Novel formulations and methods for emerging problems in single cell data analysis

Mohammadi, Ravindra, Szpankowski, Subramaniam, Goldsmith, Gleich, Grama

Purdue University

STC Annual Review, 2017

Nov 29, 2017

PURDUE UNIVERSITY

Introduction and Overview

- The wealth of data from various biological screens offers tremendous opportunities and challenges for Science of Information.
- 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 is orders of magnitude more than conventional genomic/ transcriptomic/ interactomic technologies.
- CSoI collaborations have resulted in very powerful and among the first computational/ statistical/ modeling toolkits for single cell datasets.

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 in 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

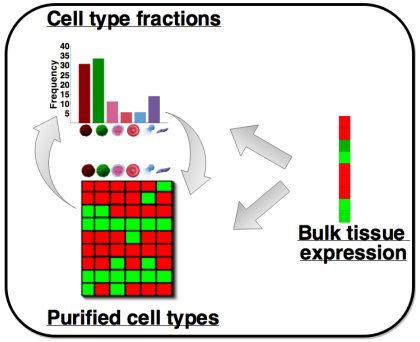
Introduction: Some Basic Terminology

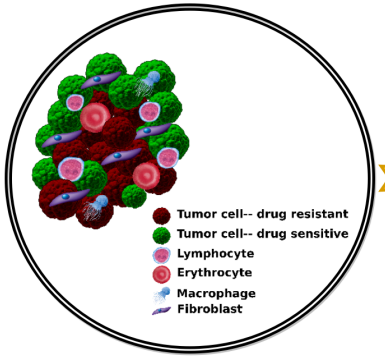
- Each cell in an organism (with some noted exceptions) inherits the same genetic code (its genome).
- Different cells exhibit different behavior (and function) as a result of different activity levels of genes.
- Cells exhibiting the same profile of genetic activity are generally believed to be of the 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.
- The most common tool 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).

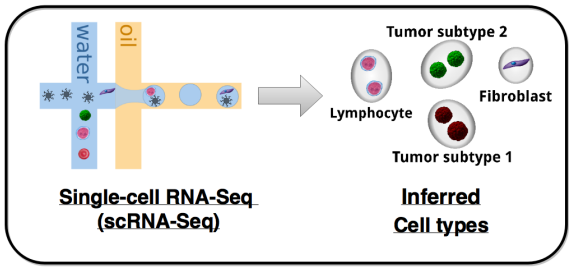
PURDUE UNIVERSITY

Overview of Presentation

Mixture of different cells





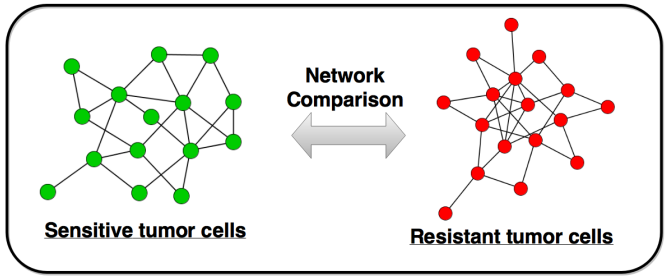


Deconvolving gene expression profiles

A Comprehensive Survey of Deconvolution Methods for Separating Cell types in Complex Tissues, 2016

Identifying cell types from single-cell profiles

Take ACTION to identify high-resolution cell types and associated transcriptional pathways, 2016



Constructing cell type-specific networks

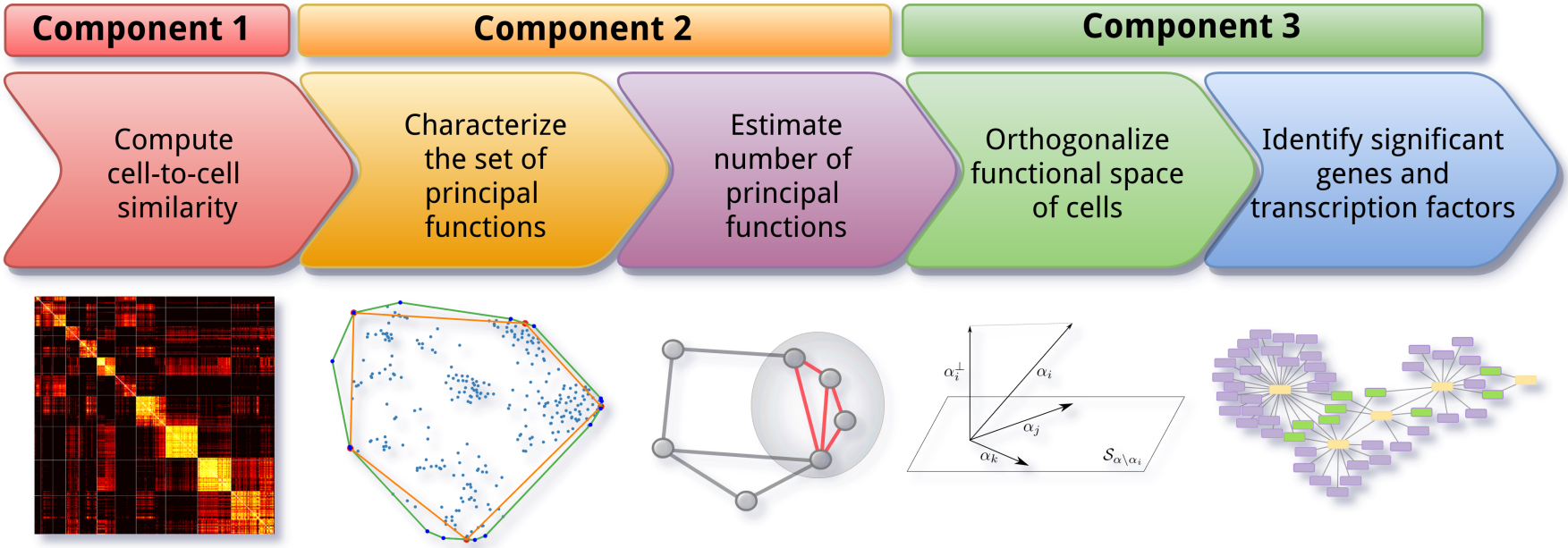
A convex optimization approach for identification of human tissue-specific interactomes, 2016

PURDUE UNIVERSITY

1. Establishing functional identity of cells
2. Part II: Constructing tissue/cell type-specific networks
3. Part III: Deconvolving expression profiles of complex tissues

PURDUE UNIVERSITY

Establishing functional identity of cells



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.

Component 1: New measures for cell-cell similarity

Supporting evidence

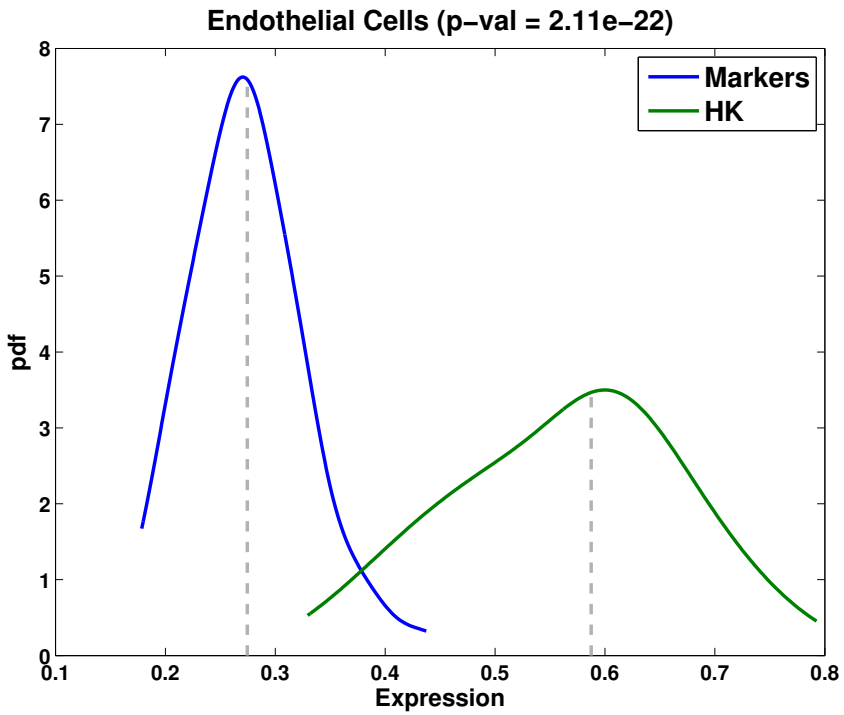


Figure: Endothelial Cells

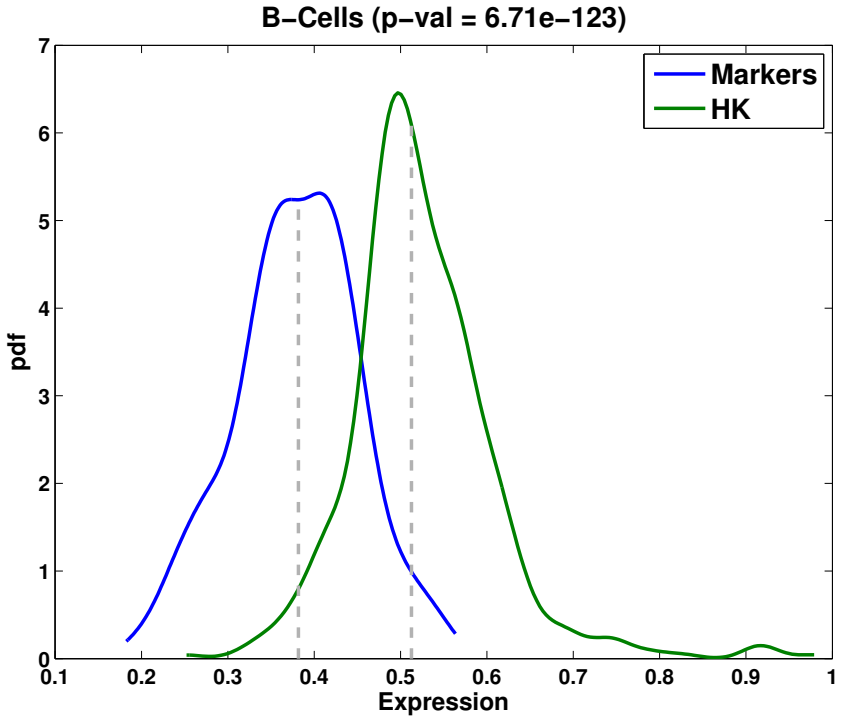


Figure: B-Cells

PURDUE UNIVERSITY

Component 1: New measures for cell-cell similarity

Supporting evidence

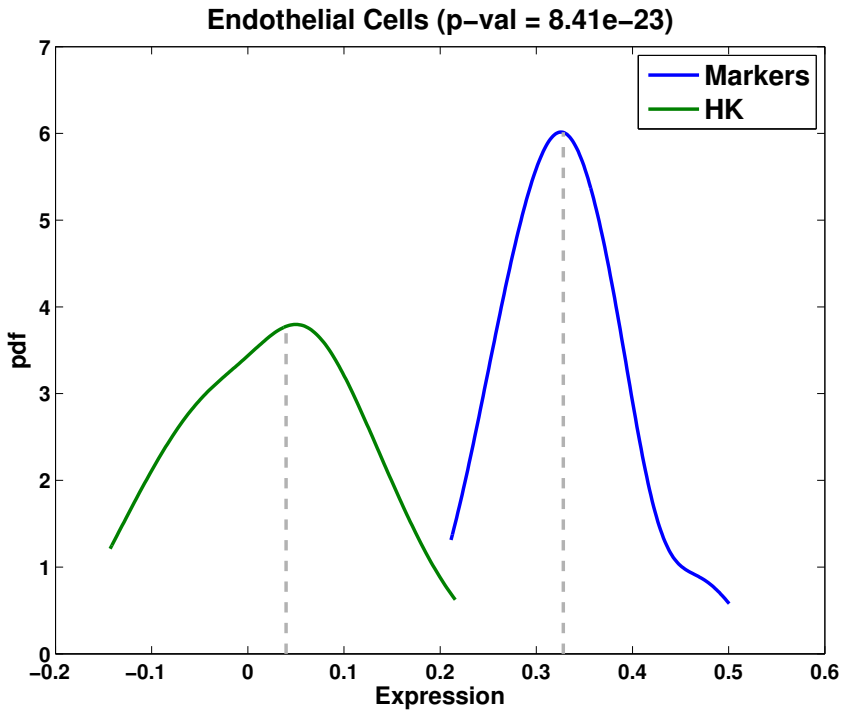


Figure: Endothelial Cells

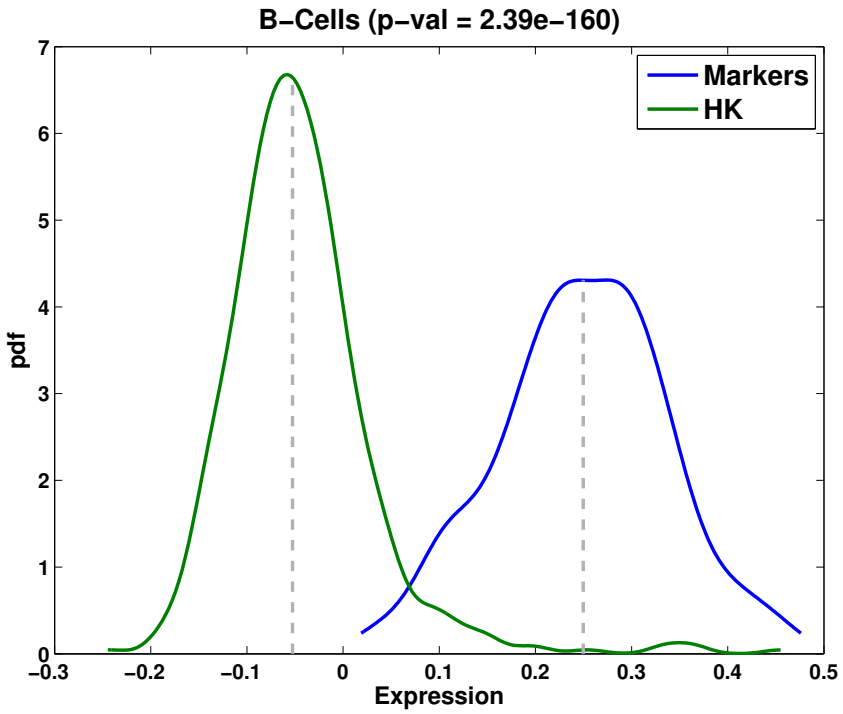
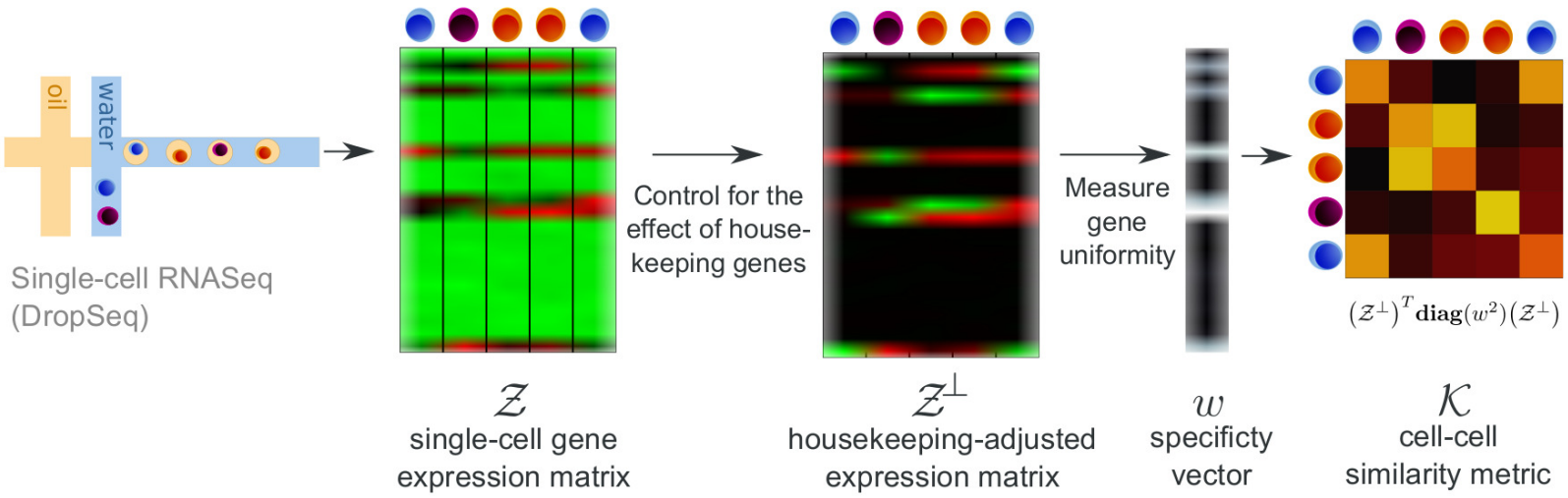


Figure: B-Cells

PURDUE UNIVERSITY

Component 1: New measures for cell-cell similarity

Cell similarity kernel in ACTION



- The main steps involved in identifying similarity between cells

Goal: Identify the shared subspace of genes

Low-rank decomposition

$$A = U_r \Sigma_r V_r = \sum_{i=1}^{r} \sigma_i u_i {v_i}^T,$$

Example decomposition choices:

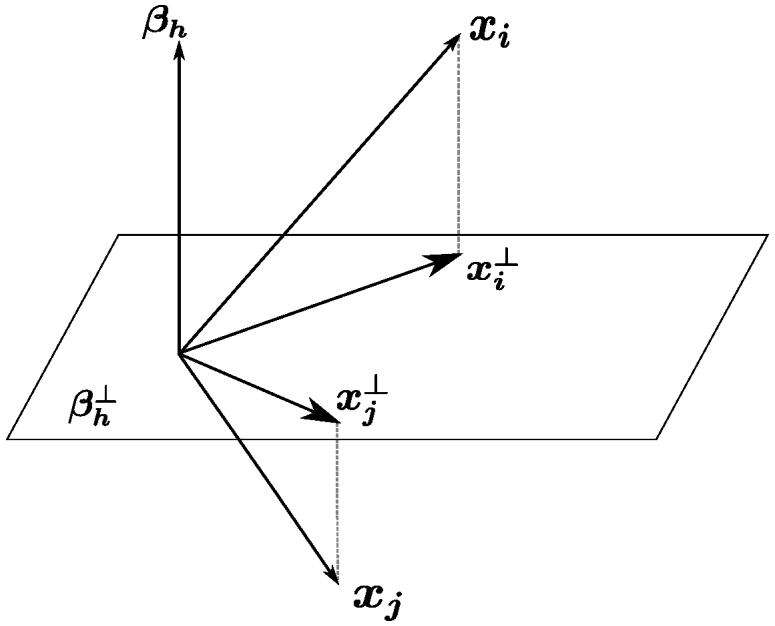
- Mean vector
  - *Optimal in a least-square sense when the chance of observing a gene is uniform across all cells.*
- Singular Value Decomposition (SVD)
- Nonnegative Matrix Underapproximation (NMU)
- Sparse NMU

PURDUE UNIVERSITY

Goal: Remove the effect of common subspace

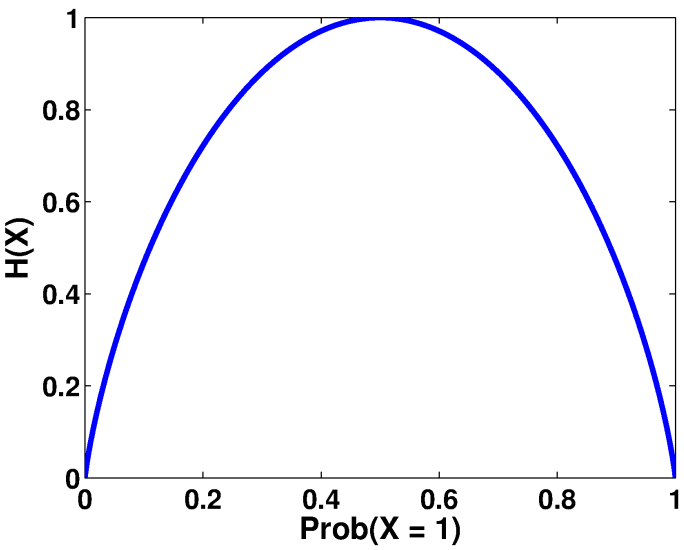
- $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$: tissues/cell types i and j
- z-score normalize $\boldsymbol{x}_i$ to compute $\boldsymbol{z}_i$
- $\boldsymbol{\beta}_h$: the common signature
- z-score normalize $\boldsymbol{\beta}_h$ to compute $\boldsymbol{z}_h$
- Project to the orthogonal subspace:

$$\boldsymbol{z}_i^{\perp} = \left(\mathbf{I} - \frac{\boldsymbol{z}_h \boldsymbol{z}_h^T}{\|\boldsymbol{z}_h\|_2^2}\right)\boldsymbol{z}_i.$$



Similar in nature to the partial Pearson's correlation

Goal: Estimate expression-specificity of genes across different cells



- Entropy as a measure of expression uniformity: $H(i) = -\sum_j p_{ij} log(p_{ij})$
- How informative is observing a gene with respect to the cell type that it came from
- Maximum entropy when probability of a gene coming from all cell types is equal
- For each gene i, compute a specificity factor w_i.

Similar formulations have been previously used for marker detection.

ACTION-adjusted cell signatures

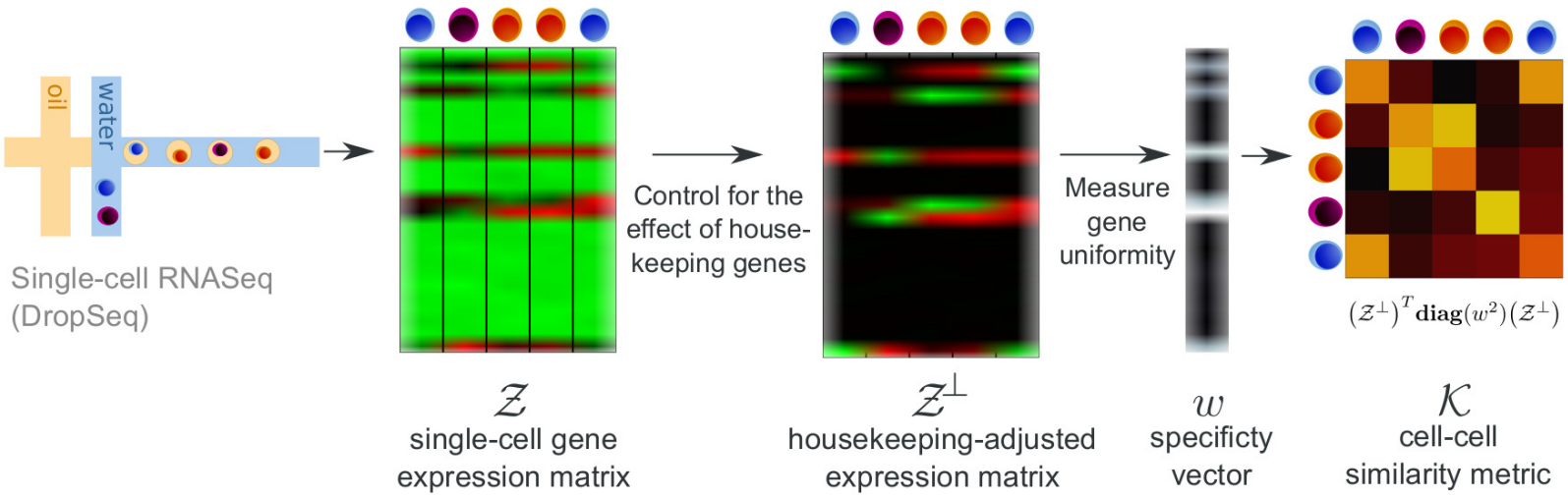
$$\mathbf{Y} = \boldsymbol{diag}(\boldsymbol{w})\mathcal{Z}^{\perp}$$

ACTION metric (kernel)

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

Component 1

Cell similarity kernel – revisited



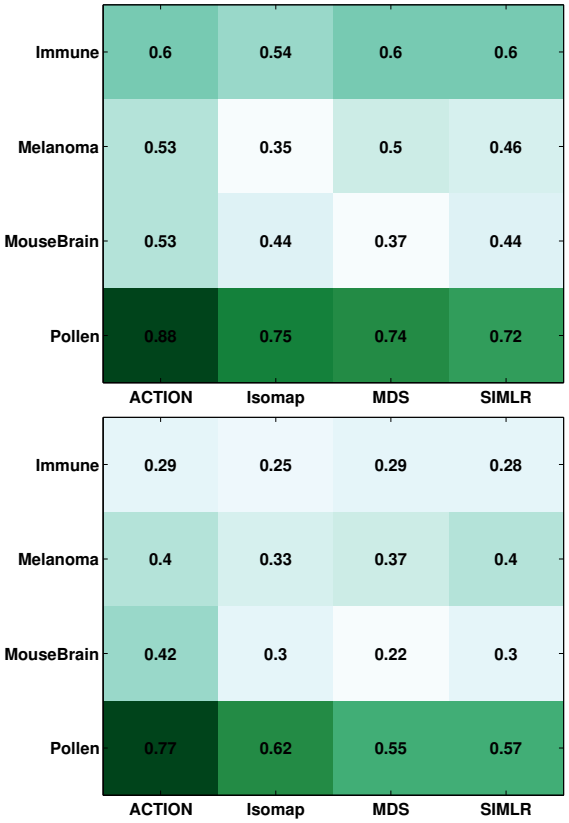
- Now we have computed the ACTION kernel

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



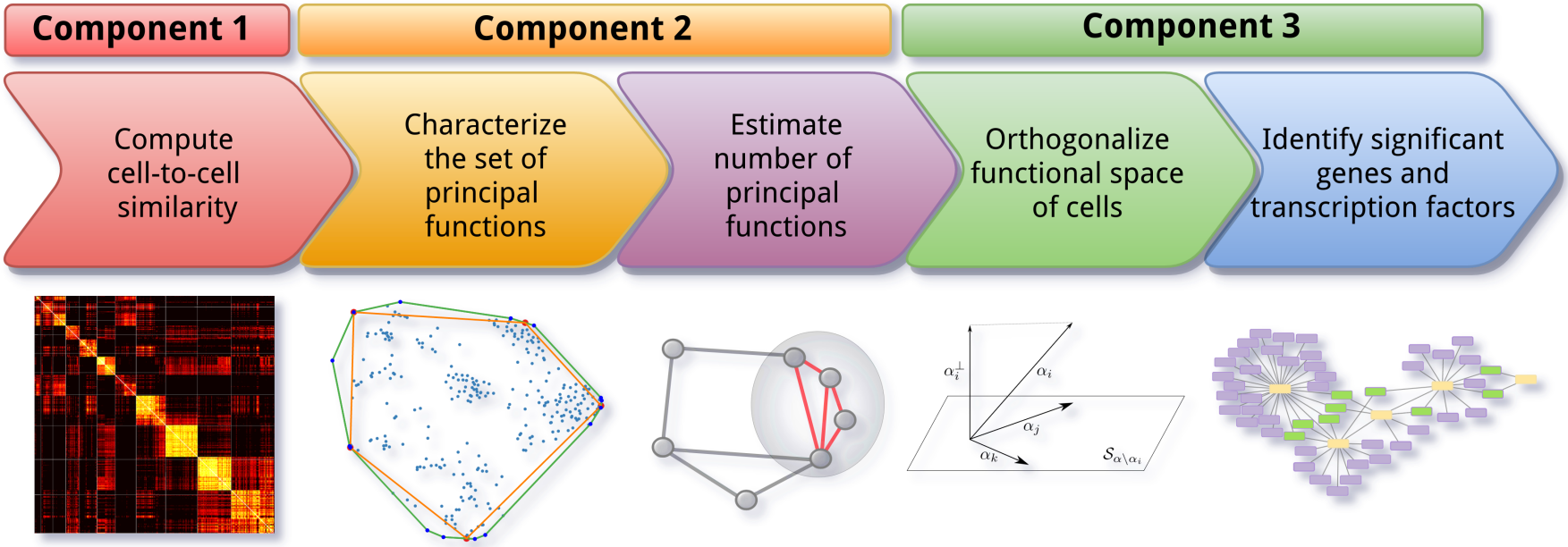
- Benchmarks:
  - SIMLR: Specifically designed for single-cell data
  - IsoMap,MDS: General purpose dimension reduction
- Tested a range of parameters (5:5:50). Reported best case for each method.
- Ties:
  - *Immune* (NMI: ACTION/MDS/SMLR, ARI: ACTON/MDS)
  - *Melanoma* (ARI: ACTION/SIML)
- In all other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

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

PURDUE UNIVERSITY

Overall Workflow

Component 2



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: $\mathbf{C} \in \{0,1\}, \mathbf{H} \in \{0,1\}$

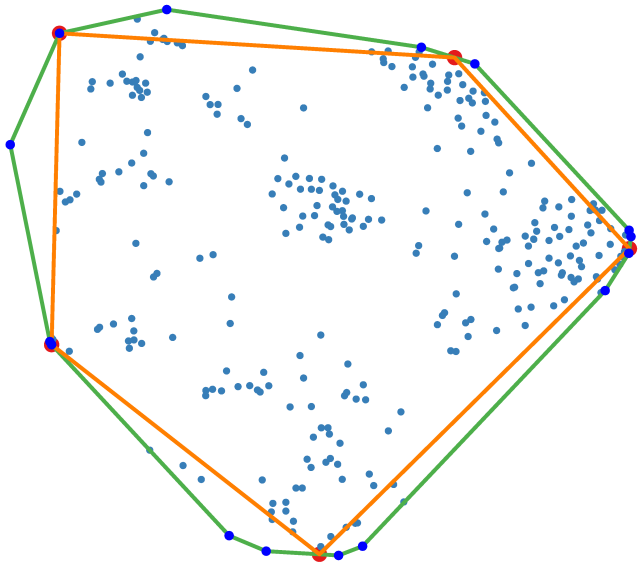
PURDUE UNIVERSITY

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.

Goal: Understand the behavior of near-separable NMF

Performance guarantee

$$\max_{1 \le j \le r} \min_{s \in \mathcal{S}} \| \mathbf{Y}(:,s) - \mathbf{W}(:,j) \| \le \mathcal{O}\Big(\epsilon \kappa^2(\mathbf{W})\Big)$$

- For any near-separable matrix, multiplying it by any nonsingular matrix $\mathbf{Q}$ preserves separability, where matrix $\mathbf{W}$ is replaced with $\mathbf{QW}$.
- In this case, we have the following modified upper bound: $\mathcal{O}\Big(\epsilon \kappa(\mathbf{W}) \kappa^3(\mathbf{QW})\Big)$.

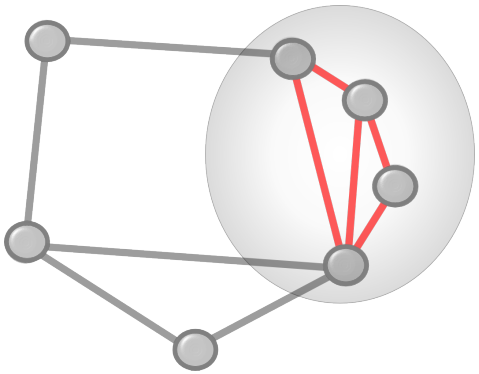
- AA further relaxes matrix $\mathbf{C}$: $\mathbf{C}, \mathbf{H} \in \mathbb{R}^+$.
- It can handle cases where pure pixel assumption is violated.
- But it no longer has global convergence guarantee $\rightarrow$ it is also dependent on the initialization
  - To address this, we use the solution of 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.

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.

Component 2: Characterizing principal functional profiles

Statistical significance of oversampling

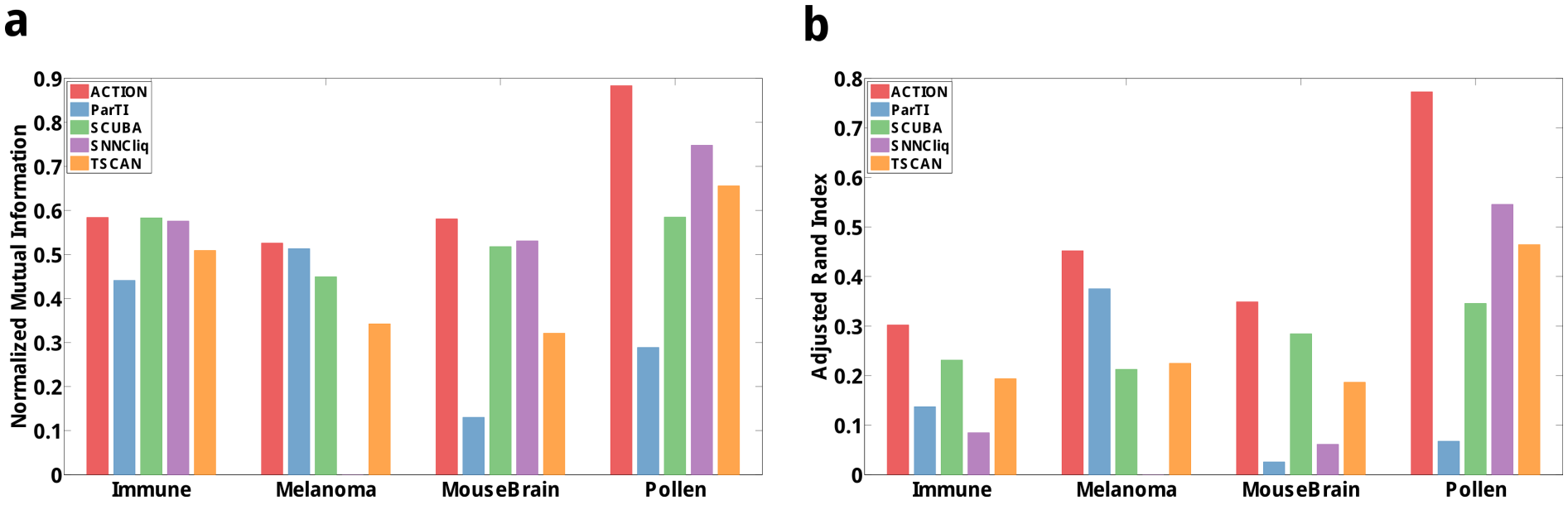


- We build and threshold an archetype-archetype similarity graph.
- For each connected component in this graph, we assess its statistical significance using an Erdos Renyi graph model.
- Probability that there exists in $\boldsymbol{G}$ a subgraph of density $\delta(Z)$ and size at least $|Z|$:

$$\Pr[\exists H \subseteq \boldsymbol{G}, |H| \geq |Z| \ : \ \delta(H) = \delta(Z)].$$

Component 2: Characterizing principal functional profiles

Test 1: Identifying cell types using closest archetype

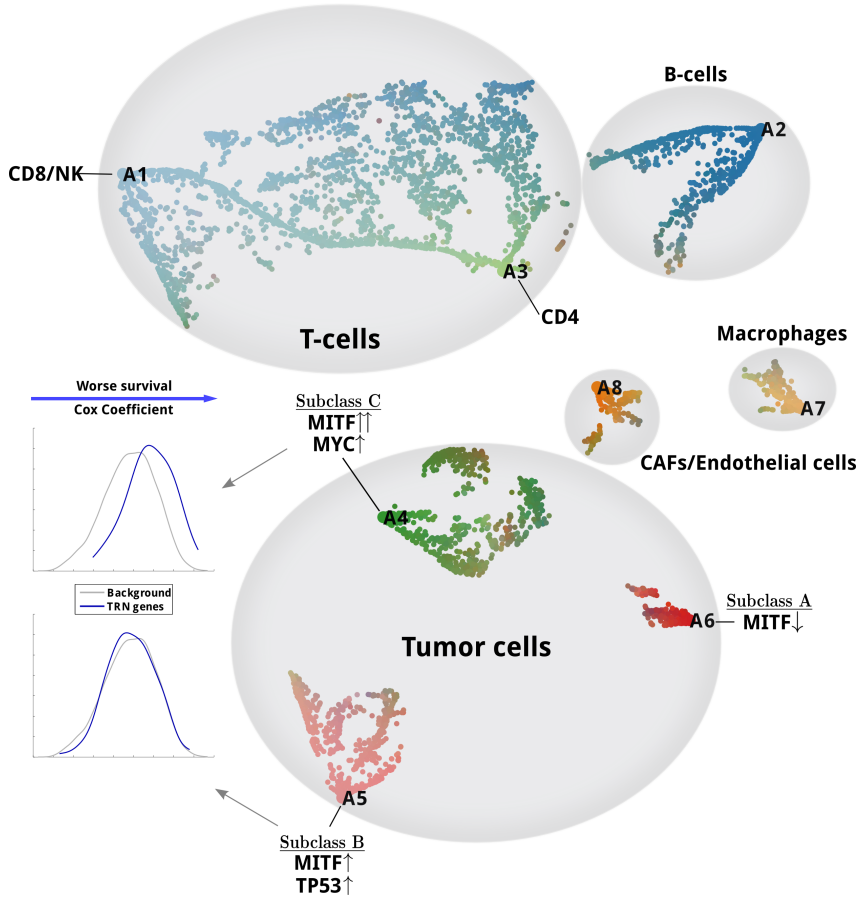


- ACTION excels at identifying underlying cell types in all cases

- 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

A continuous view of transcriptional profiles

Case study in the Melanoma dataset



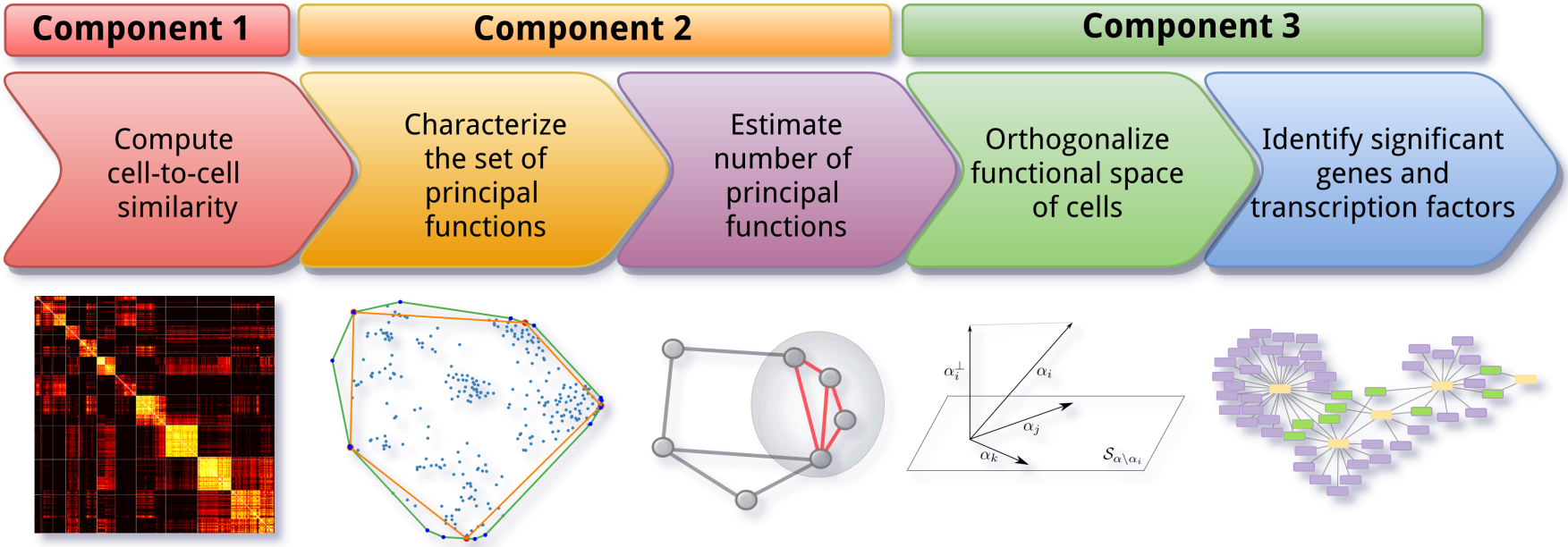
- T-cells reside in a continuum of states (Thogerson *et al.*).
- Tumor cells form compact groups.
- Two subclasses of MITF-associated tumors significantly differ in terms of their survival.

- ACTION highlights the underlying topology of cell types

PURDUE UNIVERSITY

Overall Workflow

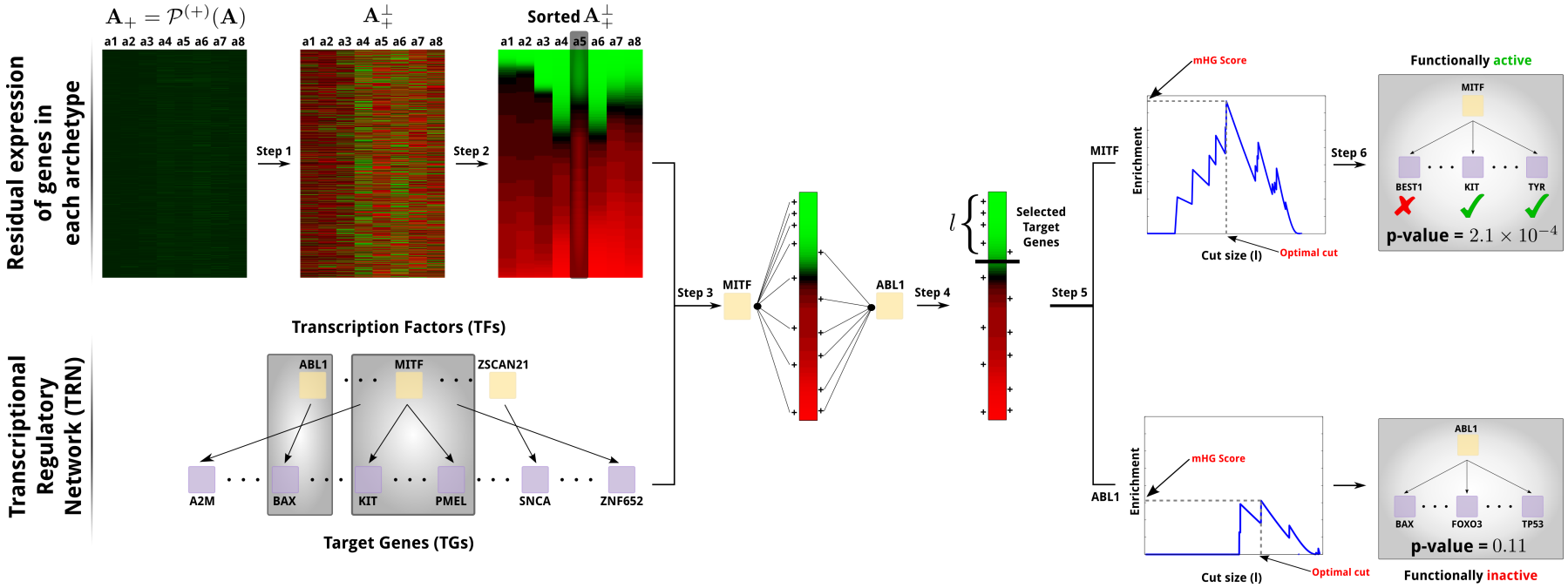
Component 3: Identifying the interactions underlying architypes



PURDUE UNIVERSITY

Component 3: Identifying the interactions underlying architypes

Constructing TRN



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.

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

Identifying transcriptional controls of Melanoma subtypes

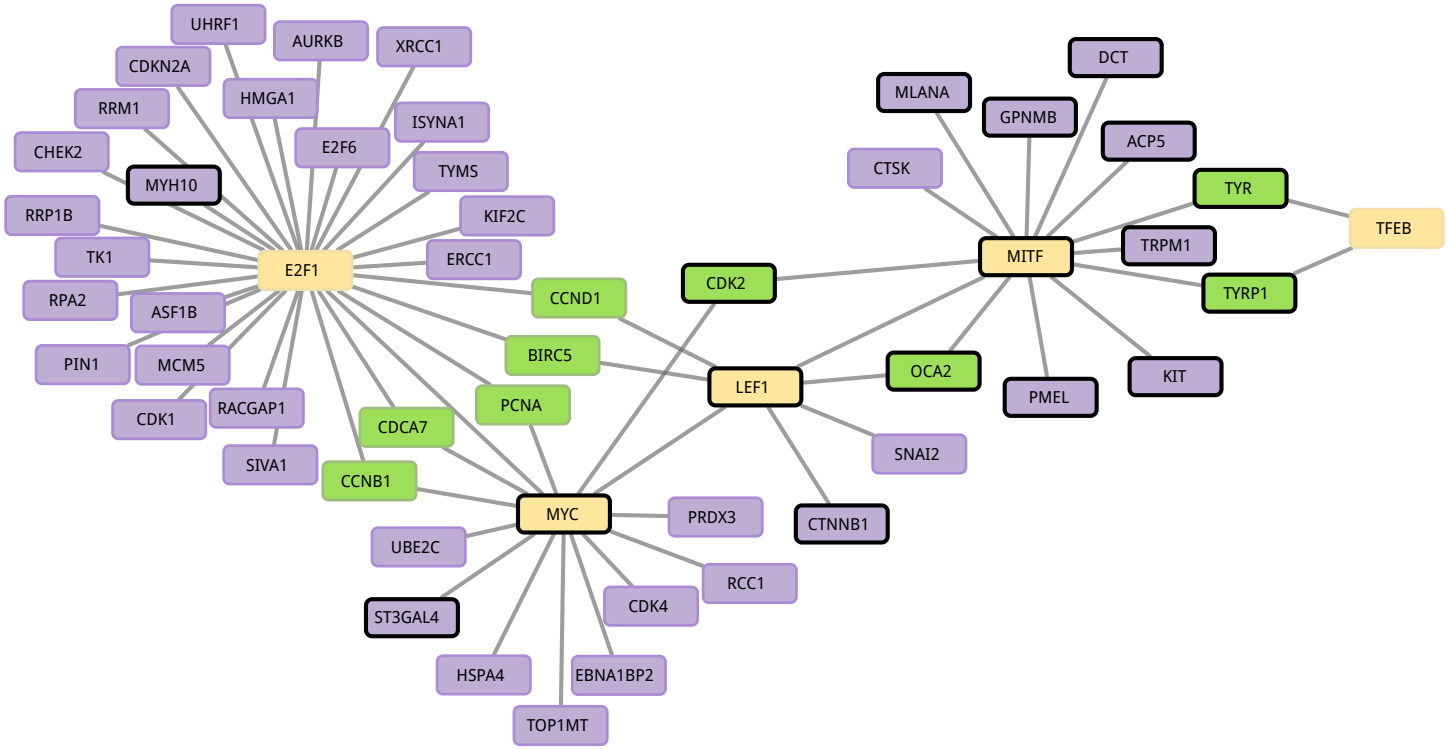
Proliferative versus invasive status

- 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

Case study in MITF⇈/MYC↑ subtype

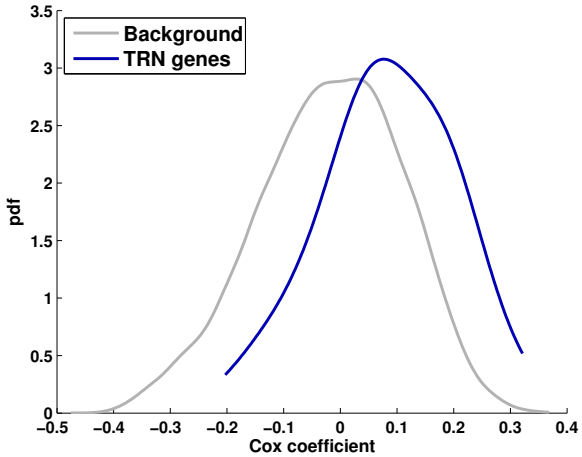
Core transcription factors

- 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).

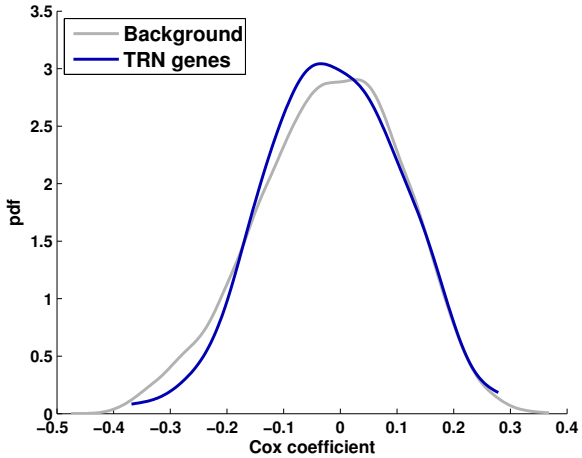
PURDUE UNIVERSITY

Inferring transcriptional controls of Melanoma subtypes

Survival analysis



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



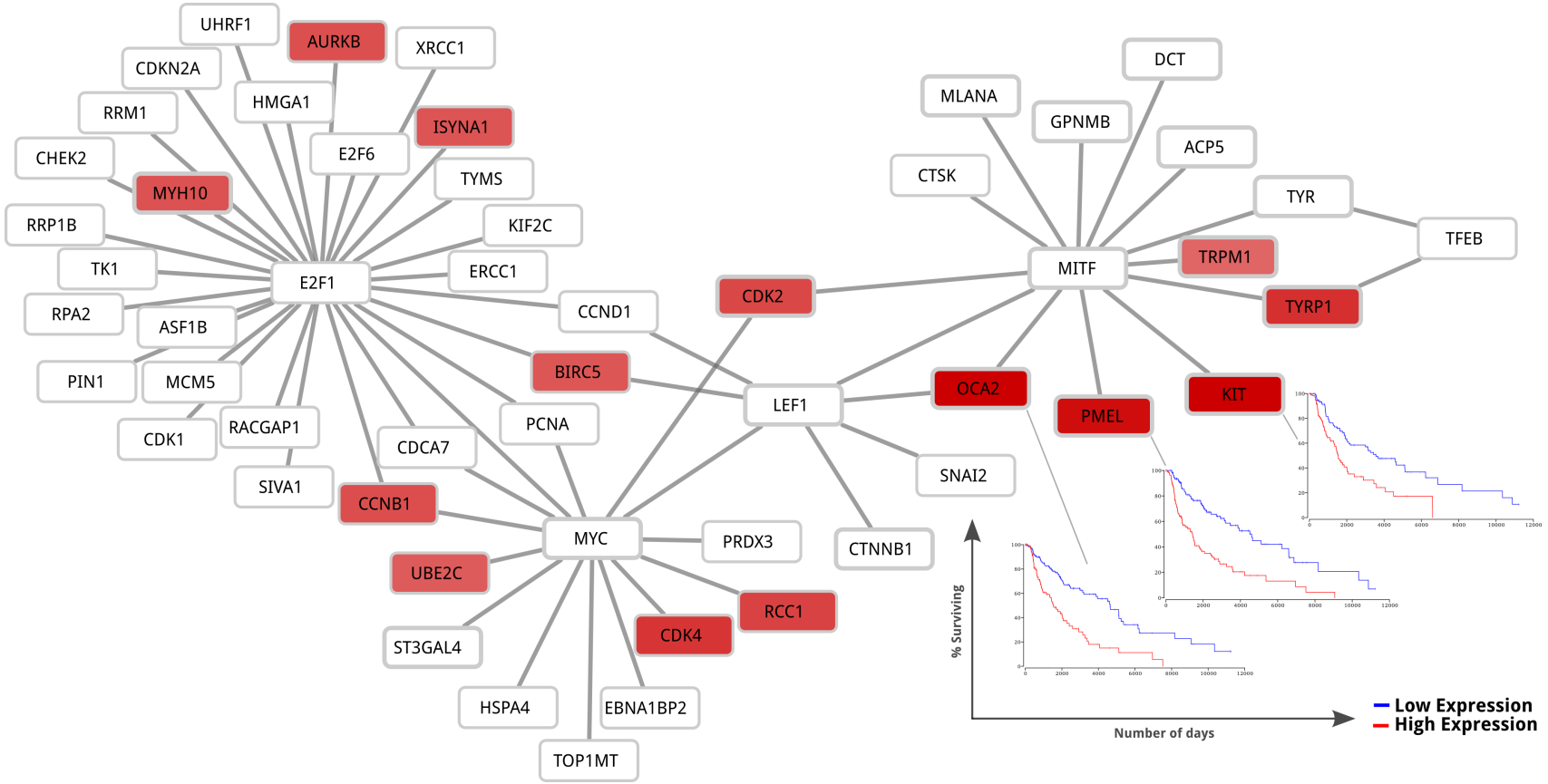
Subtype C: p-value $= 0.31$

- 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

- Use ACTION to infer cell types.
- Use inferred cell types to distinguish true zeros from missing values
  - There is a significant biological signal embedded merely within the sparsity pattern of the single-cell profiles.
- Use SVR to impute missing values.

Forthcoming Results

Use ACTION to infer lineage paths within the functional space of the cells

- Identify stable attractor states within the continuous functional space of cells.
- Trace the most likely transition paths between the states.
- Identify regulatory factors that stimulate these transitions/fate decisions

PURDUE UNIVERSITY

- Use ACTION to identify cell types in human brain, construct cell type-specific region-region gene correlation networks, and compare them with the networks constructed from the resting state fMRI (joint project with Vikram Ravindra, Purdue University)
- Impact of exposing RAW 264.7 macrophage cell line to exosomes from: (i) non-metastatic PEDF expressing A375 cells, and (ii) metastatic A375 melanoma cells (Joint project with Anindita Basu, University of Chicago).

PURDUE UNIVERSITY

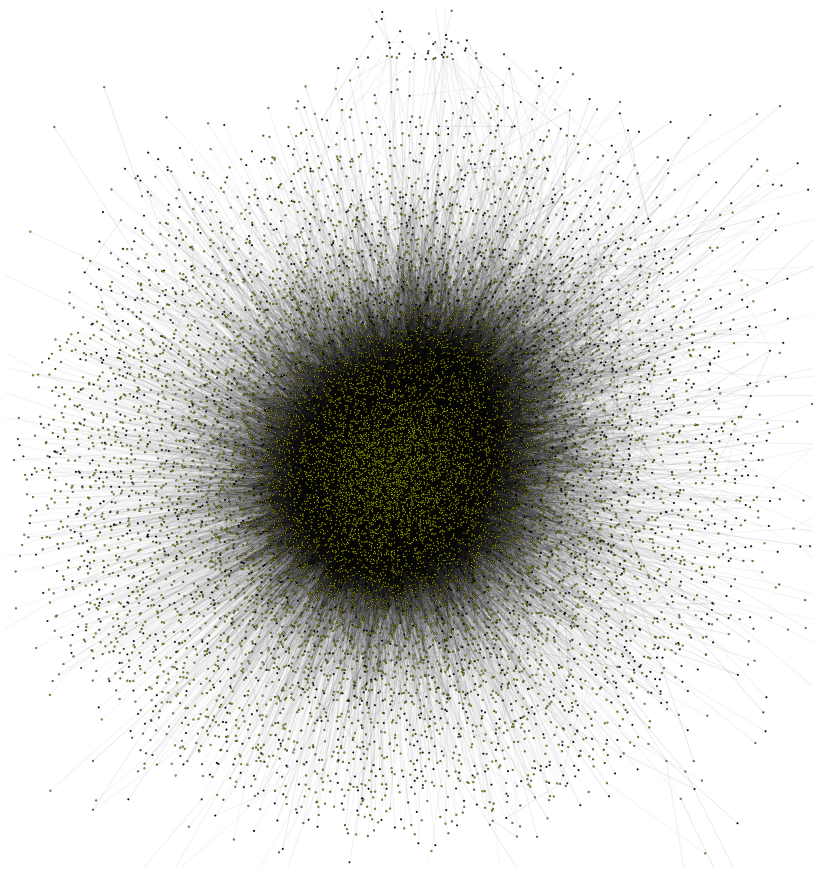
1. Establishing functional identity of cells
2. **Part II: Constructing tissue/cell type-specific networks**
3. Part III: Deconvolving expression profiles of complex tissues

PURDUE UNIVERSITY

Motivation

Global interactome is not context-specific

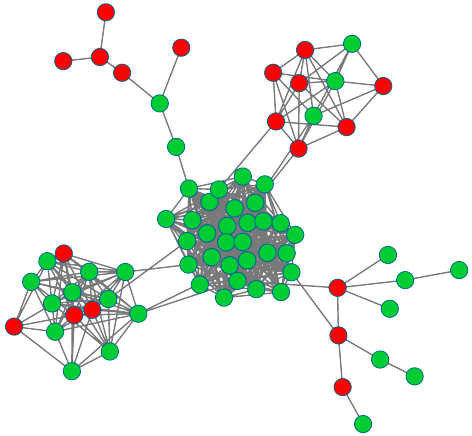
Global human interactome is a superset of all possible physical interactions that can take place in the cell. It does not provide any information as to which one of these interactions do take place in a given tissue/cell-type context.



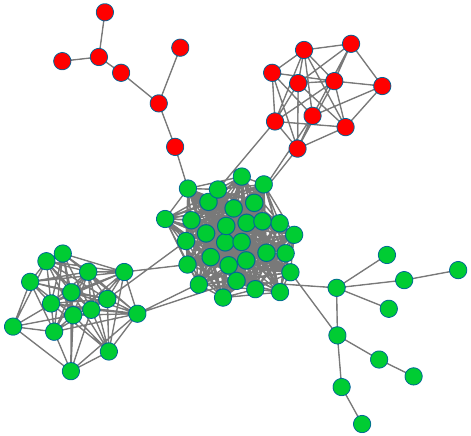
PURDUE UNIVERSITY

Key Question

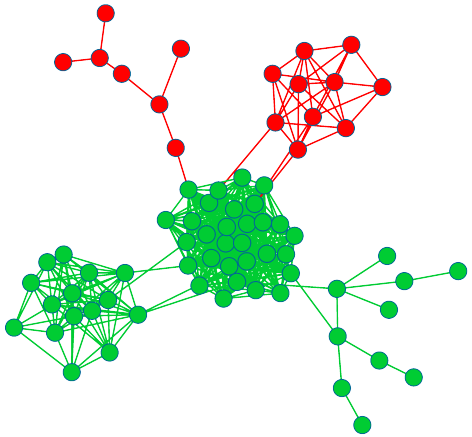
Can we predict which links/edge are active in a given context?



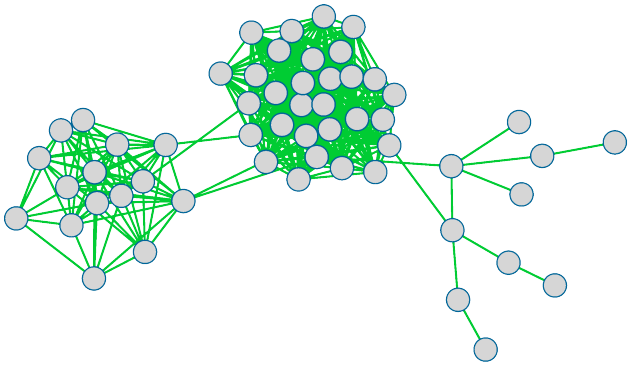
(a) Original



(b) Diffusion

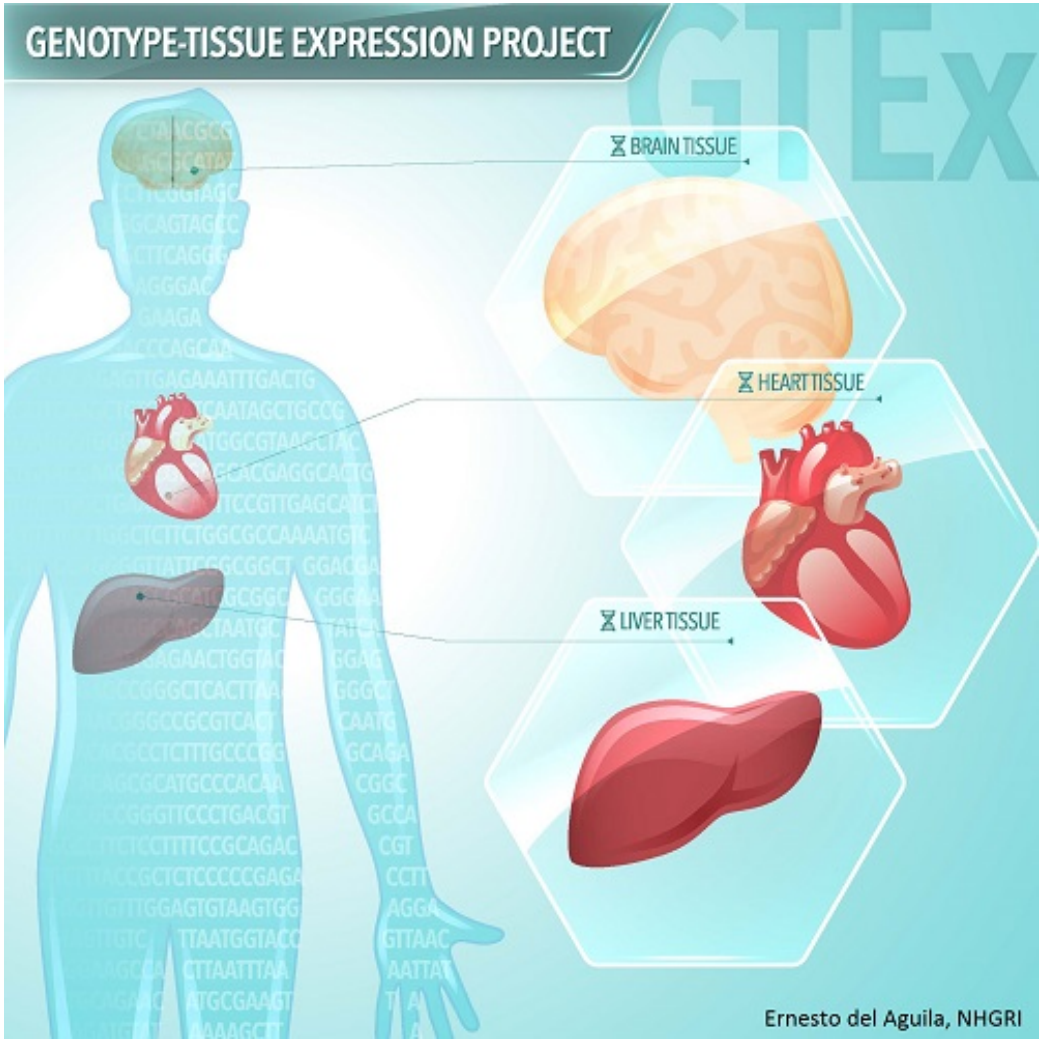


(c) Projection



(d) Pruning

Genotype-Tissue Expression (GTEx) Project



Adopted from: NIH CommonFund

- RNA-Seq dataset v4.0
- 2,916 samples
- 30 different tissues
- Processed each sample individually using UPC/SCAN

PURDUE UNIVERSITY

Activity Propagation (ActPro)

From transcriptional activity to functional activity

Goal: Estimate functional activity of genes

Convex program

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmin}} \left\{ (1-\alpha)\boldsymbol{x}^T \mathbf{L} \boldsymbol{x} + \alpha \parallel \boldsymbol{x} - \boldsymbol{z} \parallel_1 \right\}$$

$$\text{Subject to: } \begin{cases} \mathbf{1}^T \boldsymbol{x} = 1 \\ 0 \leq \boldsymbol{x} \end{cases}$$

- Vector $\boldsymbol{z}$ encodes transcriptional activity of genes, estimated by UPC
- Matrix $\mathbf{L}$ is the *Laplacian* matrix, defined as $\mathbf{A} - \mathbf{D}$, where d_{ii} is the weighted degree of ith vertex in the global interactome.
- Parameter α controls the relative importance of regularization

PURDUE UNIVERSITY

Convex program

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmin}} \left\{ (1-\alpha)\boldsymbol{x}^T \mathbf{L} \boldsymbol{x} + \alpha \parallel \boldsymbol{x} - \boldsymbol{z} \parallel_1 \right\}$$

- The Laplacian operator $\mathbf{L}$ acts on a given function defined over vertices of a graph, such as $\boldsymbol{x}$, and computes the smoothness of $\boldsymbol{x}$ over adjacent vertices.
- We can expand it as $\sum_{i,j} w_{i,j}(x_i - x_j)^2$, which is the accumulated difference of values between adjacent nodes scaled by the weight of the edge connecting them.
- First term is a diffusion kernel. It propagates activity of genes through network links.

Convex program

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmin}} \left\{ (1-\alpha)\boldsymbol{x}^T \mathbf{L} \boldsymbol{x} + \alpha \parallel \boldsymbol{x} - \boldsymbol{z} \parallel_1 \right\}$$

- The second term is a regularizer which penalizes changes or deviations
- We can expand it as $\sum_i |x_i - z_i|$, where x_i and z_i are the (inferred) functional and the transcriptional activity of gene i, respectively.
- It enforces sparsity over the vector of differences between *transcriptional* and *functional* activities.

What do we gain?

Tissue-specific networks have higher power/accuracy in predicting tissue-specific biology and pathobiology

PURDUE UNIVERSITY

Tissue-specific Pathology

Predicting disease-related genes

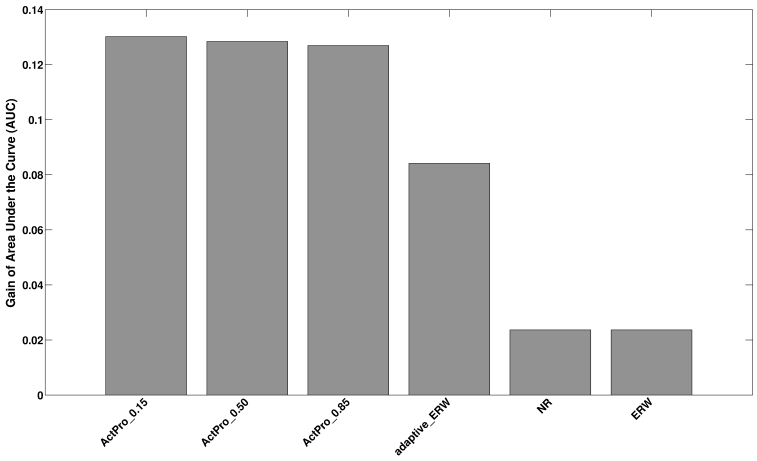
	global	ActPro_0.15	ActPro_0.50	ActPro_0.85	ERW	NR
Alzheimer's disease	**4.12E-3**	6.96E-3	5.98E-3	5.44E-3	5.32E-3	9.60E-2
breast carcinoma	1.83E-3	1.11E-3	8.40E-4	**8.30E-4**	4.09E-3	8.15E-2
chronic lymphocytic leukemia	8.20E-4	7.40E-4	**4.80E-4**	5.10E-4	8.50E-4	2.94E-2
coronary artery disease	3.95E-1	1.58E-1	1.09E-1	1.03E-1	1.33E-1	**1.93E-2**
Crohn's disease	2.56E-2	1.93E-2	1.50E-2	**1.44E-2**	8.54E-2	4.14E-1
metabolic syndrome X	1.11E-2	1.09E-2	**1.07E-2**	1.12E-2	1.02E-1	7.39E-1
Parkinson's disease	1.59E-2	1.25E-2	9.89E-3	**9.50E-3**	1.34E-2	9.62E-2
primary biliary cirrhosis	**7.20E-4**	1.32E-3	3.16E-3	3.40E-3	2.80E-2	6.86E-1
psoriasis	**2.10E-4**	1.10E-3	1.16E-3	9.50E-4	4.67E-3	3.24E-1
rheumatoid arthritis	1.70E-2	**9.28E-3**	1.06E-2	1.10E-2	6.39E-2	3.61E-1
systemic lupus erythematosus	4.98E-2	1.19E-2	7.56E-3	7.22E-3	2.55E-3	**1.60E-4**
type 1 diabetes mellitus	2.64E-2	3.01E-2	**2.38E-2**	2.40E-2	2.64E-1	9.39E-1
type 2 diabetes mellitus	1.57E-3	2.90E-4	2.40E-4	**1.80E-4**	5.60E-4	7.90E-3
vitiligo	**1.17E-3**	2.13E-3	3.04E-3	3.54E-3	1.84E-2	5.69E-1
schizophrenia	3.47E-1	2.13E-1	1.93E-1	1.84E-1	1.40E-1	**4.10E-2**
combined	1.53E-13	1.24E-17	6.62E-19	**3.70E-19**	9.03E-14	2.43E-03

1. Symmetric random-walk as a measure of distance
2. Empirical p-value for each tissue
3. p-value combination using Edgington method

- *ActPro* excels in prioritizing disease-related genes

PURDUE UNIVERSITY

Predicting tissue-specific interactions in known pathways – Average performance



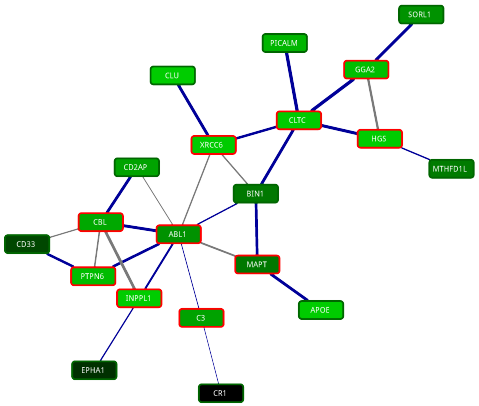
- Edge Set Enrichment Analysis (ESEA).
- Differential correlation score:

$$EdgeScore = \boldsymbol{MI}_{all}(i,j) - \boldsymbol{MI}_{control}(i,j)$$

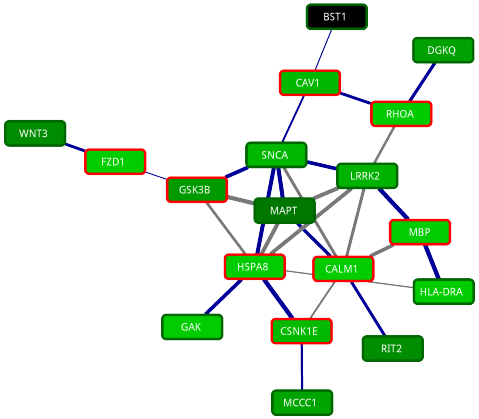
- Gain of Correlation (GoC) edges

Novel Insights

Identifying disease-related pathways in brain



Alzheimer's Disease



Parkinson's Disease

- Prize Collecting Steiner Tree (PCST)

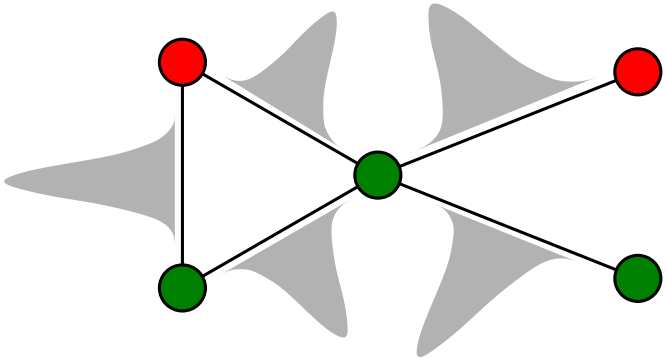
$$\underset{<v,e>\in T}{\operatorname{argmin}} \left\{ \sum_{e} c_e - \lambda \sum_{v} b_v \right\}$$

- $c_e = \frac{1}{w_e}$ and $b_v = \begin{cases} \infty; v \in markers \\ 1; O.W. \end{cases}$
- Red nodes are novel factors

- *ActPro* identifies novel disease-related pathways

PURDUE UNIVERSITY

Goal: Identify driver network perturbations that mediate drug resistance.



- Use single-cell profiles to construct an ensemble of cell type-specific networks, one for before and one for after treatment.
- Combine individual networks within each ensemble to construct a meta-network with a distribution over each edge.
- Identify differential edges that are significantly rewired across conditions.

Key idea: A majority of perturbations do not disable proteins, but they affect individual interactions.

PURDUE UNIVERSITY

Forthcoming Results

Identify intercellular signaling pathways between cells

- Traditional computational approach is to merely look at the expression of known interacting ligands/receptors pairs in adjacent cells.
- There is significant potential for an experimental technology to directly capture these transient interactions.

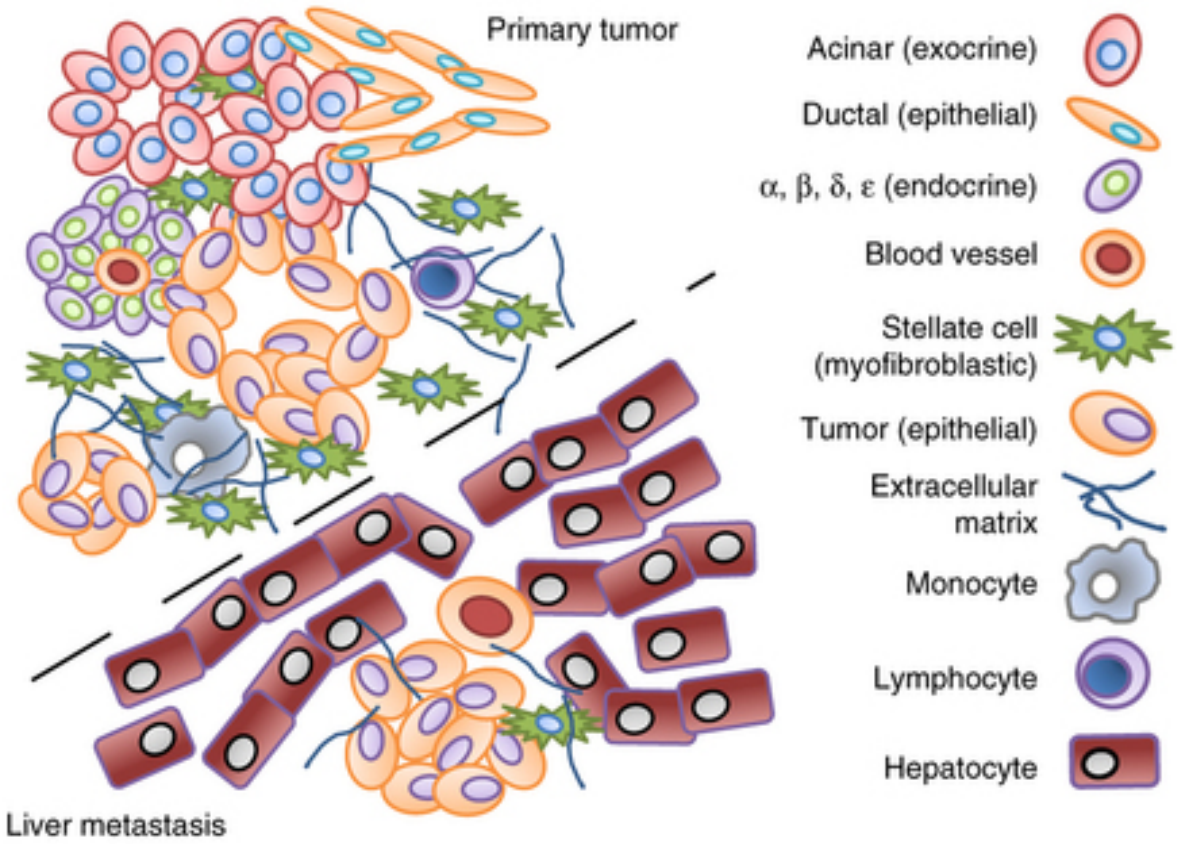
PURDUE UNIVERSITY

1. Establishing functional identity of cells
2. Part II: Constructing tissue/cell type-specific networks
3. Part III: Deconvolving expression profiles of complex tissues

PURDUE UNIVERSITY

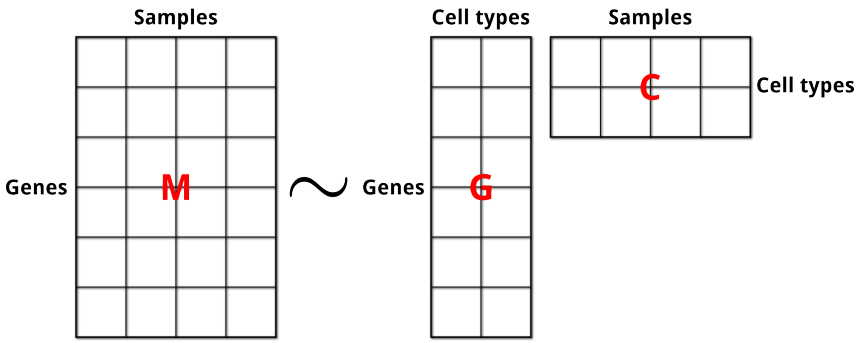
Motivation

Tumor heterogeneity, including its internal diversity, as well as interaction with surrounding microenvironment, is one of the most fundamental determinants of treatment response, drug resistance, and patient relapse.



Adopted from Moffitt *et al.*, 2015

Goal: To decompose a heterogeneous expression profile into its purified cell types



- $\mathbf{M} \in \mathbb{R}^{n \times p}$: Expression matrix of mixed samples
- $\mathbf{G} \in \mathbb{R}^{n \times q}$: Reference signature matrix of primary cell types.
- $\mathbf{C} \in \mathbb{R}^{q \times p}$: Relative proportions of each cell-type in mixture samples.

Given an observed mixture matrix $\mathbf{M}$, find optimal $\mathbf{G}$ and $\mathbf{C}$ that approximate mixture matrix as closely as possible, according to a distance function δ, while satisfying a set of desired constraints:

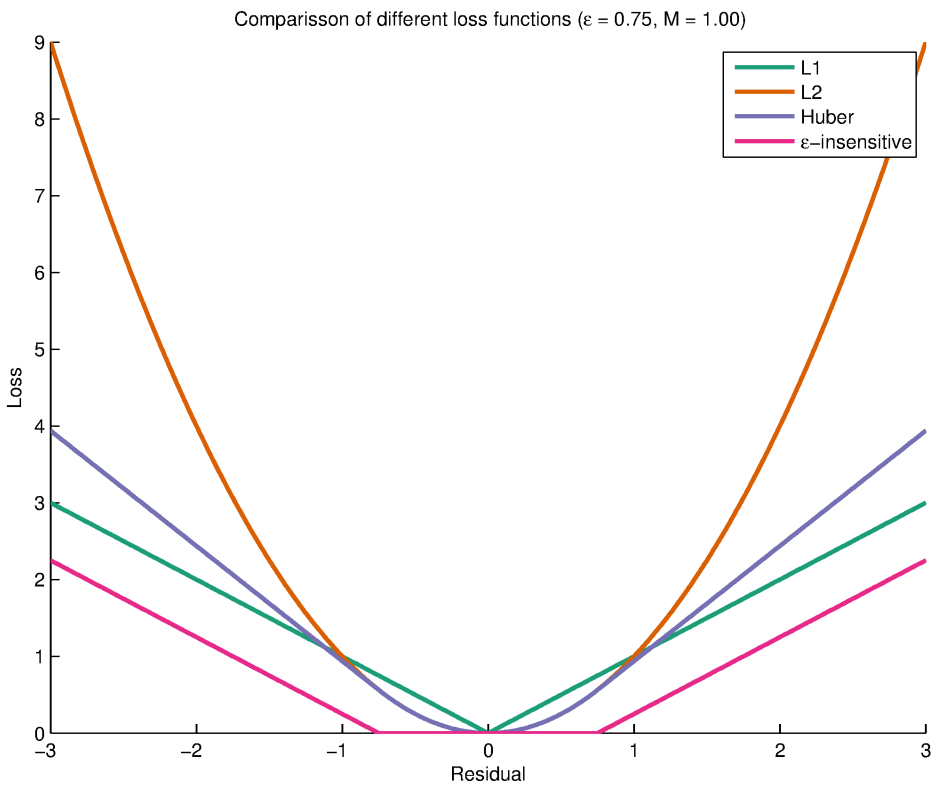
Objective

$$\min_{\mathbf{G},\mathbf{C}\in\text{feasible region}} \delta(\mathbf{GC}, \mathbf{M})$$

Goal: To systematically evaluate different configurations and their performance in gene expression deconvolution

- Different loss functions for evaluating estimation error
- Constraints on solutions
- Preprocessing and data filtering
- Feature selection
- Regularization

Loss functions



1. $\mathcal{L}_2(r_i) = r_i^2 = (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2$
2. $\mathcal{L}_1(r_i) = |r_i| = |y_i - \boldsymbol{w}^T \boldsymbol{x}_i|$
3. $\mathcal{L}_{Huber}^{(M)}(r_i) =$
$$\begin{cases} r_i^2, & \text{if } |r_i| \leq M \\ M(2|r_i| - M), & \text{otherwise} \end{cases}$$
4. $\mathcal{L}_{\epsilon}^{(\epsilon)}(r_i) =$
$$\begin{cases} 0, & \text{if } |r_i| \leq \epsilon \\ |r_i| - \epsilon, & \text{otherwise} \end{cases}$$

Regularizers

- Shrinking/smoothing regression coefficients $\boldsymbol{w}$:

$$\mathcal{R}_2(\boldsymbol{w}) = \| \boldsymbol{w} \|_2^2 = \sum_{i=1}^{k} w_i^2.$$

- Sparsifying solutions :

$$\mathcal{R}_1(\boldsymbol{w}) = \| \boldsymbol{w} \|_1 = \sum_{i=1}^{k} |w_i|.$$

- Ordinary Least Squares (OLS):

$$\begin{aligned}\mathbf{min}_{\boldsymbol{w}}\{\sum_{i=1}^{m}\mathcal{L}_2(r_i)\} &= \mathbf{min}_{\boldsymbol{w}}\{\sum_{i=1}^{m}(y_i-\boldsymbol{w}^T\boldsymbol{x}_i)^2\}\\ &= \mathbf{min}_{\boldsymbol{w}} \parallel y-\mathbf{X}\boldsymbol{w} \parallel_2^2\end{aligned}$$

- Least Absolute Selection and Shrinkage Operator (LASSO) Regression:

$$\mathbf{min}_{\boldsymbol{w}}\{\sum_{i=1}^{m}\mathcal{L}_2(r_i)+\lambda\mathcal{R}_1(\boldsymbol{w})\}$$
$$= \mathbf{min}_{\boldsymbol{w}} \parallel y-\mathbf{X}\boldsymbol{w} \parallel_2^2 +\lambda \parallel \boldsymbol{w} \parallel_1$$

- Support Vector Regression (SVR):

$$\mathbf{min}_{\boldsymbol{w}}\{\sum_{i=1}^{m}\mathcal{L}_\epsilon(y_i-\boldsymbol{w}^T\boldsymbol{x}_i)+\lambda\mathcal{R}_2(\boldsymbol{w})\} \tag{1}$$

Constraints

- Non-negativity (NN)
- Sum-to-one (STO)
- Similar cell quantity (SCQ)

PURDUE UNIVERSITY

Selecting genes to include in basis matrix

Updating **C** is highly over-determined. We try to select genes to simultaneously minimize noise and enhance conditioning of the basis matrix **G**:

- Range filtering
- Marker selection

New criteria: Sum-To-One (STO) violations

- Violating reference gene:

$$\boldsymbol{m}(i) \leq \mathbf{G}_{min}(i); \forall 1 \leq i \leq n$$

- Violating mixture gene:

$$\mathbf{G}_{max}(i) \leq \boldsymbol{m}(i); \forall 1 \leq i \leq n$$

PURDUE UNIVERSITY

Summary of results

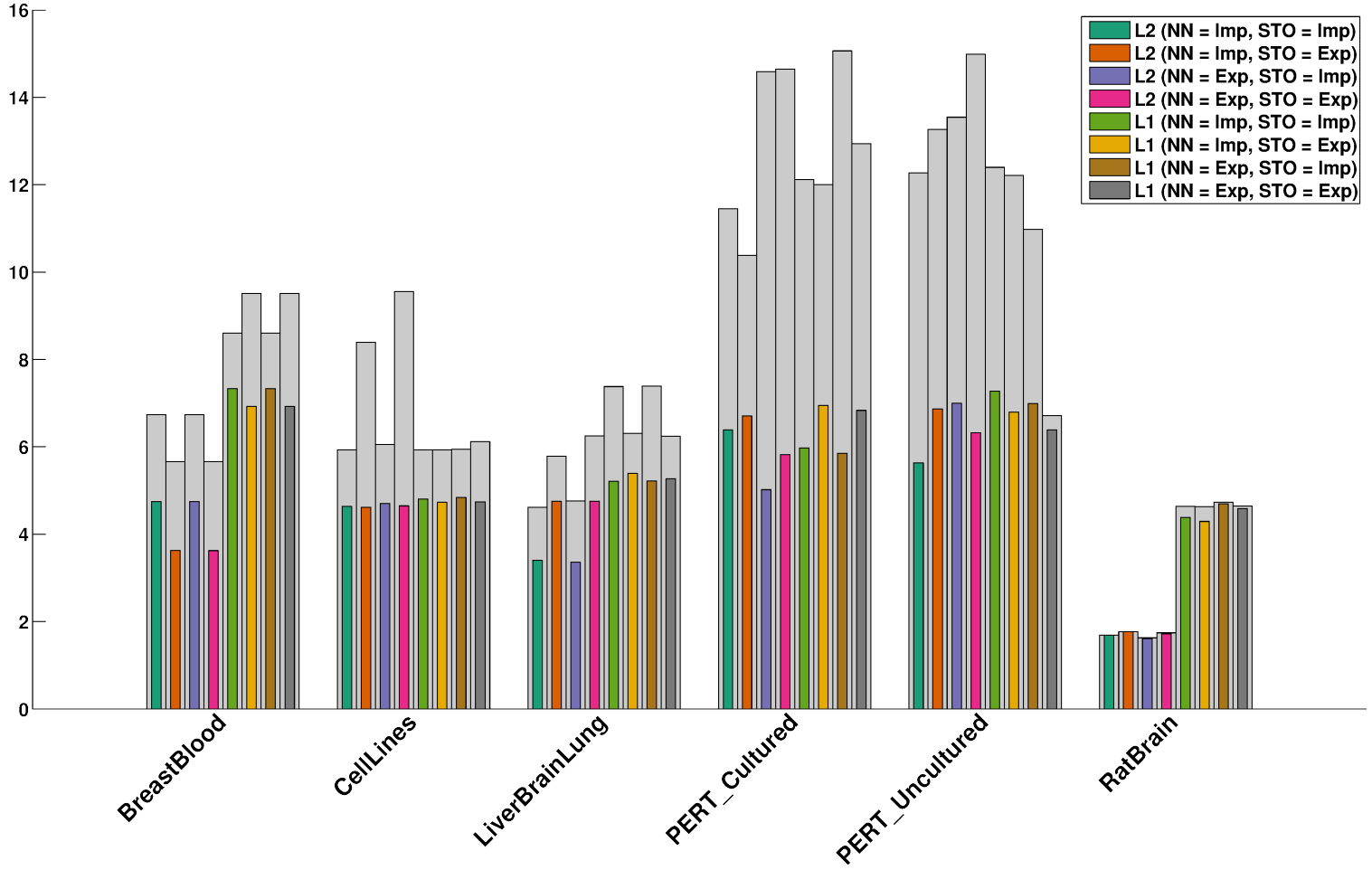
We performed comprehensive, unbiased evaluation for all combinations of these factors on the following datasets:

Dataset	# features	# samples	# references
BreastBlood	54675	9	2
CellLines	54675	12	4
LiverBrainLung	31099	33	3
PERT_Cultured	22215	2	11
PERT_Uncultured	22215	4	11
RatBrain	31099	10	4
Retina	22347	24	2

PURDUE UNIVERSITY

Summary of results

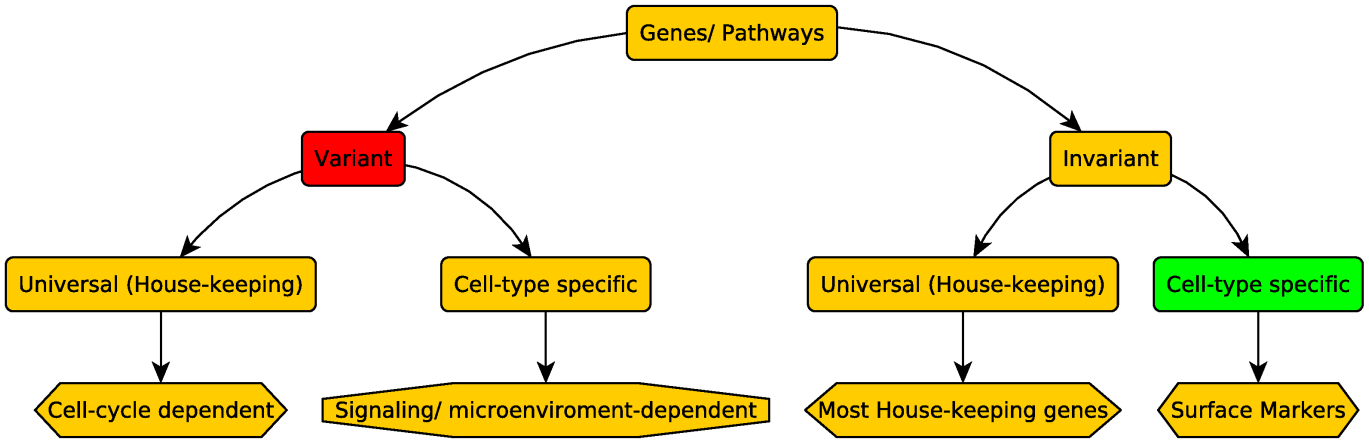
Take home message



- With the right choice of preprocessing and objective function, we can limit error levels in all test datasets

Summary of results
Key observation

Selecting the "right" set of genes for deconvolution has one of the strongest effects on the overall deconvolution performance.



- Selecting genes that are not:
  - Time-dependent, such as cell cycle genes.
  - Microenvironment-dependent factors, such as genes involved in cell signaling pathways.

PURDUE UNIVERSITY

Motivation

- Bulk-tissue RNA-seq profiling is still more cost effective and the preferred choice for large population studies.
- Fresh specimens needed for single-cell profiling is not always available (for example in archived formalin fixed paraffin embedded (FFPE) tissue samples).
- There is a significant body of knowledge in existing databases using bulk-tissue profiling.

PURDUE UNIVERSITY

Thank you!