Models and Methods for Single Cell Data Analysis

Shahin Mohammadi, Vikram Ravindra, David Gleich, Ananth Grama

Department of Computer Science
Purdue University

Sept 25, 2019

PURDUE UNIVERSITY

- Recent advances in single cell technologies enable us to probe dynamic states of individual cells.
- Single cell technologies are also redefining basic understanding of cell types, tissue organization, pathology, and response.
- Single cell technologies result in datasets, models, and information that are orders of magnitude larger than conventional genomic/ transcriptomic/ interactomic repositories.

PURDUE UNIVERSITY

Introduction: Some Basic Terminology

- DNA is the basic code that governs living systems.
- DNA is transcribed into RNA. This process of transcription is controlled by a number of transcriptional control mechanisms (Transcription Regulation, Post Transcriptional Regulation).
- RNA is translated into proteins – the workhorses of living cells. The process of translation is controlled by a number of control mechanisms (Translational Controls).
- The activity of proteins is controlled by various post translational modifications (phosphorylation, methylation).

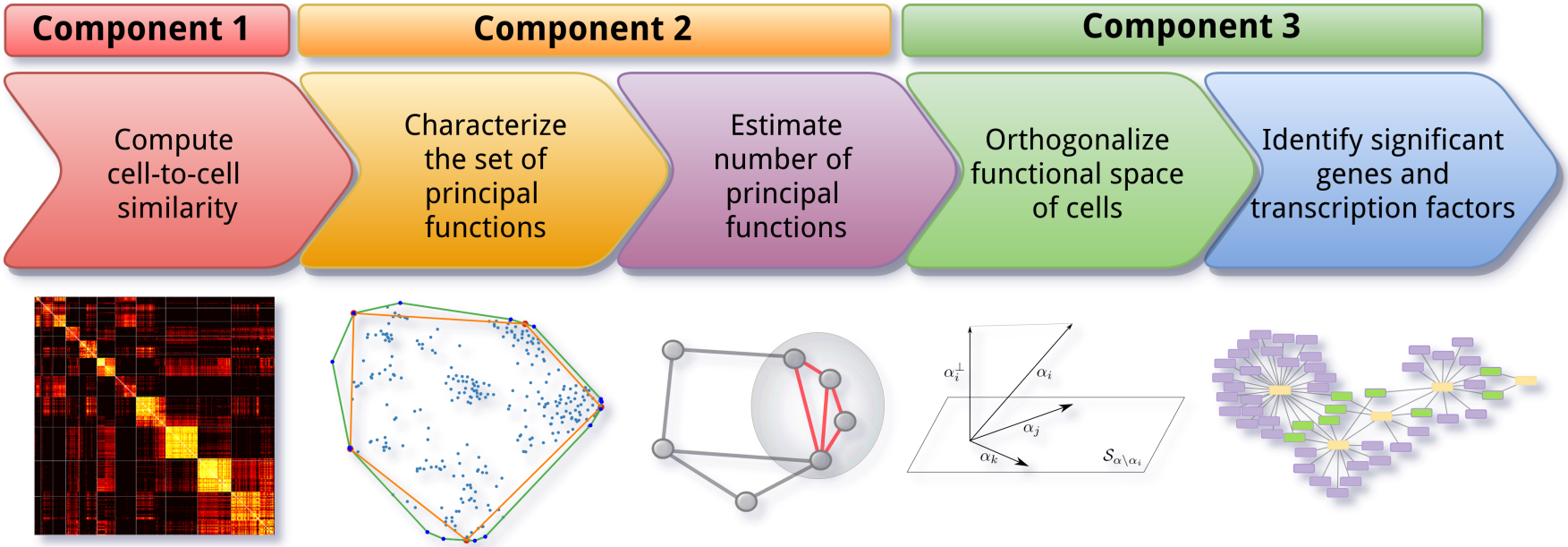
PURDUE UNIVERSITY

Introduction: Some Basic Terminology

- Each cell in an organism (with some noted exceptions) inherits the same genetic code (its genome).
- Different cells exhibit different behavior (and function) as a result of different activity levels of genes and controls.
- Cells exhibiting the same profile of genetic activity are generally believed to be of the same type.
- Within the set of genes, some genes are generally active across all cell types (housekeeping genes), other are selective to sets of cell types (tissue selective), others are specific to cell types (tissue specific).
- Genes whose activity is unique to cell types are called markers.
- The activity of genes in a cell is impacted by its state, stressors (external stimuli), disease, etc.
- One of the common tools to interrogate the state of a cell is to study gene expression using microarrays or RNA Sequencing (RNASeq).
- Among the most common single cell technologies is single cell RNA Seq (scRNASeq).

PURDUE UNIVERSITY

Establishing functional identity of cells



PURDUE UNIVERSITY

Underlying hypothesis

Transcriptional profile of cells is dominated by housekeeping genes, whereas their functional identity is determined by a combination of weak but preferentially expressed genes.

PURDUE UNIVERSITY

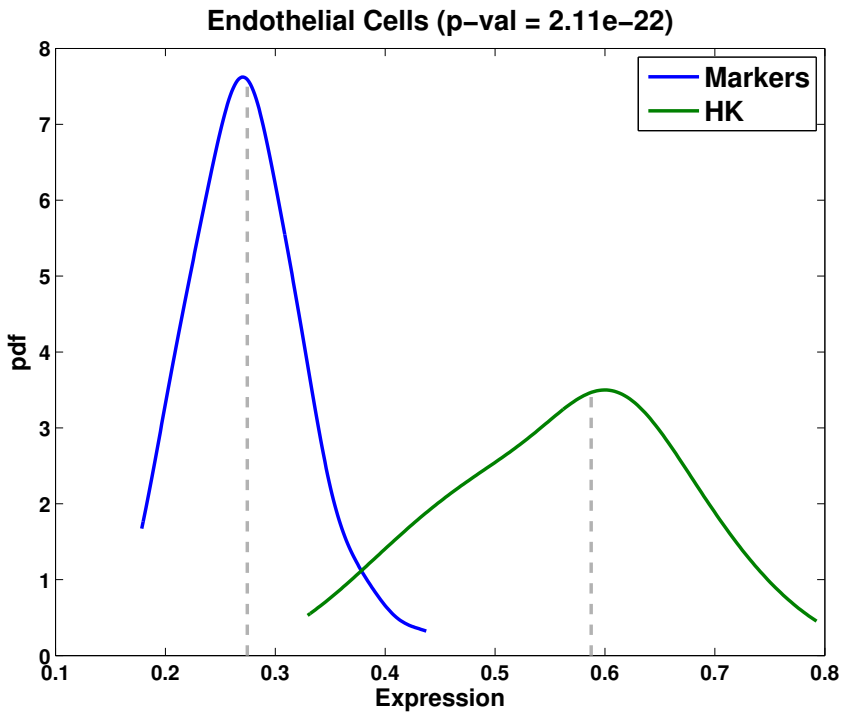


Figure: Endothelial Cells

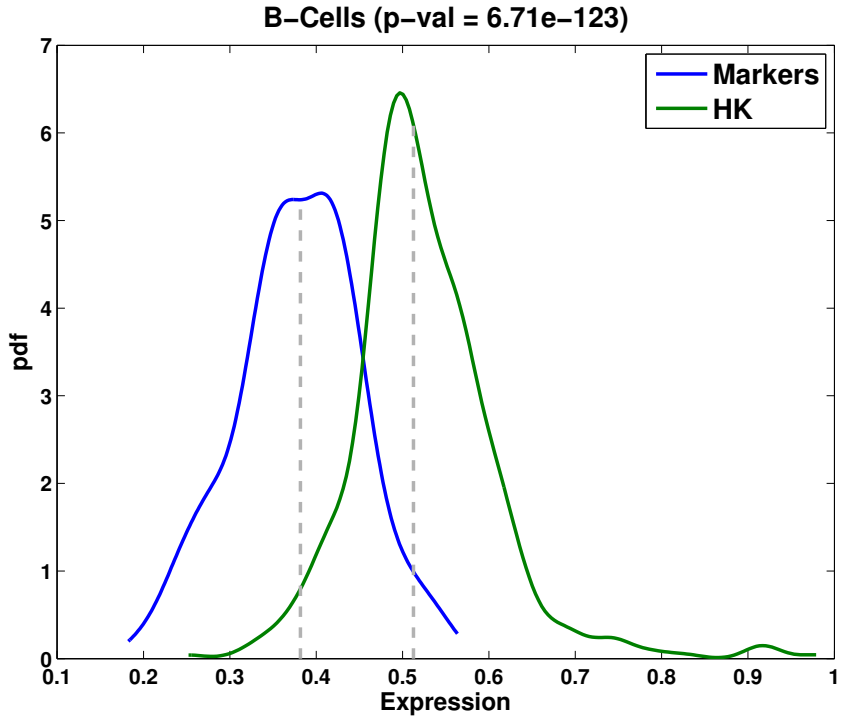
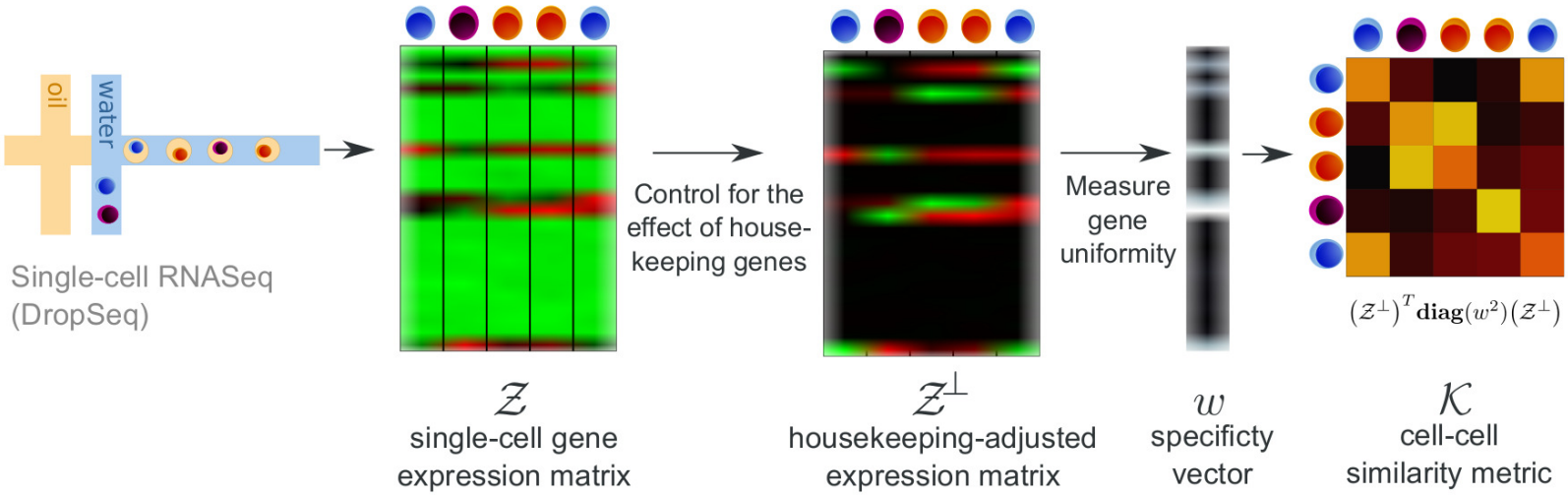


Figure: B-Cells

PURDUE UNIVERSITY

Component 1: New measures for cell-cell similarity

Cell similarity kernel in ACTION



- The main steps involved in identifying similarity between cells

PURDUE UNIVERSITY

ACTION-adjusted cell signatures

$$\mathbf{Y} = \boldsymbol{diag}(\boldsymbol{w})\mathcal{Z}^{\perp}$$

- To compute $\boldsymbol{w}$, we assess how informative is observing a gene with respect to the cell type that it came from
- For each gene i, we compute a specificity factor w_i.

ACTION metric (kernel)

$$\begin{aligned} \mathbf{K}_{ACTION} &= \mathbf{Y}^T\mathbf{Y} \\ &= \left(\mathcal{Z}^{\perp}\right)^T \boldsymbol{diag}(\boldsymbol{w}^2)\left(\mathcal{Z}^{\perp}\right) \end{aligned}$$

PURDUE UNIVERSITY

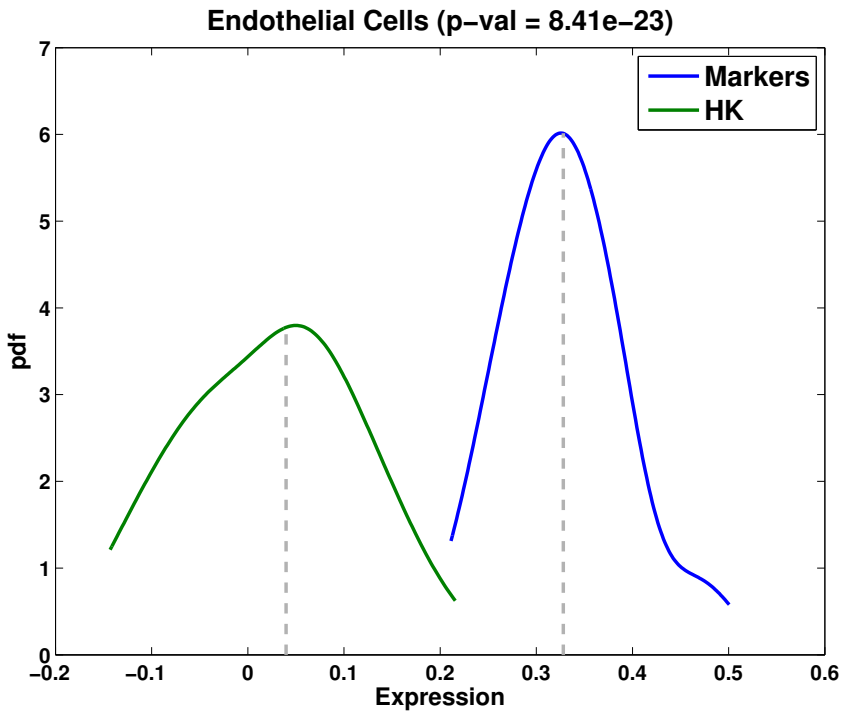


Figure: Endothelial Cells

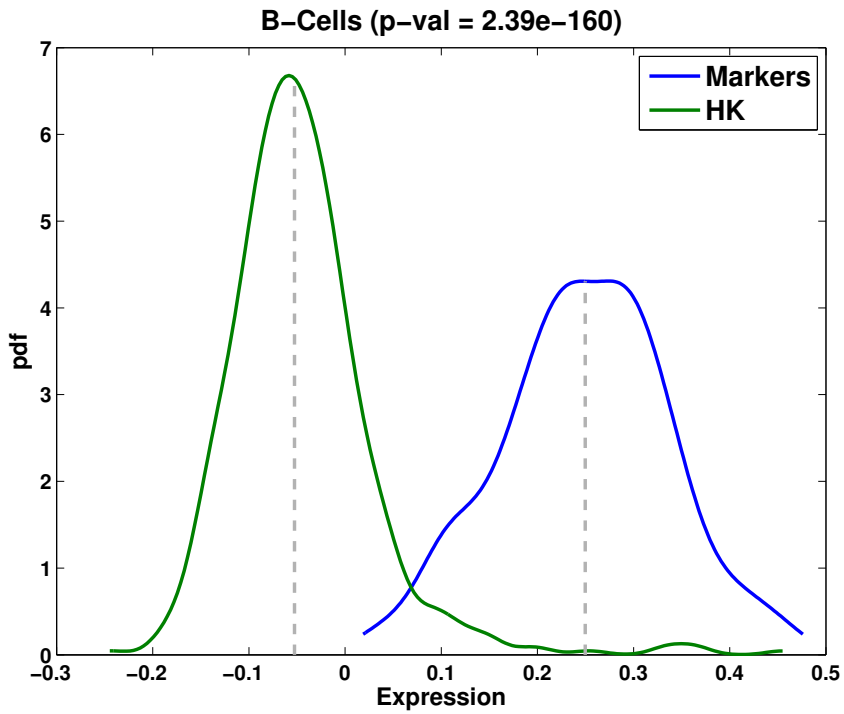


Figure: B-Cells

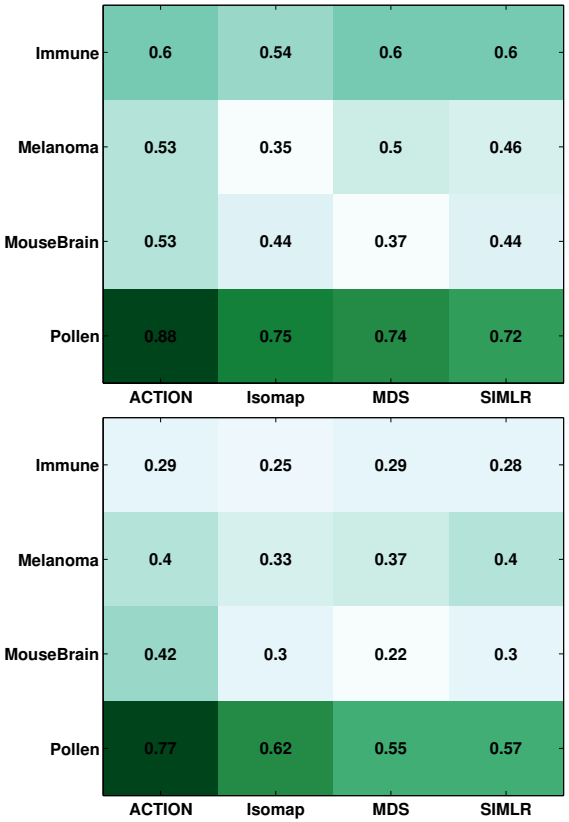
PURDUE UNIVERSITY

Benchmark datasets

- **Immune:** 1,522 immune cells from mouse hematopoietic system (30 different types of stem, progenitor, and fully differentiated cells)
- **Melanoma:** 4,645 malignant, immune, and stromal cells isolated from 19 freshly procured human melanoma tumors (7 major types, including T, B, NK, CAF, Endo, Macro, and Tumor)
- **MouseBrain:** 3005 cells from the mouse cortex and hippocampus (7 major types, including *astrocytes-ependymal, endothelial-mural, interneurons, microglia, oligodendrocytes, pyramidal CA1*, and *pyramidal SS*).
- **Pollen:** Small set of 301 cells spanning 11 different cell types in developing cerebral cortex

PURDUE UNIVERSITY

Performance of ACTION Kernel



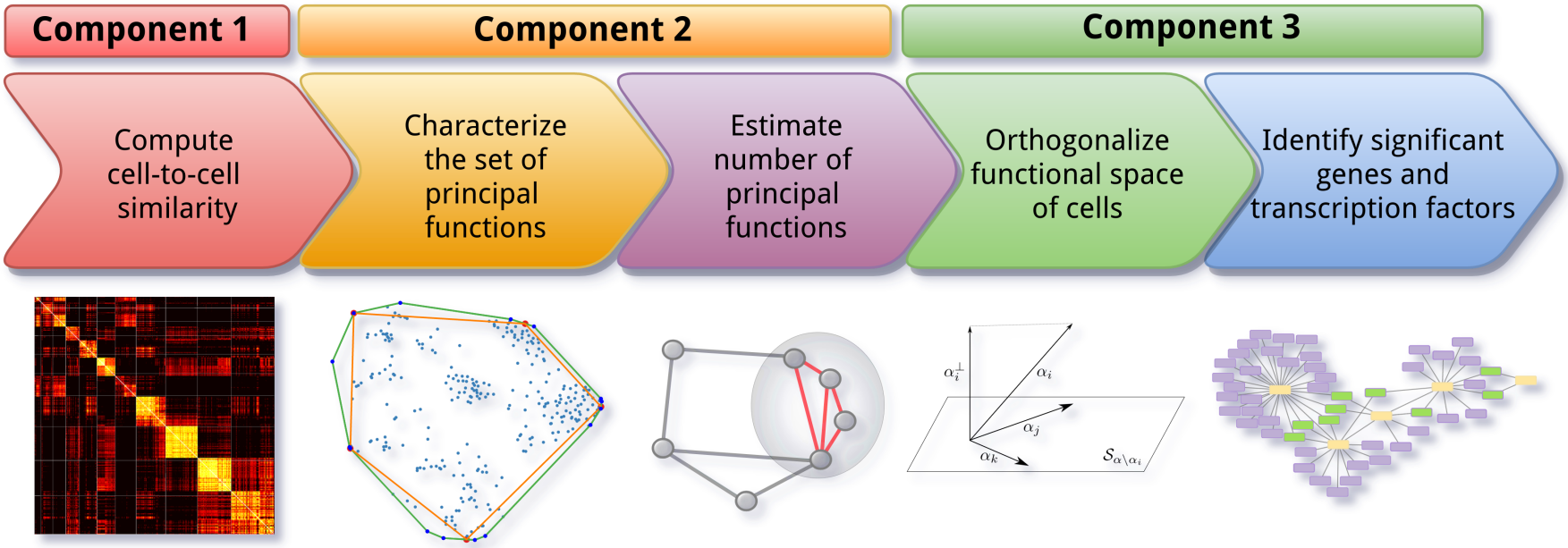
- Benchmarks:
  - SIMLR: Specifically designed for single-cell data
  - IsoMap,MDS: General purpose dimension reduction
- Tested a range of parameters (5:5:50). Reported best case for each method.
- Ties:
  - *Immune* (NMI: ACTION/MDS/SMLR, ARI: ACTON/MDS)
  - *Melanoma* (ARI: ACTION/SIML)
- In all other cases, *ACTION* metric significantly outperforms all other methods.

- Overall, *ACTION* metric performs better than other methods

PURDUE UNIVERSITY

Overall Workflow

Component 2



PURDUE UNIVERSITY

General framework

$$\begin{aligned} \underset{\mathbf{C},\mathbf{H}}{\text{argmin}} \quad & \| \mathbf{Y} - \underbrace{\mathbf{YC}}_{\mathbf{W}} \mathbf{H} \| \\ \text{subject to:} \quad & \| \mathbf{C}(:,i) \|_1 = 1. \\ & \| \mathbf{H}(:,i) \|_1 = 1. \\ & 0 \leq \mathbf{C}, 0 \leq \mathbf{H} \end{aligned}$$

Various algorithms can be cast using this formulation

- K-means: $\mathbf{C} \in \mathbb{R}^+, \mathbf{H} \in \{0,1\}$
- K-medoids: $\mathbf{C} \in \{0,1\}, \mathbf{H} \in \{0,1\}$

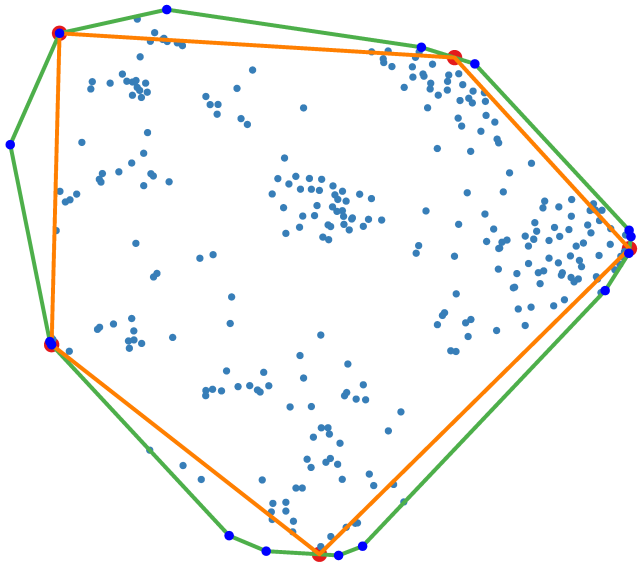
PURDUE UNIVERSITY

Convex NMF

$$\begin{aligned} \underset{\mathcal{K},\mathbf{H}}{\text{argmin}} \quad & \| \mathbf{Y} - \mathbf{Y}(:,\mathcal{S})\mathbf{H} \| \\ \text{subject to:} \quad & \| \mathbf{H}(:,i) \|_1 = 1, \mathbf{H} \in \mathbb{R}^+. \end{aligned}$$

- It uses the same formulation as k-medoid, but relaxes the hard assignment of cells: $\mathbf{C} \in \{0,1\}, \mathbf{H} \in \mathbb{R}^n$
- Unlike k-medoid and k-means, it has an optimal global solution.
  - Under near-separability assumption: there exists, for each cell type, an ideal example in the population.
- A modification of the *Gram Schmidt* process.

PURDUE UNIVERSITY



Geometry of functional space: each point is a cell and red points are the "pure cells"

- Picking k corner points/archetypes from the convex hull of the cells, such that they optimally "contain" the rest of cells.
- Each archetype is an ideal example of a cell type with a distinct set of principal functions.

PURDUE UNIVERSITY

- AA further relaxes matrix $\mathbf{C}$: $\mathbf{C}, \mathbf{H} \in \mathbb{R}^{+}$.
- It can handle cases where pure pixel assumption is violated.
- But it no longer has global convergence guarantee $\rightarrow$ it is also dependent on the initialization
  - To address this, we use the solution of convex NMF for initializing AA.
- In essence, this allows local adjustment of the Convex NMF solution.
- This can be thought of as a variant of block-coordinate descent for optimization.

PURDUE UNIVERSITY

Goal: To identify when we should stop adding new archetypes.

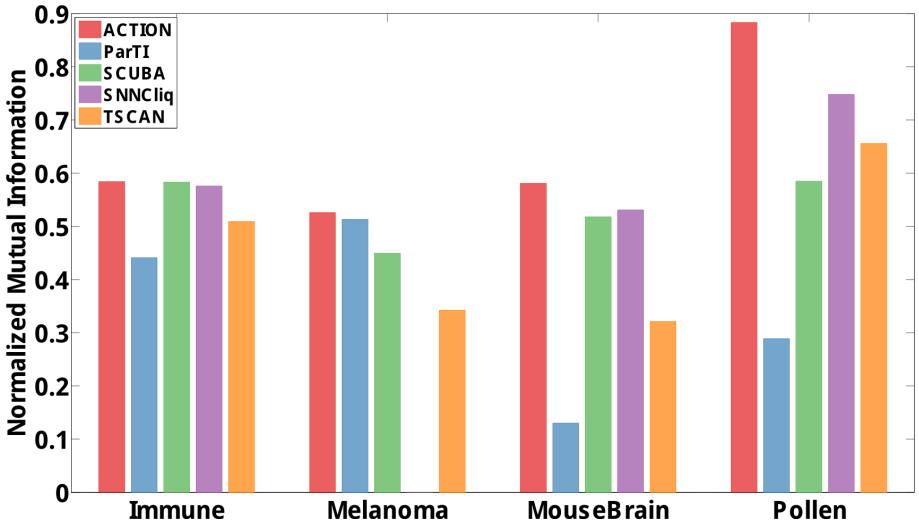
- Underlying concept: add archetypes until we sense "oversampling."
- Oversampling happens when we start adding archetypes that are "too close" to each other.
- Each archetype is a cell $\rightarrow$ we can compute their similarity of using the ACTION metric.

PURDUE UNIVERSITY

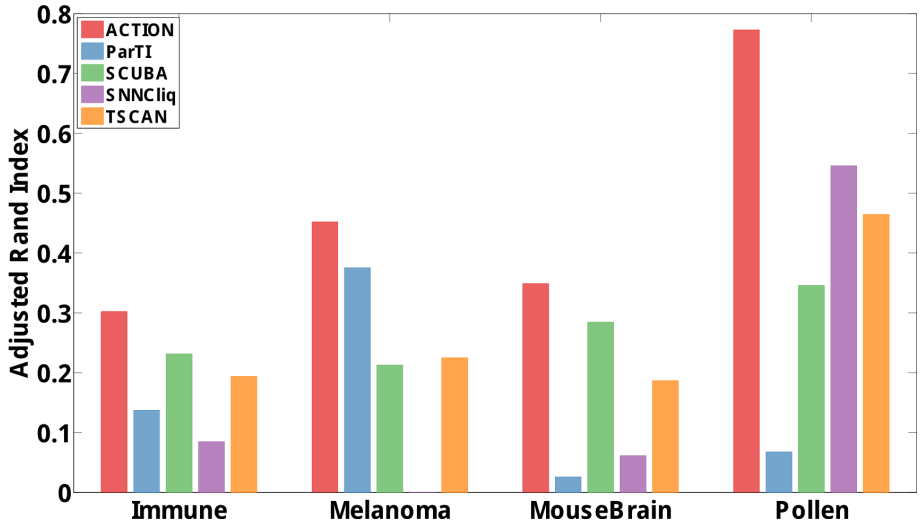
Component 2: Characterizing principal functional profiles

Test 1: Identifying cell types using closest archetype

a



b



- ACTION excels at identifying underlying cell types in all cases

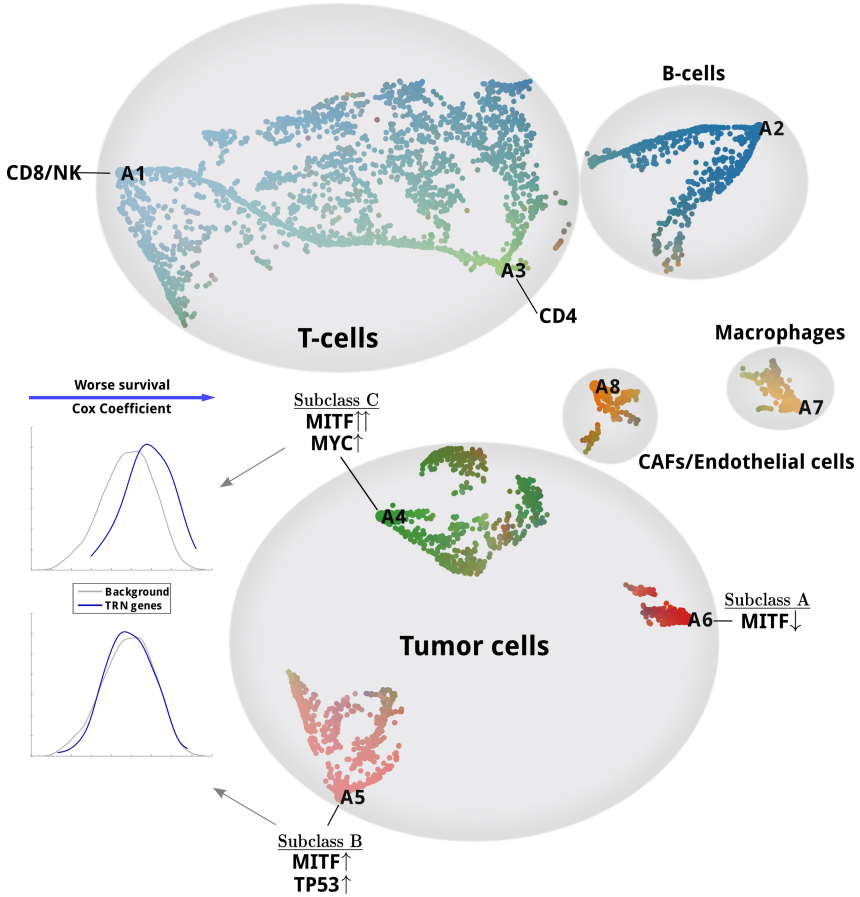
PURDUE UNIVERSITY

- Use matrix $\mathbf{H}$ instead of $\mathbf{Y}$ in visualization:
  - We are interested in the relationship between cells and their surrounding archetypes.
- Initialize using Fiedler embedding
  - Position according to the dominant eigenvectors of the Laplacian matrix: $\mathbf{L} = \mathbf{diag}(\Delta_{\mathbf{Y}}) - \mathbf{Y}$.
- Update using t-SNE

PURDUE UNIVERSITY

A continuous view of transcriptional profiles

Case study in the Melanoma dataset



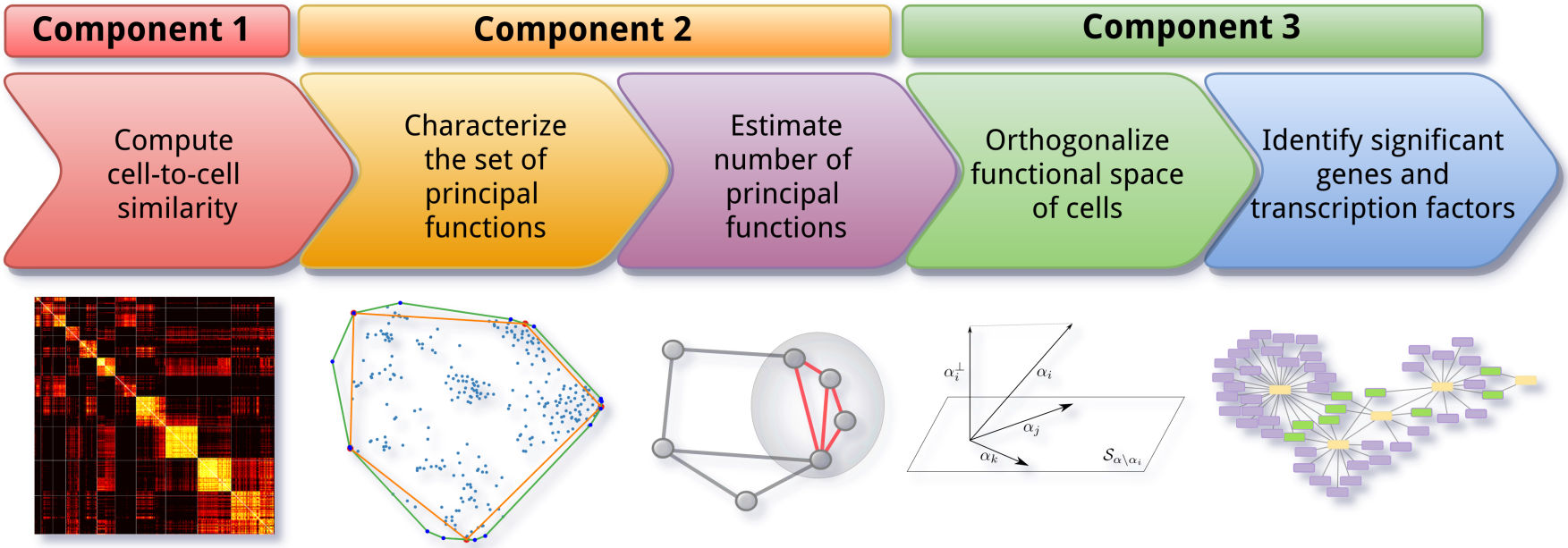
- T-cells reside in a continuum of states (Thogerson *et al.*).
- Tumor cells form compact groups.
- Two subclasses of MITF-associated tumors significantly differ in terms of their survival.

- ACTION highlights the underlying topology of cell types

PURDUE UNIVERSITY

Overall Workflow

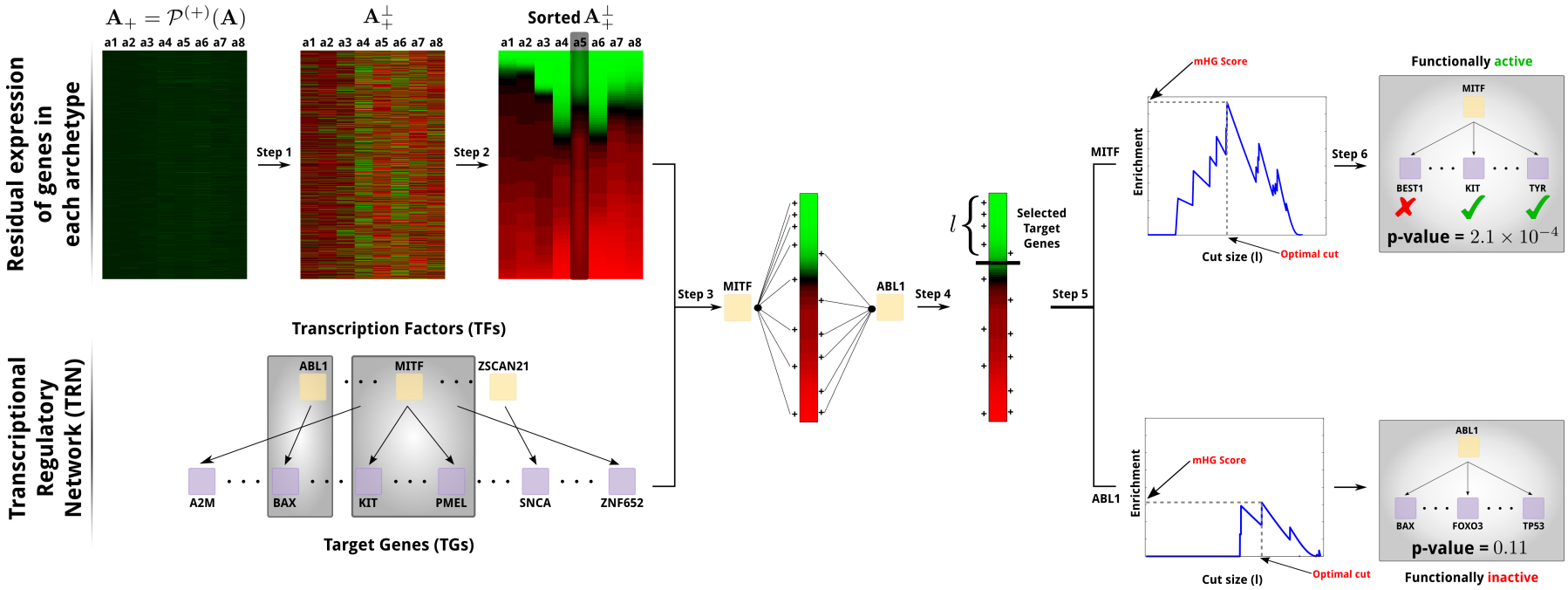
Component 3: Identifying the interactions underlying architypes



PURDUE UNIVERSITY

Component 3: Identifying the interactions underlying architypes

Constructing TRN



PURDUE UNIVERSITY

Goal: Identifying key regulatory elements that drive each cell type

1. Archetype Orthogonalization ($\rightarrow$ Only over positive projection)

$$\boldsymbol{a}_i^{\perp} = \left(\mathbf{I} - \mathbf{A}_{-i}(\mathbf{A}_{-i}^T\mathbf{A}_{-i})^{-1}\mathbf{A}_{-i}^T\right)\boldsymbol{a}_i$$

2. Assessing significance of TFs/TGs

$$\begin{aligned} p\text{-value}(Z = b_l(\lambda)) &= \text{Prob}(b_l(\lambda) \leq Z) \\ &= \sum_{x=b_l(\lambda)}^{min(T,l)} \frac{\binom{T}{x}\binom{m-T}{l-x}}{\binom{m}{l}} \end{aligned}$$

Use Dynamic Programming to compute exact p-value.

PURDUE UNIVERSITY

Functional activity of transcription factors (TFs)

Key point!

We identify "functional activity" of transcription factors (TFs) by aggregating transcriptional activity of their downstream targets, not the transcriptional level of TFs themselves. TFs can, and typically do, get regulated through post-translational mechanisms.

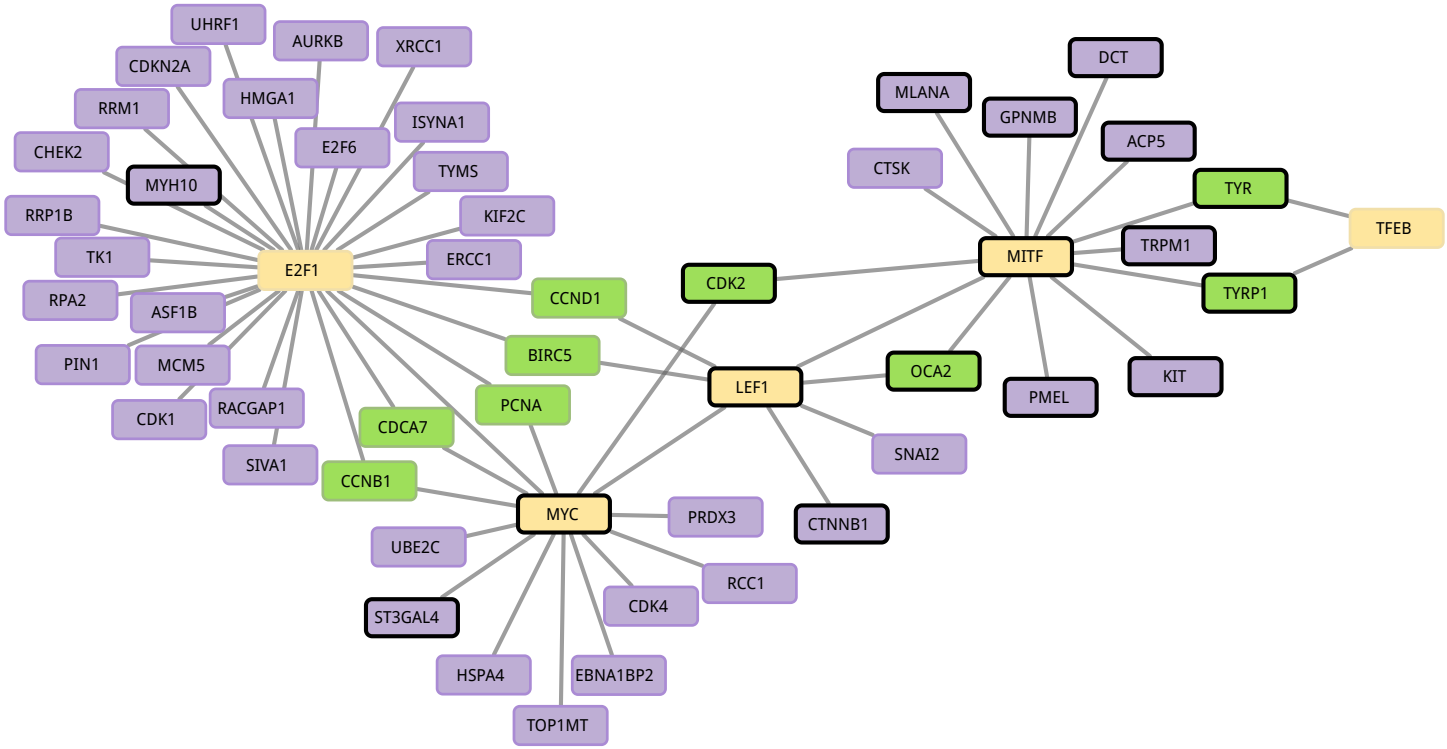
PURDUE UNIVERSITY

- Both *Subtype A* and *Subtype C* exhibit high activity of *MITF* and *Sox10* transcription factors, which are canonical markers for melanoma cells in the "proliferative" (as opposed to "invasive") state (Verfaiilie *et al.*).
- These two subtypes are significantly enriched for marker genes in the proliferative state:
  - *Subtype A:* 9.3×10^{-14}
  - *Subtype B:* 7.9×10^{-11}
- Subtype A has higher MITF activity (according to its activated targets):
  - *GPNMB*, *M1ANA*, *PMEL*, and *TYR* are shared between two subtypes.
  - *ACP5*, *CDK2*, *CTSK*, *DCT*, *KIT*, and TRPM1/P1 are uniquely upregulated in subtype A.

PURDUE UNIVERSITY

Dissecting transcriptional controls of Melanoma subclasses

Case study in MITF⇑⇑/MYC↑ subtype

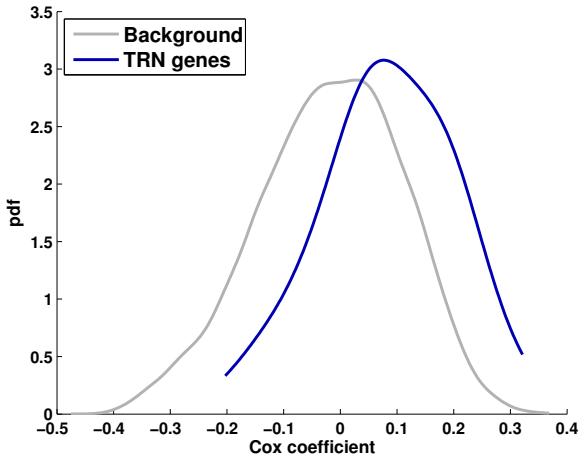


- 19 "functionally" active transcription factors in subtype A (p-value ≤ 0.05)
- We focus on the five most significant TFs and their targets (p-value $\leq 10^{-3}$)

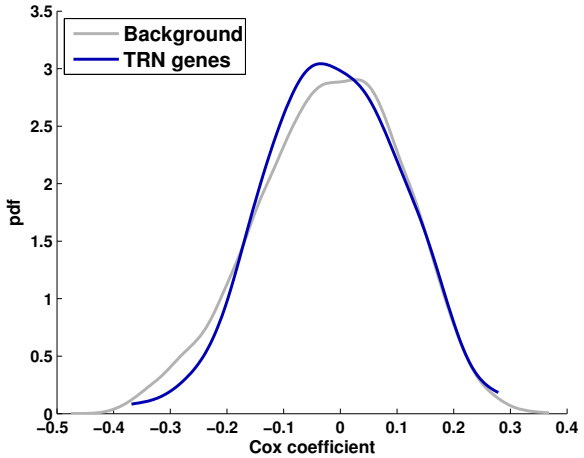
PURDUE UNIVERSITY

- MITF is among the best-known markers for classifying melanoma patients (Hartman *et al.*: MITF in melanoma: mechanisms behind its expression and activity).
- Overexpression of the E2F1 is common in high-grade tumors that are associated with poor survival in melanoma patients (Alla *et al.*: E2F1 in melanoma progression and metastasis).
- Melanoma cell phenotype switching, between proliferative an invasive states, is regulated by differential expression of LEF1/TCF4 (Eichhoff *et al.*:Differential LEF1 and TCF4 expression is involved in melanoma cell phenotype switching).
- Amplification and overexpression of the c-myc have been associated with poor outcome (Kraehn *et al.*: Extra c-myc oncogene copies in high risk cutaneous malignant melanoma and melanoma metastases).

PURDUE UNIVERSITY



Subtype A: p-value $= 5.4 \times 10^{-10}$



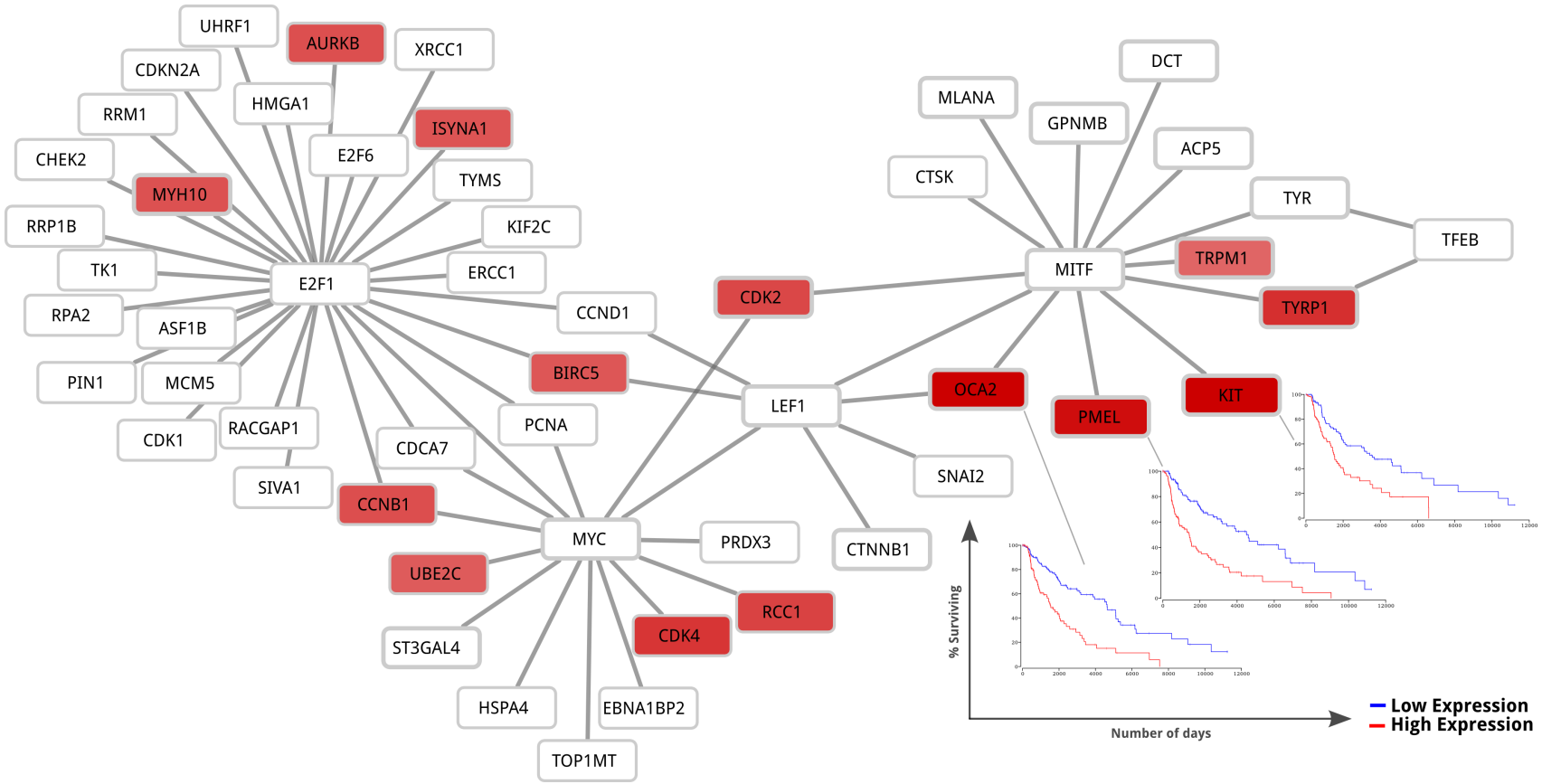
Subtype C: p-value $= 0.31$

- OncoLnc (Jordan Anaya)
- Multivariate Cox regressions
- Gene expression, sex, age, and grade or histology as factors
- Genes associated with Subclass A have significantly worse outcome, compare to the background of all genes

PURDUE UNIVERSITY

Case study in MITF⇑/MYC↑ subtype

Survival analysis revisited – Kaplan-Meier plots



Recap

1. A novel cell similarity metric that is robust to biological noise, while at the same time is sensitive enough to identify weak cell type-specific signals
2. New notion of functional identity of cells
   - Under the pure cell assumption, this metric induces a convex topology that embeds functional identity of cells
3. Use functional identity of cells to identify both discrete cell types and continuous cell states
4. Identify driving transcriptional controls that mediate the functional identity of cells

Clinical significance: Characterization of two MITF-associated subclasses of Melanoma patients, one of which has substantially worse outcomes, along with their underlying regulatory elements.

PURDUE UNIVERSITY