Deciphering the identity, composition, and interaction of highly refined cell types within complex tissues

Shahin Mohammadi

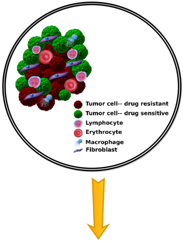
Department of Computer Science
Purdue University

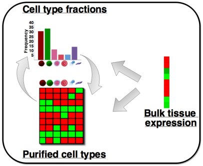
Broad Fellows Program

June 6, 2017

Overview

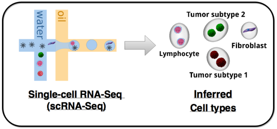
Mixture of different cells





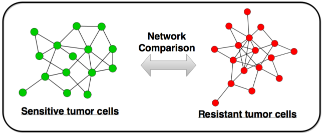
Deconvolving gene expression profiles

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



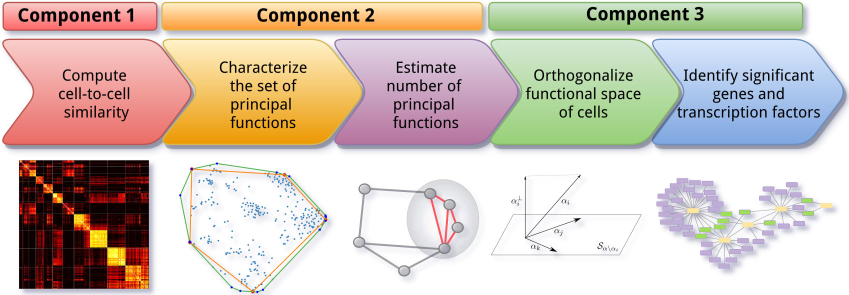
Constructing cell type-specific networks

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

Discovering the functional identity of cell types

1. Discovering functional identity of cell types
2. Constructing tissue/cell type-specific networks
3. Deconvolving expression profile of complex tissues

Overall Workflow



Component 1

Component 2

Component 3

Compute cell-to-cell similarity

Characterize the set of principal functions

Estimate number of principal functions

Orthogonalize functional space of cells

Identify significant genes and transcription factors

Component 1

Motivation

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

Supporting evidence

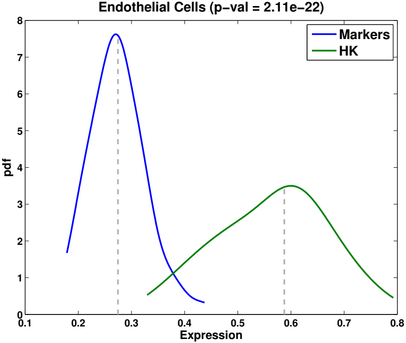


Figure : Endothelial Cells

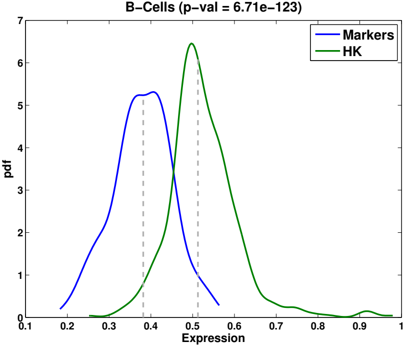


Figure : B-Cells

Component 1

Supporting evidence

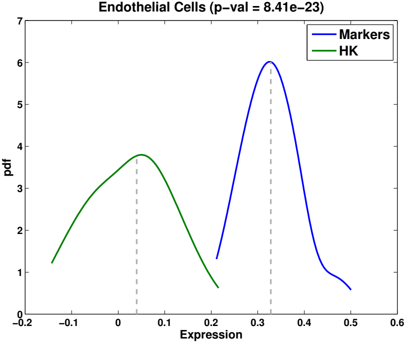


Figure : Endothelial Cells

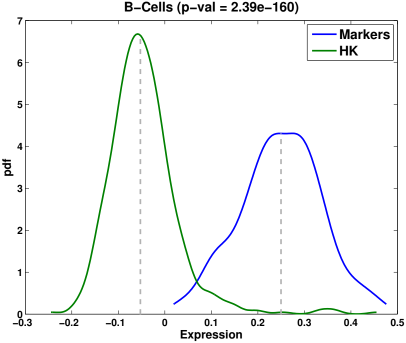
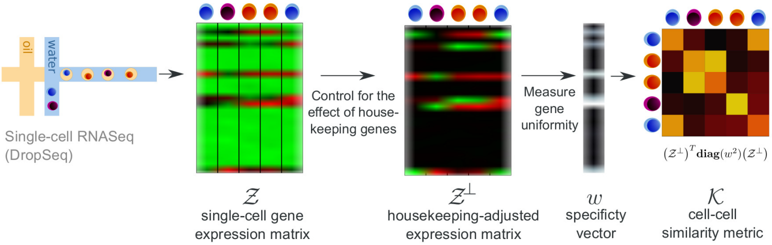


Figure : B-Cells

Component 1

Overall flow – cell similarity kernel


oil
water
Single-cell RNASeq (DropSeq)
Control for the effect of house-keeping genes
Measure gene uniformity
$(\mathcal{Z}^{\perp})^T \mathbf{diag}(w^2)(\mathcal{Z}^{\perp})$
$\mathcal{Z}$
single-cell gene expression matrix
$\mathcal{Z}^{\perp}$
housekeeping-adjusted expression matrix
w
specificty vector
$\mathcal{K}$
cell-cell similarity metric

- The main steps involved in identifying similarity between cells

Component 1

Reducing the noise contributed by highly expressed but uninformative genes

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

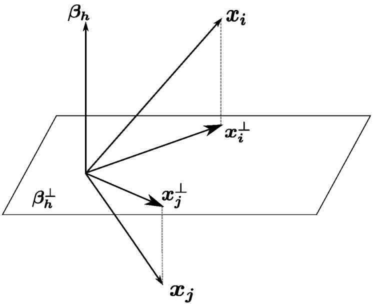
Component 1

Reducing the noise contributed by highly expressed but uninformative genes

Goal: Remove the effect of common subspace

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

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



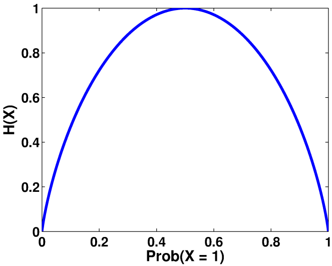
$\boldsymbol{\beta}_h$
$\boldsymbol{x}_i$
$\boldsymbol{x}_i^{\perp}$
$\boldsymbol{\beta}_h^{\perp}$
$\boldsymbol{x}_j^{\perp}$
$\boldsymbol{x}_j$

Similar in nature to the partial Pearson's correlation

Component 1

Enhancing the signal from preferentially-expressed genes

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 observing a gene is 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 formulation have been previously used for marker detection.

Component 1

Putting pieces back together

ACTION-adjusted cell signatures

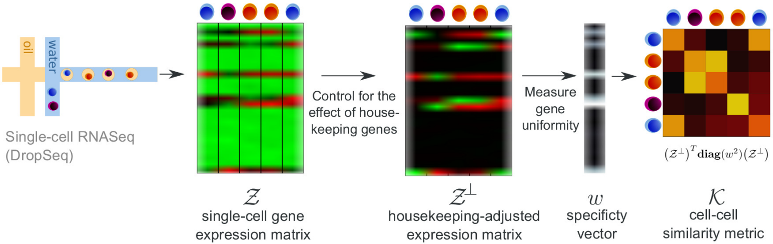
$$\mathbf{Y} = \mathbf{diag}(\mathbf{w})\mathcal{Z}^{\perp}$$

ACTION metric (kernel)

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

Component 1

Cell similarity kernel – revisited

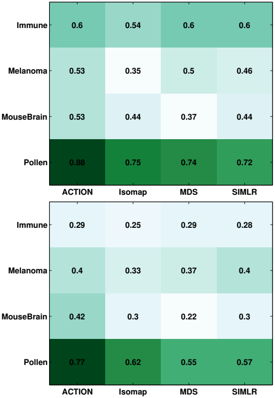


- Now we have computed the ACTION kernel

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

Performance of ACTION Kernel

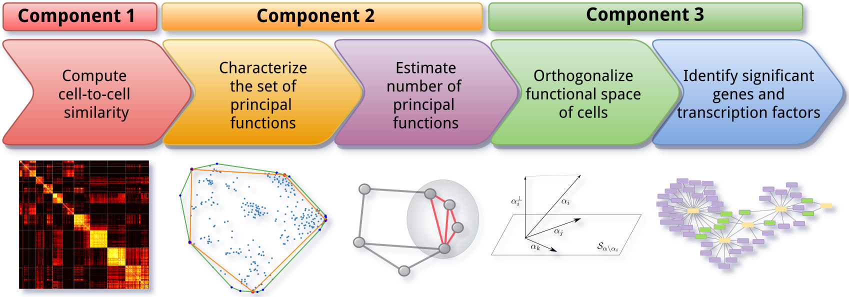


- Benchmarks:
  - SIMLR: Specifically designed for single-cell data
  - IsoMap,MDS: General purpose dimension reduction
- Tested a range of parameters (5:5:50). Reported best case.
- Ties:
  - *Immune* (NMI: ACTION/MDS/SMLR, ARI: ACTON/MDS)
  - *Melanoma* (ARI: ACTION/SIML)
- In all other cases, *ACTION*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

- *ACTION* metric performs equally good or better than other methods

Overall Workflow

Component 2



Component 1

Component 2

Component 3

Compute cell-to-cell similarity

Characterize the set of principal functions

Estimate number of principal functions

Orthogonalize functional space of cells

Identify significant genes and transcription factors

General framework

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

Various algorithms can be cast using this formulation

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

Component 2

Motivations

There are fundamental problems with K-means/medoids:

- They use hard assignment, whereas many cell types are believed to form a continuum.
- They are sensitive to initialization.
- They are dependent on k.

Component 2

Convex Nonnegative Matrix Factorization (NMF)

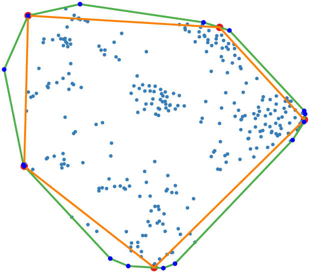
Convex NMF

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

- 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.

Component 2

Convex NMF– Geometric interpretation



<span style="color:red">Geometry of functional space:</span> 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

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

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

Component 2

Archetypal Analysis (AA)

- It further relaxes matrix **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 → it is also dependent on the initialization
    - To address this issue, we use the solution of convex NMF for initializing A.A.
- In essence, it allows local adjustment of the Convex NMF solution.
- A variant of block-coordinate descent for optimization.

Component 2

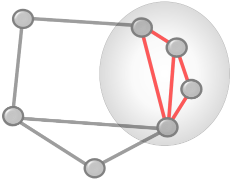
Finding the number of archetypes (k)

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

- Idea is simple: keep adding archetypes till we sense ”oversampling.”
- Oversampling happens when we start adding archetypes that are ”too close” to each other.
- Each archetype is a cell → we can compute their similarity of using the ACTION metric.

Component 2

Statistical significance of oversampling

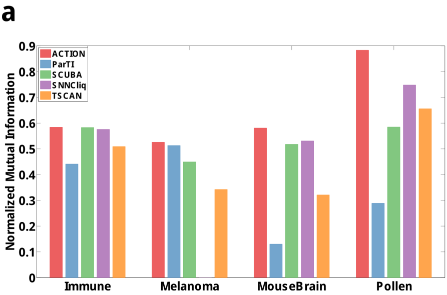


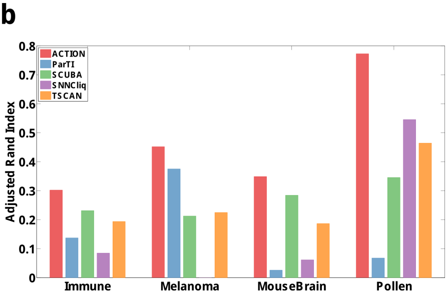
- We build and threshold an archetype-archetype similarity graph.
- For each connected component in this graph, we assess its statistical significance using ER model.
- Probability that there exists in **G** a subgraph of density $\delta(Z)$ and size at least $|Z|$:

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

Component 2

Test 1: Identifying cell types using closest archetype





- ACTION excels in identifying underlying cell types in all cases

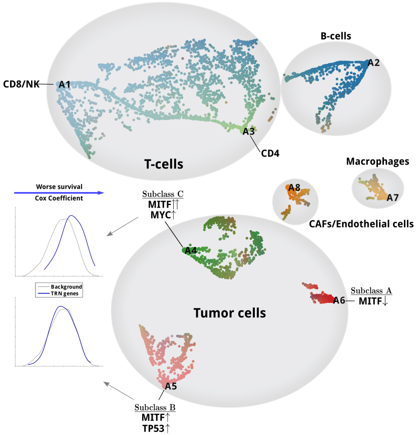
Component 2

Visualizing the functional space

- 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

Continuous view

Case study in the Melanoma dataset

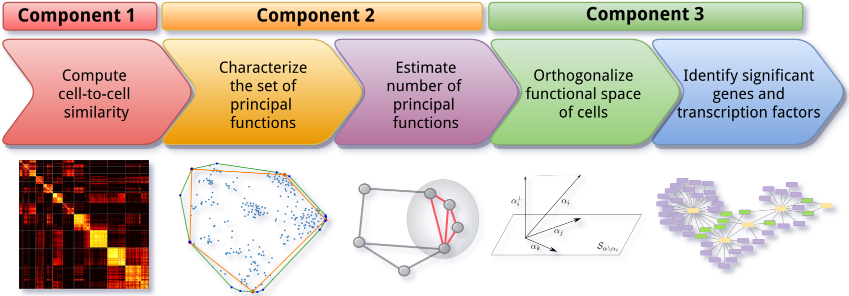


- T-cells reside in a continuum of states (Thogerson *et al.*).
- Tumor cells form compact groups.
- Two subclasses of MITF-associated tumors significantly differ in terms of their survival.

- ACTION sheds light on the underlying topology of cell types

Overall Workflow

Component 3



Component 1

Component 2

Component 3

Compute cell-to-cell similarity

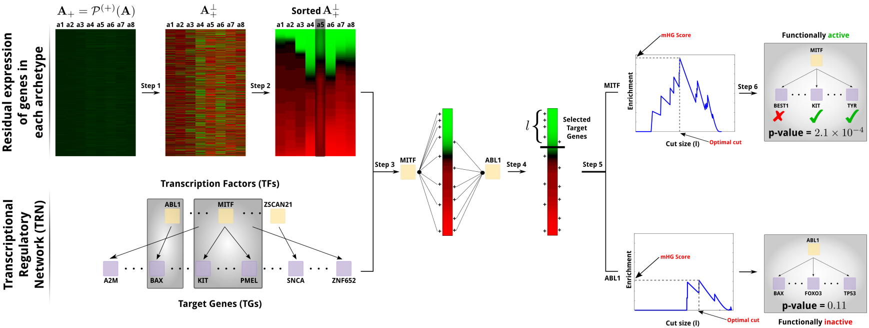
Characterize the set of principal functions

Estimate number of principal functions

Orthogonalize functional space of cells

Identify significant genes and transcription factors

Component 3
Constructing TRN



Component 3

Constructing TRN

Continued Goal: Identifying key regulatory elements that drive each cell type

1. Archetype Orthogonalization (→ Only over positive projection)

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

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.

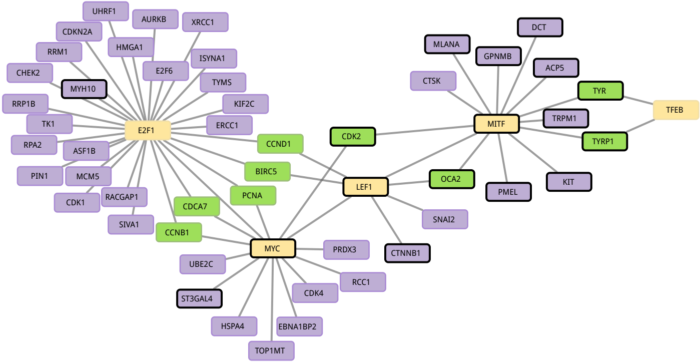
Dissecting transcriptional controls of Melanoma subtypes

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.

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 $\leq 10^{-3}$)

Case study in MITF⇈/MYC↑ subtype

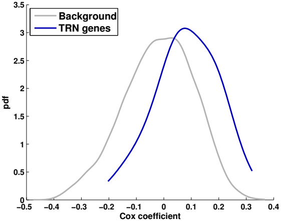
Core transcription factors

- MITF is one of the most well-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).

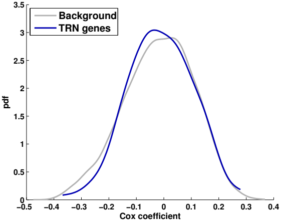
PURDUE
UNIVERSITY

Dissecting transcriptional controls of Melanoma subtypes

Survival analysis



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

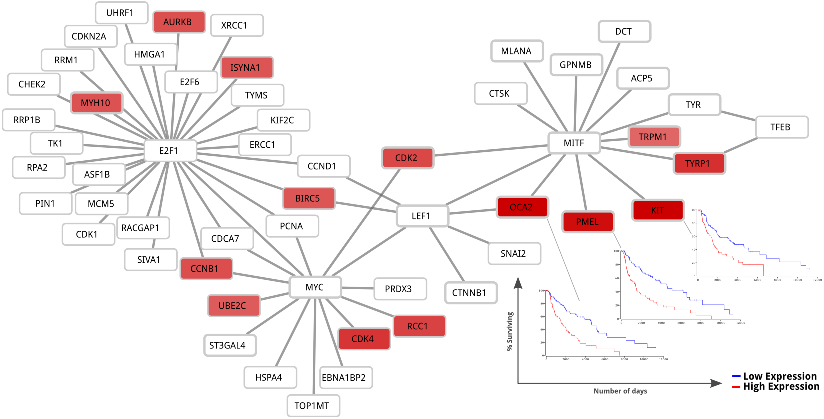


Subtype C: p-value $= 0.31$

- 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

Case study in MITF⇈/MYC↑ subtype

Survival analysis revisited – Kaplan-Meier plots



Contributions
Recap

1. Developed 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. Characterized the functional identity of cells
    - Under the pure cell assumption, this metric induces a convex topology that embeds functional identity of cells
3. Utilized functional identity of cells to identify both discrete cell types and continuous cell states
4. Identified 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.

Mystery of inflated zeros: a curse or a blessing?

- 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.

Future Directions

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

Future Directions

Applications

- 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).

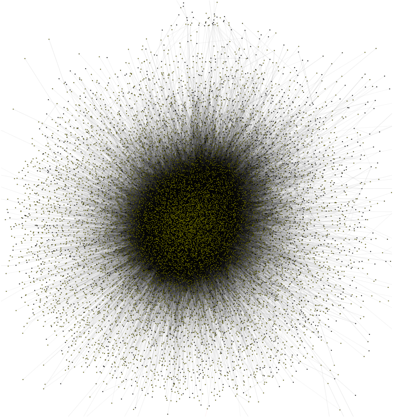
Constructing tissue-specific interactome

1. Discovering functional identity of cell types
2. **Constructing tissue/cell type-specific networks**
3. Deconvolving expression profile of complex tissues

Motivation

Global interactome is not context-specific

Global human interactome is a superset of all possible physical interactions that can take places 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.

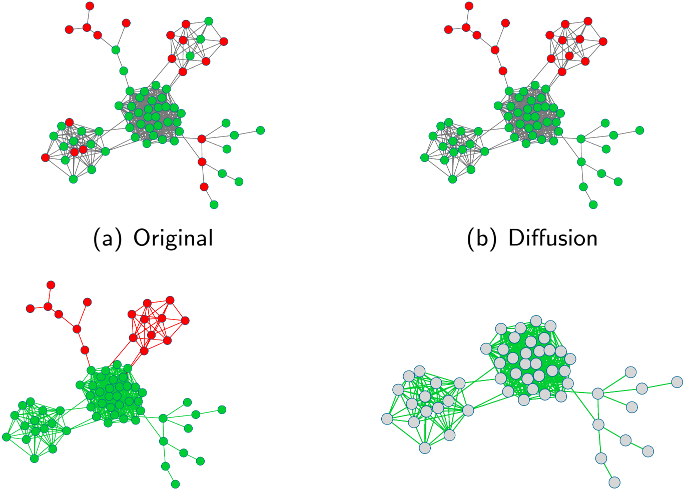


Big Question

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

Roadmap

Toy Example



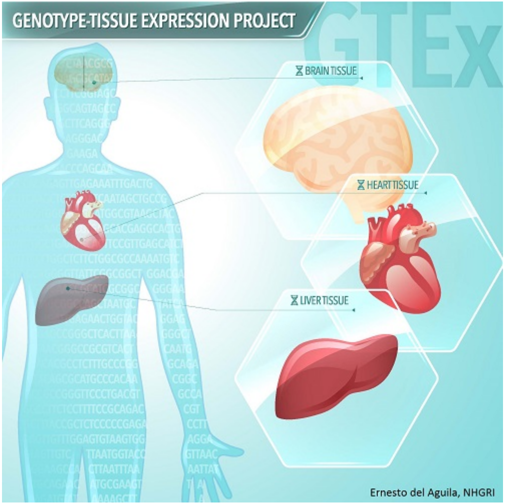
(a) Original

(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

Activity Propagation (ActPro)

From transcriptional activity to functional activity

Goal: Estimate functional activity of genes

Convex program

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

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

- Vector $\mathbf{z}$ 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 i^{th} vertex in the global interactome.
- Parameter α controls the relative importance of regularization

PURDUE UNIVERSITY

Activity Propagation (ActPro)

Interpretation – Loss function

Convex program

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

- The Laplacian operator **L** acts on a given function defined over vertices of a graph, such as **x**, and computes the smoothness of **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.

Activity Propagation (ActPro)

Interpretation – Regularizer

Convex program

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

- 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.

Results

What do we gain?

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

Tissue-specific Pathology

Predicting disease-related genes

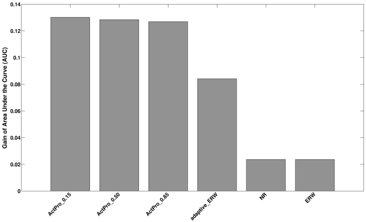
	global	**ActPro_0.15**	**ActPro_0.50**	**ActPro_0.85**	**ERW**	**NR**
Alzheimer's disease	**4.12E-3**	6.96E-3	5.98E-3	5.44E-3	5.32E-3	9.60E-2
breast carcinoma	1.83E-3	1.11E-3	8.40E-4	**8.30E-4**	4.09E-3	8.15E-2
chronic lymphocytic leukemia	8.20E-4	7.40E-4	**4.80E-4**	5.10E-4	8.50E-4	2.94E-2
coronary 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

Tissue-specific Biology

Predicting tissue-specific interactions in known pathways – Average performance

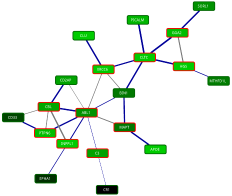


- Edge Set Enrichment Analysis (ESEA).
- Differential correlation score:

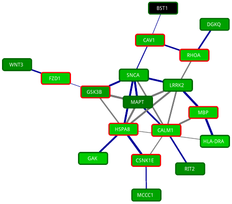
$$EdgeScore = \mathbf{MI}_{all}(i,j) - \mathbf{MI}_{control}(i,j)$$

- Gain of Correlation (GoC) edges

Novel Biology
Identifying disease-related pathways in brain



Alzheimer's Disease



Parkinson's Disease

- Prize Collecting Steiner Tree (PCST)

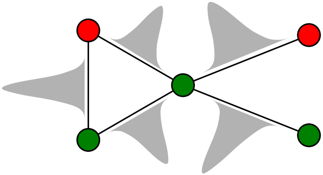
$$\underset{<v,e>\in T}{\text{argmin}} \left\{ \sum_{e} c_e - \lambda \sum_{v} b_v \right\}$$

- $c_e = \frac{1}{w_e}$ 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

Future direction
Differential network analysis

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.

Future Directions

Uncover intercellular signaling pathways between cells

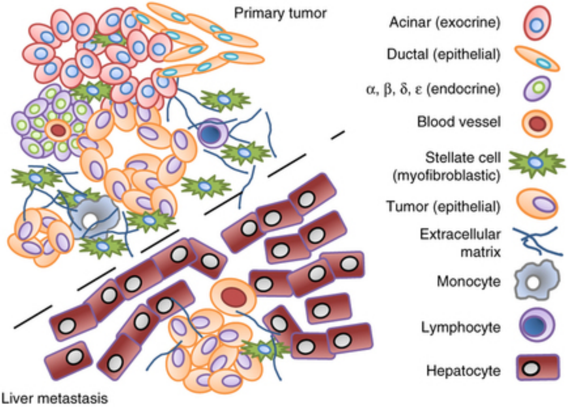
- Traditional computational approach is to merely look at the expression of known interacting ligands/receptors pairs in adjacent cells.
- There is a hope for an experimental technology to directly capture these transient interactions.

Constructing tissue-specific interactome

1. Discovering functional identity of cell types
2. Constructing tissue/cell type-specific networks
3. **Deconvolving expression profile of complex tissues**

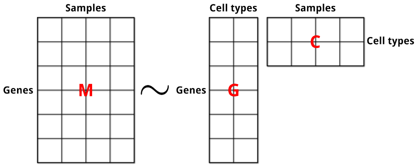
Motivation

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.

Deconvolution: Formal Definition

Problem definition

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

Objective

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

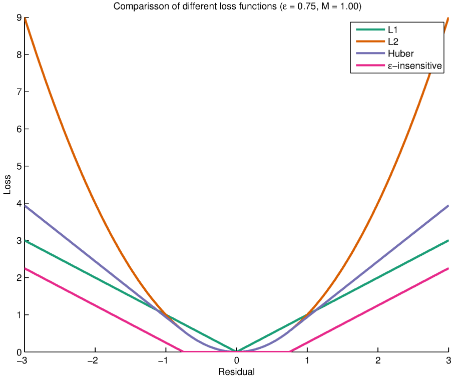
Deconvolution: Formal Definition

Scope of this study

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. $\mathcal{L}_2(r_i) = r_i^2 = (y_i - \mathbf{w}^T\mathbf{x}_i)^2$
2. $\mathcal{L}_1(r_i) = |r_i| = |y_i - \mathbf{w}^T\mathbf{x}_i|$
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 **w**:

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

- Sparsifying solutions :

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

Examples

Some of existing combinations

- Ordinary Least Squares (OLS):

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

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

$$\mathbf{min}_{\mathbf{w}}\{\sum_{i=1}^{m}\mathcal{L}_2(r_i)+\lambda\mathcal{R}_1(\mathbf{w})\}$$

$$= \mathbf{min}_{\mathbf{w}} \parallel y-\mathbf{Xw} \parallel_2^2 + \lambda \parallel \mathbf{w} \parallel_1$$

- Support Vector Regression (SVR):

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

Constraints

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

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:

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

- Violating mixture gene:

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

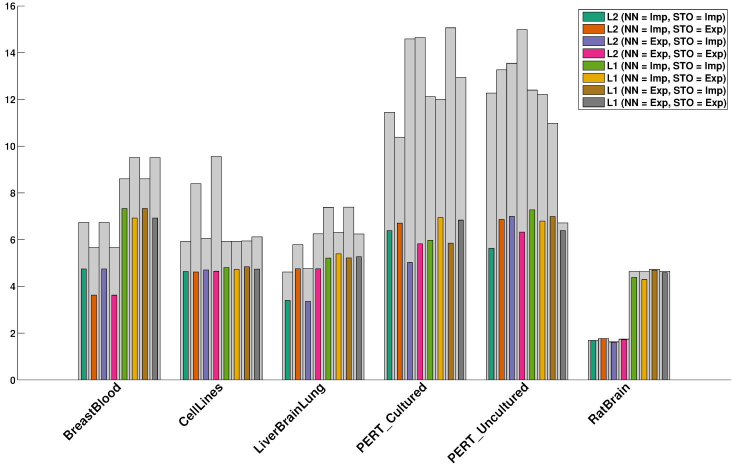
Summary of results

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

Summary of results

Take home message

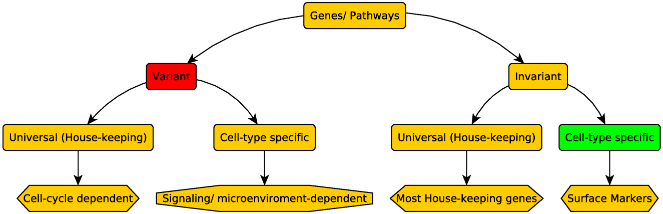


- With the right choice of preprocessing and objective function, we can limit error levels in all test datasets

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.

Future directions

Use single-cell profiles as basis for deconvolution

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.

Future directions

Use single-cell profiles as basis for deconvolution– Sketch of the method

Joint project with Yu Li @MIT.

- Bulk and single-cell each have their own unique signatures in their measurements.
- Unlike *BSEQ-sc*, we first aim to develop a deep-learning machine to map expression profiles from single-cell space to their corresponding bulk, purified projection.
- We use projected profiles as an initial estimate of **G** and jointly estimate **C** and update **G**.

Future directions

Hierarchical deconvolution

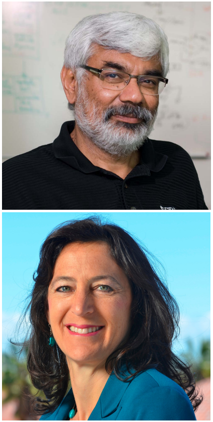
Challenge: Emergence of complex underlying structure as we increase the total number cell types.

- Infer a hierarchy for cell types, first.
- Use parent(s) of each node as a prior.
- Orthogonalize cell types w.r.t. the prior of all ancestor cells.
- Deconvolve each layer with the residual subspace.

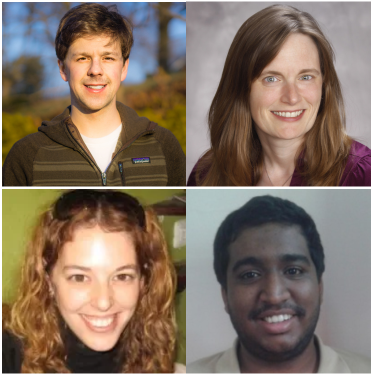
Acknowledgement

Advisors

Visiting labs



Colleagues



Relevant Publications I

- Z. R. M. Kashani, H. Ahrabian, E. Elahi, A. Nowzari-Dalini, E. S. Ansari, S. Asadi, **S. Mohammadi**, F. Schreiber, and A. Masoudi-Nejad (2009) Kavosh: a new algorithm for finding network motifs.
  BMC bioinformatics, **10**(318).

- **S. Mohammadi** and A. Grama (2011) Biological Network Alignment.
  In M. Koyutürk, S. Subramaniam, and A. Grama (eds.), *Functional Coherence of Molecular Networks in Bioinformatics*, Springer, pages 97–136.

- G. Kollias, **S. Mohammadi**, and A. Grama (2011) Network Similarity Decomposition (NSD): A Fast and Scalable Approach to Network Alignment.
  IEEE Transactions on Knowledge and Data Engineering (TKDE), **24**(12):2232–2243.

Relevant Publications II

- **S. Mohammadi**, G. Kollias, and A. Grama (2012) Role of Synthetic Genetic Interactions in Understanding Functional Interactions Among Pathways.
  In *Pacific Symposium on Biocomputing (PSB)*.
- G. Kollias, M. Sathe, **S. Mohammadi**, and A. Grama (2013) A fast approach to global alignment of protein-protein interaction networks.
  BMC research notes, **6**(1):35.
- **S. Mohammadi**, S. Subramaniam, and A. Grama (2013) Inferring the Effective TOR-Dependent Network: A Computational Study in Yeast.
  BMC Systems Biology, **7**(1):84.
- **S. Mohammadi**, B. Saberidokht, S. Subramaniam, and A. Grama (2015) Scope and limitations of baker's yeast as a model organism for studying human tissue-specific pathways.
  BMC Systems Biology, **96**(9).

Relevant Publications III

- **S. Mohammadi**, D. F. Gleich, T. G. Kolda, and A. Grama (2016) Triangular Alignment (TAME): A Tensor-based Approach for Higher-order Network Alignment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*.

- **S. Mohammadi** and A. Grama (2016) A convex optimization approach for identification of human tissue-specific interactomes. *Bioinformatics*, **32**(12):i243–i252.

- **S. Mohammadi** and A. Grama (2016) De novo identification of cell type hierarchy with application to compound marker detection. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics - BCB'16*. pages 251–261.

Relevant Publications IV

- **S. Mohammadi**, N. Zuckerman, A. Goldsmith, and A. Grama (2016) A Comprehensive Survey of Deconvolution Methods for Separating Cell types in Complex Tissues.
  IEEE, Special Issue on Principles and Applications of Science of Information.
- **S. Mohammadi**, S. Kylasa, G. Kollias, and A. Grama (2016) A Context-specific Recommendation System for Predicting Similar PubMed Articles.
  In *IEEE ICDM Workshop on Semantics-Enabled Recommender System*.
- **S. Mohammadi**, V. Ravindra, D. Gleich, and A. Grama (2016) Take ACTION to identify high resolution cell types and associated transcriptional pathways.
  Tech report on *bioRxiv*.

Questions?

where?
how?
discover
questions
why
asking questions
challenge
who?
QUESTIONS
clues
what
where
investigation
?
Knowing
clues
how
why
ask
Knowing