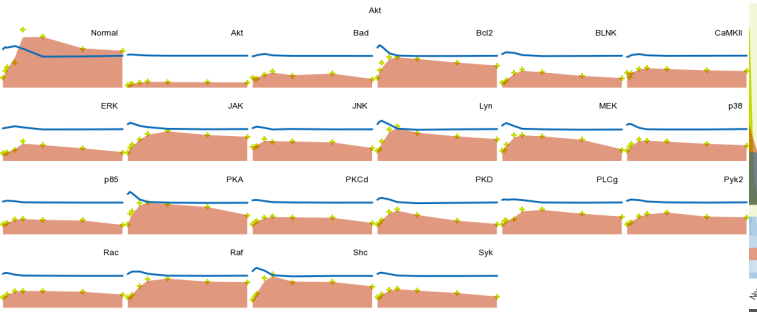
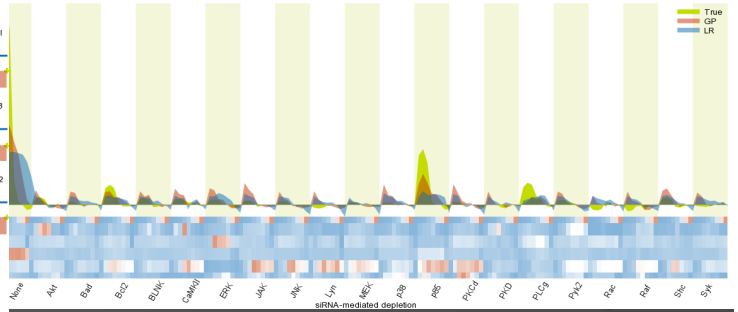
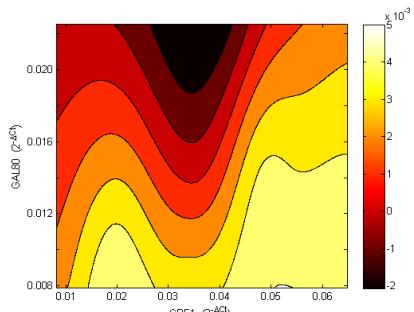
Harri Lähdesmäki – Aalto University School of Science and Technology
Department of Information and Computer Science
Tarmo Äijö – Tampere University of Technology
Department of Signal Processing

APPLICATION OF HIGH-PERFORMANCE COMPUTING IN MODELING BIOLOGICAL SYSTEMS

12th of October 2010



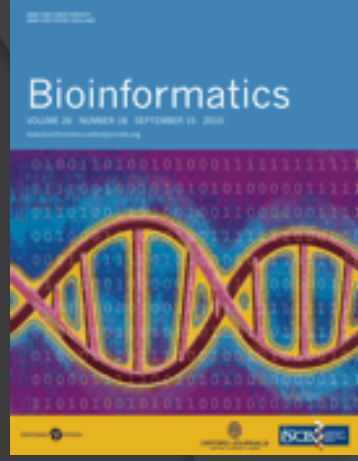




Laboratory of Regulatory Genomics

- Principal investigator: Prof. Harri Lähdesmäki (pro term), ICS/Aalto University
- 
- Aalto University School of Science and Technology
- 4 PhD students and 4 MSc students (in ICS/Aalto and SGN/TUT)
- The ReG laboratory uses computational techniques to model and understand molecular regulatory mechanisms and their effect on health and disease
- The research focuses on analysis of high-throughput data and development of statistical modeling and machine learning methods to understand transcriptional and post-transcriptional regulatory mechanisms, protein signaling pathways, and effects of mutations on regulatory mechanisms
- The ReG laboratory also develops and applies methods for biological sequence analysis and for combining heterogeneous biological information sources and high-throughput measurements
- Research projects are carried out in close collaboration with experimental groups, and the ReG laboratory collaborates on molecular immunology, cancer and type 1 diabetes systems biology research projects.
- More and most recent information: http://www.cis.hut.fi/harrila/

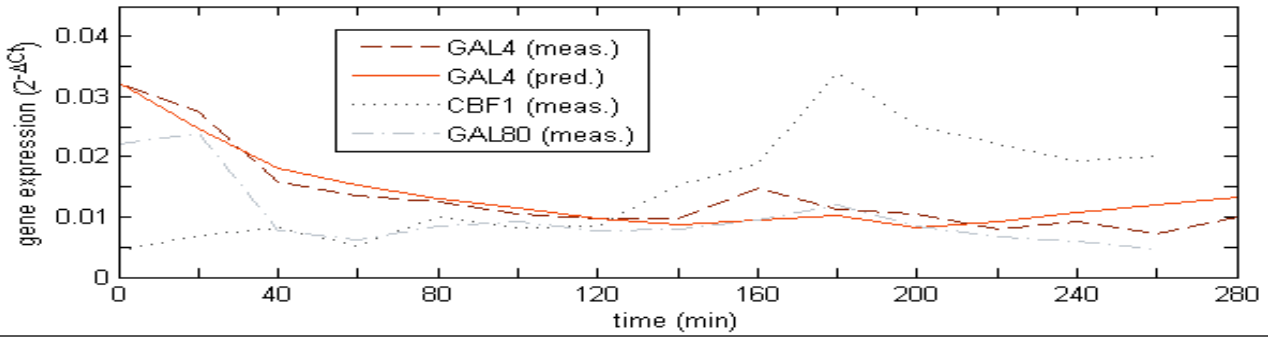
Bayesian inference for gene regulatory networks from expression data

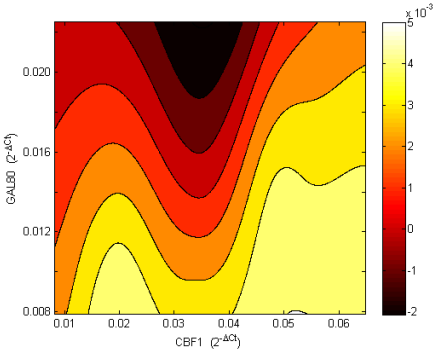


- **Reference**: Äijö T. and Lähdesmäki J., *Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics, Bioinformatics*, 25 (22):2937-2944
- Revealing the structure and dynamics of a gene regulatory network (GRN) is of great interest and represents a considerably challenging computational problem
- Our approach uses Gaussian processes and ODEs to model changes in gene expression levels, and uses a Bayesian framework to define a posterior probability distribution over the model space, i.e., interactions between the genes

Bayesian inference for gene regulatory networks from expression data (cont.)

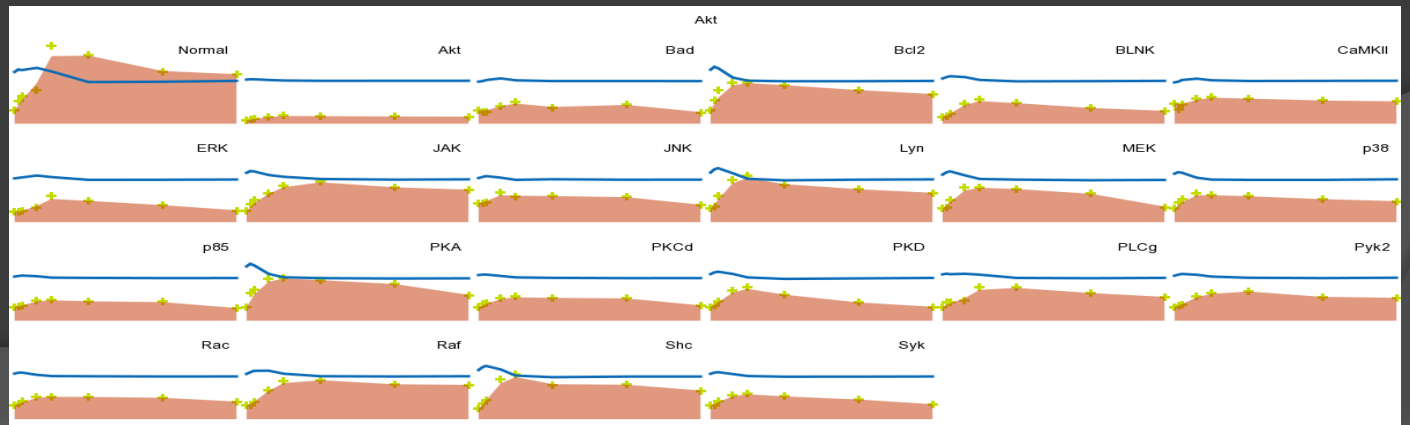
- Type 2 maximum likelihood estimation of model parameters
- The amount of models to consider grows quickly as a function of the number of genes – 2^(N^2)
  - exhaustive search of the model space
  - models are independent → embarrassingly parallel problem
- **Calculations requiring several years of CPU time are completed in a couple of days by harvesting unused capacity from IT-infrastructure using Techila** (private computing cloud)

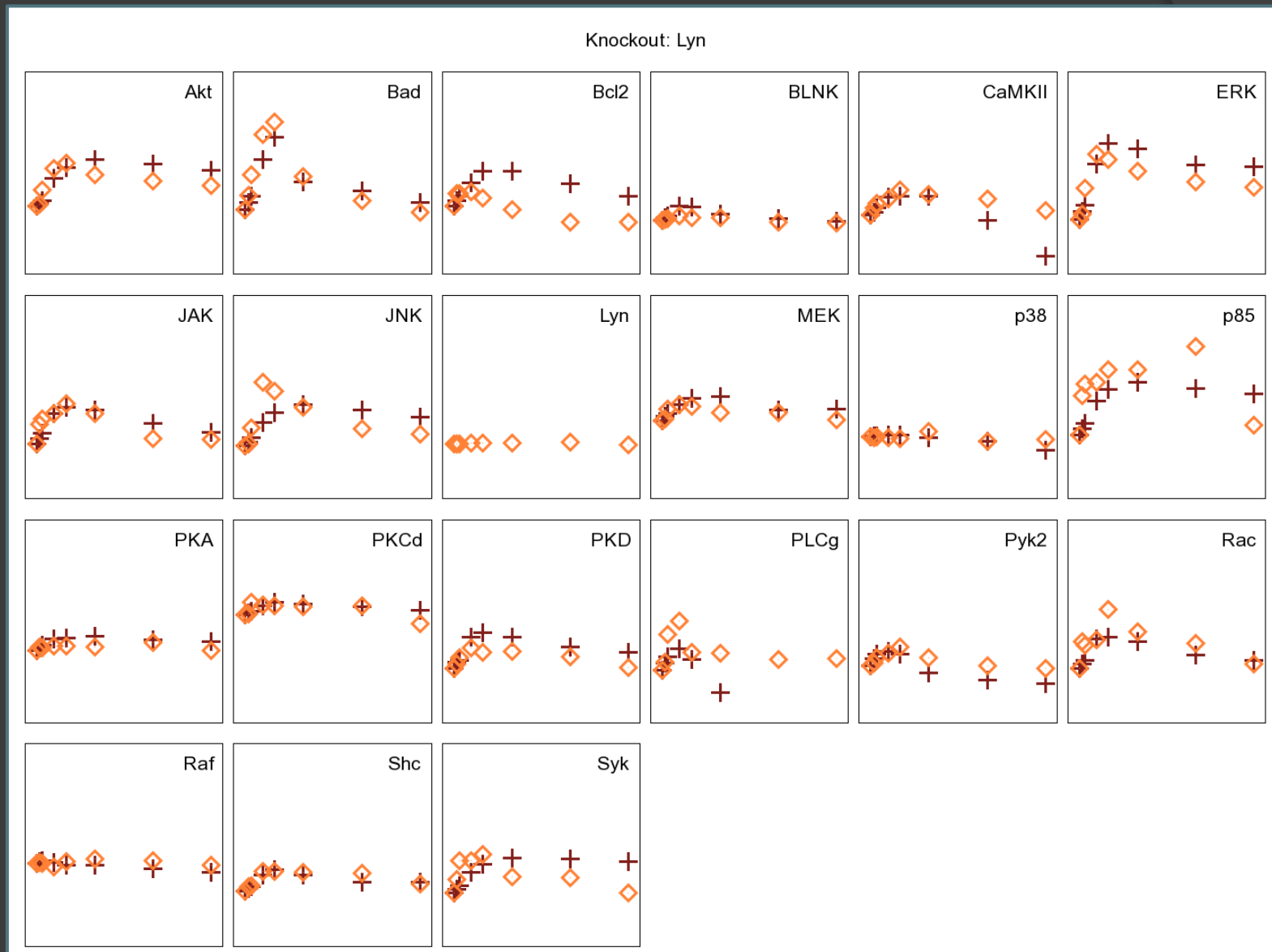




Bayesian modeling of signal processing in signaling pathways

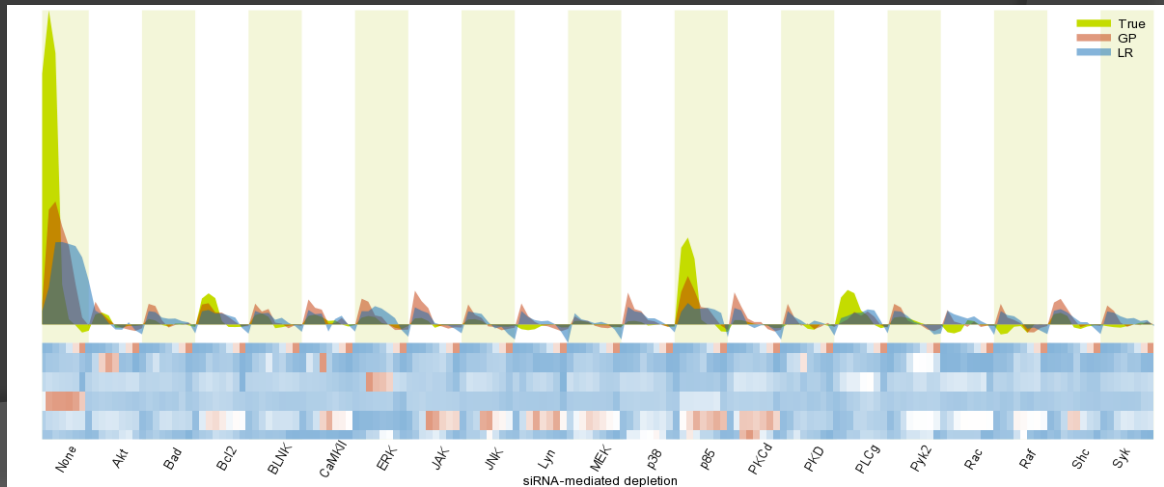
- Reference: Äijö T. et al., *Non-parametric and continuous modeling of cell dynamics, with an application to signaling networks*, in progress
- Application: B-cell receptor (BCR) signaling pathway
- Why to build a mathematical model ?
  - To understand how the signals propagate through the pathway from the BCR to the cell nucleus
  - To see how we can alter the signaling pathway in order to achieve the desired output




Knockout: Lyn
Akt
Bad
Bcl2
BLNK
CaMKII
ERK
JAK
JNK
Lyn
MEK
p38
p85
PKA
PKCd
PKD
PLCg
Pyk2
Rac
Raf
Shc
Syk

Bayesian modeling of signal processing in signaling pathways (cont.)

- Time consuming steps are
  - Model fitting (independent calculations)
  - Model selection (independent calculations)
- Models are ranked based on their predictive ability, which is assessed using cross-validation
- **Cross-validation is nice example of an embarrassingly parallel problem** (on the level of different data partitions)

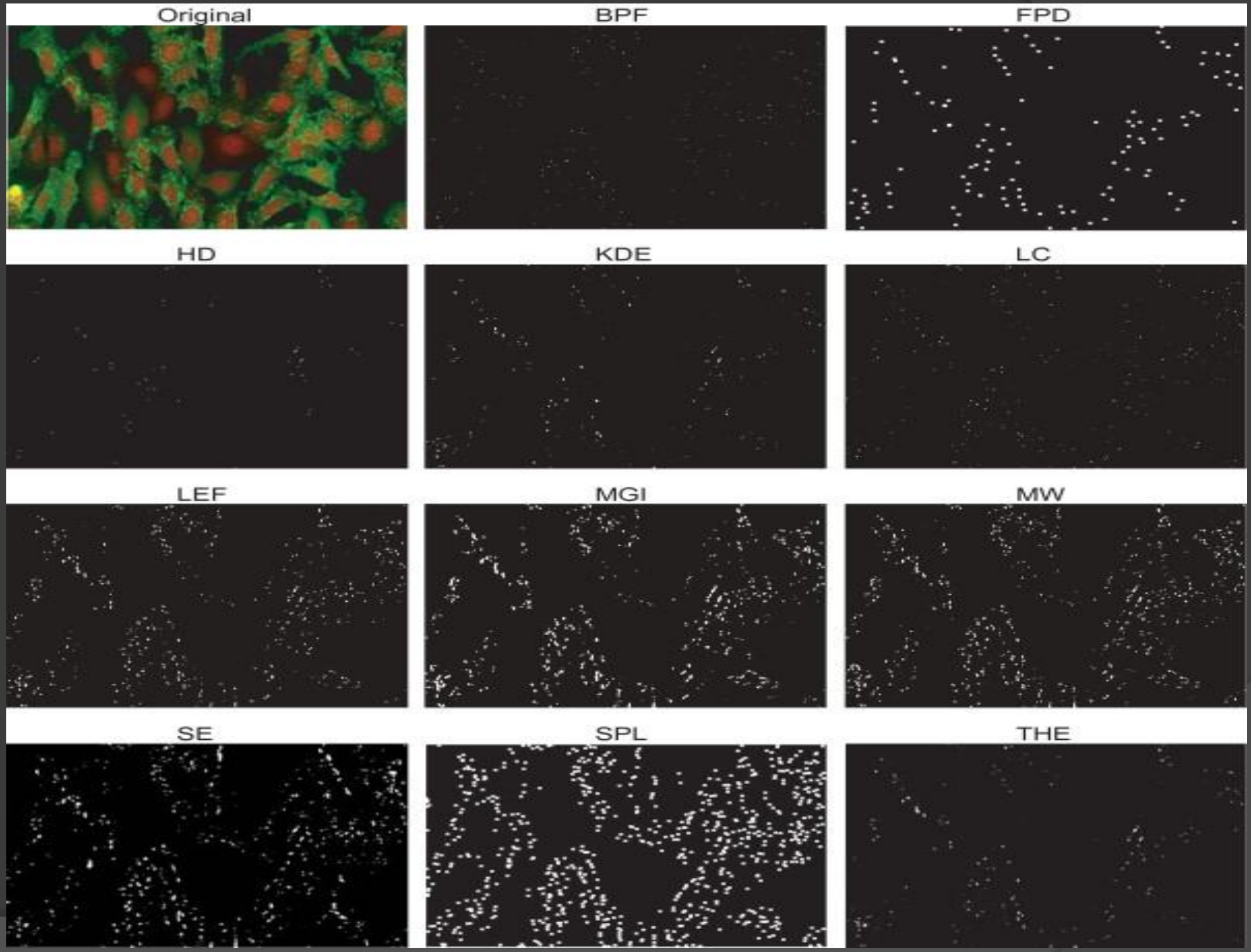


Application to image analysis

- **Reference**: Ruusuvuori P. et al., *Evaluation of methods for detection of fluorescence labeled subcellular objects in microscope images*, *BMC Bioinformatics*, 2010, 11:248

- An example from the Laboratory of Microscopic Image Analysis led by senior researcher Dr. Antti Niemistö & Dr. Heikki Huttunen
- With increased imaging throughput and large-scale data acquisition, the challenge of image interpretation and information extraction has also shifted from visual inspection or interactive analysis towards increasingly automated methods

- Accurate and automated subcellular object segmentation is an essential enabler for a variety of applications
- Evaluating the performance of image segmentation algorithms has been a long-standing challenge


Original
BPF
FPD
HD
KDE
LC
LEF
MGI
MW
SE
SPL
THE

Application to image analysis (cont.)

- In order to better understand the performances of segmentation algorithms under different conditions, we have carried out a comparative study including eleven spot detection of segmentation algorithms from various application fields

- Because the parameter values of a given segmentation algorithm have a significant effect on the detection accuracy and need to be tuned specifically
  for the applied data.

- **Parameter optimization was done applying search techniques in Techila environment. Results for a well-justified comparison required several years of CPU time. In Techila the optimization was completed in a week.**

End-User Opinion of Private Computing Cloud

- Speeds up the whole development process of code and models, for example, within MATLAB and R
  - Techila is **easy and simple to use** → we can continue using familiar tools like MATLAB as before. Techila has integrated distributed computing capabilities to e.g. MATLAB so seamlessly that we barely even realize that we use distributed computing. Results just come in fraction of time.
  - No more waiting for computing time in clusters → Techila gives us immediately a lot of computing power that speeds up the research and development a lot!
- Enables use of significantly more detailed models
- Enables the use of alternative methods → exhaustive searches as an alternative to approximate search strategies
- Frees much more time to result analysis as computations get completed in a fraction of the original time!