Personalized Causal Machine Learning Using Genomic Data

Gregory Cooper and Xinghua Lu
Department of Biomedical Informatics
University of Pittsburgh

NHGRI Workshop on Machine Learning in Genomics
April 13-14, 2021

Outline

- Causal machine learning
- Personalized causal machine learning
- Examples
- Extensions
- Comments
- Conclusions

Causal Machine Learning

Science in general and genomic science in particular are centrally concerned with the discovery of causal relationships in order to:

- Understand mechanisms
- Understand the molecular details of transcription regulation

- Predict the results of interventions
- Predict the cellular effects of a gene modification

- Control events
- Control the overexpression of a genomic driver of cancer that is due to copy number amplification

Personalized (Instance-Specific) Causal Machine Learning

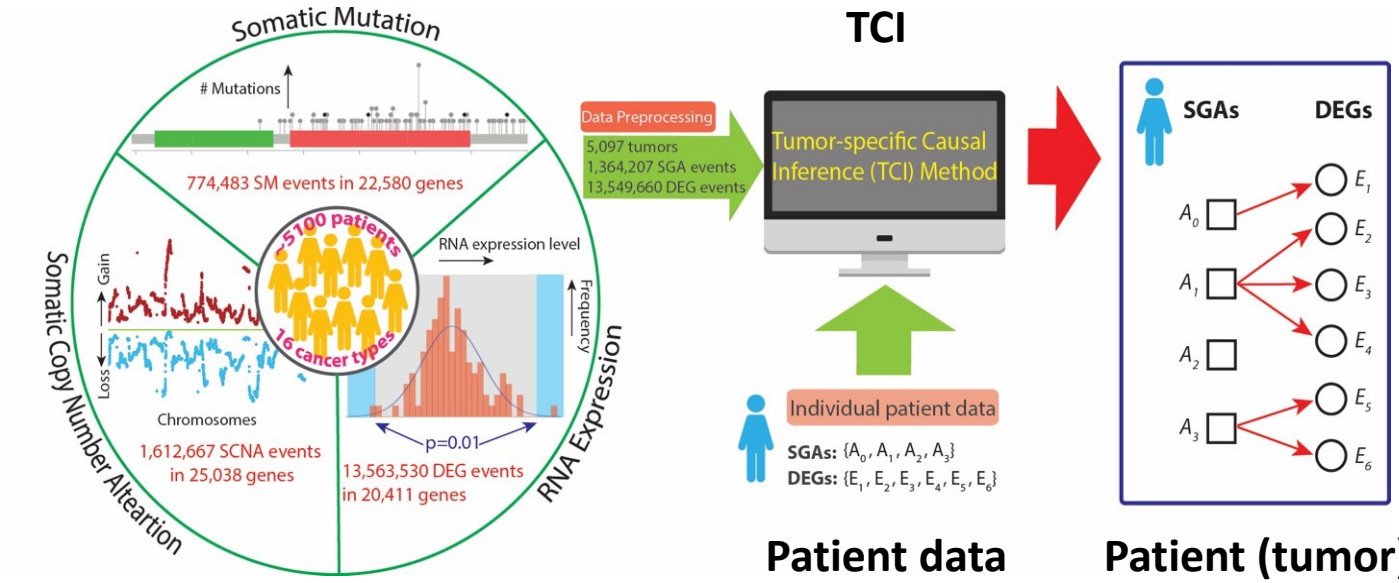
- Traditional causal structure learning algorithms learn the causal relationships that exists *in a population*
  - Often it is a mixture of causal relationships
  - Only the strongest, shared causal relationships may be learned
- Learning causal structure that is *specific to a given instance* (*e.g., a patient*) is an important but understudied problem

  Doing so allows us to understand more precisely the causal relationships of the instance, which more exactly guides how to maintain health and cure disease in the instance.

Example: Identifying Tumor-Specific Genomic Drivers of Cancer

- Identifying driver somatic genomic alterations (SGAs) of an individual tumor is an important task within personalized cancer medicine
- A tumor usually hosts hundreds to thousands of SGAs (e.g., ~400 in breast cancer)
  - A few are drivers (causes cancerous behavior in a tumor)
  - Many are passengers (no causal influence on cancer behavior)
- Current knowledge of cancer driver genes is incomplete
  - >10% of tumors have no known drivers

From Big Data to a Model of an Individual Tumor



TCGA Data that We Used

Types of data used

DNA data

- Whole exome data
- Copy number variation data

mRNA expression data

Variables we defined on TCGA data

Let SGA_i denote the somatic genomic alteration status of $gene_i$.

SGA_i = 1 if $gene_i$ contains any nonsynonymous somatic mutations or $gene_i$ has an abnormal degree of copy number variation;
otherwise SGA_i = 0.

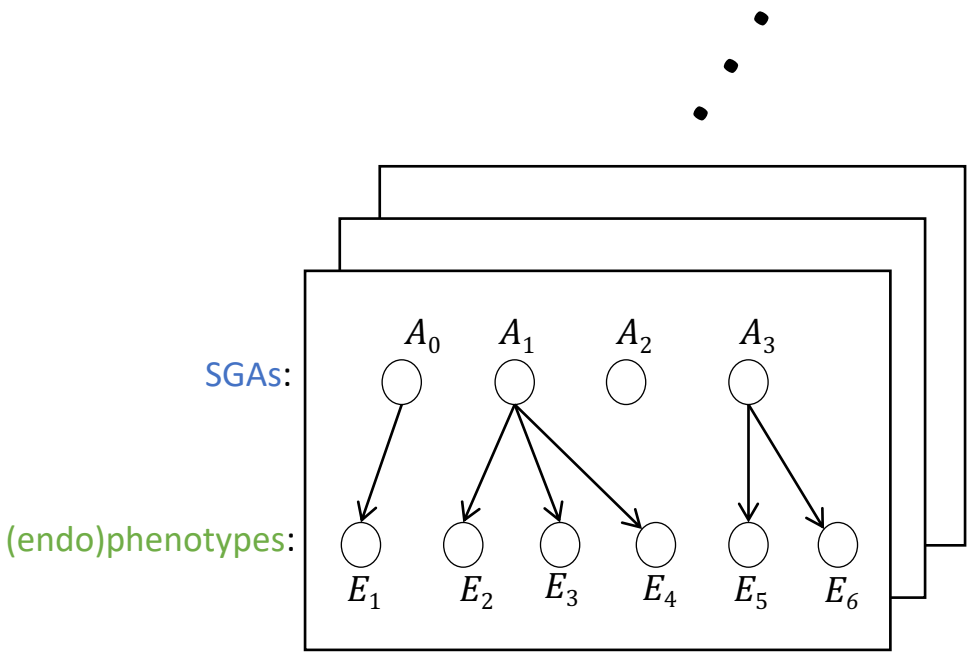
Let DEG_i denote the differentially expressed gene status of $gene_i$.

DEG_i = 1 if $gene_i$ is significantly differentially expressed, relative to a baseline;
otherwise, DEG_i = 0.

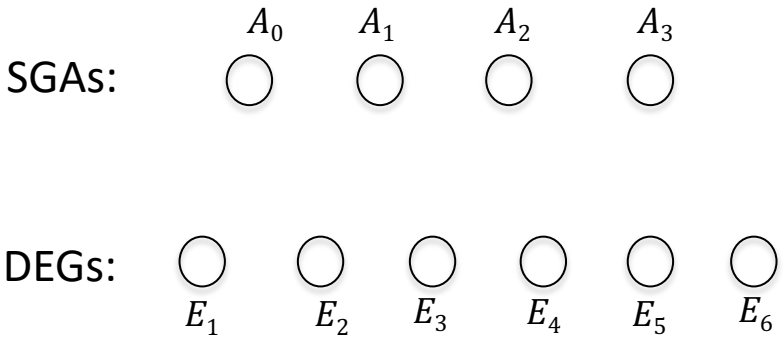
Tumor-specific Causal Inference (TCI) Algorithm

- Infer causal relationships between **SGAs** and the **molecular phenotypes** observed in a given tumor
  - Transcriptome: differentially expressed genes (DEGs)
  - Proteomics
  - Metabolomics
  - Immunology markers

- The SGAs that causally regulate oncogenic phenotypes are drivers of the tumor



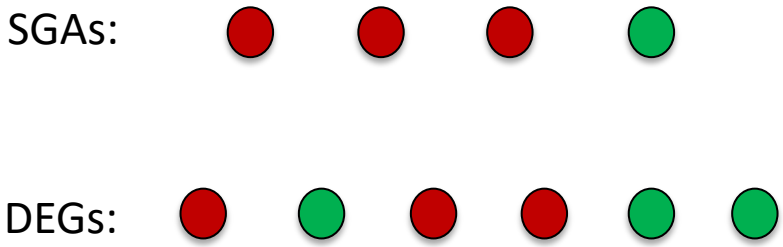
TCI Algorithm*



- Uses a bipartite graph representation
- Searches the graph for somatic genomic alterations (SGAs) that account for differentially expressed genes (DEGs) involved in an oncogenic process.
- Uses a Bayesian evaluation measure, which is tumor specific (and therefore supports precision/personalized modeling)

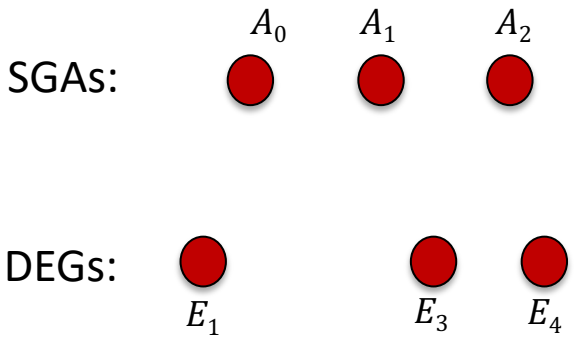
* Cai C, Cooper GF, Lu KN, Ma X, Xu S, Zhao Z, Chen X, Xue Y, Lee AV, Clark N, Chen V. Systematic discovery of the functional impact of somatic genome alterations in individual tumors through tumor-specific causal inference. PLOS Computational Biology, 15 (2019).

TCI Algorithm



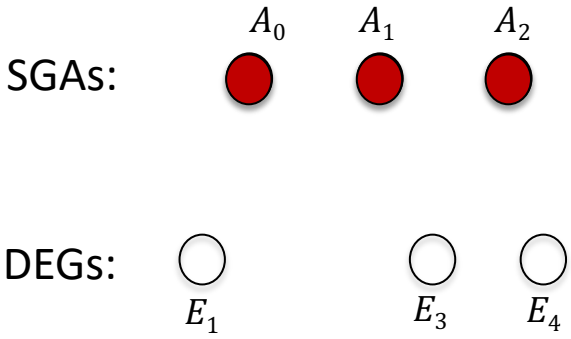
- Uses a bipartite graph representation
- Searches the graph for somatic genomic alterations (SGAs) that account for differentially expressed genes (DEGs) involved in an oncogenic process.
- Uses a Bayesian evaluation measure, which is tumor specific

TCI Algorithm



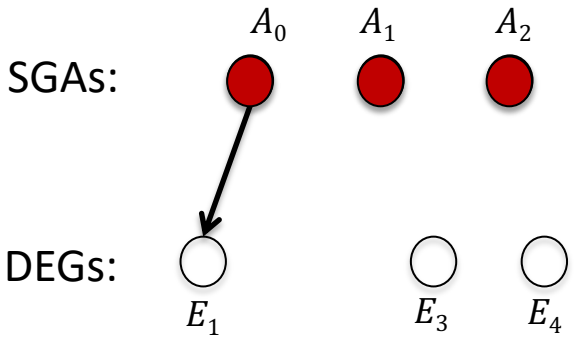
- Uses a bipartite graph representation
- Searches the graph for somatic genomic alterations (SGAs) that account for differentially expressed genes (DEGs) involved in an oncogenic process.
- Uses a Bayesian evaluation measure, which is tumor specific

TCI Algorithm



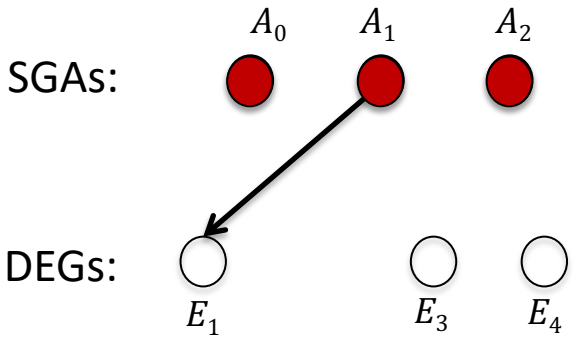
- Uses a bipartite graph representation
- Searches the graph for somatic genomic alterations (SGAs) that account for differentially expressed genes (DEGs) involved in an oncogenic process.
- Uses a Bayesian evaluation measure, which is tumor specific

TCI Algorithm



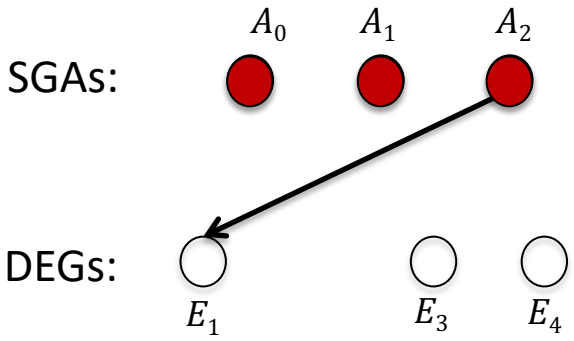
- Uses a bipartite graph representation
- Searches the graph for somatic genomic alterations (SGAs) that account for differentially expressed genes (DEGs) involved in an oncogenic process.
- Uses a Bayesian evaluation measure, which is tumor specific

TCI Algorithm



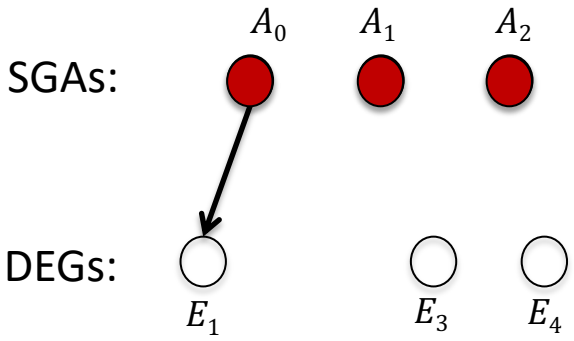
- Uses a bipartite graph representation
- Searches the graph for somatic genomic alterations (SGAs) that account for differentially expressed genes (DEGs) involved in an oncogenic process.
- Uses a Bayesian evaluation measure, which is tumor specific

TCI Algorithm



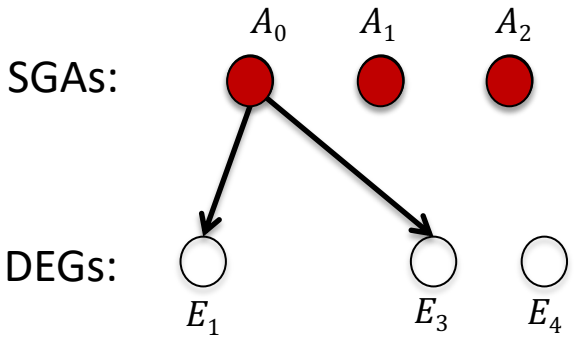
- Uses a bipartite graph representation
- Searches the graph for somatic genomic alterations (SGAs) that account for differentially expressed genes (DEGs) involved in an oncogenic process.
- Uses a Bayesian evaluation measure, which is tumor specific

TCI Algorithm



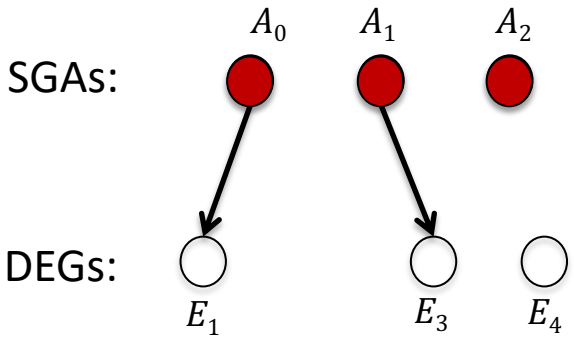
- Uses a bipartite graph representation
- Searches the graph for somatic genomic alterations (SGAs) that account for differentially expressed genes (DEGs) involved in an oncogenic process.
- Uses a Bayesian evaluation measure, which is tumor specific

TCI Algorithm



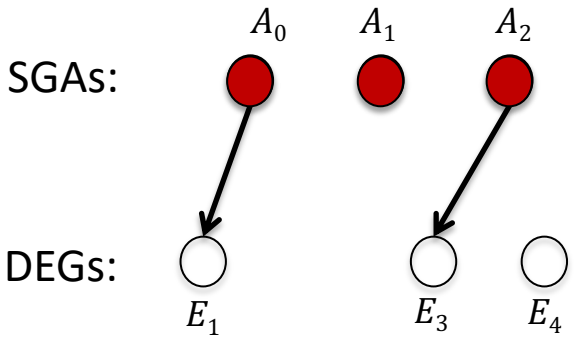
- Uses a bipartite graph representation
- Searches the graph for somatic genomic alterations (SGAs) that account for differentially expressed genes (DEGs) involved in an oncogenic process.
- Uses a Bayesian evaluation measure, which is tumor specific

TCI Algorithm



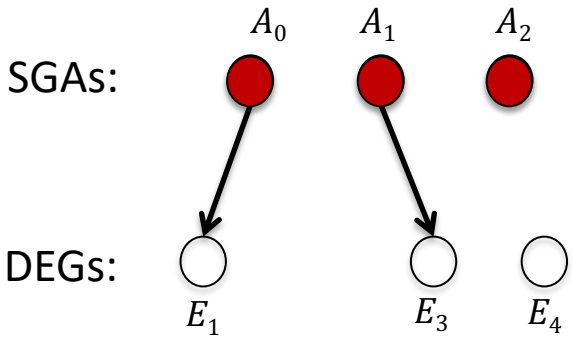
- Uses a bipartite graph representation
- Searches the graph for somatic genomic alterations (SGAs) that account for differentially expressed genes (DEGs) involved in an oncogenic process.
- Uses a Bayesian evaluation measure, which is tumor specific

TCI Algorithm



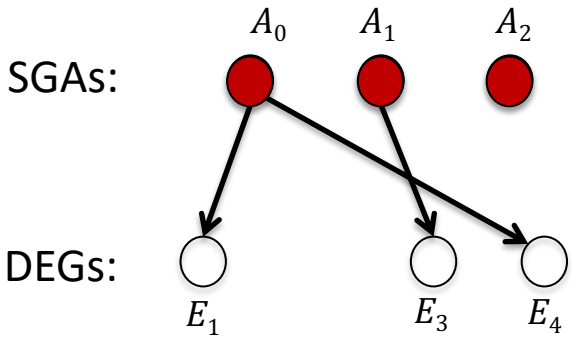
- Uses a bipartite graph representation
- Searches the graph for somatic genomic alterations (SGAs) that account for differentially expressed genes (DEGs) involved in an oncogenic process.
- Uses a Bayesian evaluation measure, which is tumor specific

TCI Algorithm



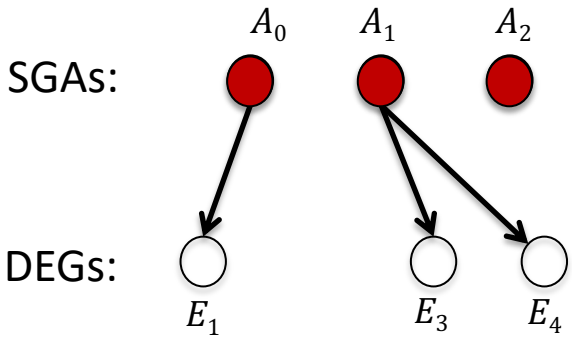
- Uses a bipartite graph representation
- Searches the graph for somatic genomic alterations (SGAs) that account for differentially expressed genes (DEGs) involved in an oncogenic process.
- Uses a Bayesian evaluation measure, which is tumor specific

TCI Algorithm



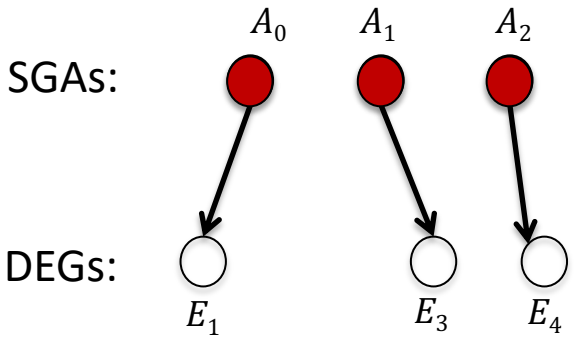
- Uses a bipartite graph representation
- Searches the graph for somatic genomic alterations (SGAs) that account for differentially expressed genes (DEGs) involved in an oncogenic process.
- Uses a Bayesian evaluation measure, which is tumor specific

TCI Algorithm



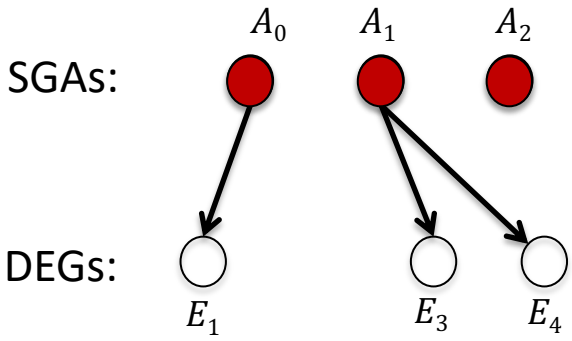
- Uses a bipartite graph representation
- Searches the graph for somatic genomic alterations (SGAs) that account for differentially expressed genes (DEGs) involved in an oncogenic process.
- Uses a Bayesian evaluation measure, which is tumor specific

TCI Algorithm



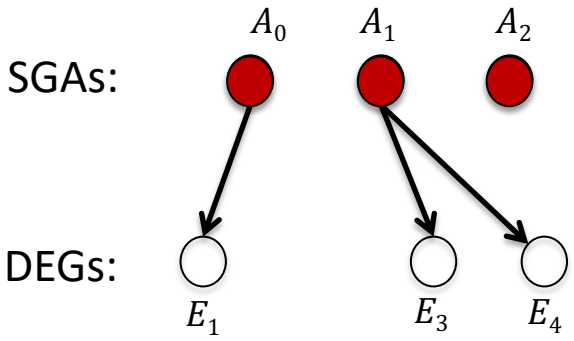
- Uses a bipartite graph representation
- Searches the graph for somatic genomic alterations (SGAs) that account for differentially expressed genes (DEGs) involved in an oncogenic process.
- Uses a Bayesian evaluation measure, which is tumor specific

TCI Algorithm



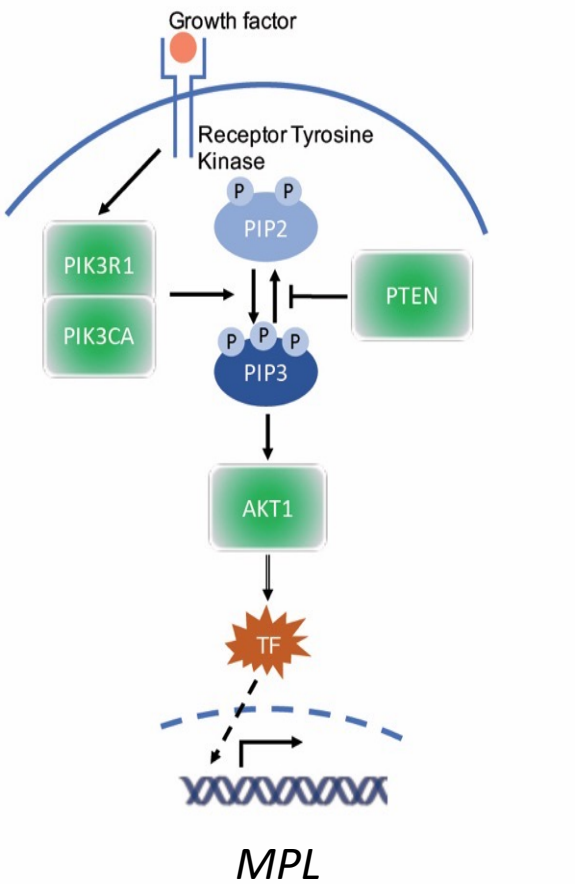
- Uses a bipartite graph representation
- Searches the graph for somatic genomic alterations (SGAs) that account for differentially expressed genes (DEGs) involved in an oncogenic process.
- Uses a Bayesian evaluation measure, which is tumor specific

TCI Algorithm

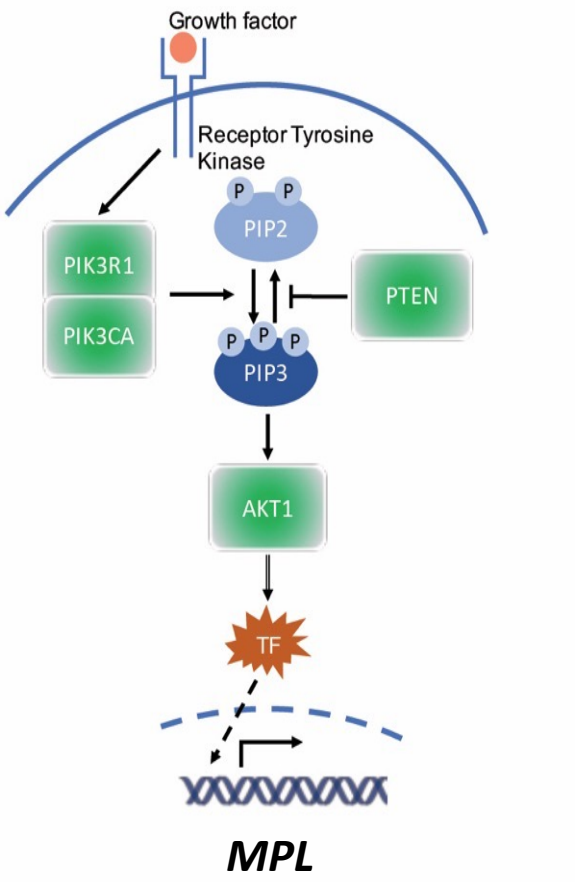


- Uses a bipartite graph representation
- Searches the graph for somatic genomic alterations (SGAs) that account for differentially expressed genes (DEGs) involved in an oncogenic process.
- Uses a Bayesian evaluation measure, which is tumor specific
- Makes the biologically plausible “mutually exclusivity” assumption that each DEG *in a single tumor* has only one driver (explaining away)

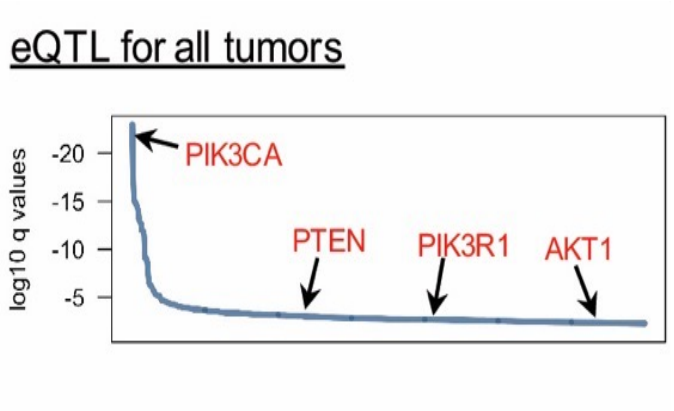
PI3K/AKT Pathway Regulating *MPL* Expression



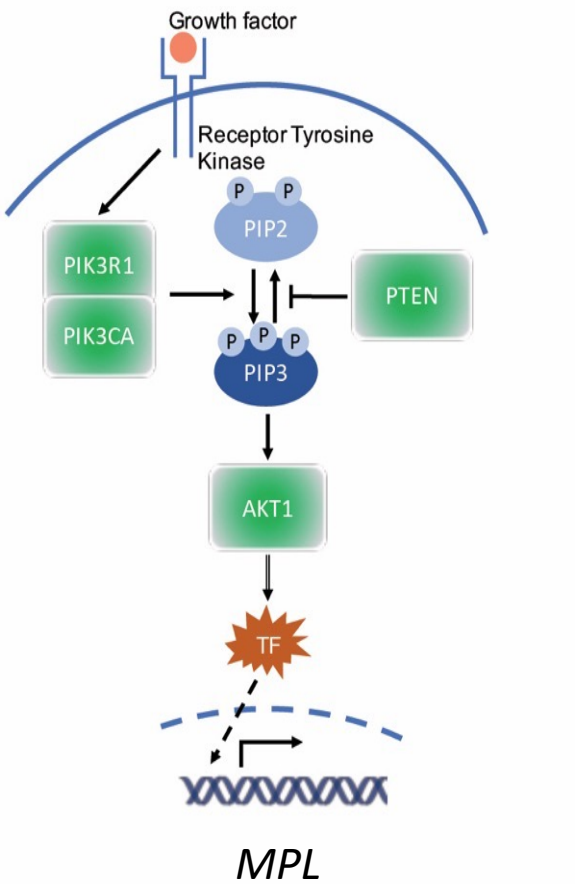
eQTL Analysis



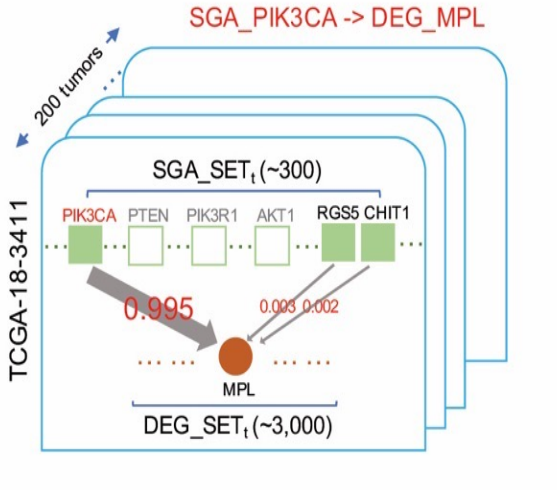
MPL



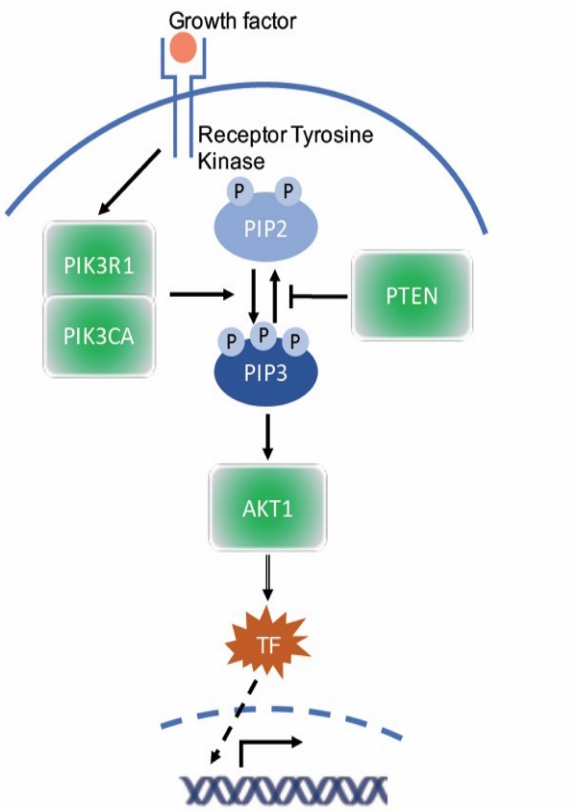
TCI Analysis of Tumor 18-3411



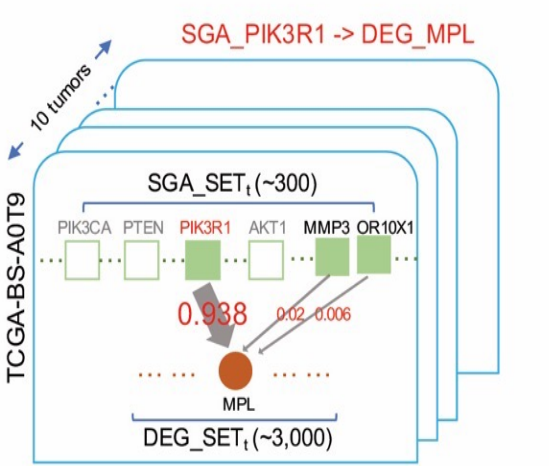
MPL



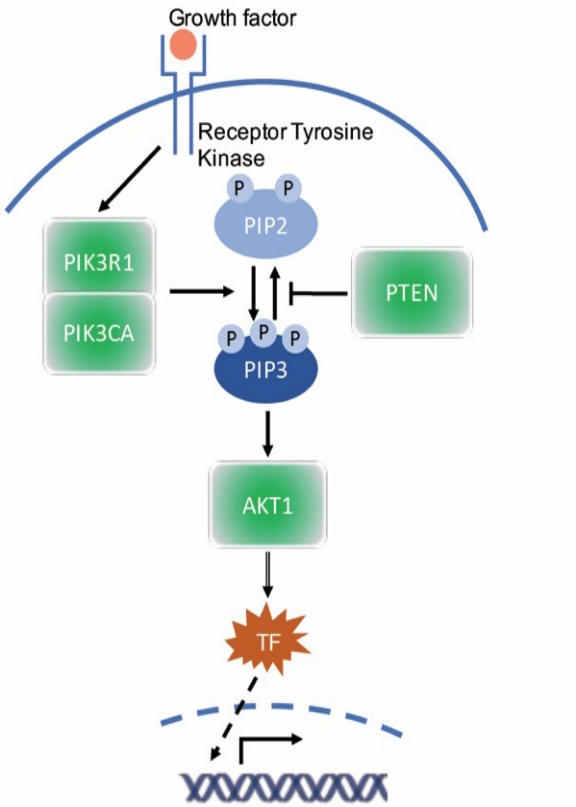
TCI Analysis of Tumor BS-A0T9



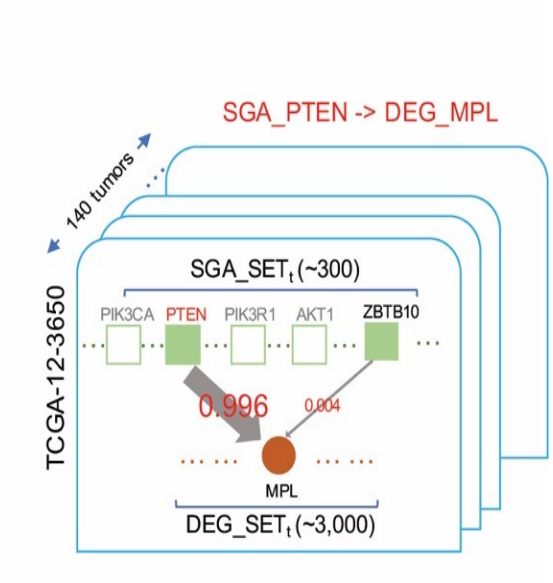
MPL



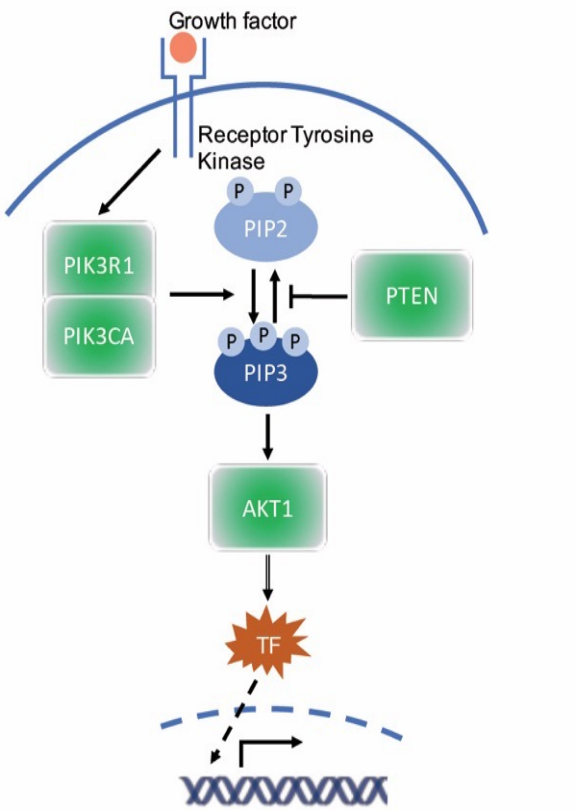
TCI Analysis of Tumor 12-3650



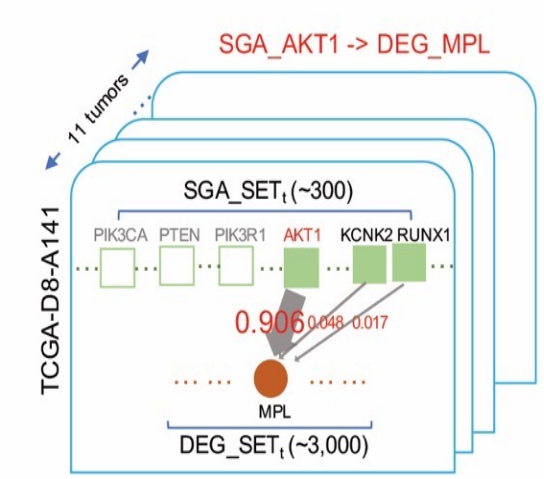
MPL



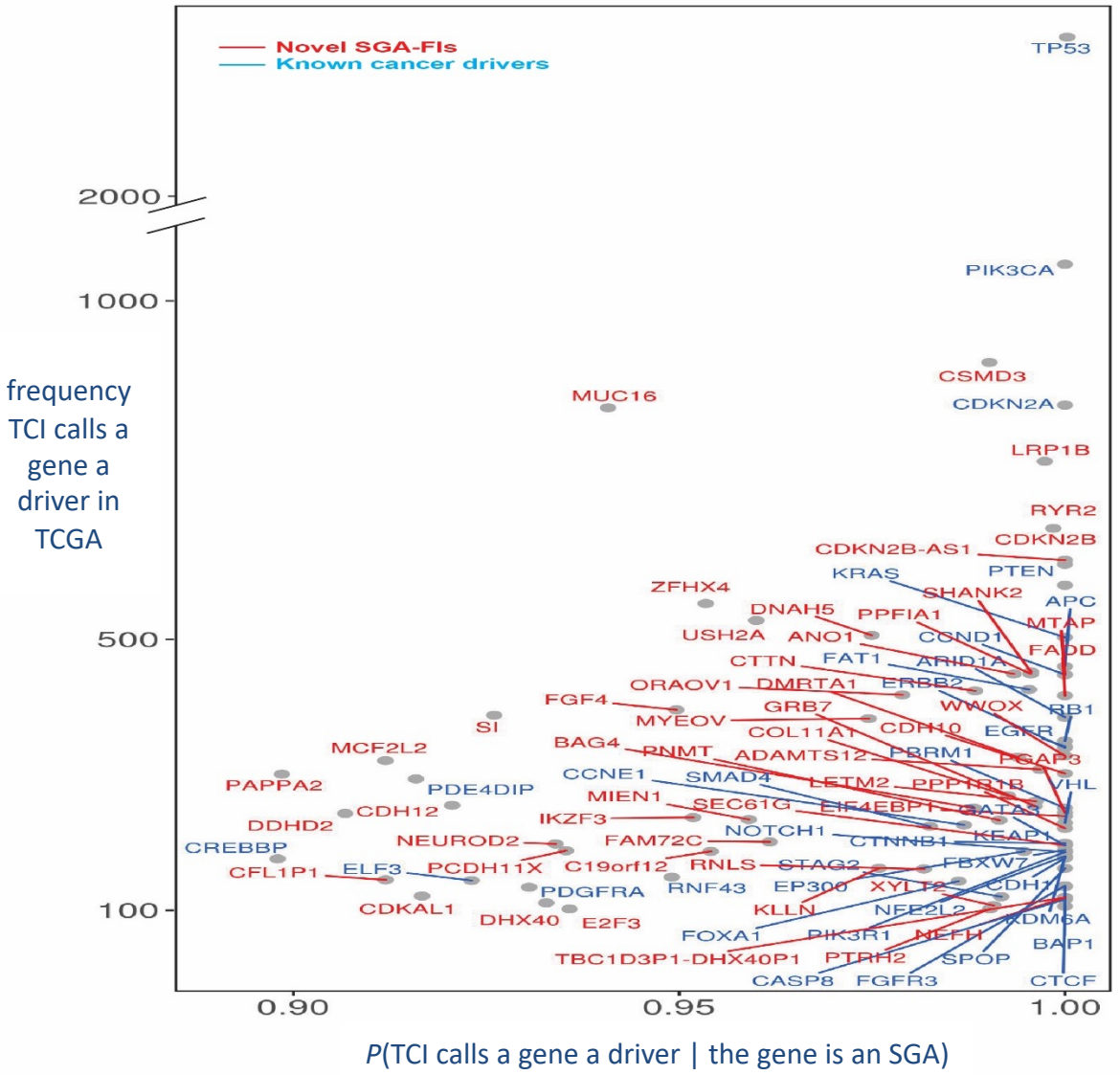
TCI Analysis of Tumor D8-A141



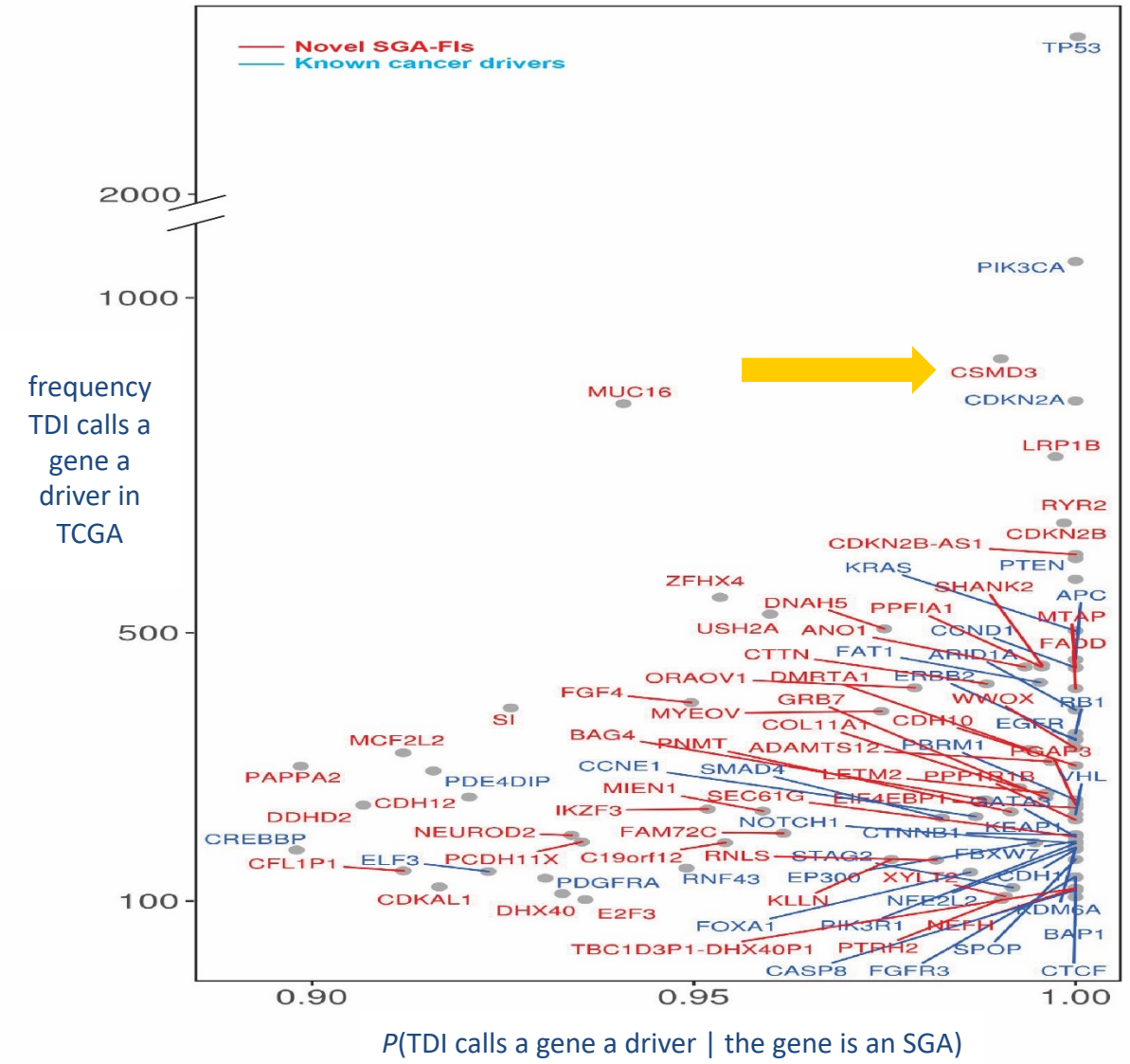
MPL



Results of Applying TCI to TCGA Pan-Cancer Data



Results of Applying TCI to TCGA Pan-Cancer Data

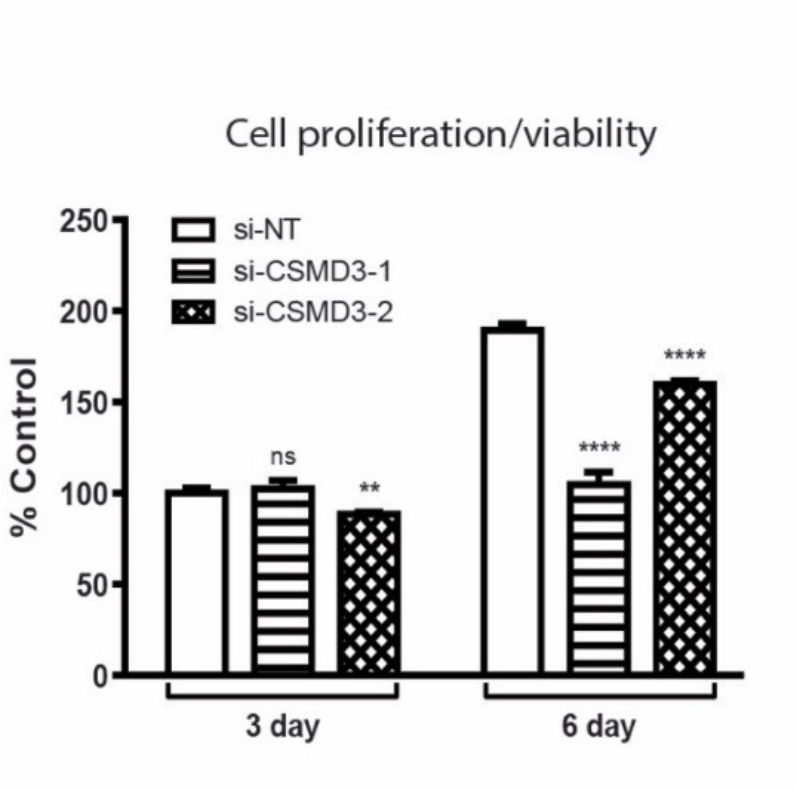


Experimental Investigation of CSMD3*

- Used a gastric cancer cell line in which CSMD3 is highly expressed
- Knocked down (attenuated) CSMD3 expression with two different siRNAs

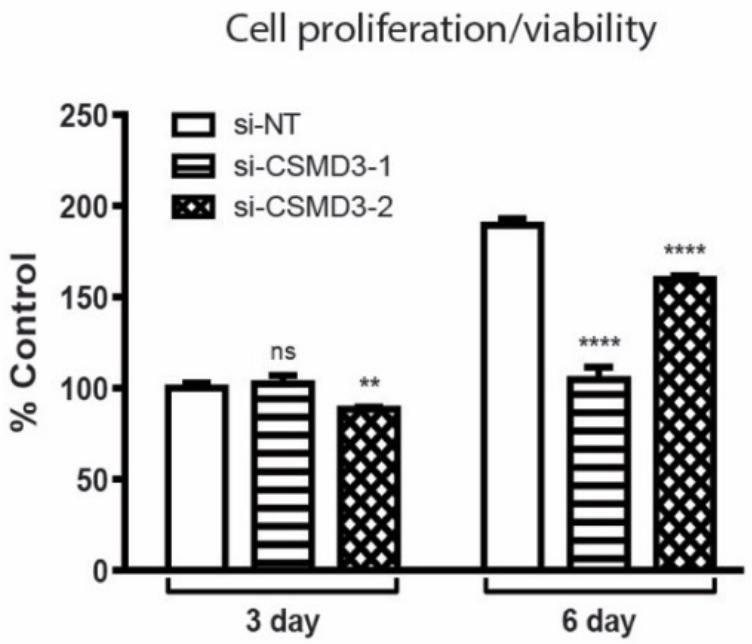
* Cai C, Cooper GF, Lu KN, Ma X, Xu S, Zhao Z, Chen X, Xue Y, Lee AV, Clark N, Chen V. Systematic discovery of the functional impact of somatic genome alterations in individual tumors through tumor-specific causal inference. PLOS Computational Biology, 15 (2019).

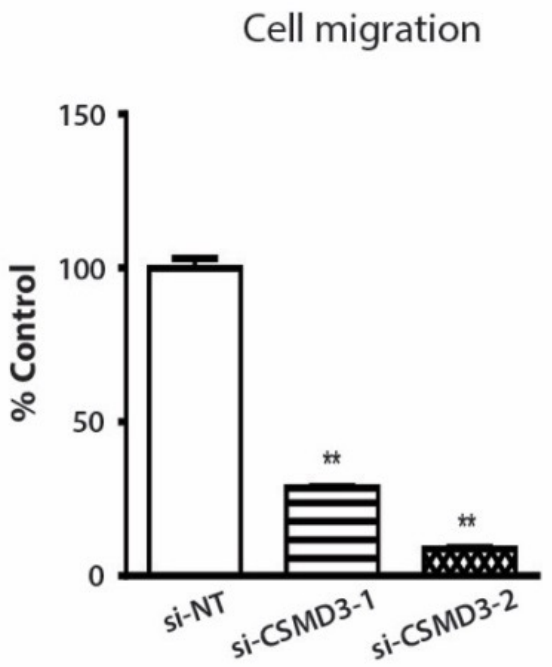
Results of Knocking Down CSMD3



** $p < 0.01$ **** $p < 0.0001$

Results of Knocking Down CSMD3





** $p < 0.01$ **** $p < 0.0001$

CSMD3 in COSMIC

- COSMIC is an online database of somatically acquired mutations found in human cancer https://cancer.sanger.ac.uk/cosmic
- CSMD3 is now classified as a Tier 2 COSMIC cancer gene.
- Tier 2 genes are genes with strong indications of a role in cancer but with less extensive available evidence. These are generally more recent targets, where the body of evidence supporting their role in cancer is still emerging.

Representative Related Work

- There has been prior work on
  - representing and learning context-specific conditional independence [1-7]
  - learning instance-specific models [8, 9]
  - learning instance-specific causal models[10,11]
- To our knowledge, there has not been prior work on Bayesian methods to learn instance-specific causal Bayesian networks, as presented here.

1. Boutilier C, et al. Context-specific independence in Bayesian networks . UAI (2013).
2. Chickering DM, et al. A Bayesian approach to learning Bayesian networks with local structure. UAI (1997).
3. Friedman N, et al. Learning Bayesian networks with local structure . UAI (1996).
4. Geiger D, et al. Knowledge representation and inference in similarity networks and Bayesian multinets . Artificial Intelligence (1996).
5. Hyttinen A, et al. Structure learning for Bayesian networks over labeled DAGs. Conference on Probabilistic Graphical Models (2018).
6. Pensar J, et al. Labeled directed acyclic graphs: a generalization of context-specific independence in directed graphical models. KDD (2015).
7. Zou Y, et al. Representing local structure in Bayesian networks by Boolean functions . Pattern Recognition Letters (2017).
8. Lengerich B, et al. Learning sample-specific models with low-rank personalized regression. NeurIPS (2019).
9. Liu X, et al. Personalized characterization of diseases using sample-specific networks. Nucleic Acids Research (2016).
10. Li X, et al. Learning subject-specific directed acyclic graphs with mixed effects structural equation models from observational data . Frontiers in Genetics (2018).
11. Kuijjer ML, et al. Estimating sample-specific regulatory networks. iScience (2019).

Extensions

- TCI estimates genomic causes of (endo)phenotypes
- We are extending it to include embeddings that provide additional control for confounding
- We have also developed a method called iGFCI for learning instance-specific pathways between genomic causes and the resulting (endo)phenotypes*.
- iGFCI models for the possibility that there are latent confounders of the measured variables.
- We have recently applied iGFCI to investigate molecular pathways involved in immune regulation.

* Jabbari F, Visweswaran S, Cooper GF. Instance-specific Bayesian network structure learning. In: *Proceedings of the Conference on Probabilistic Graphical Models* (2018).

Comments

Ideal types of data to fully leverage the potential for instance specific causal machine learning include those that are:

- tissue specific
- include cells in the microenvironment of the disease tissue
- include multiple types of measurements within single cells

Conclusions

- Instance specific causal machine learning is a promising tool for analyzing genomic data.
- The method is applicable to many types of genome-(endo)phenome analysis, not just cancer.
- Additional development and evaluation are needed and are ongoing.

References

Cai C, Cooper GF, Lu KN, Ma X, Xu S, Zhao Z, Chen X, Xue Y, Lee AV, Clark N, Chen V. Systematic discovery of the functional impact of somatic genome alterations in individual tumors through tumor-specific causal inference. *PLOS Computational Biology,* 15 (2019).

Jabbari F, Visweswaran S, Cooper GF. Instance-specific Bayesian network structure learning. In: *Proceedings of the Conference on Probabilistic Graphical Models* (2018).

Acknowledgements

- Thanks to Drs. Fattaneh Jabbari and Chunhui Cai for slides and figures that they contributed to this talk.
- Thanks to Dr. Shyam Visweswaran, Dr. Cai, and Dr. Jabbari for their central conceptual and technical contributions to this research.
- Support for this work was provided by grant U54HG008540 from the National Institutes of Health (NIH) and by grant IIS-1636786 from the National Science Foundation.

Thank you

gfc@pitt.edu
xinghua@pitt.edu

www.ccd.pitt.edu