

Normalization issues when designing small diagnostic biomarker panels from large scale clinical microarray studies

Jochen Jäger

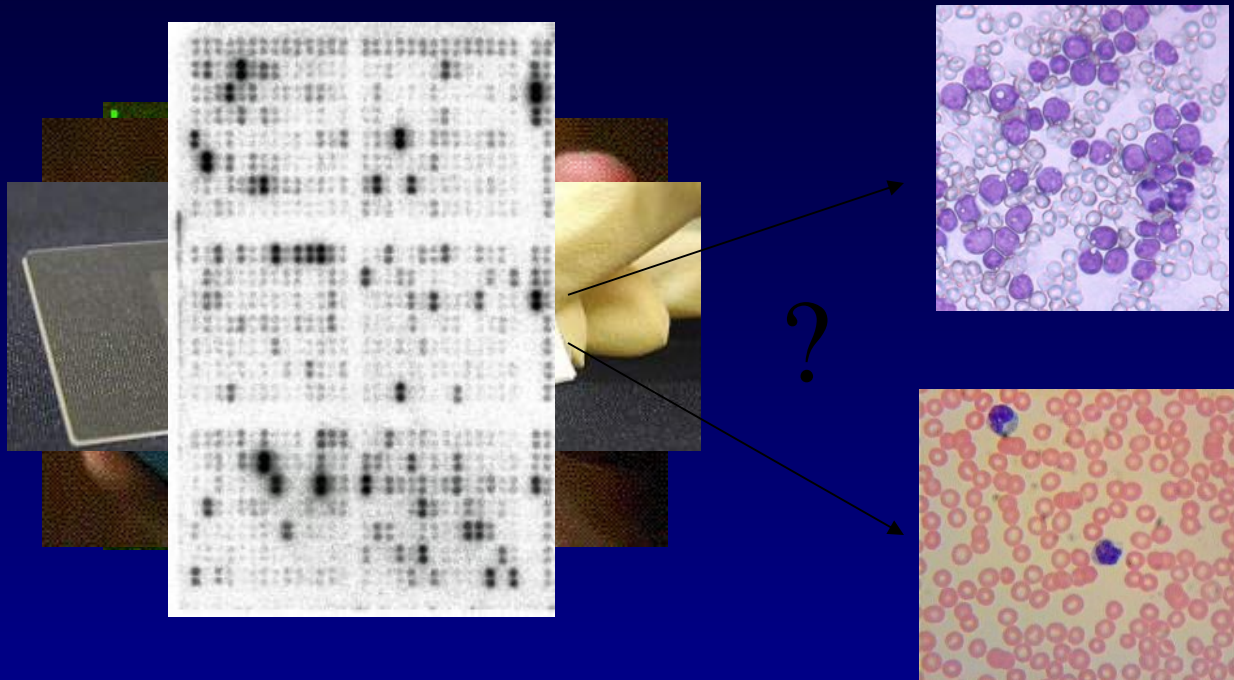


Max Planck Institute for Molecular Genetics
Dept. Computational Molecular Biology,
Berlin, Germany

Now: Hamilton Bonaduz AG
BU Robotics
CH-7402 Bonaduz
Switzerland

Motivation

Microarray expression profiles to characterize cell tissues

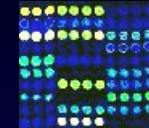
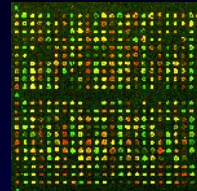
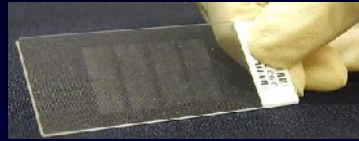


Most common causes of death

1. Heart Disease 31%
2. Cancer 23%
3. Cerebrovascular Disease (Stroke) 7%
4. Respiratory disease (COPD) 5%
5. Accidents 4.2%
6. Pneumonia and Influenza 3.9%
7. Diabetes 2.8%
8. Suicide 1.3%
9. Nephritis 1.1%
10. Liver Disease 1.1%
11. Other 20%

Trend: less cardio-vascular, slightly more cancer

Problem Statement



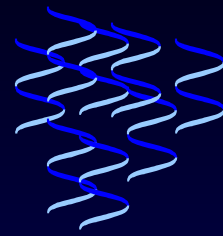
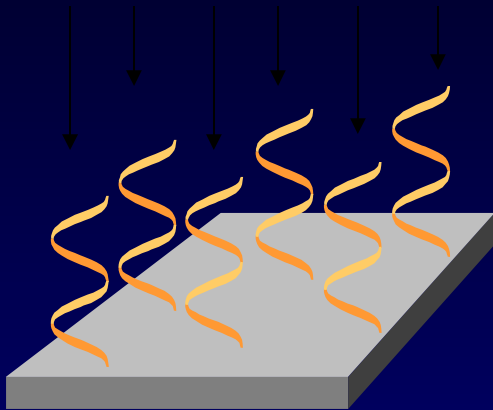
- Scenario: Big clinical study (disease \leftrightarrow control) with many, comprehensive microarrays aiming for a diagnostic chip
- Goal: reliable disease type prediction with smaller and thus cheaper and faster biomarker panels
- **Question 1: Can we switch from a large microarray to a small marker panel?**
- **Question 2: How many patients do we need?**
- **Question 3: How many markers do we need?**
- **Question 4: How do we select markers?**
- **Question 5: How do we deal with normalization issues?**

Microarrays

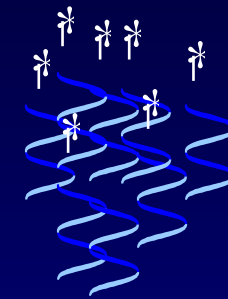
DNA 

select
genes

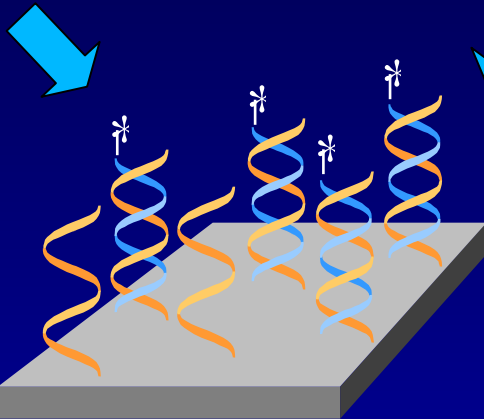
spot
genes



extract cDNA



label cDNA



Annealing phase

Normalization

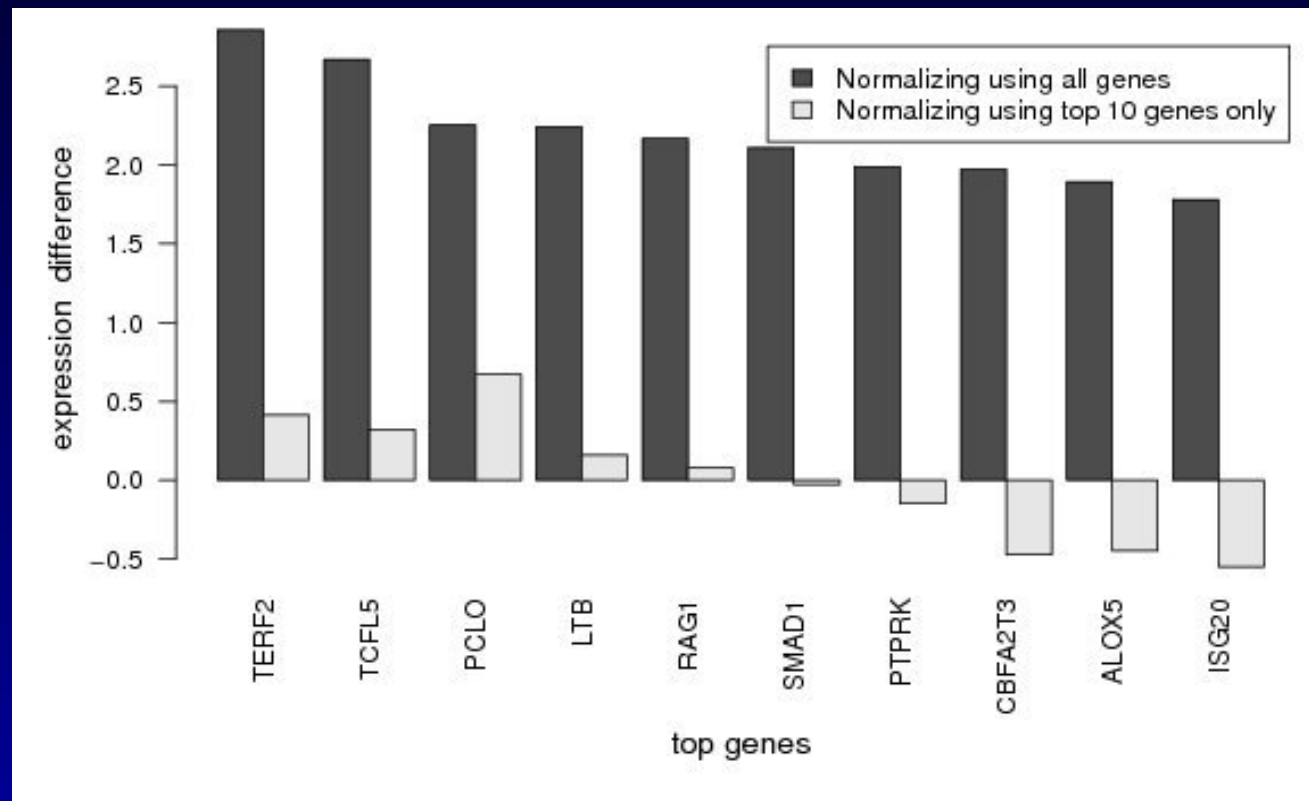
- Needed to bring microarrays on a comparable scale
- Remove technical artifacts but don't touch biological signal

Normalization issues

- Normalization methods inherently rely on assumptions, e.g.:
 - control genes not deregulated
 - majority of genes not deregulated
 - sum of all genes similar for all samples
- When normalizing a whole genome chip most of the assumptions seem valid
- For a diagnostic chip this does not necessarily hold

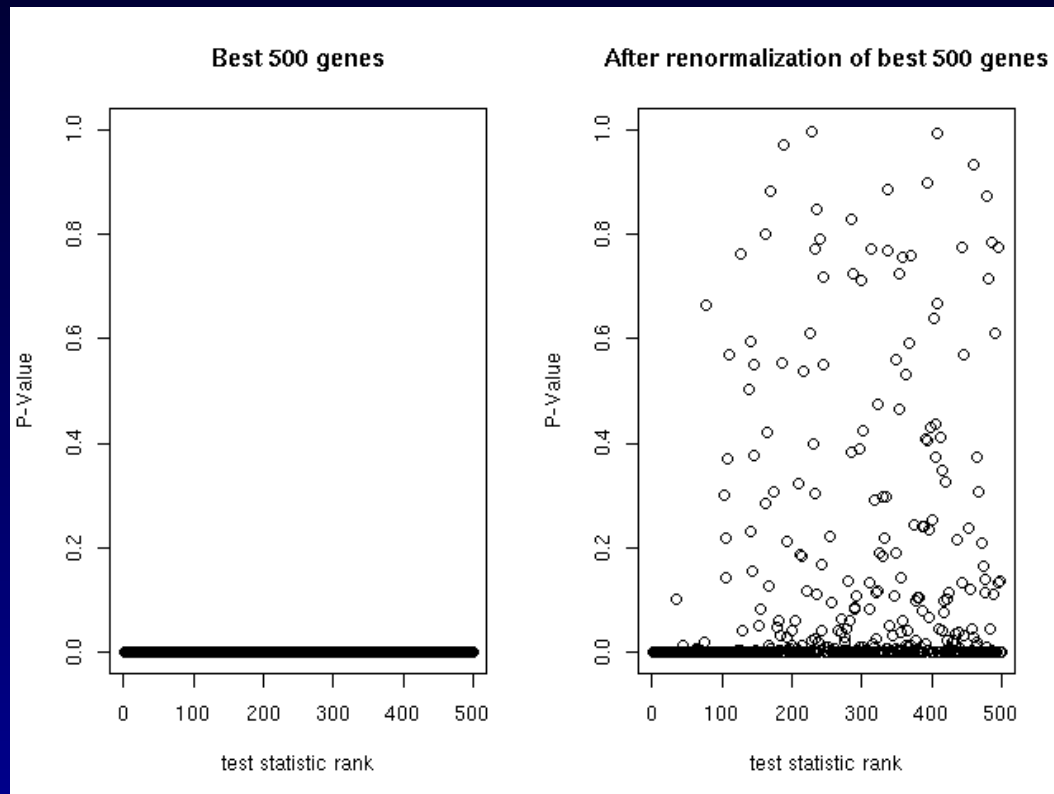
Problem Illustration

- Effects (on a log scale) for the 10 genes with the maximal difference between TEL versus BCR and E2A in the leukemia dataset (Yeoh et al. (2002)). Comparing a normalization using all genes and normalization on a diagnostic chip that uses just these top 10 genes.



Motivation Normalization effects

Leukemia BCR-ABL vs E2A-PBX1, 500 top t-scoring

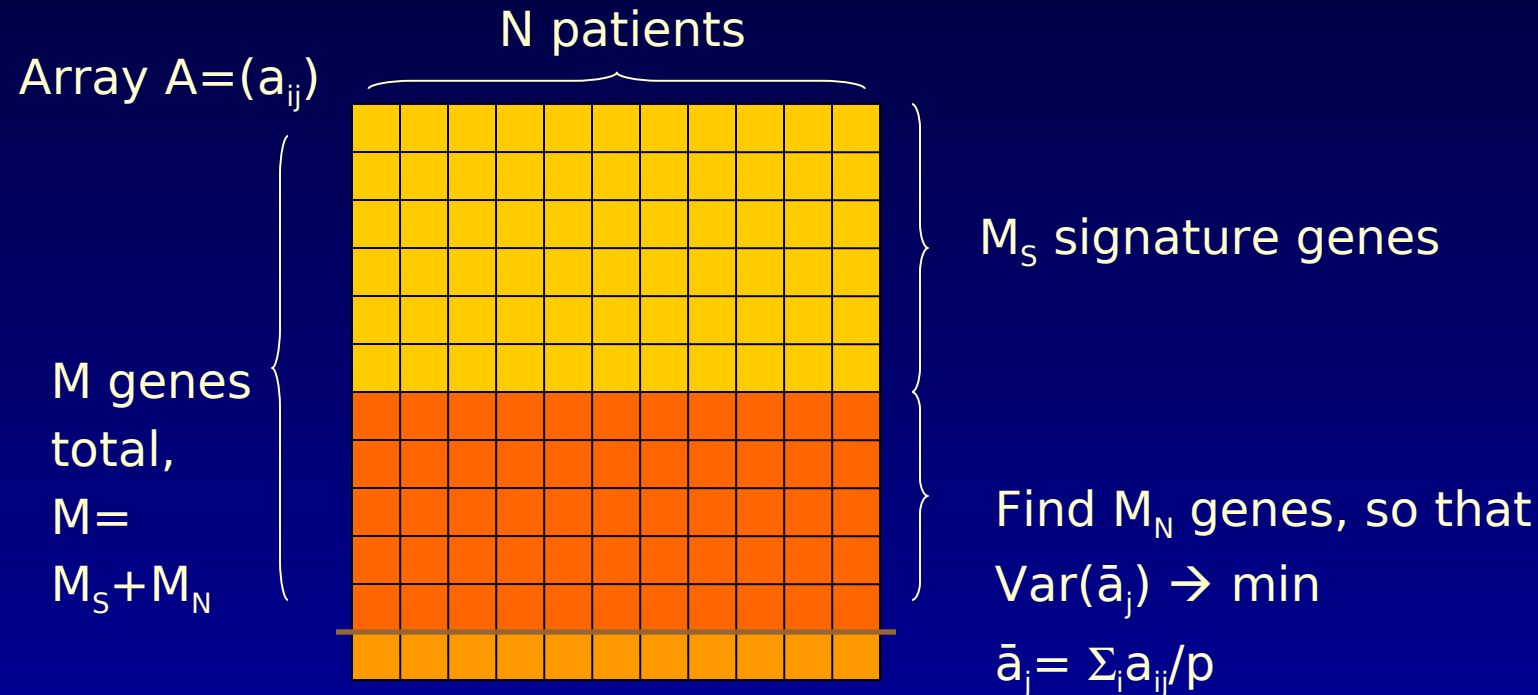


Solutions

- Add some more genes on the diagnostic chip which can be used for normalization
- Choices can be:
 - genes with smallest effect size
 - genes with smallest variance
 - genes with high variance, small effect
 - genes with low variance, high mean
 - balanced genes (will dive into this deeper)

Balanced genes

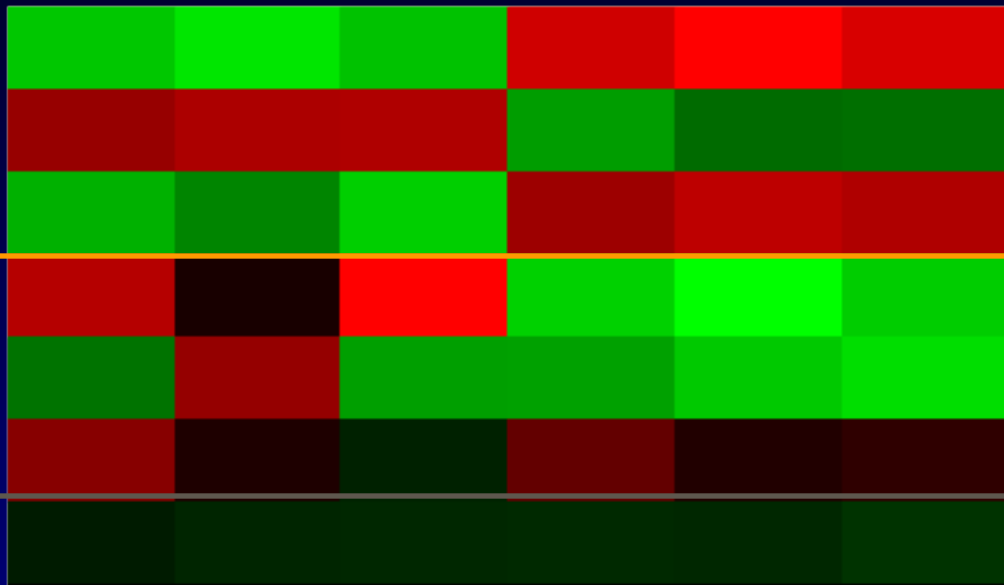
- Select additional genes so that the sums of the genes on the diagnostic chips of all samples become similar



Normalization Example

Grp A

Grp B



3 signature genes

Mean expression

Algorithm

Greedy forward selection:

Let: D be the given set of genes used for classification
(differential genes)

$N = \{ \}$, the set of genes used for normalization
(control spots)

$minvar = \infty$

for $c = 1..k_n$ (for each control spot)

for $g = 1..K$ (for each gene on the large chip)

calculate $v_g = \sum_{j \in A \cup B} (\bar{x}_j - \bar{x})^2$

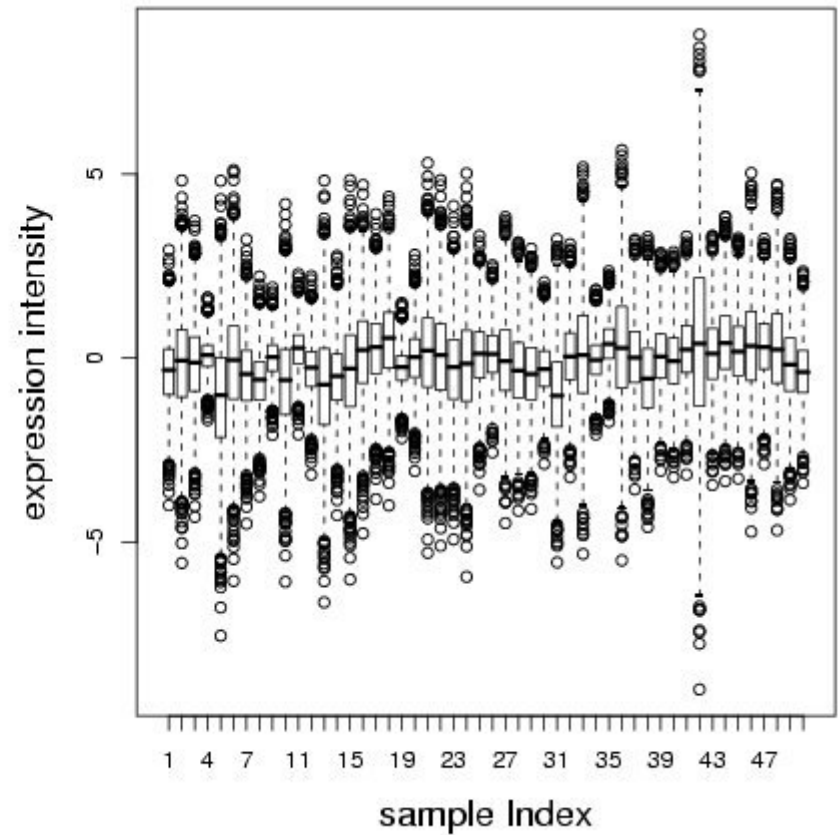
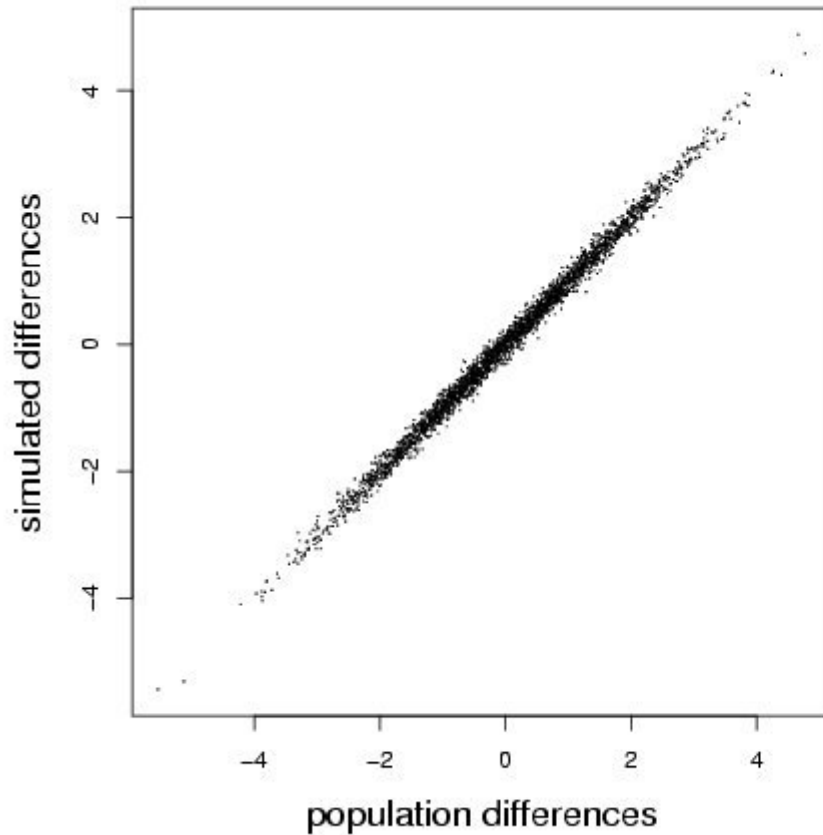
if ($v_g < minvar$) $minvar = v_g$; $bestgene = g$

$N = N \cup \{bestgene\}$

Evaluation on simulated data and real data

- Simulation of 3000x50 matrix: multivariate normal distribution
 2. generate covariance matrix Σ with inverse wishart
 3. generate underlying means μ with random draws from $N(0,1)$
 4. Draw from multivariate $N(\mu, \Sigma)$ distribution
 5. Add additive and multiplicative noise $N(0,1)$
 6. Generate train and testset

Simulated Data

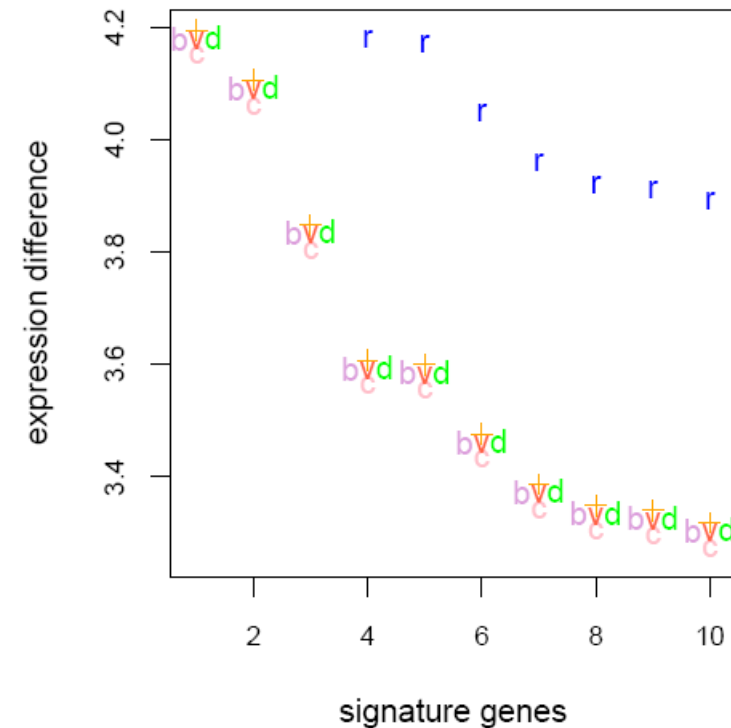
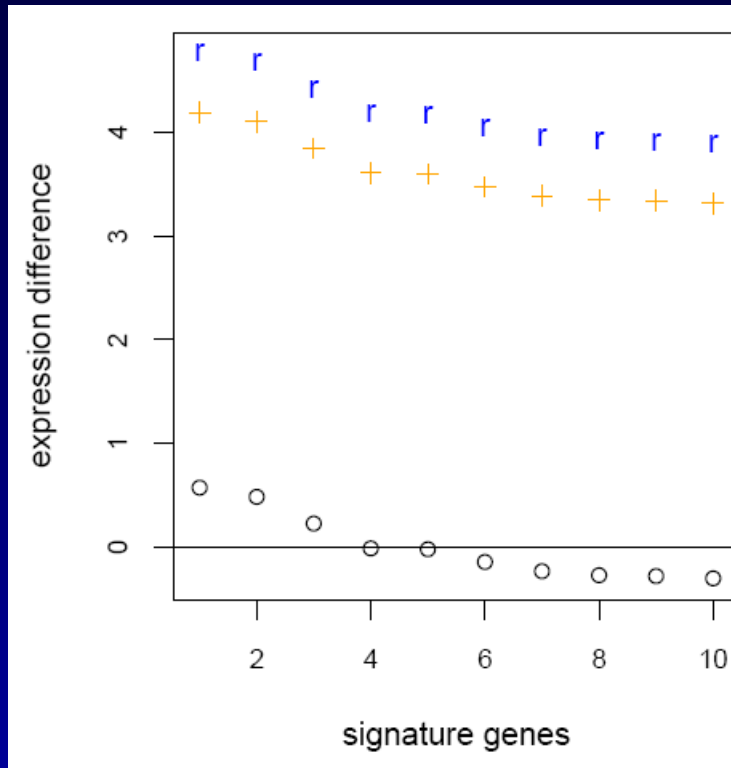


Results Simulation

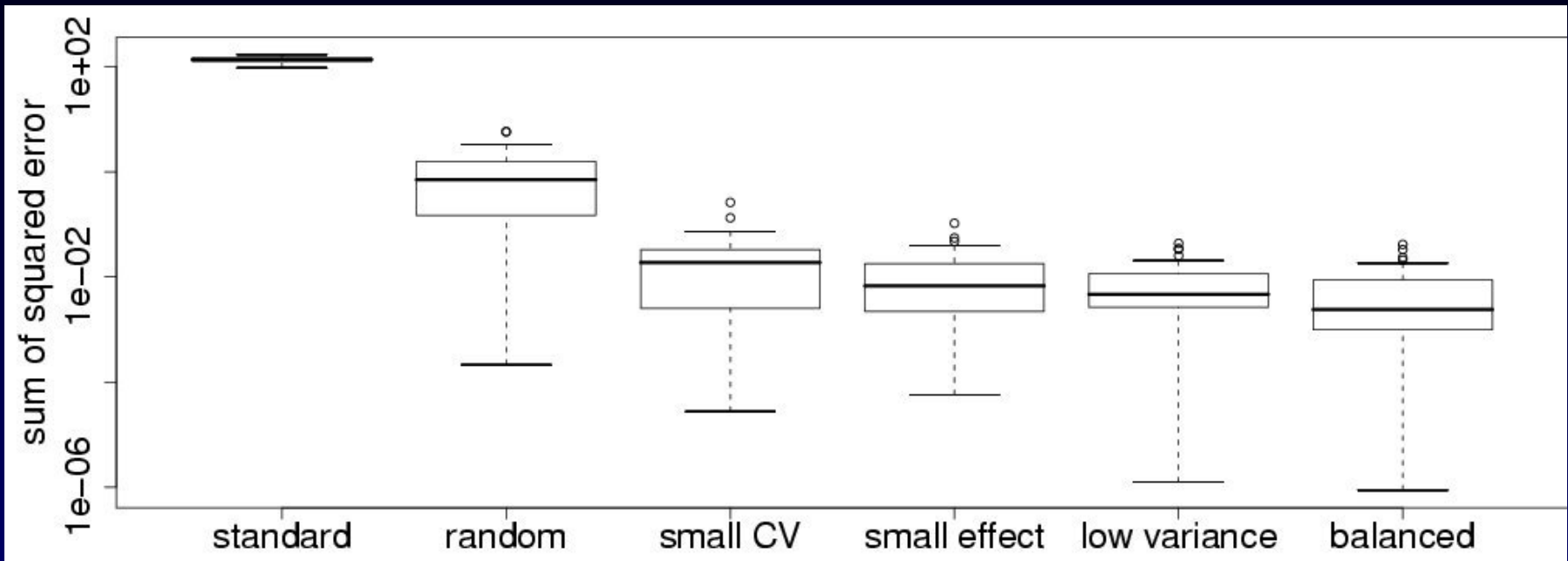
Effect of the 10 spotted genes for the testdata

Normalization with:

- O: just 10 genes
- b: balanced genes
- d: least effect
- +: all genes
- v: least variance
- r: random genes
- c: least COV



Results simulated data



Repeat data simulation 20 times

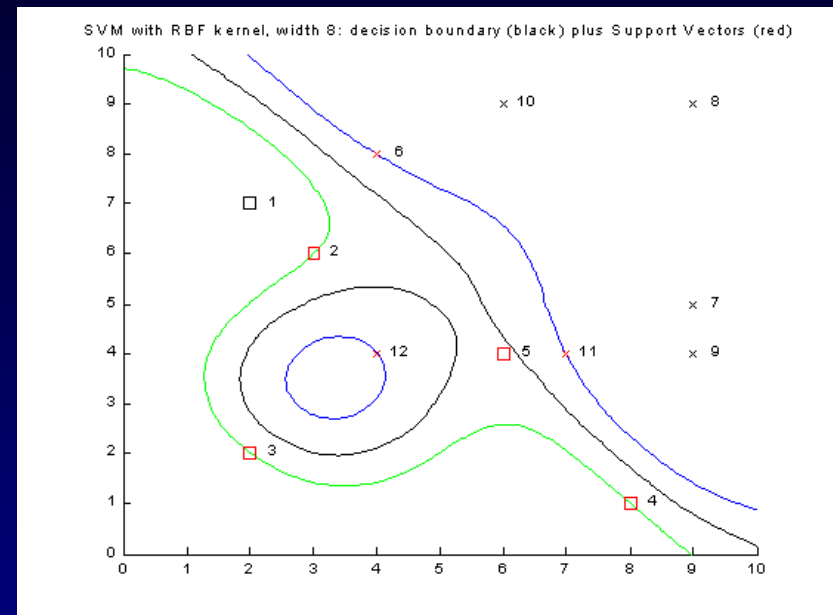
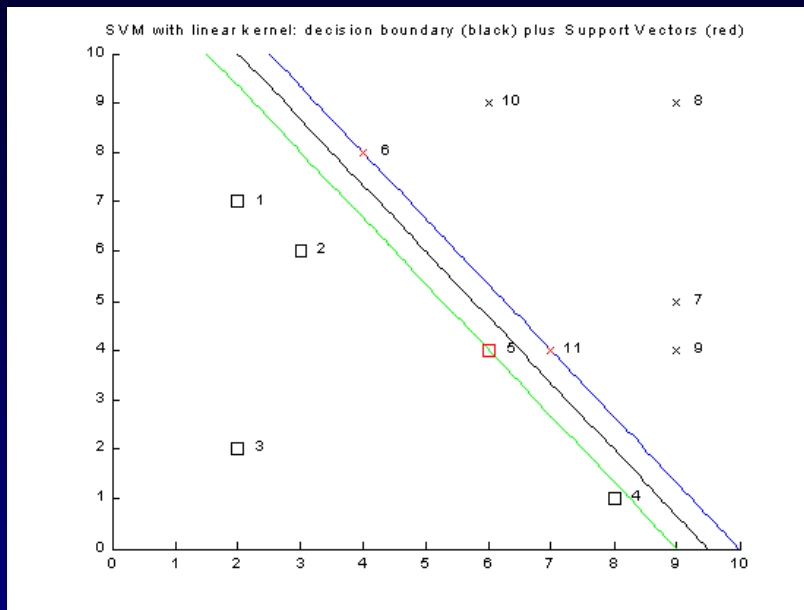
Sum of squared error of the distance of the models to the model using all genes for normalization

Real data application

- Leukemia data (Yeoh et al 2002)
- Use MCRestimate package for determining the number of signature genes needed (5)
- Design virtual array
- Randomly split data equally 100 times in test and trainset and evaluate performance with LOOCV SVM

Support Vector machines

- Find separating hyperplane with maximal distance to closest training example

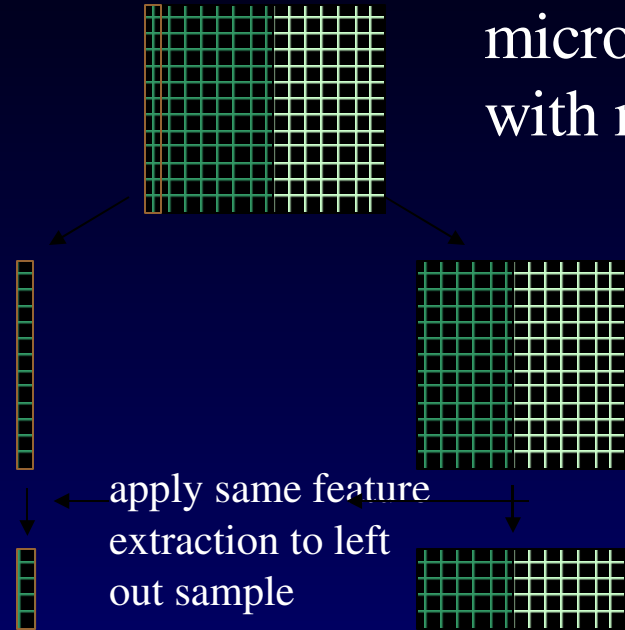


- Advantages:
 - avoids overfitting
 - can handle higher order interactions and noise using kernel functions and soft margin

Evaluation with LOOCV

Repeat for each of
 the n examples:
 leave out one
 sample

microarray data: n examples
 with m expression levels each

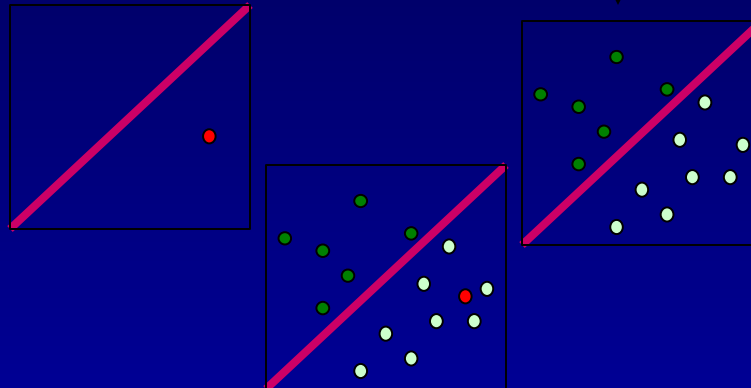


train data

extract features

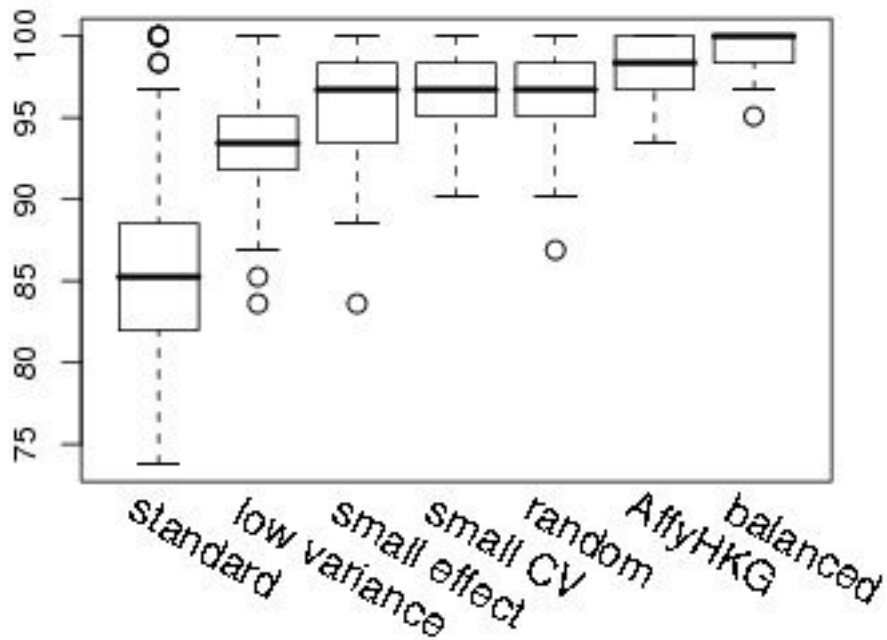
train learner

classify held-out
 sample

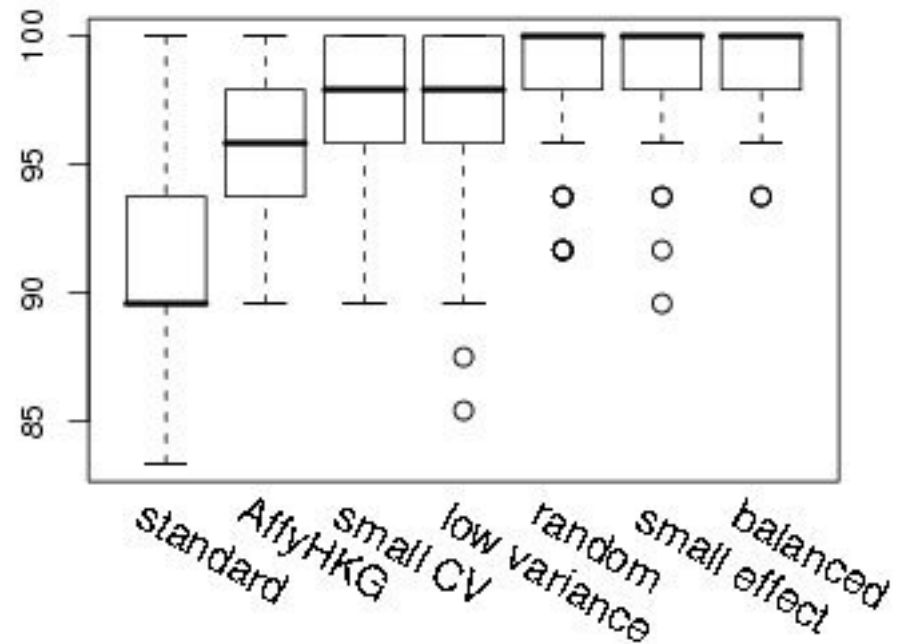


Results on 2 real datasets

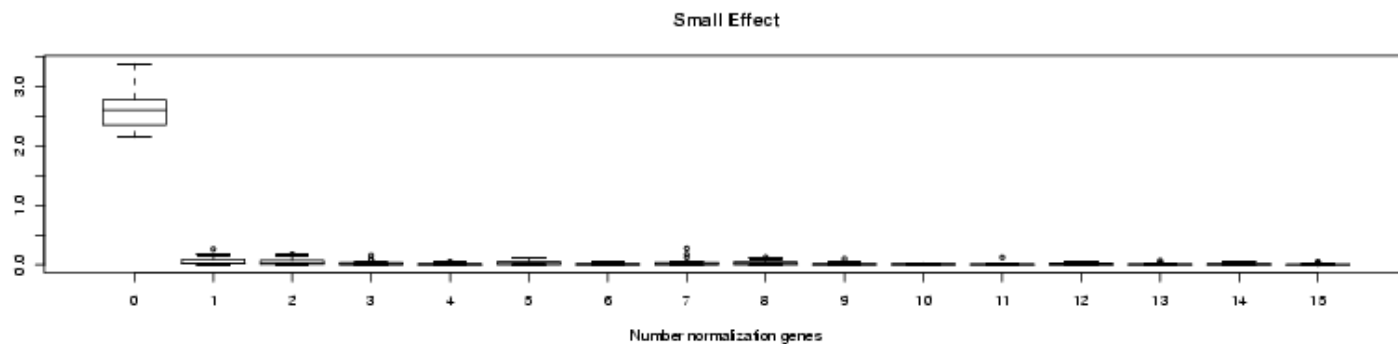
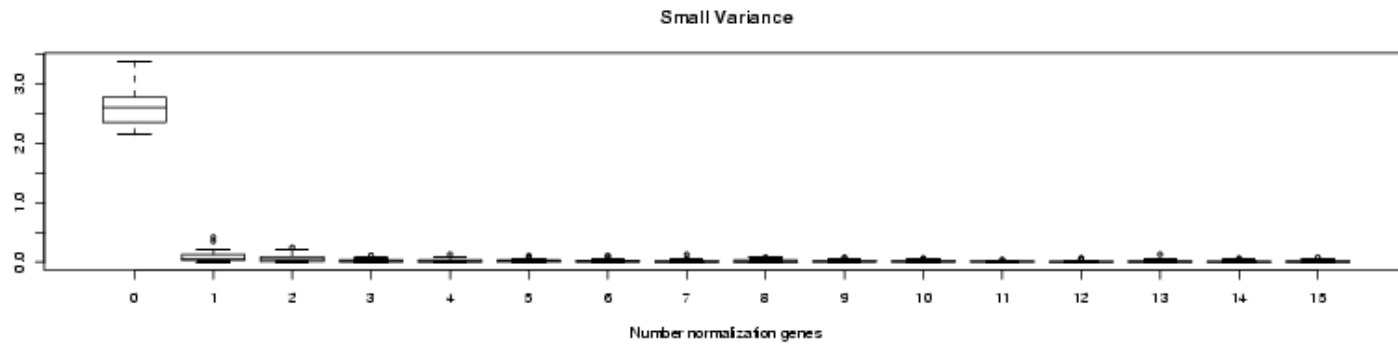
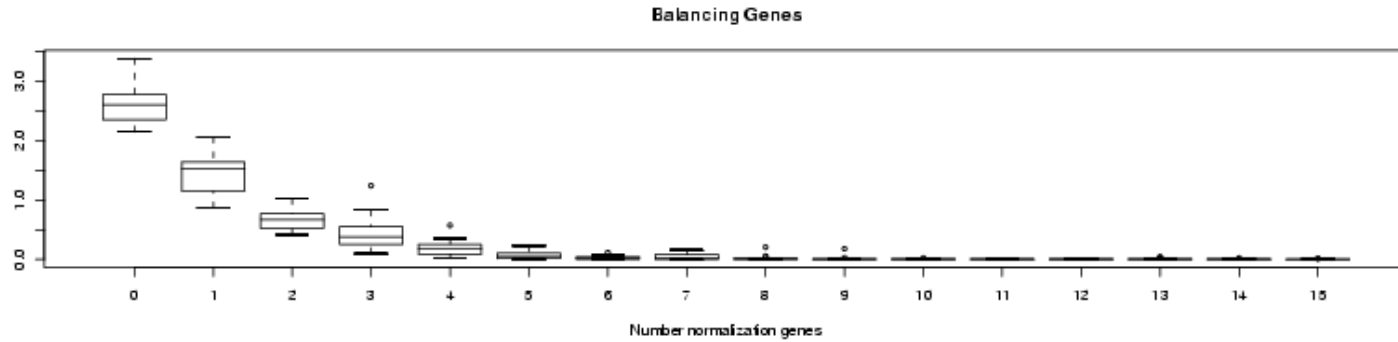
Leukemia: TEL vs Rest



Lung: Normal vs Tumor



How many normalization genes



Discussion

- How many genes are really necessary as diagnostic and normalization genes?
- What happens when changing platforms (e.g. from microarray to PCR)?
- Can we combine several diagnostic chips for different diseases into one?