Utilizing promoter pair orientations for HMM-based analysis of ChIP-chip data

Michael Seifert, Jens Keilwagen, Marc Strickert, and Ivo Grosse

seifert@ipk-gatersleben.de

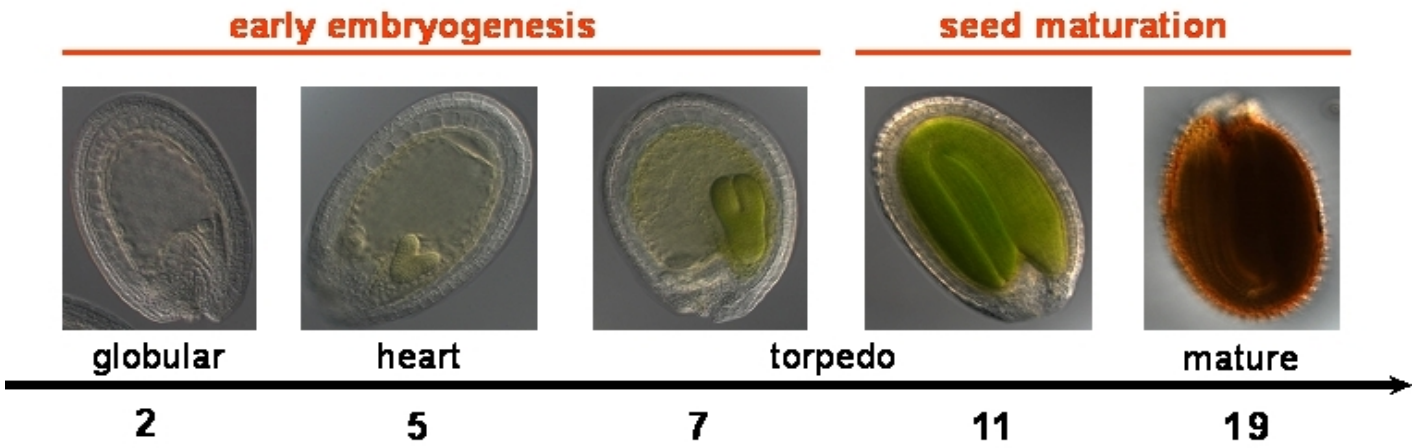
Arabidopsis Seed Development

Arabidopsis thaliana

- Model organism
- Genome sequenced
- Expression data collections

Seed Development

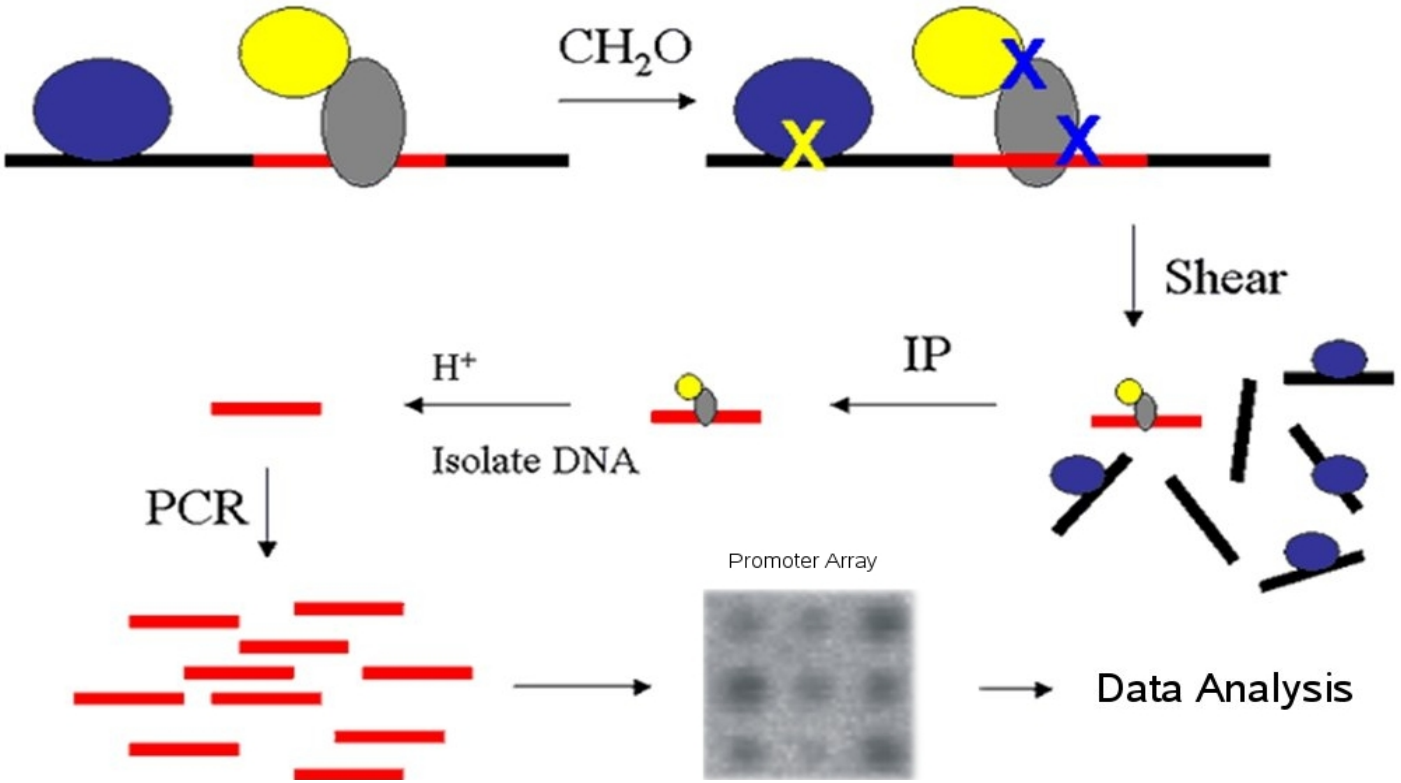
- Spatio-temporal process under control of transcription factors
- Key transcription factor: ABI3



Goal: Predict target genes of ABI3

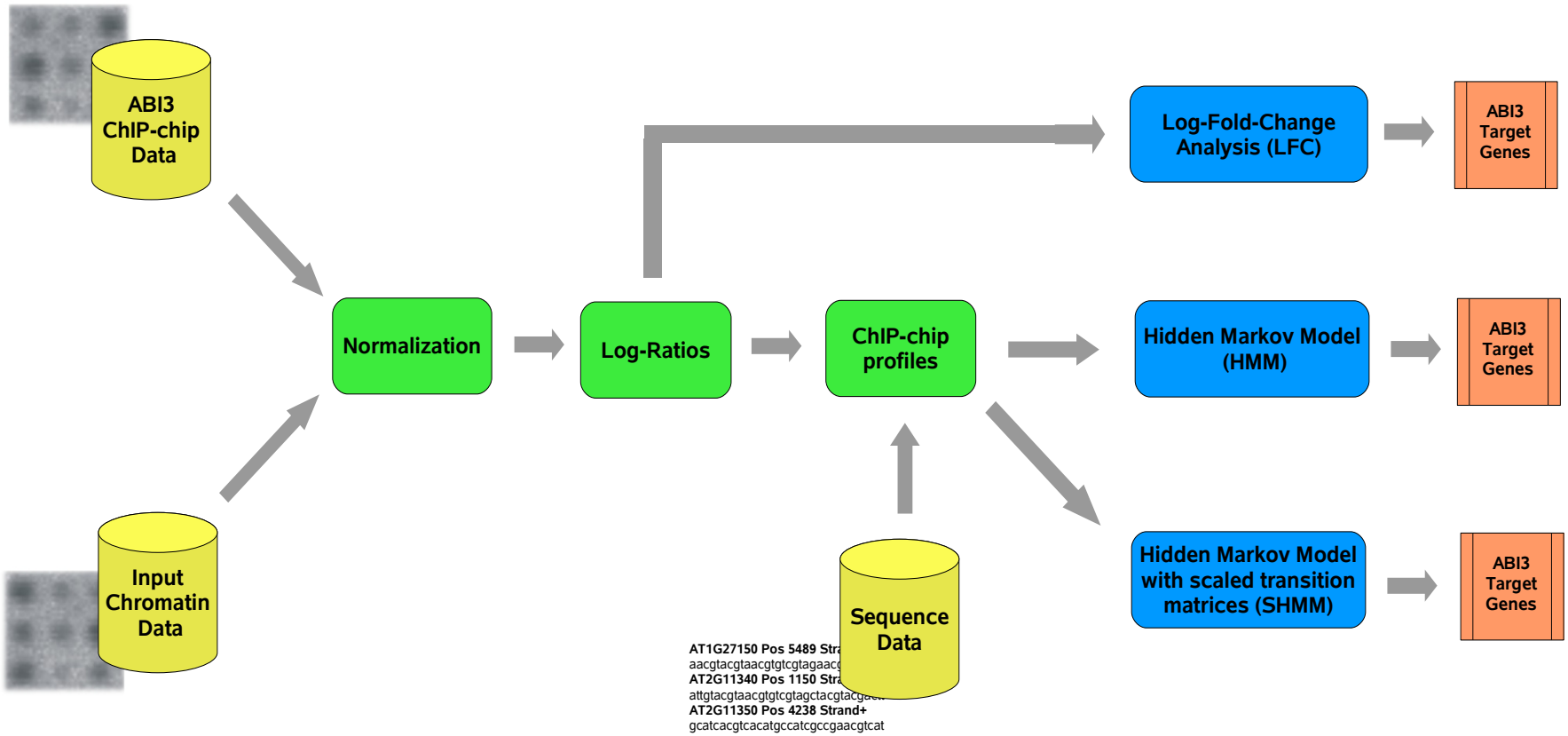
ChIP-chip

- Chromatin-immunoprecipidation (ChIP) coupled with hybridization to promoter arrays (chip) [Ren et al. (2000) Science Vol 290]



Target Gene Detection Pipeline

- Analysis of ChIP-chip data by three approaches



Log-Fold-Change Analysis (LFC)

Log-Ratios

- Log-Ratio of gene t: $o_t = \log_2 \left(\frac{\text{Intensity of promoter } t \text{ in ABI3 data}}{\text{Intensity of promoter } t \text{ in Chromatin Input}} \right)$
- Potential of a gene to be a target gene of ABI3

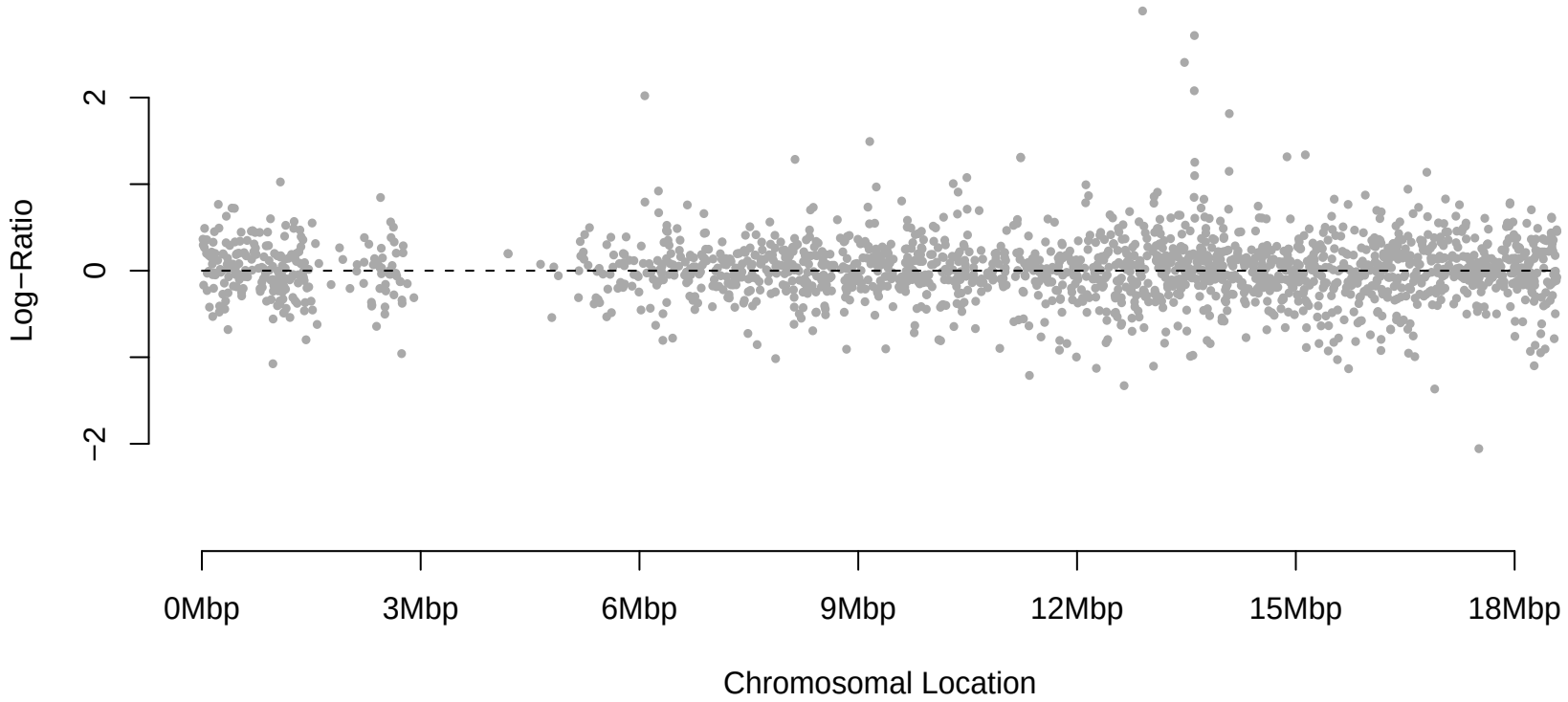
Log-Fold-Change Analysis (LFC)

Log-Fold-Change Analysis

- Sort the genes in each experiment by decreasing Log-Ratios
- Take the top k genes of each experiment and determine their intersection
- Genes in the intersection are putative ABI3 target genes

ChIP-chip Profile

- Log-Ratios in the context of chromosomal locations

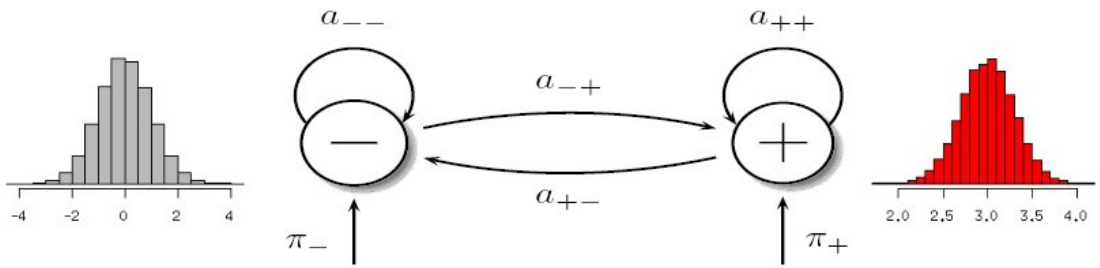


- Weak positive correlations between adjacent Log-Ratios

Hidden Markov Model (HMM)

ChIP-chip profiles modeled by HMM

- States $-$ (non-target) and $+$ (target) characterize Log-Ratios by specific Gaussian emission distributions
- Log-Ratios are processed from 'left to right' along chromosomes



Initialization

- Choose start-, transition-, and emission parameters

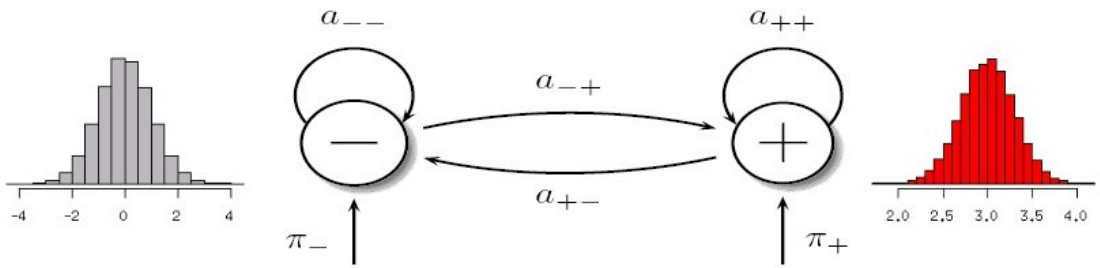
Training

- Extended Baum-Welch training using a prior to characterize the non-target and the target state

Hidden Markov Model (HMM)

ChIP-chip profiles modeled by HMM

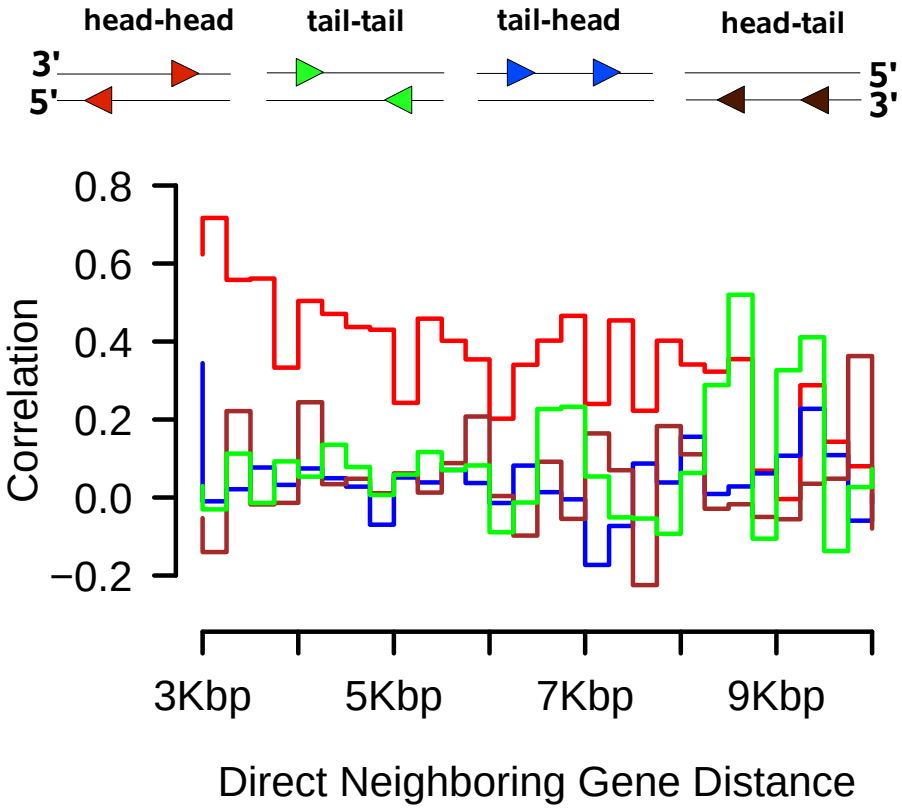
- States $-$ (non-target) and $+$ (target) characterize Log-Ratios by specific Gaussian emission distributions
- Log-Ratios are processed from 'left to right' along chromosomes



Target Gene Detection

- Sort the genes in each experiment by decreasing probability of being an ABI3 target gene
- Take the top k genes of each experiment and determine their intersection
- Genes in the intersection are putative ABI3 target genes

Analysis of Promoter Pair Orientations

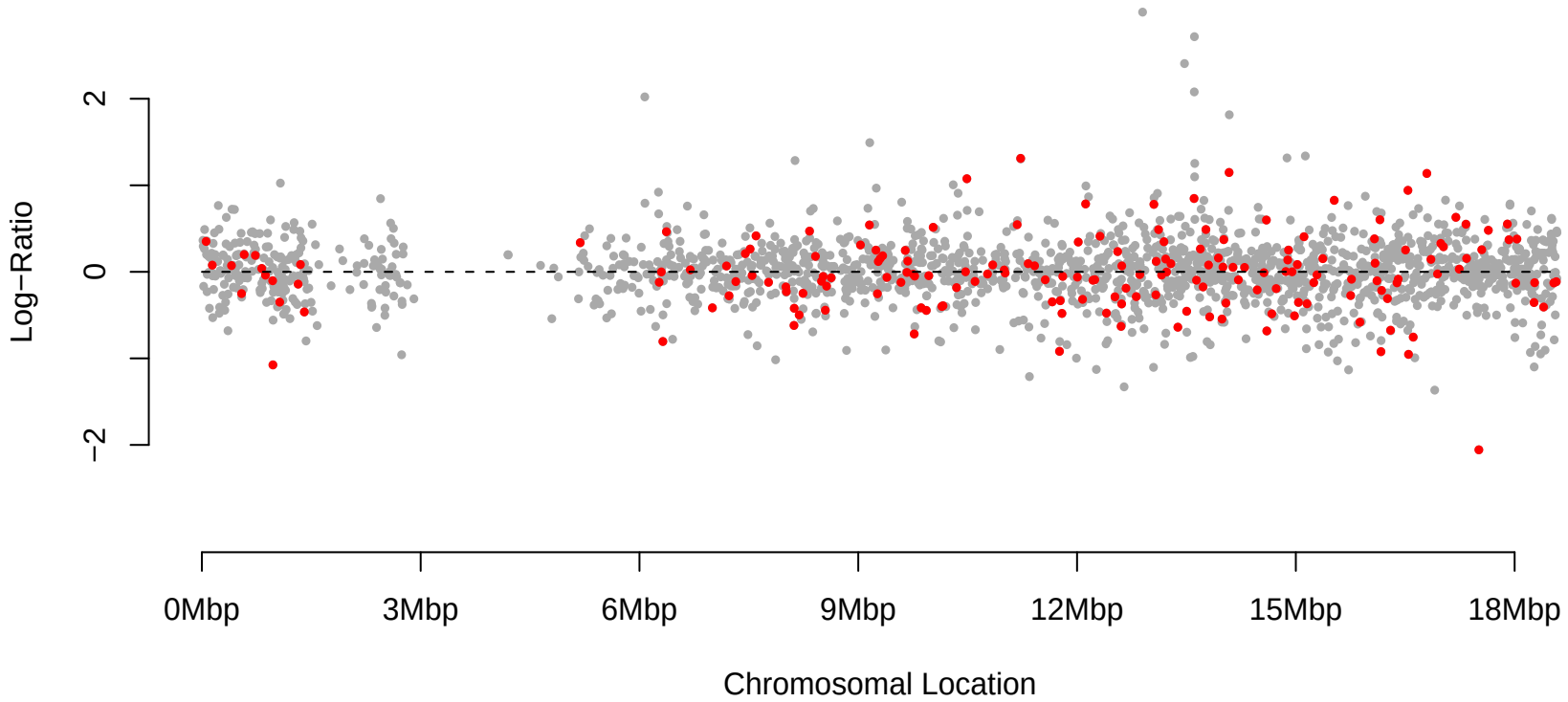


- Four gene pair orientations on DNA
- Triangle: Gene
- Tip of a triangle: Reading direction

$\Rightarrow$ Significantly higher correlations of Log-Ratios for promoter pairs in head-head orientation (limit 9Kbp)

ChIP-chip Profile

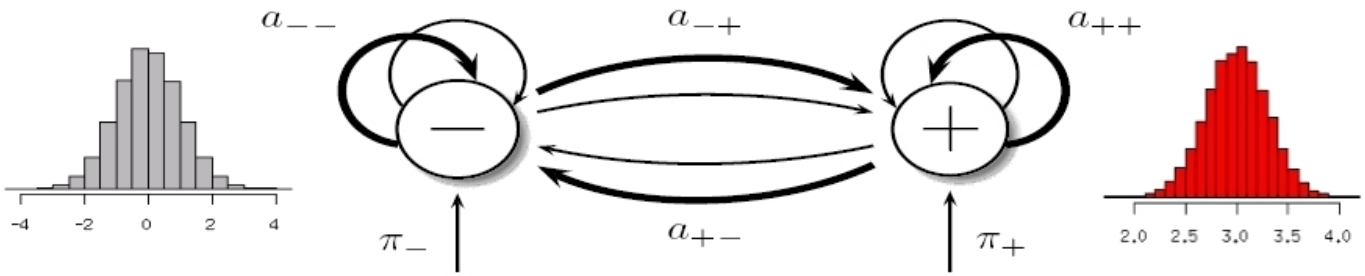
- Log-Ratios in the context of chromosomal locations differentiating head-head promoter pairs in distances $\leq$ 9Kbp from other promoter pair orientations



HMM with scaled transition matrices (SHMM)

Genes in head-head orientation

- More likely that both are either non-targets or targets of ABI3
- Increased self-transition probabilities: $i \in \{-,+\}, f_2 > 1 : a_{ii} \mapsto \frac{a_{ii}-1+f_2}{f_2}$
- Thick arrows for head-head orientations with gene distance $\leq$ 9Kbp, otherwise thin arrows



- Initialization, Training, and Target Gene Detection like for HMM

Comparison of LFC, HMM, and SHMM

Target Genes per Method

- Top k candidates of each experiment
- Intersection of Top k candidate lists

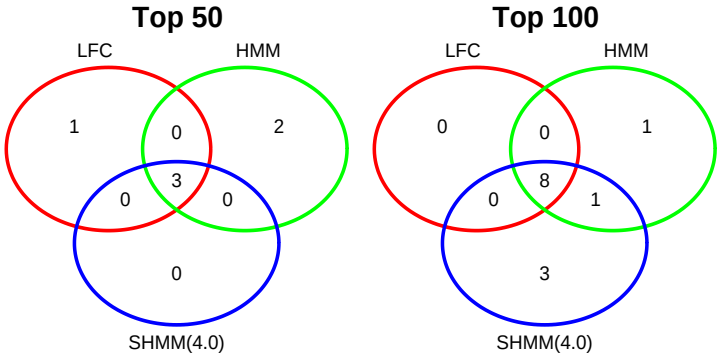
$\Rightarrow$ Target gene lists

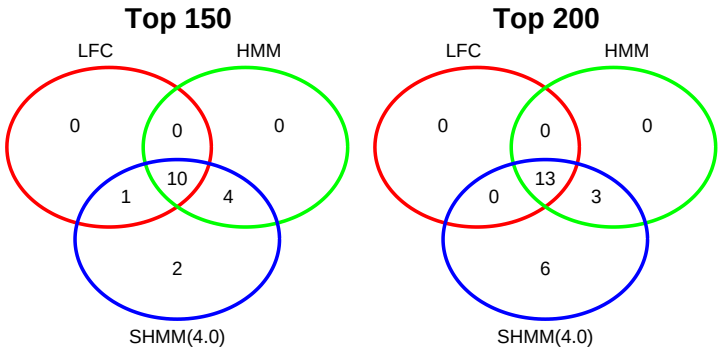
Comparison of Methods

- Venn diagrams of target gene lists
- $k \in \{50, 100, 150, 200\}$

Results

- No new target genes at $k = 300$
- SHMM tends to be more general
- Predicted target genes must be validated by independent tests





Validation of ABI3 Target Genes

Genevestigator (GV)

- Expression of ABI3: Inflorescence, Silique, and Seed
- Score $= \frac{\text{Infl.+Silique+Seed}}{\text{Expression in all Categories}}$
- Compare Score against Scores of 1,000 randomly selected genes
- Hints which genes might be activated

Transient Assay (TA)

- Target gene promoter in fusion with GUS reporter gene
- ABI3 vs. lack of ABI3
- Test reaction on ABI3

ID	LFC	HMM	SHMM	GV	TA
T1	✓	✓	✓	0.94	5
T2	✓	✓	✓	0.11	2.5
T3	✓	✓	✓	0.86	12
T6	✓	✓	✓	0.72	15
T7	✓	✓	✓	0.90	7
T12	✓	✓	✓	0.74	24
T13	✓	✓	✓	0.09	0.4
T14	✓	✓	✓	0.93	8
T16	✓	✓	✓	0.95	27
T17	✓	✓	✓	0.98	27
T19	✓	✓	✓	0.98	27
T20	✓	✓	✓	0.57	8
T22	✓	✓	✓	0.81	30
T11	×	✓	✓	0.09	2
T15	×	✓	✓	0.10	-
T18	×	✓	✓	0.98	27
T4	×	×	✓	0.03	-
T5	×	×	✓	0.39	3
T8	×	×	✓	0.46	12
T9	×	×	✓	0.07	1
T10	×	×	✓	0.95	6
T21	×	×	✓	0.20	0.6

Validation of ABI3 Target Genes

Results

- LFC and HMM missed 20% of target genes compared to SHMM
  - Significant GV scores
  - Activation in TAs
- SHMM over 40% more putative target genes than LFC
  - 9 of 22
  - 7 TAs
  - No reaction 1
  - Repression 1
  - Activation 5

ID	LFC	HMM	SHMM	GV	TA
T1	✓	✓	✓	0.94	5
T2	✓	✓	✓	0.11	2.5
T3	✓	✓	✓	0.86	12
T6	✓	✓	✓	0.72	15
T7	✓	✓	✓	0.90	7
T12	✓	✓	✓	0.74	24
T13	✓	✓	✓	0.09	0.4
T14	✓	✓	✓	0.93	8
T16	✓	✓	✓	0.95	27
T17	✓	✓	✓	0.98	27
T19	✓	✓	✓	0.98	27
T20	✓	✓	✓	0.57	8
T22	✓	✓	✓	0.81	30
T11	×	✓	✓	0.09	2
T15	×	✓	✓	0.10	-
T18	×	✓	✓	0.98	27
T4	×	×	✓	0.03	-
T5	×	×	✓	0.39	3
T8	×	×	✓	0.46	12
T9	×	×	✓	0.07	1
T10	×	×	✓	0.95	6
T21	×	×	✓	0.20	0.6

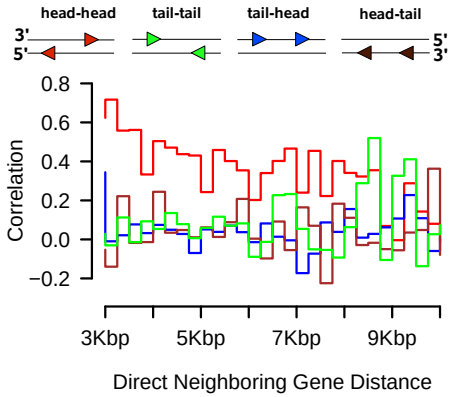
Conclusions

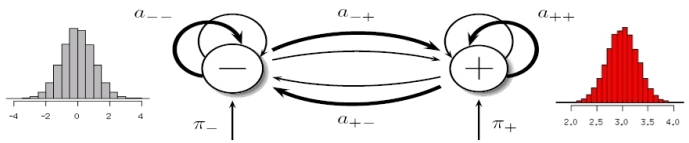
Summary

- Comparison of 3 methods for the detection of ABI3 target genes from ChIP-chip promoter array data
  1. LFC: Log-Ratios
  2. HMM: Log-Ratios in the context of chromosomal locations
  3. SHMM: Log-Ratios in the context of chromosomal locations differentiating head-head orientations from others

Main Result

- SHMM predicted the highest number of target genes validated by Genevestigator and transient assays





Acknowledgment and Software

Basics of SHMMs

- Alexander Schliep (MPI Berlin), and Stefan Posch (MLU Halle)

ArabidoSeed Project

- Groups of Lothar Altschmied, Helmut Bäumlein, and Udo Conrad (IPK Gatersleben)
- Urs Hähnel, and Gudrun Mönke
- http://arabidoseed.ipk-gatersleben.de

Financial Support

- BMBF grants 0312706A and 0313155
- Ministry of culture Saxony-Anhalt grant XP3624HP/0606T

Software

- http://dig.ipk-gatersleben.de/SHMMs/ChIPchip/ChIPchip.html