Supplementary Figures

Hub genes in a pan-cancer co-expression network show potential for predicting drug responses

Francisco Azuaje 1,+,* , Tony Kaoma 2, Céline Jeanty 1,+ , Petr V. Nazarov 2, Arnaud Muller 2, Sang-Yoon Kim 2, Gunnar Dittmar 2, Anna Golebiewska 1, Simone P. Niclou 1.

1NorLux Neuro-Oncology Laboratory, Department of Oncology, Luxembourg Institute of Health (LIH), Luxembourg, Luxembourg.
2Proteome and Genome Research Unit, Department of Oncology, Luxembourg Institute of Health (LIH), Luxembourg, Luxembourg.

+ Current affiliation: 2

Correspondence:
Dr. Francisco Azuaje

Francisco.Azuaje@lih.lu (F.A.)
Supplementary Figure S1. Statistical enrichment analysis of biological processes and pathways in the set of 47 network hubs. Related to Figure 2B. An alternative visualization of functional enrichments using G:Profiler (Reimand et al., 2007). As shown in Figure 2B, our set of 47 hub genes (columns) is significantly associated (at corrected P-value = 0.05) with a diversity of biological processes and pathways (rows). Each colored cell represents the association between an individual gene and a functional annotation. Colors are used to specify the evidence type of the observed association: ☐: Inferred from experiment, ☇: Direct assay / Mutant phenotype, ☉: Traceable author, ☪: Electronic annotation, additional information at http://biit.cs.ut.ee/gprofiler.
Supplementary Figure S2. Enriched GO terms detected in our 47-gene signature and their association with drugs on the basis of their known targets. Terms are projected onto a semantic similarity space with REViGO, in which similar terms are positioned closer to each other. Each term is represented by a bubble with color and size indicating the term’s level of statistical enrichment in our list and frequency in the GO database respectively. Detected associations between biological process and drugs are indicated in text boxes.
Supplementary Figure S3. Alternative visualizations and unsupervised clustering of CCLE and GDSC cell lines on the basis of their 47-hub gene profiles. Related to Figures 2D and 4A. Spectral clustering analysis was applied using the SNFtool (Wang et al., 2014) to independently explore the potential of the 47 genes’ expression data to segregate (cell line-drug experiment) samples. A, CCLE and B, GDSC results. In A. and B., rows and columns in each heatmap represent samples and genes respectively, and color represents gene expression intensity. To facilitate visualization, clustering results for different numbers of clusters (K) are provided as independent plots. Note that the order of the rows in each clustering (plot) is not preserved. In each plot, additional columns (right side) representing the drug sensitivity of the samples against Erlotinib ad Paclitaxel are illustrated.
Supplementary Figure S4. Predicted vs. actual sensitivity values in the CCLE and GDSC datasets (Related to Figures 3A and 4B). Alternative visualization to those shown in Figure 3A. A. CCLE plot (n=10981). B. GDSC plot (n=9984).
Supplementary Figure S5. Distribution of predicted versus observed drug sensitivity values according to specific drugs (related to Figure 3A). Upper panel: Density plot of predicted AA values, with most effective drugs (based on observed and predicted values) highlighted. Lower panel: Focused view of Figure 3A that shows the predicted vs. actual AA values for each of the top-8 most effective drugs.
Supplementary Figure S6. Comparison of the expression of our signature’s 47 genes among all samples in microarray and RNA-Seq datasets from the CCLE.

Supplementary Figure S7. Characterization of mean gene expression and variability in (CCLE) microarray and sequencing datasets. In both platforms, the 47 genes of our signature (highlighted dots) show average expression in the center of the distribution and high variability between the samples.
Supplementary Figure S8. Characterization of candidate hubs in microarray and sequencing platforms. Gene networks based on array and sequencing data, considered sum of R² as a measure of degree of each gene (node) and visualized the distribution for all genes and the selected 47-gene subset. For both platforms, our 47 genes are among the top hub genes.

Supplementary Figure S9. Robustness of the predictive performance of our model for multiple (randomly sampled) 10-fold cross-validation (CV) iterations. Prediction performance is summarized with the concordance index (CI) for CV 100 iterations. All CV iterations reported CIs between 0.765 and 0.77, and a coefficient of variation = 0.026%.
Supplementary Figure S10. Prediction performance of the 47-gene (multiple linear regression) model trained on the GDSC. Concordance index (CI) and Spearman (rs) and Kendall (rk) correlations are also reported.

Supplementary Figure S11. Summary of prediction performance of the GDSC-derived model on the CCLE dataset. Concordance indices (CIs) between the predicted and the observed sensitivity values for the drugs found in both datasets.
Supplementary Figure S12. Comparison of our drug sensitivity model to Dong et al.’s models. Our model was adapted to perform sensitivity classification as done in Dong et al’s, which was based on a predetermined subset of samples exhibiting extreme “sensitivity” and “resistant” responses. The right-side plot is an alternative representation (with drug labels) of the bottom-left plot.
Supplementary Figure S13. Comparison of our hub-based model vs. a model based on the elastic net as reported in (Jang et al., 2014).
Supplementary Figure S14. The 47-gene signature distinguishes cell types and is reproducible. Related to section: “Independent *in vitro* validation on several cell lines and compounds”. Gene expression of 47 genes in 4 GBM cell lines using microarrays (U87, NCH601, NCH421k, NCH644) and 3 GBM cell lines qPCR (U87, NCH421k, NCH644). Analysis performed to verify the robustness and platform-independent replicability of the 47-gene expression data and its capacity to distinguish between cell lines.

Supplementary Figure S15. Boxplot summary of prediction results for 4 drugs (Erlotinib, 17-AAG, Panobinostat and Paclitaxel) and 4 GBM cell lines. Related to Figure 5B. Each cell line type comprises multiple biological replicates (18 samples in total): 6 U87, 3 NCH644, 3 NCH601 and 6 NCH421k samples.
Supplementary Figure S16. Drug response curves for the 4 drugs tested on the 4 GBM cell lines. Related to Figure 5C. Drugs were tested on each cell line in triplicates, and relative viability (vs. vehicle-treated samples) was measured for 8 drug concentration values (shown here as Log[µM]).

Supplementary Figure S17. Drug sensitivity predictions for the 4 drugs tested on 4 GBM cell lines together with Topotecan. Topotecan is known to target TOP1 (DNA Topoisomerase I). In our set of GBM cell lines selected for validation, TOP1 is relatively highly expressed in NCH601 and weakly expressed in U87 (data not shown in the figure).