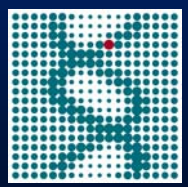
Analysis of Gene Expression Trees in Blood Cell Development

Ivan G. Costa Filho
Stefan Roepcke
Alexander Schliep

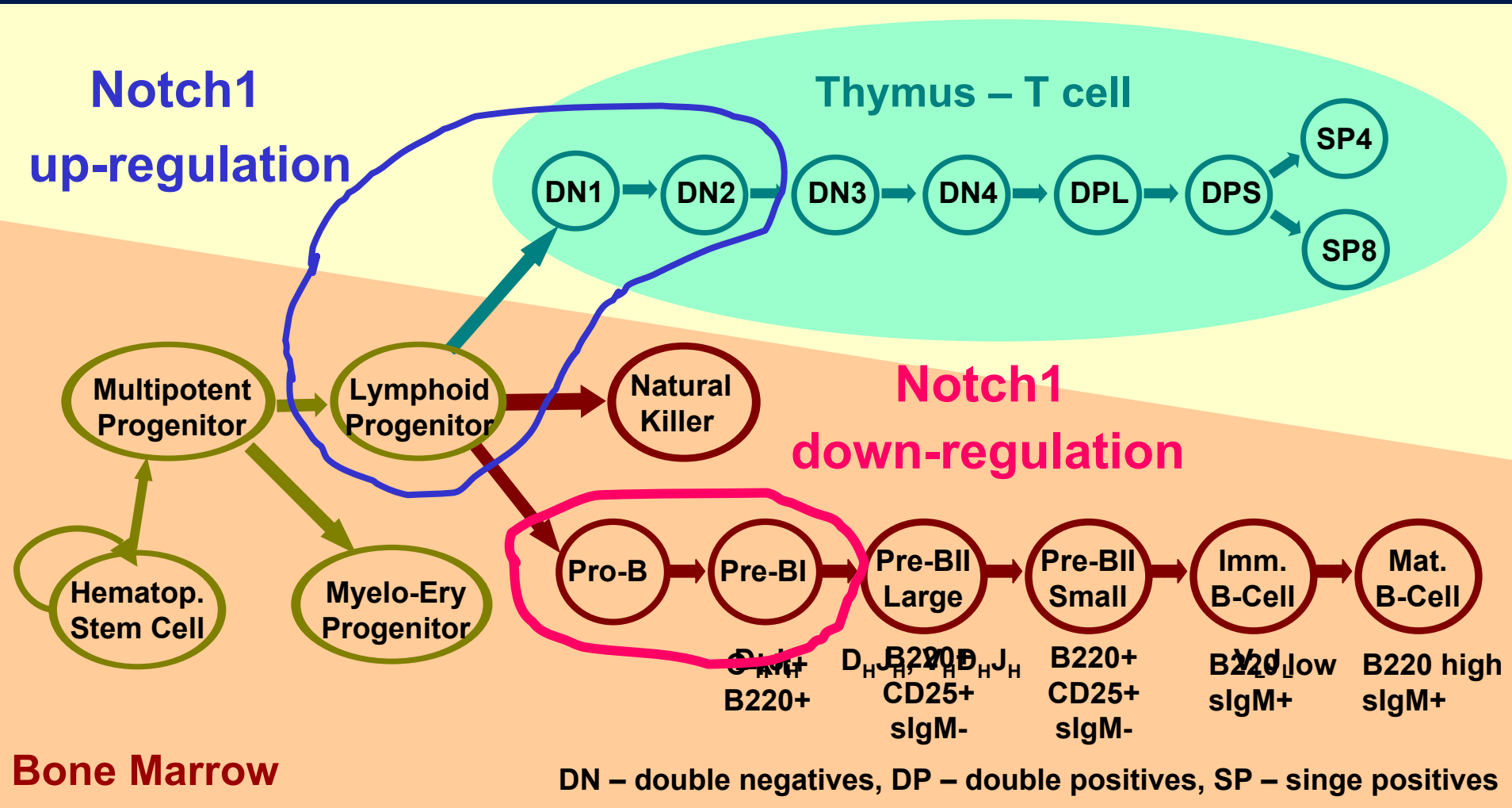


Computational Biology Department
Max Planck Institute for Molecular Genetics, Berlin

Blood Cell Development

- Motivation
  - understand development system
  - clinical interest: immune cells, leukemia
- Technical aspects
  - pure cell samples easily accessible
  - broadly studied developmental system
  - no computational framework available

Blood Cell Development



Goal

Understand gene regulation
during development of blood cells

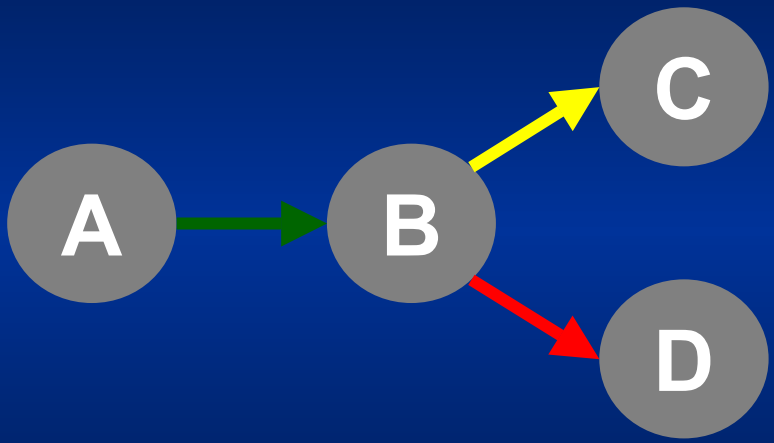
Method Outline

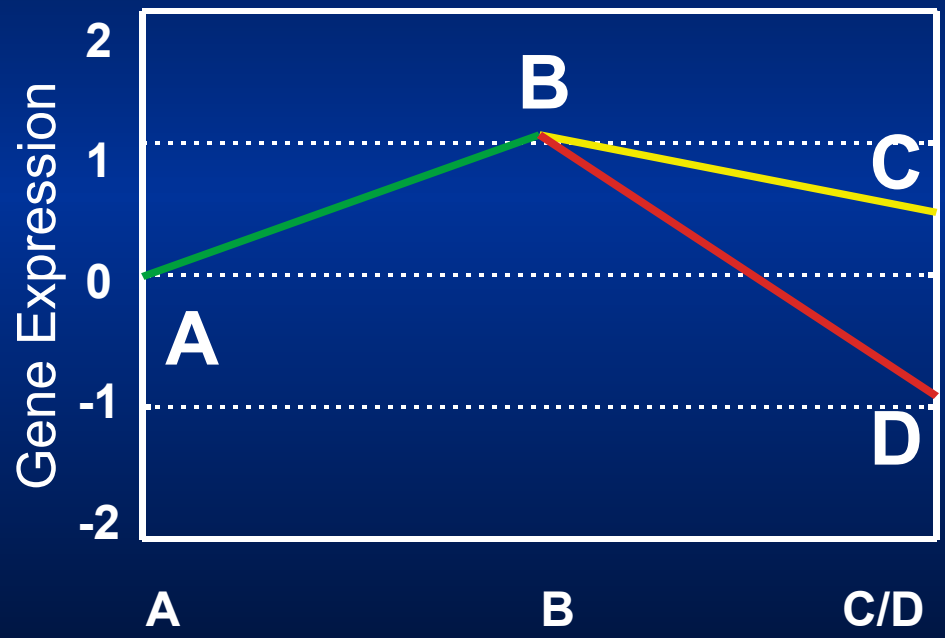
Use a statistical model for

1. finding similar ‘development profiles’
   - tree model
2. clustering ‘development profiles’
   - combine tree models in a mixture
3. interesting regulatory patterns
   - enrichment of microRNA targets

(1) Method
Tree Model

Tree Model Development Profile

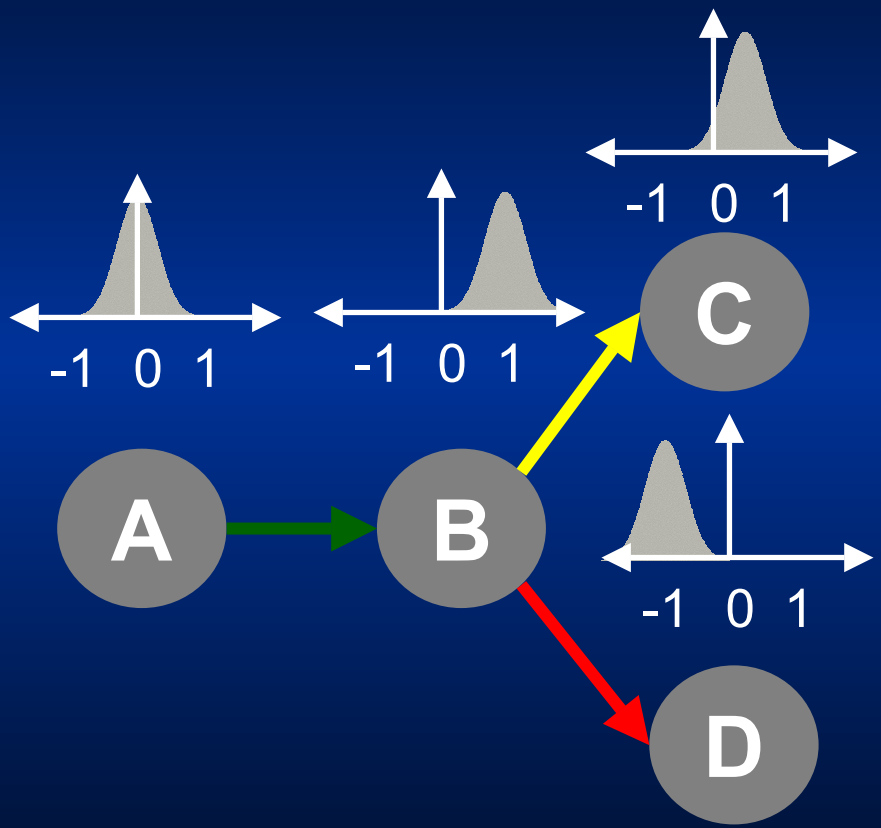


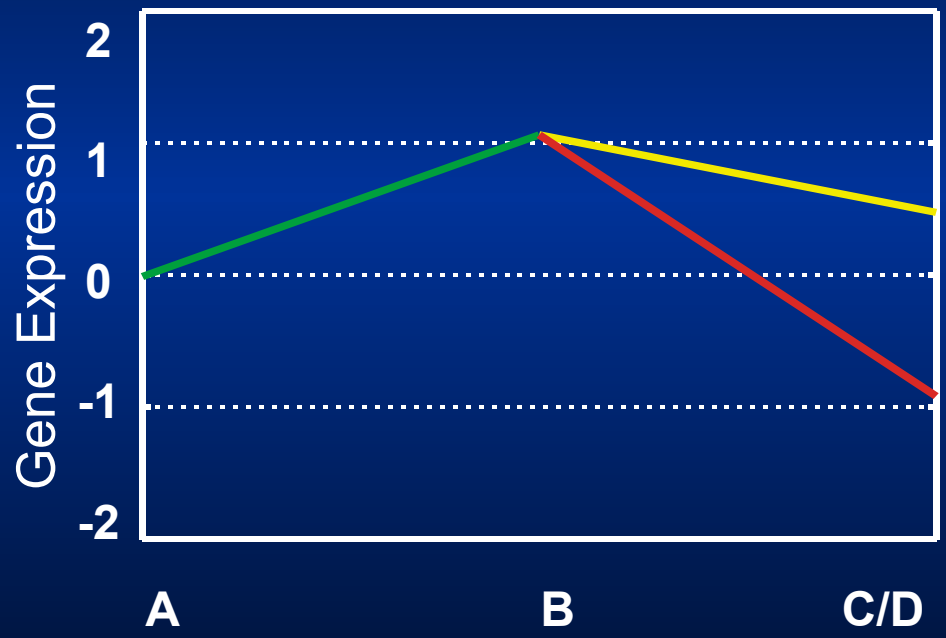


development profile:

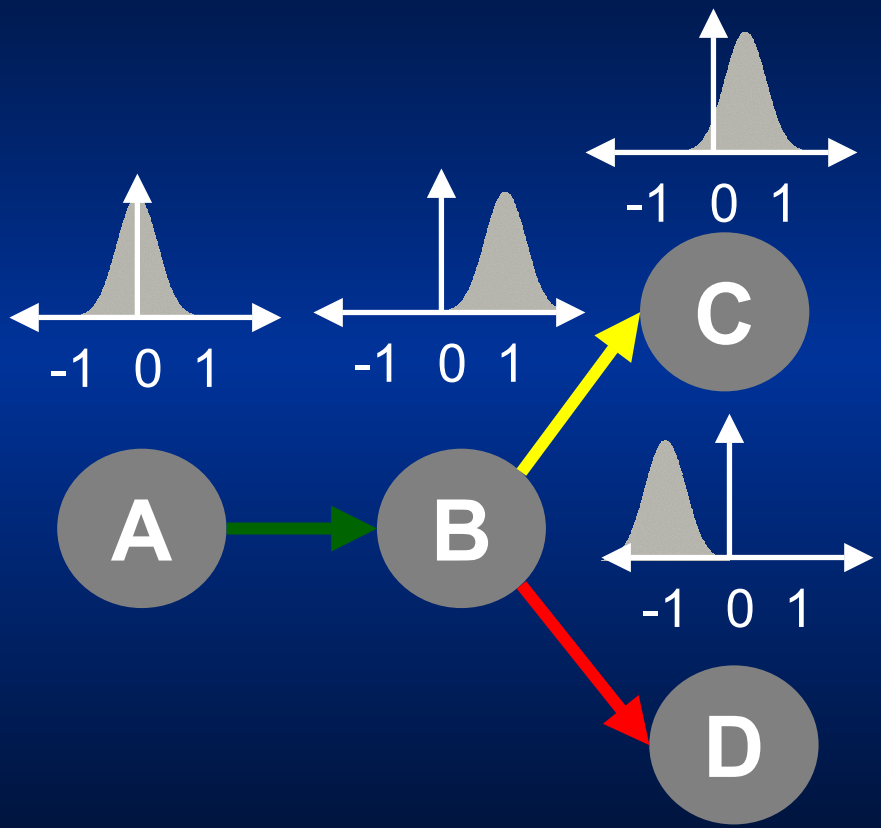
	A	B	C	D
Notch1	0	1	0.5	-1

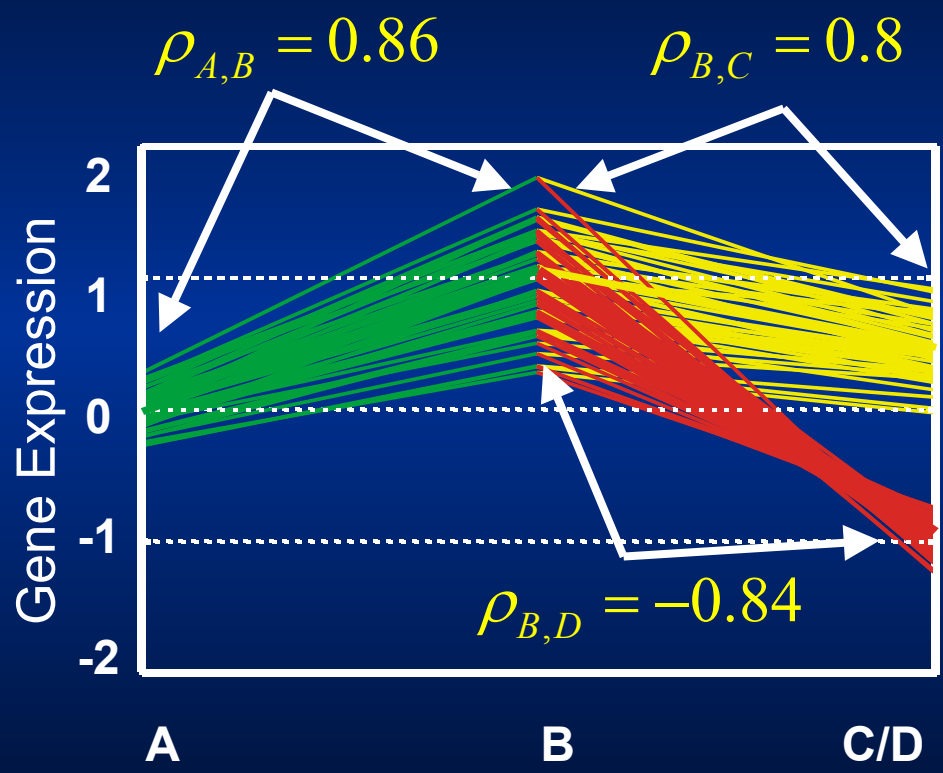
Tree Model





Tree Model

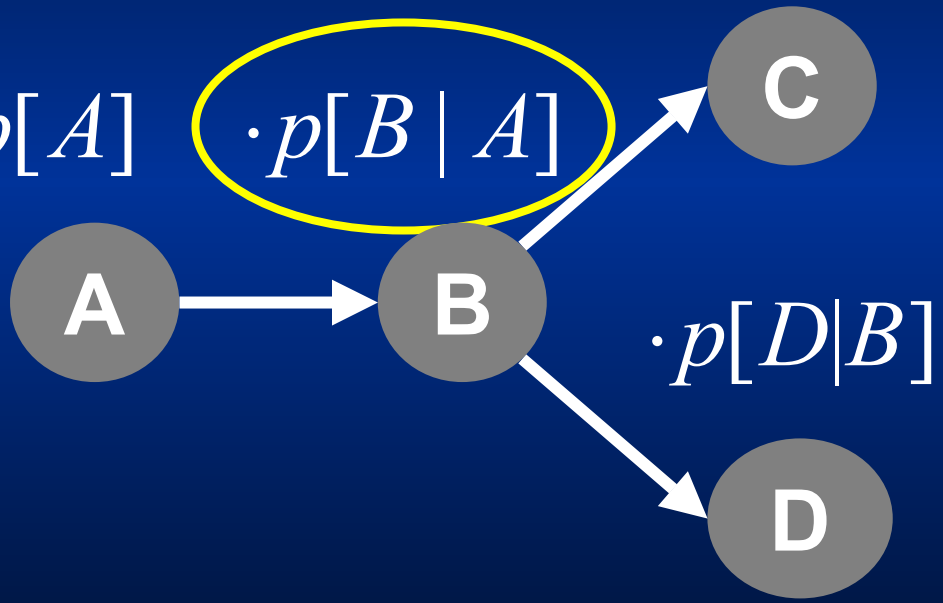




$\rho_{A,B}$ correlation of *A* and *B*

Tree Model Assumption

$$p[A,B,C,D] = p[A] \cdot p[B \mid A] \cdot p[C|B] \cdot p[D|B]$$



Tree Model
Conditional Gaussian

Conditional Gaussian pdf for $p[b|\mathrm{a}]$,

$$p[b \mid a, \theta] = \frac{1}{\sqrt{2\pi\sigma^2_{B|A}}} \exp\left(\frac{-(b - \mu_B - w_{B|A}(a - \mu_A))^2}{2\sigma^2_{B|A}} \right),$$

Maximum likelihood estimates:

$$\mu_A$$

$$w_{B|A} = \sigma_{AB} / \sigma^2_A$$

$$\sigma^2_{B|A} = \sigma^2_B - w^2_{B|A}\sigma^2_A$$

(2) Method
Mixtures of Trees

Perspective - Clustering

- ‘Classical’ methods assume independence between variables.
- ‘Complex’ models (ie. multivariate gaussians) over fit parameters.
- Methods for time-courses consider temporal dependencies, but not trees dependencies
- Mixtures of trees uses prior knowledge with the requirement of few additional parameters

Clustering Method

1. Combine *K* tree models in a mixture model

$$p[A,B,C,D \mid \alpha_1,...,\alpha_K, \text{tree}_1,...,\text{tree}_K]$$

$$= \sum_{j=1}^{K} \alpha_k \cdot p[A,B,C,D \mid \text{tree}_k]$$

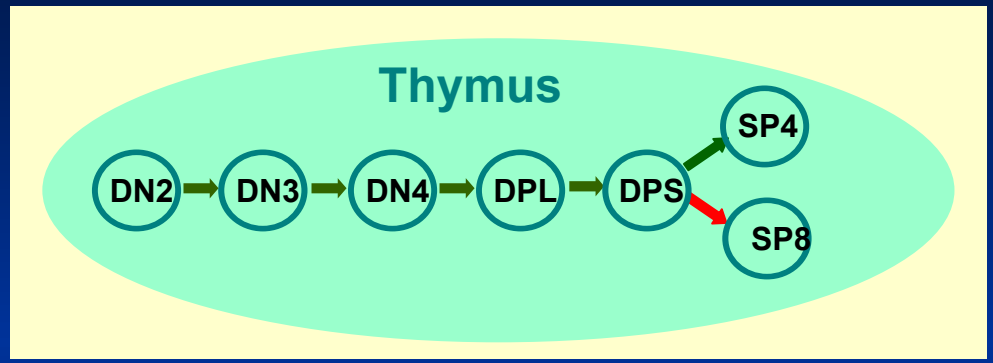
where all trees share the same topology

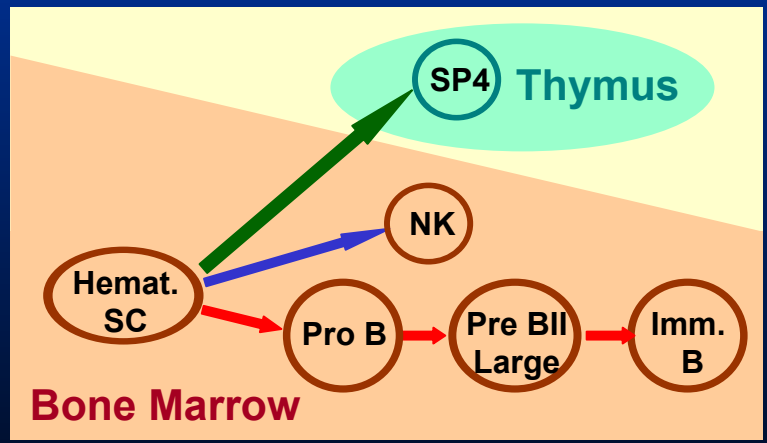
2. Estimate the mixture using EM algorithm

Results

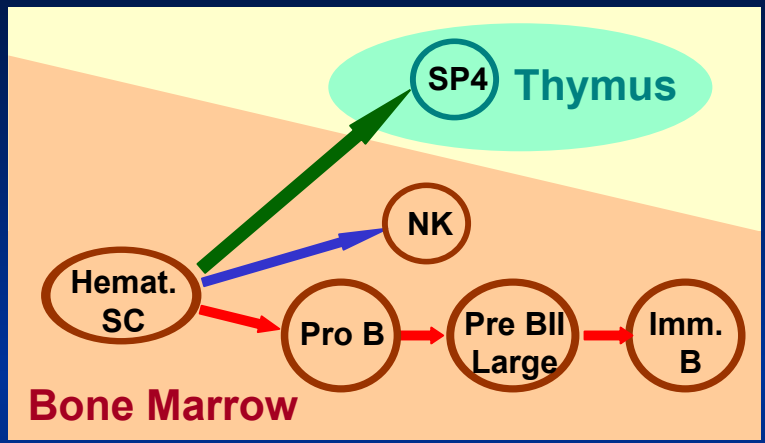
Data

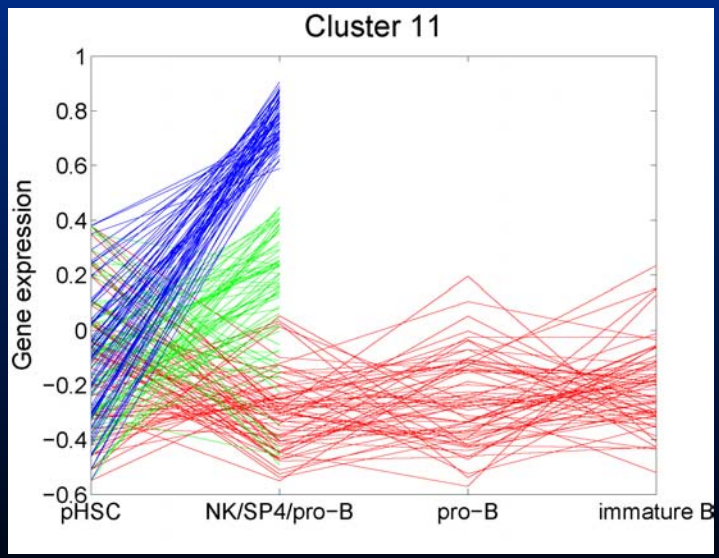
- T Cell [Hoffman at *el.*]
  - 7 detailed stages of mouse t-cell development
  - 1318 genes after filtering
- Lymphoid Tree
  - 6 stages of mouse lymphoid lineage
  - from three studies [Bystrykh at *el.*, Poirot at *el.*, Tze at *el.*]
  - 1321 genes after filtering

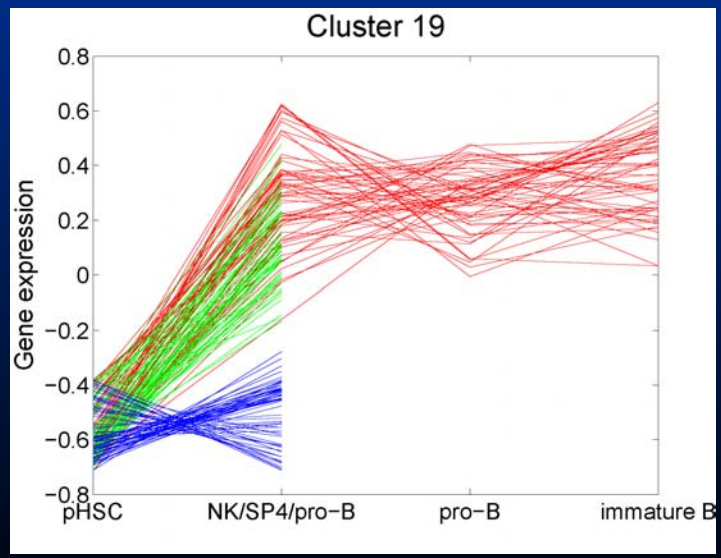




Results – Lymphoid Tree





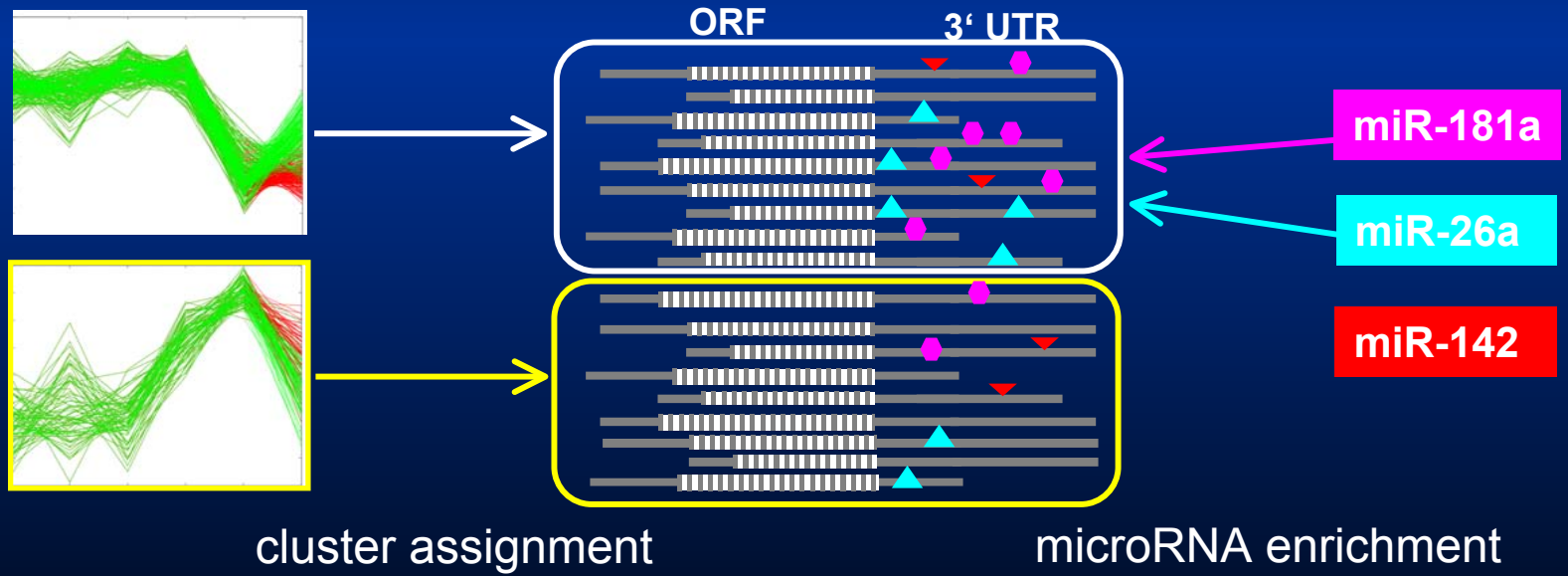


(3) Method
microRNA Target Detection

MicroRNA Target Detection

MicroRNA have post-transcriptional roles by repression of target genes

- important in hematopoiesis [Chen *et al.* 2004]
- degradation of targets transcripts [Lin, 2005, Huang, 2006]

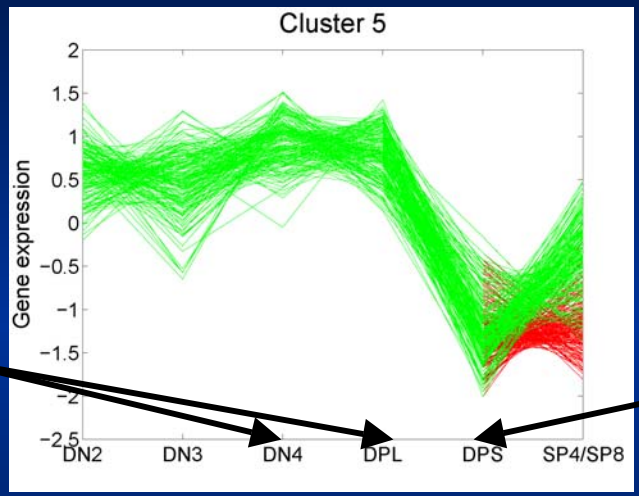


Sequence based target detection with mirRanda [Enright, 2003]

Results - microRNA Targets

Receptor rearrangements

proliferating cells



resting cells

microRNA	microRNA Targets
miR-15a, miR-181a, miR-221, miR-24, miR-26a	2410015N17Rik, Alad, Atpif1, Aurkb, Cdc25a, Chek1, Cks1b, Cks2, Eed, H2afx, Kpnb1, Mcm5, Nasp, Pex7, Psmd12, Ranbp5, Rars, Tk1, Trip13, Uchl5

Results - microRNA Targets

Some Numbers

Enrichment was found in:

- **5 out of 20 clusters**
- **11 out of 17 microRNAs**
- **39 out of 229 predicted targets**

Receptor r… prolife… …g cells

microRNA	microRNA Targets
miR-15a, miR-181a, miR-221, miR-24, miR-26a	2410015N17Rik, Alad, Atpif1, Aurkb, Cdc25a, Chek1, Cks1b, Cks2, Eed, H2afx, Kpnb1, Mcm5, Nasp, Pex7, Psmd12, Ranbp5, Rars, Tk1, Trip13, Uchl5

Summary

- Novel framework for analysis of development
  - querying, visualization, clustering, …
- Recovery of well known biological facts
- Discovery of putative regulatory elements
  - concise lists of microRNA targets and insights on microRNA function

Outlook

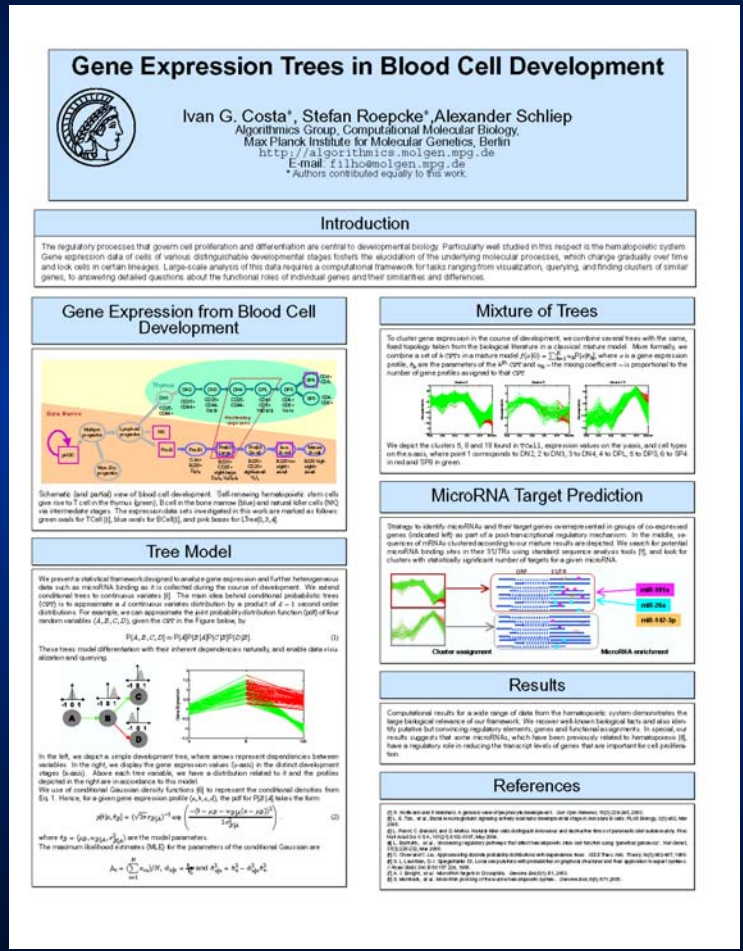
- Improve microRNA target detection
  - integrate microRNA expression in the target prediction framework [Huang *et al*, 2006]
- Structure learning of tree topologies
- Analysis of a more detailed development tree (with Fritz Melchers at MPI-Infection Biology)
  - microRNA targets validation

Acknowledgments

- *Christopher Hafemeister* - implementations
- *Fritz Melchers* (MPI for Infectious biology) for helpful discussions and encouragement.
- CNPq (Brasil) and DAAD (Germany) for funding.

Thanks.

Poster L32 - Today



Gene Expression Trees in Blood Cell Development

Ivan G. Costa*, Stefan Roepcke*, Alexander Schliep
Algorithmics Group, Computational Molecular Biology,
Max Planck Institute for Molecular Genetics, Berlin
http://algorithmics.molgen.mpg.de
E-mail: filho@molgen.mpg.de
*Authors contributed equally to this work.

Introduction

The regulatory processes that govern cell proliferation and differentiation are central to developmental biology. Particularly well studied in this respect is the hematopoietic system. Gene expression data of cells of various distinguishable developmental stages fosters the elucidation of the underlying molecular processes, which change gradually over time and lock cells in certain lineages. Large-scale analysis of this data requires a computational framework for tasks ranging from visualization, querying, and finding clusters of similar genes, to answering detailed questions about the functional roles of individual genes and their similarities and differences.

Gene Expression from Blood Cell Development

Tree Model

Mixture of Trees

MicroRNA Target Prediction

Results

References

Max Planck Molecular Genetics: **Algorithmics Group**

http://algorithmics.molgen.mpg.de

Conditional Prob. Trees Definition (1)

The prob. density function of a CPT is

$$p[X \mid \theta] = \prod_{u}^{D} p[X_u \mid X_{\mathrm{pa}(u)}, \theta]$$

where $X = \{X_1, ..., X_j, ..., X_D\}$

X_j takes expression values of stage j

$\mathrm{pa}(u) = v$ for $1 < v < u \leq D$

θ are the tree parameters

Prob. Conditional Trees Definitions (2)

$$\mathrm{P}[x_u \mid x_v, \theta] = \frac{1}{Z} \exp\left(\frac{-(x_u - \mu_u - w_{u|v}(x_v - \mu_v))^2}{2\sigma^2_{u|v}} \right)$$

- MLE estimates:

$$\mu_u = \bar{x}_u$$

$$w_{u|v} = \mathrm{cov}(x_u, x_v) / \mathrm{var}(x_v)$$

$$\sigma^2_{u|v} = \mathrm{var}(x_v) - w^2_{u|v} \mathrm{var}(x_u)$$

- MAP estimates:

$$w_{u|v} : \ Normal(0, N\beta_{u|v} \mathrm{var}(x_v)^{-1})$$

$$w^*_{u|v} = \mathrm{cov}(x_u, x_v) / (\mathrm{var}(x_v) + \beta^{-1}_{u|v})$$

- With empirical Bayes

$$\beta_{u|v} = N \Big/ \left(\frac{\mathrm{var}(x_u)\,\mathrm{var}(x_v)}{\mathrm{cov}(x_u, x_v)^2} - 1 \right)$$

Mixture Model Estimation

We want to maximize:

$$\mathrm{P}[X \mid \Theta] = \prod_{i=1}^{N} \sum_{j=1}^{K} \alpha_j \cdot \mathrm{P}[x_i \mid \theta_j]$$

By adding a hidden variable ***Y*** , we obtain:

$$\mathrm{P}[X, Y \mid \Theta] = \mathrm{P}[X \mid Y, \Theta]\ \mathrm{P}[Y \mid \Theta]$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{K} (\alpha_j \cdot \mathrm{P}[x_i \mid \theta_j])^{r_{ij}}$$

where $Y = \{y_i\}_{i=1}^{N}$ and $y_i \in \{1, \ldots, K\}$

$$r_{ij} = \mathrm{P}[y_i = j \mid x_i, \theta_j]$$

EM for Mixture Estimation

- Input:
  - genes profiles $\boldsymbol{X}$ and the number of clusters k
- Initialization:
  - Randomly assign the posterior probabilities $P[y_i = j | \boldsymbol{x}_i, \theta_j^0]$
- Iterate (until convergence):
  - Re-estimate Θ^t given $P[y_i = j | \boldsymbol{x}_i, \theta_j^{t-1}]$ **(M Estep)**
  - Calculate $P[y_i = j | \boldsymbol{x}_i, \theta_j^t]$ given Θ^t **(E Estep)**

From Mixtures to Groups

1. Maximum posterior assignment:

$$y_i = \arg\max_{1 \le j \le K} (\mathrm{P}[y_i = j \mid x_i, \theta_j])$$

2. Entropy cut-off assignment:

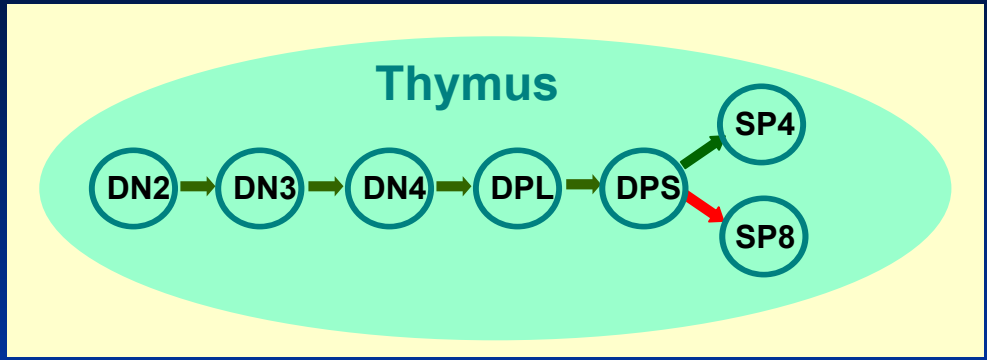
- if $\mathrm{H}(\{\mathrm{P}[y_i = j \mid x_i, \theta_j]\}_{j=1}^{K}) < \varepsilon$

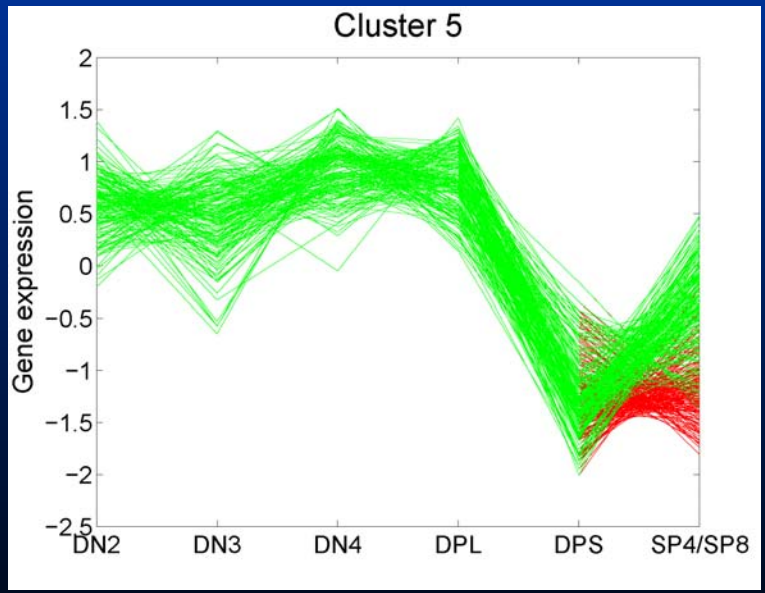
$$y_i = \arg\max_{1 \le j \le K} (\mathrm{P}[y_i = j \mid x_i, \theta_j])$$

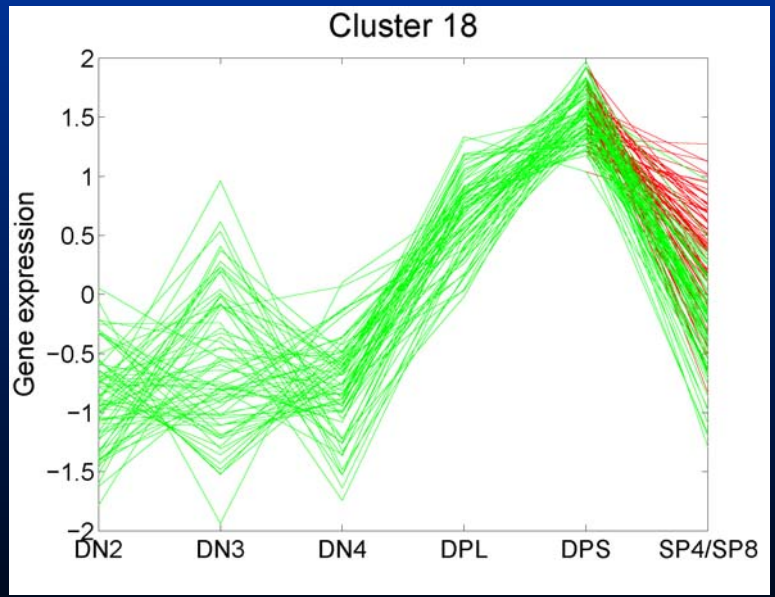
- else

$$y_i = K + 1$$

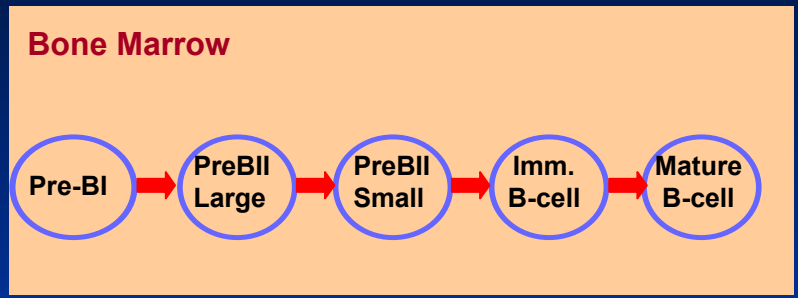
Results – T Cell

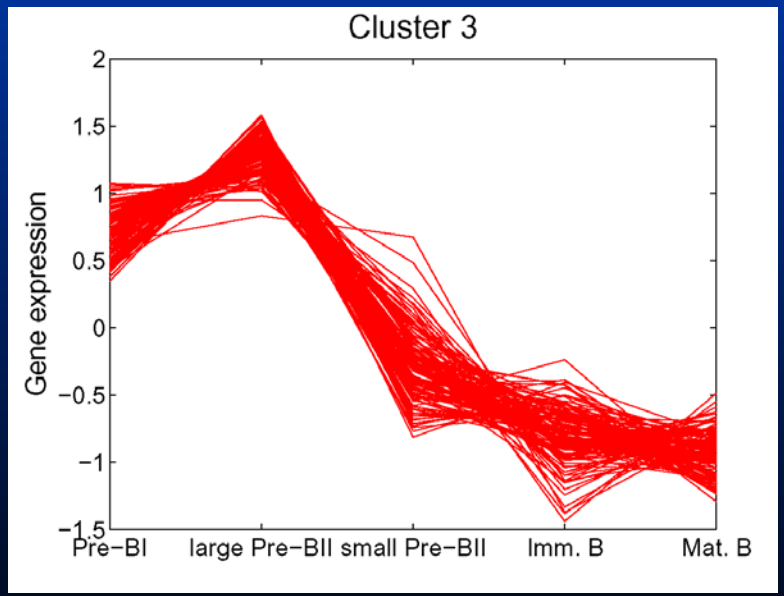


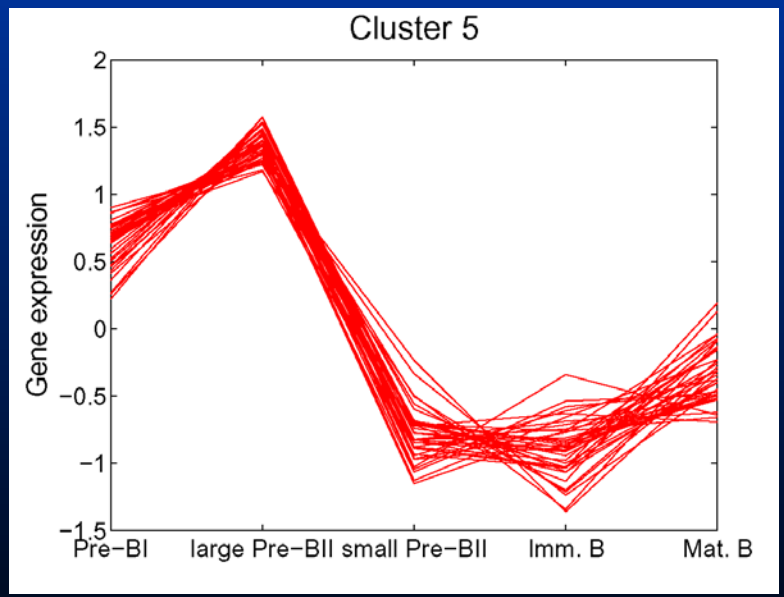




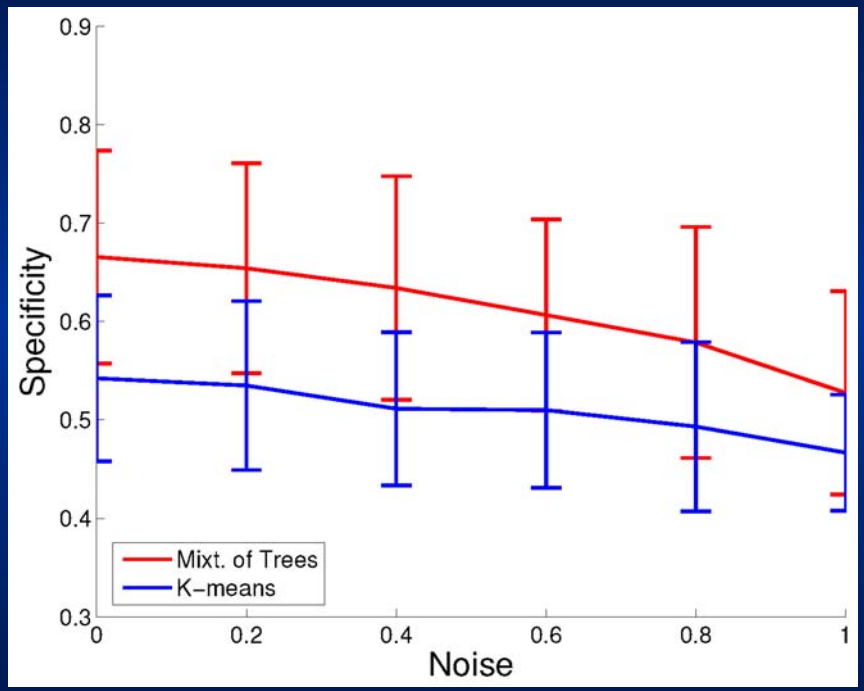
Results – Bcell

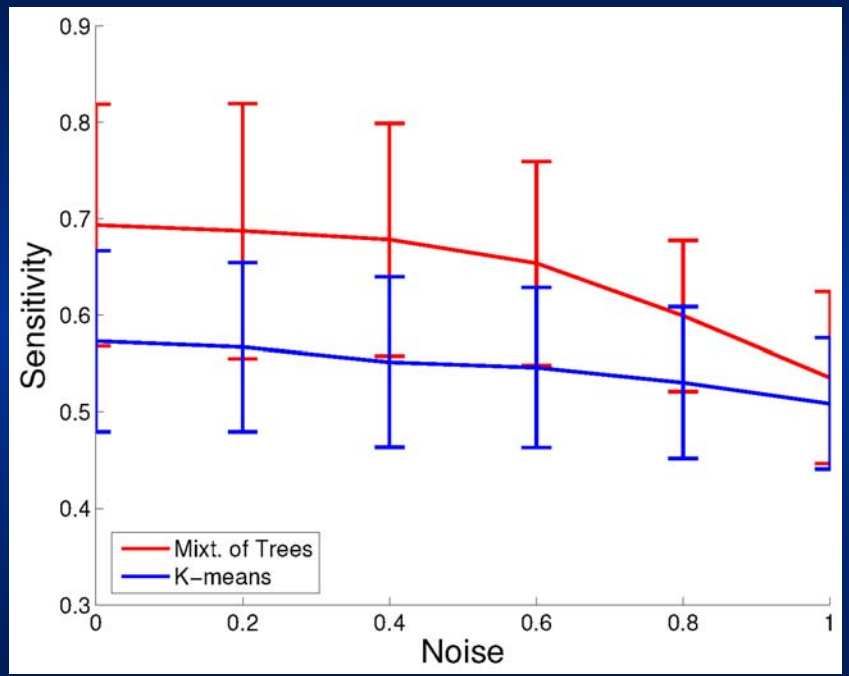






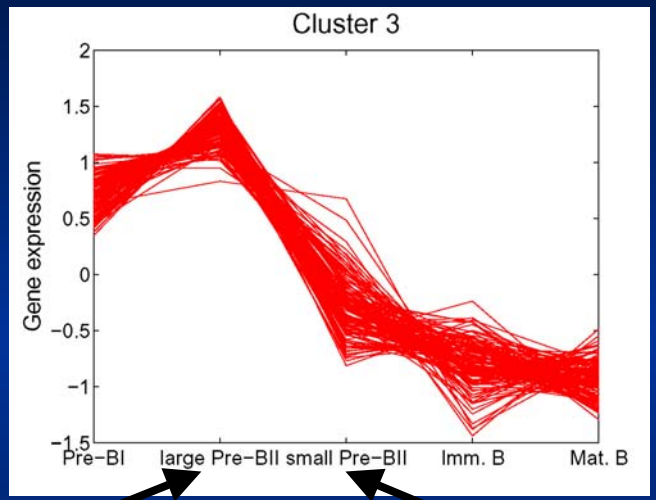
Results - Simulated Data

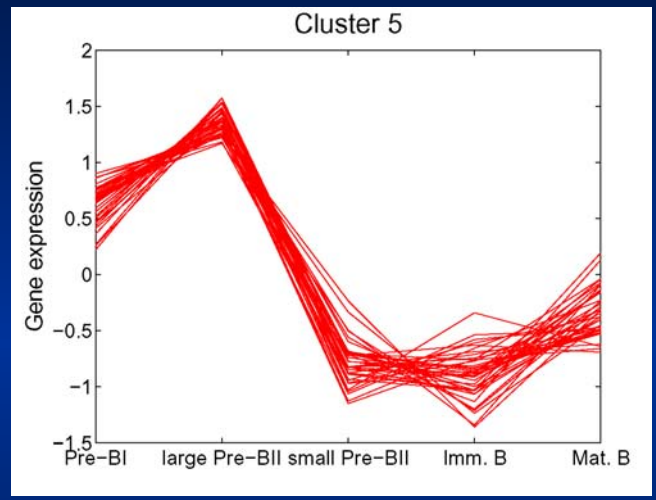




- Simulated Data
  - –random parameters for a given mixture of trees
  - –addition of independent noise

Results - microRNA Targets



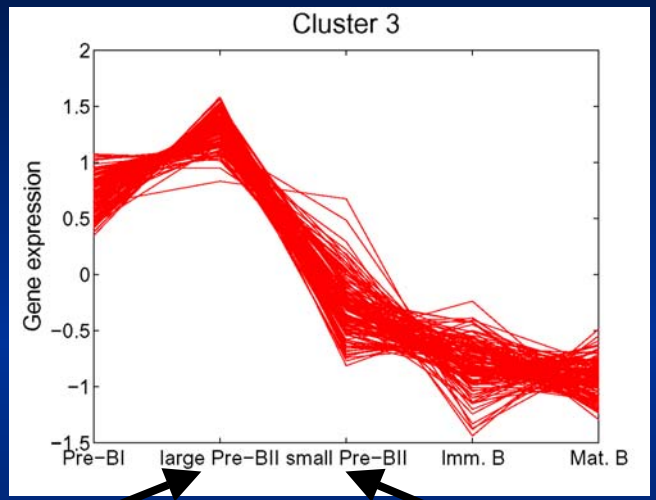


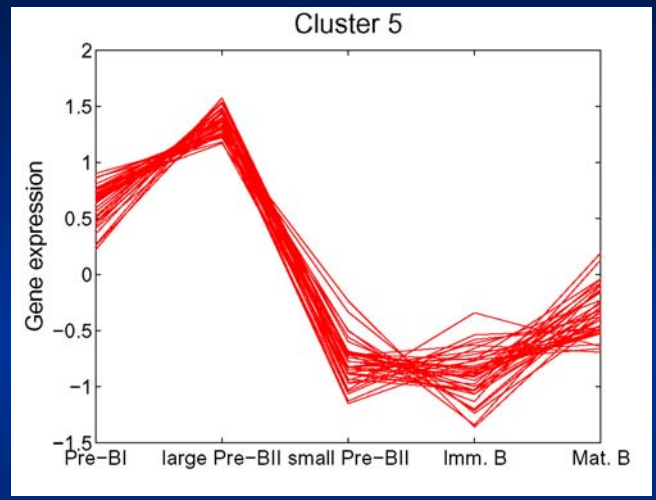
proliferating cells

resting cells

microRNA	microRNA Targets
miR-181b, miR-181c miR-26a	Atpif1, Aurkb, Cbx1, Cdc45l, Cks1b, Cks2, Cox5a, Hmgb2, Melk, Ttk, Uchl5
miR-15a, miR15b, miR-221, miR-223	Cdca4, Chek1, Mcm4, Nasp, Nfyb, Smc4l1, Tuba2[4]

Results - microRNA Targets





proliferating cells

resting cells

microRNA	microRNA Targets
miR-181b, miR-181c miR-26a	Atpif1, Aurkb, Cbx1, Cdc45l, Cks1b, Cks2, Cox5a, Hmgb2, Melk, Ttk, Uchl5
miR-15a, miR15b, miR-221, miR-223	Cdca4, Chek1, Mcm4, Nasp, Nfyb, Smc4l1, Tuba2