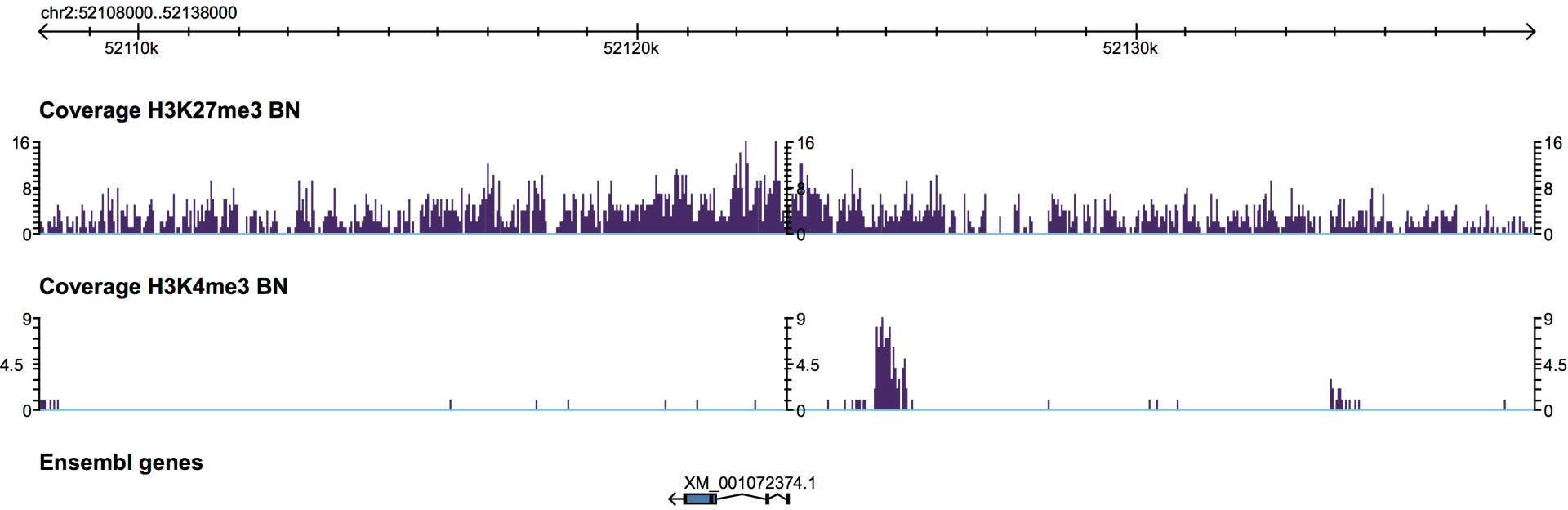
Introduction to ChIP-seq peak calling and differential peak calling

Matthias Heinig

Institute of Computational Biology

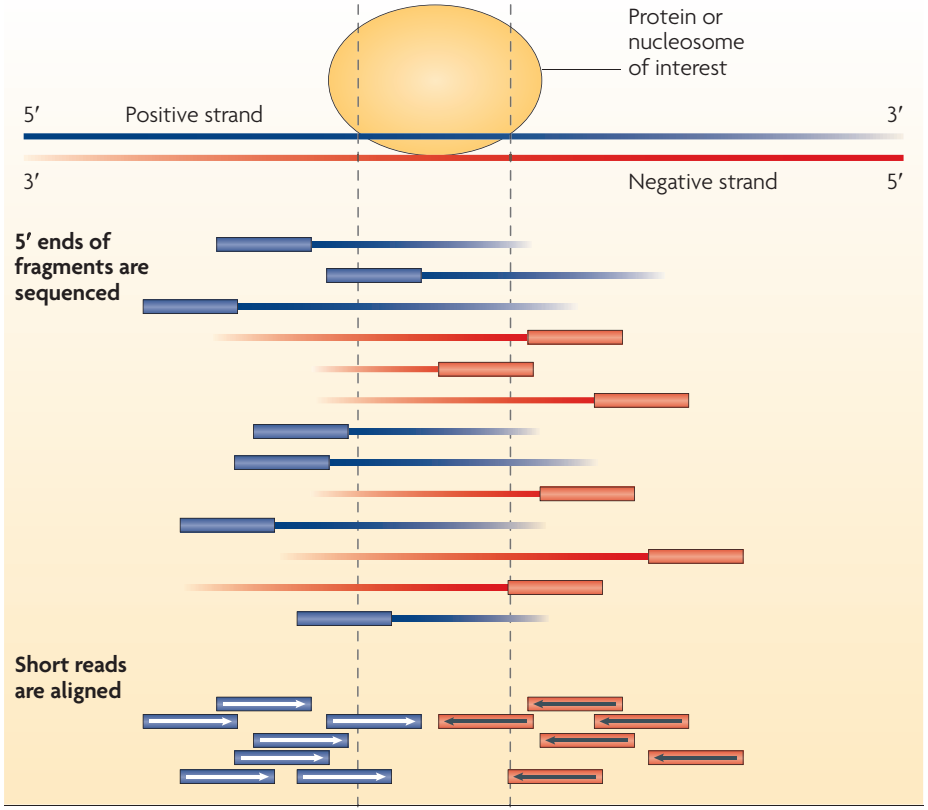
HelmholtzZentrum münchen
German Research Center for Environmental Health

ChIP-seq tracks



- Signal to noise ratio
  - High for peak like features
  - Low(er) for large domain features

Coverage / read counts

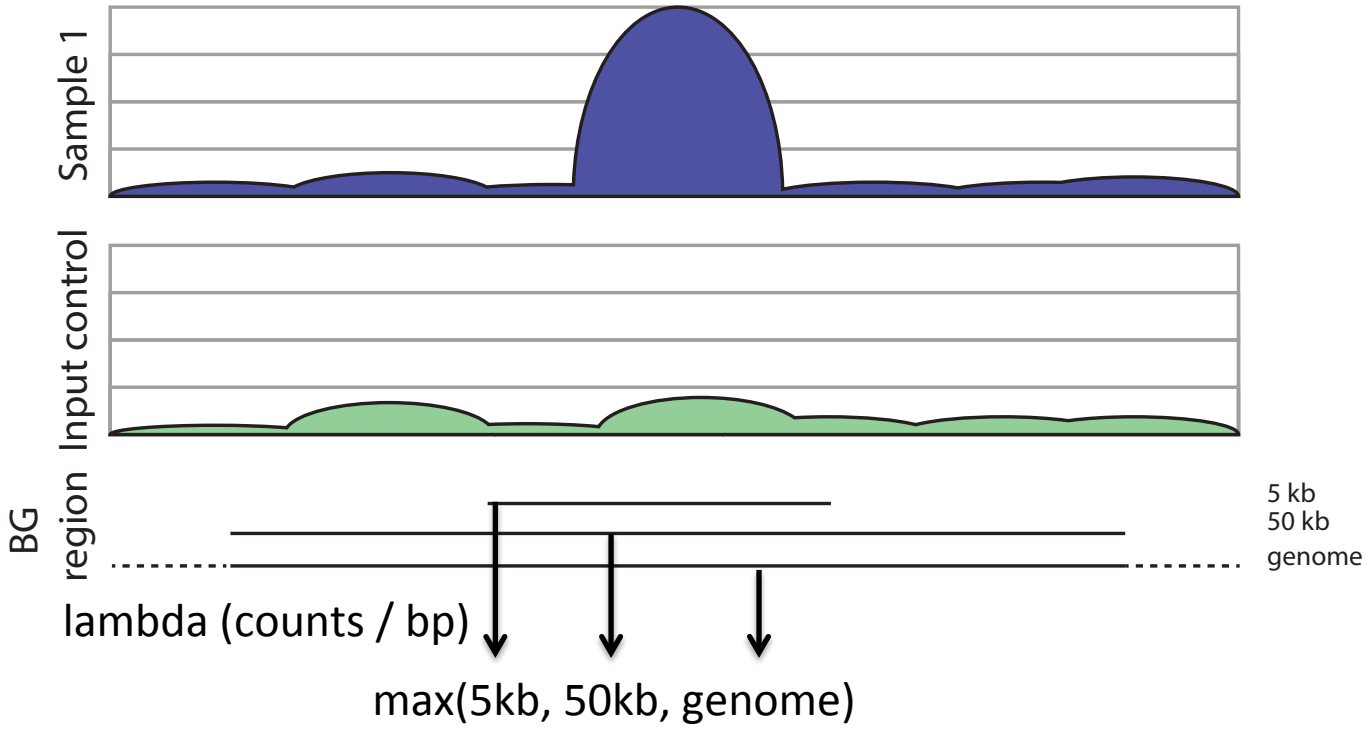


Coverage: how many fragments are aligned to each position?
Read counts: how many fragments start in each bin?

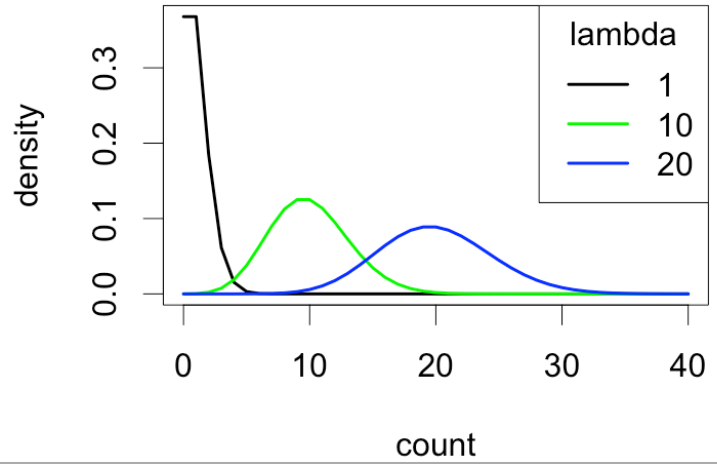
Peak calling strategies

- Statistical testing
  - Modeling the null (background) distribution
    e.g. MACS (Zhang 2008)
- Probabilistic modeling
  - Modeling the signal and the background distribution
    e.g. Zinba (Rashid 2011)
- Signal processing approaches
  - Filter
    e.g. Dfilter (Kumar 2013)

MACS



Poisson distribution



Probabilistic modeling: ZINBA

Typically NGS data show overdispersion (variance greater than Poisson lambda)

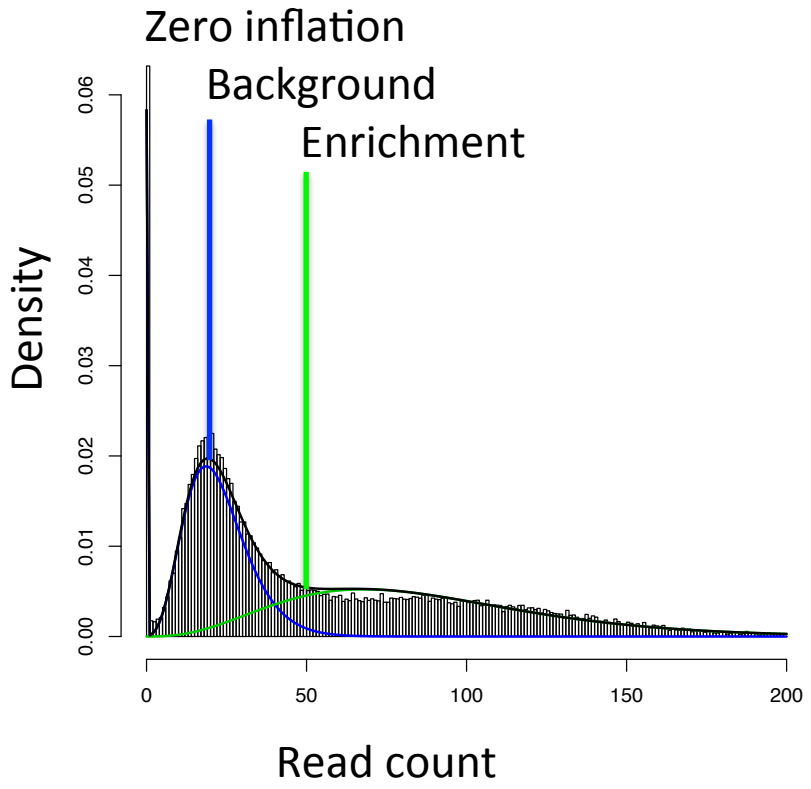
- Negative binomial distribution

Many empty bins distort estimation of background

- Zero inflation component

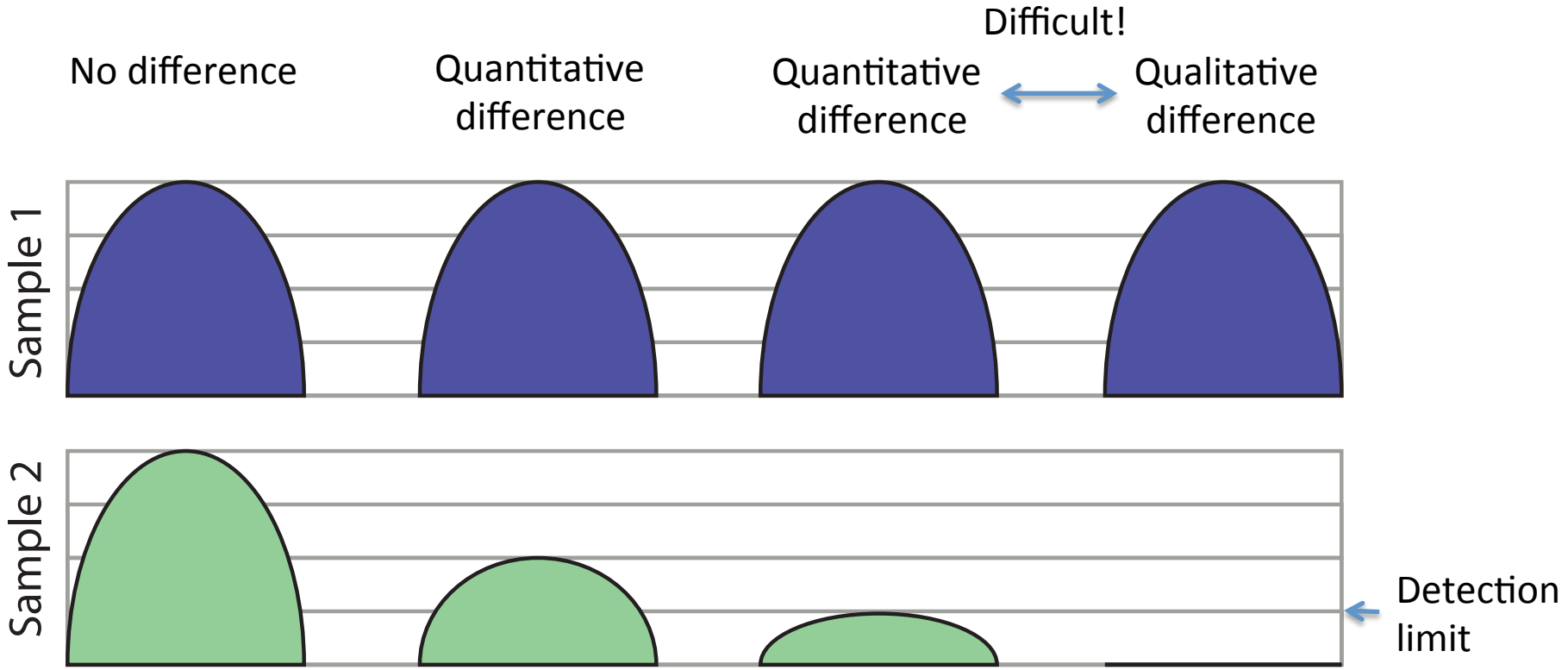
ZINBA (zero inflated negative binomial algorithm)

- 3 component mixture model



Comparison of epigenetic tracks

Goal: identify regions that differ between samples



Comparison of epigenetic tracks

Challenges:

- Actually two problems
  - Identification of features
  - Comparison of features
- Calling absence of peaks
- Use of input tracks?
- No gold standard for the evaluation of methods

Strategies

Where to look?

- Independent peak detection analyses
- Sliding window approaches
- Binning

How to compare?

- Comparison absence / presence of calls
- Quantitative comparison
- Normalization
- Hypothesis testing
- Probabilistic modeling
- Considering local dependencies

Available tools (selection)

Tool		Diffbind	PePr	diffReps	RSEG	Chipdiff	histoneHMM
where to look	peak detection	✓	✗	prescreening	✗	prescreening	✗
	sliding window	✗	✓	✓	✗	✗	✗
	binning	✗	✗	✗	✓	✓	✓
how to compare	use of input	✗	substract input	✗	✗	✗	✗
	normalization	scaling	scaling	scaling	✗	✗	✗
	hypothesis testing	negative binomial	negative binomial	negative binomial	✗	✗	✗
	probabilistic	✗	✗	✗	NBDiff	Hierarchical binomial model	Multivariate NB
	dependencies	✗	merging	merging, hotspot detection	HMM	HMM	HMM

METHOD

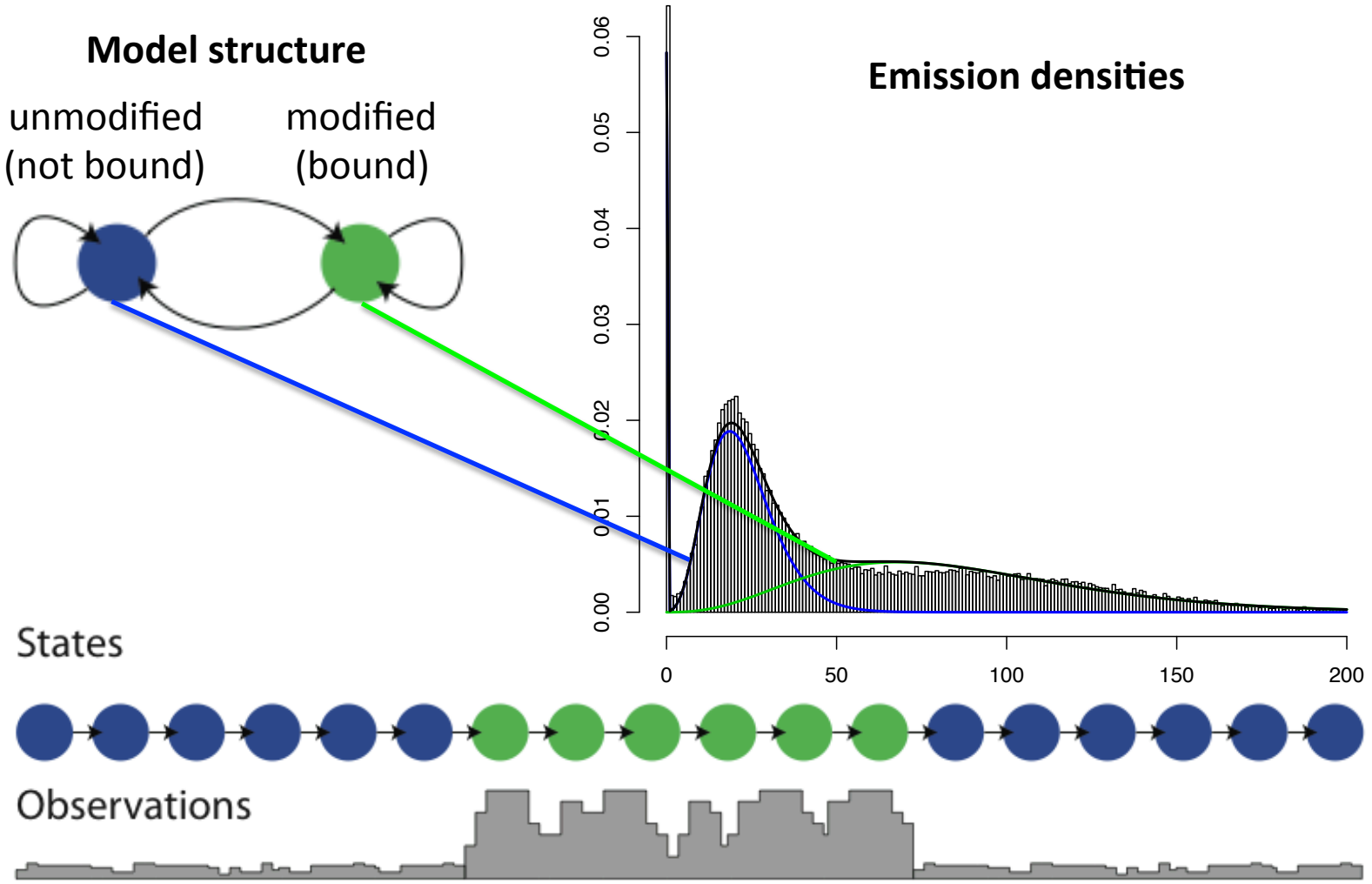
histoneHMM: Differential analysis of histone modifications with broad genomic footprints

Matthias Heinig[1*], Maria Colomé-Tatché[3], Aaron Taudt[3], Carola Rintisch[2], Sebastian Schafer[2], Michal Pravenec[4], Norbert Hubner[2], Martin Vingron[1] and Frank Johannes[5]

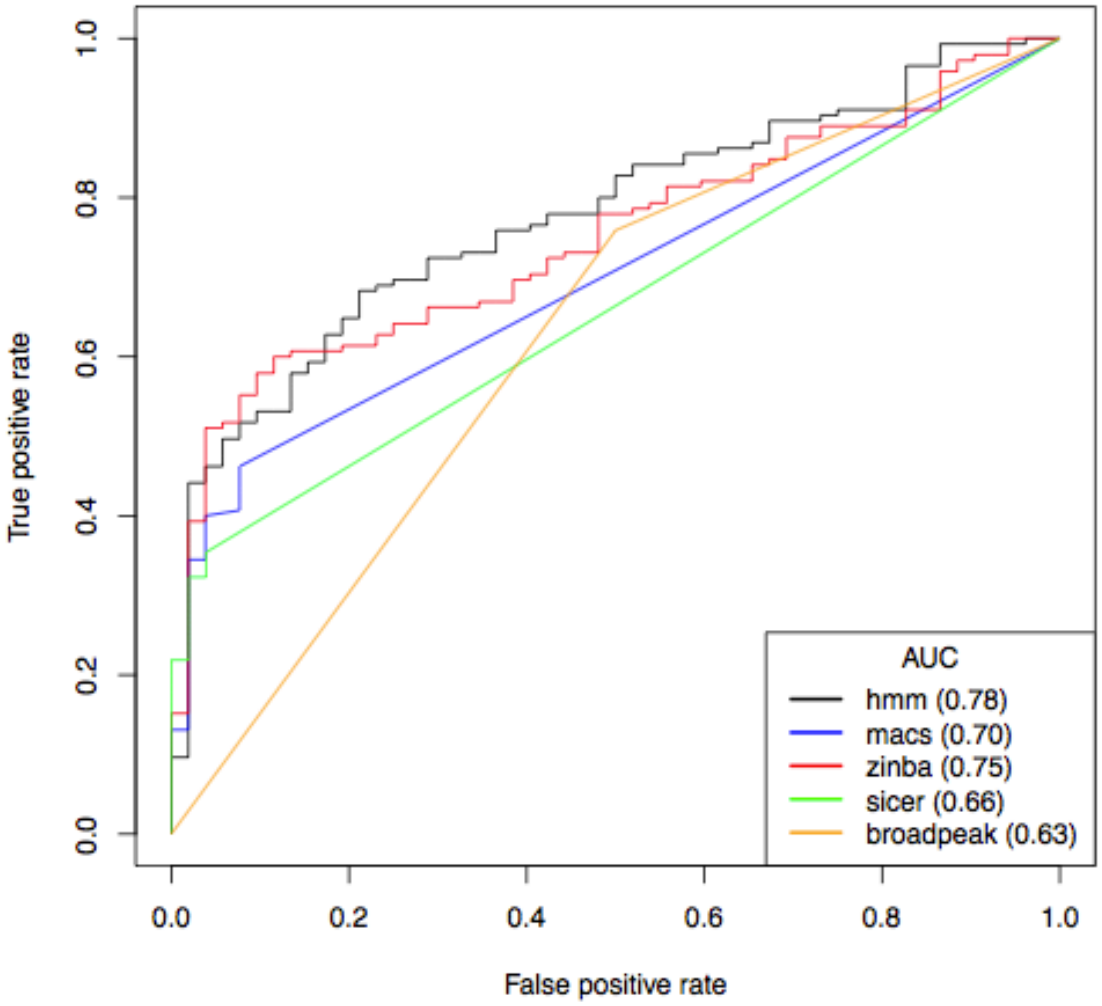
Matthias Heinig

Max Planck Institute for molecular genetics

HMM for region calling

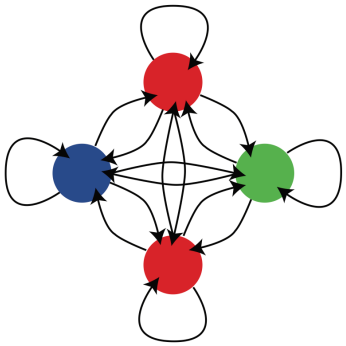


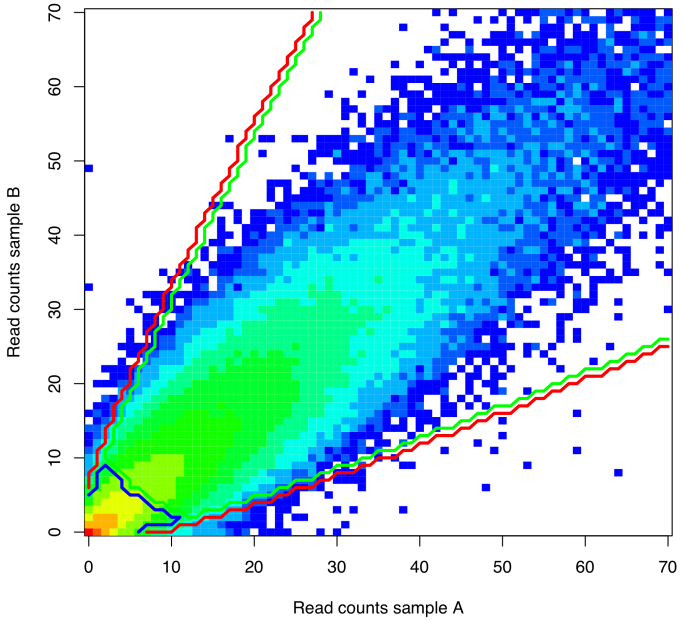
Evaluation with qPCR data

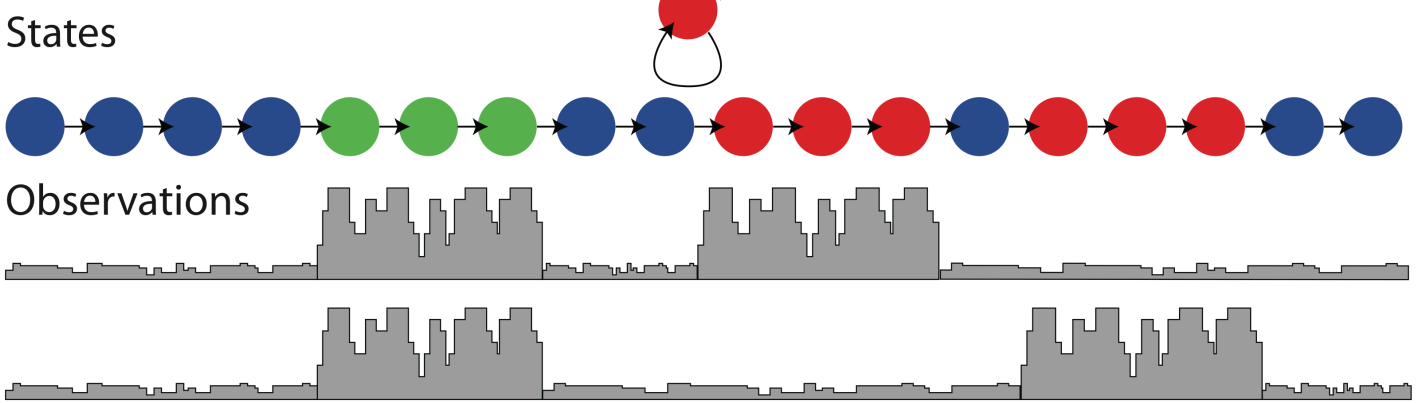


HMM for differential analysis

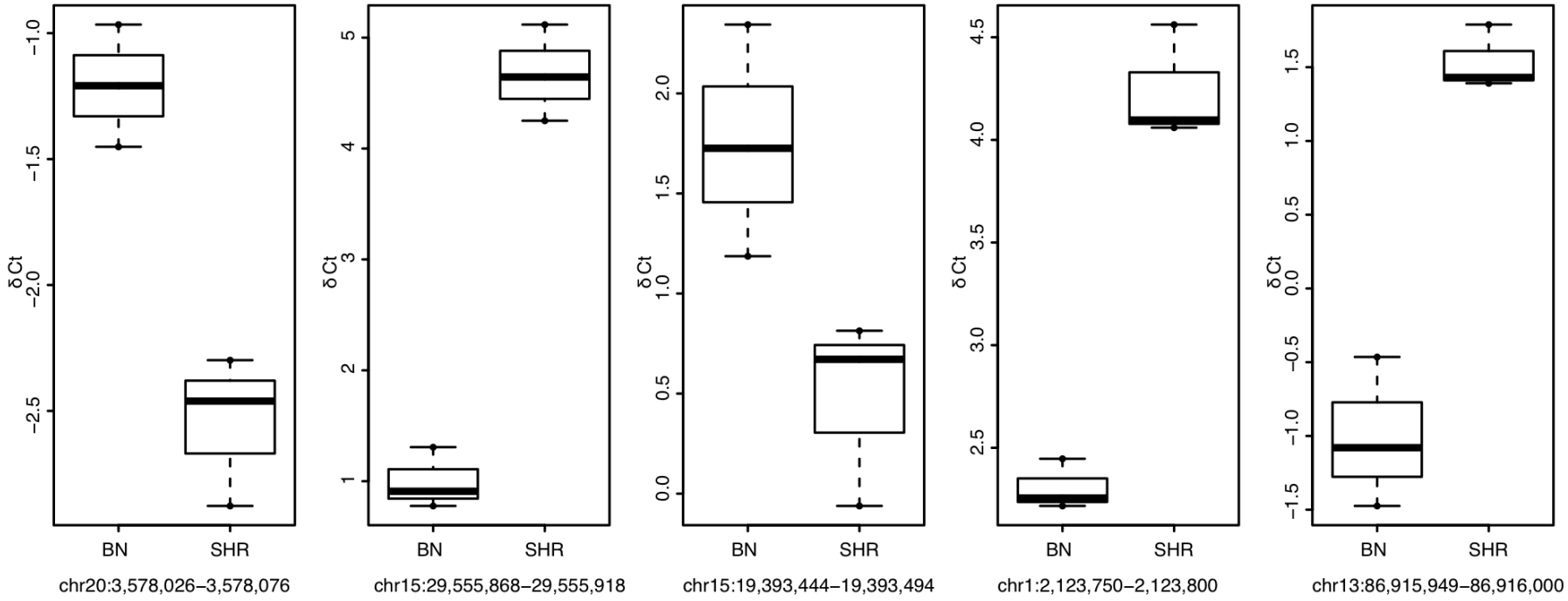
Hidden Markov model for multivariate count data







Evaluation with qPCR data



Evaluation with expression data