Hierarchical HotNet: identifying hierarchies of altered subnetworks Supplementary Information

Matthew A. Reyna, Mark D.M. Leiserson, Benjamin J. Raphael

The following sections elaborate on the data, methods, and results from the main manuscript.

1 Methods

1.1 Definition of similarity matrix S for Hierarchical HotNet

Given a vertex-weighted graph $G = (V, E, w)$ with n vertices and m edges, let $A = [a_{ij}]$ be an unweighted adjacency matrix with $a_{ij} = 1$ if there is an edge $(v_j, v_i) \in E$ and $a_{ij} = 0$ otherwise. If the graph is directed and/or weighted, then the adjacency matrix can be defined accordingly.

Let $D = \text{diag}(d_1, \ldots, d_n)$ be a diagonal degree matrix, where

$$d_j = \sum_{i=1}^{n} a_{ij}, \tag{1}$$

and let $W = [w_{ij}]$ be the transition matrix for the random walk, where

$$w_{ij} = \begin{cases} \frac{a_{ij}}{d_j}, & d_j \neq 0, \\ 0, & d_j = 0. \end{cases} \tag{2}$$

We define the random walk with restart by

$$s^{(k+1)} = (1 - \beta) W s^{(k)} + \beta f, \tag{3}$$

where $s^{(0)} \in \mathbb{R}^n$ is the initial distribution of walkers, $s^{(k)} \in \mathbb{R}^n$ is the distribution of walkers after k steps, $\beta \in (0, 1)$ is the restart probability for the random walk, and $f \in \mathbb{R}^n$ is the preference vector for the random walk.

If G is (strongly) connected, then it can be shown, e.g., with the Perron-Frobenius theorem, that the random walk with restart in (3) has a unique, non-trivial stationary distribution s, which is given by

$$s = Pf, \tag{4}$$

where

$$P = \beta \left(I - (1-\beta)W\right)^{-1}. \tag{5}$$

For each $v_j \in V$, we set $s^{(0)} = f = \mathbf{e}_j$, where $\mathbf{e}_j$ is the standard basis vector, and we scale (4) by the vertex score $w(v_j)$ to construct the jth column of the joint similarity matrix $S = [s_{ij}]$, which is given by

$$S = PF, \tag{6}$$

where $F = \operatorname{diag}\left(w(v_1), \ldots, w(v_n)\right)$. By expanding (5) into a geometric series, we have

$$s_{ij} = w(v_j)\left(\beta(1-\beta)\frac{a_{ij}}{d_j} + \beta(1-\beta)^2 \sum_{k=1}^{n} \frac{a_{ik}}{d_k}\frac{a_{kj}}{d_j} + \cdots\right), \tag{7}$$

which is the formula given for S in the main manuscript.

If $s^{(0)} = f = \mathbf{e}_j$, then it can be shown that the connectivity conditions for the graph are no longer necessary for the random walk with restart to converge to the above stationary distribution. Therefore, this definition of a similarity matrix S in (6) applies to any vertex-weighted graph whether or not the graph is undirected or directed, unweighted or weighted, or disconnected or connected.

1.2 Parameter selection for the restart probability β for the similarity matrix for Hierarchical HotNet

For $s^{(0)} = f = \mathbf{e}_j$, the restart probability β determines the locality of the random walk, where smaller values of β preserve more global properties of the graph and larger values of β preserve more local properties. Many papers that use the random walk with restart have found that their results are relatively insensitive to the choice of this parameter over the range of values that they considered, e.g., [6].

For our heuristic, we choose β to balance the stationary distribution (4) between the network neighborhood of a vertex and more distant vertices. Since we choose β based on network topology alone, i.e., with uniform vertex scores, the same choice of β can be used for the same network with different sets of vertex scores.

Given a graph $G = (V, E)$, let

$$\mathcal{N}(v) = \{v \in V : (u, v) \in E\} \tag{8}$$

be the first-order network neighborhood of $v \in V$, and let

$$P(\beta) = [p_{ij}(\beta)] = \beta \left(I - (I-\beta)W\right)^{-1} \tag{9}$$

be the topological similarity matrix for G with restart parameter β. We choose $\beta \in (0, 1)$ as the value of β that satisfies

$$\sum_{v_j \in V} \sum_{v_i \in \mathcal{N}(v_j)} p_{ij}(\beta) = \sum_{v_j \in V} \sum_{v_i \notin \mathcal{N}(v_j) \cup \{v_j\}} p_{ij}(\beta). \tag{10}$$

We can solve (10) with a nonlinear root finder. In practice, there exists a unique solution to (10), and a numerical root finder tends to converges quickly because the terms in (10) are smooth (see (7)) and few digits of the β are needed.

This procedure is applies to directed and edge-weighted graphs, but it requires minor modifications for complete edge-weighted graphs. In this case, we suggest replacing $\mathcal{N}(v)$ with $\mathcal{N}_\epsilon(v_i) = \{v_j \in V : a_{ij} \geq \epsilon\}$, where ϵ is the median edge weight of the complete weighted graph.

2 Data

We used the following datasets in our analysis.

2.1 Somatic mutation data

For our analysis, we used the following sets of pan-cancer somatic mutation gene scores, which were the most recent versions available as of February 22, 2018.

- MutSig q-value scores [5, 4]:
  `http://www.lagelab.org/wp-content/uploads/2017/06/NetSig_Code.zip`
- TCGA PanCanAtlas mutation frequency scores [5, 4]:
  `https://www.synapse.org/#!Synapse:syn7214402`

For the mutation frequency scores, we first restricted our analysis to nonsynonymous somatic variants (`Frame_Shift_Del`, `Frame_Shift_Ins`, `In_Frame_Del`, `In_Frame_Ins`, `Missense_Mutation`, `Nonsense_Mutation`, `Nonstop_Mutation`, `Translation_Start_Site`) by omitting synonymous variants (`3'Flank`, `3'UTR`, `5'Flank`, `5'UTR`, `IGR`, `Intron`, `lincRNA`, `RNA`, `Silent`, `Splice_Site`). We then removed samples with 400 or more mutated genes and genes with mutations in 2% samples for genes with MutSig q-values $q > 0.1$, i.e., genes that were frequently mutated but not statistically significant. We finally defined the mutation frequency of a gene as the fraction of samples with one or more mutations in the gene.

Altogether, we used MutSig q-values from 4,742 tumors across 21 tumor types and TCGA PanCanAtlas mutation frequency scores from 9,326 tumors (down from 10,206 tumors before removing highly mutated samples and genes) across 33 tumor types.

2.2 Interaction networks

For our analysis, we used the following interaction networks, which were the most recent versions available as of February 23, 2018.

- HINT+HI [2, 8]:
  - HINT binary
    `http://hint.yulab.org/download/HomoSapiens/binary/hq/`

  - HINT co-complex
    `http://hint.yulab.org/download/HomoSapiens/cocomp/hq/`
  - HuRI HI:
    `http://interactome.baderlab.org/download`

- iRefIndex 15.0 [7]:
  `http://irefindex.org/download/irefindex/data/archive/release_15.0/psi_mitab/MITAB2.6/9606.mitab.22012018.txt.zip`
- ReactomeFI 2016 [1, 3]:
  `http://reactomews.oicr.on.ca:8080/caBigR3WebApp2016/FIsInGene_022717_with_annotations.txt.zip`

For the ReactomeFI interaction network, we considered the set of interactions with a confidence score of 0.75 (out of 1) or larger. This step is not necessary for Hierarchical HotNet, which can analyze directed and weighted interaction networks, but we used this undirected and unweighted version of the network for each method for a more direct comparison of results between methods. For each network, we also restricted our analysis to the largest connected component of the network.

3 Results

Table S1 provides the 128 genes in the Hierarchical HotNet consensus results, Table S2 provides pathway annotations for the Hierarchical HotNet consensus results, and Fig. S1 illustrates these Hierarchical HotNet consensus subnetworks.

4 Comparison with HotNet2

HotNet2 [6] uses a random walk-based approach for inferring the pairwise similarity between vertices. It defines the same joint similarity matrix S (see (6) or (7)), which is denoted as E and described as the "exchanged heat matrix" in [6]. HotNet2 finds a set of similarity thresholds for S using an ensemble of random graphs, and it clusters the vertices at these thresholds by finding the strongly connected components of the graph corresponding to the thresholded similarity matrix. HotNet2 then compares the observed graph against (another) collection of random graphs by counting the number of strongly components above a certain size, using this statistic to evaluate statistical significance. It combines results from multiple networks and vertex scores with a multi-stage consensus procedure that allows it to remove artifacts from particular network topologies and sets of vertex scores.

Hierarchical HotNet also uses a random walk-based approach for inferring the pairwise similarity of the vertices. However, instead of clustering the vertices at a fixed set of similarity thresholds, it finds a hierarchical decomposition of the vertices that is equivalent to clustering the vertices at all such thresholds. It compares the observed graph against a collection of random graphs by

ABCA2	*CUL9*	*LAMA4*	*RASA1*
ADAMTS13	*CYTH1*	*LAMA5*	*RB1*
AKT1	*DRC1*	*LAMC3*	*RIT1*
APC	*DROSHA*	*MAGI2*	*RPE65*
ARHGAP35	*DSCAML1*	*MAP2K4*	*RPL5*
ARID1A	*E4F1*	*MAP3K1*	*RUNX1*
ARID2	*EGFR*	*MBD6*	*SCAPER*
ASXL1	*EP300*	*MCF2L2*	*SERPINB5*
ASXL2	*ERBB2*	*MEGF10*	*SMAD4*
ATM	*ERBB3*	*MERTK*	*SMARCA4*
B2M	*F5*	*MMRN1*	*SMARCB1*
BAP1	*F8*	*MTOR*	*SPEN*
BRAF	*FBXW7*	*MYD88*	*SPOP*
C3	*FGFR2*	*MYO7A*	*SPTBN4*
CADPS2	*FLT3*	*NF1*	*STK11*
CASP8	*HIPK1*	*NFASC*	*SV2A*
CBFB	*HIPK2*	*NID2*	*TBL1XR1*
CCDC40	*HLA-A*	*NOTCH1*	*TOPORS*
CD1A	*HRAS*	*NPM1*	*TP53*
CD46	*IFT122*	*NRAS*	*TTBK1*
CDH1	*IFT140*	*NRCAM*	*TTBK2*
CDKN1B	*KCNQ2*	*OTUD3*	*TTC21B*
CDKN2A	*KCNQ3*	*PBRM1*	*UHRF1BP1L*
CELSR3	*KCNQ5*	*PCDH1*	*USP24*
CFH	*KDM6A*	*PIK3CA*	*VHL*
CHD4	*KIDINS220*	*PIK3R1*	*VWA8*
CHD8	*KMT2C*	*PKD1*	*VWF*
CNKSR1	*KMT2D*	*PKD2*	*WDR19*
CNTN1	*KMT2E*	*PLCL2*	*ZNF420*
CR2	*KRAS*	*PLXNB3*	*ZNFX1*

Table S1: 128 genes in the Hierarchical HotNet consensus results

Name	Genes
BAP	*ASXL1*, *ASXL2*, *BAP1*
CBF	*CBFB*, *RUNX1*
Notch	*CD46*, *CDKN1B*, *CNTN1*, *CREBBP*, *EP300*, *FBXW7*, *NOTCH1*, *SPEN*
p53	*AKT1*, *ATM*, *CDKN2A*, *CHD4*, *EP300*, *HIPK1*, *HIPK2*, *MTOR*, *STK11*, *TP53*
PI(3)K	*AKT1*, *CDKN1B*, *EGFR*, *ERBB2*, *ERBB3*, *FGFR2*, *MTOR*, *PIK3CA*, *PIK3R1*
Ras/Raf	*BRAF*, *CNKSR1*, *EGFR*, *ERBB2*, *ERBB3*, *FGFR2*, *HRAS*, *KRAS*, *KSR2*, *NF1*, *NRAS*, *RASA1*, *SPTBN4*, *VWF*
Rb	*CDKN2A*, *RB1*
RTKs	*ERBB2*, *ERBB3*, *EGFR*
SWI/SNF	*SMARCA4*, *SMARCB1*

Table S2: Overlap between Hierarchical HotNet consensus results and several biological processes and pathways that are known to harbor driver mutations in cancer.

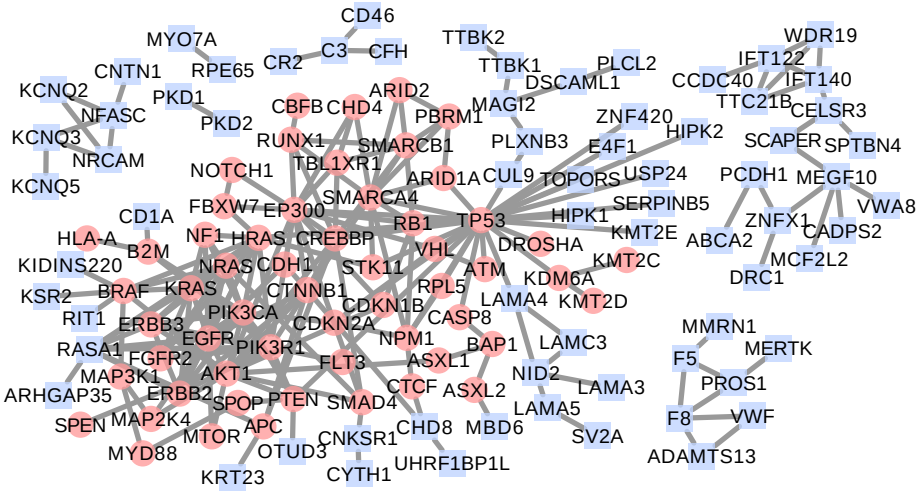

Figure S1: Hierarchical HotNet consensus subnetworks $\bar{G}_2$ for three interaction networks (HINT+HI, iRefIndex, ReactomeFI) and two gene scores (TCGA PanCanAtlas mutation frequency, MutSig q-value). Red circles indicate CGC genes and blue squares indicate non-CGC genes.

comparing the corresponding hierarchies, and it identifies a region of the hierarchy at which the observed hierarchy differs most from the expected hierarchies, allowing it to find significant clustering of high-scoring vertices at any region in the hierarchy. It combines results from multiple networks and vertex scores with a simplified consensus procedure that allows it to further remove artifacts from network and vertex scores.

Hierarchical HotNet improves upon HotNet2 in the following ways:

1. Hierarchical HotNet introduces a dendrogram of vertex sets. In a biological setting, this dendrogram allows us to observe gene sets at different biological scales and the relationships between those gene sets across biological scales. Hierarchical HotNet is able to identify statistically significant regions of the hierarchy while HotNet2 identifies statistically significant clusters at over a small set and narrow range of parameter values.

2. Hierarchical HotNet is able to identify statistically significant results on data sets for which HotNet2 could not find statistically significant results. In some of these cases, the parameter selection procedure in HotNet2 failed to identify parameters where statistically significant clustering occurred, but Hierarchical HotNet uses fewer parameters and has a simpler but more robust parameter selection procedure for the remaining parameters. In other cases, the results were not significant in HotNet2, but Hierarchical HotNet has an improved significance test that evaluates the size of the clustered sets instead of the number of clustered sets exceeding a certain size. Hierarchical HotNet also has more options for evaluating statistical significance.

3. Hierarchical HotNet can be applied more diverse data, including directed and edge-weighted graphs. HotNet2 can only be applied to graphs undirected and unweighted edges.

4. Hierarchical HotNet has a simpler, more robust consensus procedure for combining results from different networks and sets of vertex weights. This simplified procedure allows it to more easily be used with different numbers of graph topologies and vertex weights. In a biological setting, Hierarchical HotNet identifies a larger fraction of cancer genes.

5. The Hierarchical HotNet code can produce HotNet [10, 9] and HotNet2 results with less computational cost by changing the parameters and test statistics to those used by these methods. However, we recommend using Hierarchical HotNet instead of HotNet or HotNet2 because this method provides more robust results on a broader range of datasets.

References

[1] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie,

Maulik R Kamdar, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 42(D1):D472–D477, 2014.

[2] Jishnu Das and Haiyuan Yu. Hint: High-quality protein interactomes and their applications in understanding human disease. *BMC systems biology*, 6(1):92, 2012.

[3] Antonio Fabregat, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, Steven Jupe, Florian Korninger, Sheldon McKay, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 44(D1):D481–D487, 2016.

[4] Heiko Horn, Michael S Lawrence, Candace R Chouinard, Yashaswi Shrestha, Jessica Xin Hu, Elizabeth Worstell, Emily Shea, Nina Ilic, Eejung Kim, Atanas Kamburov, et al. Netsig: network-based discovery from cancer genomes. *Nature methods*, 2017.

[5] Michael S Lawrence, Petar Stojanov, Craig H Mermel, James T Robinson, Levi A Garraway, Todd R Golub, Matthew Meyerson, Stacey B Gabriel, Eric S Lander, and Gad Getz. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495, 2014.

[6] Mark DM Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, Jonathan V Eldridge, Jacob L Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, 47(2):106–114, 2015.

[7] Sabry Razick, George Magklaras, and Ian M Donaldson. irefindex: a consolidated protein interaction database with provenance. *BMC bioinformatics*, 9(1):1, 2008.

[8] Thomas Rolland, Murat Taşan, Benoit Charloteaux, Samuel J Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, Irma Lemmens, Celia Fontanillo, Roberto Mosca, et al. A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226, 2014.

[9] Fabio Vandin, Patrick Clay, Eli Upfal, and Benjamin J Raphael. Discovery of mutated subnetworks associated with clinical data in cancer. In *Pacific Symposium on Biocomputing*, volume 17, pages 55–66, 2012.

[10] Fabio Vandin, Eli Upfal, and Benjamin J Raphael. Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*, 18(3):507–522, 2011.