PNAS
www.pnas.org

Supplementary Information for

Indirect Behavioral Network Analysis Reveals An Immune Tolerance Mechanism In Cancer

James C. Mathews, Saad Nadeem, Maryam Pouryahya, Zehor Belkhatir, Joseph O. Deasy, Arnold Levine, and Allen R. Tannenbaum

Arnold Levine
Email: alevine@ias.edu

This PDF file includes:

Supplementary text
Figures S1 to S2
Tables S1 to S3
SI References

www.pnas.org/cgi/doi/10.1073/pnas.2002179117

Supplementary Information Text

Illustrative GMT analyses. Figure S1 shows the results of the GMT analysis on the PANTHER curated gene network. Figure S2 shows the results on the GTEx lung and breast tissue RNA expression datasets. The GTEx analyses take advantage of the full generality of the method by involving non-synthesized node data measurements. In these examples the network topology was inferred from the sample data with a Pearson correlation cutoff, but if a prior network of interest is available it can be used instead. We caution that our experience with curated networks (NetPath, PANTHER, HPRD) suggests that the choice of prior network strongly influences the results, the greater the sparsity the stronger the influence. So analyses involving empirical node weights and a prior network should always be compared with a control analysis based on synthetic node weights as in Figure S1.

Gene Ontology (GO) annotations may be used to explain gene clusters. The genes involved in a cluster come from an unbiased process of gene assembly and the mathematics that is being used by this approach. In fact the GO terms used belong to a separate structural hierarchy tree that goes from a general to a specific description of functions. The annotations that are used to identify a gene in a hierarchy are simply too numerous to be useful. Because of this only those identifiers of gene sets that are well clustered along the hierarchy, based upon the average pairwise graph distance between nodes in the hierarchy, are used. The annotations are assessed for statistical significance as follows. For each tagged subset of genes (i.e. each annotation term), the mean of the node-to-node pairwise weighted graph distances is calculated and regarded as a dispersal or coordination statistic. The statistical significance is measured with 10000-trial bootstrapping by random permutation of the whole gene set, with a p-value recording the fraction of the trials in which the statistic was lower than the observed value. Figure S1 and S2 show the application of this method to the PANTHER curated database of gene interactions and the GTEx gene expression profiles of normal lung and breast tissue.

Algorithms. Tables S1, S2, and S3 show the details of the Gaussian Mixture Transport calculations in algorithmic form.

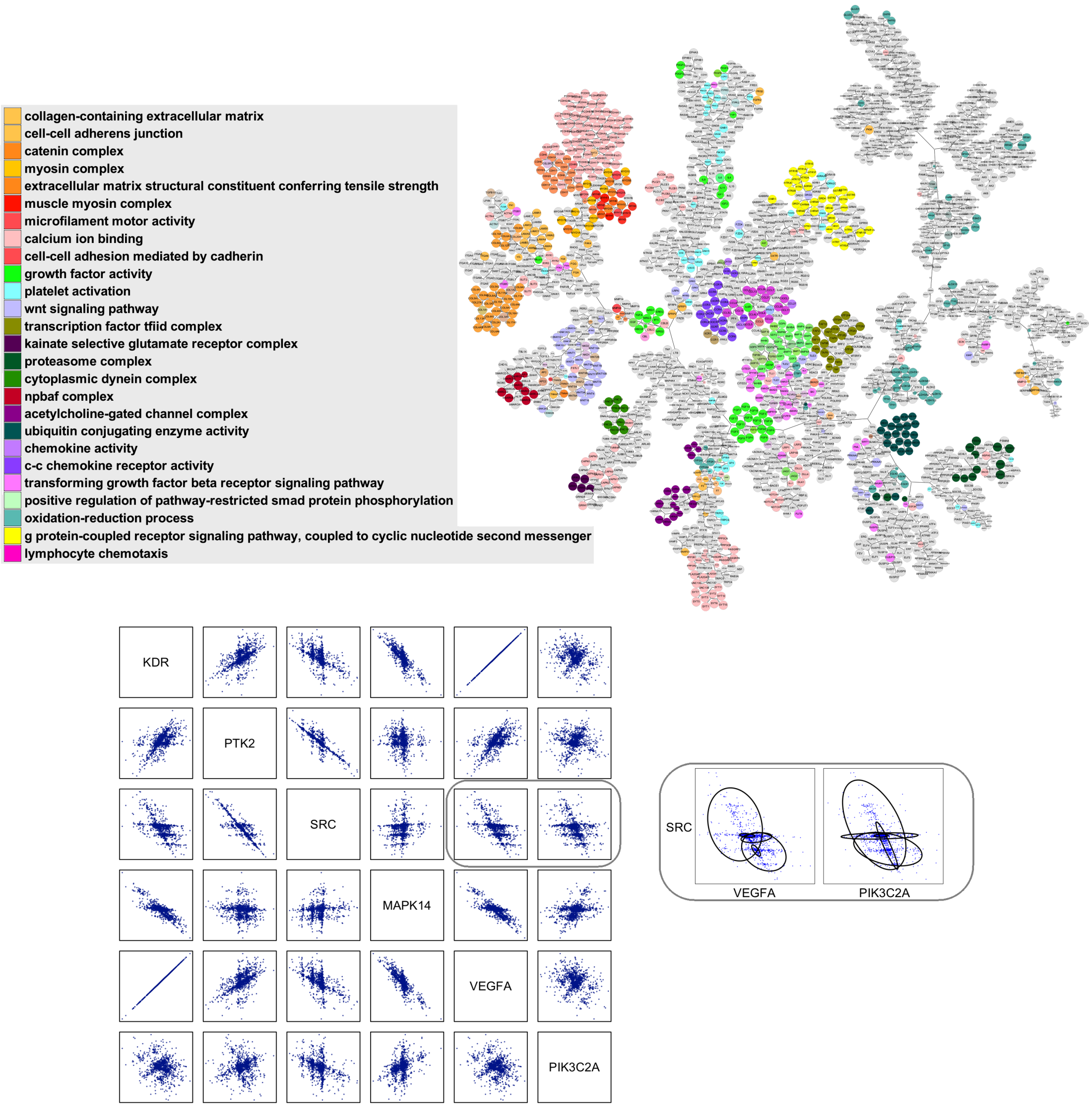

Fig. S1. (Above) The GMT hierarchy computed from the PANTHER curated gene network with 2404 nodes and 32113 edges. Approximately 40% of the network is accounted for by well-defined processes. The gene modules appearing here are showing evidence of coordination already at the level of the network topology, without any influence from empirical node weights. This plot and GO term list could be used as a control for a separate data-driven analysis, with new annotations considered significant only when sufficiently distinct from the annotations on this list. (Below) Scatter plots of the synthesized data for selected nodes/genes, for illustration. The column containing a given gene name represents the 'behavioral role profile' of the gene with respect to the other genes. The behavioral role profiles of highly correlated genes are very similar (e.g. the KDR and VEGFA columns). However, similar behavioral roles can be observed even for uncorrelated genes. This phenomenon is the key difference between the GMT hierarchy and a standard correlation-based clustering. For example, the behaviors of VEGFA and PIK3C2A with respect to SRC are similar, even though the scatter plot VEGFA-PIK3C2A shows a high degree of independence. Namely, both stand in a relation of 'inhibition' to SRC. At a low level of our hierarchy, VEGFA and PIK3CA and perhaps other SRC inhibitors remain separate. At a middle level of our hierarchy, VEGFA and PIK3CA are closer together. In general mid level clusters may represent clusters of function, while lower levels stratify each function by the specific manner in which the function is carried out.

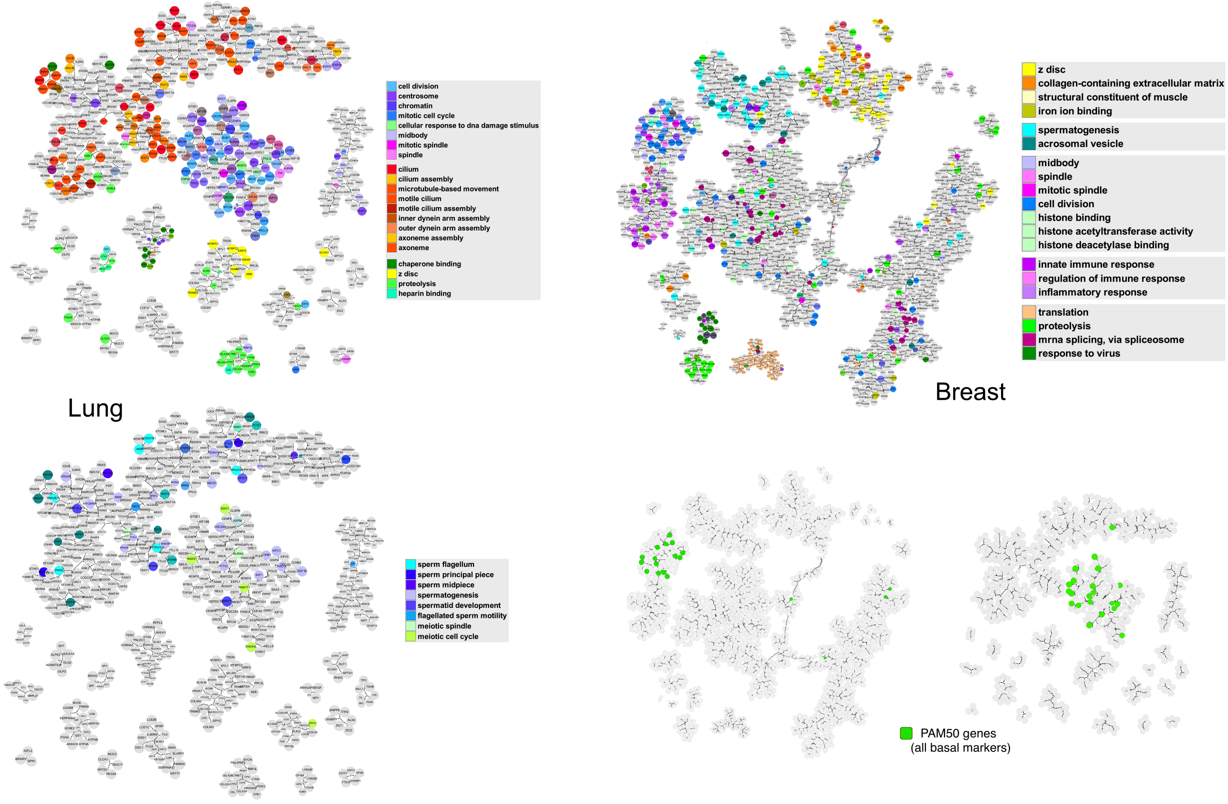

Fig. S2. The GMT hierarchy and tree-assisted Gene Set Enrichment Analysis computed from the RNA expression profiles of the GTEx lung tissue samples (left above and below) and breast tissue samples (above right). (Below right) The 18-20 genes of the 50-gene PAM50 breast-tumor prognostic signature which appear in the breast and lung analysis turn out to be exactly the basal-marker subset identified in our earlier work [2]. Thus we find that the PAM50 is representative of only one of the key gene modules for breast tissue. Also, a substantial part of the variance in breast tumors captured by the PAM50 is likely due to variance observable already in normal (non-cancer) tissue, and moreover this type of variation is observed in normal tissues other than breast. The analysis provides evidence that the basal-marker gene submodule of the PAM50 is related to the cell cycle machinery including motisis, chromatin, and the architecture of the mitotic spindle (shown in shades of blue in the two upper plots). It also suggests augmentation of these markers by other genes in the apparent module, including MCM10, HMMR, ASPM, TOP2A, POLQ, RAD54L, and AURKA. Some of these genes are known to be related to breast cancer. For example AURKA was already suggested as part of the simplified 3-gene prognostic signature for breast tumors SCMGENE [1].

Table S1. Notation for algorithms

N	number of nodes
E	number of edges
S	number of node weightings
s	source function $[1, E] \to [1, N]$
t	target function $[1, E] \to [1, N]$
w	node weight matrix $[1, N] \times [1, S] \to \mathbb{R}$
P	Gaussian mixture modeling population number
GMM	Gaussian mixture model, *mclust* R package [3]
$j \sim i$	nodes i and j are neighbors

Table S2. GMT distance node similarities algorithm

```
function GMT DISTANCES($s$,$t$,$w$,$P$)
    for $e$ in $[1, E]$ do
        Model$(e)$ =GMM$(w(s(e), -), w(t(e), -), P)$
    end for
    list = $\{\}$
    for $i$ in $[1, N]$ do
        for $j \sim i$ do
            for $k \sim i$, $k > j$ do
                M1 = Model$(e(j, i))$
                M2 = Model$(e(k, i))$
                $d$ =GMM/OMT Distance(M1,M2)
                list $\leftarrow ((j, k), d)$
            end for
        end for
    end for
    groups = group(list) by $(j, k)$ value
    similarities = $\{\}$
    for group $(j, k)$ in groups do
        similarities$(j, k)$ =average $d$ over group
    end for
    return similarities
end function
```

Table S3. GMM/OMT (GMT) Distance algorithm

```
function GMM/OMT DISTANCE(M1,M2)
    D1 = probabilities(M1)                                        ▷ Vector size $P$
    D2 = probabilities(M2)                                        ▷ Vector size $P$
    for $a$ in $[1, P]$ do
        for $b$ in $[1, P]$ do
            G1 = Gaussian model(M1, $a$)
            G2 = Gaussian model(M2, $b$)
            cost$(a, b)$ = OMT distance(M1(a),M2(b))
        end for
    end for
    return Earth Mover's Distance(D1, D2, cost)
end function
function OMT DISTANCE(G1,G2)
    $v_1$ = mean(G1)
    $v_2$ = mean(G2)
    $\Sigma_1$ = covariance matrix(G1)
    $\Sigma_2$ = covariance matrix(G2)
    $D$ = sum of squared entries of $v_1 - v_2$
    $D'$ = trace($\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}$)
    return $(D + D')^{1/2}$
end function
```

SI References

1. B. Haibe-Kains, C. Desmedt, S. Loi, A.C. Culhane, Bontempi G., J. Quackenbush, and C. Sotiriou. A three-gene model to robustly identify breast cancer molecular subtypes. *J National Cancer Inst.*, 104(4):311–325, 2012.
2. J. C. Mathews, S. Nadeem, A. J. Levine, M. Pouryahya, J. O. Deasy, and A. Tannenbaum. Robust and interpretable PAM50 reclassification exhibits survival advantage for myoepithelial and immune phenotypes. *NPJ Breast Cancer*, 5:30, 2019.
3. Luca Scrucca, Michael Fop, Thomas Brendan Murphy, and Adrian E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):205–233, 2016.