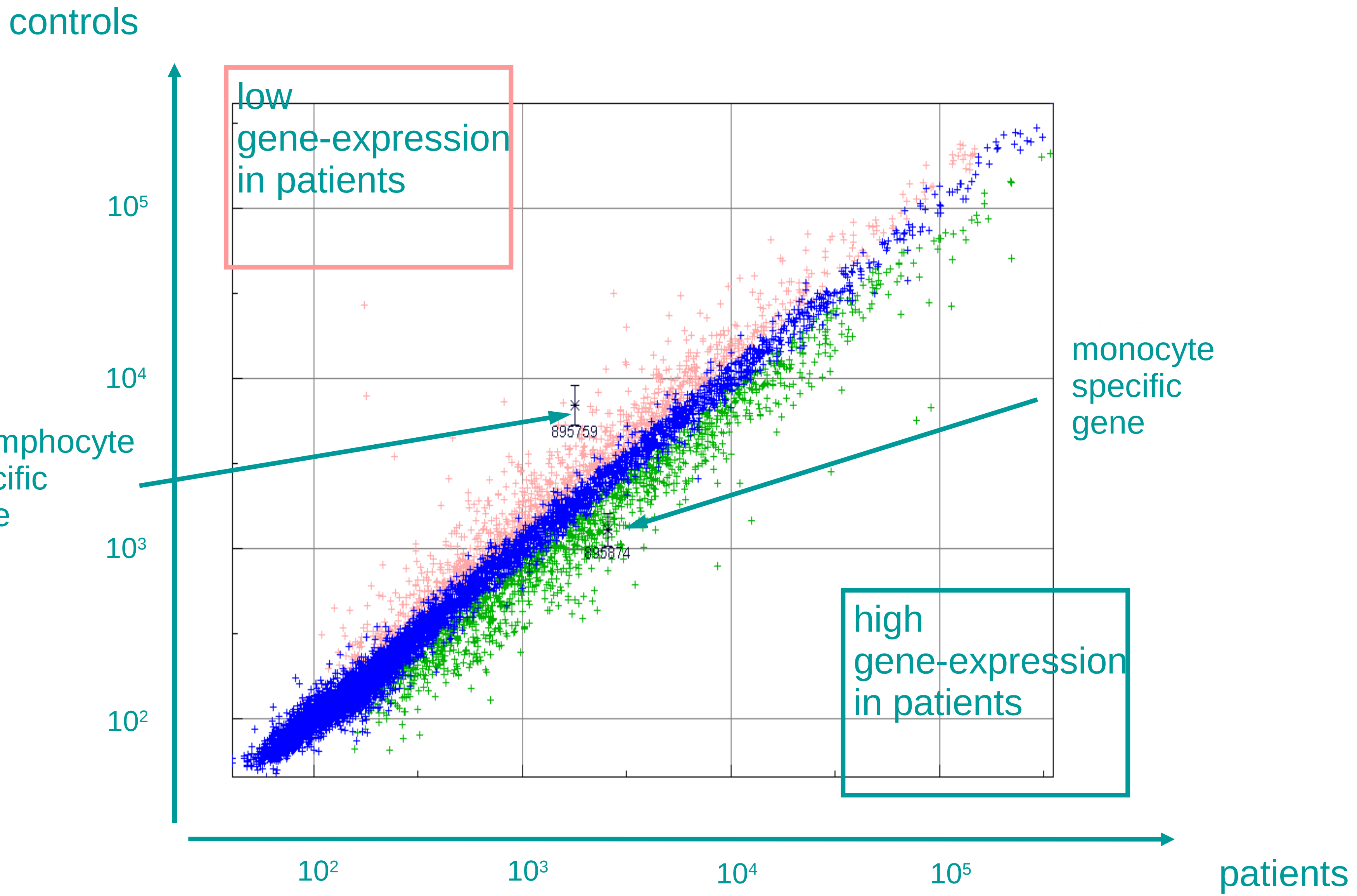
- Gene expression

TB patients (30)

blood

- Cell type proportions
- **sorted cells (CD3,others)**

case study – results: gene-expression



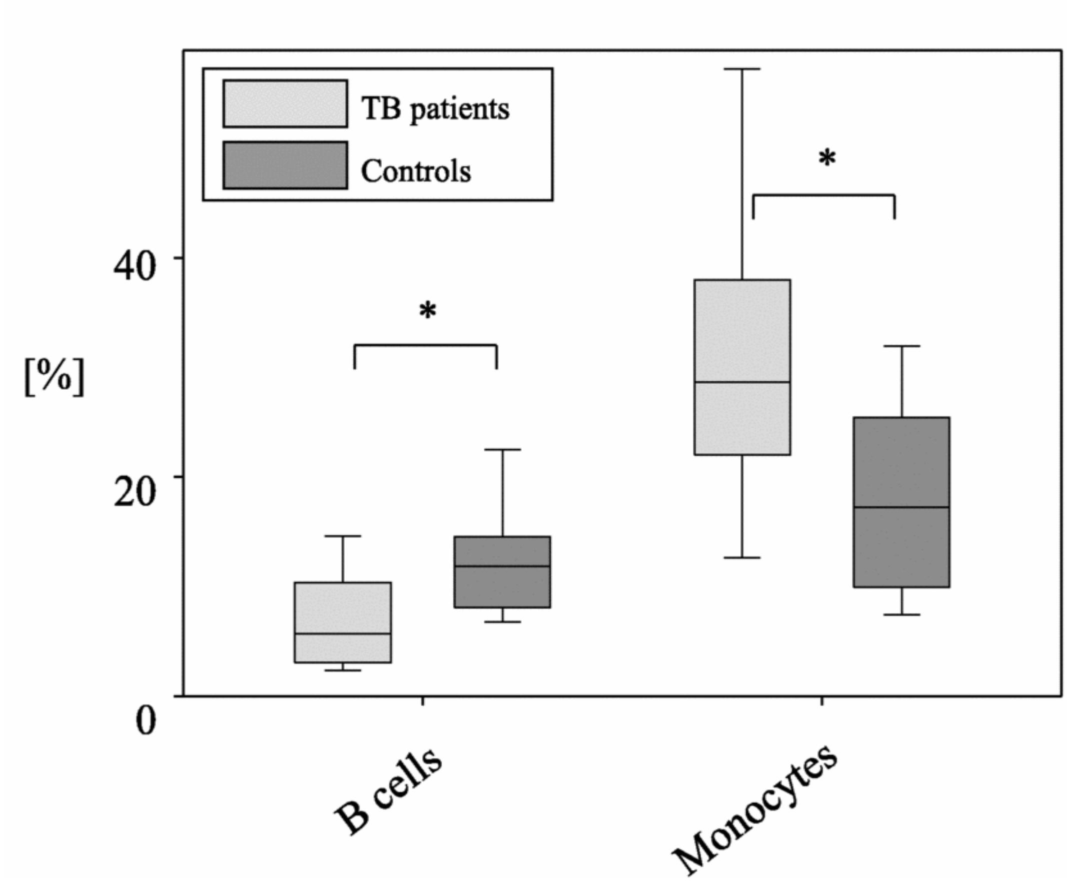
low gene-expression in patients

B lymphocyte specific gene

monocyte specific gene

high gene-expression in patients

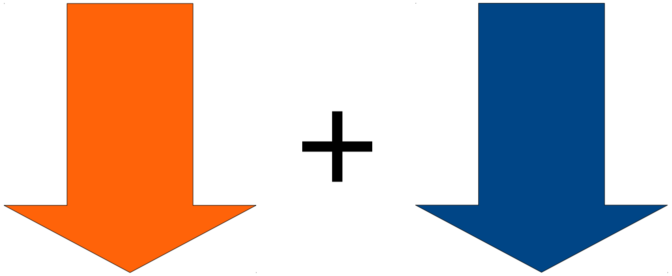
case study – results: cell counting



this is the problem :

cell-type proportions

cell-type specific gene expression



+

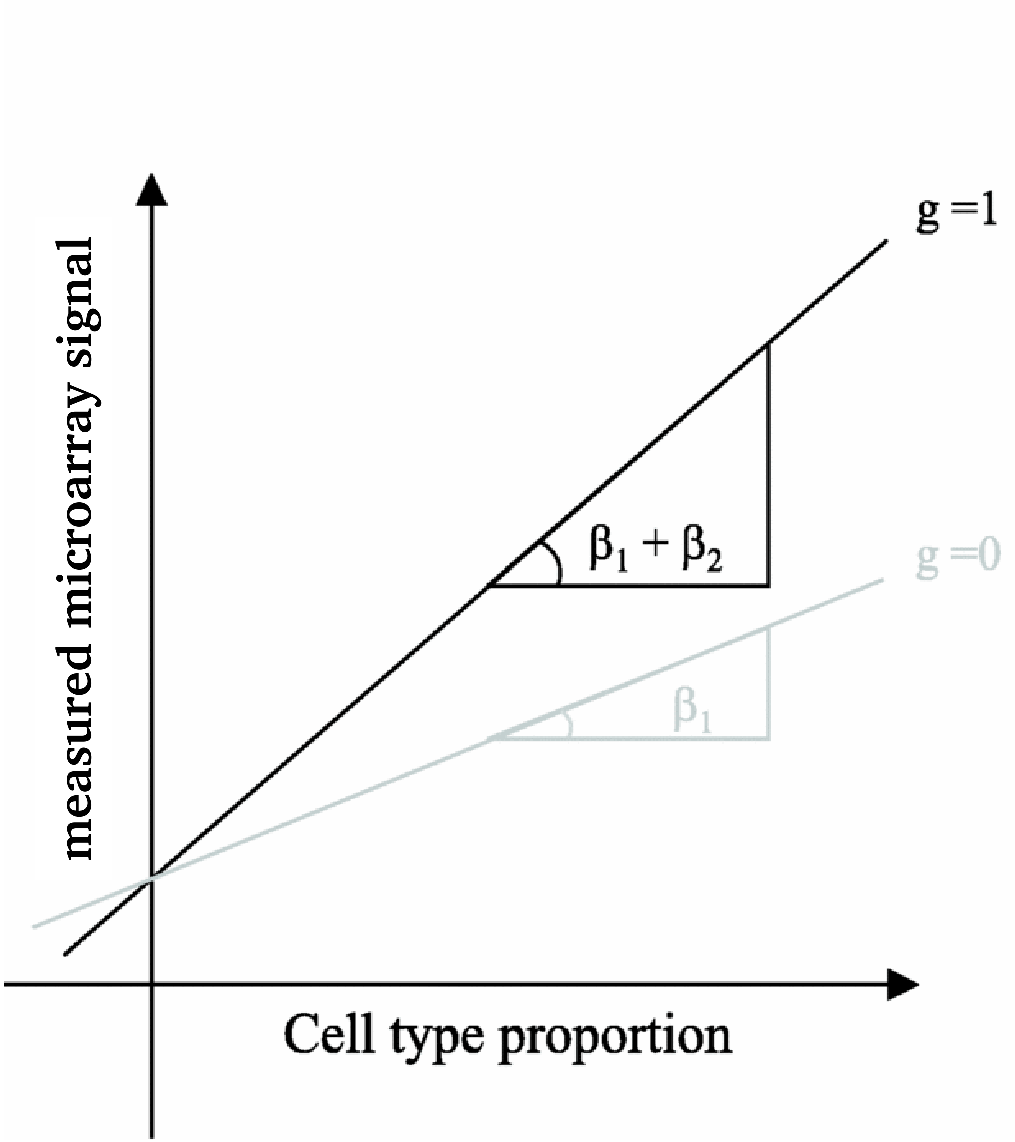
microarray results

possible cases

- simplest: cell-type specific expression
cell-type proportions measured
independency
- problematic: non-specific expression
proportions not measured
independency
- worst: expression dependent
on proportions

simplest case

quantitative model



$y = ß_0 + ß_1 \cdot p + \boxed{ß_2 \cdot g * p}$

Group models:

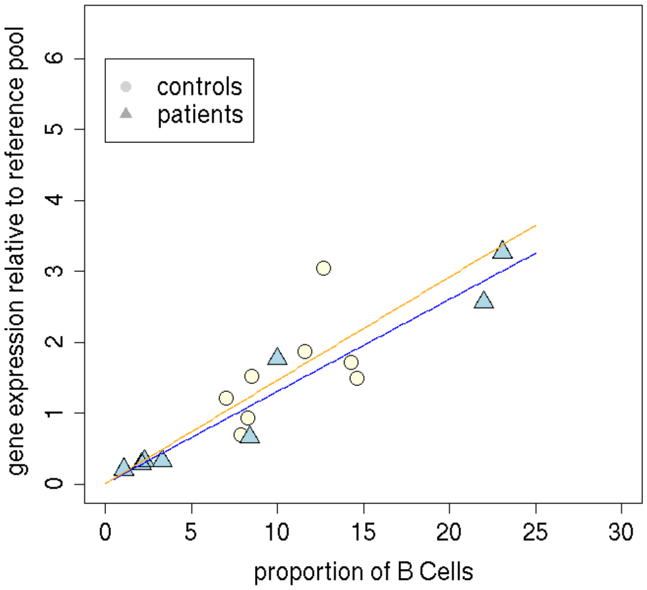
g=0, controls:
$y = ß_0 + ß_1 \cdot p + e$

g=1, patients:
$y = ß_0 + (ß_1 + ß_2) \cdot p + e$

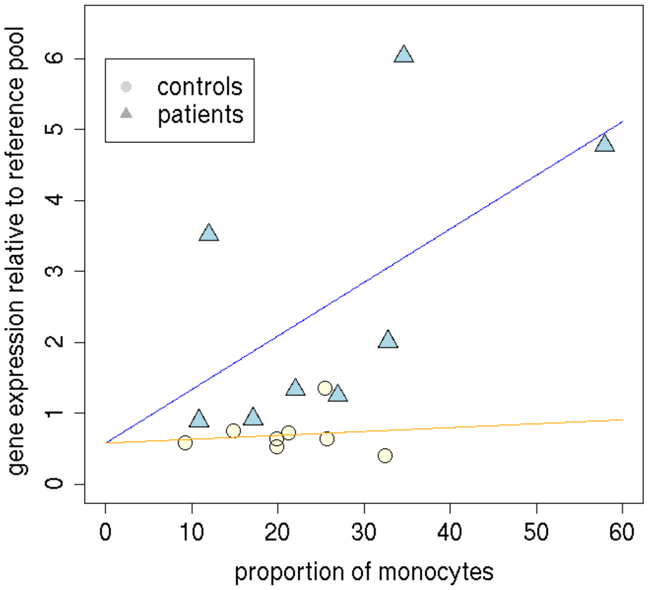
Jacobsen, Repsilber et al. MIM 2006

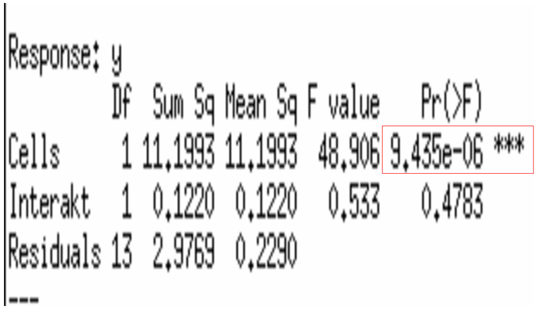
regressions

CD 20 expression vs B-cell proportion

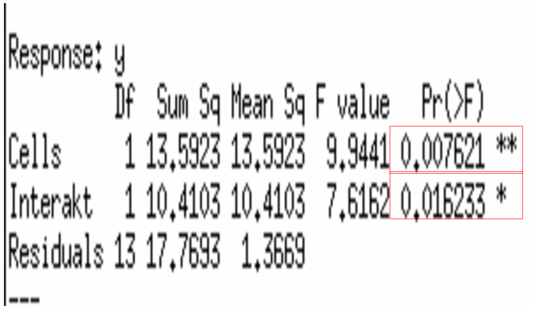


CD 64 expression vs monocytes proportion





```
Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
Cells      1 11.1993 11.1993  48.906 9.435e-06 ***
Interakt   1  0.1220  0.1220   0.533    0.4783
Residuals 13  2.9769  0.2290
---
```



```
Response: y
          Df  Sum Sq Mean Sq F value   Pr(>F)
Cells      1 13.5923 13.5923  9.9441 0.007621 **
Interakt   1 10.4103 10.4103  7.6162 0.016233 *
Residuals 13 17.7693  1.3669
---
```

experimental validation (on new samples):

- single cell qPCR
- single cell protein assay

problematic case

more realistic assumptions :

- non-specific gene expression
  (most genes expressed in all cell types)
- cell types: proportions unknown
- independence

existing approaches

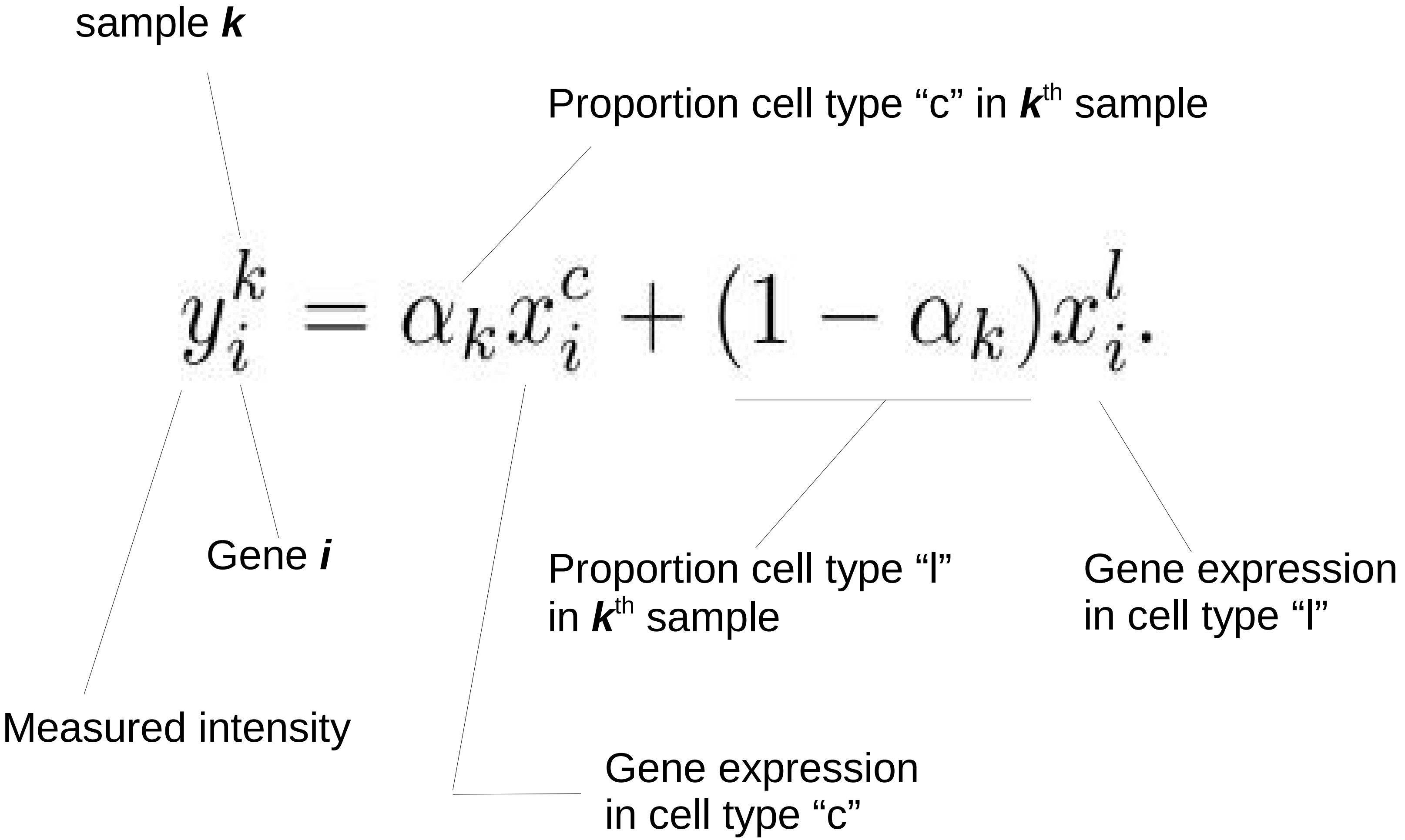
- Venet et al, Bioinf 2001
- Lahdesmaki et al, BMC Bioinf 2005

de-composition
of measured gene expression signals

non-negative matrix factorization

“deconfounding”

deconfounding



$$y_i^k = \alpha_k x_i^c + (1 - \alpha_k) x_i^l.$$

sample ***k***

Proportion cell type “c” in ***k***th sample

Gene ***i***

Proportion cell type “l” in ***k***th sample

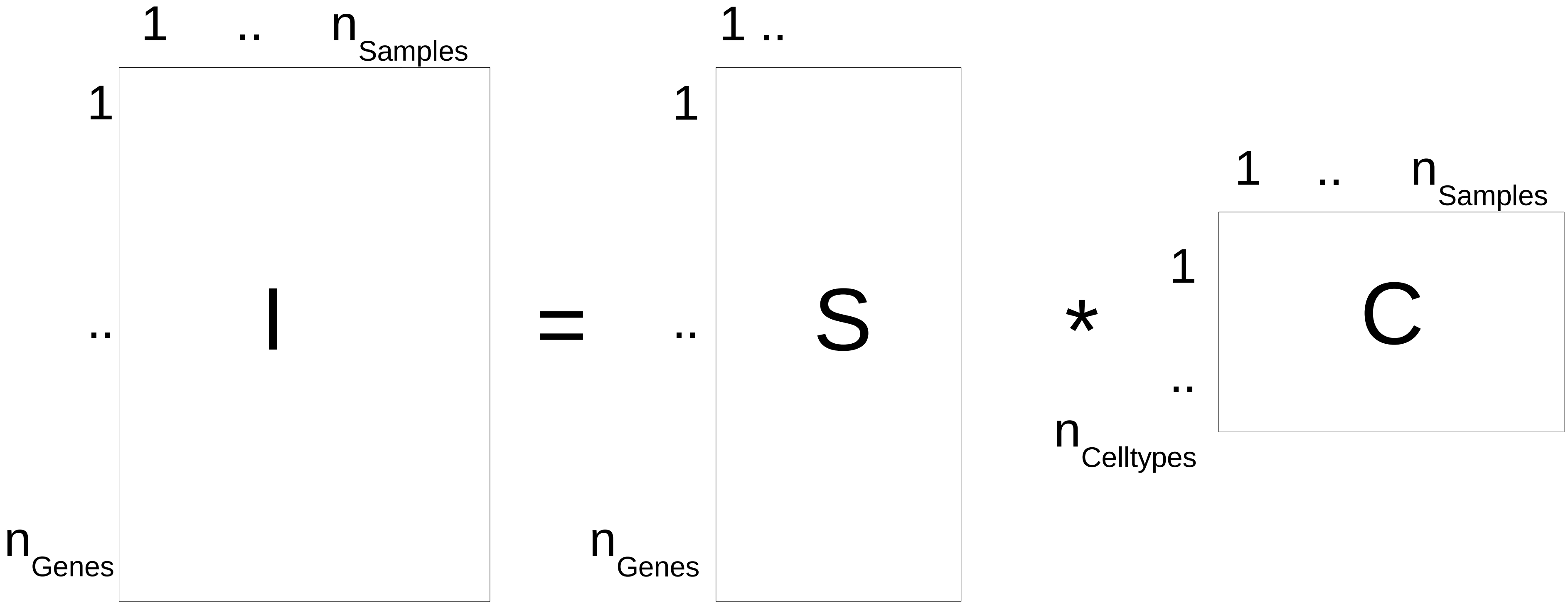
Gene expression in cell type “l”

Measured intensity

Gene expression in cell type “c”

Lahdesmaki et al., 2005

deconfounding



$1 \quad .. \quad n_{Samples}$

1

$..$

n_{Genes}

I

$=$

$1 \; ..$

1

$..$

n_{Genes}

S

$*$

$1 \quad .. \quad n_{Samples}$

1

$..$

$n_{Celltypes}$

C

measured
microarray signals

cell-type
specific expression
profiles

cell-type proportions
for all samples

Venet et al., 2001

constraints

Normalization on I:

$$\sum_{i}^{n_{genes}} I_{ij} = const.$$

column sums

Constraints for S:

$$S_{ik} \geq 0$$

$$\sum_{i}^{n_{genes}} S_{ik} = const.$$

column sums

Constraints for C:

$$C_{kj} \geq 0$$

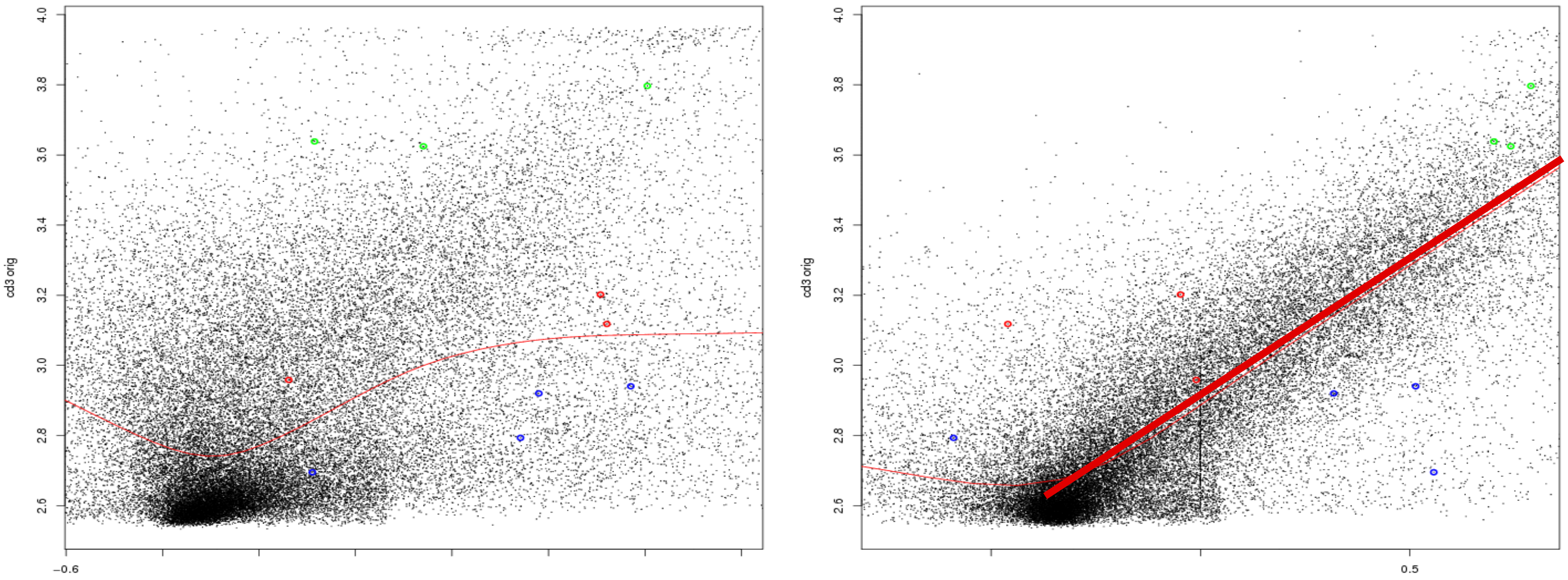
$$\sum_{k}^{n_{cell\ types}} C_{kj} = 1$$

experimental validation
# of the deconfounding approach

deconfounding at work :

- recovering cell type specific gene expression
- recovering cell type proportions

validation: gene expression profiles



cell type 1

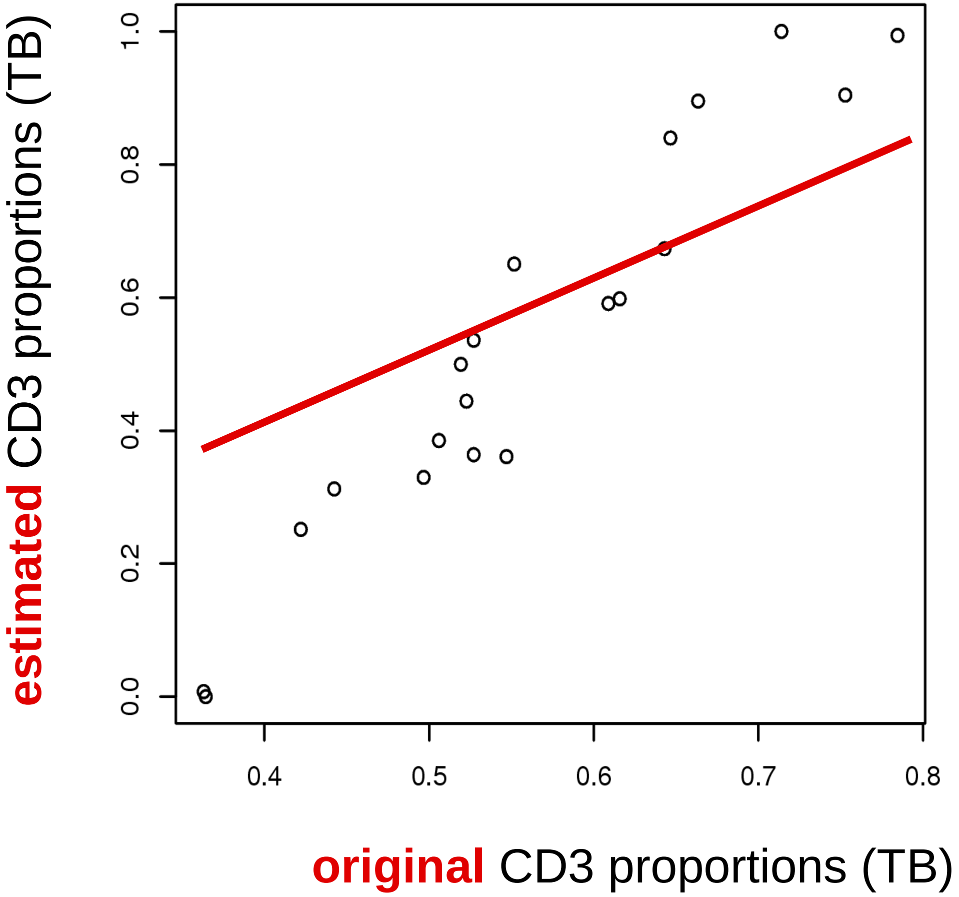
cell type 2

original gene expression profile CD3

experimental data

estimated gene expression profile

validation: cell type proportions



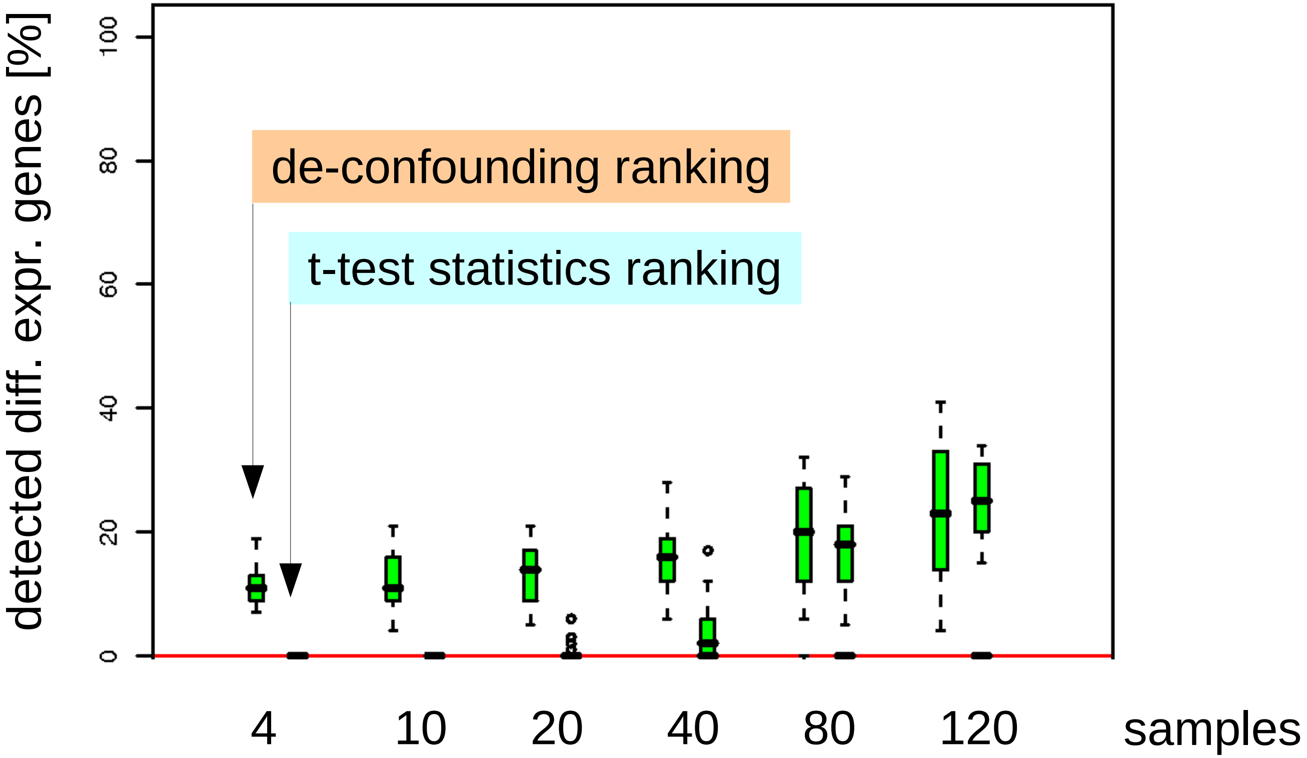
experimental data

does “deconfounding” help
for detection of
valid differential gene expression ???

does “deconfounding” help
for detection of
valid differential gene expression ???

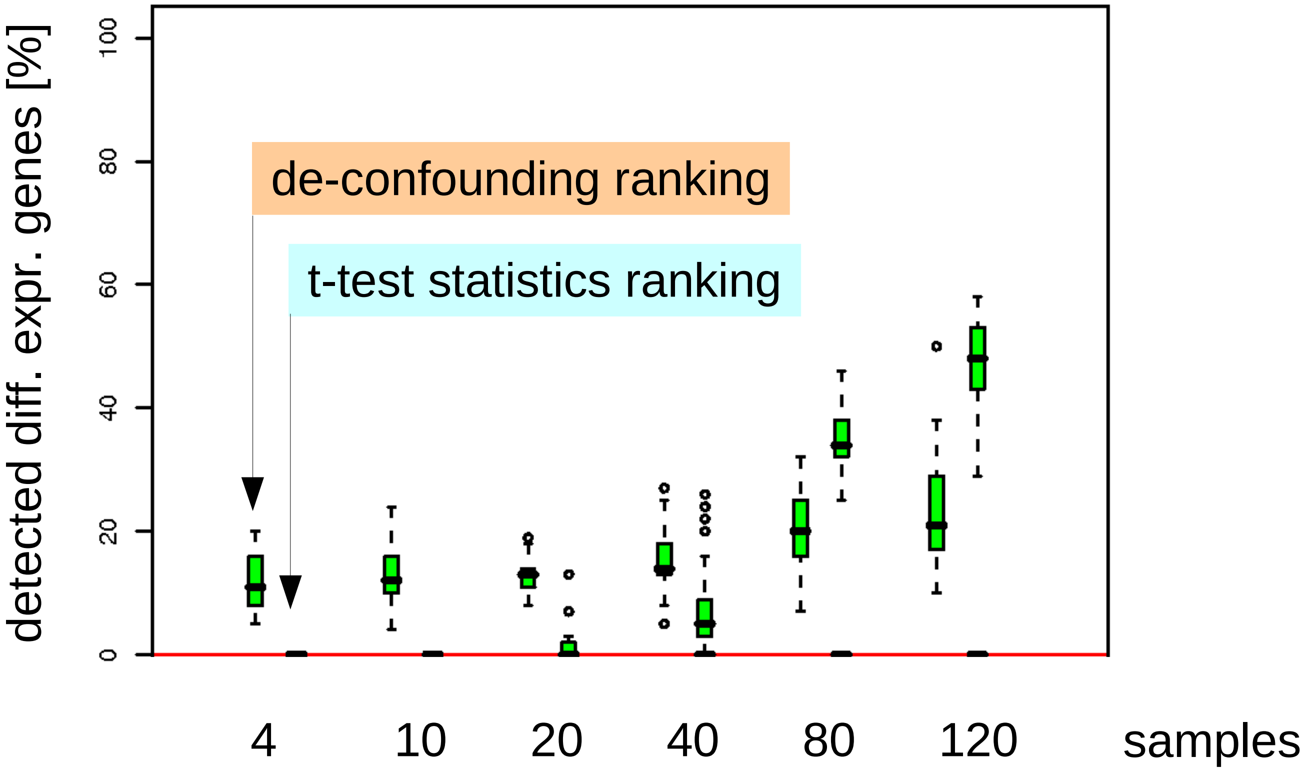
simulation study

CD3: UP / other: ---

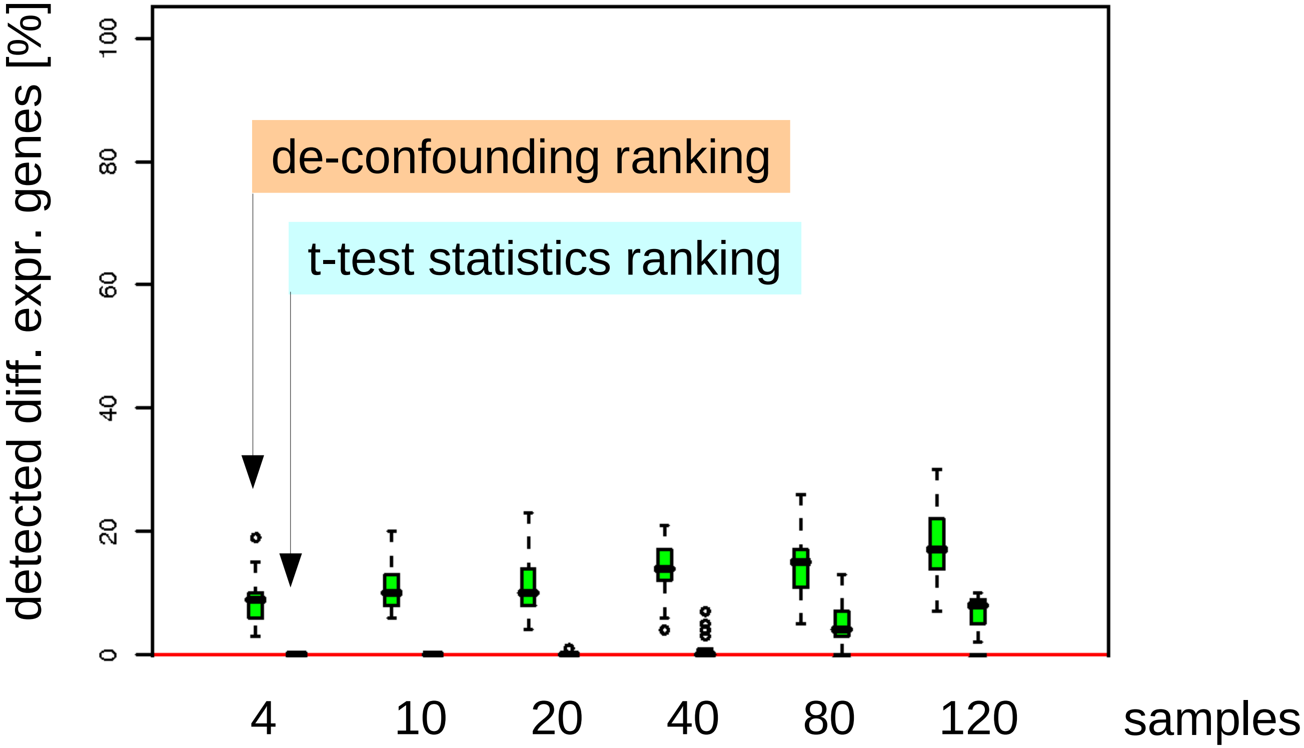


de-confounding ranking
t-test statistics ranking
detected diff. expr. genes [%]
0
20
40
60
80
100
4
10
20
40
80
120
samples

CD3: UP / other: UP



CD3: UP / other: DOWN



does “deconfounding” help
for detection of
valid differential gene expression ???

yes (it seems)

approaching screening for **biosignatures**

approaching screening for biosignatures

- **problem**:
  - sample variability is already used for estimating the non-negative factorization

approaching screening for biosignatures

- **problem**:
  - sample variability is already used for estimating the non-negative factorization

- **possible solutions**:

  1 predicting cell-type proportions for a new sample, measuring distance to estimated profiles with the same cell-type proportions

Repsilber et al., BMC Bioinformatics 2010

approaching screening for biosignatures

- **problem**:
  - sample variability is already used for estimating the non-negative factorization

- **possible solutions**:

  1 predicting cell-type proportions for a new sample, measuring distance to estimated profiles with the same cell-type proportions

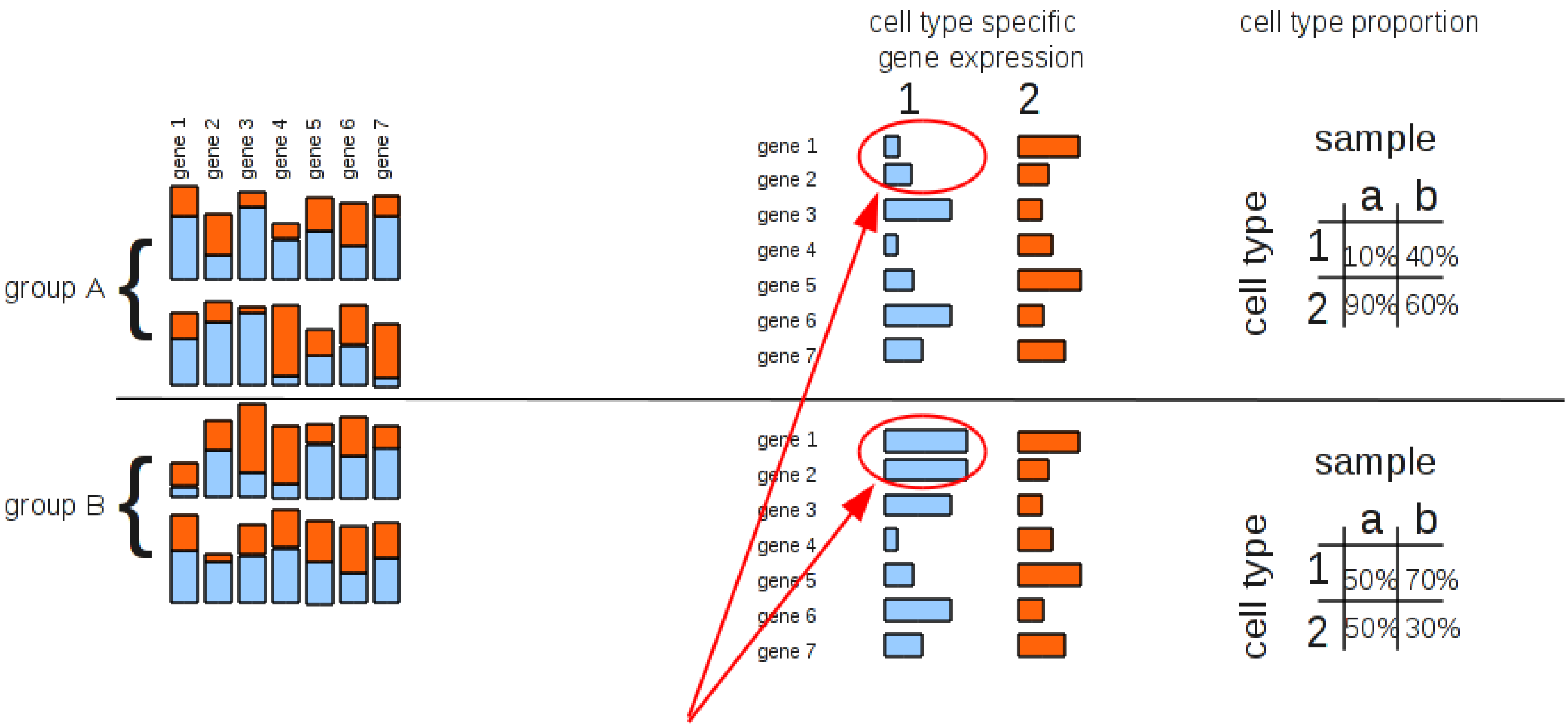
  2 estimating sample variability by substracting mean values cell-type-wise – followed by statistical learning

solution 2

STEP 1

measured data:

after factorization:



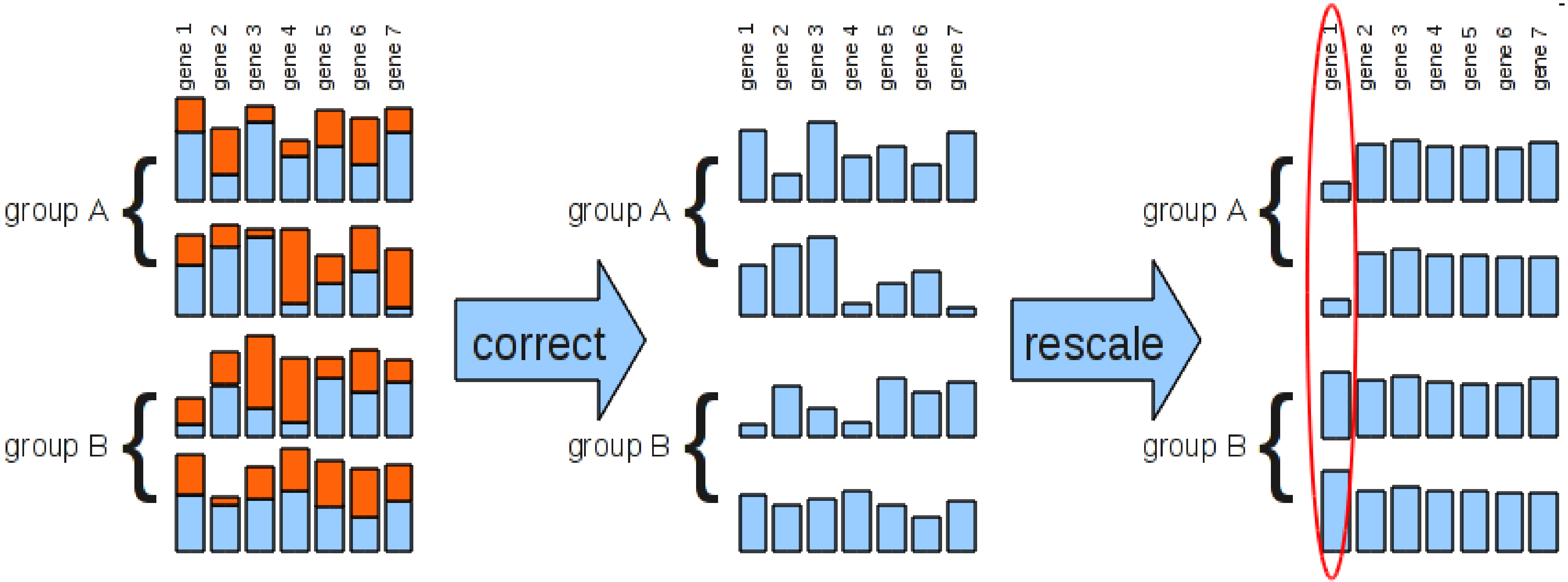
cell type	a	b
1	10%	40%
2	90%	60%

cell type	a	b
1	50%	70%
2	50%	30%

differential gene expression

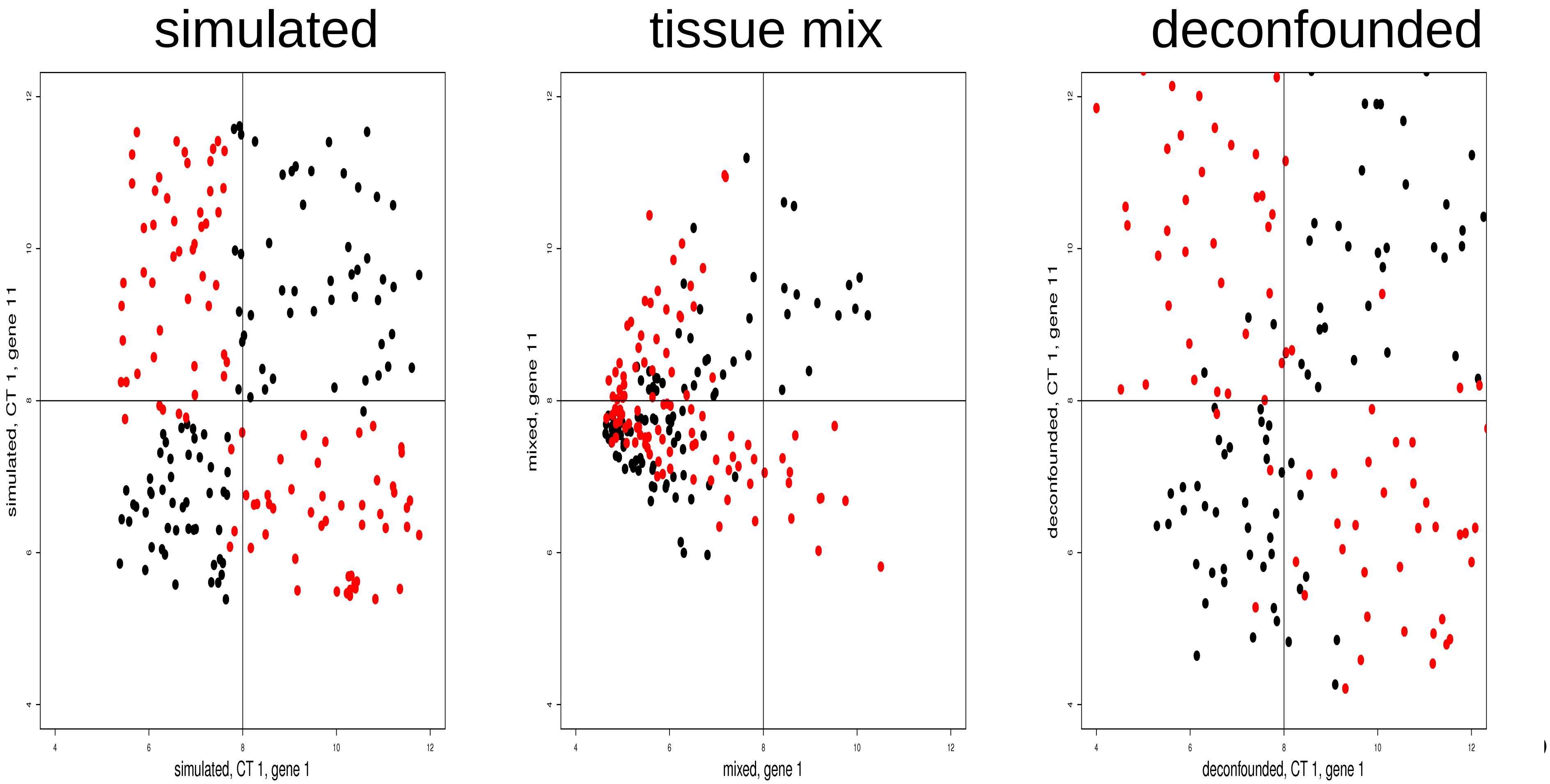
solution 2

STEP 2



solution 2

Results

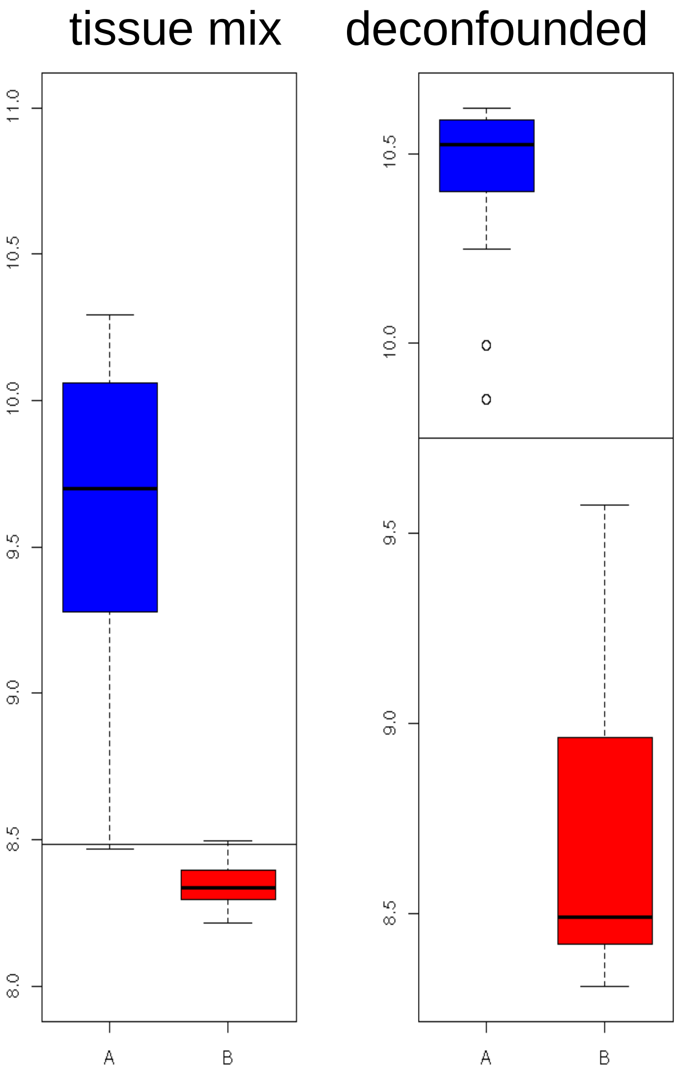


simulated

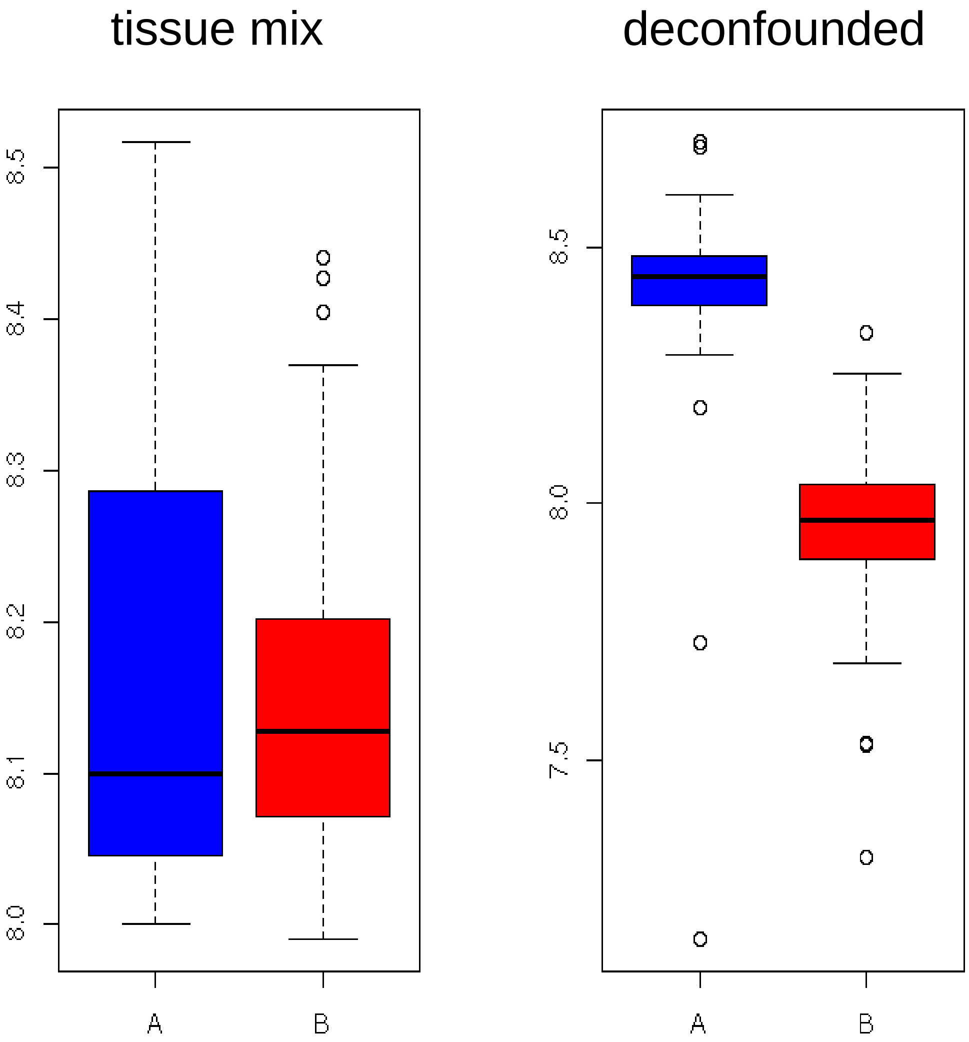
tissue mix

deconfounded

solution 2



delta = 2



delta = 0.4

worst case

worst case

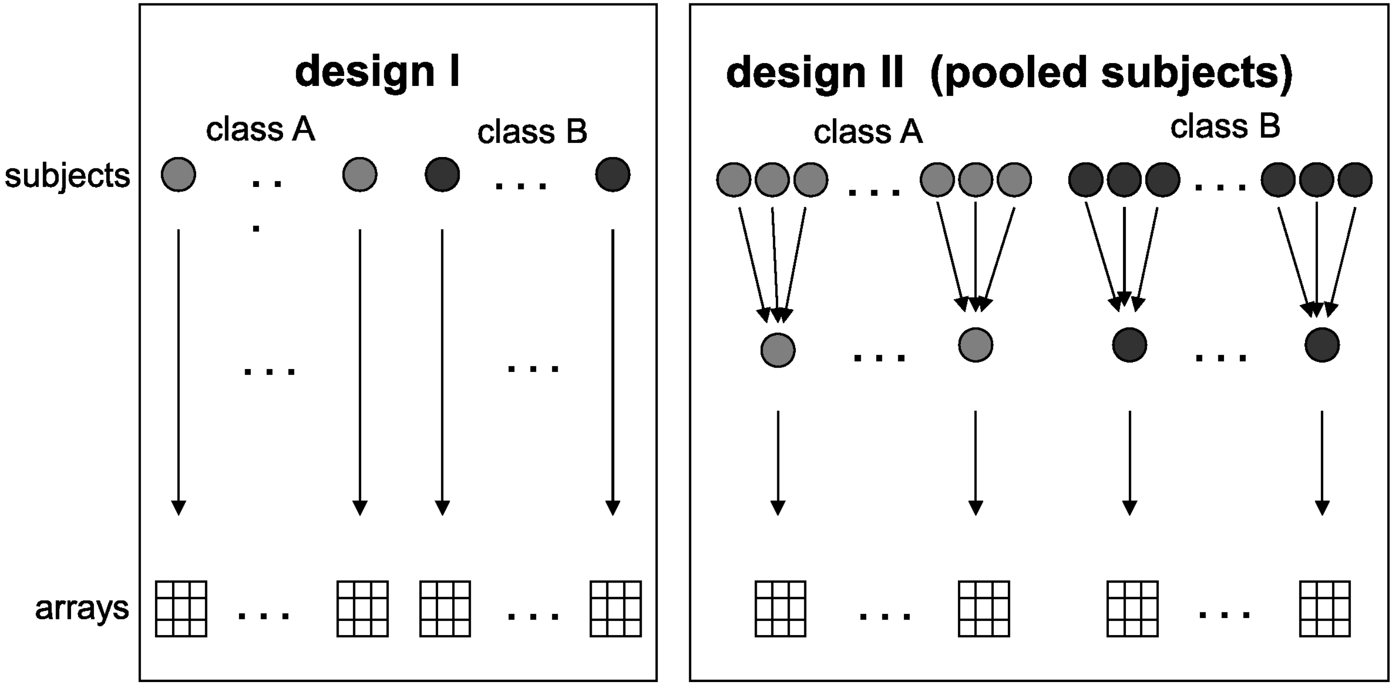
→ next time

biomarker/biosignature – problems

- part I: heterogeneous tissues (= mixtures of cell types)

- part II: pooled sample designs (= mixtures of individual samples)

pooling design



investigated pool sizes: 1(non-pooled), 2, 3, 5

advice: do not pool

The design of a classification study, like for biomarker search, should not consist of pooled samples, because data is required at the ”**individual level**”.

Kerr 2003

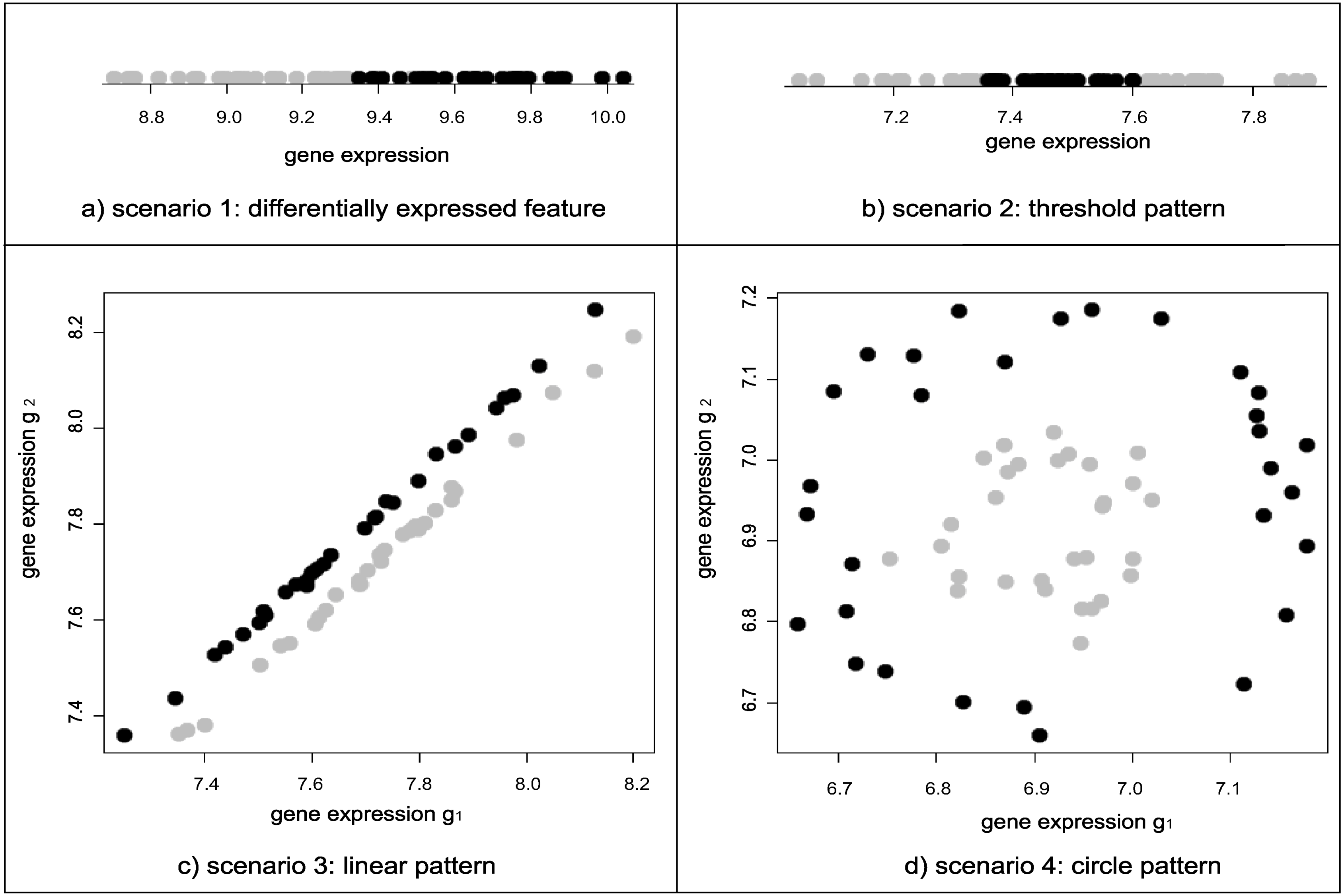
objectives

- find differences in screening methods regarding

    - prediction error minimization

    - finding the true underlying features

data I: simulated



a) scenario 1: differentially expressed feature

b) scenario 2: threshold pattern

c) scenario 3: linear pattern

d) scenario 4: circle pattern

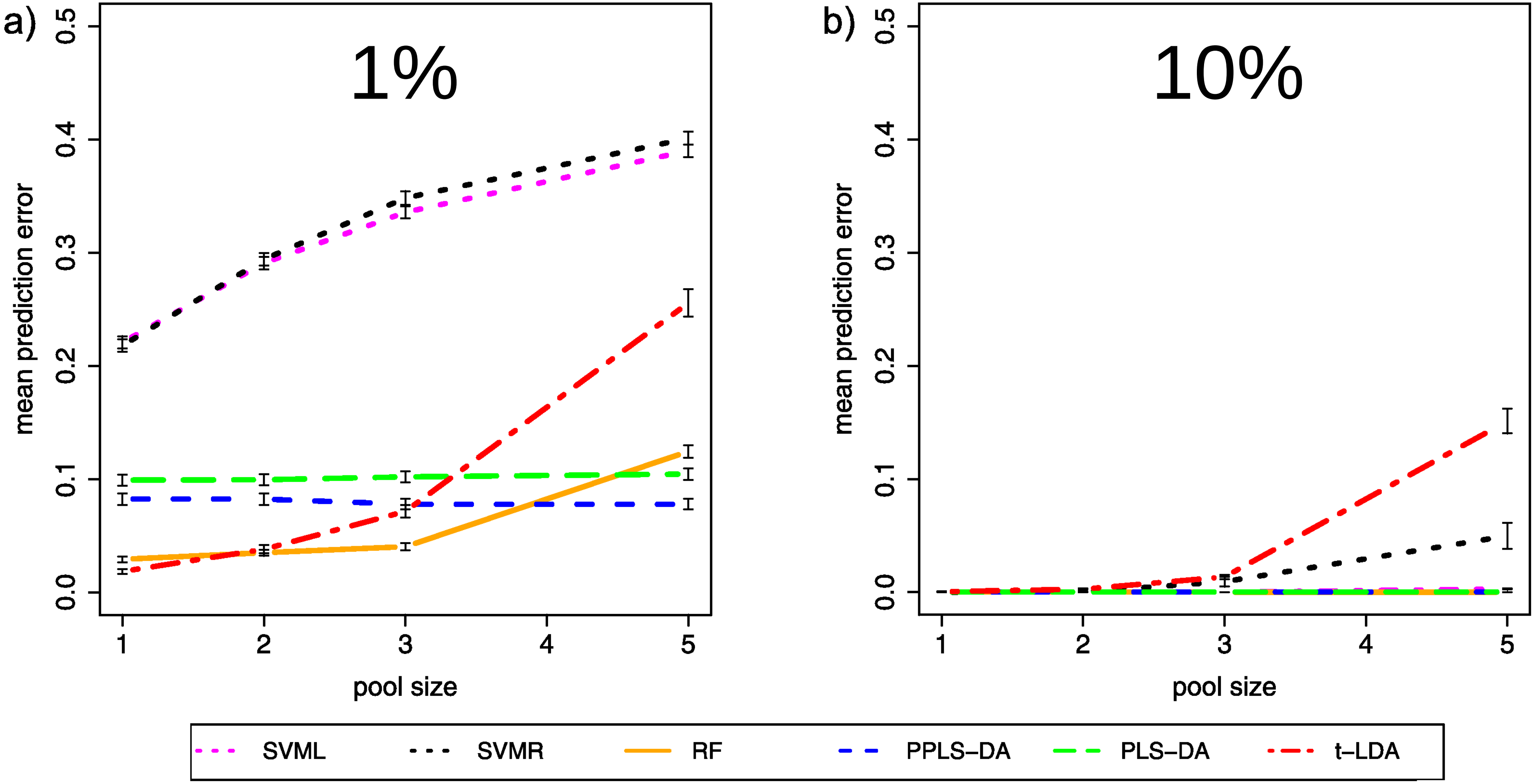
data II: experimental

- cancer gene expression studies
  - Leukemia (Golub et al., 1999)
  - Prostate 1 (Singh et al., 2002)
  - Prostate 2 (Lapointe et al., 2004)
  - Breast Cancer (van't Veer et al., 2002)

methods

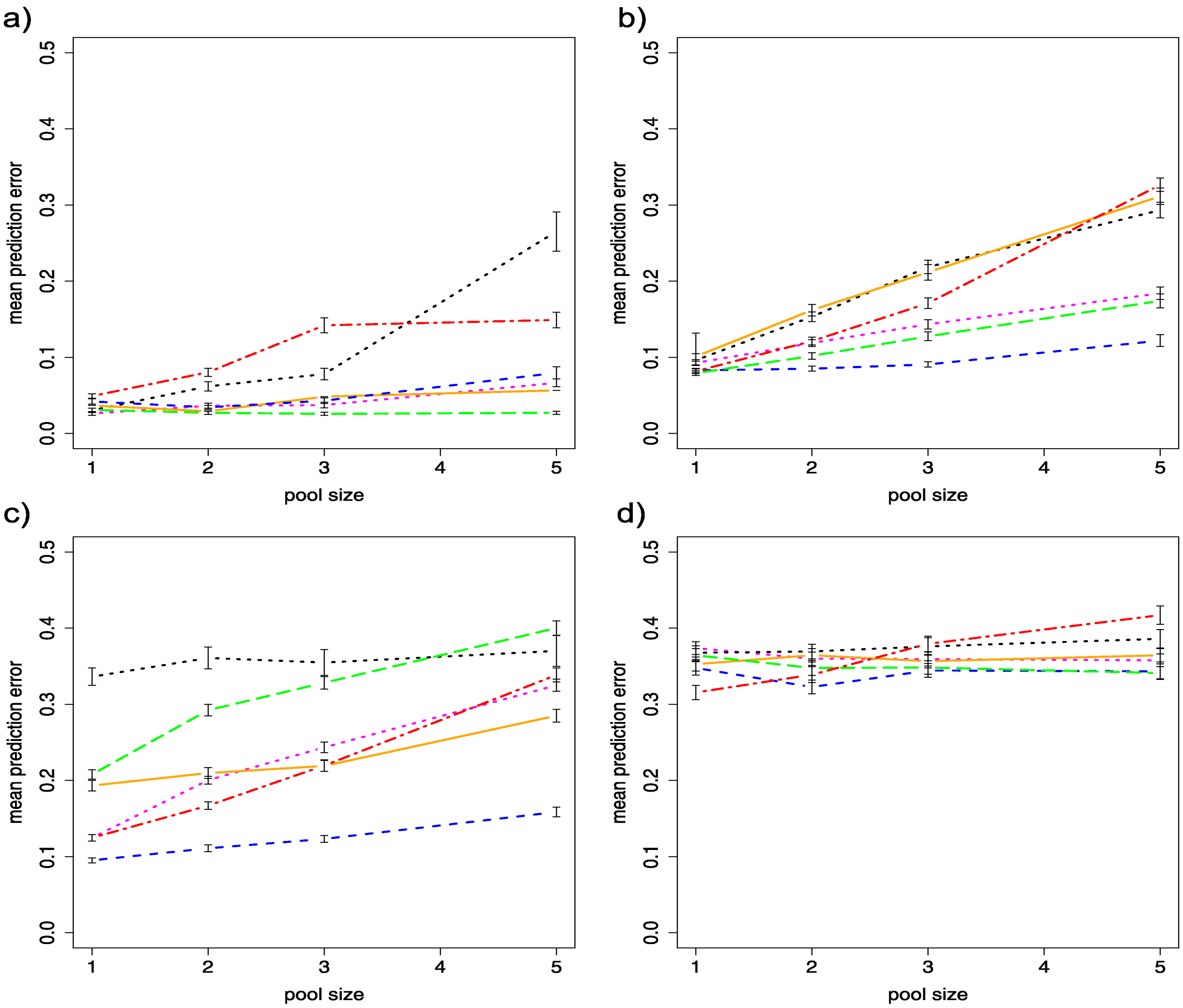
- svm (linear, radial)
- Random Forest
- t-test-filter + LDA
- (P)PLS-DA + LDA

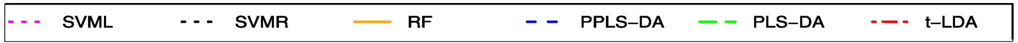
simulation results – prediction error



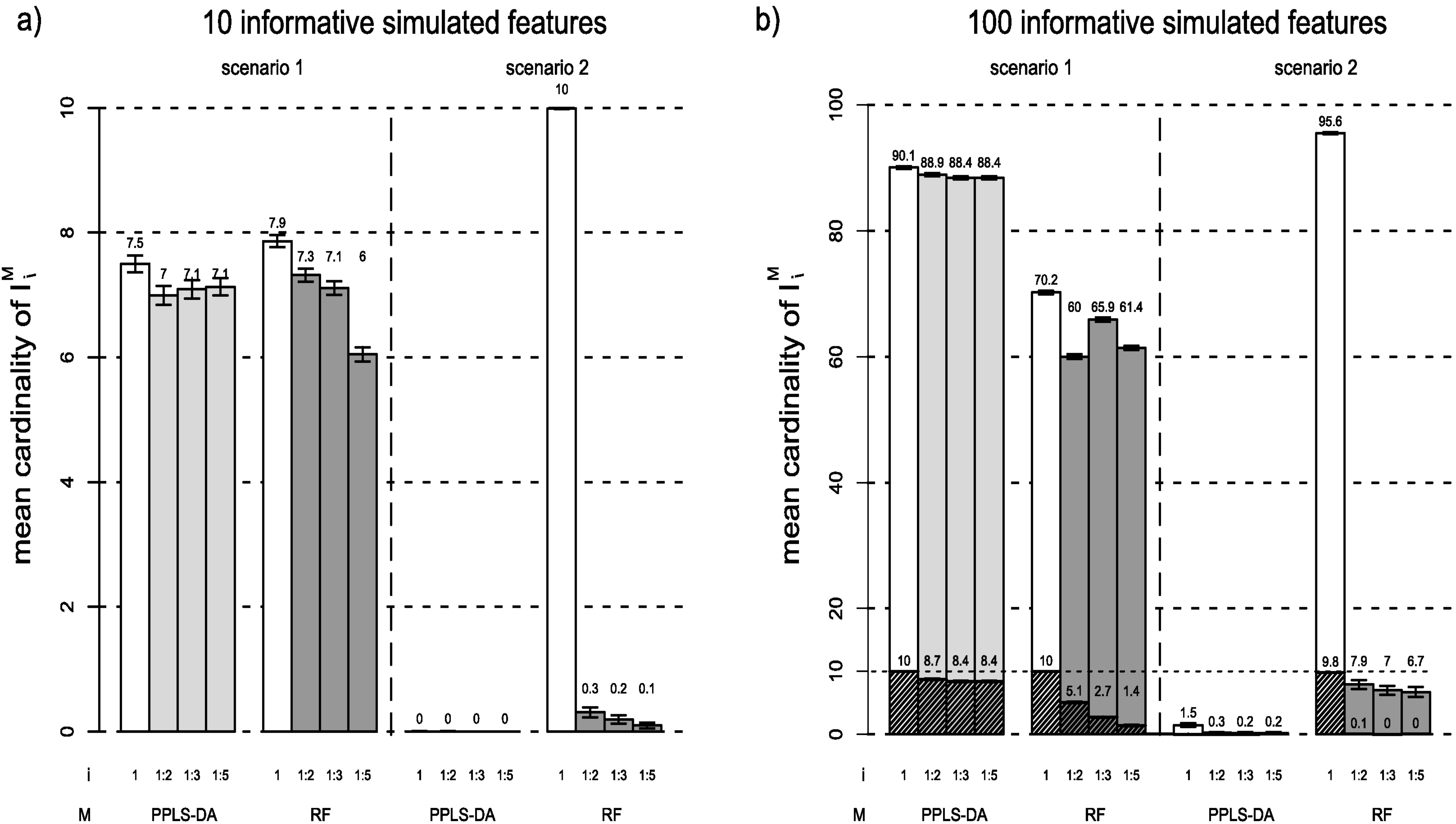
scenario 1: differentially expressed genes

experimental results – prediction error





simulation results – feature recovery



take home

take home:

- avoid heterogeneous tissues and avoid sample pooling !

take home:

- avoid heterogeneous tissues and avoid sample pooling !

- if not avoidable:

  - look for huge effects
  - try source decomposition methods
  - try methods robust for pooling effects

take home:

- avoid heterogeneous tissues and avoid sample pooling !

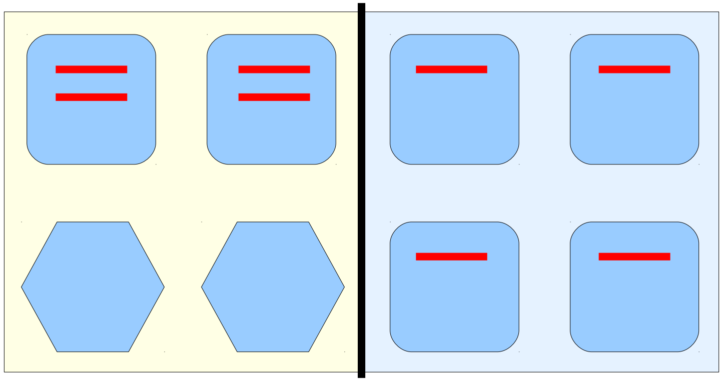
- if not avoidable:

    - look for huge effects
    - try source decomposition methods
    - try methods robust for pooling effects

- validate honestly!

references

- Venet et al.: Separation of samples into their constituents using gene expression data. Bioinformatics 2001, 17, S1, S279-S287
- Lahdesmaki et al.: In silico microdissection of microarray data from heterogeneous cell populations. BMC Bioinformatics 2005, 6, 54ff
- Stuart et al.: In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. PNAS 2004, 101, 615-620
- Ghosh: Mixture models for assessing differential expression in complex tissues using microarray data. Bioinformatics 2004, 20: 1663-1669
- Kerr, M. K. (2003). Design considerations for efficient and effective microarray studies. Biometrics 59(4), 822–8.
- Repsilber et al.: Biomarker discovery in heterogeneous tissue samples – taking the in-silico deconfounding approach. BMC Bioinformatics 2010, 11:27
- Telaar et al.: Biomarker discovery: Classification using pooled samples – A simulation study. Journal of Computational Statistics, submitted 2011



- taking questions!
- repsilber@fbn-dummerstorf.de