[Supplementary Figure] Large-scale integrative network-based analysis identifies common pathways disrupted by copy number alterations across cancers

TaeHyun Hwang†1, Gowtham Atluri2, Rui Kuang2, Vipin Kumar2, Timothy Starr1,4, Kevin Silverstein1, Peter M Haverty3, Zemin Zhang3, Jinfeng Liu1,3

1Masonic Cancer Center, 2Department of Computer Science and Engineering, University of Minnesota - Twin Cities, Minneapolis, Minnesota, United States of America; 3Department of Bioinformatics and Computational Biology, Genentech Inc., South San Francisco, California, United States of America; 4Department of Obstetrics, Gynecology & Women’s Health, University of Minnesota, Minneapolis, Minnesota, United States of America

†To whom correspondence should be addressed. E-mail: (hwang071@umn.edu, liu.jinfeng@gene.com)
Supplementary Figure 1. The fraction of annotated genes in copy number alterations. This figure describes the fraction of significantly altered genes in copy number that are annotated with existing pathway and process database. Each gene in copy number alteration is ranked by its corresponding qval from GISTIC using pooled analysis. For example, more than 230 out of 380 (top 2%) ranked genes (18,932 genes in hg18) in significantly altered copy number regions are not annotated with KEGG, Biocarta, Reactome, GO Biological Process (BP), and Molecular function (MF).
Supplementary Figure 2. The significance of copy number alterations across cancers. This figure describes the significance of amplifications (top, red), and deletions (bottom, blue) on genomes (x-axis) across different types of cancers (y-axis) using GISTIC-qval. Qval from GISTIC is -log10 transformed.
Supplementary Figure 3. TGF-beta signaling pathway is commonly disrupted by genes in copy number alterations across multiple types of cancers. This figure describes commonly disrupted TGF-beta signaling pathway across many cancer types. Red, and green colors indicate that genes are amplified, or deleted in copy number alterations, respectively. Gray color indicates that the gene is not altered by copy numbers. Diamond shape node represents that member genes in the pathway, and circular shape node represents genes that are interacting with member genes in the pathway in protein-protein interaction network. Size of nodes represents recurrent frequency of genes in copy number data (i.e. larger size of gene node indicates the gene is
frequently altered in patient samples). One interesting observation is that copy number changes of genes in the TGF-beta signaling pathway are different across many types of cancers, but they are consistently interacting with frequently altered genes in the protein-protein interaction network.
Supplementary Figure 4. Telomerase pathway is commonly disrupted by genes in copy number alterations across multiple types of cancers. This figure describes commonly disrupted Telomerase pathway across many cancer types. Red, and green colors indicate that genes are amplified, or deleted in copy number alterations, respectively. Gray color indicates that the gene is not altered by copy numbers. Diamond shape node represents that member genes in the pathway, and circular shape node represents genes that are interacting with member genes in the pathway in protein-protein interaction network. Size of nodes represents recurrent frequency of genes in copy number data (i.e. larger size of gene node indicates the gene is frequently altered in patient samples).
Supplementary Figure 5. NTRK1 (TrkA) signaling pathway is commonly disrupted by genes in copy number alterations across multiple types of cancers. This figure describes commonly disrupted TrkA signaling pathway across many cancer types. Red, and green colors indicate that genes are amplified, or deleted in copy number alterations, respectively. Gray color indicates that the gene is not altered by copy numbers. Diamond shape node represents that member genes in the pathway, and circular shape node represents genes that are interacting with member genes in the pathway in protein-protein interaction network. Size of nodes represents recurrent frequency of genes in copy number data (i.e. larger size of gene node indicates the gene is frequently altered in patient samples).
Supplementary Figure 6. Distribution of pathway activity scores. This figure describes a distribution of activity scores of pathways with a distribution of activity scores of pathways using randomization from Biocarta pathway database. A background distribution of randomization (black) is compared to a distribution of true pathway activity score (red). Pathways which have significant scores of true pathway activity score (on the right(red)) include TGF-beta, telomerase, NTRK1 and EGFR signaling pathways.
Supplementary Figure 7. Heat map describing the correlation coefficient of pathway co-disruption. This figure describes a distribution of activity scores of pathways with a distribution (red: positive correlation, green negative correlation). We collected top 3% (7 out of 217) disrupted pathways from the Biocarta pathway database in each cancer type, and found 45 disrupted pathways. We computed correlation coefficient of 47 pathways based on inferred pathway activity scores, and applied two-way hierarchical clustering using Cluster 3 with complete linkage to find co-disrupted pathway clusters.
Supplementary Figure 8. False discovery rate using decoy pathways. Decoy pathways were created by assigning a new gene name to each gene in a pathway. NetPathID and Lee Lee et al. methods were then used to compute the activity of every pathway. We computed the areas under the curve (AUCs) based on rankings of the pathways for the methods in each pathway database. Since we are most interested in whether the true biological pathway is near the top, we report the AUC up to the first 50 false positives. Specifically, we calculated the percentage of the area under the ROC curve at 50 false positives. Each bar plot in each figure shows the AUC 50 for distinguishing real from decoy pathways using the perturbation ranking. For example, in breast cancer using Biocarta pathway database, AUC 50s are 0.821 and 0.452 for NetPathID and Lee et al method, respectively.
**Supplementary Figure 9. Clustering and Kaplan-Meier analysis** This figure describes two-way hierarchical clustering of two ovarian cancer datasets and breast cancer dataset using member genes in commonly disrupted pathways Kaplan-Meier survival plots for the clusters of patient subgroups from cancer microarray gene expression datasets (left) and colors indicate patient subgroups used for Kaplan-Meier analysis (right). (A, B): We found four patient subtypes, and Group C patients (median survival 35 months) have longer survival time, compared to Group A,
B, and D patients (median survival 29.5, 24, and 24 months, respectively). For instance, Group B and C patients show significantly different survival time (p-value < 0.01984, Hazard ratio = 1.3089). (C,D): While Group A (median survival: 18 months) and D (median survival: 19 months) have short survival time. Specifically, Group C and D show significantly different survival time (Hazard ratio: 1.4222, p-value <0.00906). (E,F): We found three patient subgroups in breast cancer dataset. Group C patients have short survival time, compared to Group A, and B patients (p-value <0.00475, Hazard ratio = 1.4847, median survival of Group A and C: 36 vs. 24 months).
Supplementary Figure 10. A functional map of commonly disrupted pathways across cancers

Functional enrichment results of commonly disrupted pathways across cancers were mapped as a network of gene sets (nodes) related by mutual overlap (edges), where the color (red) indicates significance (FDR q-value) of gene sets. Node size is proportional to the total number of genes in each set and edge thickness represents the number of overlapping genes between sets. Groups of functionally related gene sets are circled and labeled with enriched GO biological processes manually. Enriched gene sets are coherent (e.g. enriched gene sets have similar biological processes and functions) and showed many cancer-related biological processes including regulation of cell death, apoptosis, cell migration, and cell differentiation.
**Supplementary Figure 11. Effects on different parameter choices of alpha.** We tested five values for hyper parameter alpha to the label propagation. One interesting observation is that the alpha parameter influences the number of commonly disrupted pathways: the larger alpha parameter, more commonly disrupted pathways. This can be explained by the optimization formulation in equation (1); when the alpha is large, we put more weights on the cluster structures in the protein-protein interaction network, and thus label propagation favors the modularity structure in the protein-protein interaction network by assigning highly consistent activity scores to the densely connected genes. In other words, those genes that are densely connected in the gene modules (e.g. member genes in the pathways) will get highly consistent activity scores. In contrast, when the alpha is close to 0, the label propagation is similar to simply overlaying activity score to genes without exploring the cluster structures in the protein-protein interaction networks.