LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
IBE MEDIZINISCHE FAKULTÄT
INSTITUT FÜR MEDIZINISCHE INFORMATIONSVERARBEITUNG
BIOMETRIE UND EPIDEMIOLOGIE

# affyPara
# Parallelized preprocessing for large Affymetrix microarray data sets

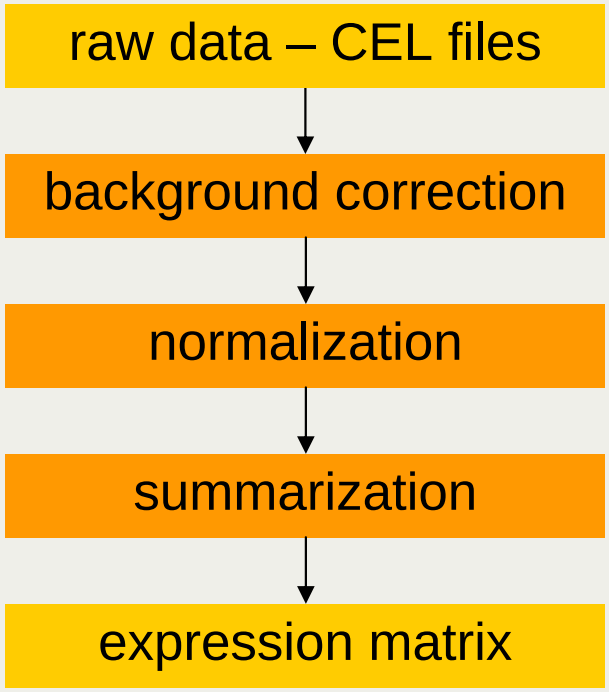
Markus Schmidberger
Ulrich Mansmann

IBE
http://ibe.web.med.uni-muenchen.de

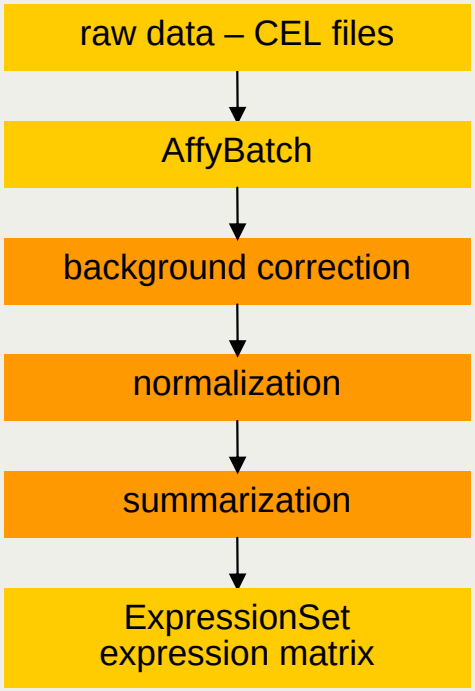
Seattle Dev Meeting 2008
July 30, 2008; Seattle, WA, USA

# 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

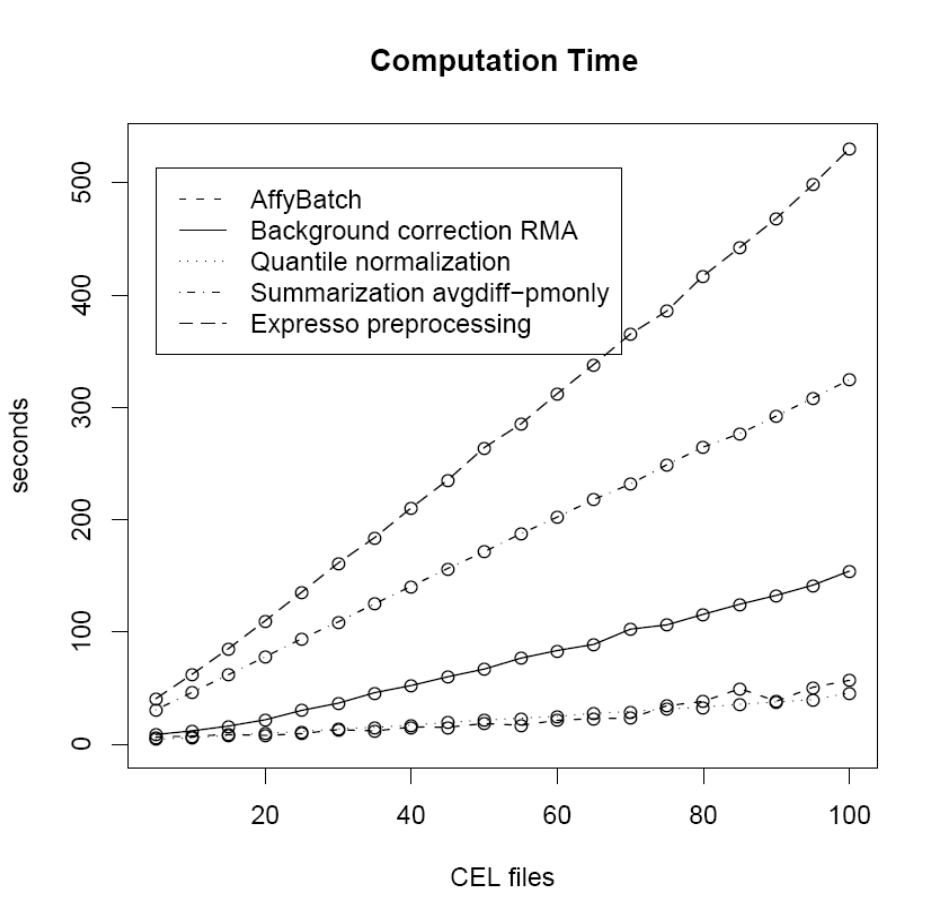


# Problems




- 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

# Problems




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

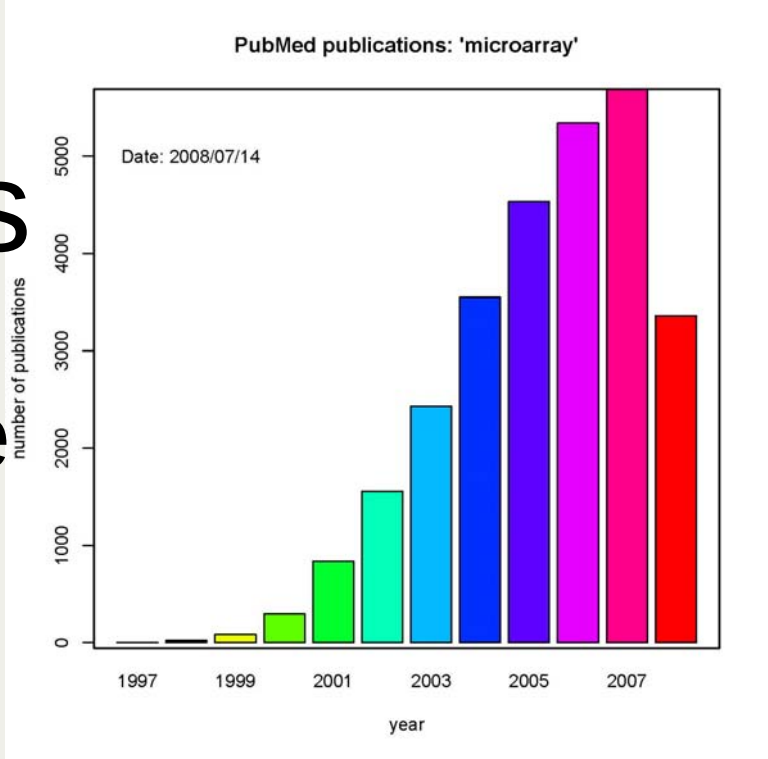
http://bmbolstad.com/misc/ComputeRMAFAQ/size.html

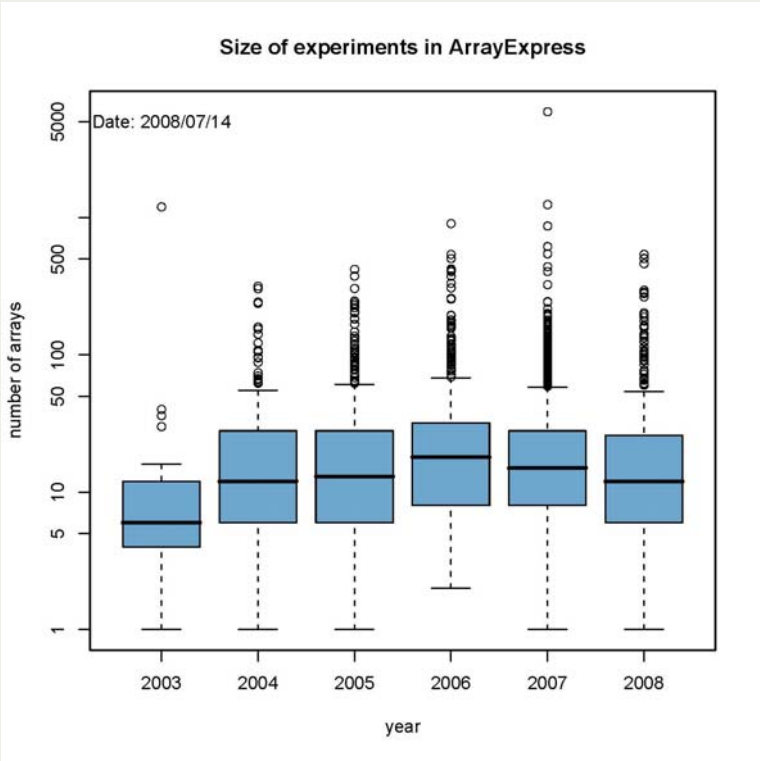
| System | max. CEL files |
|---|---|
| 64-bit linux system with 4 GB main memory | 400 |
| 32-bit linux system with 4 GB main memory | 160 |
| 32-bit Microsoft Windows XP system with 1 GB main memory | 60 |

Chip: HG-U133A

# 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

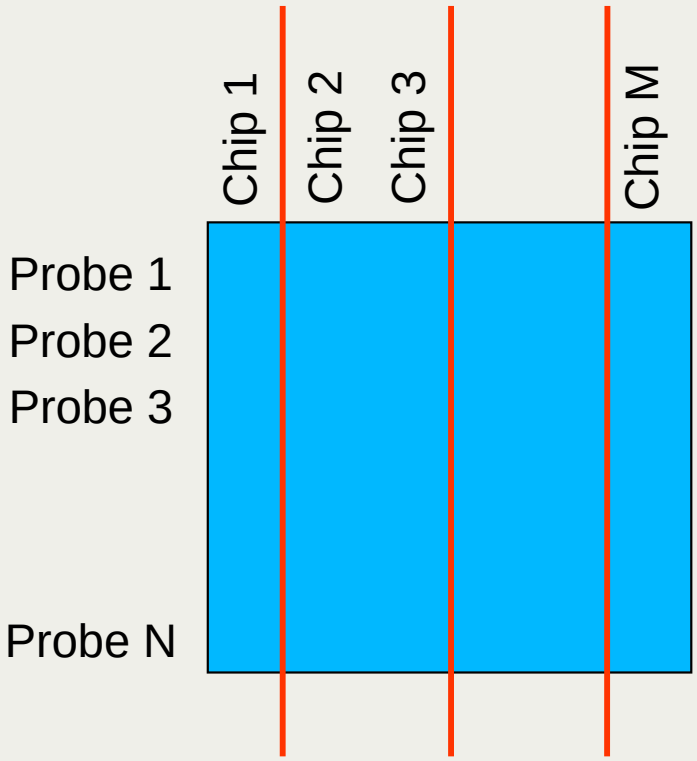
- Multiprocessors
  - the use of two or more central processing units (CPUs) within a single computer system
  - Today: Two-processors get a standard for workstations
  - OpenMP
- Multicomputers = Cluster
  - different parts of a program run simultaneously on two or more computers that are communicating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

# Software: MPI

- **M**essage **P**assing **I**nterface
- 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

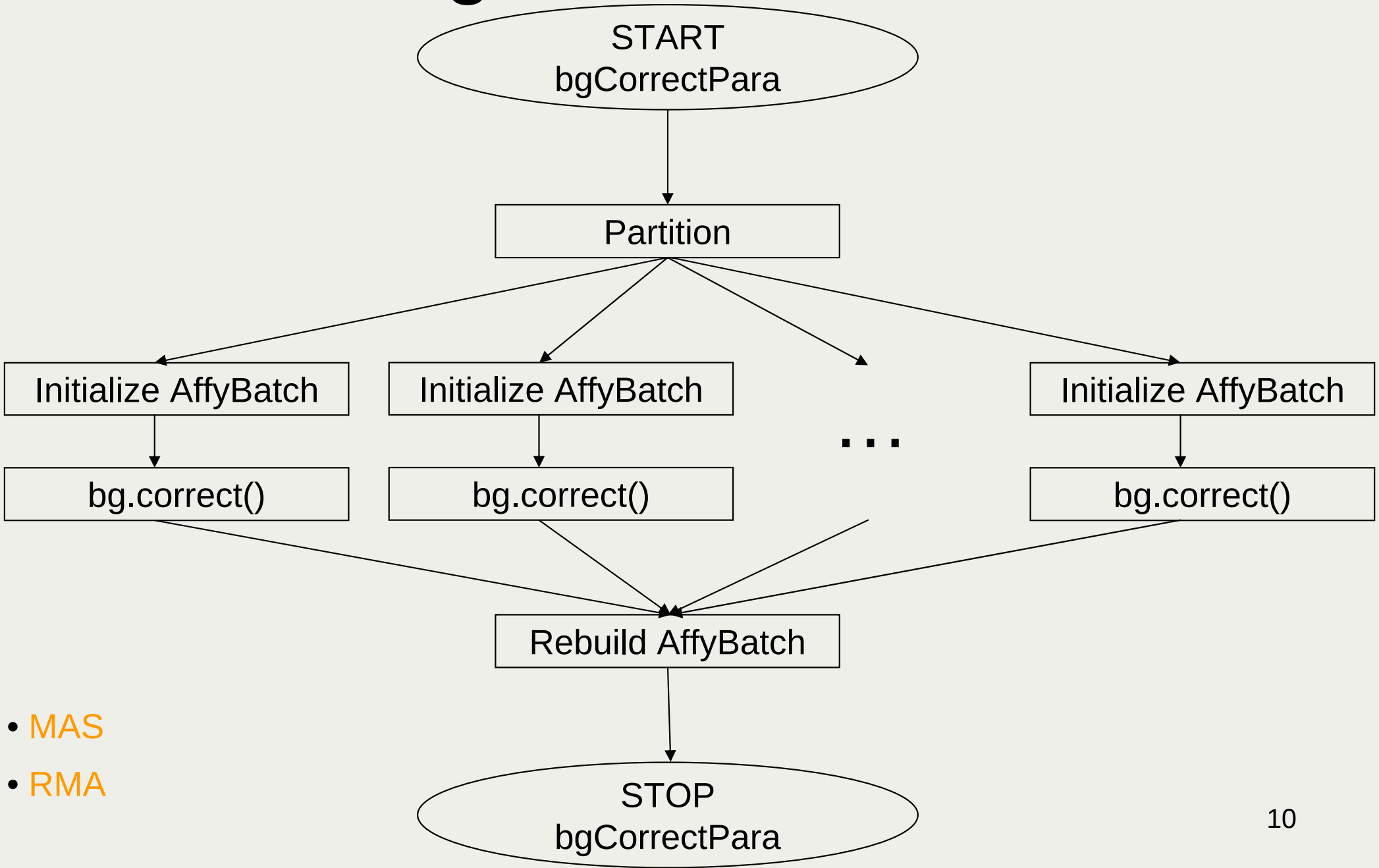
# Decomposition of AffyBatch

- AffyBatch = intensities from multiple arrays



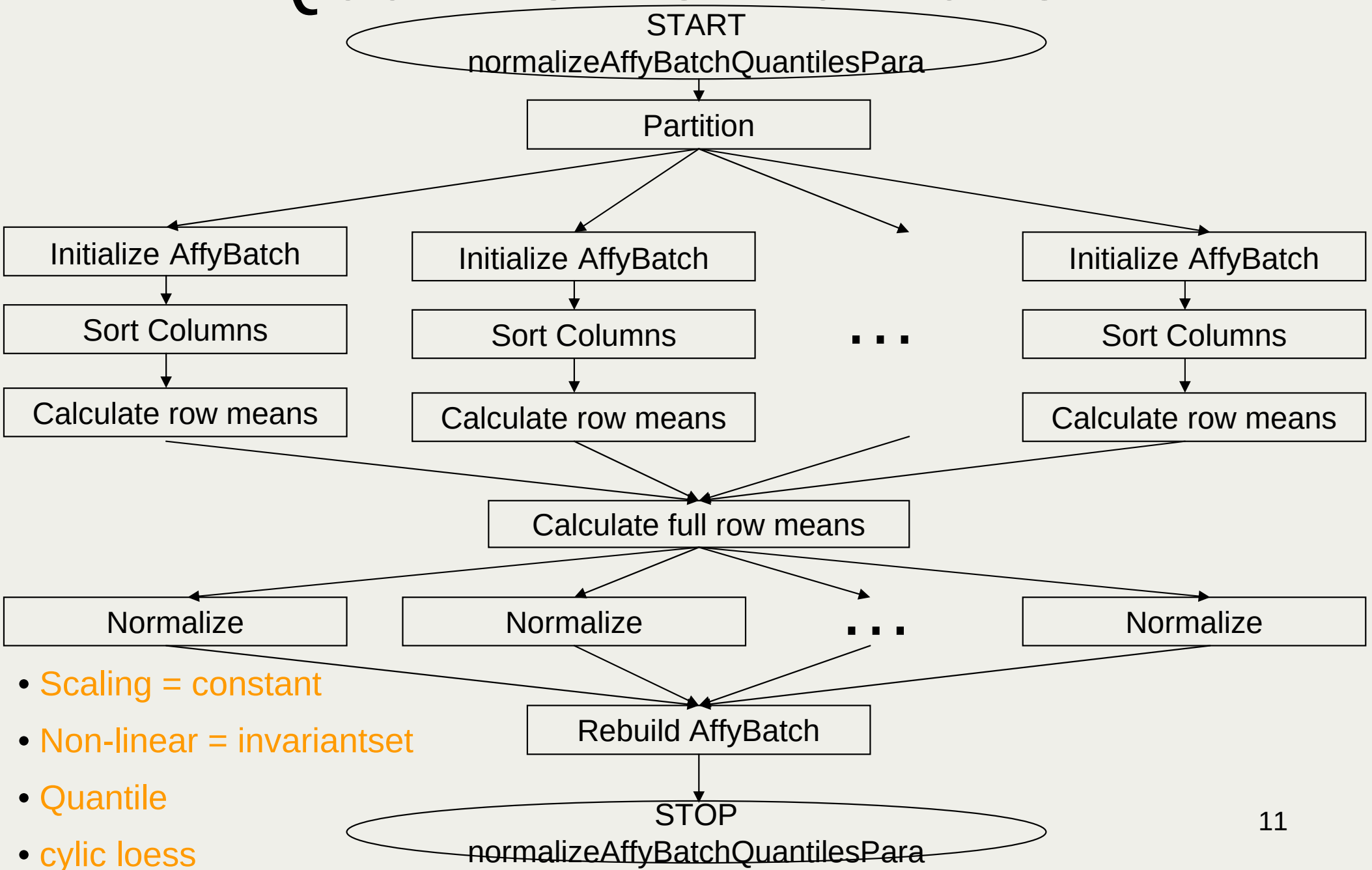

- Which decomposition is the best ?
  - Partition by chips
  - Partition by probes
  - Partition of CEL file name list
- Communication Overhead
  - A lot of data to transfer
  - Create AffyBatches at nodes
  - Complete preprocessing method: preproPara()

# Background Correction



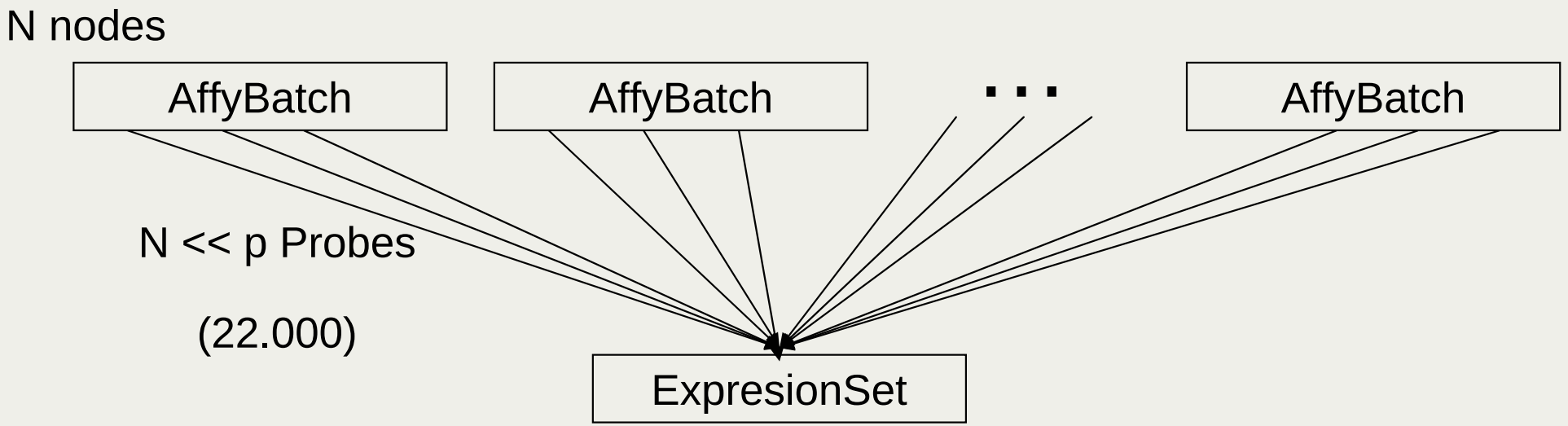

- MAS
- RMA

# Quantile Normalization




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

# Summarization



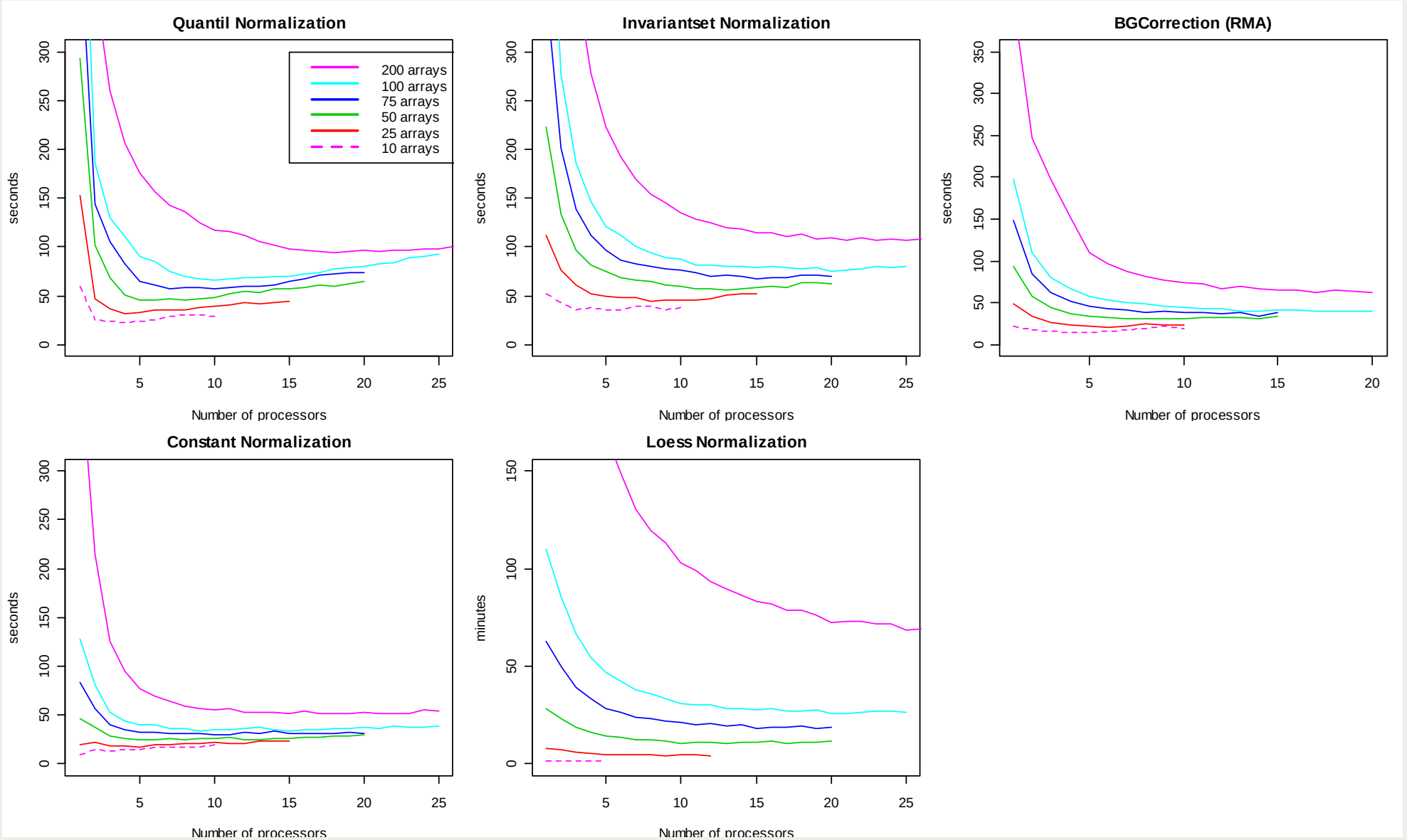

- avgdiff
- liwong
- mas
- medianpolish

# Results

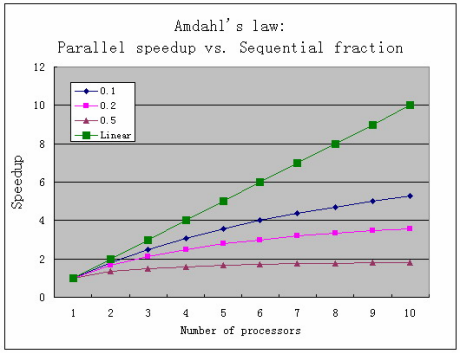
- BioConductor Package **affyPara** with parallelized affy-functions
  - Version 1.1.6
  - More CEL Files preprocessable
    - IBE Cluster: ~ 16.000 microarrays
  - Speedup

- Parallelization methods produce in view of machine accuracy the same results as serialized methods.
  - All.equal(), machine's precision.

# Results - Computation Time



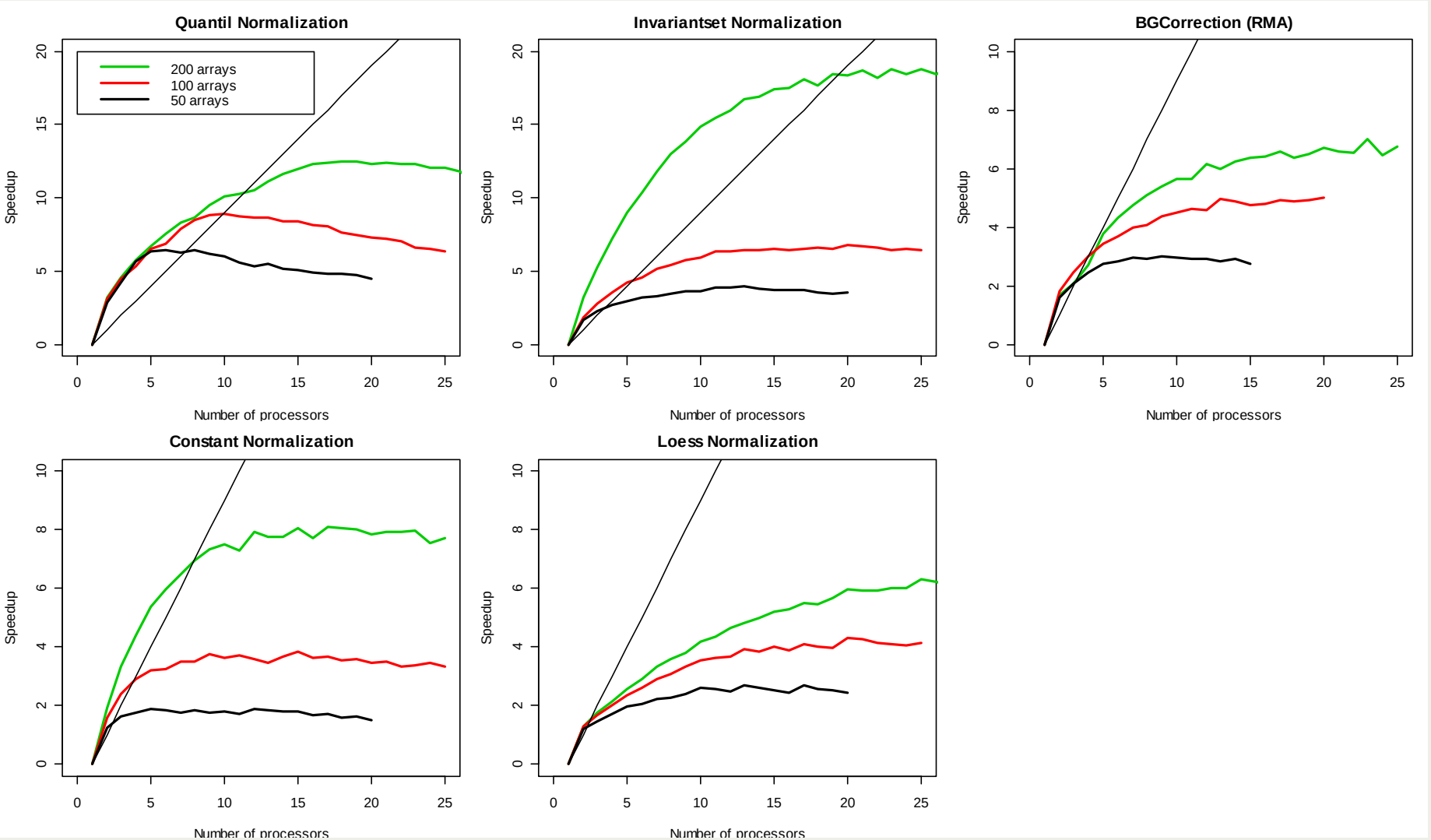
# Results – Speedup




- Speedup of the methods up to factor 15

Sp = T_1 / T_p

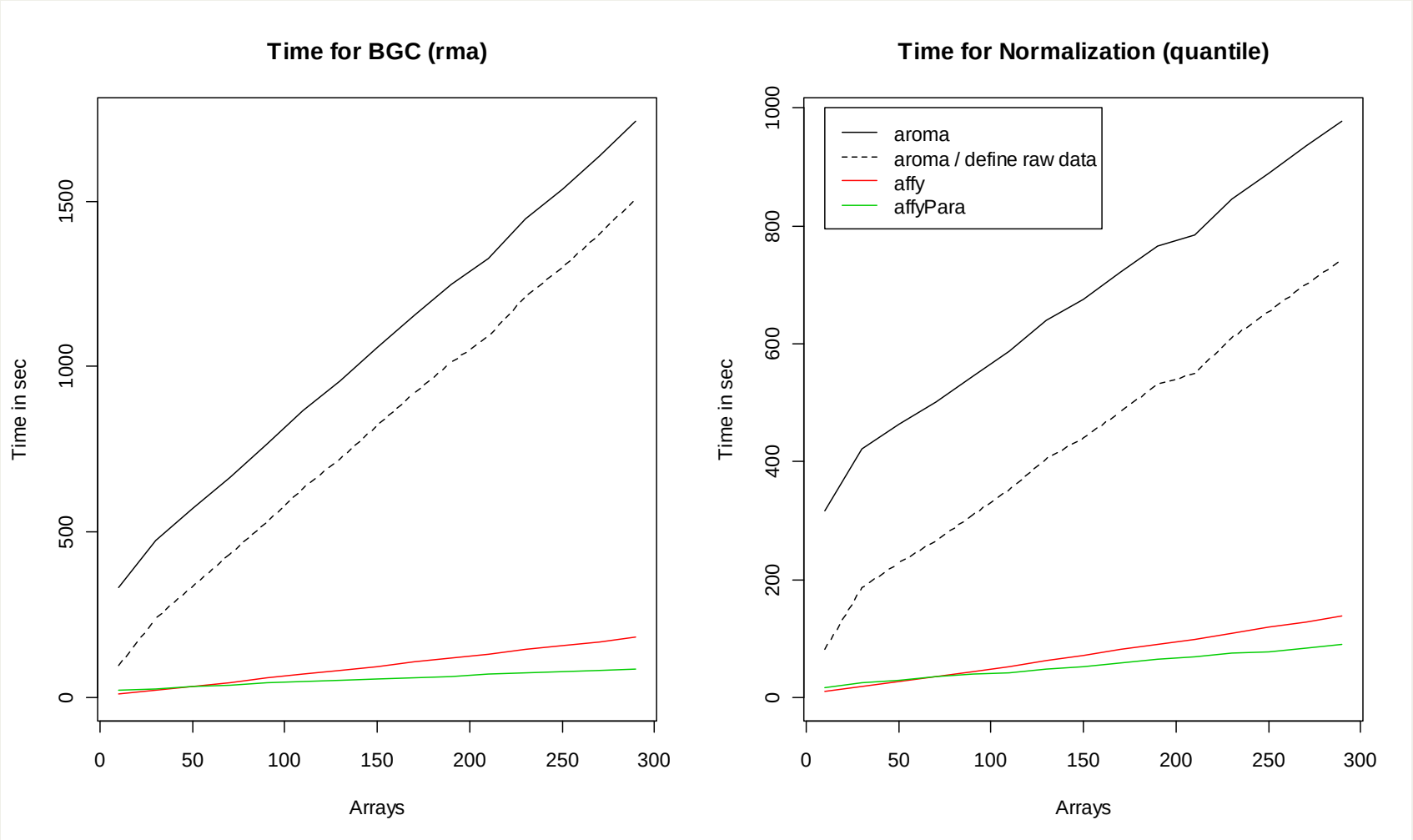
Sp ~ 1 / [ s +p/N ]



# affyPara – aroma.affymetrix

| | affyPara | aroma.affymetrix | Affymetrix packages in Bioconductor |
|---|---|---|---|
| **Maximum Number of arrays** | Unlimited (limited by number of computers) | limited by size of hard disk | limited by maim memory |
| **Programming language** | R + MPI | R | R |
| **Operating system** | Linux, Windows, OSX | Linux, Windows, OSX | Linux, Windows, OSX |
| **Data handling** | Memory | File system | Memory |
| **Third-party extensions** | Yes | Yes | Yes |
| **Open source** | Yes | Yes | Yes |
| **Usability** | Easy, very similar to other BioConductor packages | Difficult data structure, very different to other BioConductor packages | easy |
| **chip types** | expression arrays | expression arrays, SNP chips, exon arrays, … | expression arrays, SNP chips, exon arrays, … |
| **Computation time** | fast | slow | medium |

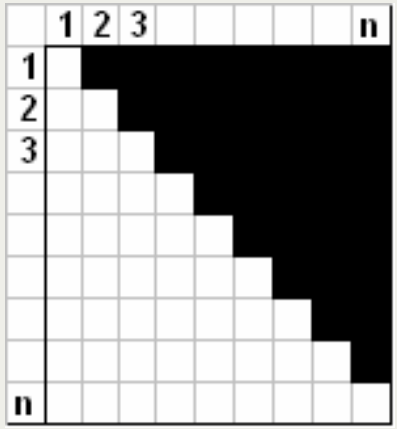
# affyPara – aroma.affymetrix Computation time



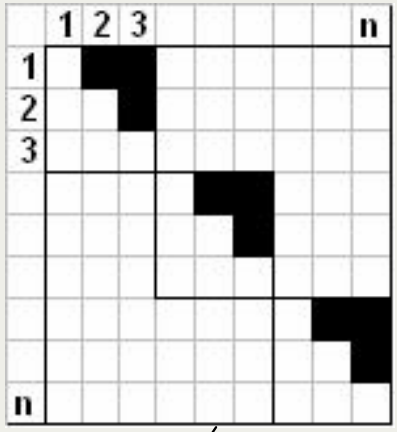
# New methods based on parallelization strategies

- 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)

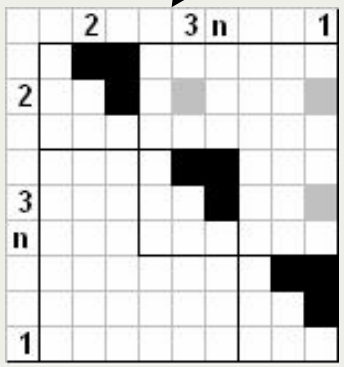
Complete cyclic loess




Partial cyclic loess on 3 nodes




Permutations of Arrays 2-3 times

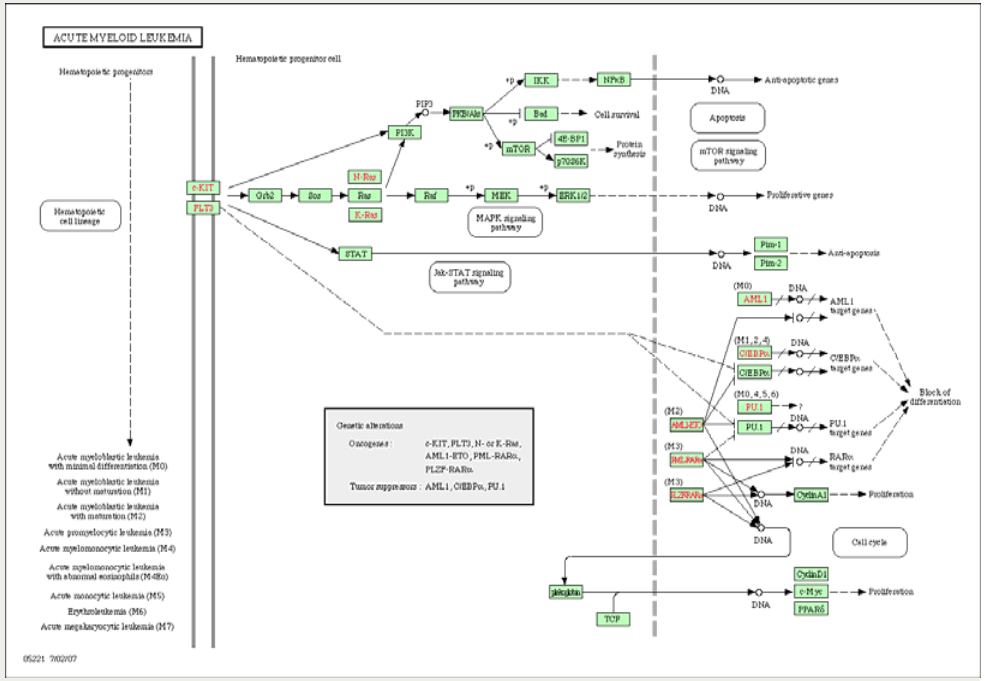


# Large project in applied bioinformatics in preparation

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

- Preprocessing all together

- Analyzing all together



# Collecting data from public libraries

- Chip Type ?
  - HG-U133A

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

- 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)

# Collecting data from public libraries - Results

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

Second / Referenz Data Set:

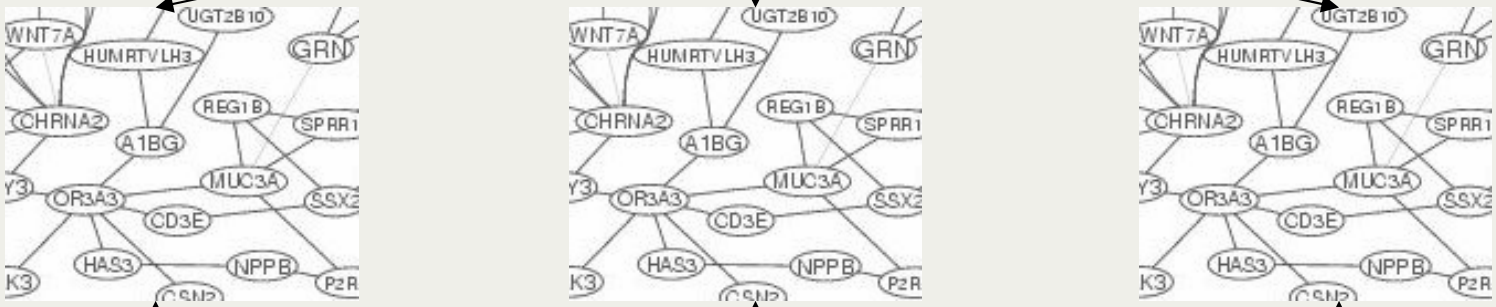
- expO (Expression Project For Oncology)

- 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

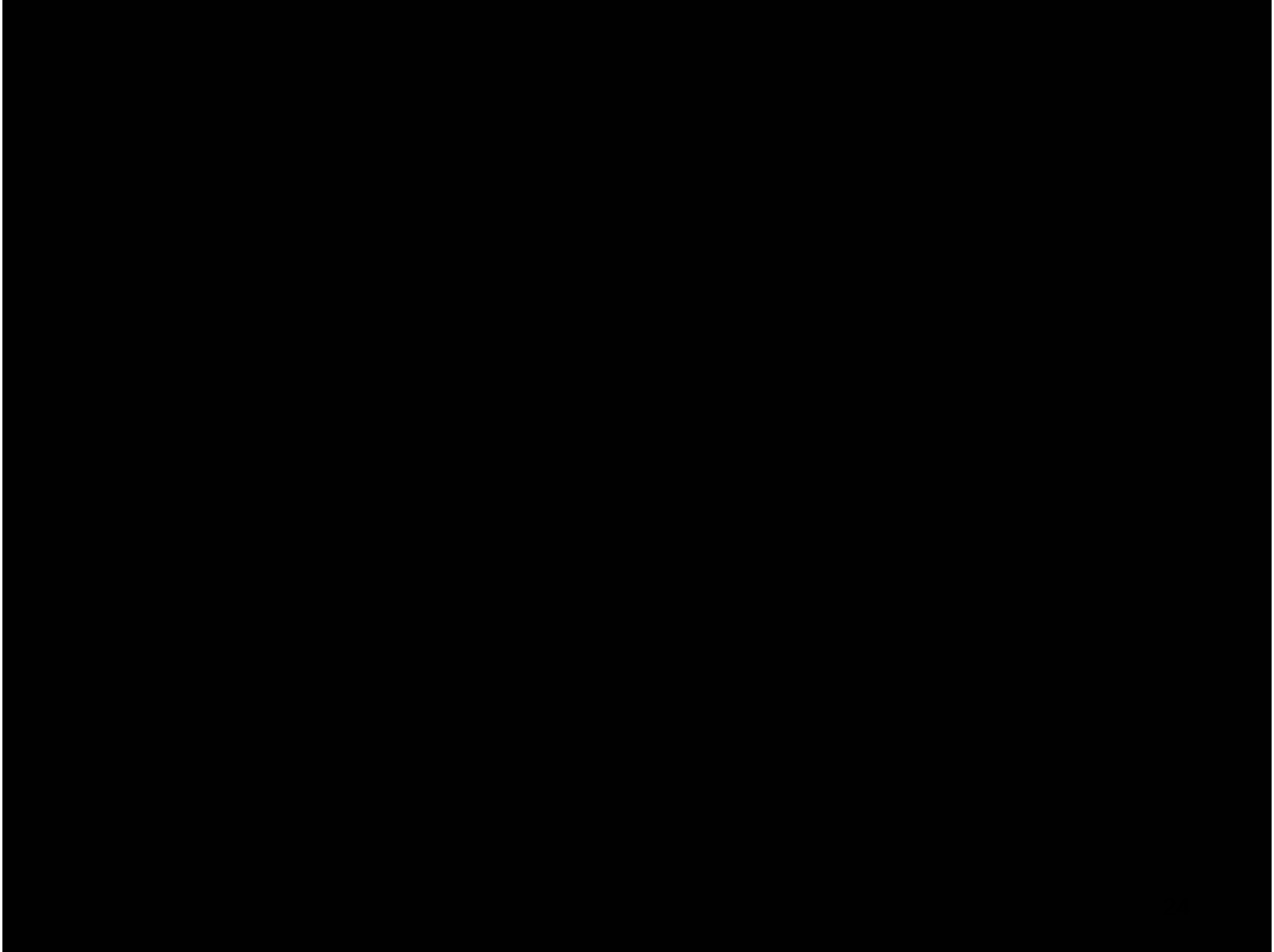
Normalization all together

# affyPara: Parallelized preprocessing for large Affymetrix microarray data sets & 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, 14-18 April 2008, Miami, Florida, USA. IEEE 2008.

**Thanks for your attention**

Dipl.-Tech. Math. Markus Schmidberger
schmidb@ibe.med.uni-muenchen.de
http://ibe.web.med.uni-muenchen.de



# 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.

# Future

- Ideal decomposition – decomposition rule?
- Improve summarization (hierarchically)
- Parallelization of other functions:
  affy, VSN, FARMS, QM,…
- R and Multiprocessors
  - openMP
- Parallelization of Analyze-Methods
- R and GPUs
  - CUDA

# IBE Workstation Cluster

- 32 workstations
  - 8 GB main memory
  - 2 dual core, Intel Xeon DP 5150
  - 1 Gbit Network
  - Shared memory: Samba
  - Linux 2.6.18, openSuse
  - LAM/MPI 7.1.3
  - R 2.6.0

# Decomposition

- Data decomposition
  - Data is broken into individual parts and distributed among processes that are essentially similar
- Task decomposition
  - Problem is divided in such a way that each process is doing a different calculation
- In practice: mixture of both

# Ideas for analyzing Differences in gene interaction

| | Classification (e.g. TNM) | | |
|---|---|---|---|
| | good | middle | bad |
| Cancer 1 | ? | ? | ? |
| Cancer 2 | ? | ? | ? |
| Cancer 3 | ? | | |
| Cancer 4 | ? | ? | ? |
| Cancer 5 | ? | ? | ? |

**Pathways**

**Differences ? / Similarities ?**

# Package affyPara

| | | affy | affyPara |
|---|---|---|---|
| BGC | RMA | bg.correct.rma | bgCorrectPara |
| | MAS 5.0 | bg.correct.mas | bgCorrectPara |
| | RMA alt | bg.correct.rma2 | bgCorrectPara |
| Normalization | Scaling | normalize.AffyBatch.constant | normalizeAffyBatchConstantPara |
| | invariantset | normalize.AffyBatch.invariantset | normalizeAffyBatchInvariantsetPara |
| | Quantiles | normalize.AffyBatch.quantiles | normalizeAffyBatchQuantilesPara |
| Summarization | | computeExprSet | computeExprSetPara |
| complete | preprocessing | threestep, expresso | preproPara |