Using Bayesian networks to identify control topography between cancer processes and immune responses via metagene constructs.

Jacob Kaiser

Thesis submitted
to the College of Medicine
at West Virginia University

in partial fulfillment of the requirements for the degree of
Masters of Science in Biomedical Sciences

David Klinke, Ph. D., Chair
John Barnett, Ph. D.
Christopher Cuff, Ph. D.
Rosana Schafer, Ph. D.

Department of Biomedical Sciences

Morgantown, West Virginia
2014

Keywords: TCGA, Bayesian Networks, Control Topography, Cancer Progression

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

ABSTRACT

Using Bayesian networks to identify control topography between cancer processes and immune responses via metagene constructs.

Cancer arises from a deregulation of both intracellular and intercellular control systems. Understanding the architecture of these control systems and how they are changed in diseases could present opportunities for therapeutic targets to restore normal control. However, since intercellular control structures only appear in intact systems, it is difficult to identify how these control structures become altered using *in vitro* models and it can be difficult to determine if an *in vivo* model system appropriately replicates what occurs in human disease. In order to overcome this, we use the diversity in normal and malignant human tissue samples from the Cancer Genome Atlas database of human breast cancer to identify intercellular control topology *in vivo*. To improve the underlying biological signals from the noisy gene expression data, we constructed Bayesian networks using metagene constructs, which represented groups of genes that are concomitantly reported with different immune and cancer states. From these directional, acyclic graphs, we found opposing relationships between cell proliferation and epithelial-to-mesenchymal transformation (EMT) with regards to macrophage polarization. Furthermore, we also found that it was possible to identify the relationship between EMT and macrophage polarization with fewer datasets when the Bayesian network was generated from malignant samples alone, while it was possible to identify the relationship between proliferation and macrophage polarization with fewer samples when the samples were taken from a combination of the normal and malignant samples. When the same technique was applied to other cancers, we found a common result that proliferation was associated with a type 1 cell-mediated anti-tumor immunity and EMT was associated with a pro-tumor anti-inflammatory response. All together, these networks give us an understanding of what relationships are occurring in human cancer progression, and this knowledge can be used to help identify model system that more closely mimic human disease progression.

Table Of Contents

Tables

Figures

1 Introduction

Homeostasis is a basic requirement for life, and can be observed existing at many levels in multi-cellular organisms, including at the cellular level (such as ion pumps maintaining an ion gradient between the cell and its environment), at the tissue level (such as angiogenesis as a means to maintain oxygen and nutrient delivery)[1], and at the organism level (such as by maintaining the internal temperature of the body). Tumorogenesis, in many ways, represents a rewiring of and subversion of normal cellular circuitry resulting in an alteration of the normal dynamic processes that would act to maintain homeostasis in the tissue microenvironment[2,3]. These subversions can be rewiring that occur entirely within the cancer cell (such as achieving replicative immortality and conversion of cells to more invasive phenotypes) or can include alterations that change the tissue in which the cancer cells are growing (such as promotion of angiogenesis)[3].

A key aspect to understanding tumor progression is to understand that a tumor is governed by the same evolutionary principles that control all life. While the idea that evolutionary principles can be applied to tumorogenesis is not a new one[4,5], it is becoming more evident that cancer needs to be viewed as a evolutionary, dynamic process. However, it should be noted that cancer growth represents somatic and not Darwinian evolution. This distinction has several implications for how a cancer can evolve. First, since the evolution is occurring in the somatic cells, individual mutations can be passed down to subsequent cancer cells, giving rise to a heterogeneous population of cells. These mutations, while random and normally negative for the cell, are selected by the same evolutionary forces that act on any organism (can it acquire resources, avoid predation, and replicate itself)[3]. Somatic evolution also has implications with regards to the time scale at which this evolutionary selection is taking place –

we have evolved methods to combat cancer on a generational timescale, whereas the tumor evolves in a single individual. As such, the person is effectively static from the view of the cancer. This, along with the fact that malignant cells begin as normal somatic cells, gives the tumor the potential to not just adapt to its environment, but to adapt the environment to its own needs[6]. This combination of normal evolutionary forces along with the ability of the cancer to manipulate its environment drives cancer progression and helps give rise to hallmarks of cancer. For example, it has been shown, both in biological and computational models, that unequal nutrient distributions within an environment would select for more mobile forms of cancer[7–10]. We have known about some of these alterations for a while, though a lot of the earlier research focused on how the tumor manipulated itself for survival, as can be seen by the original hallmarks of cancer, which focused almost entirely on intracellular alterations required for cancer progression.[11]. This can be contrasted to the new enabling and emerging hallmarks, which lists certain forms of inflammation as an enabler of tumor progression (sometimes through increased genetic damage, and also through predation of certain forms of cancer[3,12,13]), and eventual escape of the immune system as a hallmark of cancer (either through a change in the cancer itself to make itself less immunogenic[14,15], or by biasing the immune system to an unproductive response[16]).

Identification of these changes in control remains a challenge. Part of the difficulty is due to current laboratory techniques. Most cell based techniques were developed as a means to observe and manipulate processes within a cell, and while some of these techniques can be modified to observe certain interaction between cells – such as through co-cultures[17] - they are mostly limited to identifying intercellular phenomena. Furthermore, this analysis requires an *a prior* assumption about which cells and environmental conditions are needed in order to observe

the physiologically significant relationships, and while it does make it possible to determine what events are occurring within the culture, it does not offer a means to verify if these interactions are biologically relevant. Finally, most conventional experimental designs are fairly focused, and require the assumption that you are already looking for the relevant interactions. More complicated interactions can be observed by the use of *in vivo* models, such as through xenografts, the use of established cancer cell lines, or the use of genetically engineered mice (GEM). Xenografts, while they can use human cancers that have arisen naturally and can be useful for general observation of human cancers, are not ideal when trying to identify interactions between the tumor and the host, as the species gap can mask the interactions, and there is no guarantee that the same processes are relevant in both species[18]. Use of an established mouse cancer cell line or GEM mice gets around the species problem, but results in studying the a mouse disease in mice, which, even if they contain the conserved genes and proteins, may result in a fundamentally different response when the same phenomenon is viewed in humans[19,20].

These problems affect more than just cancer research, and as high throughput sample processing is becoming more cost efficient, along with the subsequent increases in available genetic, mRNA, and protein expression data available, people are turning more towards a systems view of disease[21]. One source of data that can be used to identify such interactions is analysis of human cancer biopsy data. These tissue samples contain malignant cells and infiltrating immune cells, which enables insight into the nature of the local immune response that is occurring with the tumor. It is also biologically relevant, as it is using human data, and since it comprises multiple cell types, it can be used to help generate an influence diagram that would help in observing the overall control mechanisms.

Along with the advent of high throughput data is the potential to utilize probabilistic inference methods to identify relationships out of the data that could not be observed using simpler statistical techniques[22]. One of the methods that can be used to identify the topology of an in influence diagram in an unbiased method is through the use of algorithms that identify Bayesian networks[21]. Bayesian networks are a type of directed acyclic graphs (DAG), where each node represents a random variable and each edge represents a causal relationship between two nodes. Bayesian networks have previously been used to model signaling pathways within cells[23], correctly identifying the known DNA repair networks in *E. coli* using microarray data[24] and simple phosphorylation cascades in T lymphocytes using flow cytometry data[25,26]. The use of Bayesian networks with flow cytometry data is particularly powerful, as it allows each cell to viewed as an individual event, and possible extreme cases, which are useful in identifying Bayesian networks, are not lost through averaging[25].

At the same time, there are limitations to using flow cytometry data when trying to construct a Bayesian network. For example, analyzing cells via flow cytometry severely limits the amount of proteins that can be observed at a single time[26], which proves to be a significant limitation when trying to understand networks that involve potentially hundreds of genes and proteins. Furthermore, most studies thus far have focused on identifying intracellular events that occur over a relatively short timeframe – on the order of minutes or hours. Our objective is to identify long term changes that occur in conjunction with disease progression - a process that occurs over months or years and that is reflected in changes in cell populations within a tissue and cellular development processes. In particular, we are looking at the interplay of processes that are commonly associated with oncogenesis and immune surveillance. Microarray data, on the other hand, loses some of the diversity found in flow cytometry data, but if obtained from a

complex sample can provide insight into cellular composition within a tissue. It is limited, though, by the computational expense associated with identifying a Bayesian network. While Bayesian networks can handle noisy data, which microarray data normally is, the time required to generate the network increases greatly as the number of nodes increases[27]. Given our study objective and the limitations of computational methods, we combined Bayesian network analysis with metagene constructs to identify relationships between oncogenic processes and immune surveillance.

The purpose of this study is to see if we can identify causal evidence of crosstalk between events associated with local cellular immune-surveillance during breast cancer oncogenesis given pre-existing microarray data, metagene constructs, and Bayesian networking. In short, we found that cellular proliferation and EMT had opposing relationships with macrophage polarization in invasive breast cancer, with increased proliferation being associated with classically activated macrophages (M1) and EMT being associated with alternatively activated macrophages (M2). We found that sample size and complexity affected the resulting Bayesian networks, with smaller sample sizes resulting in less complex networks, while changes in the composition of the sample influenced the relationships that were seen. When we expanded this study to other forms of cancer, we saw that overall increases in proliferation went along with increases in cell mediated anti-tumor immunity whereas increases in EMT resulted in decreases in cell mediated anti-tumor immunity.

2 Methods

2.1 Data Acquisition

Gene expression values in normal and malignant tissues were obtained as part of the Cancer Genome Atlas (TCGA)[28]. In short, homogenized samples taken from primary tumors after diagnosis but before treatment or from matched normal tissue samples were analyzed using on the Agilent G4502A 07 microarray chip. Gene expression was determined, and genes were normalized to a log2 scale using the RMA (Robust Multichip Average) method[29], with negative numbers representing lower gene expression and positive numbers representing greater gene expression. Level 3 tissue microarray data were downloaded for the invasive breast carcinoma samples (BC, tumors = 599, normals = 65), glioblastoma multiform (GBM, tumors = 482, normals = 10), lung squamous cell carcinoma (LUSC, tumors = 155, normals = 0), and colon adenocarcinoma (COAD, tumors = 174, normals = 9) samples. In this case, normals represent microarray data from normal, non-cancerous tissue. Genes of interest were identified, and samples missing any of the genes were eliminated from the study.

2.2 Metagene calculations

To represent cellular processes, we used metagene constructs. A metagene is the expression and aggregation of individual genes observed by microarray data, and can represent either cell infiltration, cell polarization, or a cellular process. Each metagene is defined *a priori* by genes that are either known to be uniquely upregulated or downregulated during a cellular process or during cell differentiation. These metagene constructs serve two purposes in this study. First, it simplifies the data, bringing it together in such a way that it can more easily understood. A DAG

Table 1: Gene list of Metagenes

Metagene	Genes	Ref. #
Proliferation	DNMT3B, MCM6, CDC25A, PFAS, MCM4, XRCC5, FAM29A, CXXC6, IGF2BP1, PLAA, DEPDC1B, TEX10, CCDC99, MSH6, DLG7, SKIV2L2, CENPE, CHEK2, SOHLH2, CCNB1, RRAS2, PRIM1, PAICS, CCNA2, CPSF3, NUSAP1, LIN28B, IMP5, KIF11, BMPR1A NDC80, BCAT1, CCNG1, ASCC3, FANCB, MCM10, HMGA2, SKP2, TRIM24, ORC1L, HDAC2, HESX1, C1orf45, INHBE, C21orf45, DCUN1D5, POLE2, MRPL3, CENPH, MYCN, CCDC5, GDF3, TBCE, RIOK2, BCKDHB, RAD1, C5orf13, ADH5, PLRG1, ROR1, RAB3B, DBC1, KIF23, DIAPH3, GNL2, FGF2, TARDBP, NMNAT2, ZNF167, KIF20A, CENPI, DDX1, C3orf21, GPR176, FBXO22, BBS9, C14orf166, FAM44B, CDC123, SNRPD3, FAM118B, PDH3, EIF2B3, KDELC1, APLP1, DACT1, PDHB, C14orf119, DTD1, SAMM50, CCL26, CCDC9-B, MED20, UTP6, RARS2, KIAA0020, ARMCX2, RARS, MTHFD2, DHX15, HTR7, HIST1H4C, MTHFD1L, ARMC9, XPOT, IARS, HDX, ARPM1, ERCC2, GARS, KIF7, HIP2, SLC25A3, ICMT, UGCGL2, ATP11C, SLC24A1, EIF2AKA, ALX1, DC2, TRPC4, HAS2, FZD2, TRNT1, SNX8, CDH6, HAT1, SEC11A, DIMT1L, TM2D2, FST, GBE1	[30]
EMT	SNAII2, COL5A2, FAP, POSTN, COL1A1, COL3A1, FBN1, TNFAIP6, MMP2, GREM1, BGN, CDH11, SPOCK1, DCN, COPZ2, THY1, PLOCE, PRRX1, PDGFRB, SPARC, INHBA, COL6A3, FN1, ACTA2, COL11A1, THBS2, COL10A1, COL5A2, LRRC15, COL5A1, MMP11, ADAM12, LOX, AEBP1, SULF1, ASPN, CTSK, HNT, EPYC, PLAU, OLFML2B, LUM, LOXL2, MXRA5, MFAP5, NUAK1, RAB31, TIMP3, CRISPLD2, ITGBL1, TMEM158, SFRP4, SERPINF1, C7orf10, NOX4, EDNRA, RCN3, C1QTNF3, COMP, LGALS1, COL6A2, GLT8D2, NID2, AXIN2, PITX2, MITF, NRCAM, TCF4, LGR5, FST, LEF1, FN1, FGF4, MMP7, RHOU, CLDN1, FGF18, MYC, MYCBP, JUN, FZD7, PPARD, WISP1, CTLA4, TNFRSF19, EN2, SP5, HNF1A, FOSL1, STRA6, VEGFA, ID2, WNT1, WNT10A, WNT10B, WNT11, WNT16, WNT2, WNT2B, WNT3, WNT3A, WNT4, WNT5A, WNT5B, WNT6, WNT7A, WNT7B, WNT8A, WNT8B, WNT9A, WNT9B	[31,32]
T-cell	CD247, CD3D, CDD3E, CD3G, ITGAL, ITGB2, ICAM1, CD2, CD28, THY1, PTPRC	[33]
Natural Killer Cells	KLRC1, KLRC2, KLRC3, KLRD1	[34]
Macrophages	CD14, MRC1, CPM, ITGAM, NOS2, HLA.DRA, HLA.DMA, HLA.DOA, HLA.DPA1, HLA.DQA1, HLA.DQA2	[35]
Th1	CD4, IFNG, IL10, FASLG, EOMES, TBX21	[36]
Th2	CD4, IL4, IL5, IL10, GATA3	[36]
Th17	CD4, IL17A, IL17F, RORA, RORC	[36]
Treg	CD4, TGFB1, IL10, IL12A, EBI3, RORC, FOXP3, TBX21, CCR6, MYB	[36]
Macrophage 1	IDO1, IL23A, IL12B, CCL17, IL1B	[35]
Macrophage 2	ARG1, TIMP2, LYVE1, KLF4, CD163, STAB1	[35]

containing fourteen nodes is much easier to make sense of then a DAG containing nearly three hundred nodes and is much more computationally traceable[27]. The reduced computational expense enables one to test hypothesis related to network topology via simulations, for instance, the statistically significance of an edge can be obtained by comparing how often an edge is inferred from the TCGA data relative to a dataset that has no information – that is, a null hypothesis. Secondly, it serves as a means of helping to eliminate error. Microarray data is noisy, with the result given being the summation of both the true gene expression as well the noise inherit to the assay (i.e., lab variability, experimenter skill, sensitivity of the machine, and batch of reagents used)[24]. The metagene helps eliminate this error as it averages across several genes. The genes that make up each metagene are identified in Table 1.

The presence of T cells, Natural Killer cells, and macrophages in the tumor microenvironment were represented by immune infiltrate metagenes[33]. The value for an immune infiltrate metagene were calculated according to the formula:

$$v = \sum_{j=1}^{n} y_j/n$$

where n represents the number of genes in the metagene and y_j equals the expression of gene j in the metagene.

The EMT[31,32] and proliferation[30] metagenes were calculated according to the formula:

$$v = \frac{n}{\sum_{j=1}^{n}((y_j - (\bar{y}_j + 3\sigma_j))/(\sigma_j))^2,}$$

where n represents the number of genes in the metagene, y_j equals the gene expression of the jth gene in the metagene, $\bar{y}_j$ represents the mean of gene y's expression in the dataset and σ_j equals the standard deviation of the gene.

Polarization of T cells into one of four subsets and of macrophages into one of two subsets[35,36] were calculated according to the equations below

$$v_i = \prod_{j=1}^{n} (\sigma_j / (y_j - (\bar{y}_j \pm 3\sigma_j)))^2,$$

$$P(M_i|Y) = \frac{v_i}{\sum_{k=1}^{m} v_k},$$

where y_j equals the gene expression of the jth gene in the metagene, $\bar{y}_j$ represents the mean of gene y's expression in the dataset and σ_y equals the standard deviation of the gene. The standard deviation is either added or subtracted from the mean depending on whether the gene is upregulated or down regulated in the polarization, with upregulated genes being added and down regulated genes being subtracted. $P(M_i|Y)$ is the probablilty of polarization state i given the data Y, and m represents the total number of possible polarizations. Products were used as we assumed the expression of the genes to be mutually inclusive, with a polarization only being considered when all of the genes for it were upregulated when compared to the gene expression of the alternative polarizations. All calculations were performed in R[37].

2.3 Bayesian Networks

Bayesian networks were generated from the metagene data using an Incremental Associated Markov Blanket (IAMB) as described by Tsamardinos[38] and implemented in R. In short, IAMB is made up of two phases – a forward stage where a network is generated in such a way that it maximizes the conditional independence of the nodes, and a backwards phase where it is removes any remaining conditionally independent connections. This results in the construction of a Markov blanket. A Markov blanket of a node is defined as the set of other nodes that contains all the information in the data set that can aid in predicting the value of the node of interest. A simple way to think of a Markov blanket is to take a set that contains all of the nodes that are related to the node of interest, then remove any nodes from that set that are conditionally independent, and add in any nodes that are conditionally dependent. Conditional independence is the situation where two variables that are partially dependent on each other when viewed in isolation become independent when combined with a third variable. This can be explained using a simple example – the genetic information of a child, their father, and their paternal grandmother. The genome of the child is correlated with that of their grandmother (as would be expected, $1/4^{th}$ of their genes came from the paternal grandmother). However, the child is conditionally independent of their paternal grandmother if we know the genome of the father – since any genetic information shared between the grandmother and the child would need to go through the father, nothing more can be inferred about the child's genome knowing both the father's and paternal grandmother's genome than can be inferred if you know just the fathers genome. Note that the reverse is not true – a child is not conditionally independent of their parent given the grandparent, we can still infer more about the child's genome. In a sense, identification of conditional independence allows for the identification of intermediates – the

genetic information of the paternal grandmother was transported to the child through the father. Another concept that goes along with conditional independence is conditional dependence – where two variables that are independent of each other when observed in isolation become dependent with the addition of a third variable. To carry the example from earlier further, the genome of the mother is independent of the genome of the father, meaning you can't infer anything about the mother's genome by knowing the father's genome. However, the two genomes become dependent if the child's genome is known – you can infer more about the mother's genome if you know the genome of both the child and the father than if you knew the genome of just the child. Thus, the Markov blanket for the father includes the father's parents, his child, and his wife, but would not include the father's grandparents, siblings, or grandchildren. It should also be noted that reversed networks are functionally equivalent. This distinction is important when it comes trying to define causal relationships; while from a Bayesian network standpoint a → b → c is functionally the same as c → b → a, it does not make sense to say that a child's genome influences their father's genome, which in turn influences the grandparent's genome[39]. To get around this limitation, one of two things must be done: first, one of the directional relationships needs to be defined *a priori* or second, the dataset used to generate the network needs to include temporal data[40].

Confidence for the node edges was calculated using a bootstrap resampling method that included 100,000 replications[39]. For each replicate, patient data was randomly sampled with replacement n times, where n is the starting number of patients in the dataset, and a network was generated from the new dataset. Lines were only included if they had a p-value of less than 0.01. Bayesian networks were generated for all cancer sets. To assess how the complexity of the TCGA study samples influenced network generation, we also generated networks for smaller

subsets of in data, including the tumor samples only or a percentage of the entire dataset (75%, 50%, 25%). These percentage subsets were generated by sampling without replacement from the entire cohort. Model generation and averaging were performed in R using the boot.strength() and the average.network() methods from the bnlearn package. Since temporal data from a single patient was not available, we used matched normals and various patients to simulate temporal data. We assume these cross-sectional samples from normal and diseased tissue represents random samples from a common temporal trajectory associated with oncogenesis.

2.4 Statistics

Expression data for only the genes of interest were analyzed in R, and heatmaps were generated using the heatmap.2() function. Patients and genes were clustered using the ward method. Principal component analysis was performed on the data using the prcomp() function, with scale set to false, and rotational data for the genes were returned. Differences in node connectivity and average Markov blanket size were compared using two way ANOVA's. Distributions between cancer group 1 and cancer group 2 were compared using the Mann-Whitney-Wilcoxon test. P-values for the Mann-Whitney-Wilcoxon test that were less than 0.01 were considered to be statistically significant.

3 Results

3.1 Identification of patient subtypes and metagenes

While the genes we used to identify the different metagenes were identified and created using human data, we wanted to verify their usefulness, both in their use to distinguish between patients, and also to see if they did, in fact, vary together. To see if the genes selected could distinguish cancer and non-cancer patients, hierarchical clustering was performed on the genes used in the metagenes (fig. 1), and the patients were divided into three groups. The patients, with a couple of exceptions, divided into three groups, one normal group and two cancer groups. This suggests that the patients, at least, can be separated by the genes chosen for the metagenes, with normal samples being separated from the tumor samples, and the tumor samples being split into two groups, one of which more closely resembles the normal samples.

In order to better understand how the different metagenes explained the variance samples observed, Principal Component Analysis was performed. It was found that a large percentage of the variance was explained in the first four principal components (fig. 2a, 54%), and that the genes associated with EMT and proliferation were separated by principal component 1 (fig. 2b – fig 2d). This can be observed by looking at the distribution of the genes, with most of the proliferation metagenes grouping towards the far right on PC1, whereas the EMT genes tended towards the center and left portions. This also means that since the behavior of 256 genes can be simplified to 4 dimensions, that many of the genes are varying consistently with each other. It should be noted that there are several genes that cluster near the origin in all 4 principal components – these genes could potentially be removed, as they are uninformative with regards to the breast cancer, but were not excluded in this case as they could be informative in other cancers.

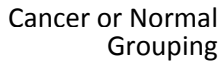


Figure 1: Hierarchical clustering of breast cancer patient data separates the patients into cancer and non-cancer patients. Hierarchical clustering of the patient gene signatures separate the patients into three groups, two cancer groups and one normal group. Hierarchal clustering was performed on patient data for the genes listed in Table 1. Patients were colored based on whether the sample came from normal breast tissue (blue) or tumor breast tissue (red) and were grouped into three groups, group 1 (black), group 2 (purple), and group 3 (green). Genes were color coded based on metagene grouping, EMT (black), Proliferation (grey), T cell infiltration (green), NK cell infiltration (red), Macrophage Infiltration (orange), T cell polarization (yellow), and Macrophage polarization (blue)

Since both of these metagenes informed principal component 1, we looked to see how the distribution of these metagenes differed between the two cancer groups and the normals. As would be expected, it was found that expression of the proliferation metagene was increased in both cancer group 1 and cancer group 2 when compared to the normal group, although the two cancer groups did not have identical distributions (fig. 3b). However, it was observed that expression of the EMT metagene was increased primarily in cancer group 1, with cancer group 2 and the normal group having a very similar distribution (fig. 3a). Similar to the group differences in proliferation the macrophage polarization was very different between the cancer and normal patients, with the cancer patients exhibiting a shift from the M2 to the M1 polarization, with group 2 exhibiting a much stronger bias(fig 3c). This is of interest, as both polarization play different and opposite roles in cancer immunology. This is not surprising, as the M1 macrophage, also known as the classically activated macrophage, is activated via inflammation while the M2 mostly plays a role in wound healing. As such, the two cancer metagenes can mostly separate the patients into the three groups found in the hierarchical clustering - at least in regards to explaining the overall variance seen in the gene expression - with the proliferation metagene primarily showing whether the biopsy came from a normal or tumor, and the EMT metagene seemingly mostly distinguishing between different cancers. Due to the role that EMT plays in invasiveness[41], however we could not identify a relationship between EMT expression and existing distal metastasis. This may be due to the fact there was a very low number of patients who had metastasized (n = 8, Supplemetal Table 8).

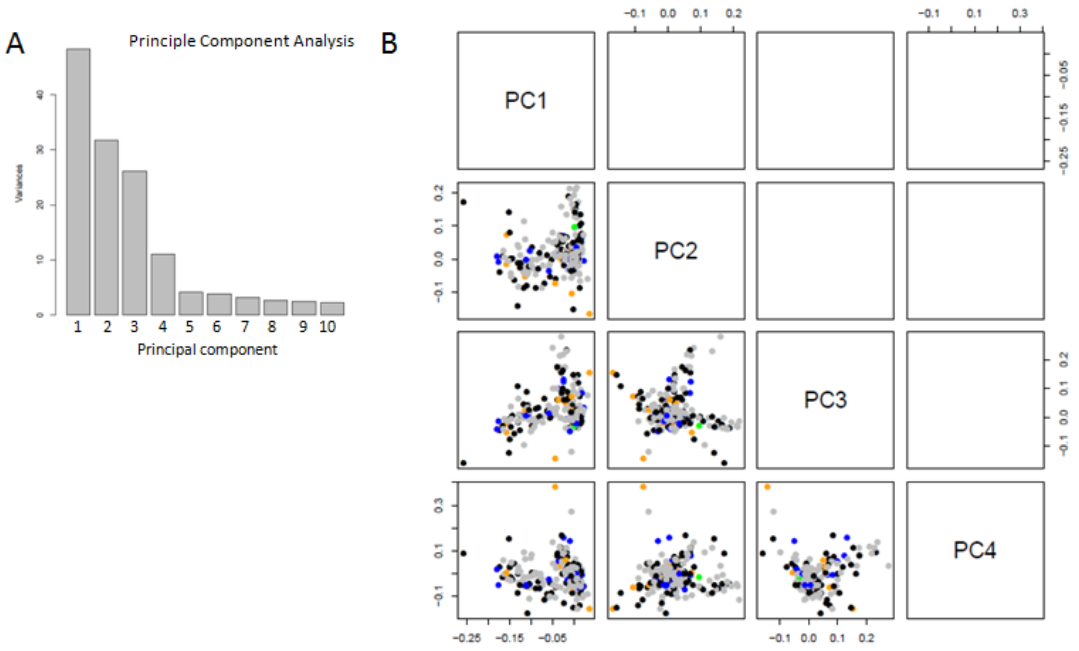


Figure 2: Most of the variability in the gene data can be captured in the first four principal components. Principal component analysis was performed on the data from fig.1, and the amount of variability explained by each component was graphed (a). Genes were color coded based on metagene grouping, EMT (black), Proliferation (grey), T cell infiltration (green), NK cell infiltration (red), Macrophage Infiltration (orange), T cell polarization (yellow), and Macrophage polarization (blue), and a rotational graph representing the role of each gene in principal components 1, 2, 3, and 4 were generated (b).

Figure 3: The distribution of the EMT (a), the proliferation (b), and the M1 macrophage polarization (c) were calculated and the group averages were displayed as a solid vertical line. Groupings were based on figure 1, with black representing cancer group 1, purple representing cancer group 2, and green representing normals.

3.2 Use of Bayesian networking as a means to identify topology of extracellular control networks

After confirming that the genes naturally cluster into the different metagenes, we asked whether we could observe evidence of crosstalk between the cancer and immune metagenes. To accomplish this, a Bayesian network of the metagenes was generated using an IAMB algorithm. Although Bayesian networks can be used to identify causal relationships between data points, causality can only be inferred if there is either temporal data, or the direction of the edges are known *a priori*. Since we could not analyze multiple cancer biopsies from a single patient from across time, we used the whole dataset along with the matched normals as a means to simulate temporal disease progression[40]. The generated network represented the averaging of 100,000 generated networks, a process that had previously been shown to be a fairly conservative method of identifying edges[39]. When the analysis was performed on the whole data set of invasive breast cancer, it was observed that the EMT metagene and the proliferation metagene had reciprocal effects on macrophage polarization, with EMT seemingly being associated with increases in macrophage type 2 polarization (M2) and proliferation being associated with macrophage type 1 polarization (M1) (fig 4a). This is of interest, as macrophage polarizations are thought to play opposing roles in cancer immunosurveillance. The M1 polarization is the classical macrophage, which serves to scavenge cell debris and is generally pro-inflammatory[42]. In contrast, the M2 polarization is associated with wound healing, suppression of inflammation, and is considered to promote tumor growth[43,44]. These relationships are captured in the generated network, with an M1 polarization also being associated with an increase in overall T cell infiltration and M2 being associated with a decrease in T cell infiltration. When the analysis was repeated using only gene expression values derived from tumor samples of invasive breast cancer cohort, the relationships

between EMT and macrophage polarization, as well as the relationship between macrophage polarization and T cell infiltration persisted (fig 4b). However, the relationship between macrophage polarization and proliferation was lost. This implies that the relationship between macrophage polarization and proliferation was mostly informed by the change from normal to cancer. It is also worth noting that not all relationships with proliferation were lost with the change from using all samples to only using cancer samples. For instance, the relationship between proliferation and the T helper Type 1 cells (Th1) polarization was maintained, and the confidence was, in fact, increased (p-value = 1.45E-20 vs. p-value = 7.0E-15). This suggests that the relationship between the Th1 polarization and proliferation is a relationship inherent to invasive breast cancer, and not simply representing a change from normal to cancer.

Since EMT and proliferation played opposing roles in the Bayesian networks with regards to macrophage polarization, we next wanted to see whether their overall distribution was different in the two cancer groups. It was found that the two groups did differ significantly with regards to their average EMT metagene expression, with group 1 having a higher average expression (p value < 0.001, Mann-Whitney Wilcox, Supplemental Figure 1i). Surprisingly, though the means were much closer together, the difference between the two groups in the average proliferation metagene expression was statistically significant, with group 2 having a higher expression (p value = 2.95e-7, Supplemental Figure 1j). As would be expected from the Bayesian network, it was found that the two groups did differ significantly with regards to their macrophage polarization, with group 2 having higher levels of M1 polarized macrophages (p value = 9.40e-15, Supplemental Figure 1h). As would be expected with the changes in macrophage polarization, group 1 also had a statistically significant increase in Th2 and Treg polarizations, with Treg being anti-inflammatory and Th2 being commonly opposed to Th1, and

Figure 4: Bayesian networks reveal cross-talk among polarized immune subsets and inverse relationships between the proliferation and EMT metagene with regards to macrophage polarization. Bayesian networks were generated for BC data using either the whole data set (a) or just the cancer dataset (b). In both cases the network was generated using an IAMB algorithm, with the network representing the average of 100,000 generated Bayesian networks. Black lines represent positive relationships while red lines represent negative relationships. The line thickness is proportional to the negative log of the confidence in the connection, with thicker lines representing a higher confidence. Confidence p values are given in the supplemental table 1.

Th1 polarization being higher in group 2. These interactions, however, did not appear to be direct effects as indicated by the Bayesian network inference results.

While the objective of using Bayesian network inference is to identify more complex multivariate relationships within the dataset, we also wanted to see if the relationships between cancer and immune metagenes could be observed directly. Since the whole data set was required to observe both relationships, we looked to see if there was much of a direct correlation between either the EMT metagene or the proliferation metagene and macrophage polarization. We found a very weak but significant correlation between EMT and macrophage polarization ($r = -.18$, p-value $< .001$, fig 5), while the correlation between proliferation and macrophage polarization was a bit stronger ($r = .34$, p-value $< .001$, fig 5). This, however, is not too surprising. First, while the proliferation and EMT metagenes exist as a continuum, the macrophage polarization metagene is mostly binary. Secondly, the Bayesian network suggested that the interactions between EMT and macrophage polarization was rather subtle, as shown by the larger p-value when compared to the p-value for proliferation and macrophage polarization (p value of 5.07e-5 vs 7.38e-16), which can be considered as a strength of evidence.

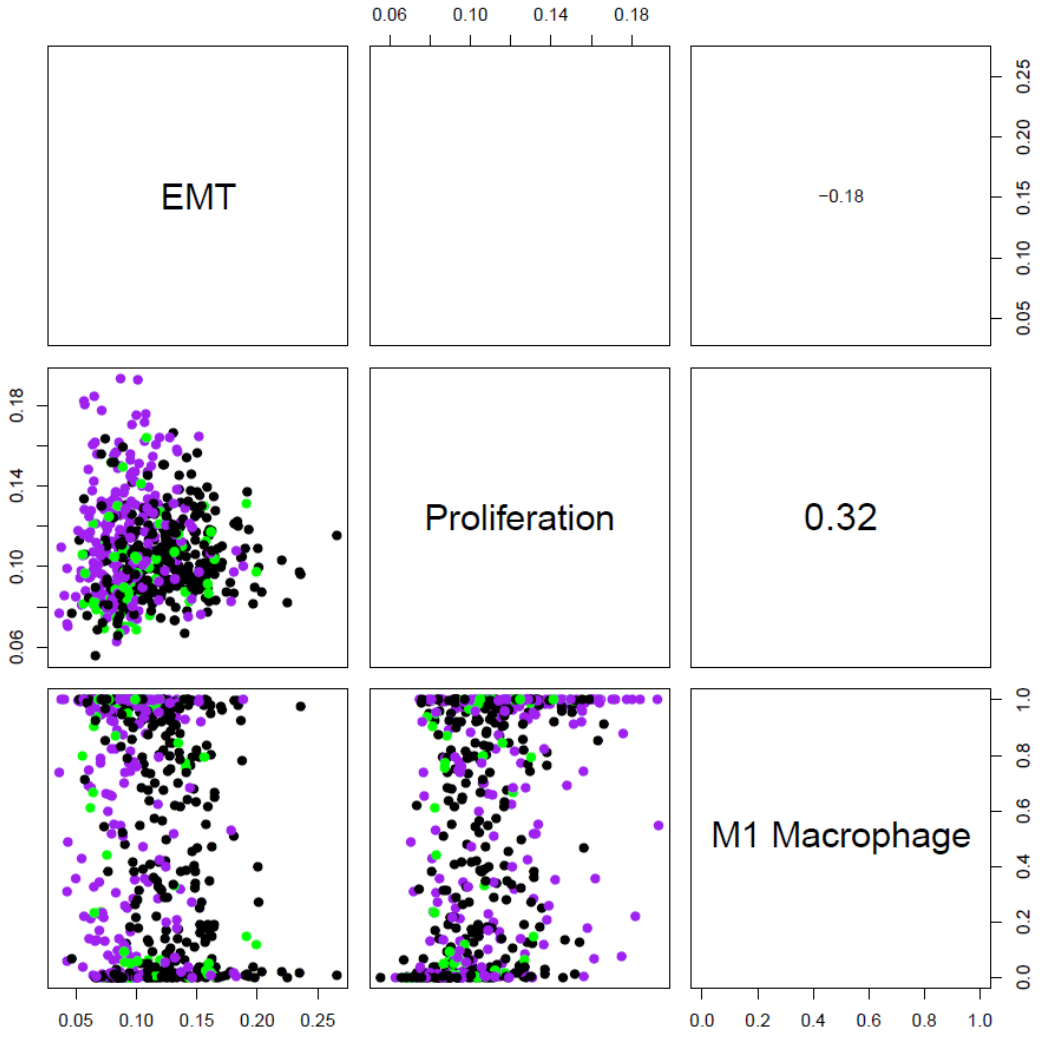


Figure 5: The relationships between EMT, Proliferation, and M1 Macrophage polarization show subtle relationships between themselves. Below the parallel correlation graphs are shown for the three metagenes. Black dots represent data from cancer group 1, purple dots show data from cancer group 2, and green shows data from matched normals. Above the parallel is the correlation coefficient, size scaled to strength of relationship.

3.3 The sample size and diversity of a data set influences network generation and verification of metagene constructs

We next determined the overall effect of sample size and diversity of the TCGA dataset with regards to the generated networks to assess how this approach could be generalized. To accomplish this, we generated mock datasets that represented 75% (fig 6a), 50% (fig 6b), or 25% (fig 6c) of the dataset by drawing randomly without replacement from the whole invasive breast cancer data set. As one would expect, the network appeared to be become progressively less complex as the dataset became smaller. It also appeared that the overall confidence levels associated with the edges fell. Interestingly, when the dataset was reduced to 25%, the relationship between EMT and macrophage polarization was lost, although the relationship between proliferation and macrophage polarization endured. This is consistent with our findings that the relationship between EMT and macrophage polarization was more subtle, and also implies this relationship might not be identifiable with a smaller dataset.

The relationship between proliferation and macrophage polarization is interesting, as it was lost with the removal of the normal breast tissue samples but persisted in the 75% and 50% subsets. To better examine the impact of the inclusion of normal samples, we repeated the experiment using only cancer samples (fig 7). In this case, the relationship between EMT and macrophage polarization was maintained through all datasets. While the relationship between proliferation and macrophage polarization was lost in all subsets, the relationship between proliferation and Th1 polarization was also maintained. It is also interesting to note that these datasets generated orphan nodes – nodes that contain no connections to any other. This suggests that the inclusion of normal patient data was required for the identification of the relationship

between proliferation and macrophage polarization. In contrast, the exclusion of normal patient data made it easier to identify the relationship between EMT and macrophage polarization.

In order to get a better idea of the overall effects of changing the complexity and size of the dataset used, 10 replicates of the earlier studies were performed, each containing 100,000 bootstrap samples, and the resulting complexity of the inferred networks was quantified by their average node connectivity and Markov blanket size (table 2). In this particular case, the amount of samples drawn from either dataset was equal to the percentage of cancer samples only, since we wanted to directly compare the tumor only and tumor and normal datasets to each other without the influence of overall difference in number between cancer and combined groups influencing the network complexity. What we found was that average Markov blanket did not change between using the entire dataset versus using cancer dataset only, though there was a statistically significant difference in average Markov blanket size and node connectivity when comparing the smallest dataset (25%) to the largest dataset (75%, p-value $<$ .01). There was, however, a decrease in node connectivity when the network was generated from cancer dataset alone (p-value $<$ .01).

One potential concern of the model that we are using is that while certain metagenes are defined independently of each other (for example, immune cell infiltration, proliferation, and EMT), other metagenes, such as immune polarization, are defined as mutually exclusive. For instance, M1 macrophage polarization is defined both by the increased expression of M1 associated genes and by a decrease in M2 associated genes. To test our inference approach, we tested whether the immune polarization networks were informed by the data or constrained by the particular model formulations. In order to test this, we focused on the T helper cell polarizations, and compared the connections generated from real data to the connections derived

Figure 6: The relationship between EMT and macrophage polarization appears to be subtle, and is lost when the network is generated from subsamples of the dataset. Bayesian networks were generated using random sampling without replacement of the whole data set, with subsamples representing 75% (a), 50% (b), or 25% (c) of the whole BC dataset. In all cases the network was generated using an IAMB algorithm, with the network representing the average of 100,000 generated Bayesian networks. Black lines represent positive relationships while red lines represent negative relationships. The line thickness is proportional to the log of the confidence in the connection, with thicker lines representing a higher confidence. Confidence p values are given in the supplemental tables 5, 6 and 7.

Figure 7: The relationship between EMT and macrophage polarization is maintained when subsamples are taken of only the cancer samples. Bayesian networks were generated using random sampling without replacement of the BC dataset containing only tumor biopsies, with subsamples representing 75% (a), 50% (b), or 25% (c) of the BC dataset. In all cases the network was generated using an IAMB algorithm, with the network representing the average of 100,000 generated Bayesian networks. Black lines represent positive relationships while red lines represent negative relationships. The line thickness is proportional to the log of the confidence in the connection, with thicker lines representing a higher confidence. Confidence p values are given in the supplemental tables 5, 6 and 7.

from metagenes generated from random genes. To accomplish this, we scrambled the genes associated with each polarization subset and repeated the analysis. What we found was a redirection in relationships (specifically Th1 no longer having connections to T helper type 17 (Th17) and T helper type 2 cells) (fig 8). Furthermore, in the repeated analysis, there was an ambiguity in the nature of the relationship between Th1 and Treg, with some models finding a positive relationship while other networks identified a negative relationship. However, the overall shape of the graph was consistent across all three gene reshufflings, which suggests that the data plays at least some role in determining the final relationships. Furthermore, prior studies had identified a reciprocal role for Th1 and Th17 in human tumor infiltrates[45].

Table 2: Average connectivity and Markov blanket size of invasive breast cancer subsets

% of Dataset	Cancer + Normals, Average node connection	Cancer Only, Average node connection **	Cancer + Normals, Markov Blanket Size	Cancer Only, Markov Blanket Size **
Whole Dataset	2.92	2.46	3.69	3.07
75% (n = 399)	2.95 ± 0.06	2.61 ± 0.10	3.67 ± 0.20	3.23± 0.17
50% (n = 266)	2.87 ± 0.13	2.43 ± 0.15	3.56± 0.18	3.21± 0.49
25%* (n = 133)	2.28 ± 0.12	2.15 ± 0.25	2.71± 0.25	2.84± 0.64

Mean and standard deviation of average node connections and Markov blanket size were calculated for the whole invasive breast cancer dataset and cancer only invasive breast cancer dataset. Percentages were based on the size of the cancer only datasets. 2-way ANOVA was performed on the data. * signifies a difference between that row and the 75% row. ** signifies a difference between that column and the total data set

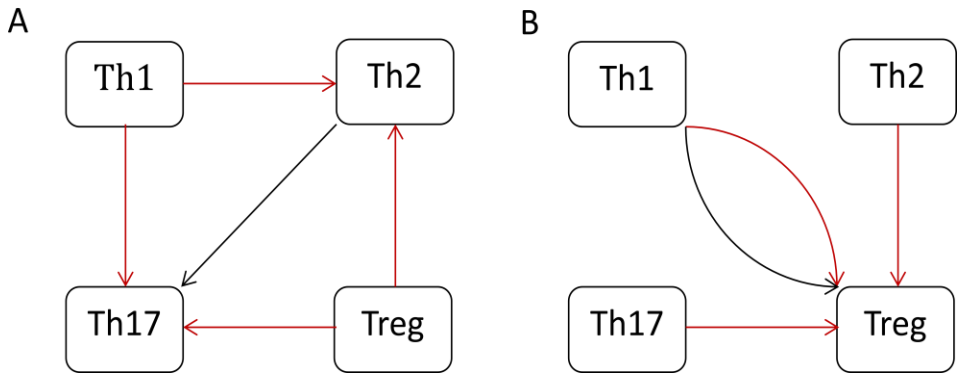


Figure 8: The relationships between the T-cell polarizations are defined by the data, and not by the model. The genes used to identify each T helper call subtype were replaced with random immune genes used in the study, and the probability of the polarizations were recalculated. Bayesian networks were then generated using an IAMB algorithm, with the network representing the average of 100,000 Bayesian networks generated from resampling. This was replicated 6 times. A) represents networks from real data and B) represents data from the random data. Black lines represent positive relationships while red lines represent negative relationships. Dual arrows represent were both relationships were observed.

3.4 Similar Bayesian networks are identified in other cancers

One of the advantages of using the TCGA study is that it spans many different cancers and samples are processed similarly. This helps minimize variation in network inference that may be introduced by study design. We wanted to determine if the relationships we observed were specific to breast cancer, or if similar relationships could be observed in other cancers. In order to do this, we downloaded complete datasets from the lung squamous cell carcinoma (fig 9a), colon adenocarcinoma (fig 9b), and glioblastoma multiform (fig 9c) arms of the TCGA study. While glioblastoma technically is not a carcinoma, it had been reported that it does undergo a shift towards a mesenchymal state, resulting in an increase in expression of EMT genes[46]. Overall the generated Bayesian networks for the other cancers were similar but not identical to the overall network inferred from the breast cancer dataset. For example, identical relationships between proliferation and macrophage polarization were observed in the colon adenocarcinoma and glioblastoma multiform datasets, but were not seen in the lung squamous cell carcinoma dataset. In addition, the relationship between EMT and macrophage polarization was observed only in the lung squamous cell carcinoma dataset. Differences in the Bayesian networks were not unexpected, given the size and composition of the datasets (Table 3). For example, the lung squamous cell carcinoma dataset, while small, only contains tumor biopsies, and identified the least complex network. In contrast, the colon adenocarcinoma dataset is small but contains a mixture of both normal and tumor biopsies and the inferred network has more connection. The glioblastoma multiform network is interesting as it was the most complex and contained more connection than the other cancer networks, but it also represents relationships that, unlike all the other networks, arise in an immune privileged area[47].

Table 3: Node connectivity and Markov blanket size of all cancer data sets used in study.

Cancer	Average node connectivity	Average Markov Blanket Size	Number of Tumor Samples	Number of Normal Samples
Breast Cancer	2.77	3.69	532	65
Glioblastoma Multiform	2.92	4.00	467	10
Lung Cancer	1.85	2.40	154	0
Colorectal Cancer	2.31	2.46	154	19

Values represent the mean node connectivity and Markov blanket size for the networks generated using the breast cancer, glioblastoma multiform, lung cancer, and colorectal cancer datasets. The total numbers of tumor and normal samples used in the analysis are also provided.

Figure 9: Bayesian networks of other cancers. Bayesian networks were generated using the metagenes for the lung cancer (a), colon adenocarcinoma (b), and glioblastoma multiform (c) datasets. In all cases the network was generated using an IAMB algorithm, with the network representing the average of 100,000 generated Bayesian networks. Black lines represent positive relationships while red lines represent negative relationships. The line thickness is proportional to the log of the confidence in the connection, with thicker lines representing a higher confidence. Confidence p values are given in the supplemental tables 2, 3 and 4.

4 Discussion, conclusions, and recommendations

The purpose of this study was to identify evidence of alterations in the normal behavior of the immune system that are causally related to oncogenic changes, namely an increase in proliferation and EMT. As a means of accomplishing this, we used gene expression data obtained from tumor and normal tissue biopsies in conjunction with defined gene signatures, called metagenes, which are indicative of immune infiltration, immune polarization, and common cancer processes, to infer relationships among these processes via Bayesian networking. We used data from the invasive breast cancer arm of the TCGA to generate directed acyclic graphs and used model averaging to establish confidence in the network topology. As a form of external validation, we found that similar network structures were observed in other cancers in a manner consistent with the size and diversity of the underlying datasets. In summary, we have outlined a novel method of identifying areas of local crosstalk between different cells within the tumor microenvironment using microarray data and prior knowledge of gene signatures.

As the overall objective was to identify relationships among biological processes associated with tissue homeostasis and immune-mediated control of multicellular tissues, the inferred networks identified some interesting crosstalk among these processes. In particular, we found that an increase in proliferation tended to coincide with increases in cell-mediated immune responses that promote cancer destruction while EMT increases tended to coincide with increases in cell-mediated immune responses that either did not kill the cancer or which would help promote tumor tolerance. For example, in the lung cancer and glioblastoma multiform datasets increased proliferation led to increases in M1 polarized macrophages, the same relationship, but in the opposite direction, was identified in the colorectal adenocarcinoma

dataset. Proliferation was also found to be associated with increases in Th1 cells in colorectal and breast cancer data sets and with increases in M1 macrophage polarization and natural killer cell infiltration in glioblastoma multiform. A type I cell-mediated immune response is generally considered to be the response that has a positive overall impact on cancer survival,[45] as it uses Th1 polarized CD4 T lymphocytes, CD8 T lymphocytes[48] and natural killer[49] cells as effector cells to help destroy the cancer. At the same time, increased EMT activity was associated with increases in M2 polarization in the lung and breast cancer and appeared to be driven by decrease in natural killer cells in glioblastoma multiform. Of all the cancers analyzed, glioblastoma multiform was the most divergent. However, this could be simply due to the fact that these interactions are occurring in an immune-privileged area, with the underlying immune processes being different that what would be observed in a non-privileged area.

These networks provided a topology and directionality of the intracellular networks at work in a tumor microenvironment. However, certain aspects of the directionality remain uncertain. From this study, we had reversed directionality with regards to macrophage polarization and EMT if the analysis was performed with either the whole breast cancer dataset or just the malignant samples, which begs the question which model most closely reflects what occurs within the patient. While the relationship between macrophage polarization and EMT had not been reported before in breast cancer, similar relationships have been observed in other forms of cancer, for instance, M2 macrophages are the most common polarization for cancer associated macrophages[50] and promote EMT *in vitro*[51]. In melanoma, tumor cells induce immune-suppression when they undergo EMT[52]. Alternatively, M2 polarized macrophages have been shown to induce EMT in certain forms of pancreatic cancer[53]. Also, feedback loops are a common motif in biological systems, but are necessarily removed in the directed acyclic graphs

used in our analysis. This would result in the generation of overly simplistic models – for example, a positive feedback loop would be interpreted as a straight forward causal relationship.

It is also important to remember the assumptions that are used in these analyses. One of the larger assumptions is that we can replicate temporal cancer progression using samples from different cancer biopsies and that cancer follows a single course. The temporal aspect of the data is limited partially by the fact the biopsies are all taken at diagnosis before treatments has begun. As such, we do have access to information from more advanced tumors. Also, while we have tumor biopsies and matched normal biopsies, we do not have any data on intermediate data between the two, and thus are missing part of the progression. Finally, we needed to assume that there is a common cancer progression, which, as more data is acquired, may not be the case.

Despite these limitations, these networks do give us an understanding of what relationships are occurring in human cancer progression. This can be used to help identify and verify model systems that more closely mimic human disease progression, resulting in the selection of more relevant models. For example, longitudinal studies using mouse models that mimic the metagene signatures associated with oncogenesis may help inform ambiguities in our causal networks as well as serve as a relevant model for testing new treatments. Furthermore, these models can be used to identify instances of feedback loops. As the amount of available data increases, it will become possible to create networks for different subsets of cancer progression, which in turn could help in the identification of model systems that better replicate certain forms of cancer progression.

One particular strength of this approach is the versatility offered by the use of metagenes. Since the generation of a metagene only requires that the relevant cells be in the biopsy and an

accepted list of genes that are differentially expressed by the cells of interest, it is possible to analyze the relationship between a wide variety of processes. Furthermore, it would also be possible to include other, global measurements as well as analysis of distal tissue, allowing for the generation of more systemic models. In summary, we have used an existing technique in a novel method to observe changes in intercellular relationships using data obtained as part of the Cancer Genome Atlas. This technique can be used to help identify more relevant disease models and can be applied to a wide range of more complicated diseases that engulf tissues, complicated processes associated with tissue development, and regenerative medicine.

5 References

1. Sung, H.-K. *et al.* Adipose vascular endothelial growth factor regulates metabolic homeostasis through angiogenesis. *Cell Metab.* **17,** 61–72 (2013).
2. Csete, M. & Doyle, J. Reverse engineering of biological complexity. *Science* **295,** 1664–1669 (2002).
3. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
4. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194,** 23–28 (1976).
5. Cairns, J. Mutation selection and the natural history of cancer. *Publ. Online 15 May 1975 Doi101038255197a0* **255,** 197–200 (1975).
6. Merlo, L. M. F., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nat. Rev. Cancer* **6,** 924–935 (2006).
7. Anderson, A. R. A., Weaver, A. M., Cummings, P. T. & Quaranta, V. Tumor morphology and phenotypic evolution driven by selective pressure from the microenvironment. *Cell* **127,** 905–915 (2006).
8. Basanta, D., Hatzikirou, H. & Deutsch, A. Studying the emergence of invasiveness in tumours using game theory. *Eur. Phys. J. B - Condens. Matter Complex Syst.* **63,** 393–397 (2008).
9. Aktipis, C. A., Maley, C. C. & Pepper, J. W. Dispersal Evolution in Neoplasms: The role of disregulated metabolism in the evolution of cell motility. *Cancer Prev. Res. (Phila. Pa.)* **5,** 266–275 (2012).
10. Mazzone, M. *et al.* Heterozygous deficiency of PHD2 restores tumor oxygenation and inhibits metastasis via endothelial normalization. *Cell* **136,** 839–851 (2009).

11. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100,** 57–70 (2000).
12. Visser, K. E. de, Eichten, A. & Coussens, L. M. Paradoxical roles of the immune system during cancer development. *Nat. Rev. Cancer* **6,** 24–37 (2006).
13. Enderling, H., Hlatky, L. & Hahnfeldt, P. Immunoediting: Evidence of the multifaceted role of the immune system in self-metastatic tumor growth. *Theor. Biol. Med. Model.* **9,** 31 (2012).
14. Seliger, B. Strategies of tumor immune evasion. *BioDrugs Clin. Immunother. Biopharm. Gene Ther.* **19,** 347–354 (2005).
15. Seliger, B. *et al.* Association of HLA class I antigen abnormalities with disease progression and early recurrence in prostate cancer. *Cancer Immunol. Immunother.* **59,** 529–540 (2010).
16. Wilson, E. B. *et al.* Human tumour immune evasion via TGF-β blocks NK cell activation but not survival allowing therapeutic restoration of anti-tumour activity. *PLoS ONE* **6,** e22842 (2011).
17. Kulkarni, Y. M. *et al.* A quantitative systems approach to identify paracrine mechanisms that locally suppress immune response to Interleukin-12 in the B16 melanoma model. *Integr. Biol.* **4,** 925 (2012).
18. Richmond, A. & Su, Y. Mouse xenograft models vs GEM models for human cancer therapeutics. *Dis. Model. Mech.* **1,** 78–82 (2008).
19. Mestas, J. & Hughes, C. C. W. Of mice and not men: differences between mouse and human Immunology. *J. Immunol.* **172,** 2731–2738 (2004).
20. Shay, T. *et al.* Conservation and divergence in the transcriptional programs of the human and mouse immune systems. *Proc. Natl. Acad. Sci.* **110,** 2946–2951 (2013).

21. Sieberts, S. K. & Schadt, E. E. Moving toward a system genetics view of disease. *Mamm. Genome* **18,** 389–401 (2007).

22. Friedman, N. Inferring cellular networks using probabilistic graphical models. *Science* **303,** 799–805 (2004).

23. Sachs, K., Gifford, D., Jaakkola, T., Sorger, P. & Lauffenburger, D. A. Bayesian network approach to cell signaling pathway modeling. *Sci. Signal.* **2002,** pe38 (2002).

24. Perrin, B.-E. *et al.* Gene networks inference using dynamic Bayesian networks. *Bioinformatics* **19,** ii138–ii148 (2003).

25. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A. & Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308,** 523–529 (2005).

26. Sachs, K. *et al.* Learning Signaling Network Structures with Sparsely Distributed Data. *J. Comput. Biol.* **16,** 201–212 (2009).

27. Zou, M. & Conzen, S. D. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* **21,** 71–79 (2005).

28. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

29. Bolstad, B. M., Irizarry, R. A., Åstrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19,** 185–193 (2003).

30. Palmer, N. P., Schmid, P. R., Berger, B. & Kohane, I. S. A gene expression profile of stem cell pluripotentiality and differentiation is conserved across diverse solid and hematopoietic cancers. *Genome Biol.* **13,** R71 (2012).

31. Cheng, W.-Y., Kandel, J. J., Yamashiro, D. J., Canoll, P. & Anastassiou, D. A multi-cancer mesenchymal transition gene expression signature is associated with prolonged time to recurrence in glioblastoma. *PLoS ONE* **7,** e34705 (2012).
32. The Wnt Homepage. (2013). at <http://www.stanford.edu/group/nusselab/cgi-bin/wnt/>
33. BioCarta. at <http://www.biocarta.com/Default.aspx>
34. Plougastel, B. & Trowsdale, J. Sequence analysis of a 62-kb region overlapping the human KLRC cluster of genes. *Genomics* **49,** 193–199 (1998).
35. Movahedi, K. *et al.* Different tumor microenvironments contain functionally distinct subsets of macrophages derived from Ly6C(high) monocytes. *Cancer Res.* **70,** 5728–5739 (2010).
36. Wei, G. *et al.* Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. *Immunity* **30,** 155–167 (2009).
37. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
38. Tsamardinos, I., Aliferis, C., Statnikov, A. & Statnikov, E. Algorithms for large scale Markov blanket discovery. in *16th Int. FLAIRS Conf. St* 376–380 (AAAI Press, 2003).
39. Friedman, N., Goldszmidt, M. & Wyner, A. *Data Analysis with Bayesian Networks: A Bootstrap Approach*. (1999).
40. Bar-Joseph, Z., Gitter, A. & Simon, I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* **13,** 552–564 (2012).
41. Chen, W. C. & Obrink, B. Cell-cell contacts mediated by E-cadherin (uvomorulin) restrict invasive behavior of L-cells. *J. Cell Biol.* **114,** 319–327 (1991).

42. Ohri, C. M., Shikotra, A., Green, R. H., Waller, D. A. & Bradding, P. Macrophages within NSCLC tumour islets are predominantly of a cytotoxic M1 phenotype associated with extended survival. *Eur. Respir. J.* **33,** 118–126 (2009).
43. Savage, N. D. L. *et al.* Human anti-inflammatory macrophages induce Foxp3+GITR+CD25+ regulatory T cells, which suppress via membrane-bound TGFβ-1. *J. Immunol.* **181,** 2220–2226 (2008).
44. Heusinkveld, M. & Burg, S. H. van der. Identification and manipulation of tumor associated macrophages in human cancers. *J. Transl. Med.* **9,** 216 (2011).
45. Tosolini, M. *et al.* Clinical impact of different classes of infiltrating T cytotoxic and helper cells (Th1, Th2, Treg, Th17) in patients with colorectal cancer. *Cancer Res.* **71,** 1263–1271 (2011).
46. Zarkoob, H., Taube, J. H., Singh, S. K., Mani, S. A. & Kohandel, M. Investigating the link between molecular subtypes of glioblastoma, epithelial-mesenchymal transition, and CD133 Cell Surface Protein. *PLoS ONE* **8,** e64169 (2013).
47. Weller, M. *et al.* CD95-dependent T-cell killing by glioma cells expressing CD95 ligand: More on tumor immune escape, the CD95 counterattack, and the immune privilege of the brain. *Cell. Physiol. Biochem.* **7,** 282–288 (1997).
48. Mahmoud, S. M. A. *et al.* Tumor-infiltrating CD8+ lymphocytes predict clinical outcome in Breast Cancer. *J. Clin. Oncol.* **29,** 1949–1955 (2011).
49. Zamai, L. *et al.* NK Cells and Cancer. *J. Immunol.* **178,** 4011–4016 (2007).
50. Sica, A., Schioppa, T., Mantovani, A. & Allavena, P. Tumour-associated macrophages are a distinct M2 polarised population promoting tumour progression: Potential targets of anti-cancer therapy. *Eur. J. Cancer* **42,** 717–727 (2006).

51. Bonde, A.-K., Tischler, V., Kumar, S., Soltermann, A. & Schwendener, R. A. Intratumoral macrophages contribute to epithelial-mesenchymal transition in solid tumors. *BMC Cancer* **12,** 35 (2012).
52. Kudo-Saito, C., Shirako, H., Takeuchi, T. & Kawakami, Y. Cancer metastasis is accelerated through immunosuppression during Snail-induced EMT of cancer cells. *Cancer Cell* **15,** 195–206 (2009).
53. Liu, C.-Y. *et al.* M2-polarized tumor-associated macrophages promoted epithelial–mesenchymal transition in pancreatic cancer cells, partially through TLR4/IL-10 signaling pathway. *Lab. Invest.* **93,** 844–854 (2013).

Supplemental Table 1: Directionality and confidence of invasive breast cancer data,

Whole Data Set				Cancer Only			
Parent	**Child**	**P-value**	**Relationship**	**Parent**	**Child**	**P-value**	**Relationship**
pM1	pM2	0	-	pM1	pM2	0	-
MPhi	Tcell	8.20E-104	+	MPhi	Tcell	6.75E-103	+
Tcell	NK	3.65E-46	+	Tcell	NK	1.57E-45	+
pTH1	pTh2	1.86E-42	-	pTH1	pTh2	3.67E-38	-
pTh2	pTh17	1.67E-36	+	pTh2	pTh17	1.04E-31	+
pTreg	pTh2	6.26E-35	-	pTreg	pTh2	3.02E-32	-
pM1	Tcell	1.32E-29	+	pM1	Tcell	7.33E-20	+
pM2	Tcell	1.32E-29	-	pM2	Tcell	7.33E-20	-
EMT	MPhi	5.43E-22	+	MPhi	EMT	1.55E-23	+
CD4	CD8	3.31E-20	+				
Proliferation	pTH1	3.14E-17	+	pTH1	Proliferation	1.00E-20	+
Proliferation	pM2	7.38E-16	-				-
pM1	Proliferation	7.00E-15	+				
pTreg	pTh17	5.08E-13	-	pTreg	pTh17	4.06E-12	-
Tcell	CD4	2.16E-06	+				
pTH1	NK	3.21E-06	-	pTH1	NK	2.45E-08	-
EMT	pM2	5.07E-05	+	pM2	EMT	4.08E-14	+
pTH1	pTh17	6.43E-05	-	pTH1	pTh17	5.25E-05	-
EMT	pM1	0.000585	-	pM1	EMT	4.08E-14	-
				Tcell	CD8	0.003424	+

Directionality and confidence of connection as generated from 100,000 bootstrap resamplings of the calculated metagene values for breast cancer. The left graph represents the whole dataset, and the right graph represen3ts the cancer samples alone. Relationship signifies whether the two had a positive or negative correlation.

Supplemental Table 2: Directionality and confidence of colon adenocarcinoma

Whole Data Set				Cancer Only			
Parent	**Child**	**P-value**	**Relationship**	**Parent**	**Child**	**P-value**	**Relationship**
pM1	pM2	0	-	pM1	pM2	0	-
Tcell	MPhi	5.13E-44	+	Tcell	MPhi	6.41E-38	+
NK	Tcell	9.10E-30	+	NK	Tcell	1.80E-23	+
CD8	CD4	1.39E-15	+	CD4	CD8	2.17E-26	+
pTH1	pTh17	2.36E-15	-	pTh17	pTH1	1.60E-15	-
pTreg	pTh17	9.68E-15	-	pTreg	pTh17	1.89E-05	-
pTH1	pTh2	5.74E-13	-	pTh2	pTH1	9.55E-14	-
pTreg	pTh2	8.13E-13	-	pTh2	pTreg	0.000173	-
pTh17	NK	1.84E-10	-				
pM1	Proliferation	2.35E-05	+				
Proliferation	pM2	2.48E-05	-				
pM1	MPhi	4.78E-05	-				
pM2	MPhi	4.78E-05	+				
				pTh17	Proliferation	0.000149	-
				EMT	pM1	0.000238	-
				EMT	pM2	0.00025	+

Directionality and confidence of connection as generated from 100,000 bootstrap resamplings of the calculated metagene values for colon adenocarcinoma. The left graph represents the whole dataset, and the right graph represents the cancer samples alone. Relationship signifies whether the two had a positive or negative correlation.

Supplemental Table 3: Directionality and confidence of lung cancer

Parent	Child	P-value	Relationship
pM1	pM2	0	-
Tcell	MPhi	4.63E-38	+
NK	Tcell	2.59E-27	+
pTh17	pTreg	3.11E-18	-
pTh2	pTreg	5.06E-17	-
EMT	Tcell	9.89E-13	+
pTh2	pTH1	9.23E-06	-
pTh17	pTH1	0.000663	-
EMT	pM1	0.001161	-
EMT	pM2	0.001205	+
pTH1	Proliferation	0.00871	-
NK	pTH1	0.017097	+

Directionality and confidence of connection as generated from 100,000 bootstrap resamplings of the calculated metagene values for lung cancer. Relationship signifies whether the two had a positive or negative correlation.

Supplemental Table 4: Directionality and confidence of glioblastoma multiform

Whole Data Set				Cancer Only			
Parent	**Child**	**P-value**	**Relationship**	**Parent**	**Child**	**P-value**	**Relationship**
pM1	pM2	0	-	pM1	pM2	0	-
pTH1	pTh17	3.00E-61	-	pTH1	pTh17	3.04E-59	-
pTreg	pTh17	3.16E-47	-	pTreg	pTh17	7.60E-46	-
Tcell	EMT	3.39E-43	+	Tcell	EMT	9.73E-49	+
Tcell	MPhi	3.11E-33	+	Tcell	MPhi	2.51E-34	+
MPhi	pTH1	6.01E-29	+	MPhi	pTH1	4.72E-27	+
pTh17	pTh2	1.76E-28	+	pTh17	pTh2	4.03E-29	+
pTH1	pTreg	2.00E-27	+	pTH1	pTreg	2.42E-26	+
Tcell	pTreg	5.65E-25	+	Tcell	pTreg	6.74E-24	+
pM1	MPhi	7.37E-22	-	pM1	MPhi	1.32E-23	-
pM2	MPhi	7.37E-22	+	pM2	MPhi	1.32E-23	+
pTreg	pTh2	3.92E-17	-	pTreg	pTh2	5.65E-16	-
MPhi	Proliferation	2.42E-13	-	MPhi	Proliferation	2.15E-12	-
pTH1	pTh2	2.81E-13	-	pTH1	pTh2	1.89E-12	-
pTh17	Proliferation	3.05E-13	-				
NK	EMT	4.45E-12	-	NK	EMT	2.34E-12	-
Proliferation	NK	9.80E-06	+	Proliferation	NK	5.24E-07	+
pM1	Proliferation	0.000487	+	pM2	Proliferation	0.000109	+
pM2	Proliferation	0.000487	-	pM1	Proliferation	0.000109	-
				pM2	EMT	5.89E-09	+
				pM1	EMT	5.89E-09	-
				pTreg	CD4	0.022077	+

Directionality and confidence of connection as generated from 100,000 bootstrap resamplings of the calculated metagene values for colon adenocarcinoma. The left graph represents the whole dataset, and the right graph represents the cancer samples alone. Relationship signifies whether the two had a positive or negative correlation.

Supplemental Table 5: Directionality and confidence of invasive breast cancer data, 75% of dataset

Whole Data Set				Cancer Only			
Parent	**Child**	**P value**	**Relationship**	**Parent**	**Child**	**P value**	**Relationship**
pM1	pM2	0	-	pM1	pM2	0	-
MPhi	Tcell	2.03E-81	+	MPhi	Tcell	1.17E-89	+
Tcell	NK	3.82E-56	+	Tcell	NK	3.85E-61	+
pTH1	pTh17	8.90E-30	-	pTH1	pTh17	2.68E-52	-
pTh17	pTh2	2.44E-28	+	pTh17	pTh2	1.04E-28	+
pM2	Tcell	3.53E-25	-	pM2	Tcell	1.15E-17	-
pM1	Tcell	3.53E-25	+	pM1	Tcell	1.15E-17	+
CD4	CD8	4.31E-19	+				
pTreg	pTh2	1.54E-14	-	pTreg	pTh2	9.38E-13	-
pTh17	pTreg	2.28E-14	-	pTreg	pTh17	4.10E-41	-
EMT	MPhi	3.01E-14	+	MPhi	EMT	7.95E-23	+
Proliferation	pTH1	1.62E-13	+	pTH1	Proliferation	1.93E-17	+
Proliferation	pM1	4.07E-12	+				
Proliferation	pM2	4.30E-12	-				
pTH1	pTh2	3.83E-09	-	pTH1	pTh2	1.89E-10	-
Tcell	CD4	1.81E-05	+				
EMT	pM1	0.00047	-	pM1	EMT	9.94E-12	-
EMT	pM2	0.000477	+	pM2	EMT	9.94E-12	+
				CD8	Proliferation	0.001114	+
				Tcell	CD8	0.004646	+

Directionality and confidence of connection as generated from 100,000 bootstrap resamplings of the calculated metagene values for 75% of the invasive breast cancer dataset. The left graph represents the whole dataset, and the right graph represents the cancer samples alone. Relationship signifies whether the two had a positive or negative correlation.

Supplemental Table 6: Directionality and confidence of invasive breast cancer data, 50% of dataset

Whole Data Set				Cancer Only			
Parent	**Child**	**P value**	**Relationship**	**Parent**	**Child**	**P value**	**Relationship**
pM1	pM2	0	-	pM1	pM2	0	-
CD8	CD4	0	+				
MPhi	Tcell	1.62E-56	+	MPhi	Tcell	8.70E-60	+
Tcell	NK	9.20E-43	+	Tcell	NK	1.35E-23	+
pTh17	pTreg	8.76E-14	-	pTh17	pTreg	1.74E-17	-
EMT	MPhi	9.60E-13	+	MPhi	EMT	1.91E-19	+
pTH1	pTh2	2.52E-11	-	pTh2	pTH1	1.02E-25	-
pTH1	Proliferation	2.86E-11	+				
Proliferation	pM1	1.11E-10	+				
Proliferation	pM2	1.19E-10	-				
pTh2	pTreg	1.32E-10	-	pTh2	pTreg	3.81E-14	-
pM1	Tcell	5.48E-10	+	pM1	Tcell	9.34E-12	+
pM2	Tcell	5.48E-10	-	pM2	Tcell	9.34E-12	-
Tcell	CD4	3.80E-07	+				
EMT	pM1	0.001047	-	pM1	EMT	4.95E-08	-
EMT	pM2	0.001068	+	pM2	EMT	4.95E-08	+
pTh2	pTh17	0.007238	-				
				pTh17	pTH1	4.71E-34	-
				pTH1	Proliferation	1.46E-10	+
				pTH1	NK	2.80E-05	+
				CD8	Proliferation	0.000486	+
				Tcell	CD8	0.032026	+

Directionality and confidence of connection as generated from 100,000 bootstrap resamplings of the calculated metagene values for 50% of the invasive breast cancer dataset. The left graph represents the whole dataset, and the right graph represents the cancer samples alone. Relationship signifies whether the two had a positive or negative correlation.

Supplemental Table 7: Directionality and confidence of invasive breast cancer data, 25% of dataset

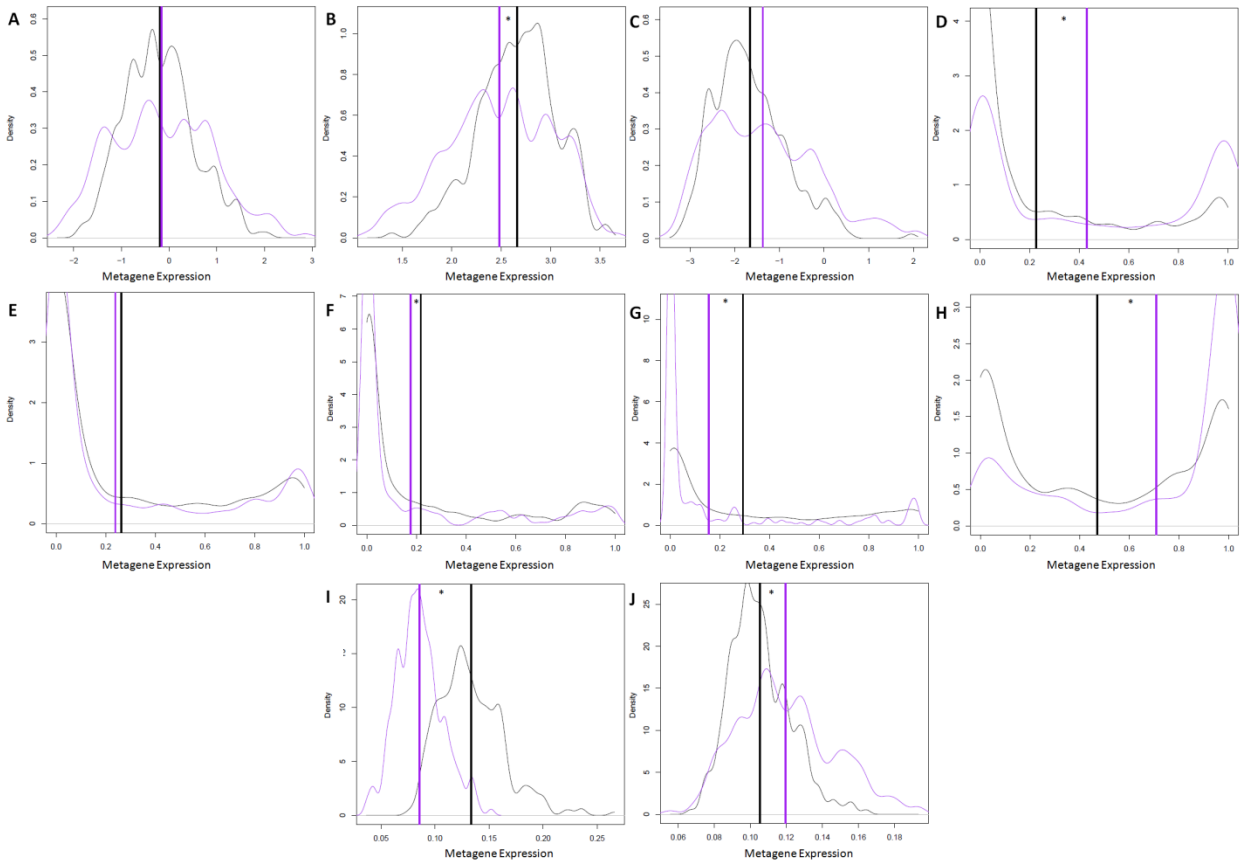
Whole Data Set				Cancer Only			
Parent	**Child**	**P value**	**Relationship**	**Parent**	**Child**	**P value**	**Relationship**
pM1	pM2	0	-	pM1	pM2	0	-
MPhi	Tcell	2.31E-26	+	MPhi	Tcell	1.96E-26	+
Tcell	NK	3.31E-19	+	Tcell	NK	3.39E-20	+
pTh17	pTH1	5.80E-14	-	pTH1	pTh17	7.06E-09	-
pTh2	pTH1	6.94E-13	-				
pTh2	pTreg	2.70E-09	-	pTh2	pTreg	1.09E-08	-
pTh17	pTreg	6.65E-09	-	pTh17	pTreg	7.28E-10	-
MPhi	EMT	1.01E-08	+				
pM2	Tcell	2.44E-08	-				
pM1	Tcell	2.44E-08	+				
pTH1	Proliferation	2.80E-05	+	pTH1	Proliferation	2.36E-06	+
Proliferation	pM1	0.000235	+				
Proliferation	pM2	0.000247	-				
Tcell	CD8	0.000594	+				
Tcell	CD4	0.011028	+				
				CD8	CD4	1.42E-30	+
				EMT	MPhi	2.23E-05	+
				pM1	EMT	0.000132	-
				pM2	EMT	0.000132	+

Directionality and confidence of connection as generated from 100,000 bootstrap resamplings of the calculated metagene values for 25% of the invasive breast cancer dataset. The left graph represents the whole dataset, and the right graph represents the cancer samples alone. Relationship signifies whether the two had a positive or negative correlation.

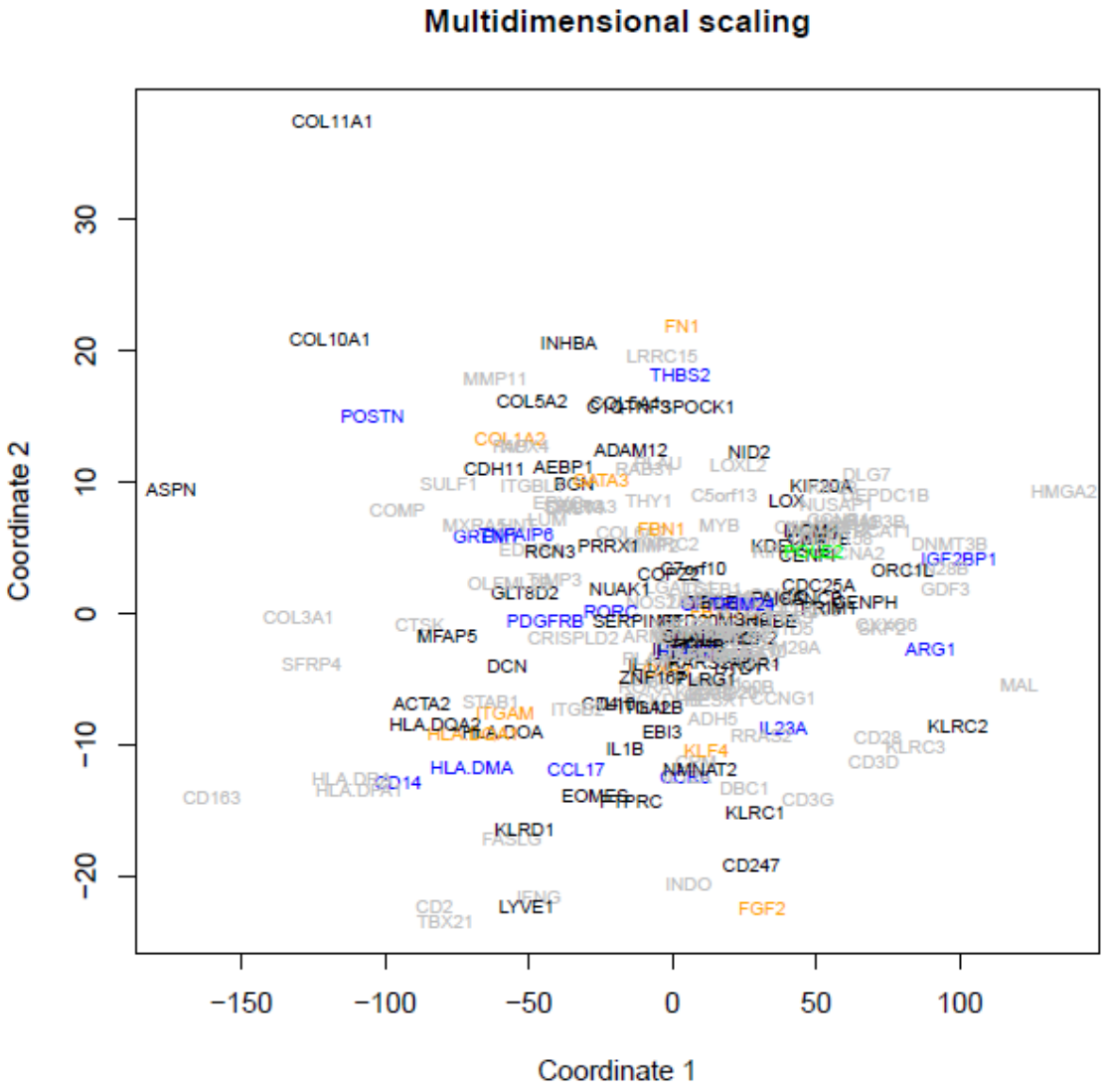
Supplemental Table 8: Breast Cancer Summary Data

	Whole Dataset	Group 1	Group 2
Gender	F = 572, M = 6	F = 304, M = 4	F = 219, M = 2
Age at Diagnosis	57.88	57.57	58.34
Metastasis Code	M0 = 374, M1 = 10, MX = 10, Null = 139	M0 = 224, M1 = 6, MX = 6, Null = 70	M0 = 148, M1 = 4, MX = 4, Null=69
Cancer Stage I	N = 42, A = 23, B = 2	N = 26, A = 13, B = 1	N = 15, A = 10, B = 1
Cancer Stage II	N = 0, A = 136, B = 84	N = 0, A = 76, B = 50	N = 0, A = 59, B = 34
Cancer Stage III	N = 0, A = 60, B = 12, C = 14	N = 0, A = 42, B =5, C = 8	N = 0, A = 18, B = 5, C = 6
Cancer Stage IV	8	6	2
Cancer Stage X	11	9	2

Table represents all data with regards to gender distribution, age at diagnosis, metastasis code and cancer stage for whole invasive breast cancer dataset as well as cancer group 1 and group 2 available when data was initially downloaded.



Supplemental Figure 1: The cancer groups differ in their expression of the EMT metagene, proliferation metagene, and macrophage polarizations. The distribution of the T cell metagene (a), the macrophage metagene (b), the NK metagene (c), pTh1 metagene (d), pTh17 metagene (e), pTh2 metagene (f), the pTreg metagene (g), the pM1 macrophage metagene (h), the EMT metagene (i), and the proliferation metagene (j) were generated and separated by grouping, with black representing cancer group 1 and yellow representing cancer group 2. * represents a p-value of less than .01.



Supplemental Figure 2: Multidimensional Scaling as principal coordinate analysis was preformed. Genes were color coded based on metagene grouping, EMT (black), Proliferation (grey), T cell infiltration (green), NK cell infiltration (red), Macrophage Infiltration (orange), T cell polarization (yellow), and Macrophage polarization (blue). Scaling was performed using the cmdscale() function.