MicroDiscovery

Methods and tools for evaluating HT data

Johannes Schuchhardt, MicroDiscovery GmbH

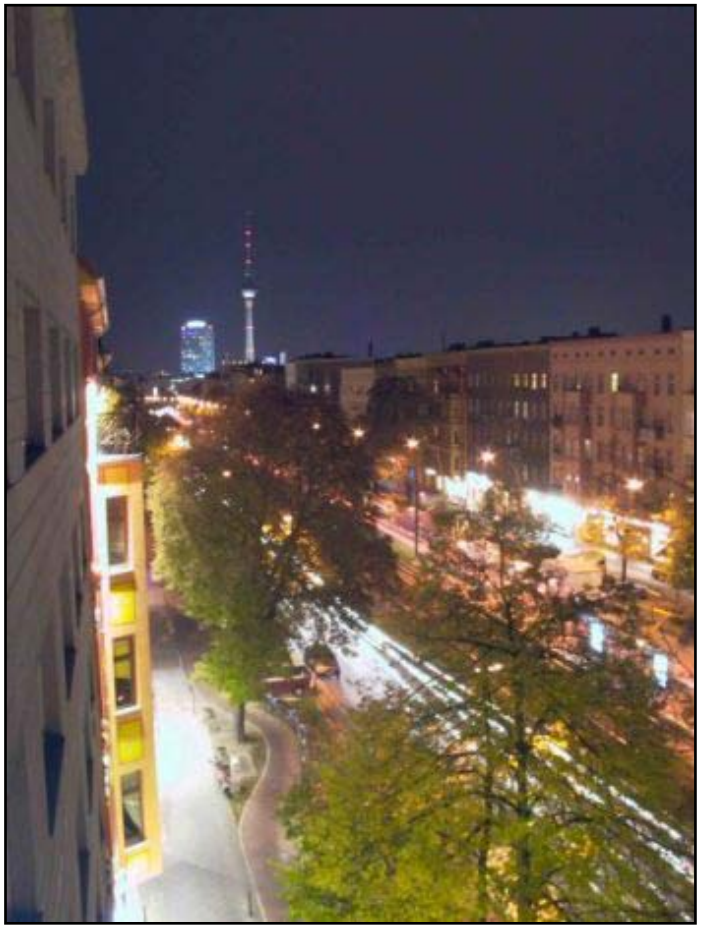
Leibniz Conference, Castle Lichtenwalde

Topics of this talk

- **Company profile**
- Error models for biomarker detection in proteomics data (cNEUPRO project)
- Statistics for pathway detection

MicroDiscovery

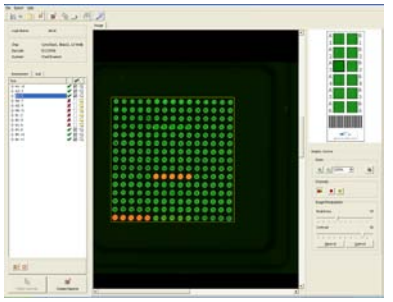
- Founded in 2000
- 20 employees (May 2009)
- Focus: Applied Bioinformatics
  - Software, regulated markets
  - Services, data analysis
- ISO 9001:2000 certified
- Reference Customers
  - Greiner Bio-One, febit biotech
  - Bayer-Schering Pharma AG

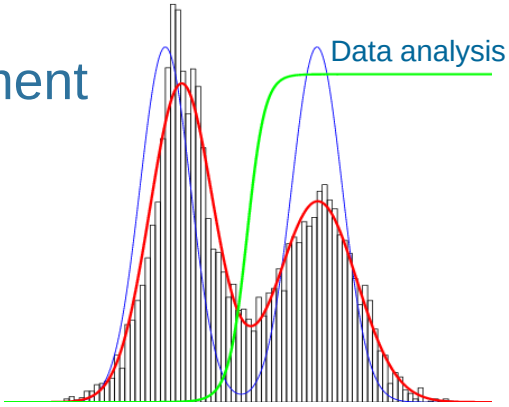


Business Areas

- Software Development
  - Software for R&D
  - Software for IVD
  - Software for medical technology
- Scientific Services
  - Evaluation of research data and clinical studies
  - Bioinformatics and explorative statistical data analysis
  - Development of algorithms and methods
  - Data integration and knowledge management

Software





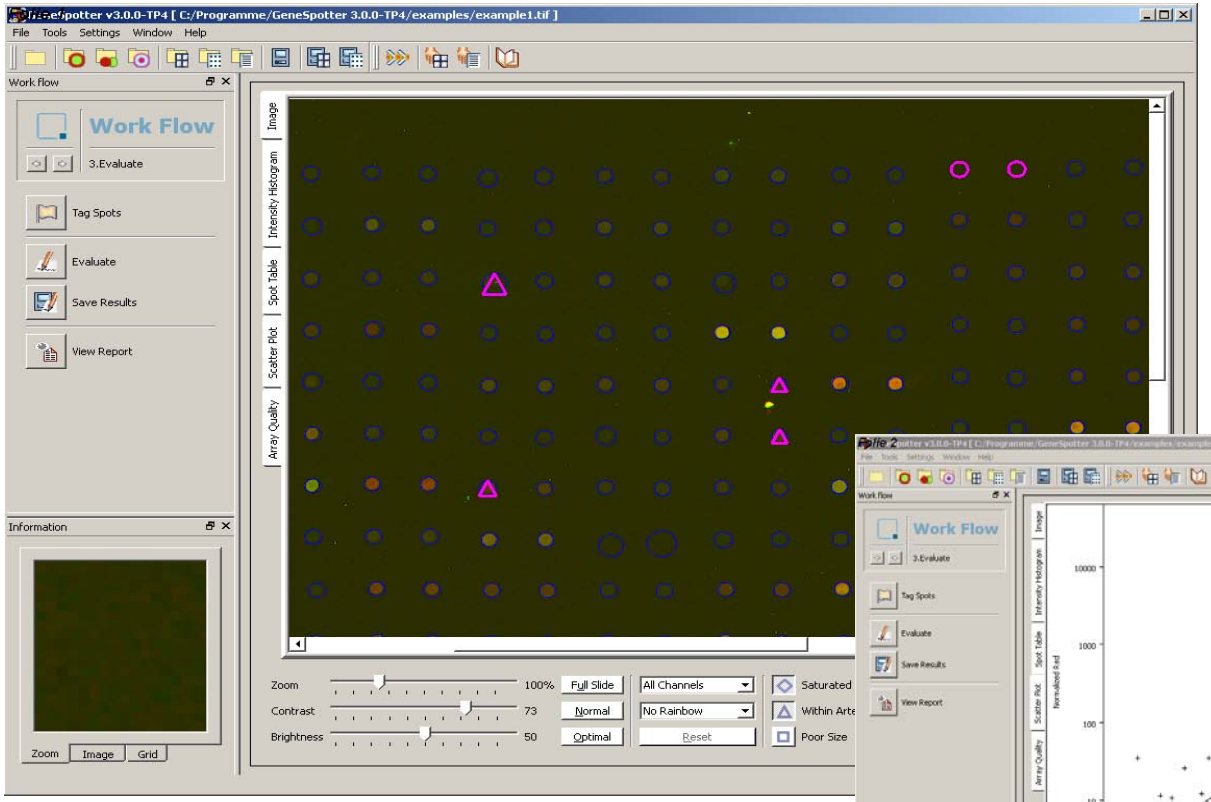
IT meets life science

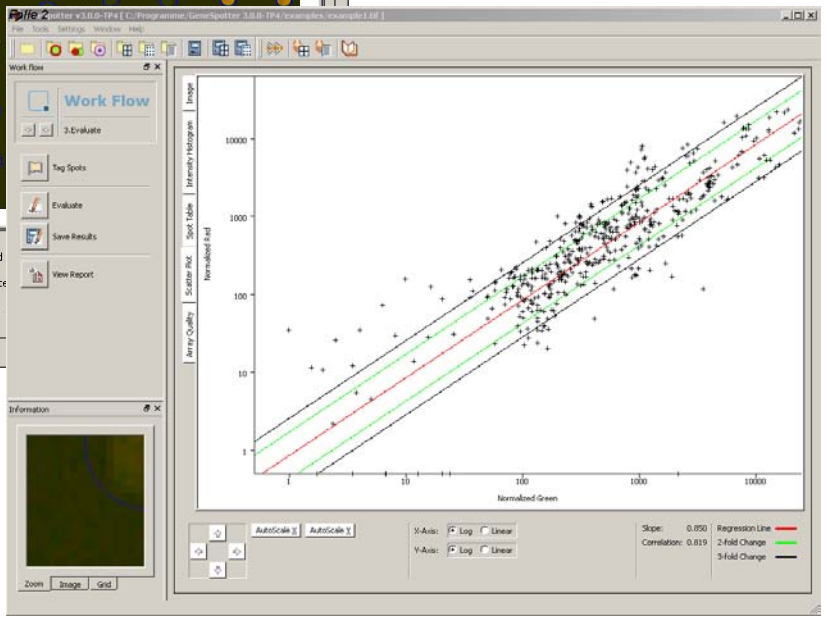
MicroDiscovery

Software Development

MicroDiscovery

Software Development for R&D





Characteristics

- Automation
- Quality control
- Ergonomics

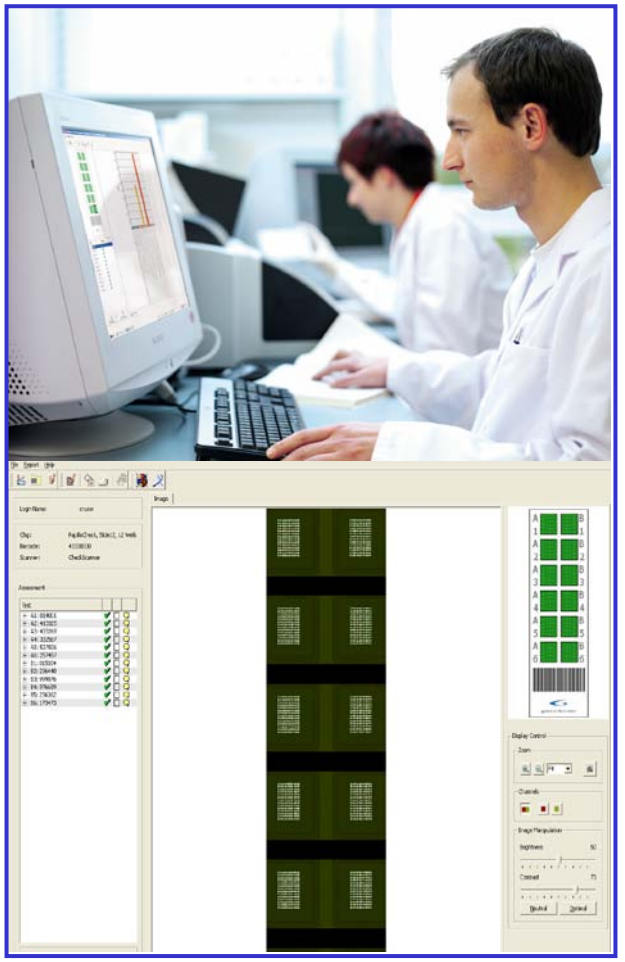
CheckReport™ – IVD

CheckReport™ – software for microarray based pathogen classification

Features:

- Data security
- Reproducibility
- Logging / System-audit-trail
- CE certification for IVD
- FDA conform (CFR21 part 11)

greiner bio-one



Test portfolio:

ParoCheck	Feb. 2005
CarnoCheck	Apr. 2005
MycoDetect	Oct. 2005
CytoCheck	Oct. 2005
PapilloCheck	Nov. 2006

Topics of this talk

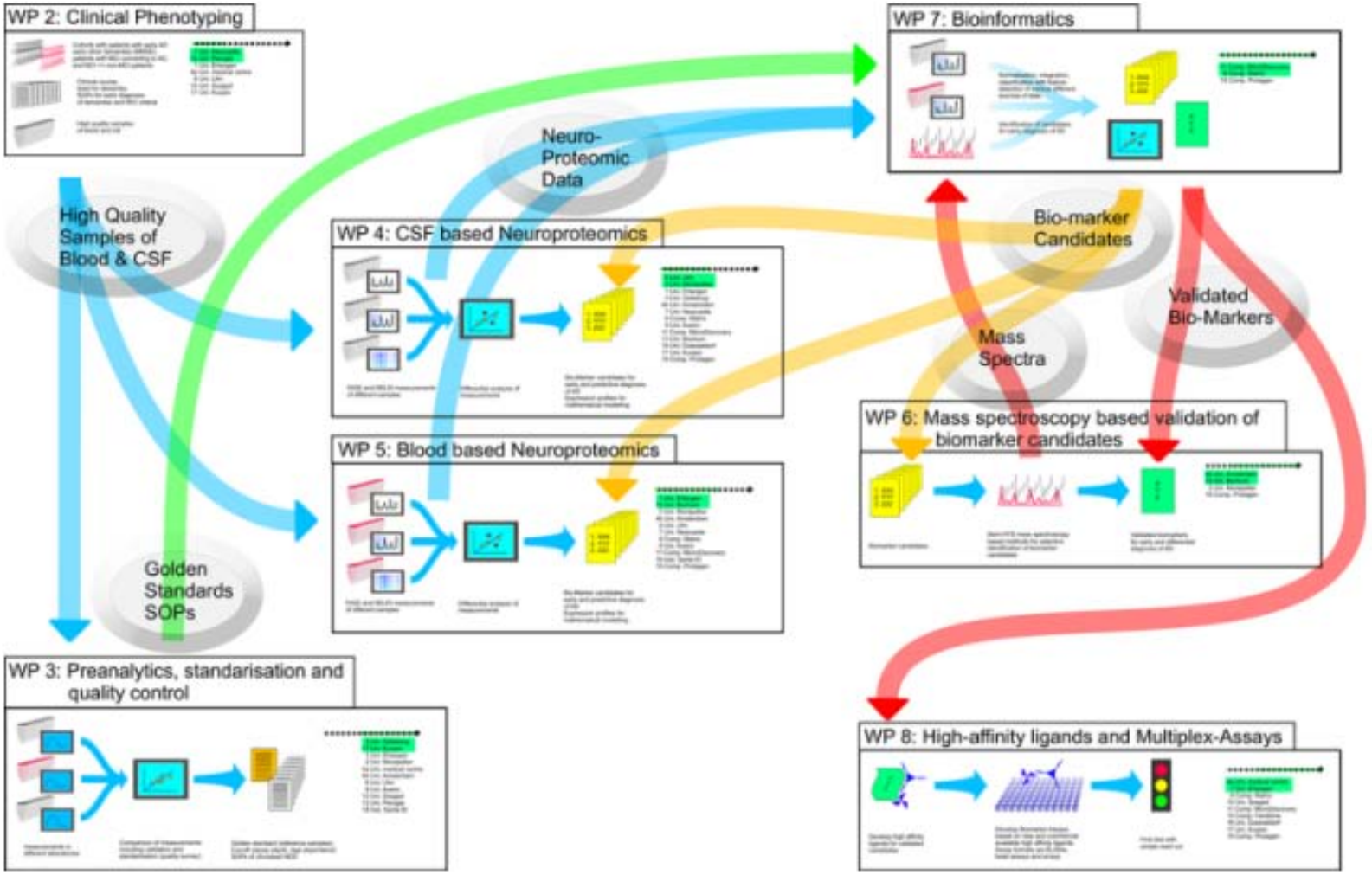
- Company profile
- **Error models for biomarker detection in proteomics data** (cNEUPRO project)
- Statistics for pathway detection

Biomarker identification: cNEUPRO Project

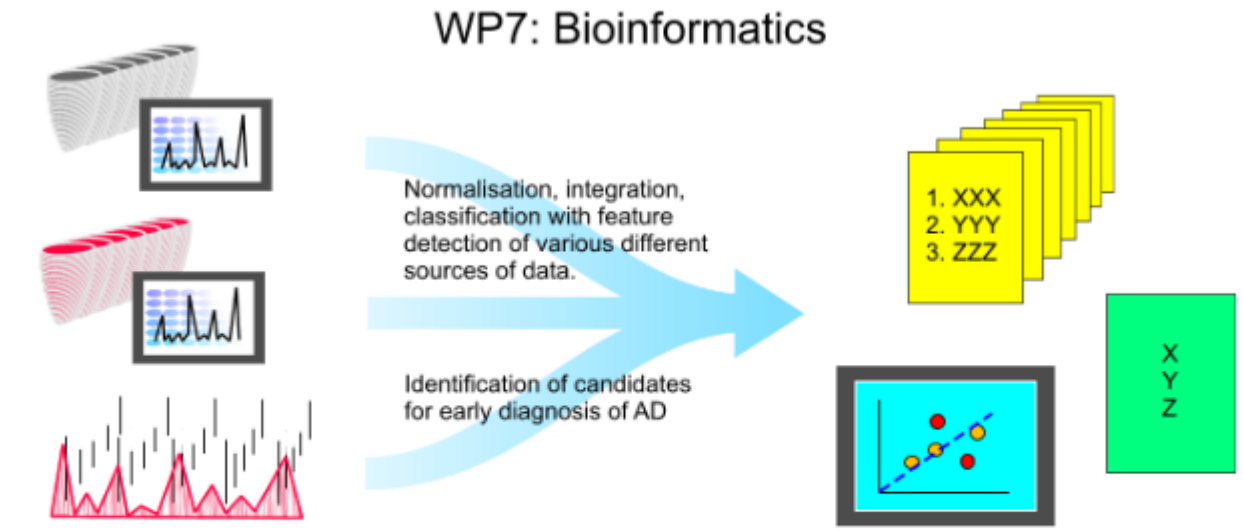
cNEUPRO: clinical neuro-proteomics (EU, FP6)

- 20 Academic and non-academic partners, coordinator: Prof. Jens Wiltfang, UK Essen
- Predictive dementia diagnostics for early stages of Alzheimer‘s disease (MCI => MCI-AD)
- Identification of biomarkers using CSF based and blood based proteomics
- Website: www.cneupro.eu

Workflow of sample preparation and analysis



Bioinformatics task: identification of marker candidates



11 Comp. MicroDiscovery
8 Comp. Matrix
19 Comp. Protagen

Proteomics data is generated in several centers

Data sets: CSF based quantitative neuro-proteomics

Amsterdam
C. Jimenez

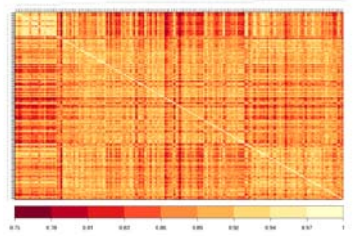
Bochum
K. Marcus

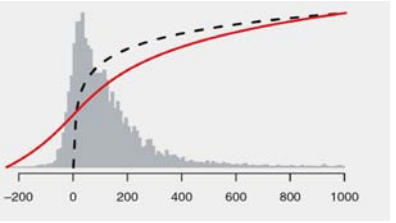
Dortmund
Protagen

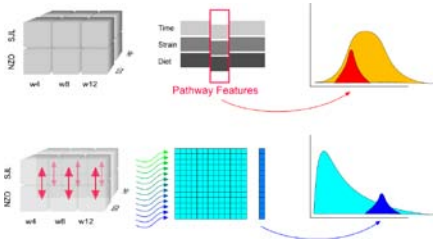
Data management
Biomarker identification

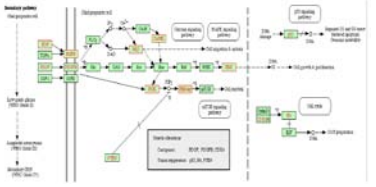
Workflow for data evaluation

- Sample preparation
- Mass spectrometry
  - Protein identification
  - Generation of protein count lists
- Statistics for hit identification:
  - Data normalization
  - Differential protein expression
  - Regression
  - ANOVA p-values
  - Pathway detection









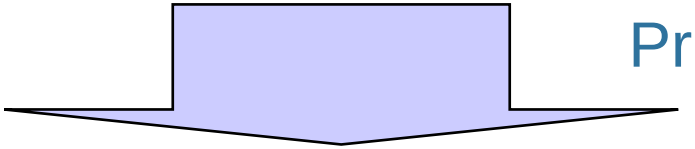
Design of exploratory study: 4 groups 5 patients

Control
5X

MCI
5X

MCI/AD
5X

AD
5X

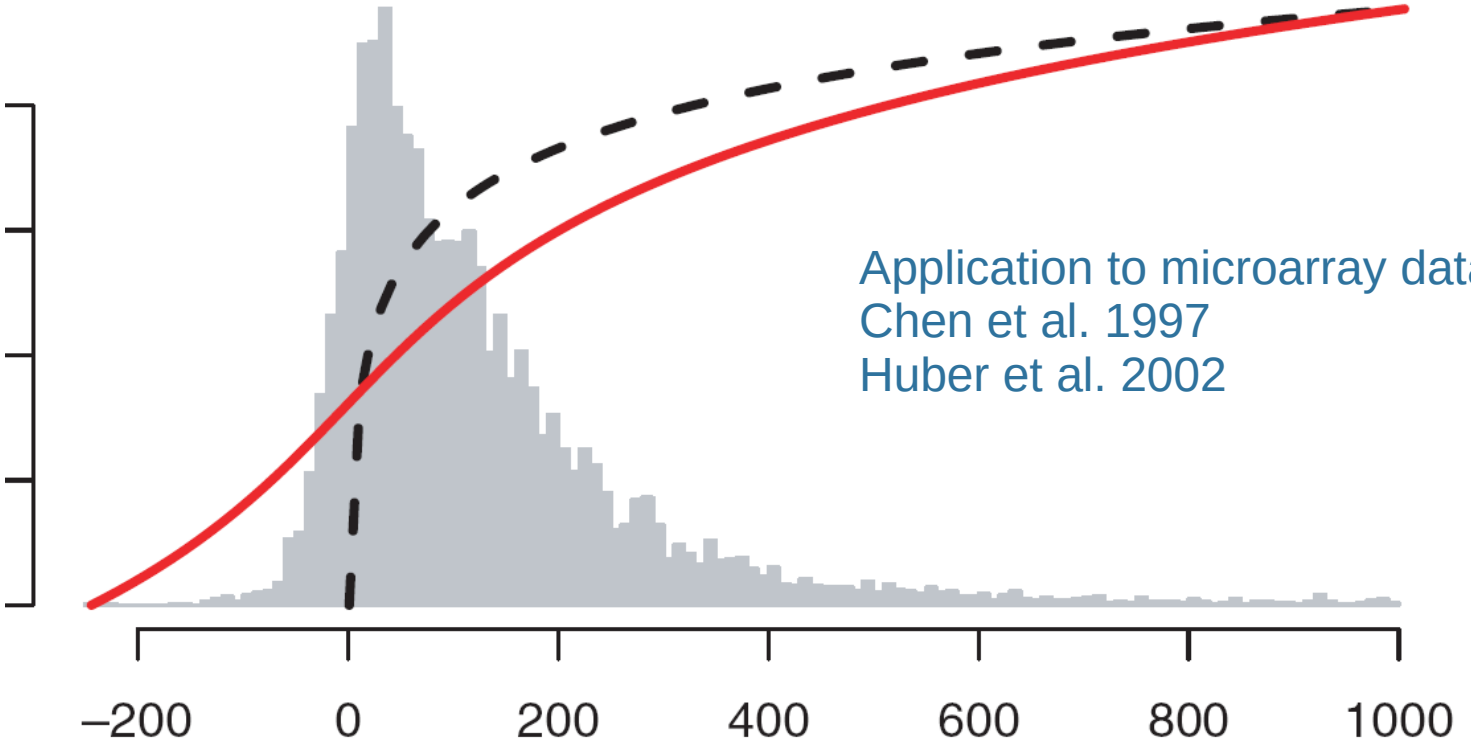


Protein count table

Proteins	Sample 1	Sample 2	Sample 3	Sample 4
Protein 1	5	7	3	….
Protein 2	340	370	322	…
…				

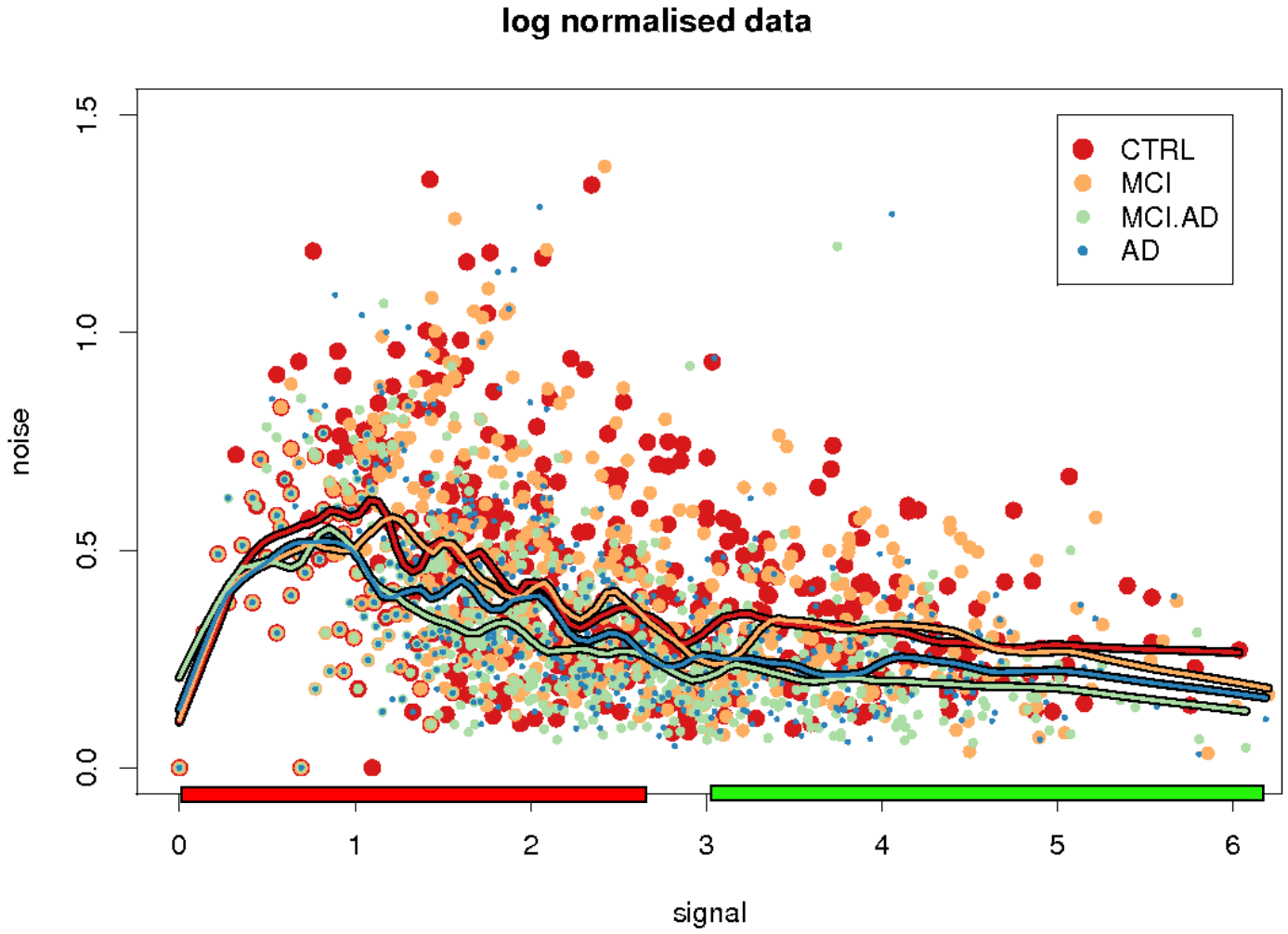
Variance stabilizing transformation

Idea: find a transformation of the original signals so that transformed data have uniform variance across all signal intensities

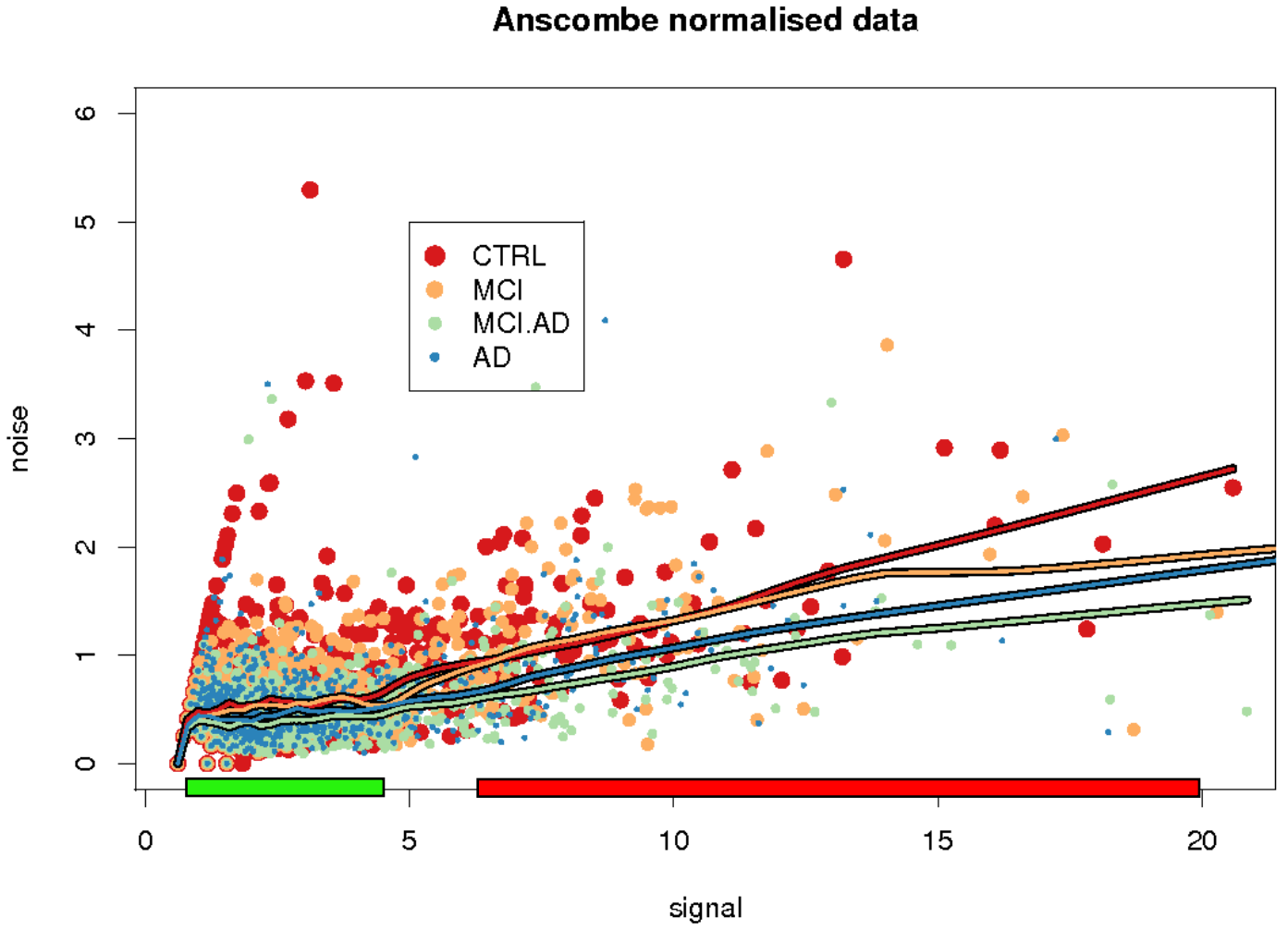


Application to microarray data:
Chen et al. 1997
Huber et al. 2002

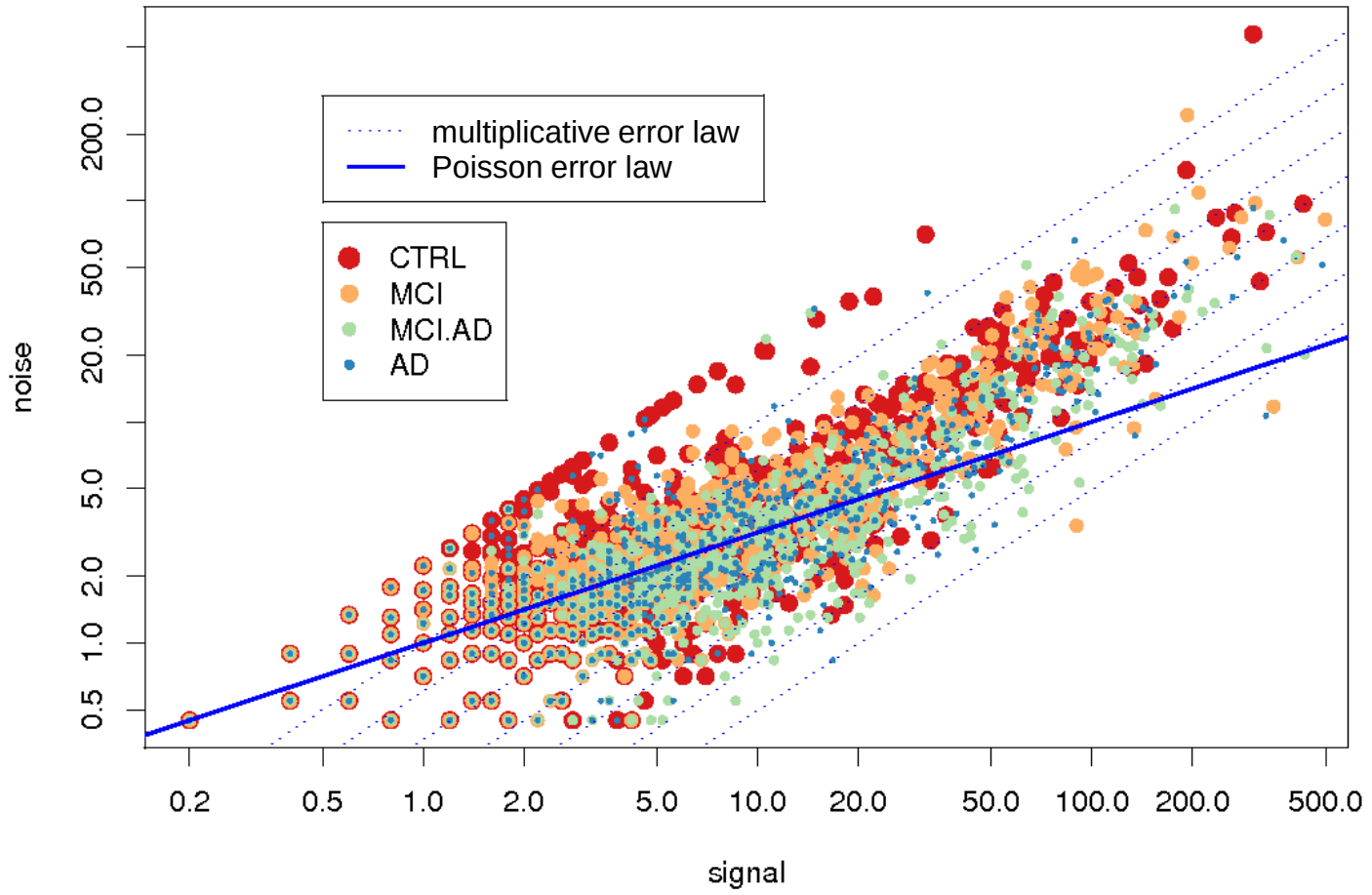
Effect of log transformation on data variance



Poisson law derived transformation



Heterogeneous behavior of variance in original data

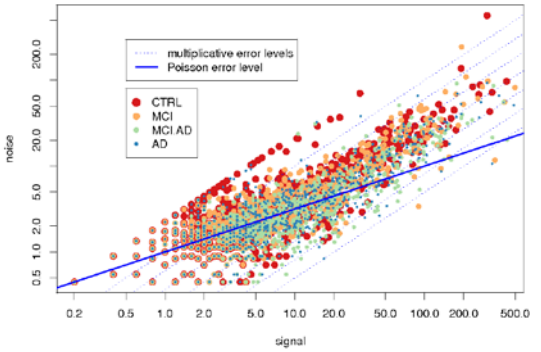


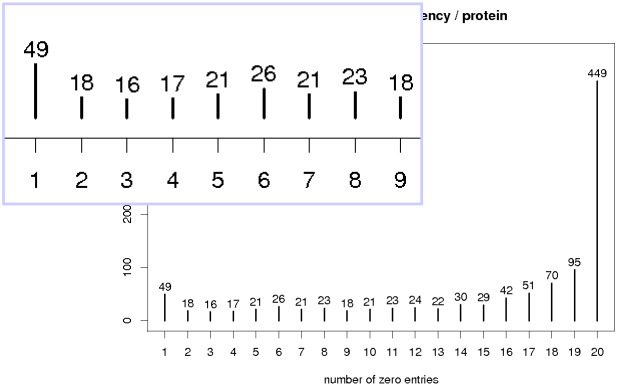
Heterogeneous behavior of variance in original data

Three effects need to be taken into account:

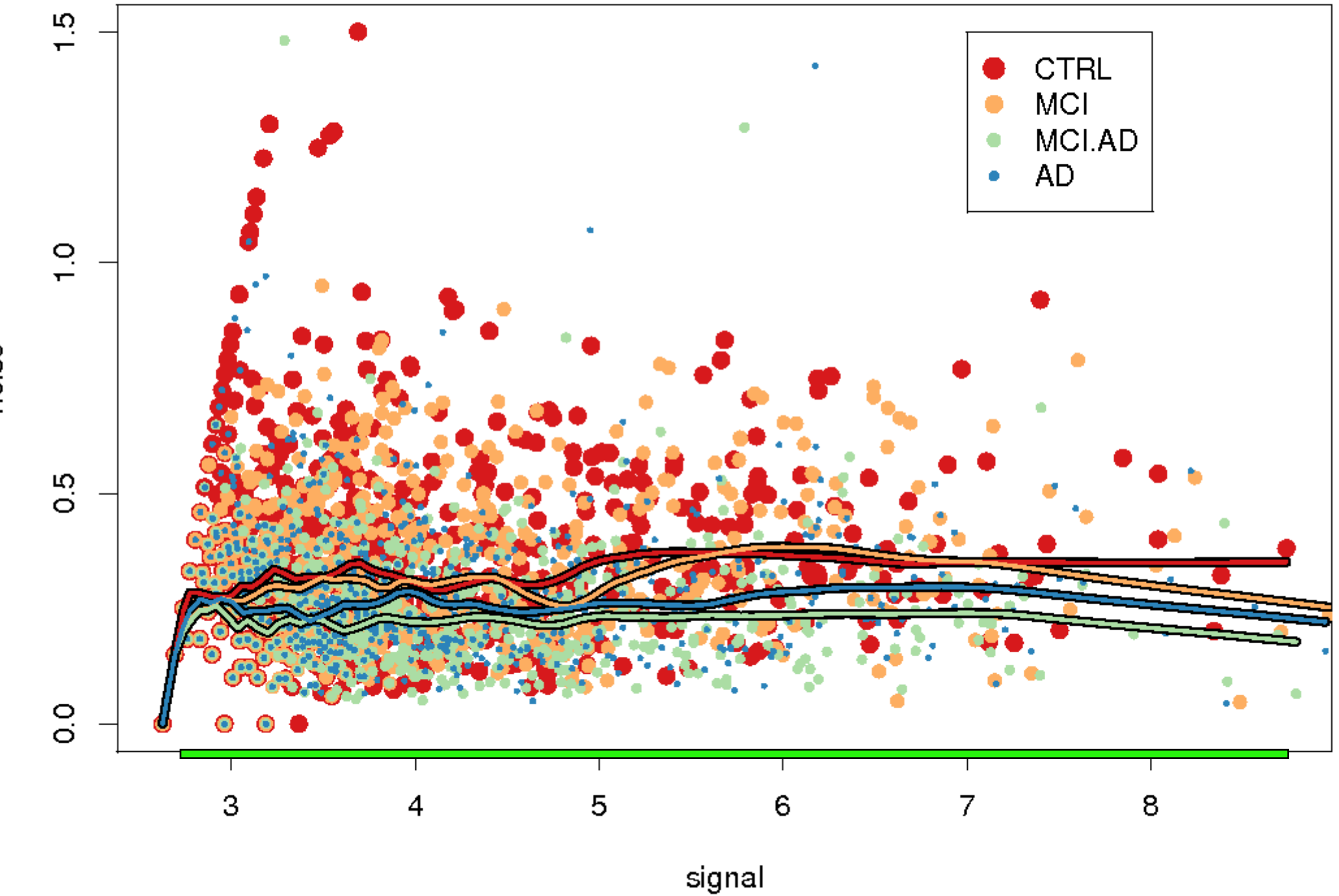
- Poisson fluctuations for weak signals (low counts)
- Biological and technological induced multiplicative noise for stronger signals
- Zero inflation (“false positive identifications”)

Solution: modified inverse binomial distribution

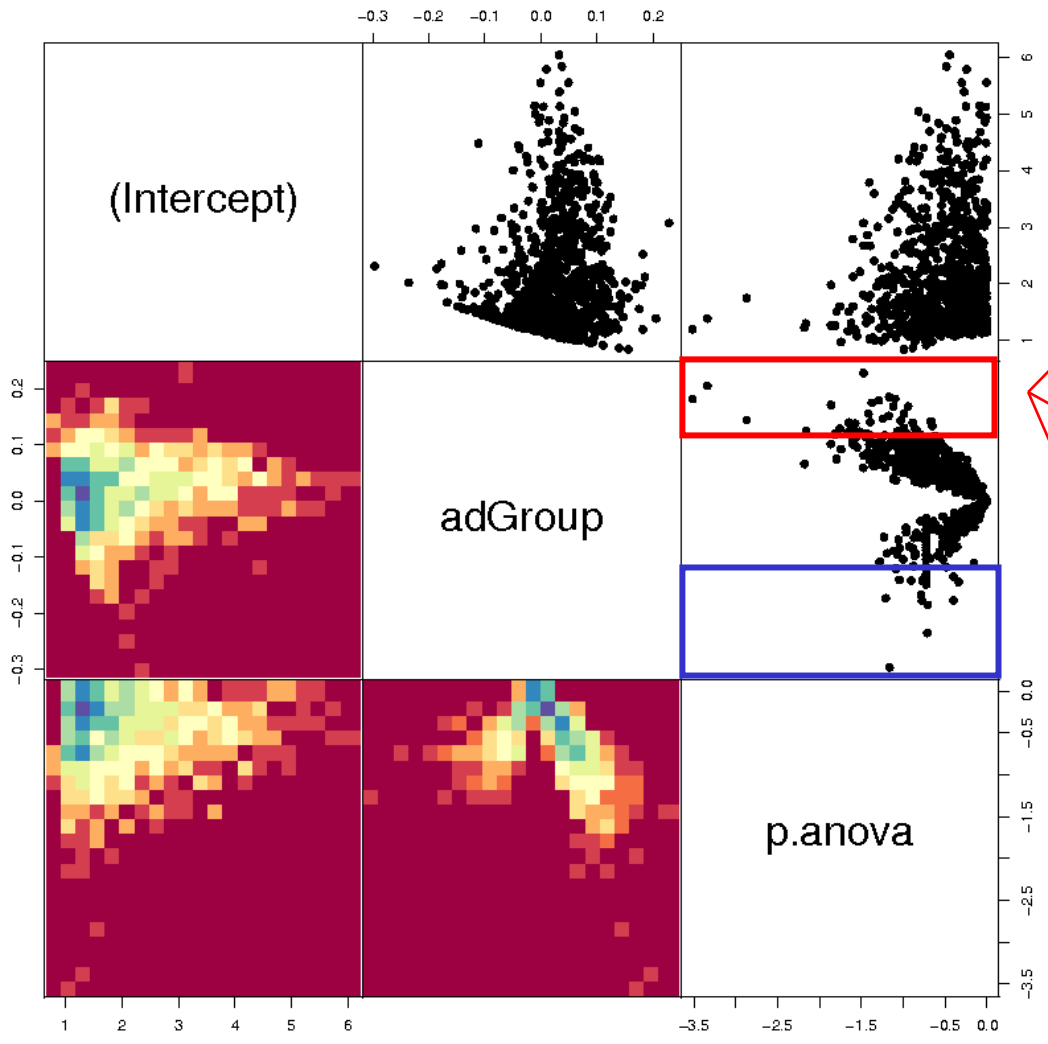


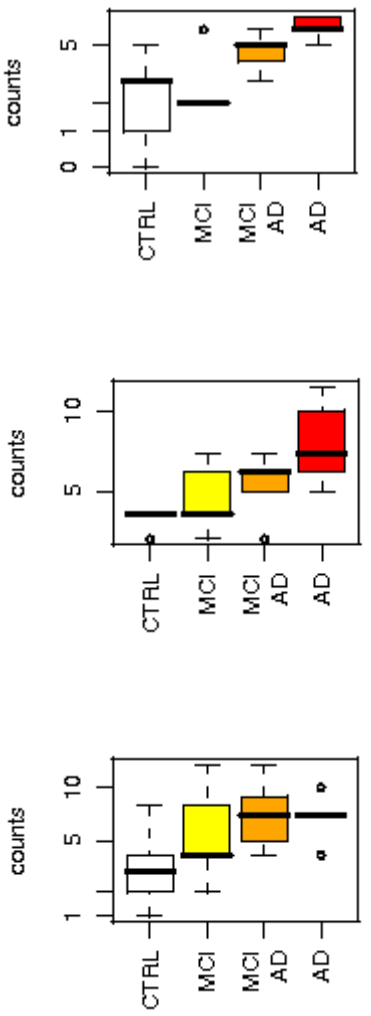


Applying an optimzed normalization strategy (mixture model)



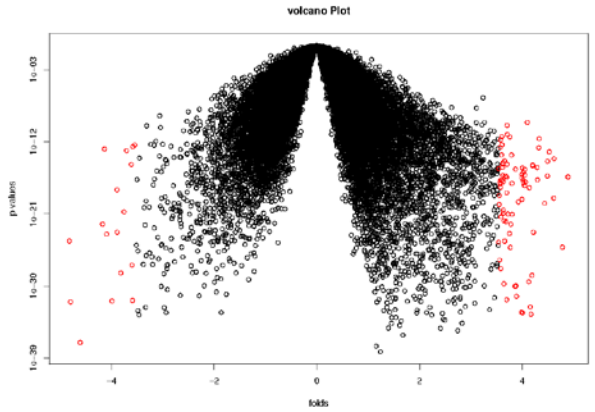
Regression and ANOVA p-values



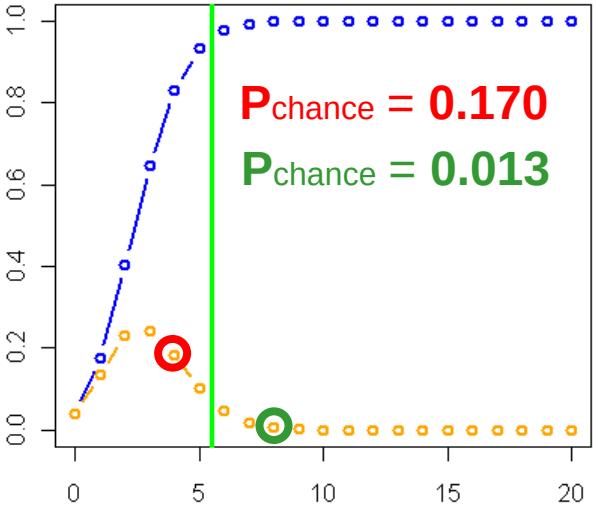


MicroDiscovery

From protein lists to pathways using Fisher's exact test



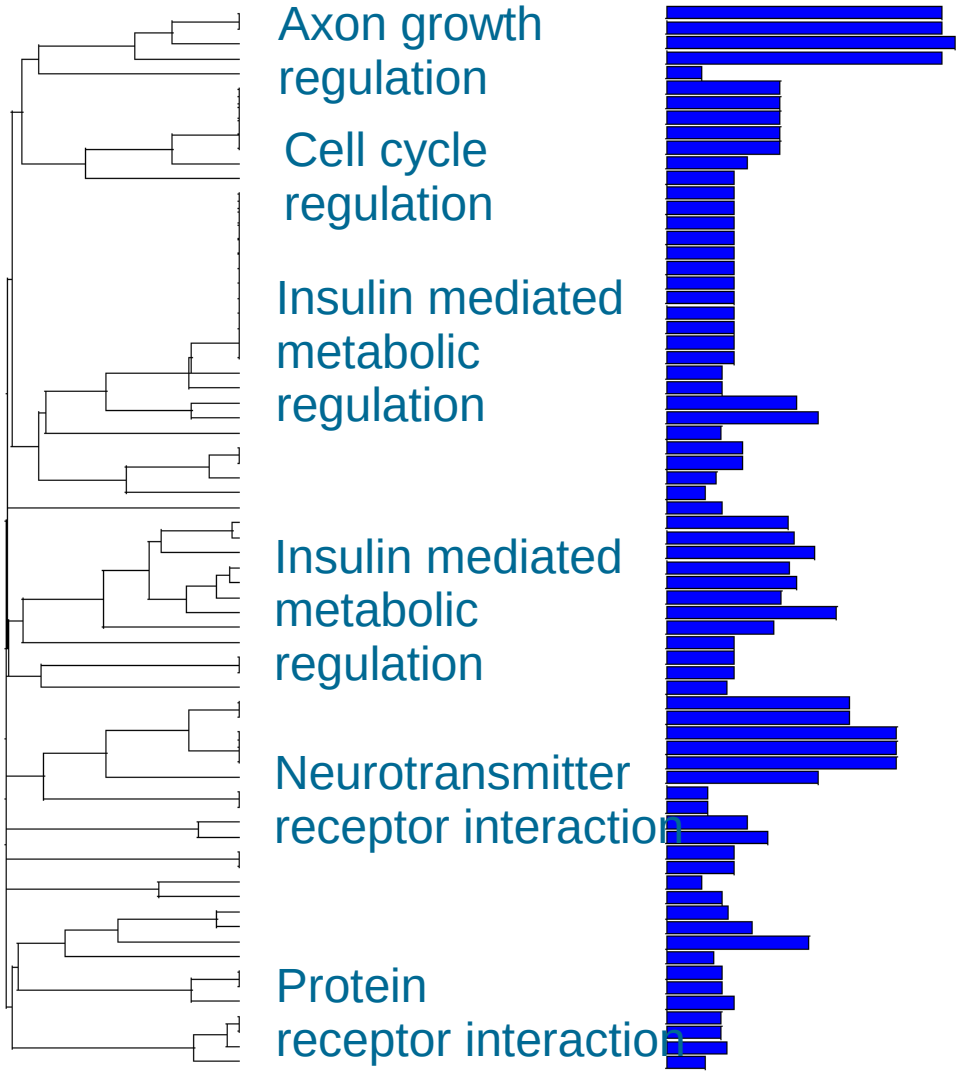
	Differential	Non diff.	Total
Pathway A	4	16	20
Gene set	150	850	1000

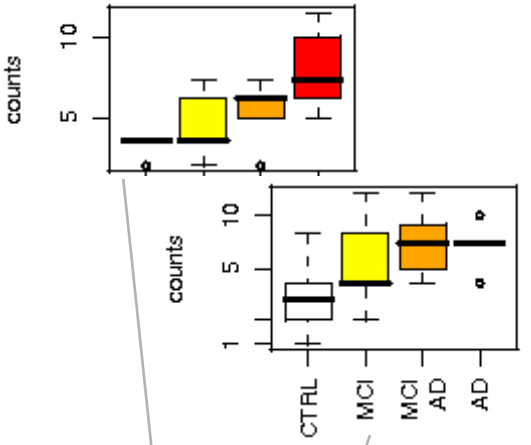


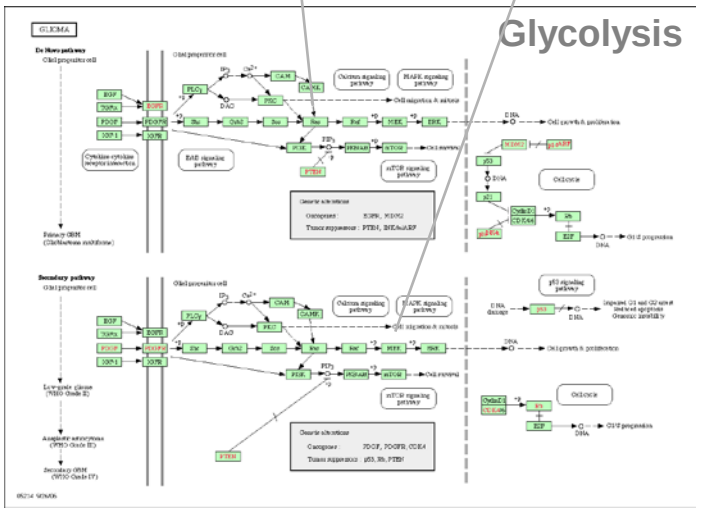
	Differential	Non diff.	Total
Pathway B	8	12	20
Gene set	150	850	1000

MicroDiscovery

Several pathways are associated with up regulated proteins

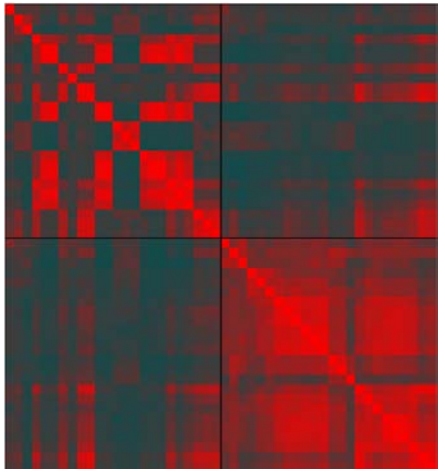




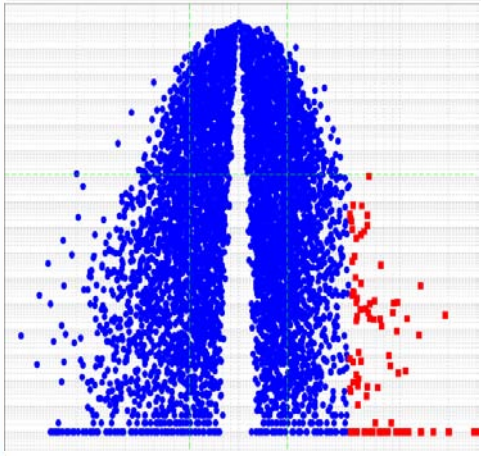


ProfileDB: typical flow of evaluation

1. Select study
2. Select assay groups to compare
3. Search for differential expression
4. Follow up
    1. Individual genes
    2. Pathways/mechanisms



Differential comparison



Set

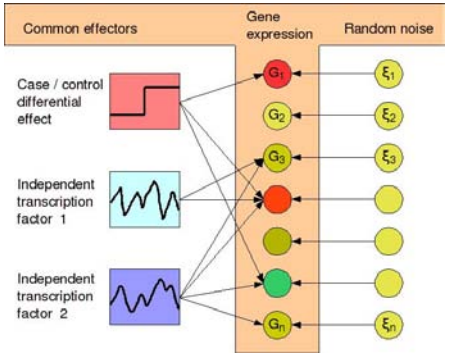
Gen 1
Gen 2
Gen 3
…

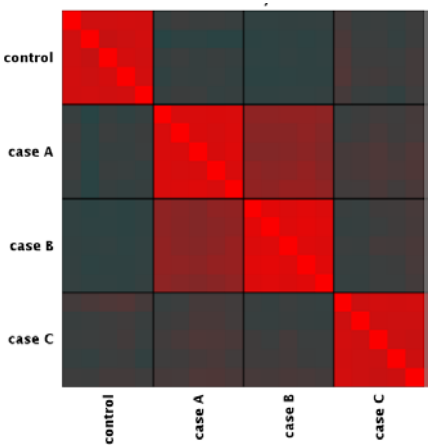
Gene set analysis:

- Pathway (KEGG/Reactome)
- Term System (GeneOntology)
- Publication (Medline abstracts)
- Models (Biomodels Database)
- …

Conclusions and outlook

- Applying the correct transformation leads to a well stabilized standard deviation even in the range of very weak signals
- Finding robust procedures for parameter estimation is a major challenge
- Understanding of the physical sources of variation can help a lot in formulating the right error model





- Novel methods for pathway detection
- Systematic evaluation of existing approaches to gene set enrichment
- Corrected p-value calculations incorporating pathway internal structure

Thank you !