





CAP-RNAseq: an integrated pipeline for functional annotation and prioritization of co-expression clusters

Merve Vural-Ozdeniz , Kubra Calisir , Rana Acar , Aysenur Yavuz, Mustafa M. Ozgur, Ertugrul Dalgic and Ozlen Konu 

Corresponding author: Ozlen Konu, Department of Molecular Biology and Genetics, Bilkent University, 06800, Ankara, Türkiye. E-mail: konu@fen.bilkent.edu.tr

Abstract

Cluster analysis is one of the most widely used exploratory methods for visualization and grouping of gene expression patterns across multiple samples or treatment groups. Although several existing online tools can annotate clusters with functional terms, there is no all-in-one webserver to effectively prioritize genes/clusters using gene essentiality as well as congruency of mRNA-protein expression. Hence, we developed CAP-RNAseq that makes possible (1) upload and clustering of bulk RNA-seq data followed by identification, annotation and network visualization of all or selected clusters; and (2) prioritization using DepMap gene essentiality and/or dependency scores as well as the degree of correlation between mRNA and protein levels of genes within an expression cluster. In addition, CAP-RNAseq has an integrated primer design tool for the prioritized genes. Herein, we showed using comparisons with the existing tools and multiple case studies that CAP-RNAseq can uniquely aid in the discovery of co-expression clusters enriched with essential genes and prioritization of novel biomarker genes that exhibit high correlations between their mRNA and protein expression levels. CAP-RNAseq is applicable to RNA-seq data from different contexts including cancer and available at <http://konulabapps.bilkent.edu.tr:3838/CAPRNAseq/> and the docker image is downloadable from <https://hub.docker.com/r/konulab/caprnaseq>.

Keywords: RNA-Seq; clustering; prioritization; essential genes; networks; annotation

INTRODUCTION

The recent use of RNA sequencing (RNA-seq) technology has generated massive amounts of data that can be re-analyzed for biomarker discovery [1]. Clustering genes with similar expression patterns [2] has also enabled distinguishing groups with differential expression in cancer [3], across different brain regions [4] or in response to environmental stressors [5]. Functional annotation of clusters can further provide mechanistic leads about the treatment effects via integration of enrichment analyses of biological/molecular terms [6, 7] or protein-protein interactions [8–10].

In the literature, many comprehensive RNA-seq analysis tools exist, such as GENAVi [11], iDEP [12], DEBrowser [13], RNfuzzy [14], ToppGene Suite [15], BEAVR [16], WebMeV [17], Omics Playground [18], 3D RNA-seq [19], FungiExpressZ [20], Clust [21] and DEGUST [22]. Many of these allow for uploading and clustering RNA-seq

data, yet only a few provide visualizations or enrichment analyses that are cluster specific [14, 18, 20, 21]. Moreover, techniques like self-organizing maps are available to help order patterns of co-expression clusters according to their similarities [23], yet identification of pairs of clusters with inverse expression patterns (aka. mirror clusters) has received less attention [24, 25].

Since mRNA and protein levels are only moderately correlated [26, 27], the use of comprehensive databases, such as Human Protein Atlas (HPA) [28] and DepMap [29, 30], which contain cell/tissue- and/or cancer-specific protein expression data, could be incorporated in the RNA-seq cluster/gene annotation pipelines. DepMap also provides information on gene essentiality from CRISPR or RNAi screens, e.g. shinyDepMap [31], potentially further enhancing the co-expression cluster annotation. However, these databases have not yet been incorporated in the existing RNA-seq tools.

Merve Vural-Ozdeniz is a Ph.D. candidate in the Department of Neuroscience, Bilkent University, Ankara, Türkiye. Her research interests include the development of bioinformatics pipelines.

Kubra Calisir graduated with an MSc degree from the Department of Molecular Biology and Genetics, Bilkent University, Ankara, Türkiye. She is currently pursuing a Ph.D. degree in Department of Biochemistry and Molecular Biology, Hollings Cancer Center, Medical University of South Carolina, Charleston, SC, USA. Her research interests include cancer biology and immunology.

Rana Acar is an MSc student majoring in Molecular Biology and Genetics at the Department of Molecular Biology and Genetics, Bilkent University, Ankara, Türkiye, with research interests in transcriptomics analyses and drug screening.

Aysenur Yavuz graduated with a BSc degree from the Department of Molecular Biology and Genetics, Bilkent University, Ankara, Türkiye. She is currently a MSc. student in the Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium.

Mustafa Mert Ozgur graduated with a BSc degree from the Department of Molecular Biology and Genetics, Bilkent University, Ankara, Türkiye.

Ertugrul Dalgic is an assistant professor in the Department of Medical Biology, School of Medicine, Zonguldak Bülent Ecevit University, Zonguldak, Türkiye, with research interests in medical biology and bioinformatics.

Ozlen Konu is an associate professor in the Department of Molecular Biology and Genetics, Bilkent University, Ankara, Türkiye. Her research interests include human disease models and comparative transcriptomics analyses.

Received: October 2, 2023. **Revised:** December 4, 2023. **Accepted:** December 21, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

For commercial re-use, please contact journals.permissions@oup.com

Moreover, it is essential to validate the identified biomarkers through wet-laboratory experiments. The predominant method for assessing biomarker expression is quantitative polymerase chain reaction (qPCR), which necessitates the design of specific primers. Despite the availability of numerous primer design applications [32–34], the seamless integration of a primer design tool into an online RNA-seq analysis pipeline is still lacking.

To fill the abovementioned gaps, we developed an online tool called CAP-RNAseq, which stands for Cluster, Annotate and Prioritize-RNA-seq data, that allows the user to (1) upload inhouse or public bulk RNA-seq gene-level raw count data to filter, and then generate, visualize and annotate co-expression clusters; (2) prioritize clusters/genes based on tissue/cell-specific gene essentiality scores and mRNA-protein expression correlation data; and (3) generate primers for prioritized genes. CAP-RNAseq has been developed as an R Shiny [35, 36] application and can be used within a diverse array of contexts, including cancer biology.

Methods and demonstration of the CAP-RNAseq pipeline

CAP-RNAseq is an online application that requires the upload of gene-level raw bulk RNA-Seq count data, in which each row is uniquely identified by a human gene name, along with a text file that contains the phenotype/condition labels corresponding to each sample using the Dataset tab (Figure 1). In the Clustering tab, hierarchical and k-means clustering are available, as well as a heatmap of dissimilarities among clusters (mirror clusters) (Figure 1). Once co-expression clusters are obtained, three other main tabs, namely, DGEA-GSEA, Cluster Prioritization, and Gene Prioritization, are available to visualize and annotate selected clusters (Figure 1). In the following sections, each tab is explained in more detail and by using a demo dataset.

Demo datasets integrated into CAP-RNAseq

Examples of these input files and how to upload and visualize them can be found on the online tutorial page of CAP-RNAseq. Data can be filtered by CPM thresholds before performing an ANOVA on raw count data upon applying a variance stabilization transformation (vst) by the *DESeq2* package [37]. The demo datasets obtained from GEO [38] were generated using Illumina short-read NGS technology and each had more than two groups. They were focused on understanding the effects of (1) overexpression of two variants of NTRK2 gene (TrkB.FL and TrkB.T1) or GFP (control) in human neuronal stem cells [39] (GSE136868; Demo 1), (2) silencing of NTRK2 gene in senescent fibroblasts [40] (GSE190998; Demo 2) and (3) therapy on blood samples from a breast cancer patient cohort with clinical data [41] (GSE201085; Demo 3). We filtered the Demo 1 dataset by vst + ANOVA (P -value <0.05) to reduce the gene number to 2407 before integration into the app and used it to demonstrate the pipeline (Figure 1; online tutorial).

Platforms integrated into CAP-RNAseq for prioritization

CAP-RNAseq integrates various datasets from platforms that help with cluster/gene annotation and/or prioritization (Figure 1). DepMap, which is from the Cancer Dependency Map (CDP) housing genetic dependencies of cancer cells [28, 29] based on large-scale siRNA and CRISPR-based functional screens, has been integrated using the *depmap* package [42] in R. It provides access

to TPM (transcripts per million) data on 19177 genes, 1393 cell lines, encompassing 33 primary diseases, and 38 tissues/lineages from the 22Q1 CCLE (Cancer Cell Line Encyclopedia) [43]; and proteomics data (20Q2 quantitative protein profiling via mass spectrometry from the Gygi lab [44]) on 12399 proteins from 375 cell lines, 24 diseases and 27 lineages. CAP-RNAseq calculates the Pearson's correlation coefficient between mRNA and protein expression for each gene across selected samples along with a linear regression coefficient, R-squared and adjusted R-squared values together with the slope, intercept and their associated P -values.

Project Score database, a component of the CDP [45], where the target priority score for a gene is based on CRISPR-Cas9 experimental evidence, target dependency and the frequency of somatic changes in patient tumors, has been integrated into CAP-RNAseq [30]. The CDP contains 2879 protein-coding genes in 15 different cancer types with therapeutic target scores ranging from 0 to 100. CAP-RNAseq calculates the number of genes having priority scores in each cluster and also incorporates the dependency scores calculated by the *shinyDepMap* tool [31] for 15847 genes from 423 cell lines, reporting two parameters: efficacy and selectivity, respectively measuring the effect of gene loss on cell growth in a sensitive cell line, and the variation in gene essentiality between sensitive and resistant cell lines.

CAP-RNAseq also contains HPA [28], which independently provides information on protein expression levels for different genes across different cancer types as well as tissues and cell types. We used the *HPAanalyze* package [46] to obtain and visualize these protein expression data (categorized as 'high', 'medium', 'low' or 'not detected') on 20082 proteins, 20 cancer types and 63 normal tissues.

Clustering types and identification of mirror clusters

CAP-RNAseq performs hierarchical clustering of expression profiles using scaled logCPM values by 'hclust' function of R and allows user-selection of linkage criteria and distance methods to help determine an optimal number of clusters in addition to generating silhouette graphs [47] (Figure 1; online tutorial). CAP-RNAseq employs a k-means algorithm [48] to assign genes to the selected number of cluster centroids based on the logCPM values using 'kmeans' function of R. The scaled data are visualized in line or boxplot graphs. The user can vary the number of clusters, maximum number of iterations and number of random starting partitions, and view the number of differentially expressed genes within each cluster for a selected pair of treatments (Figure 1). Using the 2407 differentially expressed genes in the Demo 1 dataset, we performed hierarchical clustering and then set k to eight (Supplementary Figure 1A and Figure 2A), and generated clusters using a maximum of 2000000 iterations and a random set of 20. We found that clusters 1 and 2 had more genes with high priority scores in the central nervous system (Supplementary Figure 1B). There were more differentially expressed genes between the TrkB.FL and TrkB.T1 groups in clusters 4, 7 and 8 (Supplementary Figure 1C).

The dissimilarity heatmap of Pearson's correlation coefficients among cluster centroids enables users to identify clusters whose expression patterns are the most inversely correlated, also known as *mirror clusters*. Based on the lowest correlation scores, a table is generated to display these mirror clusters (online tutorial). In the Demo 1 dataset, clusters 1 and 2 exhibited incrementally and

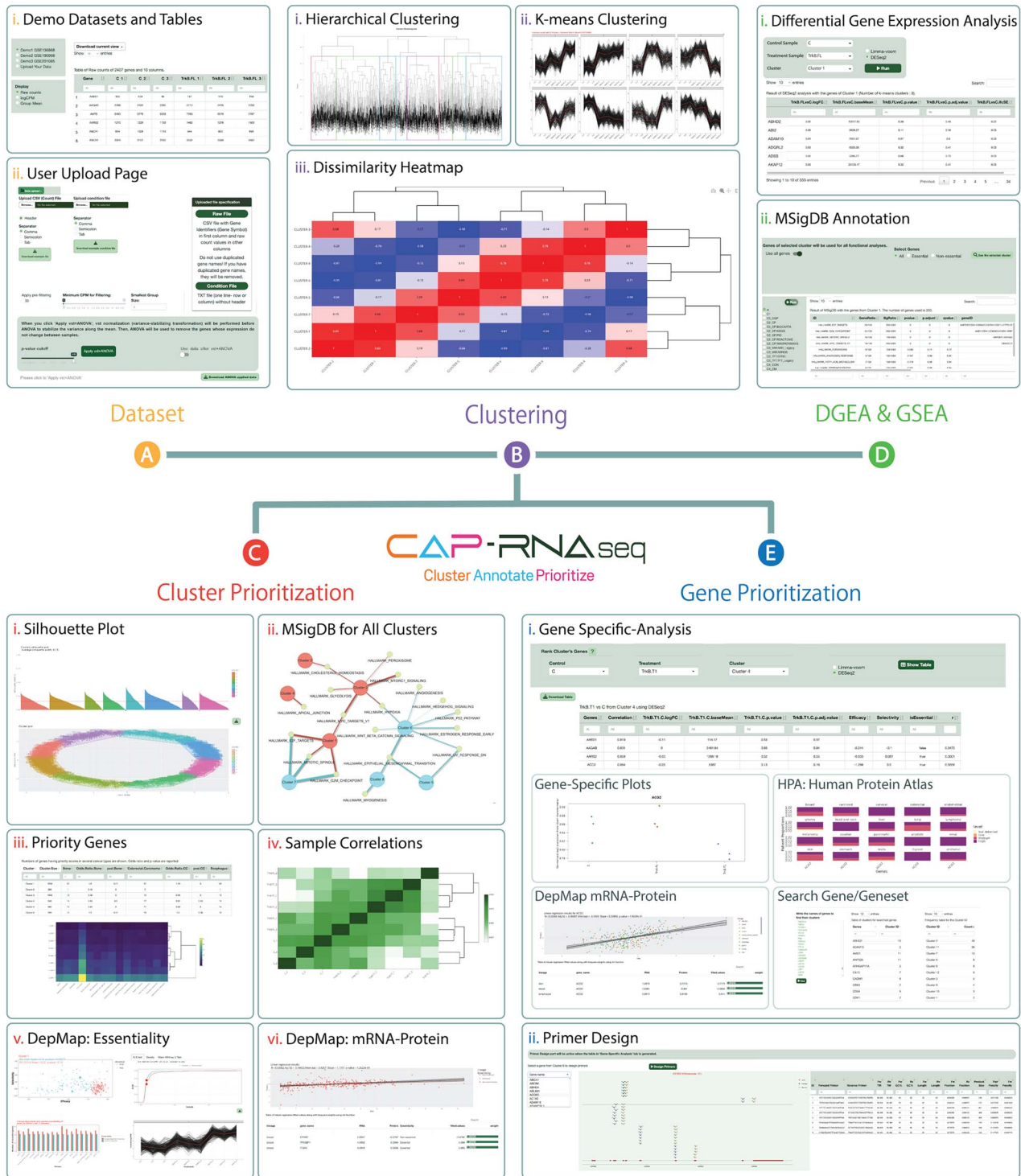


Figure 1. The workflow of CAP-RNAseq.

steady patterns of increase in expression in the groups overexpressing NTRK2 full-length and NTRK2 isoform clones, respectively (Figure 2A). On the other hand, clusters 8 and 6 exhibited a mirror image of clusters 1 and 2. In cluster 3, the increase by the full-length protein was higher than that by the isoform; and the mirror image of this expression pattern was found in cluster 5. We identified four pairs of mirror clusters that could also be deduced from the dissimilarity heatmap, marked with yellow rectangles (Figure 2B).

Cluster prioritization

To prioritize a cluster, multiple paths can be taken in CAP-RNAseq, e.g. Silhouette plots, MSigDB analysis, enrichment of priority scores or analyses based on visualization of DepMap.

Silhouette plots

CAP-RNAseq provides the user with the ability to study data consistency within a cluster using silhouette plots ('silhouette'

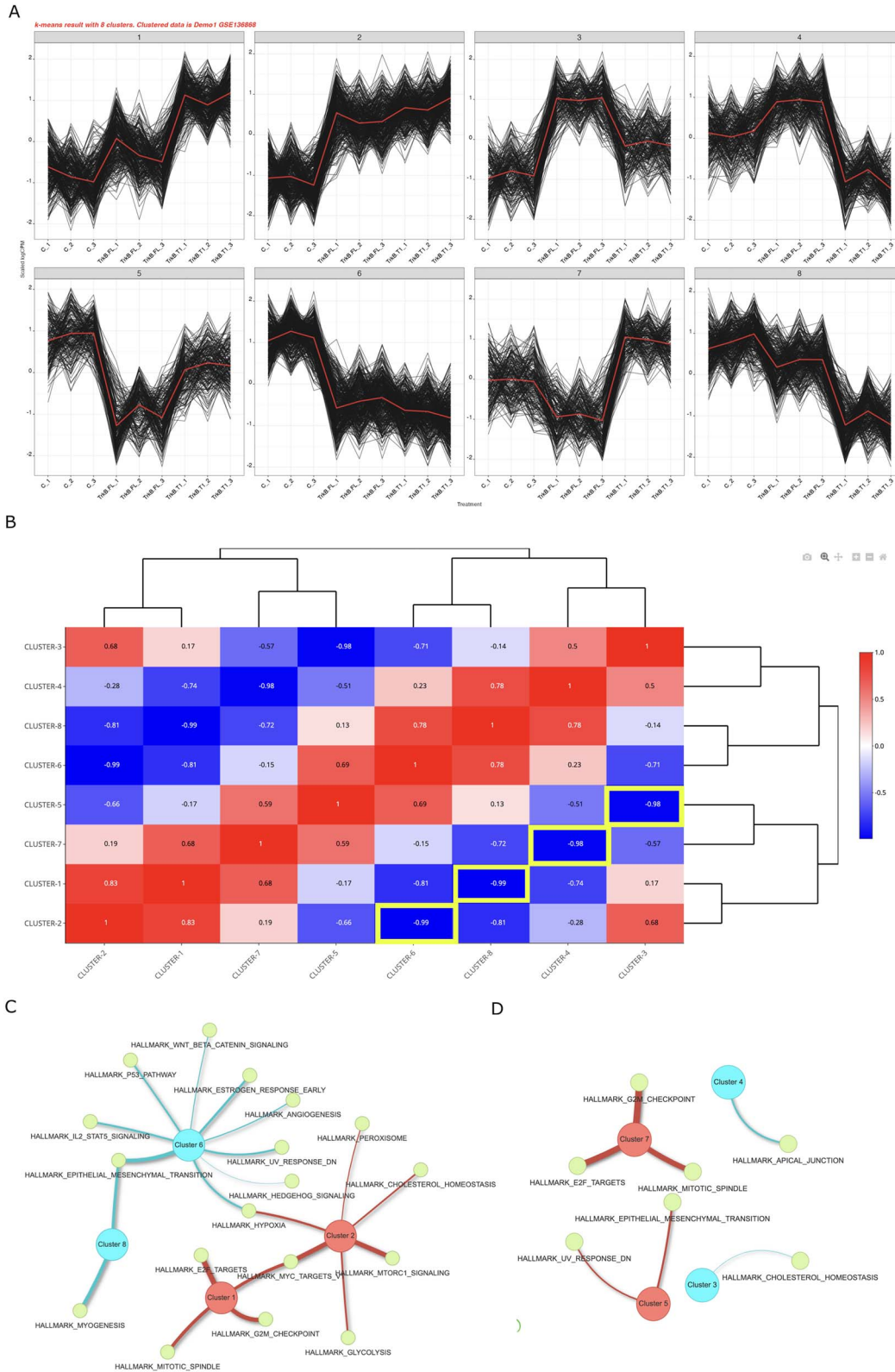


Figure 2. Clustering analysis of Demo 1 data. **(A)** *K*-means clustering result with 8 clusters where red lines highlight the centroids of clusters. **(B)** Dissimilarity heatmap where negative values show the lowest similarities which indicate the pairs of mirror clusters, while the positive values indicate most similar clusters. Rectangles were drawn to highlight mirror clusters. **(C)** Network visualization of MSigDB terms for clusters 1, 2 and their mirror clusters 6 and 8. **(D)** Network visualization of MSigDB terms for clusters 3, 4, 5 and 7. Edge widths show the number of genes overlapping with each hallmark collection term.

function of *cluster* package) [49] by displaying silhouette width scores in a table. *ggplot2*-based visualizations of silhouette widths are presented using *factoextra* package [50]. As a result, CAP-RNAseq identifies which clusters contain genes that are likely to be placed in the wrong cluster; and ranks the clusters based on their consistency scores (Figure 1; online tutorial).

Functional enrichment by MSigDB and networks of cluster terms

CAP-RNAseq uses the Molecular Signatures Database (MSigDB), one of the larger repositories of gene sets [51, 52], which can be run for all clusters at once using the genes filtered based on their essentiality (i.e. essential or non-essential) and their mRNA ~ protein correlation coefficients calculated from DepMap data. MSigDB sets were retrieved using the *msigdb* R package [53]. The user needs to select a gene set collection(s), e.g. the default selection is set to H, for Hallmark, before CAP-RNAseq performs an enrichment analysis for each cluster one by one (Figure 1; online tutorial).

These enrichment results are visualized as a network using *visnetwork* R package [54], where nodes represent either clusters or MSigDB terms, and the edge widths represent the number of genes overlapping between clusters. In networks, the user can set the *q*-value as well as the number of enriched terms to be displayed. The cluster nodes can be colored based on group comparisons (i.e. an increase (red) or a decrease (blue) in expression level) using the scaled and averaged logCPM values used in *k*-means clustering. Using the Demo 1 dataset, we used the ‘Hallmarks’ collection from MSigDB; a network of clusters was created for the top 10 terms, each having a *q*-value <0.05. The cluster nodes were colored based on the difference in logCPM values between the TrkB.T1 isoform (treatment) and full-length TrkB.FL (control) groups. Genes in cluster 1 and cluster 2, which exhibited increases by the overexpression of TrkB.T1 or TrkB.FL, were enriched in E2F and MYC targets, G2M checkpoints, mitotic spindles, glycolysis, MTORC1 signaling and hypoxia (Figure 2C). Their mirror clusters, cluster 6 and cluster 8, on the other hand, were enriched in the P53 pathway, myogenesis, hedgehog signaling and hypoxia. The remaining clusters did not share terms to the same extent with each other (Figure 2D).

Moreover, one or more clusters can be prioritized based on the enrichment of CDP target priority scores of genes in a cancer type(s) out of 15 in total (Figure 1; online tutorial) using a table that provides statistics based on the ‘oddsratio’ function from the *epitools* package [55] and helps prioritize a cluster with a higher association score in a cancer type(s) of interest. Using CAP-RNAseq, the Pearson’s correlation coefficients among samples within a chosen cluster are also visualized as an interactive heatmap with the *heatmaply* package [56] (Figure 1; online tutorial).

Prioritization by essentiality and efficacy/selectivity statistics of a cluster

CAP-RNAseq can display in a table the mean (+/– std) efficacy as well as selectivity scores from DepMap along with the total number and percentage of essential genes observed in each cluster and the associated odds ratios from Fisher’s exact tests. The efficacy and selectivity scores are shown in a scatter plot, where observations are colored with blue, to indicate essentiality; or red, for non-essentiality. This plot can be redrawn using other options, such as coefficient of variation (CV) on CPM, log₂ transformed average CPM or log₂ of range CPM values. Moreover, density distributions, Kolmogorov–Smirnov test for comparing densities

[57] and the Mann–Whitney U-test to compare medians [58, 59] are made available (Figure 1).

For the Demo 1 dataset, the mean and standard deviation values of efficacy and selectivity for each gene in a cluster along with Fisher’s exact test enrichment statistics were obtained (Supplementary Table 1) and cluster 1 consisted of significantly more essential genes than random, while cluster 6 was significantly depleted of essential genes. In cluster 6 (Figure 3A), highly down-regulated genes (higher CV) had low to moderate expression values and were likely to be non-essential (Figure 3B). The difference in the median CVs of essential and non-essential genes in cluster 6 was significant (Figure 3C). Furthermore, cluster 6’s mirror, i.e. cluster 2, had significantly high number of essential genes whose expressions were upregulated by both TrkB.T1 and TrkB.FL groups when compared to control samples (Figure 3D). Interestingly, the majority of these essential genes showed high levels of expression yet low CVs (Figure 3E–F).

Cluster-specific differential gene expression analysis and functional annotations

The user can further perform differential gene expression analysis (DGEA) for a pair of conditions on a selected cluster’s genes using DESeq2 [37] or limma-voom [60]. This action provides the names of genes significantly modulated by a treatment within the selected cluster and relevant statistical values, such as log₂ fold change (logFC), *P*-value and adjusted *P*-value for each gene’s expression.

Moreover, the user can perform functional analysis through MSigDB in a cluster-specific manner (Figure 1; online tutorial).

Gene prioritization

A section of CAP-RNAseq has been allocated to the prioritization of genes within a selected cluster and several options were made available as explained below.

Distance correlation with the cluster’s centroid

A distance correlation measure is calculated between the expression profile of each gene within a chosen cluster and that of its cluster centroid using ‘dcor’ function of *energy* package of R [61] before ranking the genes (online tutorial). This step allows for selecting a gene that can be the best representative of the selected cluster’s average expression profile. In the Demo 1 dataset, the genes modulated by NTRK2 overexpression were ranked based on different attributes (Supplementary Table 2), e.g. cluster 6 genes using (1) correlation with the cluster centroid, (2) the significance of logFC and/or (3) correlation between mRNA and protein levels. Accordingly, we prioritized the CCND2, an essential gene with a high efficacy score, with decreased expression in the overexpressed groups when compared to the control group. The selectivity of CCND2 was found to be 0.68, which was higher than the average selectivity score of cluster 6.

Protein expression levels in HPA across diseases and tissues

The HPA tab incorporates the ‘pathology’ and ‘normal tissue’ datasets of HPA in which protein expression was determined through IHC staining. CAP-RNAseq displays the protein expression level of a selected gene when the user enters the names of cancer and normal tissues from a selection menu using the stacked bar plots (Figure 1). The protein expression level of CCND2 in the HPA database was low except the thyroid, carcinoid and

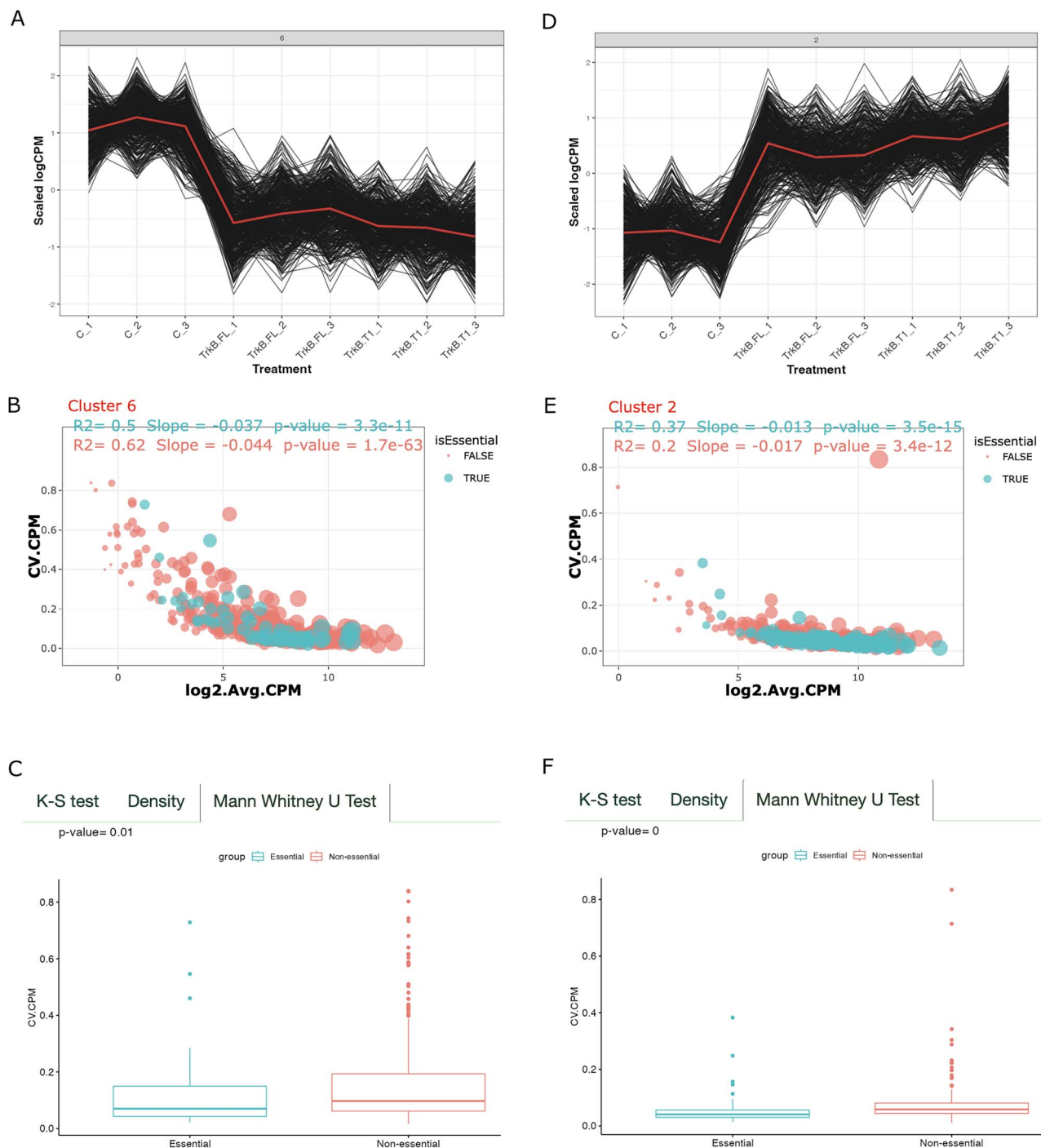


Figure 3. Analysis of cluster 6 (A–C) and cluster 2 (D–F). (A, B) Cluster visualization of each cluster as a result of k-means clustering. (B, E) Graphs showing the log₂ of Avg CPM in x-axis versus CV of CPM in y-axis of genes in clusters. Point size shows the log₂ of range CPM values. (C, F) Graphs showing the median CV for both essential and non-essential genes based on Mann–Whitney U-test. Avg: average, CV: coefficient of variation, CPM: counts per million.

cervical cancers, and not found at high levels in normal brain tissues (Figure 4A and B).

Correlation between mRNA and protein levels based on DepMap data

Using DepMap [62, 63], CAP-RNAseq can plot TPM values against protein z-scores, performs robust regression with bisquare weighting [64], assigns weights to find potential outlier samples [65], and estimates the intercept and slope values before the

selected gene is used for the primer design module in CAP-RNAseq (Figure 1; online tutorial). In the Demo 1 dataset, DepMap mRNA versus protein correlation plot across all lineages showed a significant correlation for CCND2 (Figure 4C), which is a cell cycle gene, associated with ERBB2-negative tumors and poor differentiation [66, 67]; and aberrantly expressed in a variety of malignancies [68–70]. Moreover, CCND2 expression is low in normal brain and low-grade gliomas while being significantly upregulated in GBM [71–73].

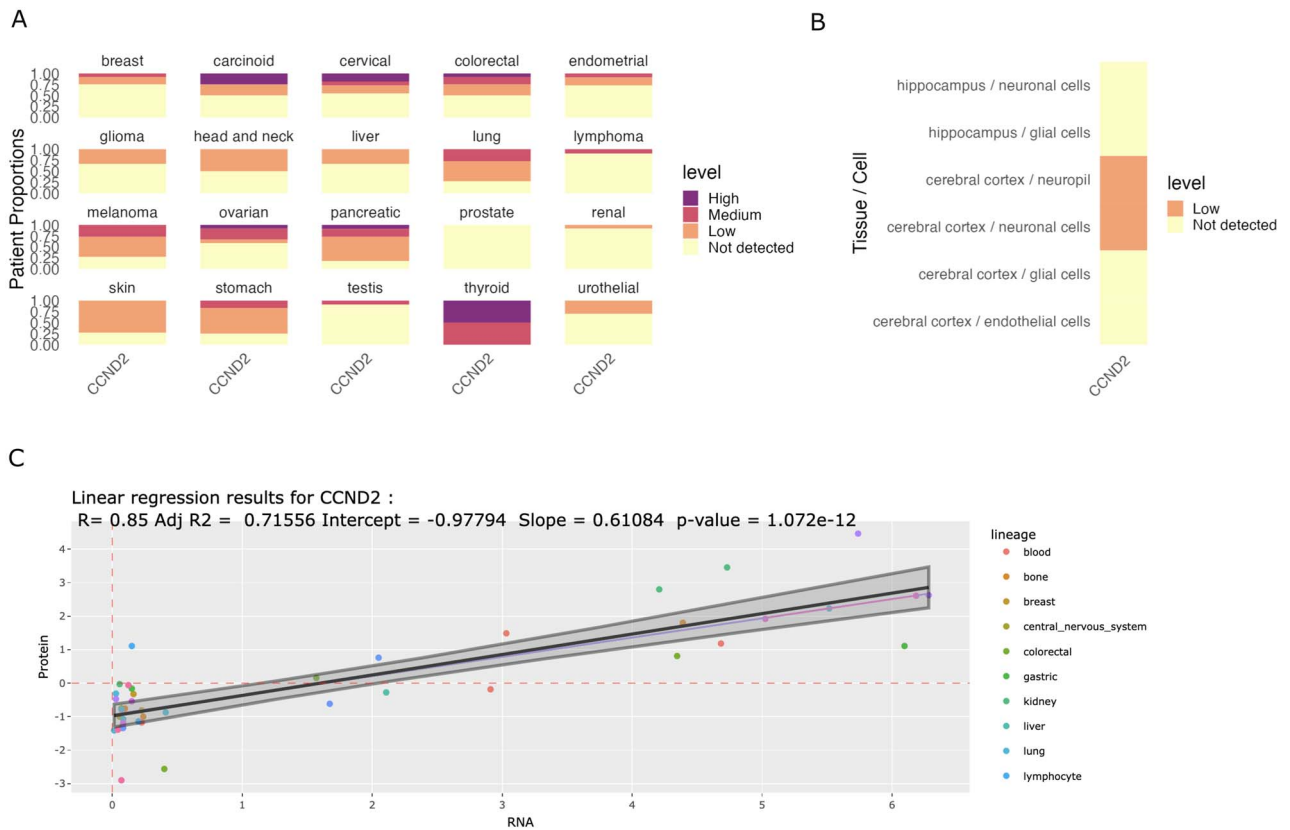


Figure 4. CCND2 gene prioritization revealed in cluster 6. **(A)** Protein expression of CCND2 in many cancers using the HPA dataset. **(B)** Protein expression of CCND2 in normal cells of the hippocampus and cerebral cortex using the HPA dataset. **(C)** mRNA versus protein correlation plot across all lineages using DepMap data.

Primer design

For a selected transcript whose exon-intron structural information is obtained from Biomart [74, 75], CAP-RNAseq (Figure 1) designs primers, flanking an intron within randomly selected two consecutive exons, ten times among all exons of the selected gene [76]. The primer design module of the *scAmpDesign* package [77] utilizing Primer3 [32, 78] is used with specific changes in the source code, which include updates to the latest version of the files retrieved from Primer3, adjustments to the number of returned primer pairs and improvements to the table's architecture. CAP-RNAseq confirms whether the forward and reverse primers are in different exons before presenting the list of primers in table format and plots an interactive ggplot graph showing exons and primer pair locations. Primers for CCND2 are shown with their exon and primer positions in the CCND2 gene (Supplementary Figure 2A). In addition, CAP-RNAseq provides a table consisting of various features of the designed primers, e.g. melting temperature (T_m), GC content values, positions within the gene sequence, and penalty scores (Supplementary Figure 2B).

Comparisons with the existing tools

We have compared CAP-RNAseq in detail with iDEP [12], a highly comprehensive expression analysis tool that contains a set of cluster-specific analysis features (Supplementary Table 3A) and also across a larger set of RNA-seq tools collected to the best of our knowledge (Supplementary Table 3B). We have found that CAP-RNAseq has novel features not present in iDEP or other tools, which are marked with NA/No. For example, the vst + ANOVA [37, 79, 80] filtering feature is unique and reduces

data volume for clustering in a supervised manner as opposed to using an unsupervised approach such as selecting the top variable genes as in iDEP (Supplementary Figure 3). In CAP-RNAseq, the user can filter their data using vst + ANOVA upon uploading any short-read gene-level raw RNA-seq count dataset with multiple groups (e.g. time-series) that are applicable for use with other expression analysis methods such as limma and masigpro [81].

CAP-RNAseq utilizes the k -means algorithm [82], and via hierarchical clustering helps the user visualize and decide on the optimal numbers (k) of clusters. Several online tools perform k -means clustering on normalized RNA-seq data as well [12, 21, 83]. However, they either do not have the ability to test the optimality of the selected number of clusters or do not perform differential gene expression analysis and further gene prioritization on clusters (Supplementary Table 3B).

Moreover, CAP-RNAseq uses line or box plots for visualization of expression patterns, whereas iDEP and WebMeV use heatmaps and PCA, respectively. Although GENAVi can cluster samples against each other for all or a selected set of genes, it does not for the genes of a cluster obtained from k -means method. On the other hand, CAP-RNAseq can do so and in addition identifies mirror clusters based on comparing cluster centroids, a prominent feature of CAP-RNAseq, that is not present in any other application.

The enrichment analysis of a cluster is performed by only a few tools (iDEP, RNFuzzy, Omics Playground and FungiExpresZ), but CAP-RNAseq also provides a bipartite network to visualize the shared and unique enriched terms among clusters, enhancing iDEP's tree-like display (Supplementary Table 3A). CAP-RNAseq

uniquely integrates DepMap datasets and provides a statistical evaluation of gene essentiality for each co-expression cluster in a dataset with respect to gene efficacy and selectivity for further cluster prioritization.

To prioritize genes, the user can sort the gene table with parameters unique to CAP-RNAseq, such as gene essentiality and the degree of mRNA-protein correlation (Supplementary Table 3A). While HPA is used within RNafuzzy, it is limited to enrichment analyses unlike CAP-RNAseq that provides a detailed visual representation of protein levels in user-selected cancers and tissues. Moreover, CAP-RNAseq's interactive and modular nature makes possible the incorporation of other analysis methods in the future. Overall, CAP-RNAseq not only distinguishes itself from the current RNA-seq applications, but also successfully complements them for the analysis of both custom and publicly available bulk RNA-seq data.

Application of CAP-RNAseq on publicly available RNA-seq datasets

CAP-RNAseq can be applied to any bulk RNA-seq gene-level raw count data, with each row identified by a unique human gene symbol and having two or more groups, but it is not recommended for the analysis of single cell or non-coding RNA-seq data. In the following sections, we further exemplify the potential use of CAP-RNAseq in different contexts: (1) cluster comparisons between complementary datasets (Demo 2 versus Demo 1) and (2) prioritization of biomarkers in clinical cancer patient datasets (Demo 3).

Case study 1: cluster comparisons between complementary datasets

The Demo 2 raw data were obtained from GEO. i.e. GSE190998, in which proliferating WI-38 fibroblast cells were profiled using short-read RNA-seq against senescent WI-38 fibroblasts transfected with a control siRNA or an siRNA against NTRK2 or BDNF [40] (from which we did not use BDNF samples). Filtering with *vst*+ANOVA using a *P*-value threshold of 0.05 resulted in 10151 significant genes, and to be comparable with the Demo 1 dataset, a more stringent *P*-value threshold (i.e. < 0.001) was chosen, resulting in 4412 genes that were assigned to 12 clusters (Supplementary Figure 4). From the table of essential gene counts in each cluster, under the 'DepMap: Essentiality' tab (Supplementary Table 4), cluster 2 that contained the highest percentage of essential genes and cluster 5, highly enriched with non-essential genes, were analyzed further. Cluster 2 showed that essential genes exhibited a higher CV of CPM and log₂ of range CPM (Figure 5A and B) supported by a Mann-Whitney U-test (Figure 5C). However, in cluster 5, where the senescent cells had higher expression levels when compared to proliferating cells (Figure 5D), the non-essential genes had the highest CV of CPM (Figure 5E), also supported by Mann-Whitney U-test (Figure 5F).

The essential genes were upregulated in only one cluster (cluster 10) and were enriched with peroxisome and UV response terms, while the downregulated essential genes were enriched with MYC targets, G2M checkpoints, mitotic spindle and E2F targets and spread out to multiple clusters (Figure 6A). Moreover, clusters enriched with non-essential genes, e.g. downregulated clusters 3, 6 and 11 and upregulated clusters 5 and 8, had the epithelial mesenchymal transition term in common (Figure 6B). These findings suggested that senescence, regardless of presence or absence of NTRK2, resulted in downregulation of a subset of

essential genes with importance in cell cycle and MYC and E2F signaling while non-essential genes were largely upregulated.

In cluster 5 (Supplementary Table 5A), the CCND2 gene ranked high and was significantly upregulated by senescent siControl and even further by siNTRK2 treatment when compared to the proliferation control siRNA. CCND2 exhibited a moderate expression and the highest CV on CPM (0.91). Hence, CAP-RNAseq identified a novel association between NTRK2 and CCND2 such that inhibition of NTRK2 increased and overexpression of NTRK2 decreased the expression of the CCND2 gene, respectively. We mapped the Demo 2 cluster 5 genes on the Demo 1 clusters using the 'Search Gene/Geneset' box of CAP-RNAseq (Figure 1; online tutorial) and found that many of them were found in the Demo 1 clusters 6 and 8 (Supplementary Table 5B; Figure 7A). The patterns of co-expression were reversed between datasets for the searched gene set (Supplementary Figure 4; Figure 7B-C).

Case study 2: prioritization of novel biomarkers in clinical cancer patient datasets

GSE201085 [41], an Illumina short-read RNA-seq dataset from blood samples of breast cancer patients (*n*=53 with four groups: residual disease (*n*=23 with or without recurrence); pathologic complete response (pCR, *n*=9); and those who did not receive NAC (*n*=21)), was re-analyzed for the discovery of novel biomarkers with high correlation coefficients between the mRNA and protein levels. After filtering at *P*-value <0.05, 4846 genes were retained and integrated as Demo 3. For clustering into 8 clusters, we used the group means and for visualization, we selected the boxplot option (Figure 8A). The cluster-term network prioritized three clusters whose genes were upregulated in the pCR group (Figure 8A; Supplementary Figure 5A). pCR group members had a heightened immune response and increased activity in apoptosis, hypoxia, TP53 and NFκB signaling (Supplementary Figure 5A). Moreover, CAP-RNAseq provided candidate biomarkers of recurrence, e.g. ENPP5 (cluster 4; Supplementary Figure 5B) whose upregulation has been demonstrated in triple-negative breast cancer tumors [84]. In addition, ENPP5 also showed moderate to high protein expression levels in breast cancers but not in normal breast tissue, using the HPA module (Figure 8B) and a significantly positive correlation between its mRNA and protein expression levels (Figure 8C).

DISCUSSION

The statistical analysis of RNA-seq data is central to better understand how genes and signaling pathways are modulated by different perturbations [85–87]. Although there are many tools for online RNA-seq analysis, CAP-RNAseq provides an all-in-one novel pipeline starting from filtering/clustering raw count data to prioritization of co-expression clusters/genes based on gene essentiality and congruency of mRNA-protein expression levels and some of its unique features include the annotation of co-expression clusters as 'mirror clusters' and generation of bipartite cluster-term networks.

CAP-RNAseq prioritizes co-expression clusters using DepMap [29, 30] and MSigDB [51, 52], and genes by ranking them based on their correlation between mRNA-protein levels, and/or gene essentiality scores. Although previous studies [88, 89] and RNA-seq tools, e.g. iDEP, have utilized MSigDB analysis for enriching differentially expressed gene sets, CAP-RNAseq also uses DepMap for enrichment and prioritization of co-expression clusters.

DepMap has already been integrated into apps such as NetControl4BioMed [90] and shinyDepMap [31], which enable researchers

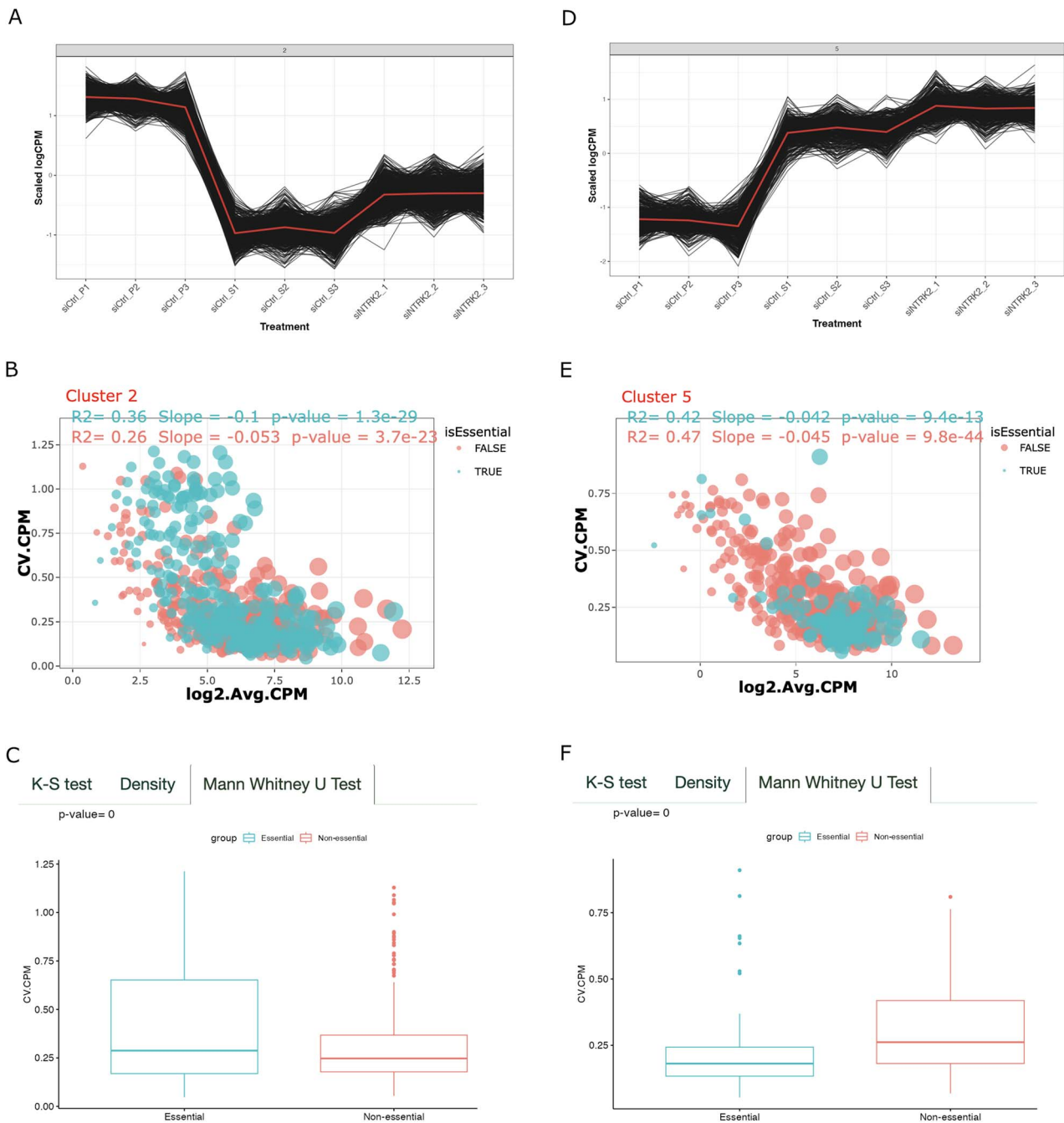


Figure 5. Analysis of cluster 2 (A–C) and cluster 5 (D–F). (A, B) Cluster visualization of each cluster as a result of k-means clustering. (B, E) Graphs showing the \log_2 of Avg CPM in x-axis versus CV of CPM in y-axis of genes in clusters. Point size shows the \log_2 of range CPM values. (C, F) Graphs showing the median CV for both essential and non-essential genes based on Mann–Whitney U-test. Avg: average, CV: coefficient of variation, CPM: counts per million.

to identify essential genes and ultimately discover druggable targets as well as used in other studies focusing on drug resistance [91] and identification of therapeutic targets [92]. However, CAP-RNAseq is the only RNA-seq data clustering, annotation and prioritization tool that incorporates DepMap and hence allows for the identification of expression patterns associated with the overrepresentation of essential genes in a given treatment/cluster. It also enables comparisons between essential and non-essential genes with respect to mean and variance of their expression levels in an experiment- or condition-specific manner.

Indeed, the essential genes might have higher expression levels and lower variance. For instance, in *Caulobacter crescentus* [93]

and in multiple *Escherichia coli* species [94], the gene essentiality and expression levels were positively correlated. DepMap data were also used to construct predictive models of gene essentiality based on modifier gene expression variation [95]. The presence of a transcriptional program conserved across different organisms, including *E. coli*, yeast and humans, showed that essential genes but not non-essential genes needed to be expressed above a certain threshold mRNA number to maintain cellular functions [96]. Indeed, many essential genes were transcriptionally and mutationally robust via use of multiple transcription start sites within a promoter [97], and hence low transcriptional noise and low variability of promoters were associated with gene

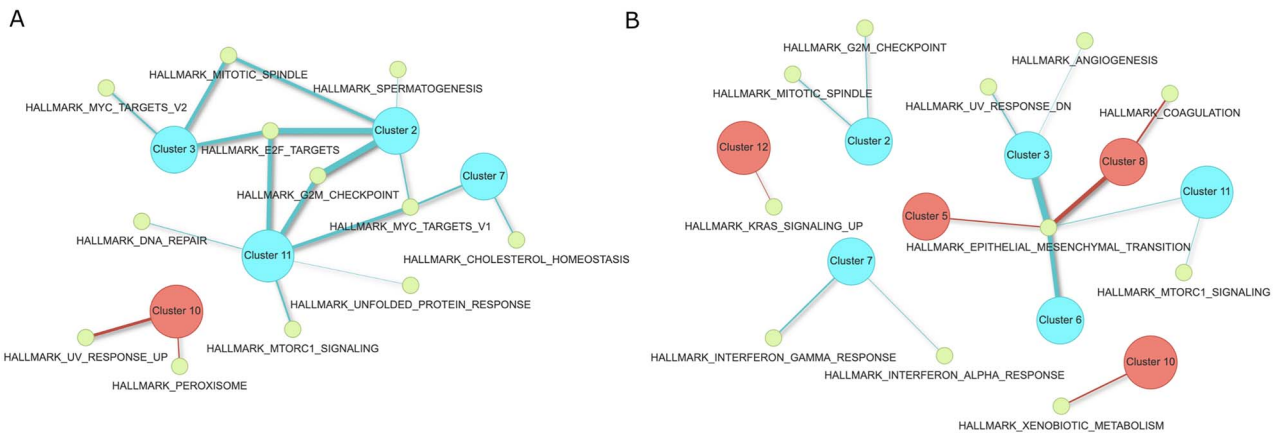


Figure 6. MSigDB networks for ‘Hallmark’ terms. Functional network for only essential genes (A) and non-essential genes (B) in clusters including terms that have q -values < 0.05 . Edge widths show the number of genes overlapping with each hallmark collection. Cluster nodes indicate the increase or decrease in the average expression of genes with respect to the proliferating control siRNA and siNTRK2 groups for each cluster.

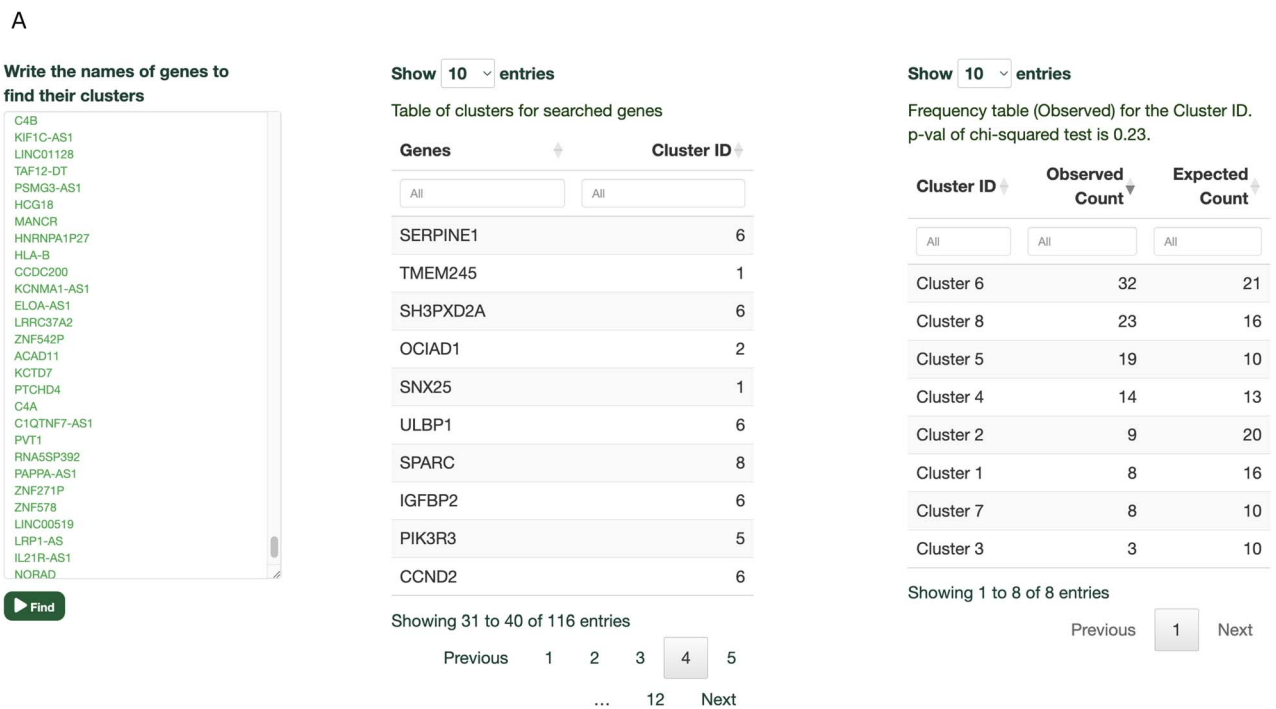


Figure 7. Comparative analysis of co-expression patterns across datasets. (A) The genes comprising cluster 5 in the Demo 2 dataset were queried using the ‘Search Gene/Geneset’ tab in the Demo 1 analysis. (B) Expression pattern of cluster 6 in Demo 1 dataset and (C) expression pattern of cluster 8 in Demo 1 dataset, depicting the co-expression patterns of the identified genes from cluster 5 in Demo 2 dataset.

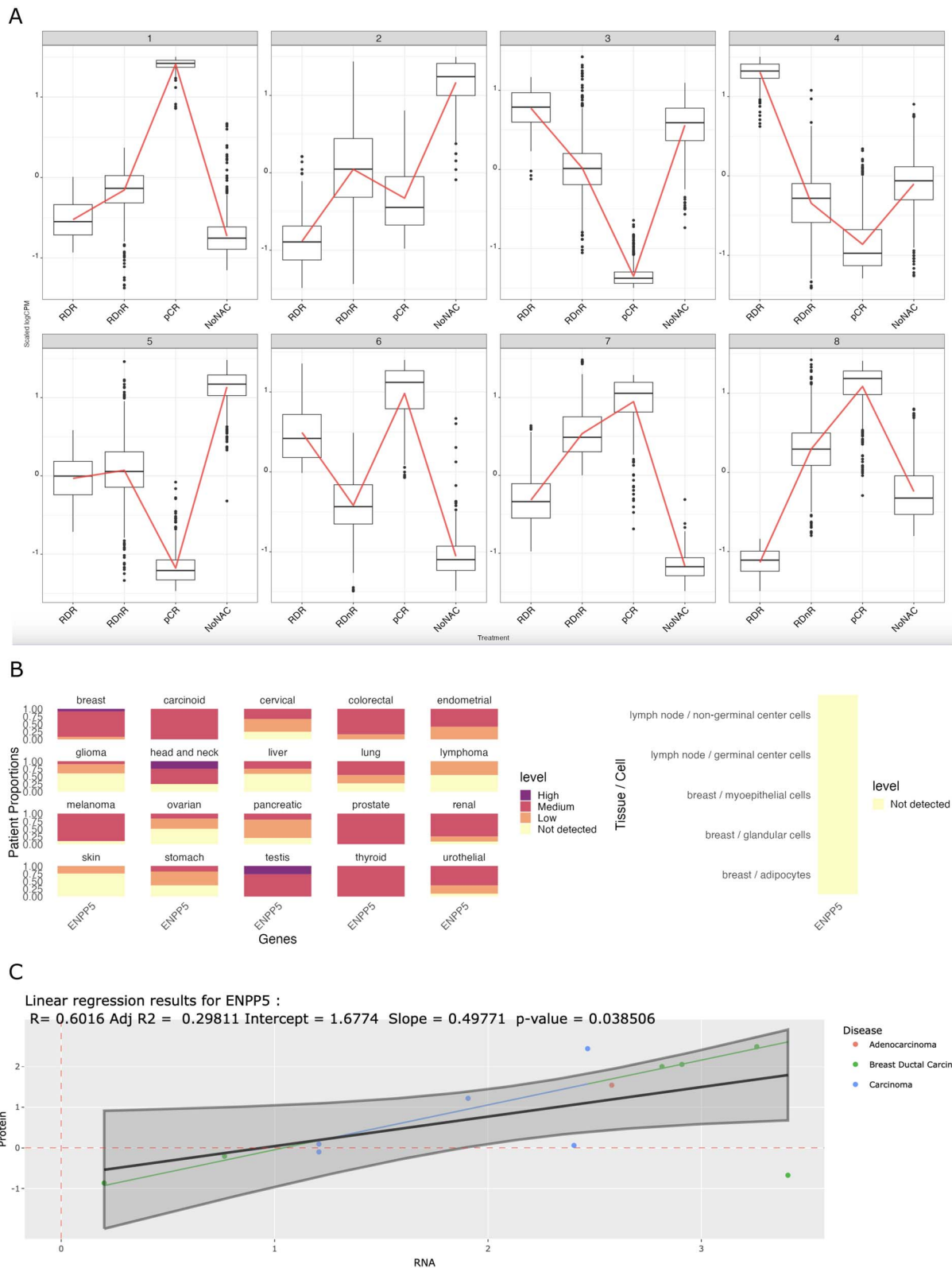


Figure 8. Analysis of GSE201085 data. **(A)** K-means clustering ($k=8$) with group means visualized as boxplot. **(B)** Protein expression of ENPP5 in many cancers and normal cells (breast and lymph nodes) using the HPA dataset. **(C)** mRNA versus protein correlation plot across diseases in breast using DepMap data.

essentiality. A statistical method called CEDA [98] also showed that the percentage of essential genes increased when gene expression data were integrated with the CRISPR screen data.

In the present study, we demonstrated that overexpression/depletion models of NTRK2 led to modulation of subsets of essential genes, whose ranges of expression as well as degrees of modulation exhibited variability. Indeed, very highly expressed genes,

regardless of being essential or not, exhibited lower coefficients of variation across different treatments. The essential genes were enriched in co-expression clusters in which NTRK2 overexpression led to increased expression supporting NTRK2's role as an oncogene [99], while senescent cells with or without the treatment with siNTRK2 decreased highly the mRNA expression of a subset of essential genes [40]. Accordingly, we could demonstrate via CAP-RNAseq that a subset of essential genes could be highly significantly modulated in fibroblast senescence and such tests can be extended to RNA-seq data from other senescent cells including those of the liver.

Moreover, novel associations discovered by CAP-RNAseq, as in the case of NTRK2 and CCND2, can be further studied in silico. Indeed, we also found a negative correlation between the expression level of CCND2 and that of NTRK2 in the lower grade glioma patient dataset of TCGA (TCGA-LGG) ($r = -0.17$, $P = 8.005e-5$), although not in higher grade glioblastoma patients (TCGA-GBM) ($r = 0.02$, $P = 0.777$) based on cbioportal.org [100–102]. By using the 'Search Gene/Geneset' module, we found other cluster 5 genes expressed like CCND2 (Supplementary Table 5B), which can be further tested in other TCGA or clinical cohorts.

In addition, CAP-RNAseq offers the ability to prioritize genes for future validation [103, 104]. While mRNA expression levels are highly informative in cancer diagnosis and classification [105, 106], true causative agents in the cell are most likely the levels and interactions of proteins [107]. Since the correlation between mRNA and protein expression levels is not always high due to post-transcriptional regulation [108] and/or the basal levels of protein abundance [109], the discovery of genes, e.g. ENPP5, which exhibit high mRNA-protein level correlation, could provide more reliable mRNA biomarkers in clinic to test by RT-qPCR primers that could be generated by CAP-RNAseq.

CONCLUSIONS

CAP-RNAseq serves as an invaluable tool for the annotation and prioritization of co-expression clusters using custom or publicly available bulk RNA-seq data, as it facilitates a deeper understanding of the patterns and relationships within and between datasets. It also allows for biomarker discovery and gene prioritization and can be applied across a wide range of research contexts that includes cancer. The tool offers unique features and capabilities that distinguish it from other similar applications and is available at <http://konulabapps.bilkent.edu.tr:3838/CAPRNAseq/> and the docker image is downloadable from <https://hub.docker.com/r/konulab/capmaseq>.

Key Points

- CAP-RNAseq is the first all-in-one R Shiny application on which users can filter gene-level RNA-seq raw count data by *vst* + ANOVA and prioritize co-expression clusters/genes based on gene essentiality and the degree of congruence between the mRNA-protein levels.
- CAP-RNAseq utilizes the *k*-means algorithm to cluster expression profiles, and effectively identifies mirror clusters expressed inversely of each other, and generates cluster-term networks.
- CAP-RNAseq uniquely integrates key databases and applications such as DepMap, Human Protein Atlas,

MSigDB and Primer3 using relevant R packages and employs them for RNA-seq expression data analysis.

ACKNOWLEDGEMENTS

We thank Zeynep Yucekaya who designed CAP-RNAseq logo and helped with the app's pipeline figure. We also thank the anonymous reviewers for their helpful suggestions that led to improvements in the application and manuscript.

FUNDING

This project was funded in part by The Health Institutes of Türkiye (TUSEB) [4405]. Moreover, this project has received funding from the European Horizon's research and innovation program HORIZON-HLTH-2022-STAYHLTH-02 under agreement No 101095679. Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

DATA AVAILABILITY

CAP-RNAseq is a web server available at <http://konulabapps.bilkent.edu.tr:3838/CAPRNAseq/> and the docker image is downloadable from <https://hub.docker.com/r/konulab/capmaseq>. Demo 1 dataset was downloaded from <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE136868>. Demo 2 dataset was downloaded from <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE202951>. Demo 3 dataset was downloaded from <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE201085>. Within CAP-RNAseq, processed data for demo datasets can be downloaded. Dataset for target priority scores was downloaded from <https://score.depmap.sanger.ac.uk/downloads>. Datasets for efficacy/selectivity values and dependency scores were downloaded from <https://doi.org/10.6084/m9.figshare.13653260.v1> and <https://doi.org/10.6084/m9.figshare.13653257.v1>, respectively.

REFERENCES

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63. <https://doi.org/10.1038/nrg2484>.
2. Lee JK, Williams PD, Cheon S. Data mining in genomics. *Clin Lab Med* 2008;**28**(1):145–viii. <https://doi.org/10.1016/j.cll.2007.10.010>.
3. Wu Y, Wang X, Wu F, et al. Transcriptome profiling of the cancer, adjacent non-tumor and distant normal tissues from a colorectal cancer patient by deep sequencing. *PLoS One* 2012;**7**(8):e41001. <https://doi.org/10.1371/journal.pone.0041001>.
4. Lancour D, Dupuis J, Mayeux R, et al. Analysis of brain region-specific co-expression networks reveals clustering of established and novel genes associated with Alzheimer disease. *Alzheimers Res Ther* 2020;**12**:103. <https://doi.org/10.1186/s13195-020-00674-7>.
5. Yang L, Jin Y, Huang W, et al. Full-length transcriptome sequences of ephemeral plant *Arabidopsis pumila* provides insight into gene expression dynamics during continuous salt

- stress. *BMC Genomics* 2018;**19**(1):717. <https://doi.org/10.1186/s12864-018-5106-y>.
6. Mikolajewicz N, Gacesa R, Aguilera-Urbe M, et al. Multi-level cellular and functional annotation of single-cell transcriptomes using scPipeline. *Commun Biol* 2022;**5**:1142. <https://doi.org/10.1038/s42003-022-04093-2>.
 7. Iida K, Kondo J, Wibisana JN, et al. ASURAT: functional annotation-driven unsupervised clustering of single-cell transcriptomes. *Bioinformatics* 2022;**38**:4330–6. <https://doi.org/10.1093/bioinformatics/btac541>.
 8. Jiang W, Zhang L, Guo Q, et al. Identification of the pathogenic biomarkers for hepatocellular carcinoma based on RNA-seq analyses. *Pathol Oncol Res* 2019;**25**:1207–13. <https://doi.org/10.1007/s12253-019-00596-2>.
 9. Karimizadeh E, Sharifi-Zarchi A, Nikaein H, et al. Analysis of gene expression profiles and protein-protein interaction networks in multiple tissues of systemic sclerosis. *BMC Med Genomics* 2019;**12**:199. <https://doi.org/10.1186/s12920-019-0632-2>.
 10. Pane K, Affinito O, Zanfardino M, et al. An integrative computational approach based on expression similarity signatures to identify protein–protein interaction networks in female-specific cancers. *Front Genet* 2020;**11**:612521. <https://doi.org/10.3389/fgene.2020.612521>.
 11. Reyes ALP, Silva TC, Coetzee SG, et al. GENAVi: a shiny web application for gene expression normalization, analysis and visualization. *BMC Genomics* 2019;**20**(1):745. <https://doi.org/10.1186/s12864-019-6073-7>.
 12. Ge SX, Son EW, Yao R. iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC Bioinformatics* 2018;**19**(1):534. <https://doi.org/10.1186/s12859-018-2486-6>.
 13. Kucukural A, Yukselen O, Ozata DM, et al. DEBrowser: interactive differential expression analysis and visualization tool for count data. *BMC Genomics* 2019;**20**:6. <https://doi.org/10.1186/s12864-018-5362-x>.
 14. Haering M, Habermann BH. RNfuzzyApp: an R shiny RNA-seq data analysis app for visualisation, differential expression analysis, time-series clustering and enrichment analysis. *F1000Research* 2021;**10**:654. <https://doi.org/10.12688/f1000research.54533.1>.
 15. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* 2009;**37**:W305–11. <https://doi.org/10.1093/nar/gkp427>.
 16. Perampalam P, Dick FA. BEAVR: a browser-based tool for the exploration and visualization of RNA-seq data. *BMC Bioinformatics* 2020;**21**(1):221. <https://doi.org/10.1186/s12859-020-03549-8>.
 17. Wang YE, Kutnetsov L, Partensky A, et al. WebMeV: a cloud platform for analyzing and visualizing cancer genomic data. *Cancer Res* 2017;**77**:e11–4. <https://doi.org/10.1158/0008-5472.CAN-17-0802>.
 18. Akhmedov M, Martinelli A, Geiger R, Kwee I. Omics playground: a comprehensive self-service platform for visualization, analytics and exploration of big omics data. *NAR Genomics Bioinforma* 2020;**2**(1):lqz019. <https://doi.org/10.1093/nargab/lqz019>.
 19. Guo W, Tzioutziou NA, Stephen G, et al. 3D RNA-seq: a powerful and flexible tool for rapid and accurate differential expression and alternative splicing analysis of RNA-seq data for biologists. *RNA Biol* 2021;**18**(11):1574–87. <https://doi.org/10.1080/15476286.2020.1858253>.
 20. Parsania C, Chen R, Sethiya P, et al. FungiExpresZ: an intuitive package for fungal gene expression data analysis, visualization and discovery. *Brief Bioinform* 2023;PMID: 36806894;**24**(2):bbad051. <https://doi.org/10.1093/BIB/BBAD051>.
 21. Abu-Jamous B, Kelly S. Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data. *Genome Biol* 2018;**19**(1):172. PMID: 30359297. <https://doi.org/10.1186/s13059-018-1536-8>.
 22. Powell D. Degust: interactive RNA-seq analysis, Zenodo; 2015. doi: 10.5281/zenodo.3258933. GitHub Repository Available at <https://github.com/drpowell/degust/tree/v3.2.0>.
 23. Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A* 1999;**96**:2907–12. <https://doi.org/10.1073/pnas.96.6.2907>.
 24. Yaslianifard S, Movahedi M, Yaslianifard S, Mozhgani S-H. The mirror like expression of genes involved in the FOXO signaling pathway could be effective in the pathogenesis of human lymphotropic virus type 1 (HTLV-1) through disruption of the downstream pathways. *BMC Res Notes* 2023;**16**(1):1–7. PMID: 37461070. <https://doi.org/10.1186/s13104-023-06423-x>.
 25. Baldessari D, Shin Y, Krebs O, et al. Global gene expression profiling and cluster analysis in *Xenopus laevis*. *Mech Dev* 2005;**122**:441–75. PMID: 15763214. <https://doi.org/10.1016/j.mod.2004.11.007>.
 26. Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* 2017;**8**:1826. <https://doi.org/10.1038/s41467-017-01261-5>.
 27. Jain A, Tuteja G. TissueEnrich: tissue-specific gene enrichment analysis. *Bioinformatics* 2019;**35**:1966–7. <https://doi.org/10.1093/bioinformatics/bty890>.
 28. Uhlén M, Fagerberg L, Hallström BM, et al. Tissue-based map of the human proteome. *Science* 2015;**347**(6220):1260419. <https://doi.org/10.1126/science.1260419>.
 29. Tsherniak A, Vazquez F, Montgomery PG, et al. Defining a cancer dependency map. *Cell* 2017;**170**:564–576.e16. <https://doi.org/10.1016/j.cell.2017.06.010>.
 30. Behan FM, Iorio F, Picco G, et al. Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature* 2019;**568**:511–6. <https://doi.org/10.1038/s41586-019-1103-9>.
 31. Shimada K, Bachman JA, Muhlich JL, Mitchison TJ. Shinydepmap, a tool to identify targetable cancer genes and their functional connections from cancer dependency map data. *Elife* 2021;**10**:10. <https://doi.org/10.7554/eLife.57116>.
 32. Untergasser A, Cutcutache I, Koressaar T, et al. Primer3-new capabilities and interfaces. *Nucleic Acids Res* 2012;**40**:e115. <https://doi.org/10.1093/nar/gks596>.
 33. Yao E, Blake VC, Cooper L, et al. GrainGenes: a data-rich repository for small grains genetics and genomics. *Database (Oxford)* 2022;**2022**:baac034. PMID: 35616118. <https://doi.org/10.1093/DATABASE/BAAC034>.
 34. Döring M, Pfeifer N. openPrimeRui: Shiny Application for Multiplex PCR Primer Design and Analysis. 2023. <https://doi.org/10.18129/B9.bioc.openPrimeRui>, R package version 1.24.0, <https://bioconductor.org/packages/openPrimeRui>.
 35. Team RC. R: A language and environment for statistical computing. *R Found Stat Comput* 2021.
 36. Chang W, Cheng J, Allaire J, et al. shiny: Web Application Framework for R. R package version 1.7. 2.9000., Retrieved Febr. 2022;**23**. <https://shiny.rstudio.com/>

37. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550. <https://doi.org/10.1186/s13059-014-0550-8>.
38. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2017;**45**:D12–7. <https://doi.org/10.1093/nar/gkv1290>.
39. Pattwell SS, Arora S, Cimino PJ, et al. A kinase-deficient NTRK2 splice variant predominates in glioma and amplifies several oncogenic signaling pathways. *Nat Commun* 2020;**11**:2977. <https://doi.org/10.1126/sciadv.abo6789>.
40. Anerillas C, Herman AB, Munk R, et al. Author correction: a BDNF-TrkB autocrine loop enhances senescent cell viability. *Nat Commun* 2022;**13**:7540. <https://doi.org/10.1038/s41467-022-35154-z>.
41. Axelrod ML, Wang Y, Xu Y, et al. Peripheral blood monocyte abundance predicts outcomes in patients with breast cancer. *Cancer Res Commun* 2022;**25**:286–92 PMID: 36304942. <https://doi.org/10.1158/2767-9764.CRC-22-0023>.
42. Gatto L, Killian T. *depmap: Cancer Dependency Map Data Package*, 2023. R package version 1.16.0, <https://bioconductor.org/packages/depmap>. (01 Sep 2023, date last accessed).
43. DepMap, Broad. DepMap 22Q1 public. Figshare. *Dataset* 2022. <https://doi.org/10.6084/m9.figshare.19139906.v1>.
44. Nusinow DP, Szpyt J, Ghandi M, et al. Quantitative proteomics of the Cancer Cell Line Encyclopedia. *Cell* 2020;**180**:387–402.e16. <https://doi.org/10.1016/j.cell.2019.12.023>.
45. Dwane L, Behan FM, Gonçalves E, et al. Project Score database: a resource for investigating cancer cell dependencies and prioritizing therapeutic targets. *Nucleic Acids Res* 2021;**49**:D1365–72. <https://doi.org/10.1093/nar/gkaa882>.
46. Tran AN, Dussaq AM, Kennell T, et al. HPAanalyze: an R package that facilitates the retrieval and analysis of the human protein atlas data. *BMC Bioinformatics* 2019;**20**(1):463. <https://doi.org/10.1186/s12859-019-3059-z>.
47. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987;**20**:53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
48. Hartigan JA, Wong MA. Algorithm AS 136: a K-means clustering algorithm. *Appl Stat* 1979;**28**:100. <https://doi.org/10.2307/2346830>.
49. Maechler M, Rousseeuw P, Struyf A, et al. Cluster: cluster analysis basics and extensions. R Package Version 2.1.4. 2023. <https://CRAN.R-project.org/package=cluster> (01 Sep 2023, date last accessed).
50. Kassambara A, Mundt F. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R Package Version 1.0.7. <https://CRAN.R-project.org/package=factoextra> (01 Sep 2023, date last accessed).
51. Liberzon A, Subramanian A, Pinchback R, et al. Molecular Signatures Database (MSigDB) 3.0. *Bioinformatics* 2011;**27**:1739–40. <https://doi.org/10.1093/bioinformatics/btr260>.
52. Liberzon A, Birger C, Thorvaldsdóttir H, et al. The molecular signatures database Hallmark gene set collection. *Cell Syst* 2015;**1**:417–25. <https://doi.org/10.1016/j.cels.2015.12.004>.
53. Dolgalev I. *msigdb: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format*. R package version 7.5.1, <https://CRAN.R-project.org/package=msigdb>, (01 Sep 2023, date last accessed).
54. Almende BV, Thieurmel B. *visNetwork: Network Visualization using “vis.js” Library*. <https://CRAN.R-project.org/package=visNetwork> (01 Sep 2023, date last accessed).
55. Aragon T. *epitools: Epidemiology Tools*. R package version 0.5-10.1. <https://CRAN.R-project.org/package=epitools>, (01 Sep 2023, date last accessed).
56. Galili T, O’Callaghan A, Sidi J, Sievert C. Heatmaply: an R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics* 2018;**34**(9):1600–1602. <https://doi.org/10.1093/bioinformatics/btx657>.
57. Massey FJ. The Kolmogorov-Smirnov test for goodness of fit. *J Am Stat Assoc* 1951;**46**:68–78. <https://doi.org/10.2307/2280095>.
58. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* 1947;**18**:50–60. <https://www.jstor.org/stable/2236101>.
59. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics* 1945;**1**:80. <https://doi.org/10.2307/3001968>.
60. Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47. <https://doi.org/10.1093/nar/gkv007>.
61. Rizzo M, Szekely G. *energy: E-Statistics: Multivariate Inference via the Energy of Data*. R package version 1.7-11. <https://CRAN.R-project.org/package=energy> (01 Sep 2023, date last accessed).
62. Meyers RM, Bryan JG, McFarland JM, et al. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat Genet* 2017;**49**:1779–84. <https://doi.org/10.1038/ng.3984>.
63. Ghandi M, Huang FW, Jané-Valbuena J, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* 2019;**569**:503–8. <https://doi.org/10.1038/s41586-019-1186-3>.
64. Hubert M, Rousseeuw PJ. Robust regression with both continuous and binary regressors. *J Stat Plan Inference* 1997;**57**:153–63. [https://doi.org/10.1016/S0378-3758\(96\)00041-9](https://doi.org/10.1016/S0378-3758(96)00041-9).
65. Nie L, Wu G, Zhang W. Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in *Desulfovibrio vulgaris*: a quantitative analysis. *Genetics* 2006;**174**:2229–43. <https://doi.org/10.1534/genetics.106.065862>.
66. Wang D, Zhang Y, Che YQ. CCND2 mRNA expression is correlated with R-CHOP treatment efficacy and prognosis in patients with ABC-DLBCL. *Front Oncol* 2020;**10**:1180. <https://doi.org/10.3389/fonc.2020.01180>.
67. Shan YS, Hsu HP, Lai MD, et al. Cyclin D1 overexpression correlates with poor tumor differentiation and prognosis in gastric cancer. *Oncol Lett* 2017;**14**:4517–26. <https://doi.org/10.3892/ol.2017.6736>.
68. Park SY, Lee CJ, Choi JH, et al. The JAK2/STAT3/CCND2 axis promotes colorectal cancer stem cell persistence and radioresistance. *J Exp Clin Cancer Res* 2019;**38**:399. <https://doi.org/10.1186/s13046-019-1405-7>.
69. Zhu C, Shao P, Bao M, et al. MiR-154 inhibits prostate cancer cell proliferation by targeting CCND2. *Urol Oncol Semin Orig Investig* 2014;**32**:31.e9–16. <https://doi.org/10.1016/j.urolonc.2012.11.013>.
70. Yu G, Zhou H, Yao W, et al. lncRNA TUG1 promotes cisplatin resistance by regulating CCND2 via epigenetically silencing miR-194-5p in bladder cancer. *Nucleic Acids* 2019;**16**:257–71. <https://doi.org/10.1016/j.omtn.2019.02.017>.
71. Zhang X, Zhao M, Huang AY, et al. The effect of cyclin D expression on cell proliferation in human gliomas. *J Clin Neurosci* 2005;**12**:166–8. <https://doi.org/10.1016/j.jocn.2004.03.036>.
72. Koyama-Nasu R, Nasu-Nishimura Y, Todo T, et al. The critical role of cyclin D2 in cell cycle progression and tumorigenicity of glioblastoma stem cells. *Oncogene* 2013;**32**:3840–5. <https://doi.org/10.1038/onc.2012.399>.
73. Kheirollahi M, Mehr-Azin M, Kamalian N, Mehdipour P. Expression of cyclin D2, P53, Rb and ATM cell cycle genes in

- brain tumors. *Med Oncol* 2011;**28**:7–14. <https://doi.org/10.1007/s12032-009-9412-8>.
74. Durinck S, Moreau Y, Kasprzyk A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 2005;**21**:3439–40. <https://doi.org/10.1093/bioinformatics/bti525>.
 75. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. *Nat Protoc* 2009;**4**:1184–91. <https://doi.org/10.1038/nprot.2009.97>.
 76. You FM, Huo N, Gu YQ, et al. ConservedPrimers 2.0: a high-throughput pipeline for comparative genome referenced intron-flanking PCR primer design and its application in wheat SNP discovery. *BMC Bioinformatics* 2009;**10**:331. <https://doi.org/10.1186/1471-2105-10-331>.
 77. Cribbs A. *scAmpDesign: What the Package Does (Title Case)*. R package version 0.1.0. 2022. <https://github.com/Acribbs/scAmpDesign> (01 Sep 2023, date last accessed).
 78. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. *Bioinformatics* 2007;**23**:1289–91. <https://doi.org/10.1093/bioinformatics/btm091>.
 79. Gibson MG, Jnr RGM. Beyond ANOVA: basics of applied statistics. *Stat* 1986;**35**:566.
 80. Durbin BP, Hardin JS, Hawkins DM, Rocke DM. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* 2002;**18**:S105–10. https://doi.org/10.1093/bioinformatics/18.suppl_1.s105.
 81. Nueda MJ, Tarazona S, Conesa A. Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics* 2014;**30**:18:2598PMID: 24894503–602.<https://doi.org/10.1093/bioinformatics/btu333>.
 82. Celebi ME, Kingravi HA. Deterministic initialization of the K-means algorithm using hierarchical clustering. *Int J Pattern Recognit Artif Intell* 2012;**26**:1250018. <https://doi.org/10.1142/S0218001412500188>.
 83. Helmy M, Agrawal R, Ali J, et al. GeneCloudOmics: a data analytic cloud platform for high-throughput gene expression analysis. *Front Bioinform* 2021;**1**:693836. <https://doi.org/10.3389/fbinf.2021.693836>.
 84. Wu W, Warner M, Wang L, et al. Drivers and suppressors of triple-negative breast cancer. *Proc Natl Acad Sci U S A* 2021;**118**:33:e2104162118. <https://doi.org/10.1073/pnas.2104162118>.
 85. Hoshida Y, Brunet JP, Tamayo P, et al. Subclass mapping: identifying common subtypes in independent disease data sets. *PloS One* 2007;**2**:e1195. <https://doi.org/10.1371/journal.pone.0001195>.
 86. Johnson NT, Dhroso A, Hughes KJ, Korkin D. Biological classification with RNA-seq data: can alternatively spliced transcript expression enhance machine learning classifiers? *RNA* 2018;**24**:1119–32. <https://doi.org/10.1261/rna.062802.117>.
 87. Risso D, Pagnotta SM. Per-sample standardization and asymmetric winsorization lead to accurate clustering of RNA-seq expression profiles. *Bioinformatics* 2021;**37**:2356–64.<https://doi.org/10.1093/bioinformatics/btab091>.
 88. Chen J, Fu Y, Hu J, He J. Hypoxia-related gene signature for predicting LUAD patients' prognosis and immune microenvironment. *Cytokine* 2022;**152**:155820. <https://doi.org/10.1016/j.cyto.2022.155820>.
 89. Li S, Liu W, Chen Y, et al. Transcriptome analysis of cepharanthine against a SARS-CoV-2-related coronavirus. *Brief Bioinform* 2021;**22**:1378–86. <https://doi.org/10.1093/bib/bbaa387>.
 90. Popescu VB, Sánchez-Martín JÁ, Schacherer D, et al. NetControl4BioMed: a web-based platform for controllability analysis of protein-protein interaction networks. *Bioinformatics* 2021;**37**:3976–8. <https://doi.org/10.1093/bioinformatics/btab570>.
 91. Yang Z, Ye Z, Qiu J, et al. A mutation-induced drug resistance database (MdrDB). *Commun Chem* 2023;**6**:123. <https://doi.org/10.1038/s42004-023-00920-7>.
 92. Wong KK. Integrated transcriptomics and proteomics data analysis identifies CDH17 as a key cell surface target in colorectal cancer. *Comput Biol Chem* 2023;**105**:107897. <https://doi.org/10.1016/j.compbiolchem.2023.107897>.
 93. Fang G, Passalacqua KD, Hocking J, et al. Transcriptomic and phylogenetic analysis of a bacterial cell cycle reveals strong associations between gene co-expression and evolution. *BMC Genomics* 2013;**14**:450. <https://doi.org/10.1186/1471-2164-14-450>.
 94. Rousset F, Cabezas-Caballero J, Piastra-Facon F, et al. Publisher correction: the impact of genetic diversity on gene essentiality within the Escherichia coli species. *Nat Microbiol* 2021;**6**:65. <https://doi.org/10.1038/s41564-021-00893-0>.
 95. Rosenski J, Shifman S, Kaplan T. Predicting gene knockout effects from expression data. *BMC Med Genomics* 2023;**16**:26. <https://doi.org/10.1186/s12920-023-01446-6>.
 96. Lo TW, Choi HKJ, Huang D, Wiggins PA. The one-message-per-cell-cycle rule: a conserved minimum transcription level for essential genes. Preprint. ArXiv 2023; arXiv:2307.03324v1. Published 2023 Jul 6.
 97. Einarsson H, Salvatore M, Vaagensø C, et al. Promoter sequence and architecture determine expression variability and confer robustness to genetic variants. *Elife* 2022;**11**:11. <https://doi.org/10.7554/eLife.80943>.
 98. Zhao Y, Yu L, Wu X, et al. CEDA: integrating gene expression data with CRISPR-pooled screen data identifies essential genes with higher expression. *Bioinformatics* 2022;**38**:5245–52. <https://doi.org/10.1093/bioinformatics/btac668>.
 99. Hu J, Huang Y, Wu Y, et al. NTRK2 is an oncogene and associated with microRNA-22 regulation in human gastric cancer cell lines. *Tumor Biol* 2016;**37**:15115–23. <https://doi.org/10.1007/s13277-016-5337-y>.
 100. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012;**2**:401–4. <https://doi.org/10.1158/2159-8290.CD-12-0095>.
 101. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013;**6**:pl1. <https://doi.org/10.1126/scisignal.2004088>.
 102. de Bruijn I, Kundra R, Mastrogiacomo B, et al. Analysis and visualization of longitudinal genomic and clinical data from the AACR Project GENIE Biopharma Collaborative in cBioPortal [published online ahead of print, 2023 Sep 5]. *Cancer Res* 2023;**83**:3861–7. <https://doi.org/10.1158/0008-5472.CAN-23-0816>.
 103. Chandraratna PK, Cikic S, Baddoo MC, et al. Transcriptome analysis reveals sexual disparities in gene expression in rat brain microvessels. *J Cereb Blood Flow Metab* 2021;**41**:2311–28. <https://doi.org/10.1177/0271678X21999553>.
 104. Kim J, Xu Z, Marignani PA. Single-cell RNA sequencing for the identification of early-stage lung cancer biomarkers from circulating blood. *Npj Genomic Med* 2021;**6**:87. <https://doi.org/10.1038/s41525-021-00248-y>.
 105. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by

- gene expression monitoring. *Science* 1999;**286**:531–7. <https://doi.org/10.1126/science.286.5439.531>.
106. Macgregor PF, Squire JA. Application of microarrays to the analysis of gene expression in cancer. *Clin Chem* 2002;**48**:1170–7.
107. Greenbaum D, Colangelo C, Williams K, Gerstein M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* 2003;**4**(9):117. <https://doi.org/10.1186/gb-2003-4-9-117>.
108. Anderson L, Seilhamer J. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* 1997;**18**:533–7. <https://doi.org/10.1002/elps.1150180333>.
109. Ørntoft TF, Thykjaer T, Waldman FM, et al. Genome-wide study of gene copy numbers, transcripts, and protein levels in pairs of non-invasive and invasive human transitional cell carcinomas. *Mol Cell Proteomics* 2002;**1**:37–45. <https://doi.org/10.1074/mcp.m100019-mcp200>.