Causal networks of immune cells defined by multivariate gene expression signatures

Ali S. Bakhtiari PhD[1,2#], Kristin Wallace PhD[2], David N. Lewin[3], Shaoli Sun[3], Jennifer D. Wu[4*], Alexander V. Alekseyenko PhD[1,2,6]

[1]Biomedical Informatics Center, [2]Department of Public Health Sciences, [3]Department of Pathology and Laboratory Medicine, [4]Department of Microbiology and Immunology, College of Medicine, [4]Department of Oral Health Sciences, College of Dental Medicine, Medical University of South Carolina, Charleston, South Carolina, [*]Now at Feinberg School of Medicine, Northwestern University, Chicago, [#]Corresponding author bakhtiaa@musc.edu

Abstract

An understanding of immune cell interactions in the tumor microenvironment is necessary to gain insight into cancer pathogenesis and immunotherapy mechanisms. Immune gene expression from cancer tissues can serve as a viable proxy for direct measurement of immune cells. Using gene expression, we define multivariate immune cell type signatures in an essentially unbiased and non-parametric way, based on literature-driven gene lists. We study the interactions between different immune cell subtypes in colorectal cancer based on these signatures using multivariate distance correlation approach. We infer a relevance network of these genes and refine it into a draft causal network using the Fast Causal Inference algorithm. The methodological value of our approach is in demonstrating the ability to draw causal inference between multivariate vectors. The association study reveals a complex network of interactions between different immune cell types, plausibly explainable by current immunology literature. The causal analysis between various cell types reveals a causal relationship between natural killer and follicular helper T cells. The same analysis also reveals internal causal interactions between various T cells.

Introduction

The human body has evolved various immune measures for suppressing cancer cells [1]. The understanding of this process has led to novel cancer immunotherapy treatment approaches [2]. In essence, cancer cells take two counter strategies against host immune cells [3]: (i) they either avoid recognition by the immune system or (ii) manage to suppress the immune activity within their microenvironment. In virtually every tumor microenvironment, there is a network of active immune cells [4].

To understand the interaction of immune cells in the tumor microenvironment, it is necessary to identify the population of immune subsets with known functions. The traditional pathology approach involves a direct count of immunohistochemically stained slides, a laborious and error prone process that results in low throughput. We propose that immune cell type assays based on gene expression signatures is a desirable alternative method to speed up discovery. The NanoString immunology panel [5] is capable of identifying the required information. The NanoString panel digitally queries the expression of specific genes known to be involved in cancer pathways[6].

A gene expression-based immune cell analysis involves drawing inferences between immune cell signatures defined as lists of genes whose expression is characteristic of the given immune cells. Thus, the identification of the relationships between the immune cells relies on our ability to draw association between multivariate vectors corresponding to expression of genes comprising the signatures. This presents several methodological challenges: (i) a general framework is needed to establish multivariate associations and (ii) the network effects we would like to uncover are seldom linear, thus requiring nonlinear methods, because the later are prone to overfitting. Distance correlation provides a viable solution that resolves both of these challenges [7]. Distance correlations are strongly robust toward the nature of association between two multivariate vectors, as described further in our methods.

In this paper, we use gene expression profiling to draw inferences about immune cell interactions in colorectal cancer (CRC). Using distance correlation approach, we draw a relevance network of these immune cells and further refine it to identify the potential causal relationships, using the Fast Causal Inference (FCI) algorithm on the computed distance correlations between the multivariately defined immune cell signatures. This approach is generalizable to other multivariate features from high-throughput assays and from unstructured biomedical data.

Materials and Methods

Study Data

Nineteen patients with CRC were examined (9 Caucasians, 10 African Americans; 11 males, 8 females; mean age 65.42±10.36; 7 distal, 12 proximal). Each patient's NanoString data consists of 579 gene expressions.

Signatures of immune cell types

The existing literature identifies gene expression markers of tumor infiltrating leukocytes [6, 8]. Based on those findings, we gathered a set of 57 genes known to be related to distinguishable immune cell types. In total, 11 different cell types are associated with the latter subset. Each immune cell type is defined as a vector of gene expressions of known genes (Table 1).

Table 1. Cell types and their respective known genes considered in our analysis.

Cell Type	Signature Gene
Cytotoxic	CD8A,CD8B,PRF1
NK	E4BP4,IL2,IL12,IL15,IL18,GranzymeB,IL17a,IL22,CCL3,CCL4,MIP1B,CCL5,perforin
Th1	TBX21,IL2,IL12B,IL12A,IL27,IFNG,TNF,IL2,IL12B,IL12A,STAT1,CSF2,IFNG,LTA,STAT4
Th17	STAT3,RORC,IL1B,TGFB1,IL6,IL17A,IL17F,IL22,IL23A,CSF1,CSF2,TNF
Treg	FOXP3,IL2,TGFB1,IL10
B	GATA3,CD20,TLR9,CD19,CR2,MS4A1,TNFRSF17
Innate	TLR2,TLR4,TLR9,CD33
Tcm	ATM
Tem	CCR2
Tfh	CXCL13,MAF,PDCD1,BCL6
Th2	CXCR6,GATA3,IL26,STAT6

Data Handling and Preparation

We normalized the NanoString data using nSolver Analysis Software Version 3.0 with geometric mean for positive control and code set content normalization (by housekeeping genes). We loaded the resulting values into R package phyloseq [9] to streamline distance calculations and sample data handling. Euclidean distances were calculated and used for distance measurements. We used R package energy [10] to calculate the distance correlations between the cell type subsets and to determine their statistical significance at alpha level 0.05, using the distance correlation t-test. We used the R package pcalg [11] for causal analysis of the distance correlations, using the fast causal inference (FCI) algorithm [12].

Multivariate Distance Correlation

We use multivariate distance correlation (MDC) [7] to infer the association between different cell types within our subject samples. Let X and Y be observations of l through m variables respectively, measured in n subjects. Let X_j denote the observations in subject j. We define distance covariance between X and Y as:

$$\mathrm{dCov}_n^2(X, Y) := \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} A_{j,k} B_{j,k}.$$

Where

$$A_{j,k} := a_{j,k} - \overline{a}_{j\cdot} - \overline{a}_{\cdot k} + \overline{a}_{\cdot\cdot}, B_{j,k} := b_{j,k} - \overline{b}_{j\cdot} - \overline{b}_{\cdot k} + \overline{b}_{\cdot\cdot},$$

In above $\overline{a}_{j\cdot}$ is the j-th row mean and $\overline{a}_{\cdot\cdot}$ is the grand mean and

$$a_{j,k} \quad = \| X_j - X_k \|, j, k = 1,2, \dots, n,$$
$$b_{j,k} \quad = \| Y_j - Y_k \|, j, k = 1,2, \dots, n,$$

Finally, we define distance correlation as

$$\mathrm{dCor}(X, Y) = \frac{\mathrm{dCov}(X, Y)}{\sqrt{\mathrm{dVar}(X)\, \mathrm{dVar}(Y)}}.$$

In the equation above $\mathrm{dVar}(X) = abs(\, \mathrm{dCov}(X, X))$.

Note that A and B are doubly-centered Euclidean distance matrices. Other distances can be applicable and individual applications may warrant a benchmark and some prior intuition to determine the best metric to use. When the causal relationships are captured by multivariate profiles the mutual information-based techniques become confusing and computationally cumbersome. Distance correlation is a value between 0 and 1, with 0 showing independence and 1 showing almost surely one to one linear relationship between X and Y.

Fast Causal Inference

The Fast Causal Inference (FCI) algorithm allows for estimation of causal relationships between immune cell types. The algorithm steps are fully described in [4]. The FCI method uses partial correlations for testing the linear independence of variables X and Y. Partial correlation between X and Y conditioned on Z is defined as follows:

$$\hat{\rho}_{XY\cdot\mathbf{Z}} = \frac{N \sum_{i=1}^{N} r_{X,i} r_{Y,i} - \sum_{i=1}^{N} r_{X,i} \sum_{i=1}^{N} r_{Y,i}}{\sqrt{N \sum_{i=1}^{N} r_{X,i}^2 - (\sum_{i=1}^{N} r_{X,i})^2} \sqrt{N \sum_{i=1}^{N} r_{\square,i}^2 - (\sum_{i=1}^{N} r_{Y,i})^2}}.$$

Where $r_{X,i} = x_i - \langle \mathbf{w}_X^*, \mathbf{z}_i \rangle$, which is the residual of linear regression of X on Z.

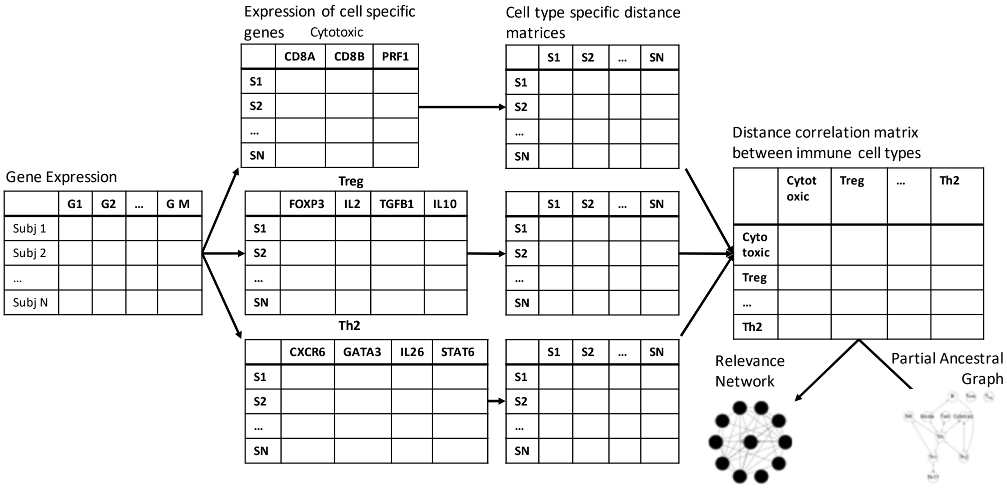
In our analysis, we used the test of independence by Fisher's Z-transformation. We set the significance threshold at 0.05.

The FCI algorithm results in a network where the edges follow the description of partial ancestral graphs (PAG) [13]. The following nomenclature is used in PAG.

- *A o-o B indicates non d-separated nodes*
- *A->B indicates that A is a cause of B*
- *A o->B edges indicate that B is not an ancestor of A*
- *A<->B edges indicate that there is a latent common cause between A and B*

Figure 1 depicts the step by step algorithm for inferring the causal relationships.

Figure 1. Data transformation steps in multivariate distance correlation based causal inference



Results

The expression based immune cell signatures corroborate pathology observation.

It is known from pathology that stroma and deep stroma observations are associated with cytotoxic and Treg immune cell presence [14, 15]. A trained pathologist counted the CD8 and FOXp3 stained cell concentrations per high power field in our samples. Expression based signatures for cytotoxic and Treg cells are expected to have significant, but not necessarily perfect correlations with the counts of CD8+ and FOXp3 stained cells respectively. We have calculated MDC between the pathology counts and expression signatures, as well as individual genes comprising the signatures (Table 2). Our results indicate that the expression-based immune cell signatures are significantly correlated with pathology counts. As expected the genes corresponding to histological markers have strong correlations with the counts as well.

Table 2. The MDC between cytotoxic and Treg cell signatures and pathology counts in stroma and deep stroma.

Pathology CD8+ cell counts per high power field		
	Deep Stroma	**Stroma**
Cytotoxic cells signature	**0.25**	**0.24**
CD8A	**0.339**	**0.198**
CD8B	-0.03	**0.24**
PRF1	**0.31**	**0.19**
Pathology FOXp3 cell counts per high power field		
	Deep Stroma	**Stroma**
Treg cells signature	**0.21**	**0.23**
FOXP3	**0.4**	**0.43**
IL2	0.16	**0.5**
TGFB1	-0.05	0.06
IL10	0.03	-0.02
Significant MDCs (P<0.05) are in bold.		

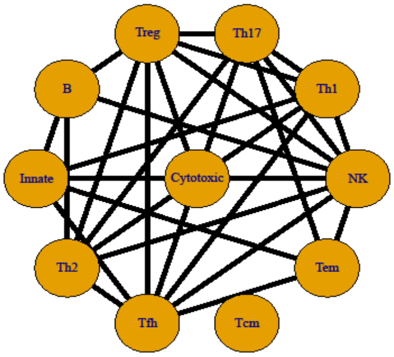


Figure 2. Association Network between different immune cell types.

The immune cell signatures demonstrate dense relevance network relationships

We observed a complex network of association between different cell types, Figure 2. The observed MDC values are shown in Table 3. Most of the relationships observed in Figure 2 have known scientific basis.

For instance, it is known that T memory cells (Tcm) act independently from other cell types. The distance correlation clearly reaffirms the independence.

FCI reveals causal relationship between NK and Tfh cells

FCI analysis concluded that most of the observed associations are non-causal; however, the relationship between NK and Tfh cells is inferred as causal in nature. We note here that the graph equivalency of the directed acyclic graphs (DAG) used in the FCI algorithm is theoretically incapable of identifying cause and effect in paired nodes.

Discussion

To our knowledge, this paper is the first to describe fast causal inference on multivariate vectors using distance correlation approach as the means of establishing correlations between the vectors. The significance of this is in the enhanced ability to causally describe multivariate phenomena without the need to explicitly measure or define univariate proxies for them. Distance correlations can be used in lieu of linear correlations to capture multivariate correlations, which can form the basis for computational causal inference.

Discovering the nature of interactions between immune cell types is an important step in discovering the mechanisms that those cells employ for targeting cancer cells, and understanding the tumor microenvironment biology and cancer pathogenesis. In this work, we showed that combining prior information about the genetic signatures of each of the immune cell types and potential interactions with advanced learning techniques has the potential of revealing the aforementioned interactions.

Table 3. Multivariate distance correlations of immune cell types with significant edges (P< 0.05).

	B	*Cytotoxic*	*Innate*	*NK*	*Tem*	*Tfh*	*Th1*	*Th17*	*Th2*
Innate	0.39	*N.E.*							
NK	0.27	0.32	*N.E.*						
Tem	*N.E.*	*N.E.*	0.17	0.26					
Tfh	*N.E.*	0.43	0.31	0.46	0.38				
Th1	*N.E.*	0.25	0.16	0.41	0.12	0.40			
Th17	*N.E.*	*N.E.*	*N.E.*	0.30	*N.E.*	0.26	0.35		
Th2	0.30	0.38	*N.E.*	0.26	*N.E.*	0.39	0.32	0.22	
Treg	0.21	0.16	*N.E.*	0.28	*N.E.*	0.30	0.20	0.22	0.28

N.E. - no edge

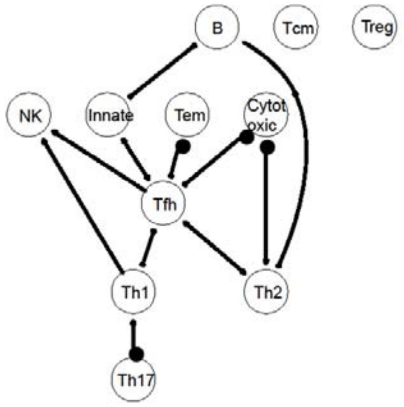


Figure 3. Hypothetical network revealed by artificial 10-fold increase of the sample size.

The limited quantity of the available data is a limiting factor in utilizing the full potential of the causal techniques used in this study. In order to demonstrate the effect of sample size, we propose the following hypothetical scenario, in which the same MDC is observed in a dataset 10 times the current actual size.

The choice of 10 fold inflation of the dataset size is in accordance with the eventual number of data points that will be available to us. The causal network inferred by the FCI algorithm under these assumptions reveals the hypothetical network, which is much denser and potentially contains more mechanistic knowledge, Figure 3. Speculatively some of the revealed inferences from the hypothetical network have known scientific basis. Based on our expertise, we saw no contradiction in the revealed network with what is corroborated by the state of the art in immunology, with the exception of NK cells relationships. NK cells are known to influence Th1 differentiation, but not Tfh [16]. This network, however, shows the inverse relationships exists between Th1 and NK and that there is a causal link between Tfh and NK cells. These hypothetical inferences may represent aspects of new biology, which may be confirmed or refuted once a larger sample size will be available to for actual analysis.

We used the linear partial correlation assumption in our FCI analysis. It has been shown in [7] that this assumption leads to decreased power. We intend to develop the FCI model to work with distance partial correlations so as to harness the full potential of the algorithm.

Acknowledgements

This work was supported by College of Medicine Enhancing Team Science (COMETS) grant from MUSC College of Medicine, AVA was supported in part by the NIH R34DE025085, R01CA164964, AVA and ASB were supported in part by the NIH R21AR067459, JW is supported by R01CA204021 and R01CA208246.

References

1. Cavallo, F., et al., *2011: the immune hallmarks of cancer.* Cancer Immunol Immunother, 2011. **60**(3): p. 319-26.
2. Couzin-Frankel, J., *Breakthrough of the year 2013. Cancer immunotherapy.* Science, 2013. **342**(6165): p. 1432-3.
3. Gajewski, T.F., H. Schreiber, and Y.X. Fu, *Innate and adaptive immune cells in the tumor microenvironment.* Nat Immunol, 2013. **14**(10): p. 1014-22.
4. Grivennikov, S.I., F.R. Greten, and M. Karin, *Immunity, inflammation, and cancer.* Cell, 2010. **140**(6): p. 883-99.
5. Cesano, A., *nCounter® PanCancer Immune Profiling Panel (NanoString Technologies, Inc., Seattle, WA).* Journal for ImmunoTherapy of Cancer, 2015. **3**(1): p. 42.
6. Dennis, L., et al., *Multiplexed Cancer Immune Response Analysis*. 2015.
7. Szekely, G.J., M.L. Rizzo, and N.K. Bakirov, *Measuring and testing dependence by correlation of distances.* Ann. Statist., 2007. **35**(6): p. 2769-2794.
8. Danaher, P., et al., *Gene expression markers of Tumor Infiltrating Leukocytes.* Journal for ImmunoTherapy of Cancer, 2017. **5**(1): p. 18.
9. McMurdie, P.J. and S. Holmes, *phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data.* PLoS One, 2013. **8**(4): p. e61217.
10. Szekely, M.L.R.a.G.J., *energy: E-Statistics: Multivariate Inference via the Energy of Data.* 2016.
11. Kalisch, M., et al., *Causal Inference Using Graphical Models with the R Package pcalg.* 2012, 2012. **47**(11): p. 26.
12. Spirtes, P., C. Meek, and T. Richardson, *Causal inference in the presence of latent variables and selection bias*, in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. 1995, Morgan Kaufmann Publishers Inc.: Montr&#233;al, Qu&#233;, Canada. p. 499-506.
13. Zhang, J., *On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias.* Artificial Intelligence, 2008. **172**(16): p. 1873-1896.
14. Conti, J. and G. Thomas, *The Role of Tumour Stroma in Colorectal Cancer Invasion and Metastasis.* Cancers, 2011. **3**(2): p. 2160-2168.
15. Deschoolmeester, V., et al., *Immune Cells in Colorectal Cancer: Prognostic Relevance and Role of MSI.* Cancer Microenvironment, 2011. **4**(3): p. 377-392.
16. Cook, K.D., S.N. Waggoner, and J.K. Whitmire, *NK Cells and Their Ability to Modulate T Cells during Virus Infections.* Critical reviews in immunology, 2014. **34**(5): p. 359-388.