

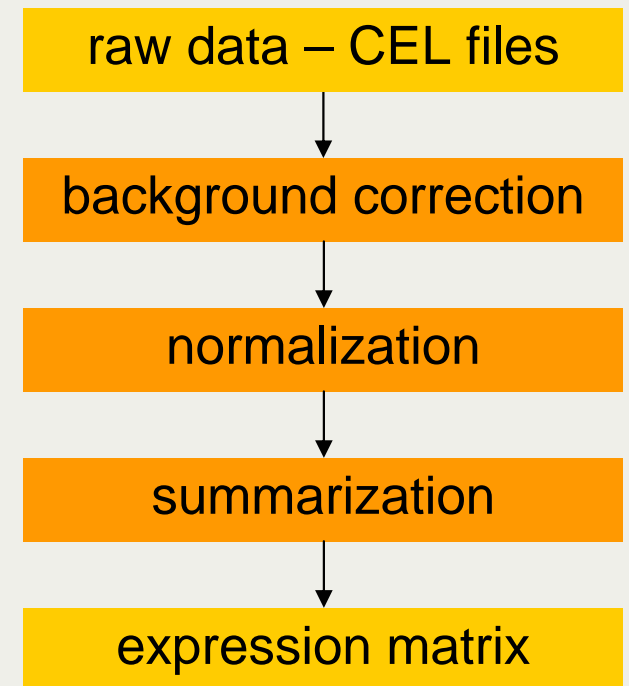
# affyPara: Parallelized preprocessing algorithms for high-density oligonucleotide array data

Markus Schmidberger  
Ulrich Mansmann



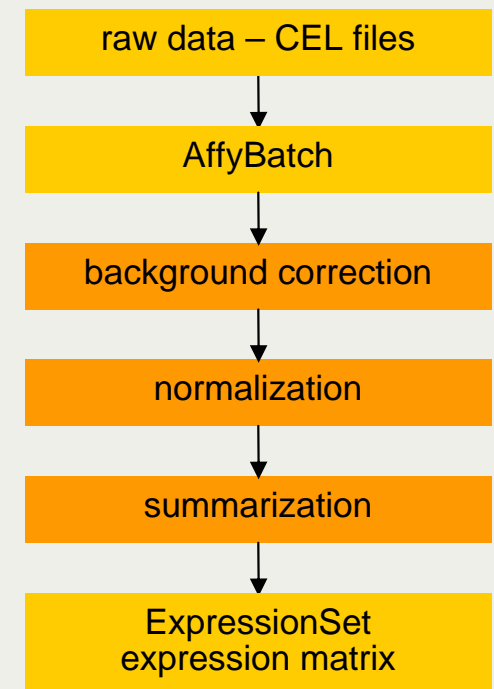
# Preprocessing

- **Background correction**
  - remove noise of optical detection system
- **Normalization**
  - make measurements comparable from different array hybridizations
- **Summarization**
  - transcripts are represented in multiple probes
- R and BioConductor mostly used in research



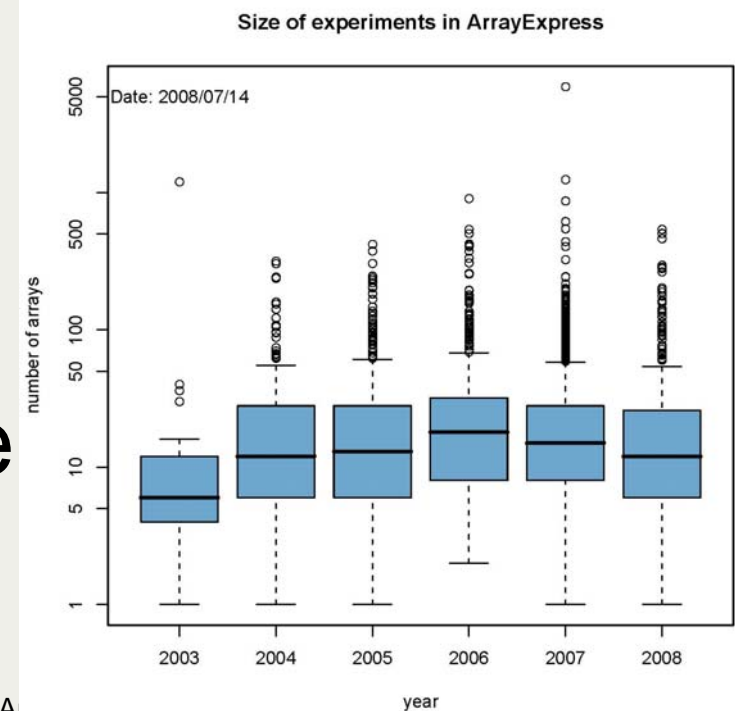
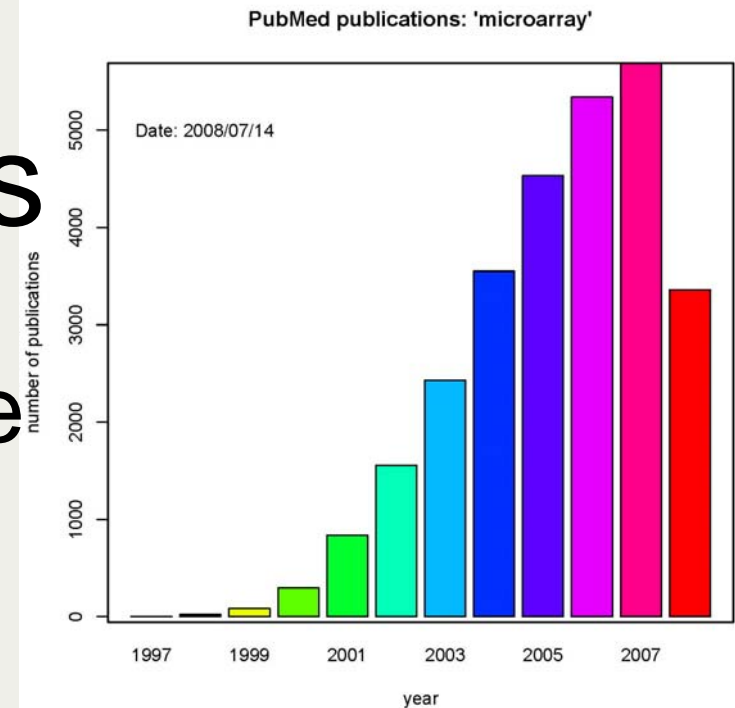
# Problems

- **Data-structure of R**
  - data are stored in class 'AffyBatch'
  - complex class with a lot of different slots
  - AffyBatch is memory intensiv
- **Performance of algorithms**
  - Inefficient program structure
  - Long computation time



# Challenges

- Microarray experiments more and more popular
- Microarray chips become cheaper
  - Experiments grow in size
  - EBI experiment: 6000 Arrays
- Microarray chips grow in size
  - More genes per chip



# Possible Solutions

- Business applications
  - Expensive, not adaptable
- Faster and bigger computers
  - Expensive, limited
  - Main memory 256 GB: 60t €
- Better coding
  - C, DB, hard disk
    - hard disk as main memory -> aroma.affymetrix (Bengtsson)
- **Distribution to several computers / processors**
  - Concurrent calculation of parts at different processors
  - Main memory 8 GB: 2000 € -> 60t € = 30 computers

# Parallelization



- Multiprocessors
  - the use of two or more central processing units (CPUs) within a single computer system
  - Today: Two-processors get a standard for workstations
  - OpenMP
- Multicomputers = Cluster
  - different parts of a program run simultaneously on two or more computers that are communicating with each other over a network
  - Computer, network, software
  - MPI

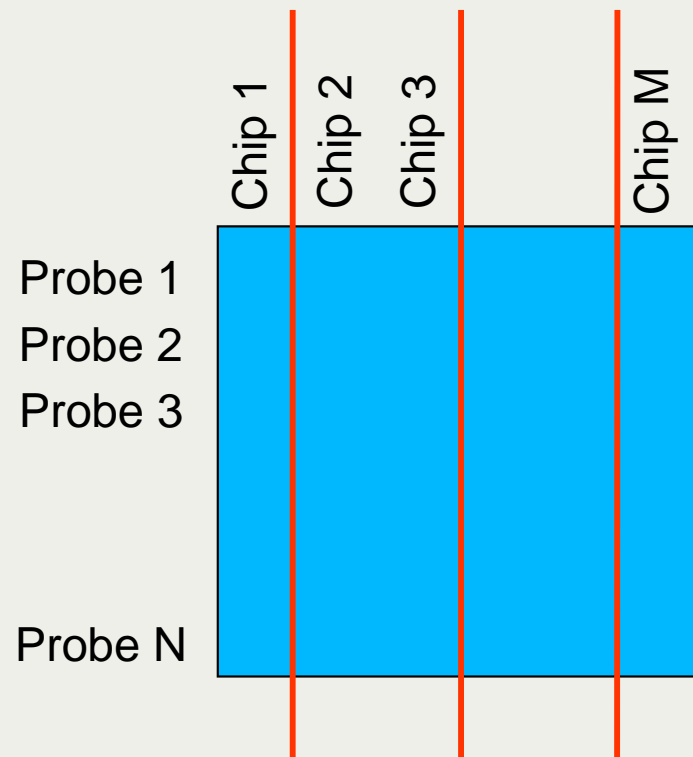
# Software: MPI

- **Message Passing Interface**
- MPI is an API for parallel programming based on the message passing model
- MPI processes execute in parallel
- MPI is a standard for libraries
- Libraries exists for
  - FORTRAN, C, C++
  - R: **Rmpi**, **Snow**, papply
- IBE Cluster: LAM/MPI 7.1.3



# Decomposition of AffyBatch

- AffyBatch = intensities from multiple arrays



- Which **decomposition** is the best ?

- Partition by chips
- Partition by probes
- Partition of CEL file name list

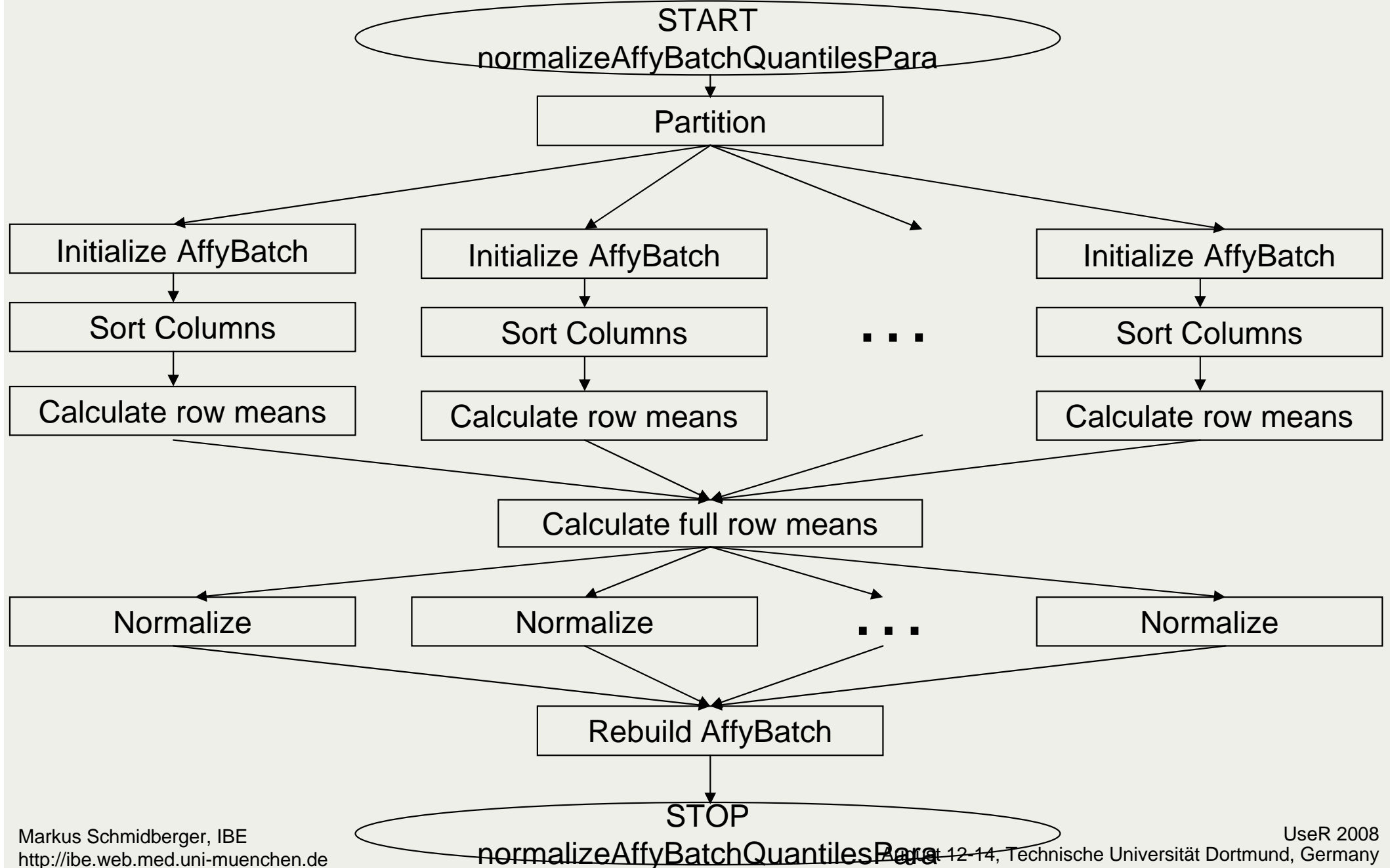
- **Communication Overhead**

- A lot of data to transfer
- Create AffyBatches at nodes
- Complete preprocessing method: `preproPara()`



- Scaling = constant
- Non-linear = invariantset
- Quantile
- cyclic loess

# Quantile Normalization



# affyPara – Code Usability

```
R> library(affy)
R> AB <- ReadAffy()
R> AB_bgc <- bg.correct(AB, method="rma")
R> AB_norm <- normalize.AffyBatch.quantiles(AB_bgc, type="pmonly")
```

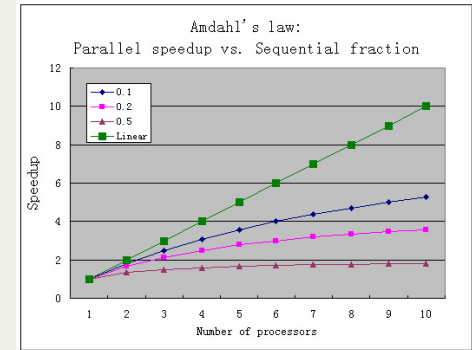
```
R> library(affyPara)
R> c1 <- makeCluster(5, type="MPI") # type="nws"
R> AB <- ReadAffy()
R> AB_bgc <- bgCorrectPara(c1, AB, method="rma")
R> AB_norm <- normalizeAffyBatchQuantilesPara(c1, AB_bgc, type="pmonly", verbose=TRUE)
R> stopCluster(c1)
```

```
Build hard disk file structure (/rawdata, /annotationData)
R> library(aroma.affymetrix)
R> cdf <- AffymetrixCdfFile$fromChipType("HG-U133A")
R> cs <- AffymetrixCelSet$fromName(name, tags, chipType=cdf)
R> bc <- RmaBackgroundCorrection(cs)
R> csBC <- process(bc)
R> AB_bgc <- extractAffyBatch(csBC)
```

# Package **affyPara**

- BioConductor Package **affyPara** with parallelized affy-functions
  - Version 1.1.7
  - Solves main memory problems
  - More CEL Files preprocessable
    - IBE Cluster: ~ 16.000 microarrays
  - Speedup
- Parallelization methods produce in view of machine accuracy the **same results** as serialized methods.
  - All.equal(), machine's precision.

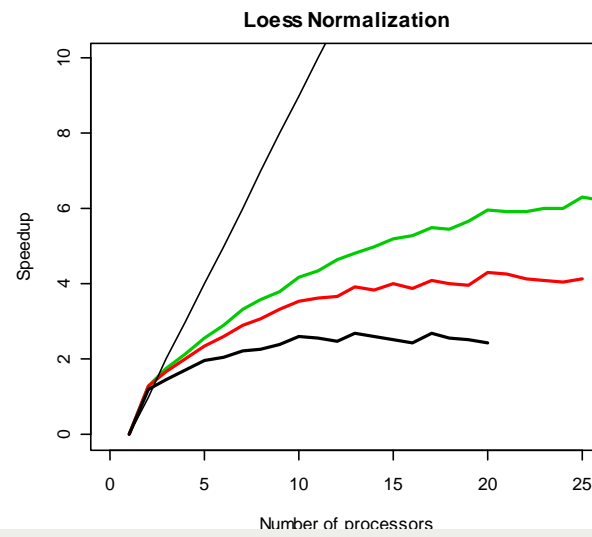
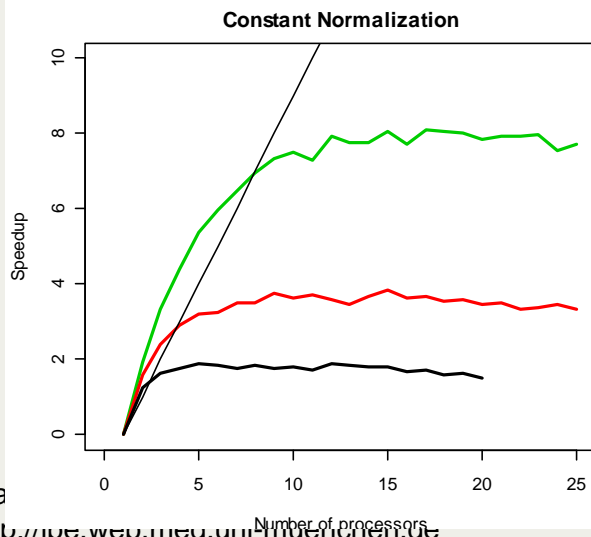
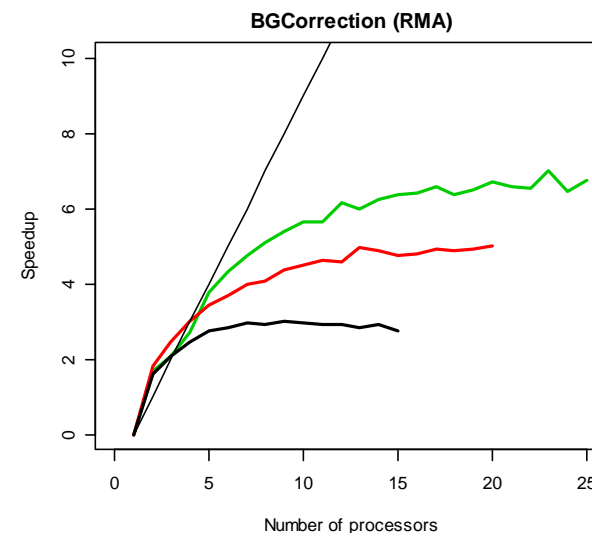
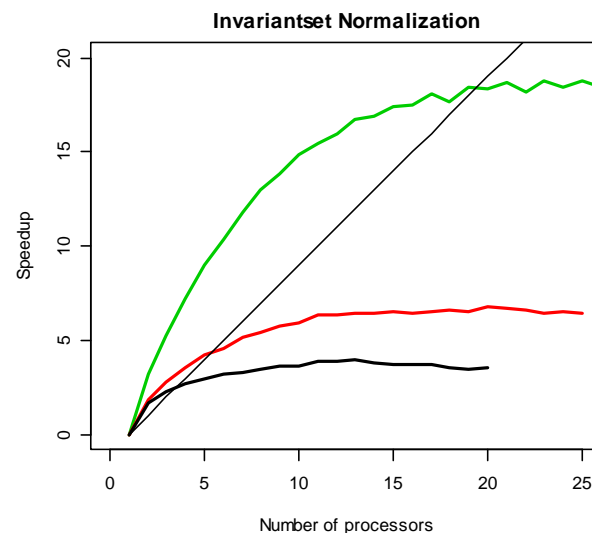
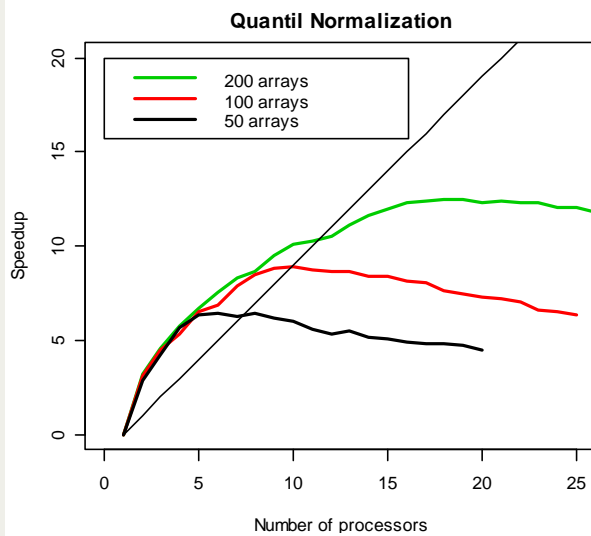
# Results – Speedup



- Speedup of the methods up to factor 15

$$Sp = T_1 / T_p$$

$$Sp \sim 1 / [s + p/N]$$

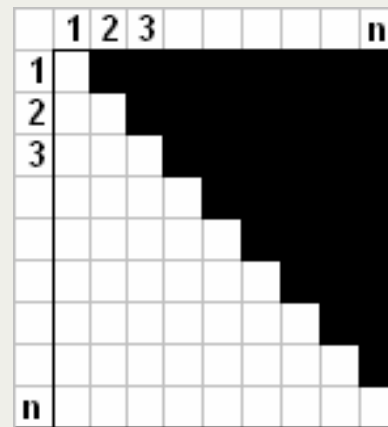


# New methods based on parallelization strategies

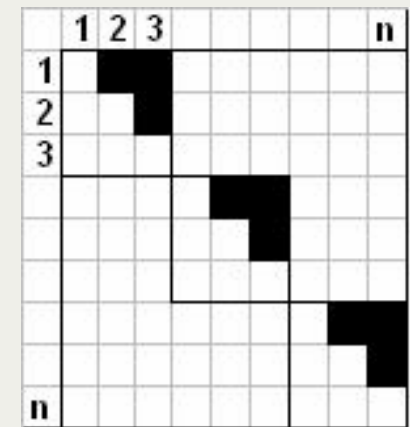
- **Partial Cyclic Loess Normalization with Permutation**

- Array Permutations
- ~ 75% of complete loess normalization
- Same (good) results
  - Using Boxplot and Histogramm
- Speedup: 6-7 (100 arrays)

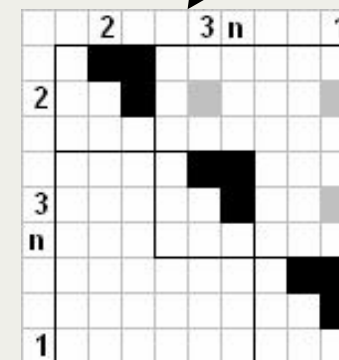
Complete cyclic loess



Partial cyclic loess on 3 nodes



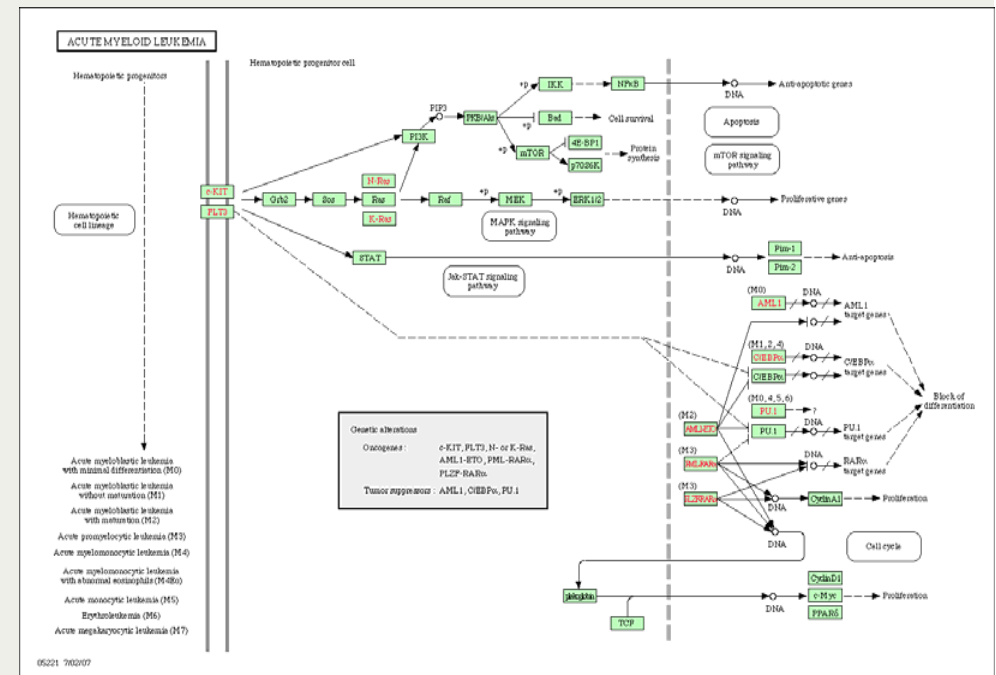
Permutations of Arrays 2-3 times



# Large project in applied bioinformatics in preparation

- Collecting cancer data from public libraries
  - ArrayExpress, GEO, ...
  - > more than **5000** microarrays ??

- Preprocessing **all together**
- Analyzing **all together**



# Collecting data from public libraries - Results

HG-U133A	# cancerous Arrays
Lymphoma	673
Colon	128
Breast	2109
Prostate	300
Lung	256
Leukaemia	1163
SUM	4629

Second / Referenz Data Set:

- **expO**  
(Expression Project For Oncology)
- cancer patients from the expO project
- 1973 HG-U133 Plus 2.0 Chips

# Ideas for analyzing Differences in gene interaction

Comparing networks



modelling and estimating gene interaction

Cancer 1 / Group 1

Cancer 2 / Group 2

Cancer 3 / Group 3

Normalization all together





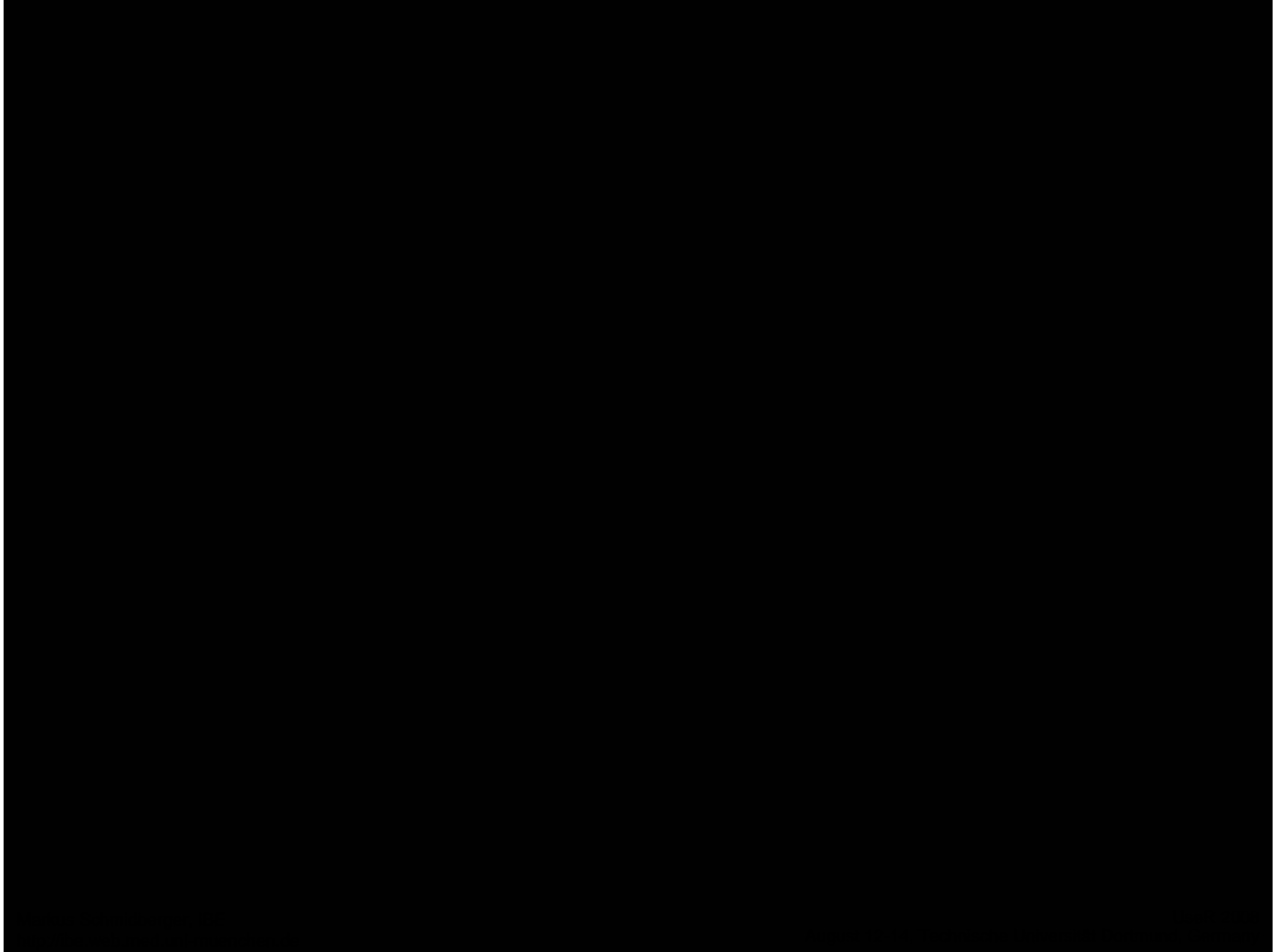
# Parallelization in R & BioC

- Multiprocessors  $\leftrightarrow$  pnmath, ?? ROMP ??
  - Multiprocessors available for everyone
  - Difficult to use in packages
  - **Good for R base**
- Multicomputers  $\leftrightarrow$  Rmpi, snow, papply, ( pvm )
  - Cluster not available for everyone
    - Cloud Computing  $\leftrightarrow$  ?? AmazonEC2 ??
  - Cluster Management necessary
    - sfCluster or slurm or ?? RSunGridEngine ??
  - **Good for R packages**
- GPU
  - New and promising technology
  - Probably available for everyone (graphic board - cheap)
  - NVIDIA CUDA  $\leftrightarrow$  ?? RCUDA ??

affyPara:  
Parallelized preprocessing algorithms  
for high-density oligonucleotide array data  
&  
Large applied study for evaluation

M. Schmidberger, U. Mansmann; *Parallelized preprocessing algorithms for high-density oligonucleotide array data*; 22th International Parallel and Distributed Processing Symposium (IPDPS 2008), Proceedings, ISBN: 978-1-4244-1693-6, 14-18 April 2008, Miami, FL, USA. IEEE 2008

**Thanks for your attention**



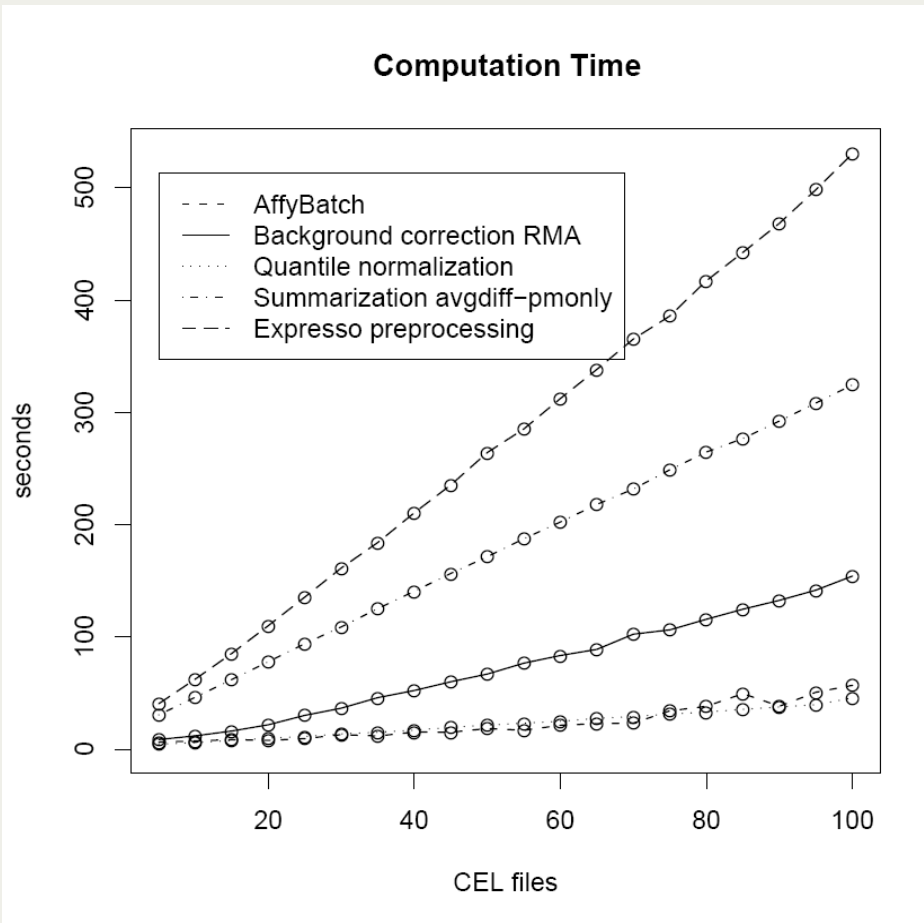
# Literature R and Preprocessing

- Gentleman et. All; Bioinformatics and Computational Biology Solutions, *Using R and Bioconductor*; Springer, 2005 (Statistics for Biology and Health)
- Berrar et. All.; *A Practical Approach to Microarray Data Analysis*; Kluwer Academic Publishers, 2004
- Bolstad, Irizarry, Astrand, Speed; *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias*; *Bioinformatics*, 2003, 19(2), 185-193
- Irizarry, Wu, Jaffee; *Comparison of Affymetrix GeneChip expression measures*; *Bioinformatics*, 2006, 22 no. 7, 789-794

# Literature Parallel Computing

- Sloan, J. D.: *High Performance Linux Clusters with OSCAR, Rocks, OpenMosix, and MPI* O'Reilly, 2004
- Tierney, Luke: *Code Analysis and Parallelizing Vector Operations in R*. In: DSC 2007. Auckland, New Zealand, February 2007
- Rossini, Anthony: *Simple Parallel Statistical Computing in R*. In: UW Biostatistics Working Paper Series 193 (2003)
- Sevcikova, Hana: *Statistical Simulations on Parallel Computers*. In: Journal of Computational and Graphical Statistics 13 (2003), Nr. 4, 886-906.

# Problems



**How many arrays can I RMA process?**  
(Ben Bolstad)

<http://bmbolstad.com/misc/ComputeRMAFAQ/size.html>

System	max. CEL files
64-bit linux system with 4 GB main memory	400
32-bit linux system with 4 GB main memory	160
32-bit Microsoft Windows XP system with 1 GB main memory	60

**Chip: HG-U133A**  
45.000 Probes  
~ 5\*1e5 rows

# Collecting data from public libraries

- Chip Type ?
  - **HG-U133A**

			3.6.2008		
GEO ID	AE ID	Beschreibung	# GEO	# AE	SUMME
GPL69	A-AFFY-33	HG-U133A	16490	17161	<b>33651</b>
GPL570	A-AFFY-44	HG-U133 Plus 2.0	14323	6888	<b>21211</b>
GPL2641	A-AFFY-65	Mapping 10K 2.0 Array Xba 142	7823	45	<b>7868</b>
GPL91	A-AFFY-9	HG-U95A	4708	1429	<b>6137</b>
GPL97	A-AFFY-34	HG-U133B	3830	4017	<b>7847</b>

- Search criteria

- Size of experiments **> 25**
- Available data: **CEL files, annotation files**
- **Cancer** data
  - Lung, leukaemia, breast, prostate, colon, lymphoma

- Databases

- ArrayExpress (AE)
- Gene Expression Omnibus (GEO)