

TG9: Key topics for guiding design and analysis of high-dimensional data

Harald Binder

**Division Biostatistics and Bioinformatics
University Medical Center Mainz, Germany**



UNIVERSITÄTS**medizin.**

MAINZ

Topic group 9: High-dimensional data

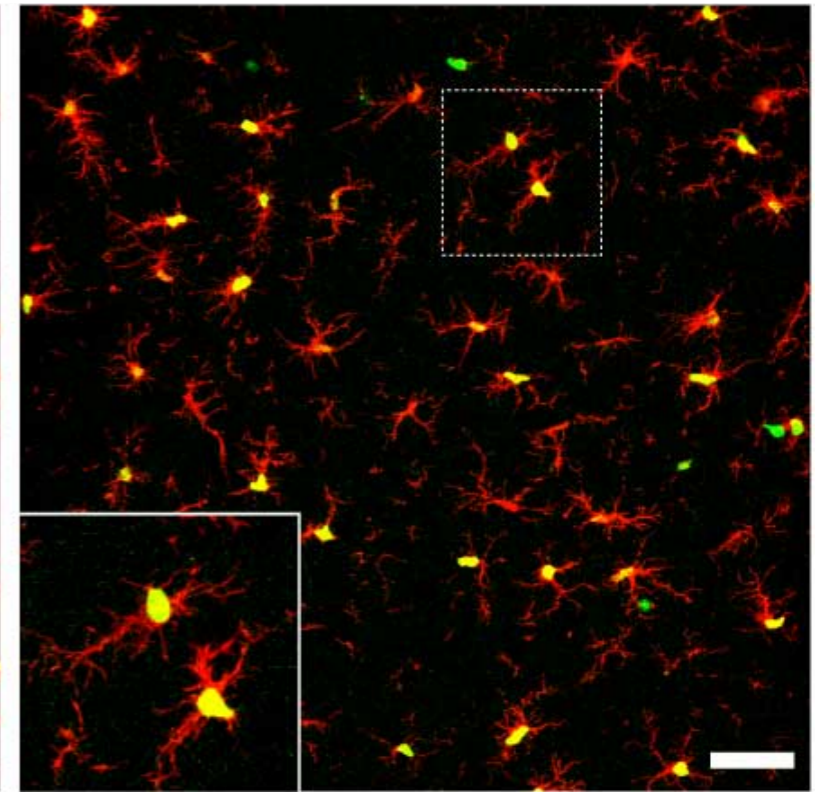
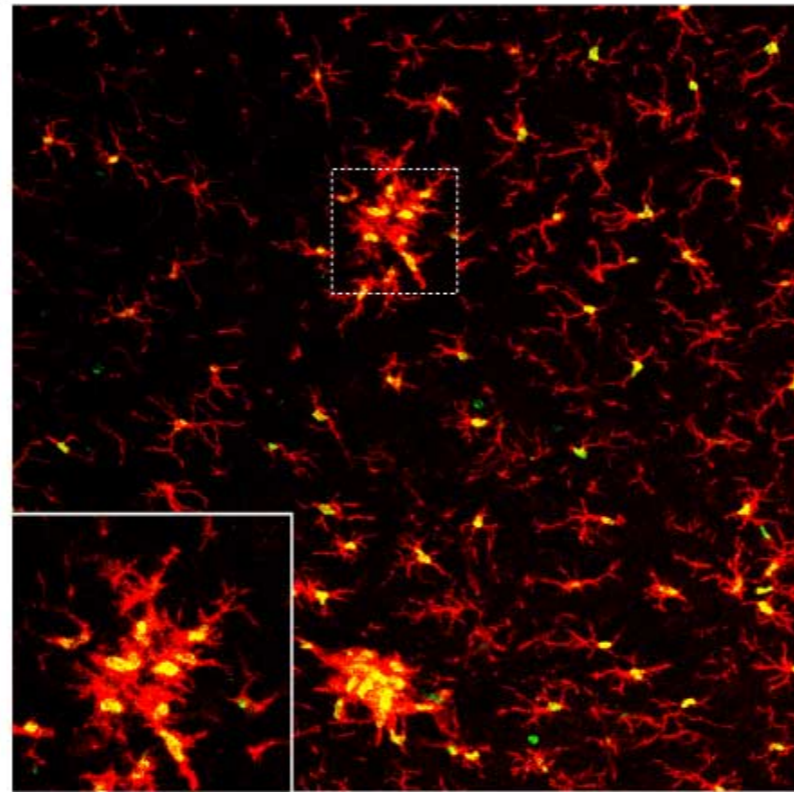
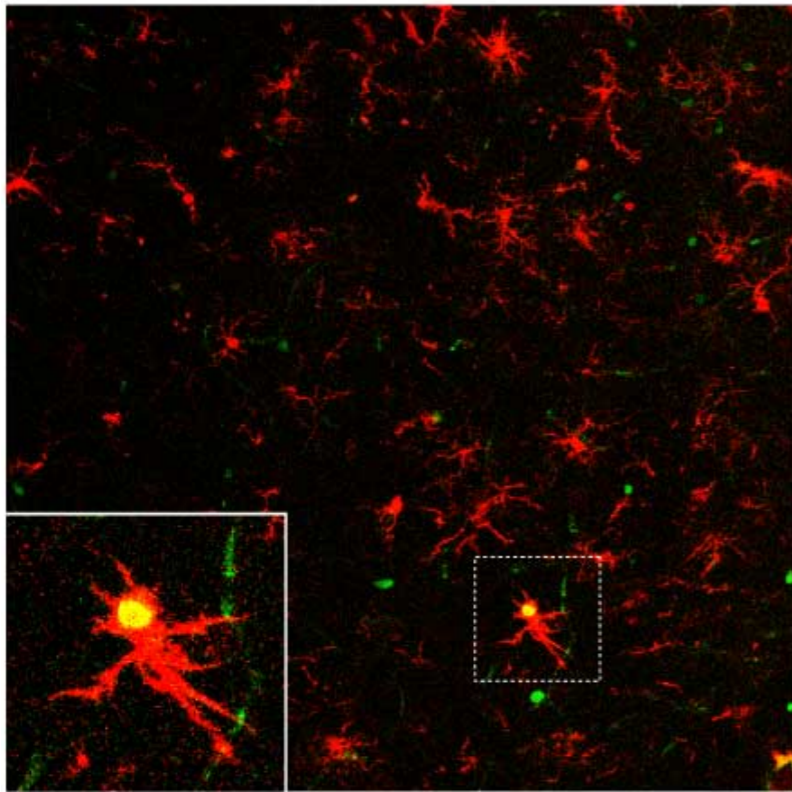
- Chairs: Lisa McShane (NCI, USA), Jörg Rahnenführer (TU Dortmund, Germany)
- Members:
 - Axel Benner (DKFZ Heidelberg, Germany)
 - Harald Binder (University Medical Center Mainz, Germany)
 - Anne-Laure Boulesteix (LMU Munich, Germany)
 - Tomasz Burzykowski (Hasselt University, Belgium)
 - W. Evan Johnson (Boston University, USA)
 - Lara Lusa (University of Ljubljana, Slovenia)
 - Stefan Michiels (University Paris-Sud, France)
 - Sherri Rose (Harvard Medical School, USA)

Day 3

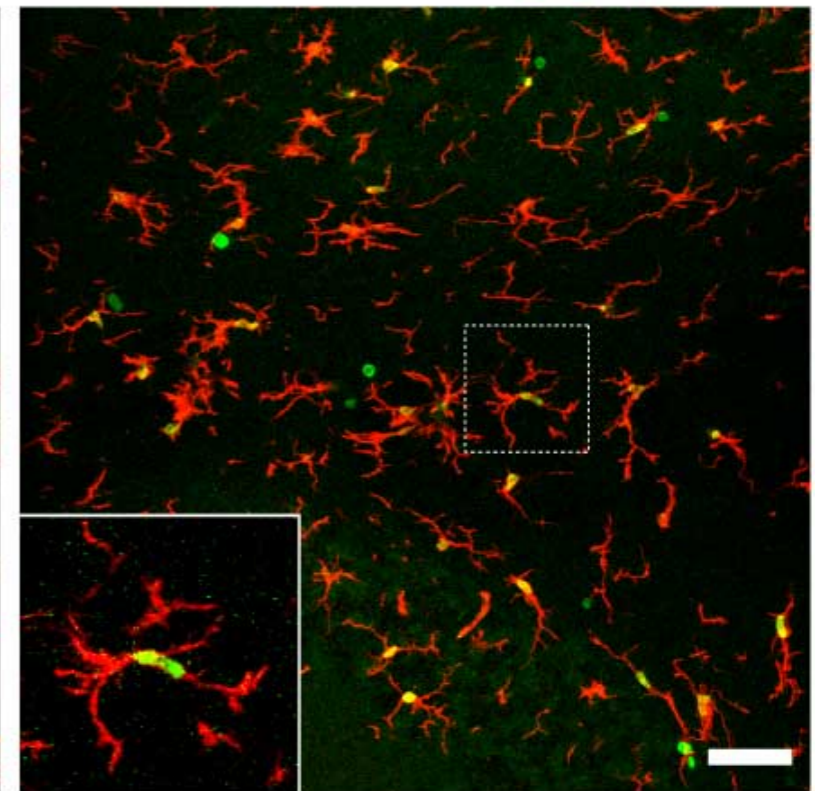
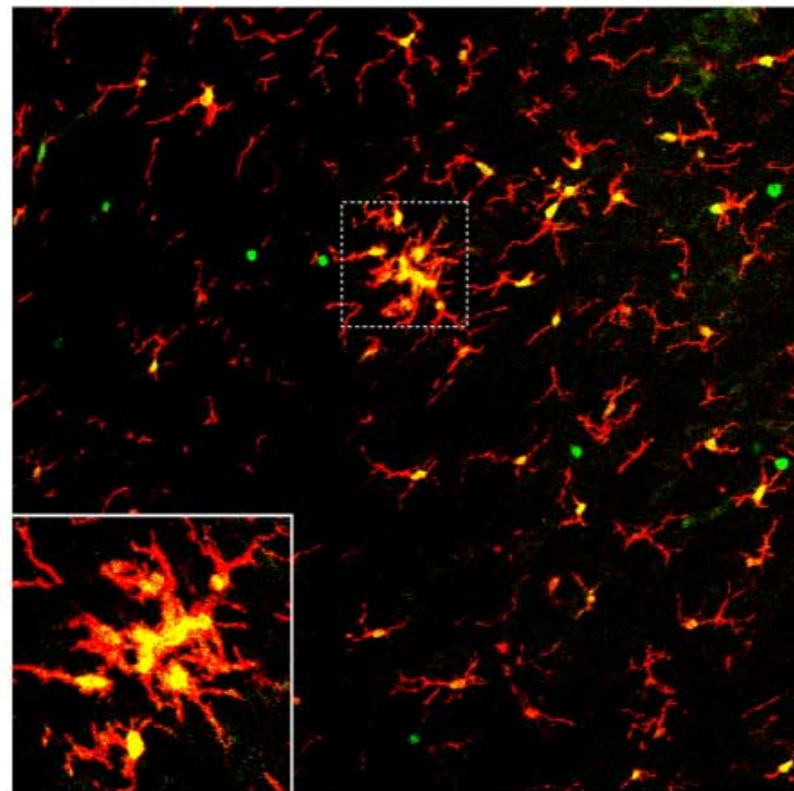
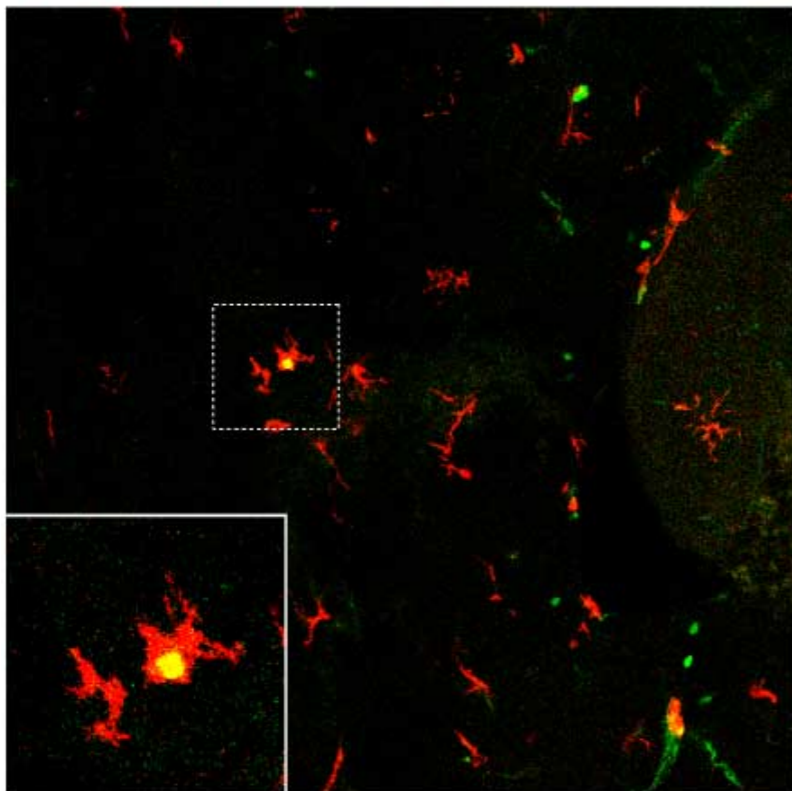
Day 7

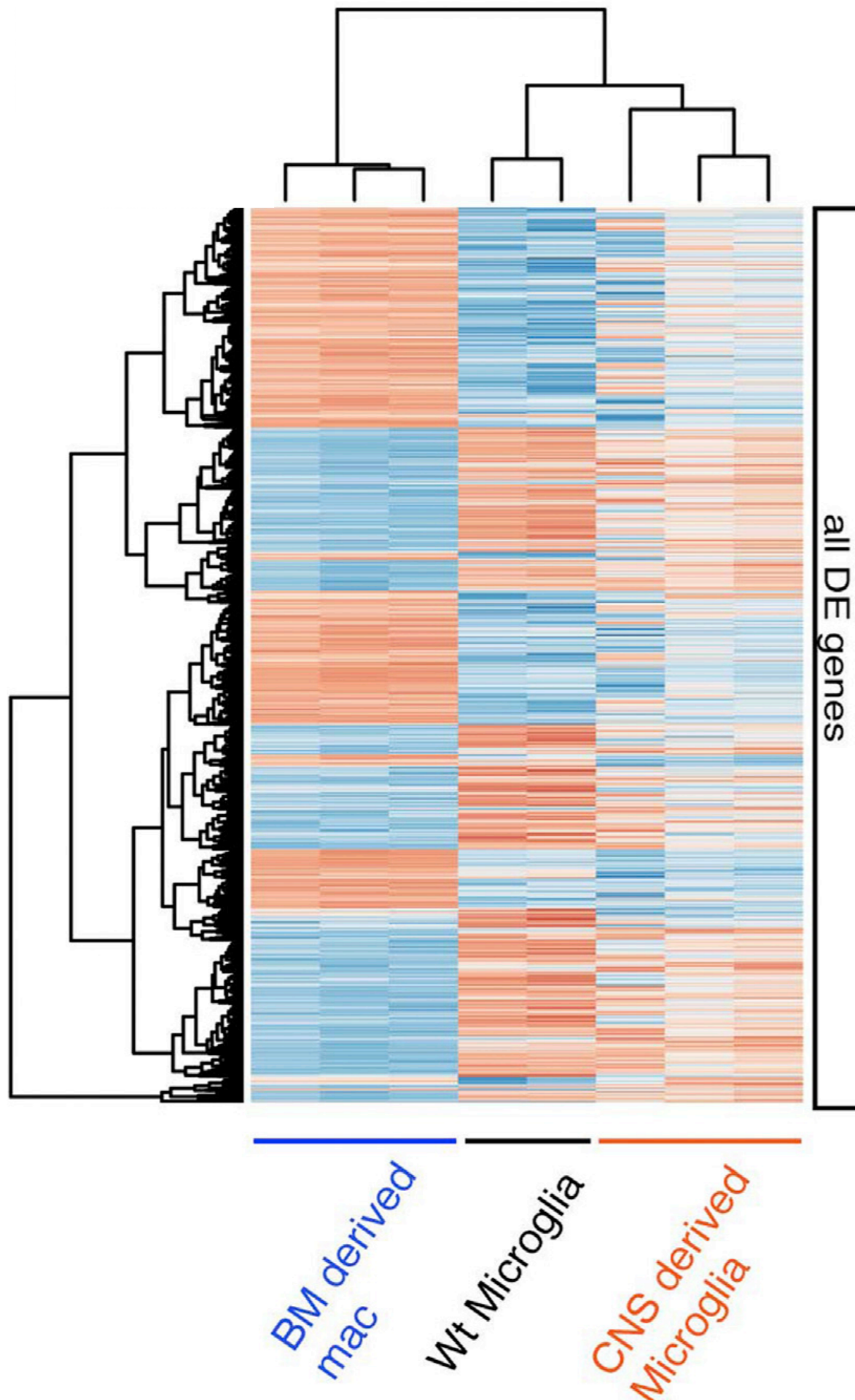
Day 14

Cortex



Cerebellum



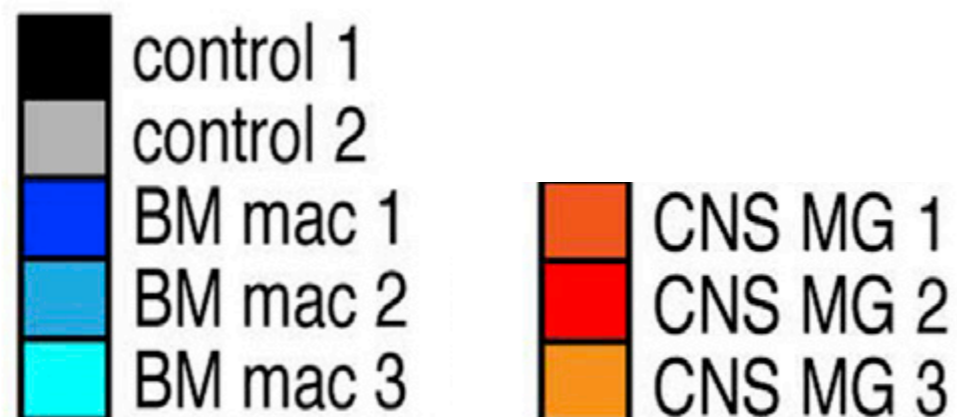
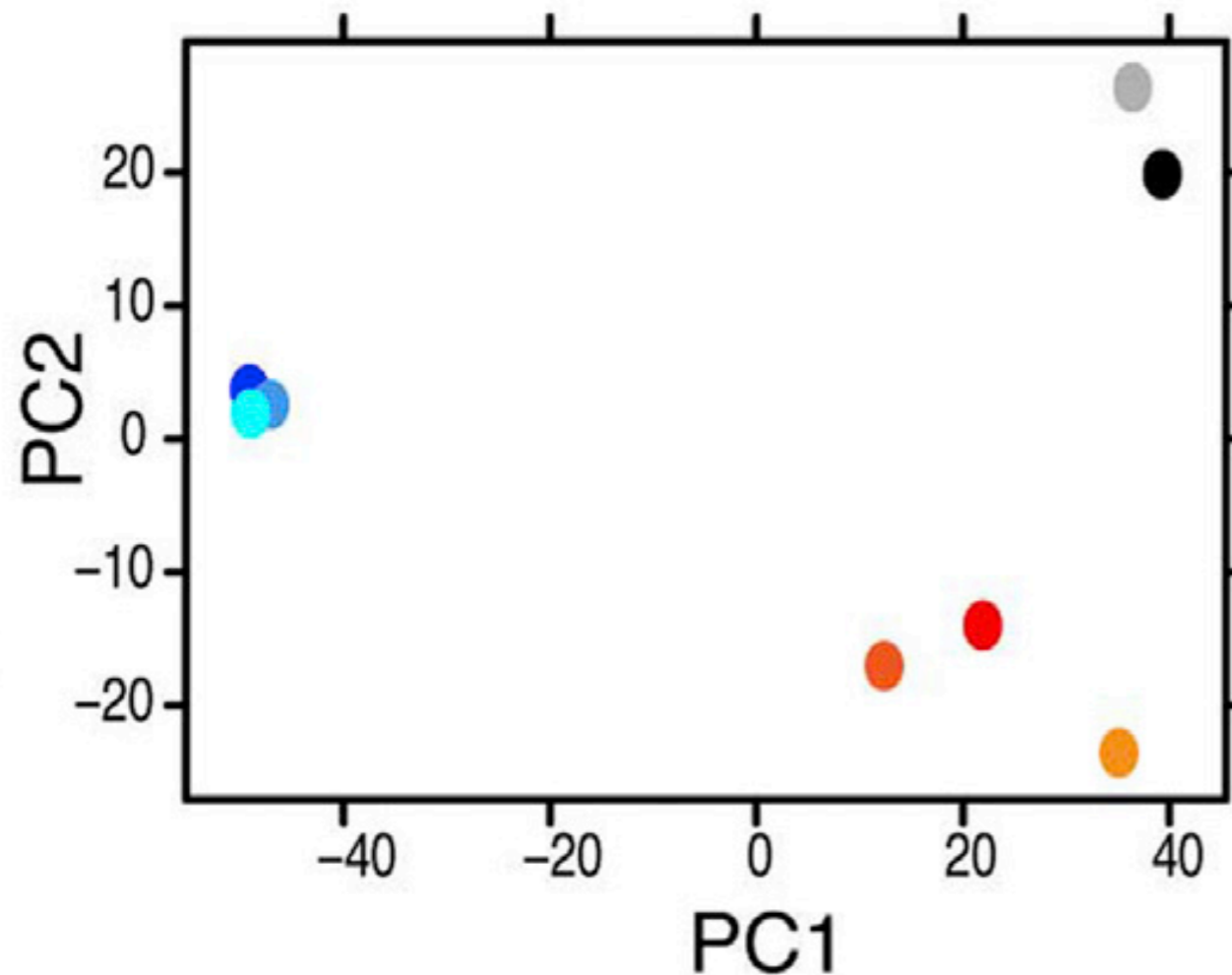


How to ...

- check for batch effects
- perform normalization
- group similar samples
- group similar genes
- select representative genes
- deal with multiple testing

Subtopics (1)

- Data preprocessing
 - normalization/calibration
 - identification of outliers/errors
- Exploratory data analysis
 - graphical displays
 - clustering approaches
 - integrative analysis of different data types
- Multiple testing
 - biomarkers differentially expressed between groups
 - common set of explanatory variables on a large set of outcomes

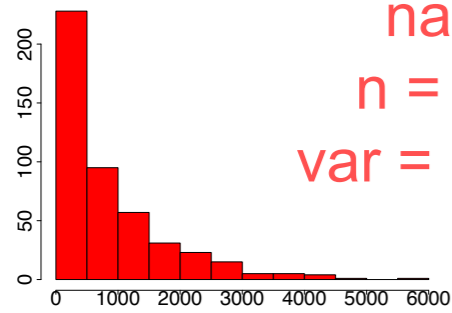


How to ...

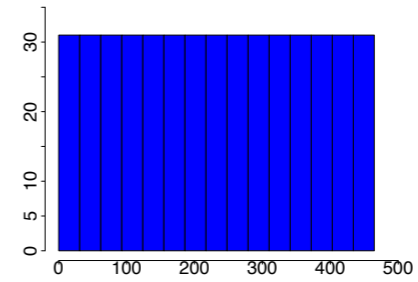
- find a low-dimensional representation
- perform inference there
- perform simulations for checking properties

Subtopics (2)

- Data reduction
 - Tasks: visualization of samples or variables, building/finding prototypical samples, building new features
 - Traditional approaches: principal components, multidimensional scaling, correspondence analysis, cluster analysis
 - Current research: representation learning, deep learning
- Simulation
 - Distributions (RNA-seq, methylation microarrays, ...)
 - Simulation using extracted parameters
 - Simulation approaches
 - Simulation based on real data



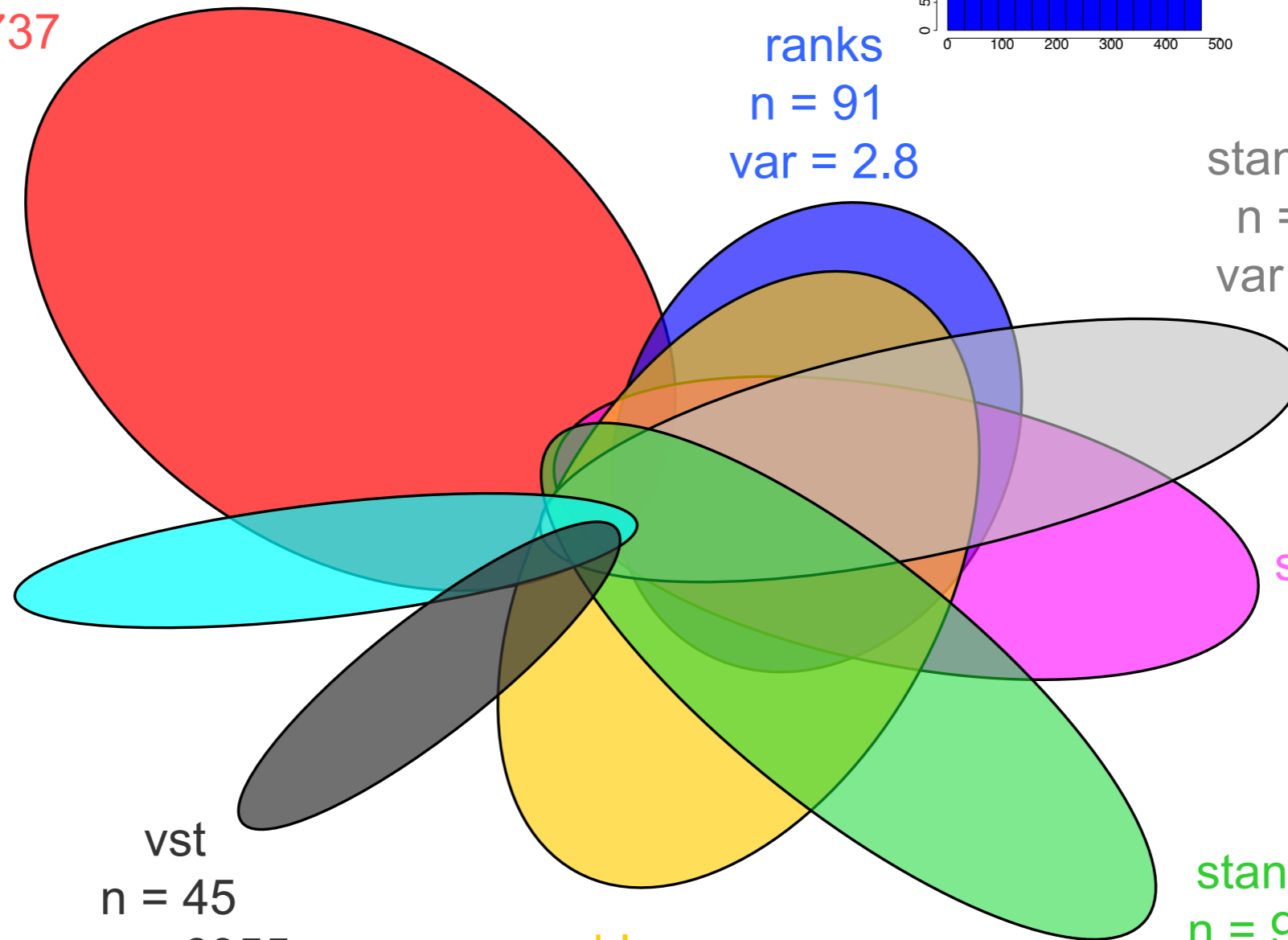
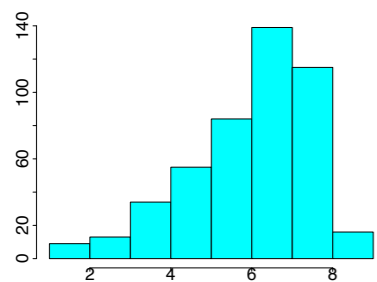
naive
n = 126
var = 17737



ranks
n = 91
var = 2.8

stand. vst
n = 83
var = 1.9

logs
n = 55
var = 588

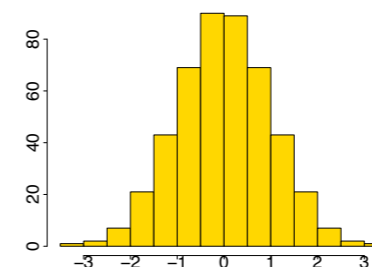
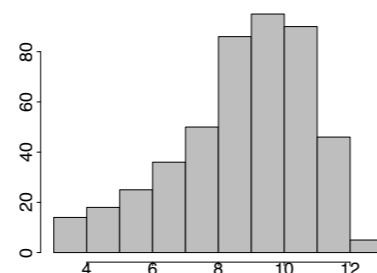


stand. logs
n = 78
var = 2.0

vst
n = 45
var = 6955

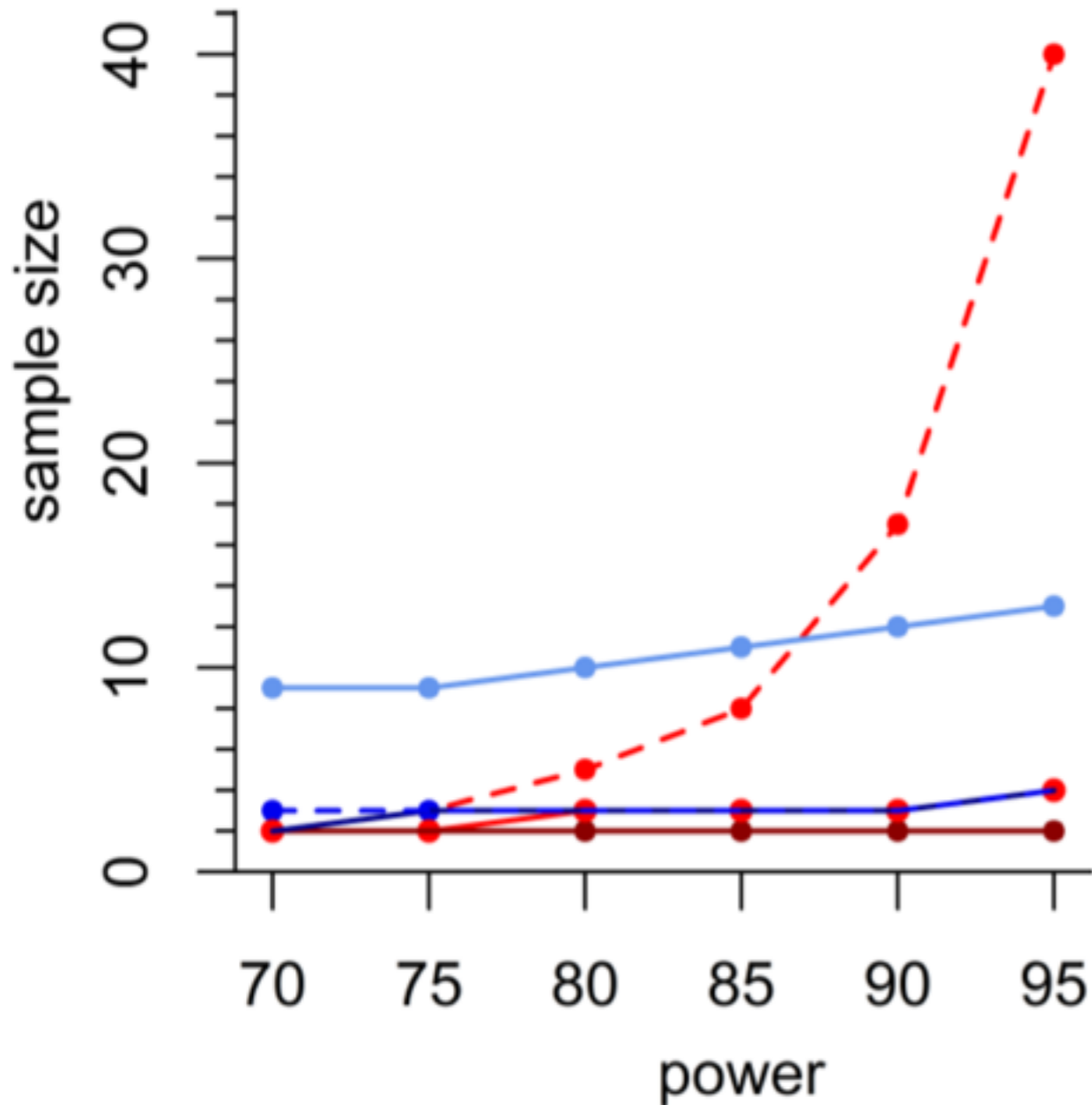
stand.
n = 92
var = 2.6

blom
n = 112
var = 1.1

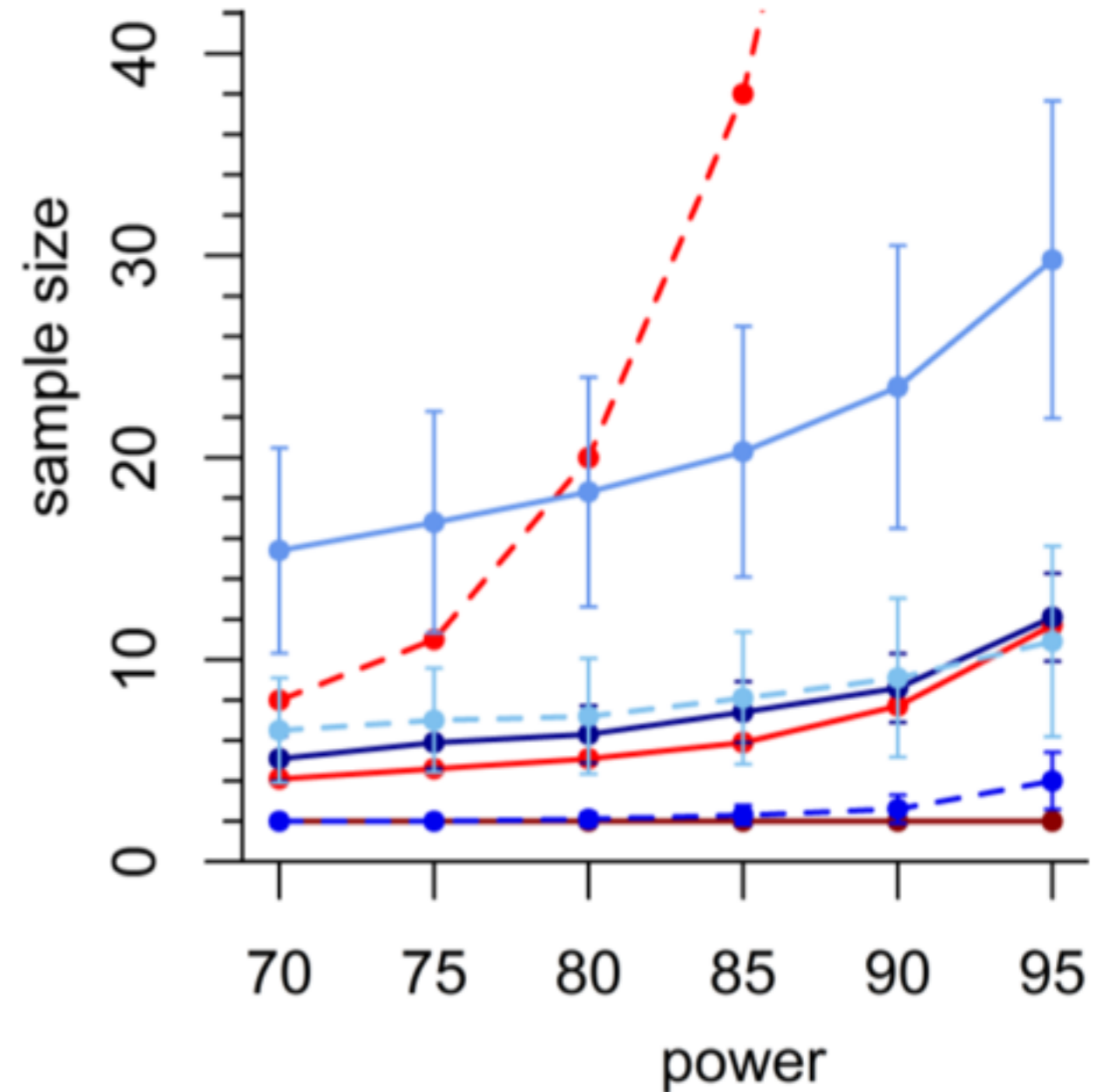


Subtopics (3)

- Prediction models
 - Machine learning methods
 - Penalized regression
 - Evaluation
- Categorical and ordinal data
- Comparative effectiveness and causal inference
- Design considerations
 - Sample size planning and power calculation
 - Experimental design for observational studies
- Publicly available data sets

a Sample size for mouse data

- powerSampleSizeCalculator
- - - PROPER Bottomly/Cheung
- PROPER pilot

b Sample size for human data

- RnaSeqSampleSize
- Scotty
- - - ssizeRNA
- - - SSPA

Outlook

- Large topic with links to Bioinformatics and Systems Biology
- But: high-dimensional challenges also in non-omics settings
- Overlap with many other topic groups, but always with high-dimensional flavor
- Subtopics make progress feasible
- First drafts of papers by end of 2016