Integrating transcriptomics with proteomics to construct causal regulatory networks in cancer

Simon Koplev[1,2]; Katie Lin[3]; Anders B. Dohlman[1,2]; Avi Ma'ayan[1,2].

[1]Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY.

[2]BD2K-LINCS Data Coordination and Integration Center (DCIC).

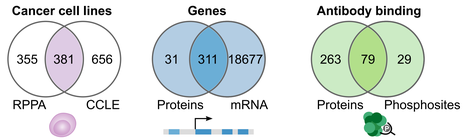
[3]Department of Computer Science, Columbia University, New York, NY.

BD2K - LINCS
DATA COORDINATION AND INTEGRATION CENTER

Icahn School of Medicine at Mount Sinai

Introduction

Oncogenic cell signaling involves aberrant disruption to multiple regulatory layers, highjacking physiological programs of coordinated transcription. Prominently, epithelial-mesenchymal transition (EMT) is a normal cellular event that occurs during vertebrate development and wound healing. In cancer, EMT drives tumor progression through promotion of metastasis. Most pan-cancer profiling and analyses have focused on transcriptomics and genomics. Integrating protein-based data with other regulatory layers could reveal in greater detail the mechanisms of cancer-related processes such as EMT.

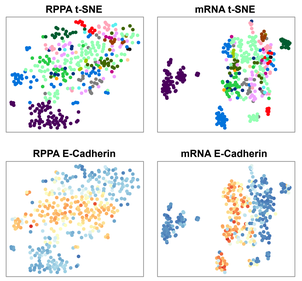


We integrated protein expression and phosphosite data on 736 pan-cancer cell lines, measured by reverse phase protein arrays (RPPA), with matching transcriptomic data.

Main results

1) Causal discovery algorithms and bimodal expression analysis of pan-cancer data uncover oncogenic regulatory dependencies exemplified by the EMT process.
2) EMT-associated proteins are distributed in multiple clusters across cancer cell lines with core proteins in the vicinity of E-cadherin--a hallmark marker of EMT.
3) E-cadherin is primarily transcriptionally regulated with important roles for H3 acetylation and methylation.
4) E-cadherin is also translationally or post-translationally regulated in a subset of cancer cell lines.
5) EMT can be induced in cancer cell lines by specific histone deacetylase (HDAC) inhibitors and reverted by kinase inhibitors.

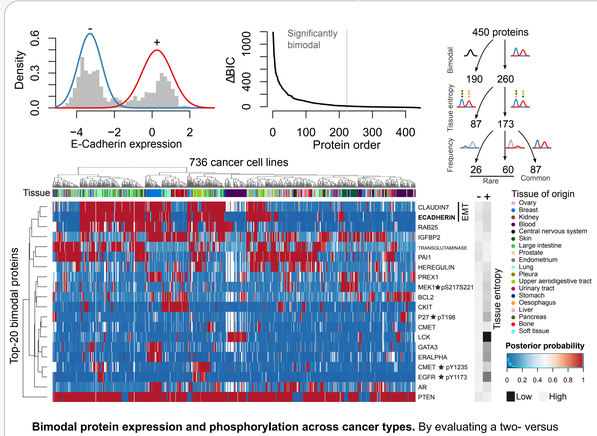
Euclidean distances



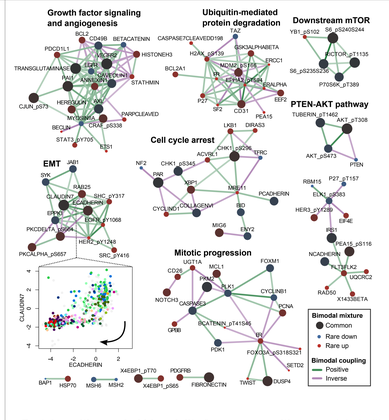
Overview of pan-cancer cell line data quantifying abundances of transcripts, proteins, and protein phosphorylations. The antibody-based protein measurements (RPPA) were matched to mRNA data from the Cancer Cell Line Encyclopedia (CCLE). Dimensionality reduction by t-SNE shows the overall structure of transcript and protein data with respect to the tissue of origin for each cell line.

The structure of both protein and transcript data is highly correlated with epithelial and mesenchymal cellular states as indicated by the expression of E-cadherin, which forms a cell line band and a gradient in both data types.

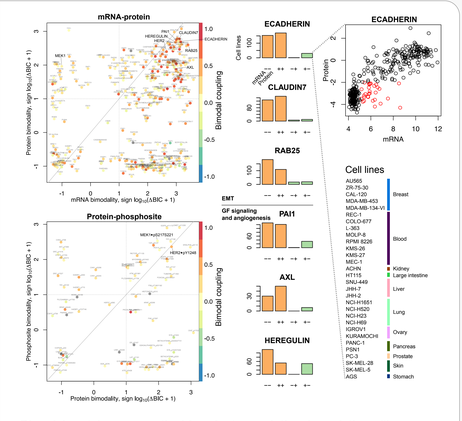
Pan-cancer bimodality of RPPA and EMT



Bimodal protein expression and phosphorylation across cancer types. By evaluating a two- versus one-component Gaussian fit using the Bayesian Information Criterion (BIC), 260 out of 450 antibody-based measurements were bimodal across cancer cell lines. The heatmap shows the probabilities of each cell line belonging to the up (+, red) or down (-, blue) distribution for the top-20 most bimodal proteins. The EMT-associated proteins E-cadherin and Claudin7 were among the most bimodal proteins suggesting a prominent role of EMT across the cancer cell lines, while validating the approach for identifying novel target proteins involved in cancer progression.

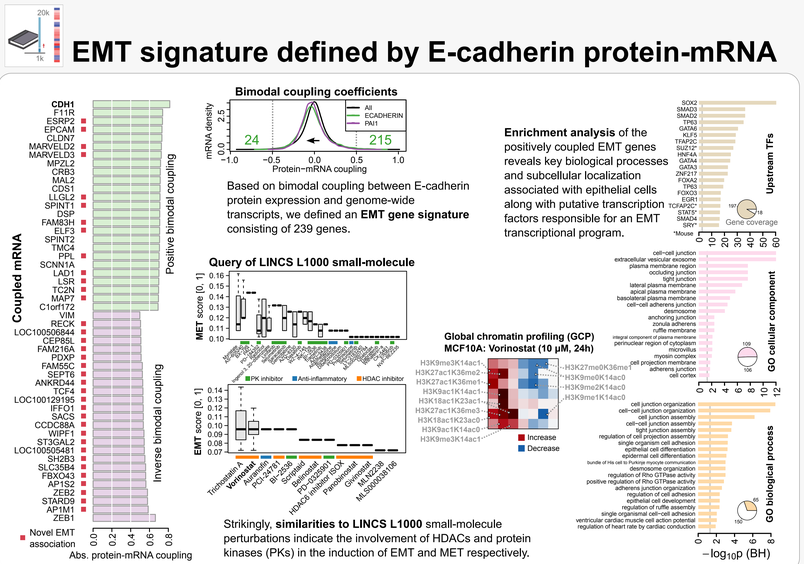


Proteins with coupled bimodality form communities associated with shared cancer-related hallmarks and include both protein abundances and phosphorylation states. Pan-cancer protein communities detected by Spearman's correlation (|r| > 0.3) of the posterior down (-) probabilities followed by calculating the leading non-negative eigenvector. Each community was manually named by investigating the biological function of its core components.

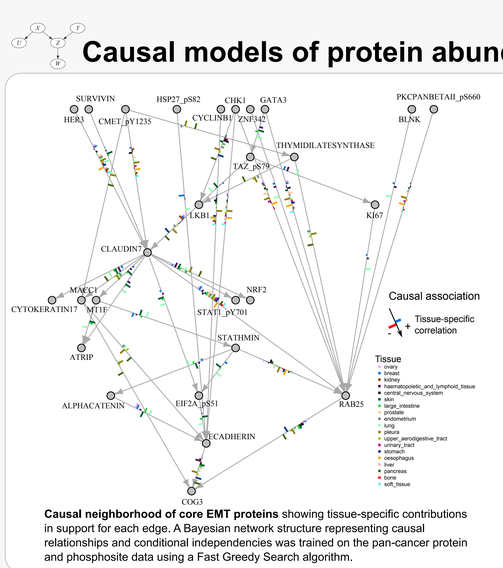


Bimodal proteins are mostly driven by transcriptional regulation with exceptions indicating the extent of translational or post-translational regulation. Comparing transcript and protein expression, the core members of the identified EMT community replicate both their bimodality and have coupled posterior probabilities of the low (-) and high (+) distribution. Nonetheless, in 30 cancer cell lines, high abundance of E-cadherin mRNA was associated with low protein expression indicating context-dependent transcriptional or post-translational regulation of E-cadherin and EMT.

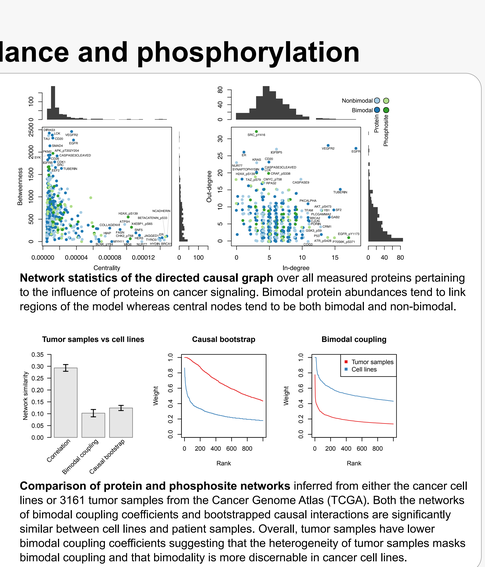
EMT signature defined by E-cadherin protein-mRNA



Causal models of protein abundance and phosphorylation



Causal neighborhood of core EMT proteins showing tissue-specific contributions in support for each edge. A Bayesian network structure representing causal relationships and conditional independencies was trained on the pan-cancer protein and phosphosite data using a Fast Greedy Search algorithm.



Network statistics of the directed causal graph over all measured proteins pertaining to the influence of proteins on cancer signaling. Bimodal protein abundances tend to link regions of the model whereas central nodes tend to be both bimodal and non-bimodal.

Comparison of protein and phosphosite networks inferred from either the cancer cell lines or 3161 tumor samples from the Cancer Genome Atlas (TCGA). Both the networks of bimodal coupling coefficients and bootstrapped causal interactions are significantly similar between cell lines and patient samples. Overall, tumor samples have lower bimodal coupling coefficients suggesting that the heterogeneity of tumor samples masks bimodal coupling and that bimodality is more discernable in cancer cell lines.

This work is supported by the NIH Common Fund grant #U54HL127624, BD2K-LINCS DCIC.