scDECO: A Novel Statistical Framework to Identify Differential Co-Expression Gene Combinations Systematically Using Single-Cell RNA Sequencing Data

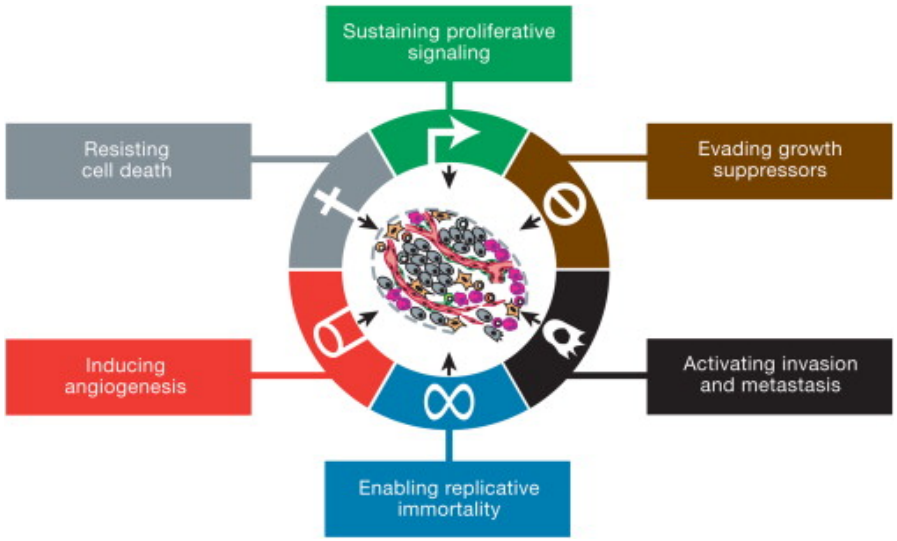
Yen-Yi Ho[1], Hexin Chen[1], Chun-Liang Chen[2], Evan Keller[3]

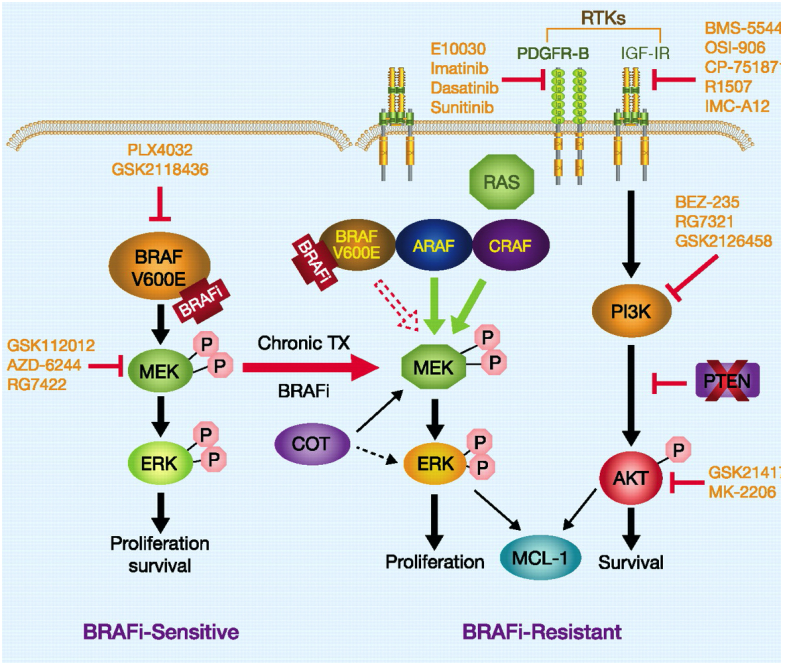
[1] University of South Carolina,
[2]UT Health San Antonio,
[3]University of Michigan

September 14, 2022

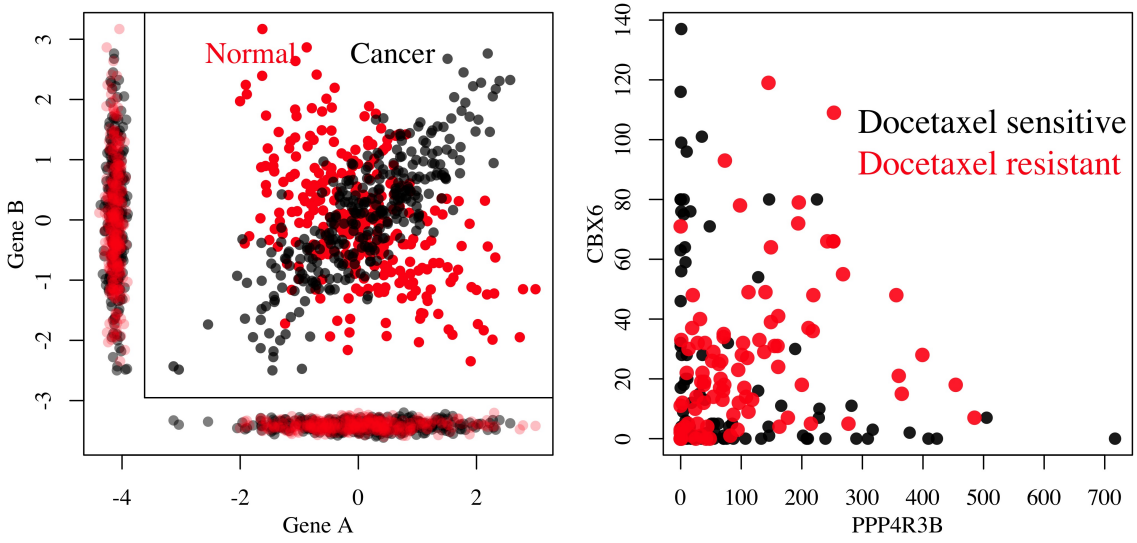
Motivation: Cancer Pathways

- Cancer-specific pathway activities that enable tumor growth and metastatic dissemination
- Alternative signalling pathways in response to anti-cancer treatments



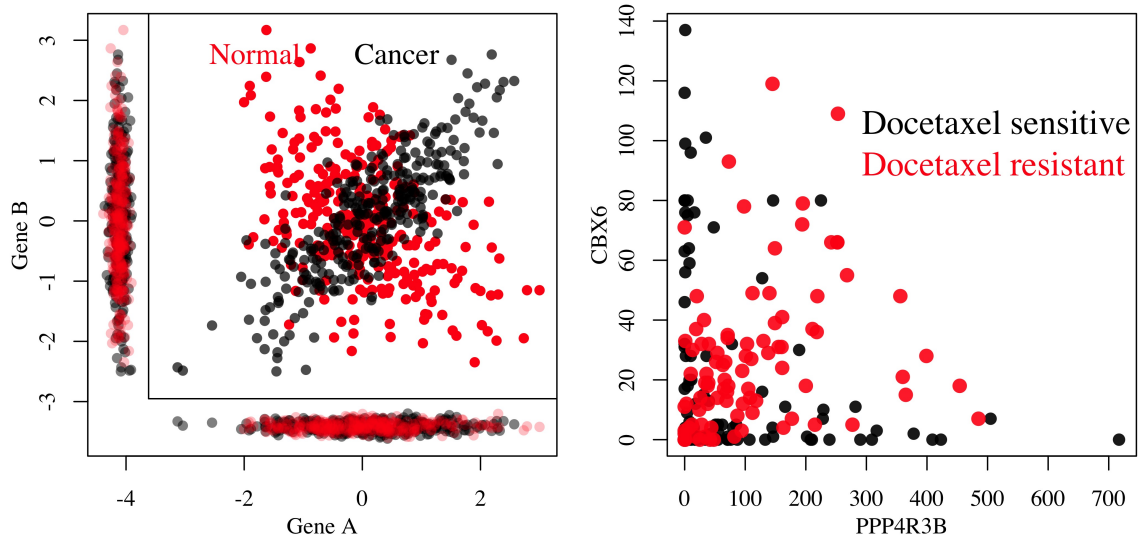


Differential Co-Expression (DC)



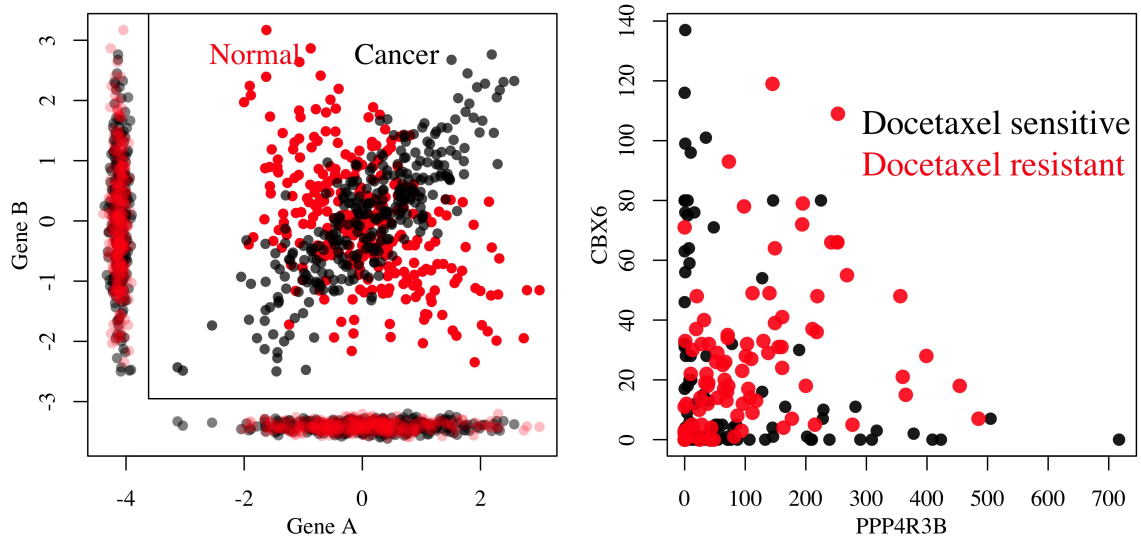
- Differential expression (DE) analysis is likely to miss meaningful genetic interactions.

Differential Co-Expression (DC)



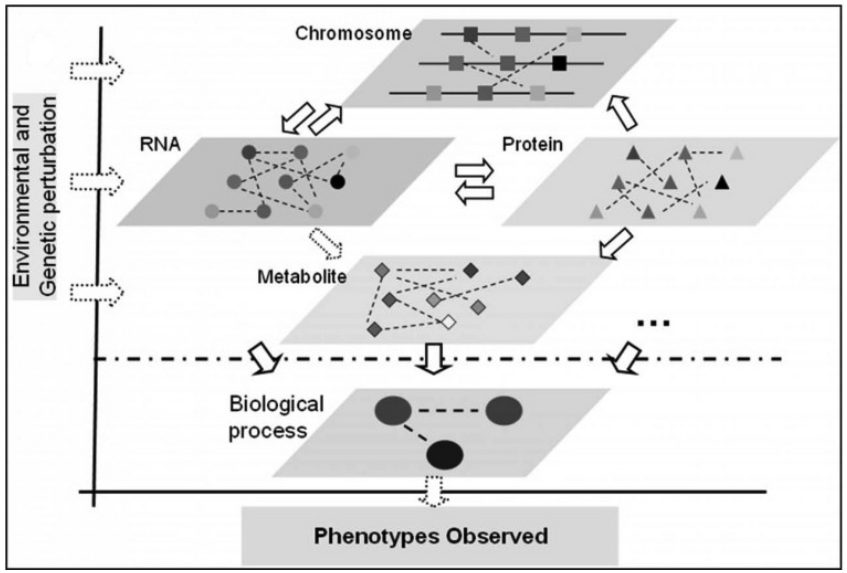
- Differential expression (DE) analysis is likely to miss meaningful genetic interactions.
- Differential co-expression (DC) analysis addresses this issue by evaluating whether there are correlated changes between pairs of genes across different modulating conditions.

Differential Co-Expression (DC)



- Differential expression (DE) analysis is likely to miss meaningful genetic interactions.
- Differential co-expression (DC) analysis addresses this issue by evaluating whether there are correlated changes between pairs of genes across different modulating conditions.
- scRNAseq data are count-based and exhibit characteristics such as overdispersion and zero-inflation

Our Vision & Long-term Goal



To develop tools for identifying alterations of interactions within/between various molecular layers in cancer.

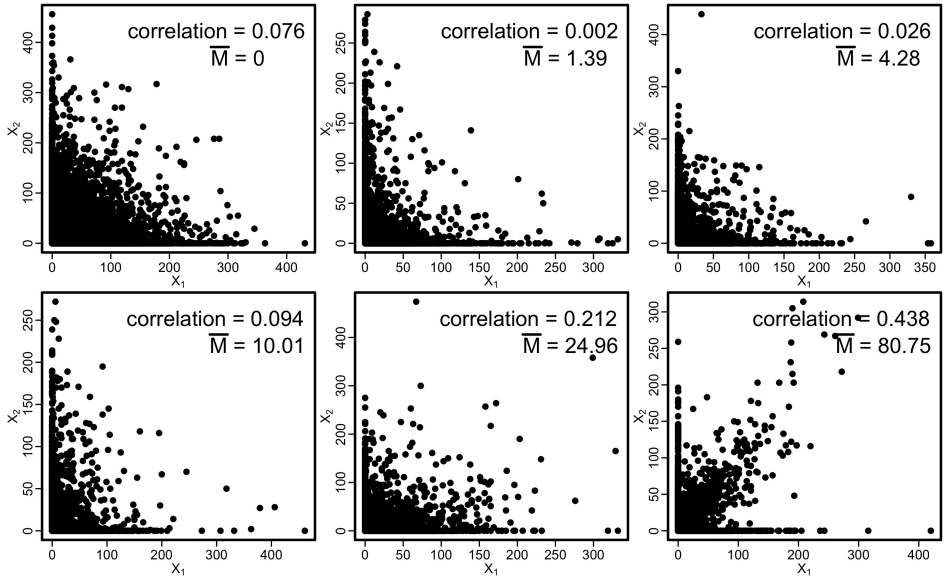
Our Recent Works

- LiquidAssociation R package for expression data from microarray experiments [Ho et al., 2011]
- Network construction and latent pathway identification [Ho et al., 2014, Baek et al., 2020]
- Fast search algorithm for identifying DC [Gunderson and Ho, 2014]
- Meta-Analysis [Kinzy et al., 2019, Wang et al., 2017]
- Correlated Count Data for bulk RNA-seq data ([Ma et al., 2020])

Short-Term Goal

To develop a flexible Single-Cell RNAseq Differential COExpression (scDECO) analysis framework and apply the proposed algorithm to identify sets of clinically relevant DC gene pairs using scRNAseq datasets.

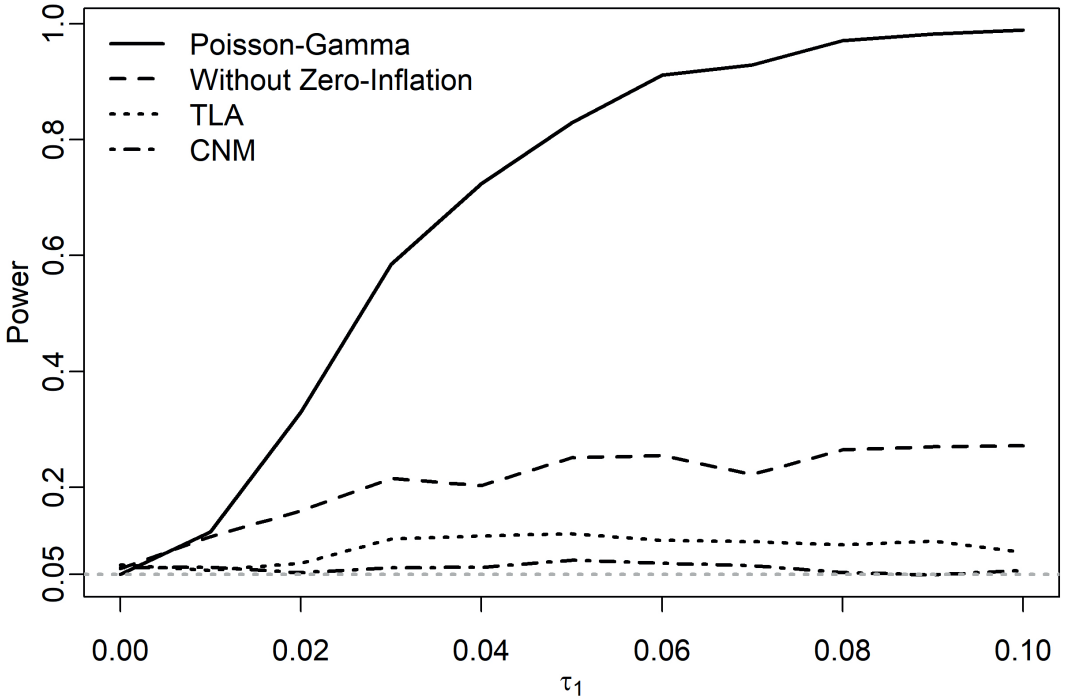
Aim 1: Develop a novel model-based framework for detecting DC using scRNAseq data



τ_1

- To develop a system to simulate data with differential co-expression patterns that mimic experimental scRNAseq datasets generated from various experiment protocols.
- Analyses to compare the performance of the proposed approaches to current DC analysis approaches based on experimental scRNAseq datasets and to evaluate the effect of read depth per cell in the comparisons

Findings


Poisson-Gamma
Without Zero-Inflation
TLA
CNM
Power
0.05
0.2
0.4
0.6
0.8
1.0
0.00
0.02
0.04
0.06
0.08
0.10
τ_1

Findings

Table: Poisson-Gamma: Coverage probability of 95% credible intervals (CIs), interval lengths, Mean square errors (MSE), and mean bias errors (MBE) based on 1,000 MCMC simulations ($\tau_0 = 0$, $\tau_1 = 0.05$) using ZENCO with Re. 50 participants, 100 cells per participant.

Parameter	95% Coverage probability	CI length	MSE	MBE
μ_1	0.946	2.356	0.366	-0.057
μ_2	0.944	2.356	0.382	-0.075
ϕ_1	0.943	0.672	0.030	-0.007
ϕ_2	0.951	0.667	0.028	-0.003
τ_0	0.960	1.054	0.066	-0.011
τ_1	0.951	0.044	0.000	0.000

Table: Copula-Based Model: Coverage probability of 95% credible intervals (CIs), interval lengths, Mean square errors (MSE), and mean bias errors (MBE) based on 1,000 MCMC simulations ($\tau_0 = 0$, $\tau_1 = 0.05$) using Copula with Re (new simulation setting). 20 participants, 500 cells per participants.

Parameter	Coverage probability	CI length	MSE	MBE
μ_1	0.962	0.422	0.011	-0.002
μ_2	0.959	0.422	0.011	-0.005
τ_0	0.956	0.963	0.052	0.007
τ_1	0.872	0.011	0.000	0.000
τ_y	0.959	1.436	0.132	0.003

Findings: scDECO with Individual Random Effects

Table: Coverage probability of 95% credible intervals (CIs), interval lengths, Mean square errors (MSE), and mean bias errors (MBE) based on 1,000 MCMC simulations for model with random effect

Parameter	Coverage probability	CI length	MSE	MBE
τ_0	0.97	0.48	0.0112	-0.0031
τ_1	0.96	0.47	0.0117	-0.0080

Table: Robustness: coverage probability of 95% credible intervals (CIs), interval lengths, Mean square errors (MSE), and mean bias errors (MBE) based on 1,000 MCMC simulations for model with random effect.

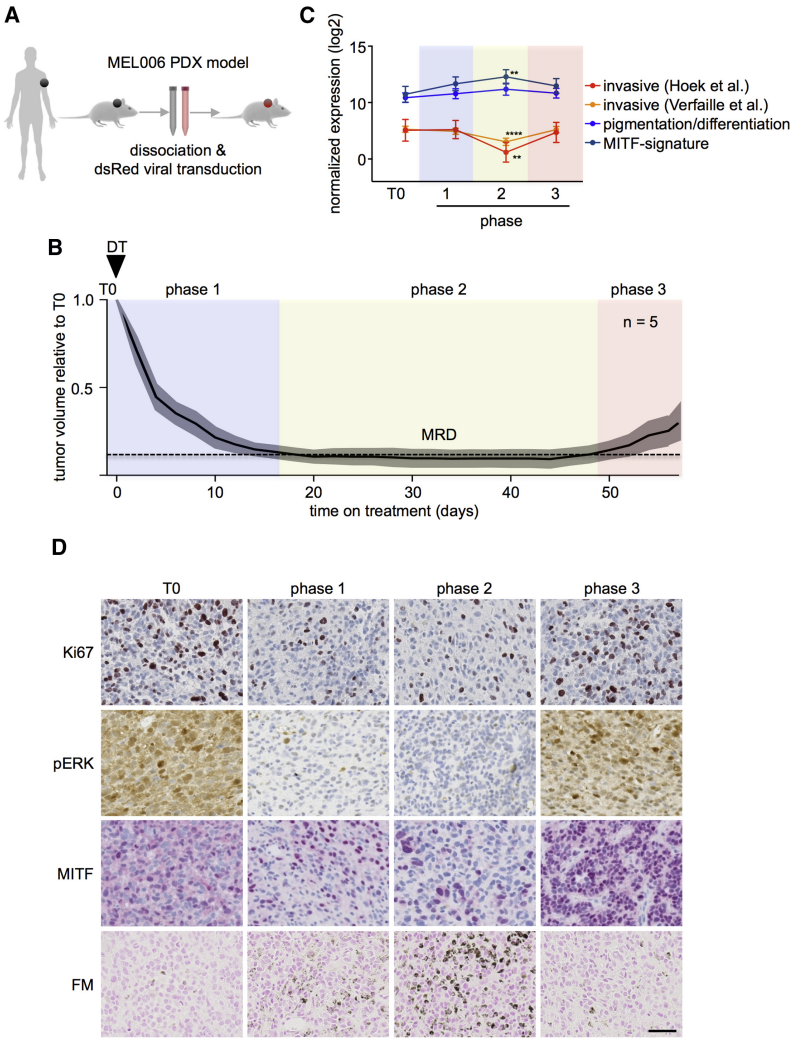
Parameter	Coverage probability	CI length	MSE	MBE
τ_0	0.96	0.47	0.0121	-0.0001
τ_1	0.98	0.47	0.0106	-0.0090

Findings: Fast Search Algorithm

Table: *Comparison of ES, SPSL and C-SPSL model based on 100 simulation iterations in scenario I (sparsity = 70%). The true values of τ_1 are set at $(0, 0, 0, 0, 0, 0, 0, 1, 1, 1)^T$ and the true values τ_0 are set to 0. The false discovery rate (FDR) and false negative rate (FNR) are reported.*

Sample size	Method	FDR	FNR	Run Time
	ES	0.1790	0.0067	11,050.47
$n = 200$	SPSL	0.0200	0.0652	292.63
	C-SPSL	0.0108	0.0903	355.15
	ES	0.0530	0.0000	30,540.85
$n = 500$	SPSL	0.0150	0.0012	573.35
	C-SPSL	0.0100	0.0012	671.10
	ES	0.0175	0.0000	69,904.12
$n = 1,000$	SPSL	0.0025	0.0000	1,038.18
	C-SPSL	0.0025	0.0000	1,183.61

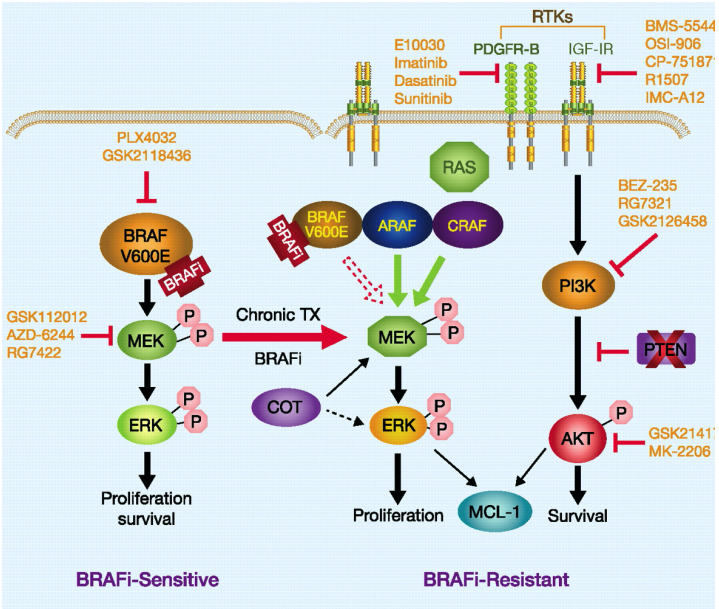
Aim 2: Identify sets of clinically relevant DC gene pairs using scRNAseq datasets from melanoma and advanced prostate cancer patients



Findings and Validation

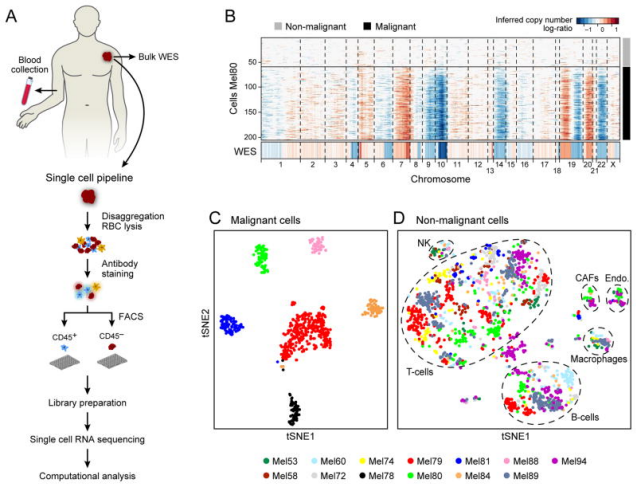
Table: Top table of dynamic correlations differences. $\Delta\tau_1$ is the difference between τ_1 estimates in Phase 3 (P3) and Phase 1 (P1).

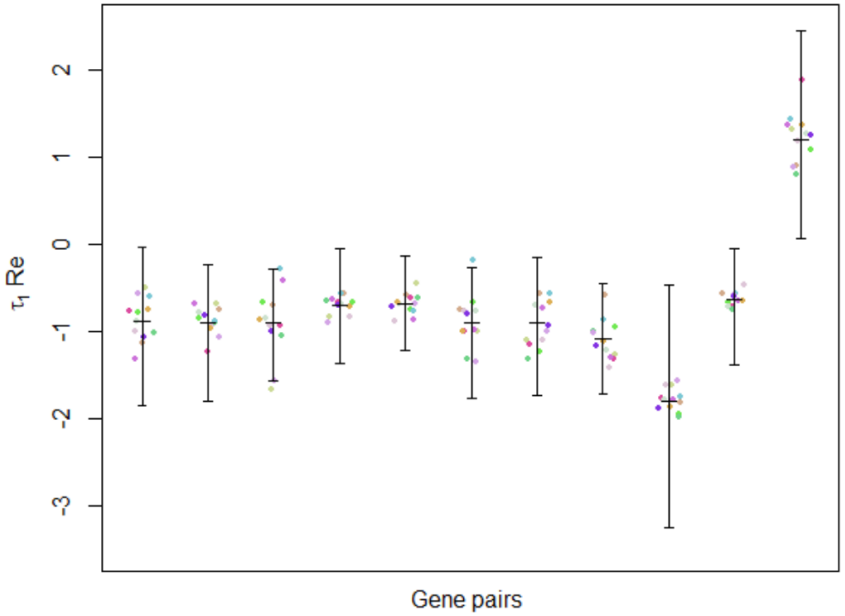
#	Gene1	Gene2	$\tau_1(P1)$	$\tau_1(P3)$	$\Delta\tau_1$
1	PDGFC	FGFR1	0.045 (0.021,0.068)	-0.003 (-0.010, 0.005)	-0.047 (-0.072,-0.023)
2	AKT1	BAX	0.040 (0.008,0.071)	-0.003 (-0.014, 0.008)	-0.043 (-0.075,-0.010)
3	AKT1	PIK3R1	-0.016 (-0.035,0.004)	0.024 (0.009, 0.038)	0.040 (0.015, 0.062)
4	PDGFC	MAP2K2	0.016 (-0.002,0.032)	-0.023 (-0.036,-0.006)	-0.039 (-0.059,-0.013)
5	IGF1R	FGFR1	-0.024 (-0.048,0.000)	0.007 (0.000, 0.014)	0.032 (0.006, 0.056)



Findings

- 4,645 cells isolated from 19 freshly resected melanoma tumors using Smart-Seq2
- We will develop risk scoring algorithms using top scoring DC gene pairs for patients clinical outcome prediction.





Aim 3: Implement a freely accessible software package for DC analysis.

We will release R/Bioconductor packages for implementing the scDECO algorithm. The R packages will provide the functionality to

- Simulate datasets that exhibit DC patterns based on parameter settings calculated from experimental scRNAseq datasets;
- Implement the algorithm using the Poisson-Gamma and the Gaussian copula model with and without zero-inflation, respectively;
- Perform goodness of fit and model selection based on the scRNAseq data under study;
- Calculate risk scores based on DC gene pairs.
- The scDECO framework will be provided as open-source software packages under the BSD 3-Clause License. The software will be distributed and maintained via the GitHub or R/Bioconductor repository.

Goals and Time Line

	Aim1	Aim2
Year 1	1. Implement and test scDECO	1.Implement scDECO
	2. Submit results for publication	using scRNA-seq datasets
Year 2	3. Evaluate risk score function	2. Prediction using scDECO
	4. Submit results for publication	3. Submit results for publication
	5. Release R packages	

Manuscript and Software

- Software: R/Bioconductor and GitHub software packages under the BSD 3-Clause License.
  - LiquidAssociation R package `https://www.bioconductor.org/packages/release/bioc/html/LiquidAssociation.html`
  - fastLiquidAssociation R package `https://www.bioconductor.org/packages/release/bioc/html/fastLiquidAssociation.html`
  - nPARS `https://people.stat.sc.edu/hoyen/research.html`
  - Correlated Count Data for bulk RNA-seq data `https://github.com/ZichenMa-USC/Correlated-bivariate-count-data-regression`
  - ZENCO for single-cell RNA-seq data `https://github.com/zheny714/ZENCO`
  - Integrating correlated multi-omics data from single-cell experiments. `https://github.com/ZichenMa-USC/FlexibleCopulaModel`
  - Fast Search Algorithm (SPSL, C-SPSL) `https://github.com/zhangwenda1990/DGCspsl`

Manuscripts and Software

- Manuscripts
  1. Yang Z, Ho YY. Modeling dynamic correlation in zero-inflated bivariate count data with applications to single-cell RNA sequencing data. *Biometrics. 2021 PubMed PMID: 33720414; PubMed Central PMCID: PMC8477913; DOI: 10.1111/biom.13457.*
  2. *Flexible copula model for integrating correlated multi-omics data from single-cell experiments. Biometrics (To appear)*
  3. *Zhang W., Wang L., Fan D., Ho, Y.-Y. (2022+) Fast Search Algorithms for Identifying Dynamic Gene Co-expression via Bayesian Variable Selection (Under Review: Statistics in Medicine)*
  4. *Yang Z., Chen H., Ho Y.-Y. (2022+) Use sufficient direction factor model to classify cell types using single-cell RNA sequencing data. (In Preparation)*
  5. *Yang Z., Ho Y.-Y. (2022+) scDECO: A novel statistical framework to identify differential co-expression gene combinations systematically using single-cell RNA sequencing data. (In Preparation)*

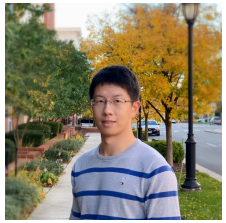
Our Team

Yen-Yi Ho
Department of Statistics
University of South Carolina

Chun-Liang Chen
The University of Texas
Health Science Center at San Antonio

Evan Keller,
Department of Urology and Pathology
University of Michigan Medical School

Hexin Chen
Department of Biological Sciences
University of South Carolina



Zhen Yang, PhD Student Recently Graduated
Data Scientist for Walmart Labs, SF CA

Questions and Feedback

- Contact: Yen-Yi Ho: hoyen@stat.sc.edu

Thank you!!

References I

Baek, S., Ho, Y.-Y., and Ma, Y. (2020).
Using sufficient direction factor model to analyze latent activities associated with breast cancer survival.
Biometrics, 76(4):1340–1350.

Gunderson, T. and Ho, Y.-Y. (2014).
An efficient algorithm to explore liquid association on a genome-wide scale.
BMC bioinformatics, 15(1):371.

Ho, Y.-Y., Cope, L. M., and Parmigiani, G. (2014).
Modular network construction using eqtl data: an analysis of computational costs and benefits.
Frontiers in genetics, 5:40.

References II

Ho, Y.-Y., Parmigiani, G., Louis, T. A., and Cope, L. M. (2011).
Modeling liquid association.
Biometrics, 67(1):133–141.

Kinzy, T. G., Starr, T. K., Tseng, G. C., and Ho, Y.-Y. (2019).
Meta-analytic framework for modeling genetic coexpression dynamics.
Statistical applications in genetics and molecular biology, 18(1):1–12.

Ma, Z., Hanson, T., and Ho, Y.-Y. (2020).
Flexible bivariate correlated count data regression.
Statistics in Medicine, 39:3476–3490.

References III

Wang, L., Liu, S., Ding, Y., Yuan, S., Ho, Y.-Y., and Tseng, G. C. (2017).
Meta-analytic framework for liquid association.
Bioinformatics, 33(14):2140–2147.