4-4-2019

Simulating Pathway-Based Steady States to Prevent Epithelial-Mesenchymal Transition in Ovarian Cancer

K Jeselle Clark
kjc2872@rit.edu

Simulating Pathway-Based Steady States to Prevent Epithelial-Mesenchymal Transition in Ovarian Cancer

K. Jeselle Clark

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Bioinformatics

Thomas Gosnell School of Life Sciences
College of Science

Rochester Institute of Technology
Rochester, NY
April 4, 2019

Rochester Institute of Technology
Thomas H. Gosnell School of Life Sciences
Bioinformatics Program

To: Head, Thomas H. Gosnell School of Life Sciences

The undersigned state that Kathy Jeselle Clark, a candidate for the Master of Science degree in Bioinformatics, has submitted her thesis and has satisfactorily defended it.

This completes the requirements for the Master of Science degree in Bioinformatics at Rochester Institute of Technology.

Thesis committee members:

Name	**Date**
Gary R. Skuse, Ph.D. Thesis Advisor	
Gordon Broderick, Ph.D.	
Matthew Morris, Ph.D.	
Maureen Ferran, Ph.D.	

Feng Cui, Ph.D.
Director of Bioinformatics MS Program

475-4115 (voice)
fxcsbi@rit.edu

CONTENTS

Table of Abbreviations

Abbreviation (alphabetical)	Term
BC	Betweenness Centrality
Co1	Cardinality of One
EMT	Epithelial-Mesenchymal Transition
GOI	Genes of Interest
HGSOC	High-Grade Serous Ovarian Cancer
IBAQ	Intensity-Based Absolute Quantification
MaxQB	MaxQuant DataBase
MIS	Minimal Intervention Set
MS	Mass Spectrometry
OC	Ovarian Cancer
PS	Pathway Studio

ABSTRACT

Ovarian cancer is a complex disease that involves gene regulatory dysfunction and that requires a systemic viewpoint to fully understand. Applying executable biology to ovarian cancer research and leveraging documented regulatory protein interactions, one can efficiently inform the prediction of characteristic gene-product activation using a logical model checking approach. Using this innovative approach to reducing terms and satisfying constraints, this thesis presents a strategy for applying regulatory systems biology to cancer research. By viewing ovarian cancer pathways like an electrical circuit, and constructing a pathway model with natural language processing tools, gene product expression patterns that have not been explained by traditional wet-bench biology are able to be predicted *in silico*. This research yields seven gene products whose perturbation is predicted to be sufficient to induce the epithelial-mesenchymal transition of ovarian cancer.

INTRODUCTION

I. Cancer Research Fundamentals

Simply described by Dr. Sol Efroni, cancer is a "family of gene-based diseases," and therefore must be studied at the genomic level in order to full understand its initiation, progression, and general pathology.[1] There are many barriers to entry for this direction of research, which all reduce down to the seemingly overwhelming amount of data. Researchers like Dr. Efroni utilize bioinformatics and computational approaches in order to overcome this hurdle. Besides computationally analyzing genetic sequences, there are other ways to rationalize the analysis of cancer data. Efroni et al. use a discrete state approach, working under the assumption that the exact levels of gene expression are less important than the knowledge of how they interact logically as a network. To study a whole family of genes at exact levels can be overwhelming and difficult to represent, therefore values are reduced to qualitative descriptions such as up and down regulated. For example, a qualitative probability of an interaction with gene A is determined by the probability of active an interaction multiplied by the probability of an upregulation of A plus the probability of inactive an interaction multiplied by the downregulation of A.[1] Researchers use this equation on each interaction in a network of genes in order to investigate whether a specific combination of active interactions leads to a cancerous state. This process is laid out like a pipeline, or guidebook, that should explain how to apply this approach to different cancers and different cancer stages.

In the study of pancreatic cancer, Gong reports on specific mutations of cell cycle regulators that increase the likelihood of cancer progression.[2] There are many ways to promote the progression of disease, but each person’s cancer will behave uniquely. The lack of a direct causal link is often attributed to the complexity of cancer as a collection of diseases. Unlike infectious disease studies, cancer research does not have such a standard by which to prove that a specific gene or mutation is the cause of a cancer. A microbe is relatively easy to prove as the cause for a given disease by following

Koch's postulates: isolate the same microbe from different infected individuals, culture the microbe *in vitro*, and successfully reproduce the disease in a healthy model by inoculating with the cultured microbe. Cancer is not understood in a similar manner because there are too many interacting entities involved; many genes are associated with cancer, but cannot be individually proven as a root cause. Additionally, cancers' very nature is a result of mutations, thereby making it somewhat difficult to reliably study long-term. Cancer is rarely caused by one instigator and a tumor can originate from multiple initiating cells or mutations. Finally, cancer is an aberration of self, and therefore is unique in every individual case. All of these qualities make cancer an extraordinary challenge for traditional reductionist and time invariant research methods.

II. Ovarian Cancer

Ovarian cancer (OC) is a lethal disease categorized by heterogeneous gynecological growth.[3,4] However, this definition does not adequately describe the far-reaching effects of this disease. For all of the emotion and widespread awareness, there is so much not fully understood about preventative measures, diagnostic avenues, and treatment protocols. Society does, on the other hand, understand how quickly it can spread to neighboring organs, and how it afflicts thousands of women in the United States every year. The Centers for Disease Control states that 21,429 new cases were diagnosed, and 13,920 women died in 2015 in the United States alone.[5] Around 90% of all deaths caused by cancer are do to metastatic disease,[6] and OC is referred to as the most lethal cancer. It is incredibly difficult to diagnose in an early stage because the ovary is so buried within the internal organs and there are so few symptoms.[7] For these reasons, the overall mortality rate is 60%, making it the most lethal gynecological malignancy.[8] Additionally, it is highly interconnected by vascularized tissue and the extracellular matrixes of other organs.

Therefore, metastatic spread is likely to occur, often before detection of the tumor. The field of OC research is desperate for new studies and meaningful breakthroughs. Because of the aggressive metastasis and complex heterogeneity, research into OC stands to provide insites into many cancers benefitting millions of families, not just those individuals who are afflicted by OC.

III. Ovarian Cancer Metastasis and Epithelial-Mesenchymal Transition

OC is most lethal once it metastasizes to other regions of the body, as surgical treatment of localized cancer usually results in patient remission.[6] Preventing spread is therefore an important step towards lowering the mortality rate of the disease. Metastasis is an extremely complex event in the pathology of any cancer and is not fully understood. In brief, metastasis can be described as the translocation of a tumorigenic cell to a different part of the body. Tumorigenic meaning a cell capable of replicating into a tumor in a different environment from where it originates.[9,10] However, that movement involves several activation steps, including: disengagement and escape from the original cancerous tissue, physical movement to a new tissue usually utilizing blood or lymphatic vessels, and attachment to a new tissue upon which the tumor is able to grow and divide.[6] Each of these steps involves numerous protein interactions that may inhibit or promote the process.[7]

This first step of metastasis is called the epithelial-mesenchymal transition (EMT) and it is the generalized focus of this project. Encompassed within that process is a cascade of events that convert a tumor cell from an epithelial (local, differentiated) state to a mesenchymal (loosely associated, undifferentiated) state. Unfortunately, a mesenchymal cell is more capable of surviving away from the tissue from which it is derived, thus making metastatic migration more likely. This is the reason for studying EMT, in the hopes that further understanding may revolutionize the way researchers and clinicians think about cancer diagnostics. If cancerous spread is detected earlier or even predicted by means of risk, our understanding of cancer could be dramatically altered. This project hopes to identify

protein drug targets for anti-metastatic OC treatments in order to increase the effectiveness of surgical treatments.

IV. Clinical Improvements to Current Treatment Methods

The most common screening method is a pelvic exam, performed via external palpation. Using this methodology, a physician may not detect the tumor if it is not large enough, or if the tumor is large enough to detect, it may already be invasive. Late-stage diagnosis usually means the treatment must be aggressive, and will rely partially on luck to find and kill every metastatic cell. Once the tumor is successfully diagnosed, a label is associated with the tumor that denotes which stage of the EMT the tumor has progressed into.[3] An epithelial tumor is localized and does not show signs of invasion or angiogenesis, and can often be treated quickly with surgery or radiation therapy. An advanced stage OC will usually undergo chemotherapy and cytoreduction, a form of surgery that attempts to remove as much of the malignant tumor as possible, with the understanding that the physicians will likely not excise every cell.[11] The term “cancer free” has a loophole. It is currently impossible to be sure that a tumor has not metastasized before surgical treatment, because it only takes one mesenchymal cell to spread a tumor.

To illustrate this project’s goal with regards to current treatment methods, one can imagine OC like a forest fire. When fighting a wildfire, the first goal is containment, not merely spraying water on the flames; if the fire cannot spread, it is more easily extinguished while also limiting the damage it causes. Similarly, when treating a wildfire of ovarian metastasis, it is essential to first limit the spread of the disease. If positively limited to just the initial site of the tumor, surgery and targeted radiation therapies become more effective. This project focuses on

the containment of illness and will attempt to further understand the EMT pathway, in the hopes of finding druggable target(s) directed at inhibiting metastasis in the future.

V. Computational Improvements to Current Research Methods

In the interest of improving experimental *in vitro* models of OC, Bowtell et al. studied the genetic and transcriptomic composition of specifically high-grade serous ovarian cancer (HGSOC) cell lines.[11] HGSOCs are generally identified as the most advanced stage, metastatic OCs. This research allowed them to assign cell types to clusters based on phenotypic behavior and genotypic irregularities and mutations. They studied and organized cell lines based on microRNA regulation, tumor microenvironments, and angiogenesis presence, among other analyses. Researchers like Bowtell et al. cite the lack of integration of these methods as a limiting factor in the continued progress of clinical research. All of these methods of study are so different, and do not have much overlap in reliability or focus. This is one limitation of current experimental methods.

Francavilla et al. utilized the breakthroughs of improved mass spectrometry (MS) accuracy and methods in order to model the proteomics of OC, including how proteins are activated or inactivated via phosphorylation.[4] Using patients' cells, they are able to compare how proteins are differentially present in epithelial cancer cells versus HGSOC. Whereas many other datasets compare healthy to cancerous data, this research attempts to look more specifically at cells to determine what entities are causing the development into more aggressive cancer types. Adding in systems biology modeling, researchers are able to analyze how entities differentially interact, in addition to their differential abundance. They found that CDK7 (cyclin dependent kinase 7) affects the proliferation of epithelial OC by negatively regulating the cell cycle. None of this analysis can be considered conventional, because it uses standards from so many sources and prior publications, but it combines them in a novel way. There is no standard or guideline agreed upon for the analysis of proteomic cancer data so there is

no accepted way to determine how reliable their analysis is. In the absence of an accepted standard of practice, research papers like this must exhaustively demonstrate their point.

Many established methodologies are based on protocols proposed by papers that became the accepted norme. One example of this widespread agreement is that of high-throughput RNA-Seq. Many researchers follow the same protocols proposed by Zhong et al. and Wang et al. to acquire and pre-process data.[12,13] The contrast, however, is what worked for the establishment of lab-bench methods that researchers accept around the world, is not working with computational methods. There are so many different ways to analyze high-throughput data, and yet there is no consensus of what works best.

The last few decades of cancer research have been transformative for key research technologies. As data becomes easier to collect through the implementation of omics tools, there is an ever-growing need for efficient processing of large datasets. The most efficient way to process data is with computational methods like next-generation sequencing analysis pipelines, algorithmic prediction modeling, and the systemic representation of disease. These have all been applied to ovarian cancer studies, with varying degrees of success.

VI. Application of Systems Biology

Clinical research often naturally takes on a narrow focus, such as identifying the gene that causes a genetic disorder, or isolating the bacterium that causes an infectious disease. That same viewpoint cannot be used when studying cancer as there does not appear to be a single cause that results in cancerous growth or malignancies. If that were the case, cancer would be much easier to study. Unfortunately, the cause and effect relationships are more complex.

Therefore, cancer researchers must apply a more broad-scope view of the issue, hence the increasingly common application of bioinformatics and computational tools.

Systems biology is an integrated research methodology that studies cancer from a bird's eye view. A systems biology approach attempts to look at as much of the biology as possible all at once, and it uses dynamic analyses of cell regulation to illustrate exactly how a cell changes as cancer progresses. By mapping the interacting components of patients' cells as nodes of a modified circuit diagram, one can visualize how the system changes from healthy to cancerous to metastatic. Viewing biology as a circuit of chemical interactions allows even small perturbations of the system's components to be studied. This is arguably the best way to study cancer since several small perturbations are able to combine and lead to a progressively cancerous state. However, the systems biology approach may rely heavily on a pathway schema as it exists while the cell or patient is healthy.[6] This is not enough to study disease, as disease can often dramatically change how the components of some pathways interact. Some newer studies have modeled what happens when entities are inactivated showing how that can lead to a disease state,[14,15] which is part of the inspiration for this research into OC.

VII. Steady States

When relating cancer to a circuit of molecules and interactions, it is important to understand that biology often follows Occam's razor. In other words, cell biology can seem like a chaotic system while attempting to understand it, but there are points of equilibrium that are found naturally. This phenomenon of biology existing as a series of low-energy equilibrium states is defined by Wooten et al. as "basins of attraction."[16] Wooten's basins can be described by the metaphor of a ball rolling around on an uneven surface with lots of hills and valleys interspersed. The ball (a patient's cell) will stay in the most local valley (basin of attraction) until moved by a significant enough force to push it into a neighboring valley. These valleys are essentially states of being where the patient's health is not

significantly changing, or is at equilibrium. If each valley is a state of being, some valleys may be labelled healthy, and others in varying degrees of diseased. In order to get over a mountain and into a neighboring diseased valley, a significant change must occur to the cell or to the landscape. The cell could be altered by mutation, such as the activation of an oncogene or inactivation of a tumor suppressor. In theory, the landscape may be altered by the administration of a drug or a dramatic change in lifestyle, and that new landscape can alter the likelihood of moving to or staying in diseased valleys. This project will refer to these valleys (aka. basins of attraction, equilibrium states, etc) as “steady states.”

VIII. Modeling Ovarian Cancer

Table 1: Ovarian Cancer Descriptors

Type	Epithelial = outer lining of the ovaries affected	Germ = gametic pre-cursors	Stromal = hormone controlling cells connected to the ovaries	
Stage	1 = Confined to the ovaries	2 = Migration to the surrounding organs	3 = Tumors found in the abdomen smaller than 2cm	4 = Cancerous cells are found in the lungs, liver, or spleen.
Grade	GB = low malignant potential	G1 = many differentiated cells	G2 = moderately differentiated cells	G3 – G4 = mostly undifferentiated cells

There are many ways in which people describe the progression of cancer. OC is commonly labeled with stages, grades, or types, as summarized by Table 1. The type of OC predominantly describes the origin type of cell. The stage of cancer is a generalized proxy for how invasive it has become, where a stage 1 is the lowest and typically most easily treated and stage 4 is considered difficult because it requires a system-wide treatment such as chemotherapy. The grade of OC also describes how much the original cells look like their predecessors, utilizing

the concept that cancer cells generally progress towards a less differentiated state. Therefore, a higher grade indicates less differentiated cells, and more abnormal-looking tissue.

This project moves away from these models that do not fully describe OC on the cellular level. Instead, each tissue sample is organized into a group based on cellular protein levels. Datasets that are able to be clustered in this way are the desired types to follow this workflow and can be applied directly to pathway modeling. Essentially, each clustered group of tissue samples is able to be treated as steady states. This project simulates how a cell may convert from one clustered tissue type into another.

IX. Proteomics

This project is focused on the analysis of protein data, as opposed to RNA quantification for a few reasons. Although RNA quantification methods like RNA-Seq provide useful insights into what machinery the cell is using and what genes are being expressed and to what degree, RNA strands may be used many times by the cell to create protein products. This means that the final counts for products cannot be fully extrapolated by RNA studies. Further, when studying cancer, researchers should assume that any number of failed cell machineries can lead to a diseased steady state. For example, post translational modifications and aberrant translational machinery lead to changes in final protein counts, both of which could increase the likelihood of cancer progression. Oncogenes and tumor suppressor genes can conceivably be activated or suppressed during translation, a concept that cannot be studied with RNA quantification alone. For these reasons, proteomics databases are considered favorable.

Unlike RNA-Seq data, where RNA molecule abundance is recorded as a count, there is not a widely standardized a way of analyzing proteomic data. This is because mass spectrometry (MS) does not return molecular counts, but spectra. Each spectrum must be normalized in order to compare to any other MS spectra. Many programs use slightly different cutoffs while analyzing noise and peak widths, leading to the lack of consensus in proteomics research. This project uses intensity-based absolute quantification (iBAQ) scores, an absolute measurement of protein abundance normalized by peptide

detection intensity levels, as a proxy for RNA-Seq expression counts.[17] This method is referred to as a label-free quantification method.

X. Discretization

IBAQ scores are a continuous scale on the order of hundreds of millions and make simulation trials computationally exhausting. Therefore, it is necessary to simplify the problem by reducing iBAQ scores into discrete terms, a process referred to as discretization. This is done using simple statistical significance tests. IBAQ scores that are not significantly different between the two cancer groups for any given protein are left unconstrained in the model. Meanwhile significantly different groups are given a representative binary code, where 0 represents significantly lowered level of abundance, and 1 represents heightened abundance.

XI. Employing Constraint Satisfaction Techniques

This project uses this discretized proteomics data as constraints for determining what steady states exist within the pathway model. Simplifying the problem by introducing constraint satisfaction programming is currently the preferred way to minimize the computational time and resource dependancy when solving exhaustively.[18] These constraints ensure alignment of model predicted behvaior with the proteomic data. This is not an easy task especiually in larger networks because the problem is exponentially complex; it is possible that every entity in a pathway model of size n interacts as an activator or an inhibitor with every other entity, therefore the size of the operation space is represented as $O(2^n)$. If this project did not reduce resolution into a binary representation, but instead discretized the iBAQ scores to d levels of expression, the problem would be $O(d^n)$ large. Using proteomics data to constrain the expected behavior of the regulatory circuit, a simulation algorithm based in backtracking logic can

compute how each entity must influence the circuit. This is output as a solution set consisting of logical decision weights and the expression threshhold values for activation tthese regulatory actions. For example, if a node is known to stay activated and is known to be able to activate two other nodes, the solution set is partially solved. Exponential time problems can be extremely difficult to work with but there is no current way to avoid this complexity without compromising the thoroughness of the search for these parameter estimates.

XII. Parameterization and Populating Models

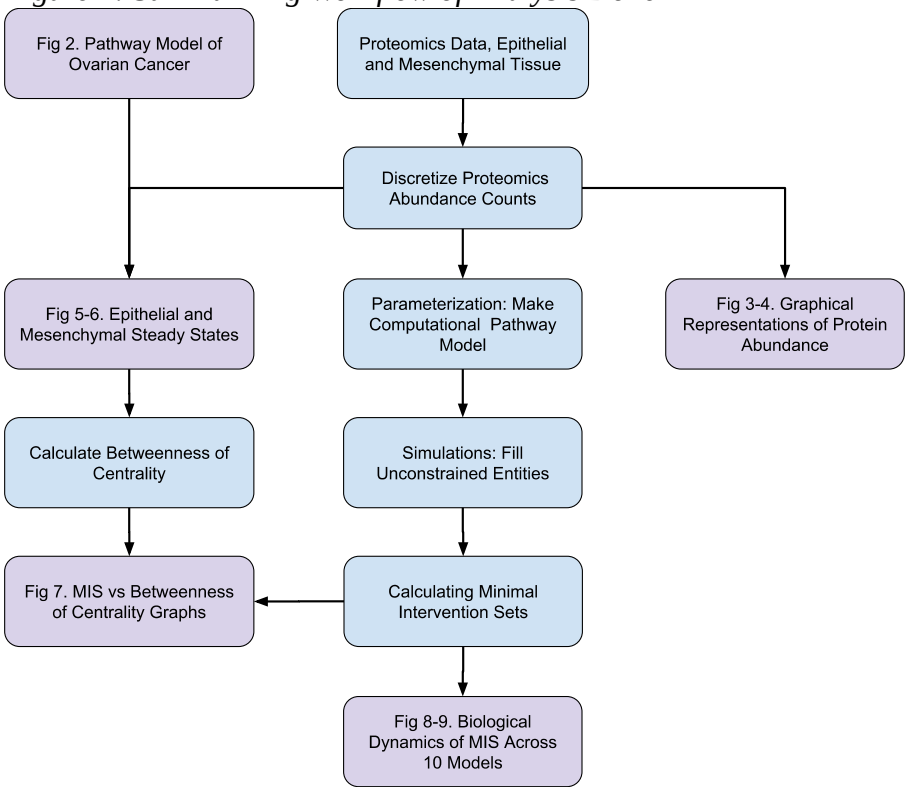
This research uses the tool Bio-ModelChecker, developed by the Center for Clinical Systems Biology at Rochester Regional Health for the constraint satisfaction of logic models and their simulaiton of how healthy and disease states behave.[19] Requiring a pathway map in the form of a JSON with an adjacency matrix, transition parameters, and incomplete steady states defining initial constraints, Bio-ModelChecker can simulate how perturbations in the network determine the state of unconstrained nodes. This robust simulation uses a decision tree to describe the biological plausibility of possible outcomes while reducing terms to manage computational efficiency and optimize informative power.

XIII. Minimal Intervention Sets

Casting these results into a more clinical perspective, it is understood that differences between cell types does not immediately offer a novel treatment solution. In order to understand how one might prevent EMT in a patient, it is first necessary to understand how EMT occurs naturally. Minimal intervention set (MIS) simulations attempt to answer this question by introducing a certain amount of noise into the circuit and simulating how a cell might transition from the epithelial steady state to the mesenchymal state.[20] The goal of a “minimal intervention set,” consists of finding a node in the pathway map or a small subset of nodes that influence the predicted dynamic behavior such that a new stable behavior or phenotype is acheived. For instance, if it is known that node A controls node B, and

B is responsible for metastasis, then one MIS solution is the activation of node A, as that one change is able to induce metastasis. Ideally, this program finds a singular node that is able to switch the network single-handedly from epithelial to mesenchymal and back. Returning to the idea that pathways in biology can be modeled like an electrical circuit diagram, an MIS is like a switch that activates or inactivates the rest of the circuit. Therefore, the simpler the MIS, the greater potential for a good drug target.

Figure 1: Summarizing Workflow of Analysis Done



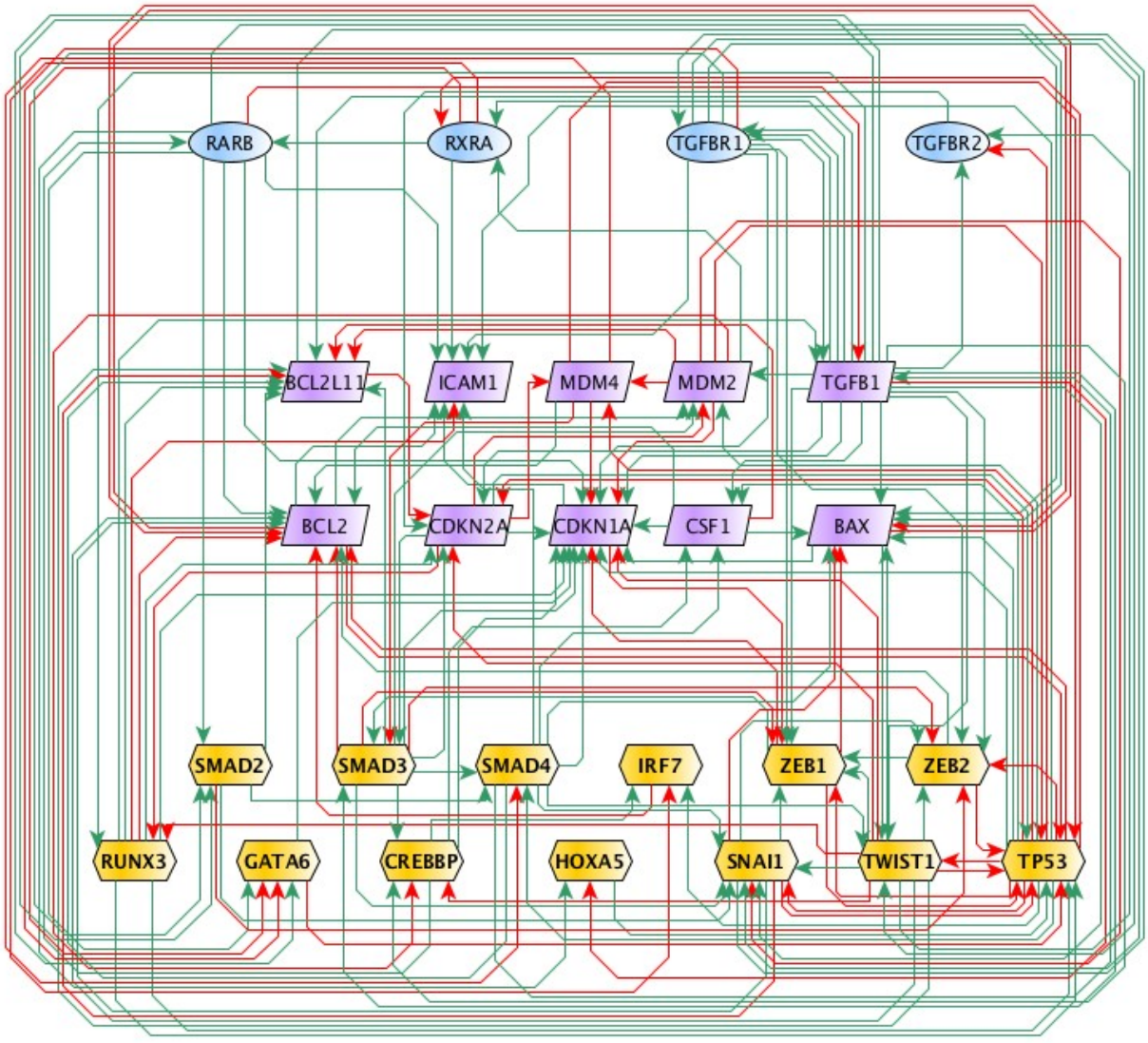
RESULTS AND DISCUSSION

I. Genes of Interest

This project focuses on the gene product interactions that cause EMT. To study this underlying genomic influence, pathway modeling tools were used to generate a map of

interacting proteins. The proteins in this map make up the list of gene products that will be focused on throughout this project, referred to as the genes of interest (GOI) (Figure 2). Protein types are represented by color and shape: blue circles are receptors, purple parallelograms are signaling molecules, and yellow hexagons are transcription factors. This list of proteins was generated stepwise, starting with the core set of genes proposed by Lu et al., including SNAIL, ZEB, SMAD3/4, P53, MDM2, and TGFβ.[10] This list of six applies generally to all cancerous EMT, and serve as the "core genes" for this project.

Figure 2: Core Genes Pathway Model



Various language processors, databases, and pathway modeling tools were used to expand the core genes into a more comprehensive map of cellular interactions, while narrowing the scope to only OC. First, core genes were confirmed by verifying connections between entities using Pathway Studio.[21] Pathway Studio was also used as a natural language processing tool to search the Elsevier database for any further connections between the core genes. Next, Mogrify was used to add a list of transcription factor encoding genes shown to have significantly altered transcript concentration between OC mesenchymal precursor cells and metastatic cells: HOXA5, ICAM1, GATA6, RARB, IRF7, and RXRA.[22] Reactome and String were also used to contribute interconnected entities from their respective databases. Only two entities were added to the dataset by hand based on new literature: CSF1 and TWIST1.[6,23–25] Every entity was assembled into one pathway in Pathway Studio, and was supported by references in the Pathway Studio database. This yielded the project a complex dataset of 27 genes, all supported by at least 2 sources (Figure 2). Even though the connections in this pathway were verified by at least two sources, they were still often found via natural language processing and thus were also checked by hand.

II. Reference Checking

It is important to remember that if natural language processing software or text-mining in general is used to build a pathway model, the references cited behind interactions must be human verified. This pathway originally consisted of 8,788 references spread unevenly across 176 total connections. After hand-checking each connection for accurate supporting references, 21 relations were reversed or deleted, resulting in a text-mining success rate of 95.5% (Appendix A).

III. Proteomic Data

Acquired from the MaxQuant DataBase (MaxQB) under ID P017, extensive proteomic data describing OC cell lines was applied to the regulatory network containing the 27 GOI. This data was analyzed by Coscia et al., carefully separating the cell lines into clusters of epithelial and mesenchymal samples in 2016.[17] This project utilizes the clustered groups defined within the database: Group 1 being epithelial, Group 2 and 3 being clumped together as they are both contain mesenchymal cell lines.

Included in this dataset are proteomic surveys of 67 distinct cell lines, including 26 OC cell lines, some HGSOCs, some immortal ovarian epithelium, and some fallopian tube epithelial cell lines. The products of approximately 21,000 genes are included in this collection, surveyed by single run MS. Protein expression profiles of all of these cell lines were grouped by unsupervised clustering into epithelial and mesenchymal phenotypic classes. The groupings were verified by Coscia et al. against a separate principal component analysis of clinical samples from the Cancer Genome Atlas, before being added into the MaxQB online database tool.

Figure 3: Histograms of iBAQ Scores of Tissue Samples (35 bins)

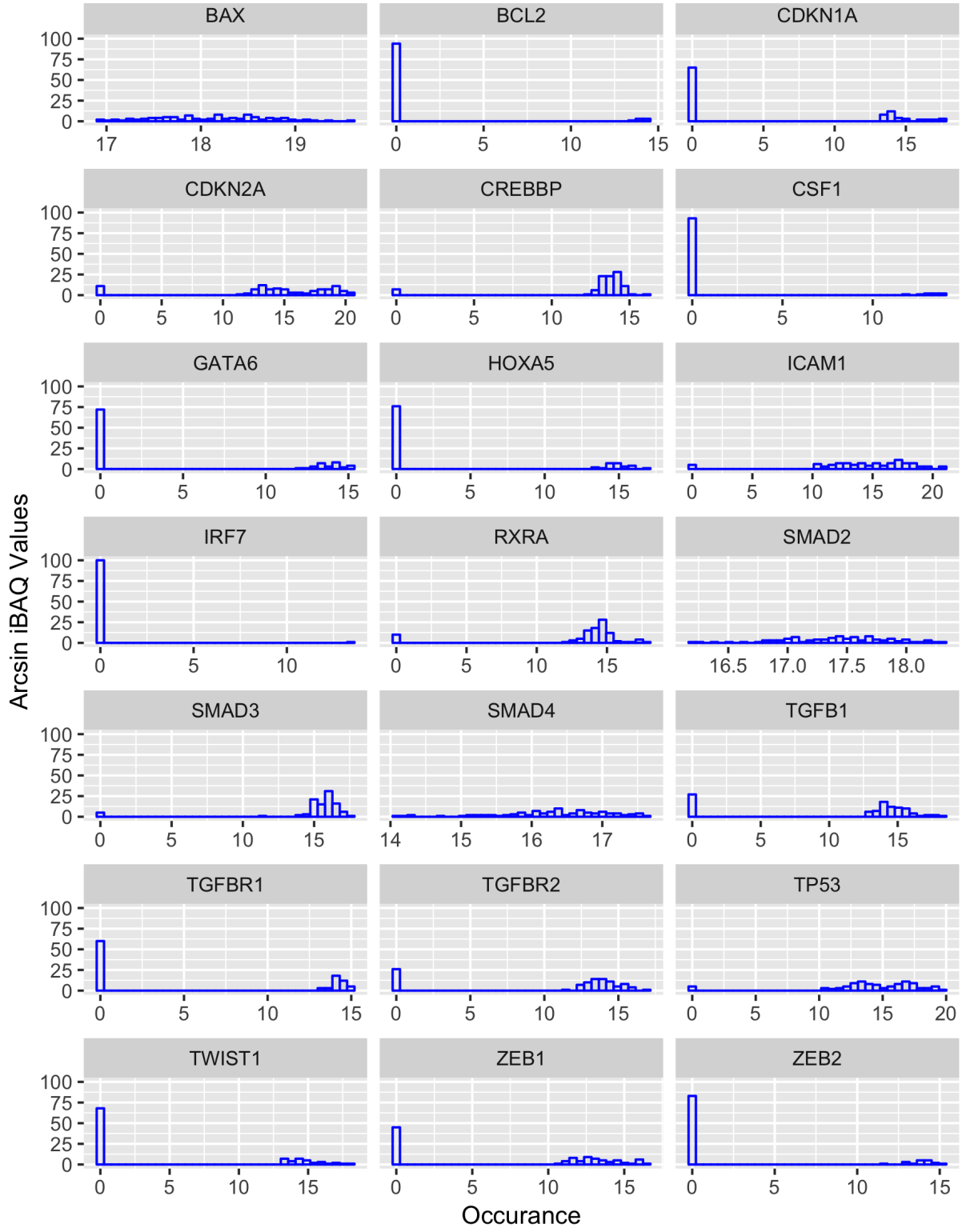
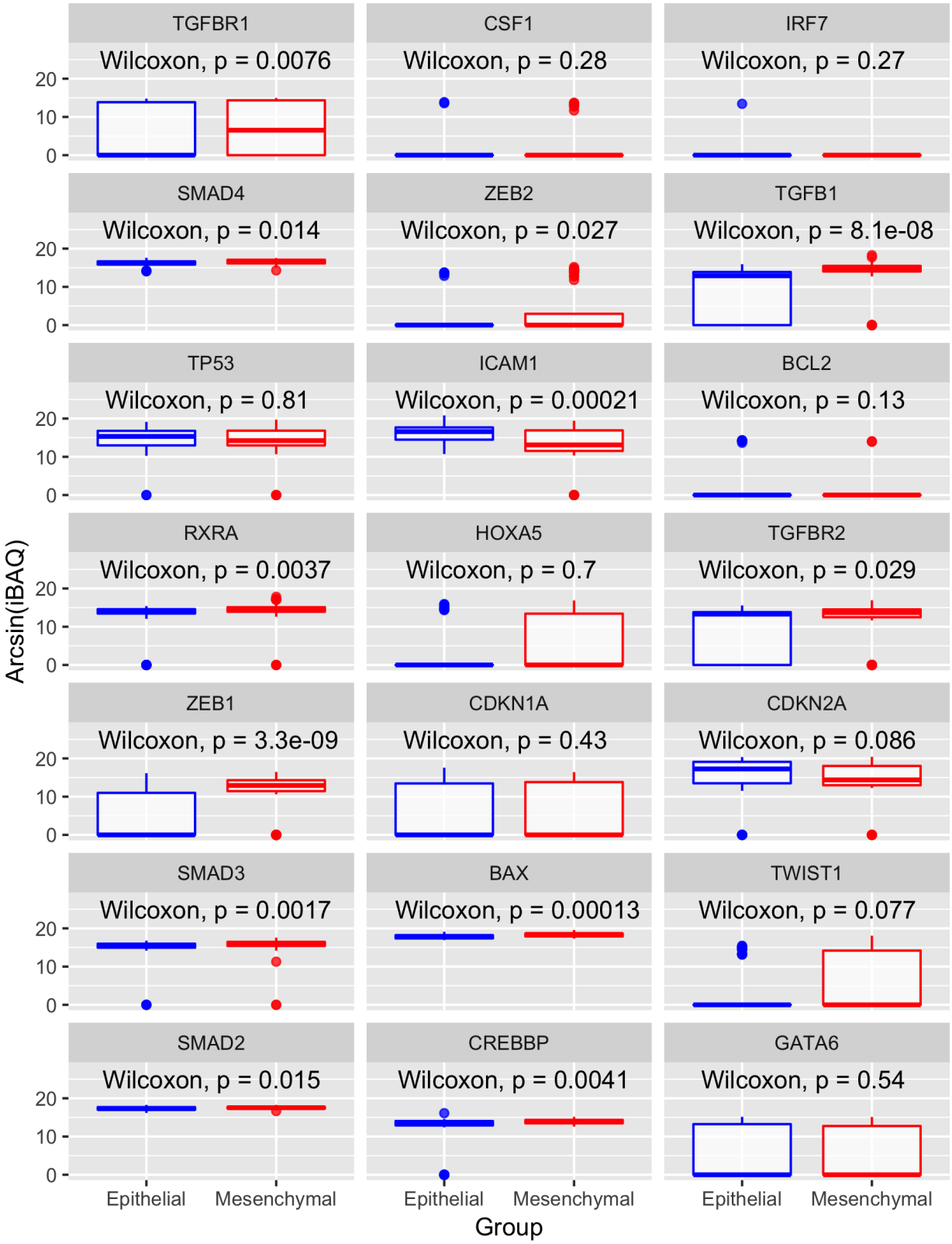


Figure 4: iBAQ Score Boxplots and Wilcox Significance Tests



IV. Discretization

As shown in Figure 3, the distribution of each GOI varied, but often followed an apparent bimodal curve. This is seen most clearly in proteins TGFβ1, ZEB1, and CDKN2A. This supports the hypothesis that protein abundance changes significantly for the GOI in epithelial versus mesenchymal OC. However, there is not yet proof that the cancerous cell type is correlated with the proteins showing bimodal distributions. In Figure 4 the GOI are segregated into database-defined cell types and plotted in side-by-side boxplots. The Wilcox significance tests are given above each panel.

Based on these significance tests, 12 of the 27 GOI are quantitatively different across the two cancer cell types: TGFβR1, ICAM1, SMAD3, RXRA, BAX, SMAD4, ZEB2, TGFβR2, SMAD2, TGFβ1, ZEB1, and CREBBP. Biologically speaking, this subset of 12 is more likely to play a causal role in the progression of OC, with the caveat that approximately six proteins were not included in the proteomics database. From this point forward, this subset of genes is constrained to experimental data when estimating parameter values for this model to predict the unknown protein abundance levels.

V. Parameterization

Because this data is clustered into two groups, and is not time delineated, it is not possible to score returned steady states. Without a way to order solutions by a time-oriented objective function, all steady states returned are therefore equally likely. The steady state determination was allowed to run for 45 min, and returned 339 solutions, all of which supported identical steady states. This fact grants validity to this model since the solution set could easily have been composed of 339 unique sets of steady states. It was assumed that 339 solutions was a large enough sample to be representative of all solutions, especially since all steady states

coincided. These two steady states are representative of the epithelial and mesenchymal states represented in Figures 5 and 6 as the color of each entity.

The second parameter set was used to determine which interactions were most crucial for supporting stable behavior for the two cell phenotypes of interest (i.e. epithelial and mesenchymal) as steady states. This was done by allowing the algorithm to solve the problem of reproducing these two measured steady states with the added complexity of promoting the simplest possible model. This was achieved by modifying the algorithm to prune regulatory interactions in these pathways as often as possible while still supporting stable dynamics at these phenotypic steady states. This exercise does not result in the most informative survey of steady states, but it does highlight what pathway edges are generally the most informative for the stability of all steady states. This core model analysis is more computationally complex due to increased variability associated with the iterative editing of the model circuit architecture. The “ordering” parameter is then reset to indicate no ordering, thus dropping the connections which are uninformative for the current stage of simulated models. This minimization of pathway connectivity is done to determine the minimal necessary set of pathway interactions capable of supporting the requisite behavior. This larger and more complex optimization problem requires much more time to run. The maximum allotted run time was therefore set for 12 hours in order to yield a sufficiently large sample of edges, from which 42 solution sets were returned. This second solution output was used to calculate the frequency at which each edge appears. The frequency by which an edge is included in the pathway circuit is a proxy for its relative importance. This metric of interaction confidence or criticality is represented in Figures 5 and 6 as the width and blue-grey color gradient of the edges, where thicker blue lines indicate the highest relative importance for the model’s stability.

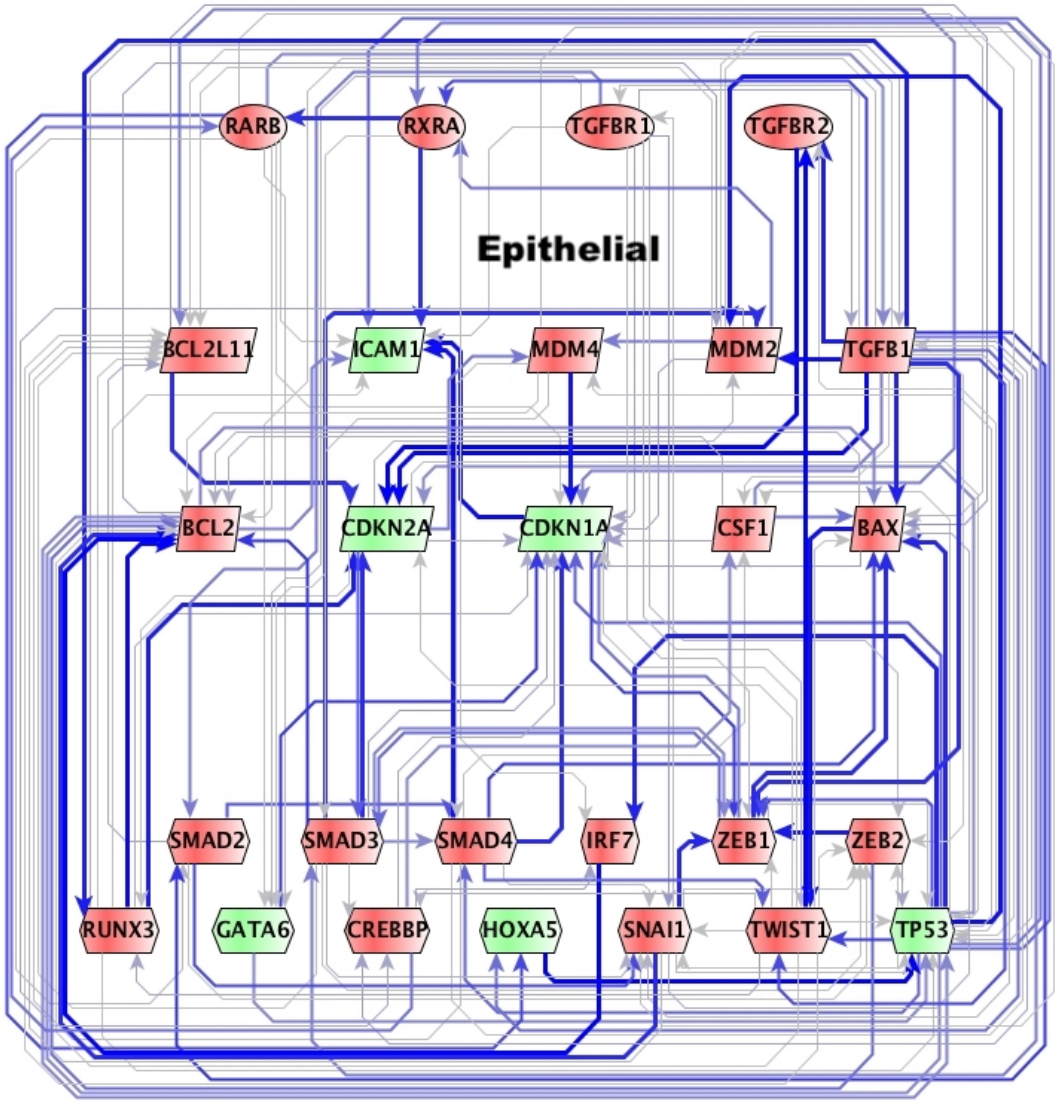
One caveat to this methodology is the assumption that these solutions are a representative sample from a population of all possible solutions. With unlimited resources and time, perhaps thousands of additional solutions may have been calculated and returned. It is assumed, however, that

since all solutions agreed completely on the end steady states, the overwhelming majority of the population would have have supported this pathway circuit architecture with the differences in decisional logic supporting multiple transition dynamics that nonetheless all settle in the same terminal steady states.

VI. Steady State Solutions

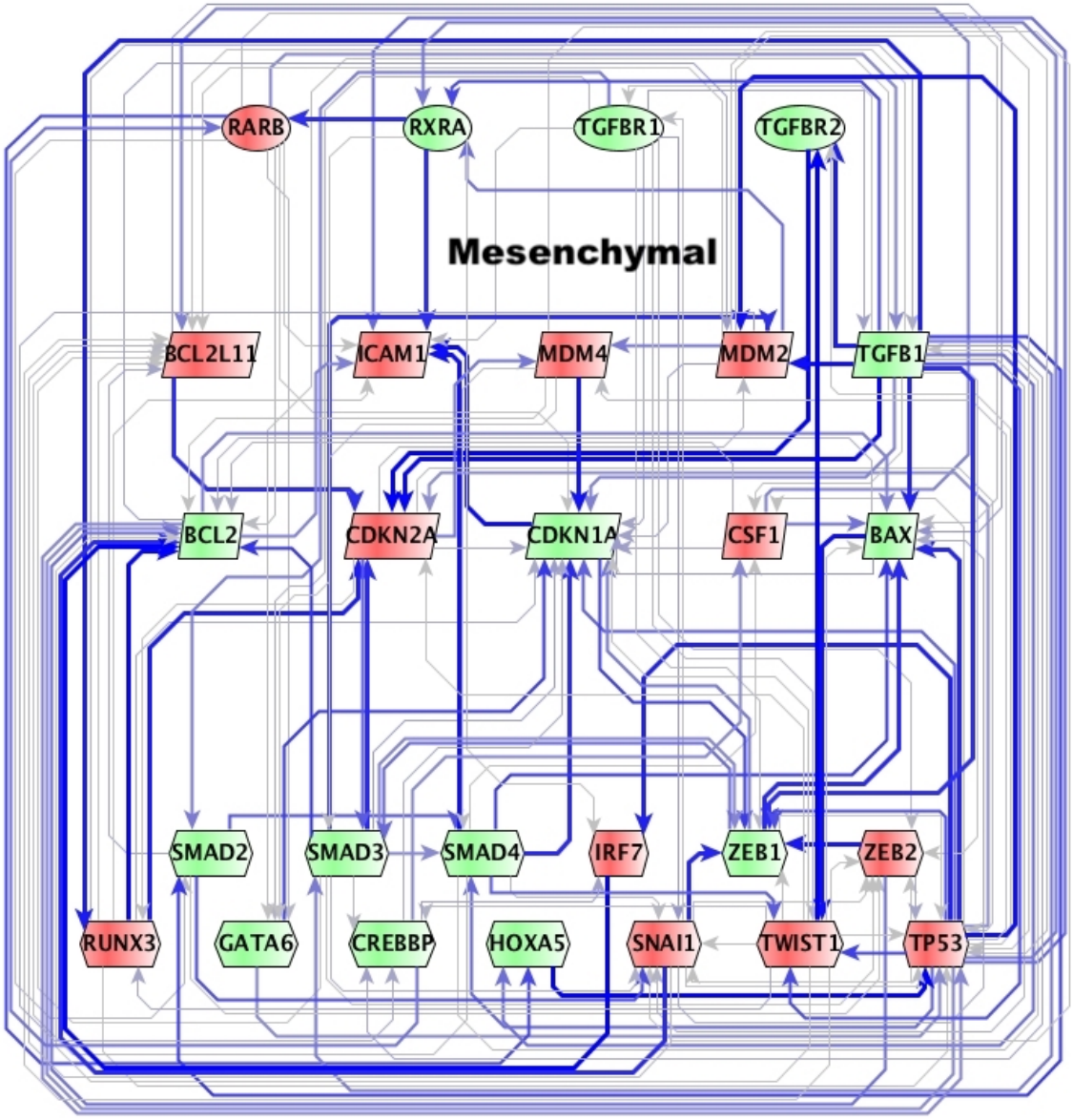
The epithelial steady state is shown in Figure 5 based on constraints derived from experimental data applied to 12 of the 27 entities in the model. The stable expression of the other 15 markers were predicted such that a stable phenotype was supported. Green entities are activated genes, and red entities are inactivated genes. The blue color vibrancy and thickness of the connections represent how informative those interactions are on the model, as described above in Section V. The model shows a much easier way to visualize the differences in protein abundance that were seen in Figure 4. This allows researchers to understand the circuit logically, at the expense of some resolution. For example, now it is easier to see that the SMAD and TGFβ circuits are activated during EMT.

Figure 5: Epithelial Steady State



The mesenchymal steady state is shown in Figure 6, following the same color scheme as Figure 5. The differences between the two graphs are the activation of RXRA, TGFβ1, TGFβR1-2, SMA2-4, ZEB1, CREBBP, BCL2 and BAX. Additionally, TP53, ICAM1, and CDKN2A were inactivated in the mesenchymal state. Most of theses differences were observed phenomenon, but BCL2, TP53, and CDKN2A were predicted by the model.

Figure 6: Mesenchymal Steady State



VII. Solution Sampling

While all of the 339 model solutions in the steady state determination agreed, each was a unique solution because there are subtle variations in regulatory logic. Those variations support different phenotypes outside the narrow context of the constraints. This speaks to the defining influence of pathway connectivity in determining available steady states. Each of the allowable transition paths between steady states supported by these different models is expected to vary analysis of individual model dynamics would detract from the primary focus of this work, namely terminal cell phenotype. Therefore, a random sample of 10 models was taken to proceed with further analysis. The next steps consist of identifying which molecular pathway elements might constitute potential triggers of metastatic transformation. Simulations are run to answer this question, wherein each model is evaluated for how it arrives at the same end state, which elucidates the key potential mediators of EMT.

VIII. Minimal Intervention Sets for Inducing EMT

The MIS routine available in BioModelChecker was used to search for the most influential trigger nodes in the network. The network was input and the MIS objectives were defined as a set of pathway or gene products that could drive the state of the pathway model towards EMT in less than 20 steps, given that the number of simultaneous targets had to be less than or equal to 6. These choices were made under a few assumptions. First, any potential EMT trigger this research may identify must be specific so as to not occur simply by biological noise and not cause undesired regulation of downstream entities. If 6 or more entities are needed to trigger EMT, then metastasis is not likely and a drug target will not be needed. Second, 20 transition steps should be more than enough for a EMT trigger to be influential in changing the network. Any additional transition steps would suggest an important contribution of biological noise on the outcome or that the trigger does not occupy a central enough role to act as a programmable switch for the network. Upon analysis of this network, it was found that these assumptions were conservative and did not limit the results. Further, the MISs were so

extensive, numbering 288,408 unique solutions. This was such a large number of solutions, that the set of EMT triggering proteins was filtered to a maximum size of 5. This resulted in 69,301 possible unique solutions triggering EMT. Within the MIS solutions, seven proteins were able to individually trigger EMT.

IX. Ranking of MIS Candidates via Betweenness Centrality

Betweenness centrality (BC) is a concept in circuit diagrams that equates to the criticality of a specific node to the interconnectedness of the network. However, BC is not just a count of the number of connections leading to or away from a node, but it also takes into account a node's ability to be a linker and shorten all other paths traversing the pathway. Applying this to systems biology, BC is a calculable value representing how much a single node is traversed when navigating the pathway model. A node with a high BC is one that increases the relative accessibility of the rest of the pathway network. This is biologically relevant if researchers are trying to target a specific pathway with a drug. A protein with a higher BC is more likely to be involved in the drug interaction, whether direct or indirect through other nodes. The BC was calculated for each protein in GOI set and represented graphically on a horizontal axis (Figure 7). A vertical axis was added to stratify the frequency at which proteins were present in the MIS candidates. In other words, the horizontal axis is a proxy for how important the proteins are to the network stability and drug targeting, and the vertical axis is a proxy for how important the proteins are to causing EMT.

Figure 7: MIS Abundance vs Betweenness Centrality

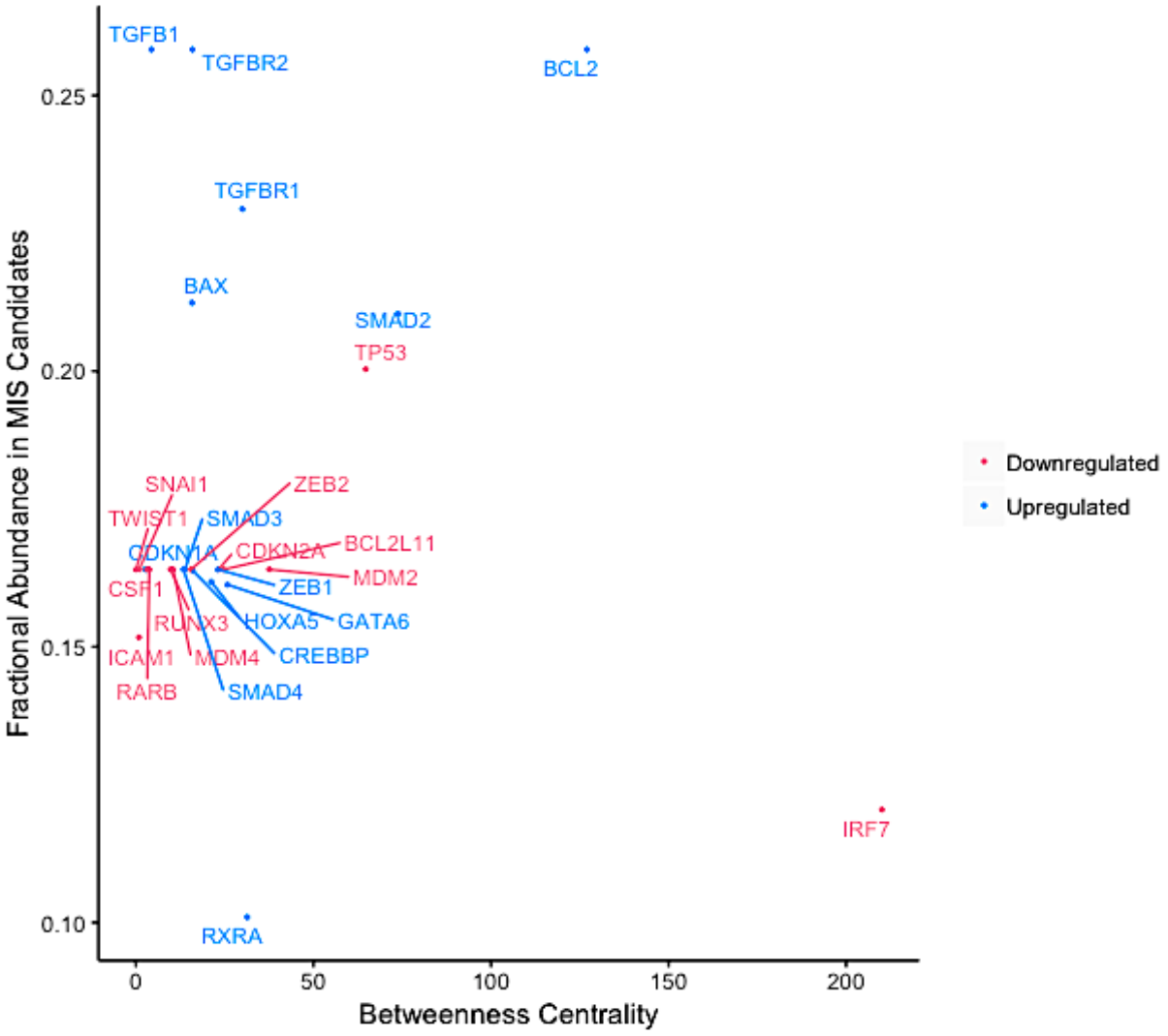


Figure 7 shows that upregulation of TGFβ1, TGFβR2, and BCL2 are part of the solution set in just over 25% of MIS candidates for the random sample of 10 models. TGFβR1, BAX, SMAD2, and TP53 are present in 20-25% of the MIS candidates. Most proteins are present in approximately 16% of solutions, indicating a basal level wherein it is assumed that the noise of the network could be responsible for triggering a metastatic equilibrium state. These are all nodes that were previously seen to be activated when going from the epithelial state to the mesenchymal state. However, some of the nodes that were also switched are not found in many MIS solutions such as RXRA, SMAD3-4, ZEB1, CREBBP, ICAM1, and CDKN2A. RXRA, however, stands out as being present in far fewer solutions.

This could mean that it requires many biological steps to be regulated and it was therefore rarely included by the algorithm's step cutoff.

The interconnectedness of the proteins in the model are depicted in Figure 7. IRF7 stands out as being extremely well traversed, yet seemingly less important to the process of EMT at first glance. Applying biological context to this phenomenon, it is already understood that the inactivation of IRF7 is associated with the initiation of many types of cancer.[26] As this pathway model represents the development and progression of cancer, as opposed to its initiation, the lack of EMT influence is understandable. Based on Figures 3, 4, 5, and 6, IRF7 is inactivated during oncogenesis, but does not appear to be further up- or down-regulated during EMT. This is hugely encouraging for future research into the systemic changes that occur during cancer initiation, but is not the current focus of this project.

It is also notable that BCL2 has a moderately high BC relative to other proteins in the model, and is seemingly one of the most important, or at least frequently selected, nodes for inducing EMT. This indicates that BCL2 is highly involved in the conversion of cells into a more invasive subtype, a notion that is supported by literature. BCL2 is currently defined as a proto-oncogene with an anti-apoptotic function, an attribute that is already being targeted by the chemotherapy treatment Venetoclax for chronic lymphocytic leukemia.[27] For these reasons, BCL2 is likely the single best candidate for anti-metastatic OC drug treatment as it seems to be so influential in ovarian EMT, and already has an approved targeting drug currently not being employed for OC treatment.

X. Analysis of MIS Candidates as Stepwise Time Series

This research points to seven single, molecular entities that are sufficient to induce EMT: TP53, SMAD2, BAX, TGFβR1, TGFβ1, TGFβR2, and BCL2. Each of these entities was returned as a cardinality of one (Co1) by the minimal intervention set program of Bio-ModelChecker, exhibiting defining control over the network to single-handedly switch a cell from epithelial to mesenchymal. This indicates that controlling EMT is even more difficult than previously believed. If any of the seven single cardinality entities are even persistently altered, EMT can occur in less than 20 steps (usually closer to ~7 steps). However, those steps are unit-less and could take any amount of time to occur, as this dataset does not involve time-series data. To definitively prevent EMT all of these genes must be up or down regulated respectively to not trigger EMT, in addition to controlling most other genes in the pathway, and that is not currently clinically possible. The problem now becomes one of identifying a pathway entity which when modulated renders the entire network insensitive to changes by these 7 triggers.

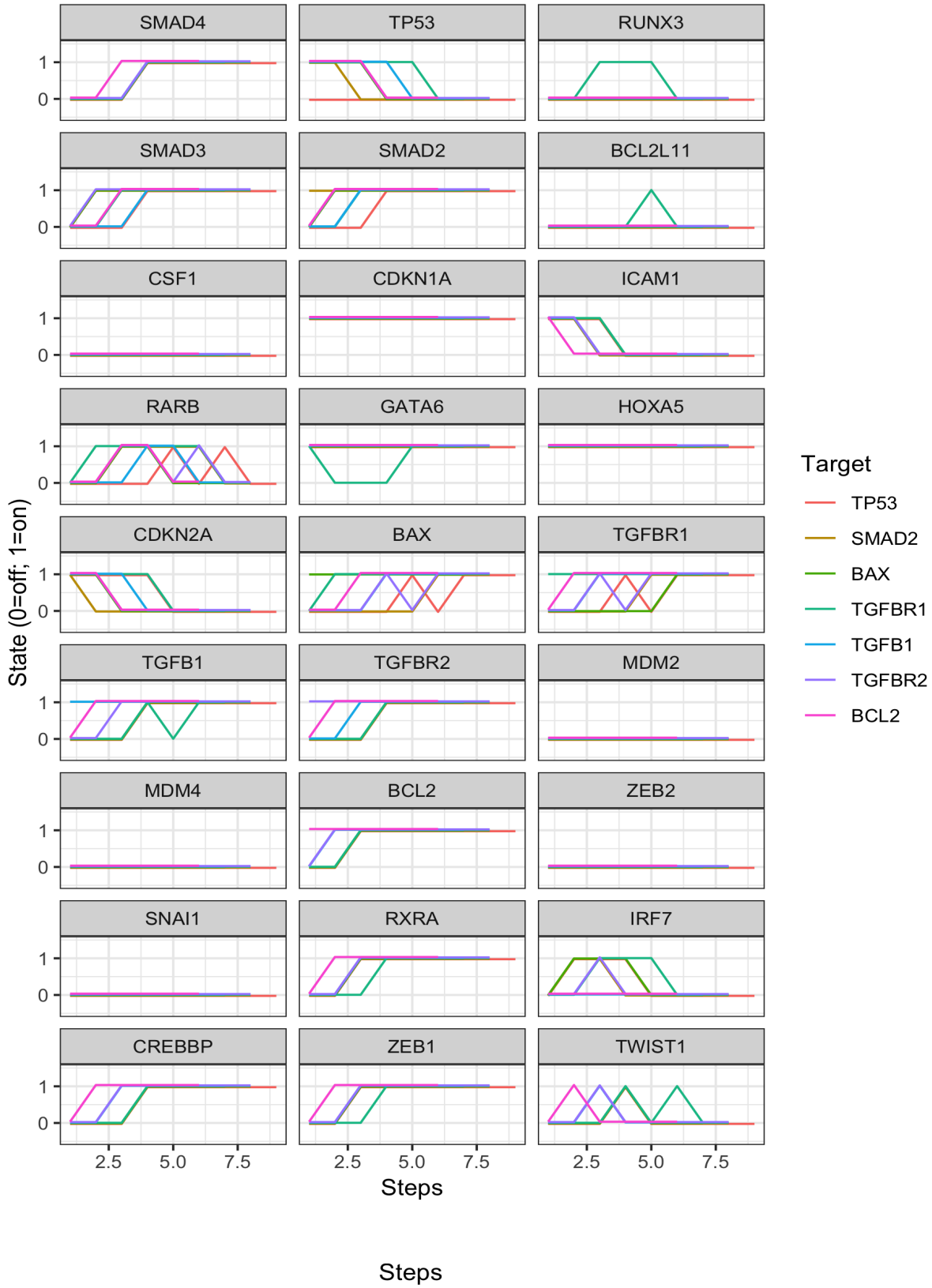
These 7 MIS candidate triggers are plotted as facets in Figures 8 and 9. For each of the 10 models sampled from the 339 solution set of steady states, almost all of the corresponding MIS candidates induced a similar series of state transitions during EMT. Models 1, 2, 4, 5, and 6 supported a common response dynamic (Figure 8), while models 3, 7, 8, 9, and 10 commonly supported a slightly different response were identical (Figure 9). Transitions predicted for all models in response to each MIS for all models can be found in Appendix C. The only protein response path that differed between theses models was that of SMAD4.

It is easiest to process these figures as time-series graphs, where each line is a gene changing regulation state over time. The gene products progress from their state in an epithelial state (left) over a series of steps, eventually resulting in a mesenchymal end state (right). In this way, the horizontal axis represents a proxy for time, with units of the number of steps until mesenchymal resolution. Remembering the parameters for this program, the horizontal axis could have had 0-20 steps, but was

minimized since the mesenchymal state was always found within 8 steps. The vertical axis is a descriptive scale where 0 is a down-regulated state, and 1 is an up-regulated state. Each line represents the discrete expression level of the gene product labelled at the top of the panel, as it is controlled by one of the Co1 trigger genes. The Co1 trigger genes are represented by the colored lines, defined in the legend. Looking at Figure 8, all colored lines (Co1s) control SMAD4's expression so that it is changed from a down-regulated state (0) to an up-regulated state (1) by the end of the graph. SMAD4 is the only gene MIS that is different in its stepwise transition between all 10 models sampled, however in Figure 9, the two groups of models merely show a slightly different path resulting in the same outcome for SMAD4. Taking into account biological noise levels, it is likely that all 10 models are extremely similar.

It is noteworthy that some colored lines end one or two steps earlier indicating that they lead GOIs to the resolution state in fewer steps. This is the case for most of the graphs, where some Co1s result in stable mesenchymal resolution quicker. Something worth studying further is the transient activation and inactivation of RARB that is seen in all test models and influenced by all Co1s. This means that RARB is deactivated in both steady states, but must be active to transition between the two. If this is as a persistent occurrence as it seems, it follows that RARB is transiently responsible for EMT. This is also shown to be the case for IRF7 or TWIST1 to a lesser degree as not all Co1s flip twice.

Figure 8: Stepwise Transitions of MIS for Models 1,2,4,5, & 6



CONCLUSIONS

Cancer must be considered an epigenetic state of being rather than an affliction layered on top of a healthy or normal state. The progression of cancer is caused by numerous perturbations of proteins, which result in a systemic disease. It is the accumulation of perturbations that lead to cancer development, which are triggered by carcinogens, normal biological errors, genetic aberrations, along with many other factors being discovered regularly. It is increasingly apparent that the key to cancer progression is not any one change or even a series of changes, but a large-scale shift in how entities interact. This is why chemotherapy is often equated to guesswork; just because it worked on one patient does not mean it will work in the same way for another patient that has cancer of the same original location. Drug treatments need to start addressing the differences in patients, in separate tumors, and even in individual cells.

Even more, cancer is commonly heterogeneous and individual cells within the same tumor will not react to treatment in the same way. There exist selection pressures for cells in different orientations within any stage tumor that must be understood in order to administer a treatment that is positively known to be effective. If molecular interactions are not understood, then drug interactions are not understood and treatments are made up of guesswork. Herein lies the purpose of studying cancer from a systemic point of view.

The first half of this thesis presents a pipeline for the formation of system models looking at protein-protein interactions for any specific severity and type of cancer. Any type of expression data, RNA-Seq or MS, with no limit to size can inform the formation of cancer clusters. This project focusses on OC data, segregated into two clusters based on histology, but

this workflow can be used for future research of other cancers or other subtypes of OC. The adaptability of this pipeline is what grants it amazing power for future use.

It is not enough to simply build a generic pathway model, however. An in-depth model is likely too large and too complicated to read effectively with the naked eye, and it does not necessarily inform treatment decisions or inspire clinical relevance. It is therefore extremely important to further visualize how the pathway model can be changed or manipulated. This can occur by means of mutations, changes to the environment, and drug administration. All of these changes can lead to disease change and progression and are taken into account by the large scale clustering of biological data. Employing bioinformatics to study cancer big data allows for the detection of small perturbations in a network. The more data used to inform the model, the more sensitive it is to changes in protein abundance or interactions. Computational advancements have raised the ceiling on what is possible for the study of large datasets. In this case, these computational advancements enable systems biology to track how dozens of proteins interact, and predict how those interactions may change over time.

Applying this novel computational systems biology approach to modern cancer research, we can model how cancer changes over time in order to understand why those changes occur. OC is a model system to demonstrate the power of this network because it is characterized by cellular modifications that consistently lead to increasingly invasive subtypes. These OC subtypes are well characterized by tissue morphology, and therefore create somewhat obvious clustering rules to inform the model. Essentially, by understanding how the pathway is exploited into a cancerous state, we can start addressing ways to prevent such exploitation and preserve the normal function and interaction of proteins.

In the context of this project, ovarian oncogenesis has already occurred, and therefore it is difficult to decipher what is a perturbation of generalized OC and what is causal for EMT. Future work should include recreating an OC model with data informing our understanding of initiation and early

progression, however this data is much more difficult to acquire due to the often late diagnosis of OC. A heightened understanding of ovarian oncogenesis can help prune assumptions made about this network. For example, IRF7 is a deeply interconnected protein in this EMT pathway based on BC and interaction confidence, although it is predominantly inactivated in both data clusters, epithelial and mesenchymal. It is observed that IRF7 is transiently activated by all Co1 proteins except TP53 and TGFβ1 in order to result in EMT. This indicates that IRF7 plays some hidden role in EMT, but is likely a large player in the initiation of disease. It is also worth noting that while the counts of IRF7 are effectively inactive, the gene does not appear to be knocked out in progressive stages of cancer because it is transiently activated during EMT. Making comparative steady states for healthy vs early cancerous ovarian protein interactions may elucidate a better understanding of where IRF7 affects disease progression.

At the heart of this thesis, seven potential drug targets are presented for future analysis and research (BCL2, TP53, TGFβ1, TGFβR1, TGFβR2, SMAD2, and BAX), and future work is finding clinical ways to prevent EMT by controlling theses gene products. For example, TP53 and SMAD2 are known to be so inherently intertwined in pathways in every cell that it would be virtually impossible to control off-target effects. There are potentially many drug treatments already approved by the FDA for the treatment of various diseases that one can repurpose for the treatment of OC, perhaps by targeting proteins that might desensitize the network to these triggers. Such resilience inducing interventions could be applied prophylactically in women with high risk of developing OC. For example, the chemotherapeutic drug Venetoclax should be investigated further as it is used to inhibit BCL2 in severe types of leukemia.[27] This research indicates that BCL2 interactions are highly informative of EMT

progression in OC, and therefore inhibitors of BCL2 are worth investigating clinically as repurposed treatments. Therefore, it is time for a shift in the way researchers and clinicians view chemotherapy, to move away from shotgun approaches attempting to kill the cancer faster than the patient. It is now possible to computationally simulate how a drug with a known method of action will affect a cancerous cell of a specific type within a patient, an idea linked into the philosophy of precision medicine. The repurposing of established cancer drugs is highly advantageous towards expedited improvements for the treatment of OC, and can now be done efficiently on a patient to patient basis.

Chemotherapy is often seen as a last-resort method to treating cancer, after a tumor is invasive enough to begin metastasizing. This research proposes the idea that chemotherapeutic drugs, like Venetoclax, could work as a temporary preventative measure against metastasis. Thibault et al. would likely argue that this could lead to chemo-resistance,[28] but this proposal is only a measure to decrease the likelihood of metastasis while awaiting other treatment options. This would only be a temporary addition to existing surgical treatments but could delay progression of disease while planning other treatments. There may be many other drugs like Venetoclax that may be repurposed successfully. Therefore, a detailed database of available and approved chemotherapeutic agents, their methods of action, and their current clinical targets would prove useful in the future towards the advancement of this workflow. Several databases of chemotherapy drugs exist but they rarely include any methods of action that are readily searchable, partially due to the fact that the actions of some approved chemotherapy drugs are still not fully understood. Unfortunately, this extends into a cultural dilemma wherein researchers have little motivation to research how an approved drug actually works as they would receive little compensation or reward. Long term motivation and ability to do this research is difficult to find, but could yield hundreds of new applications for existing drugs and revolutionize current treatment understanding.

METHODS

I. Pathway Modeling for Genes of Interest

The pathway databases used to build regulatory networks linking GOI are Pathway Studio (PS), Mogrify, Reactome, and String.[21,22,29,30] PS uses a text-mining engine, called MedScan, and natural language processing to read through a growing collection of over four million journal articles and to extract biological entities and their interactions into a searchable database.[31] Once the pathway model is built with corroborating evidence for each protein addition, references must be checked to verify that the machine learning algorithm correctly associates entities in the network. From an original machine generated network, relations are manually condensed or pruned due to duplicate or sparse reference lists. Two relations are condensed into one if the nodes and directionality of the edges were the same. A relation is manually deleted if its associated references do not support the polarity of the relation. Relations are dropped if there were duplicate types of interactions, such as “positive regulation” and “positive expression”.

II. Data to Apply to the Network

It is recommended that proteomics data, as opposed to RNA-Seq data, be used for future use of this workflow as it is more informative of what gene products are actually present. From the proteomic data, IBAQ scores should be calculated or extracted, as means of label-free quantification from the estimated range of 0 to ~200,000,000. Future iterations of this pipeline may also view the distribution of iBAQ scores before proceeding. This step serves as a way to offer preliminary confirmation that GOI are differentially present across cancer subtypes.

III. Discretization of Proteomic Data

This step qualifies gene quantities into levels of expression by comparing distributions on a macro scale. Distributions of protein abundance for each cell type are considered different based on the significance t-tests of their medians, specifically using Wilcox tests with a 95% confidence interval. The median is used instead of the mean expression level to control for possible skewness or multimodality. If the p-values resulting from the t-tests are less than 0.05, then it is concluded with 95% confidence that those medians of clustered expression values are different and the group with the higher median is labeled with a 1, while the lower median group is labeled a 0. These do not equate to a binary on/off switch, but rather a higher/lower abundance qualification. Any distributions that did not yield statistically different medians (p-value greater than 0.05) were left unconstrained in the model and given a value of -1. In this way, every entity has three relative possible states: on, off, or unknown. This data is submitted to BioModelChecker as a JSON file, an example of which is included in Appendix B.

IV. Simulation-Based Determination of Steady States

The dynamics of the GOI network are first supported by decisional logic similar to that used by computers, and are used to predict, via simulation, the location and dynamics of steady states on either side of EMT.[32] Several elements of regulatory logic are combined within BioModelChecker to predict the propagation of protein expression states. These include the directionality of connections in the pathway, the discrete levels of protein quantity (e.g. low/no or high) for different groups delineated by k-means clustering of experimental data, known information about mutations and knockouts via literature, and backtracking simulations that will attempt to fill in the map with possible states that validate all parameters.

BioModelChecker's parameter selection for this type of research includes several choice-based parameters in addition to the JSON. A JSON requires an adjacency list labelled "interactions" including sources, targets, trajectories, polarities, and confidence levels); a key for the GOI at each index called "titles"; a list of levels for each node where all are level 1 if the pathway is simple; a starting steady

state; and an ending steady state. An example is provided in Appendix C. This project utilizes the solver OR-Tools, with strong edge constraints, and no post combination constants for all program runs. Edge constraints refers to interactions between K values, where a strong positive edge must influence it's target(s) at least half the time. An ordering parameter is filled once with the selection lexicographical ordering, yielding a rapidly growing list of all possible steady states the model could arrive at. A simulation is run a second time with the ordering parameter set to unordered. The first run was used to determine the possible steady states, and it preserves all edges and constraints. The second run's parameters dictate which edges should be dropped if they do not inform a decision in the pathway. This adds another level of complexity to the backtracking algorithm and creates a relatively unbound problem, which is able to inform the frequency at which edges are maintained or pruned in the model. These two runs together result in a map of steady states that are informed about what edges are most prevalent.

Each solution set from Bio-ModelChecker yields a deterministic set of logic gates that inform the behavior of the network, and result in steady states. The logical rules are defined in a K matrix explaining how the algorithm filled in the unconstrained entities, and the starting and ending steady states are represented as lists of binary coded regulations for each node. The K matrix contains further details including the set of logical weights representing the relative strength of activators and inactivators as well as activation thresholds representing differences in biological receptor affinity. Each set of these parameter values supported different response dynamics in migrating any small deviation to the same steady state, which was registered as a different solution. The output of theses runs are in files consisting of rows of lists. The SS1 and SS2 lists are the steady states for each model output. The EM list indicates the presence of a

connection between entities in a given model. These two solution lists yield a pathway model that is easy to understand and use for future study. All unique steady states determined by Bio-ModelChecker should be extracted if there are more than one start and end.

These connections and steady state solutions are used to create pathway models of the cancer subtypes that the data originates from. The steady state solution list is the regulation level of the proteins in each cancer subtype, and frequencies calculated from the connections list is used to represent the likelihood that the interaction takes place. There are as many steady state maps as there is variability in the network. Each steady state map should be drawn in such a way that it is obvious which entities are active and which ones are relatively inactivated, such as color coding. This is fairly easy to do using the program yED. The connections should indicate the likelihood an interaction between entities takes place, and therefore should be visibly more prominent when it appears more often in models. The greater the number of solutions form the unordered run, the lesser the variability and higher the specificity of edge frequency.

V. Betweenness Centrality

Betweenness centrality is a method of measuring how interconnected each node is in the matrix. Matlab, R, and most other statistical analysis platforms will have a method for calculating BC that may require reformatting of the adjacency matrix used from Bio-ModelChecker's original JSON.

VI. Minimal Intervention Sets

At this stage of the workflow, we are attempting to induce EMT *in silico* in order to understand its possible causes. Every steady state determination is unique, therefore even though the steady state solutions may be identical, there will be variability in how that steady state is found. This variation is found in the K matrix of the solution files which is built as the simulation progresses. It represents every decision the solver makes towards determining in the unknown entities orientation. This variation across models, however, is presumably insignificant if the steady states are identical. So, it is

considered redundant to look at every single solution's K matrix, and a random sample is taken if there are an exhaustive number of solutions. This representative sample is then passed back into Bio-ModelChecker, which begins at each node following every connection possible to navigate through the entire network using the K matrix as the guidebook determining what is activated or inhibited. The end goal is to arrive at a specific attractor (in this case the mesenchymal steady state). The function will accept parameters wherein a researcher can define how many nodes are allowed to be arbitrarily perturbed by Bio-ModelChecker at one time (maximum cardinality), and how many steps the program is allowed to take throughout the pathway (path length). A cardinality is defined as the number of stabilized nodes that, when perturbed, are able to induce metastasis alone. Cardinality is limited because simpler perturbations are more likely. If EMT is induced by altering 1 gene or 6 genes, it is more likely that the 1 gene will end up causing it most of the time. Additionally, a drug should ideally effect as few entities as possible in order to decrease the chance of off-target effects. It is favorable for these reasons to limit the cardinality to a small a number as the network allows. The path length is limited because the longer a signaling chain is, the more effect biological noise will have on the system and the less specific the drug interaction. This depends on the complexity of the system, but this project limited the path length to 20, which was more than enough. Each MIS is able to be viewed essentially as a solution set of what minimalist value(s) are informative enough to convert a cell to the mesenchymal steady state over a series of steps. Those steps serve as a time proxy and can be key to understanding hidden effects nodes may have on a circuit.

REFERENCES

1. Efroni, S., Schaefer, C. F. & Buetow, K. H. Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PLoS One* **2,** (2007).

2. Gong, H. Analysis of intercellular signal transduction in the tumor microenvironment. *BMC Syst. Biol.* (2013). doi:10.1186/1752-0509-7-S3-S5

3. Beaufort, C. M. *et al.* Ovarian cancer cell line panel (OCCP): Clinical importance of in vitro morphological subtypes. *PLoS One* **9,** (2014).

4. Francavilla, C. *et al.* Phosphoproteomics of Primary Cells Reveals Druggable Kinase Signatures in Ovarian Cancer. *Cell Rep.* **18,** 3242–3256 (2017).

5. U.S. Cancer Statistics Working Group. U.S. Cancer Statistics Data Visualizations Tool, based on November 2017 submission data (1999-2015). *U.S. Department of Health and Human Services, Center for Disease Control and Prevention and National Cancer Institute.* (2018).

6. Spano, D., Heck, C., De Antonellis, P., Christofori, G. & Zollo, M. Molecular networks that regulate cancer metastasis. *Seminars in Cancer Biology* (2012). doi:10.1016/j.semcancer.2012.03.006

7. Yeung, T.-L. *et al.* Cellular and molecular processes in ovarian cancer metastasis. A Review in the Theme: Cell and Molecular Processes in Cancer Metastasis. *Am. J. Physiol. - Cell Physiol.* **309,** C444–C456 (2015).

8. Huang, J. *et al.* The PARP1 inhibitor BMN 673 exhibits immunoregulatory effects in a Brca1-/-murine model of ovarian cancer. *Biochem. Biophys. Res. Commun.* (2015). doi:10.1016/j.bbrc.2015.05.083

9. Valastyan, S. & Weinberg, R. A. Tumor metastasis: Molecular insights and evolving paradigms. *Cell* (2011). doi:10.1016/j.cell.2011.09.024

10. Lu, M., Jolly, M. K., Onuchic, J. & Ben-Jacob, E. Toward decoding the principles of cancer metastasis circuits. *Cancer Res.* (2014). doi:10.1158/0008-5472.CAN-13-3367

11. Bowtell, D. D. *et al.* Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer. *Nat. Rev. Cancer* (2015). doi:10.1038/nrc4019

12. Zhong, S. *et al.* High-throughput illumina strand-specific RNA sequencing library preparation. *Cold Spring Harb. Protoc.* (2011). doi:10.1101/pdb.prot5652

13. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics* (2009). doi:10.1038/nrg2484

14. Chuang, R. *et al.* Drug target optimization in chronic myeloid leukemia using innovative computational platform. *Sci. Rep.* (2015). doi:10.1038/srep08190

15. Silverbush, D. *et al.* Cell-specific computational modeling of the PIM pathway in acute myeloid leukemia. *Cancer Res.* (2017). doi:10.1158/0008-5472.CAN-16-1578

16. Wooten, D. J. & Quaranta, V. Mathematical models of cell phenotype regulation and reprogramming: Make cancer cells sensitive again! *Biochim. Biophys. Acta - Rev. Cancer* (2017). doi:10.1016/j.bbcan.2017.04.001

17. Coscia, F. *et al.* Integrative proteomic profiling of ovarian cancer cell lines reveals precursor cell associated proteins and functional status. *Nat. Commun.* **7,** 1–14 (2016).

18. Russel, S. & Norvig, P. *Artificial intelligence—a modern approach 3rd Edition*. *The Knowledge Engineering Review* (2012). doi:10.1017/S0269888900007724

19. Sedghamiz, H., Morris, M., Craddock, T. J., Whitley, D. & Broderick, G. Bio-ModelChecker: Using Bounded Constraint Satisfaction to Seamlessly Integrate Observed Behavior with Prior Knowledge of Biological Networks. *Front. Bioinforma. Biotechnol.* (2019).

20. Sedghamiz, H. *et al.* Computation of Robust Minimal Intervention Sets in Multi-valued Biological Regulatory Networks. *Front. Physiol.* (2019). doi:10.3389/fphys.2019.00241

21. Nikitin, A., Egorov, S., Daraselia, N. & Mazo, I. Pathway studio--the analysis and navigation of molecular networks. *Bioinformatics* (2003). doi:10.1093/BIOINFORMATICS/BTG290

22. Rackham, O. J. L. *et al.* A predictive computational framework for direct reprogramming between human cell types. *Nat. Genet.* (2016). doi:10.1038/ng.3487

23. Toy, E. P. *et al.* Enhanced Ovarian Cancer Tumorigenesis and Metastasis by the Macrophage Colony-Stimulating Factor. *Neoplasia* **11,** 136–144 (2009).

24. Chambers, S. K. Role of CSF-1 in progression of epithelial ovarian cancer. *Futur. Oncol.* **5,** 1429–1440 (2009).

25. Grither, W. R. *et al.* TWIST1 induces expression of discoidin domain receptor 2 to promote ovarian cancer metastasis. *Oncogene* **37,** 1714–1729 (2018).

26. Ning, S., Pagano, J. S. & Barber, G. N. IRF7: Activation, regulation, modification and function. *Genes and Immunity* (2011). doi:10.1038/gene.2011.21

27. Mihalyova, J. *et al.* Venetoclax: A new wave in hematooncology. *Experimental Hematology* (2018). doi:10.1016/j.exphem.2018.02.002

28. Thibault, B., Castells, M., Delord, J. P. & Couderc, B. Ovarian cancer microenvironment: Implications for cancer dissemination and chemoresistance acquisition. *Cancer and Metastasis Reviews* (2014). doi:10.1007/s10555-013-9456-2

29. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gkx1132

30. Jensen, L. J. *et al.* STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* (2009). doi:10.1093/nar/gkn760

31. Humphreys, K. *et al.* MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics* (2009). doi:10.1093/bioinformatics/btg452

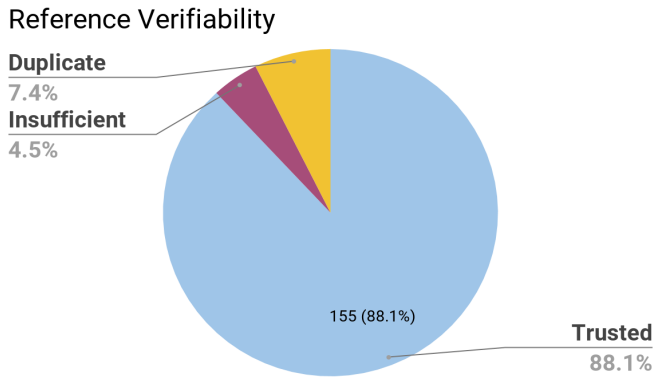
32. Sedghamiz, H., Morris, M., Craddock, T. J. A., Whitley, D. & Broderick, G. High-fidelity discrete modeling of the HPA axis: A study of regulatory plasticity in biology. *BMC Syst. Biol.* **12,** 1–16 (2018).

APPENDIX A: PATHWAY REFERENCE DETAILS

I. Reference Statistics

Of 176 reference sentences used to generate the map of all core gene relations, 21 were hand pruned due to duplicate information, inaccurate language processing, or false conclusions. 95.5% of the relations were successfully verified, including duplicates. These are represented in .

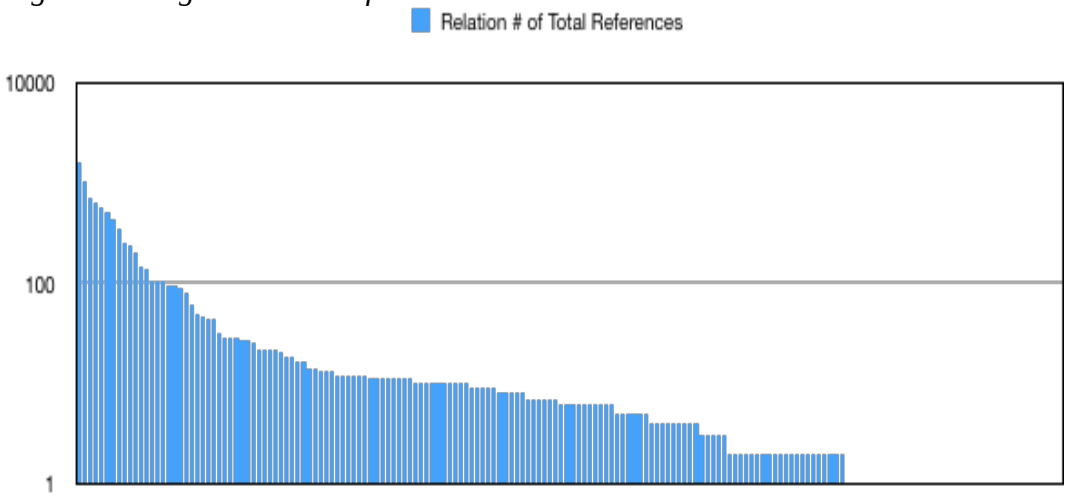
Figure 10: Reference Verification Statistics



II. Reference Distribution

8,788 original references existed across 176 total relations in the original pathway map of 27 entities. Each of the 176 relations had a list of sentences from which Pathway Studio concluded the relations' existence. Each relation's reference list had to be human-verified. Supplemental Figure 2 is the log-scale distribution of the number of references supporting each relation.

Figure 11: Log-10 Based Reference Distribution



APPENDIX B – JSON FILE FORMAT

Below is the JSON used for this research, the format is described and explained in the methods section Simulation-Based Determination of Steady States. The first line variable name "_FILENAME_" should be the name of the file excluding file extension for continuity use of BioModelChecker.


```
{
  "_FILENAME_": {
    "interaction": [
        [2, 4, 25, 1, 10, 12, 24, 1, 5, 10, 26, 4, 12, 13, 15, 18, 24, 4, 7, 11,
14, 1, 2, 3, 4, 4, 24, 24, 2, 3, 4, 6, 13, 16, 20, 24, 2, 4, 18, 19, 24, 1, 2, 4,
8, 9, 10, 17, 21, 2, 3, 4, 5, 10, 21, 24, 24, 8, 10, 11, 14, 3, 19, 20, 27, 2, 3,
4, 5, 6, 12, 19, 20, 26, 27, 1, 3, 4, 7, 10, 20, 21, 25, 1, 2, 3, 22, 24, 26, 27,
1, 3, 4, 5, 6, 7, 8, 9, 12, 15, 16, 17, 18, 20, 22, 22, 23, 25, 26, 27, 4, 8, 10,
17, 20, 21, 25, 26, 27, 5, 1, 2, 4, 5, 7, 10, 11, 12, 13, 16, 19, 20, 23, 25, 26,
27, 1, 2, 3, 4, 5, 6, 15, 20, 23, 24, 26, 27, 1, 4, 18, 24, 2, 24, 26],
        [1, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7,
8, 8, 9, 10, 11, 11, 11, 11, 11, 11, 11, 12, 12, 12, 12, 12, 13, 13, 13, 13, 13,
13, 13, 13, 14, 14, 14, 14, 14, 14, 14, 14, 15, 15, 15, 15, 16, 16, 16, 16, 17, 17,
17, 17, 17, 17, 17, 17, 17, 17, 18, 18, 18, 18, 18, 18, 18, 18, 19, 19, 19, 19, 19,
19, 19, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,
20, 21, 21, 21, 21, 21, 21, 21, 21, 21, 22, 23, 23, 23, 23, 23, 23, 23, 23, 23, 23,
23, 23, 23, 23, 23, 23, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 24, 25, 25, 25,
25, 26, 26, 26],
        [-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -
1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1],
        [0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1,
1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1,
0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1,
1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
```

1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0,
0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1],
        [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1]
      ],
    "titles": ["SMAD4", "TP53", "RUNX3", "SMAD3", "SMAD2", "BCL2L11", "CSF1",
"CDKN1A", "ICAM1", "RARB", "GATA6", "HOXA5",
"CDKN2A", "BAX", "TGFBR1", "TGFB1", "TGFBR2", "MDM2", "MDM4", "BCL2", "ZEB2",
"SNAI1", "RXRA", "IRF7", "CREBBP", "ZEB1",
"TWIST1"],
    "L": [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1],
    "SS": [
      [[0, 1, -1, 0, 0, -1, -1, -1, 1, -1, -1, -1, -1, 0, 0, 0, 0, -1, -1, -1,
-1, -1, 0, -1, 0, 0, -1]],
      [[1, 0, -1, 1, 1, -1, -1, -1, 0, -1, -1, -1, -1, 1, 1, 1, 1, -1, -1, -1,
-1, -1, 1, -1, 1, 1, -1]]
    ]
  }
}

APPENDIX C: ALL MODELS STEPWISE TRANSITION GRAPHS

Figure 12: Model 1 Stepwise MIS Transitions

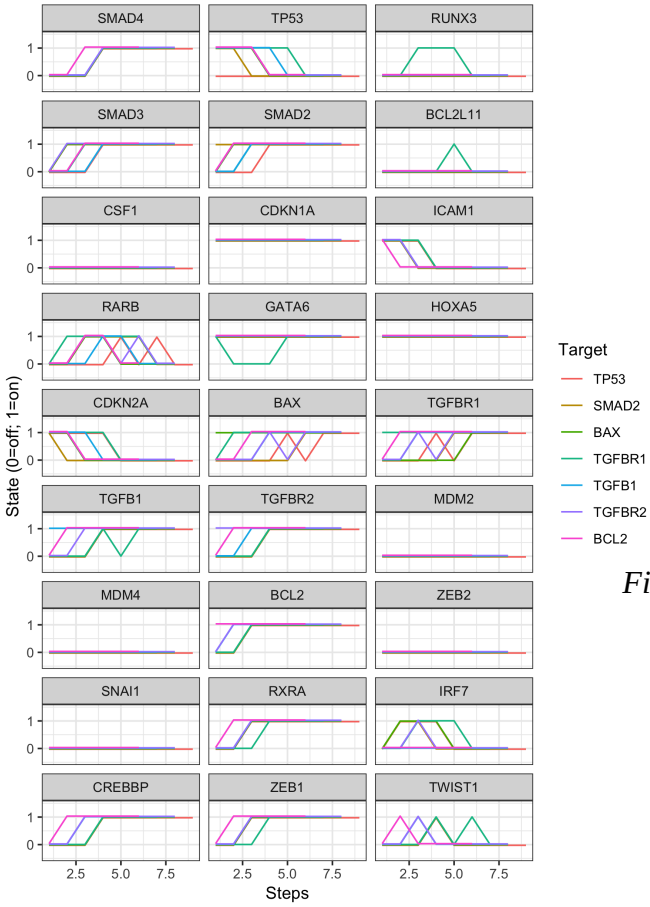


Figure 13: Model 2 Stepwise MIS Transitions

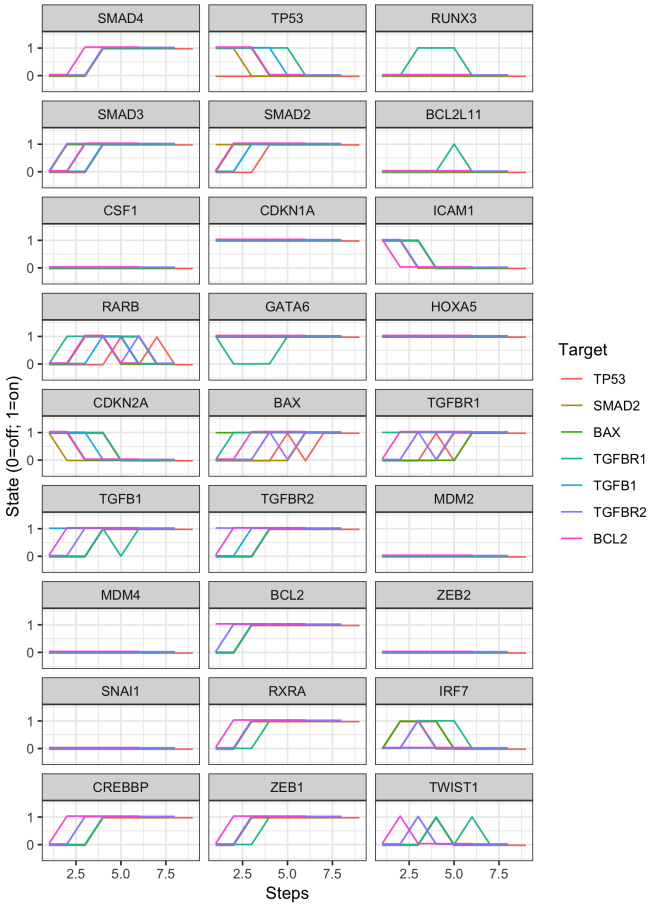


Figure 14: Model 3 Stepwise MIS Transitions

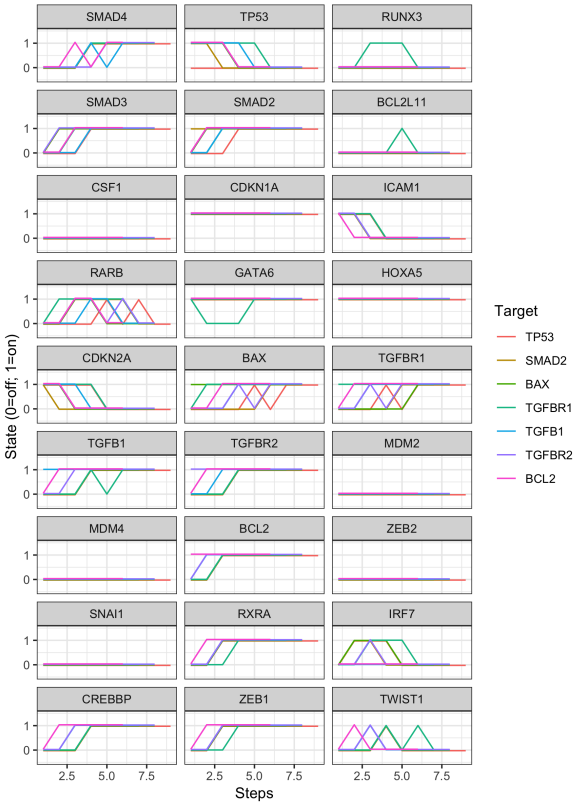


Figure 15: Model 4 Stepwise MIS Transitions

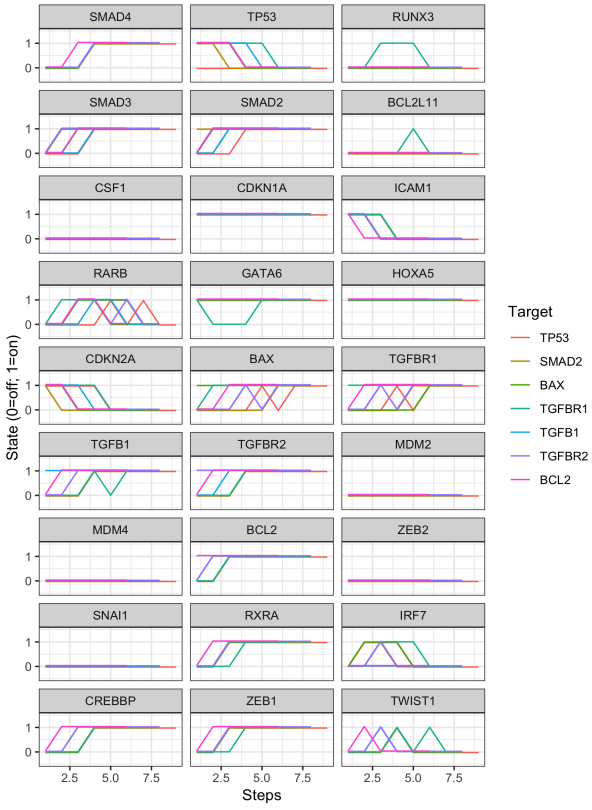


Figure 16: Model 5 Stepwise MIS Transitions

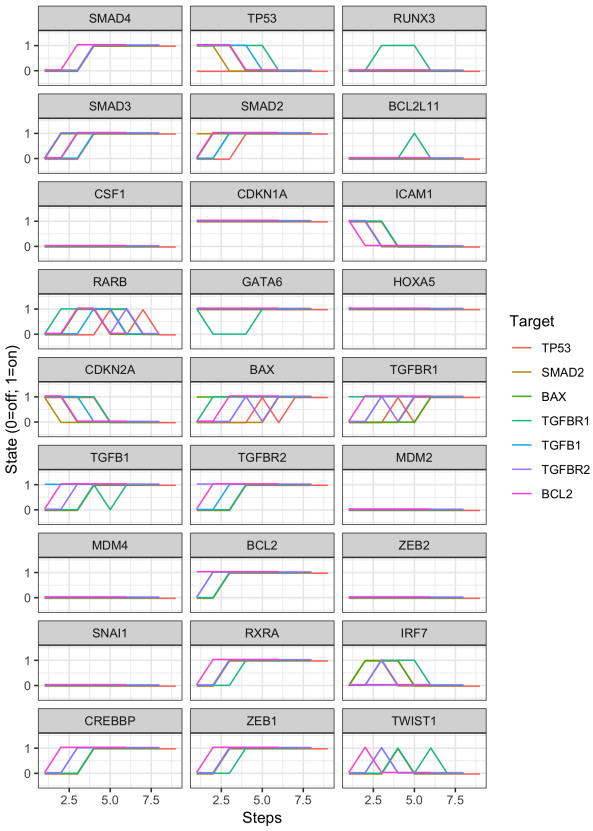


Figure 17: Model 6 Stepwise MIS Transitions

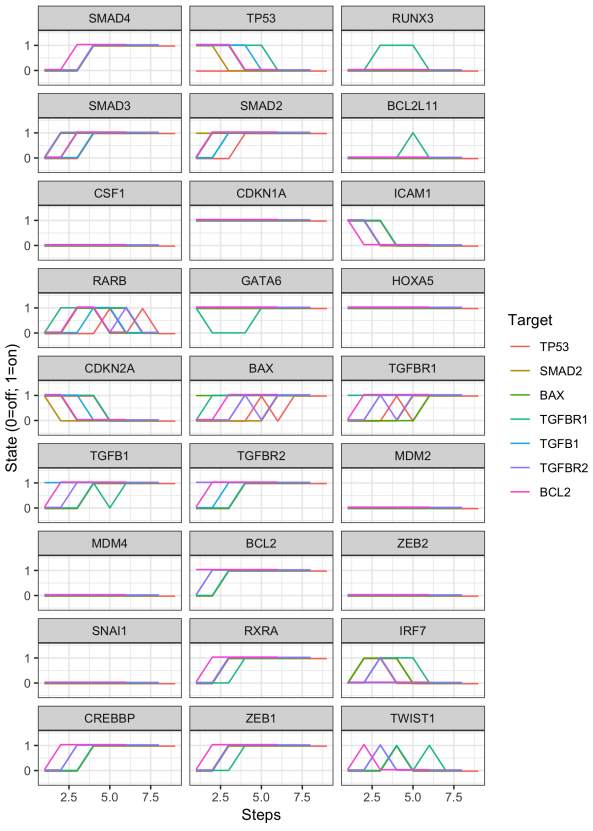


Figure 18: Model 7 Stepwise MIS Transitions

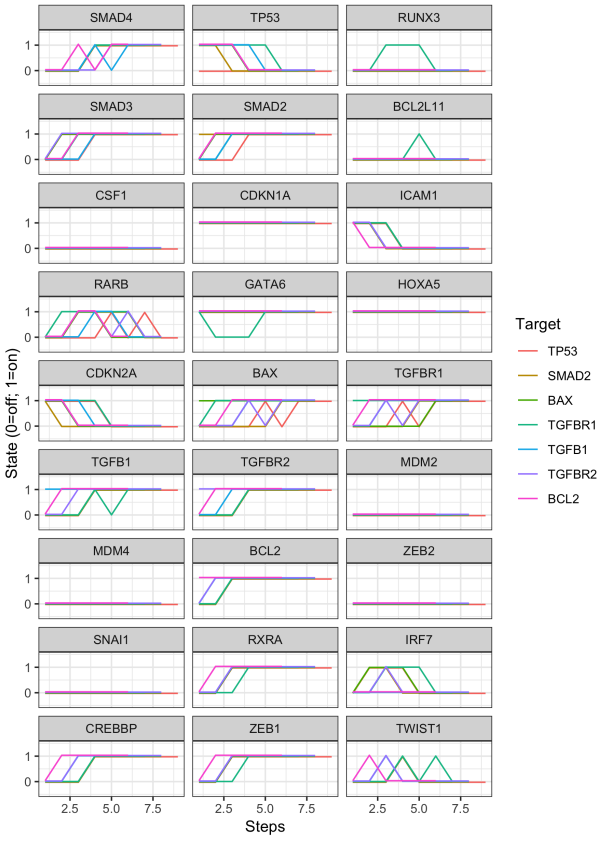


Figure 19: Model 8 Stepwise MIS Transitions

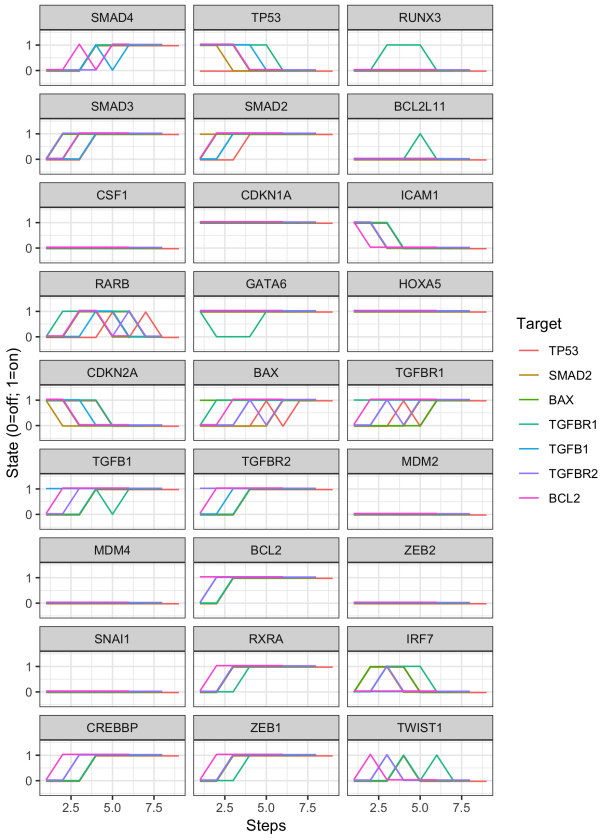


Figure 20: Model 9 Stepwise MIS Transitions

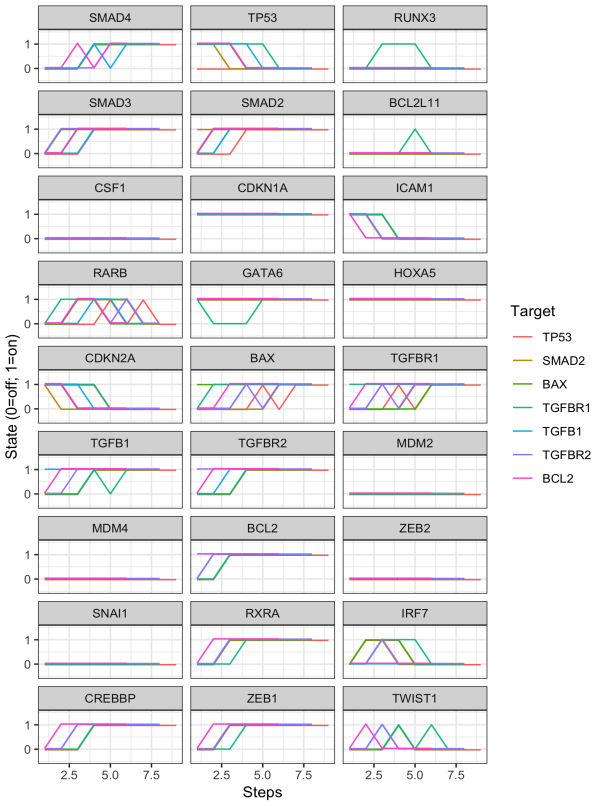


Figure 21:Model 10 Stepwise MIS Transitions