Journal of Visualized Experiments

Analyzing tumor gene expression factors with the CorExplorer web portal

--Manuscript Draft--

Article Type:	Invited Methods Article - JoVE Produced Video
Manuscript Number:	JoVE60431R1
Full Title:	Analyzing tumor gene expression factors with the CorExplorer web portal
Section/Category:	JoVE Cancer Research
Keywords:	Cancer; tumor RNAseq; precision oncology; mutual information; correlation explanation
Corresponding Author:	Shirley Pepke Lyrid LLC South Pasadena, CA UNITED STATES
Corresponding Author's Institution:	Lyrid LLC
Corresponding Author E-Mail:	spepke@lyridllc.com
Order of Authors:	Shirley Pepke
	William M. Nelson
	Greg Ver Steeg
Additional Information:	
Question	**Response**
Please indicate whether this article will be Standard Access or Open Access.	Open Access (US$4,200)
Please indicate the **city, state/province, and country** where this article will be **filmed**. Please do not use abbreviations.	Marina del Rey, California, USA

TITLE:

Analyzing Tumor Gene Expression Factors with the CorExplorer Web Portal

AUTHORS AND AFFILIATIONS:

Shirley Pepke[1], William M. Nelson[2], and Greg Ver Steeg[3]

[1]Lyrid LLC, South Pasadena, CA
[2]Independent consultant, Tucson, AZ
[3]Information Sciences Institute, University of Southern California, Los Angeles, CA

Corresponding author:

Shirley Pepke (spepke@lyridllc.com)

Email addresses of co-authors:

William M. Nelson (wmn@protonmail.com)
Greg Ver Steeg (gregv@isi.edu)

KEYWORDS:

Correlation explanation, tumor RNA-seq, computational oncology, mutual information, gene expression, cancer

SUMMARY:

We introduce the CorExplorer web portal, an resource for exploration of tumor RNA sequencing factors found by the machine learning algorithm CorEx (Correlation Explanation), and show how factors can be analyzed relative to survival, database annotations, protein-protein interactions, and one another to gain insight into tumor biology and therapeutic interventions.

ABSTRACT:

Differential gene expression analysis is an important technique for understanding disease states. The machine learning algorithm CorEx has shown utility in analyzing differential expression of groups of genes in tumor RNA-seq in a way that may be helpful for advancing precision oncology. However, CorEx produces many factors that can be challenging to analyze and connect to existing understanding. To facilitate such connections, we have built a website, CorExplorer, that allows users to interactively explore the data and answer common questions related to its analysis. We trained CorEx on RNA-seq gene expression data for four tumor types: ovarian, lung, melanoma, and colorectal. We then incorporated corresponding survival, protein-protein interactions, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichments, and heatmaps into the website for association with the factor graph visualization. Here we employ example protocols to illustrate use of the database for comprehending the significance of the learned tumor factors in the context of this external data.

INTRODUCTION:

Since its introduction just over a decade ago, RNA-seq has become a ubiquitous tool for measuring gene expression[1]. This is because it enables rapid and cheap de novo profiling of the

entire transcriptome of a sample. However, RNA-seq tumor data reflects an underlying biology that is intrinsically complex and often under-sampled, while the data itself is high-dimensional and noisy. This presents a significant challenge for extracting reliable signals. The CorEx algorithm leverages multivariate mutual information to find subtle patterns in such situations[2,3] . This technique was previously adapted to analyze ovarian tumor RNA-seq samples from the The Cancer Genome Atlas (TCGA) and in this context it appeared to have significant advantages over more commonly used analysis methods[4].

Though the use of RNA-seq is enormously widespread in research applications, including in oncology, those efforts have not led to broad utilization for the purposes of clinical interventions[5]. Part of the reason for this is a lack of user-friendly algorithms and software targeted to these specific problems. To help bridge this gap, we have designed the CorExplorer web portal to enable researchers from a variety of backgrounds to study gene expression factors of tumor RNA-seq samples as found by the CorEx machine learning algorithm. The CorExplorer portal supports interactive visualization and querying of factors from several different tumor types including lung, colon, melanoma, and ovarian[6–10], with the intent of helping researchers to sift through the data correlations and identify candidate pathways to stratify patients for therapeutic purposes.

We expect the CorExplorer portal may be useful to several types of users. The portal was designed with the user in mind who wishes to understand the broad factors driving tumoral gene expression differences in public databases and possibly also place individual gene expression profiles in the context of tumors with similar characteristics. In addition to the representative protocols outlined here, CorExplorer investigations may serve as a starting point to suggest hypotheses for further testing, to compare and contrast CorEx findings on datasets outside of the CorExplorer, and to connect pathological expression signatures of one or a few genes in an individual tumor to larger groups that may be coordinately affected. Finally, it may serve as a user-friendly introduction to the application of machine learning to RNA-seq for those getting started in the field.

PROTOCOL:

1. Exploring factors containing a gene of interest

1.1. Open a web browser and go to http://corex.isi.edu, the CorExplorer home page.

1.2. On the right side under **Quick Links**, click on the **+** expand button next to **Ovarian (TCGA-OV)** to see a summary of the CorEx factor graph that was trained on the TCGA ovarian cancer data (shown in **Figure 1**). Optionally, click on others to compare.

1.3. Once finished inspecting the factor graphs, click on **Lung (TCGA-LUAD)** to access the CorExplorer page for lung cancer RNA-seq.

1.3.1. Explore the CorEx factor graph for a gene of interest using the CorExplorer ‘Factor Graph’ window.

1.3.1.1. Move the mouse cursor over the factor graph display window. Zoom into the factor graph using the mouse scroll wheel or trackpad to see details of the graph such as the most important genes in each factor and the connections between nodes at different layers. Alternatively, click and drag to move the view area or any node.

1.3.1.2. To find a target gene (here we'll use BRCA1), click on the **Gene** dropdown menu at the top of the factor graph window. Type 'BRCA1' to select it in the dropdown list and press **Return** to make the view zoom to factor 26, the factor with which BRCA1 is most strongly correlated.

1.3.1.3. Reposition the mouse over the graph display and scroll to zoom out to see the Level 2 node, L2_8, and its associated factors that are neighbors to factor 26. Note that only genes with weight greater than the threshold indicated on the **Min link weight** slider are shown.

1.3.1.4. To see all of the genes associated with the factor, click on the L1_26 node and select **Load additional genes** in the pop-up window. When the word 'Done' appears, close the pop-up window.

1.3.1.5. Now go back to the header section above the factor graph window and grab and drag the **Min link weight** modifier. Now, as the link weight slider is moved down to 0.05, other genes in factor L1_26, including BRCA2, will appear in weight order. Optionally, reposition nodes by grabbing and dragging to improve layout.

1.3.2. Determine how stratification of patients with respect to the factor affects survival by querying in the survival window.

1.3.2.1. In the survival window, uncheck **Sort by p-val**, then select factor 26 in the **Single Factor** dropdown menu in order to show survival curves for factor 26.

1.3.2.2. Scroll down the survival graph to show the number of patients at risk along the x-axis.

1.3.3. Find associations with biological function by querying within the Annotation window.

1.3.3.1. In the annotation window, to sort the **Factor** dropdown menu by factor number rather than False Discovery Rate (FDR), uncheck **FDR sort**.

1.3.3.2. Scroll and click to select factor 26 in the annotation window dropdown to show enrichment annotations for the factor.

1.3.3.3. Scroll down the annotation list until **DNA repair** is visible and click on it to immediately see associated genes highlighted in yellow on the graph display. See the middle panel of **Figure 2**.

1.3.3.4. Note that factors disappear or appear as different GO terms are selected, according to

whether or not they are enriched for genes with the selected annotation, e.g. 'intrinsic apoptotic signaling pathway in response to DNA damage.'

1.3.4. Explore the factors further by adding windows with different functionality.

1.3.4.1. From the top menu bar, add a protein-protein interaction network (PPI) window by selecting **PPI** from the **Add Window** dropdown, then click the **Add** button to add a PPI graph window to the display area. In the PPI graph window, choose factor 'Layer1: 26' to show the protein-protein interactions. Note the density of connections.

1.3.4.2. From top menu bar, instead of **PPI**, select **Heatmap** from the **Add Window** dropdown, then click the **Add** button to add a heatmap window to the display area. In the heatmap window, choose factor 'Layer1: 26' to show the gene expression patterns.

1.3.4.3. Grab and reposition the heatmap window so that the survival window is also visible. Along the top of the heatmap, observe how the orange/blue/grey colored bar corresponds to patient risk strata on survival graph. Results are shown in the bottom of **Figure 2**.

2. Filtering and interpreting CorEx factors using gene weight, survival, and annotation data

2.1. Filter for factors of interest using survival and cluster quality.

2.1.1. From the Dataset dropdown menu at the top, select **TCGA_OVCA** to go to the CorExplorer page for the TCGA ovarian cancer RNA-seq.

2.1.2. Once the page has loaded, note from the survival window that the factor with the largest survival differential for different strata is 114.

2.1.3. At the top of the factor graph window select 'Layer1: 114' from the **Factor** dropdown.

2.1.4. Grab the link weight slider with the mouse and move it up to 0.5. Note that the large number of genes in factor 114 (1609), with none having weight >0.35, indicates a relatively weak clustering.

2.1.5. Next, expand the list of factors in the survival window and select the next best factor in the survival window dropdown, factor 39, to show its associated survival curves.

2.1.6. Select factor 39 in the annotation window by clicking on it. The significant GO and KEGG annotations are shown.

2.2. To gain a better understanding of the biological role of genes in factor 39, interpret the factors using neighborhood annotation information as follows.

2.2.1. At the top of the factor graph window, select factor 'Layer1: 39' in the factor dropdown.

Then, move the mouse over the factor graph window and zoom out to reveal the entire L2_14 cluster with 6 factors: 14, 32, 39, 42, 52, and 82 (shown in **Figure 3**).

2.2.2. To understand the relative significance of the factors linked to the L2_14 node, start by viewing survival differentials for each of the L2_14 factors. Uncheck **Sort by p-val** in the survival window and then click on each of the factor numbers in succession. Doing this, note that only factors 14, 32, and 39 display a survival association.

2.2.3. Now from the top menu bar, select **PPI** from the **Add Window** dropdown once again. Press **Add** to add a PPI graph window to the display area. In the PPI graph window, select factor 'Layer1: 52' to show the protein-protein interactions that are significant. An example layout of windows at this point is shown in **Figure 3**.

2.2.4. Click the **View at StringDB** link at the bottom of the PPI window to link out to the StringDB online database. Click **Continue** from the first screen, then select the **Analysis** tab below the network graph as before to get an online GO analysis for the PPI network genes. The top cellular component is 'MHC class II protein complex.'

2.2.5. Return to the CorExplorer tab and PPI window and select factor 32, this time from the factor dropdown. Click the link **View at StringDB** out to the StringDB analysis. The top cellular component is 'MHC class I protein complex,' in contrast to class II for factor 52 in the previous step!

2.2.6. Finally, go back to the PPI window and select 'Layer1: 39' from the factor dropdown menu at the top. Click the link **View at StringDB** to link out to the StringDB analysis.

2.2.7. Click **Continue** from the first screen, then select the **Analysis** tab below the network graph to get an online GO analysis for the PPI network genes. Observe that the top molecular function is 'CXCR3 chemokine receptor binding.'

3. Using survival and database annotations to look for promising therapeutic combinations

3.1. Switch to the TCGA melanoma CorExplorer by selecting **TCGA_SKCM** from the **Dataset** dropdown menu.

3.2. Note that the factor with the largest survival differential is factor 171. Examine the factor 171 annotations by scrolling and note that 'immune response' and 'cytokine-mediated signaling pathway' are near the top (as they were for the top ovarian factor).

3.3. To find a complementary factor, examine the top survival-associated factors along with their top annotation terms. To do this, click on the **Dataset overview** link in the top menu bar to open a separate tab containing a table with dataset processing details as well as a summary of top factors according to p-value of the survival differential. Note that the first non-immune factor is 88.

3.4. Return to the **TCGA_SKCM** browser tab.

3.5. Select factor 88 in the survival, annotation, and graph windows. The top several GO terms are related to 'rRNA processing' and 'mitochondrion organization,' confirming it as distinct from the immune-related factors.

3.6. In the survival window, on the paired factors dropdown, select '88_171' to see how survival is improved for patients in the middle stratum for the combined 171 and 88 expression factors. Annotation and survival comparisons are illustrated in **Figure 4**.

4. Finding commonalities and differences of gene expression variation across tumor types using the Search page

4.1. Click on the **CorExplorer** heading to return to the front page.

4.2. Click on **Search** on the top menu bar to go to a page allowing searching over all the datasets on the CorExplorer site.

4.3. In the **Gene** search box, enter 'FLT1' (VEGFR1) and hit **Return** or press **Search**. FLT1 is found with a relatively high weight in the following factors: OVCA - 76, LUAD - 162, SKCM - 195 and SKCM - 184, as well as COAD - 112 and COAD - 74.

4.4. Alternatively, search for a related GO term across all the datasets. Try this in the 'GO Search' box by typing 'angiogenesis' and hitting **Return** or pressing **Search**. All FLT1 factors, with the exception of SKCM-195, are listed as statistically enriched for 'angiogenesis' genes–factor 195 does, in fact, have the annotation, but below the default 10-8 threshold. Search results for this and the prior step are shown in **Figure 5**.

4.5. As further examples, in the GO search box, first type 'epidermal growth factor receptor.' Only LUAD is enriched for this term, a well-known stratification factor for lung cancer. Next, type 'mesenchymal' in the search box. This term is enriched in gene expression groups for OVCA, where it is a well-studied stratification factor.

REPRESENTATIVE RESULTS:

Searching for the gene 'BRCA1' in the lung cancer dataset reveals it to be most strongly associated with CorEx factor 26 (**Figure 2**). GO term enrichment for this factor is seen to be extremely high, with DNA repair exhibiting an FDR of only 1×10^{-19}. The selection also draws attention to the second level cluster L2_8 that has six closely related factors as children. Selecting 'DNA repair' in either the GO term annotations or the factor graph's GO enriched dropdown highlights associated genes in each of the factors, with the factor 26 having by far the most, as expected[11]. The protein-protein interaction network is strongly connected, further supporting the tightly linked functionality of the genes in factor 26. The associated survival graph suggests a possible association with patient survival, but this would have to be confirmed in a larger dataset.

Starting with survival can allow dissection of reasons for improved survival associated with particular gene expression groups. As an example, the top factor influencing survival for ovarian cancer is seen to be number 39, which is strongly enriched for genes associated with the immune system (**Figure 3**). Five other factors associated with the same level 2 node are also indicated to be immune-related, however the survival impact appears to be strongly variable among them, with 39 being the highest and 52 being the lowest. Adding a protein-protein interaction window for a factor shows the immediate interaction network and allows for link out to the StringDB[12] website to query various enrichments for the PPI network genes. By doing this for each of the L2_14 factors in turn, one finds that StringDB enrichments for the PPI network genes suggest the following possible explanation for the associations with survival. Factor 32 contains genes that make up the major histocompatibility complex (MHC) class I protein complex, which is recognized by cytotoxic T lymphocytes. Factor 39 corresponds to cytokine signaling and CXCR3 receptor binding, related to CD8+ T lymphocytes. Both of these factors appear to confer a significant survival advantage for patients exhibiting relatively high expression of the corresponding genes. Cytotoxic CD8+ T lymphocytes are primarily responsible for anti-tumor immunity. Factor 52, on the other hand, is comprised of genes coding for proteins in the MHC class II complex which are recognized primarily by CD4+ T helper cells rather than directly by cytotoxic T lymphocytes. The remaining L2_14 factors reflect generalized immune system activation that doesn't differentiate the two types of lymphocyte populations. A survival association specific to cytotoxic T lymphocyte recognition of MCH class I cellular antigens is consistent with our understanding of antitumor immunity in general and from other cancers such as melanoma[13,14].

The web portal supports the discovery of pairs of factors with complementary functions that may suggest effective tumor-specific combination therapies. The dataset overview can be scanned for factors that show a correlation with survival yet have distinct GO enrichments. For melanoma (TCGA_SKCM; **Figure 4**), it is seen that the top survival factor 171 is immune related, while factor 88 down the list shows enrichment for genes related to mitochondrion organization. Indeed, this has been suggested as a target in melanoma[15]. Adding survival windows to the CorExplorer page allows comparison of stratification using the factor pair to that of each factor individually, showing that favorable gene expression patterns from both groups exhibits a trend of survival better than that for either factor alone. The top stratum does not appear to be improved however, suggesting immunotherapy only may be the best option for some patients.

Commonalities and differences among tumors can be seen by searching across datasets for genes or GO terms (**Figure 5**). As an example, FLT1 (aka VEGFR1) is a well-studied pro-angiogenic marker[16,17]. When it is put into the search bar, all of the tumors have factors in which FLT1 plays a major role. Conversely, when the GO term ‘angiogenesis’ is input on the search page, 5 out of 6 of the FLT1 groups appear with that enrichment. All FLT1 factors, with the exception of SKCM-195, are listed as statistically enriched for ‘angiogenesis’ genes. The sixth factor does, in fact, have the annotation, but below the default 10-8 threshold. When the weighting within the factor list is utilized in an alternative enrichment calculator, e.g., Gene Set Enrichment Analysis (GSEA)[18], the sixth factor is found to be significantly enriched for ‘angiogenesis’ genes as well.

It is important to check the heatmaps to ensure the gene expression pattern is of adequate quality to support biological interpretations. Heatmaps that show strong clear variation may exhibit either coordinated expression of the factor genes ranging from low to high or more complex patterns with some genes having low expression correlated with others having high (**Figure 6**). A key marker of a high-quality grouping is the presence of several genes with a smooth variation in expression as a function of factor score. The factor heatmaps show samples ordered according to factor score, thus there should be a smooth gradient moving from left to right. However, this can fail to happen in at least two different ways. Most commonly, the correlations may be extremely noisy (**Figure 5C**), calling into question the robustness and utility of any inferences regarding survival and/or biological function. Also, patterns that happen only in a small minority of samples may not conform to the model of three expression states assumed by the CorEx algorithm, resulting in a misleading classification of the samples (right side of **Figure 5D**).

FIGURE AND TABLE LEGENDS:

Figure 1: CorExplorer front page. After clicking on **+** next to **Ovarian Cancer** under **Quick Links**, factor graph details are shown. The CorEx hierarchical model is made up of input variables (gene expression in this case) on the bottom layer and inferred latent factors in the higher layers.

Figure 2: Using a gene name to guide exploration. The figure shows a series of screenshots illustrating exploration of CorEx lung cancer factors strongly related to BRCA1. First, selecting 'BRCA1' in the **Gene** dropdown box for the factor graph causes the graph view to zoom in on the factor for which BRCA1 has greatest weight. Zooming out a bit frames the layer two node L2_8 connecting that factor to other related ones. Survival and annotations can be compared: clicking on the GO term **DNA repair** highlights annotated genes. A PPI window is added to show the network interactions for genes in the factor. Using the **Add Window** button to add a heat map shows association of expression patterns with survival, suggesting increased expression of DNA repair genes may be associated with decreased survival.

Figure 3: Using clinical data (survival) to guide exploration. Exploring the top survival-associated factor (39) for ovarian cancer reveals interesting relationships among neighboring factors. After selecting factor 39 in the factor graph and zooming out a bit, the layer two factor linked to factor 39 is seen to have five other associated factors. An additional survival window allows direct comparison of the associated survival differentials. Factors 39 and 32 both show a positive survival correlation, in contrast to factor 52, which does not. The protein-protein interaction networks are all well-defined. Linking out to StringDB allows comparison of the GO annotations (not shown): Factor 39 is associated with a cytokine signaling network related to cytotoxic CD8+ T lymphocyte activation and factor 32 is dominated by MHC class I antigen presenting proteins that trigger recognition by such lymphocytes; the neighboring factors, however, are dominated by other immune system components such as CD4+ helper T cells and show no survival correlation.

Figure 4: Exploring top survival factors suggests potential therapeutic combinations. The 'Datasets' link on the home page menu bar leads to a concise table of survival factors ordered by

p-value, along with the top GO annotation (not shown). Using this information for melanoma, the combination of factor 171 for immune function with factor 88 for mitochondrion organization appears complementary. The figure shows annotation windows for each of the factors side-by-side to contrast them. Survival curves for patients stratified by the two factors individually or together indicate that the combination increases the survival differential compared to either factor alone.

Figure 5: The Search page facilitates pan-cancer analysis. Genes or GO biological process terms can be searched for across all datasets using the **Search** link from the home page. The figure shows search results for the gene FLT1 and the GO term 'angiogenesis'. The results show the presence of FLT1 in factors annotated with the term 'angiogenesis' across cancers.

Figure 6: Heatmaps can be used to qualitatively assess correlations among genes and samples according to factor score. High quality gene expression relationships are shown by smooth gradation when patients are ordered by factor score in the heatmaps. The leftmost heatmap for factor 18 is one example. The patterns may also encompass complex signatures of up and down expression as in the middle large heatmap for factor 11. Lower quality patterns sometimes show abrupt changes in expression for a subgroup of patients as in the factor 9 heatmap on the right or simple very noisy correlations as in the factor 161 heatmap at the lower right.

DISCUSSION:

We have presented the CorExplorer site, a publicly accessible web server for interactive exploration of maximally correlated gene expression factors learned from tumor RNA-seq by the CorEx algorithm. We have shown how the website may be used to stratify patients according tumor gene expression, and how such stratification corresponds to biological function and survival.

Other webservers for RNA-seq analysis have been built. Differential and co-expression analysis for tumors can be examined and integrated with other data types in cbioPortal[19,20]. The servers GenePattern[21], Mev[22], and Morpheus[23], incorporate established clustering techniques such as principal component analysis (PCA), kmeans, or self-organizing maps (SOMs). More innovative efforts include CamurWeb[24], based upon an automated rule-generating classifier, and TACCO[25], which implements random forest classifiers and lassos. The CorEx algorithm used here optimizes multivariate information in order to find a hierarchy of factors that explain patterns in data. The nonlinear and hierarchical factor learning appears to yield improved interpretability relative to the linear global factors found via PCA[4]. Additionally, the technique's fine-grained parsing of sample signals allows precise tumor comparisons vis-à-vis more commonly used broad subtypes. This combination of overlapping and hierarchical factor analysis distinguishes the CorExplorer from most other approaches and necessitates new tools for visualization and summarization.

A critical part of the CorExplorer factor analysis is the capability to explore not just several, but over 100 factors with informative gene patterns that are placed within an overlapping hierarchy. The CorExplorer facilitates the mining of these myriad factors for biological and clinical associations and allows for exceptionally detailed characterization of individual tumors.

The unsupervised learning of such a large number of factors means that not all will be relevant to disease biology. In such a case, it is essential to either use annotations or known genes to pull out factors of interest or search for factors associated with clinical data such as survival. Thus, the CorExplorer allows users to implement this very important filtering step. The presence of factor gene patterns in a tumor may even suggest an approach to personalized oncology treatment. Further, the multiplicity of factor scores for each tumor that allows for discovery of potentially useful therapeutic combinations.

It is sometimes the case that no significant GO annotations appear for factors highly correlated with survival. While this may occur due to noisy or under sampled data, there are other possible causes such as a cluster size that is too small to register significant enrichment scores or the group being a 'basket' of single genes from diverse pathways without coherent biological association. Additionally, a category of annotation different from the KEGG and GO biological process, e.g. cellular compartment, may be appropriate. These can be accessed by linking out to StringDB as demonstrated in the protocol. The Gene Ontology enrichment analysis on the CorExplorer site currently does not account for the gene weighting in a factor, though this will likely be remedied in the near future. Note a gene list option is available under 'Add Window' that allows for download of the complete factor gene list for further analysis with external tools.

For the purposes of the website, CorEx was run on each of the datasets five times and the run that resulted in the greatest overall Total Correlation was retained. Having a statistical representation of the results of multiple runs may be more informative and is a goal for future work. Additionally, the set of tumor types available on the server is rather small but we expect this to expand over time according to user interest.

As outlined above, the CorExplorer visualizes CorEx RNA-seq factor relationships along with clinical and database information, thus enabling a variety of different modes of interrogation. We are hopeful that this tool will lead to further work to utilize the power of RNA-seq analysis for discovery and clinical application in oncology.

ACKNOWLEDGMENTS:

GV was supported by DARPA award W911NF-16-0575.

DISCLOSURES:

The authors declare that they have no competing financial interests.

REFERENCES:

1. Petryszak, R. et al. The RNASeq-er API—a gateway to systematically updated analysis of public RNA-seq data. *Bioinformatics.* **33**, 2218–2220 (2017).

2. Steeg, G. V., Galstyan, A. Maximally Informative Hierarchical Representations of High-Dimensional Data. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*. San Diego, CA (2015).

3. Ver Steeg, G., Galstyan, A. Discovering structure in high-dimensional data through correlation explanation. *Advances in Neural Information Processing Systems.* Montreal, Canada (2014).

4. Pepke, S., Ver Steeg, G. Comprehensive discovery of subsample gene expression components by information explanation: therapeutic implications in cancer. *BMC medical Genomics.* **10**, 12 (2017).

5. Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D., Craig, D. W. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nature Reviews Genetics.* **17**, 257 (2016).

6. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature.* **511**, 543 (2014).

7. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature.* **487**, 330 (2012).

8. Akbani, R. et al. Genomic classification of cutaneous melanoma. *Cell.* **161**, 1681–1696 (2015).

9. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature.* **474**, 609 (2011).

10. Grossman, R. L. et al. Toward a shared vision for cancer genomic data. *New England Journal of Medicine.* **375**, 1109–1112 (2016).

11. Moynahan, M. E., Chiu, J. W., Koller, B. H. & Jasin, M. Brca1 controls homology-directed DNA repair. *Molecular Cell.* **4**, 511–518 (1999).

12. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research.* **47**, D607–D613 (2018).

13. Durgeau, A., Virk, Y., Corgnac, S., Mami-Chouaib, F. Recent advances in targeting CD8 T-cell immunity for more effective cancer immunotherapy. *Frontiers in Immunology.* **9**, 14 (2018).

14. Sato, E. et al. Intraepithelial CD8+ tumor-infiltrating lymphocytes and a high CD8+/regulatory T cell ratio are associated with favorable prognosis in ovarian cancer. *Proceedings of the National Academy of Sciences of the United States of America.* **102**, 18538–18543 (2005).

15. De Moura, M. B. et al. Mitochondrial respiration-an important therapeutic target in melanoma. *PLoS One.* **7**, e40690 (2012).

16. Folkman, J., Merler, E., Abernathy, C., Williams, G. Isolation of a tumor factor responsible for angiogenesis. *Journal of Experimental Medicine.* **133**, 275–288 (1971).

17. Takahashi, S. Vascular endothelial growth factor (VEGF), VEGF receptors and their inhibitors for antiangiogenic tumor therapy. *Biological and Pharmaceutical Bulletin.* **34**, 1785–1788 (2011).

18. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America.* **102**, 15545–15550 (2005).

19. Cerami, E. et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery.* **2**, 401–404 (2012).

20. Gao, J. et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Science Signalling.* **6**, pl1 (2013).

21. Reich, M. et al. GenePattern 2.0. *Nature Genetics.* **38**, 500 (2006).

22. Wang, Y. E., Kutnetsov, L., Partensky, A., Farid, J. & Quackenbush, J. WebMeV: A Cloud Platform for Analyzing and Visualizing Cancer Genomic Data. *Cancer Research.* **77**, e11–e14 (2017).

23. Morpheus. Available at: https://software.broadinstitute.org/morpheus.

24. Weitschek, E., Lauro, S. D., Cappelli, E., Bertolazzi, P., Felici, G. CamurWeb: a classification software and a large knowledge base for gene expression data of cancer. *BMC Bioinformatics.* **19**, 354 (2018).

25. Chou, P.-H. et al. tACCo, a Database Connecting transcriptome Alterations, pathway Alterations and Clinical outcomes in Cancers. *Scientific Reports.* **9**, 3877 (2019).

Figure 1 pdf

Click here to access/download;Figure;figure1.front.pdf

CorEx Portal

Not Secure | corex.isi.edu

CorExplorer

Overview Datasets Search How-To Download Publications

Welcome to the CorEx portal for gene expression analysis!

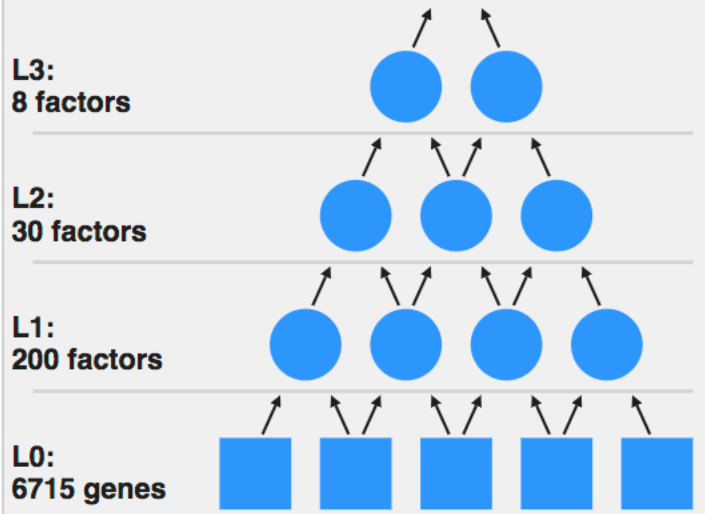
This website began as an extension of work applying the CorEx machine learning algorithm to gene expression data from ovarian cancer tumors as presented in this paper.

CorEx searches for statistical patterns of gene expression present in subgroups of patients. These subgroups can then be mined for significance in terms of survival or database annotations. The primary goal in the original article was to enable meaningful personalized expression profiling for the purposes of both target discovery and treatment selection in ovarian cancer patients.

This website is intended to facilitate the rapid analysis and interpretation of gene expression data for large numbers of tumors across various cancer types. Here you can interactively view and explore the original published dataset, and eventually data for other cancers as well. Additionally, we plan to add the ability to upload custom tumor data in the future. If you are interested in seeing results from a specific GDC study, please let us know.

Quick Links

- Ovarian (TCGA-OV)
- Lung (TCGA-LUAD)
- Melanoma (TCGA-SKCM)
- Colon (TCGA-COAD)
- Ovarian (BMC Paper)

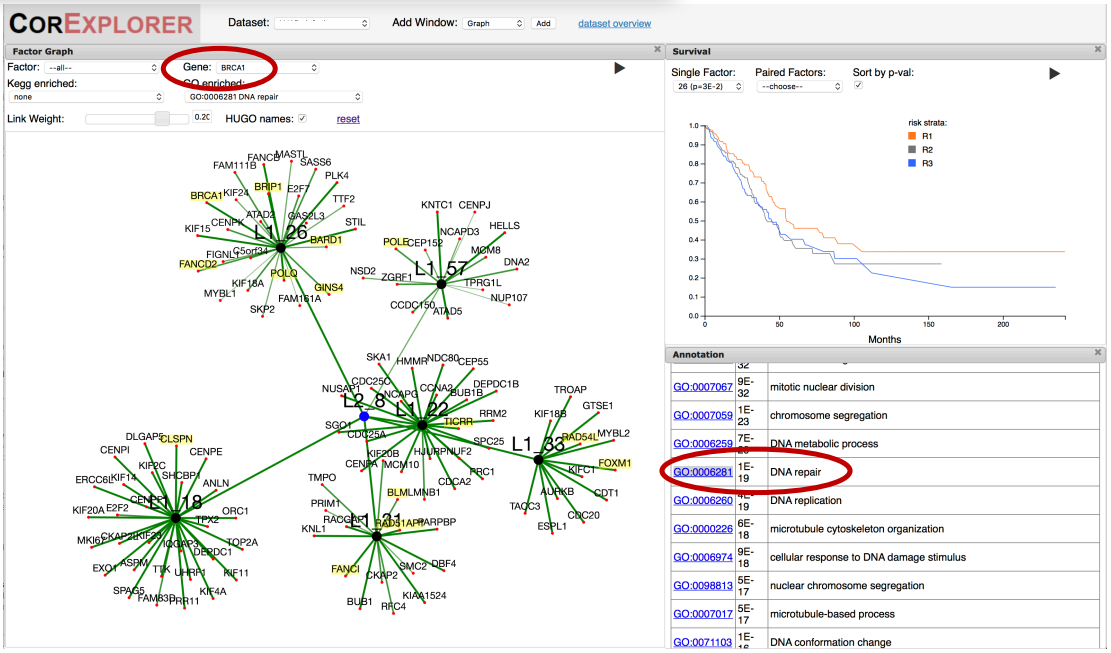


We love feedback! gregv@isi.edu

login

Figure 2 pdf

Click here to access/download;Figure;figure2.LUAD_DNArepair.pdf



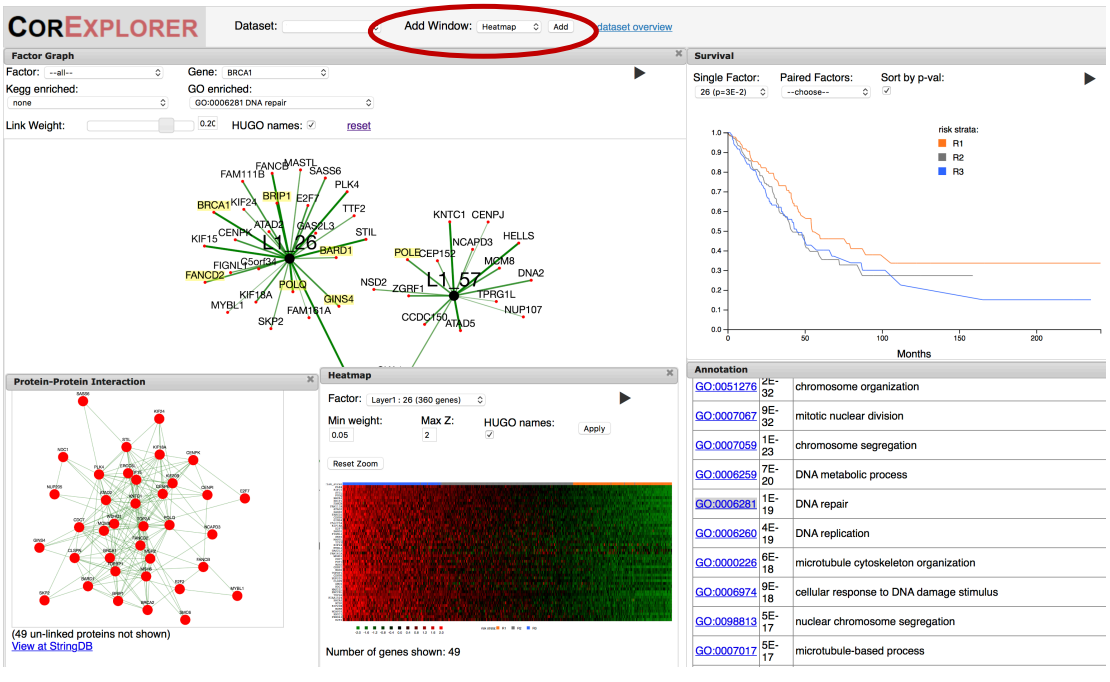


Figure 3 pdf

Click here to access/download;Figure;figure3.ovca_immuno.pdf

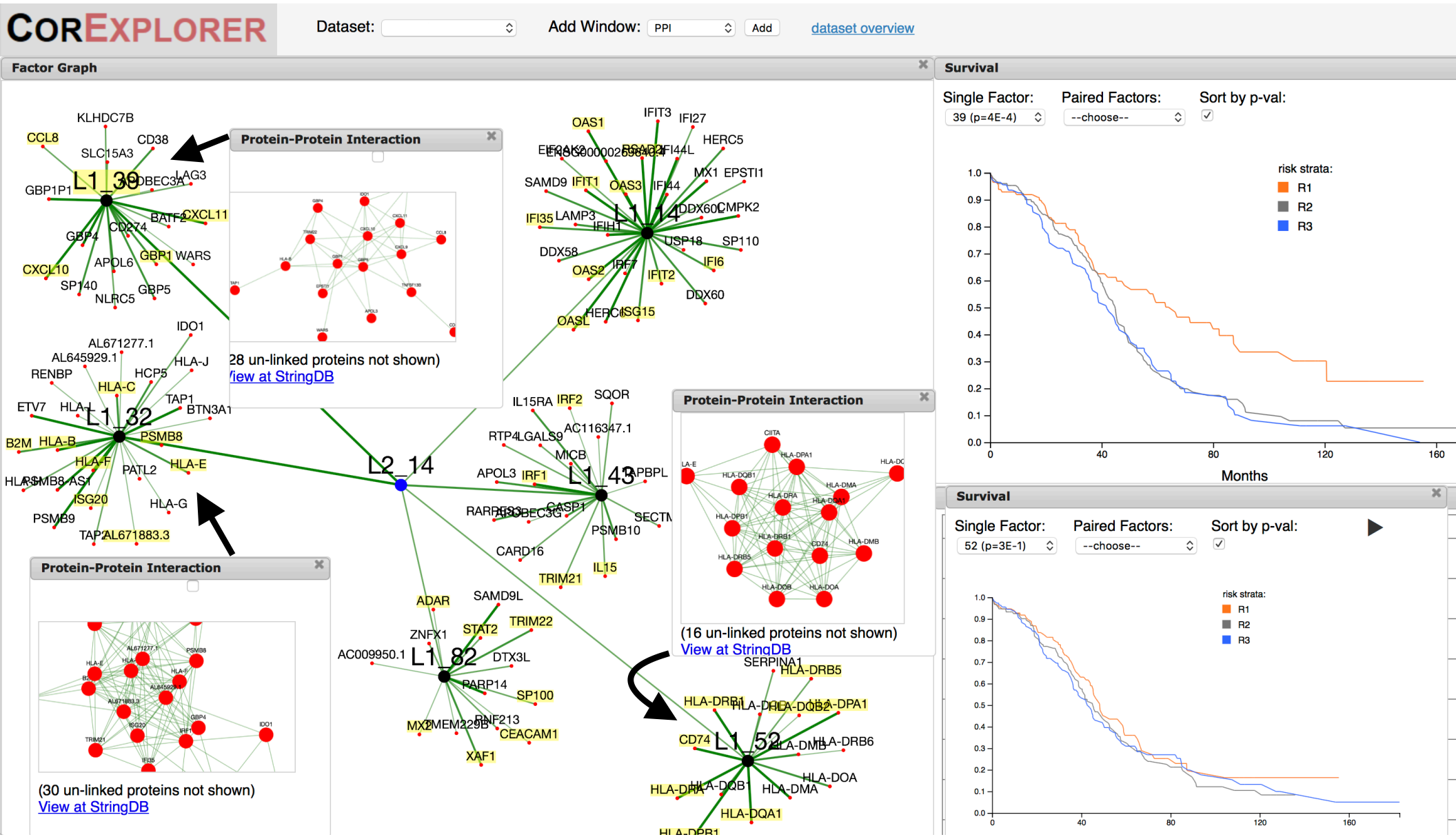


Figure 4 pdf

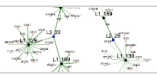
Click here to access/download;Figure;figure4.pdf

CorExplorer Dataset: Add Window: Survival Add dataset overview

Annotation

Factor: 171 (4E-22) GO and Kegg FDR sort

GO	FDR	Description
GO:0006955	4E-22	immune response
GO:0002682	3E-19	regulation of immune system process
GO:0045087	4E-15	innate immune response
GO:0050776	1E-14	regulation of immune response
GO:0002252	1E-14	immune effector process
GO:0006952	3E-14	defense response
GO:0098542	1E-13	defense response to other organism
GO:0051607	2E-12	defense response to virus
GO:0009615	5E-12	response to virus
GO:0002684	2E-11	positive regulation of immune system process
GO:0051707	1E-9	response to other organism
GO:0043207	1E-9	response to external biotic stimulus
GO:0009607	3E-9	response to biotic stimulus
GO:0031347	8E-9	regulation of defense response
GO:0001817	1E-8	regulation of cytokine production
GO:0043122	5E-8	regulation of I-kappaB kinase/NF-kappaB signaling
GO:0045088	5E-8	regulation of innate immune response
GO:0019221	1E-7	cytokine-mediated signaling pathway



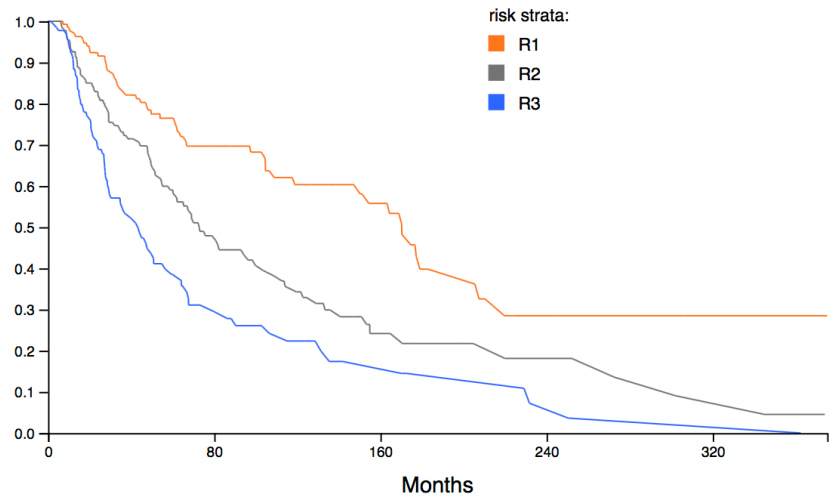
Annotation

Factor: 88 (3E-14) GO and Kegg FDR sort

GO	FDR	Description
GO:0034660	3E-14	ncRNA metabolic process
GO:0034470	7E-11	ncRNA processing
GO:0006396	3E-10	RNA processing
GO:0022613	9E-10	ribonucleoprotein complex biogenesis
GO:0016072	8E-9	rRNA metabolic process
GO:0006364	2E-8	rRNA processing
GO:0042254	2E-8	ribosome biogenesis
GO:0007005	7E-8	mitochondrion organization
GO:0032259	1E-6	methylation
GO:0051276	1E-5	chromosome organization
GO:0006974	1E-5	cellular response to DNA damage stimulus
GO:0033554	2E-5	cellular response to stress
GO:0007006	3E-5	mitochondrial membrane organization
GO:0006399	3E-5	tRNA metabolic process
GO:0043414	5E-5	macromolecule methylation
GO:1901566	5E-5	organonitrogen compound biosynthetic process
GO:0009451	7E-5	RNA modification
GO:0043065	1E-4	positive regulation of apoptotic process
GO:0000122	1E-4	negative regulation of transcription from RNA polymerase II promoter
GO:0044711	1E-4	single-organism biosynthetic process

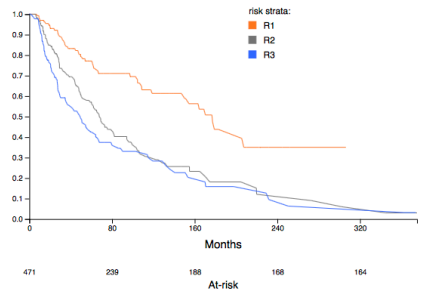
Survival

--choose-- 88_171 (p=5E-10)



Survival

Single Factor: 171 (p=9E-9) Paired Factors: --choose-- Sort by p-val:



Survival

Single Factor: 88 (p=3E-7) Paired Factors: --choose-- Sort by p-val:

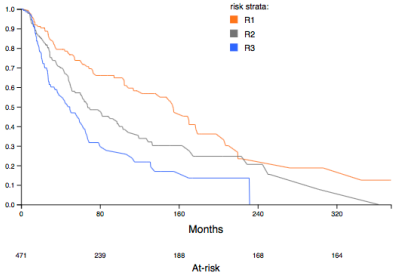


Figure 5 pdf

Click here to access/download;Figure;figure5.search.pdf

CorExplorer | Overview | Datasets | Search | How-To | Download | Publications

Quick Links

- Ovarian (TCGA-OV)
- Lung (TCGA-LUAD)
- Melanoma (TCGA-SKCM)
- Colon (TCGA-COAD)
- Ovarian (BMC Paper)

What's New

- Additions
- Papers

Gene Search

Enter gene name, or part of name (case-insensitive).

Gene: FLT1 Dataset: All datasets Search Clear

Weight: 0.05 MI: 0.05

Gene	Hugo	Dataset	Factor	Weight	MI
FLT1	FLT1	tcga_ov.shrinkage2	108	0.534	0.534
ENSG00000102755.9	FLT1	TCGA_OVCA	76	0.449	0.449
ENSG00000102755.9	FLT1	TCGA_LUAD	162	0.517	0.517
ENSG00000102755.9	FLT1	TCGA_SKCM	195	0.618	0.618
ENSG00000102755.9	FLT1	TCGA_SKCM	184	0.108	0.489
ENSG00000102755.9	FLT1	TCGA_COAD	112	0.606	0.606
ENSG00000102755.9	FLT1	TCGA_COAD	74	0.055	0.373

GO Search

Search for GO-enriched factors. Enter GO term or part of the GO description.

GO: angiogenesis Dataset: All datasets Search Clear

FDR: 1.0E-

GO	GO description	Dataset	Factor	FDR
GO:0001525	angiogenesis	tcga_ov.shrinkage2	7	4E-26
GO:0001525	angiogenesis	tcga_ov.shrinkage2	16	1E-19
GO:0001525	angiogenesis	tcga_ov.shrinkage2	4	4E-18
GO:0001525	angiogenesis	tcga_ov.shrinkage2	133	5E-18
GO:0001525	angiogenesis	tcga_ov.shrinkage2	13	1E-17
GO:0045765	regulation of angiogenesis	tcga_ov.shrinkage2	7	3E-16
GO:0001525	angiogenesis	tcga_ov.shrinkage2	108	1E-13
GO:0001525	angiogenesis	tcga_ov.shrinkage2	106	1E-13
GO:0045765	regulation of angiogenesis	tcga_ov.shrinkage2	16	5E-13
GO:0001525	angiogenesis	tcga_ov.shrinkage2	19	6E-12
GO:0045765	regulation of angiogenesis	tcga_ov.shrinkage2	1	7E-12
GO:0001525	angiogenesis	tcga_ov.shrinkage2	12	2E-11

Figure 6 pdf

Click here to access/download;Figure;figure6.heatmaps.pdf

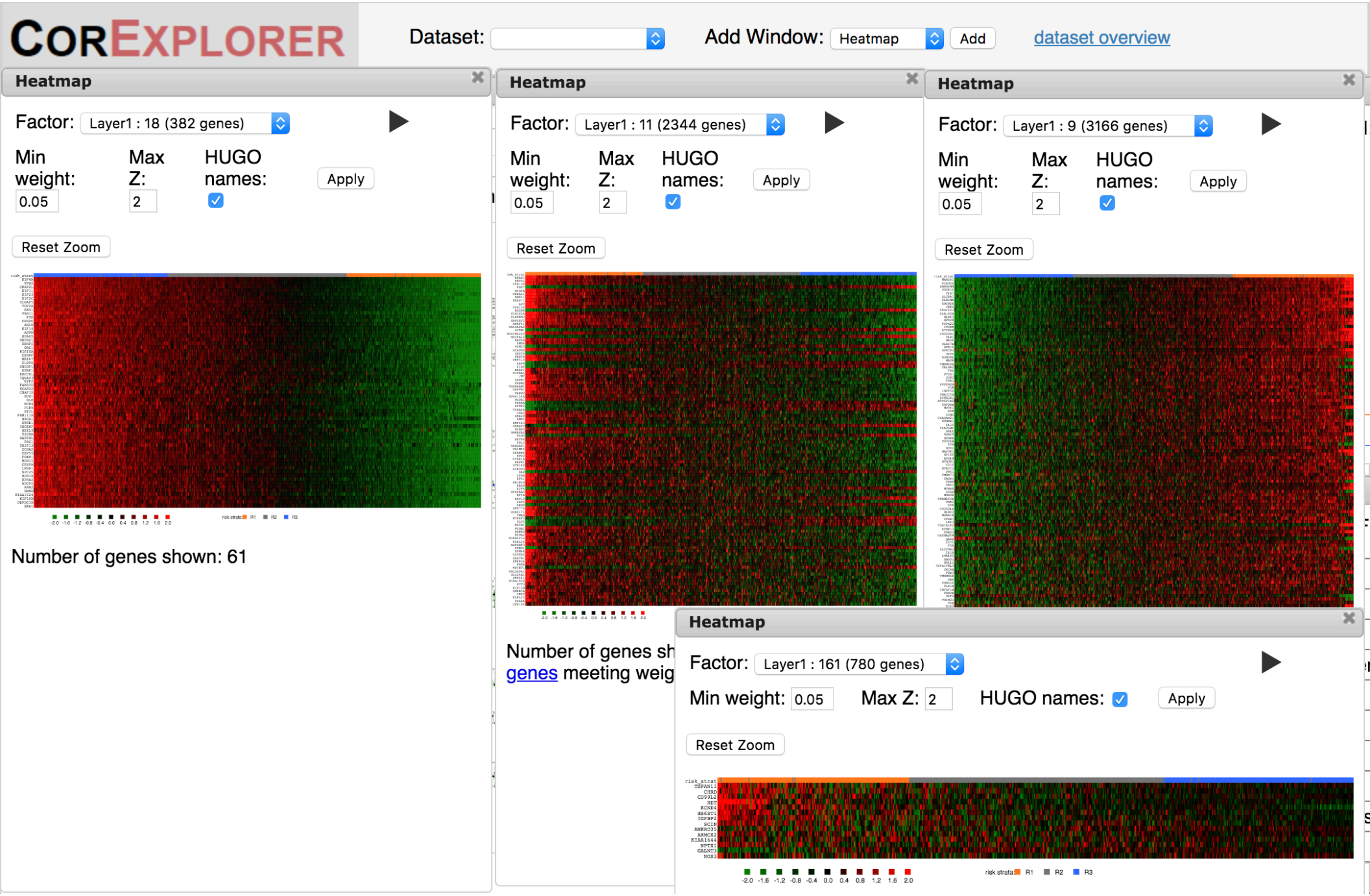


Figure 1


Click here to access/download
Video or Animated Figure
figure1.front.svg

Figure 2

Click here to access/download

Video or Animated Figure

figure2.LUAD_DNArepair.svg

Figure 3


Click here to access/download
Video or Animated Figure
figure3.ovca_immuno.svg

Figure 4


Click here to access/download
Video or Animated Figure
figure4.svg

Figure 5


Click here to access/download
Video or Animated Figure
figure5.search.svg

Figure 6

Click here to access/download

Video or Animated Figure

figure6.heatmaps.svg

Name of Material/Equipment	Company	Catalog Number	Comments/Description
Public server for CorExplorer website	USC	http://corex.isi.edu	Intel Xeon E5-2690 4-core 2.6 GHz, 8GB RAM. Backend architecture is LAMP: Linux, Apache, MySQL, PHP.
Web browser	Google/Apple	Chrome/Safari	Verified web browsers.

1 Alewife Center #200
Cambridge, MA 02140
tel. 617.945.9051
www.jove.com

ARTICLE AND VIDEO LICENSE AGREEMENT

Title of Article:	Analyzing tumor gene expression factors with the CorExplorer web portal
Author(s):	Shirley Pepke, William M. Nelson, and Greg Ver Steeg

Item 1: The Author elects to have the Materials be made available (as described at http://www.jove.com/publish) via:

☐ Standard Access ☒ Open Access

Item 2: Please select one of the following items:

☒ The Author is **NOT** a United States government employee.

☐ The Author is a United States government employee and the Materials were prepared in the course of his or her duties as a United States government employee.

☐ The Author is a United States government employee but the Materials were NOT prepared in the course of his or her duties as a United States government employee.

ARTICLE AND VIDEO LICENSE AGREEMENT

1. **Defined Terms.** As used in this Article and Video License Agreement, the following terms shall have the following meanings: "**Agreement**" means this Article and Video License Agreement; "**Article**" means the article specified on the last page of this Agreement, including any associated materials such as texts, figures, tables, artwork, abstracts, or summaries contained therein; "**Author**" means the author who is a signatory to this Agreement; "**Collective Work**" means a work, such as a periodical issue, anthology or encyclopedia, in which the Materials in their entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole; "**CRC License**" means the Creative Commons Attribution-Non Commercial-No Derivs 3.0 Unported Agreement, the terms and conditions of which can be found at: http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode; "**Derivative Work**" means a work based upon the Materials or upon the Materials and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Materials may be recast, transformed, or adapted; "**Institution**" means the institution, listed on the last page of this Agreement, by which the Author was employed at the time of the creation of the Materials; "**JoVE**" means MyJove Corporation, a Massachusetts corporation and the publisher of The Journal of Visualized Experiments; "**Materials**" means the Article and / or the Video; "**Parties**" means the Author and JoVE; "**Video**" means any video(s) made by the Author, alone or in conjunction with any other parties, or by JoVE or its affiliates or agents, individually or in collaboration with the Author or any other parties, incorporating all or any portion of the Article, and in which the Author may or may not appear.

2. **Background.** The Author, who is the author of the Article, in order to ensure the dissemination and protection of the Article, desires to have the JoVE publish the Article and create and transmit videos based on the Article. In furtherance of such goals, the Parties desire to memorialize in this Agreement the respective rights of each Party in and to the Article and the Video.

3. **Grant of Rights in Article.** In consideration of JoVE agreeing to publish the Article, the Author hereby grants to JoVE, subject to **Sections 4** and **7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Article in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Article into other languages, create adaptations, summaries or extracts of the Article or other Derivative Works (including, without limitation, the Video) or Collective Works based on all or any portion of the Article and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and(c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. If the "Open Access" box has been checked in **Item 1** above, JoVE and the Author hereby grant to the public all such rights in the Article as provided in, but subject to all limitations and requirements set forth in, the CRC License.

4. **Retention of Rights in Article.** Notwithstanding the exclusive license granted to JoVE in **Section 3** above, the Author shall, with respect to the Article, retain the non-exclusive right to use all or part of the Article for the non-commercial purpose of giving lectures, presentations or teaching classes, and to post a copy of the Article on the Institution's website or the Author's personal website, in each case provided that a link to the Article on the JoVE website is provided and notice of JoVE's copyright in the Article is included. All non-copyright intellectual property rights in and to the Article, such as patent rights, shall remain with the Author.

5. **Grant of Rights in Video – Standard Access.** This **Section 5** applies if the "Standard Access" box has been checked in **Item 1** above or if no box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby acknowledges and agrees that, Subject to **Section 7** below, JoVE is and shall be the sole and exclusive owner of all rights of any nature, including, without limitation, all copyrights, in and to the Video. To the extent that, by law, the Author is deemed, now or at any time in the future, to have any rights of any nature in or to the Video, the Author hereby disclaims all such rights and transfers all such rights to JoVE.

6. **Grant of Rights in Video – Open Access.** This **Section 6** applies only if the "Open Access" box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby grants to JoVE, subject to **Section 7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Video in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Video into other languages, create adaptations, summaries or extracts of the Video or other Derivative Works or Collective Works based on all or any portion of the Video and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. For any Video to which this **Section 6** is applicable, JoVE and the Author hereby grant to the public all such rights in the Video as provided in, but subject to all limitations and requirements set forth in, the CRC License.

7. **Government Employees.** If the Author is a United States government employee and the Article was prepared in the course of his or her duties as a United States government employee, as indicated in **Item 2** above, and any of the licenses or grants granted by the Author hereunder exceed the scope of the 17 U.S.C. 403, then the rights granted hereunder shall be limited to the maximum rights permitted under such statute. In such case, all provisions contained herein that are not in conflict with such statute shall remain in full force and effect, and all provisions contained herein that do so conflict shall be deemed to be amended so as to provide to JoVE the maximum rights permissible within such statute.

8. **Protection of the Work.** The Author(s) authorize JoVE to take steps in the Author(s) name and on their behalf if JoVE believes some third party could be infringing or might infringe the copyright of either the Author's Article and/or Video.

9. **Likeness, Privacy, Personality.** The Author hereby grants JoVE the right to use the Author's name, voice, likeness, picture, photograph, image, biography and performance in any way, commercial or otherwise, in connection with the Materials and the sale, promotion and distribution thereof. The Author hereby waives any and all rights he or she may have, relating to his or her appearance in the Video or otherwise relating to the Materials, under all applicable privacy, likeness, personality or similar laws.

10. **Author Warranties.** The Author represents and warrants that the Article is original, that it has not been published, that the copyright interest is owned by the Author (or, if more than one author is listed at the beginning of this Agreement, by such authors collectively) and has not been assigned, licensed, or otherwise transferred to any other party. The Author represents and warrants that the author(s) listed at the top of this Agreement are the only authors of the Materials. If more than one author is listed at the top of this Agreement and if any such author has not entered into a separate Article and Video License Agreement with JoVE relating to the Materials, the Author represents and warrants that the Author has been authorized by each of the other such authors to execute this Agreement on his or her behalf and to bind him or her with respect to the terms of this Agreement as if each of them had been a party hereto as an Author. The Author warrants that the use, reproduction, distribution, public or private performance or display, and/or modification of all or any portion of the Materials does not and will not violate, infringe and/or misappropriate the patent, trademark, intellectual property or other rights of any third party. The Author represents and warrants that it has and will continue to comply with all government, institutional and other regulations, including, without limitation all institutional, laboratory, hospital, ethical, human and animal treatment, privacy, and all other rules, regulations, laws, procedures or guidelines, applicable to the Materials, and that all research involving human and animal subjects has been approved by the Author's relevant institutional review board.

11. **JoVE Discretion.** If the Author requests the assistance of JoVE in producing the Video in the Author's facility, the Author shall ensure that the presence of JoVE employees, agents or independent contractors is in accordance with the relevant regulations of the Author's institution. If more than one author is listed at the beginning of this Agreement, JoVE may, in its sole

1 Alewife Center #200
Cambridge, MA 02140
tel. 617.945.9051
www.jove.com

ARTICLE AND VIDEO LICENSE AGREEMENT

discretion, elect not take any action with respect to the Article until such time as it has received complete, executed Article and Video License Agreements from each such author. JoVE reserves the right, in its absolute and sole discretion and without giving any reason therefore, to accept or decline any work submitted to JoVE. JoVE and its employees, agents and independent contractors shall have full, unfettered access to the facilities of the Author or of the Author's institution as necessary to make the Video, whether actually published or not. JoVE has sole discretion as to the method of making and publishing the Materials, including, without limitation, to all decisions regarding editing, lighting, filming, timing of publication, if any, length, quality, content and the like.

12. **Indemnification.** The Author agrees to indemnify JoVE and/or its successors and assigns from and against any and all claims, costs, and expenses, including attorney's fees, arising out of any breach of any warranty or other representations contained herein. The Author further agrees to indemnify and hold harmless JoVE from and against any and all claims, costs, and expenses, including attorney's fees, resulting from the breach by the Author of any representation or warranty contained herein or from allegations or instances of violation of intellectual property rights, damage to the Author's or the Author's institution's facilities, fraud, libel, defamation, research, equipment, experiments, property damage, personal injury, violations of institutional, laboratory, hospital, ethical, human and animal treatment, privacy or other rules, regulations, laws, procedures or guidelines, liabilities and other losses or damages related in any way to the submission of work to JoVE, making of videos by JoVE, or publication in JoVE or elsewhere by JoVE. The Author shall be responsible for, and shall hold JoVE harmless from, damages caused by lack of sterilization, lack of cleanliness or by contamination due to the making of a video by JoVE its employees, agents or independent contractors. All sterilization, cleanliness or decontamination procedures shall be solely the responsibility of the Author and shall be undertaken at the Author's expense. All indemnifications provided herein shall include JoVE's attorney's fees and costs related to said losses or damages. Such indemnification and holding harmless shall include such losses or damages incurred by, or in connection with, acts or omissions of JoVE, its employees, agents or independent contractors.

13. **Fees.** To cover the cost incurred for publication, JoVE must receive payment before production and publication of the Materials. Payment is due in 21 days of invoice. Should the Materials not be published due to an editorial or production decision, these funds will be returned to the Author. Withdrawal by the Author of any submitted Materials after final peer review approval will result in a US$1,200 fee to cover pre-production expenses incurred by JoVE. If payment is not received by the completion of filming, production and publication of the Materials will be suspended until payment is received.

14. **Transfer, Governing Law.** This Agreement may be assigned by JoVE and shall inure to the benefits of any of JoVE's successors and assignees. This Agreement shall be governed and construed by the internal laws of the Commonwealth of Massachusetts without giving effect to any conflict of law provision thereunder. This Agreement may be executed in counterparts, each of which shall be deemed an original, but all of which together shall be deemed to me one and the same agreement. A signed copy of this Agreement delivered by facsimile, e-mail or other means of electronic transmission shall be deemed to have the same legal effect as delivery of an original signed copy of this Agreement.

A signed copy of this document must be sent with all new submissions. Only one Agreement is required per submission.

CORRESPONDING AUTHOR

Name:	Shirley Pepke
Department:	Computational Biology
Institution:	Lyrid LLC
Title:	Principal Scientist
Signature:	
Date:	June 14, 2019

Please submit a **signed** and **dated** copy of this license by one of the following three methods:

1. Upload an electronic version on the JoVE submission site
2. Fax the document to +1.866.381.2236
3. Mail the document to JoVE / Attn: JoVE Editorial / 1 Alewife Center #200 / Cambridge, MA 02140

Lyrid, LLC
Computational Genomics, Systems Biology
and Complex Disease Research

Philip Steindel, Ph.D.
Review Editor
JoVE

Dear Dr. Steindel,

My co-authors and I have revised our manuscript, JoVE60431 "Analyzing tumor gene expression factors with the CorExplorer web portal," in response to the editorial and peer reviews. A detailed list of changes to address specific points is given below.

We greatly appreciate the constructive feedback and believe the concomitant changes have improved the manuscript significantly. We look forward to working with the JoVE staff on the next steps for publication and video presentation.

Best regards,

Shirley Pepke, Ph.D.
Principal Scientist
Lyrid, LLC

Editorial comments:
General:
1. We have more thoroughly proofread the completed manuscript for spelling and grammar errors.
2. William M. Nelson performed the work as an independent consultant (perhaps more common in bioinformatics than other fields?).
3. Abbreviations are now defined at first use.

Protocol:
1. We have modified the protocol to ensure it explicitly addresses "how" each step is performed. The individual step instructions have been limited to 2-3 actions and 4 sentences.

Discussion:
1. We have rewritten the discussion sections with the following points in mind.
a) Critical steps within the protocol - This is addressed in the 3rd paragraph. Generally, using clinical correlation (survival) and included biological knowledge from databases to filter the myriad factors for those most therapeutically relevant are particularly significant steps to make best use of the portal information.
b) Any modifications and troubleshooting of the technique - We have addressed

potential reasons for factors that lack annotation enrichments in paragraph 4 and how they might be investigated further for biological significance.
c) The significance with respect to existing methods - This is addressed in the paragraph 2, where we have now cited several related works and state how the CorExplorer differs.

References
1. The references have been uniformly formatted as requested.

Table of Materials:
1. The Table of Materials now lists the web browsers that the site was tested on and also the details of the server configuration.

Reviewers' comments:
Reviewer #1:
Minor Concerns:
I can think of two minor issues:
- Introduction section is brief and may miss important studies in the field. There are no mentions of other linear, nonlinear and hierarchical factor learning in cancer. So, the third paragraph of introduction needs to be expanded.

We have cited additional other works in the second paragraph of the Discussion section. We feel it is appropriate to restrict the comparison to online tools for exploration of RNA-seq data based upon machine learning techniques, since the manuscript is primarily about the CorExplorer interactive server.

- It is customary to specify the configuration of the hosted server in the manuscript for performance concerns.

This is now listed in the Table of Materials.

- Factors related to specific genes can be searched, and they are highlighted and can be clicked on, but that factor is not highlighted in the new window. So, it is better, if possible, to highlight the factor in the new window

This should now be fixed.

Reviewer #2:

- Authors should extend Section "Introduction" by taking into account other machine learning algorithms, well known gene expression analysis methods, and alternative portals dedicated to cancer studies.

Paragraph 2 of the Discussion now cites some of the suggested references as well as a few others.

- Section "PROTOCOL" is too technical and reflects more a user manual. Authors should describe the procedure in a more readable way.

We have attempted to tweak the protocol language so that it reads a bit more like a narrative. For instance, many connector phrases and words have been added to smooth out the presentation. It is also more hierarchically structured for readability. For instance, the text of the Protocol steps in sections 1 and 2 reflect the organization level of the steps, thus are more pedagogical.

- The manuscript presents many figures, several in section "Discussion". The figures must be better integrated in the text and comprehensively described. Figure 1 must be improved and the graph must be reduced so that it does not overlap to the text.
In general, I suggest to improve the presentation of the web portal by writing the manuscript in a more clear way.

Five of the six figures are first referenced in the steps of the protocol. We have elaborated on the relevance to the specific steps. The figure captions have been expanded to be more explanatory. Figure 1 has been redone so there is no overlap.